Acta Numerica (2009), pp. 1–110
doi: 10.1017/S0962492906390010
© Cambridge University Press, 2009
Printed in the United Kingdom

Recent trends in the numerical solution of retarded functional differential equations*

Alfredo Bellen, Stefano Maset and Marino Zennaro
Dipartimento di Matematica e Informatica,
Università degli Studi di Trieste, I-34100 Trieste, Italy
E-mail: {bellen}{maset}{zennaro}@units.it

Nicola Guglielmi
Dipartimento di Matematica Pura e Applicata,
Università degli Studi di L'Aquila, I-67100 L'Aquila, Italy
E-mail: guglielm@univaq.it

Retarded functional differential equations (RFDEs) form a wide class of evolution equations which share the property that, at any point, the rate of the solution depends on a discrete or distributed set of values attained by the solution itself in the past. Thus the initial problem for RFDEs is an infinite-dimensional problem, taking its theoretical and numerical analysis beyond the classical schemes developed for differential equations with no functional elements. In particular, numerically solving initial problems for RFDEs is a difficult task that cannot be founded on the mere adaptation of well-known methods for ordinary, partial or integro-differential equations to the presence of retarded arguments. Indeed, efficient codes for their numerical integration need specific approaches designed according to the nature of the equation and the behaviour of the solution.

By defining the numerical method as a suitable approximation of the solution map of the given equation, we present an original and unifying theory for the convergence and accuracy analysis of the approximate solution. Two particular approaches, both inspired by Runge–Kutta methods, are described. Despite being apparently similar, they are intrinsically different. Indeed, in the presence of specific types of functionals on the right-hand side, only one of them can have an explicit character, whereas the other gives rise to an overall procedure which is implicit in any case, even for non-stiff problems.

In the panorama of numerical RFDEs, some critical situations have been recently investigated in connection to specific classes of equations, such as the accurate location of discontinuity points, the termination and bifurcation of

* This work was supported by INdAM–GNCS.

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

the solutions of neutral equations, with state-dependent delays, the regularization of the equation and the generalization of the solution behind possible termination points, and the treatment of equations stated in the implicit form, which include singularly perturbed problems and delay differential-algebraic equations as well. All these issues are tackled in the last three sections.

In this paper we have not considered the important issue of stability, for which we refer the interested reader to the comprehensive book by Bellen and Zennaro (2003).

CONTENTS

1 Introduction 2
2 Some particular RFDEs 7
3 Discontinuity points and vanishing delays 11
4 Existence and uniqueness 15
5 Numerical methods for RFDEs 21
6 Functional continuous Runge–Kutta methods 32
7 Order conditions for FCRK methods 39
8 The standard approach 62
9 Implementation issues in the standard approach 80
10 Neutral problems with state-dependent delays 88
11 Implicit problems with state-dependent delays 102
References 108

1. Introduction

In this paper, we present methods for numerically solving the *Cauchy problem*, or *initial problem* (IP), for the very general *retarded functional differential equation* (RFDE)

$$y'(t) = F(t, y_t), \tag{1.1}$$

where y is an $\mathbb{R}^d$-valued function of a real variable, $F : \mathbb{R} \times X \to \mathbb{R}^d$, X being a subset of the set $\mathcal{C}$ of the continuous functions $(-\infty, 0] \to \mathbb{R}^d$, and, according to the Hale–Krasovski notation, $y_t \in X$ is given by

$$y_t(\vartheta) = y(t + \vartheta), \quad \vartheta \in (-\infty, 0].$$

The set X is called the *data set* of the RFDE (1.1) and the function y_t is called the *state* at time t, since, under minimal assumptions, it uniquely determines the future evolution $y(s)$, $s \geq t$.

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

In order to define the IP for RFDEs, we must associate to (1.1) an initial point $t_0 \in \mathbb{R}$ and initial data $\phi \in X$. The resulting problem takes the form

$$\begin{aligned} y'(t) &= F(t, y_t), \quad t \geq t_0, \\ y_{t_0} &= \phi, \end{aligned} \tag{1.2}$$

where the function $\phi \in X$ represents the initial state of the system.

Equation (1.1), also called the *Volterra functional differential equation*, provides a powerful tool for modelling many phenomena in applied mathematics and, in the literature, is often referred to by different terminology, such as *time delay system*, *hereditary system*, *system with memory*, *system with after-effect*, *etc.*

There are many kinds of RFDEs characterized by the action of the functional F on the state y_t. In particular, we include in (1.1) the *neutral functional differential equations*, where F also acts on the derivative of the state y_t, *i.e.*,

$$F(t, y_t) = G(t, y_t, y'_t).$$

In any case, they are all evolution systems which share the property that, at any t, the dynamic depends not only on the current value $y(t)$, but also on a discrete or distributed set of values of the solution y in the past. This fact, together with the need for an initial function rather than an initial value, makes the theoretical analysis, as well as the numerical approximation of the IP (1.2), much more complicated than the initial value problem for ordinary differential equations it formally resembles.

The general theory of RFDEs is widely developed, and we refer the reader to the classical books by Bellman and Cooke (1963), El'sgol'ts and Norkin (1973), Hale (1977), Driver (1977), Kolmanovskii and Nosov (1986), Kolmanovskii and Myshkis (1992), Hale and Verduyn Lunel (1993), Kuang (1993) and Diekmann, van Gils, Verduyn Lunel and Walther (1995), which also include many real-life examples of RFDEs and more general retarded functional differential equations.

As for the numerics, apart from some isolated earlier papers, the analysis of numerical methods for RFDEs started in the early 1960s. Since then, specific methods have been separately developed by adapting the well-known methods for ordinary differential equations to the presence of delays. An exhaustive collection of methods and related references up to the beginning of 1970s was given by Cryer (1972). Other papers reporting the state of the art for general or particular classes of RFDEs appeared from time to time in the subsequent decades. In particular, Bellen (1985) and Meinardus and Nürnberger (1985) surveyed papers up to the 1980s, followed by Zennaro (1995), Baker, Paul and Willé (1995*a*, 1995*b*) and Baker (1996, 2000) up to the publication of the monograph by Bellen and Zennaro (2003), which was the first book completely devoted to the numerical analysis of the Cauchy

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

problem for differential equations with delays. After some historical remarks on theoretical and numerical methods for RFDEs, the book provides a detailed analysis of continuous Runge–Kutta (RK) methods $(A, b(\theta), c)$, in view of their application in the following general procedure, called the *standard approach*, for IPs of the form

$$\begin{aligned} y'(t) &= \hat{F}(t, y(t), y_t, y'_t), \quad t \geq t_0, \\ y_{t_0} &= \phi, \end{aligned} \tag{1.3}$$

where $\hat{F}$ explicitly separates the dependence on $y(t)$ from that on y_t.

Given a mesh $\Delta = \{t_0, t_1, \ldots, t_n, \ldots\}$, the standard approach for (1.3) consists in solving step by step, by means of a continuous RK method, the local problems

$$\begin{aligned} w'_{n+1}(t) &= \hat{F}\big(t, w_{n+1}(t), x_t, x'_t\big), \quad t_n \leq t \leq t_{n+1}, \\ w_{n+1}(t_n) &= y_n, \end{aligned} \tag{1.4}$$

where

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

and $\eta(s)$ is the continuous approximate solution computed by the method itself up to t_n.

The philosophy underlying the standard approach consists in considering (1.3) as an ODE, where the states x_t and x'_t, acting as forcing terms, are virtually known and given by the approximate solution itself, either having been or to be computed. It is clear that this approach relies on the availability of a *continuous numerical method*, that is, a numerical method which provides a continuous approximate solution. It is also clear that, whenever the right-hand side functional in (1.4) requires values of the functions x and x' at some points lying in the current interval, the method becomes implicit even if the underlying RK method is explicit. This makes the procedure more suited for stiff problems, for which the RK method is itself expected to be implicit. The standard approach, even before being so-named, was the most widely adopted method for RFDEs in the literature from the 1970s to the 1990s, also using continuous approximations other than continuous RK methods. In particular, for continuous RK methods, significant results on convergence, variable step-size implementation and stability analysis were achieved. A detailed presentation of such results up to 2002 is available in Bellen and Zennaro (2003), along with an exhaustive bibliography.

Since the publication of Bellen and Zennaro (2003), important work on the numerical solution of RFDEs has appeared. Part of it was devoted to further analyses of specific problems using well-known and consolidated

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

techniques, especially as far as stability issues are concerned (see Wang and Li (2004) and the book by Kuang and Cong (2005)). Another part was devoted to developing new methods and to tackling some topics that, due to their more intrinsic difficulty, had not yet been well investigated, namely equations of neutral type and equations with state-dependent delays.

This paper, rather than reporting the state of the art, aims at providing, in an original and unifying approach, results on well-posedness and the error analysis of different numerical schemes based on continuous RK methods for as large as possible classes of RFDEs. The paper also reports some recent results on specific issues that needed, and still need, further investigation. In particular, we consider the termination and bifurcation of the solution at some critical point for RFDEs of neutral type, with state-dependent delays and the possible generalization of the solution beyond such points as well.

The paper is organized as follows. In Section 2 we provide some particular classes of RFDEs together with suitable data sets X, where they are naturally defined and where the true and the approximate solutions will be sought. For other classes of RFDEs and bibliographic references, see also Brunner (2004).

In Section 3 we analyse a specific phenomenon, typical of RFDEs, which is not present in ordinary differential equations, namely the appearance of so-called *discontinuity points*, often called *breaking points*. These originate in the possible lack of continuity in the derivative of the solution of (1.1) at the initial point t_0, that is,

$$\phi'^{-}(0) \neq y'(t_0)^{+} = F(t_0, \phi).$$

This event implies that the solution, as defined in the forthcoming Definition 1.1, must be considered in the 'almost everywhere' sense.

In Section 4 we provide a short review of existence and uniqueness results for the solutions of the various classes of equations considered in Section 2. In order to do that, it is essential to establish what we actually mean by a solution of (1.2) in a right neighbourhood of t_0.

Definition 1.1. Let $T > 0$. A solution of (1.2) on $(-\infty, t_0 + T]$ is a continuous function $y : (-\infty, t_0 + T] \to \mathbb{R}^d$ such that:

- $y_t \in X$ for all $t \in [t_0, t_0 + T]$;
- the function $t \mapsto F(t, y_t)$, $t \in [t_0, t_0 + T]$, is measurable and bounded;
- for all $t \in (-\infty, t_0 + T]$, we have

$$y(t) = \begin{cases} \phi(0) + \int_{t_0}^{t} F(s, y_s)\,\mathrm{d}s & \text{if } \ t \in [t_0, t_0 + T], \\ \phi(t - t_0) & \text{if } \ t \in (-\infty, t_0]. \end{cases}$$

Note that the third condition in the above definition is equivalent to requiring that y is differentiable almost everywhere in $[t_0, t_0 + T]$, $y'(t) = F(t, y_t)$ for almost all $t \in [t_0, t_0 + T]$ and $y_{t_0} = \phi$.

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

In Section 5 we develop an original and unifying approach which allows us to analyse well-posedness and convergence for most of the methods developed so far for the whole class of equations stated in the form (1.2).

Sections 6, 7 and 8 are devoted to the construction and accuracy analysis of two classes of methods, both based on suitable continuous RK methods $(A, b(\theta), c)$, for which the continuous approximation $\eta(s)$ is given step by step, by

$$\eta(t_n+\theta h_{n+1}) = y_n+h_{n+1}\sum_{i=1}^{\nu} b_i(\theta)F\Big(t^i_{n+1}, Y_{t^i_{n+1}}, Y'_{t^i_{n+1}}\Big), \quad 0 \le \theta \le 1, \tag{1.5}$$

for the problem (1.2), and by

$$\eta(t_n+\theta h_{n+1}) = y_n+h_{n+1}\sum_{i=1}^{\nu} b_i(\theta)\hat{F}\Big(t^i_{n+1}, Y^i_{n+1}, Y_{t^i_{n+1}}, Y'_{t^i_{n+1}}\Big), \quad 0 \le \theta \le 1, \tag{1.6}$$

for the class of problems in the form (1.3).

The stages Y^i_{n+1} in (1.6) are the classical stage values of the RK method, whereas the stages $Y_{t^i_{n+1}}$ in (1.5) and (1.6) are *states* and the two methods differ from each other in how they are defined.

In Section 6 we consider the first class of methods, called *functional continuous Runge–Kutta* (FCRK) methods and denoted by $(A(\theta), b(\theta), c)$, where, for each i, the stage $Y_{t^i_{n+1}}$ is a polynomial determined by the coefficients $a_{ij}(\theta)$. A particular class of FCRK methods based on a predictor–corrector version of the collocation method, proposed by Tavernini (1971), is reported as a prototype of the class. Although FCRK methods seem the most natural and direct way to extend RK formulas to RFDEs (1.1), they were neglected for a long time and have been investigated, in their general form, only recently; see the error analysis by Maset, Torelli and Vermiglio (2005). The merit of such a class of methods is that, contrary to the standard approach, they are available in explicit form when values of y_t or y'_t are required at points of the current integration step. In this section the methods are introduced and their well-posedness, proved by Maset (2009), is reported.

In Section 7 the error analysis and order conditions are developed for the class of FCRK methods and explicit schemes of order up to four are constructed.

In Section 8 we consider the second class of methods, namely the standard approach based on continuous RK methods as described in (1.4) for initial problems of the form (1.3). In this case all the states $Y_{t^i_{n+1}}$ are given by the same function η, *i.e.*, $Y_{t^i_{n+1}} = \eta_{t^i_{n+1}}$ for all i. The numerical analysis of this approach is now consolidated and available in the cited book by Bellen and Zennaro (2003). Here we report some general results on the discrete and

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

uniform order of continuous RK methods and on the corresponding methods for the solution of (1.3). These results also serve as a background for the subsequent sections, where only the standard approach is considered.

In Section 9 we address some implementation problems arising in the use of the standard approach, in connection with the accurate computation of the breaking points. In particular, we face the paradoxical situation caused by the state-dependent delays, where the accuracy in the calculation of a breaking point depends on the accuracy of the approximate solution used to detect it, which, in turn, depends on the accuracy of the same breaking point we are trying to locate. The strategy adopted in the earlier and in the last releases of the code RADAR5 by Guglielmi and Hairer (2001, 2008) is described with details and numerical comparisons.

In Section 10 we consider RFDEs of neutral type with state-dependent delays for which the derivative of the solution is discontinuous at the initial point t_0, that is,

$$\phi'^{-}(0) \neq y'(t_0)^{+} = G(t_0, \phi, \phi').$$

Such an inequality produces a sequence of breaking points where, the delay being state-dependent, the solution may either cease to exist or bifurcate. These occurrences are investigated from both the theoretical and numerical point of view. Possible regularizations of the equation, leading to weak (or generalized) solutions defined beyond such termination and bifurcation points, are proposed and compared.

Finally, in Section 11, we address our attention to a special class of state-dependent problems in the implicit form

$$M\,u'(t) = F\big(u(t), u\big(\alpha(u(t))\big)\big),$$

where the matrix M is constant, and possibly singular. Besides including neutral state-dependent RFDEs, such problems also include singularly perturbed problems and a variety of delay differential-algebraic equations. Since these problems often have a stiff character, they usually need to be integrated by an implicit method, and therefore a Newton or quasi-Newton iterative process is needed. The efficient implementation of such iterations is investigated in detail in the case of overlapping.

2. Some particular RFDEs

Now we introduce some particular and important RFDEs (1.1). In our presentation, we divide the RFDEs into two classes defined by different data sets and corresponding to non-neutral and neutral types.

Let us first consider RFDEs (1.1) with data set $X = \mathcal{C}$.

- *Delay differential equations* (DDEs):

$$y'(t) = f\big(t, y(t), y(t-\tau_1(t)), \ldots, y(t-\tau_s(t))\big), \tag{2.1}$$

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

where $f : \mathbb{R} \times \mathbb{R}^d \times (\mathbb{R}^d)^s \to \mathbb{R}^d$ and $\tau_i : \mathbb{R} \to [0, +\infty)$, $i = 1, \dots, s$. The functions τ_i, $i = 1, \dots, s$, are called *delays*. For such equations, the functional F in (1.1) is given by

$$F(t, \varphi) = f\big(t, \varphi(0), \varphi(-\tau_1(t)), \dots, \varphi(-\tau_s(t))\big), \quad (t, \varphi) \in \mathbb{R} \times \mathcal{C}. \tag{2.2}$$

- *Delay integro-differential equations* (DIDEs):

$$y'(t) = f\left(t, y(t), \int_{t-\tau_1(t)}^{t-\tau_2(t)} k(t, t-s, y(s))\, \mathrm{d}s\right), \tag{2.3}$$

where $f : \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$, $\tau_1, \tau_2 : \mathbb{R} \to [0, +\infty)$ and $k : \mathbb{R} \times (0, +\infty) \times \mathbb{R}^d \to \mathbb{R}^d$. The functions τ_i, $i = 1, 2$, are the delays and the function k is called the *kernel*. The functional F takes the form

$$F(t, \varphi) = f\left(t, \varphi(0), \int_{-\tau_1(t)}^{-\tau_2(t)} k(t, -\vartheta, \varphi(\vartheta))\, \mathrm{d}\vartheta\right), \quad (t, \varphi) \in \mathbb{R} \times \mathcal{C}. \tag{2.4}$$

We assume that:

(K) The kernel k is measurable and, for some norm $|\cdot|$ on $\mathbb{R}^d$, for any bounded subset B of $\mathbb{R} \times \mathbb{R}^d$, the function M_B given by

$$M_B(\vartheta) = \sup_{(t,y) \in B} |k(t, -\vartheta, y)|, \quad \vartheta \in (-\infty, 0),$$

is locally integrable.

Under assumption (K), the integral in (2.4) exists and is finite for any $(t, \varphi) \in \mathbb{R} \times \mathcal{C}$. Such an assumption is satisfied if the kernel is continuous or weakly singular, *i.e.*,

$$k(t, x, y) = x^{-\alpha} \cdot a(t, x, y), \quad (t, x, y) \in \mathbb{R} \times (0, +\infty) \times \mathbb{R}^d,$$

or

$$k(t, x, y) = \log x \cdot a(t, x, y), \quad (t, x, y) \in \mathbb{R} \times (0, +\infty) \times \mathbb{R}^d,$$

where $\alpha \in [0, 1)$ and $a : \mathbb{R} \times [0, +\infty) \times \mathbb{R}^d \to \mathbb{R}^d$ is continuous.

Other RFDEs (1.1) with data set $\mathcal{C}$ are DDEs and DIDEs, where the delays also depend on the value $y(t)$.

- *State-dependent delay differential equations* (SDDDEs):

$$y'(t) = f\big(t, y(t), y\big(t - \tau_1(t, y(t))\big), \dots, y\big(t - \tau_s(t, y(t))\big)\big), \tag{2.5}$$

where $\tau_i : \mathbb{R} \times \mathbb{R}^d \to [0, +\infty)$, $i = 1, \dots, s$, and

$$F(t, \varphi) = f\big(t, \varphi(0), \varphi\big(-\tau_1(t, \varphi(0))\big), \dots, \varphi\big(-\tau_s(t, \varphi(0))\big)\big), \quad \text{for } (t, \varphi) \in \mathbb{R} \times \mathcal{C}. \tag{2.6}$$

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

- *State-dependent delay integro-differential equations* (SDDIDEs):

$$y'(t) = f\bigg(t, y(t), \int_{t-\tau_1(t,y(t))}^{t-\tau_2(t,y(t))} k(t, t-s, y(s))\,\mathrm{d}s\bigg), \tag{2.7}$$

where $\tau_1, \tau_2 : \mathbb{R} \times \mathbb{R}^d \to [0, +\infty)$ and

$$F(t,\varphi) = f\bigg(t, \varphi(0), \int_{-\tau_1(t,\varphi(0))}^{-\tau_2(t,\varphi(0))} k(t, -\vartheta, \varphi(\vartheta))\,\mathrm{d}\vartheta\bigg), \quad (t,\varphi) \in \mathbb{R} \times \mathcal{C}.$$

Another possible choice of the data set X is the set $\mathcal{LC}$ of the locally Lipschitz-continuous functions $(-\infty, 0] \to \mathbb{R}^d$, which are known to be differentiable *almost everywhere*. For $\varphi \in \mathcal{LC}$, by defining

$$\varphi'(\vartheta) = \frac{1}{2}\bigg(\limsup_{h\to 0} \frac{\varphi(\vartheta+h) - \varphi(\vartheta)}{h} + \liminf_{h\to 0} \frac{\varphi(\vartheta+h) - \varphi(\vartheta)}{h}\bigg)$$

at any point $\vartheta \in (-\infty, 0]$, the derivative φ' belongs to the set $\mathcal{B}$ of the measurable and locally bounded functions $(-\infty, 0] \to \mathbb{R}^d$.

Particular RFDEs (1.1) with the data set $\mathcal{LC}$ are the *neutral functional differential equations* (NFDEs),

$$y'(t) = G(t, y_t, y'_t), \tag{2.8}$$

where $G : \mathbb{R} \times \mathcal{C} \times \mathcal{B} \to \mathbb{R}^d$ and the functional F in (1.1) is given by

$$F(t,\varphi) = G(t, \varphi, \varphi'), \quad (t,\varphi) \in \mathbb{R} \times \mathcal{LC}.$$

Particular examples of NFDEs are as follows.

- *Neutral delay differential equations* (NDDEs):

$$\begin{aligned} y'(t) = f\Big(&t, y(t), y(t-\tau_1(t)), \dots, y(t-\tau_s(t)),\\ &y'(t-\tau_1^*(t)), \dots, y'(t-\tau_{s^*}^*(t))\Big), \end{aligned} \tag{2.9}$$

where $f : \mathbb{R} \times \mathbb{R}^d \times (\mathbb{R}^d)^s \times (\mathbb{R}^d)^{s^*} \to \mathbb{R}^d$, $\tau_i : \mathbb{R} \to [0, +\infty)$, $i = 1, \dots, s$, and $\tau_i^* : \mathbb{R} \to [0, +\infty)$, $i = 1, \dots, s^*$. The functional G in (2.8) is given by

$$\begin{aligned} G(t,\varphi,\psi) = f\Big(&t, \varphi(0), \varphi(-\tau_1(t)), \dots, \varphi(-\tau_s(t)),\\ &\psi(-\tau_1^*(t)), \dots, \psi(-\tau_{s^*}^*(t))\Big),\\ &\text{for } \ (t,\varphi,\psi) \in \mathbb{R} \times \mathcal{C} \times \mathcal{B}. \end{aligned}$$

- *Neutral delay integro-differential equations* (NDIDEs):

$$y'(t) = f\bigg(t, y(t), \int_{t-\tau_1(t)}^{t-\tau_2(t)} k\big(t, t-s, y(s), y'(s)\big)\,\mathrm{d}s\bigg), \tag{2.10}$$

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

where $f : \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$, $\tau_1, \tau_2 : \mathbb{R} \to [0, +\infty)$ and $k : \mathbb{R} \times (0, +\infty) \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$. The functional G takes the form

$$G(t, \varphi, \psi) = f\Big(t, \varphi(0), \int_{-\tau_1(t)}^{-\tau_2(t)} k\big(t, -\vartheta, \varphi(\vartheta), \psi(\vartheta)\big)\, \mathrm{d}\vartheta\Big), \quad \text{for } (t, \varphi, \psi) \in \mathbb{R} \times \mathcal{C} \times \mathcal{B}.$$

As for the non-neutral case, we assume that:

(NK) The kernel k is measurable, and, for some norm $|\cdot|$ on $\mathbb{R}^d$, for any bounded subset B of $\mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^d$, the function M_B given by

$$M_B(\vartheta) = \sup_{(t,y,z) \in B} |k(t, -\vartheta, y, z)|, \quad \vartheta \in (-\infty, 0),$$

is locally integrable.

Other RFDEs with data set $\mathcal{LC}$ are NDDEs and NDIDEs with state-dependent delays.

- *Neutral state-dependent delay differential equations* (NSDDDEs):

$$y'(t) = f\Big(t, y(t), y\big(t - \tau_1(t, y(t))\big), \dots, y\big(t - \tau_s(t, y(t))\big), y'\big(t - \tau_1^*(t, y(t))\big), \dots, y'\big(t - \tau_{s^*}^*(t, y(t))\big)\Big), \tag{2.11}$$

where $\tau_i : \mathbb{R} \times \mathbb{R}^d \to [0, +\infty)$, $i = 1, \dots, s$, $\tau_i^* : \mathbb{R} \times \mathbb{R}^d \to [0, +\infty)$, $i = 1, \dots, s^*$, and

$$G(t, \varphi, \psi) = f\Big(t, \varphi(0), \varphi\big(-\tau_1(t, \varphi(0))\big), \dots, \varphi\big(-\tau_s(t, \varphi(0))\big), \psi\big(-\tau_1^*(t, \varphi(0))\big), \dots, \psi\big(-\tau_{s^*}^*(t, \varphi(0))\big)\Big), \quad \text{for } (t, \varphi, \psi) \in \mathbb{R} \times \mathcal{C} \times \mathcal{B}.$$

- *Neutral state-dependent delay integro-differential equations* (NSDDIDEs):

$$y'(t) = f\Big(t, y(t), \int_{t - \tau_1(t, y(t))}^{t - \tau_2(t, y(t))} k\big(t, t - s, y(s), y'(s)\big)\, \mathrm{d}s\Big), \tag{2.12}$$

where $\tau_1, \tau_2 : \mathbb{R} \times \mathbb{R}^d \to [0, +\infty)$ and

$$G(t, \varphi, \psi) = f\Big(t, \varphi(0), \int_{-\tau_1(t, \varphi(0))}^{-\tau_2(t, \varphi(0))} k\big(t, -\vartheta, \varphi(\vartheta), \psi(\vartheta)\big)\, \mathrm{d}\vartheta\Big), \quad \text{for } (t, \varphi, \psi) \in \mathbb{R} \times \mathcal{C} \times \mathcal{B}. \tag{2.13}$$

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

3. Discontinuity points and vanishing delays

In this section we describe two particular situations caused by the presence of two kinds of points: those where some derivative of the solution is not continuous and those where the delay vanishes.

3.1. Discontinuity points

In this section we briefly analyse the propagation of *discontinuity points*, also called *breaking points*, for the derivatives of the solution of (1.2) along the integration interval.

