Summer School on Delay Differential Equations and Applications

Dobbiaco (BZ), Italy, June 26–30, 2006

The numerical solution of delay differential equations

M. Zennaro
Dipartimento di Matematica e Informatica
Università di Trieste

Main reference for these lecture notes:

A. Bellen and M. Zennaro, Numerical Methods for Delay Differential Equations, Numerical Mathematics and Scientific Computation, Oxford Science Publications, Oxford University Press, 2003

Contents

1 The problems to be solved **6**

2 Existence and uniqueness of solutions **7**

3 Location of discontinuities and smoothing of the solution **12**

3.1 Primary and secondary discontinuities . 14
3.2 Vanishing and non-vanishing delays . . . 22
3.3 Bounded and unbounded delays 24
3.4 State dependent delays 29
3.5 Multiple delays 30
3.6 Propagation of discontinuities in systems 31

4 Some of the main differences **33**

5 Numerical ODE theory is not enough for DDEs **45**

5.1 On the order of the methods 48
5.2 On the stability properties of the methods 51
5.3 A good method for DDEs 59

6 Continuous extensions of RK methods **63**

7 Interpolants of the first class **74**

7.1 Collocation methods 77
7.2 Natural continuous extensions 81

7.3 An application of the NCEs 90

8 Direct construction of continuous RK methods 93

9 The standard approach in the general delay case 99

10 DDEs with constant delay: natural methods and superconvergence 115

10.1 Bellman's method of steps 120

11 DDEs with non-vanishing time dependent delay: constrained mesh and superconvergence 122

12 DDEs with vanishing time dependent delay 128

13 RK methods for NDDEs 130

14 The test equations 140

15 Analysis of the test equations (14.1) and (14.2) 142

16 Analysis of the test equation (14.3) 151

16.1 Description of the asymptotic stability region $\mathcal{S}_\tau$ for real coefficients 155

16.2 Description of the asymptotic stability region $\mathcal{S}_\tau$ for complex coefficients 157

17 Analysis of the test equation (14.4) 161

18 Generalizations of A-stability to DDEs 165

18.1 P-stability 173

18.2 D-stability 183

18.3 FP-stability and FP-contractivity 190

19 Generalizations of A-stability to ND-DEs 204

19.1 NP-stability 207

Summer School on Delay Differential Equations and Applications

Dobbiaco (BZ), Italy, June 26–30, 2006

The numerical solution of delay differential equations

M. Zennaro
Dipartimento di Matematica e Informatica
Università di Trieste

Lecture 1: Existence, uniqueness and regularity of solutions

Main reference for this Lecture:
A. Bellen and M. Zennaro, Numerical Methods for Delay Differential Equations, Numerical Mathematics and Scientific Computation, Oxford Science Publications, Oxford University Press, 2003 (Chapter 2)

1. The problems to be solved

We consider systems of DDEs

$$\begin{cases} y'(t) = f(t, y(t), y(t-\tau)), & t_0 \le t \le t_f, \\ y(t) = \phi(t), & t \le t_0, \end{cases} \tag{1.1}$$

where $f : [t_0, t_f] \times \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}^d$, and systems of NDDEs

$$\begin{cases} y'(t) = f(t, y(t), y(t-\tau), y'(t-\sigma)), & t_0 \le t \le t_f, \\ y(t) = \phi(t), & t \le t_0, \end{cases} \tag{1.2}$$

where $f : [t_0, t_f] \times \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}^d$.

Since for some $t \ge t_0$ it can be that $t - \tau < t_0$, an *initial function* $\phi(t)$ is needed for the wellposedness of the problems rather than a simple initial value y_0, as happens for ODEs.

We report some results on existence and uniqueness of solutions and fix our attention on the nature and location of derivative jump discontinuities, if any, and their propagation along the integration interval under different hypotheses on the *delays* τ and σ.

The delays τ and σ, which always are non-negative, may be just constants (the *constant delay* case), or functions of t, $\tau = \tau(t)$ and $\sigma = \sigma(t)$ (the *variable* or *time dependent delay* case), or even functions of t and y itself, $\tau = \tau(t, y(t))$ and $\sigma = \sigma(t, y(t))$ (the *state dependent delay* case).

Whereas for constant and time dependent delays these problems have been widely investigated, a complete and satisfactory analysis for the most general case of state dependent delay and, in particular, for NDDEs is not available yet.

2. Existence and uniqueness of solutions

As with ordinary differential equations, existence and uniqueness theorems for the IVPs (1.1) and (1.2) are essentially based on the continuity of the functions $f(t,u,v)$ and $f(t,u,v,w)$ with respect to t and Lipschitz continuity with respect to u, v and w.

In particular, for the equation

$$\begin{cases} y'(t) = f\Big(t, y(t), y\big(t-\tau(t,y(t))\big), y'\big(t-\sigma(t,y(t))\big)\Big), \\ y(t) = \phi(t), \quad t \le t_0, \end{cases} \tag{2.1}$$

the local existence and uniqueness analysis is almost trivial whenever

$$\inf_{[t_0,t_f]\times\mathbb{R}^d} \tau(t,x) = \tau_0 > 0$$

and

$$\inf_{[t_0,t_f]\times\mathbb{R}^d} \sigma(t,x) = \sigma_0 > 0.$$

In fact, in the interval $[t_0, t_0+H]$, $H = \min\{\tau_0, \sigma_0\}$, the equation to solve reduces to the ODE

$$\begin{cases} y'(t) = f\Big(t, y(t), \phi\big(t-\tau(t,y(t))\big), \phi'\big(t-\sigma(t,y(t))\big)\Big), \\ y(t_0) = \phi(t_0), \end{cases}$$

for which well-known standard results may be used to prove existence and uniqueness of the solution. In particular, the continuity of $f(t,u,v,w)$ with respect to t and Lipschitz continuity with respect to u, v and w, along with the Lipschitz continuity of ϕ, ϕ', τ and σ, guarantee the local existence and uniqueness of the solution in $[t_0, t_0+\delta]$ for some $\delta > 0$.

For the existence of the solution in a finite interval $[t_0, t_f]$ we may proceed by successive integrations on the intervals $[t_0 + iH, t_0 + (i+1)H]$, $i = 0, 1, \ldots$, where the equations to be solved are still ordinary, the solution being known up to $[t_0 + iH]$. This method is known in the literature as the *method of steps* and is one of the basic methods for the theoretical analysis as well as numerical integration of DDEs and NDDEs.

When the delays τ or σ vanish at some point t^*, the method of steps does not apply in a neighborhood of t^* and results on existence and uniqueness become more difficult to prove.

Some known results for equations (1.1) under more general conditions on the delays follow.

****

Theorem 2.1 (Local existence) *Consider the equation*

$$\begin{cases} y'(t) = f\Big(t, y(t), y\big(t - \tau(t)\big)\Big), & t_0 \le t < t_f, \\ y(t_0) = y_0, \end{cases} \tag{2.2}$$

and assume that the function $f(t, u, v)$ is continuous on $A \subseteq [t_0, t_f) \times \mathbb{R}^d \times \mathbb{R}^d$ and locally Lipschitz continuous with respect to u and v. Moreover, assume that the delay function $\tau(t) \ge 0$ is continuous in $[t_0, t_f)$, $\tau(t_0) = 0$ and, for some $\xi > 0$, $t - \tau(t) > t_0$ in the interval $(t_0, t_0 + \xi]$. Then the problem (2.2) has a unique solution in $[t_0, t_0 + \delta)$ for some $\delta > 0$ and this solution depends continuously on the initial data.

****

It can be shown that, under the same hypotheses, the solution can be continued until a *maximal solution* defined in the interval $[t_0, b)$, with $t_0 < b \leq t_f$. This allows us to prove the following global existence theorem.

****

Theorem 2.2 (Global existence) *If, under the hypotheses of Theorem 2.1, the unique maximal solution of (2.2) is bounded, then it exists on the entire interval* $[t_0, t_f)$.

****

In order to apply the global existence theorem, we need an a priori bound for the solution. This is given in the following corollary.

****

Corollary 2.1 *Besides the hypotheses of Theorem 2.1, assume that the function* $f(t, u, v)$ *satisfies the condition*

$$\|f(t, u, v)\| \leq M(t) + N(t)(\|u\| + \|v\|)$$

in $[t_0, t_f) \times \mathbb{R}^d \times \mathbb{R}^d$, *where* $M(t)$ *and* $N(t)$ *are continuous positive functions on* $[t_0, t_f)$. *Then the solution of (2.2) exists and is unique on the entire interval* $[t_0, t_f)$.

****

The following result extends the existence result of Theorem 2.1 to the more general case of state dependent delays.

****

Theorem 2.3 (Local existence) *Consider the equation*

$$\begin{cases} y'(t) = f\Big(t, y(t), y\big(t - \tau(t, y(t))\big)\Big), & t \geq t_0, \\ y(t) = \phi(t), & t \leq t_0. \end{cases} \tag{2.3}$$

Let $U \subseteq \mathbb{R}^d$ and $V \subseteq \mathbb{R}^d$ be neighborhoods of $\phi(t_0)$ and $\phi(t_0 - \tau(t_0, \phi(t_0)))$, respectively, and assume that the function $f(t, u, v)$ is continuous with respect to t and Lipschitz continuous with respect to u and v in $[t_0, t_0 + h] \times U \times V$ for some $h > 0$. Moreover, assume that the initial function $\phi(t)$ is Lipschitz continuous for $t \leq t_0$ and that the delay function $\tau(t, y) \geq 0$ is continuous with respect to t and Lipschitz continuous with respect to y in $[t_0, t_0 + h] \times U$. Then the problem (2.3) has a unique solution in $[t_0, t_0 + \delta)$ for some $\delta > 0$ and this solution depends continuously on the initial data.

****

As for NDDEs of the type (2.1), things are slightly different and additional conditions must be imposed on the function $f(t,u,v,w)$. For example, observe that, if $\tau(t_0) = \sigma(t_0) = 0$, then equation (2.1) yields $y'(t_0) = f\big(t_0, y(t_0), y(t_0), y'(t_0)\big)$. Therefore, if the initial value $y(t_0)$ is such that the equation

$$z = f\big(t_0, y(t_0), y(t_0), z\big) \tag{2.4}$$

has no solutions for z, then the IVP (2.1) has no solutions through the point $(t_0, y(t_0))$. The following theorem holds for the case of vanishing delay.

****

Theorem 2.4 (Local existence) *Consider the NDDE*

$$\begin{cases} y'(t) = f\Big(t, y(t), y\big(t-\tau(t)\big), y'\big(t-\tau(t)\big)\Big), & t \geq t_0, \\ y(t_0) = y_0, \\ y'(t_0) = y'_0, \end{cases} \tag{2.5}$$

where $\tau(t_0) = 0$ and, for some $\xi > 0$, $t - \tau(t) > t_0$ in the interval $(t_0, t_0 + \xi]$. Assume $y'(t_0)$ to be a real solution of (2.4) and let $\mu_\xi = \inf_{[t_0, t_0+\xi]} \tau'(t) > 0$. Moreover, assume the function $f(t,u,v,w)$ to be continuous with respect to t and Lipschitz continuous with respect to u, v and w in a neighborhood of the point $(t_0, y(t_0), y(t_0), y'(t_0))$. If L, the Lipschitz constant with respect to w, satisfies

$$L(1 - \mu_\xi) < 1,$$

then a unique continuously differentiable solution of (2.5) exists in $[t_0, t_0 + \delta)$ for some $\delta > 0$.

****

3. Location of discontinuities and smoothing of the solution

The presence of deviated arguments in y, or y', can cause the appearance of jump discontinuities in y' or in higher derivatives of y at subsequent points.

It is known that every step by step numerical method for the initial value problems achieves its own accuracy order provided that the solution is sufficiently smooth at each step interval $[t_n, t_{n+1}]$. More precisely, for a method to be of order p, we usually ask the solution to be at least C^{p+1}-continuous on $[t_n, t_{n+1}]$.

Therefore, it is important to analyze how the discontinuity points propagate through the integration interval $[t_0, t_f]$ and how smoothness possibly increases at any discontinuity point with respect to its *ancestor*, the discontinuity point it originates from.

The number and location of discontinuity points essentially depends on the behavior of the so-called *deviated arguments*

$$\alpha(t) = t - \tau\big(t, y(t)\big)$$

and

$$\beta(t) = t - \sigma\big(t, y(t)\big),$$

viewed as functions of t either for constant or variable or even state dependent delays τ and σ.

We shall assume $\alpha(t) \le t$ and $\beta(t) \le t$ because the delays are always non-negative.

In particular, if $\alpha(t) \ge t_0$ and $\beta(t) \ge t_0$ for all $t \ge t_0$, then no values of y are needed in (1.1) and (1.2) behind t_0 and, therefore, no discontinuities propagate from t_0. Thus the solution is regular according to the regularity of f, α and β. This is the case, for example, in the following equation,

sometimes called *generalized pantograph equation* or *equation with proportional delay*:

$$\begin{cases} y'(t) = f\big(t, y(t), y(qt), y'(pt)\big), & t \geq 0, \\ y(0) = y_0, \end{cases} \tag{3.1}$$

where $0 < q < 1$ and $0 < p < 1$.

3.1. Primary and secondary discontinuities

Consider the scalar instance of equation (1.1) (i.e. $m = 1$) and assume that for the deviated argument $\alpha(t) = t - \tau$ it is $\alpha(t) < t_0$ for some points $t \in [t_0, t_f]$. Moreover, assume that the solution $y(t)$ does not link smoothly to the initial function $\phi(t)$ at t_0, i.e.

$$\phi'(t_0)^- \neq y'(t_0)^+ = f\big(t_0, \phi(t_0), \phi(\alpha(t_0))\big).$$

If the functions f, ϕ and α are continuous, then it is obvious that $y'(t)$ is also continuous for any $t > t_0$. On the other hand, if f, ϕ and α are differentiable, then $y''(t)$ exists for any t except for the points $\xi_{1,i} (> t_0)$ such that

$$\alpha(\xi_{1,i}) = t_0$$

and

$$\alpha'(\xi_{1,i}) \neq 0,$$

i.e. for the simple roots, if any, of the equation

$$\alpha(t) = t_0.$$

In fact, for any smooth function $f(t, y, x)$ we can formally write

$$\begin{aligned} y''(t)^{\pm} &= \frac{\partial f}{\partial t}\Big(t, y(t), y\big(\alpha(t)\big)\Big) + \frac{\partial f}{\partial y}\Big(t, y(t), y\big(\alpha(t)\big)\Big) y'(t) \\ &\quad + \frac{\partial f}{\partial x}\Big(t, y(t), y\big(\alpha(t)\big)\Big) y'\big(\alpha(t)\big)^{\pm} \alpha'(t), \end{aligned} \tag{3.2}$$

and hence

$$\begin{aligned} y''(\xi_{1,i})^{+} &= \frac{\partial f}{\partial t}\big(\xi_{1,i}, y(\xi_{1,i}), y(t_0)\big) + \frac{\partial f}{\partial y}\big(\xi_{1,i}, y(\xi_{1,i}), y(t_0)\big) y'(\xi_{1,i}) \\ &\quad + \frac{\partial f}{\partial x}\big(\xi_{1,i}, y(\xi_{1,i}), y(t_0)\big) y'(t_0)^{+} \alpha'(\xi_{1,i}) \end{aligned} \tag{3.3}$$

and

$$y''(\xi_{1,i})^- = \frac{\partial f}{\partial t}\big(\xi_{1,i}, y(\xi_{1,i}), y(t_0)\big) + \frac{\partial f}{\partial y}\big(\xi_{1,i}, y(\xi_{1,i}), y(t_0)\big)y'(\xi_{1,i}) + \frac{\partial f}{\partial x}\big(\xi_{1,i}, y(\xi_{1,i}), y(t_0)\big)\phi'(t_0)^-\alpha'(\xi_{1,i}). \tag{3.4}$$

Since $\alpha'(\xi_{1,i}) \neq 0$ and $\phi'(t_0)^-$ is assumed to be different from $y'(t_0)^+$, y'' does not exist at $\xi_{1,i}$ and its prolongation by $y''(\xi_{1,i}) = y''(\xi_{1,i})^+$ has a jump discontinuity.

These jump discontinuities in y'' are called 1-*level primary discontinuities*. By differentiating (3.2), one easily checks that each 1-level primary discontinuity point $\xi_{1,i}$ gives rise in turn to 2-*level primary discontinuities* in y''' at any point $\xi_{2,j}(> \xi_{1,i})$ which is a simple root of

$$\alpha(t) = \xi_{1,i} \qquad \text{for some } i.$$

In general, any k-*level primary discontinuity* point $\xi_{k,i}$ gives rise to $(k+1)$-*level primary discontinuities* in $y^{(k+2)}$ at subsequent points $\xi_{k+1,j}$, where the solution of (1.1) becomes smoother and smoother as the primary discontinuity level increases. This increase in the regularity of $y(t)$ will be referred to as *smoothing of the solution*.

Example 3.1 *Consider the equation*

$$\begin{cases} y'(t) = -y(t-1), & t \geq 0, \\ y(t) = 1, & t \leq 0, \end{cases} \tag{3.5}$$

whose solution is depicted in Figure 1. Since $y'(0)^- = 0$ *and* $y'(0)^+ = -y(-1) = -1$, *the derivative function* $y'(t)$ *has a jump at* $t = 0$. *The second derivative* $y''(t)$ *is given by*

$$y''(t) = -y'(t-1),$$

and therefore it has a jump at $t = 1$. *The third derivative* $y'''(t)$ *is given by*

$$y'''(t) = -y''(t-1) = y'(t-2),$$

and hence it has a jump at $t = 2$, *and so forth at multiples of the delay* $t = 3, 4, \ldots$ ◊

On the contrary, the same argument applied to (1.2) reveals that for NDDEs smoothing does not occur and, in general, the solution remains C^0-continuous at any primary discontinuity point.

Example 3.2 *Consider the equation*

$$\begin{cases} y'(t) = -y'(t-1), & t \geq 0, \\ y(t) = t, & t \leq 0, \end{cases} \tag{3.6}$$

whose solution is depicted in Figure 2.

Since $y'(0)^- = 1$ *and* $y'(0)^+ = -y'(-1) = -1$, *the derivative* $y'(t)$ *has a jump discontinuity at* $t = 0$. *Moreover, since* $y'(t) = -y'(t-1)$ *for every* $t \geq 0$, *the derivative* $y'(t)$ *is discontinuous at* $t = 1$ *and at all its multiples* $t = 2, 3, \ldots$ *as well.* ◊

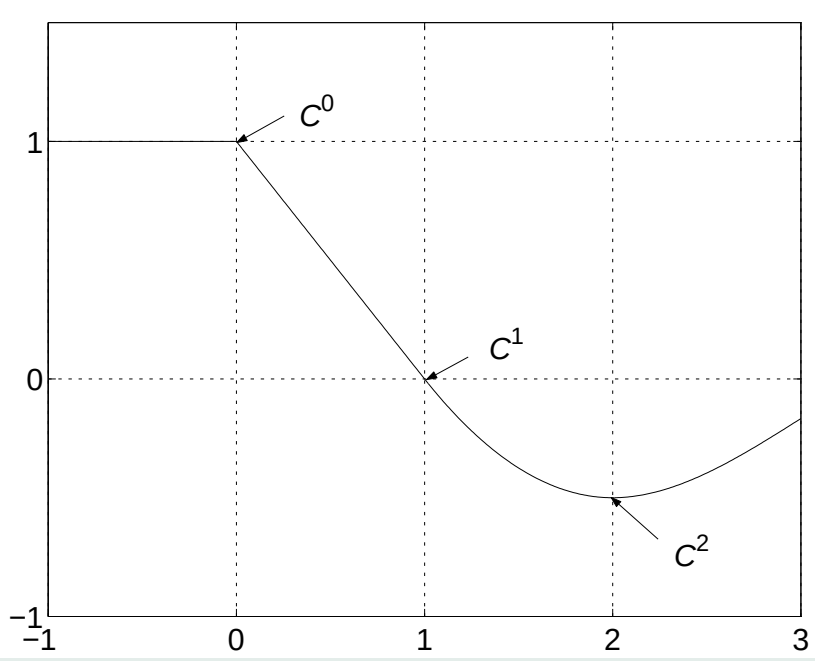

Figure 1: Solutions of (3.5).

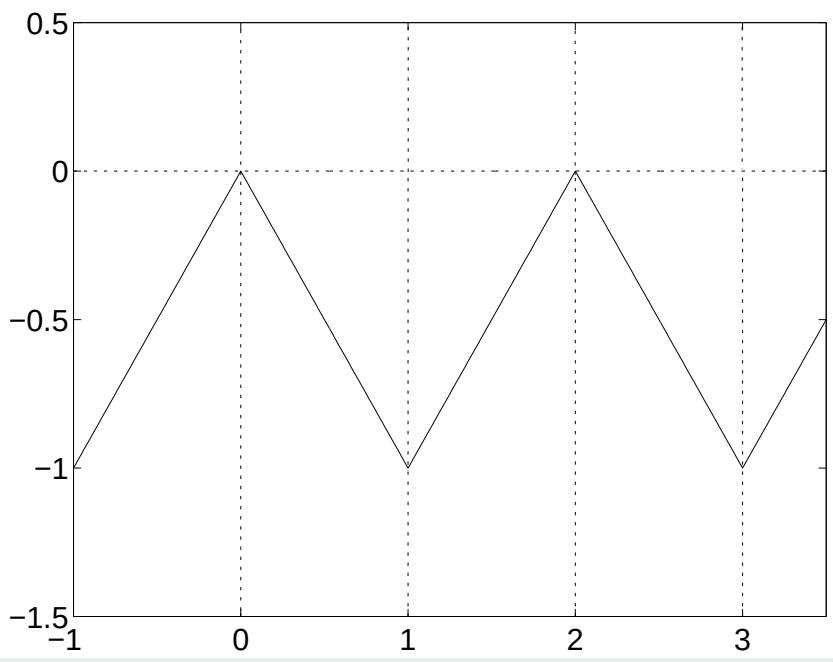

Figure 2: Solutions of (3.6).

In the particular case that a discontinuity point is a root of

$$\alpha(t) = \xi_{j,i} \tag{3.7}$$

for some j, i with *odd* multiplicity $z \geq 3$, smoothing of the solution is faster than for $z = 1$ and possibly applies to neutral equations too. In fact, by (3.3) and (3.4) it is evident that for $\alpha'(\xi_{1,i}) = 0$ the solution y is at least of class C^2 at $\xi_{1,i}$. This phenomenon, called *generalized smoothing*, is specified in the following theorem.

****

Theorem 3.1 (For DDEs) *If $\xi_{j,i}$ is a primary discontinuity point where the function $y(t)$ has continuous derivatives up to the order $\omega - 1$, then $y(t)$ is continuously differentiable at the propagated point $\xi_{j+1,k}$ at least to the order $z \cdot \omega$, provided $\xi_{j+1,k}$ is a root of (3.7) with odd multiplicity z.*

****

We have already observed that, in general, for NDDEs smoothing of the solution does not occur. More precisely, the solution $y(t)$ preserves, at any propagated point, the same regularity as its ancestor and hence as the initial point t_0. Nevertheless, for $\tau = \sigma$, generalized smoothing may actually occur, according to the following theorem, provided that the *splicing condition*

$$\phi'(t_0)^- = y'(t_0)^+ = f\big(t_0, \phi(t_0), \phi(\alpha(t_0), \phi'(\alpha(t_0))\big)$$

holds.

****

Theorem 3.2 (For NDDEs) *If $\xi_{j,i}$ is a primary discontinuity point where the function $y(t)$ has continuous derivatives up to the order $\omega - 1$, then $y(t)$ is continuously differentiable at the propagated point $\xi_{j+1,k}$ at least to the order $z \cdot (\omega - 1)$, provided that $\xi_{j+1,k}$ is a root of (3.7) with odd multiplicity z.*

****

Other discontinuities can appear if the functions f, τ and ϕ in (1.1) and (1.2) have some discontinuities with respect to t in some of their derivatives. Then such discontinuities are also propagated by the deviated arguments $\alpha(t)$ and $\beta(t)$ according to the primary discontinuity propagation rule and are called *secondary discontinuities*. As with the primary discontinuities, to preserve the accuracy order of a numerical method they must also be included in the mesh.

Henceforth, primary and secondary discontinuities will often be referred to as *discontinuities*. However, for the sake of simplicity, we assume that all the functions in (1.1) and (1.2) are C^∞-continuous. Therefore, in the interior of each interval between consecutive primary discontinuity points the solution $y(t)$ is C^∞-continuous as well, and no secondary discontinuities are present.

Definition 3.1 *A discontinuity point ξ is said to be of* order k *if $y^{(s)}(\xi)$ exists for $s = 0, \ldots, k$ and $y^{(k)}$ is Lipschitz continuous at ξ.*

Of course, for DDEs, any p-level primary discontinuity point has order $k \geq p$.

Apart from the trivial case of constant delay, where the discontinuity points are given by $t_0 + k\tau$, $k = 1, 2, \ldots$, the behavior of the deviated arguments $\alpha(t)$ and $\beta(t)$ may be more complicated and even unpredictable for equations with state dependent delay. By speculating on the properties of the deviated argument, one might build up equations with discontinuities that propagate in an arbitrarily wild manner on $[t_0, t_f]$, where possibly $t_f = +\infty$.

3.2. Vanishing and non-vanishing delays

First let us investigate, for (1.1) and for (1.2) with $\tau = \sigma$, how the primary discontinuities are located near the points where the delay $\tau(t)$ vanishes. In this case, called the *vanishing delay* case, a point $\xi > t_0$ is assumed to exist such that $\alpha(\xi) = \xi$. Owing to the continuity of $\alpha(t)$, it is evident that, for any k-level discontinuity point $\xi_{k,i} < \xi$ such that $\alpha(\xi_{k,i}) < \xi_{k,i}$, a $(k+1)$-level discontinuity point, say $\xi_{k+1,j}$, exists such that $\alpha(\xi_{k+1,j}) < \xi_{k+1,j}$ and $\xi_{k,i} < \xi_{k+1,j} < \xi$. In other words, there are infinitely many discontinuity points in any left neighborhood of ξ (see Figure 3).

On the other hand, for DDEs smoothing of the solution takes place, and then a left neighborhood of ξ exists which includes only discontinuity points of arbitrarily large order, that is where the solution is as smooth as needed. This is not the case for NDDEs, where smoothing of the solution does not take place.

In order to avoid the clustering of discontinuities due to vanishing delays, the following hypothesis will often be assumed to be satisfied:

(H_1) There exists a constant $\tau_0 > 0$ such that $\tau = t - \alpha(t) \geq \tau_0$ for all $t \in [t_0, t_f]$.

It is obvious that, under hypothesis (H_1), the distance between a discontinuity point and its ancestor is at least τ_0. Therefore, in any bounded interval $[t_0, t_f]$ the number of discontinuity points is finite.

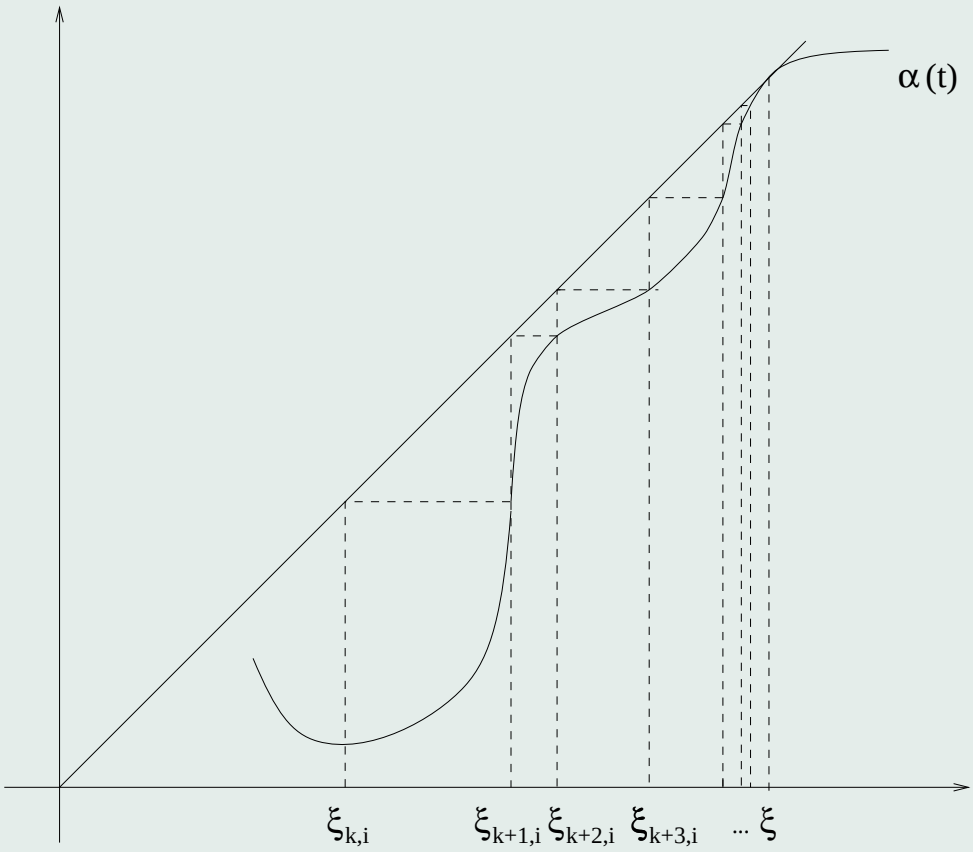

Figure 3: Accumulation of discontinuity points in a left neighborhood of a vanishing delay point ξ.

3.3. Bounded and unbounded delays

Here we investigate how the primary discontinuities propagate according to the general rule

$$\alpha(\xi_{k,j}) = \xi_{k-1,i} \qquad \text{for some } i, \tag{3.8}$$

where, for $k > 0$ and any j, $\xi_{k,j}$ is a k-level primary discontinuity point and $\xi_{0,1} = t_0$ is the only 0-level primary discontinuity point. In particular, when the integration interval is unbounded, i.e. $t_f = +\infty$, it is worth distinguishing between models with a *bounded* or an *unbounded after-effect*, that is with a bounded or an unbounded delay function τ. We shall consider the following hypotheses:

(H_2) $\lim_{t\to+\infty} \alpha(t) = +\infty$.

(H_3) There exists a constant $\tau_1 > 0$ such that $\tau = t - \alpha(t) \le \tau_1$ for all $t \in [t_0, t_f]$.

With respect to the propagation of discontinuities, (H_2) means that the solution is smoothed out indefinitely and, as k increases, the discontinuity points $\xi_{k,i}$ must either coalesce to a vanishing delay point (see Figure 3) or diverge to $+\infty$ (see Figure 4). In both cases, any k-level primary discontinuity point is met after a sufficiently large t.

As for the boundedness hypothesis (H_3), it evidently implies (H_2) but not vice versa. For instance, in the pantograph equation $y'(t) = f\big(t, y(t), y(qt)\big)$, $t \ge 0$, we have

$$\alpha(t) = qt, \quad 0 < q < 1,$$

and

$$\tau(t) = (1 - q)t,$$

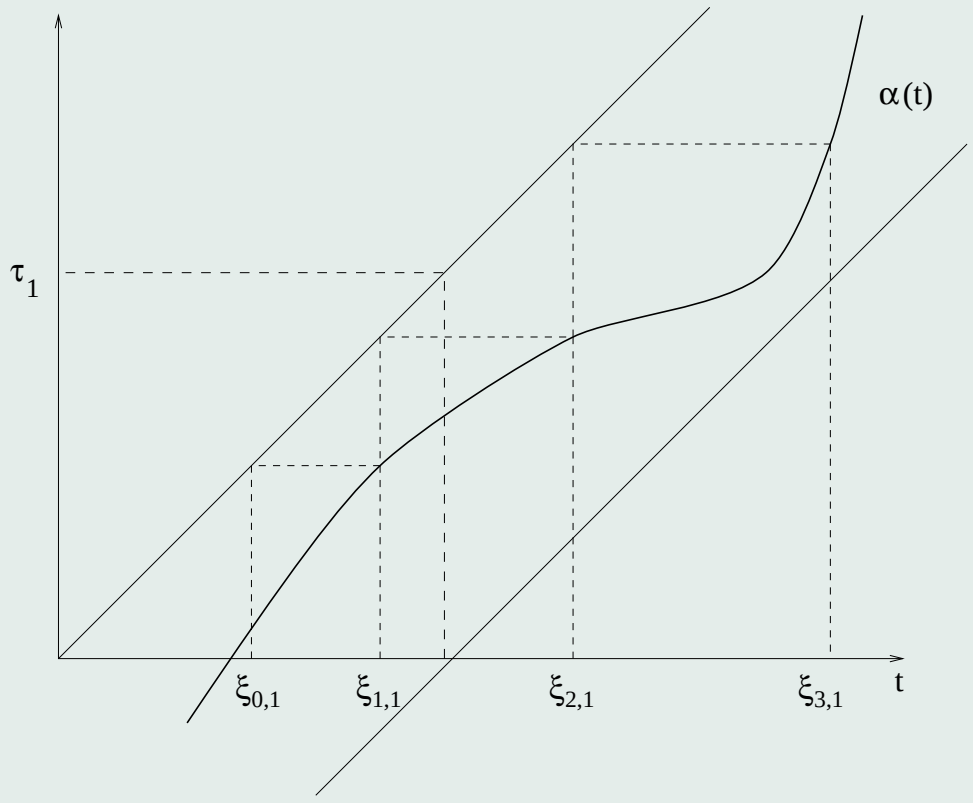

Figure 4: Divergent sequence of discontinuity points.

which are both unbounded.

In particular, when (H_3) holds, the model is said to have *fading memory*. This means that for any t, after a sufficiently long but uniformly bounded elapsed time interval, the solution value $y(t)$ will not influence the right-hand side of (1.1) or (1.2). In other words, to integrate the DDE it is sufficient to store a finite segment of the last past history.

On the contrary, for unbounded delays $\tau(t)$ the deviated argument $\alpha(t) = t - \tau(t)$ may be bounded or not. When both $\tau(t)$ and $\alpha(t)$ are unbounded, as with the pantograph equation, the solution $y(t)$ eventually depends on an arbitrarily large segment of the past history which must be stored for numerical

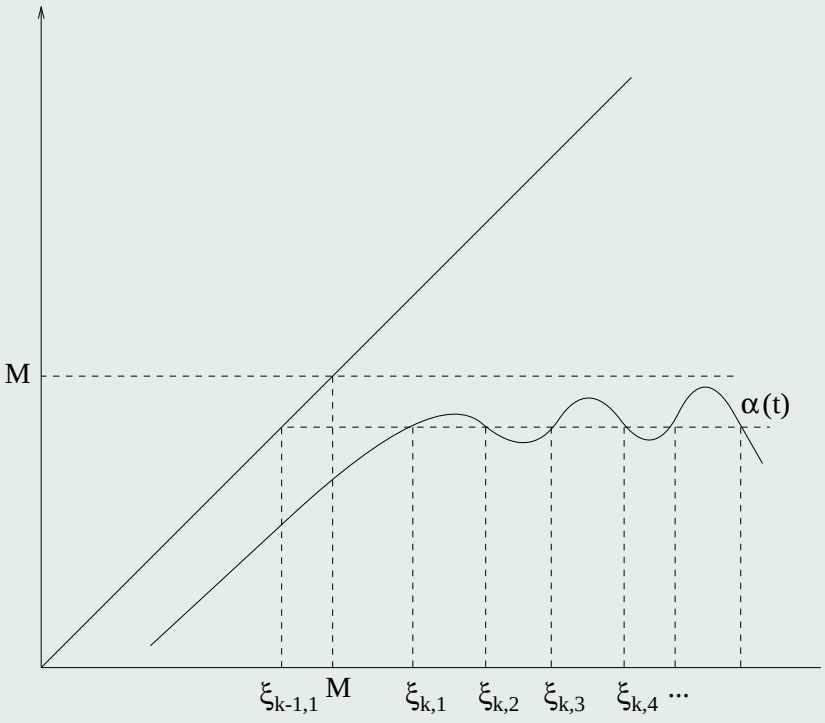

Figure 5: Discontinuity points of the same level generated by a bounded deviated argument.

integration.

If the deviated argument is instead bounded, say $\alpha(t) \leq M$, then infinitely many primary discontinuities might lie to the right of M, but none of them, if any, can increase its level in $[M, +\infty)$ (see Figure 5). This prevents smoothing of the solution beyond a certain class of regularity. Moreover, the solution $y(t)$ definitively depends on the history up to M.

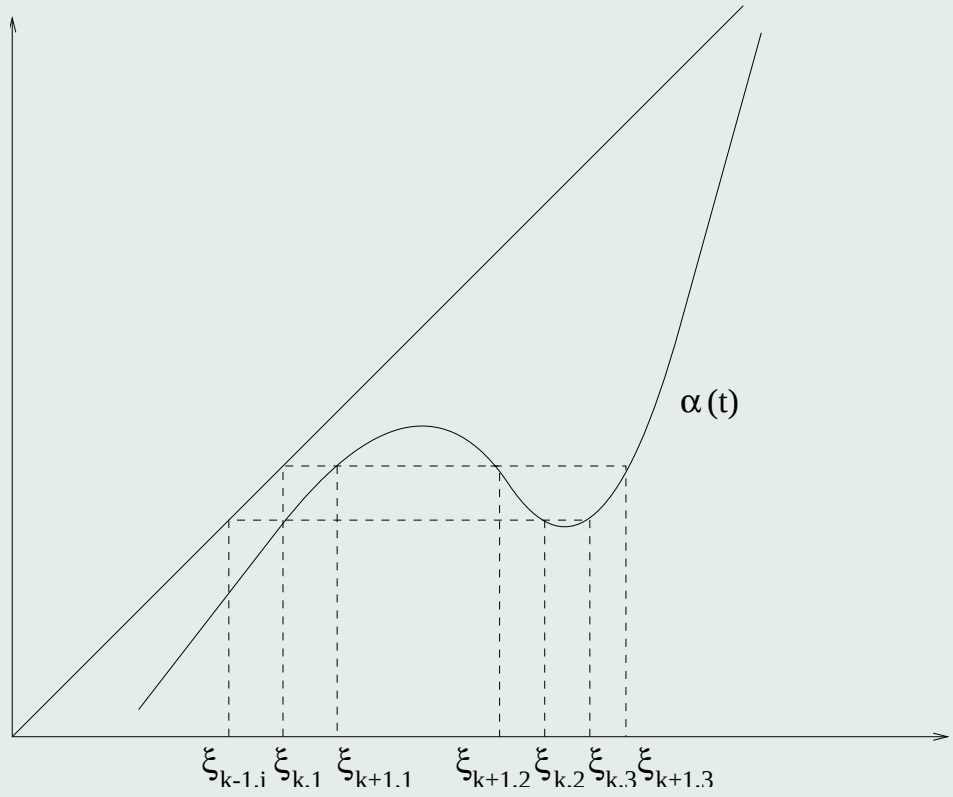

Figure 6: Interlacing of different level discontinuity points.

In many applications the following hypothesis is satisfied:

(H_4) The deviated argument $\alpha(t)$ is a strictly increasing function for all $t \in [t_0, t_f]$.

This is the case, for instance, for all models with constant delay τ. In general, if (H_4) holds and $\alpha(t_0) < t_0$, then the primary discontinuity points form an increasing sequence $\xi_1 < \xi_2 < \cdots < \xi_j < \cdots$, where, for any j, $\xi_j = \xi_{j,1}$ is the sole j-level discontinuity point. That is, as time passes the solution becomes smoother and smoother. On the contrary, oscillating deviated arguments can cause the interlacing of different level discontinuities, as shown in Figure 6.

However, in many cases it is sufficient to locate the *principal discontinuity points* defined as follows.

Definition 3.2 *The one-index subset of primary discontinuity points* $\bar{\xi}_i$ *defined inductively by* $\bar{\xi}_0 = t_0$ *and, for* $i \geq 0$, *by the minimum root* $\bar{\xi}_{i+1}$ *of*

$$\alpha(t) = \bar{\xi}_i$$

with odd multiplicity, is called the set of principal discontinuity points.

The principal discontinuity points are of interest because, for any i,

$$\alpha(t) \leq \bar{\xi}_i \quad \forall t \in [\bar{\xi}_i, \bar{\xi}_{i+1}].$$

Note that $\bar{\xi}_i$ is nothing but $\min_j \xi_{i,j}$. In particular, if (H_4) holds, then all the primary discontinuity points are principal.

3.4. State dependent delays

We briefly consider the difficulties related to the case when the delay has the form $\tau(t, y(t))$. First of all, note that hypothesis (H_1), which aims to avoid vanishing delays, is now modified to

(H_1^*) There exists a constant $\tau_0 > 0$ such that $\tau(t, z) \geq \tau_0$ for all $t \in [t_0, t_f]$ and $z \in \mathbb{R}^d$.

Of course, the monotonicity hypothesis (H_4) cannot be considered.