In order to illustrate this phenomenon, for the sake of simplicity we confine ourselves to the particular class of DDEs,

$$\begin{aligned} y'(t) &= f\big(t, y(t), y\big(t-\tau(t,y(t))\big)\big), \quad t_0 \le t \le t_0+T, \\ y(t) &= \phi(t), \quad t \le t_0, \end{aligned} \tag{3.1}$$

and NDDEs,

$$\begin{aligned} y'(t) &= f\big(t, y(t), y\big(t-\tau(t,y(t))\big), y'\big(t-\tau(t,y(t))\big)\big), \quad t_0 \le t \le t_0+T, \\ y(t) &= \phi(t), \quad t \le t_0. \end{aligned} \tag{3.2}$$

First consider equation (3.1) and assume that the *deviated argument*

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

satisfies $\alpha(t) < t_0$ for some points $t \in [t_0, t_0+T]$. Moreover, assume that the solution $y(t)$ does not link smoothly to the initial function $\phi(t)$ at t_0, that is,

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

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

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

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

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

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

and hence

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

and

$$y''(\xi_{1,i})^- = \frac{\partial f}{\partial t}\big(\xi_{1,i}, y(\xi_{1,i}), y(t_0)\big) + \frac{\partial f}{\partial y}\big(\xi_{1,i}, y(\xi_{1,i}), y(t_0)\big)y'(\xi_{1,i}) + \frac{\partial f}{\partial x}\big(\xi_{1,i}, y(\xi_{1,i}), y(t_0)\big)\phi'(t_0)^-\alpha'(\xi_{1,i}). \tag{3.5}$$

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 at $\xi_{1,i}$.

These jump discontinuities in y'' are called 1-*level primary discontinuities*. By differentiating (3.3), 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

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

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 (3.1) becomes increasingly smooth as the primary discontinuity level increases. This increase in the regularity of $y(t)$ will be referred to as *smoothing of the solution*.

Definition 3.1. Every point where some derivative $y^{(s)}$ jumps will be called a *discontinuity point* or *breaking point*. We also say that a breaking point ξ *has order* k if the solution is C^k-continuous at ξ. In particular, by $k = -1$ we mean that the solution is discontinuous at ξ.

On the contrary, the same argument applied to (3.2) reveals that, for neutral DDEs, smoothing does not occur and, in general, the solution remains C^0-continuous at any primary discontinuity point where the derivative y' jumps. This motivates the weaker definition of the solution in the 'almost everywhere' sense given in Definition 1.1. Obviously, if the splicing condition

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

holds, no discontinuities propagate from t_0 and the solution is meant in the classical sense.

It is also remarkable that, if $\alpha(t) \geq t_0$ for all $t \geq t_0$, then no values of y and/or y' are needed in (3.1) and (3.2) behind t_0 and, therefore, no primary discontinuities propagate from t_0.

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

Other discontinuities can appear if the functions f, τ and ϕ in (3.1) and (3.2) have some discontinuities with respect to t in some of their derivatives. Then such discontinuities are also propagated by the deviated argument $\alpha(t)$, according to the primary discontinuity propagation rule, and are called *secondary discontinuities*.

From the numerical point of view, it is important to analyse how the discontinuity points propagate through the integration interval $[t_0, t_0+T]$, and how smoothness possibly increases at any discontinuity point with respect to its *ancestor*, the discontinuity point from which it originates. In fact, it is known that every step-by-step numerical method for initial value problems (IVPs) 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, discontinuity points of a suitable level ought to be included in the mesh.

To finish with, we briefly consider the difficulties related to the case when the delay $\tau(t, y(t))$ is state-dependent. In order to locate the discontinuities, one should in principle apply the general propagation rule

$$\xi_{k,j} - \tau\big(\xi_{k,j}, y(\xi_{k,j})\big) = \xi_{k-1,i} \quad \text{for some} \quad i, \tag{3.6}$$

and solve it for $\xi_{k,j}$. Because the delay is dependent on $y(t)$, this cannot be done *a priori* without any 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 $\xi_{k,j}$.

In conclusion, the impossibility of locating the discontinuity points *a priori* makes the implementation and convergence analysis of numerical methods for (3.1) and (3.2) a rather complicated task, which will be further investigated in Section 9.

For a more detailed analysis of the propagation of discontinuity points in systems of DDEs with different delays we refer the reader to Bellen and Zennaro (2003) and the references therein.

3.2. Non-vanishing delays

The theoretical analysis of (1.2) for the classes of RFDEs considered in Section 2, as well as the development of numerical methods for its approximate solution, is considerably simplified if the problem reduces to a finite sequence of IVPs for ordinary differential equations (ODEs) on any interval $[t_0, t_0 + T]$. In order to characterize such an occurrence, let us consider the following condition on a delay τ.

(H_1^*) There exists a constant $\tau_0 > 0$ such that $\tau(t) \geq \tau_0$ for all $t \in \mathbb{R}$ or, for state-dependent delays, $\tau(t, z) \geq \tau_0$ for all $t \in \mathbb{R}$ and $z \in \mathbb{R}^d$.

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

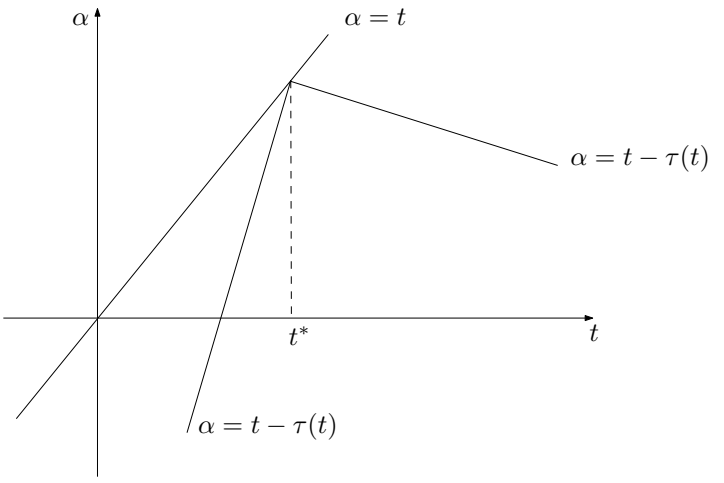


Figure 3.1. An example of a vanishing delay satisfying condition (HH_1^*).

A delay satisfying the condition (H_1^*) is said to be *non-vanishing*; otherwise it is called *vanishing*. It is evident that if all the delays in the given equation are non-vanishing, then the IP reduces on any interval to a finite sequence of IVPs for ODEs.

However, to be non-vanishing is not a necessary condition for the delays in order to reduce the problem to a finite sequence of ordinary ones. In fact, this occurs if and only if any delay τ fulfils this weaker condition.

(HH_1^*) For any $\sigma \in \mathbb{R}$ and $T > 0$, there exist $\sigma_0, \sigma_1, \ldots, \sigma_{K-1}, \sigma_K$ such that

$$\sigma = \sigma_0 < \sigma_1 < \cdots < \sigma_{K-1} < \sigma_K = \sigma + T$$

and

$$t - \tau(t) \le \sigma_k, \quad t \in [\sigma_k, \sigma_{k+1}], \quad \text{and} \quad k = 0, 1, \ldots, K-1,$$

or

$$t - \tau(t, z) \le \sigma_k, \quad t \in [\sigma_k, \sigma_{k+1}], \quad z \in \mathbb{R}^d \quad \text{and} \quad k = 0, 1, \ldots, K-1,$$

in the case of a state-dependent delay.

A delay τ satisfying (HH_1^*) is called *weakly non-vanishing*. A delay which is not weakly non-vanishing is called *strongly vanishing*. An example of a vanishing delay which is weakly non-vanishing is $\tau(t) = t - [t]$, where $[\cdot]$ denotes the greatest integer function. Such a delay occurs in some retarded differential real-life models (see Cooke and Wiener (1984)). On the other hand, an example of a vanishing delay τ which is strongly vanishing is depicted in Figure 3.1, where $\tau(t^*) = 0$. For an equation involving such a delay, an IP (1.2) with $t_0 < t^*$ may reduce to a finite sequence of IVPs for ODEs only on intervals $[t_0, t_0 + T]$ where $t_0 + T < t^*$.

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

Whether or not the problem is reducible to a sequence of IVPs for ODEs is reflected in the different hypotheses required for the existence and uniqueness of the solution.

Two significant cases of equations involving a strongly vanishing delay are the classical Volterra equation

$$y'(t) = f\left(t, y(t), \int_0^t k(t, t-s, y(s))\,\mathrm{d}s\right), \quad t \geq 0,$$

and the *pantograph equation*

$$y'(t) = f(t, y(t), y(qt)), \quad t \geq 0, \quad 0 < q < 1.$$

The former corresponds to equation (2.3), with $\tau_1(t) = t$ and $\tau_2(t) = 0$ at any point t; the latter corresponds to (2.1), with $\tau_1(t) = (1-q)t$, which equals zero only at $t = 0$.

4. Existence and uniqueness

Now, for the different types of RFDEs we give conditions under which the IP (1.2) has a *unique local solution*, that is, there exists $T > 0$ such that the IP has a unique solution on $(-\infty, t_0 + T]$. Most of the results of this and the next section are taken from Maset (2009). To this end, we let L_T denote the linear space $L^\infty([0,T], \mathbb{R}^d)$, *i.e.*, the linear space of the equivalence classes of the Lebesgue-measurable and essentially bounded functions $[0,T] \to \mathbb{R}^d$, and let $L_T^\diamond$ be the linear space of the Lebesgue-measurable and bounded functions $[0,T] \to \mathbb{R}^d$. Note that $L_T^\diamond$ can be embedded in L_T by identifying any function with its equivalence class.

Moreover, we assume the data set X satisfies the following assumptions.

(DS1) For any $\varphi \in X$ and $s \in (-\infty, 0]$, we have $\varphi_s \in X$.

(DS2) For any $\varphi \in X$, $T > 0$ and $z \in L_T$, we have $v(\varphi, z)_T \in X$, where $v(\varphi, z) : (-\infty, T] \to \mathbb{R}^d$ is the continuous function given by

$$v(\varphi, z)(t) = \begin{cases} \varphi(0) + \int_0^t z(s)\,\mathrm{d}s & \text{if } \ t \in [0,T], \\ \varphi(t) & \text{if } \ t \in (-\infty, 0]. \end{cases} \tag{4.1}$$

Since assumption (DS1) holds, we also have

$$v(\varphi, z)_t = \big(v(\varphi, z)_T\big)_{t-T} \in X, \quad t \leq T, \tag{4.2}$$

in (DS2).

It is clear that both data sets $\mathcal{C}$ and $\mathcal{LC}$ satisfy (DS1) and (DS2).

Finally, as well as assumptions (DS1) and (DS2) on the data set, we assume that the RFDE (1.1) satisfies the following *Boundedness Assumption*.

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

(BA) For any $\sigma \in \mathbb{R}$, $\varphi \in X$, $T > 0$ and $z \in L_T$, the function

$$t \mapsto F\big(\sigma + t, v(\varphi, z)_t\big), \quad t \in [0, T],$$

belongs to $L_T^\diamond$.

For the various types of RFDEs presented above, the assumption (BA) holds under minimal conditions. This is stated in the following two propositions.

Proposition 4.1. An NSDDDE (an NDDE) satisfies property (BA) if:

- the function f is measurable and locally bounded;
- the delays τ_j, $j = 1, \dots, s$, and τ_j^*, $j = 1, \dots, s^*$, are measurable and locally bounded.

An analogous proposition holds for SDDDEs and DDEs (which are not particular NSDDDEs or NDDEs, since the data set for the former is larger).

Proposition 4.2. An NSDDIDE (an NDIDE) satisfies property (BA) if:

- the function f is measurable and locally bounded;
- the delays τ_j, $j = 1, 2$, are measurable and locally bounded;
- the kernel k satisfies assumption (NK).

A similar proposition holds for SDDIDEs and DIDEs.

Since the Boundedness Assumption holds, for given $\sigma \in \mathbb{R}$, $\varphi \in X$ and $T > 0$, we can introduce the map

$$Q_T(\sigma, \varphi) : L_T \to L_T^\diamond$$

defined by

$$\big[Q_T(\sigma, \varphi)(z)\big](t) = F\big(\sigma + t, v(\varphi, z)_t\big), \quad t \in [0, T] \ \text{ and } \ z \in L_T. \tag{4.3}$$

For the analysis of the numerical methods introduced in the next section, it is useful to introduce the map

$$Q_T^\diamond(\sigma, \varphi) = Q_T(\sigma, \varphi)|_{C_T} : C_T \to L_T^\diamond, \tag{4.4}$$

where C_T is the subspace of L_T (and of $L_T^\diamond$) of the continuous functions.

In the following, we consider $Q_T(\sigma, \varphi)$ as a map $L_T \to L_T$ by embedding $L_T^\diamond$ in L_T.

There is a link between the map $Q_T(t_0, \phi)$ and a solution of (1.2) on $(-\infty, t_0 + T]$, which is given in the following basic theorem.

Theorem 4.1. Let $y : (-\infty, t_0 + T] \to \mathbb{R}^d$ and let x be the shift function given by

$$x(t) = y(t_0 + t), \quad t \in (-\infty, T]. \tag{4.5}$$

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

The function y is a solution of the IP (1.2) on $(-\infty, t_0+T]$ if and only if

$$x = v(\phi, z^*)$$

for some fixed point z^* of the map $Q_T(t_0, \phi)$.

Proof. Let $y : (-\infty, t_0+T] \to \mathbb{R}^d$ be a solution of the IP (1.2). Then the shift function x is continuous, $x_t \in X$ for all $t \in [0,T]$, the function

$$z^*(t) = F(t_0+t, x_t), \quad t \in [0,T],$$

belongs to $L_T^\diamond$ (and, then, to L_T) and, for $t \in (-\infty, T]$, we have

$$\begin{aligned} x(t) &= \begin{cases} \phi(0) + \int_0^t F(t_0+s, x_s)\,\mathrm{d}s & \text{if } \ t \in [0,T], \\ \phi(t) & \text{if } \ t \in (-\infty, 0] \end{cases} \\ &= v(\phi, z^*)(t). \end{aligned}$$

Since

$$\begin{aligned} \big[Q_T(t_0,\phi)(z^*)\big](t) &= F\big(t_0+t, v(\phi,z^*)_t\big) \\ &= F(t_0+t, x_t) = z^*(t), \quad t \in [0,T], \end{aligned}$$

z^* turns out to be a fixed point of the map $Q_T(t_0,\phi)$.

Vice versa, let $z^* \in L_T$ be a fixed point of the map $Q_T(t_0,\phi)$, and let

$$x(t) = v(\phi, z^*)(t), \quad t \in (-\infty, T].$$

The function x is continuous, $x_t = v(\phi,z^*)_t \in X$ for all $t \in [0,T]$ (recall (4.2)), the function

$$t \mapsto F(t_0+t, x_t) = F\big(t_0+t, v(\phi,z^*)_t\big), \quad t \in [0,T],$$

belongs to $L_T^\diamond$ by the Boundedness Assumption, and, for $t \in (-\infty, T]$, we have

$$\begin{aligned} x(t) &= v(\phi,z^*)(t) \\ &= v\big(\phi, Q_T(t_0,\phi)(z^*)\big)(t) \\ &= \begin{cases} \phi(0) + \int_0^t \big[Q_T(t_0,\phi)(z^*)\big](s)\,\mathrm{d}s & \text{if } \ t \in [0,T], \\ \phi(t) & \text{if } \ t \in (-\infty, 0] \end{cases} \\ &= \begin{cases} \phi(0) + \int_0^t F\big(t_0+s, v(\phi,z^*)_t\big)(s)\,\mathrm{d}s & \text{if } \ t \in [0,T], \\ \phi(t) & \text{if } \ t \in (-\infty, 0] \end{cases} \\ &= \begin{cases} \phi(0) + \int_0^t F(t_0+s, x_s)\,\mathrm{d}s & \text{if } \ t \in [0,T], \\ \phi(t) & \text{if } \ t \in (-\infty, 0]. \end{cases} \end{aligned}$$

Hence,

$$y(t) = x(t-t_0), \quad t \in (-\infty, t_0+T],$$

is a solution of (1.2) on $(-\infty, t_0+T]$. □

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

By the previous result, it is clear that there exists a unique solution of (1.2) on $(-\infty, t_0 + T]$ if and only if the map $Q_T(t_0, \phi)$ has a unique fixed point.

The reduction of the problem of existence and uniqueness of a solution on $(-\infty, t_0 + T]$ to the existence and uniqueness of fixed points of the map $Q_T(t_0, \phi)$ allows us to prove the following theorems.

Theorem 4.2. Consider a DDE (2.1). If:

- the function $f(t, y_0, y_1, \ldots, y_s)$ is measurable, locally bounded, locally Lipschitz-continuous with respect to the argument y_0 and locally Lipschitz-continuous with respect to those arguments y_j, $j = 1, \ldots, s$, such that the delay τ_j is strongly vanishing;
- the delays τ_j, $j = 1, \ldots, s$, are measurable and locally bounded;

then any IP (1.2) for the DDE has a unique local solution $\forall \phi \in \mathcal{C}$.

Theorem 4.3. Consider an SDDDE (2.5). If:

- the function $f(t, y_0, y_1, \ldots, y_s)$ is measurable, locally bounded and locally Lipschitz-continuous with respect to any argument y_j, $j = 0, 1, \ldots, s$;
- the delays $\tau_j(t, y)$, $j = 1, \ldots, s$, are measurable, locally bounded and locally Lipschitz-continuous with respect to the argument y;

then any IP (1.2) for the SDDDE has a unique local solution $\forall \phi \in \mathcal{LC}$.

Theorem 4.4. Consider an NDDE (2.9). If:

- the function $f(t, y_0, y_1, \ldots, y_s, z_1, \ldots, z_{s^*})$ is measurable, locally bounded, locally Lipschitz-continuous with respect to the argument y_0, locally Lipschitz-continuous with respect to those arguments y_j, $j = 1, \ldots, s$, such that the delay τ_j is strongly vanishing, and globally Lipschitz-continuous of constant l_j with respect to those arguments z_j, $j = 1, \ldots, s^*$, such that the delay τ_j^* is strongly vanishing;
- $$\sum_{\substack{j=1,\ldots,s^* \\ \tau_j^* \text{ is strongly vanishing}}} l_j < 1;$$
- the delays τ_j, $j = 1, \ldots, s$, are measurable and locally bounded;
- the delays τ_j^*, $j = 1, \ldots, s^*$, are measurable, locally bounded and each strongly vanishing delay τ_j^* is such that:
  - for each $\sigma \in \mathbb{R}$, there exists $T^* > 0$ such that, for any subset A of $(0, T^*]$ of zero measure, the set
    $$\left\{ t \in [0, T^*] \mid t - \tau_j^*(\sigma + t) \in A \right\}$$
    has measure zero;

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

then any IP (1.2) for the NDDE has a unique local solution $\forall \phi \in \mathcal{LC}$.

It is remarkable that, in order to obtain an existence theorem for the more restricted class of NSDDDEs, it is not sufficient to impose simultaneously the conditions required in the previous theorems for the DDEs with state-dependent delay and of neutral type. In fact, the following theorem, besides considering more restrictive conditions on the functional F, requires additional specific conditions on the initial function depending on the functional itself.

Theorem 4.5. Consider an NSDDDE (2.11). If:

- the function $f(t, y_0, y_1, \dots, y_s, z_1, \dots, z_{s^*})$ is measurable, locally bounded and locally Lipschitz-continuous with respect to the arguments $y_j, j = 0, 1, \dots, s$ and $z_j, j = 1, \dots, s^*$;
- the delays $\tau_j(t, y)$, $j = 1, \dots, s$, are measurable, locally bounded and locally Lipschitz-continuous with respect to the argument y;
- the delays $\tau_j^*(t, y)$, $j = 1, \dots, s^*$, are weakly non-vanishing, measurable, locally bounded and locally Lipschitz-continuous with respect to the argument y;

then any IP (1.2) for the NSDDDE has a unique local solution $\forall \phi \in \mathcal{LC}^1$, $\mathcal{LC}^1$ being the set of continuously differentiable functions with locally Lipschitz-continuous derivative, provided ϕ satisfies the splicing condition

$$\varphi'(0) = f\Big(t_0, \varphi(0), \varphi(-\tau_1(t_0)), \dots, \varphi(-\tau_s(t_0)), \varphi'(-\tau_1^*(t_0)), \dots, \varphi'(-\tau_{s^*}^*(t_0))\Big). \tag{4.6}$$

Now we consider integro-differential equations.

Theorem 4.6. Consider an NDIDE (2.3). If:

- the function $f(t, y_0, y_1)$ is measurable, locally bounded, locally Lipschitz-continuous with respect to the argument y_0 and, if τ_1 or τ_2 is strongly vanishing, also locally Lipschitz-continuous with respect to the arguments y_1;
- the delays τ_1 and τ_2 are measurable and locally bounded;
- the kernel k satisfies assumption (NK) and, if τ_1 or τ_2 is strongly vanishing, also the assumption:

  (NK1) There exists $\widehat{T} > 0$ such that, for any bounded subset B of $\mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^d$, the Lipschitz constants

$$\theta \mapsto \sup_{\substack{(t,y_1,z)\in B \\ (t,y_2,z)\in B \\ y_1 \neq y_2}} \frac{|k(t, -\theta, y_1, z) - k(t, -\theta, y_2, z)|}{|y_1 - y_2|}, \quad \theta \in [-\widehat{T}, 0),$$

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

and

$$\theta \mapsto \sup_{\substack{(t,y,z_1)\in B\\(t,y,z_2)\in B\\z_1\neq z_2}} \frac{|k(t,-\theta,y,z_1)-k(t,-\theta,y,z_2)|}{|z_1-z_2|}, \quad \theta\in[-\widehat{T},0),$$

are integrable;

then any IP (1.2) for the NDIDE has a unique local solution $\forall\phi\in\mathcal{LC}$.

Theorem 4.7. Consider an NSDDIDE (2.12). If:

- the function $f(t,y_0,y_1)$ is measurable, locally bounded, locally Lipschitz-continuous with respect to the arguments y_0 and y_1;
- the delays $\tau_1(t,y)$ and $\tau_2(t,y)$ are measurable, locally bounded and locally Lipschitz-continuous with respect to the argument y;
- the kernel $k(t,\theta,y,z)$ satisfies conditions (NK), (NK1) and:

  (NK2) For any bounded subset B of $\mathbb{R}\times\mathbb{R}^d\times\mathbb{R}^d$, the function M_B defined in assumption (NK) is essentially locally bounded on the images $\tau_j(\mathbb{R}\times\mathbb{R}^d)$, $j=1,2$;

then any IP (1.2) for the NSDDIDE has a unique local solution $\forall\phi\in\mathcal{LC}$.

Analogous theorems are valid for DIDEs and SDDIDEs by replacing assumption (NK) with (K) and by modifying assumptions (NK1) and (NK2) in an obvious way. Note that for RFDEs of integral type the existence and uniqueness of the solution is achieved in their own data set.

4.1. The solution map

In the foregoing theorems we have seen that existence and uniqueness of the local solutions is not always guaranteed for any initial data in the data set. In particular, when the delay is state-dependent, the existence and uniqueness is guaranteed on a restriction of the data set, namely $\mathcal{LC}$ for SDDDEs and the subset of $\mathcal{LC}^1$ satisfying the splicing conditions (4.6) for NSDDDEs.

Therefore, it is worth introducing the concept of *state set*, as the subset of the data set for which the local solution uniquely exists and can be prolonged to a maximal solution.

Definition 4.1. A subset Y of the data set X is called a *state set* if, for any $t_0\in\mathbb{R}$ and $\phi\in Y$, the IP (1.2) has a unique local solution y, defined on $(-\infty,t_0+T]$, and $y_t\in Y$ holds for all $t\in[t_0,t_0+T]$.

If Y is a state set for the RFDE (1.1), then, for any $\sigma\in\mathbb{R}$ and $\varphi\in Y$, the IP (1.2) with $t_0=\sigma$ and $\phi=\varphi$ has a unique local solution. Such a solution

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

can be prolonged to a unique maximal solution $y(\sigma,\varphi):(-\infty,\sigma+T_{\max})\to\mathbb{R}^d$, where $T_{\max}=T_{\max}(\sigma,\varphi)\in(0,+\infty]$.

Summarizing, under the conditions stated in the previous theorems, we have the following state sets for the various classes of RFDEs:

- the set $\mathcal{C}$ is a state set for DDEs, DIDEs and SDDIDEs,
- the set $\mathcal{LC}$ is a state set for SDDDEs, NDDEs, NDIDEs, and NSDDIDEs,
- the subset of $\mathcal{LC}^1$ satisfying the splicing condition (4.6) is the state set for NSDDDEs.

The *solution map* for the RFDE is the map V associating the state

$$V(\sigma,\varphi,T)=y(\sigma,\varphi)_{\sigma+T}\in Y$$

at the point $\sigma+T$ to the triple (σ,φ,T), where $\sigma\in\mathbb{R}$, $\varphi\in Y$ and $T\in[0,T_{\max}(\sigma,\varphi))$. This state determines the future states $y(\sigma,\varphi)_{\sigma+T+\Delta}$, where $\Delta\in[0,T_{\max}(\sigma,\varphi)-T)$, by

$$V(\sigma,\varphi,T+\Delta)=V\big(\sigma+T,V(\sigma,\varphi,T),\Delta\big).$$

We can conclude this section by the following basic proposition based on Theorem 4.1.

Proposition 4.3. The solution map can be expressed as

$$V(\sigma,\varphi,T)=v\big(\varphi,z^*(\sigma,\varphi,T)\big)_T, \tag{4.7}$$

where $v(\cdot,\cdot)$ is defined in (4.1) and $z^*(\sigma,\varphi,T)$ is the unique fixed point of the map $Q_T(\sigma,\varphi)$ defined in (4.3).

5. Numerical methods for RFDEs

A numerical method for a RFDE with state set Y provides a map $\widetilde{V}$ approximating the solution map V. More precisely, the map $\widetilde{V}$ associates data

$$\widetilde{V}(\sigma,\varphi,h)\in Y$$

approximating the state $V(\sigma,\varphi,h)$ to the triple (σ,φ,h), where $\sigma\in\mathbb{R}$, $\varphi\in Y$ and $h\in[0,H_{\max}(\sigma,\varphi))$. Here, the third argument of $\widetilde{V}$ is denoted by h and not by T, since it has the meaning of *step-size* in the integration process described below. Moreover, note that its definition domain is $[0,H_{\max}(\sigma,\varphi))$, which can be different from the definition domain of the third argument T of V.

We deal with the problem of the computation of the solution $y=y(t_0,\phi)$ of the IP (1.2), where $\phi\in Y$, on the integration window $[t_0,t_0+T]$, where $T\in(0,T_{\max}(t_0,\phi))$. We refer to it as the *integration problem* (t_0,ϕ,T).

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

When a numerical method providing a map $\widetilde{V}$ is applied to the integration problem (t_0, ϕ, T) with the mesh

$$\Delta = \{t_n \mid n = 0, 1, 2, \ldots, N_\Delta\}, \quad t_0 < t_1 < t_2 < \cdots < t_{N_\Delta} = t_0 + T, \tag{5.1}$$

it yields the finite sequence of states

$$\{\phi_n\}_{n=0,1,2,\ldots,N_\Delta},$$

where $\phi_n \in Y$ is an approximation of the exact state y_{t_n} at the mesh point t_n, given by the recursion

$$\begin{aligned} \phi_{n+1} &= \widetilde{V}(t_n, \phi_n, h_{n+1}), \quad n = 0, 1, \ldots, N_\Delta - 1, \\ \phi_0 &= \phi. \end{aligned} \tag{5.2}$$

Here, $h_{n+1} = t_{n+1} - t_n$ is the $(n+1)$th step-size and satisfies $h_{n+1} \in [0, H_{\max}(t_n, \phi_n))$. Note that the sequence

$$\{y_{t_n}\}_{n=0,1,2,\ldots,N_\Delta}$$

of the exact states satisfies the recursion

$$\begin{aligned} y_{t_{n+1}} &= V(t_n, y_{t_n}, h_{n+1}), \quad n = 0, 1, \ldots, N_\Delta - 1, \\ y_{t_0} &= \phi. \end{aligned}$$

In this paper we consider numerical methods such that

$$\widetilde{V}(\sigma, \varphi, h) = v\big(\varphi, \widetilde{z}^*(\sigma, \varphi, h)\big)_h \tag{5.3}$$

(recall Proposition 4.3), where $\widetilde{z}^*(\sigma, \varphi, h) \in L_h^\diamond$ is a suitable function approximating the fixed point $z^*(\sigma, \varphi, h)$ of the map $Q_h(\sigma, \varphi)$. We define the *local error functions* $e(\sigma, \varphi, h)$ and $E(\sigma, \varphi, h)$ at (σ, φ) with step-size h by

$$e(\sigma, \varphi, h) = \widetilde{z}^*(\sigma, \varphi, h) - z^*(\sigma, \varphi, h) \tag{5.4}$$

and

$$\begin{aligned} E(\sigma, \varphi, h) &= \big(v\big(\varphi, \widetilde{z}^*(\sigma, \varphi, h)\big) - v\big(\varphi, z^*(\sigma, \varphi, h)\big)\big)|_{[0,h]} \\ &= v\big(0, e(\sigma, \varphi, h)\big)|_{[0,h]} \\ &= \int e(\sigma, \varphi, h), \end{aligned} \tag{5.5}$$

where, for $z \in L_h$, $\int z$ denotes the primitive of z, *i.e.*,

$$\Big(\int z\Big)(t) = \int_0^t z(s)\,\mathrm{d}s, \quad t \in [0, h].$$

5.1. Convergence analysis

Now, we tackle the study of the error in the numerical solution of integration problems by a numerical method providing approximations of type (5.3).

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

To this end, we consider:

- a norm $|\cdot|$ on $\mathbb{R}^d$;
- a norm on the spaces $L_h^\diamond$ and $C_h \subseteq L_h^\diamond$,

$$\|z\| = \sup_{t\in[0,T]} |z(t)|, \quad z \in L_h^\diamond.$$

We consider an integration problem (t_0, ϕ, T) and the numerical computation of the solution $y = y(t_0, \phi)$ on the integration window $[t_0, t_0 + T]$ given by the process (5.2). Since we need to measure the error between the states y_{t_n} and the approximated states ϕ_n, $n = 1, 2, \ldots, N_\Delta$, we introduce the distance

$$d(\varphi, \psi) = \sup_{\theta\in(-\infty,0]} |\varphi(\theta) - \psi(\theta)| \in [0, +\infty]$$

between two data $\varphi, \psi \in \mathcal{C}$.

Note that, for $n = 0, 1, \ldots, N_\Delta - 1$, we have

$$\phi_n = (\phi_{n+1})_{-h_{n+1}} \quad \text{and} \quad y_{t_n} = (y_{t_{n+1}})_{-h_{n+1}}$$

and then

$$d(\phi_n, y_{t_n}) \le d(\phi_{n+1}, y_{t_{n+1}}).$$

Hence, we define

$$\mathbf{E}(\Delta) = d(\phi_{N_\Delta}, y_{t_0+T})$$

the *global error* in the numerical solution of the integration problem (t_0, ϕ, T) with mesh Δ. As we will see below, the global error is linked to the local errors at the mesh points

$$\begin{aligned} \mathbf{E}_{n+1}(\Delta) &= \|E(t_n, y_{t_n}, h_{n+1})\|, \\ \mathbf{E}^0_{n+1}(\Delta) &= |E(t_n, y_{t_n}, h_{n+1})(h_{n+1})|, \\ \mathbf{e}_{n+1}(\Delta) &= \|e(t_n, y_{t_n}, h_{n+1})\|, \end{aligned} \tag{5.6}$$

where $n = 0, 1, \ldots, N_\Delta - 1$.

Our aim is to study the infinitesimal order of $\mathbf{E}(\Delta)$ as $h_\Delta \to 0$, where

$$h_\Delta = \max_{n=1,2,\ldots,N_\Delta} h_n$$

is the maximum step-size in the mesh Δ.

In the next definition, we introduce the concept of global order of the given numerical method for RFDEs.

Definition 5.1. Let r be a positive integer and let $\mathcal{F}$ be a family of integration problems (t_0, ϕ, T) such that $y(t_0, \phi)|_{[t_0,t_0+T]}$ is piecewise smooth (*i.e.*, piecewise C^m for some positive integer m). The method has *global order* r on $\mathcal{F}$ if, for any integration problem $(t_0, \phi, T) \in \mathcal{F}$, we have

$$\mathbf{E}(\Delta_k) = \mathcal{O}(h^r_{\Delta^{(k)}}), \quad k \to +\infty,$$

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

for any sequence $\{\Delta^{(k)}\}$ of meshes on $[t_0, t_0+T]$ such that

$$h_{\Delta^{(k)}} \to 0, \quad k \to +\infty,$$

and, for any k, $\Delta^{(k)}$ includes all the breaking points of $y(t_0,\phi)|_{[t_0,t_0+T]}$.

Now, we distinguish between RFDEs with data set $\mathcal{C}$ and RFDEs with data set $\mathcal{LC}$.

RFDEs with data set $\mathcal{C}$

For RFDEs with data set $\mathcal{C}$, we introduce the concepts of uniform order and discrete order of the given method.

Definition 5.2. Let q be a positive integer and let $\mathcal{F}$ be a family of integration problems (t_0,ϕ,T) for which $y(t_0,\phi)|_{[t_0,t_0+T]}$ is piecewise smooth. The method has *uniform order* q on $\mathcal{F}$ if, for any integration problem $(t_0,\phi,T)\in\mathcal{F}$, there exist constants $H>0$ and $C>0$ such that

$$\|E(t,y_t,h)\| \le Ch^{q+1},$$

for any $t\in[t_0,t_0+T)$ and $h\in[0,H_{\max}(t,y_t))$ such that $h\le T-t$, and the interval $(t,t+h)$ does not contain breaking points of $y(t_0,\phi)$ and $h<H$.

Definition 5.3. Let p be a positive integer and let $\mathcal{F}$ be a family of integration problems (t_0,ϕ,T) for which $y(t_0,\phi)|_{[t_0,t_0+T]}$ is piecewise smooth. The method has *discrete order* p on $\mathcal{F}$ if, for any integration problem $(t_0,\phi,T)\in\mathcal{F}$, there exist constants $H>0$ and $C>0$ such that

$$|E(t,y_t,h)(h)| \le Ch^{p+1},$$

for any $t\in[t_0,t_0+T)$ and $h\in[0,H_{\max}(t,y_t))$ such that $h\le T-t$, and the interval $(t,t+h)$ does not contain breaking points of $y(t_0,\phi)$ and $h<H$.

In order to link the uniform and discrete orders to the global order, we introduce the concept of stability of the method.

Definition 5.4. Let $\mathcal{F}$ be a family of integration problems (t_0,ϕ,T). The method is *stable* on $\mathcal{F}$ if, for any integration problem (t_0,ϕ,T), there exist $\delta>0$, $\overline{H}>0$ and $L\ge 0$ such that

$$\overline{H} \le H_{\max}(t,\varphi), \quad t\in[t_0,T_0+T),$$

and

$$\|\widetilde{z}(t,\varphi,h)-\widetilde{z}(t,y_t,h)\| \le L\cdot d(\varphi,y_t), \quad h\in[0,\overline{H}).$$

for any $t\in[t_0,t_0+T)$ and $\varphi\in Y$ such that $d(\varphi,y_t)\le\delta$.

Here is the convergence theorem for RFDEs with data set $\mathcal{C}$.

Theorem 5.1. (Convergence) Consider the numerical solution of integration problems given by the recursive process (5.2) and assume that

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

the numerical method is of type (5.3). Let $\mathcal{F}$ be a family of integration problems. If the method has uniform order q, discrete order p and it is stable on $\mathcal{F}$, then it has global order $q' = \min\{q+1, p\}$ on $\mathcal{F}$.

Proof. Consider the numerical solution of an integration problem (t_0, ϕ, T) belonging to the family $\mathcal{F}$. Let $\{\Delta^{(k)}\}$ be a sequence of meshes (with mesh points $t_n^{(k)}$, $n = 0, 1, \ldots, N_{\Delta^{(k)}}$, and step-sizes $h_{n+1}^{(k)}$, $n = 0, 1, \ldots, N_{\Delta^{(k)}} - 1$) such that $h_{\Delta^{(k)}} \to 0$, $k \to \infty$.

Let K_0 be such that

$$h_{\Delta^{(k)}} \leq \overline{H}, \quad k \geq K_0,$$

where $\overline{H}$ is given in Definition 5.4. Hence, for $k \geq K_0$, we have

$$h_{n+1}^{(k)} \leq \overline{H} \leq H_{\max}\big(t_n^{(k)}, y_{t_n^{(k)}}\big), \quad n = 0, 1, \ldots, N_{\Delta^{(k)}} - 1,$$

and so the local errors in (5.6) are defined.

Now, we prove the following relations for the errors.

(i) If

$$\max_{n=1,\ldots,N_{\Delta^{(k)}}} \mathbf{E}_n(\Delta^{(k)}) \to 0, \quad k \to \infty,$$

and

$$\max_{n=1,\ldots,N_{\Delta^{(k)}}-1} \frac{\mathbf{E}_n^0(\Delta^{(k)})}{h_n^{(k)}} \to 0, \quad k \to \infty,$$

then there exists K_1, $K_1 \geq K_0$, such that, for $k \geq K_1$, the sequence $\{\phi_n^{(k)}\}$ is defined, *i.e.*,

$$h_{n+1}^{(k)} \leq H_{\max}\big(t_n^{(k)}, \phi_n^{(k)}\big), \quad n = 0, 1, \ldots, N_{\Delta^{(k)}} - 1,$$

and

$$\mathbf{E}(\Delta^{(k)}) = \mathcal{O}\Big(\max_{n=1,\ldots,N_{\Delta^{(k)}}} \mathbf{E}_n(\Delta^{(k)})\Big) + \mathcal{O}\left(\max_{n=1,\ldots,N_{\Delta^{(k)}}-1} \frac{\mathbf{E}_n^0(\Delta^{(k)})}{h_n^{(k)}}\right), \quad k \to \infty.$$

In order to prove (i), let K_2 be such that $K_2 \geq K_1$ and

$$h_{\Delta^{(k)}} < \frac{1}{L}, \quad k \geq K_2 \tag{5.7}$$

and

$$\frac{\mathrm{e}^{\frac{L}{1-h_\Delta^{(k)} L}T}}{1 - h_\Delta^{(k)} L} \cdot \max_{n=1,\ldots,N_{\Delta^{(k)}}} \mathbf{E}_n(\Delta^{(k)}) + \frac{\mathrm{e}^{\frac{L}{1-h_\Delta^{(k)} L}T} - 1}{L} \cdot \max_{n=1,\ldots,N_{\Delta^{(k)}}-1} \frac{\mathbf{E}_n^0(\Delta^{(k)})}{h_n^{(k)}} \leq \delta. \tag{5.8}$$

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

Now, we fix an index k such that $k \geq K_2$ (it will be dropped in the notation).

We define, for $n = 1, \ldots, N_\Delta$,

$$\overline{\mathbf{E}}_n = \max_{i=1,\ldots,n} \mathbf{E}_i(\Delta),$$
$$\overline{\mathbf{E}}^0_n = \max_{i=1,\ldots,n} \frac{\mathbf{E}^0_i(\Delta)}{h_i}.$$

Moreover, if ϕ_n is defined, $n = 0, \ldots, N_\Delta$, we set

$$\mathbf{d}_n = d(\phi_n, y_{t_n}),$$
$$\mathbf{d}^0_n = |\phi_n(0) - y_{t_n}(0)|,$$
$$\overline{\mathbf{d}}^0_n = \max_{i=0,\ldots,n} \mathbf{d}^0_i,$$

noting that

$$\mathbf{d}_n = \max_{i=0,\ldots,n} \mathbf{d}_i.$$

For a given $N \in \{1, \ldots, N_\Delta - 1\}$, we assume that the sequence $\{\phi_n\}_{n=0}^N$ is defined and satisfies

$$\mathbf{d}_N \leq \frac{\mathrm{e}^{\frac{L}{1-h_\Delta L}(t_{N-1}-t_0)}}{1-h_\Delta L} \cdot \overline{\mathbf{E}}_N + \frac{\mathrm{e}^{\frac{L}{1-h_\Delta L}(t_{N-1}-t_0)} - 1}{L} \cdot \overline{\mathbf{E}}^0_{N-1}, \tag{5.9}$$

setting $\overline{\mathbf{E}}^0_{N-1} = 0$ for $N = 1$. To complete the induction, we prove that ϕ_{N+1} is defined and satisfies

$$\mathbf{d}_{N+1} \leq \frac{\mathrm{e}^{\frac{L}{1-h_\Delta L}(t_N-t_0)}}{1-h_\Delta L} \cdot \overline{\mathbf{E}}_{N+1} + \frac{\mathrm{e}^{\frac{L}{1-h_\Delta L}(t_N-t_0)} - 1}{L} \cdot \overline{\mathbf{E}}^0_N. \tag{5.10}$$

Since (5.8) and (5.9) imply $\mathbf{d}_N = d(\phi_n, y_{t_n}) \leq \delta$, we have

$$h_{N+1} \leq h_\Delta < \overline{H} \leq H_{\max}(t_N, \phi_N),$$

recalling Definition 5.4. Hence, ϕ_{N+1} is defined. Moreover, we have

$$\|\widetilde{z}(t_n, \phi_n, h_{n+1}) - \widetilde{z}(t_N, y_{t_N}, h_{N+1})\| \leq L\mathbf{e}_N. \tag{5.11}$$

Let $n = 0, 1, \ldots, N$. By (5.11), we obtain

$$\begin{aligned}
\mathbf{d}^0_{n+1} &= |\phi_{n+1}(0) - y_{t_{n+1}}(0)| \\
&\leq \big|v\big(\phi_n, \widetilde{z}^*(t_n, \phi_n, h_{n+1})\big)(h_{n+1}) \\
&\qquad - v\big(y_{t_n}, \widetilde{z}^*(t_n, y_{t_n}, h_{n+1})\big)(h_{n+1})\big| \\
&\quad + \big|v\big(y_{t_n}, \widetilde{z}^*(t_n, y_{t_n}, h_{n+1})\big)(h_{n+1}) \\
&\qquad - v\big(y_{t_n}, z^*(t_n, y_{t_n}, h_{n+1})\big)(h_{n+1})\big| \\
&\leq \mathbf{d}^0_n + h_{n+1} L \mathbf{d}_n + \mathbf{E}^0_{n+1}(\Delta).
\end{aligned}$$

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

Moreover, for $\theta \le -h_{n+1}$, we have

$$\begin{aligned}|\phi_{n+1}(\theta) - y_{t_{n+1}}(\theta)| &= \big|v\big(\phi_n, \widetilde{z}^*(t_n, \phi_n, h_n)\big)(h_{n+1}+\theta) \\ &\qquad - v\big(y_{t_n}, z^*(t_n, y_{t_n}, h_{n+1})\big)(h_{n+1}+\theta)\big| \\ &\le |\phi_n(h_{n+1}+\theta) - y_{t_n}(h_{n+1}+\theta)| \\ &\le \mathbf{d}_n\end{aligned}$$

and, for $\theta \in [-h_{n+1}, 0]$, again by (5.11), we have

$$\begin{aligned}|\phi_{n+1}(\theta) - y_{t_{n+1}}(\theta)| &= \big|v\big(\phi_n, \widetilde{z}^*(t_n, \phi_n, h_{n+1})\big)(h_{n+1}+\theta) \\ &\qquad - v\big(y_{t_n}, z^*(t_n, y_{t_n}, h_{n+1})\big)(h_{n+1}+\theta)\big| \\ &\le \big|v\big(\phi_n, \widetilde{z}^*(t_n, \phi_n, h_{n+1})\big)(h_{n+1}+\theta) \\ &\qquad - v\big(y_{t_n}, \widetilde{z}^*(t_n, y_{t_n}, h_{n+1})\big)(h_{n+1}+\theta)\big| \\ &\quad + \big|v\big(y_{t_n}, \widetilde{z}^*(t_n, y_{t_n}, h_{n+1})\big)(h_{n+1}+\theta) \\ &\qquad - v\big(y_{t_n}, z^*(t_n, y_{t_n}, h_{n+1})\big)(h_{n+1}+\theta)\big| \\ &\le \mathbf{d}_n^0 + h_{n+1}L\mathbf{d}_n + \mathbf{E}_{n+1}(\Delta).\end{aligned}$$

Thus, we have

$$\mathbf{d}_{n+1}^0 \le \mathbf{d}_n^0 + h_{n+1}L\mathbf{d}_n + \mathbf{E}_{n+1}^0(\Delta), \quad n = 0, 1, \dots, N, \tag{5.12}$$

and

$$\mathbf{d}_{n+1} \le \max\big\{\mathbf{d}_n, \mathbf{d}_n^0 + h_{n+1}L\mathbf{d}_n + \mathbf{E}_{n+1}(\Delta)\big\}, \quad n = 0, 1, \dots, N. \tag{5.13}$$

Now, for $n = 0, 1, \dots, N$, we have

$$\overline{\mathbf{d}}_{n+1}^0 \le \overline{\mathbf{d}}_n^0 + h_{n+1}L\mathbf{d}_n + \mathbf{E}_{n+1}^0(\Delta) \tag{5.14}$$

since $\overline{\mathbf{d}}_{n+1}^0 = \max\big\{\mathbf{d}_{n+1}^0, \overline{\mathbf{d}}_n^0\big\}$ and

$$\mathbf{d}_{n+1}^0 \le \overline{\mathbf{d}}_n^0 + h_{n+1}L\mathbf{d}_n + \mathbf{E}_{n+1}^0(\Delta)$$

by (5.12). Moreover, for $i = 0, 1, \dots, N$, (5.13) yields

$$\mathbf{d}_{i+1} \le \mathbf{d}_{i-k}^0 + h_{i+1-k}L\mathbf{d}_{i-k} + \mathbf{E}_{i+1-k}(\Delta)$$

for some $k = 0, 1, 2, \dots$. Hence,

$$\mathbf{d}_{i+1} \le \frac{1}{1 - h_\Delta L}\big(\overline{\mathbf{d}}_i^0 + \overline{\mathbf{E}}_{i+1}\big), \quad i = 0, 1, \dots, N. \tag{5.15}$$

By inserting (5.15) with $i + 1 = n$ in (5.14), we obtain

$$\begin{aligned}\overline{\mathbf{d}}_{n+1}^0 \le &\left(1 + h_{n+1}\frac{L}{1 - h_\Delta L}\right)\overline{\mathbf{d}}_n^0 \\ &+ h_{n+1}\left(\frac{L}{1 - h_\Delta M}\overline{\mathbf{E}}_{n+1} + \overline{\mathbf{E}}_{n+1}^0\right), \quad n = 0, 1, \dots, N.\end{aligned} \tag{5.16}$$

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

The recursion (5.16) yields

$$\overline{\mathbf{d}}_N^0 \le \Big(\mathrm{e}^{\frac{L}{1-h_\Delta L}(t_N-t_0)}-1\Big)\overline{\mathbf{E}}_N + \frac{\mathrm{e}^{\frac{L}{1-h_\Delta L}(t_N-t_0)}-1}{\frac{L}{1-h_\Delta L}}\cdot\overline{\mathbf{E}}_N^0.$$

Thus, by (5.15) with $i+1=N+1$ we obtain (5.10).

Since ϕ_N is defined and (5.9) holds with $N=1$, we obtain

$$\mathbf{E}(\Delta)=\mathbf{d}_{N_\Delta} \le \frac{\mathrm{e}^{\frac{L}{1-h_\Delta L}(T-h_{N_\Delta})}}{1-h_\Delta L}\cdot\overline{\mathbf{E}}_{N_\Delta} + \frac{\mathrm{e}^{\frac{L}{1-h_\Delta L}(T-h_{N_\Delta})}-1}{L}\cdot\overline{\mathbf{E}}_{N_\Delta-1}^0,$$

and then (i) is proved.

Now, the theorem follows in a straightforward way from (i). □

RFDEs with data set $\mathcal{LC}$

For RFDEs (1.1) with data set $\mathcal{LC}$, we introduce the concept of order of the method.

Definition 5.5. Let q be a positive integer and let $\mathcal{F}$ be a family of integration problems (t_0,ϕ,T) such that $y(t_0,\phi)|_{[t_0,t_0+T]}$ is piecewise smooth. The method has *order* q on $\mathcal{F}$ if, for any integration problem $(t_0,\phi,T)\in\mathcal{F}$, there exist constants $H>0$ and $C>0$ such that

$$\|e(t,y_t,h)\| \le Ch^q$$

for any $t\in[t_0,t_0+T)$ and $h\in[0,H_{\max}(t,y_t))$ such that $h\le T-t$, the interval $(t,t+h)$ does not contain breaking points of $y(t_0,\phi)$ and $h<H$.

As for RFDEs with data set $\mathcal{C}$, the order is linked to the global order by the concept of stability.

Definition 5.6. Let $\mathcal{F}$ be a family of integration problems (t_0,ϕ,T). The method is *stable* on $\mathcal{F}$ if, for any integration problem (t_0,ϕ,T), there exist $\delta>0$, $\overline{H}>0$, $L\ge 0$, $M\ge 0$ and $P\in[0,1)$ such that

$$\overline{H}\le H_{\max}(t,y_t),\quad t\in[t_0,t_0+T),$$

and

$$\|\widetilde{z}(t,\varphi,h)-\widetilde{z}(t,y_t,h)\| \le L\cdot d(\varphi,y_t)+M\cdot d\big(\varphi'_{-\tau(t)},y'_{t-\tau(t)}\big) + P\cdot d(\varphi',y'_t),\quad h\in[0,\overline{H}),$$

for any $t\in[t_0,t_0+T)$ and $\varphi\in Y$ such that $d(\varphi,y_t)\le\delta$. Here, τ is a function $[t_0,t_0+T]\to[0,+\infty)$ for which there exist $\xi_0,\xi_1,\dots,\xi_{K-1},\xi_K$ such that

$$t_0=\xi_0<\xi_1<\cdots<\xi_{K-1}<\xi_K=t_0+T$$

and

$$t-\tau(t)\le\xi_k,\quad t\in[\xi_k,\xi_{k+1}]\ \text{ and }\ k=0,1,\dots,K-1.$$

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

Here is the convergence theorem for RFDEs with data set $\mathcal{LC}$.

Theorem 5.2. (Convergence) Consider the numerical solution of integration problems given by the recursive process (5.2) and assume that the numerical method is of type (5.3). Let $\mathcal{F}$ be a family of integration problems. If the method has order q and it is stable on $\mathcal{F}$, then it has global order q on $\mathcal{F}$.

Proof. Consider the numerical solution of an integration problem (t_0, ϕ, T) belonging to the the family $\mathcal{F}$. Let $\{\Delta^{(k)}\}$ be a sequence of meshes such that $h_{\Delta^{(k)}} \to 0$, $k \to \infty$.

Let K_0 be such that

$$h_{\Delta^{(k)}} \leq \overline{H}, \quad k \geq K_0,$$

where $\overline{H}$ is given in Definition 5.6. Hence, for $k \geq K_0$, the local errors in (5.6) are defined. As in the proof of Theorem 5.1, we now address the errors.

(i) If

$$\max_{n=1,\ldots,N_{\Delta^{(k)}}} \mathbf{e}_n(\Delta^{(k)}) \to 0, \quad k \to \infty,$$

then there exists K_1, $K_1 \geq K_0$, such that, for $k \geq K_1$, the sequence $\{\phi_n^{(k)}\}$ is defined and

$$\mathbf{E}(\Delta^{(k)}) = \mathcal{O}\Big(\max_{n=1,\ldots,N_{\Delta^{(k)}}} \mathbf{e}_n(\Delta^{(k)})\Big).$$

In order to prove (i), let K_2 be such that $K_2 \geq K_1$ and (5.7) hold, *i.e.*,

$$h_{\Delta^{(k)}} < \frac{1}{L}, \quad k \geq K_2,$$

and let

$$\frac{\mathrm{e}^{\frac{L}{1-h_\Delta^{(k)}L}T}}{1-h_\Delta^{(k)}L} \cdot \max_{n=1,\ldots,N_{\Delta^{(k)}}} \mathbf{E}_n(\Delta^{(k)}) \leq \delta. \tag{5.17}$$

Now, we fix an index k such that $k \geq K_2$ (it will be dropped in the notation).