In order to locate the discontinuities, in principle one should apply the general propagation rule (3.8) which, for the case of state dependent delay, is

$$\xi_{k,j} - \tau\big(\xi_{k,j}, y(\xi_{k,j})\big) = \xi_{k-1,i} \quad \text{for some } i, \tag{3.9}$$

and solve it for $\xi_{k,j}$. Because the delay is dependent of $y(t)$, this cannot be done a priori without knowledge of the solution. Moreover, it is evident that even assuming some approximation of $y(t)$ is available, we must be satisfied with an approximation of the discontinuity point $\xi_{k,j}$. However, finding the multiple roots is an inherently ill-conditioned problem. In conclusion, the impossibility of locating the discontinuity points a priori makes the implementation and convergence analysis of numerical methods for (1.1) and (1.2) a rather complicated task.

3.5. Multiple delays

Sometimes in applications we find DDEs or NDDEs where the right-hand side depends on more than one retarded argument, that is equations such as

$$\begin{cases} y'(t) = f\Big(t, y(t), y\big(\alpha_1(t)\big), \dots, y\big(\alpha_r(t)\big)\Big), & t \geq t_0, \\ y(t) = \phi(t), & t \leq t_0, \end{cases}$$

or

$$\begin{cases} y'(t) = f\Big(t, y(t), y\big(\alpha_1(t)\big), \dots, y\big(\alpha_r(t)\big), \\ \qquad\qquad y'\big(\beta_1(t)\big), \dots, y'\big(\beta_r(t)\big)\Big), & t \geq t_0, \\ y(t) = \phi(t), & t \leq t_0, \end{cases}$$

respectively.

In principle, the presence of many delays does not introduce any particular additional difficulties over those in the one-delay case. The main complication lies in the possibly more chaotic proliferation of the discontinuity points. For example, even if all the deviated arguments are strictly increasing, each discontinuity point generates, in turn, another r such points. Thus, if $\xi_{k-1,j}$ is a discontinuity point of level $k-1$, then we find the corresponding discontinuity points of level k by solving the r equations

$$\alpha_i(\xi_{k,(j-1)r+i}) = \xi_{k-1,j}, \quad i = 1, \dots, r.$$

It may happen that two or more discontinuity points, possibly of different levels, coincide. Similar chaotic proliferation occurs for the principal discontinuities. However, in most cases the difference with respect to the one-delay case is technical rather than conceptual.

3.6. Propagation of discontinuities in systems

Consider equations (1.1) and (1.2) with $d > 1$ and a unique delay $\tau = \sigma$.

If we assume that each component y_i' depends on all components of the delayed solution $y(t-\tau)$, the propagation of discontinuities, as well as the smoothing and generalized smoothing of the solution, takes place according to the results of the previous subsections.

On the contrary, if some components are not coupled to all delayed components, the way the discontinuities propagate is different and some higher order smoothing of solutions might actually occur.

Summer School on Delay Differential Equations and Applications

Dobbiaco (BZ), Italy, June 26–30, 2006

The numerical solution of delay differential equations

M. Zennaro
Dipartimento di Matematica e Informatica
Università di Trieste

Lecture 2: Passing from ODEs to DDEs

Main reference for this Lecture:

A. Bellen and M. Zennaro, Numerical Methods for Delay Differential Equations, Numerical Mathematics and Scientific Computation, Oxford Science Publications, Oxford University Press, 2003 (Chapter 1)

4. Some of the main differences

We want to stress some of the main differences in the behaviour of the solutions of DDEs such as

$$\begin{cases} y'(t) = f\big(t, y(t), y(t-\tau)\big), & t \geq t_0, \\ y(t) = \phi(t), & t \leq t_0, \end{cases} \tag{4.1}$$

with respect to those of ODEs such as

$$\begin{cases} y'(t) = g\big(t, y(t)\big), & t \geq t_0, \\ y(t_0) = y_0. \end{cases} \tag{4.2}$$

As already mentioned in Lecture 1, since for some $t \geq t_0$ it can be that $t - \tau < t_0$, a first difference between equations (4.2) and (4.1) is that the solution of the former is usually determined by an *initial function* $\phi(t)$ rather than by a simple initial value y_0, as happens for the latter.

Moreover, in general, the right-hand derivative $y'(t_0)^+$, that is $f(t_0, \phi(t_0), \phi(t_0 - \tau))$, does not equal the left-hand derivative $\phi'(t_0)^-$ and hence the solution y is not smoothly linked to the initial function $\phi(t)$ at the point t_0, where only C^0-continuity can be assured. Moreover, such a derivative jump discontinuity propagates from the initial point t_0 along the integration interval and gives rise to subsequent discontinuity points where the solution is smoothed out more and more. As a consequence, even if the functions $f(t, y, x)$, $\tau(t, y)$ and $\phi(t)$ in (4.1) are C^∞-continuous, in general the solution $y(t)$ is simply C^1-continuous in $[t_0, t_f]$.

The presence of an initial function in the problem (4.1) has various other unexpected consequences on the solutions. Some of them are illustrated by the following examples.

Example 4.1 *Unlike the ordinary equations, there is no longer injectivity between the set of initial data and the set of solutions $y(t)$, $t \geq t_0$. In fact, the equation*

$$y'(t) = y(t-1)\big(y(t)-1\big), \qquad t \geq 0,$$

has the constant solution $y(t) = 1$ in $[0, +\infty)$ for any initial function $\phi(t)$ defined in $[-1, 0]$ such that $\phi(0) = 1$. $\Diamond$

The next three examples show that, in the state dependent delay case, the lack of regularity of the initial function $\phi(t)$ may cause the *non-uniqueness* of the solution of (4.1) or its *termination* after some bounded interval. Moreover, in the latter case also strange behaviors of the numerical methods may be observed, as the raise of *ghost solutions*.

Example 4.2 (non-uniqueness) *As an example of non-uniqueness, consider the equation*

$$\begin{cases} y'(t) = y\big(t - |y(t)| - 1\big) + \frac{1}{2}, & t \geq 0, \\ y(t) = \phi(t), & t \leq 0, \end{cases} \tag{4.3}$$

where

$$\phi(t) = \begin{cases} 1, & t < -1, \\ 0, & -1 \leq t \leq 0. \end{cases} \tag{4.4}$$

It is easy to see that in $[0, 2]$ both functions

$$y(t) = \tfrac{3}{2}t$$

and

$$y(t) = \tfrac{1}{2}t$$

are solutions of (4.3). $\Diamond$

Example 4.3 (termination) *As an example of termination of the solution, consider the equation*

$$\begin{cases} y'(t) = -y\big(t-2-y(t)^2\big)+5, & t \geq 0, \\ y(t) = \phi(t), & t \leq 0, \end{cases} \tag{4.5}$$

where

$$\phi(t) = \begin{cases} \frac{9}{2}, & t < -1, \\ \\ -\frac{1}{2}, & -1 \leq t \leq 0. \end{cases} \tag{4.6}$$

The solution in $[0, \frac{125}{121}]$ *is given by*

$$y(t) = \begin{cases} \frac{1}{2}(t-1), & 0 \leq t \leq 1, \\ \\ \frac{11}{2}(t-1), & 1 \leq t \leq \frac{125}{121}. \end{cases} \tag{4.7}$$

It is not difficult to see that the solution cannot be continued beyond the point $t = \frac{125}{121}$. *In fact, at* $t = \frac{125}{121}$ *the deviated argument* $t-2-y(t)^2$ *is equal to* -1 *and therefore, in a right neighborhood of such a point,* $y(t-2-y(t)^2)$ *is given by one of the two values of* $\phi(t)$. *Thus the solutions of (4.5) should take the form*

$$y(t) = c\Big(t - \tfrac{125}{121}\Big) + \tfrac{2}{11},$$

with

$$c = \tfrac{1}{2} \quad \text{if } t-2-y(t)^2 < -1$$

and

$$c = \tfrac{11}{2} \quad \text{if } t-2-y(t)^2 \geq -1.$$

Now, each choice of c *leads to a solution* $y(t)$ *that contradicts the assumption made on* $t-2-y(t)^2$ *and hence the solution does not exist for* $t > \frac{125}{121}$. $\Diamond$

Example 4.4 (ghost solutions) *From a numerical point of view, termination of the solution is a very delicate issue. In fact, it may result in surprising and misleading behavior in the implementation of the numerical method. For instance, with reference to the previous Example 4.3, in a right neighborhood of the termination point* $t_N = \frac{125}{121}$*, where* $y_N \approx \frac{2}{11}$*, the* forward Euler *method reads*

$$y_{n+1} = y_n + h_{n+1}(-\tfrac{9}{2} + 5) \quad \text{if } t_n - 2 - y_n^2 < -1$$

and

$$y_{n+1} = y_n + h_{n+1}(\tfrac{1}{2} + 5) \quad \text{if } t_n - 2 - y_n^2 \geq -1,$$

and for no reason it stops integrating at any $n \geq N$*. The resulting approximation is plotted in Figure 7 where, for* $t \geq \frac{125}{121}$*, a* ghost solution *appears that approximates the function* $\sqrt{t-1}$*. Such a function is not a solution of (4.5) but fulfills the equation* $t - 2 - y(t)^2 = -1$*. In fact, the numerical solution is forced to attain values* y_n *such that the delayed arguments* $t_n - 2 - y_n^2$ *oscillate around the equilibrium value* -1*.*

On the other hand, the backward Euler *method reads*

$$y_{n+1} = y_n + h_{n+1}(-\tfrac{9}{2} + 5) \quad \text{if } t_{n+1} - 2 - y_{n+1}^2 < -1$$

and

$$y_{n+1} = y_n + h_{n+1}(\tfrac{1}{2} + 5) \quad \text{if } t_{n+1} - 2 - y_{n+1}^2 \geq -1.$$

It is not difficult to see that, for small values of h_{n+1}*, no solution* y_{n+1} *exists. This agrees perfectly with the theory, but may result in a very large number of rejected steps in the root-finding mechanism before the overall procedure stops.* $\Diamond$

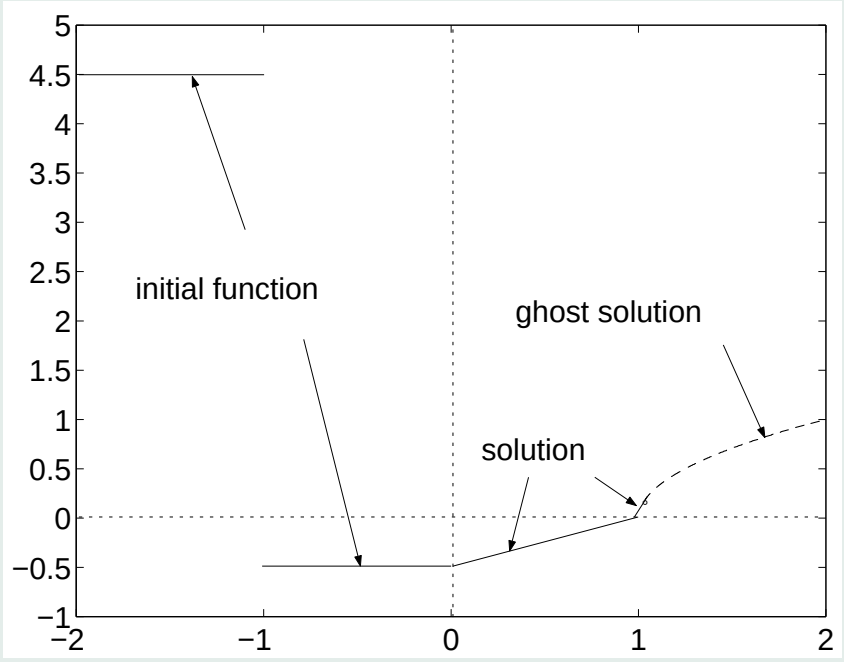

Figure 7: Solutions (solid) and ghost solution (dashed) of (4.5).

As we have seen in Lecture 1, for equations (4.1), where the delay does not depend on $y(t)$, the theory of existence and uniqueness of solutions does not present substantial additional difficulties with respect to the ordinary case (4.2) as long as the deviated argument $t - \tau(t)$ is uniformly strictly positive. Nevertheless, even in the case of constant delay, DDEs possess a dynamical structure that is richer than ODEs. In fact, whereas bounded solutions of autonomous ODEs may oscillate only if the system has at least two components and may behave chaotically only if there are at least three components (Poincaré–Bendixon theorem), DDEs may already exhibit oscillatory and even chaotic behavior in the scalar case.

Example 4.5 *Consider the following* delay logistic equation*:*

$$y'(t) = ay(t)\big(1 - y(t-1)\big), \tag{4.8}$$

which models the dynamics of populations. It improves the Verhulst–Pearl model $y'(t) = ay(t)\big(1 - y(t)\big)$ *in that the* growing factor $1 - y(t)$ *does not act instantaneously but after some time lag.*

Whereas the solutions of the Verhulst–Pearl equation are monotonic, the positive solutions of (4.8) are monotonic for $a \in (0, 1/e)$*, oscillate for* $a \in [1/e, \pi/2)$ *and approach periodic orbits for* $a > \pi/2$ *(Figures 8 and 9).* $\Diamond$

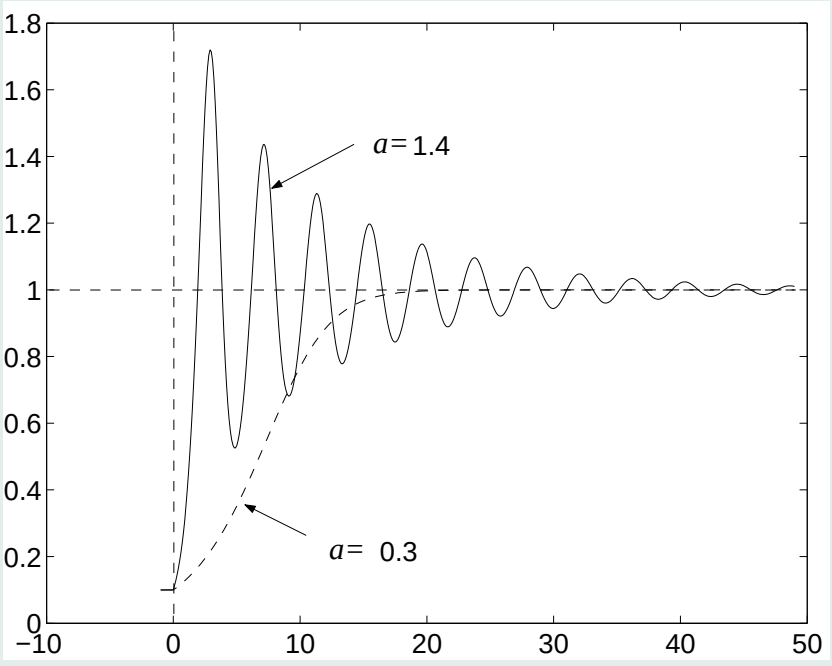

Figure 8: Solutions of (4.8) with $y(t) = 0.1$, $t \leq 0$, for $a = 1.4$ and 0.3.

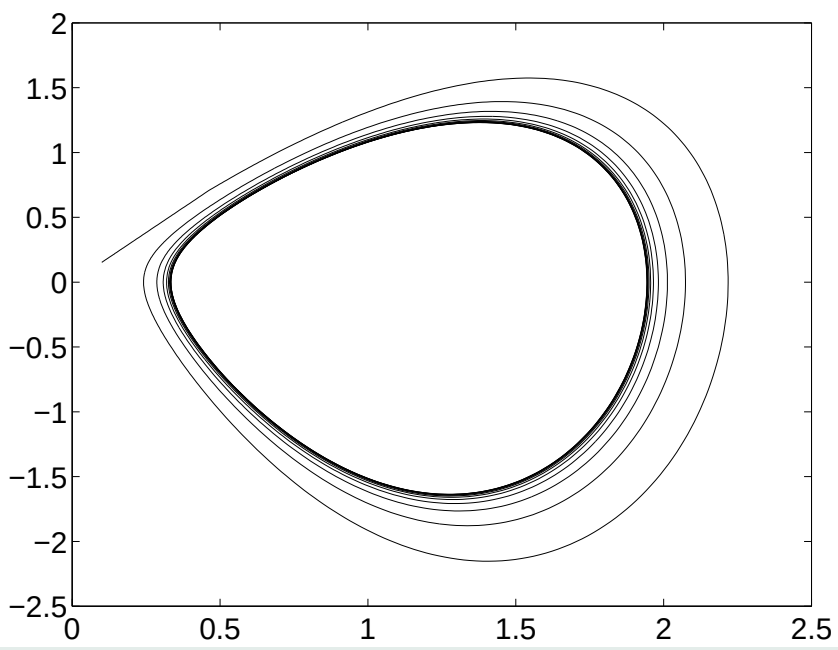

Figure 9: Solution of (4.8) for $a = 1.7$ in the phase plane.

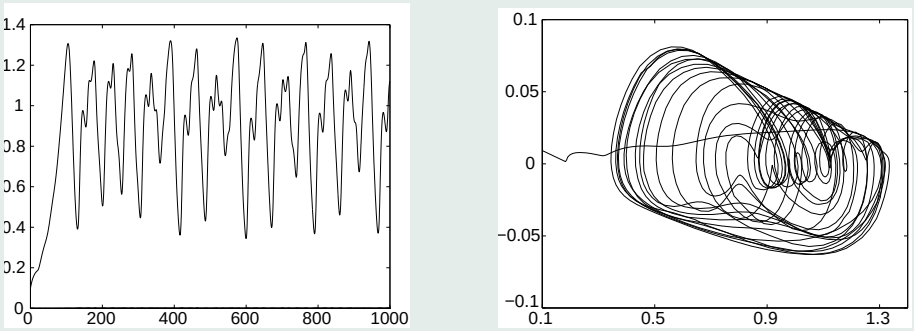

Figure 10: Solutions of equation (4.9) for $t \leq 1000$ in the (t, y)-plane (left) and in the phase plane (right).

Example 4.6 *Consider the following equation, connected with the release of mature cells into the blood stream:*

$$y'(t) = \frac{by(t-\tau)}{1+[y(t-\tau)]^n} - ay(t). \tag{4.9}$$

For certain values of the parameters and of the delay, the solution is oscillatory, and sometimes it even oscillates chaotically. This is the case in patients with leukemia. In particular, for $a = 0.1$, $b = 0.2$, $n = 10$ and $\tau = 20$, the model behaves chaotically, as shown in Figure 10. $\Diamond$

Finally, observe that the presence of a delayed term may drastically change the qualitative behavior of the solution by acting as a stabilizer or a destabilizer of models governed by ODEs.

Example 4.7 *Consider the linear scalar equation*

$$\begin{cases} y'(t) = \lambda y(t) + \mu y(t-1), & t \geq 0, \\ y(t) = -t + 1, & t \leq 0, \end{cases} \tag{4.10}$$

with real constant coefficients λ and μ. It is known that, for $\mu = 0$, the solution of equation (4.10), which reads

$$\begin{cases} y'(t) = \lambda y(t), & t \geq 0, \\ y(0) = 1, \end{cases} \tag{4.11}$$

vanishes asymptotically for any negative λ, whereas it blows up for any positive λ. Moreover, in the former case it remains bounded by the initial value 1. On the other hand, for $\mu \neq 0$ the delayed term $\mu y(t-1)$ in (4.10) acts as a forcing term *and the above-mentioned properties of the solution might not hold. In particular, for any $\mu > 0$ there exists $\lambda < 0$ for which the solution does not vanish asymptotically, and another $\lambda < 0$ for which the solution does, but is not bounded by the initial value $y(0) = 1$. Figure 11 illustrates such situations for $\mu = 4$ and for $\lambda = -3.5$ and -5. Also, for $\lambda = 0.5$ and $\mu = -1$ the delayed term $-y(t-1)$ acts as a* stabilizer *of the model whose solution behaves stably despite the positivity of λ (see Figure 12).* $\Diamond$

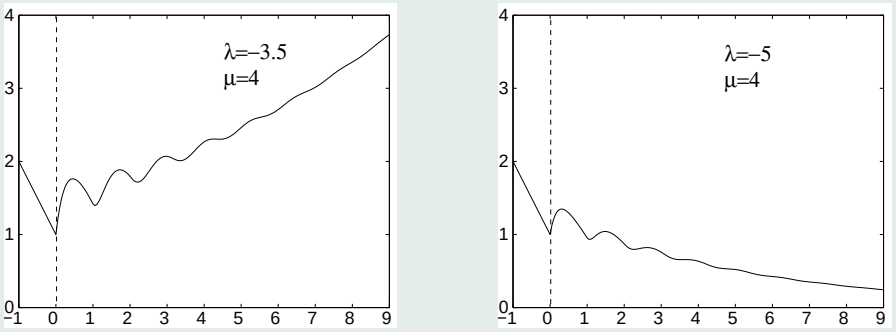

Figure 11: Stable and unstable solutions of (4.10) for $\lambda < 0$.

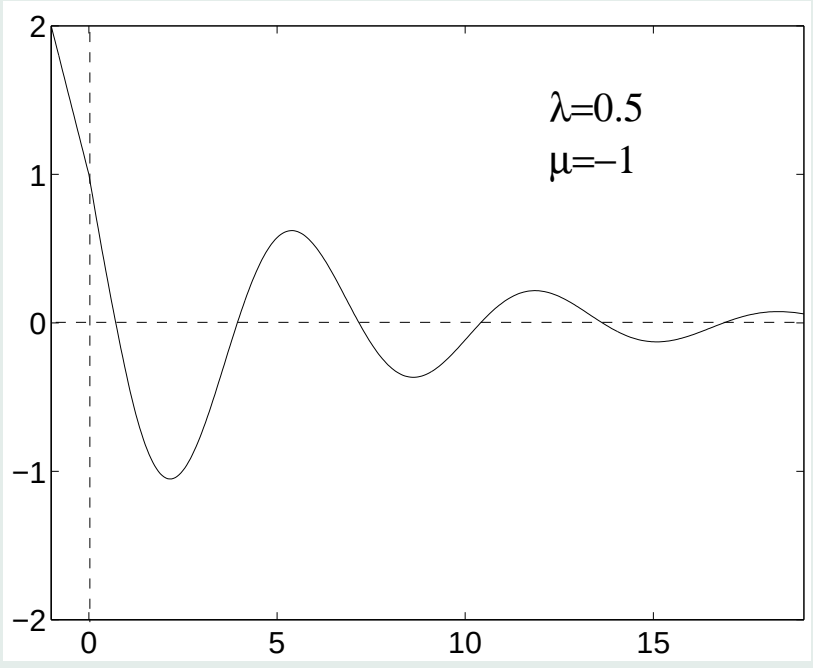

Figure 12: Stable solution of (4.10) for $\lambda > 0$.

Example 4.8 *An example showing that small delays can produce a large effect is given by the NDDE*

$$y'(t) = -1.5y'(t-\tau) + \lambda y(t), \quad \lambda < 0,$$

the solution of which is asymptotically stable for $\tau = 0$ *and unstable for any* $\tau > 0$. *In this case the delayed term acts as a* destabilizer. $\Diamond$

5. Numerical ODE theory is not enough for DDEs

In order to illustrate some basic features of numerical methods for DDEs and the differences they exhibit with respect to ODE methods, consider the constant delay differential equation

$$\begin{cases} y'(t) = f\big(t, y(t), y(t-1)\big), & t \geq 0, \\ y(t) = \phi(t), & t \leq 0. \end{cases} \tag{5.1}$$

The most natural, but not unique, approach for solving (5.1) numerically is to assign integration steps less than or equal to the delay $\tau = 1$ and to integrate step by step the ODEs obtained from (5.1) by substituting the delayed term $y(t-1)$ by a function $\eta(t-1)$ given, according to the value of t, either by the initial function $\phi(t-1)$ or by a continuous extension of the approximate solution previously computed by the method itself. Thus, at the $(n+1)$st step the equation to be solved is

$$\begin{cases} w'_{n+1}(t) = f\big(t, w_{n+1}(t), x(t-1)\big), & t_n \leq t \leq t_{n+1}, \\ w_{n+1}(t_n) = y_n, \end{cases} \tag{5.2}$$

where

$$x(s) = \begin{cases} \phi(s) & \text{for } s \leq 0, \\ \eta(s) & \text{for } 0 \leq s \leq t_n. \end{cases}$$

The integration formula provides the value y_{n+1} and the approximate solution η of (5.1) is then continued in $[t_n, t_{n+1}]$ in such a way that $\eta(t_{n+1}) = y_{n+1}$.

A peculiarity of this approach is that, whereas the numerical ODE method furnishes approximate values of the solution at nodal points only, implementation of the numerical method for the solution of (5.2) may require knowledge of the approximate solution $\eta(t)$ at some points $t-1$ possibly other than the nodal points. Therefore, in general, the DDE methods will be based on *continuous extensions* of numerical ODE schemes. This can be done either by a posteriori *interpolation* of the values y_n given by the underlying *discrete* ODE method or, preferably, by *continuous* ODE methods, that is methods that provide step by step a continuous approximation of the solution (see Figure 13). As we shall see, the success of the resulting DDE method, in terms of accuracy and stability, depends on the particular choice of the discrete method and of the continuous extension as well.

We have already pointed out that the presence of a delayed term can drastically modify some boundedness or stability properties and, in general, the dynamics of the simpler ODE model. Now we want to illustrate, by means of some examples, that also in the implementation of numerical schemes some desirable accuracy and stability properties of the underlying ODE method can be destroyed when the method is applied to a DDE.

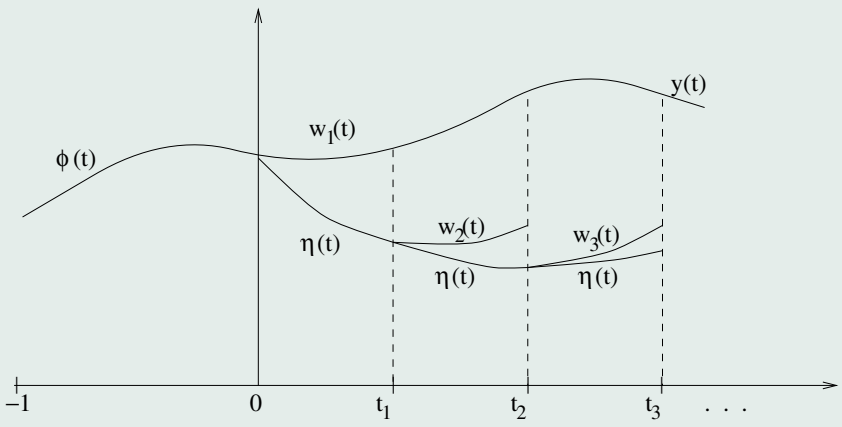

Figure 13: Approximate solution of (5.1) by the continuous ODE method.

5.1. On the order of the methods

In order to illustrate the possible loss of accuracy, consider the class of constant coefficient linear equations

$$\begin{cases} y'(t) = ay(t) - \frac{\pi}{2}\mathrm{e}^a y(t-1), & t \geq 0, \\ y(t) = \phi(t) = \mathrm{e}^{at}\sin(\frac{\pi}{2}t), & t \leq 0, \end{cases} \tag{5.3}$$

whose solutions, $y(t) = \mathrm{e}^{at}\sin(\frac{\pi}{2}t)$, are of class C^∞ in $[-1, +\infty)$.

According to (5.2), for $n = 0, 1, \ldots$, we shall solve the ODE

$$\begin{cases} w'_{n+1}(t) = aw_{n+1}(t) - \frac{\pi}{2}\mathrm{e}^a x(t-1), & t_n \leq t \leq t_{n+1}, \\ w_{n+1}(t_n) = y_n, \end{cases} \tag{5.4}$$

where

$$x(s) = \begin{cases} \phi(s) = \mathrm{e}^{as}\sin(\frac{\pi}{2}s) & \text{for } s \leq 0, \\ \eta(s) & \text{for } 0 \leq s \leq t_n. \end{cases}$$

A good class of methods for integrating (5.4) is given by *collocation at ν Gaussian points*, which can be viewed as ν-stage Runge–Kutta methods of order 2ν. Since they are projection methods based on piecewise polynomial approximations of degree ν, they also furnish a continuous extension $\eta(t)$ of uniform accuracy order $\nu+1$. Therefore, Gaussian collocation appears to be an attractive class of continuous methods for integrating DDEs such as (5.3).

For $\nu = 1$ the method is known as the *midpoint rule*, and for the general equation (5.2) it takes the form

$$y_{n+1} = y_n + hf\left(t_n + \frac{h}{2}, \frac{y_n + y_{n+1}}{2}, x\Big(t_n + \frac{h}{2} - 1\Big)\right). \quad (5.5)$$

Application of (5.5) to (5.4) with constant integration stepsize $h = 1/(m-\delta)$, $m \geq 2$, m integer, and $0 \leq \delta < 1$, yields

$$y_{n+1} = \begin{cases} y_n + h\Big(a\frac{y_n+y_{n+1}}{2} - \mathrm{e}^a\frac{\pi}{2}\phi(t_n + \frac{h}{2} - 1)\Big), & t_n + \frac{h}{2} - 1 \leq 0, \\ y_n + h\Big(a\frac{y_n+y_{n+1}}{2} - \mathrm{e}^a\frac{\pi}{2}\eta(t_n + \frac{h}{2} - 1)\Big), & t_n + \frac{h}{2} - 1 > 0, \end{cases}$$

where $\eta\big(t_n + \frac{h}{2} - 1\big)$ is given by linear interpolation, that is

$$\begin{aligned} \eta\Big(t_n + \frac{h}{2} - 1\Big) &= \eta\Big(t_n + \frac{h}{2} - (m-\delta)h\Big) \\ &= \begin{cases} (\frac{1}{2} - \delta)y_{n-m} + (\delta + \frac{1}{2})y_{n-m+1}, & 0 \leq \delta \leq \frac{1}{2}, \\ (\frac{3}{2} - \delta)y_{n-m+1} + (\delta - \frac{1}{2})y_{n-m+2}, & \frac{1}{2} < \delta < 1. \end{cases} \end{aligned}$$

Summing up, the midpoint rule for (5.3) takes the form

$$y_{n+1} = \frac{(1 + \frac{1}{2}ha)y_n - \frac{\pi}{2}h\mathrm{e}^{a(n+\frac{1}{2})h}\sin\big(\pi/2((n+\frac{1}{2})h - 1)\big)}{1 - \frac{1}{2}ha} \quad (5.6)$$

for $n \leq m - \delta - \frac{1}{2}$ and

$$y_{n+1} = \begin{cases} \dfrac{(1 + \frac{1}{2}ha)y_n - \frac{\pi}{2}\mathrm{e}^a h\big((\frac{1}{2} - \delta)y_{n-m} + (\delta + \frac{1}{2})y_{n-m+1}\big)}{1 - \frac{1}{2}ha}, & 0 \leq \delta \leq \frac{1}{2}, \\ \dfrac{(1 + \frac{1}{2}ha)y_n - \frac{\pi}{2}\mathrm{e}^a h\big((\frac{3}{2} - \delta)y_{n-m+1} + (\delta - \frac{1}{2})y_{n-m+2}\big)}{1 - \frac{1}{2}ha}, & \frac{1}{2} < \delta < 1, \end{cases} \quad (5.7)$$

for $n > m - \delta - \frac{1}{2}$.

For $\nu > 1$ the method takes a more complicated form involving the solution of a linear system in $\mathbb{R}^\nu$ and the use of a polynomial of degree ν at each integration step. Let us restrict ourselves to the investigation of $\nu = 1$ and $\nu = 2$ and recall that, whereas the uniform and discrete accuracy orders of the midpoint rule ($\nu = 1$) for ODEs are both equal to 2, the two-stage Gaussian collocation ($\nu = 2$) has nodal order equal to 4 and uniform order equal to 3 (see Section 7.1).

In order to check the nodal accuracy order p of the resulting DDE methods, consider the solution of (5.3) in the interval $[0, 10]$ and remember that the solution is of class C^∞. Therefore, no discontinuities are present which can spoil the accuracy order of the method. Let us denote by e_h the maximum absolute error at the nodal points for the integration stepsize $h = 1/(m - \delta)$. By halving the stepsize, the asymptotic value of the ratio $r_h = e_h/e_{\frac{h}{2}}$ is expected to be a power of 2, where the exponent is the order p.

The behavior of r_h is shown in Figure 14 for integer ($\delta = 0$) and non-integer ($\delta > 0$) values of $m - \delta$, corresponding to integration stepsizes h which are and which are not submultiples of the delay $\tau = 1$. In both cases the midpoint rule turns out to preserve the expected nodal order 2. On the contrary, the two-stage Gaussian collocation method exhibits nodal order 4 or 3 according to whether or not $m-\delta$ is integer, while the uniform approximation is accurate to order 3. More generally, we shall see that for arbitrary stepsizes the DDE method based on ν-stage Gaussian collocation exhibits nodal and uniform accuracy orders equal to the uniform order $\nu+1$ of the continuous ODE method whereas, for a constant stepsize $h = 1/m$, with m integer, it preserves the nodal accuracy order 2ν.

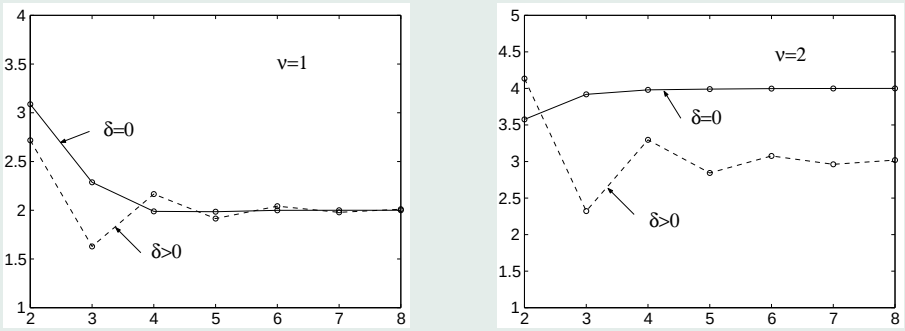

Figure 14: Logarithmic plots of ratios r_h, $h = 1/(m-\delta)$, $m - \delta = 2^i * k$, as a function of i, for the numerical solutions of equation (5.3) with $a = 1$ obtained by collocation at $\nu = 1$ and $\nu = 2$ Gaussian points. The values of $m - \delta$ are determined by $k = 1$ (solid) and $k = 5/3$ (dotted).

5.2. On the stability properties of the methods

As for the stability analysis of DDE methods, consider the class of constant coefficient linear test equations

$$\begin{cases} y'(t) = \lambda y(t) - \frac{4}{5}\lambda y(t-1), & t \geq 0, \\ y(t) = 1, & t \leq 0, \end{cases} \tag{5.8}$$

whose solutions, depicted in Figure 15 for some values of λ, are asymptotically stable for any $\lambda < 0$.

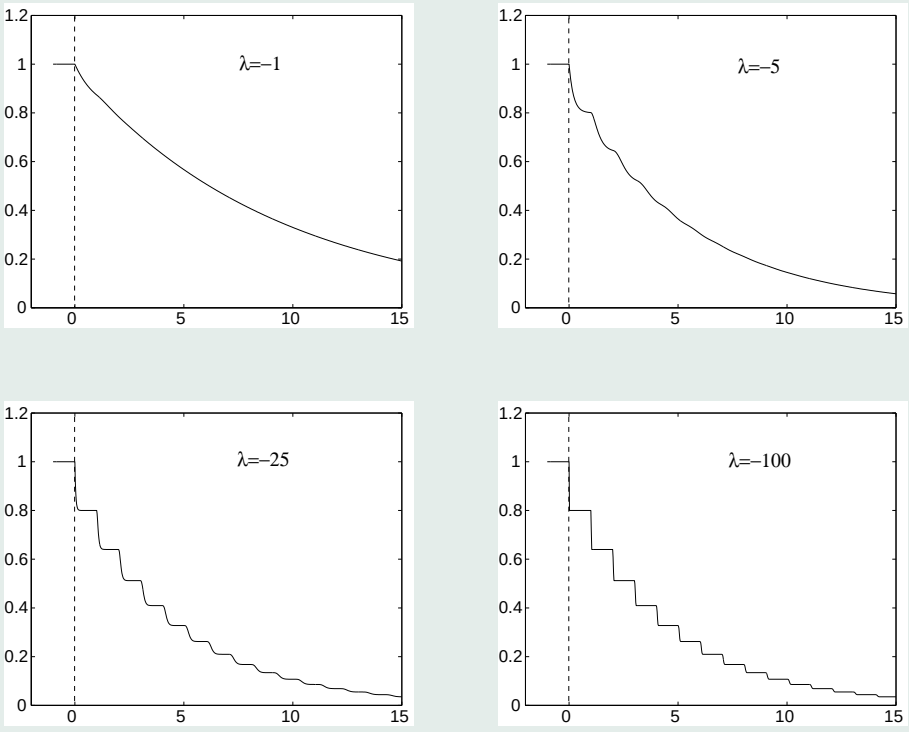

Figure 15: Solutions of (5.8) for some $\lambda < 0$.

The midpoint rule (5.5), extended by linear interpolation, for equation (5.8) takes the form

$$y_{n+1} = \frac{(1+\frac{1}{2}h\lambda)y_n - \frac{4}{5}h\lambda}{1-\frac{1}{2}h\lambda}$$

for $n \le m - \delta - \frac{1}{2}$ and

$$y_{n+1} = \begin{cases} \dfrac{(1+\frac{1}{2}h\lambda)y_n - \frac{4}{5}h\lambda\big((\frac{1}{2}-\delta)y_{n-m} + (\delta+\frac{1}{2})y_{n-m+1}\big)}{1-\frac{1}{2}h\lambda}, & 0 \le \delta \le \frac{1}{2}, \\ \dfrac{(1+\frac{1}{2}h\lambda)y_n - \frac{4}{5}h\lambda\big((\frac{3}{2}-\delta)y_{n-m+1} + (\delta-\frac{1}{2})y_{n-m+2}\big)}{1-\frac{1}{2}h\lambda}, & \frac{1}{2} < \delta < 1, \end{cases} \tag{5.9}$$

for $n > m - \delta - \frac{1}{2}$.

It is well known that the midpoint rule is *A-stable*, that is for all equations $y' = \lambda y$, $\Re(\lambda) < 0$, it provides numerical solutions that vanish asymptotically for any integration stepsize $h > 0$. Since the solutions of (5.8) are asymptotically stable for any $\lambda < 0$, one expects the same behavior for the numerical solution given by the midpoint rule independently of the stepsize h. On the contrary, whereas application of (5.9) provides stable numerical solutions of (5.8) whenever $m - \delta$ is integer ($\delta = 0$), somewhat surprisingly, non-integer values of $m - \delta$ may produce numerical instability. Figure 16 shows the numerical solutions of (5.8) with $\lambda = -50$, given by (5.9) for $m - \delta = 10$ and $m - \delta = 12.5$. Despite the latter being obtained by a smaller integration stepsize, it behaves unstably.

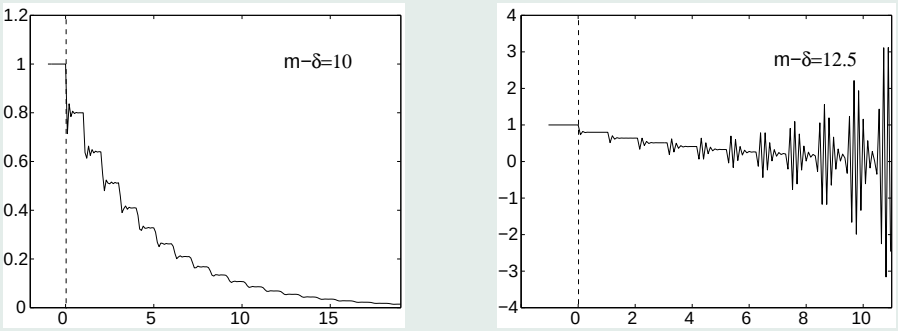

Figure 16: Numerical solutions of equation (5.8) with $\lambda = -50$ by the midpoint rule with $h = 1/(m-\delta)$ for integer and non-integer $m - \delta$.

Now consider another very popular A-stable method, namely the *trapezoidal rule*, which, for the general equation (5.2), is

$$y_{n+1} = y_n + \frac{h}{2}\big(f(t_n, y_n, x(t_n - 1)) + f(t_{n+1}, y_{n+1}, x(t_{n+1} - 1))\big). \tag{5.10}$$

Application of (5.10) to (5.8) with constant stepsize $h = 1/(m-\delta)$, m integer, $m \geq 2$, and $0 \leq \delta < 1$, and linear interpolation between nodal points yields

$$y_{n+1} = \frac{(1 + \frac{1}{2}h\lambda)y_n - \frac{4}{5}h\lambda}{1 - \frac{1}{2}h\lambda}$$

for $n \leq m - 2$,

$$y_{n+1} = \frac{(1 + \frac{1}{2}h\lambda)y_n - \frac{1}{2}h\frac{4}{5}\lambda(2 - \delta + \delta y_1)}{1 - \frac{1}{2}h\lambda}$$

for $n = m - 1$ and

$$y_{n+1} = \frac{(1 + \frac{1}{2}h\lambda)y_n - \frac{1}{2}h\frac{4}{5}\lambda\big((1-\delta)y_{n-m} + y_{n-m-1} + \delta y_{n-m+2}\big)}{1 - \frac{1}{2}h\lambda} \tag{5.11}$$

for $n \geq m$.