We define, for $n = 1, \ldots, N_\Delta$,

$$\overline{\mathbf{E}}_n = \max_{i=1,\ldots,n} \mathbf{E}_i(\Delta),$$
$$\overline{\mathbf{e}}_n = \max_{i=1,\ldots,n} \mathbf{e}_i(\Delta).$$

Moreover, if ϕ_n is defined, $n = 0, \ldots, N_\Delta$, we set

$$\mathbf{d}_n = d(\phi_n, y_{t_n}),$$
$$\mathbf{d}'_n = d(\phi'_n, y'_{t_n}),$$

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

noting that

$$\mathbf{d}_n = \max_{i=0,\dots,n} \mathbf{d}_i \quad \text{and} \quad \mathbf{d}'_n = \max_{i=0,\dots,n} \mathbf{d}'_i.$$

We define the sets of indices

$$\begin{aligned} &I_0 = \{0\}, \\ &I_k = \big\{n \in \{0,\dots,N_\Delta\} \mid t_n \in [\xi_{k-1}, \xi_k]\big\}, \quad k = 1,\dots,K, \end{aligned}$$

and set, for $k = 0,\dots,K$,

$$\widehat{\mathbf{d}}_k = \mathbf{d}_{\max I_k}, \quad \widehat{\mathbf{d}}'_k = \mathbf{d}'_{\max I_k} \quad \text{and} \quad \widehat{\overline{\mathbf{e}}}_k = \overline{\mathbf{e}}_{\max I_k}.$$

For a given $N \in \{1,\dots,N_\Delta - 1\}$, we assume that the sequence $\{\phi_n\}_{n=0}^N$ is defined and satisfies

$$\mathbf{d}_N \le \frac{\mathrm{e}^{LC_1(t_N - t_0)} - 1}{LC_1} \cdot \overline{\mathbf{e}}_N, \tag{5.18}$$

where

$$C_1 = 1 + \frac{M+P}{1-P}\left(1 + \frac{MC_0}{1-P}\right) \tag{5.19}$$

and

$$C_0 = \sum_{i=1}^{k} \left(\frac{M}{1-P}\right)^{k-i}. \tag{5.20}$$

We complete the inductive proof by showing that ϕ_{N+1} is defined and satisfies

$$\mathbf{d}_{N+1} \le \frac{\mathrm{e}^{LC_1(t_{N+1} - t_0)} - 1}{LC_1} \cdot \overline{\mathbf{e}}_{N+1}. \tag{5.21}$$

Since $\mathbf{d}_N = d(\phi_n, y_{t_n}) \le \delta$ holds by (5.17) and (5.18), we obtain that ϕ_{N+1} is defined. Moreover, recalling Definition 5.6, we have

$$\begin{aligned} &\|\widetilde{z}(t_n, \phi_n, h_{n+1}) - \widetilde{z}(t_N, y_{t_N}, h_{N+1})\| \\ &\qquad \le L\mathbf{d}_N + Md\big((\phi'_n)_{-\tau(t_n)}, y'_{t_n - \tau(t_n)}\big) + P\mathbf{d}'_n. \end{aligned} \tag{5.22}$$

Let $n = 0, 1, 2, \dots, N$. We look for a bound for $\mathbf{d}'_{n+1}$.
For $\theta \le -h_{n+1}$, we have

$$\begin{aligned} |\phi'_{n+1}(\theta) - y'_{t_{n+1}}(\theta)| &= \big|v'\big(\phi_n, \widetilde{z}^*(t_n, \phi_n, h_{n+1})\big)(h_{n+1} + \theta) \\ &\qquad - v'\big(y_{t_n}, z^*(t_n, y_{t_n}, h_{n+1})\big)(h_{n+1} + \theta)\big| \\ &= |\phi'_n(h_{n+1} + \theta) - y'_{t_n}(h_{n+1} + \theta)| \\ &\le \mathbf{d}'_n. \end{aligned}$$

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

For $\theta \in [-h_{n+1}, 0]$, by (5.22), we have

$$\begin{aligned}
|\phi'_{n+1}(\theta) - y'_{t_{n+1}}(\theta)| &= \big|v'\big(\phi_n, \widetilde{z}^*(t_n, \phi_n, h_{n+1})\big)(h_{n+1}+\theta) \\
&\qquad - v'\big(y_{t_n}, z^*(t_n, y_{t_n}, h_{n+1})\big)(h_{n+1}+\theta)\big| \\
&\le \big|\widetilde{z}^*(t_n, \phi_n, h_{n+1})(h_{n+1}+\theta) \\
&\qquad - \widetilde{z}^*(t_n, y_{t_n}, h_{n+1})(h_{n+1}+\theta)\big| \\
&\le \big|\widetilde{z}^*(t_n, \phi_n, h_{n+1})(h_{n+1}+\theta) \\
&\qquad - \widetilde{z}^*(t_n, y_{t_n}, h_{n+1})(h_{n+1}+\theta)\big| \\
&\quad + \big|\widetilde{z}^*(t_n, y_{t_n}, h_{n+1})(h_{n+1}+\theta) \\
&\qquad - z^*(t_n, y_{t_n}, h_{n+1})(h_{n+1}+\theta)\big| \\
&\le L\mathbf{d}_n + Md\big((\phi'_n)_{-\tau(t_n)}, (y'_{t_n})_{-\tau(t_n)}\big) \\
&\qquad + P\mathbf{d}'_n + \mathbf{e}_{n+1}(\Delta).
\end{aligned}$$

Now, if $t_n \in [\xi_{k-1}, \xi_k]$, $k = 1, \dots, K$, then $t_n - \tau(t_n) \le \xi_{k-1}$ and so

$$d\big((\phi'_n)_{-\tau(t_n)}, y'_{t_n-\tau(t_n)}\big) \le d(\phi'_m, y'_{t_m}) = \widehat{\mathbf{d}}'_k,$$

where m is some index in the set I_{k-1}. Thus, if $t_n \in [\xi_{k-1}, \xi_k]$, we have

$$\mathbf{d}'_{n+1} \le \max\big\{\mathbf{d}'_n, L\mathbf{d}_n + M\widehat{\mathbf{d}}'_{k-1} + P\mathbf{d}'_n + \mathbf{e}_{n+1}(\Delta)\big\}.$$

As a consequence, we obtain

$$\mathbf{d}'_{n+1} \le L\mathbf{d}_n + M\widehat{\mathbf{d}}'_{k-1} + P\mathbf{d}'_n + \overline{\mathbf{e}}_{n+1}$$

and then

$$\mathbf{d}'_{n+1} \le \frac{L}{1-P}\mathbf{d}_n + \frac{M}{1-P}\widehat{\mathbf{d}}'_{k-1} + \frac{1}{1-P}\overline{\mathbf{e}}_{n+1}. \tag{5.23}$$

Inequality (5.23) yields

$$\widehat{\mathbf{d}}'_k \le \frac{M}{1-P}\widehat{\mathbf{d}}'_{k-1} + \frac{L}{1-P}\widehat{\mathbf{d}}_k + \frac{1}{1-P}\widehat{\overline{\mathbf{e}}}_k. \tag{5.24}$$

and a recursive use of (5.24) gives

$$\widehat{\mathbf{d}}'_k \le \frac{C_0 L}{1-P}\widehat{\mathbf{d}}_k + \frac{C_0}{1-P}\widehat{\overline{\mathbf{e}}}_k,$$

where C_0 is given by (5.20). Then, by (5.23),

$$\mathbf{d}'_{n+1} \le \frac{L}{1-P}\left(1 + \frac{MC_0}{1-P}\right)\mathbf{d}_n + \frac{1}{1-P}\left(1 + \frac{MC_0}{1-P}\right)\overline{\mathbf{e}}_{n+1}. \tag{5.25}$$

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

Now, we look for a bound for $\mathbf{d}_{n+1}$. For $\theta \le -h_{n+1}$, we have

$$\begin{aligned}|\phi_{n+1}(\theta) - y_{t_{n+1}}(\theta)| &= \big|v\big(\phi_n, \widetilde{z}^*(t_n, \phi_n, h_n)\big)(h_{n+1}+\theta) \\ &\qquad - v\big(y_{t_n}, z^*(t_n, y_{t_n}, h_{n+1})\big)(h_{n+1}+\theta)\big| \\ &\le |\phi_n(h_{n+1}+\theta) - y_{t_n}(h_{n+1}+\theta)| \\ &\le \mathbf{d}_n\end{aligned}$$

and, for $\theta \in [-h_{n+1}, 0]$, by (5.22), we have

$$\begin{aligned}|\phi_{n+1}(\theta) - y_{t_{n+1}}(\theta)| &= \big|v\big(\phi_n, \widetilde{z}^*(t_n, \phi_n, h_{n+1})\big)(h_{n+1}+\theta) \\ &\qquad - v\big(y_{t_n}, z^*(t_n, y_{t_n}, h_{n+1})\big)(h_{n+1}+\theta)\big| \\ &\le \big|v\big(\phi_n, \widetilde{z}^*(t_n, \phi_n, h_{n+1})\big)(h_{n+1}+\theta) \\ &\qquad - v\big(y_{t_n}, \widetilde{z}^*(t_n, y_{t_n}, h_{n+1})\big)(h_{n+1}+\theta)\big| \\ &\qquad + \big|v\big(y_{t_n}, \widetilde{z}^*(t_n, y_{t_n}, h_{n+1})\big)(h_{n+1}+\theta) \\ &\qquad\quad - v\big(y_{t_n}, z^*(t_n, y_{t_n}, h_{n+1})\big)(h_{n+1}+\theta)\big| \\ &\le \mathbf{d}_n + h_{n+1}L\mathbf{d}_n \\ &\qquad + h_{n+1}Md\big((\phi'_n)_{-\tau(t_n)}, (y'_{t_n})_{-\tau(t_n)}\big) \\ &\qquad + P\mathbf{d}'_n + \mathbf{E}_{n+1}(\Delta).\end{aligned}$$

Thus,

$$\mathbf{d}_{n+1} \le (1 + h_{n+1}L)\mathbf{d}_n + h_{n+1}(M+P)\mathbf{d}'_n + \mathbf{E}_{n+1}(\Delta). \tag{5.26}$$

By summarizing, we have proved the inequalities (5.25) and (5.26) for $n = 0, \dots, N$. Hence, we obtain

$$\mathbf{d}_{n+1} \le (1 + h_{n+1}LC_1)\mathbf{d}_n + h_{n+1}C_1\overline{\mathbf{e}}_{n+1}, \quad n = 0, \dots, N, \tag{5.27}$$

where C_1 is given by (5.19). The recursion (5.27) yields (5.21). Moreover, since ϕ_N is defined and (5.18) holds with $N = 1$, we obtain

$$\mathbf{E}(\Delta) = \mathbf{d}_{N_\Delta} \le \frac{\mathrm{e}^{LC_1T} - 1}{LC_1} \cdot \overline{\mathbf{e}}_{N_\Delta},$$

and then (i) and the theorem follow. □

6. Functional continuous Runge–Kutta methods

The development of methods based on suitable modifications of RK methods for the numerical integration of IPs for RFDEs began in the late 1960s/early 1970s, and was essentially due to the pioneering papers of Feldstein (1964), Tavernini (1971) and Cryer and Tavernini (1972). In particular, Cryer and Tavernini (1972) considered the following generalization of the Euler and

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

Heun method, and proved that they are convergent of order one and two respectively. Specifically, we have the following methods.

Euler method:

$$\eta(t_n + \theta h_{n+1}) = \eta(t_n) + h_{n+1}\theta F\big(t_n, \eta_{t_n}\big), \quad \theta \in [0,1];$$

Heun method:

$$\eta(t_n + \theta h_{n+1}) = \eta(t_n) + h_{n+1}\left[\left(\theta - \frac{1}{2}\theta^2\right)F\big(t_n, \eta_{t_n}\big) + \frac{1}{2}\theta^2 F\big(t_{n+1}, Y_{t_{n+1}}\big)\right], \quad \text{for } \theta \in [0,1],$$

where

$$Y\big(t_n + \theta h\big) = \eta(t_n) + h_{n+1}\theta F\big(t_n, \eta(t_n)\big), \quad \theta \in [0,1],$$
$$Y(t) = \eta(t), \quad t \in (-\infty, t_n].$$

Tavernini (1971) also proposed higher-order explicit methods obtained by a predictor–corrector implementation of polynomial collocation. First, he introduced a sequence $\{\eta^{(s)}\}_{s=2,3,\ldots,\bar{s}}$ of implicit methods given by

$$\eta^{(s)}(t_n + \theta h_{n+1}) = \eta(t_n) + h_{n+1}\sum_{i=1}^{s} b_i^{(s)}(\theta)F\Big(t_n + c_i^{(s)}h_{n+1}, \eta^{(s)}_{t_n + c_i^{(s)}h_{n+1}}\Big), \quad \text{for } \theta \in [0,1],$$

where, for $i = 1, \ldots, s$, $c_i^{(s)} \in [0,1]$ are distinct points, and $b_i^{(s)}(\cdot) : [0,1] \to \mathbb{R}$, $i = 1, \ldots, s$, are polynomials of degree s defined by the collocation conditions

$$(\eta^{(s)})'\big(t_n + c_i^{(s)}h_{n+1}\big) = F\Big(t_n + c_i^{(s)}h_{n+1}, \eta^{(s)}_{t_n + c_i^{(s)}h_{n+1}}\Big), \quad i = 1, \ldots, s.$$

In particular, he proposed the three equispaced nodes collocation methods, given in Table 6.1.

Then Tavernini considered the explicit method given by the recurrence relation

$$\eta^{(1)}(t_n + \theta h_{n+1}) = \eta(t_n) + h_{n+1}\theta F\big(t_n, \eta_{t_n}\big), \quad \theta \in [0,1],$$
$$\eta^{(1)}(t) = \eta(t), \quad t \le t_n,$$
$$\eta^{(s)}(t_n + \theta h_{n+1}) = \eta(t_n) + h_{n+1}\sum_{i=1}^{s} b_i^{(s)}(\theta)F\Big(t_n + c_i^{(s)}h_{n+1}, \eta^{(s-1)}_{t_n + c_i^{(s)}h_{n+1}}\Big), \quad \text{for } \theta \in [0,1],$$
$$\eta^{(s)}(t) = \eta(t), \quad t \le t_n,$$

for $s = 2, \ldots, \bar{s}$, and finally,

$$\eta(t_n + \theta h_{n+1}) = \eta^{(\bar{s})}(t_n + \theta h_{n+1}), \quad \theta \in [0,1].$$

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

Table 6.1. Abscissae and continuous weights of the collocation methods proposed by Tavernini (1971).

s	$c^{(s)}$	$b^{(s)}$
2	$(0, 1)$	$b^{(2)}(\theta) = \left(\theta - \frac{1}{2}\theta^2,\ \frac{1}{2}\theta^2\right)$
3	$\left(0, \frac{1}{2}, 1\right)$	$b^{(3)}(\theta) = \left(\theta - \frac{3}{2}\theta^2 + \frac{2}{3}\theta^3,\ 2\theta^2 - \frac{4}{3}\theta^3,\ -\frac{\theta^2}{2} + \frac{2}{3}\theta^3\right)$
4	$\left(0, \frac{1}{3}, \frac{2}{3}, 1\right)$	$b^{(4)}(\theta) = \Big(\theta - \frac{11}{4}\theta^2 + 3\theta^3 - \frac{9}{8}\theta^4,\ \frac{9}{2}\theta^2 - \frac{15}{2}\theta^3 + \frac{27}{8}\theta^4,$ $-\frac{9}{4}\theta^2 + 6\theta^3 - \frac{27}{8}\theta^4,\ \frac{1}{2}\theta^2 - \frac{3}{2}\theta^3 + \frac{9}{8}\theta^4\Big)$

For $\overline{s} = 1$ and $\overline{s} = 2$ the previous Euler and Heun methods, respectively, are obtained. In this way, explicit methods of arbitrary global order r for functional equations were obtained at the cost of $1 + \frac{r(r-1)}{2}$ evaluations of the functional F. Indeed, Tavernini also found a particular explicit method of global order 4 by using only 6 evaluations of F, instead of 7 as required by the approach described above. Surprisingly, after Tavernini this approach was not further investigated. Only recently, a general class of RK methods for RFDEs, including all implicit and explicit methods considered by Tavernini as particular instances, was proposed and investigated by Maset, Torelli and Vermiglio (2005). These methods, denoted by $(A(\theta), b(\theta), c)$ and called *functional continuous Runge–Kutta* methods (FCRK), are considered in the following section.

6.1. The general form of FCRK methods

Definition 6.1. Let ν be a positive integer. A ν-stage functional continuous Runge–Kutta method is a triple $(A(\theta), b(\theta), c)$, where $A(\theta)$ is an $\mathbb{R}^{\nu\times\nu}$-valued polynomial function such that $A(0) = 0$, $b(\theta)$ is an $\mathbb{R}^{\nu}$-valued polynomial function such that $b(0) = 0$, and $c \in \mathbb{R}^{\nu}$ with $0 \le c_i \le 1$, $i = 1, \dots, \nu$.

The FCRK method $(A(\theta), b(\theta), c)$ provides the approximation

$$\widetilde{V}(\sigma, \varphi, h) = \widetilde{v}_h, \tag{6.1}$$

where the function $\widetilde{v} : (-\infty, h] \to \mathbb{R}^d$ is given by

$$\widetilde{v}(\theta h) = \varphi(0) + h \sum_{i=1}^{\nu} b_i(\theta) K_i, \quad \theta \in [0, 1],$$
$$\widetilde{v}(t) = \varphi(t), \quad t \in (-\infty, 0],$$

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

the *derivatives* $K_i \in \mathbb{R}^d$, $i = 1, \dots, \nu$, are given by

$$K_i = F(\sigma + c_i h, Y^i_{c_i h})$$

and the *stage functions* $Y^i : (-\infty, h] \to \mathbb{R}^d$, $i = 1, \dots, \nu$, are given by

$$\begin{aligned} Y^i(\theta h) &= \varphi(0) + h \sum_{j=1}^{\nu} a_{ij}(\theta) K_j, \quad \theta \in [0,1], \\ Y^i(t) &= \varphi(t), \quad t \in (-\infty, 0]. \end{aligned}$$

Note that the conditions $A(0) = 0$ and $b(0) = 0$ guarantee the continuity of the functions $\widetilde{v}$ and Y^i, $i = 1, \dots, \nu$.

The FCRK method $(A(\theta), b(\theta), c)$ will be denoted by the tableau

$$\begin{array}{c|c} c & A(\theta) \\ \hline & b(\theta) \end{array}.$$

In particular, the Euler and Heun methods are given by

$$\begin{array}{c|c} 0 & 0 \\ \hline & \theta \end{array} \tag{6.2}$$

and

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & \theta & 0 \\ \hline & \theta - \frac{1}{2}\theta^2 & \frac{1}{2}\theta^2 \end{array}, \tag{6.3}$$

respectively.

Moreover, we partition the set $I = \{1, \dots, \nu\}$ of indices in the subsets

$$I^+ = \{c_i > 0 \mid i = 1, \dots, \nu\}$$

and

$$I^0 = \{c_i = 0 \mid i = 1, \dots, \nu\}.$$

Note that, if $i \in I^0$, then $K_i = F(\sigma, \varphi)$.

To show that the FCRK method $(A(\theta), b(\theta), c)$ provides an approximation of the solution map in the form (5.3), we introduce the following.

- The *prolongation linear operators*

$$\begin{aligned} \pi &: (\mathbb{R}^d)^\nu \to C_h = C([0,h], \mathbb{R}^d), \\ \Pi_i &: (\mathbb{R}^d)^\nu \to C_h, \quad i \in I, \\ \Pi &: (\mathbb{R}^d)^\nu \to (C_h)^\nu, \end{aligned} \tag{6.4}$$

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

given by

$$(\pi U)(t) = \sum_{i=1}^{\nu} b_i'\left(\frac{t}{h}\right) U_i,$$

$$(\Pi_i U)(t) = \sum_{j=1}^{\nu} a_{ij}'\left(\frac{t}{h}\right) U_j,$$

$$t \in [0,h] \quad \text{and} \quad U = (U_1, \dots, U_\nu) \in (\mathbb{R}^d)^\nu,$$

where the derivatives $b_i'(\theta)$ and $a_{ij}'(\theta)$ of the polynomial functions $b_i(\theta)$ and $a_{ij}(\theta)$ appear, and

$$\Pi U = (\Pi_1 U, \dots, \Pi_\nu U), \quad U \in (\mathbb{R}^d)^\nu.$$

- The *restriction linear operator*

$$R : (C_h)^\nu \to (\mathbb{R}^d)^\nu$$

defined by

$$RZ = \big(Z_1(c_1 h), \dots, Z_\nu(c_\nu h)\big), \quad Z \in (C_h)^\nu.$$

- The map

$$\mathbf{Q}_h^\diamond(\sigma, \varphi) : (C_h)^\nu \to (L_h^\diamond)^\nu$$

defined by

$$\mathbf{Q}_h^\diamond(\sigma, \varphi) Z = \big(Q_h^\diamond(\sigma, \varphi)(Z_1), \dots, Q_h^\diamond(\sigma, \varphi)(Z_\nu)\big), \quad Z \in (C_h)^\nu,$$

where $L_h^\diamond$ is the space of the measurable and bounded functions $[0,h] \to \mathbb{R}^d$ and $Q_h^\diamond(\sigma, \varphi)$ is given by (4.4).

Proposition 6.1. The FCRK method $(A(\theta), b(\theta), c)$ yields an approximation of the type (5.3), where

$$\widetilde{z}^*(\sigma, \varphi, h) = \pi K,$$

and $K = (K_1, \dots, K_\nu) \in (\mathbb{R}^d)^\nu$ is a fixed point of the map

$$R\mathbf{Q}_h^\diamond(\sigma, \varphi)\Pi : (\mathbb{R}^d)^\nu \to (\mathbb{R}^d)^\nu.$$

Proof. For the function $\widetilde{v}$ in (6.1), we have

$$\begin{aligned} \widetilde{v}(t) &= \varphi(0) + h \sum_{i=1}^{\nu} b_i\left(\frac{t}{h}\right) K_i \\ &= \varphi(0) + \int_0^t \sum_{i=1}^{\nu} b_i'\left(\frac{s}{h}\right) K_i \, \mathrm{d}s \\ &= v(\varphi, \pi K), \quad t \in [0,h]. \end{aligned}$$

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

Analogously, for the stage functions we obtain

$$\begin{aligned} Y^i(t) &= \varphi(0) + \int_0^t \sum_{j=1}^{\nu} a'_{ij}\left(\frac{s}{h}\right) K_j \, \mathrm{d}s \\ &= v(\varphi, \Pi_i K)(t), \quad t \in [0,h], \end{aligned}$$

and then, for $i = 1, \dots, \nu$,

$$\begin{aligned} K_i &= F\big(\sigma + c_i h, v(\varphi, \Pi_i K)_{c_i h}\big) \\ &= [Q_h^{\diamond}(\sigma,\varphi)(\Pi_i K)](c_i h) \\ &= R\big(Q_h^{\diamond}(\sigma,\varphi)(\Pi_1 K), \dots, Q_h^{\diamond}(\sigma,\varphi)(\Pi_\nu K)\big)_i \\ &= [R\mathbf{Q}_h^{\diamond}(\sigma,\varphi)\Pi](K)_i. \end{aligned}$$

□

An FCRK method $(A(\theta), b(\theta), c)$ is called *explicit* if $c_1 = 0$ and, for $i = 2, \dots, \nu$, $a_{ij}(\theta) = 0$ for $j = i, \dots, \nu$. The method is called *implicit* if it is not explicit.

If the method is explicit, the derivatives K_i, $i = 1, \dots, \nu$, can be explicitly obtained in a recursive way: we compute

$$K_1 = F(\sigma, \varphi)$$

and then, successively for $i = 2, \dots, \nu$,

$$K_i = F(\sigma + c_i h, Y^i_{c_i h}),$$

where

$$\begin{aligned} Y^i(\theta h) &= \varphi(0) + h \sum_{j=1}^{i-1} a_{ij}(\theta) K_j, \quad \theta \in [0,1], \\ Y^i(t) &= \varphi(t), \quad t \in (-\infty, 0]. \end{aligned}$$

If the method is implicit, the derivative vector $K = (K_1, \dots, K_\nu)$ is obtained as a solution of the fixed point equation

$$K = [R\mathbf{Q}_h^{\diamond}(\sigma,\varphi)\Pi](K)$$

on $(\mathbb{R}^d)^\nu$.

6.2. Well-posedness of FCRK methods

In this section, we study fixed points of the map $R\mathbf{Q}_h^{\diamond}(\sigma,\varphi)\Pi$. In order to avoid unwieldy notation, we omit the ordered pair (σ, φ).

First of all, observe that we cannot hope to have a unique fixed point of the map $R\mathbf{Q}_h^{\diamond}\Pi$, even for small h. In fact, consider the scalar IP for ODEs

$$\begin{aligned} y'(t) &= \lambda y(t)^2, \quad t \geq 0, \\ y(0) &= y_0 \in \mathbb{R}. \end{aligned}$$

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

The derivative K for the implicit Euler method is an approximation of $y'(h)$ and satisfies the equation

$$K = \lambda(y_0 + hK)^2.$$

Such an equation has the two solutions,

$$K_+ = \frac{2\lambda y_0^2}{1 - 2h\lambda y_0 + \sqrt{1 - 4h\lambda y_0}} \to \lambda y_0^2, \quad h \to 0,$$

and

$$K_- = \frac{2\lambda y_0^2}{1 - 2h\lambda y_0 - \sqrt{1 - 4h\lambda y_0}} \to \infty, \quad h \to 0.$$

The solution K_+ is an approximation of $y'(h)$, whereas K_- is a spurious solution which diverges as $h \to 0$.

In order to study the fixed points of the map $R\mathbf{Q}_h^\diamond\Pi$, we introduce the following.

- A norm $|\cdot|$ on $\mathbb{R}^d$.
- A norm on $(\mathbb{R}^d)^\nu$,
$$\|U\|_\infty = \max_{i=1,\dots,\nu} |U_i|, \quad U \in (\mathbb{R}^d)^\nu.$$
- A norm on the spaces $L_h^\diamond$ and $C_h \subseteq L_h^\diamond$,
$$\|z\| = \sup_{t\in[0,T]} |z(t)|, \quad z \in L_h^\diamond.$$
- For any $\rho > 0$, the closed ball in $(\mathbb{R}^d)^\nu$,
$$B_\rho = \big\{U \in (\mathbb{R}^d)^\nu \mid \|U\|_\infty \le \rho\big\},$$
the closed ball in C_h,
$$C_{h,\rho} = \{z \in C_h \mid \|z\| \le \rho\},$$
and the Lipschitz constant,
$$k_{h,\rho}^\diamond = \sup_{\substack{z_1,z_2\in C_{h,\rho}\\ z_1\neq z_2}} \frac{\|Q_h^\diamond(z_1) - Q_h^\diamond(z_2)\|}{\|z_1 - z_2\|}$$
of the map $Q_h^\diamond$ on $C_{h,\rho}$.

The next theorem concerns the fixed points of $R\mathbf{Q}_h^\diamond\Pi$ (see Maset (2009)).