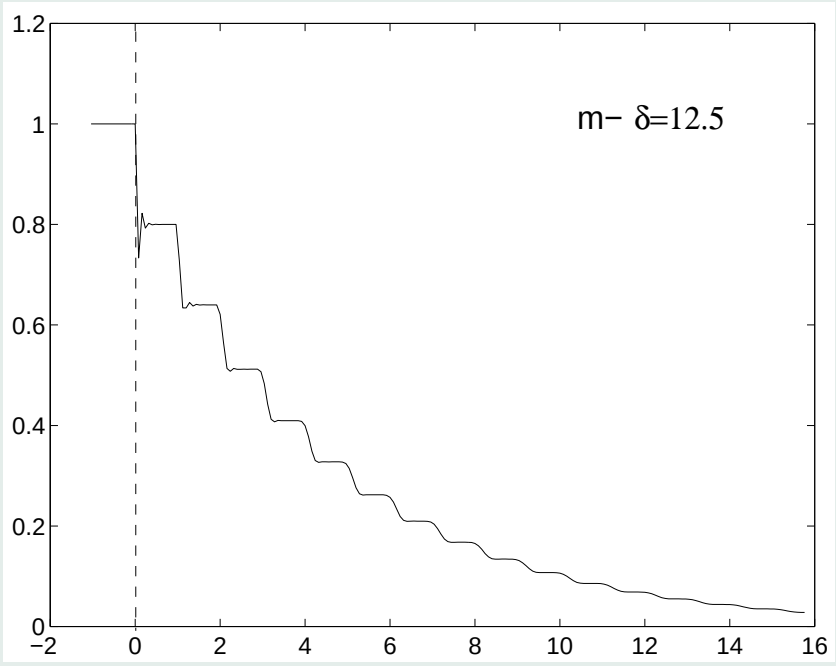

Figure 17: Numerical solution of equation (5.8) with $\lambda = -50$ by the trapezoidal rule with $m - \delta = 12.5$.

Unlike the midpoint rule, the trapezoidal rule also provides stable solutions for integration stepsizes that are not submultiples of the delay. Figure 17 illustrates the solution behavior of (5.11) for $\lambda = -50$ and $m - \delta = 12.5$, where the midpoint rule failed.

Although the trapezoidal rule appears to be more robust than the midpoint rule for constant coefficient linear DDEs, they may both be inadequate for the larger class of variable coefficient linear DDEs, even for integer values of $m - \delta$.

In order to illustrate this further discrepancy between the methods as applied to ODEs and DDEs, consider the equation

$$\begin{cases} y'(t) = \lambda(t)y(t) - \frac{4}{5}\lambda(t)y(t-1), & t \geq 0, \\ y(t) = t+1, & t \leq 0, \end{cases} \tag{5.12}$$

where $\lambda(t) = -50\sin^2(\frac{2\pi}{3}(t - \frac{1}{4}))$, whose solution, plotted in Figure 18, is asymptotically stable.

Now, for the variable coefficient linear equation $y'(t) = \lambda(t)y(t)$, $y(0) = y_0$, it is known that the solutions go to zero for any real function $\lambda(t) \leq 0$ such that $\int_0^t \lambda(s)\mathrm{d}s \to -\infty$ as $t \to +\infty$. It is also known that the midpoint rule is *algebraically stable* and, consequently, the numerical solution is bounded by the initial value $|y_0|$ for any constant stepsize $h > 0$. Unlike the case of constant coefficients, now the midpoint rule gives rise to unstable solutions even for some constant integration stepsizes that are submultiples of the delay. Results for $h = 0.5$, corresponding to $m - \delta = 2$, are plotted in Figure 19 (left).

As the final negative result in this section, we show that the trapezoidal rule may also behave unstably for equations such as (5.12). To this end, consider equation (5.12) with the slightly different coefficient $\lambda(t) = -50\sin^2(\frac{2\pi}{3}t)$, whose solution is plotted in Figure 18 (right). The result, for the same integration stepsize $h = 0.5$, is plotted in Figure 19 (right).

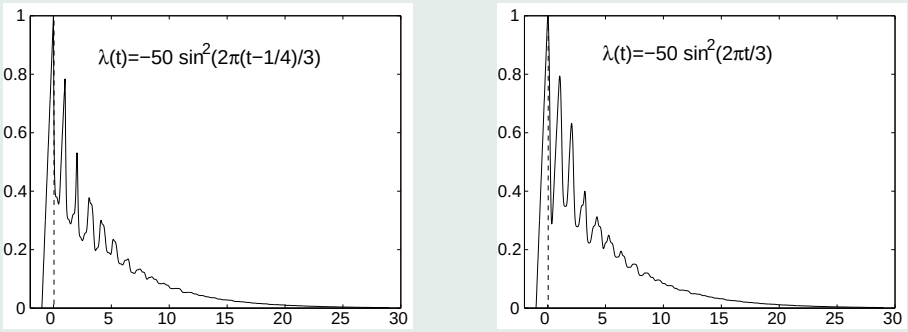

Figure 18: Solutions of equation (5.12) with some $\lambda(t) \leq 0$.

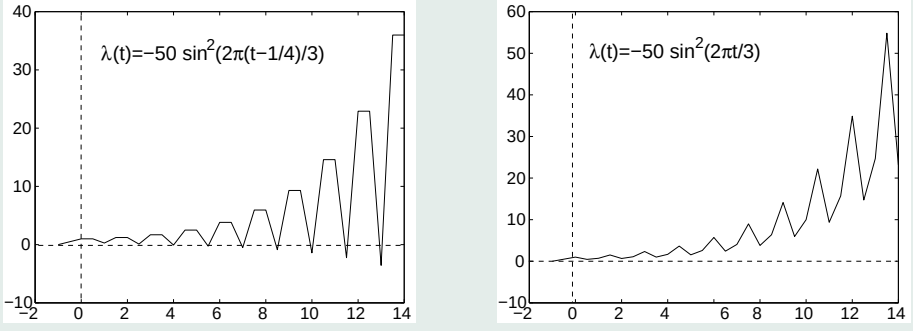

Figure 19: Numerical solutions of equation (5.12) with some $\lambda(t) \leq 0$ obtained by the midpoint rule (left) and the trapezoidal rule (right) with integer $m - \delta = 2$.

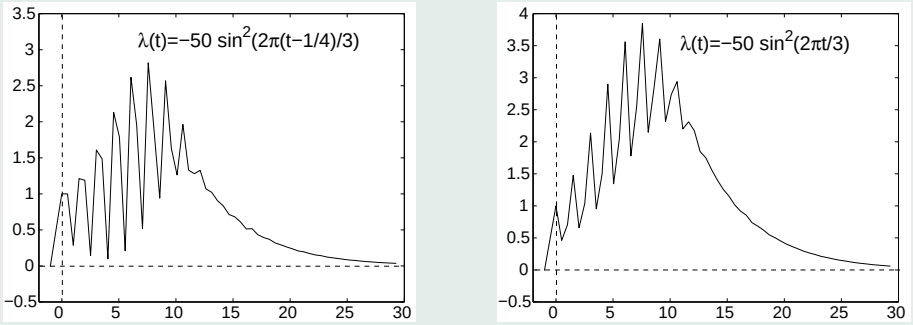

Figure 20: Numerical solutions of equation (5.12) with some $\lambda(t) \leq 0$ obtained by the midpoint rule (left) and the trapezoidal rule (right) with $m - \delta = 1.98$.

It is worth pointing out that the last two examples, designed to spoil the algebraic stability properties of the ODE methods as applied to DDEs, are pathological for the methods we considered. A slight modification of either the coefficient $\lambda(t)$ in (5.12) or the stepsize may lead to models for which the midpoint rule and the trapezoidal rule become stable. Numerical results given by the midpoint rule and the trapezoidal rule for the last two examples with stepsize $h = 0.505$, corresponding to $m - \delta = 1.98$, are depicted in Figure 20 where, after some spurious oscillations, they appear asymptotically stable.

5.3. A good method for DDEs

In order to get a second-order approximation for a class of stable problems including (5.12) which is stable for any stepsize, we may choose the two-stage *Lobatto IIIC* method with linear interpolation between nodal points.

For equation (5.12) with constant integration stepsize $h = 1/(m-\delta)$, m integer, $m \geq 2$, and $0 \leq \delta < 1$, the method reads

$$\begin{aligned} Y_1^n &= y_n + \tfrac{1}{2}h\big(\lambda(t_n)Y_1^n - \lambda(t_{n+1})Y_2^n - \tfrac{4}{5}\lambda(t_n)\eta(t_n-1) + \tfrac{4}{5}\lambda(t_{n+1})\eta(t_{n+1}-1)\big), \\ Y_2^n &= y_n + \tfrac{1}{2}h\big(\lambda(t_n)Y_1^n + \lambda(t_{n+1})Y_2^n - \tfrac{4}{5}\lambda(t_n)\eta(t_n-1) - \tfrac{4}{5}\lambda(t_{n+1})\eta(t_{n+1}-1)\big), \\ y_{n+1} &= Y_2^n, \end{aligned}$$

where

$$\left.\begin{aligned} \eta(t_n-1) &= t_n \\ \eta(t_{n+1}-1) &= t_{n+1} \end{aligned}\right\} \quad \text{for } n \leq m-2,$$

$$\left.\begin{aligned} \eta(t_n-1) &= t_n \\ \eta(t_{n+1}-1) &= 1-\delta+\delta y_1 \end{aligned}\right\} \quad \text{for } n = m-1,$$

$$\left.\begin{aligned} \eta(t_n-1) &= (1-\delta)y_{n-m} + \delta y_{n-m+1} \\ \eta(t_{n+1}-1) &= (1-\delta)y_{n-m+1} + \delta y_{n-m+2} \end{aligned}\right\} \quad \text{for } n \geq m.$$

For $\lambda(t) = -50\sin^2(\frac{2\pi}{3}(t-\frac{1}{4}))$, it yields stable solutions for any stepsize. In particular, for $h = 0.5$, where the midpoint rule failed, the solution remains stable (see Figure 21).

The superiority of the Lobatto IIIC method also with respect to the trapezoidal rule is illustrated in Figure 22, where the solution of equation (5.8) with $\lambda = -50$ is plotted with the same stepsize as in Figure 17. The spurious oscillations exibited by the trapezoidal rule near the corners have disappeared.

On the basis of the previous examples showing the failure of accurate and stable methods as applied to the simplest class

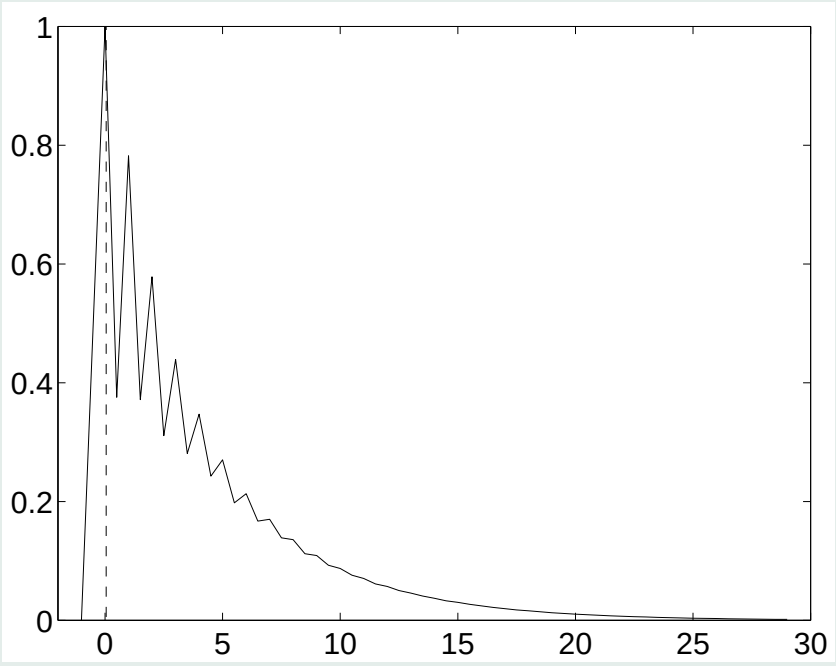

Figure 21: Numerical solution obtained by the two-stage Lobatto IIIC method for equation (5.12) with $\lambda(t) = -50sin^2(\frac{2\pi}{3}(t - \frac{1}{4}))$ and with $m - \delta = 2$.

of scalar linear equations with constant delay, we conclude this section by stressing that *integration of DDEs cannot be based on the mere adaptation of some standard ODE code to the presence of delayed terms. Integration of DDEs actually requires the use of specifically designed methods, according to the nature of the equation and the behavior of the solution.*

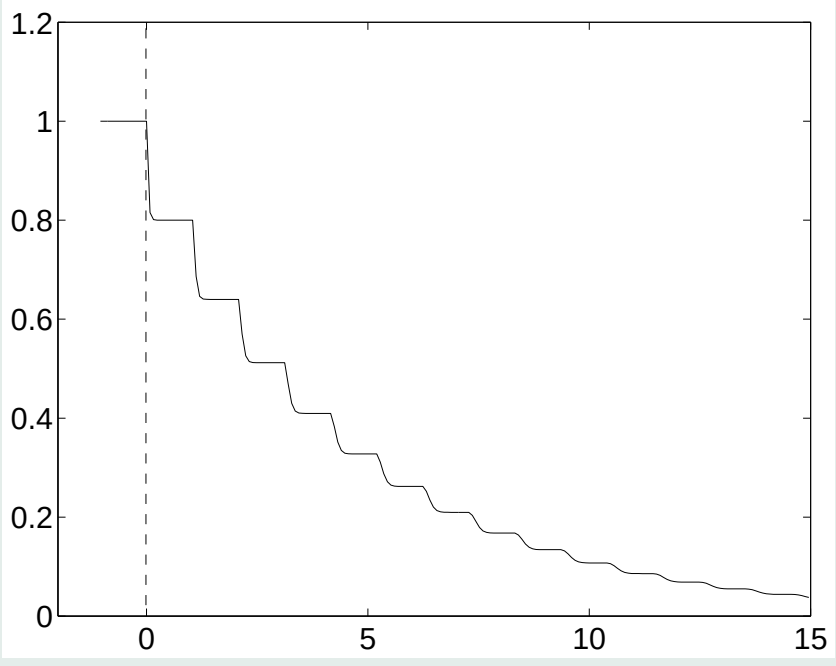

Figure 22: Numerical solution of equation (5.8) with $\lambda = -50$ by the two-stage Lobatto IIIC method with $m - \delta = 12.5$.

Summer School on Delay Differential Equations and Applications

Dobbiaco (BZ), Italy, June 26–30, 2006

The numerical solution of delay differential equations

M. Zennaro
Dipartimento di Matematica e Informatica
Università di Trieste

Lecture 3: Continuous Runge-Kutta methods

Main reference for this Lecture:
A. Bellen and M. Zennaro, Numerical Methods for Delay Differential Equations, Numerical Mathematics and Scientific Computation, Oxford Science Publications, Oxford University Press, 2003 (Chapter 5)

6. Continuous extensions of RK methods

We consider only one-step methods and, in particular, Runge–Kutta (RK) methods.

Given a mesh $\Delta = \{t_0, t_1, \ldots, t_n, \ldots, t_N = t_f\}$, a ν-stage RK method for the numerical solution of the ODE

$$\begin{cases} y'(t) = g\big(t, y(t)\big), & t_0 \le t \le t_f, \\ y(t_0) = y_0, \end{cases} \tag{6.1}$$

has the form (in the so-called Y *notation*)

$$Y^i_{n+1} = y_n + h_{n+1} \sum_{j=1}^{\nu} a_{ij} g(t^j_{n+1}, Y^j_{n+1}), \quad i = 1, \ldots, \nu, \tag{6.2}$$

$$y_{n+1} = y_n + h_{n+1} \sum_{i=1}^{\nu} b_i g(t^i_{n+1}, Y^i_{n+1}), \tag{6.3}$$

where $t^i_{n+1} = t_n + c_i h_{n+1}, c_i = \sum_{j=1}^{\nu} a_{ij}, i = 1, \ldots, \nu$, $h_{n+1} = t_{n+1} - t_n$ and ν is referred to as the number of *stages*. The b_i's are called *weights* of the quadrature formula (6.3) and the c_i's are called *abscissae* and, for most of common methods, they belong to $[0, 1]$. Since the RK method (6.2), (6.3) is characterized by the weights b_i and the matrix coefficients $A = [a_{ij}]_{i,j=1}^{\nu}$, it will be denoted by (A, b). It is worth observing that in many papers and books the RK formulae are written in an equivalent different form, the so-called K *notation*. So the RK method (6.2), (6.3) takes the form

$$K^i_{n+1} = g\Big(t^i_{n+1}, y_n + h_{n+1} \sum_{j=1}^{\nu} a_{ij} K^j_{n+1}\Big), \quad i = 1, \ldots, \nu,$$

$$y_{n+1} = y_n + h_{n+1} \sum_{i=1}^{\nu} b_i K^i_{n+1}.$$

Note that the K notation is obtained from formulae (6.2), (6.3) by setting

$$K^i_{n+1} = g(t^i_{n+1}, Y^i_{n+1}), \quad i = 1, \ldots, \nu.$$

Although in developing and implementing RK methods for ODEs the two notations are basically equivalent, in the application of RK methods to DDEs it will often be preferable to adopt the K notation.

The computational complexity of the method is mainly determined by the number of stages and by the form of the coefficient matrix A. It is well known that when the matrix A is lower triangular with zero diagonal elements, the method is called *explicit* and the computational cost is lower, whereas when the matrix A is full, the method is called *implicit* and the computational cost is higher.

The one-step interpolants of the RK method (6.2), (6.3) are constructed step by step by making use of information from the underlying mesh interval $[t_n, t_{n+1}]$ only, possibly by including some additional stages, that is by some extra evaluations of the right-hand-side function $g(t, y)$ in (6.1).

Interpolants constructed using no extra stages are called *interpolants of the first class* and the resulting continuous extension $\eta(t)$ is defined, in each subinterval of the mesh Δ, by a one-step continuous quadrature rule of the form

$$\eta(t_n + \theta h_{n+1}) = y_n + h_{n+1} \sum_{i=1}^{\nu} b_i(\theta) g(t_{n+1}^i, Y_{n+1}^i), \quad 0 \le \theta \le 1, \tag{6.4}$$

or, in the K notation,

$$\eta(t_n + \theta h_{n+1}) = y_n + h_{n+1} \sum_{i=1}^{\nu} b_i(\theta) K_{n+1}^i, \quad 0 \le \theta \le 1,$$

where the $b_i(\theta)$'s are polynomials of suitable degree $\le \delta$ satisfying

$$b_i(0) = 0 \quad \text{and} \quad b_i(1) = b_i, \quad i = 1, \ldots, \nu, \tag{6.5}$$

so as to define a continuous piecewise polynomial function.

Interpolants constructed by means of additional stages are called *interpolants of the second class* and the continuous extension is given by

$$\eta(t_n + \theta h_{n+1}) = y_n + h_{n+1} \sum_{i=1}^{s} b_i(\theta) g(t_{n+1}^i, Y_{n+1}^i), \quad 0 \le \theta \le 1, \tag{6.6}$$

or, in the K notation, by

$$\eta(t_n + \theta h_{n+1}) = y_n + h_{n+1} \sum_{i=1}^{s} b_i(\theta) K_{n+1}^i, \quad 0 \le \theta \le 1, \tag{6.7}$$

where the $b_i(\theta)$'s are again polynomials of suitable degree $\le \delta$ satisfying the continuity conditions

$$b_i(0) = 0, \quad i = 1, \ldots, s;$$

$$b_i(1) = b_i, \ \ i = 1, \ldots, \nu; \qquad b_i(1) = 0, \ \ i = \nu+1, \ldots, s. \tag{6.8}$$

The additional $s - \nu$ stages are given by

$$Y_{n+1}^i = y_n + h_{n+1} \sum_{j=1}^{s} a_{ij} g(t_{n+1}^j, Y_{n+1}^j), \quad i = \nu+1, \ldots, s, \tag{6.9}$$

or, in the K notation, by

$$K_{n+1}^i = g\Big(t_{n+1}^i, y_n + h_{n+1} \sum_{j=1}^{s} a_{ij} K_{n+1}^j\Big), \quad i = \nu+1, \ldots, s,$$

so that the original coefficient matrix $A = [a_{ij}]_{i,j=1}^{\nu}$ is embedded into the block lower triangular matrix

$$A' = \begin{bmatrix} A & 0 \\ [a_{ij}]_{i=\nu+1,j=1}^{s,\nu} & [a_{ij}]_{i,j=\nu+1}^{s} \end{bmatrix}. \tag{6.10}$$

The overall *continuous Runge–Kutta* (CRK) methods (6.2), (6.3), (6.4) and (6.2), (6.3), (6.9), (6.6), denoted by $\big(A, b(\theta)\big)$ and $\big(A', b(\theta)\big)$, respectively, are the continuous extensions of the RK method (A, b) and δ will be referred to as the *degree of the interpolant*. In contrast, the method (A, b) will be called the *underlying (discrete)* RK method.

It is worth remarking that, in general, $\eta(t_n + c_i h_{n+1}) \neq Y^i_{n+1}$. Nevertheless, equality holds for every right-hand side $g(t, y)$ whenever $b_i(c_j) = a_{ji}$, as appears evident by comparing (6.6) and (6.9). So we have

$$\{\eta(t_n + c_i h_{n+1}) = Y^i_{n+1} \quad \forall i\} \Longleftrightarrow \{b_i(c_j) = a_{ji} \quad \forall i, j\}. \tag{6.11}$$

An interpolant, either of the first or second class, determines a matrix B, whose elements are $b_{ij} = b_j(c_i)$.

Definition 6.1 *A CRK method is called* natural *if* $A = B$ $(A' = B)$.

Now consider the *increment functions* Φ and Ψ of the discrete and continuous RK methods, that are defined by

$$\Phi(t_n, y_n, h_{n+1}; g) = \sum_{i=1}^{\nu} b_i g(t^i_{n+1}, Y^i_{n+1})$$

and

$$\Psi(t_n, y_n, h_{n+1}, \theta; g) = \sum_{i=1}^{s} b_i(\theta) g(t^i_{n+1}, Y^i_{n+1}), \tag{6.12}$$

respectively.

****

Proposition 6.1 *For each Lipschitz continuous right-hand-side $g(t,y)$ in (6.1) there exist a stepsize $h_g > 0$ and a constant $Q_g > 0$, depending only on the Lipschitz constant L_g, such that the increment function Ψ (and thus also Φ) of an RK method (6.2), (6.3) and its interpolants (6.4) and (6.6) satisfies the Lipschitz condition*

$$\|\Psi(t_n, y_n, h_{n+1}, \theta; g) - \Psi(t_n, \tilde{y}_n, h_{n+1}, \theta; g)\| \leq Q_g \|y_n - \tilde{y}_n\|$$

for all $t_n \in [t_0, t_f]$, $\theta \in [0,1]$, $h_{n+1} \leq h_g$ and $y_n, \tilde{y}_n \in \mathbb{R}^d$.

****

****

Proposition 6.2 *For each Lipschitz continuous right-hand-side $g(t,y)$ in (6.1) there exist a stepsize $h_g > 0$ and a constant $\delta_g > 0$, depending only on the Lipschitz constant L_g, such that the increment function Ψ (and thus also Φ) of an RK method (6.2), (6.3) and its interpolants (6.4) and (6.6) satisfies the continuity condition*

$$\begin{aligned} &\|\Psi(t_n, y_n, h_{n+1}, \theta; \tilde{g}) - \Psi(t_n, y_n, h_{n+1}, \theta; g)\| \\ &\leq \delta_g \sup_{t_n \leq t \leq t_{n+1},\ y \in \mathbb{R}^d} \|\tilde{g}(t,y) - g(t,y)\| \end{aligned}$$

for all $t_n \in [t_0, t_f]$, $\theta \in [0,1]$ and $h_{n+1} \leq h_g$ and for any other right-hand-side function $\tilde{g}$.

****

We have the following definition for the class of RK methods and their interpolants.

Definition 6.2 *We say that the RK method (6.2), (6.3) is* consistent of order *(or, equivalently, has* order*) p if $p \geq 1$ is the largest integer such that, for all C^p-continuous right-hand-side functions $g(t, y)$ in (6.1) and for all mesh points, we have that*

$$\|z_{n+1}(t_{n+1}) - y_{n+1}\| = O(h_{n+1}^{p+1}),$$

uniformly with respect to y_n^* in any bounded subset of $\mathbb{R}^d$ and to $n = 0, \ldots, N-1$, where $z_{n+1}(t)$ is the local solution *to the* local problem

$$\begin{cases} z'_{n+1}(t) = g\big(t, z_{n+1}(t)\big), & t_n \leq t \leq t_{n+1}, \\ z_{n+1}(t_n) = y_n^*. \end{cases} \tag{6.13}$$

We say that the interpolant (6.4) or (6.6) is consistent of uniform order *(or, equivalently, has* uniform order*) q if $q \geq 1$ is the largest integer such that, for all C^q-continuous right-hand side functions $g(t, y)$ and for all mesh points, we have that*

$$\max_{t_n \leq t \leq t_{n+1}} \|z_{n+1}(t) - \eta(t)\| = O(h_{n+1}^{q+1}).$$

The convergence results are summarized by the following theorem.

****

Theorem 6.1 *Let the RK method (6.2), (6.3) be consistent of* order p *and let the right-hand-side function $g(t, y)$ in (6.1) be C^p-continuous. Then the method is convergent of order (or, equivalently, has global order) p on any bounded interval $[t_0, t_f]$, that is*

$$\max_{1\le n\le N} \|y(t_n) - y_n\| = O(h^p), \tag{6.14}$$

where $h = \max_{1\le n\le N} h_n$.

If the interpolant (6.4) or (6.6) has uniform order q, then the CRK method (6.2), (6.3), (6.4) or (6.2), (6.3), (6.9), (6.6) is uniformly convergent of order (or, equivalently, has uniform global order) $q' = \min\{p, q+1\}$; that is

$$\max_{t_0\le t\le t_f} \|y(t) - \eta(t)\| = O(h^{q'}). \tag{6.15}$$

****

We shall often refer to the order of consistency and to the order of convergence of the RK method (6.2), (6.3) as the *discrete order* and the *discrete global order* of the CRK method (6.2), (6.3), (6.4) or (6.2), (6.3), (6.9), (6.6).

The following theorem provides additional results on the derivatives of the continuous extension.

****

Theorem 6.2 *If, in addition to the hypotheses of Theorem 6.1, the interpolant is a piecewise polynomial of degree $\delta \geq q$ and the right-hand-side function $g(t, y)$ in (6.1) is $C^{\max\{\delta,p\}}$-continuous, then the following convergence, boundedness and unboundedness estimates hold for the derivatives of the global error function:*

$$\max_{t_0 \leq t \leq t_f} \|y^{(j)}(t) - \eta^{(j)}(t)\| = O(h^{q+1-j}), \quad j = 1, \ldots, \delta, \quad (6.16)$$

where the derivatives of $\eta(t)$ at the mesh points are meant in the left/right sense.

****

The estimates (6.15) and (6.16) show that the first derivative retains the global uniform order of the interpolant if and only if the interpolant has the maximum attainable uniform order p. It is also evident that, in order to get the uniform order q, the interpolant must be of degree $\delta \geq q$. On the other hand, polynomials of degree $\delta > q$ are unnecessary, as shown by the following theorem.

****

Theorem 6.3 *Assume that the RK method (6.2), (6.3) has a continuous extension $\eta(t)$ of order q and degree $d > q$. Then there exists another continuous extension $\tilde{\eta}(t)$ of order q whose degree is also q.*

****

Remark 6.1 *By Theorems 6.2 and 6.3 we may observe that, not only is the employment of interpolants of degree higher than q unnecessary, but interpolants of degree $\delta > q+1$ are even dangerous in that the derivatives of order k, with $q+2 \leq k \leq \delta$, may diverge as $h \to 0$. For these reasons we shall assume that continuous extensions of order q will be always made by interpolants of degree $\delta = q$.*

It is important to give an answer to the following two questions.

Question 1 What is the maximum uniform order an RK method of order p can achieve by means of an interpolant of the first class?

Question 2 What is the (mininum) number of stages necessary to construct a CRK method of uniform order $p-1$ or even p?

We shall consider both of them in the forthcoming sections. So far, we can give the following upper bound to the uniform order of an interpolant (for both classes).

****

Theorem 6.4 *Assume that the RK method (6.2), (6.3) has a continuous extension $\eta(t)$ given by (6.6). Then its uniform order q cannot exceed s^*, the number of distinct abscissae of the extended RK method represented by (6.10).*

****

The above result is obvious after observing that formula (6.6) is a continuous quadrature rule based exactly on s^* distinct abscissae.

Since the construction of interpolants of the second class is a quite technical matter, in this lecture we shall confine ourselves to analyze in some detail only the interpolants of the first class. However, we shall briefly consider the direct construction of continuous RK methods, without passing necessarily through interpolants of the second class of a given discrete RK formula.

Table 1: Order conditions for continuous RK methods.

Order	Conditions
1	$\sum_{i=1}^{\nu} b_i(\theta) = \theta$
2	$\sum_{i=1}^{\nu} b_i(\theta) c_i = \frac{1}{2}\theta^2$
3	$\sum_{i=1}^{\nu} b_i(\theta) c_i^2 = \frac{1}{3}\theta^3$ $\sum_{i,j=1}^{\nu} b_i(\theta) a_{ij} c_j = \frac{1}{6}\theta^3$
4	$\sum_{i=1}^{\nu} b_i(\theta) c_i^3 = \frac{1}{4}\theta^4$ $\sum_{i,j=1}^{\nu} b_i(\theta) c_i a_{ij} c_j = \frac{1}{8}\theta^4$ $\sum_{i,j=1}^{\nu} b_i(\theta) a_{ij} c_j^2 = \frac{1}{12}\theta^4$ $\sum_{i,j,k=1}^{\nu} b_i(\theta) a_{ij} a_{jk} c_k = \frac{1}{24}\theta^4$

7. Interpolants of the first class

A general analysis of the uniform order for the continuous extension (6.4) is based on the property that, for any $0 < \theta \leq 1$, it can be viewed as the discrete method $(\frac{A}{\theta}, \frac{b(\theta)}{\theta})$ with stepsize θh_{n+1}. So, we immediately get the uniform order conditions for the polynomials $b_i(\theta)$ from the well-known order conditions of the RK methods. The conditions up to order $p = 4$ are shown in Table 1.

In order to answer Question 1, each method has to be analyzed individually by checking the order conditions. In general, we can give only a partial answer by means of the following theorem, the proof of which does not directly involve the order conditions.

****

Theorem 7.1 *Every RK method (6.2), (6.3) of order $p \geq 1$ has a continuous extension $\eta(t)$ of order (and degree) $q = 1, \ldots, \lfloor \frac{p+1}{2} \rfloor$.*

****

****

Theorem 7.2 *If an RK method (6.2), (6.3) has a continuous extension $\eta(t)$ of order (and degree) $q \geq 2$, then it also has another continuous extension $\tilde{\eta}(t)$ of order (and degree) $\tilde{q}$ for each $\tilde{q} \leq q - 1$.*

****

In conclusion, we can answer Question 1 by saying that, in general, only interpolants up to order $\lfloor \frac{p+1}{2} \rfloor$ are assured to exist. On the other hand, it might well be that the maximum uniform order reachable by means of interpolants of the first class is actually $> \lfloor \frac{p+1}{2} \rfloor$ and, possibly, even $= p$.

Definition 7.1 *We say that an RK method (6.2), (6.3) of discrete order p is* superconvergent *if the maximum uniform order q reachable by means of interpolants of the first class is $\leq p - 1$.*

In other words, *superconvergence* is attained at the endpoint of the step-interval with respect to the maximum uniform accuracy order q. Of course, it might well be that the interpolant attains a higher order $p' > q$, not necessarily equal to the discrete order p, also at some additional points inside the step-interval. They will be called *inner superconvergence points*.

7.1. Collocation methods

A particular class of continuous RK methods that has been studied extensively are the *one-step collocation* methods. However, the interest in piecewise collocation is mostly due to the simplicity in determining the order of convergence and superconvergence via the non-linear variation-of-constants formula and to the optimal stability properties as discrete methods, rather than to its intrinsically continuous nature.

The one-step collocation method can be defined as follows. Chose ν distinct abscissae $c_1, \dots, c_\nu \in [0,1]$ and, in each mesh interval $[t_n, t_{n+1}]$, compute the polynomial $\eta(t)$ of degree $\leq \nu$ satisfying

$$\eta'(t_{n+1}^i) = g\big(t_{n+1}^i, \eta(t_{n+1}^i)\big), \quad i = 1, \dots, \nu, \quad \eta(t_n) = y_n.$$

It is easy to check that such methods can be rewritten as a continuous implicit RK method (6.2), (6.4), where

$$a_{ij} = \int_0^{c_i} \ell_j(\xi) \mathrm{d}\xi, \quad i, j = 1, \dots, \nu,$$

$$b_i(\theta) = \int_0^{\theta} \ell_i(\xi) \mathrm{d}\xi, \quad i = 1, \dots, \nu,$$

$\ell_i(\xi)$ being the Lagrange polynomial coefficient $\prod_{k=1, k\neq i}^{\nu} \frac{\xi - c_k}{c_i - c_k}$.

In particular, we have $b_i(c_j) = a_{ji}$ and, therefore, any collocation method is a natural CRK method.

For any choice of the abscissae $c_1, \dots, c_\nu \in [0,1]$, the collocation method has order $p \geq \nu$ and the uniform order of the interpolant (6.4) is $q = \nu$. Consequently, by Theorem 6.1, the collocation method is a continuous RK method of global uniform order $q' = \nu$ (if $p = \nu$) or $q' = \nu + 1$ (if $p > \nu$). In this sense the collocation method is optimal in that it achieves the maximum attainable uniform order for the given number

of stages. In particular, if the abscissae are the shifted roots of the Legendre orthogonal polynomial of degree ν, then the method has order $p = 2\nu$. This is the most famous example of superconvergence.

The following are some examples of superconvergent collocation methods.

- Gaussian methods: discrete order $p = 2\nu$, uniform order $q = \nu$, global uniform order $q' = \min\{p, q+1\}$;

$\nu = 1$ (*midpoint rule*): $p = 2, \quad q = 1, \quad q' = 2;$

$$\begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array} \qquad b_1(\theta) = \theta;$$

$\nu = 2$ (*Hammer–Hollingsworth* method): $p = 4, \quad q = 2, \quad q' = 3;$

$$\begin{array}{c|cc} \frac{1}{2} - \frac{\sqrt{3}}{6} & \frac{1}{4} & \frac{1}{4} - \frac{\sqrt{3}}{6} \\ \frac{1}{2} + \frac{\sqrt{3}}{6} & \frac{1}{4} + \frac{\sqrt{3}}{6} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \qquad \begin{aligned} b_1(\theta) &= -\tfrac{\sqrt{3}}{2}\theta(\theta - 1 - \tfrac{\sqrt{3}}{3}), \\ b_2(\theta) &= \tfrac{\sqrt{3}}{2}\theta(\theta - 1 + \tfrac{\sqrt{3}}{3}). \end{aligned}$$

- Radau IIA methods: discrete order $p = 2\nu - 1$, uniform order $q = \nu$, global uniform order $q' = \min\{p, q + 1\}$;

$\nu = 1$ (*backward Euler* method): $p = 1, \quad q = 1, \quad q' = 1;$

$$\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array} \qquad b_1(\theta) = \theta;$$

$\nu = 2$ (*Ehle* method): $p = 3, \quad q = 2, \quad q' = 3;$

$$\begin{array}{c|cc} \frac{1}{3} & \frac{5}{12} & -\frac{1}{12} \\ 1 & \frac{3}{4} & \frac{1}{4} \\ \hline & \frac{3}{4} & \frac{1}{4} \end{array} \qquad \begin{aligned} b_1(\theta) &= -\tfrac{3}{4}\theta(\theta - 2), \\ b_2(\theta) &= \tfrac{3}{4}\theta(\theta - \tfrac{2}{3}). \end{aligned}$$

- Lobatto IIIA methods: discrete order $p = 2\nu-2$, uniform order $q = \nu$, global uniform order $q' = \min\{p, q+1\}$;

$\nu = 2$ (*trapezoidal rule*): $p = 2, \quad q = 2, \quad q' = 2;$

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & \frac{1}{2} & \frac{1}{2} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \qquad \begin{aligned} b_1(\theta) &= -\tfrac{1}{2}\theta(\theta-2), \\ b_2(\theta) &= \tfrac{1}{2}\theta^2; \end{aligned}$$

$\nu = 3$ (*Ehle* method): $p = 4, \quad q = 3, \quad q' = 4;$

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{5}{24} & \frac{1}{3} & -\frac{1}{24} \\ 1 & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \\ \hline & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \end{array} \qquad \begin{aligned} b_1(\theta) &= 2\theta(\tfrac{1}{3}\theta^2 - \tfrac{3}{4}\theta + \tfrac{1}{2}), \\ b_2(\theta) &= -4\theta^2(\tfrac{1}{3}\theta - \tfrac{1}{2}), \\ b_3(\theta) &= 2\theta^2(\tfrac{1}{3}\theta - \tfrac{1}{4}). \end{aligned}$$

7.2. Natural continuous extensions

They are characterized by an additional *asymptotic orthogonality condition*, which allows preservation of the superconvergence of RK methods as applied to particular functional differential equations including DDEs and NDDEs.

Definition 7.2 *We say that the interpolant $\eta(t)$ in (6.4) of order (and degree) q is a* natural continuous extension *(NCE) of the RK method (6.2), (6.3) of order p if the polynomials $b_i(\theta), i = 1, \dots, \nu$, are such that $\eta(t)$ satisfies the additional asymptotic orthogonality condition*

$$\left\| \int_{t_n}^{t_{n+1}} G(t)[z'_{n+1}(t) - \eta'(t)]\mathrm{d}t \right\| = O(h_{n+1}^{p+1}) \qquad (7.1)$$

for every sufficiently smooth matrix-valued function G, uniformly with respect to $n = 0, \dots, N-1$, where $z_{n+1}(t)$ is the solution to the local problem (6.13).

It is easy to see that the NCEs necessarily also satisfy

$$\left\| \int_{t_n}^{t_{n+1}} G(t)[z_{n+1}(t) - \eta(t)]\mathrm{d}t \right\| = O(h_{n+1}^{p+1}) \qquad (7.2)$$

for every sufficiently smooth matrix-valued function G.

Moreover, by (7.1) and (7.2), we easily get

$$\max_{0\le n\le N-1} \left\| \int_{t_n}^{t_{n+1}} G(x)[y^{(j)}(x) - \eta^{(j)}(x)]\mathrm{d}x \right\| = O(h^{p+1}), \quad j = 0, 1, \qquad (7.3)$$

for every sufficiently smooth matrix-valued function G, where $h = \max_{1\le n\le N} h_n$.

With respect to the results of Theorems 6.1 and 6.2, condition (7.3) states the additional global convergence property of NCEs. It is worth remarking that, for any one-step collocation method, the collocation polynomial is an NCE of degree $q = \nu$.

With regard to the existence of NCEs and their attainable uniform order, the following theorems hold.

****

Theorem 7.3 *If the interpolant $\eta(t)$ in (6.4) of order (and degree) q is an NCE of the RK method (6.2), (6.3) of order p, then $q \geq \lfloor \frac{p+1}{2} \rfloor$.*

****

****

Theorem 7.4 *For every RK method (6.2), (6.3) of order p, there exists an NCE $\eta(t)$ of minimal order (and degree) $q = \lfloor \frac{p+1}{2} \rfloor$.*

****

Note that the NCE of an RK method is not unique. In general, for the construction of an NCE, one can use the order conditions of Table 1 along with the orthogonality condition (7.1).

Here are the NCEs of some explicit RK (ERK) methods.

- One-stage ERK method of order $p = 1$ (Euler method):
  - $q = \lfloor \frac{p+1}{2} \rfloor = \nu = p = 1$

$$b_1(\theta) = \theta.$$

- Two-stage ERK methods of order $p = 2$:
  - $q = \lfloor \frac{p+1}{2} \rfloor = 1$

$$b_i(\theta) = b_i\theta, \quad i = 1, 2;$$

  - $q = \nu = p = 2$

$$b_1(\theta) = (b_1 - 1)\theta^2 + \theta,$$
$$b_2(\theta) = b_2\theta^2.$$

- Three-stage ERK methods of order $p = 3$ (with $c_2,\ c_3 \neq 0$):

  * $q = \lfloor \frac{p+1}{2} \rfloor = 2$

    $$b_i(\theta) = w_i\theta^2 + (b_i - w_i)\theta, \quad i = 1, 2, 3,$$

    where

    $$w_1 = -\frac{1}{2c_3} - (c_3 - c_2)\lambda,$$
    $$w_2 = c_3\lambda,$$
    $$w_3 = \frac{1}{2c_3} - c_2\lambda,$$

    and $\lambda \in \mathbb{R}$. Here we have a one-parameter family of NCEs, that is an example of non-uniqueness of NCEs of minimal order q. In particular, we mention

    $$b_i(\theta) = 3(2c_i - 1)b_i\theta^2 + 2(2 - 3c_i)b_i\theta, \quad i = 1, 2, 3,$$

    which corresponds to the value $\lambda = \frac{3(2c_2-1)}{c_3}b_2$;

  * $q = \nu = p = 3$
    there are no NCEs of order $q = 3$.

- Four-stage ERK methods of order $p = 4$:

  * $q = \lfloor \frac{p+1}{2} \rfloor = 2$

    $$b_i(\theta) = 3(2c_i - 1)b_i\theta^2 + 2(2 - 3c_i)b_i\theta, \quad i = 1, 2, 3, 4;$$

  * $q = 3$

    $$b_1(\theta) = 2(1 - 4b_1)\theta^3 + 3(3b_1 - 1)\theta^2 + \theta,$$
    $$b_i(\theta) = 4(3c_i - 2)b_i\theta^3 + 3(3 - 4c_i)b_i\theta^2, \quad i = 2, 3, 4;$$

  * $q = \nu = p = 4$
    there are no NCEs of order $q = 4$.

Example 7.1 *Consider the test equation*

$$\begin{cases} y'(t) = -y(t), & t \geq t_0, \\ y(0) = 1, \end{cases} \tag{7.4}$$

and compute the approximate solution in the interval $[0, h]$ *using the following well-known ERK method of order* $p = 4$ *(the classical four-stage RK method), interpolated by its unique NCE of order* $q = 2$*:*

$$\begin{array}{c|cccc} 0 & & & & \\ \frac{1}{2} & \frac{1}{2} & & & \\ \frac{1}{2} & 0 & \frac{1}{2} & & \\ 1 & 0 & 0 & 1 & \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array} \qquad \begin{aligned} b_1(\theta) &= (-\tfrac{1}{2}\theta + \tfrac{2}{3})\theta, \\ b_2(\theta) &= \tfrac{1}{3}\theta, \\ b_3(\theta) &= \tfrac{1}{3}\theta, \\ b_4(\theta) &= (\tfrac{1}{2}\theta - \tfrac{1}{3})\theta. \end{aligned}$$

Since the integration is made on the interval $[0, h]$*, the expected nodal error* $|y_1 - y(h)|$ *is* $O(h^5)$ *and the uniform error* $\max_{0\leq\theta\leq 1} |\eta(\theta h) - y(\theta h)|$ *is* $O(h^3)$*. This means that the order conditions of Table 1 are fulfilled up to order 4 at* $\theta = 1$*, and up to order 2 for all* $\theta \in [0, 1]$*. On the other hand, it is easy to verify that, for* $\theta = \frac{1}{2}$*, the conditions of order 3 are fulfilled as well, and this is the sole inner superconvergence point, where the error is* $O(h^4)$*. Table 2 illustrates these occurrences.*

For the NCE of order $q = 3$ *given by*

$$\begin{aligned} b_1(\theta) &= (\tfrac{2}{3}\theta^2 - \tfrac{3}{2}\theta + 1)\theta, \\ b_2(\theta) &= (-\tfrac{2}{3}\theta + 1)\theta^2, \\ b_3(\theta) &= (-\tfrac{2}{3}\theta + 1)\theta^2, \\ b_4(\theta) &= (\tfrac{2}{3}\theta - \tfrac{1}{2})\theta^2, \end{aligned}$$

Table 2: Numerical results for the NCE of order $q = 2$ of the ERK method of order $p = 4$ applied to equation (7.4).

h	$\lvert y_1 - y(h)\rvert$	$\max_{0\le\theta\le1} \lvert\eta(\theta h) - y(\theta h)\rvert$	$\lvert\eta(\frac{1}{2}h) - y(\frac{1}{2}h)\rvert$
1	$9.95 \cdot 10^{-3}$	$2.12 \cdot 10^{-2}$	$1.33 \cdot 10^{-2}$
0.1	$8.47 \cdot 10^{-8}$	$9.25 \cdot 10^{-6}$	$1.30 \cdot 10^{-6}$
0.01	$8.35 \cdot 10^{-13}$	$8.14 \cdot 10^{-9}$	$1.30 \cdot 10^{-10}$

Table 3: Numerical results for the NCE of order $q = 3$ of the ERK method of order $p = 4$ applied to equation (7.4).

h	$\lvert y_1 - y(h)\rvert$	$\max_{0\le\theta\le1} \lvert\eta(\theta h) - y(\theta h)\rvert$
1	$9.95 \cdot 10^{-3}$	$1.57 \cdot 10^{-2}$
0.1	$8.47 \cdot 10^{-8}$	$1.46 \cdot 10^{-6}$
0.01	$8.35 \cdot 10^{-13}$	$1.45 \cdot 10^{-10}$

the uniform error is $O(h^4)$ and no inner superconvergence points exist. In fact, the first of the conditions of order 4 in Table 1 reads $\theta^2(\theta^2 - 2\theta + 1) = 0$, which is satisfied only for $\theta = 0$ and $\theta = 1$. Numerical experiments are shown in Table 3.

The continuous errors for $q = 2$ and $q = 3$ are plotted in Figure 23. $\Diamond$

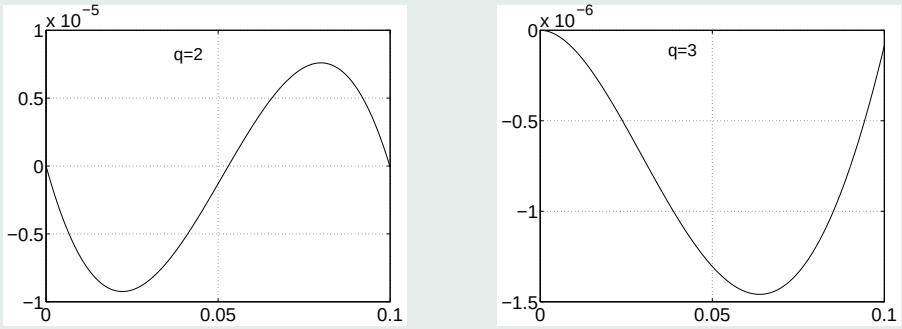

Figure 23: Continuous errors of the NCEs of order $q = 2, 3$ of the ERK method of order $p = 4$ applied to equation (7.4) in the interval $[0, 0.1]$.

Here are the NCEs of general ν-stage RK methods (also implicit) of order $p \geq \nu$ up to $\nu = 3$ with distinct abscissae $c_1, \ldots, c_\nu$.

- One-stage RK methods of order $p = 1, 2$:
  the discrete weight is

$$b_1 = 1;$$

  $* \; q = \lfloor \frac{p+1}{2} \rfloor = \nu = 1$

$$b_1(\theta) = \theta.$$

- Two-stage RK methods of order $p = 2, 3, 4$: the discrete weights are

$$b_1 = \frac{2c_2-1}{2(c_2-c_1)},$$
$$b_2 = \frac{1-2c_1}{2(c_2-c_1)};$$

  * $q = \lfloor \frac{p+1}{2} \rfloor = 1$ (only for $p = 2$)

$$b_1(\theta) = b_1\theta,$$
$$b_2(\theta) = b_2\theta;$$

  * $q = \nu = 2$

$$b_1(\theta) = -\frac{1}{2(c_2-c_1)}\theta^2 + \frac{c_2}{c_2-c_1}\theta,$$
$$b_2(\theta) = \frac{1}{2(c_2-c_1)}\theta^2 - \frac{c_1}{c_2-c_1}\theta.$$

- Three-stage RK methods of order $p = 3, 4, 5, 6$: the discrete weights are

$$b_1 = \frac{6c_2c_3-3(c_2+c_3)+2}{6(c_2-c_1)(c_3-c_1)},$$
$$b_2 = \frac{-6c_1c_3+3(c_1+c_3)-2}{6(c_2-c_1)(c_3-c_2)},$$
$$b_3 = \frac{6c_1c_2-3(c_1+c_2)+2}{6(c_3-c_1)(c_3-c_2)};$$

* $q = \lfloor \frac{p+1}{2} \rfloor = 2$ (only for $p = 3$)

$$b_i(\theta) = w_i\theta^2 + (b_i - w_i)\theta, \qquad i = 1, 2, 3,$$

where

$$\begin{aligned} w_1 &= -\tfrac{1}{2(c_3-c_1)} - (c_3 - c_2)\lambda, \\ w_2 &= (c_3 - c_1)\lambda, \\ w_3 &= \tfrac{1}{2(c_3-c_1)} - (c_2 - c_1)\lambda, \end{aligned}$$

and $\lambda \in \mathbb{R}$. Here we have a one-parameter family of NCEs.

For $p = 4$ there exists a unique NCE of order $q = 2$, that is

$$b_i(\theta) = 3(2c_i - 1)b_i\theta^2 + 2(2 - 3c_i)b_i\theta, \qquad i = 1, 2, 3,$$

which corresponds to the value $\lambda = \frac{(6c_1c_3 - 3(c_1+c_3)+2)(1-2c_2)}{2(c_3-c_1)(c_2-c_1)(c_3-c_2)}$;

* $q = \nu = 3$

$$\begin{aligned} b_1(\theta) &= \tfrac{1}{3(c_2-c_1)(c_3-c_1)}\theta^3 - \tfrac{c_2+c_3}{2(c_2-c_1)(c_3-c_1)}\theta^2 + \tfrac{c_2c_3}{(c_2-c_1)(c_3-c_1)}\theta, \\ b_2(\theta) &= -\tfrac{1}{3(c_2-c_1)(c_3-c_2)}\theta^3 + \tfrac{c_1+c_3}{2(c_2-c_1)(c_3-c_2)}\theta^2 - \tfrac{c_1c_3}{(c_2-c_1)(c_3-c_2)}\theta, \\ b_3(\theta) &= \tfrac{1}{3(c_3-c_1)(c_3-c_2)}\theta^3 - \tfrac{c_1+c_2}{2(c_2-c_1)(c_3-c_1)}\theta^2 + \tfrac{c_1c_2}{(c_3-c_1)(c_3-c_2)}\theta; \end{aligned}$$

it exists if and only if the RK method satisfies the simplifying assumptions

$$\sum_{j=1}^{3} a_{ij}c_j = \frac{1}{2}c_i^2, \qquad i = 1, 2, 3,$$

which always hold for $p = 5, 6$.

7.3. An application of the NCEs

Consider an interval $[a, b]$ and an integer q, $\lfloor \frac{p+1}{2} \rfloor \le q \le p$, and assume that, to any mesh $\Delta = \{t_0 = a, t_1, \dots, t_n, \dots, t_N = b\}$, two piecewise C^{p+1}-continuous functions $u_\Delta(t)$ and $v_\Delta(t)$ are associated which fulfill the following properties:

$$\max_{0\le n\le N} \|u_\Delta(t_n) - v_\Delta(t_n)\| = O(h^p),$$

$$\max_{a\le t\le b} \|u_\Delta(t) - v_\Delta(t)\| = O(h^{q'}), \tag{7.5}$$

where $q' = \min\{q+1, p\}$,

$$\max_{a\le t\le b} \|u_\Delta^{(j)}(t) - v_\Delta^{(j)}(t)\| = O(h^{q-j+1}), \quad j = 1, \dots, q, \tag{7.6}$$

$$\max_{0\le n\le N-1} \left\| \int_{t_n}^{t_{n+1}} G(t)[u_\Delta^{(j)}(t) - v_\Delta^{(j)}(t)]\mathrm{d}t \right\| = O(h^{p+1}), \quad j = 0, 1, \tag{7.7}$$

for every sufficiently smooth matrix-valued function G.

Moreover, assume that all the derivatives of $u_\Delta(t)$ and $v_\Delta(t)$ are uniformly bounded as $h \to 0$.

Then consider another interval $[\tilde{a}, \tilde{b}]$ and, for any mesh Δ in $[a, b]$, consider the following two IVPs for ODEs with the same initial condition but with the right-hand side depending on $u_\Delta(t)$ and $v_\Delta(t)$, respectively:

$$\begin{cases} z'(x) = f\Big(x, z(x), u_\Delta\big(\alpha(x)\big), u'_\Delta\big(\alpha(x)\big)\Big), & \tilde{a} \le x \le \tilde{b}, \\ z(\tilde{a}) = z_a, \end{cases} \tag{7.8}$$

$$\begin{cases} w'(x) = f\Big(x, w(x), v_\Delta\big(\alpha(x)\big), v'_\Delta\big(\alpha(x)\big)\Big), & \tilde{a} \le x \le \tilde{b}, \\ w(\tilde{a}) = z_a, \end{cases} \tag{7.9}$$

where $\alpha(x)$ is one-to-one between the intervals $[\tilde{a}, \tilde{b}]$ and $[a = \alpha(\tilde{a}), b = \alpha(\tilde{b})]$. Moreover, denote by $\beta(t)$ the inverse of $\alpha(x)$ and assume that the functions $f(x, z, y, w)$, $\alpha(x)$ and $\beta(t)$ are sufficiently smooth. Finally, let $\tilde{\Delta} = \{x_0 = \tilde{a}, x_1, \ldots, x_n, \ldots, x_N = \tilde{b}\}$ be the corresponding mesh of $[\tilde{a}, \tilde{b}]$ such that $x_n = \beta(t_n)$, $n = 0, 1, \ldots, N$.

Then we have the following result.

****

Theorem 7.5 *The solutions $z(x)$ and $w(x)$ of the ODEs (7.8) and (7.9) satisfy, in the interval $[\tilde{a}, \tilde{b}]$ and with respect to the mesh $\tilde{\Delta}$, the same properties satisfied by $u_\Delta(t)$ and $v_\Delta(t)$ in the interval $[a, b]$ with respect to the mesh Δ, that is:*

$$\max_{0 \le n \le N} \|z(x_n) - w(x_n)\| = O(h^p), \tag{7.10}$$

$$\max_{\tilde{a} \le x \le \tilde{b}} \|z(x) - w(x)\| = O(h^{q'}), \tag{7.11}$$

where $q' = \min\{q + 1, p\}$,

$$\max_{\tilde{a} \le x \le \tilde{b}} \|z^{(j)}(x) - w^{(j)}(x)\| = O(h^{q-j+1}), \quad j = 1, \ldots, q, \tag{7.12}$$

$$\max_{0 \le n \le N-1} \left\| \int_{x_n}^{x_{n+1}} G(x)[z^{(j)}(x) - w^{(j)}(x)]\mathrm{d}x \right\| = O(h^{p+1}), \quad j = 0, 1, \tag{7.13}$$

for every sufficiently smooth matrix-valued function G.

Moreover, all the derivatives of $z(x)$ and $w(x)$ are uniformly bounded as $h \to 0$.

****

Besides the forthcoming application to DDEs and NDDEs, Theorem 7.5 may be applied directly to the numerical solution of the ODE

$$\begin{cases} z'(x) = f\Big(x, z(x), y\big(\alpha(x)\big), y'\big(\alpha(x)\big)\Big), \quad x_0 \le x \le x_f, \\ z(x_0) = z_0, \end{cases} \tag{7.14}$$

where $y(t)$ is the solution of the *driving equation*

$$\begin{cases} y'(t) = g\big(t, y(t)\big), \quad t_0 \le t \le t_f, \\ y(t_0) = y_0. \end{cases} \tag{7.15}$$

In order to solve (7.14) numerically, we first solve (7.15) on a given mesh $\Delta = \{t_0, t_1, \ldots, t_n, \ldots, t_N = t_f\}$ by an RK method of order p with an NCE $\eta(t)$ of order q. Then we solve the perturbed ODE

$$\begin{cases} w'(x) = f\Big(x, w(x), \eta\big(\alpha(x)\big), \eta'\big(\alpha(x)\big)\Big), \quad x_0 \le x \le x_f, \\ w(x_0) = z_0. \end{cases} \tag{7.16}$$

Since the NCE $\eta(t)$ satisfies conditions (6.15), (6.16) and (7.3), Theorem 7.5 applies with $[a, b] = [t_0, t_f]$, $[\tilde{a}, \tilde{b}] = [x_0, x_f]$, $u_\Delta(t) = y(t)$ and $v_\Delta(t) = \eta(t)$. In particular, (7.10) holds for the points of the corresponding mesh $\tilde{\Delta}$. Therefore, for the numerical solution of (7.16), any discrete RK method of order p across the mesh $\tilde{\Delta}$ preserves the nodal order p, despite the possibly lower uniform accuracy order q of the employed NCE $\eta(t)$.

8. Direct construction of continuous RK methods

So far we have considered continuous extensions of a priori given discrete RK methods. In this section we present the other philosophy of constructing directly a CRK method, without necessarily starting from a given discrete formula.

As pointed out in Section 7, a general analysis of the uniform order for the continuous extensions (6.4) is based on the property that, for any $0 < \theta \leq 1$, it can be viewed as a discrete method $(\frac{A}{\theta}, \frac{b(\theta)}{\theta})$ with scaled stepsize θh_{n+1}. So we immediately get the uniform order conditions for the parameters c_i and a_{ij} and for the polynomials $b_i(\theta)$ from the well-known order conditions of the RK methods. The conditions up to order $p = 4$ have already been given in Table 1.

Let $N(p)$ and $CN(q)$ be the minimum number of stages for which there exist RK methods of (discrete) order p and CRK methods of uniform order q, respectively. Similarly, let $EN(p)$ and $CEN(q)$ be the same quantities restricted to the class of ERK methods and continuous ERK (CERK) methods.

In the general case, it is well known that

$$N(p) = \left\lfloor \frac{p+1}{2} \right\rfloor \quad \text{and} \quad CN(q) = q$$

and that these optimal bounds are attained, for instance, by collocation methods (see Section 7.1).

For ERK and CERK methods the results are often obtained by making somewhat sophisticated analyses of the continuous order conditions.

All the order barriers for explicit methods are summarized in Tables 4 and 5.

Table 4: The minimum number of stages necessary for an ERK method to attain the discrete order p.

p	$EN(p)$
1	1
2	2
3	3
4	4
5	6
6	7
7	9
8	11
$r \geq 9$	$\geq r+3$

Now we concentrate on CERK methods with a minimum number of stages $CEN(q)$. It easily turns out that, for given $q \geq 2$, a whole family of such methods exists, which depends on a certain number of parameters. So the parameters can be selected in order to guarantee some nice properties of the method, such as minimization of a suitable estimate of the local error constant and maximization of the absolute stability region of the underlying discrete method.

Another nice characteristic of some CERK methods is the FSAL (*first same as last*) property. The FSAL property means that the last stage can be reused as the first stage $K^1_{n+1} = g(t_{n+1}, y_{n+1})$ of the next step. This implies that the actual cost of the method is reduced by one function evaluation per step. Of course, because the method is explicit, the reusable stage can be involved only for computation of the interpolant $\eta(t_n + \theta h_{n+1})$ for $\theta \neq 1$ and not for computation of $y_{n+1} =$

Table 5: The minimum number of stages necessary for a CERK method to attain the uniform order q.

q	$CEN(q)$
1	1
2	2
3	4
4	6
5	8
6	11
$r \geq 7$	$\geq 2r-2$

$\eta(t_{n+1})$.

In Tables 6, 7 and 8 we report three examples of CERK methods with a minimum number of stages $CEN(q)$ for uniform orders $q = 3, 4$ and 5, respectively. Such methods satisfy the FSAL property and minimize a suitable measure of the local error constant.

Table 6: Butcher's tableau and continuous weights of a CERK method of order 3 with four stages.

0				
$\frac{12}{23}$	$\frac{12}{23}$			
$\frac{4}{5}$	$-\frac{68}{375}$	$\frac{368}{375}$		
1	$\frac{31}{144}$	$\frac{529}{1152}$	$\frac{125}{384}$	
	$b_1(\theta)$	$b_2(\theta)$	$b_3(\theta)$	$b_4(\theta)$

$$
\begin{aligned}
b_1(\theta) &= \tfrac{41}{72}\theta^3 - \tfrac{65}{48}\theta^2 + \theta,\\
b_2(\theta) &= -\tfrac{529}{576}\theta^3 + \tfrac{529}{384}\theta^2,\\
b_3(\theta) &= -\tfrac{125}{192}\theta^3 + \tfrac{125}{128}\theta^2,\\
b_4(\theta) &= \theta^3 - \theta^2.
\end{aligned}
$$

Table 7: Butcher's tableau and continuous weights of a CERK method of order 4 with six stages.

0						
$\frac{1}{6}$	$\frac{1}{6}$					
$\frac{11}{37}$	$\frac{44}{1369}$	$\frac{363}{1369}$				
$\frac{11}{17}$	$\frac{3388}{4913}$	$-\frac{8349}{4913}$	$\frac{8140}{4913}$			
$\frac{13}{15}$	$-\frac{36764}{408375}$	$\frac{767}{1125}$	$-\frac{32708}{136125}$	$\frac{210392}{408375}$		
1	$\frac{1697}{18876}$	0	$\frac{50653}{116160}$	$\frac{299693}{1626240}$	$\frac{3375}{11648}$	
	$b_1(\theta)$	$b_2(\theta)$	$b_3(\theta)$	$b_4(\theta)$	$b_5(\theta)$	$b_6(\theta)$

$$
\begin{aligned}
b_1(\theta) &= -\tfrac{866577}{824252}\theta^4 + \tfrac{1806901}{618189}\theta^3 - \tfrac{104217}{37466}\theta^2 + \theta,\\
b_2(\theta) &= 0,\\
b_3(\theta) &= \tfrac{12308679}{5072320}\theta^4 - \tfrac{2178079}{380424}\theta^3 + \tfrac{861101}{230560}\theta^2,\\
b_4(\theta) &= -\tfrac{7816583}{10144640}\theta^4 + \tfrac{6244423}{5325936}\theta^3 - \tfrac{63869}{293440}\theta^2,\\
b_5(\theta) &= -\tfrac{624375}{217984}\theta^4 + \tfrac{982125}{190736}\theta^3 - \tfrac{1522125}{762944}\theta^2,\\
b_6(\theta) &= \tfrac{296}{131}\theta^4 - \tfrac{461}{131}\theta^3 + \tfrac{165}{131}\theta^2.
\end{aligned}
$$

Table 8: Butcher's tableau and continuous weights of a CERK method of order 5 with eight stages.

0								
$\frac{1}{6}$	$\frac{1}{6}$							
$\frac{1}{4}$	$\frac{1}{16}$	$\frac{3}{16}$						
$\frac{1}{2}$	$\frac{1}{4}$	$-\frac{3}{4}$	1					
$\frac{1}{2}$	$-\frac{3}{4}$	$\frac{15}{4}$	-3	$\frac{1}{2}$				
$\frac{9}{14}$	$\frac{369}{1372}$	$-\frac{243}{343}$	$\frac{297}{343}$	$\frac{1485}{9604}$	$\frac{297}{4802}$			
$\frac{7}{8}$	$-\frac{133}{4512}$	$\frac{1113}{6016}$	$\frac{7945}{16544}$	$-\frac{12845}{24064}$	$-\frac{315}{24064}$	$\frac{156065}{198528}$		
1	$\frac{83}{945}$	0	$\frac{248}{825}$	$\frac{41}{180}$	$\frac{1}{36}$	$\frac{2401}{38610}$	$\frac{6016}{20475}$	
	$b_1(\theta)$	$b_2(\theta)$	$b_3(\theta)$	$b_4(\theta)$	$b_5(\theta)$	$b_6(\theta)$	$b_7(\theta)$	$b_8(\theta)$

$$
\begin{aligned}
b_1(\theta) &= \tfrac{596}{315}\theta^5 - \tfrac{4969}{819}\theta^4 + \tfrac{17893}{2457}\theta^3 - \tfrac{3292}{819}\theta^2 + \theta,\\
b_2(\theta) &= 0,\\
b_3(\theta) &= -\tfrac{1984}{275}\theta^5 + \tfrac{1344}{65}\theta^4 - \tfrac{43568}{2145}\theta^3 + \tfrac{5112}{715}\theta^2,\\
b_4(\theta) &= \tfrac{118}{15}\theta^5 - \tfrac{1465}{78}\theta^4 + \tfrac{3161}{234}\theta^3 - \tfrac{123}{52}\theta^2,\\
b_5(\theta) &= 2\theta^5 - \tfrac{413}{78}\theta^4 + \tfrac{1061}{234}\theta^3 - \tfrac{63}{52}\theta^2,\\
b_6(\theta) &= -\tfrac{9604}{6435}\theta^5 + \tfrac{2401}{1521}\theta^4 + \tfrac{60025}{50193}\theta^3 - \tfrac{40817}{33462}\theta^2,\\
b_7(\theta) &= -\tfrac{48128}{6825}\theta^5 + \tfrac{96256}{5915}\theta^4 - \tfrac{637696}{53235}\theta^3 + \tfrac{18048}{5915}\theta^2,\\
b_8(\theta) &= 4\theta^5 - \tfrac{109}{13}\theta^4 + \tfrac{75}{13}\theta^3 - \tfrac{18}{13}\theta^2.
\end{aligned}
$$

Summer School on Delay Differential Equations and Applications

Dobbiaco (BZ), Italy, June 26–30, 2006

The numerical solution of delay differential equations

M. Zennaro
Dipartimento di Matematica e Informatica
Università di Trieste

Lecture 4: The standard approach for delay differential equations using continuous Runge-Kutta methods

Main reference for this Lecture:
A. Bellen and M. Zennaro, Numerical Methods for Delay Differential Equations, Numerical Mathematics and Scientific Computation, Oxford Science Publications, Oxford University Press, 2003 (Chapters 3, 4 and 6)

9. The standard approach in the general delay case

We define and develop the standard approach based on the class of CRK methods discussed in Lecture 3.

The standard approach for solving the DDE

$$\begin{cases} y'(t) = f\Big(t, y(t), y\big(t - \tau(t, y(t))\big)\Big), & t_0 \le t \le t_f, \\ y(t) = \phi(t), & t \le t_0, \end{cases} \tag{9.1}$$

consists in solving step by step the local problems

$$\begin{cases} w'_{n+1}(t) = f\Big(t, w_{n+1}(t), x\big(t - \tau(t, w_{n+1}(t))\big)\Big), & t_n \le t \le t_{n+1}, \\ w_{n+1}(t_n) = y_n, \end{cases} \tag{9.2}$$

where

$$x(s) = \begin{cases} \phi(s) & \text{for } s \le t_0, \\ \eta(s) & \text{for } t_0 \le s \le t_n, \\ w_{n+1}(s) & \text{for } t_n \le s \le t_{n+1}, \end{cases}$$

and $\eta(s)$ is the continuous approximate solution computed by the method itself up to t_n. The continuous RK method $\big(A, b(\theta)\big)$ being given by (6.2) and (6.4) or, for interpolants based on extra stages, by (6.2), (6.6) and (6.9), the overall method for DDEs turns out to be

$$Y^i_{n+1} = y_n + h_{n+1} \sum_{j=1}^{s} a_{ij} f\Big(t^j_{n+1}, Y^j_{n+1}, \eta\big(t^j_{n+1} - \tau(t^j_{n+1}, Y^j_{n+1})\big)\Big), \quad i = 1, \ldots, s, \tag{9.3}$$

$$\eta(t_n + \theta h_{n+1}) = y_n + h_{n+1} \sum_{i=1}^{s} b_i(\theta) f\Big(t^i_{n+1}, Y^i_{n+1}, \eta\big(t^i_{n+1} - \tau(t^i_{n+1}, Y^i_{n+1})\big)\Big),$$

$$0 \le \theta \le 1. \tag{9.4}$$

This method will be called the *RK method for DDEs* or, in short, the *DDE method*. Moreover, we shall say that $\big(A, b\big)$ is the *underlying discrete RK method*, whereas $\big(A, b(\theta)\big)$ is the *underlying interpolant*. The pair formed by the underlying discrete RK method and by the underlying interpolant is called the *underlying continuous RK method*.

For simplicity, we have put both formula (6.2) for the stages of the discrete RK method and formula (6.9) for the possible additional stages of the interpolant together in (9.3).

Note that the use of RK methods with an abscissa $c_i > 1$ could lead to an advanced deviated argument $t^i_{n+1} - \tau(t^i_{n+1}, Y^i_{n+1}) > t_{n+1}$, where the continuous extension $x(s)$ should be computed in some subsequent step. Therefore, in order to avoid such a disappointing situation, we make the following assumption.

Assumption 9.1 *The RK method for DDEs (9.3), (9.4) is such that the abscissae satisfy the constraint* $0 \le c_i \le 1$, $i = 1, \ldots, s$.

However, even under Assumption 9.1, it may well be that, for some index i, the argument $t^i_{n+1} - \tau(t^i_{n+1}, Y^i_{n+1})$ of $\eta(s)$ lies in the current interval $[t_n, t_{n+1}]$. We shall call this occurence *overlapping*. It is convenient to define the *spurious stage*

$$\tilde{Y}^i_{n+1} = \eta(t^i_{n+1} - \tau(t^i_{n+1}, Y^i_{n+1}))$$

which, in case of overlapping, is given by formula (9.4) itself for

$$\theta = \theta^i_{n+1} = c_i - \frac{\tau(t^i_{n+1}, Y^i_{n+1})}{h_{n+1}}.$$

It is worth remarking that the overall method becomes implicit even if the underlying CRK method is explicit.

On the contrary, if overlapping does not occur, the spurious stage is simply given by the interpolant $\eta(t)$ as computed in the past.

In any case, in the mesh interval $[t_n, t_{n+1}]$ the method takes the form (in the Y notation)

$$\eta(t_n+\theta h_{n+1}) = y_n+h_{n+1}\sum_{i=1}^{s} b_i(\theta) f(t^i_{n+1}, Y^i_{n+1}, \tilde{Y}^i_{n+1}), \quad 0 \le \theta \le 1, \tag{9.5}$$

$$Y^i_{n+1} = y_n + h_{n+1}\sum_{j=1}^{s} a_{ij} f(t^j_{n+1}, Y^j_{n+1}, \tilde{Y}^j_{n+1}), \quad i = 1, \dots, s, \tag{9.6}$$

where the spurious stages $\tilde{Y}^i_{n+1}$ are given by

$$\tilde{Y}^i_{n+1} = y_n + h_{n+1}\sum_{j=1}^{s} b_j(\theta^i_{n+1}) f(t^j_{n+1}, Y^j_{n+1}, \tilde{Y}^j_{n+1}) \tag{9.7}$$

if $t^i_{n+1} - \tau(t^i_{n+1}, Y^i_{n+1}) > t_n$, and by

$$\tilde{Y}^i_{n+1} = \eta\big(t^i_{n+1} - \tau(t^i_{n+1}, Y^i_{n+1})\big) \tag{9.8}$$

otherwise.

Note that, whereas the system (9.5), (9.6), (9.8) has to be solved only for the stage values Y^j_{n+1}, $j = 1, \dots, s$, the system enlarged by (9.7) for some i has to be solved also for the relevant spurious stages $\tilde{Y}^i_{n+1}$.

Indeed, the dimension of the system is not increased. In fact, by using the K notation

$$K^i_{n+1} = f(t^i_{n+1}, Y^i_{n+1}, \tilde{Y}^i_{n+1}),$$

we get the following system to be solved for K^i_{n+1}, $i = 1, \ldots, s$:

$$\eta(t_n + \theta h_{n+1}) = y_n + h_{n+1} \sum_{i=1}^{s} b_i(\theta) K^i_{n+1}, \quad 0 \le \theta \le 1, \quad (9.9)$$

$$K^i_{n+1} = f\left(t^i_{n+1}, y_n + h_{n+1} \sum_{j=1}^{s} a_{ij} K^j_{n+1}, \tilde{Y}^i_{n+1}\right), \quad i = 1, \ldots, s, \quad (9.10)$$

where

$$\tilde{Y}^i_{n+1} = y_n + h_{n+1} \sum_{j=1}^{s} b_j \left(c_i - \frac{\tau(t^i_{n+1}, y_n + h_{n+1} \sum_{k=1}^{s} a_{ik} K^k_{n+1})}{h_{n+1}} \right) K^j_{n+1} \quad (9.11)$$

if $t^i_{n+1} - \tau(t^i_{n+1}, y_n + h_{n+1} \sum_{k=1}^{s} a_{ik} K^k_{n+1}) > t_n$, and

$$\tilde{Y}^i_{n+1} = \eta\left(t^i_{n+1} - \tau(t^i_{n+1}, y_n + h_{n+1} \sum_{k=1}^{s} a_{ik} K^k_{n+1}) \right) \quad (9.12)$$

otherwise.

From a practical point of view, it would be important to answer a priori the following two questions:

(Q_1) Whether or not, for the local problem (9.2), overlapping occurs and, in particular, whether or not the approximated delayed function $\eta(t-\tau)$ is known at a given point t of the current interval $[t_n, t_{n+1}]$.

(Q_2) Whether or not, for the stepsize h_{n+1}, the current interval $[t_n, t_{n+1}]$ includes some discontinuity point ξ and, more specifically, how to tune the stepsize h_{n+1} in order to have $t_{n+1} = \xi$, as we would like to have in the proximity of ξ for accuracy reasons (see the forthcoming convergence theorem).

As far as question (Q_1) is concerned, overlapping can be avoided for a sufficiently small stepsize by assuming the hypothesis (H_1) or, for state dependent delays, the more restrictive form (H_1^*), introduced in Lecture 1.

Indeed, (H_1^*) prevents the state dependent delay from vanishing even if it is computed in a perturbation of the true local solution $w_{n+1}(t)$. Moreover, observe that (H_1^*) is equivalent to (H_1) whenever the delay is not state dependent.

Under the hypothesis (H_1^*) on the delay, for sufficiently small stepsize, namely $h_{n+1} = t_{n+1} - t_n \leq \tau_0$, overlapping does not occur and the function $\eta(s)$ is known for every $s = t - \tau(t, z)$ with $t \in [t_n, t_{n+1}]$ and for all $z \in \mathbb{R}^d$. In fact,

$$t - t_n \leq \tau_0 \leq \tau(t, z)$$

for all $t \in [t_n, t_{n+1}]$ and, hence,

$$t - \tau(t, z) \leq t_n.$$

Thus overlapping is avoided for any approximation of the local solution $w_{n+1}(t)$.

On the other hand, when $\tau_0 < h_0$, the minimum stepsize the method would like to use, it might well be that overlapping occurs. In fact, in such a case, for some n there might exist $t \in [t_n, t_{n+1}]$ such that

$$t - t_n > \tau_0 \quad \text{and} \quad \tau(t, w_{n+1}(t)) = \tau_0,$$

which yield

$$t - \tau(t, w_{n+1}(t)) > t_n.$$

When the hypothesis (H_1), or (H_1^*), does not hold, the delay τ necessarily vanishes at some point ξ and thus overlapping inevitably occurs whenever the integration interval $[t_n, t_{n+1}]$ includes ξ.

As for question (Q_2), which is particularly hard for state dependent delays, the location of discontinuity points has been discussed by various authors from both the theoretical and implementational points of view. Two approaches have been pursued in the literature.

The first, usually referred to as the *tracking of discontinuities*, is based on finding the discontinuities $\xi_{k,j}$ satisfying

$$\xi_{k,j} - \tau\big(\xi_{k,j}, w_{n+1}(\xi_{k,j})\big) = \xi_{k-1,i} \quad \text{for some } i$$

(see (3.9)) and to include them as mesh points.

An alternative approach that relies on stepsize control gives up tracking the discontinuities, which are instead assumed to be automatically included in the mesh by suitable variable stepsize strategies based on the estimation of the local error or on the computation of the defect. In general, the codes are simpler but undergo a larger number of rejected steps and may lead to a sequence of very small stepsizes in the neighborhood of a low order discontinuity ξ, as one easily guesses if the method repeatedly fails in placing the end point of the current interval at the discontinuity ξ.

Most of the methods known in the literature and the relevant existing software for DDEs (and NDDEs) can be framed within the standard approach.

Despite it not being possible to express all RK methods for DDEs in terms of the stage values Y^i_{n+1} only, there are particular classes, essentially collocation methods (see Section 7.1), that allow us to express the spurious stages $\tilde{Y}^i_{n+1}$ in the system (9.6) in terms of the Y^i_{n+1}. This is the case for any *natural* CRK method (see Definition 6.1) with s distinct abscissae $c_1, \ldots, c_s$ such that $c_i \neq 0$, $i = 1, \ldots, s$, and a continuous extension $\eta(t_n + \theta h_{n+1})$ of degree s. In fact, in this case the polynomial $\eta(t)$ may be written using the Lagrange interpolation formula through the $s+1$ values $y_n (= \eta(t_n))$ and $Y^i_{n+1} (= \eta(t_n + c_i h_{n+1}))$, $i = 1, \ldots, s$, that is

$$\eta(t_n + \theta h_{n+1}) = \ell_0(\theta) y_n + \sum_{i=1}^{s} \ell_i(\theta) Y^i_{n+1}, \tag{9.13}$$

where ℓ_j, $j = 0, \ldots, s$ are the Lagrange polynomial coefficients relevant to the nodes $c_0 = 0$ and c_i, $i = 1, \ldots, s$. Therefore $\tilde{Y}^i_{n+1}$, which is equal to $\eta\big(t^i_{n+1} - \tau(t^i_{n+1}, Y^i_{n+1})\big)$, may be written using (9.13) for $\theta = \theta^i_{n+1} = c_i - \frac{\tau(t^i_{n+1}, Y^i_{n+1})}{h_{n+1}}$. The Gaussian collocation and Radau IIA methods satisfy the above condition and are natural choices for the construction of DDE methods.

For both the Y and K notations, the method is well-posed for any sufficiently small h_{n+1}, as stated by the following theorem.

****

Theorem 9.1 (Well-posedness) *Assume that the local problem (9.2) possesses a unique solution $w_{n+1}(t)$. Then, for sufficiently small stepsize h_{n+1}, equations (9.3)–(9.4) admit a unique solution $\eta(t)$.*

****

As for the convergence analysis of the DDE methods, we have the following result. However, in connection with the previous question (Q_2) about the location of discontinuity points, we assume to be able to compute and include them as mesh points, even in the state dependent delay case.

****

Theorem 9.2 (Convergence) *Consider the DDE*

$$\begin{cases} y'(t) = f\Big(t, y(t), y\big(t - \tau(t, y(t))\big)\Big), & t_0 \le t \le t_f, \\ y(t) = \phi(t), & t \le t_0, \end{cases} \tag{9.14}$$

where $f(t, y, x)$ is C^p-continuous in $[t_0, t_f] \times \mathbb{R}^d \times \mathbb{R}^d$, the initial function $\phi(t)$ is C^p-continuous and the delay $\tau(t, y)$ is C^p-continuous in $[t_0, t_f] \times \mathbb{R}^d$. Moreover, assume that the mesh $\Delta = \{t_0, t_1, \ldots, t_n, \ldots, t_N = t_f\}$ includes all the discontinuity points of order $\le p$ lying in $[t_0, t_f]$. If the underlying CRK method has discrete order p and uniform order q, then the DDE method (9.5), (9.6), (9.7), (9.8) has discrete global order and uniform global order $q' = \min\{p, q + 1\}$; that is

$$\max_{1 \le n \le N} \|y(t_n) - y_n\| = O(h^{q'})$$

and

$$\max_{t_0 \le t \le t_f} \|y(t) - \eta(t)\| = O(h^{q'}),$$

where $h = \max_{1 \le n \le N} h_n$.

****

The proof of the theorem is based on the following lemma.

****

Lemma 9.1 *Let $\zeta(t)$ and $\eta(t)$ be the numerical solutions of the initial value problems*

$$\begin{cases} z'_{n+1}(t) = f\Big(t, z_{n+1}(t), u\big(t - \tau(t, z_{n+1}(t))\big)\Big), & t_n \le t \le t_{n+1}, \\ z_{n+1}(t_n) = z_n, \end{cases}$$

and

$$\begin{cases} w'_{n+1}(t) = f\Big(t, w_{n+1}(t), v\big(t - \tau(t, w_{n+1}(t))\big)\Big), & t_n \le t \le t_{n+1}, \\ w_{n+1}(t_n) = w_n, \end{cases}$$

respectively, obtained by the DDE method (9.5), (9.6), (9.7), (9.8).