Theorem 6.1. (Well-posedness) If:

(A) There exist a function $a : (0, h_0] \times [\rho_0, +\infty) \to [0, +\infty)$, where $h_0 > 0$ and $\rho_0 > 0$, and a constant $b \in [0, +\infty)$ such that:

(1) $k_{h,\rho}^\diamond \le a(h, \rho) + b$, $(h, \rho) \in (0, h_0] \times [\rho_0, +\infty)$,

(2) $\lim_{h\downarrow 0} a(h, \rho) = 0$, $\rho \in [\rho_0, +\infty)$,

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

(3) $b\Lambda < 1$, where

$$\Lambda = \max_{i\in I^+} \max_{\theta\in[0,c_i]} \sum_{j=1}^{\nu} |a'_{ij}(\theta)|, \tag{6.5}$$

then there exist $\overline{\rho}_0 > 0$ and $\overline{h} > 0$ such that, for $0 < h < \overline{h}$, the map $R\mathbf{Q}_h^{\diamond}\Pi$ has a unique fixed point in $B_{\overline{\rho}_0}$, which is contained in the interior of $B_{\overline{\rho}_0}$. Moreover, any other fixed point of $R\mathbf{Q}_h^{\diamond}\Pi$ diverges as $h \to 0$. Finally, for any $\rho \geq \overline{\rho}_0$, there exists $\widehat{h} = \widehat{h}(\rho) > 0$ such that

$$k^{\diamond}_{h,\Lambda\rho}\Lambda < 1 \tag{6.6}$$

for $0 < h < \widehat{h}$.

The hypothesis (A) in Theorem 6.1 holds for any $(\sigma,\varphi) \in \mathbb{R}\times Y$ for DDEs, SDDDEs, DIDEs, SDDIDEs, NDIDEs and NSDDIDEs, whenever the equations satisfy the conditions stated in the existence theorems of Section 4. For NDDEs, we have to require, in addition, that all the delays τ_j^*, $j = 1,\dots,s^*$, are non-vanishing. In all these cases, (1) holds with $b = 0$ and then (3) is satisfied with no restrictions on Λ.

For an NDDE with some strongly vanishing delay τ_j^*, the hypothesis (A) holds for any $(\sigma,\varphi) \in \mathbb{R}\times Y$ under the restriction $\Lambda \leq 1$. In fact, (1) holds for some $b \in [0,1)$ and then (3) is fulfilled if $\Lambda \leq 1$.

Henceforth, we assume that the hypothesis (A) of Theorem 6.1 is satisfied for any $(\sigma,\varphi) \in \mathbb{R}\times Y$. Under this assumption, the FCRK method $(A(\theta), b(\theta), c)$ provides the approximation

$$\widetilde{V}(\sigma,\varphi,h) = v\big(\varphi, \widetilde{z}^*(\sigma,\varphi,h)\big), \quad (\sigma,\varphi) \in \mathbb{R}\times Y \ \text{ and } \ h \in \big[0, H_{\max}(\sigma,\varphi)\big),$$

where $\widetilde{z}^*(\sigma,\varphi,h) = pK$, $K = K(\sigma,\varphi,h)$ is the unique fixed point of the map $R\mathbf{Q}_h^{\diamond}(\sigma,\varphi)\Pi$ on $B_{\overline{\rho}_0}$, and $\overline{\rho}_0 = \overline{\rho}_0(\sigma,\varphi)$ and $H_{\max}(\sigma,\varphi) = \overline{h} = \overline{h}(\sigma,\varphi)$ are defined in Theorem 6.1.

7. Order conditions for FCRK methods

In this section, which is also based on the results by Maset, Torelli and Vermiglio (2005), we give an expansion of the local error functions defined in Section 5 in terms of the step-size h, and then we develop conditions for obtaining a given uniform or discrete order for RFDEs with data set $\mathcal{C}$ and a given order for RFDEs with data set $\mathcal{LC}$.

7.1. Study of the local error functions

We analyse the local error functions $e(\sigma,\varphi,h)$ and $E(\sigma,\varphi,h)$ given by (5.4) and (5.5). Here, we have $(\sigma,\varphi) \in \mathbb{R}\times Y$ and

$$h \in \big[0, \min\{T_{\max}(\sigma,\varphi), H_{\max}(\sigma,\varphi)\}\big),$$

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

and so both the functions $z^*(\sigma,\varphi,h)$ and $\widetilde{z}^*(\sigma,\varphi,h) = pK(\sigma,\varphi,h)$ are defined. In this study, we assume that $z^*(\sigma,\varphi,h)$ is at least continuous. In order to avoid cumbersome notation, we omit the dependence on (σ,φ,h).

Let

$$Z^* = (z^*,\dots,z^*) \in (C_h)^\nu.$$

It is clear that Z^* is a fixed point of the map $\mathbf{Q}_h^\diamond$.

The local error functions e and E can be written as

$$e = \gamma + \pi\Delta \tag{7.1}$$

and

$$E = \int \gamma + \int \pi\Delta, \tag{7.2}$$

where

$$\gamma = \pi R Z^* - z^* \in C_h$$

and

$$\Delta = K - RZ^* \in (\mathbb{R}^d)^\nu.$$

As for the error γ, we can give the following bound.

Proposition 7.1. If z^* is of class C^m, where m is non-negative integer, then

$$\max_{\theta\in[0,1]}\left|\gamma(\theta h) - \sum_{k=0}^{m-1}\gamma_k(\theta)\frac{h^k}{k!}(z^*)^{(k)}(0)\right| \le \frac{h^m}{m!}\max_{\theta\in[0,1]}\left(\sum_{i=1}^{\nu}|b_i'(\theta)|c_i^m + \theta^m\right)\max_{t\in[0,h]}|z^{(m)}(t)|,$$

where, for $k = 0,1,\dots,m-1$,

$$\gamma_k(\theta) = \sum_{i=1}^{\nu} b_i'(\theta)c_i^k - \theta^k, \quad \theta\in[0,1].$$

Proof. We have, for $\theta\in[0,1]$,

$$\begin{aligned}\gamma(\theta h) &= \sum_{i=1}^{\nu} b_i'(\theta)z^*(c_i h) - z^*(\theta h)\\ &= \sum_{i=1}^{\nu} b_i'(\theta)\left(\sum_{k=0}^{m-1}\frac{c_i^k h^k}{k!}(z^*)^{(k)}(0)\right.\\ &\qquad\left. + \frac{1}{(m-1)!}\int_0^1 (1-s)^{m-1}(z^*)^{(m)}(sc_i h)c_i^m h^m\,\mathrm{d}s\right)\end{aligned}$$

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

$$
\begin{aligned}
&- \Bigg(\sum_{k=0}^{m-1} \frac{\theta^k h^k}{k!} (z^*)^{(k)}(0) \\
&\qquad + \frac{1}{(m-1)!} \int_0^1 (1-s)^{m-1} (z^*)^{(m)}(s\theta h) \theta^m h^m \, \mathrm{d}s \Bigg) \\
&= \sum_{k=0}^{m-1} \frac{h^k}{k!} \Bigg(\sum_{i=1}^{\nu} b_i'(\theta) c_i^k - \theta^k \Bigg) (z^*)^{(k)}(0) + \frac{h^m}{(m-1)!} \\
&\quad \cdot \int_0^1 (1-s)^{m-1} \Bigg(\sum_{i=1}^{\nu} b_i'(\theta) c_i^m (z^*)^{(m)}(s c_i h) - \theta^m (z^*)^{(m)}(s\theta h) \Bigg) \, \mathrm{d}s,
\end{aligned}
$$

whenever z^* is of class C^{m+1}. □

As for the other error Δ in (7.1), we have

$$\Delta = R\big(\mathbf{Q}_h^\diamond(Z^* + \Pi\Delta + \Gamma) - \mathbf{Q}_h^\diamond(z^*)\big), \tag{7.3}$$

where

$$\Gamma = \Pi R Z^* - Z^* \in (C_h)^\nu.$$

Note that, for $i \in I^0$, $\Delta_i = 0$.

In the next proposition, we establish that

$$\|\Delta\|_\infty = \mathcal{O}(\|\Gamma\|_\infty^+), \quad \|\Gamma\|_\infty^+ \to 0, \tag{7.4}$$

where

$$\|\Gamma\|_\infty^+ = \max_{i \in I^+} \|\Gamma_i|_{[0, c_i h]}\|.$$

Proposition 7.2. If $h < \widehat{h}(\Lambda\overline{\rho}_1)$, where $\widehat{h}(\cdot)$ is defined in Theorem 6.1, Λ is defined in (6.5), $\overline{\rho}_1 = \max\{\overline{\rho}_0, \frac{1}{\Lambda}\|z^*\|^\diamond\}$ and $\overline{\rho}_0$ is defined in Theorem 6.1, then

$$\|\Delta\|_\infty \le \frac{k_{h,\Lambda\overline{\rho}_1}^\diamond}{1 - k_{h,\Lambda\overline{\rho}_1}^\diamond \Lambda} \cdot \|\Gamma\|_\infty^+. \tag{7.5}$$

Proof. Since $K \in B_{\overline{\rho}_0}$ and

$$\Pi K = Z^* + \Pi\Delta + \Gamma,$$

we have

$$z^* + \Pi_i\Delta + \Gamma_i = \Pi_i K \in C_{h,\Lambda\overline{\rho}_0}, \quad i \in I^+.$$

Thus, we have

$$z^* + \Pi_i\Delta + \Gamma_i \in C_{h,\Lambda\overline{\rho}_1}, \quad i \in I^+,$$

and

$$z^* \in C_{h,\Lambda\overline{\rho}_1}.$$

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

Let $h < \widehat{h}(\Lambda\overline{\rho}_1)$. For any $i \in I^+$, we have

$$\begin{aligned}
|\Delta_i| &= \left|Q^\diamond_h(z^* + \Pi_i\Delta + \Gamma_i)(c_ih) - Q^\diamond_h(z^*)(c_ih)\right| \\
&= \left|Q^\diamond_{c_ih}\big((z^* + \Pi_i\Delta + \Gamma_i)|_{[0,c_ih]}\big)(c_ih) - Q^\diamond_{c_ih}(z^*|_{[0,c_ih]})(c_ih)\right| \\
&\le \left\|Q^\diamond_{c_ih}\big((z^* + \Pi_i\Delta + \Gamma_i)|_{[0,c_ih]}\big) - Q^\diamond_{c_ih}(z^*|_{[0,c_ih]})\right\|^\diamond \\
&\le k_{c_ih,\Lambda\overline{\rho}_1}\|(\Pi_i\Delta + \Gamma_i)|_{[0,c_ih]}\| \\
&\le k_{c_ih,\Lambda\overline{\rho}_1}\left(\max_{\theta\in[0,c_i]}\sum_{j=1}^{\nu}|a'_{ij}(\theta)|\|\Delta\|_\infty + \|\Gamma_i|_{[0,c_ih]}\|\right) \\
&\le k_{h,\Lambda\overline{\rho}_1}(\Lambda\|\Delta\|_\infty + \|\Gamma\|^+_\infty).
\end{aligned}$$

Thus, (7.5) follows since $k_{h,\Lambda\overline{\rho}_1}\Lambda < 1$ holds by (6.6). □

As for the errors $\Gamma_i \in C_h$, $i \in I^+$, the following proposition holds.

Proposition 7.3. If z^* is of class C^m, where m is a non-negative integer, then, for $i \in I^+$,

$$\begin{aligned}
&\max_{\theta\in[0,c_i]}\left|\Gamma_i(\theta h) - \sum_{k=0}^{m-1}\Gamma_{ik}(\theta)\frac{h^k}{k!}(z^*)^{(k)}(0)\right| \\
&\qquad\le \frac{h^m}{m!}\max_{\theta\in[0,c_i]}\left(\sum_{i=1}^{\nu}|a'_{ij}(\theta)|c_i^m + \theta^m\right)\max_{t\in[0,c_ih]}|z^{(m)}(t)|,
\end{aligned}$$

where, for $k = 0, 1, \dots, m-1$,

$$\Gamma_{ik}(\theta) = \sum_{i=1}^{\nu}a'_{ij}(\theta)c_j^k - \theta^k, \quad \theta \in [0, c_i].$$

Proof. The proof is analogous to the proof of Proposition 7.1. □

Now, we look for an expansion of the error Δ in terms of Γ. For our purposes, it is sufficient to consider the first-order expansion given in the next proposition.

Proposition 7.4. Let us assume that the map $Q^\diamond_h$ is of class C^2. Then:

(i) the map $\mathbf{Q}^\diamond_h$ is of class C^2;

and, under the hypothesis in Proposition 7.2,

(ii) the linear map

$$R(\mathbf{Q}^\diamond_h)'(z^*)\Pi : (\mathbb{R}^d)^\nu \to (\mathbb{R}^d)^\nu$$

has norm less than 1;

(iii)

$$\Delta = L_h\Gamma + \mathbf{R}(\Delta, \Gamma), \tag{7.6}$$

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

where L_h is the linear operator $(C_h)^\nu \to (\mathbb{R}^d)^\nu$ given by

$$L_h Z = \big(I_{(\mathbb{R}^d)^\nu} - R(\mathbf{Q}_h^\diamond)'(z^*)\Pi\big)^{-1} R(\mathbf{Q}_h^\diamond)'(Z^*)Z, \quad Z \in (C_h)^\nu,$$

and

$$\mathbf{R}(\Delta,\Gamma) = \frac{1}{2}\big(I_{(\mathbb{R}^d)^\nu} - R(\mathbf{Q}_h^\diamond)'(z^*)\Pi\big)^{-1} \cdot R\int_0^1 (1-s)(\mathbf{Q}_h^\diamond)''\big(Z^* + s(\Pi\Delta+\Gamma)\big)(\Pi\Delta+\Gamma,\Pi\Delta+\Gamma)\,\mathrm{d}s.$$

Proof. Point (i) is obvious. Note that

$$(\mathbf{Q}_h^\diamond)'(z)Y = \big((Q_h^\diamond)'(Z_1)Y_1,\dots,(Q_h^\diamond)'(Z_\nu)Y_\nu\big), \quad \text{for } Z,Y \in (C_h)^\nu,$$

and

$$(\mathbf{Q}_h^\diamond)''(z)(Y,X) = \big((Q_h^\diamond)''(Z_1)(Y_1,X_1),\dots,(Q_h^\diamond)''(Z_\nu)(Y_\nu,X_\nu)\big), \quad \text{for } Z,Y,X \in (C_h)^\nu.$$

Now, we prove (i) and (ii). Under the hypothesis of Proposition 7.2, *i.e.*, $h < \widehat{h}(\Lambda\overline{\rho}_1)$, we have

$$\|z^*\| < \Lambda\overline{\rho}_1$$

and

$$k_{h,\Lambda\overline{\rho}_1}\Lambda < 1.$$

The map $Q_h^\diamond$ is differentiable at z^*. Note that

$$[(Q_h^\diamond)'(z^*)u](0) = 0, \quad u \in C_h.$$

Moreover, for $0 < h_1 \le h$, the map $Q_{h_1}^\diamond$ is differentiable at $z^*|_{[0,h_1]}$ and

$$(Q_{h_1}^\diamond)'(z^*|_{[0,h_1]})u|_{[0,h_1]} = \big((Q_{h_1}^\diamond)'(z^*)u\big)|_{[0,h_1]}, \quad u \in C_h,$$

and

$$\big\|(Q_{h_1}^\diamond)'\big(z^*|_{[0,h_1]}\big)\big\| \le k_{h_1,\Lambda\overline{\rho}_1}^\diamond.$$

Hence, for $U \in (\mathbb{R}^d)^\nu$, we have

$$\begin{aligned}
\big|\big(R(\mathbf{Q}_h^\diamond)'(Z^*)\Pi U\big)_i\big| &= \big|\big[(Q_h^\diamond)'(z^*)\Pi_i U\big](c_i h)\big| \\
&= \big|\big[(Q_{c_i h}^\diamond)'(z^*|_{[0,c_i h]})\Pi_i U|_{[0,c_i h]}\big](c_i h)\big| \\
&\le \big\|(Q_{c_i h}^\diamond)'(z^*|_{[0,c_i h]})\Pi_i U|_{[0,c_i h]}\big\| \\
&\le k_{c_i h,\Lambda\overline{\rho}_1}\big\|\Pi U|_{[0,c_i h]}\big\| \\
&\le k_{c_i h,\Lambda\overline{\rho}_1}\max_{\theta\in[0,c_i]}\sum_{j=1}^{\nu}|a_{ij}(\theta)|\|U\|_\infty \\
&\le k_{h,\Lambda\overline{\rho}_1}\Lambda\|U\|_\infty
\end{aligned}$$

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

if $i \in I^+$ and

$$\left|\left(R(\mathbf{Q}_h^\diamond)'(Z^*)\Pi U\right)_i\right| = \left|\left[(Q_h^\diamond)'(z^*)\Pi_i U\right](0)\right| = 0$$

if $i \in I^0$. Thus

$$\left\|R(\mathbf{Q}_h^\diamond)'(Z^*)\Pi U\right\| \le k_{h,\Lambda\overline{p}_1}\Lambda < 1.$$

Finally, we prove (iii). Since the map $\mathbf{Q}_h^\diamond$ is of class C^2, we obtain

$$\mathbf{Q}_h^\diamond(Z^* + \Pi\Delta + \Gamma) - \mathbf{Q}_h^\diamond(z^*) = (\mathbf{Q}_h^\diamond)'(z^*)(\Pi\Delta + \Gamma) + \int_0^1 (1-s)(\mathbf{Q}_h^\diamond)''(Z^* + s\Pi\Delta + \Gamma)(\Pi\Delta + \Gamma, \Pi\Delta + \Gamma)\,\mathrm{d}s,$$

and so, by (7.3),

$$\left(I_{(\mathbb{R}^d)^\nu} - R(\mathbf{Q}_h^\diamond)'(z^*)\Pi\right)\Delta = R(\mathbf{Q}_h^\diamond)'(z^*)\Gamma + R\int_0^1 (1-s)(\mathbf{Q}_h^\diamond)''(Z^* + s\Pi\Delta + \Gamma)(\Pi\Delta + \Gamma, \Pi\Delta + \Gamma)\,\mathrm{d}s.$$

Since $\|R(\mathbf{Q}_h^\diamond)'(Z^*)\Pi\| < 1$, we obtain (7.6). □

Under the assumption that the map $(Q_h^\diamond)''$ is bounded in a neighbourhood of z^*, we obtain

$$\Delta = L_h\Gamma + \mathcal{O}\left((\|\Gamma\|_\infty^+)^2\right), \quad \|\Gamma\|_\infty^+ \to 0.$$

Moreover, we can write

$$L_h\Gamma = R(\mathbf{Q}_h^\diamond)'(Z^*)\Gamma + \sum_{k=1}^{\infty}\left(R(\mathbf{Q}_h^\diamond)'(z^*)\Pi\right)^k R(\mathbf{Q}_h^\diamond)'(Z^*)\Gamma.$$

In general, the first term $R(\mathbf{Q}_h^\diamond)'(z^*)\Gamma$ on the right-hand side does not have infinitesimal order, with respect to h, lower than the other terms. However, there are important cases where this happens. This is the subject of the next subsection, where we consider RFDEs (1.1) satisfying the following *Regularity Condition*.

(RC) For any $(\sigma, \varphi) \in \mathbb{R} \times X$, there exists a map $\overline{Q}_h^\diamond(\sigma, \varphi) : C_h \to L_h^\diamond$ such that

$$Q_h^\diamond(\sigma, \varphi)(z) = \overline{Q}_h^\diamond(\sigma, \varphi)\left(\int z\right), \quad z \in C_h.$$

Clearly, RFDEs with data set $\mathcal{C}$ satisfy the Regularity Condition. As for RFDEs with data set $\mathcal{LC}$, the Regularity Condition is satisfied for NDDEs or NSDDIDEs whenever all the delays are non-vanishing.

On local error functions for equations satisfying the Regularity Condition
For equations fulfilling the Regularity Condition, we introduce the map

$$\overline{\mathbf{Q}}_h^\diamond : (C_h)^\nu \to (L_h)^\nu$$

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

defined by

$$\overline{\mathbf{Q}}_h^{\diamond} Z = \big(\overline{Q}_h^{\diamond}(Z_1), \dots, \overline{Q}_h^{\diamond}(Z_\nu)\big), \quad Z \in (C_h)^\nu.$$

Moreover, note that

$$(Q_h^{\diamond})'(z)v = (\overline{Q}_h^{\diamond})'\bigg(\int z\bigg)\int v, \quad z, v \in C_h,$$
$$(Q_h^{\diamond})''(z)v = (\overline{Q}_h^{\diamond})''\bigg(\int z\bigg)\bigg(\int v, \int w\bigg), \quad z, v, w \in C_h.$$

When the Regularity Condition holds, Proposition 7.4 can be restated as follows.

Proposition 7.5. Let us assume that the map $\overline{Q}_h^{\diamond}$ is of class C^2 . Under the hypothesis in Proposition 7.2, we have

$$\Delta = L_h\Gamma + \mathbf{R}(\Delta, \Gamma),$$

where

$$L_h\Gamma = \bigg(I - R(\overline{\mathbf{Q}}_h^{\diamond})'\bigg(\int Z^*\bigg)\int \Pi\bigg)^{-1} R(\overline{\mathbf{Q}}_h^{\diamond})'\bigg(\int Z^*\bigg)\int \Gamma$$

and

$$\mathbf{R}(\Delta, \Gamma) = \frac{1}{2}\bigg(I - R(\overline{\mathbf{Q}}_h^{\diamond})'\bigg(\int Z^*\bigg)\int \Pi\bigg)^{-1}$$
$$\cdot R\int_0^1 (1-s)(\overline{\mathbf{Q}}_h^{\diamond})''\bigg(\int Z^* + s\int(\Pi\Delta + \Gamma)\bigg)\bigg(\int(\Pi\Delta + \Gamma), \int(\Pi\Delta + \Gamma)\bigg)\,\mathrm{d}s$$

with

$$\int Z^* = \bigg(\int z^*, \dots, \int z^*\bigg) \in (C_h)^\nu$$

and

$$\int \Pi : (\mathbb{R}^d)^\nu \to (C_h)^\nu, \ \bigg(\int \Pi\bigg)U = \bigg(\int \Pi_1 U_1, \dots, \int \Pi_\nu U_\nu\bigg).$$

As a consequence of the previous proposition, we obtain

$$\Delta = L_h\Gamma + h^2 \cdot \mathcal{O}\big((\|\Gamma\|_\infty^+)^2\big), \quad \|\Gamma\|_\infty^+ \to 0, \tag{7.7}$$

under the assumption that the map $(\overline{Q}_h^{\diamond})''$ is bounded in a neighbourhood of $\int z^*$.

Now, we refine equation (7.7) by giving an expansion of the components Δ_i, where $i \in I^+$, in terms of powers of h. To this end, we let $c_1^*, \dots, c_{\nu^*}^*$

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

denote the distinct positive abscissae. Moreover, we introduce:

- the continuous functions $g_k : [0,h] \to \mathbb{R}$, $k = 0,1,2,\ldots$, given by

$$g_k(\theta h) = \int_0^\theta \gamma_k(\beta)\,\mathrm{d}\beta = \sum_{i=1}^{\nu} b_i(\theta) c_i^k - \frac{\theta^{k+1}}{k+1}, \quad \theta \in [0,1]; \tag{7.8}$$

- for any $i \in I^+$, the continuous functions $G_{ik} : [0,h] \to \mathbb{R}$, $k = 0,1,2,\ldots$, given by

$$G_{ik}(\theta h) = \int_0^\theta \Gamma_{ik}(\beta)\,\mathrm{d}\beta = \sum_{i=1}^{\nu} a_{ij}(\theta) c_j^k - \frac{\theta^{k+1}}{k+1}, \quad \theta \in [0,1]. \tag{7.9}$$

Finally, we write $(\overline{Q}_h^\diamond)'$ and $(\overline{\mathbf{Q}}_h^\diamond)'$ instead of $(\overline{Q}_h^\diamond)'\big(\int z^*\big)$ and $(\overline{\mathbf{Q}}_h^\diamond)'\big(\int Z^*\big)$, respectively.