Then there exist a stepsize $h_u > 0$ and constants $P_u > 0$ and $Q_u > 0$ such that, for $h_{n+1} \le h_u$,

$$\begin{aligned} \max_{t_n \le t \le t_{n+1}} \|\zeta(t) - \eta(t)\| \;\le\; & (1 + h_{n+1}Q_u)\|z_n - w_n\| \\ & + h_{n+1}P_u \max_{t \le t_{n+1}} \|u(t) - v(t)\|. \end{aligned}$$

****

Proof of Lemma 9.1. By subtracting the resulting two formulae (9.5) and by using the notation (6.12), we get

$$\begin{aligned}&\|\zeta(t_n+\theta h_{n+1})-\eta(t_n+\theta h_{n+1})\|\le\|z_n-w_n\|\\&+h_{n+1}\|\Psi(t_n,z_n,h_{n+1},\theta;f_u)-\Psi(t_n,w_n,h_{n+1},\theta;f_v)\|,\end{aligned}$$

where

$$f_u(t,y)=f(t,y,u(t-\tau(t,y)))$$

and

$$f_v(t,y)=f(t,y,v(t-\tau(t,y))).$$

Thus, by Propositions 6.1 and 6.2, there exist constants $Q_u>0$ and $\delta_u>0$ such that, for $h_{n+1}\le h_u$,

$$\begin{aligned}&\|\Psi(t_n,z_n,h_{n+1},\theta;f_u)-\Psi(t_n,w_n,h_{n+1},\theta;f_v)\|\\&\le\|\Psi(t_n,z_n,h_{n+1},\theta;f_u)-\Psi(t_n,w_n,h_{n+1},\theta;f_u)\|\\&+\|\Psi(t_n,w_n,h_{n+1},\theta;f_u)-\Psi(t_n,w_n,h_{n+1},\theta;f_v)\|\\&\le Q_u\|z_n-w_n\|+\delta_u\sup_{t_n\le t\le t_{n+1},\ y\in\mathbb{R}^d}\|f_u(t,y)-f_v(t,y)\|\\&\le Q_u\|z_n-w_n\|+\delta_uM\sup_{t_n\le t\le t_{n+1},\ y\in\mathbb{R}^d}\|u(t-\tau(t,y))-v(t-\tau(t,y))\|,\end{aligned}$$

where M is the Lipschitz constant of $f(t,y,x)$ with respect to x. Therefore, since $\tau(t,y)\ge 0$, with $P_u=\delta_uM$ the proof is complete. ■

Proof of Theorem 9.2. For each $n = 0, \ldots, N-1$, consider the local problems

$$\begin{cases} z'_{n+1}(t) = f\Big(t, z_{n+1}(t), y\big(t - \tau(t, z_{n+1}(t))\big)\Big), \quad t_n \le t \le t_{n+1}, \\ z_{n+1}(t_n) = y(t_n), \\ y(t) = \phi(t), \quad t \le t_0, \end{cases} \tag{9.15}$$

whose solution evidently is $z_{n+1}(t) = y(t)$, i.e. the solution of (9.14). Moreover, consider the auxiliary local problem

$$\begin{cases} w'_{n+1}(t) = f\Big(t, w_{n+1}(t), \eta\big(t - \tau(t, w_{n+1}(t))\big)\Big), \quad t_n \le t \le t_{n+1}, \\ w_{n+1}(t_n) = \eta(t_n), \\ \eta(t) = \phi(t), \quad t \le t_0, \end{cases} \tag{9.16}$$

where, for $s \le t_{n+1}$, $\eta(s)$ is the continuous numerical solution given by the DDE method itself.

By Lemma 9.1 with $u(x) = y(x)$, $v(x) = \eta(x)$, $z_n = y(t_n)$ and $w_n = \eta(t_n)$, for $h \le h_y$, the numerical solutions $\zeta(t)$ and $\eta(t)$ of (9.15) and (9.16), respectively, satisfy the inequality

$$\begin{aligned} \max_{t_n \le t \le t_{n+1}} \|\zeta(t) - \eta(t)\| \;&\le\; (1 + h_{n+1}Q)\|y(t_n) - \eta(t_n)\| \\ &+\; h_{n+1}P \max_{t \le t_{n+1}} \|y(t) - \eta(t)\|, \end{aligned} \tag{9.17}$$

where the constants P and Q depend just on the solution y.

Now consider the inequality

$$\|y(t_{n+1}) - \eta(t_{n+1})\| \le \|y(t_{n+1}) - \zeta(t_{n+1})\| + \|\zeta(t_{n+1}) - \eta(t_{n+1})\|.$$

Since the discontinuity points of order $\le p$ are included in the mesh, the solution $z_{n+1}(t)$ of (9.15) is sufficiently smooth in $[t_n, t_{n+1}]$. Therefore, since the DDE method has discrete order p (see Definition 6.2), (9.17) yields

$$\begin{aligned} \|y(t_{n+1}) - \eta(t_{n+1})\| \;&\le\; M h_{n+1}^{p+1} + (1 + h_{n+1}Q)\|y(t_n) - \eta(t_n)\| \\ &\quad + h_{n+1}P \max_{t \le t_{n+1}} \|y(t) - \eta(t)\| \end{aligned}$$

for some constant $M > 0$. Consequently, with

$$e_n = \max_{i \le n} \|y(t_i) - \eta(t_i)\|$$

and

$$E_n = \max_{t \le t_n} \|y(t) - \eta(t)\|$$

for $n = 0, \ldots, N$, we obtain

$$e_{n+1} \le M h_{n+1}^{p+1} + (1 + h_{n+1}Q)e_n + h_{n+1}PE_{n+1} \qquad (9.18)$$

for $n = 0, \ldots, N-1$.

Similarly, for the interpolant consider the inequality

$$\max_{t_n \le t \le t_{n+1}} \|y(t) - \eta(t)\| \le \max_{t_n \le t \le t_{n+1}} \|y(t) - \zeta(t)\| + \max_{t_n \le t \le t_{n+1}} \|\zeta(t) - \eta(t)\|.$$

Again by the smoothness of $z_{n+1}(t)$ in $[t_n, t_{n+1}]$ and since the DDE method has uniform order q (see again Definition 6.2), we get

$$\begin{aligned} \max_{t_n \le t \le t_{n+1}} \|y(t) - \eta(t)\| \;\le\; & M h_{n+1}^{q+1} + (1 + h_{n+1}Q)\|y(t_n) - \eta(t_n)\| \\ & + h_{n+1}P \max_{t \le t_{n+1}} \|y(t) - \eta(t)\| \end{aligned}$$

for the same constant $M > 0$ (it is not restrictive). Thus we obtain

$$\max_{t_n \le t \le t_{n+1}} \|y(t) - \eta(t)\| \le M h_{n+1}^{q+1} + (1 + h_{n+1}Q)e_n + h_{n+1}PE_{n+1} \qquad (9.19)$$

for $n = 0, \ldots, N-1$.

With $R = \max\{M, P, Q\}$ the inequalities (9.18) and (9.19) yield

$$e_{n+1} \le (1 + h_{n+1}R)e_n + h_{n+1}RE_{n+1} + h_{n+1}Rh^p \qquad (9.20)$$

and

$$\max_{t_n \le t \le t_{n+1}} \|y(t) - \eta(t)\| \le (1+h)Re_n + hRE_{n+1} + Rh^{q+1} \qquad (9.21)$$

for $n = 0, \ldots, N-1$.

Now assume, without any restriction, that $h \leq h^*$, where $h^* = \min\{1, 1/(2R)\}$. Since both e_n and E_n are monotonic, (9.21) implies

$$E_{n+1} \leq (1+h)Re_n + hRE_{n+1} + Rh^{q+1}$$

and, hence,

$$E_{n+1} \leq \frac{(1+h)R}{1-hR}e_n + \frac{R}{1-hR}h^{q+1} \tag{9.22}$$

for $n = 0, \ldots, N-1$. With $S = R + 4R^2$, by substituting (9.22) into (9.20), we get

$$\begin{aligned} e_{n+1} &\leq \left[1 + h_{n+1}\left(R + \frac{(1+h)R^2}{1-hR}\right)\right] e_n + h_{n+1}\left(R + \frac{R^2}{1-hR}\right) h^{q'} \\ &\leq (1 + Sh_{n+1})e_n + h_{n+1}Sh^{q'} \\ &\leq e^{Sh_{n+1}}e_n + h_{n+1}Sh^{q'} \end{aligned}$$

for $n = 0, \ldots, N-1$, where $q' = \min\{p, q+1\}$.

Now we are in a position to prove by standard arguments that

$$e_n \leq \left(\sum_{i=1}^{n} \mathrm{e}^{S(t_n - t_i)} h_i\right) Sh^{q'} \leq \left(\mathrm{e}^{S(t_f - t_0)} - 1\right) h^{q'},$$

that completes the proof. ■

According to Theorem 9.2, if the underlying CRK method has discrete order p and uniform order q, then we can either be satisfied with a DDE method with, possibly lower, uniform global order $q' = \min\{p, q+1\}$, or increase the uniform order of the underlying interpolant up to at least $p-1$ in order to preserve the uniform global order p.

We can summarize the last option in the following corollary.

****

Corollary 9.1 *Under the hypotheses of Theorem 9.2 with* $q \geq p-1$*, the continuous numerical solution* $\eta(t)$ *is such that*

$$\max_{t_0 \leq t \leq t_f} \|y(t) - \eta(t)\| = O(h^p).$$

****

Theorem 9.2 and Corollary 9.1 just guarantee that, by using an interpolant of order $p-1$, the global order p of the discrete method is preserved for any choice of the mesh. A sharper error estimate and convergence analysis of the standard approach reveals that, under some restrictions on the mesh, the condition $q = p-1$ is no longer necessary for the method to preserve the global order p. In other words, superconvergence is possible. On the other hand, an efficient DDE code ought to be implemented in a variable stepsize mode by performing the error control. In this case, if we try to estimate the local error by a method of higher order $p+1$, uniform approximation of order $p-1$ for the deviated arguments $y(t-\tau)$ is not sufficient and must be raised to p.

For a deeper insight into the convergence properties it is still worth distinguishing between constant, vanishing and non-vanishing delays.

10. DDEs with constant delay: natural methods and superconvergence

Assume that the DDE (9.1) reduces to

$$\begin{cases} y'(t) = f\Big(t, y(t), y\big(t-\tau\big)\Big), & t_0 \le t \le t_f, \\ y(t) = \phi(t), & t \le t_0, \end{cases} \tag{10.1}$$

where τ is a constant delay.

Then the RK method (9.5), (9.6), (9.8) takes the simpler form

$$\eta(t_n + \theta h_{n+1}) = y_n + h_{n+1} \sum_{i=1}^{s} b_i(\theta) f\big(t_{n+1}^i, Y_{n+1}^i, \eta(t_{n+1}^i - \tau)\big), \tag{10.2}$$

$$Y_{n+1}^i = y_n + h_{n+1} \sum_{j=1}^{s} a_{ij} f\big(t_{n+1}^j, Y_{n+1}^j, \eta(t_{n+1}^j - \tau)\big), \quad i = 1, \dots, s, \tag{10.3}$$

where, for $h_{n+1} \le \tau$, $\eta(t_{n+1}^j - \tau)$ is known for any j.

In particular, if the algorithm proceeds with constant stepsize $h = \tau/m$ for some integer $m \ge 1$, then the deviated arguments take the values

$$t_{n+1}^j - \tau = t_{n+1-m}^j = t_{n+1-m} + c_j h, \quad j = 1, \dots, s.$$

If the underlying CRK method is natural, then $\eta(t_{n+1}^j) = Y_{n+1}^j$ and also $\eta(t_{n+1}^j - \tau) = Y_{n+1-m}^j$ (see (6.11)). Therefore, we get the discrete method

$$y_{n+1} = y_n + h \sum_{i=1}^{s} b_i f\left(t_{n+1}^i, Y_{n+1}^i, Y_{n+1-m}^i\right), \tag{10.4}$$

$$Y_{n+1}^i = y_n + h \sum_{j=1}^{s} a_{ij} f\left(t_{n+1}^j, Y_{n+1}^j, Y_{n+1-m}^j\right), \quad i = 1, \ldots, s. \tag{10.5}$$

Summarizing, any DDE method (10.2), (10.3) based on a natural CRK method implemented with constant stepsize $h_{n+1} = h = \tau/m$, $m \geq 1$ integer, takes the form (10.4), (10.5) which does not make any explicit calls on interpolation. We shall call it the *natural RK method for DDEs*.

As for the convergence and accuracy order of (10.4), (10.5), we could still use the general results of Theorem 9.2. On the basis of that theorem, any RK method for DDEs is convergent with uniform global order $q' = \min\{p, q+1\}$. Therefore, the DDE method preserves the discrete global order p if the underlying interpolant has uniform order $q \geq p-1$. Unfortunately, in general, the interpolant of a natural CRK method has uniform order $q < p-1$. Nevertheless, it can be proved that, for (10.4), (10.5), the discrete global order p of the underlying natural CRK method is preserved no matter what the uniform order q of the interpolant is.

****

Theorem 10.1 *If the function $f(t, y, x)$ is C^p-continuous in $[t_0, t_f] \times \mathbb{R}^d \times \mathbb{R}^d$, then the natural RK method for DDEs (10.4), (10.5) of discrete order p is convergent of discrete order p, that is*

$$\max_{1 \leq n \leq N} \|y(t_n) - y_n\| = O(h^p),$$

for any initial function $\phi(t)$ of class C^p in $[t_0 - \tau, t_0]$ and for constant stepsize $h = \tau/m$, $m \geq 1$ integer.

****

The proof is obtained by simply noting that the nodal approximations given by the scheme (10.4), (10.5) coincide with those given by the underlying discrete RK scheme implemented using *Bellman's method of steps* with the same stepsize $h = \tau/m$ (see the forthcoming Section 10.1).

Theorem 10.1 says, in particular, that a superconvergent CRK method, which is also natural, preserves the superconvergence for DDEs. That is, the attained order at the nodal points is larger than the global uniform order. In view of more general cases involving variable delays, it is worth remarking that this preservation of superconvergence is essentially due to two occurrences which are intrinsic to implementations of natural CRK schemes with constant stepsize $h = \tau/m$, $m \geq 1$ integer:

(i) The deviated argument $t - \tau$ maps any mesh interval $[t_n, t_{n+1}]$ into a previous mesh interval $[t_{n-m}, t_{n+1-m}]$ and, hence, in the local problem

$$\begin{cases} w'_{n+1}(t) = f\big(t, w_{n+1}(t), \eta(t-\tau)\big), & t_n \leq t \leq t_{n+1}, \\ w_{n+1}(t_n) = y_n, \end{cases}$$

the function $\eta(t - \tau)$ is of class C^∞.

(ii) For every abscissa c_i, the deviated argument $t - \tau$ maps the point t^i_{n+1} into the point $t^i_{n+1} - \tau = t^i_{n+1-m}$, so that

$$\eta(t^i_{n+1} - \tau) = Y^i_{n+1-m}.$$

On the other hand, it is easy to check that, for an arbitrary choice of the mesh points $\{\xi_0 = t_0 < t_1 < \cdots < t_m = \xi_1\}$ in the first macro interval $[\xi_0, \xi_1]$, conditions (i) and (ii) still hold in $[t_0, t_f]$ provided that, in the subsequent macro intervals $[\xi_k, \xi_{k+1}]$, $k = 1, 2, \ldots$, the mesh points are defined by

$$t_{km+i} = t_i + k\tau, \quad i = 1, \ldots, m, \quad k = 1, 2, \ldots \tag{10.6}$$

A mesh Δ satisfying condition (10.6) is called *constrained mesh* (see the forthcoming Definition 11.1 relevant to more general delays).

Example 10.1 *Consider equation (5.3) and the DDE method (5.6), (5.7) based on the midpoint rule with constant stepsize* $h = 1/(m-\delta)$. *Note that, for* $\delta = 0$, *the corresponding mesh* Δ *is a constrained mesh, whereas, for* $\delta \neq 0$, *it is not. Therefore, since the midpoint rule with linear interpolation is a collocation method of order* $p = 2$, *it is also a natural CRK method, and the resulting DDE method for* $\delta = 0$ *converges with global order* 2. *On the other hand, the midpoint rule with linear interpolation is a CRK method of uniform order* $q = 1$ *and, hence, Theorem 9.2 guarantees the global order* 2 *for any sequence of meshes with constant (or even variable) stepsize. This is the case of* $h = 1/m_i$, *with* $m_i = \frac{5}{3}2^i$, *which leads to a non-constrained mesh for any* i. *The numerical evidence for both constrained and non-constrained meshes is given in Figure 14.*

The DDE method based on collocation at $\nu > 1$ *Gaussian points is different because it is a superconvergent method with global order* $p = 2\nu$ *and local uniform order* $q = \nu$. *For* $\delta = 0$ *the DDE method still preserves the global order* 2ν. *On the contrary, for* $\delta \neq 0$ *the mesh is no longer constrained and, hence, by Theorem 9.2, the global order decays to the uniform global order* $\nu + 1$. *For* $\nu = 2$, *Figure 14 illustrates the numerical results with stepsizes that are integer and non-integer submultiples of the delay.* $\Diamond$

10.1. Bellman's method of steps

The most elegant approach aimed at avoiding the need for interpolation in the numerical solution of the DDE with constant delay (10.1) is known as the *method of steps* and was first developed by Bellman. Bellman's method is not the most convenient approach for solving DDEs numerically, but it is certainly attractive because it allows the use of variable stepsizes without any interpolation.

The discontinuity points are $\xi_k = t_0 + k\tau$. In the first interval $[t_0, t_0 + \tau]$ the DDE (10.1) has the form

$$\begin{cases} y'(t) = f\big(t, y(t), \phi(t-\tau)\big), \\ y(t_0) = \phi(t_0). \end{cases}$$

In the second interval $[t_0 + \tau, t_0 + 2\tau]$ we can define $y_1(t) = y(t-\tau)$ and $y_2(t) = y(t)$, and thus we can write the DDE (10.1) as the $2d$-dimensional system of ODEs

$$\begin{cases} y_1'(t) = f\big(t-\tau, y_1(t), \phi(t-2\tau)\big), \\ y_2'(t) = f\big(t, y_2(t), y_1(t)\big), \\ y_1(t_0+\tau) = \phi(t_0), \\ y_2(t_0+\tau) = y(t_0+\tau). \end{cases}$$

In general, in the interval $[t_0 + (k-1)\tau, t_0 + k\tau]$ we can write the DDE (10.1) as the kd-dimensional system of ODEs

$$\begin{cases} y_i'(t) = f\big(t-(k-i)\tau, y_i(t), y_{i-1}(t)\big), \quad i = 1, \dots, k, \\ y_i(t_0 + (k-1)\tau) = y(t_0 + (i-1)\tau), \quad i = 1, \dots, k, \end{cases} \tag{10.7}$$

where we have set $y_0(t) = \phi(t - k\tau)$ and $y_i(t) = y(t-(k-i)\tau)$, $i = 1, \dots, k$.

Passing from k to $k+1$ in (10.7) means shifting the integration interval from $[t_0+(k-1)\tau, t_0+k\tau]$ to $[t_0+k\tau, t_0+(k+1)\tau]$ and extending the solution from $[t_0, t_0+k\tau]$ to $[t_0, t_0+(k+1)\tau]$ by adding the component $y_{k+1}(t) = y(t)$ in the current interval. Therefore, we can choose a standard numerical method for ODEs and solve, for increasing k, the larger and larger systems (10.7). For each step k, the numerical solution of the kd-dimensional system (10.7) aims to provide an approximate value of $y_k(t_0+k\tau) = y(t_0+k\tau)$ to be used as the initial value of the new component $y_{k+1}(t) = y(t)$ in the next step. The process ends when, for some k, $t_0 + k\tau \geq t_f$.

In this way we are obliged to solve a system of ever-growing dimension and, in principle, to recompute many times the same pieces of solution related to the previous intervals. On the other hand, the reduction to a system of ODEs avoids the typical complications due to the presence of the delayed argument, that is storing and interpolating the computed solution throughout the interval $[t_0, t_f - \tau]$.

As for the convergence of the process, observe that, at each step, any numerical method of global order p provides a numerical approximation of order p for any perturbation of the initial values in (10.7) of order $\geq p$. Therefore, in the next step, we have to solve an IVP with initial values perturbed to the same order as that of the method, and so forth. The whole process performs to order p for any finite number of steps.

11. DDEs with non-vanishing time dependent delay: constrained mesh and superconvergence

Assume that the DDE (9.1) reduces to

$$\begin{cases} y'(t) = f\Big(t, y(t), y\big(t-\tau(t)\big)\Big), & t_0 \le t \le t_f, \\ y(t) = \phi(t), & t \le t_0, \end{cases} \tag{11.1}$$

where the delay depends on t and satisfies hypotheses (H_1) and (H_4). Then the standard approach consists in computing the discontinuity points $\xi_0 = t_0 < \xi_1 < \cdots < \xi_l = t_f$ by using (3.8), and then solving the ODE

$$\begin{cases} w'(t) = f\Big(t, w(t), \eta\big(t-\tau(t)\big)\Big), & \xi_{k-1} \le t \le \xi_k, \\ w(\xi_{k-1}) = \eta(\xi_{k-1}), \end{cases}$$

at each macro interval $[\xi_{k-1}, \xi_k]$.

The RK method (9.5), (9.6), (9.7), (9.8) takes the form

$$Y^i_{n+1} = y_n + h_{n+1} \sum_{j=1}^{s} a_{ij} f\Big(t^j_{n+1}, Y^j_{n+1}, \eta\big(t^j_{n+1} - \tau(t^j_{n+1})\big)\Big), \quad i = 1, \dots, s, \tag{11.2}$$

$$\eta(t_n + \theta h_{n+1}) = y_n + h_{n+1} \sum_{i=1}^{s} b_i(\theta) f\Big(t^i_{n+1}, Y^i_{n+1}, \eta\big(t^i_{n+1} - \tau(t^i_{n+1})\big)\Big), \tag{11.3}$$

where hypothesis (H_1) ensures that, for h_{n+1} small enough, the terms $\eta\big(t^i_{n+1} - \tau(t^i_{n+1})\big)$ are known.

In connection with the constrained mesh selection introduced in the previous section for preserving the superconvergence, consider the following generalization of the definition of a constrained mesh to the case of variable delay.

Definition 11.1 *A mesh Δ related to the deviated argument $\alpha(t) = t - \tau(t)$ is called a* constrained mesh *if, for an arbitrary set of mesh points $\{\xi_0 = t_0 < t_1 < \cdots < t_m = \xi_1\}$ in $[\xi_0, \xi_1]$, the mesh is recursively determined in the subsequent intervals $[\xi_k, \xi_{k+1}]$, $k = 1, \ldots, l-1$, by*

$$t_{km+i} = t_{(k-1)m+i} + \tau(t_{km+i}), \qquad i = 1, \ldots, m. \tag{11.4}$$

Note that, by starting from $[\xi_0, \xi_1]$, the *forward* construction of the constrained mesh requires the solution of equation (11.4) for t_{km+i}, $i = 1, \ldots, m$, $k = 1, \ldots, l-1$. A cheaper construction of a constrained mesh can be done in a *backward* fashion by starting from an arbitrary mesh selection in the last macro interval $[\xi_{l-1}, \xi_l]$ and by explicitly computing the mesh points $t_{(k-1)m+i}$, $k = l-1, l-2, \ldots, 1$, $i = 1, \ldots, m$, in the previous macro intervals.

It is evident that, for a constrained mesh Δ, the characteristic property (i) is still fulfilled. On the contrary, property (ii) is not satisfied even if $b_j(c_i) = a_{ij}$, unless the deviated argument $\alpha(t)$ is a function of the form $\alpha(t) = rt - d$, $0 < r < 1$, $d \geq 0$, $t \geq t_0 > 0$ such as, for example, in the pantograph equation (3.1). In fact, in general, we have

$$t^j_{n+1} - \tau(t^j_{n+1}) \neq t_{n+1-m} + c_j h_{n+1-m} = t^j_{n+1-m},$$

and hence, even if $\eta(t^j_{n+1-m}) = Y^j_{n+1-m}$, we have $\eta\big(t^j_{n+1} - \tau(t^j_{n+1})\big) \neq Y^j_{n+1-m}$.

Despite the lack of property (ii), which appeared to be crucial in the proof of Theorem 10.1 via Bellman's approach, the preservation of superconvergence for CRK methods of uniform order $q < p - 1$ is still possible under the constrained mesh selection strategy.

However, the asymptotic orthogonality condition (7.1), rather than $\eta(t_{n+1}^i) = Y_{n+1}^i$, is the essential property for interpolation in order to preserve the superconvergence in the constrained mesh implementation of the DDE method.

The following result on superconvergence is a corollary to Theorem 7.5.

****

Theorem 11.1 *Assume that the DDE method (11.2), (11.3) is applied to (11.1) with a constrained mesh Δ. If the underlying CRK method has discrete order p and the interpolant is an NCE of order q, then the continuous numerical solution $\eta(t)$ is such that*

$$\max_{t_0 \le t \le t_f} \|y(t) - \eta(t)\| = O(h^{q'}),$$

where $q' = \min\{q+1, p\}$,

$$\max_{0 \le n \le N} \|y(t_n) - \eta(t_n)\| = O(h^p),$$

$$\max_{t_0 \le t \le t_f} \|y^{(j)}(t) - \eta^{(j)}(t)\| = O(h^{q-j+1}), \quad j = 1, \ldots, q,$$

$$\max_{0 \le n \le N-1} \left\| \int_{t_n}^{t_{n+1}} G(x)[y^{(j)}(x) - \eta^{(j)}(x)]\mathrm{d}x \right\| = O(h^{p+1}), \quad j = 0, 1,$$

for every sufficiently smooth matrix-valued function $G(x)$.

****

Proof of Theorem 11.1. The proof is carried out by induction on the macro intervals $[\xi_{k-1}, \xi_k]$. The result is clearly true in the first interval $[\xi_0, \xi_1]$ because of the property (6.14) of the discrete RK method and of the properties (6.15), (6.16) and (7.3) of the NCE $\eta(t)$.

Then assume that the result is true up to the interval $[\xi_{k-2}, \xi_{k-1}]$ and solve

$$\begin{cases} w'(t) = f\Big(t, w(t), \eta\big(t - \tau(t)\big)\Big), & \xi_{k-1} \le t \le \xi_k, \\ w(\xi_{k-1}) = \eta(\xi_{k-1}). \end{cases} \tag{11.5}$$

Moreover, consider the auxiliary DDE

$$\begin{cases} z'(t) = f\big(t, z(t), y(t - \tau(t))\big), & \xi_{k-1} \le t \le \xi_k, \\ z(\xi_{k-1}) = \eta(\xi_{k-1}). \end{cases}$$

Since we have assumed $\|y(\xi_{k-1}) - \eta(\xi_{k-1})\| = O(h^p)$, it is clear that

$$\max_{\xi_{k-1} \le t \le \xi_k} \|y(t) - z(t)\| = O(h^p). \tag{11.6}$$

Because the mesh is constrained, by setting $\alpha(x) = x - \tau(x)$ and by identifying the intervals $[\xi_{k-1}, \xi_k]$ with $[\tilde{a}, \tilde{b}]$ and $[\xi_{k-2}, \xi_{k-1}]$ with $[a, b]$ and the functions $y(t)$ with $u_\Delta(t)$ and $\eta(t)$ with $v_\Delta(t)$, we are in the hypotheses of Theorem 7.5. Therefore, formulae (7.10), (7.11), (7.12) and (7.13) hold for $w(t)$ and $z(t)$.

Since the RK method and its NCE satisfy (6.14), (6.15), (6.16) and (7.3) with respect to the solution $w(t)$ of (11.5), the estimate (11.6) eventually implies that the result is true also in the interval $[\xi_{k-1}, \xi_k]$. ■

Example 11.1 *Consider the equation*

$$\begin{cases} y'(t) = \lambda \frac{t-1}{t} y\big(t - \log(t) - 1\big) y(t), & t \geq 1, \\ y(t) = 1, & 0 \leq t \leq 1. \end{cases} \tag{11.7}$$

We want to integrate it in the interval $[1, \xi_2]$*, where the breaking points are* $\xi_1 = 3.1461932206205825852$ *and* $\xi_2 = 5.9254498245082464926$ *and the solution is*

$$y(t) = \begin{cases} \exp\left(\lambda(t - \log(t) - 1)\right), & 1 \leq t \leq \xi_1, \\ \exp\left(\lambda + \int_{\xi}^{t} \lambda \frac{s-1}{s} \mathrm{e}^{\lambda(s - \log(s(s - \log(s) - 1)) - 2)} \mathrm{d}s\right), & \xi_1 \leq t \leq \xi_2. \end{cases}$$

To this end, consider the following family of CERK methods:

$$\begin{array}{c|ccc|l} 0 & & & & b_1(\theta) = -x_1\theta^2 + (x_1 + \frac{1}{6})\theta, \\ \frac{1}{2} & \frac{1}{2} & & & b_2(\theta) = -x_2\theta^2 + (x_2 + \frac{1}{3})\theta, \\ \frac{1}{2} & 0 & \frac{1}{2} & & b_3(\theta) = -x_3\theta^2 + (x_3 + \frac{1}{3})\theta, \\ 1 & 0 & 0 & 1 & b_4(\theta) = -x_4\theta^2 + (x_4 + \frac{1}{6})\theta, \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array}$$

of discrete order $p = 4$*, uniform order* $q = 2$ *and uniform global order* $q' = 3$*, where the parameters* x_1, x_2, x_3 *and* x_4 *satisfy the conditions*

$$\begin{aligned} x_1 + x_2 + x_3 + x_4 &= 0, \\ x_4 - x_1 &= 1, \\ x_2 + x_3 + 2x_4 &= -1, \end{aligned}$$

derived from Table 1.

In particular, for $x_1 = \frac{1}{2}$, $x_4 = -\frac{1}{2}$ *and* $x_2 = x_3 = 0$*, the interpolant is an NCE (see Section 7.2) with matrix* $B =$

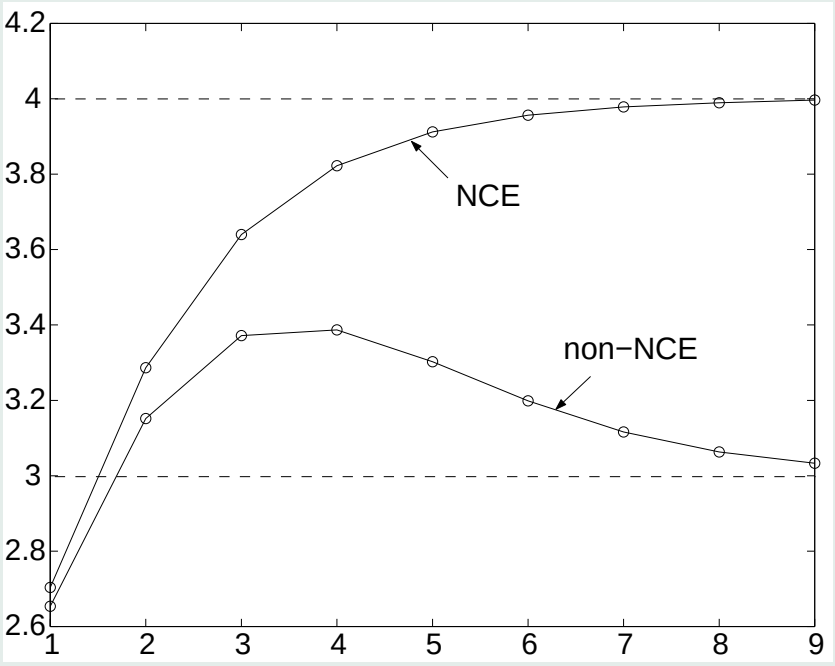

Figure 24: Logarithmic plots of ratios between consecutive errors for the solution of (11.7) with $\lambda = 1$, obtained by doubling the number of steps for the explicit RK method of order 4 interpolated to the uniform global order 3 by an NCE and a non-NCE interpolant.

$\big(b_j(c_i)\big) \neq A$. *Any other choice of the parameters* x_i *leads to an interpolant that is not an NCE.*

The constrained mesh is built up in a backward fashion with N *equidistant mesh points in the last macro interval* $[\xi_1, \xi_2]$ *and the corresponding images in* $[1, \xi_1]$. *We have compared the solutions obtained by using the NCE and the interpolant determined by the parameters* $x_1 = \frac{1}{2}$, $x_4 = -\frac{1}{2}$ *and* $x_2 = -x_3 = 1$. *In Figure 24 the nodal order of the two methods are shown by successively doubling* N. *As expected, the method performs to order 4 when the interpolant is the NCE, whereas the order is only 3 in the other case.* $\Diamond$

12. DDEs with vanishing time dependent delay

When the delay $\tau(t)$ in (11.1) vanishes at some point $t^* \in [t_0, t_f]$, overlapping necessarily occurs for any mesh interval $[t_n, t_{n+1}]$ including t^*. Therefore, for the method

$$\eta(t_n+\theta h_{n+1}) = y_n+h_{n+1}\sum_{i=1}^{s} b_i(\theta) f(t^i_{n+1}, Y^i_{n+1}, \tilde{Y}^i_{n+1}), \quad 0 \le \theta \le 1, \tag{12.1}$$

$$Y^i_{n+1} = y_n + h_{n+1}\sum_{j=1}^{s} a_{ij} f(t^j_{n+1}, Y^j_{n+1}, \tilde{Y}^j_{n+1}), \quad i = 1, \ldots, s, \tag{12.2}$$

where

$$\tilde{Y}^i_{n+1} = y_n + h_{n+1}\sum_{j=1}^{s} b_j \left(c_i - \frac{\tau(t^i_{n+1})}{h_{n+1}} \right) f(t^j_{n+1}, Y^j_{n+1}, \tilde{Y}^j_{n+1}) \tag{12.3}$$

if $t^i_{n+1} - \tau(t^i_{n+1}) > t_n$, and

$$\tilde{Y}^i_{n+1} = \eta\big(t^i_{n+1} - \tau(t^i_{n+1})\big) \tag{12.4}$$

if $t^i_{n+1} - \tau(t^i_{n+1}) \le t_n$, the option (12.3) is possibly involved and some values $\tilde{Y}^i_{n+1}$ might actually be unknown.

This does not cause much trouble and the general results of Theorems 9.1 and 9.2 extend to the method (12.1), (12.2), (12.3), (12.4).

As for possible superconvergence phenomena, the results of the previous section hold as long as we avoid overlapping. On the contrary, when we pass a vanishing delay point, the previous theory collapses. However, there is a particular case of vanishing delay, namely the *proportional delay* $\tau(t) = (1 - r)t$, $0 < r < 1$, typical of the pantograph equation, where the preservation of superconvergence is still possible (but we do not treat this case in these Lectures).

13. RK methods for NDDEs

Consider the DDE of neutral type

$$\begin{cases} y'(t) = f\Big(t, y(t), y\big(t-\tau(t,y(t))\big), y'\big(t-\sigma(t,y(t))\big)\Big), \quad t_0 \le t \le t_f, \\ y(t) = \phi(t), \qquad t \le t_0. \end{cases} \tag{13.1}$$

The standard approach consists in solving step by step the local problems

$$\begin{cases} w'_{n+1}(t) = f\Big(t, w_{n+1}(t), x\big(t-\tau(t,w_{n+1}(t))\big), z\big(t-\sigma(t,w_{n+1}(t))\big)\Big), \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad t_n \le t \le t_{n+1}, \\ w_{n+1}(t_n) = \eta(t_n), \end{cases}$$

where

$$x(s) = \begin{cases} \phi(s) & \text{for } s \le t_0, \\ \eta(s) & \text{for } t_0 \le s \le t_n, \\ w_{n+1}(s) & \text{for } t_n \le s \le t_{n+1}, \end{cases}$$

$$z(s) = \begin{cases} \phi'(s) & \text{for } s \le t_0, \\ \lambda(s) & \text{for } t_0 \le s \le t_n, \\ w'_{n+1}(s) & \text{for } t_n \le s \le t_{n+1}, \end{cases}$$

$\eta(t)$ is the continuous approximation of $y(t)$ and $\lambda(t)$ is an approximation of $y'(t)$ given by

$$\lambda(t) = \eta'(t) \tag{13.2}$$

or by

$$\lambda(t) = \mathcal{P}\Big(f\big(\cdot, \eta(\cdot), \eta\big(\,\cdot - \tau(\cdot,\eta(\cdot))\big), \lambda\big(\,\cdot - \sigma(\cdot,\eta(\cdot))\big)\big)\Big)(t), \tag{13.3}$$

where, in each mesh interval $[t_k, t_{k+1}]$, $\mathcal{P}$ is an interpolation operator in a suitable polynomial space of degree possibly other than $\deg(\eta')$ and nodes in $[t_k, t_{k+1}]$.

For the choice (13.2), the DDE method (9.3), (9.4) in the Y notation modifies to the following *RK method for NDDEs*:

$$Y^i_{n+1} = y_n + h_{n+1} \sum_{j=1}^{s} a_{ij} f\big(t^j_{n+1}, Y^j_{n+1}, \tilde{Y}^j_{n+1}, \tilde{Z}^j_{n+1}\big), \quad i = 1, \ldots, s, \tag{13.4}$$

$$\eta(t_n + \theta h_{n+1}) = y_n + h_{n+1} \sum_{i=1}^{s} b_i(\theta) f\big(t^i_{n+1}, Y^i_{n+1}, \tilde{Y}^i_{n+1}, \tilde{Z}^i_{n+1}\big), \; 0 \le \theta \le 1, \tag{13.5}$$

$$\lambda(t_n + \rho h_{n+1}) = \sum_{i=1}^{s} b'_i(\rho) f\big(t^i_{n+1}, Y^i_{n+1}, \tilde{Y}^i_{n+1}, \tilde{Z}^i_{n+1}\big), \quad 0 \le \rho \le 1, \tag{13.6}$$

where

$$\tilde{Y}^j_{n+1} = \eta\big(t^j_{n+1} - \tau(t^j_{n+1}, Y^j_{n+1})\big),$$
$$\tilde{Z}^j_{n+1} = \lambda\big(t^j_{n+1} - \sigma(t^j_{n+1}, Y^j_{n+1})\big).$$

Note that, for the arguments $s_j = t^j_{n+1} - \tau(t^j_{n+1}, Y^j_{n+1})$ and $s'_j = t^j_{n+1} - \sigma(t^j_{n+1}, Y^j_{n+1})$, the values $\eta(s_j)$ and $\lambda(s'_j)$ may or may not be known.

If overlapping occurs, that is if, for some index i, the arguments $s_i > t_n$ and/or $s'_i > t_n$, then the *spurious stages* $\tilde{Y}^i_{n+1}$ and/or $\tilde{Z}^i_{n+1}$ are unknown and are given by (13.5) and (13.6) for

$$\theta = \theta^i_{n+1} = c_i - \frac{\tau(t^i_{n+1}, Y^i_{n+1})}{h_{n+1}},$$

and

$$\rho = \rho^i_{n+1} = c_i - \frac{\sigma(t^i_{n+1}, Y^i_{n+1})}{h_{n+1}},$$

respectively, that is

$$\tilde{Y}^i_{n+1} = y_n + h_{n+1} \sum_{j=1}^{s} b_j(\theta^i_{n+1}) f\big(t^j_{n+1}, Y^j_{n+1}, \tilde{Y}^j_{n+1}, \tilde{Z}^j_{n+1}\big),$$

$$\tilde{Z}^i_{n+1} = \sum_{j=1}^{s} b'_j(\rho^i_{n+1}) f\big(t^j_{n+1}, Y^j_{n+1}, \tilde{Y}^j_{n+1}, \tilde{Z}^j_{n+1}\big).$$

On the contrary, if the arguments of $\eta(s)$ and $\lambda(s)$ lie outside the current interval $[t_n, t_{n+1}]$, then the values $\tilde{Y}^j_{n+1}$ and $\tilde{Z}^j_{n+1}$ are given by the interpolants $\eta(s)$ and $\eta'(s)$ as computed at the past points

$$\begin{aligned} t^i_{n+1} - \tau(t^i_{n+1}, Y^i_{n+1}) &= t_{n+1-m} + \theta h_{n+1-m}, \\ t^i_{n+1} - \sigma(t^i_{n+1}, Y^i_{n+1}) &= t_{n+1-m'} + \rho h_{n+1-m'}, \end{aligned}$$

for suitable values of m, m', θ and ρ.

As with DDEs with no neutral terms, the spurious stages $\tilde{Y}^i_{n+1}$ and $\tilde{Z}^i_{n+1}$, if any, only apparently increase the dimension of the system to be solved at each step. In fact, by using the K notation

$$K^i_{n+1} = f\big(t^i_{n+1}, Y^i_{n+1}, \tilde{Y}^i_{n+1}, \tilde{Z}^i_{n+1}\big),$$

all the stages Y^i_{n+1}, $\tilde{Y}^i_{n+1}$ and $\tilde{Z}^i_{n+1}$, as well as the arguments θ^i_{n+1} and ρ^i_{n+1}, turn out to depend on K^i_{n+1} only.

Remark 13.1 *According to the arguments of Section 9, for any natural CRK method with s distinct abscissae $c_1, \dots, c_s$ such that $c_i \neq 0$, $i = 1, \dots, s$, and continuous extension $\eta(t)$ of degree s, the system to be solved at each step may be stated in terms of the sole Y^i_{n+1}'s. In fact, the polynomial $\eta(t_n + \theta h_{n+1})$ may be written using the Lagrange interpolation formula through the $s+1$ values $y_n (= \eta(t_n))$ and $Y^i_{n+1} (= \eta(t_n + c_i h_{n+1}))$, $i = 1, \dots, s$, that is*

$$\eta(t_n + \theta h_{n+1}) = \ell_0(\theta) y_n + \sum_{i=1}^{s} \ell_i(\theta) Y^i_{n+1}, \tag{13.7}$$

where ℓ_j, $j = 0, \dots, s$ are the Lagrange polynomial coefficients on the nodes $c_0 = 0$ and c_i, $i = 1, \dots, s$. Therefore, $\tilde{Y}^i_{n+1} = \eta\big(t^i_{n+1} - \tau(t^i_{n+1}, Y^i_{n+1})\big)$ may be written by (13.7) for $\theta = \theta^i_{n+1} = c_i - \frac{\tau(t^i_{n+1}, Y^i_{n+1})}{h_{n+1}}$. Similarly, $\tilde{Z}^i_{n+1} = \lambda\big(t^i_{n+1} - \sigma(t^i_{n+1}, Y^i_{n+1})\big)$ may be written by the derivative of (13.7) for $\theta = \rho^i_{n+1} = c_i - \frac{\sigma(t^i_{n+1}, Y^i_{n+1})}{h_{n+1}}$.

For the choice (13.3), the RK method for NDDEs (in the Y notation) is given by (13.4), (13.5) along with

$$\lambda(t_n+\rho h_{n+1}) = \sum_{i=0}^{s^*} \ell_i(\rho) f\big(\bar{t}^i_{n+1}, U^i_{n+1}, \tilde{U}^i_{n+1}, \tilde{V}^i_{n+1}\big), \quad 0 \le \rho \le 1, \tag{13.8}$$

where $\bar{t}^i_{n+1} = t_n + \bar{c}_i h_{n+1}$ and $\ell_i(\theta)$, $i = 0, \dots, s^*$, are the nodes and the Lagrange polynomial coefficients of the interpolation operator $\mathcal{P}$. Here, besides the values

$$\tilde{Y}^j_{n+1} = \eta\big(t^j_{n+1} - \tau(t^j_{n+1}, Y^j_{n+1})\big),$$
$$\tilde{Z}^j_{n+1} = \lambda\big(t^j_{n+1} - \sigma(t^j_{n+1}, Y^j_{n+1})\big),$$

there are additional values

$$U^j_{n+1} = \eta\big(\bar{t}^j_{n+1}\big),$$
$$\tilde{U}^j_{n+1} = \eta\big(\bar{t}^j_{n+1} - \tau(\bar{t}^j_{n+1}, U^j_{n+1}))\big),$$
$$\tilde{V}^j_{n+1} = \lambda\big(\bar{t}^j_{n+1} - \sigma(\bar{t}^j_{n+1}, U^j_{n+1}))\big).$$

Note that, according to the arguments $s_j = t^j_{n+1} - \tau(t^j_{n+1}, Y^j_{n+1})$ and $s'_j = t^j_{n+1} - \sigma(t^j_{n+1}, Y^j_{n+1})$, the values $\eta(s_j)$ and $\lambda(s'_j)$ may or may not be known. If $s_j > t_n$ and/or $s'_j > t_n$, then $\tilde{Y}^j_{n+1}$ and/or $\tilde{Z}^j_{n+1}$ are unknown and must be computed by (13.5) and (13.8), respectively. In particular, if $s'_j > t_n$, then for the application of (13.8) in the current interval U^j_{n+1}, $\tilde{U}^j_{n+1}$ and $\tilde{V}^j_{n+1}$ need to be known. Here the U^j_{n+1}'s are certainly unknown, whereas knowledge of $\tilde{U}^j_{n+1}$ and $\tilde{V}^j_{n+1}$ depends on the location of the further arguments $\bar{t}^j_{n+1} - \tau(\bar{t}^j_{n+1}, U^j_{n+1})$ and $\bar{t}^j_{n+1} - \sigma(\bar{t}^j_{n+1}, U^j_{n+1})$.

Summarizing, if for some index j some of the arguments are $> t_n$, then the relevant *spurious stages* $\tilde{Y}^j_{n+1}$, $\tilde{Z}^j_{n+1}$, U^j_{n+1}, $\tilde{U}^j_{n+1}$ or $\tilde{V}^j_{n+1}$ are unknown and are given by (13.5) and (13.8) for suitable values of θ and ρ. More precisely,

$$\tilde{Y}^j_{n+1} = \eta(t_n + \theta^j_{n+1} h_{n+1}) \qquad \text{with } \theta^j_{n+1} = c_j - \frac{\tau(t^j_{n+1}, Y^j_{n+1})}{h_{n+1}},$$

$$\tilde{Z}^j_{n+1} = \lambda(t_n + \rho^j_{n+1} h_{n+1}) \qquad \text{with } \rho^j_{n+1} = c_j - \frac{\sigma(t^j_{n+1}, Y^j_{n+1})}{h_{n+1}},$$

$$U^j_{n+1} = \eta(t_n + \bar{c}^j h_{n+1}),$$

$$\tilde{U}^j_{n+1} = \eta(t_n + \bar{\theta}^j_{n+1} h_{n+1}) \qquad \text{with } \bar{\theta}^j_{n+1} = \bar{c}_j - \frac{\tau(\bar{t}^j_{n+1}, U^j_{n+1})}{h_{n+1}},$$

$$\tilde{V}^j_{n+1} = \lambda(t_n + \bar{\rho}^j_{n+1} h_{n+1}) \qquad \text{with } \bar{\rho}^j_{n+1} = \bar{c}_j - \frac{\sigma(\bar{t}^j_{n+1}, U^j_{n+1})}{h_{n+1}}.$$

The dimension of the system may still be reduced by using the K notation but, unlike the option (13.2), besides the K values

$$K^j_{n+1} = f\big(t^j_{n+1}, Y^j_{n+1}, \tilde{Y}^j_{n+1}, \tilde{Z}^j_{n+1}\big), \quad j = 1, \ldots, s,$$

we have the additional values

$$H^j_{n+1} = f\big(\bar{t}^j_{n+1}, U^j_{n+1}, \tilde{U}^j_{n+1}, \tilde{V}^j_{n+1}\big), \quad j = 0, \ldots, s^*.$$

Remark 13.2 *Also with the option (13.3), the number of unknowns in the system to be solved at each step may be reduced. In fact, if the underlying CRK method is natural and if the interpolation formula (13.8) is based on the nodes* $\bar{c}_i = c_i$, $i = 1, \dots, s^* = s$, *and on another node* $\bar{c}_0 \neq c_i$, *then, for* $j = 1, \dots, s$,

$$Y_{n+1}^j = U_{n+1}^j,$$
$$\tilde{Y}_{n+1}^j = \tilde{U}_{n+1}^j,$$
$$\tilde{Z}_{n+1}^j = \tilde{V}_{n+1}^j$$

and, therefore, also

$$H_{n+1}^j = K_{n+1}^j.$$

In this case the spurious stages reduce to just Y_{n+1}^j, $\tilde{Y}_{n+1}^j$ *and* $\tilde{Z}_{n+1}^j$ *in the* Y *notation, and to just*

$$K_{n+1}^j = f\big(t_{n+1}^j, Y_{n+1}^j, \tilde{Y}_{n+1}^j, \tilde{Z}_{n+1}^j\big)$$

in the equivalent K *notation. Note also that, for the new set of stage values*

$$Z_{n+1}^j = \lambda(t_n + c_j h_{n+1}), \quad j = 1, \dots, s,$$

by (13.8) we have

$$Z_{n+1}^j = K_{n+1}^j.$$

On the other hand, independently of the choice of the $\bar{c}_i$*'s, if* $c_i \neq 0$, $i = 1, \dots, s$, *the arguments of Section 9 allow us to express each* $\tilde{Y}_{n+1}^j$ *in terms of the* Y_{n+1}^j*'s and, hence, the overall method is based on the stage values* Y_{n+1}^j *and* $\tilde{Z}_{n+1}^j$. *However, in no case the RK method can reduce to just the* Y *values.*

We give up providing a detailed analysis for the special cases of constant and time dependent delays, as made in the foregoing sections for DDEs with no neutral terms. The analysis is quite similar and, in particular, the convergence result that extends Theorem 9.2 may be stated as follows.

****

Theorem 13.1 *Consider the state dependent NDDE (13.1), where $f(t,y,x,w)$ is C^p-continuous in $[t_0,t_f]\times\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d$, the delays $\tau(t,y)$ and $\sigma(t,y)$ are C^p-continuous in $[t_0,t_f]\times\mathbb{R}^d$ and the initial function $\phi(t)$ is C^p-continuous. Moreover, assume that the mesh $\Delta=\{t_0,t_1,\ldots,t_n,\ldots,t_N=t_f\}$ includes all the discontinuity points of order $\le p$ lying in $[t_0,t_f]$. If the underlying CRK method $\big(A,b(\theta)\big)$ has discrete order p and uniform order q, and the approximation $\lambda(t)$ has uniform order r, then the resulting RK method for NDDEs has discrete global order and uniform global order $q'=\min\{p,q+1,r+1\}$; that is*

$$\max_{1\le n\le N}\|y(t_n)-y_n\|=O(h^{q'})$$

and

$$\max_{t_0\le t\le t_f}\|y(t)-\eta(t)\|=O(h^{q'}),$$

where $h=\max_{1\le n\le N}h_n$. In particular, if $\lambda(t)$ is given by the option (13.2), then $r=q-1$ and, hence, $q'=\min\{p,q\}$.

****

Note that, for the option (13.3), $\lambda(t)$ is given by (13.8) and the interpolation operator $\mathcal{P}$ has order $r = s^*$. Therefore, on the basis of Theorem 13.1, it is useless to take $s^* > q$. On the other hand, the choice $s^* = q$ preserves the optimal order $q' = \min\{p, q+1\}$ and makes the option (13.3), along with the conditions in Remark 13.2, preferable to (13.2).

As for possible superconvergence phenomena, similar results as with the non-neutral equation can be proved. They are still based on the general result about the preservation of superconvergence of RK methods stated in Theorem 7.5.

Summer School on Delay Differential Equations and Applications

Dobbiaco (BZ), Italy, June 26–30, 2006

The numerical solution of delay differential equations

M. Zennaro
Dipartimento di Matematica e Informatica
Università di Trieste

Lecture 5: Stability analysis of some test delay equations

Main reference for this Lecture:
A. Bellen and M. Zennaro, Numerical Methods for Delay Differential Equations, Numerical Mathematics and Scientific Computation, Oxford Science Publications, Oxford University Press, 2003 (Chapters 8 and 9)

14. The test equations

The stability analysis of DDEs have been investigated extensively, although not completely developed for the more complicated cases yet. Here we resume with some definitions and results on *contractivity* (or *dissipativity*) and *asymptotic stability* of DDEs and NDDEs that are useful for deriving classes of numerical methods that preserve the same property. In the stability analysis of DDEs, there are two possible approaches that have been considered in the literature in view of the construction of correspondingly stable numerical methods. One approach consists in finding conditions on the right-hand-side function f such that the problem is stable for all or for some classes of delays, typically for all constant delays. The second approach consists in finding weaker conditions on f such that the desired stability property is guaranteed for the specific given (in general constant) delay. The two concepts of stability are actually different and we shall refer to the former as *stability for all delays*, or *delay independent stability*, and to the latter as *stability for fixed delay*, or *delay dependent stability*.

It is evident that the class of problems that are stable for all delays is smaller than that for fixed delay. Moreover, characterizing the class of problems that are stable for all delays and finding numerical methods that are stable on that class is easier than for fixed delay. On the contrary, the analysis for fixed delay is sharper and the resulting class of stable problems is larger. Characterizing this class and finding stable numerical methods turns out to be much harder.

In these lectures we consider only scalar linear test equations as follows.

We distinguish the class of scalar linear DDEs with variable coefficients

$$\begin{cases} y'(t) = \lambda(t)y(t) + \mu(t)y(t-\tau(t)), \quad t \geq t_0, \\ y(t) = \phi(t), \quad t \leq t_0, \end{cases} \tag{14.1}$$

and with constant coefficients

$$\begin{cases} y'(t) = \lambda y(t) + \mu y(t-\tau(t)), \quad t \geq t_0, \\ y(t) = \phi(t), \quad t \leq t_0, \end{cases} \tag{14.2}$$

respectively. Observe that, if the delay τ is constant, then equation (14.2) reduces to the linear autonomous equation

$$\begin{cases} y'(t) = \lambda y(t) + \mu y(t-\tau), \quad t \geq t_0, \\ y(t) = \phi(t), \quad t \leq t_0, \end{cases} \tag{14.3}$$

As far as the neutral equations are concerned, we consider the constant coefficient and constant delay case

$$\begin{cases} y'(t) = \lambda y(t) + \mu y(t-\tau) + \nu y'(t-\tau), \quad t \geq t_0, \\ y(t) = \phi(t), \quad t \leq t_0, \end{cases} \tag{14.4}$$

Our aim is to find conditions, necessary and/or sufficient, on the coefficients of the above equations in order that they are *contractive*, that is

$$|y(t)| \leq \max_{x \leq t_0} |\phi(x)|, \quad t \geq t_0, \tag{14.5}$$

and/or *asymptotically stable*, that is

$$\lim_{t \to +\infty} y(t) = 0. \tag{14.6}$$

15. Analysis of the test equations (14.1) and (14.2)

In order to carry out our analysis, we need the following preliminary results regarding ODEs with *forcing terms*.

****

Proposition 15.1 *Consider the scalar ODE with the forcing term*

$$\begin{cases} y'(t) = \lambda(t)y(t) + \Re(\lambda(t))g(t), & t \geq t_0, \\ y(t_0) = y_0, \end{cases} \tag{15.1}$$

where $\lambda(t)$ and $g(t)$ are continuous complex valued functions. If

$$\Re(\lambda(t)) \leq 0, \quad t \geq t_0, \tag{15.2}$$

then

$$|y(t)| \leq E(t_0,t)|y_0| + (1 - E(t_0,t)) \max_{t_0 \leq x \leq t} |g(x)|, \quad t \geq t_0, \tag{15.3}$$

where $E(t_1,t_2) = \exp\left(\int_{t_1}^{t_2} \Re(\lambda(x))\mathrm{d}x\right)$.

****

****

Corollary 15.1 *Under the hypotheses of Proposition 15.1, the solution $y(t)$ of (15.1) satisfies*

$$|y(t)| \leq \max\left\{|y_0|, \max_{t_0 \leq x \leq t} |g(x)|\right\}, \quad t \geq t_0. \tag{15.4}$$

****

****

Proposition 15.2 *Consider the scalar ODE with the forcing term*

$$\begin{cases} y'(t) = \lambda y(t) + \Re(\lambda) g(t), & t \geq t_0, \\ y(t_0) = y_0, \end{cases} \tag{15.5}$$

where $\lambda \in \mathbb{C}$ and $g(t)$ is a continuous complex valued function. If $\Re(\lambda) \leq 0$, then

$$|y(t)| \leq \mathrm{e}^{\Re(\lambda)(t-t_0)}|y_0| + \left(1 - \mathrm{e}^{\Re(\lambda)(t-t_0)}\right) \max_{t_0 \leq x \leq t} |g(x)|, \quad t \geq t_0. \tag{15.6}$$

****

****

Corollary 15.2 *Under the hypotheses of Proposition 15.2, the solution $y(t)$ of (15.5) satisfies (15.4).*

****

Here are the contractivity and asymptotic stability results, which are proved by using induction on the macro intervals $[\bar{\xi}_{k-1}, \bar{\xi}_k]$, where the $\bar{\xi}_k$'s are the principal discontinuity points (see Definition 3.2).

****

Theorem 15.1 *If the coefficients of equations (14.1) and (14.2) are such that*

$$\Re(\lambda(t)) + |\mu(t)| \le 0, \quad t \ge t_0, \tag{15.7}$$

and

$$\Re(\lambda) + |\mu| \le 0, \tag{15.8}$$

respectively, then the equations are contractive (i.e. (14.5) holds) for any initial function $\phi(t)$ and for all delays $\tau(t)$ satisfying hypothesis (H_1).

****

Proof of Threorem 15.1. It can be easily seen that condition (15.7) is entirely equivalent to conditions (15.2) and

$$|\mu(t)| = r(t)\Re(\lambda(t)), \quad t \ge t_0, \tag{15.9}$$

with $-1 \le r(t) \le 0$.

It is clear that $|y(t)| \le \max_{x \le t_0} |\phi(x)|$ holds for $t \le t_0$. Now assume by induction that (14.5) holds for $t \le \bar{\xi}_{k-1}$ and consider the solution $y(t)$ for $t \in [\bar{\xi}_{k-1}, \bar{\xi}_k]$. Since $|r(x)| \le 1$, by Corollary 15.1 we get

$$\begin{aligned} |y(t)| &\le \max\left\{ |y(\bar{\xi}_{k-1})|, \sup_{\bar{\xi}_{k-1} \le x \le \bar{\xi}_k} (|r(x)| \cdot |y(x - \tau(x))|) \right\} \\ &\le \max\left\{ |y(\bar{\xi}_{k-1})|, \max_{x \le \bar{\xi}_{k-1}} |y(x)| \right\}. \end{aligned}$$

Hence, by the inductive hypothesis, (14.5) also holds for $t \in [\bar{\xi}_{k-1}, \bar{\xi}_k]$ and, thus, the proof is complete. ■

****

Theorem 15.2 *If the coefficients of equation (14.1) are such that*

$$\Re(\lambda(t)) \leq \Lambda_0 < 0, \quad t \geq t_0, \tag{15.10}$$

and

$$R \cdot \Re(\lambda(t)) + |\mu(t)| \leq 0, \quad t \geq t_0, \tag{15.11}$$

for some non-negative real number $R < 1$, and the coefficients of equation (14.2) are such that

$$\Re(\lambda) + |\mu| < 0, \tag{15.12}$$

then the solution $y(t)$, besides being contractive, is asymptotically stable (i.e. (14.6) holds) for any initial function $\phi(t)$ and for all delays $\tau(t)$ satisfying hypotheses (H_1) and (H_2). Moreover, if also hypothesis (H_3) is satisfied, then the convergence rate is at least of exponential type, i.e. at least like $\mathrm{e}^{-\alpha(t-t_0)}$ for some $\alpha > 0$.

****

Proof of Threorem 15.2. The proof proceeds inductively on the macro intervals between the principal discontinuity points.

Set $\bar{\xi}_0 = t_0$ and, for $t \leq \bar{\xi}_0$, define the constant function

$$\delta_0(t) = \max_{x \leq \bar{\xi}_0} |\phi(x)|$$

and the numbers

$$\Delta_{-1} = \Delta_0 = \delta_0(\bar{\xi}_0).$$

Observe that the interesting case is $\Delta_{-1} = \Delta_0 > 0$, otherwise (14.6) is obvious.

Then, for $k \geq 1$, on each interval $[\bar{\xi}_{k-1}, \bar{\xi}_k]$ inductively define the functions

$$\delta_k(t) = R\Delta_{h(k-1)} + \mathrm{e}^{\Lambda_0(t-\bar{\xi}_{k-1})}(\Delta_{k-1} - R\Delta_{h(k-1)}) \qquad (15.13)$$

and the numbers

$$\Delta_k = \delta_k(\bar{\xi}_k), \qquad (15.14)$$

where

$$h(k-1) = \max\{h \mid \bar{\xi}_h \leq t - \tau(t)\ \forall t \in [\bar{\xi}_{k-1}, \bar{\xi}_k]\}.$$

Observe that, since (H_1) implies

$$\bar{\xi}_k - \bar{\xi}_{k-1} \geq \tau_0, \qquad (15.15)$$

by (H_2) we have

$$\bar{\xi}_k \to +\infty \quad \text{and} \quad h(k) \to \infty \quad \text{as } k \to \infty. \qquad (15.16)$$

Of course, we have $\delta_k(\bar{\xi}_{k-1}) = \Delta_{k-1}$, so that the piecewise exponential function $\delta(t)$, defined on $[\xi_{-1}, +\infty)$ by setting $\delta(t) = \delta_k(t)$ if $t \in [\bar{\xi}_{k-1}, \bar{\xi}_k]$, $k \geq 0$, is continuous.

Now, by the inductive hypothesis, assume that $\Delta_{k-1} - R\Delta_{h(k-1)} > 0$ and $\Delta_{j-1} \geq \Delta_j, \forall j \leq k-1$ (these inequalities clearly hold for $k = 1$). Then, by (15.13) and (15.14), we have

$$\begin{aligned}
\Delta_k - R\Delta_{h(k)} &= R\Delta_{h(k-1)} + \mathrm{e}^{\Lambda_0(\bar{\xi}_k - \bar{\xi}_{k-1})}(\Delta_{k-1} - R\Delta_{h(k-1)}) - R\Delta_{h(k)} \\
&= R(\Delta_{h(k-1)} - \Delta_{h(k)}) + \mathrm{e}^{\Lambda_0(\bar{\xi}_k - \bar{\xi}_{k-1})}(\Delta_{k-1} - R\Delta_{h(k-1)}) \\
&> 0
\end{aligned}$$

and

$$\begin{aligned}
\Delta_{k-1} - \Delta_k &= \Delta_{k-1} - R\Delta_{h(k-1)} - \mathrm{e}^{\Lambda_0(\bar{\xi}_k - \bar{\xi}_{k-1})}(\Delta_{k-1} - R\Delta_{h(k-1)}) \\
&= (\Delta_{k-1} - R\Delta_{h(k-1)})(1 - \mathrm{e}^{\Lambda_0(\bar{\xi}_k - \bar{\xi}_{k-1})}) \\
&> 0.
\end{aligned}$$

Therefore, we have proved that the sequence $\{\Delta_k\}$ and, consequently, the function $\delta(t)$ are monotonically decreasing and that

$$\Delta_k - R\Delta_{h(k)} > 0, \quad k \geq 0. \tag{15.17}$$

Now we want to prove that

$$|y(t)| \leq \delta(t), \quad t \geq \bar{\xi}_{-1}. \tag{15.18}$$

The above inequality clearly holds for $t \in [\bar{\xi}_{-1}, \bar{\xi}_0]$. Now assume by induction that it holds for $t \in [\bar{\xi}_{-1}, \bar{\xi}_{k-1}]$ and consider the function $y(t)$ for $t \in [\bar{\xi}_{k-1}, \bar{\xi}_k]$. It can easily be seen that condition (15.11) implies condition (15.9) with $-R \leq r(t) \leq 0$. Therefore, by Proposition 15.1 and by the inductive hypothesis, we get

$$\begin{aligned}
|y(t)| &\leq E(\bar{\xi}_{k-1}, t)|y(\bar{\xi}_{k-1})| \\
&\quad + \left(1 - E(\bar{\xi}_{k-1}, t)) \sup_{\bar{\xi}_{k-1} \leq x \leq \bar{\xi}_k} (|r(x)| \cdot |y(x - \tau(x))|\right) \\
&\leq E(\bar{\xi}_{k-1}, t)|y(\bar{\xi}_{k-1})| \\
&\quad + (1 - E(\bar{\xi}_{k-1}, t))R \sup_{\bar{\xi}_{h(k-1)} \leq x \leq \bar{\xi}_{k-1}} |y(x)| \\
&\leq E(\bar{\xi}_{k-1}, t)\Delta_{k-1} + (1 - E(\bar{\xi}_{k-1}, t))R \sup_{\bar{\xi}_{h(k-1)} \leq x \leq \bar{\xi}_{k-1}} \delta(x) \\
&= E(\bar{\xi}_{k-1}, t)\Delta_{k-1} + (1 - E(\bar{\xi}_{k-1}, t))R\Delta_{h(k-1)} \\
&= R\Delta_{h(k-1)} + E(\bar{\xi}_{k-1}, t)(\Delta_{k-1} - R\Delta_{h(k-1)}).
\end{aligned}$$

Finally, by (15.10), (15.13) and (15.17), we obtain

$$|y(t)| \leq \delta_k(t), \quad \bar{\xi}_{k-1} \leq t \leq \bar{\xi}_k,$$

and, hence, (15.18) holds also for $t \in [\bar{\xi}_{k-1}, \bar{\xi}_k]$.

To conclude the proof, we have to show that the function $\delta(t)$ vanishes as $t \to +\infty$. Because of (15.16), it is sufficient

to prove that $\Delta_k \to 0$. To this end, observe that, by (15.13), (15.14), (15.17) and (15.15), the sequence $\{\Delta_k\}_{k\geq -1}$ is such that

$$\begin{cases} \Delta_k \leq S\Delta_{k-1} + R(1-S)\Delta_{h(k-1)}, & k \geq 1, \\ \Delta_{-1} = \Delta_0 > 0, \end{cases} \tag{15.19}$$

where $S = \mathrm{e}^{\Lambda_0 \tau_0} < 1$. This is a variable order linear difference inequality with two positive constant coefficients, the sum of which is < 1. Therefore, since the sequence $\{\Delta_k\}$ is convergent (because it is decreasing and positive), its limit can not be strictly positive. Thus $\Delta_k \to 0$ as $k \to \infty$.

Now, assume hypothesis (H_3) is satisfied. In this case, by (15.15) it turns out that $h(k-1) \geq k - m - 1$, where $m = \lceil \frac{\tau_1}{\tau_0} \rceil$. So, since the sequence $\{\Delta_k\}$ is decreasing, the difference inequality (15.19) yields

$$\Delta_k \leq S\Delta_{k-1} + R(1-S)\Delta_{k-m-1}, \quad k \geq m,$$

which is a linear difference inequality of order $m+1$ with constant coefficients. Since the sum of the (positive) coefficients in the right-hand side is < 1, all the characteristic roots are < 1 in modulus. Thus, if β is the greatest one, then we can conclude that $\Delta_k \to 0$ at least as $|\beta|^k$.

Finally, for a given $t \geq t_0$, we have that $t \in [\bar{\xi}_{k-1}, \bar{\xi}_k]$ for some $k \geq \frac{t-t_0}{\tau_1}$. Therefore, by (15.13) and (15.14), it follows that $\delta(t) \to 0$ at least as $\mathrm{e}^{-\alpha(t-t_0)}$, where $\alpha = -(1/\tau_1)\log|\beta|$. ■

We want to stress the fact that the bound (15.18) is sharp with respect to the whole class of DDEs satisfying the hypotheses (15.10) and (15.11). In order to see this, consider the piecewise linear delay

$$\tau^*(t) = \begin{cases} t-(k-2) & \text{for } k-1 \le t < k, \\ t-(k-1) & \text{for } t = k, \end{cases} \quad k = 1, 2, \dots, \tag{15.20}$$

so that the deviating argument is the piecewise constant function

$$t - \tau^*(t) = \begin{cases} k-2 & \text{for } k-1 \le t < k, \\ k-1 & \text{for } t = k, \end{cases} \quad k = 1, 2, \dots.$$

Now it is easy to see that, for the linear equation with constant coefficients

$$\begin{cases} y'(t) = \Lambda_0 y(t) - R \cdot \Lambda_0 y(t - \tau^*(t)), & t \ge t_0, \\ y(t) \equiv 1, & t \le t_0, \end{cases}$$

the solution $y(t)$ is given just by the corresponding function $\delta(t)$ introduced in the proof of Theorem 15.2.

The delay $\tau^*(t)$ fulfils hypothesis (H_1), but is not continuous. Nevertheless, it is clear that it can be viewed as the pointwise limit of a sequence of continuous delays. In other words, there exist continuous delays that fulfill (H_1) and are arbitrarily close (pointwise) to (15.20). As a consequence, there exist linear DDEs whose solutions $y(t)$ are arbitrarily close to the corresponding function $\delta(t)$.

It is also worth remarking that hypothesis (H_2) is necessary for the asymptotic stability with respect to the whole class of DDEs satisfying the hypotheses (15.10) and (15.11). In fact, if it does not hold, there exist only a finite number of discontinuity points $\bar{\xi}_k$. So, if $\bar{\xi}_K$ is the last one, on the last interval $[\bar{\xi}_K, +\infty)$ the function $\delta(t) = \delta_{K+1}(t)$ is bounded from below by the quantity $R\Delta_{h(K)} > 0$ (see (15.13)).

Since conditions (15.10) and (15.11) are more restrictive than (15.7), they guarantee asymptotic stability along with contractivity. On the other hand, contractivity is not necessary for asymptotic stability even in the linear constant coefficient case, as shown by the equation

$$\begin{cases} y'(t) = \frac{1}{2}y(t) - y(t-1), & t \geq 0, \\ y(t) = \phi(t) = t + 1, & -1 \leq t \leq 0. \end{cases} \tag{15.21}$$

According to the theory developed in the forthcoming Section 16.1, the solution $y(t)$ is asymptotically stable but, in a right neighborhood of 0, the solution turns out to be larger than $\max_{x \leq 0} |\phi(x)| = 1$. In fact, $y(0) = 1$ and $y'(0)^+ = \frac{1}{2} > 0$.

16. Analysis of the test equation (14.3)

It is useful to point out that, for the constant delay equation (14.3) the stability analysis may be done directly by studying the roots of the *characteristic equation*

$$\zeta - \lambda - \mu e^{-\tau\zeta} = 0. \tag{16.1}$$

It is known that such an equation has infinitely many roots ζ_i, each of which has a certain multiplicity m_i. They lie in the complex half-plane $\Re(\lambda) < \alpha$ for some real α and their real parts accumulate at $-\infty$. Therefore, in any vertical strip of the complex plane there are only a finite number of roots.

It is also known that the solution to (14.2) has an expansion of the form

$$y(t) = \sum_{i=1}^{\infty} \sum_{n_i=0}^{m_i-1} \alpha_{i,n_i} t^{n_i} e^{\zeta_i t}, \tag{16.2}$$

where the coefficients α_{i,n_i} are determined by the initial function $\phi(t)$. In view of the representation (16.2), it is easy to understand that a necessary and sufficient condition for the asymptotic stability of (14.2) is that all the roots ζ_i of (16.1) be such that $\Re(\zeta_i) < 0$.

Now, we shall see that such a condition is guaranteed if (15.12) holds. In fact, if we assume by contradiction that there exists a root ζ^* of (16.1) such that $\Re(\zeta^*) \geq 0$, we have

$$\begin{aligned} 0 \leq \Re(\zeta^*) &= \Re(\lambda) + \Re(\mu e^{-\tau\zeta^*}) \\ &\leq \Re(\lambda) + |\mu| \cdot |e^{-\tau\zeta^*}| \\ &\leq \Re(\lambda) + |\mu|, \end{aligned}$$

which gives a contradiction.

So, we can say that results obtained by Theorem 15.2 on asymptotic stability for all delays of (14.2) are not worse than those obtained by the characteristic equation (in case of constant delay).

We have seen that Theorems 15.1 and 15.2 give conditions that are sufficient for contractivity and asymptotic stability independently of the particular delay $\tau(t)$. From a theoretical and practical point of view, it is interesting to investigate whether or not all or some of the above conditions are also necessary.

Again we must distinguish two cases: either when a particular delay $\tau(t)$ is considered or when the contractivity and/or asymptotic stability properties are requested for all delays of a certain class (for example, constant delays).

We confine ourselves to considering the test equation (14.2), where the delay $\tau(t)$ satisfies the additional hypothesis (H_4).

****

Proposition 16.1 *Assume that the contractivity property (14.5) holds for equation (14.2) with a fixed delay $\tau(t)$ satisfying hypotheses (H_1) and (H_4). Then the coefficients λ and μ satisfy condition (15.8).*

****

Proof of Proposition 16.1. Since hypotheses (H_1) and (H_4) are satisfied, the deviating argument $\alpha(t) = t - \tau(t)$ is a one-to-one map from $[t_0, \xi_1]$ to $[t_0 - \tau(t_0), t_0]$. Thus, we can consider the inverse function $\beta(t)$ from $[t_0 - \tau(t_0), t_0]$ onto $[t_0, \xi_1]$ and define the initial function

$$\phi^*(t) = \begin{cases} \mathrm{e}^{i\Im(\lambda)(\beta(t)-t_0)} & \text{for } t_0 - \tau(t_0) \le t < t_0, \\ \mu/|\mu| & \text{for } t = t_0, \end{cases}$$

which may be discontinuous at t_0. The solution in the first interval $[t_0, \xi_1]$ is given by

$$\begin{aligned} y^*(t) &= \mathrm{e}^{\lambda(t-t_0)} \left(\phi^*(t_0) + \mu \int_{t_0}^{t} \mathrm{e}^{-\lambda(x-t_0)} \phi^*(x - \tau(x)) \mathrm{d}x \right) \\ &= \mathrm{e}^{\lambda(t-t_0)} \left(\mu/|\mu| + \mu \int_{t_0}^{t} \mathrm{e}^{-\Re(\lambda)(x-t_0)} \mathrm{d}x \right) \end{aligned}$$

and, therefore, we get

$$|y^*(\xi_1)| = \begin{cases} 1 + |\mu|(\xi_1 - t_0) & \text{if } \Re(\lambda) = 0, \\ \left| \dfrac{\Re(\lambda)\mathrm{e}^{\Re(\lambda)(\xi_1 - t_0)} + |\mu|(\mathrm{e}^{\Re(\lambda)(\xi_1 - t_0)} - 1)}{\Re(\lambda)} \right| & \text{if } \Re(\lambda) \neq 0. \end{cases}$$

Now it can be easily seen that, if (15.8) does not hold, i.e. $\Re(\lambda) + |\mu| > 0$, then $|y^*(\xi_1)| > 1$. As we observed, the initial function $\phi^*(t)$ may not be continuous at t_0. However, it is clear that, for any $\epsilon > 0$, we can find a continuous initial function $\phi_\epsilon(t)$ such that the corresponding solution $y_\epsilon(t)$ satisfies the inequality $\max_{t_0 \le t \le \xi_1} |y_\epsilon(t) - y^*(t)| < \epsilon$. This completes the proof. ■

Table 9: Contractivity scheme for $y'(t) = \lambda y(t) + \mu y(t-\tau(t))$ with $\lambda, \mu \in \mathbb{C}$.

$\lvert y(t)\rvert \leq \max_{x\leq t_0} \lvert\phi(x)\rvert$ for all delays $\tau(t)$ satisfying (H_1)
$\Updownarrow$
$\Re(\lambda) + \lvert\mu\rvert \leq 0$
$\Updownarrow$
$\lvert y(t)\rvert \leq \max_{x\leq t_0} \lvert\phi(x)\rvert$ for fixed delay $\tau(t)$ satisfying (H_1) and (H_4)

The foregoing proposition states that condition (15.8) is necessary for contractivity for a particular fixed delay satisfying (H_1) and (H_4) and, *a fortiori*, for all delays satisfying (H_1). We can summarize the results on contractivity for equation (14.2) in Table 9.

A different case is that of asymptotic stability where, in order to analyze the necessity of condition (15.12), we confine our analysis to the case of constant delays. To do this, we need a complete description of the stability region $\mathcal{S}_\tau$ defined as follows.

Definition 16.1 *For any fixed τ, the* asymptotic stability region $\mathcal{S}_\tau$ *of equation (14.3) is the set of pairs (λ, μ) such that the corresponding solution is asymptotically stable for any initial function $\phi(t)$.*

16.1. Description of the asymptotic stability region $\mathcal{S}_\tau$ for real coefficients

Let us first consider the case of real λ and μ. A deeper analysis of the roots of the characteristic equation (16.1) reveals that, for a fixed value of the delay τ, the region of stability is larger than the cone $\lambda + |\mu| < 0$ derived by the inequality (15.12). In fact, the region of asymptotic stability $\mathcal{S}_\tau$ is given by the set of pairs (λ, μ) such that

$$\lambda < -\mu \quad \text{and} \quad \sqrt{\mu^2 - \lambda^2} < \frac{1}{\tau} \arccos(-\lambda/\mu).$$

In Figure 25 we show, for real λ and μ, the form of the asymptotic stability region in the (λ, μ)-plane.

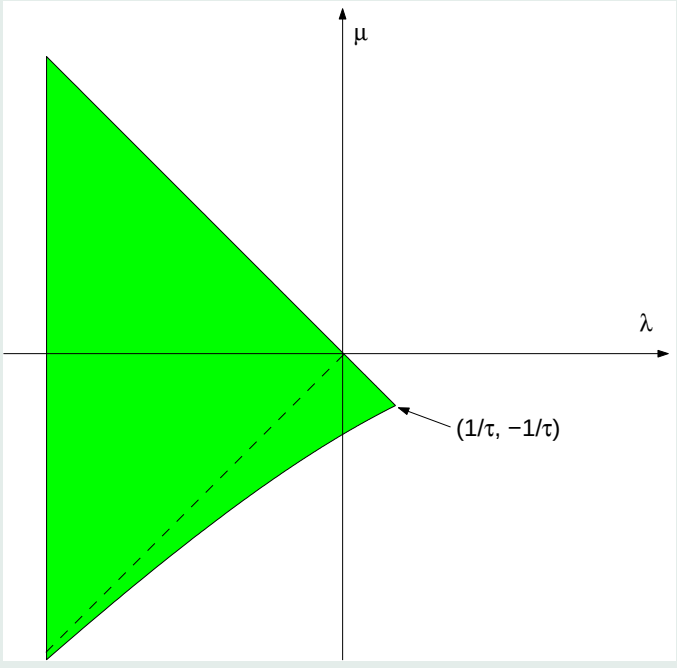

Figure 25: Asymptotic stability region $\mathcal{S}_\tau$ for equation (14.2) with constant delay τ in the real (λ, μ)-plane.

Table 10: Asymptotic stability scheme for $y'(t) = \lambda y(t) + \mu y(t-\tau)$ with $\lambda, \mu \in \mathbb{R}$.

$$\begin{array}{c} \lim_{t\to+\infty} y(t) = 0 \text{ for all constant delays } \tau \\ \Updownarrow \\ \lambda \leq \mu < -\lambda \\ \Downarrow \\ (\lambda, \mu) \in \mathcal{S}_\tau \\ \Updownarrow \\ \lim_{t\to+\infty} y(t) = 0 \text{ for fixed constant delay } \tau \end{array}$$

Equation (15.21), where $\lambda = \frac{1}{2}$, $\mu = -1$ and $\tau = 1$, provides an example where the solution goes to zero despite the coefficients not fulfilling condition (15.12). Thus, for asymptotic stability condition (15.12) is not necessary for a fixed constant delay τ. However, when we let the delay τ go to $+\infty$, we see that the region of asymptotic stability tends to the region described by

$$\lambda \leq \mu < -\lambda. \tag{16.3}$$

Therefore, condition (16.3), which is slightly weaker than condition (15.12), is necessary for the asymptotic stability of equation (14.2) for all constant delays τ. The results are summarized in Table 10.

16.2. Description of the asymptotic stability region $\mathcal{S}_\tau$ for complex coefficients

Now we describe the set $\mathcal{S}_\tau$ for complex λ and μ. Remember that, by using the characteristic equation (16.1), it turns out that

$$\mathcal{S}_\tau = \left\{(\lambda,\mu)\in\mathbb{C}^2 \mid \zeta-\lambda-\mu e^{-\tau\zeta}=0 \quad\Rightarrow\quad \Re(\zeta)<0\right\}.$$

Let us fix $\mu\in\mathbb{C}$ and consider the set

$$S_\tau(\mu) = \left\{\zeta-\mu e^{-\tau\zeta} \mid \zeta\in\mathbb{C},\quad \Re(\zeta)\geq 0\right\}.$$

It is evident that

$$(\lambda,\mu)\in S_\tau \quad \text{if and only if} \quad \lambda\notin S_\tau(\mu).$$

Long and sophisticated calculations show that the set $S_\tau(\mu)$ is the union of the open half-plane $\Re(\lambda)<-|\mu|$ and of a 2π-periodic "crest", as depicted in Figures 26 and 27.

The peak points of the crest have a real part $\Re(\lambda)=D_\tau(\mu)$, where the pair $(D_\tau(\mu),-|\mu|)$ belongs to the border of the real stability region depicted in Figure 25.

As for the region of asymptotic stability for all delays, observe that the value $D_\tau(|\mu|)$ tends to $-|\mu|$ as $\tau\to+\infty$. So, the border of the region moves to the line $\Re(\lambda)=-|\mu|$ and then, for any pair (λ,μ) such that $\Re(\lambda)<-|\mu|$, the solution asymptotically vanishes for all delays. The complete region is obtained by adding the following subset of the border:

$$\{(\lambda,\mu)\in\mathbb{C}^2 \mid \lambda\in\mathbb{R},\ |\mu|=-\lambda,\ \lambda+\mu\neq 0\}. \tag{16.4}$$

The results are summarized in Table 11.

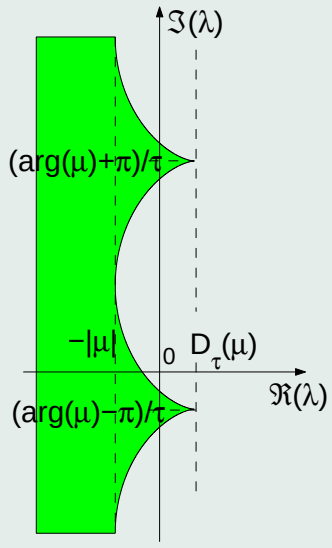

Figure 26: The restriction to the complex λ-plane of the asymptotic stability region $\mathcal{S}_\tau$ for a fixed value of μ with $1/\tau < |\mu| < \pi/(2\tau)$.