Proposition 7.6. Let us assume that:

- the map $\overline{Q}_h^\diamond$ is of class C^2 and $(\overline{Q}_h^\diamond)''$ is bounded in a neighbourhood of $\int z^*$;
- the fixed point z^* of the map Q_h is of class C^3;
- $$\sum_{j=1}^{\nu} a_{ij}(\theta) = \theta, \quad \theta \in [0, c_i] \ \text{ and } \ i \in I^+. \tag{7.10}$$

Then, for any $i \in I^+$, we have

$$\begin{aligned}\Delta_i = {} & h^2\big[(\overline{Q}_h^\diamond)'\; G_{i1}\cdot (z^*)'(0)\big](c_i h) \\ & + \frac{h^3}{2}\big[(\overline{Q}_h^\diamond)'\; G_{i2}\cdot (z^*)''(0)\big](c_i h) + h^3[(\overline{Q}_h^\diamond)' u_i](c_i h) \\ & + \mathcal{O}(h^4),\end{aligned} \tag{7.11}$$

where

$$u_i(t) = \sum_{k=1}^{\nu^*}\left[(\overline{Q}_h^\diamond)' \sum_{\substack{j=1 \\ c_j = c_k^*}}^{\nu} a_{ij}\left(\frac{t}{h}\right) G_{j1}\cdot (z^*)'(0)\right](c_k^* h), \quad t \in [0,h].$$

Proof. Since (7.10) holds, we have, for any $i \in I^+$,

$$\Gamma_{i0}(\theta) = 0, \quad \theta \in [0, c_i],$$

in Proposition 7.3. Hence $\|\Gamma\|_\infty^+ = O(h)$ and so

$$\Delta = L_h \Gamma + \mathcal{O}(h^4).$$

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

Now, we write

$$\begin{aligned} L_h\Gamma &= \left(I - R(\overline{\mathbf{Q}}_h^{\diamond})' \int \Pi\right)^{-1} R(\overline{\mathbf{Q}}_h^{\diamond})' \int \Gamma \\ &= R(\overline{\mathbf{Q}}_h^{\diamond})' \int \Gamma + \left[R(\overline{\mathbf{Q}}_h^{\diamond})' \int \Pi\right] R(\overline{\mathbf{Q}}_h^{\diamond})' \int \Gamma \\ &\quad + \sum_{k=2}^{\infty} \left[R(\overline{\mathbf{Q}}_h^{\diamond})' \int \Pi\right]^k R(\overline{\mathbf{Q}}_h^{\diamond})' \int \Gamma. \end{aligned}$$

Since $\| \int \Gamma \|_{\infty}^{+} = \mathcal{O}(h^2)$ and $\| \int \Pi \| = \mathcal{O}(h)$, we obtain

$$\sum_{k=2}^{\infty} \left[R(\overline{\mathbf{Q}}_h^{\diamond})' \int \Pi\right]^k R(\overline{\mathbf{Q}}_h^{\diamond})' \int \Gamma = \mathcal{O}(h^4).$$

Moreover, for any $i \in I^+$, we have

$$\left(R(\overline{\mathbf{Q}}_h^{\diamond})' \int \Gamma\right)_i = \left[(\overline{Q}_h^{\diamond})' \int \Gamma_i\right](c_i h),$$

where, by Proposition 7.3,

$$\begin{aligned} (\overline{Q}_h^{\diamond})' \int \Gamma_i &= (\overline{Q}_h^{\diamond})' \int \left(\Gamma_{i1}\left(\frac{\cdot}{h}\right) h(z^*)'(0) + \Gamma_{i2}\left(\frac{\cdot}{h}\right)\frac{h^2}{2}(z^*)''(0) + \mathcal{O}(h^3)\right) \\ &= (\overline{Q}_h^{\diamond})' \int_0^{\frac{\cdot}{h}} \left(\Gamma_{i1}(\theta) h^2 (z^*)'(0) + \Gamma_{i2}(\theta)\frac{h^3}{2}(z^*)''(0) + \mathcal{O}(h^4)\right) \mathrm{d}\theta \\ &= h^2 (\overline{Q}_h^{\diamond})' G_{i1} \cdot (z^*)'(0) \\ &\quad + \frac{h^3}{2}(\overline{Q}_h^{\diamond})' G_{i2} \cdot (z^*)''(0) \\ &\quad + \mathcal{O}(h^4). \end{aligned}$$

Finally, for any $i \in I^+$,

$$\left(\left[R(\overline{\mathbf{Q}}_h^{\diamond})' \int \Pi\right] R(\overline{\mathbf{Q}}_h^{\diamond})' \int \Gamma\right)_i = \left[(\overline{Q}_h^{\diamond})' \int \Pi_i \; R(\overline{\mathbf{Q}}_h^{\diamond})' \left(\int Z^*\right) \int \Gamma\right](c_i h),$$

where

$$\begin{aligned} &(\overline{Q}_h^{\diamond})' \int \Pi_i \; R(\mathbf{Q}_h^{\diamond})' \int \Gamma \\ &\quad = (\overline{Q}_h^{\diamond})' \int \left(\sum_{j=1}^{\nu} a'_{ij}\left(\frac{\cdot}{h}\right)\left[(\overline{Q}_h^{\diamond})' \int \Gamma_j\right](c_j h)\right) \\ &\quad = h(\overline{Q}_h^{\diamond})' \int_0^{\frac{\cdot}{h}} \sum_{j=1}^{\nu} a'_{ij}(\theta)\left[(\overline{Q}_h^{\diamond})' \int \Gamma_j\right](c_j h)\, \mathrm{d}\theta. \end{aligned}$$

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

Since

$$(\overline{Q}_h^{\diamond})' \int \Gamma_j = h^2 (\overline{Q}_h^{\diamond})' \int_0^{\dot{h}} \Gamma_{j1}(\beta)\,\mathrm{d}\beta (z^*)'(0) + \mathcal{O}(h^3)$$
$$= h^2 (\overline{Q}_h^{\diamond})' G_{j1} \cdot (z^*)'(0) + \mathcal{O}(h^3),$$

we conclude that

$$\int_0^{\dot{h}} \sum_{j=1}^{\nu} a'_{ij}(\theta) \left[(\overline{Q}_h^{\diamond})' \int \Gamma_j \right] (c_j h)\,\mathrm{d}\theta$$
$$= h^2 \int_0^{\dot{h}} \sum_{j=1}^{\nu} a'_{ij}(\theta) \big[(\overline{Q}_h^{\diamond})' G_{j1} \cdot (z^*)'(0) \big] (c_j h)\,\mathrm{d}\theta + \mathcal{O}(h^3)$$
$$= h^2 \int_0^{\dot{h}} \sum_{k=1}^{\nu^*} \sum_{\substack{j=1 \\ c_j = c_k^*}}^{\nu} a'_{ij}(\theta) \big[(\overline{Q}_h^{\diamond})' G_{j1} \cdot (z^*)'(0) \big] (c_k^* h)\,\mathrm{d}\theta + \mathcal{O}(h^3)$$
$$= h^2 \int_0^{\dot{h}} \sum_{k=1}^{\nu^*} \left[(\overline{Q}_h^{\diamond})' \sum_{\substack{j=1 \\ c_j = c_k^*}}^{\nu} a'_{ij}(\theta) G_{j1} \cdot (z^*)'(0) \right] (c_k^* h)\,\mathrm{d}\theta + \mathcal{O}(h^3),$$

and then

$$(\overline{Q}_h^{\diamond})' \int \Pi_i\ R(\mathbf{Q}_h^{\diamond})' \int \Gamma$$
$$= h^3 (\overline{Q}_h^{\diamond})' \int_0^{\dot{h}} \sum_{k=1}^{\nu^*} \left[(\overline{Q}_h^{\diamond})' \sum_{\substack{j=1 \\ c_j = c_k^*}}^{\nu} a'_{ij}(\theta) G_{j1} \cdot (z^*)'(0) \right] (c_k^* h)\,\mathrm{d}\theta$$
$$+ \mathcal{O}(h^4)$$
$$= h^3 (\overline{Q}_h^{\diamond})' \sum_{k=1}^{\nu^*} \left[(\overline{Q}_h^{\diamond})' \sum_{\substack{j=1 \\ c_j = c_k^*}}^{\nu} \int_0^{\dot{h}} a'_{ij}(\theta)\,\mathrm{d}\theta\ G_{j1} \cdot (z^*)'(0) \right] (c_k^* h)$$
$$+ \mathcal{O}(h^4).$$

Thus, the expansion (7.11) follows. □

Now, we are able to give the following expansions for the local error functions e and E in terms of powers of h.

Theorem 7.1. Let us assume that:

- the map $\overline{Q}_h^{\diamond}$ is of class C^2 and $(\overline{Q}_h^{\diamond})''$ is bounded in a neighbourhood of z^*;
- the fixed point z^* of the map Q_h is of class C^4;

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

- $$\sum_{j=1}^{\nu} a_{ij}(\theta) = \theta, \quad \theta \in [0, c_i] \text{ and } i \in I^+. \tag{7.12}$$

Then, for $\theta \in [0,1]$,

$$\begin{aligned} e(\theta h) = {} & \gamma_0(\theta) \cdot z(0) \\ & + h\gamma_1(\theta) \cdot (z^*)'(0) \\ & + \frac{h^2}{2}\gamma_2(\theta) \cdot (z^*)''(0) \\ & + h^2 \sum_{l=1}^{\nu^*} \left[(\overline{Q}_h^{\diamond})' \sum_{\substack{i=1 \\ c_i = c_l^*}}^{\nu} b_i'(\theta) G_{i1} \cdot (z^*)'(0) \right] (c_l^* h) \\ & + \frac{h^3}{6}\gamma_3(\theta) \cdot (z^*)'''(0) \\ & + \frac{h^3}{2} \sum_{l=1}^{\nu^*} \left[(\overline{Q}_h^{\diamond})' \sum_{\substack{i=1 \\ c_i = c_l^*}}^{\nu} b_i'(\theta) G_{i2} \cdot (z^*)''(0) \right] (c_l^* h) \\ & + h^3 \sum_{l=1}^{\nu^*} \left[(\overline{Q}_h^{\diamond})' w_l(\theta) \right] (c_l^* h) \\ & + \mathcal{O}(h^4), \end{aligned} \tag{7.13}$$

where

$$w_l(\theta)(t) = \sum_{k=1}^{\nu^*} \left[(\overline{Q}_h^{\diamond})' \sum_{\substack{i=1 \\ c_i = c_l^*}}^{\nu} \sum_{\substack{j=1 \\ c_j = c_k^*}}^{\nu} b_i'(\theta) a_{ij}\left(\frac{t}{h}\right) G_{j1} \cdot (z^*)'(0) \right] (c_k^* h),$$

$$\text{for } t \in [0, h],$$

and

$$\begin{aligned} E(\theta h) = {} & h g_0(\theta h) \cdot z^*(0) \\ & + h^2 g_1(\theta h) \cdot (z^*)'(0) \\ & + \frac{h^3}{2} g_2(\theta h) \cdot (z^*)''(0) \\ & + h^3 \sum_{l=1}^{\nu^*} \left[(\overline{Q}_h^{\diamond})' \sum_{\substack{i=1 \\ c_i = c_l^*}}^{\nu} b_i(\theta) G_{i1} \cdot (z^*)'(0) \right] (c_l^* h) \\ & + \frac{h^4}{6} g_3(\theta h) \cdot (z^*)'''(0) \end{aligned}$$

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

$$+\frac{h^4}{2}\sum_{l=1}^{\nu^*}\left[(\overline{Q}_h^{\diamond})'\sum_{\substack{i=1\\ c_i=c_l^*}}^{\nu} b_i(\theta)G_{i2}\cdot(z^*)''(0)\right](c_l^*h)$$
$$+h^4\sum_{l=1}^{\nu^*}\left[(\overline{Q}_h^{\diamond})'\,W_l(\theta)\right](c_l^*h)$$
$$+\mathcal{O}(h^5), \tag{7.14}$$

where

$$W_l(\theta)(t)=\sum_{k=1}^{\nu^*}\left[(\overline{Q}_h^{\diamond})'\sum_{\substack{i=1\\ c_i=c_l^*}}^{\nu}\sum_{\substack{j=1\\ c_j=c_k^*}}^{\nu} b_i(\theta)a_{ij}\left(\frac{t}{h}\right)G_{j1}\cdot(z^*)'(0)\right](c_k^*h),$$

for $t\in[0,h]$.

Proof. The expansion (7.13) follows by (7.1) and Propositions 7.1 and 7.6. The expansion (7.14) follows by integrating (7.13). □

7.2. Order conditions

In this section, we establish conditions on FCRK methods to obtain, for RFDEs with data set $\mathcal{C}$, a prescribed uniform or discrete order and, for RFDEs with data set $\mathcal{LC}$, a prescribed order. Moreover, by using such order conditions, we construct explicit methods attaining a given global order.

RFDEs with data set $\mathcal{C}$

Since RFDEs with data set $\mathcal{C}$ satisfy the Regularity Assumption, we can use the expansion (7.14) in Theorem 7.1 for the local error $E(\sigma,\varphi,h)$.

We consider an FCRK method satisfying the condition (7.12) and a family $\mathcal{F}$ of integration problems (t_0,ϕ,T) such that:

(i) for all $t\in[t_0,t_0+T)$, the map $\overline{Q}_h^{\diamond}(t,y_t)$ is of class C^2;

(ii) there exists $\varepsilon>0$, $\overline{H}>0$ and $M\geq 0$ such that, for all $t\in[t_0,t_0+T)$ and $h\in(0,T-t]$,
$$\left\|\left(\overline{Q}_h^{\diamond}(t,y_t)\right)''(z)\right\|\leq M$$
whenever $z\in C_h$ is such that $\|z-z^*(t,y_t,h)\|\leq\varepsilon$;

(iii) $y|_{[t_0,t_0+T]}$ is piecewise C^5.

Note that conditions (i), (ii) and (iii) permit us to use expansion (7.14) and to assume that the term $\mathcal{O}(h^5)$ is uniformly bounded with respect to $t\in[t_0,t_0+T)$. Moreover, we remark that a family of integration problems of DDEs, DIDEs, SDDDEs or SDDIDEs fulfils conditions (i), (ii) and (iii) if the function f, the delays and the kernel are sufficiently smooth (with respect to the variable t only piecewise smoothness is required).

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

Table 7.1. Uniform order conditions.

Order	Conditions
1	$\sum_{i=1}^{\nu} b_i(\theta) = \theta, \quad \theta \in [0,1]$
2	$\sum_{i=1}^{\nu} b_i(\theta) c_i = \frac{\theta^2}{2}, \quad \theta \in [0,1]$
3	$\sum_{i=1}^{\nu} b_i(\theta) c_i^2 = \frac{\theta^3}{3}, \quad \theta \in [0,1]$ For any $k = 1, \dots, \nu^*$, $\sum_{\substack{i=1 \\ c_i = c_k^*}}^{\nu} b_i(\theta) \left(\sum_{j=1}^{\nu} a_{ij}(\beta) c_j - \frac{\beta^2}{2} \right) = 0, \quad \theta \in [0,1] \text{ and } \beta \in [0, c_k^*]$
4	$\sum_{i=1}^{\nu} b_i(\theta) c_i^3 = \frac{\theta^4}{4}, \quad \theta \in [0,1]$ For any $k = 1, \dots, \nu^*$, $\sum_{\substack{i=1 \\ c_i = c_k^*}}^{\nu} b_i(\theta) \left(\sum_{j=1}^{\nu} a_{ij}(\beta) c_j^2 - \frac{\beta^3}{3} \right) = 0, \quad \theta \in [0,1] \text{ and } \beta \in [0, c_k^*]$ For any $l, k = 1, \dots, \nu^*$, $\sum_{\substack{i=1 \\ c_i = c_l^*}}^{\nu} \sum_{\substack{j=1 \\ c_j = c_k^*}}^{\nu} b_i(\theta) a_{ij}(\beta) \left(\sum_{k=1}^{\nu} a_{jk}(\gamma) c_k - \frac{\gamma^2}{2} \right) = 0,$ for $\theta \in [0,1], \quad \beta \in [0, c_l^*] \text{ and } \gamma \in [0, c_k^*]$

In Table 7.1, we give the conditions for getting uniform order two, three and four on the family $\mathcal{F}$.

The order conditions are obtained by expansion (7.14) by recalling (7.8) and (7.9). Such conditions are not only sufficient for getting the prescribed order, but they are also necessary.

As for the discrete order, the conditions for getting discrete order two, three and four are obtained by replacing $b_i(\theta)$ with $b_i = b_i(1)$ in Table 6.1.

The convergence Theorem 5.1 guarantees that a global order r on the family $\mathcal{F}$ is attained if the method has uniform order $r-1$, has discrete order r and is stable on $\mathcal{F}$. We remark that, exactly as in the case of

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

the above conditions (i), (ii) and (iii), the method is stable on a family of integration problems of DDEs, DIDEs, SDDEs and SDDIDEs if the function f, the delays and the kernel are sufficiently smooth.

It is clear that, for any $r \in \{1, 2, 3, 4\}$, by taking

$$a_{ij}(\theta) = b_j(\theta), \quad \theta \in [0, 1], \quad i, j = 1, \ldots, \nu, \tag{7.15}$$

we obtain global order r if

$$\sum_{i=1}^{\nu} b_i(\theta) c_i^{k-1} = \frac{\theta^{k-1}}{k}, \quad \theta \in [0, 1] \ \text{ and } \ k = 1, \ldots, r-1, \tag{7.16}$$

and

$$\sum_{i=1}^{\nu} b_i c_i^{r-1} = \frac{1}{r}. \tag{7.17}$$

Indeed, for methods of type (7.15), these conditions guarantee global order r even if $r > 4$.

As an example, the conditions for $r = 2$ show that global order two is attained by the one-stage method

$$\begin{array}{c|c} \frac{1}{2} & \theta \\ \hline & \theta \end{array},$$

which can be called the *functional midpoint* method. We also note that a one-stage method,

$$\begin{array}{c|c} c_1 & \theta \\ \hline & \theta \end{array}, \tag{7.18}$$

has uniform order one and global order one. The global order two is attained only if $c_1 = \frac{1}{2}$.

By using conditions (7.15) and (7.16), only implicit methods can be constructed. On the other hand, by considering methods not satisfying (7.15), we can obtain explicit methods or semi-implicit methods (*i.e.*, methods such that $a_{ij}(\theta) = 0$ for $j > i$). Here, we will consider only the case of explicit methods.

Now, for each $r \in \{1, 2, 3, 4\}$, we construct explicit methods (satisfying the condition (7.12) and attaining global order r.

$r = 1$

Global order one is obtained by one-stage explicit methods,

$$\begin{array}{c|c} 0 & 0 \\ \hline & b_1(\theta) \end{array},$$

with discrete order one. The condition for the discrete order one is $b_1 = 1$. Hence, the functional explicit Euler method has global order one.

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

$r = 2$

Global order two is obtained by two-stage explicit methods,

$$\begin{array}{c|cc} 0 & 0 & 0 \\ c_2 & \theta & 0 \\ \hline & b_1(\theta) & b_2(\theta) \end{array},$$

with discrete order two and uniform order one. The conditions for discrete order two and uniform order one are

$$b_1(\theta) + b_2(\theta) = \theta, \quad \theta \in [0,1],$$

$$b_2 = \frac{1}{2c_2}.$$

In particular, the family of FCRK methods

$$\begin{array}{c|cc} 0 & 0 & 0 \\ c_2 & \theta & 0 \\ \hline & \theta - \frac{\theta^2}{2c_2} & \frac{\theta^2}{2c_2} \end{array}, \tag{7.19}$$

with uniform order two, which satisfies

$$b_1(\theta)c_1 + b_2(\theta)c_2 = \frac{\theta^2}{2}, \quad \theta \in [0,1],$$

has global order two. Particular elements of this family are the functional Heun method (6.3) (obtained for $c = 1$) and the method

$$\begin{array}{c|cc} 0 & 0 & 0 \\ \frac{1}{2} & \theta & 0 \\ \hline & \theta - \theta^2 & \theta^2 \end{array}$$

(obtained for $c = \frac{1}{2}$), which reduces to the Runge method for an ODE.

$r = 3$

Global order three is obtained by three-stage explicit methods,

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ c_2 & \theta & 0 & 0 \\ c_3 & \theta - a_{32}(\theta) & a_{32}(\theta) & 0 \\ \hline & b_1(\theta) & b_2(\theta) & b_3(\theta) \end{array},$$

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

with discrete order three and uniform order two. The conditions for obtaining such orders are

$$\begin{aligned} b_1(\theta) &= \theta - b_2(\theta) - b_3(\theta), \quad \theta \in [0,1], \\ b_2(\theta)c_2 + b_3(\theta)c_3 &= \frac{\theta^2}{2}, \quad \theta \in [0,1], \\ b_2c_2^2 + b_3c_3^2 &= \frac{1}{3}, \end{aligned}$$

and

$$\begin{aligned} b_2 &= 0, \\ b_3\left(a_{32}(\beta)c_2 - \frac{\beta^2}{2}\right) &= 0, \quad \beta \in [0,c_3], \end{aligned}$$

if $c_2 \neq c_3$ and

$$b_2\left(\beta - \frac{\beta^2}{2}\right) + b_3\left(a_{32}(\beta)c_2 - \frac{\beta^2}{2}\right) = 0, \quad \beta \in [0,c_3],$$

if $c_2 = c_3$.

By choosing $b_2(\theta)=0$, the previous conditions select the family of methods

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ c_2 & \theta & 0 & 0 \\ \frac{2}{3} & \theta - \frac{\theta^2}{2c_2} & \frac{\theta^2}{2c_2} & 0 \\ \hline & \theta - \frac{3}{4}\theta^2 & 0 & \frac{3}{4}\theta^2 \end{array} .$$

On the other hand, by taking $c_2 = c_3$, the previous conditions reduce to

$$\begin{aligned} b_1(\theta) &= \theta - b_2(\theta) - b_3(\theta), \quad \theta \in [0,1], \\ b_2(\theta) + b_3(\theta) &= \frac{\theta^2}{2c_2}, \quad \theta \in [0,1], \\ b_2 + b_3 &= \frac{1}{3c_2^2}, \\ b_3 &\neq 0, \\ a_{32}(\beta) &= \frac{\beta^2}{2c_2}\left(1 + \frac{b_2}{b_3}\right) - \frac{b_2}{c_2 b_3}, \quad \beta \in [0,1], \end{aligned}$$

which are equivalent to

$$\begin{aligned} c_2 &= \frac{2}{3}, \\ b_1(\theta) &= \theta - b_2(\theta) - b_3(\theta), \quad \theta \in [0,1], \end{aligned}$$

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

$$b_2(\theta) = \frac{3\theta^2}{4} - b_3(\theta), \quad \theta \in [0,1],$$
$$a_{32}(\beta) = \frac{9\beta^2}{16 b_3}, \quad \beta \in [0, c_3].$$

$r = 4$

Global order four is obtained by explicit methods of discrete order four and uniform order three.

It is known that explicit continuous RK methods attain uniform order three if they have at least four stages (see Table 8.3 in Section 8). Hence, at least four stages have to be used for an FCRK method of uniform order three. One can prove that there do not exist explicit five-stage FCRK methods of discrete order four. Hence, we consider six-stage explicit methods:

$$\begin{array}{c|cccccc}
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
c_2 & \theta & 0 & 0 & 0 & 0 & 0 \\
c_3 & \theta - a_{32}(\theta) & a_{32}(\theta) & 0 & 0 & 0 & 0 \\
c_4 & \theta - \sum_{j=2}^{3} a_{4j}(\theta) & a_{42}(\theta) & a_{43}(\theta) & 0 & 0 & 0 \\
c_5 & \theta - \sum_{j=2}^{4} a_{5j}(\theta) & a_{52}(\theta) & a_{53}(\theta) & a_{54}(\theta) & 0 & 0 \\
c_6 & \theta - \sum_{j=2}^{5} a_{6j}(\theta) & a_{62}(\theta) & a_{63}(\theta) & a_{64}(\theta) & a_{65}(\theta) & 0 \\
\hline
 & b_1(\theta) & b_2(\theta) & b_3(\theta) & b_4(\theta) & b_5(\theta) & b_6(\theta)
\end{array} \quad . \tag{7.20}$$

Proposition 7.7. An explicit six-stage RK method (7.20) is of uniform order three and discrete order four if (and only if for distinct abscissae)

$$\frac{c_5 + c_6}{3} - \frac{c_5 c_6}{2} = \frac{1}{4},$$
$$b_1(\theta) = \theta - b_3(\theta) - b_4(\theta) - b_5(\theta) - b_6(\theta), \quad \theta \in [0,1],$$
$$b_2(\theta) = 0, \quad \theta \in [0,1],$$
$$b_3 = b_4 = 0,$$
$$b_3(\theta) c_3 + b_4(\theta) c_4 + b_5(\theta) c_5 + b_6(\theta) c_6 = \frac{\theta^2}{2}, \quad \theta \in [0,1],$$
$$b_3(\theta) c_3^2 + b_4(\theta) c_4^2 + b_5(\theta) c_5^2 + b_6(\theta) c_6^2 = \frac{\theta^3}{3}, \quad \theta \in [0,1],$$
$$a_{32}(\beta) = \frac{\beta^2}{2c_2}, \quad \beta \in [0, c_2],$$

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

$$
\begin{aligned}
&a_{42}(\beta)c_2 + a_{43}(\beta)c_3 = \frac{\beta^2}{2}, \quad \beta \in [0, c_4],\\
&a_{52}(\beta) = 0, \quad \beta \in [0, c_5],\\
&a_{53}(\beta) = \frac{\beta^2 c_4}{2c_3(c_4 - c_3)} - \frac{\beta^3}{3c_3(c_4 - c_3)}, \quad \beta \in [0, c_5],\\
&a_{54}(\beta) = -\frac{\beta^2 c_3}{2c_4(c_4 - c_3)} + \frac{\beta^3}{3c_4(c_4 - c_3)}, \quad \beta \in [0, c_5],\\
&a_{62}(\beta) = 0, \quad \beta \in [0, c_6],\\
&a_{63}(\beta)c_3 + a_{64}(\beta)c_4 + a_{65}(\beta)c_5 = \frac{\beta^2}{2}, \quad \beta \in [0, c_6],\\
&a_{63}(\beta)c_3^2 + a_{64}(\beta)c_4^2 + a_{65}(\beta)c_4^2 = \frac{\beta^3}{3}, \quad \beta \in [0, c_6].
\end{aligned} \tag{7.21}
$$

Proof. Sufficient conditions (and also necessary in the case of distinct abscissae) for uniform order three and discrete order four are as follows, divided into three blocks (see Table 7.1).

(1)
$$
\begin{aligned}
&b_1(\theta) + b_2(\theta) + b_3(\theta) + b_4(\theta) + b_5(\theta) + b_6(\theta) = \theta, \quad \theta \in [0, 1],\\
&b_2(\theta)c_2 + b_3(\theta)c_3 + b_4(\theta)c_4 + b_5(\theta)c_5 + b_6(\theta)c_6 = \frac{\theta^2}{2}, \quad \theta \in [0, 1],\\
&b_2(\theta)c_2^2 + b_3(\theta)c_3^2 + b_4(\theta)c_4^2 + b_5(\theta)c_5^2 + b_6(\theta)c_6^2 = \frac{\theta^3}{3}, \quad \theta \in [0, 1],\\
&b_2c_2^3 + b_3c_3^3 + b_4c_4^3 + b_5c_5^3 + b_6c_6^3 = \frac{1}{4}.
\end{aligned}
$$

(2)
$$
\begin{aligned}
&b_2(\theta)\left(-\frac{\beta^2}{2}\right) = 0, \quad \theta \in [0, 1],\ \beta \in [0, c_2],\\
&b_2 \cdot \left(-\frac{\beta^3}{3}\right) = 0, \quad \beta \in [0, c_2],\\
&b_3(\theta)\left(a_{32}(\beta)c_2 - \frac{\beta^2}{2}\right) = 0, \quad \theta \in [0, 1],\ \beta \in [0, c_3],\\
&b_3 \cdot \left(a_{32}(\beta)c_2^2 - \frac{\beta^3}{3}\right) = 0, \quad \beta \in [0, c_3],\\
&b_4(\theta)\left(a_{42}(\beta)c_2 + a_{43}(\beta)c_3 - \frac{\beta^2}{2}\right) = 0, \quad \theta \in [0, 1],\ \beta \in [0, c_4],\\
&b_4 \cdot \left(a_{42}(\beta)c_2^2 + a_{43}(\beta)c_3^2 - \frac{\beta^3}{3}\right) = 0, \quad \beta \in [0, c_4],
\end{aligned}
$$

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

$$b_5(\theta)\left(a_{52}(\beta)c_2 + a_{53}(\beta)c_3 + a_{54}(\beta)c_4 - \frac{\beta^2}{2}\right) = 0,$$
$$\text{for} \quad \theta \in [0,1],\ \beta \in [0, c_5],$$

$$b_5 \cdot \left(a_{52}(\beta)c_2^2 + a_{53}(\beta)c_3^2 + a_{54}(\beta)c_4^2 - \frac{\beta^3}{3}\right) = 0, \quad \beta \in [0, c_5],$$

$$b_6(\theta)\left(a_{62}(\beta)c_2 + a_{63}(\beta)c_3 + a_{64}(\beta)c_4 + a_{65}(\beta)c_5 - \frac{\beta^2}{2}\right) = 0,$$
$$\text{for} \quad \theta \in [0,1],\ \beta \in [0, c_6],$$

$$b_6 \cdot \left(a_{62}(\beta)c_2^2 + a_{63}(\beta)c_3^2 + a_{64}(\beta)c_4^2 + a_{65}(\beta)c_4^2 - \frac{\beta^3}{3}\right) = 0,$$
$$\text{for} \quad \beta \in [0, c_6].$$

(3) $$b_3 a_{32}(\beta)\left(-\frac{\gamma^2}{2}\right) = 0, \quad \beta \in [0, c_3],\ \gamma \in [0, c_2],$$

$$b_4 a_{42}(\beta)\left(-\frac{\gamma^2}{2}\right) = 0, \quad \beta \in [0, c_4],\ \gamma \in [0, c_2],$$

$$b_4 a_{43}(\beta)\left(a_{32}(\gamma)c_2 - \frac{\gamma^2}{2}\right) = 0, \quad \beta \in [0, c_4],\ \gamma \in [0, c_3],$$

$$b_5 a_{52}(\beta)\left(-\frac{\gamma^2}{2}\right) = 0, \quad \beta \in [0, c_5],\ \gamma \in [0, c_2],$$

$$b_5 a_{53}(\beta)\left(a_{32}(\gamma)c_2 - \frac{\gamma^2}{2}\right) = 0, \quad \beta \in [0, c_5],\ \gamma \in [0, c_3],$$

$$b_5 a_{54}(\beta)\left(a_{42}(\gamma)c_2 + a_{43}(\gamma)c_3 - \frac{\gamma^2}{2}\right) = 0, \quad \beta \in [0, c_5],\ \gamma \in [0, c_4],$$

$$b_6 a_{62}(\beta)\left(-\frac{\gamma^2}{2}\right) = 0, \quad \beta \in [0, c_6],\ \gamma \in [0, c_2],$$

$$b_6 a_{63}(\beta)\left(a_{32}(\gamma)c_2 - \frac{\gamma^2}{2}\right) = 0, \quad \beta \in [0, c_6],\ \gamma \in [0, c_3],$$

$$b_6 a_{64}(\beta)\left(a_{42}(\gamma)c_2 + a_{43}(\gamma)c_3 - \frac{\gamma^2}{2}\right) = 0, \quad \beta \in [0, c_6],\ \gamma \in [0, c_4],$$

$$b_6 a_{65}(\beta)\left(a_{52}(\gamma)c_2 + a_{53}(\gamma)c_3 + a_{54}(\gamma)c_4 - \frac{\gamma^2}{2}\right) = 0,$$
$$\text{for} \quad \beta \in [0, c_6],\ \gamma \in [0, c_5].$$

The first condition in block (2) implies $b_2(\theta) = 0$, the third and fourth

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

conditions in (2) imply $b_3 = 0$, and the fifth and sixth conditions, together with the second condition in block (3), imply $b_4 = 0$.