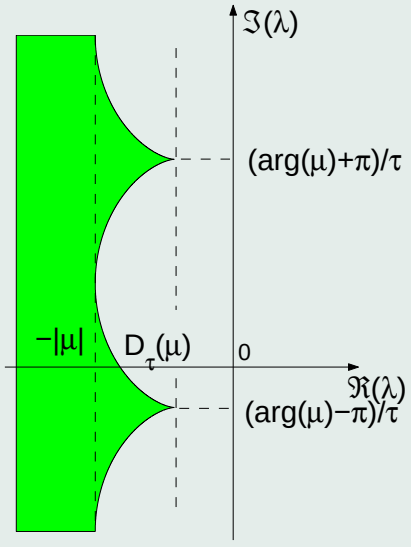

Figure 27: The restriction to the complex λ-plane of the asymptotic stability region $\mathcal{S}_\tau$ for a fixed value of μ with $\pi/(2\tau) < |\mu|$.

Table 11: Asymptotic stability scheme for $y'(t) = \lambda y(t) + \mu y(t-\tau)$ with $\lambda, \mu \in \mathbb{C}$.

$$\begin{array}{c} \lim_{t\to+\infty} y(t) = 0 \text{ for all constant delays } \tau \\ \Updownarrow \\ \Re(\lambda) < -|\mu| \text{ or } \lambda \in \mathbb{R},\ |\mu| = -\lambda,\ \lambda + \mu \neq 0 \\ \Downarrow \\ (\lambda, \mu) \in \mathcal{S}_\tau \\ \Updownarrow \\ \lim_{t\to+\infty} y(t) = 0 \text{ for fixed constant delay } \tau \end{array}$$

An alternative approach for describing the set $\mathcal{S}_\tau$ consists in fixing a value of λ and looking for the possible values of μ. The analysis, based on the boundary locus technique, is equivalent to the previous approach.

The parametric equations of the boundary of the stability regions are

$$\begin{cases} \Re(\mu) = -\Re(\lambda)\cos\theta + \Im(\lambda)\sin\theta - \frac{\theta\sin\theta}{\tau}, \\ \Im(\mu) = -\Im(\lambda)\cos\theta - \Re(\lambda)\sin\theta + \frac{\theta\cos\theta}{\tau}, \end{cases}$$

where the parameter θ varies in $[0, +\infty)$. Observe that the boundary is the "sum" of a circle centered in the origin with radius $|\lambda|$ and of the transcendental curve $\left(-\frac{\theta\sin\theta}{\tau}, \frac{\theta\cos\theta}{\tau}\right)$, which does not depend on λ.

The form of the stability region in the complex μ-plane is reported in Figure 28 for a given complex value of λ.

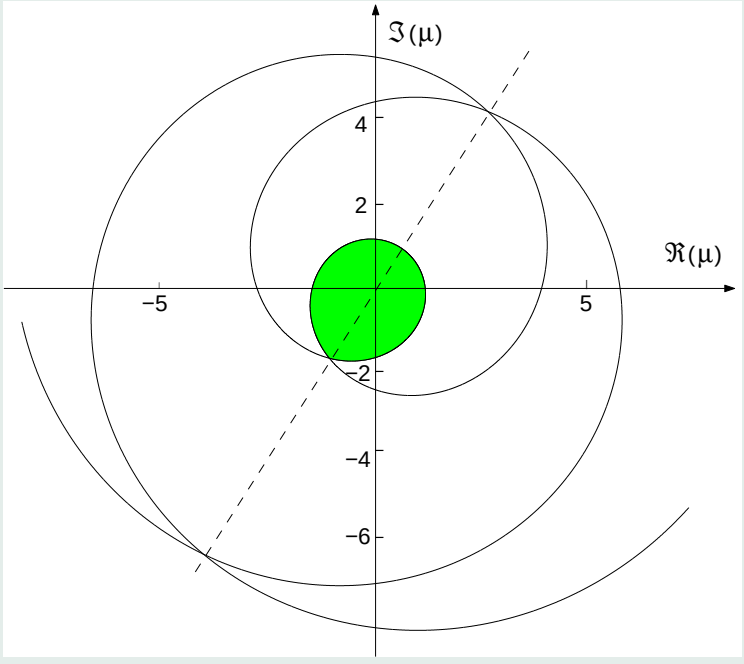

Figure 28: Asymptotic stability region for equation (14.2) with constant delay $\tau = 1$ in the complex μ-plane for $\lambda = -1 + i$.

17. Analysis of the test equation (14.4)

For the constant coefficient and constant delay case (14.4) with $\lambda, \mu, \nu \in \mathbb{R}$ and $|\nu| < 1$, the necessary and sufficient condition for asymptotic stability for all delays is again

$$\lambda \leq \mu < -\lambda.$$

For complex coefficients with $|\nu| < 1$, the asymptotic stability condition can be stated in some equivalent forms, such as

$$\begin{cases} \Re(\lambda) < 0, \\ \big(\Im(\mu\nu + \lambda)\big)^2 + (1 - |\nu|^2)(|\mu|^2 - |\lambda|^2) \leq 0, \\ \mu \neq -\lambda, \end{cases}$$

or

$$\begin{cases} \Re(\lambda) < 0, \\ |\lambda\nu + \mu| < \big|2\Re(\lambda) + |\lambda\bar{\nu} - \bar{\mu}|\big|, \\ \mu \neq -\lambda. \end{cases} \tag{17.1}$$

Despite appearing different, the two expressions are equivalent.

By choosing the option

$$2\Re(\lambda) + |\lambda\bar{\nu} - \bar{\mu}| \leq 0$$

in (17.1), we get the following sufficient condition for asymptotic stability:

$$|\lambda\bar{\nu} - \bar{\mu}| + |\lambda\nu + \mu| < -2\Re(\lambda). \tag{17.2}$$

It is not difficult to verify that

$$\lambda \in \mathbb{R} \quad \Longrightarrow \quad 2\Re(\lambda) + |\lambda\bar{\nu} - \bar{\mu}| \leq 0,$$

so that (17.2) characterizes the asymptotic stability for $\lambda \in \mathbb{R}$ and $\mu, \nu \in \mathbb{C}$. Observe that (17.2) turns out to be equivalent to

$|\mu| < -\lambda$ and $|\nu| < 1$ for real coefficients, and to $|\mu| < -\Re(\lambda)$ if $\nu = 0$, both of which are also necessary for asymptotic stability apart from some points on the border.

As for the asymptotic stability analysis of (14.4) for fixed delay, we have the following definition.

Definition 17.1 *For any fixed τ, the* asymptotic stability region $\mathcal{NS}_\tau$ *of equation (14.4) is the set of triplets (λ, μ, ν) such that the corresponding solution is asymptotically stable for any initial function $\phi(t)$.*

For the real case, the region $\mathcal{NS}_\tau$ turns out to be the subset of $\mathbb{R}^3$ bounded by the planes of equations $\nu = 1$, $\nu = -1$ and $\lambda + \mu = 0$, and by the transcendental surface

$$\Gamma_\tau = \left\{ (\lambda, \mu, \nu) \;\mid\; \nu x^2 + \mu^2 = \lambda^2 + x^2,\; x = \frac{1}{\tau} \arctan \frac{x(\mu + \lambda\nu)}{\lambda\mu - \nu x^2},\; x \in \mathbb{R} \right\}.$$

For any fixed value of ν with $|\nu| < 1$, the corresponding section of Γ_τ is

$$\Gamma_\tau(\nu) = \Big\{ \big(\lambda(\nu, \theta), \mu(\nu, \theta)\big) \;\mid\; \theta \in (0, \pi) \Big\},$$

where

$$\begin{cases} \lambda(\nu, \theta) = \dfrac{\theta}{\tau \sin\theta}(\cos\theta - \nu), \\ \\ \mu(\nu, \theta) = \dfrac{\theta}{\tau \sin\theta}(\nu \cos\theta - 1). \end{cases}$$

Note that the curve $\Gamma_\tau(\nu)$ intersects the line $\lambda + \mu = 0$ at the point

$$\left(\frac{1 - \nu}{\tau}, -\frac{1 - \nu}{\tau} \right),$$

where the border of the stability region has a cusp (see Figure 29).

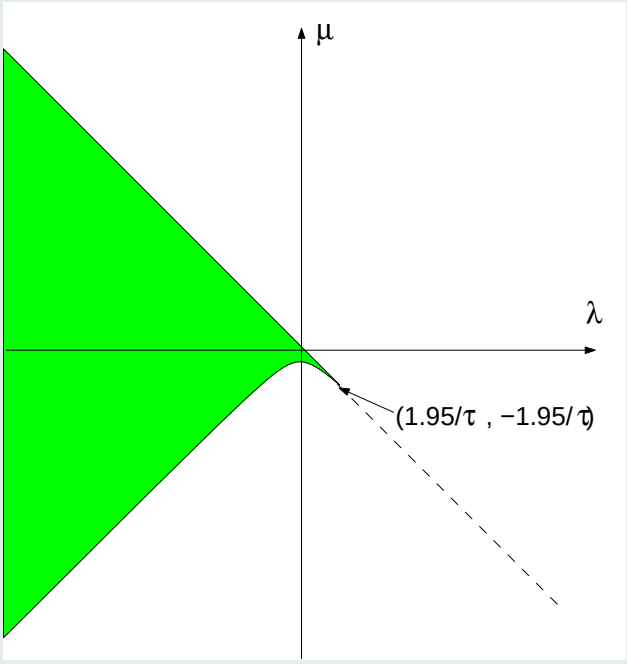

Figure 29: Section of the asymptotic stability region $\mathcal{NS}_\tau$ of equation (14.4) for $\nu = -0.95$ in the real (λ, μ)-plane.

Summer School on Delay Differential Equations and Applications

Dobbiaco (BZ), Italy, June 26–30, 2006

The numerical solution of delay differential equations

M. Zennaro
Dipartimento di Matematica e Informatica
Università di Trieste

Lecture 6: Stability analysis of DDE methods

Main reference for this Lecture:
A. Bellen and M. Zennaro, Numerical Methods for Delay Differential Equations, Numerical Mathematics and Scientific Computation, Oxford Science Publications, Oxford University Press, 2003 (Chapters 8 and 10)

18. Generalizations of A-stability to DDEs

In this lecture we analyze the stability of the RK methods for DDEs proposed in Lecture 4.

The analysis is confined to the case of time dependend delays $\tau(t)$, even if some results could be extended to state dependent delays. Moreover, we confine ourselves to the simplest test equation

$$\begin{cases} y'(t) = \lambda y(t) + \mu y(t-\tau), & t \geq t_0, \\ y(t) = \phi(t), & t \leq t_0, \end{cases} \tag{18.1}$$

where $\lambda, \mu \in \mathbb{C}$ and τ is a constant delay, which leads to generalizations of the concept of A-stability for ODEs that, in turn, is based on the simplest scalar linear autonomous test equation $y'(t) = \lambda y(t)$, $y(t_0) = y_0$.

This concept spreads out in many directions as we are interested, for example, in *contractivity* or *asymptotic stability*, which for DDEs do not necessarily coincide as it happens in the ODE case. The various stability definitions may also differ according to whether the constant stepsize is required to be a submultiple of the delay or not. These two occurrences are not equivalent.

Finally, we are interested in characterizing methods that are contractive and/or stable *for all delays* (*delay independent* stability) or *for fixed delay* (*delay dependent* stability). As it was explained in Lecture 5, whereas for contractivity the two cases do not differ from each other, the classes of equations that are asymptotically stable in the two senses are actually different and, therefore, the same is expected for numerical methods.

More precisely, we have seen that the condition

$$\Re(\lambda) + |\mu| \leq 0 \tag{18.2}$$

characterizes those equations of the type (18.1) which fulfil the contractivity property

$$|y(t)| \leq \max_{x \leq t_0} |\phi(x)|, \quad t \geq t_0, \tag{18.3}$$

for any fixed value of the delay τ. Because condition (18.2) is independent of τ, it also characterizes the contractivity for all delays.

Moreover, we have seen that the stronger condition

$$\Re(\lambda) + |\mu| < 0 \tag{18.4}$$

essentially characterizes the asymptotic stability of (18.1) for all delays τ, i.e.

$$\lim_{t \to +\infty} y(t) = 0$$

for all initial functions $\phi(t)$.

On the contrary, the asymptotic stability of (18.1) for a fixed value of τ is ensured if and only if the coefficients λ and μ belong to the larger region $\mathcal{S}_\tau$ (see Definition 16.1) for both the real and complex cases.

Therefore, it is natural to ask the numerical method for the preservation of the asymptotic stability property

$$\lim_{n \to \infty} y_n = 0$$

and/or the contractivity property

$$|y_n| \leq \max_{x \leq t_0} |\phi(x)|, \quad n \geq 0,$$

under the same conditions that guarantee such properties for the exact solution.

We start with the definitions regarding asymptotic stability.

Definition 18.1 *The* P-stability region *of a numerical method for DDEs is the set S_P of pairs of complex numbers (α, β), $\alpha = h\lambda$, $\beta = h\mu$, such that the discrete numerical solution $\{y_n\}_{n\geq 0}$ of (18.1), obtained with constant stepsize h under the constraint*

$$h = \tau/m, \quad m \geq 1, \ m \text{ integer}, \tag{18.5}$$

satisfies

$$\lim_{n\to\infty} y_n = 0 \tag{18.6}$$

for all constant delays τ and all initial functions $\phi(t)$.

Definition 18.2 *A numerical method for DDEs is* P-stable *if*

$$S_P \supseteq \{(\alpha, \beta) \in \mathbb{C}^{\,2} \mid \Re(\alpha) + |\beta| < 0\}.$$

In other words, a numerical method is P-stable if, for any stepsizes h satisfying (18.5), it preserves the asymptotic stability of (18.1) whenever $\Re(\lambda) < -|\mu|$, i.e. whenever (18.1) is asymptotically stable for all delays. Note that this definition disregards the part of the stability region of (18.1) lying on the boundary (see (16.4)).

Removing the constraint (18.5) leads to the following stronger concept of stability.

Definition 18.3 *The* GP-stability region *of a numerical method for DDEs is the set S_{GP} of pairs of complex numbers (α, β), $\alpha = h\lambda$, $\beta = h\mu$, such that the discrete numerical solution $\{y_n\}_{n\geq 0}$ of (18.1), obtained with constant stepsize h, satisfies (18.6) for all constant delays τ and all initial functions $\phi(t)$.*

Definition 18.4 *A numerical method for DDEs is* GP-stable *if*

$$S_{GP} \supseteq \{(\alpha, \beta) \in \mathbb{C}^2 \mid \Re(\alpha) + |\beta| < 0\}.$$

Observe that the constraint (18.5) defines, indeed, a constrained mesh (see Definition 11.1). Therefore, the study of the P-stability properties is confined to methods implemented under the constrained mesh strategy. On the other hand, when we remove constraint (18.5) the study of the GP-stability properties might be more appropriate. However, in both cases, the stepsize is assumed to be constant, and hence the stability properties of the numerical method implemented with varying stepsize are not completely captured by such investigations.

In order to fill in this gap, we give a definition of stability in which the stepsize is allowed to be variable. Moreover, in order to generalize our stability analysis further, in this new definition we also allow the delay to be variable. Therefore, we consider the more general test equation with variable delay

$$\begin{cases} y'(t) = \lambda y(t) + \mu y(t - \tau(t)), & t \geq t_0, \\ y(t) = \phi(t), & t \leq t_0. \end{cases} \tag{18.7}$$

In Section 14 we saw that, under hypotheses (H_1), (H_2) and (H_4), condition (18.4) also implies asymptotic stability for the variable delay equation (18.7).

Definition 18.5 *A numerical method for DDEs is* fully P-stable *(in short,* FP-stable*) if the discrete numerical solution of (18.7), obtained with any mesh Δ, satisfies (18.6) for all initial functions $\phi(t)$ and all delays $\tau(t)$ satisfying (H_1), (H_2) and (H_4), whenever (18.4) holds.*

It is clear that $S_{GP} \subseteq S_P$, that an FP-stable method is also GP-stable and that a GP-stable method is P-stable, too. Moreover, a P-stable method for DDEs is A-stable for ODEs.

$$\boxed{\text{FP-stability} \Rightarrow \text{GP-stability} \Rightarrow \text{P-stability} \Rightarrow \text{A-stability}}$$

It is worth remarking that, in the foregoing definitions of a stability region, the conditions on the parameters h, λ and μ were given in terms of the pair $(h\lambda, h\mu)$. The correctness of such a choice is not obvious a priori despite being the straight generalization of the definition of the A-stability region for ODEs. Indeed, we shall see that, since the conditions on the numerical solution $\{y_n\}_{n\geq 0}$ are required for all delays τ, the choice is consistent.

This is not the case in the definition of the stronger property of asymptotic stability for fixed delay. Contrary to the concept of P-stability, it is relevant to the larger class of equations (18.1) which are asymptotically stable for a fixed value of the delay (see Definition 16.1).

Definition 18.6 *For any fixed τ, the* D_τ-stability region *of a numerical method for DDEs is the set S_{D_τ} of triplets $(h, \lambda, \mu) \in \mathbb{R}^+ \times \mathbb{C}^2$, where h fulfils the constraint (18.5), such that the discrete numerical solution $\{y_n\}_{n\geq 0}$ of (18.1) obtained with constant stepsize h satisfies*

$$\lim_{n\to\infty} y_n = 0$$

for all initial functions $\phi(t)$.

Definition 18.7 *A numerical method for DDEs is* D-stable *if, for any fixed τ,*

$$(\lambda, \mu) \in \mathcal{S}_\tau \quad \Longrightarrow \quad (h, \lambda, \mu) \in S_{D_\tau}$$

for any constant stepsize h satisfying (18.5).

Observe that the concept of D-stability is stronger than P-stability. In fact, $S_{D_\tau} \supset \{(h, \lambda, \mu) \in \mathbb{R}^+ \times \mathbb{C}^2 \mid \Re(\lambda) + |\mu| < 0\}$ for all $\tau > 0$ and, hence, D-stability requires preservation of the asymptotic stability on a larger subclass of equations (18.1).

D-stability $\Rightarrow$ P-stability

Now we give the definitions regarding contractivity.

Definition 18.8 *The* P-contractivity region *of a numerical method for DDEs is the set C_P of pairs of complex numbers (α, β), $\alpha = h\lambda$, $\beta = h\mu$, such that the discrete numerical solution $\{y_n\}_{n\geq 0}$ of (18.1), obtained with constant stepsize h under the constraint (18.5), satisfies*

$$|y_n| \leq \max_{x\leq t_0} |\phi(x)|, \quad n \geq 0, \tag{18.8}$$

for all constant delays τ and all initial functions $\phi(t)$.

Definition 18.9 *A numerical method for DDEs is* P-contractive *if*

$$C_P \supseteq \{(\alpha, \beta) \in \mathbb{C}^{\,2} \mid \Re(\alpha) + |\beta| \leq 0\}.$$

Removing constraint (18.5) leads to the following stronger concept of contractivity.

Definition 18.10 *The* GP-contractivity region *of a numerical method for DDEs is the set C_{GP} of pairs of complex numbers (α, β), $\alpha = h\lambda$, $\beta = h\mu$, such that the discrete numerical solution $\{y_n\}_{n\geq 0}$ of (18.1) obtained with constant stepsize h satisfies (18.8) for all constant delays τ and all initial functions $\phi(t)$.*

Definition 18.11 *A numerical method for DDEs is* GP-contractive *if*

$$C_{GP} \supseteq \{(\alpha, \beta) \in \mathbb{C}^{\,2} \mid \Re(\alpha) + |\beta| \leq 0\}.$$

Recall that in Section 14 we saw that, under hypotheses (H_1) and (H_4), the solution $y(t)$ of (18.7) satisfies (18.3) if (18.2) holds.

Definition 18.12 *A numerical method for DDEs is* fully P-contractive *(in short,* FP-contractive*) if the discrete numerical solution* $\{y_n\}_{n\geq 0}$ *of (18.7), obtained with any mesh* Δ*, satisfies (18.8) for all initial functions* $\phi(t)$ *and all variable delays* $\tau(t)$ *under hypotheses* (H_1) *and* (H_4)*, whenever (18.2) holds.*

It is clear that $C_{GP} \subseteq C_P$, that an FP-contractive method is also GP-contractive and that a GP-contractive method is P-contractive, too.

$$\boxed{\text{FP-contractivity} \Rightarrow \text{GP-contractivity} \Rightarrow \text{P-contractivity}}$$

All the definitions throughout this section may be restricted to equations with real coefficients λ and μ and, in this case, the notations modify according to the notations for real ODEs, e.g. the *P(0)-stability region*, the *D(0)-stable* method, etc.

Not so obvious are the relationships among the various concepts of asymptotic stability and contractivity. However, in what follows we shall see that the sufficient conditions for contractivity are essentially sufficient to also guarantee asymptotic stability.

We treat in detail three cases, that we consider representative of three substantially different situations for equations with constant coefficients: P-stability (for all constant delays, using constant stepsize), D-stability (for fixed delay, using constant stepsize) and FP-contractivity/FP-stability (variable delay, using variable stepsize).

18.1. P-stability

The main difficult in the investigation of the P-stability properties of numerical methods for DDEs is that the final product is always a difference equation of arbitrary order m depending on the ratio τ/h.

The RK method for DDEs (10.2), (10.3), applied to the test equation (18.1) with constant stepsize h satisfying the constraint (18.5), takes the form

$$Y_{n+1}^i = y_n + h\sum_{j=1}^{s} a_{ij}(\lambda Y_{n+1}^j + \mu\eta(t_{n-m+1}^j)), \quad i = 1,\dots,s, \tag{18.9}$$

$$\eta(t_n+\theta h) = y_n + h\sum_{i=1}^{s} b_i(\theta)(\lambda Y_{n+1}^i + \mu\eta(t_{n-m+1}^i)). \tag{18.10}$$

With $b = [b_1,\dots,b_s]^T$, $b(\theta) = [b_1(\theta),\dots,b_s(\theta)]^T$, $e = [1,\dots,1]^T$ the unit s vector, I the s-dimensional identity matrix, $A = [a_{ij}]_{i,j=1}^s$, $\alpha = h\lambda$ and $\beta = h\mu$, after elimination of the stage values Y_{n+1}^i from (18.9) and computation of (18.10) for $\theta = c_1,\dots,c_s,1$, we get

$$\eta(t_{n+1}^i) = R_i(\alpha)y_n + \beta\sum_{j=1}^{s}(b(c_i)^T(I-\alpha A)^{-1})_j\eta(t_{n-m+1}^j), \quad i = 1,\dots,s, \tag{18.11}$$

where

$$R_i(\alpha) = 1 + \alpha b(c_i)^T(I-\alpha A)^{-1}e \tag{18.12}$$

and $(x)_j$ is to be understood as the jth component of the row vector x. Moreover, by (18.10) for $\theta = 1$, we get

$$y_{n+1} = R(\alpha)y_n + \beta\sum_{j=1}^{s}(b^T(I-\alpha A)^{-1})_j\eta(t_{n-m+1}^j), \tag{18.13}$$

where

$$R(\alpha) = 1 + \alpha b^T (I - \alpha A)^{-1} e \tag{18.14}$$

is the so-called *A-stability function* of the underlying discrete RK method (6.2), (6.3).

This pair of equations reduces to the recurrence relation with constant coefficients

$$\mathbf{H}_{n+1} = P(\alpha)\mathbf{H}_n + \beta Q(\alpha)\mathbf{H}_{n-m+1} \tag{18.15}$$

for the sequence of $(s+1)$-dimensional vectors

$$\mathbf{H}_n = [\eta(t_n^1), \ldots, \eta(t_n^s), y_n]^T,$$

where

$$P(\alpha) = \begin{bmatrix} 0 & e + \alpha B(I - \alpha A)^{-1} e \\ 0^T & 1 + \alpha b^T (I - \alpha A)^{-1} e \end{bmatrix},$$

$$Q(\alpha) = \begin{bmatrix} B(I - \alpha A)^{-1} & 0 \\ b^T (I - \alpha A)^{-1} & 0 \end{bmatrix},$$

and

$$B = [b_j(c_i)]_{i,j=1}^s.$$

The asymptotic behavior of the solutions of (18.15) is determined by the roots ζ of its characteristic equation

$$\det[\zeta^m I - \zeta^{m-1} P(\alpha) - \beta Q(\alpha)] = 0, \tag{18.16}$$

where, this time, I is the $(s+1)$-dimensional identity matrix.

By a small direct calculation, it is easy to check that, for all $\zeta \neq 0$ such that $\det[I - \alpha A - (\beta/\zeta^m)B] \neq 0$, the left-hand side of (18.16) can be factorized as follows:

$$\begin{aligned} &\det[\zeta^m I - \zeta^{m-1} P(\alpha) - \beta Q(\alpha)] \\ &\quad = \zeta^{ms+m-1} \det[I - \alpha A - (\beta/\zeta^m)B] \\ &\qquad \cdot \big(\zeta - R^*(\alpha, \beta/\zeta^m)\big). \end{aligned} \tag{18.17}$$

Therefore, instead of the roots ζ of the characteristic equation (18.16), we can equivalently consider the solutions of the algebraic equation

$$\zeta = R^*(\alpha, \beta/\zeta^m), \tag{18.18}$$

where the rational function

$$R^*(\alpha, z) = 1 + (\alpha + z)b^T(I - \alpha A - zB)^{-1}e$$

is called the *P-stability function*, provided that the following very mild assumption on the RK method holds.

Assumption 18.1 *The matrix $I - \alpha A - z^*B$ is singular if and only if z^* is a pole of $R^*(\alpha, z)$.*

Since $R^*(\alpha, 0) = R(\alpha)$, the *A-stability region* of the underlying discrete RK method is

$$S_A = \{\alpha \in \mathbb{C} \mid |R^*(\alpha, 0)| < 1\}.$$

Now, in the complex plane, consider the curve

$$\Gamma_\alpha = \{z \in \mathbb{C} \mid |R^*(\alpha, z)| = 1\}$$

and let

$$\sigma_\alpha = \inf_{z \in \Gamma_\alpha} |z|$$

be the distance of the curve Γ_α from the origin of the complex plane.

The following lemma, the proof of which is very long and technical, is the main tool of this P-stability analysis.

****

Lemma 18.1 *Consider the following three statements:*

(a) $\alpha \in S_A$ *and* $|\beta| < \sigma_\alpha$;

(b) *all the roots* ζ *of (18.18) are inside the unit circle for all* $m \geq 1$;

(c) $\alpha \in S_A$ *and* $|\beta| \leq \sigma_\alpha$.

Then (a) $\Longrightarrow$ (b) $\Longrightarrow$ (c).

****

If we define the set

$$\Sigma_P = \{(\alpha, \beta) \in \mathbb{C}^2 \mid \alpha \in S_A \quad \text{and} \quad |\beta| < \sigma_\alpha\},$$

we get the following characterization of the P-stability region.

****

Theorem 18.1 *The P-stability region* S_P *of the RK method for DDEs (10.2), (10.3) is such that* $\Sigma_P \subseteq S_P$. *Moreover, under the mild assumption on the method that* $\lim_{n\to\infty} \mathbf{H}_n = 0$ *if and only if* $\lim_{n\to\infty} y_n = 0$, *it also holds that* $S_P \subseteq \overline{\Sigma_P}$.

****

Proof of Theorem 18.1. Let $(\alpha, \beta) \in \Sigma_P$ and assume, by contradiction, that $(\alpha, \beta) \notin S_P$. Then the characteristic equation (18.16) has a root $\overline{\zeta}$ with $|\overline{\zeta}| \geq 1$ for some $m \geq 1$. On the other hand, by Lemma 18.1 all the roots of (18.18) are inside the unit circle for all $m \geq 1$. Moreover, by the definition of σ_α and since $|R^*(\alpha, 0)| < 1$, $\beta/\overline{\zeta}^m$ cannot be a pole of $R^*(\alpha, z)$. Therefore, by Assumption 18.1, equality (18.17) gives a contradiction.

Conversely, let $(\alpha, \beta) \in S_P$. Since $\lim_{n\to\infty} \mathbf{H}_n = 0$ for all possible solutions $\{\mathbf{H}_n\}_{n\geq 0}$ of (18.15), all the roots of the characteristic equation (18.16) are inside the unit circle for all $m \geq 1$. If, by contradiction, we assume that $(\alpha, \beta) \notin \overline{\Sigma_P}$, by Lemma 18.1 there exists a root $\overline{\zeta}$ of (18.18) with $|\overline{\zeta}| \geq 1$ for some $m \geq 1$. It is clear that $\beta/\overline{\zeta}^m$ cannot be a pole of $R^*(\alpha, z)$ and thus, again, by Assumption 18.1 equality (18.17) gives a contradiction. ■

We can conclude that, in general, the study of the P-stability properties of the RK methods for DDEs is reduced to a constrained minimum problem in the complex plane, namely to compute σ_α for $\alpha \in S_A$.

Observe that, if the method is natural, i.e. $A = B$ (see Definition 6.1), we have $R^*(\alpha, z) = R^*(\alpha + z, 0)$ and, hence,

$$R^*(\alpha, z) = R(\alpha + z).$$

Then, the set Γ_α is given by the complex numbers z such that $\alpha + z$ lies on ∂S_A, the boundary of the stability region S_A. Now, it is easy to see that

$$\sigma_\alpha = \mathrm{dist}(\alpha, \partial S_A),$$

where $\mathrm{dist}(\cdot, \cdot)$ denotes the Euclidean distance in the complex plane. Moreover, if the underlying discrete RK method is A-stable, that is $S_A \supseteq \{\alpha \in \mathbb{C} \mid \Re(\alpha) < 0\}$, then clearly it holds that

$$\sigma_\alpha \geq -\Re(\alpha),$$

and we have the following result.

****

Corollary 18.1 *A natural RK method for DDEs (10.4), (10.5) is P-stable if the underlying discrete RK method (6.2), (6.3) is A-stable.*

****

By virtue of Corollary 18.1, we can claim that the one-step collocation method at Gaussian points for DDEs is P-stable, since it is obviously natural and, as is well known, A-stable for ODEs.

Example 18.1 *As an application of the previous analysis, let us determine the P(0)-stability regions of the Θ-method*

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1-\Theta & \Theta \\ \hline & 1-\Theta & \Theta \end{array}$$

extended by linear interpolation. For constant stepsize h, it provides the nodal approximation

$$y_{n+1} = y_n + h\lambda\Big((1-\Theta)y_n + \Theta y_{n+1}\Big) + h\mu\Big((1-\Theta)\eta(t_n-\tau) + \Theta\eta(t_{n+1}-\tau)\Big). \tag{18.19}$$

In particular, with $\alpha = h\lambda$ and $\beta = h\mu$, for $h = \tau/m$ it takes the form

$$y_{n+1} = y_n + \alpha\Big((1-\Theta)y_n + \Theta y_{n+1}\Big) + \beta\Big((1-\Theta)y_{n-m} + \Theta y_{n+1-m}\Big).$$

As with linear ODEs, the Θ-method coincides with the collocation method at the point Θ

$$y_{n+1} = y_n + \alpha\Big((1-\Theta)y_n + \Theta y_{n+1}\Big) + \beta\eta(t_n + \Theta h - \tau), \tag{18.20}$$

where, by using linear interpolation, the retarded value $\eta(t_n + \Theta h - \tau)$ is still given by $\Big((1-\Theta)\eta(t_{n-m}) + \Theta\eta(t_{n+1-m})\Big)$.

Because the one-stage collocation method is a natural RK method with $A = B = \Theta$, the resulting P(0)-stability function turns out to be

$$R^*(\alpha, z) = 1 + \frac{\alpha + z}{1 - \alpha\Theta - z\Theta} = \frac{1 + (\alpha + z)(1-\Theta)}{1 - (\alpha + z)\Theta}$$

and satisfies Assumption 18.1.

The A(0)-stability region of the Θ-method is

$$S_{A(0)} = \left(\frac{2}{2\Theta - 1}, 0\right) \quad \text{for} \quad 0 \le \Theta < \tfrac{1}{2}$$

and

$$S_{A(0)} = (-\infty, 0) \cup \left(\frac{2}{2\Theta - 1}, \infty \right) \quad \text{for} \quad \tfrac{1}{2} \leq \Theta \leq 1.$$

Moreover, for any $\Theta \in [0, 1]$*, the boundary* ∂S_A *of the A-stability region is given, in the complex plane, by the circle centered in the real point* $1/(2\Theta - 1)$ *with radius equal to* $|1/(2\Theta - 1)|$.

Therefore, for real α*,*

$$\sigma_\alpha = \min \left\{ |\alpha|, |\alpha - \frac{2}{2\Theta - 1}| \right\}$$

and

$$\Sigma_{P(0)} = \left\{ (\alpha, \beta) \in \mathbb{R}^2 \mid \alpha \in S_{A(0)} \text{ and } |\beta| < \sigma_\alpha \right\}.$$

The regions of P(0)-stability are depicted in Figure 30 for $0 \leq \Theta < \frac{1}{2}$*, and in Figure 31 for* $\frac{1}{2} \leq \Theta \leq 1$*. We can conclude that both the* Θ*-method and collocation at one point* Θ *are P(0)-stable for all* $\frac{1}{2} \leq \Theta \leq 1$. $\Diamond$

A-stable CRK methods that are also P-stable are not necessarily confined to the subclass of natural CRK methods. If the CRK method $(A, b(\theta))$ is not natural, it is $A \neq B$ and the computation of σ_α for $\alpha \in S_A$ becomes more complicated.

However, it can be proved that Radau IA and Lobatto IIIC methods, as well as some *singly implicit RK* (SIRK) and *singly diagonally implicit RK* (SDIRK) methods that are not natural, are P-stable.

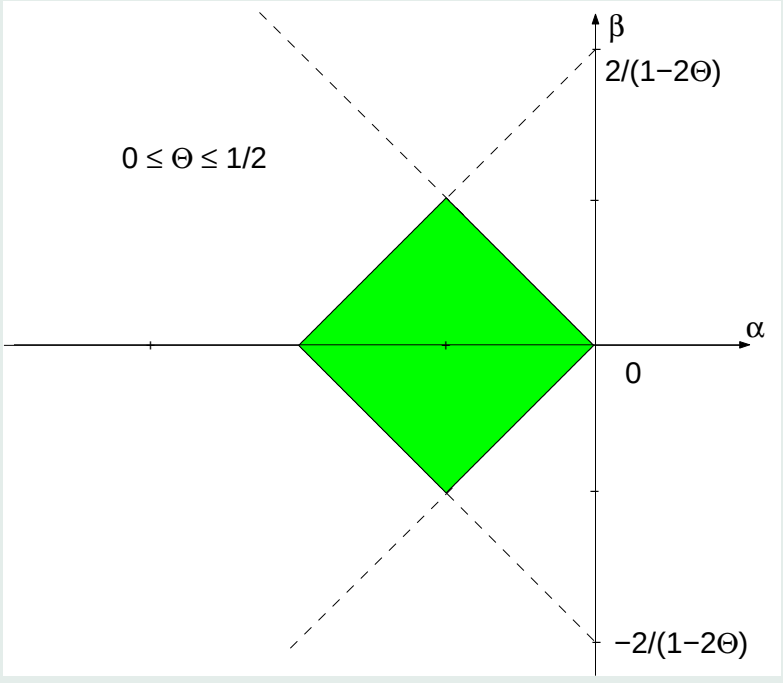

Figure 30: P(0)-stability region of the Θ-methods and the one-stage collocation method at the point Θ for $0 \leq \Theta < \frac{1}{2}$.

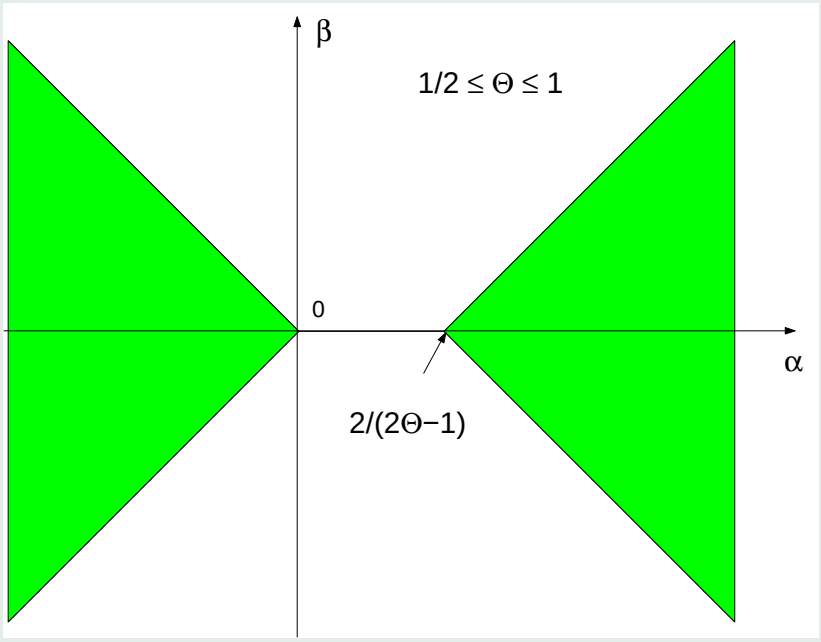

Figure 31: P(0)-stability region of the Θ-methods and the one-stage collocation method at the point Θ for $\frac{1}{2} \leq \Theta \leq 1$.

18.2. D-stability

Compared with the delay independent analysis, the asymptotic stability analysis for a fixed value of the delay is much more difficult. We confine ourselves to the class of natural RK methods for DDEs, for which equation (18.18) takes the form

$$\zeta = R\left(\alpha + \beta/\zeta^m\right),$$

where $R(\alpha)$ is the A-stability function of the underlying discrete RK method (see (18.14)) and hence, for $h = \tau/m$,

$$\zeta = R\left(\frac{1}{m}\left(\tau\lambda + \frac{\tau\mu}{\zeta^m}\right)\right).$$

So, the D_τ-stability region S_{D_τ} is described by the triplets (h, λ, μ), with $h = \tau/m$, such that

$$\zeta = R\left(\frac{1}{m}\left(\tau\lambda + \frac{\tau\mu}{\zeta^m}\right)\right) \quad \Longrightarrow \quad |\zeta| < 1.$$

By a change of variable $\xi = \tau\lambda + \tau\mu/\zeta^m$, the last condition is equivalent to

$$\xi - \tau\lambda - \frac{\tau\mu}{\left(R\left(\frac{\xi}{m}\right)\right)^m} = 0 \quad \Longrightarrow \quad \left|R\left(\frac{\xi}{m}\right)\right| < 1. \qquad (18.21)$$

Now consider the following property for the A-stability function $R(\alpha)$ of an A-stable RK method:

(P_m) The stability condition (18.21) holds for all $(\lambda, \mu) \in \mathcal{S}_\tau$.

We then observe that the RK method is D-stable if and only if (P_m) holds for all $m \geq 1$.

Note that P-stability requires the validity of (18.21) for all $m \geq 1$ on the restricted set of pairs (λ, μ) such that $|\mu| < -\Re(\lambda)$.

The increase in difficulty in passing from the P-stability to the D-stability analysis essentially originates from the fact that the set $\mathcal{S}_\tau$ is larger, depends on τ and is more complicated to describe. Actually, the success of the analysis depends on the parameterization adopted for describing the set $\mathcal{S}_\tau$ and the D_τ-stability regions S_{D_τ}.

Two alternative approaches have been pursued in the literature.

The first is based on the set

$$Q_*[\tau\lambda] = \{\tau\mu \mid (\lambda,\mu) \in \mathcal{S}_\tau\}$$

and on the sets

$$Q_m[\tau\lambda] = \{\tau\mu \mid (18.21) \text{ holds}\},$$

for which D-stability is equivalent to

$$Q_*[\tau\lambda] \subseteq \bigcap_{m=1}^{\infty} Q_m[\tau\lambda] \quad \text{for all } \lambda \in \mathbb{C}.$$

By following this approach, that seems more suitable for analyzing the real case $(\lambda,\mu) \in \mathbb{R}^2$, it can be proved that the class of Θ-methods is D(0)-stable for all $\frac{1}{2} \le \Theta \le 1$ and that the trapezoidal rule is not D-stable. Moreover, a necessary condition for D(0)-stability based on the analysis of the behavior of the $D_\tau(0)$-stability region in a neighborhood of the cusp point $(1/\tau, -1/\tau)$ on the border of the stability set $\mathcal{S}_\tau$ was found. On this basis it was proved that the Lobatto IIIC method is not D(0)-stable (see Figure 32).