For a method with $b_2(\theta) = 0$ and $b_3 = b_4 = 0$, the conditions in block (1) are satisfied only if $b_5, b_6 \neq 0$ and are equivalent to

$$b_1(\theta) = \theta - b_3(\theta) - b_4(\theta) - b_5(\theta) - b_6(\theta), \quad \theta \in [0,1],$$

$$b_3(\theta)c_3 + b_4(\theta)c_4 + b_5(\theta)c_5 + b_6(\theta)c_6 = \frac{1}{2}, \quad \theta \in [0,1],$$

$$b_3(\theta)c_3^2 + b_4(\theta)c_4^2 + b_5(\theta)c_5^2 + b_6(\theta)c_6^2 = \frac{1}{3}, \quad \theta \in [0,1],$$

$$\frac{c_5 + c_6}{3} - \frac{c_5 c_6}{2} = \frac{1}{4}.$$

For a method with $b_2(\theta) = 0$, $b_3 = b_4 = 0$ and $b_5, b_6 \neq 0$, the conditions in block (2) are equivalent to

$$b_3(\theta)\left(a_{32}(\beta)c_2 - \frac{\beta^2}{2}\right) = 0, \quad \theta \in [0,1],\ \beta \in [0, c_3],$$

$$b_4(\theta)\left(a_{42}(\beta)c_2 + a_{43}(\beta)c_3 - \frac{\beta^2}{2}\right) = 0, \quad \theta \in [0,1],\ \beta \in [0, c_4],$$

$$a_{52}(\beta)c_2 + a_{53}(\beta)c_3 + a_{54}(\beta)c_4 = \frac{\beta^2}{2}, \quad \beta \in [0, c_5],$$

$$a_{52}(\beta)c_2^2 + a_{53}(\beta)c_3^2 + a_{54}(\beta)c_4^2 = \frac{\beta^3}{3}, \quad \beta \in [0, c_5],$$

$$a_{62}(\beta)c_2 + a_{63}(\beta)c_3 + a_{64}(\beta)c_4 + a_{65}(\beta)c_5 = \frac{\beta^2}{2}, \quad \beta \in [0, c_6],$$

$$a_{62}(\beta)c_2^2 + a_{63}(\beta)c_3^2 + a_{64}(\beta)c_4^2 + a_{65}(\beta)c_4^2 = \frac{\beta^3}{3}, \quad \beta \in [0, c_6],$$

and the conditions in block (3) are equivalent to

$$a_{52}(\beta) = 0, \quad \beta \in [0, c_5],$$

$$a_{53}(\beta)\left(a_{32}(\gamma)c_2 - \frac{\gamma^2}{2}\right) = 0, \quad \beta \in [0, c_5],\ \gamma \in [0, c_3],$$

$$a_{54}(\beta)\left(a_{42}(\gamma)c_2 + a_{43}(\gamma)c_3 - \frac{\gamma^2}{2}\right) = 0, \quad \beta \in [0, c_5],\ \gamma \in [0, c_4],$$

$$a_{62}(\beta) = 0, \quad \beta \in [0, c_6],$$

$$a_{63}(\beta)\left(a_{32}(\gamma)c_2 - \frac{\gamma^2}{2}\right) = 0, \quad \beta \in [0, c_6],\ \gamma \in [0, c_3],$$

$$a_{64}(\beta)\left(a_{42}(\gamma)c_2 + a_{43}(\gamma)c_3 - \frac{\gamma^2}{2}\right) = 0, \quad \beta \in [0, c_6],\ \gamma \in [0, c_4],$$

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

$$a_{65}(\beta)\left(a_{52}(\gamma)c_2 + a_{53}(\gamma)c_3 + a_{54}(\gamma)c_4 - \frac{\gamma^2}{2}\right) = 0,$$
$$\text{for} \quad \beta \in [0, c_6], \ \gamma \in [0, c_5].$$

So, the conditions in blocks (2) and (3) are equivalent to

$$a_{32}(\beta)c_2 = \frac{\beta^2}{2}, \quad \beta \in [0, c_3],$$
$$a_{42}(\beta)c_2 + a_{43}(\beta)c_3 = \frac{\beta^2}{2}, \quad \beta \in [0, c_4],$$
$$a_{52}(\beta) = 0, \quad \beta \in [0, c_5],$$
$$a_{53}(\beta)c_3 + a_{54}(\beta)c_4 = \frac{\beta^2}{2}, \quad \beta \in [0, c_5],$$
$$a_{53}(\beta)c_3^2 + a_{54}(\beta)c_4^2 = \frac{\beta^3}{3}, \quad \beta \in [0, c_5],$$
$$a_{62}(\beta) = 0, \quad \beta \in [0, c_6],$$
$$a_{63}(\beta)c_3 + a_{64}(\beta)c_4 + a_{65}(\beta)c_5 = \frac{\beta^2}{2}, \quad \beta \in [0, c_6],$$
$$a_{63}(\beta)c_3^2 + a_{64}(\beta)c_4^2 + a_{65}(\beta)c_4^2 = \frac{\beta^3}{3}, \quad \beta \in [0, c_6].$$

Now, conditions (7.21) follow. □

The set $(c_5, c_6) \in [0, 1]^2$ satisfying the first of conditions (7.21) is shown in Figure 7.1. For example, by taking

$$c_2 = 1, \quad c_3 = \frac{1}{2}, \quad c_4 = 1, \quad c_5 = \frac{1}{2}, \quad c_6 = 1,$$
$$b_2(\theta) = 0, \quad b_3(\theta) = 0, \quad b_4(\theta) = 0,$$
$$a_{43}(\beta) = 0, \quad a_{65}(\beta) = 0,$$

we obtain the particular method of global order four which was proposed in Tavernini (1971).

We remark that there does not exist an explicit six-stage RK method (7.20) with distinct abscissae of uniform order four.

RFDEs with data set $\mathcal{LC}$

First, we consider equations satisfying the Regularity Assumption. For such equations, we can use expansion (7.13) in Theorem 7.1 for the local error function $e(\sigma, \varphi, h)$.

We consider an FCRK method satisfying condition (7.12) and a family $\mathcal{F}$ of integration problems (t_0, ϕ, T) such that conditions (i), (ii) and (iii) of

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

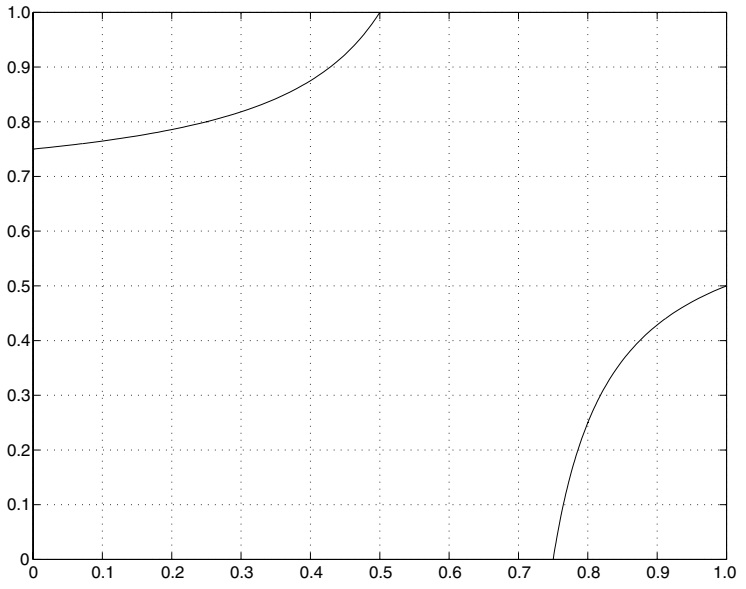


Figure 7.1. The curves are the set of couples $(c_5, c_6) \in [0,1]^2$ satisfying the first of conditions (7.21).

the previous subsection hold. A family of integration problems of NDDEs, NDIDEs and NSDDDEs with non-vanishing delays fulfils conditions (i), (ii) and (iii) if the function f, the delays and the kernel are sufficiently smooth.

In Table 7.2, we give the conditions for getting order one, two, three and four on the family $\mathcal{F}$.

It is clear that Table 7.2 reduces to Table 7.1 in the case of RFDEs with data set $\mathcal{C}$. Hence, the functional explicit Euler method has order one and the two-stage methods in the family (7.19) have order two.

Now, we consider equations not satisfying the Regularity Assumption and a family $\mathcal{F}$ of integration problems such that only condition (iii) of the previous subsection holds. By (7.1), (7.4) and Propositions 7.1 and 7.3, methods of type (7.15) attain order q if

$$\sum_{i=1}^{\nu} b_i'(\theta) c_i^{k-1} = \theta^{k-1}, \quad \theta \in [0,1] \ \text{ and } \ k = 1, \dots, q,$$

or, equivalently,

$$\sum_{i=1}^{\nu} b_i(\theta) c_i^{k-1} = \frac{\theta^k}{k}, \quad \theta \in [0,1] \ \text{ and } \ k = 1, \dots, q.$$

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

Table 7.2. Order conditions.

Order	Conditions
1	$\sum_{i=1}^{\nu} b_i'(\theta) = 1, \quad \theta \in [0,1]$
2	$\sum_{i=1}^{\nu} b_i'(\theta) c_i = \theta, \quad \theta \in [0,1]$
3	$\sum_{i=1}^{\nu} b_i'(\theta) c_i^2 = \theta^2, \quad \theta \in [0,1]$ For any $k = 1, \dots, \nu^*$, $\sum_{\substack{i=1 \\ c_i = c_k^*}}^{\nu} b_i'(\theta) \left(\sum_{j=1}^{\nu} a_{ij}(\beta) c_j - \frac{\beta^2}{2} \right) = 0, \quad \theta \in [0,1], \quad \beta \in [0, c_k^*]$
4	$\sum_{i=1}^{\nu} b_i'(\theta) c_i^3 = \theta^3, \quad \theta \in [0,1]$ For any $k = 1, \dots, \nu^*$, $\sum_{\substack{i=1 \\ c_i = c_k^*}}^{\nu} b_i'(\theta) \left(\sum_{j=1}^{\nu} a_{ij}(\beta) c_j^2 - \frac{\beta^3}{3} \right) = 0, \quad \theta \in [0,1], \quad \beta \in [0, c_k^*]$ For any $l, k = 1, \dots, \nu^*$, $\sum_{\substack{i=1 \\ c_i = c_l^*}}^{\nu} \sum_{\substack{j=1 \\ c_j = c_k^*}}^{\nu} b_i'(\theta) a_{ij}(\beta) \left(\sum_{k=1}^{\nu} a_{jk}(\gamma) c_k - \frac{\gamma^2}{2} \right) = 0,$ for $\theta \in [0,1], \quad \beta \in [0, c_l^*], \quad \gamma \in [0, c_k^*]$

Hence, the one stage methods (7.18) (including the functional explicit Euler method) have order one and the two-stage semi-implicit methods

$$\begin{array}{c|cc} 0 & 0 & 0 \\ c_2 & \theta - \frac{\theta^2}{2c_2} & \frac{\theta^2}{2c_2} \\ \hline & \theta - \frac{\theta^2}{2c_2} & \frac{\theta^2}{2c_2} \end{array}$$

have order two.

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

We conclude by remarking that, in both cases of equations satisfying and not satisfying the Regularity Assumption, the convergence Theorem 5.2 guarantees that a global order r on the family $\mathcal{F}$ is attained if the method has order r and is stable on $\mathcal{F}$. The method turns out to be stable on a family of integration problems of NDDEs, NDIDEs and NSDDIDEs if the function f, the delays and the kernel are sufficiently smooth.

8. The standard approach

In this section we outline the *standard approach* based on continuous RK methods, as described in the Introduction, applied to the specific classes of DDEs,

$$\begin{aligned} y'(t) &= f\big(t, y(t), y\big(t-\tau(t,y(t))\big)\big), \quad t_0 \le t \le t_f, \\ y(t) &= \phi(t), \quad t \le t_0, \end{aligned} \tag{8.1}$$

and NDDEs

$$\begin{aligned} y'(t) &= f\big(t, y(t), y\big(t-\tau(t,y(t))\big), y'\big(t-\tau(t,y(t))\big)\big), \quad t_0 \le t \le t_f, \\ y(t) &= \phi(t), \quad t \le t_0. \end{aligned} \tag{8.2}$$

In order to simplify the notation, we consider one single delay. Moreover, from now on, the end of the integration interval will be denoted by t_f instead of $t_0 + T$, which was used in the previous sections.

Given a mesh $\Delta = \{t_0, t_1, \ldots, t_n, \ldots, t_N = t_f\}$, the standard approach for (8.1) consists in solving step by step, by means of the chosen continuous RK method, the local problems

$$\begin{aligned} w'_{n+1}(t) &= f\big(t, w_{n+1}(t), x\big(t-\tau(t,w_{n+1}(t))\big)\big), \quad t_n \le t \le t_{n+1}, \\ w_{n+1}(t_n) &= y_n, \end{aligned} \tag{8.3}$$

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.

Analogously, the standard approach for (8.2) consists in solving step by step the local problems

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

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

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

$$\lambda(t) = \eta'(t) \tag{8.4}$$

or by

$$\lambda(t) = \mathcal{P}\Big(f\big(\cdot, \eta(\cdot), \eta\big(\cdot - \tau(\cdot, \eta(\cdot))\big), \lambda\big(\cdot - \tau(\cdot, \eta(\cdot))\big)\big)\Big)(t), \tag{8.5}$$

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 $[t_k, t_{k+1}]$.

Here we report a condensed survey on continuous RK methods for ODEs as a basic tool for the implementation of the standard approach for RFDEs. Then we provide the main results on the error analysis of the resulting methods for equations (8.1) and (8.2), also in view of the particular issues treated in the forthcoming Sections 9, 10 and 11. These topics are covered in the book by Bellen and Zennaro (2003) and, hence, neither proofs nor bibliographic references are given here.

8.1. Continuous RK methods for ODEs

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

$$\begin{aligned} y'(t) &= g(t, y(t)), \quad t_0 \le t \le t_f, \\ y(t_0) &= y_0, \end{aligned} \tag{8.6}$$

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

$$Y^i_{n+1} = y_n + h_{n+1} \sum_{j=1}^{\nu} a_{ij} g(t^j_{n+1}, Y^j_{n+1}), \quad i = 1, \ldots, \nu, \tag{8.7}$$

$$y_{n+1} = y_n + h_{n+1} \sum_{i=1}^{\nu} b_i g(t^i_{n+1}, Y^i_{n+1}), \tag{8.8}$$

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_is are called *weights* of

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

the quadrature formula (8.8) and the c_is are called *abscissae* and, for most common methods, they belong to $[0,1]$. Since the RK method (8.7), (8.8) 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,c). 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 (8.7), (8.8) takes the form

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

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

Note that K notation is obtained by setting

$$K_{n+1}^i = g(t_{n+1}^i, Y_{n+1}^i), \quad i=1,\dots,\nu,$$

in (8.7), (8.8).

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 (8.7), (8.8) 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 (8.6).

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

$$\eta(t_n+\theta h_{n+1}) = y_n + h_{n+1}\sum_{i=1}^{\nu} b_i(\theta)g(t_{n+1}^i, Y_{n+1}^i), \quad 0\le\theta\le 1, \tag{8.9}$$

or, in 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

$$b_i(0)=0 \quad \text{and} \quad b_i(1)=b_i, \quad i=1,\dots,\nu, \tag{8.10}$$

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

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

$$\eta(t_n + \theta h_{n+1}) = y_n + h_{n+1} \sum_{i=1}^{s} b_i(\theta) g(t_{n+1}^i, Y_{n+1}^i), \quad 0 \le \theta \le 1, \tag{8.11}$$

or, in K notation, by

$$\eta(t_n + \theta h_{n+1}) = y_n + h_{n+1} \sum_{i=1}^{s} b_i(\theta) K_{n+1}^i, \quad 0 \le \theta \le 1, \tag{8.12}$$

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

$$\begin{aligned} b_i(0) &= 0, \quad i = 1, \dots, s, \\ b_i(1) &= b_i, \quad i = 1, \dots, \nu, \\ b_i(1) &= 0, \quad i = \nu + 1, \dots, s. \end{aligned} \tag{8.13}$$

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

$$Y_{n+1}^i = y_n + h_{n+1} \sum_{j=1}^{s} a_{ij} g(t_{n+1}^j, Y_{n+1}^j), \quad i = \nu + 1, \dots, s, \tag{8.14}$$

or, in K notation, by

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

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

$$A' = \begin{pmatrix} A & 0 \\ (a_{ij})_{i=\nu+1, j=1}^{s,\nu} & (a_{ij})_{i,j=\nu+1}^{s} \end{pmatrix}. \tag{8.15}$$

The overall *continuous Runge–Kutta* methods (8.7), (8.8), (8.9) and (8.7), (8.8), (8.14), (8.11), denoted by $(A, b(\theta), c)$ and $(A', b(\theta))$, respectively, are the continuous extensions of the RK method (A, b, c) and δ will be referred to as the *degree of the interpolant*. In contrast, the method (A, b, c) will be called the *underlying* (*discrete*) RK method.

It is worth remarking that, in general, $\eta(t_n + c_i h_{n+1}) \ne Y_{n+1}^i$. Nevertheless, equality holds for every right-hand side $g(t, y)$ whenever $b_i(c_j) = a_{ji}$, as appears evident by comparing (8.11) and (8.14). So we have

$$\left\{ \eta(t_n + c_i h_{n+1}) = Y_{n+1}^i \quad \forall i \right\} \iff \left\{ b_i(c_j) = a_{ji} \quad \forall i, j \right\}. \tag{8.16}$$

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

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

Definition 8.1. A continuous RK method is called natural if $A = B$ ($A' = B$).

As for the order of RK methods and their interpolants, we have the following definition.

Definition 8.2. We say that the RK method (8.7), (8.8) 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 (8.6) and for all mesh points, we have that

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

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*

$$\begin{aligned} z'_{n+1}(t) &= g(t, z_{n+1}(t)), \quad t_n \leq t \leq t_{n+1}, \\ z_{n+1}(t_n) &= y_n^*. \end{aligned} \tag{8.17}$$

We say that the interpolant (8.9) or (8.11) 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

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

The convergence results are summarized by the following theorem.

Theorem 8.1. Let the RK method (8.7), (8.8) be consistent of *order* p and let the right-hand side function $g(t, y)$ in (8.6) 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,

$$\max_{1 \leq n \leq N} |y(t_n) - y_n| = \mathcal{O}(h^p), \tag{8.18}$$

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

If the interpolant (8.9) or (8.11) has uniform order q, then the continuous RK method (8.7), (8.8), (8.9) or (8.7), (8.8), (8.14), (8.11) is uniformly convergent of order (or, equivalently, has uniform global order) $q' = \min\{p, q+1\}$, that is,

$$\max_{t_0 \leq t \leq t_f} |y(t) - \eta(t)| = \mathcal{O}(h^{q'}). \tag{8.19}$$

We shall often refer to the order of consistency and to the order of convergence of the RK method (8.7), (8.8) as the *discrete order* and the *discrete global order* of the continuous RK method (8.7), (8.8), (8.9) or (8.7), (8.8), (8.14), (8.11).

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

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

Theorem 8.2. If, in addition to the hypotheses of Theorem 8.1, the interpolant is a piecewise polynomial of degree $\delta \geq q$ and the right-hand side function $g(t, y)$ in (8.6) is $C^{\max\{\delta,p\}}$-continuous, then the following convergence, boundedness and unboundedness estimates hold for the derivatives of the global error function:

$$\max_{t_0 \leq t \leq t_f} |y^{(j)}(t) - \eta^{(j)}(t)| = \mathcal{O}(h^{q+1-j}), \quad j = 1, \ldots, \delta, \tag{8.20}$$

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

The estimates (8.19) and (8.20) 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 8.3. Assume that the RK method (8.7), (8.8) 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 8.1. By Theorems 8.2 and 8.3 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 (minimum) number of stages necessary to construct a continuous RK method of uniform order $p - 1$ or even p?

So far, we can give the following upper bound to the uniform order of an interpolant (for both classes).

Theorem 8.4. Assume that the RK method (8.7), (8.8) has a continuous extension $\eta(t)$ given by (8.11). Then its uniform order q cannot exceed s^*, the number of distinct abscissae of the extended RK method represented by (8.15).

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

Table 8.1. 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	$\sum_{i=1}^{\nu} b_i(\theta) c_i^3 = \frac{1}{4}\theta^4$ $\sum_{i,j=1}^{\nu} b_i(\theta) c_i a_{ij} c_j = \frac{1}{8}\theta^4$ $\sum_{i,j=1}^{\nu} b_i(\theta) a_{ij} c_j^2 = \frac{1}{12}\theta^4$ $\sum_{i,j,k=1}^{\nu} b_i(\theta) a_{ij} a_{jk} c_k = \frac{1}{24}\theta^4$

The above result is obvious after observing that formula (8.11) is a continuous quadrature rule based on exactly s^* distinct abscissae.

Since the construction of interpolants of the second class is a rather technical matter, here we confine ourselves to analysing only the interpolants of the first class and, consequently, we shall not give an answer to Question 2. 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.

A general analysis of the uniform order for the continuous extension (8.9) is based on the property that, for any $0 < \theta \leq 1$, it can be viewed as the discrete method $(\frac{A}{\theta}, \frac{b(\theta)}{\theta}, \frac{c}{\theta})$ with step-size 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 8.1.

In order to answer Question 1, each method has to be analysed individually by checking the order conditions. In general, we can give only a

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

partial answer by means of the following theorem, the proof of which does not directly involve the order conditions.

Theorem 8.5. Every RK method (8.7), (8.8) 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 8.6. If an RK method (8.7), (8.8) 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 ensured 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 8.3. We say that an RK method (8.7), (8.8) 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 end-point 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*.

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. Choose ν distinct abscissae $c_1, \ldots, c_\nu \in [0, 1]$ and, in each mesh interval $[t_n, t_{n+1}]$, compute the polynomial $\eta(t)$ of degree $\leq \nu$ satisfying

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

It is easy to check that such methods can be rewritten as a continuous implicit RK method (8.7), (8.9), where

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

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

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

$\ell_i(\xi)$ 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 continuous RK method.

For any choice of the abscissae $c_1, \ldots, c_\nu \in [0,1]$, the collocation method has order $p \geq \nu$ and the uniform order of the interpolant (8.9) is $q = \nu$. Consequently, by Theorem 8.1, 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.

Direct construction of continuous RK methods

So far we have considered continuous extensions of *a priori* given discrete RK methods. Now we consider the other philosophy of constructing directly a continuous RK method, without necessarily starting from a given discrete formula.

As already pointed out in this section, a general analysis of the uniform order for the continuous extensions (8.9) is based on the property that, for any $0 < \theta \leq 1$, it can be viewed as a discrete method $\left(\frac{A}{\theta}, \frac{b(\theta)}{\theta}, \frac{c}{\theta}\right)$ with scaled step-size 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 (see Table 8.1 for conditions up to order $p = 4$).

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

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

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

Table 8.2. The minimum number of stages necessary for an explicit RK method to attain the discrete order p.

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

Table 8.3. The minimum number of stages necessary for a continuous explicit RK method to attain the uniform order q.

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

All the order barriers for explicit methods are summarized in Tables 8.2 and 8.3.

Now we concentrate on continuous explicit RK 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 continuous explicit RK methods is the FSAL (*first same as last*) property. The FSAL property means that the last stage can be re-used 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 re-usable 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 $y_{n+1} = \eta(t_{n+1})$.

8.2. RK methods for DDEs

Once the continuous RK method $\big(A, b(\theta), c\big)$ is chosen, the standard approach for the DDE (8.1) turns out to be

$$\eta(t_n + \theta h_{n+1}) \tag{8.21}$$
$$= y_n + h_{n+1} \sum_{i=1}^{s} b_i(\theta) f\big(t^i_{n+1}, Y^i_{n+1}, \eta\big(t^i_{n+1} - \tau(t^i_{n+1}, Y^i_{n+1})\big)\big),$$

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

for $0 \le \theta \le 1$, and

$$Y^i_{n+1} = y_n + h_{n+1} \sum_{j=1}^{s} a_{ij} f\big(t^j_{n+1}, Y^j_{n+1}, \eta\big(t^j_{n+1} - \tau(t^j_{n+1}, Y^j_{n+1})\big)\big), \quad (8.22)$$

for $i = 1, \dots, s$.