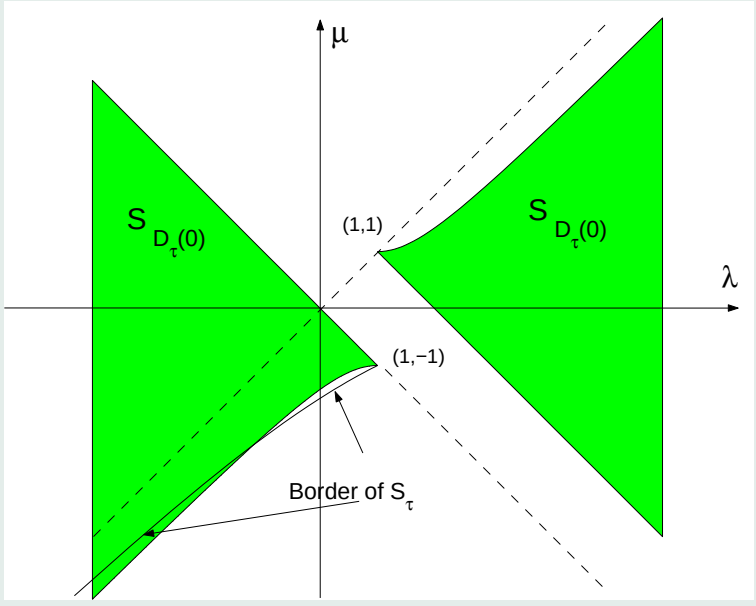

Figure 32: The stability region $\mathcal{S}_{D_\tau(0)}$ of the Lobatto IIIC method for $h = \tau/m$, $\tau = 1$, $m = 1$, versus the asymptotic stability region $\mathcal{S}_\tau$ in the real (λ, μ)-plane.

The following theorem regards A-stable methods with the symmetric A-stability function $R(\alpha)$ (i.e. $R(-\alpha)R(\alpha) = 1$).

****

Theorem 18.2 *Let us consider an A-stable RK method with the symmetric A-stability function $R(\alpha) = P(\alpha)/Q(\alpha)$ such that $P(\alpha)$ and $Q(\alpha)$ are polynomials of degree $\leq s$. If $R(\alpha) = \mathrm{e}^{\alpha} - C\alpha^{p+1} + O(\alpha^{p+2})$, with $p \geq 2s-2$, and if the error constant satisfies $(-1)^{p/2}C > 0$, then the corresponding natural RK method for DDEs is D(0)-stable. In particular, all Gaussian collocation methods are D(0)-stable.*

****

For Radau methods we have the following two results, the former of positive type on D(0)-stability and the latter of negative type on D-stability.

****

Theorem 18.3 *The s-stage Radau IIA methods are D(0)-stable for $s = 2, 3$.*

****

****

Theorem 18.4 *A natural RK method for DDEs with symmetric A-stability function cannot be D-stable. In particular, no Gaussian collocation method is D-stable.*

****

The second approach describes the set $\mathcal{S}_\tau$ and the regions S_{D_τ} in terms of $\tau\lambda$. With this approach it is possible to prove the following important result.

****

Theorem 18.5 *If (P_m) holds with $m = 1$, then it holds for any integer $m \geq 1$.*

****

In other words, for any asymptotically stable equation, the method is D-stable if the numerical solution y_n tends to zero for the largest admissible stepsize $h = \tau$.

By following this approach, Theorem 18.5 was used to prove these further results.

****

Theorem 18.6 *The backward Euler method satisfies the property (P_m) for $m = 1$ and, hence, it is D-stable.*

****

****

Theorem 18.7 *In the class of Radau IIA methods, the only one that is D-stable is the backward Euler method.*

****

Figure 33 illustrates the D_τ-stability region of the backward Euler method as the intersection of the sets for which the property (P_m) holds for $m \geq 1$.

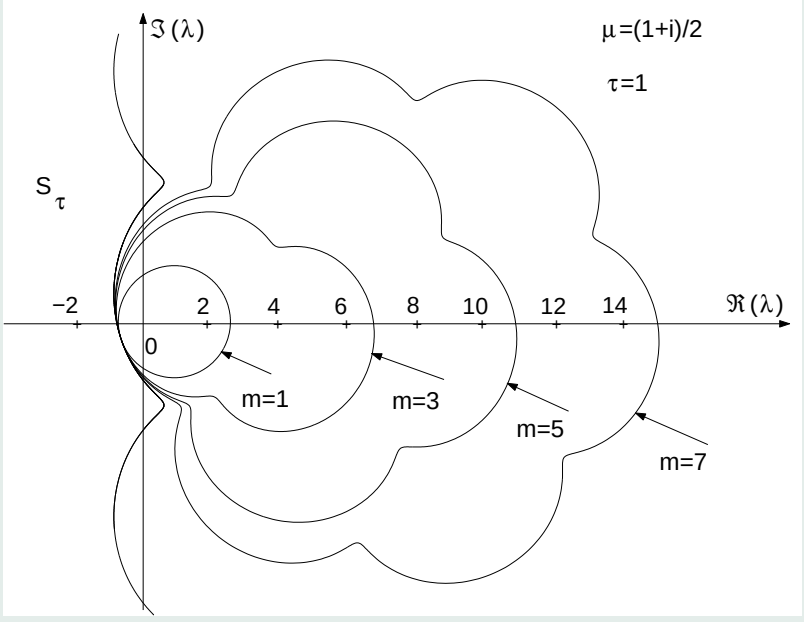

Figure 33: Regions of the complex λ-plane with fixed delay $\tau = 1$ and $\mu = 1 + i$, where the property (P_m) holds for the backward Euler method for various $m \geq 1$. For each m such regions include S_τ.

18.3. FP-stability and FP-contractivity

In this section we analyze the contractivity and asymptotic stability properties of RK methods for DDEs applied to the test equation

$$y'(t) = \lambda y(t) + \mu y(t - \tau(t)). \tag{18.22}$$

As in Section 15, we need some preliminary results regarding the stability properties of discrete and continuous RK methods with respect to the test ODE with forcing term (15.5).

Definition 18.13 *The RK method (6.2), (6.3) is* $\mathrm{A_f}$-stable *if the numerical solution* $\{y_n\}_{n\geq 0}$ *of (15.5) satisfies*

$$|y_{n+1}| \leq \max\left\{|y_n|, \max_{1\leq i\leq \nu} |g(t_{n+1}^i)|\right\} \tag{18.23}$$

for any mesh Δ*, whenever* $\Re(\lambda) \leq 0$.

Observe that an $\mathrm{A_f}$-stable RK method is necessarily A-stable, too.

Definition 18.14 *The CRK method (6.2), (6.9), (6.6) is* $\mathrm{A_f}$-stable *if the numerical solution* $\eta(t)$ *of (15.5) satisfies*

$$|\eta(t_n + \theta h_{n+1})| \leq \max\left\{|y_n|, \max_{1\leq i\leq s} |g(t_{n+1}^i)|\right\} \tag{18.24}$$

for all $\theta \in [0, 1]$ *and for any mesh* Δ*, whenever* $\Re(\lambda) \leq 0$.

The set of A_f-stable CRK methods is very meager.

In the class of one-stage CRK methods of order 1, the only one that is A_f-stable is the *backward Euler* method with linear interpolation.

In the class of two-stage CRK methods of order 2, we have the following class of A_f-stable methods:

$$\begin{array}{c|cc} c_1 & b_1 & c_1 - b_1 \\ 1 & b_1 & 1 - b_1 \\ \hline & b_1 & 1 - b_1 \end{array}$$

where $0 \leq c_1 \leq 1/3$ and $b_1 = \frac{1}{2(1-c_1)}$, with linear interpolation. This class contains the *Lobatto IIIC* method, obtained for $c_1 = 0$ (see Section 5.3).

So far, no A_f-stable discrete RK method of order $p = 3$ is known.

The next result regards the behaviour of the A-stability function at infinity.

****

Proposition 18.1 *Let the RK method (6.2), (6.3) be of order $p \geq 1$ and A_f-stable. Then its A-stability function satisfies the condition*

$$|R(\infty)| = \lim_{|\alpha| \to +\infty} |R(\alpha)| < 1. \tag{18.25}$$

In particular, if $c_i \in [-1, 1]$, $i = 1, \ldots, \nu$, then $R(\infty) = 0$.

****

Now, for a given $\lambda_0 \leq 0$, consider the condition

$$\Re(\lambda) \leq \lambda_0 \tag{18.26}$$

on the coefficient λ of equation (15.5). Then, for a discrete RK method (6.2), (6.3), define the *A-error growth function*

$$\Phi_A(\lambda_0, h) = \sup_{\Re(\alpha) \leq h\lambda_0} |R(\alpha)|, \tag{18.27}$$

that is a function of the stepsize h.

It satisfies the following properties.

****

Theorem 18.8 *The A-error growth function $\Phi_A(\lambda_0, h)$ of an A-stable RK method (6.2), (6.3) is a superexponential function of h for all $\lambda_0 \leq 0$, that is*

- $\Phi_A(\lambda_0, 0) = 1$;
- $\Phi_A(\lambda_0, h_1)\Phi_A(\lambda_0, h_2) \leq \Phi_A(\lambda_0, h_1 + h_2)$ *for all* $h_1, h_2 \geq 0$.

Moreover, if (18.25) holds and $\lambda_0 < 0$, then it is asymptotically negative superexponential, that is

- $\Phi_A(\lambda_0, h) < 1$ *for all* $h > 0$;
- $\limsup_{h \to +\infty} \Phi_A(\lambda_0, h) < 1$;

hold too.

****

The discrete analogue of (15.6) holds.

****

Theorem 18.9 *Let the RK method (6.2), (6.3) be A_f-stable. Then, for any mesh Δ, the numerical solution $\{y_n\}_{n\geq 0}$ of (15.5) with $\Re(\lambda) \leq 0$ satisfies*

$$|y_{n+1}| \leq \Psi^A_{n+1}|y_n| + (1 - \Psi^A_{n+1}) \max_{1\leq i\leq \nu} |g(t^i_{n+1})|,$$

where $\Psi^A_{n+1} = |R(h_{n+1}\lambda)| \leq 1$, $R(\alpha)$ being the A-stability function.

****

Observe that, by (18.27), under condition (18.26) we have

$$0 \leq \Psi^A_{n+1} \leq \Phi_A(\lambda_0, h_{n+1}). \tag{18.28}$$

Now go back to the test DDE (18.22).

For the moment, as in Section 15, we assume that the delay $\tau(t)$ satisfies the sole hypothesis (H_1). Then we can prove the discrete analogue of Theorem 15.1.

****

Theorem 18.10 *If the underlying CRK method (6.2), (6.9), (6.4) is A_f-stable, then the RK method for DDEs (10.2), (10.3) is FP-contractive.*

****

Proof of Theorem 18.10. We apply the RK method (6.2), (6.9), (6.4) to the test equation (18.22) that satisfies (18.2). Thus, the numerical solution in the step $[t_n, t_{n+1}]$ is the numerical solution of the local problem

$$\begin{cases} w'_{n+1}(t) = \lambda w_{n+1}(t) + \mu x(t - \tau(t)), \quad t_n \le t \le t_{n+1}, \\ w_{n+1}(t_n) = y_n, \end{cases} \tag{18.29}$$

where

$$x(s) = \begin{cases} \phi(s) & \text{for } s \le t_0, \\ \eta(s) & \text{for } t_0 \le s \le t_n, \\ w_{n+1}(s) & \text{for } t_n \le s \le t_{n+1}. \end{cases}$$

We assume by induction that

$$|\eta(t)| \le \max_{x \le t_0} |\phi(x)|, \quad t \le t_n, \tag{18.30}$$

which obviously holds for $n = 0$.

Since (18.2) is satisfied, we have that $|\mu| = -R \cdot \Re(\lambda)$, with $0 \le R \le 1$, and that $\Re(\lambda) \le 0$. Therefore, since the underlying CRK method is A_f-stable, we have

$$|\eta(t)| \le \max \left\{ |y_n|, R \max_{1 \le i \le s} |\eta(t^i_{n+1} - \tau(t^i_{n+1}))| \right\}, \quad t_n \le t \le t_{n+1}.$$

If overlapping does not occur, i.e. if $t^i_{n+1} - \tau(t^i_{n+1})) \le t_n$, $i = 1, \ldots, s$, then (18.30) clearly holds also for $t \le t_{n+1}$ and the proof is complete.

On the contrary, if overlapping occurs, then the inductive hypothesis (18.30) just implies

$$|\eta(t)| \le \max \left\{ \max_{x \le t_0} |\phi(x)|, R \max_{t_n \le x \le t_{n+1}} |\eta(x)| \right\}, \quad t_n \le t \le t_{n+1}.$$

Now, if $R < 1$, again (18.30) clearly holds also for $t \le t_{n+1}$. Otherwise, if $R = 1$, we can consider the one-parameter family

of local problems

$$\begin{cases} w'_{n+1,\rho}(t) = \lambda w_{n+1,\rho}(t) + \rho\mu x(t-\tau(t)), & t_n \le t \le t_{n+1}, \\ w_{n+1,\rho}(t_n) = y_n, \end{cases} \tag{18.31}$$

where

$$x(s) = \begin{cases} \phi(s) & \text{for } s \le t_0, \\ \eta(s) & \text{for } t_0 \le s \le t_n, \\ w_{n+1,\rho}(s) & \text{for } t_n \le s \le t_{n+1}, \end{cases}$$

for $1 - \epsilon \le \rho < 1$, $\epsilon > 0$.

For the local problem (18.31) it holds that $R = \rho < 1$. Thus, in view of the previous case, the numerical solution $\eta_\rho(t)$ satisfies

$$|\eta_\rho(t)| \le \max_{x \le t_0} |\phi(x)|, \quad t_n \le t \le t_{n+1}.$$

Since continuity arguments easily show that $\eta_\rho(t)$ converges uniformly in $[t_n, t_{n+1}]$ to the numerical solution $\eta(t)$ of (18.29) as $\rho \to 1$, again we can conclude that (18.30) holds also for $t \le t_{n+1}$. ■

Remark 18.1 *From the proof of Theorem 18.10, we can see that A_f-stability of the underlying CRK method implies even more than FP-contractivity, namely*

$$|\eta(t)| \le \max_{x \le t_0} |\phi(x)|, \quad t \ge t_0.$$

In other words, the contractivity property (18.8) extends to the continuous numerical solution.

Now we assume that, besides hypothesis (H_1), the delay $\tau(t)$ satisfies hypotheses (H_2) and (H_4) and, possibly, also (H_3). However, as in Section 15, some of the forthcoming results could be proved just with some increase in complexity without assuming the monotonicity hypothesis (H_4).

We recall that, under hypotheses (H_1), (H_2) and (H_4), the delay $\tau(t)$ determines a set of discontinuity points $\{\xi_k\}_{k\geq 0}$, $\xi_0 = t_0$, which is monotonically increasing and diverging as $k \to \infty$. Moreover, if also the hypothesis (H_3) is verified, then $\xi_{k+1} - \xi_k \leq \tau_1$ for all $k \geq 0$.

In order to prove the forthcoming Theorem 18.11, which is the discrete analogue of Theorem 15.2, we need to analyze the possible images in the deviated argument $\alpha(t) = t - \tau(t)$ of a mesh interval $[t_n, t_{n+1}]$ for an arbitrary mesh Δ, possibly allowing overlapping.

To this aim, define $n_0 = k_0 = 0$ and, for each $l \geq 1$, consider the integer $n_l \geq 1$ for which there exists $k_l \geq 1$ such that

$$t_{n_l} \leq \xi_{k_l} < t_{n_l+1} \leq \xi_{k_l+1}. \tag{18.32}$$

It is evident that, under the assumption of hypotheses (H_1) and (H_4), the integers n_l and k_l are uniquely determined for any $l \geq 1$. Therefore, a subset of discontinuity points ξ_{k_l} is defined, which are *significant* for the mesh Δ.

Moreover, it is also easy to see that, if the mesh Δ does not allow overlapping, then all the discontinuity points ξ_k are significant for Δ (i.e. $k_l = l$ for all $l \geq 1$).

Remark 18.2 *If the mesh Δ includes all the discontinuity points, in which case overlapping never occurs, then it is clear that $[t_n, t_{n+1}] \subseteq [\xi_{k-1}, \xi_k]$ for some $k \geq 1$ and that, if $[t_{m_1}, t_{m_2}]$ is the minimum union of mesh intervals that contain all the internal points of the image $\alpha([t_n, t_{n+1}])$, then we have $[t_{m_1}, t_{m_2}] \subseteq [\xi_{k-2}, \xi_{k-1}]$.*

On the contrary, if the mesh Δ does not necessarily include the discontinuity points, then the mesh interval $[t_n, t_{n+1}]$ may contain at most one significant discontinuity point, say $\xi_{k_{l-1}}$, as an internal point. Moreover, $[t_{m_1}, t_{m_2}]$ may also contain internal points of $[\xi_{k_{l-4}}, \xi_{k_{l-3}}]$, but the right-hand extremum t_{m_1+1} of the first image-step $[t_{m_1}, t_{m_1+1}]$ is always $> \xi_{k_{l-3}}$.

Solely for technical reasons, define the additional (significant) discontinuity points $\xi_{-1} = \xi_0 - \tau(\xi_0)$ and $\xi_{-2} = \xi_{-1}$ and extend backwards any mesh Δ defined in $[t_0, +\infty)$ to the points $t_{-2} = t_{-1} = \xi_{-2} = \xi_{-1}$.

****

Theorem 18.11 *Assume that the delay $\tau(t)$ satisfies hypotheses (H_1), (H_2) and (H_4). If the underlying CRK method (6.2), (6.9), (6.4) is A_f-stable, then the RK method for DDEs (10.2), (10.3) is FP-stable.*

Moreover, if the additional hypothesis (H_3) is verified and if there exists $h_0 > 0$ such that $h_{n+1} \leq h_0$ for all $n \geq 0$, then the convergence rate in (18.6) is at least of exponential type, i.e. at least like $\mathrm{e}^{-\alpha(t-t_0)}$ for some $\alpha > 0$.

****

Proof of Theorem 18.11. We apply the RK method (6.2), (6.9), (6.4) to the test equation (18.22) that satisfies (18.4). Observe that the numerical solution in the mesh interval $[t_n, t_{n+1}]$ is the numerical solution of the local problem (18.29).

Since (18.4) is satisfied, we have that $|\mu| = -R \cdot \Re(\lambda)$, with $0 \le R < 1$, and that $\Re(\lambda) < 0$. Therefore, since the underlying CRK method is A_f-stable, we have

$$|\eta(t)| \le \max\left\{|y_n|, R \max_{1\le i\le s} |\eta(t^i_{n+1} - \tau(t^i_{n+1}))|\right\}, \quad t_n \le t \le t_{n+1},$$

and, moreover, Theorem 18.9 yields

$$|y_{n+1}| \le \Psi^A_{n+1}|y_n| + (1 - \Psi^A_{n+1}) R \max_{1\le i\le s} |\eta(t^i_{n+1} - \tau(t^i_{n+1}))|.$$

Assuming that $t_{n+1} \in (\xi_{k_{l-1}}, \xi_{k_l}]$, as observed in Remark 18.2, the delayed argument $t^i_{n+1} - \tau(t^i_{n+1}) \in [t_{m_1}, t_{m_2}]$ with $t_{m_1+1} > \xi_{k_{l-3}}$ and, of course, $t_{m_2} \le t_{n+1}$.

Therefore, with

$$G_n = \max_{t_{n-1}\le t\le t_n} |\eta(t)|,$$

we have

$$|y_n| \le G_n,$$
$$G_{-1} = G_0,$$

$$|y_{n+1}| \le \Psi^A_{n+1}|y_n| + (1 - \Psi^A_{n+1}) R \max_{\xi_{k_{l-3}} < t_\nu \le t_{n+1}} G_\nu \tag{18.33}$$

and

$$G_{n+1} \le \max\left\{|y_n|, R \max_{\xi_{k_{l-3}} < t_\nu \le t_{n+1}} G_\nu\right\}. \tag{18.34}$$

Thus our aim is to prove that

$$\lim_{n\to\infty} |y_n| = \lim_{n\to\infty} G_n = 0. \tag{18.35}$$

Define $n_{-2} = -2$, $n_{-1} = -1$ and the numbers ϵ_{-2}, ϵ_{-1}, ϵ_0, E_{-2}, E_{-1} and E_0 to be all equal to G_0. Observe that the interesting case is $G_0 > 0$, otherwise (18.35) is obvious, since (18.33) and (18.34) would imply $|y_n| = G_n = 0$ for all n.

In view of the definition of the significant discontinuity points in (18.32), for $l \geq 1$ define inductively (with respect to l) the numbers

$$E_l = \epsilon_{n_l}, \tag{18.36}$$

where, for $n_{l-1} + 1 \leq n \leq n_l$,

$$\epsilon_n = RE_{l-3} + \left(\prod_{\nu=n_{l-1}+1}^{n} \Psi_\nu^A \right) (E_{l-1} - RE_{l-3}). \tag{18.37}$$

Now, as an inductive hypothesis, assume that $E_{l-1} - RE_{l-3} \geq 0$, $E_{l-3} - E_{l-2} \geq 0$ and $E_{l-2} - E_{l-1} \geq 0$. These inequalities clearly hold for $l = 1$. Then, with

$$\Pi_l = \prod_{\nu=n_{l-1}+1}^{n_l} \Psi_\nu^A, \tag{18.38}$$

by (18.37) and (18.36) we have

$$\begin{aligned} E_l - RE_{l-2} &= RE_{l-3} + \Pi_l(E_{l-1} - RE_{l-3}) - RE_{l-2} \\ &= R(E_{l-3} - E_{l-2}) + \Pi_l(E_{l-1} - RE_{l-3}) \geq 0. \end{aligned}$$

Since condition (18.4) implies (18.26) for some $\lambda_0 < 0$, Theorem 18.8 and inequality (18.28) yield

$$\Psi_\nu^A < 1 \quad \text{for all } \nu.$$

Therefore, for $n_{l-1}+1 \le n \le n_l$ we have also

$$\begin{aligned} E_{l-1}-\epsilon_n &= E_{l-1}-RE_{l-3}-\left(\prod_{\nu=n_{l-1}+1}^{n}\Psi_\nu^A\right)(E_{l-1}-RE_{l-3}) \\ &= (E_{l-1}-RE_{l-3})\left(1-\prod_{\nu=n_{l-1}+1}^{n}\Psi_\nu^A\right)\ge 0. \end{aligned}$$

In particular, for $n=n_l$ we obtain

$$E_{l-1}-E_l\ge 0,\quad l\ge 0. \tag{18.39}$$

Therefore, we have proved that the sequence $\{\epsilon_n\}$ is non-increasing and that

$$E_l-RE_{l-2}\ge 0,\quad l\ge 0. \tag{18.40}$$

Now we want to prove that

$$G_\nu\le\epsilon_{\nu-1},\quad \nu\ge -1,$$

and

$$|y_\nu|\le\epsilon_\nu,\quad \nu\ge -1.$$

The above inequalities clearly hold for $\nu=-1,0$. Now assume by induction that they hold for $\nu\le n$ with $n_{l-1}\le n\le n_l-1$ for some $l\ge 1$. Therefore, by the inductive hypothesis and by (18.34), we get

$$G_{n+1}\le\max\left\{\epsilon_n, RG_{n+1}, R\max_{\xi_{k_{l-3}}<t_\nu\le t_n}\epsilon_{\nu-1}\right\}$$

and thus, since $\xi_{k_{l-3}}<t_\nu\le t_n$ implies $n_{l-3}+1\le\nu\le n$, since the sequence $\{\epsilon_n\}$ is non-increasing and since $R<1$, by (18.36) and (18.37) we obtain

$$G_{n+1}\le\max\{\epsilon_n, RG_{n+1}, RE_{l-3}\}=\epsilon_n. \tag{18.41}$$

Analogously, by (18.33), by the inductive hypothesis and by (18.41), we get

$$|y_{n+1}| \leq \Psi^A_{n+1}\epsilon_n + (1 - \Psi^A_{n+1})R \max_{\xi_{k_{l-3}} < t_\nu \leq t_{n+1}} \epsilon_{\nu-1}$$

which, by (18.36) and (18.37), yields

$$|y_{n+1}| \leq \Psi^A_{n+1}\epsilon_n + (1 - \Psi^A_{n+1})RE_{l-3} = \epsilon_{n+1}.$$

To conclude the proof, we have to show that, under hypothesis (H_2), the sequence $\{\epsilon_n\}$ vanishes as $n \to \infty$. Indeed, hypothesis (H_2) implies that, as $n \to \infty$, $t_n \in (\xi_{k_{l-1}}, \xi_{k_l}]$ with $l \to \infty$. Therefore, it is sufficient to prove that $E_l \to 0$.

To this end, observe that, by (18.37), (18.36) and (18.38), the sequence $\{E_l\}$ is such that

$$\begin{cases} E_l = \Pi_l E_{l-1} + R(1 - \Pi_l)E_{l-3}, & l \geq 1, \\ E_{-2} = E_{-1} = E_0 > 0. \end{cases} \tag{18.42}$$

We need to give a uniform upper bound $\Pi < 1$ to the sequence $\{\Pi_l\}$.

Since inequality (18.28) holds and $\Phi_A(\lambda_0, h)$ is asymptotically negative superexponential (see Theorem 18.8), no problems are caused if the stepsizes h_{n+1} are not uniformly bounded from above.

On the contrary, one can easily construct a mesh Δ containing a subsequence of steps that vanishes at infinity, with steps of the type $[t_{n_{l-1}}, t_{n_l}]$, where $n_l = n_{l-1} + 1$. In this case the corresponding subsequence in $\{\Pi_l\}$ also goes to 1 at infinity, and hence the uniform upper bound $\Pi < 1$ does not exist.

In order to overcome this inconvenience, observe that (18.32) implies that the sum of the lengths of two consecutive intervals $[t_{n_{l-1}}, t_{n_l}]$ and $[t_{n_l}, t_{n_{l+1}}]$ is greater than the length of the interval

$[\xi_{k_{l-1}}, \xi_{k_l}]$, which is $\geq \tau_0$ by hypothesis (H_1). Therefore, again since $\Phi_A(\lambda_0, h)$ is asymptotically negative superexponential, we obtain

$$\begin{aligned}\Pi_{l+1}\Pi_l &= \prod_{\nu=n_{l-1}+1}^{n_{l+1}} \Psi_\nu^A \leq \prod_{\nu=n_{l-1}+1}^{n_{l+1}} \Phi_A(\lambda_0, h_\nu) \\ &\leq \Phi_A(\lambda_0, t_{n_{l+1}} - t_{n_{l-1}}) \leq \Pi = \max_{h\geq\tau_0} \Phi_A(\lambda_0, h) < 1.\end{aligned}$$

Consequently, by (18.39) and (18.40), iterating (18.42) twice easily yields

$$\begin{cases} E_{l+1} \leq \Pi E_{l-1} + R(1-\Pi)E_{l-3}, & k \geq 1, \\ E_{-2} = E_{-1} = E_0 > 0, \quad E_1 > 0. \end{cases}$$

This is a fourth-order difference inequality with constant coefficients. Since the coefficients are positive and also all the numbers E_l are positive, it follows that $E_l \leq \Gamma_l$, $l \geq -2$, where the sequence $\{\Gamma_l\}$ satisfies the fourth-order difference equation with constant coefficients

$$\begin{cases} \Gamma_{l+1} = \Pi\Gamma_{l-1} + R(1-\Pi)\Gamma_{l-3}, & l \geq 1, \\ \Gamma_i = E_i, \quad i = -2, \ldots, 1. \end{cases}$$

Since the sum of the (positive) coefficients in the right-hand side is < 1, all four characteristic roots are < 1 in modulus. Thus, if β is the biggest of them, we can conclude that $E_l \to 0$ at least as $|\beta|^l$.

Finally, if hypothesis (H_3) is satisfied, for a given $t_n \geq t_0$, we have that $t_n \in (\xi_{k-1}, \xi_k]$ for some

$$k \geq \frac{t_n - t_0}{\tau_1}.$$

On the other hand, since hypothesis (H_1) holds, if there exists $h_0 > 0$ such that $h_{n+1} \leq h_0$ for all $n \geq 0$, then the macro interval $[\xi_{k-1}, \xi_k]$ is necessarily included between two consecutive

significant discontinuity points $\xi_{k_{l-1}}$ and ξ_{k_l}, where

$$l \geq \frac{k}{r_0},$$

with

$$r_0 = \left\lceil \frac{h_0}{\tau_0} \right\rceil .$$

Therefore, since $\{\epsilon_n\}$ is non-increasing, it follows that $\epsilon_n \to 0$ at least like $\mathrm{e}^{-\alpha(t_n - t_0)}$, where $\alpha = -\log(|\beta|)/(r_0\tau_1)$. ■

19. Generalizations of A-stability to NDDEs

A-stability has also been generalized to DDEs of neutral type

$$\begin{cases} y'(t) = \lambda y(t) + \mu y(t-\tau) + \nu y'(t-\tau), & t \geq t_0, \\ y(t) = \phi(t), & t \leq t_0, \end{cases} \tag{19.1}$$

where $\lambda, \mu, \nu \in \mathbb{C}$ and τ is a constant delay.

Also in this case we have two concepts of asymptotic stability, one for all delays and another for fixed delay (see Definition 17.1).

Definition 19.1 *The* NP-stability region *of a numerical method for NDDEs is the set S_{NP} of triplets of complex numbers (α, β, ν), $\alpha = h\lambda$, $\beta = h\mu$, such that the discrete numerical solution $\{y_n\}_{n\geq 0}$ of (19.1), obtained with constant stepsize h under the constraint*

$$h = \tau/m, \quad m \geq 1, \ m \text{ integer}, \tag{19.2}$$

satisfies

$$\lim_{n\to\infty} y_n = 0 \tag{19.3}$$

for all constant delays τ and all initial functions $\phi(t)$.

Definition 19.2 *A numerical method for NDDEs is* NP-stable *if*

$$S_{NP} \supseteq \left\{ (\alpha, \beta, \nu) \in \mathbb{C}^3 \;\middle|\; \Re(\alpha) < 0 \text{ and } |\alpha\nu + \beta| < \left| 2\Re(\alpha) + |\alpha\bar{\nu} - \bar{\beta}| \right| \right\}.$$

In other words, a numerical method is NP-stable if, for any stepsize h satisfying (19.2), it preserves the asymptotic stability on the whole class of asymptotically stable equations (19.1) characterized by condition (17.1).

Removing constraint (19.2) leads to the following stronger concept of stability.

Definition 19.3 *The* GNP-stability region *of a numerical method for NDDEs is the set S_{GNP} of triplets of complex numbers (α, β, ν), $\alpha = h\lambda$, $\beta = h\mu$, such that the discrete numerical solution $\{y_n\}_{n\geq 0}$ of (19.1), obtained with constant stepsize h, satisfies (19.3) for all constant delays τ and all initial functions $\phi(t)$.*

Definition 19.4 *A numerical method for NDDEs is* GNP-stable *if*

$$S_{GNP} \supseteq \left\{ (\alpha, \beta, \nu) \in \mathbb{C}^{\,3} \;\middle|\; \Re(\alpha) < 0 \text{ and } |\alpha\nu + \beta| < \left| 2\Re(\alpha) + |\alpha\bar{\nu} - \bar{\beta}| \right| \right\}.$$

GNP-stability $\Rightarrow$ NP-stability $\Rightarrow$ P-stability

GNP-stability $\Rightarrow$ GP-stability

Definition 19.5 *For any fixed τ, the* ND$_\tau$-stability region *of a numerical method for NDDEs is the set S_{ND_τ} of quaternaries $(h, \lambda, \mu, \nu) \in \mathbb{R}^+ \times \mathbb{C}^3$, where h fulfils the constraint (19.2), such that the discrete numerical solution $\{y_n\}_{n\geq 0}$ of (19.1), obtained with constant stepsize h, satisfies (19.3) for all initial functions $\phi(t)$.*

Definition 19.6 *A numerical method for NDDEs is* ND-stable *if, for any fixed τ,*

$$(\lambda, \mu, \nu) \in \mathcal{NS}_\tau \quad \Longrightarrow \quad (h, \lambda, \mu, \nu) \in S_{ND_\tau}$$

for any constant stepsize h satisfying (19.2).

Observe that, for $\nu = 0$, NP-stability reduces to P-stability and ND-stability reduces to D-stability. Moreover, the concept of ND-stability is stronger than NP-stability.

ND-stability $\Rightarrow$ D-stability

ND-stability $\Rightarrow$ NP-stability

In this lecture we analyze in some detail only the simple case of NP-stability. However, very few results are available in the literature for the other concepts of stability.

19.1. NP-stability

By extending in a natural way the arguments developed in Section 18.1 for the P-stability analysis, we consider the NDDE (19.1) and, according to the notation of Section 13, consider the class of RK methods for NDDEs based on the option (13.2), i.e. $y'(t-\tau) \approx \eta'(t-\tau)$, which reads

$$Y_{n+1}^i = y_n + h\sum_{j=1}^{s} a_{ij}(\lambda Y_{n+1}^j + \mu\eta(t_{n-m+1}^j) + \nu\eta'(t_{n-m+1}^j)), \quad i = 1,\dots,s, \tag{19.4}$$

$$\eta(t_n+\theta h) = y_n + h\sum_{i=1}^{s} b_i(\theta)(\lambda Y_{n+1}^i + \mu\eta(t_{n-m+1}^i) + \nu\eta'(t_{n-m+1}^i)), \tag{19.5}$$

$$\eta'(t_n+\theta h) = \sum_{i=1}^{s} b_i'(\theta)(\lambda Y_{n+1}^i + \mu\eta(t_{n-m+1}^i) + \nu\eta'(t_{n-m+1}^i)). \tag{19.6}$$

With $b = [b_1,\dots,b_s]^T$, $b(\theta) = [b_1(\theta),\dots,b_s(\theta)]^T$, $e = [1,\dots,1]^T$ the unit s-vector, I the s-dimensional identity matrix, $A = [a_{ij}]_{i,j=1}^s$, $\alpha = h\lambda$ and $\beta = h\mu$, after elimination of the stage values Y_{n+1}^i from (19.4) and computation of (19.5) and (19.6) for $\theta = c_1,\dots,c_s,1$, we get

$$\begin{aligned}\eta(t_{n+1}^i) &= R_i(\alpha)y_n + \beta\sum_{j=1}^{s}(b(c_i)^T(I-\alpha A)^{-1})_j\eta(t_{n-m+1}^j) \\ &\quad +\nu\sum_{j=1}^{s}(b(c_i)^T(I-\alpha A)^{-1})_j h\eta'(t_{n-m+1}^j), \quad i = 1,\dots,s,\end{aligned}$$

and

$$\begin{aligned} h\eta'(t^i_{n+1}) &= \bar{R}_i(\alpha)y_n + \beta\sum_{j=1}^{s}(b'(c_i)^T(I-\alpha A)^{-1})_j\eta(t^j_{n-m+1}) \\ &\quad +\nu\sum_{j=1}^{s}(b'(c_i)^T(I-\alpha A)^{-1})_j h\eta'(t^j_{n-m+1}), \quad i=1,\dots,s, \end{aligned}$$

where

$$\begin{aligned} R_i(\alpha) &= 1+\alpha b(c_i)^T(I-\alpha A)^{-1}e, \\ \bar{R}_i(\alpha) &= \alpha b'(c_i)^T(I-\alpha A)^{-1}e, \end{aligned}$$

and $(x)_j$ is understood to be the jth component of the row vector x. Moreover, by (19.5) for $\theta = 1$, we get

$$\begin{aligned} y_{n+1} &= R(\alpha)y_n + \beta\sum_{j=1}^{s}(b^T(I-\alpha A)^{-1})_j\eta(t^j_{n-m+1}) \\ &\quad +\nu\sum_{j=1}^{s}(b(c_i)^T(I-\alpha A)^{-1})_j h\eta'(t^j_{n-m+1}), \end{aligned}$$

where $R(\alpha)$ is the A-stability function (18.14).

Then consider the sequence of $(2s+1)$-dimensional vectors

$$\mathbf{H}_n = [h\eta'(t^1_n),\dots,h\eta'(t^s_n),\eta(t^1_n),\dots,\eta(t^s_n),y_n]^T$$

and the relevant recurrence relation with constant coefficients

$$\mathbf{H}_{n+1} = P(\alpha)\mathbf{H}_n + (\beta Q(\alpha)+\nu S(\alpha))\mathbf{H}_{n-m+1}, \qquad (19.7)$$

where

$$P(\alpha) = \begin{bmatrix} 0 & 0 & \alpha C(I-\alpha A)^{-1}e \\ 0 & 0 & e+\alpha B(I-\alpha A)^{-1}e \\ 0^T & 0^T & 1+\alpha b^T(I-\alpha A)^{-1}e \end{bmatrix},$$

$$Q(\alpha) = \begin{bmatrix} 0 & C(I-\alpha A)^{-1} & 0 \\ 0 & B(I-\alpha A)^{-1} & 0 \\ 0^T & b^T(I-\alpha A)^{-1} & 0 \end{bmatrix},$$

$$S(\alpha) = \begin{bmatrix} C(I-\alpha A)^{-1} & 0 & 0 \\ B(I-\alpha A)^{-1} & 0 & 0 \\ b^T(I-\alpha A)^{-1} & 0^T & 0 \end{bmatrix},$$

$B = [b_j(c_i)]_{i,j=1}^s$ and $C = [b_j'(c_i)]_{i,j=1}^s$.

Observe that the inclusion of the elements $h\eta'(t_n^i)$ in the vector $\mathbf{H}_n$ is simply due to the formal request to reproduce a three-term recurrence relation (19.7), as was done in the non-neutral case.

The characteristic equation of (19.7) is

$$\det\big[\zeta^m I - \zeta^{m-1}P(\alpha) - \beta Q(\alpha) - \nu S(\alpha)\big] = 0 \qquad (19.8)$$

and, hence, direct computations, something longer than for the non-neutral case, lead to the factorization

$$\begin{aligned} &\det\big[\zeta^m I - \zeta^{m-1}P(\alpha) - \beta Q(\alpha) - \nu S(\alpha)\big] \\ &\quad = \zeta^{(2s+1)m-1}\det\big[I - \alpha A - (\beta/\zeta^m)B - (\nu/\zeta^m)C\big] \\ &\qquad \times\big(\zeta - R^{**}(\alpha, \beta/\zeta^m, \nu/\zeta^m)\big), \end{aligned} \qquad (19.9)$$

where

$$R^{**}(\alpha, w, z) = 1 + (\alpha + w)b^T\big(I - \alpha A - wB - zC\big)^{-1}e$$

is called the *NP-stability function*.

Now assume $\nu \neq 0$ and define

$$q = \frac{\beta}{\nu}$$

so that the factor $\zeta - R^{**}(\alpha, \beta/\zeta^m, \nu/\zeta^m)$ in (19.9) can be rewritten as $\zeta - \hat{R}^{**}(\alpha, q, \nu/\zeta^m)$, where

$$\hat{R}^{**}(\alpha, q, z) = 1 + (\alpha + qz)b^T\big(I - \alpha A - z(qB + C)\big)^{-1}e.$$

Therefore, instead of the roots ζ of the characteristic equation (19.8), we can equivalently consider the solutions of the algebraic equation

$$\zeta = \hat{R}^{**}(\alpha, q, \nu/\zeta^m), \tag{19.10}$$

provided the following assumption on the RK method holds.

Assumption 19.1 *The matrix* $I-\alpha A-z^*(qB+C)$ *is singular if and only if* z^* *is a pole of* $\hat{R}^{**}(\alpha, q, z)$.

Alternatively, assume $\beta \neq 0$ and define

$$\tilde{q} = \frac{\nu}{\beta}.$$

Observe that $\tilde{q} = q^{-1}$ for $\nu \neq 0$. Then the factor $\zeta - R^{**}(\alpha, \beta/\zeta^m, \nu/\zeta^m)$ in (19.9) can be rewritten as $\zeta - \tilde{R}^{**}(\alpha, \beta/\zeta^m, \tilde{q})$, where

$$\tilde{R}^{**}(\alpha, w, \tilde{q}) = 1 + (\alpha + w)b^T\big(I - \alpha A - w(B + \tilde{q}C)\big)^{-1}e,$$

so that (19.10) changes to

$$\zeta = \tilde{R}^{**}(\alpha, \beta/\zeta^m, \tilde{q}) \tag{19.11}$$

and Assumption 19.1 modifies as follows.

Assumption 19.2 *The matrix* $I - \alpha A - w^*(B + \tilde{q}C)$ *is singular if and only if* w^* *is a pole of* $\tilde{R}^{**}(\alpha, w, \tilde{q})$.

Observe that, for $\beta \neq 0$ and $\nu = 0$, equation (19.11) equals (18.18) and, hence, the approach is consistent with the non-neutral case.

The following theorem generalizes the result given by Corollary 18.1.

****

Theorem 19.1 *A natural RK method for NDDEs (19.4), (19.5), (19.6) for which $C = I$ (i.e., the derivatives $b_i'(\theta)$ of the continuous weights are the Lagrange polynomial coefficients of the interpolation scheme at the abscissae c_i) is NP-stable if the underlying discrete RK method (6.2), (6.3) is A-stable.*

****