In this section, the method will be called the *RK method for DDEs* or, in short, the *DDE method*, and $(A, b(\theta), c)$ will be referred to as the *underlying continuous RK method*.

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 assume that the abscissae satisfy the constraint

$$0 \le c_i \le 1, \quad i = 1, \dots, s. \quad (8.23)$$

However, even under condition (8.23), 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 occurrence *overlapping*. It is convenient to define the *spurious stage*

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

which, in the case of overlapping, is given by formula (8.21) 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 continuous RK method is explicit. This makes a remarkable difference with respect to the explicit FCRK methods, described in Section 6, which preserve their explicitness even in the case of overlapping.

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 Y notation)

$$\eta(t_n + \theta h_{n+1}) = y_n + h_{n+1} \sum_{i=1}^{s} b_i(\theta) f\big(t^i_{n+1}, Y^i_{n+1}, \tilde{Y}^i_{n+1}\big), \quad 0 \le \theta \le 1, \quad (8.24)$$

$$Y^i_{n+1} = y_n + h_{n+1} \sum_{j=1}^{s} a_{ij} f\big(t^j_{n+1}, Y^j_{n+1}, \tilde{Y}^j_{n+1}\big), \quad i = 1, \dots, s, \quad (8.25)$$

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

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

$$\tilde{Y}^i_{n+1} = y_n + h_{n+1} \sum_{j=1}^{s} b_j(\theta^i_{n+1}) f\big(t^j_{n+1}, Y^j_{n+1}, \tilde{Y}^j_{n+1}\big) \tag{8.26}$$

if the overlapping condition $t^i_{n+1} - \tau(t^i_{n+1}, Y^i_{n+1}) > t_n$ holds, and by the known value

$$\tilde{Y}^i_{n+1} = \eta\big(t^i_{n+1} - \tau(t^i_{n+1}, Y^i_{n+1})\big) \tag{8.27}$$

otherwise.

Note that, whereas the system (8.24), (8.25), (8.27) has to be solved only for the stage values Y^j_{n+1}, $j = 1, \ldots, s$, the system enlarged by (8.26) 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 K notation

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

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

$$\eta(t_n + \theta h_{n+1}) = y_n + h_{n+1} \sum_{i=1}^{s} b_i(\theta) K^i_{n+1}, \quad 0 \le \theta \le 1, \tag{8.28}$$

$$K^i_{n+1} = f\left(t^i_{n+1}, y_n + h_{n+1} \sum_{j=1}^{s} a_{ij} K^j_{n+1}, \tilde{Y}^i_{n+1}\right), \quad i = 1, \ldots, s, \tag{8.29}$$

where

$$\tilde{Y}^i_{n+1} = y_n + h_{n+1} \sum_{j=1}^{s} b_j \left(c_i - \frac{\tau(t^i_{n+1}, y_n + h_{n+1} \sum_{k=1}^{s} a_{ik} K^k_{n+1})}{h_{n+1}}\right) K^j_{n+1} \tag{8.30}$$

if the overlapping condition

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

holds, and

$$\tilde{Y}^i_{n+1} = \eta\left(t^i_{n+1} - \tau\left(t^i_{n+1}, y_n + h_{n+1} \sum_{k=1}^{s} a_{ik} K^k_{n+1}\right)\right) \tag{8.31}$$

otherwise.

Despite it being impossible 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, that allow us to express the spurious stages $\tilde{Y}^i_{n+1}$ in the system (8.25) in terms of the Y^i_{n+1}. This is the case for any *natural* con-

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

tinuous RK method (see Definition 8.1) 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,

$$\eta(t_n + \theta h_{n+1}) = \ell_0(\theta) y_n + \sum_{i=1}^{s} \ell_i(\theta) Y^i_{n+1}, \tag{8.32}$$

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 (8.32) 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 Y and K notation, the method is well-posed for any sufficiently small h_{n+1}, as stated by the following theorem.

Theorem 8.7. (Well-posedness) Assume that the local problem (8.3) possesses a unique solution $w_{n+1}(t)$. Then, for sufficiently small step-size h_{n+1}, equations (8.21)–(8.22) admit a unique solution $\eta(t)$.

As for the convergence analysis of the DDE methods, we have the following result, assuming we are able to compute and include the discontinuity points in the mesh, even in the state-dependent delay case.

Theorem 8.8. (Convergence) Consider the DDE

$$\begin{aligned} y'(t) &= f\big(t, y(t), y\big(t - \tau(t, y(t))\big)\big), \quad t_0 \le t \le t_f, \\ y(t) &= \phi(t), \quad t \le t_0, \end{aligned}$$

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 lying in $[t_0, t_f]$ where the solution $y(t)$ is not at least C^p-continuous. If the underlying continuous RK method has discrete order p and uniform order q, then the DDE method (8.24), (8.25), (8.26), (8.27) has discrete global order and uniform global order $q' = \min\{p, q+1\}$, that is,

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

and

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

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

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

According to Theorem 8.8, if the underlying continuous RK 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 8.1. Under the hypotheses of Theorem 8.8 with $q \ge p-1$, the continuous numerical solution $\eta(t)$ is such that

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

Theorem 8.8 and Corollary 8.1 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 step-size 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 deep analysis of these aspects, we again refer the interested reader to Bellen and Zennaro (2003).

We remark that the DDE method with underlying continuous RK method $\big(A, b(\theta), c\big)$ provides an approximation of the solution map of form (5.3). In particular, we have

$$V(t_n, \eta_{t_n}, h_{n+1}) = \pi K_{n+1},$$

where π is the prolongation operator defined in (6.4) and

$$K_{n+1} = (K^1_{n+1}, \dots, K^\nu_{n+1}) \in (\mathbb{R}^d)^\nu.$$

Consequently, Theorem 8.8 above may also be obtained as a corollary to the general convergence Theorem 5.1.

It is also worth remarking that a DDE method based on a natural continuous RK method $(A, b(\theta), c)$ (see Definition 8.1) provides the same approximation of the solution map as the one provided by the particular implicit FCRK method $(A(\theta), b(\theta), c)$, where

$$a_{ij}(\theta) = b_j(\theta), \quad i, j = 1, \dots, \nu.$$

In fact, for such an FCRK method, all the stage functions Y^i, $i = 1, \dots, \nu$, coincide with the function $\eta(t-t_n)$, $t \in (-\infty, h_{n+1}]$. On the contrary, if the

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

underlying continuous RK method is not natural, then the DDE method does not fall into the class of FCRK methods introduced in Section 6.

Consequently, the well-posedness result expressed by Theorem 8.7 may also be obtained as a corollary to Theorem 6.1 only when the underlying continuous RK method is natural.

RK methods for NDDEs

With respect to the NDDE (8.2), for the choice (8.4), the DDE method (8.22) and (8.21) in Y notation modifies to the following *RK method for NDDEs*:

$$Y^i_{n+1} = y_n + h_{n+1}\sum_{j=1}^{s} a_{ij} f\big(t^j_{n+1}, Y^j_{n+1}, \tilde{Y}^j_{n+1}, \tilde{Z}^j_{n+1}\big), \tag{8.33}$$

for $i = 1, \dots, s$,

$$\eta(t_n + \theta h_{n+1}) = y_n + h_{n+1}\sum_{i=1}^{s} b_i(\theta) f\big(t^i_{n+1}, Y^i_{n+1}, \tilde{Y}^i_{n+1}, \tilde{Z}^i_{n+1}\big), \tag{8.34}$$

for $0 \le \theta \le 1$, and

$$\lambda(t_n + \theta h_{n+1}) = \sum_{i=1}^{s} b'_i(\theta) f\big(t^i_{n+1}, Y^i_{n+1}, \tilde{Y}^i_{n+1}, \tilde{Z}^i_{n+1}\big), \tag{8.35}$$

for $0 \le \theta \le 1$, where

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

Note that, for the arguments $s_j = t^j_{n+1} - \tau(t^j_{n+1}, Y^j_{n+1})$, 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 argument $s_i > t_n$, then the *spurious stages* $\tilde{Y}^i_{n+1}$ and $\tilde{Z}^i_{n+1}$ are unknown, and are given by (8.34) and (8.35) for

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

that is,

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

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

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

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

interpolants $\eta(s)$ and $\eta'(s)$ as computed at the past points

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

for suitable values of m and θ.

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 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}, turn out to depend on K^i_{n+1} only.

Remark 8.2. As in the non-neutral case, for any natural continuous RK method with s distinct abscissae $c_1, \ldots, c_s$ such that $c_i \neq 0,\ i = 1, \ldots, s$, 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 $Y^i_{n+1} (= \eta(t_n + c_i h_{n+1}))$, $i = 1, \ldots, s$, that is,

$$\eta(t_n + \theta h_{n+1}) = \ell_0(\theta) y_n + \sum_{i=1}^{s} \ell_i(\theta) Y^i_{n+1}, \tag{8.36}$$

where $\ell_j,\ j = 0, \ldots, s$ are the Lagrange polynomial coefficients on the nodes $c_0 = 0$ and c_i, $i = 1, \ldots, s$. 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 (8.36) for $\theta = \theta^i_{n+1} = c_i - \tau(t^i_{n+1}, Y^i_{n+1})/h_{n+1}$. Similarly, $\tilde{Z}^i_{n+1} = \lambda(t^i_{n+1} - \tau(t^i_{n+1}, Y^i_{n+1}))$ may be written by using the derivative of (8.36) for $\theta = \theta^i_{n+1}$.

For the choice (8.5), the RK method for NDDEs (in Y notation) is given by (8.33), (8.34) along with

$$\lambda(t_n + \theta h_{n+1}) = \sum_{i=0}^{s^*} \ell_i(\theta) f\big(\bar{t}^i_{n+1}, U^i_{n+1}, \tilde{U}^i_{n+1}, \tilde{V}^i_{n+1}\big), \quad 0 \leq \theta \leq 1, \tag{8.37}$$

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

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

there are additional values

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

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

Note that, according to the argument $s_j = t_{n+1}^j - \tau(t_{n+1}^j, Y_{n+1}^j)$, the values $\eta(s_j)$ and $\lambda(s_j)$ may or may not be known. If $s_j > t_n$, then $\tilde{Y}_{n+1}^j$ and $\tilde{Z}_{n+1}^j$ are unknown and must be computed by (8.34) and (8.37), respectively. In particular, for the application of (8.37) in the current interval, U_{n+1}^j, $\tilde{U}_{n+1}^j$ and $\tilde{V}_{n+1}^j$ need to be known. Here the U_{n+1}^js are certainly unknown, whereas knowledge of $\tilde{U}_{n+1}^j$ and $\tilde{V}_{n+1}^j$ depends on the location of the further argument $\bar{t}_{n+1}^j - \tau(\bar{t}_{n+1}^j, U_{n+1}^j)$.

Summarizing, if for some index j some of the arguments are $> t_n$, then the relevant *spurious stages* $\tilde{Y}_{n+1}^j$, $\tilde{Z}_{n+1}^j$, U_{n+1}^j, $\tilde{U}_{n+1}^j$ or $\tilde{V}_{n+1}^j$ are unknown and are given by (8.34) and (8.37) for suitable values of θ. More precisely,

$$\tilde{Y}_{n+1}^j = \eta(t_n + \theta_{n+1}^j h_{n+1}) \quad \text{and} \quad \tilde{Z}_{n+1}^j = \lambda(t_n + \theta_{n+1}^j h_{n+1})$$

with $\theta_{n+1}^j = c_j - \tau(t_{n+1}^j, Y_{n+1}^j)/h_{n+1}$,

$$U_{n+1}^j = \eta(t_n + \bar{c}^j h_{n+1}),$$
$$\tilde{U}_{n+1}^j = \eta(t_n + \bar{\theta}_{n+1}^j h_{n+1}) \quad \text{and} \quad \tilde{V}_{n+1}^j = \lambda(t_n + \bar{\theta}_{n+1}^j h_{n+1})$$

with $\bar{\theta}_{n+1}^j = \bar{c}_j - \tau(\bar{t}_{n+1}^j, U_{n+1}^j)/h_{n+1}$.

The dimension of the system may still be reduced by using K notation but, unlike option (8.4), as well as the K values

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

we have the additional values

$$H_{n+1}^j = f\big(\bar{t}_{n+1}^j, U_{n+1}^j, \tilde{U}_{n+1}^j, \tilde{V}_{n+1}^j\big), \quad j = 0, \dots, s^*.$$

Remark 8.3. As with option (8.5), the number of unknowns in the system to be solved at each step may be reduced. In fact, if the underlying continuous RK method is natural and if the interpolation formula (8.37) 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 only Y_{n+1}^j, $\tilde{Y}_{n+1}^j$ and $\tilde{Z}_{n+1}^j$ in Y notation, and to merely

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

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

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

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

by (8.37) 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 $c_i \neq 0$, $i = 1, \ldots, s$, as in the non-neutral case, we can express each $\tilde{Y}_{n+1}^j$ in terms of the Y_{n+1}^js 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 can the RK method reduce to just the Y values.

The convergence result extending Theorem 8.8 may be stated as follows.

Theorem 8.9. Consider the state-dependent NDDE (8.2), where the right-hand side $f(t, y, x, w)$ is C^p-continuous in $[t_0, t_f] \times \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d$, the delay $\tau(t, y)$ is 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 lying in $[t_0, t_f]$ where the solution $y(t)$ is not at least C^p-continuous. If the underlying continuous RK method $\big(A, b(\theta), c\big)$ 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,

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

and

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

where $h = \max_{1 \leq n \leq N} h_n$. In particular, if $\lambda(t)$ is given by the option (8.4), then $r = q - 1$ and, hence, $q' = q$.

Note that, for the option (8.5), $\lambda(t)$ is given by (8.37) and the interpolation operator $\mathcal{P}$ has order $r = s^*$. Therefore, on the basis of Theorem 8.9, 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 (8.5), along with the conditions in Remark 8.3, preferable to (8.4).

Finally, it is worth remarking that, when the option (8.4) is adopted, the RK method for NDDEs yields an approximation of the solution of the form (5.3). On the contrary, this is not the case when the other option (8.5) is used.

Therefore, the above convergence result can also be obtained as a corollary to the general convergence Theorem 5.2 only if option (8.4) is adopted.

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

9. Implementation issues in the standard approach

We have seen in the previous section that two points are important for an efficient and accurate implementation of the standard approach, namely location of the breaking points, if any, as defined in Section 3 and detection of possible overlapping. From a practical point of view, it would be important to answer *a priori* the following two questions which are particularly difficult when the delay is state-dependent:

(Q_1) Whether or not, for the local problem (8.3), 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 $[t_n, t_{n+1}]$, so as to make the right choice between (8.26) and (8.27) (or between (8.30) and (8.31)) for the computation of the $\tilde{Y}^i_{n+1}$s.

(Q_2) Whether or not, for the step-size h_{n+1}, the current interval $[t_n, t_{n+1}]$ includes some breaking point ξ and, more specifically, how to tune the step-size h_{n+1} in order to have $t_{n+1} = \xi$, as we would like to have in the proximity of ξ, as required by Theorems 8.8 and 8.9.

As far as question (Q_1) is concerned, overlapping can be avoided for a sufficiently small step-size by assuming that:

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

Moreover, when the delay is actually state-dependent, it is often welcome to be able to assume the stronger condition (already considered in Section 3):

(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 $z \in \mathbb{R}^d$.

In fact, (H_1^*) prevents the state-dependent delay from vanishing even if it is computed in a perturbation of the true local solution $w_{n+1}(t)$ of (8.3).

Under the hypothesis (H_1^*) on the delay, for sufficiently small step-size, namely $h_{n+1} = t_{n+1} - t_n \leq \tau_0$, the function $\eta(s)$ is known for every $s = t - \tau(t, z)$ with $t \in [t_n, t_{n+1}]$ and for all $z \in \mathbb{R}^d$. Therefore, overlapping is avoided for any approximation of the local solution $w_{n+1}(t)$.

On the contrary, if $h_{n+1} > \tau_0$, it might well be that

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

for some $t \in (t_n, t_{n+1}]$ and, consequently, overlapping could occur.

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 interval $(t_n, t_{n+1}]$ includes ξ. This occurrence leads to some complications in the case of state-dependent delays. In fact, in this case, we cannot choose between (8.26) or (8.27) ((8.30) or (8.31)) *a priori*, and the choice

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

may vary during the computation of the $\tilde{Y}^i_{n+1}$s. This point will be deeply investigated in Section 11, as well as considering the neutral case.

As for question (Q_2), the location of breaking points has been discussed by various authors from both the theoretical and implementational perspectives. Two approaches have been pursued in the literature. They are based, respectively, on discontinuity tracking, that is, on a direct computation of discontinuities from the deviated arguments (see, *e.g.*, Willé and Baker (1992)) and on defect control (see, *e.g.*, Enright, Jackson, Nørsett and Thomsen (1988) and Shampine and Thompson (2000)).

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

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

(see (3.6)) and to include them as mesh points. It is just worth mentioning that for state-dependent delays the task appears very hard to accomplish. How to do this, and how to achieve the accuracy necessary for preserving the order of the overall integration procedure, is presented later in this section. Although expensive, this strategy appears the most robust.

The second approach, relying on step-size control, gives up tracking the discontinuities, which are instead assumed to be automatically included in the mesh by suitable variable step-size 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 step-sizes in the neighbourhood of a low-order discontinuity point ξ.

9.1. Tracking the breaking points

An accurate tracking of breaking points is important in order to compute and automatically insert them into the mesh of integration. We shall discuss this topic here in the context of DDEs and, in Section 10, in the even more challenging context of NDDEs.

Although the approach can be extended to any DDE method, here we consider collocation methods with ν distinct abscissae $c_1, \ldots, c_\nu$ such that $c_i \neq 0$, $i = 1, \ldots, \nu$, whose continuous approximation at the nth step is expressed in the form

$$\eta(t_n + \theta h_{n+1}) = \sum_{i=0}^{\nu} \ell_i(\theta) Y^i_{n+1}, \quad \theta \in [0, 1],$$

where $Y^0_n = y_n$ and the abscissa $c_0 = 0$ is added to the other abscissae of the method (see (8.32)). The algorithms we are going to describe here are extensively discussed in Guglielmi and Hairer (2008) and implemented in the code Radar5 by Guglielmi and Hairer (2001). They are mainly intended

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

for state-dependent delays, but they may be used in general because they are designed with the aim of computing only those breaking points which are important in terms of the required accuracy.

To compute the set $\mathcal{B}$ of breaking points recursively, we start by initializing $\mathcal{B} = \{t_0\}$. Then the iteration step consists in finding the zeros of the function

$$d^*_\zeta(t) = \alpha\big(t, y(t)\big) - \zeta, \tag{9.1}$$

where $\zeta \in \mathcal{B}$ is a previous breaking point. However, since $y(t)$ cannot be found exactly, we have to consider a suitable approximation $\eta(t)$, *e.g.*, the continuous extension of the collocation method, and solve the approximate equation

$$d_\zeta(t) = \alpha\big(t, \eta(t)\big) - \zeta. \tag{9.2}$$

The solution ξ of this equation is added to $\mathcal{B}$.

The novel method we present is split into two phases: a first one, where the presence of a breaking point is detected, and a second one, where the breaking point is actually computed.

Detection and accurate computation of breaking points

For a given step-size h, the first phase consists in checking the possible presence of a breaking point in the interval $[t_n, t_n + h]$. To this end we consider the continuous extension computed at the previous accepted step,

$$\hat\eta(t_{n-1} + \theta h_n) = \sum_{i=0}^{\nu} \ell_i(\theta) Y_n^i, \quad \theta \geq 1,$$

to be used for extrapolation in the current step.

After setting h as a predicted new step-size, we look for zeros of the functions

$$d_\zeta(\theta) = \alpha\big(t_{n-1} + \theta h_n, \hat\eta(t_{n-1} + \theta h_n)\big) - \zeta, \quad \theta \in [1, 1 + h/h_n],$$

for all previously computed breaking points $\zeta \in \mathcal{B}$. The presence of a new breaking point is guessed if $d_\zeta(t_n) \cdot d_\zeta(t_n + h) < 0$ for some $\zeta \in \mathcal{B}$. This idea is related to that used by Enright and Hayashi (1997) in their explicit solver.

Let $\xi^{[0]}$ be the *detected* breaking point, that is, the solution of the equation

$$\alpha\big(\xi^{[0]}, \hat\eta(\xi^{[0]})\big) - \zeta = 0. \tag{9.3}$$

In general, $\xi^{(0)}$ provides a poor approximation to the exact breaking point due to the fact that we are making use of an extrapolation of the collocation polynomial $\hat\eta$. Note that a better approximation of the solution in $[t_n, t_n + h]$ could reveal the absence of breaking points, *i.e.*, of solutions for (9.3) in that interval.

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

In any case, once a breaking point is detected inside the interval $[t_n, t_n+h]$, we assume it actually exists and, hence, we try to compute it accurately in order to preserve the high order and accuracy of the numerical method. The heuristic, which we shall explain theoretically and illustrate experimentally, is that of coupling the RK equations and the equation for the breaking point (9.2).

Therefore, we consider the system of the RK equations (see (8.24)–(8.27)) coupled with (9.2), that is,

$$0 = \alpha\big(t_n + h_{n+1}, \eta(t_n + h_{n+1})\big) - \zeta, \tag{9.4}$$

$$Y^i_{n+1} = y_n + h_{n+1} \sum_{j=1}^{\nu} a_{ij} f\big(t_n + c_j h_{n+1}, Y^j_{n+1}, \tilde{Y}^j_{n+1}\big), \quad i = 1, \dots, \nu, \tag{9.5}$$

which is solved with respect to both the stages $Y^1_{n+1}, \dots, Y^\nu_{n+1}$ and the step-size h_{n+1}.

If equations (9.4)–(9.5) are solved successfully, the point $\xi = t_n + h_{n+1}$ is inserted into the set of computed breaking points $\mathcal{B}$.

Since we are considering collocation methods, we have to solve (9.4)–(9.5) by a Newton process, especially if the problem is stiff. However, instead of applying it to the whole system, it is convenient to split the problem in order to take advantage of its structure.

Solving (9.4)–(9.5) *by an iterative scheme*

For given h_{n+1} the system (9.5) is usually solved by the well-known simplified Newton iteration that exploits the structure of the Jacobian (see, *e.g.*, Hairer and Wanner (1996)). In order to solve (9.4)–(9.5) efficiently with respect to the unknowns $\{Y^i_{n+1}\}$ and h_{n+1}, it would be important not to lose such a structure (see Guglielmi and Hairer (2001, 2008)).

Aiming to preserve the block-diagonal structure, it is possible to solve the system (9.4)–(9.5) in an iterative way. In particular, we denote by $Y^{j\,[k]}_{n+1}$, $j = 1, \dots, \nu$, and $h^{[k]}_{n+1}$ the stage values and the step-size at the kth iteration of the iterative process (this means that $t_n + h^{[k]}_{n+1}$ gives the current approximation of the breaking point).

Starting with $h^{[0]}_{n+1} = \xi^{[0]} - t_n$, where $\xi^{[0]}$ is the approximation to the breaking point obtained solving (9.3) in the detection phase, and using some initial approximation to the stage values obtained, for example, by extrapolation from the previous step, we consider the following two-step iteration.

(I1) Solve equation (9.4) with respect to the unknown $h^{[k+1]}_{n+1}$, *i.e.*,

$$0 = \alpha\big(t_n + h^{[k+1]}_{n+1}, \eta^{[k]}(t_n + h^{[k+1]}_{n+1})\big) - \zeta, \tag{9.6}$$

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

with fixed stage values $\{Y_{n+1}^{j\,[k]}\}_{j=1}^{\nu}$, that is, with a fixed vector-valued polynomial $\eta^{[k]}$ given by

$$\eta^{[k]}\big(t_n + \theta h_{n+1}^{[k]}\big) = \sum_{i=0}^{\nu} \ell_i(\theta) Y_{n+1}^{i\,[k]}, \quad \theta \geq 0.$$

(I2) Solve the system (9.5) with respect to the unknowns $\{Y_{n+1}^{j\,[k+1]}\}_{j=1}^{\nu}$ with fixed step-size $h_{n+1}^{[k+1]}$ by means of a simplified Newton iteration, *i.e.*,

$$Y_{n+1}^{i\,[k+1]} = y_n + h_{n+1}^{[k+1]} \sum_{j=1}^{\nu} a_{ij} f\big(t_n + c_j h_{n+1}^{[k+1]}, Y_{n+1}^{j\,[k+1]}, \tilde{Y}_{n+1}^{j\,[k+1]}\big), \quad \text{for} \quad i = 1, \ldots, \nu, \quad (9.7)$$

where $\tilde{Y}_{n+1}^{j\,[k+1]} = \eta\big(\alpha(t_n + c_j h_{n+1}^{[k+1]}, Y_{n+1}^{j\,[k+1]})\big)$.

It is clear that, assuming that the iterative scheme converges, its efficiency depends on the speed of convergence. The following lemma shows that this iterative method converges (see Guglielmi and Hairer (2008) for the proof). Although linear, the convergence turns out to be fast, since the convergence ratio depends on the νth power of the step-size.

Lemma 9.1. Assume that the solution $h_{n+1}^{[k+1]}$ of (9.6) is simple for all k and that the simplified Newton iteration applied to (9.7) converges in a suitable neighbourhood of h_{n+1}. Then,

$$\big|h_{n+1}^{[k+1]} - h_{n+1}\big| \leq C \cdot \big(h_{n+1}^{[k]}\big)^{\nu} \cdot \big|h_{n+1}^{[k]} - h_{n+1}\big|,$$

where h_{n+1} is the exact solution of (9.4)–(9.5) and C a suitable constant.

In practice we have experienced that the required accuracy is achieved after very few iterations. This implies that the cost is only slightly greater than that of the standard step (for which no breaking point is detected), since it essentially consists of solving (9.5) a small number of times.

In fact, in the code Radar5, which implements the described procedure, the possible presence of a breaking point is not checked at every step but, instead, the function $d_\zeta(t)$ is monitored only in the following cases:

(i) if the Newton process does not converge,
(ii) if the estimated error is not under the given required tolerance,
(iii) if the estimated error increases with respect to the previous step of a factor larger than a prescribed value (the default value is 5).

Unlike most of the previous strategies, in this way one computes only those breaking points that are relevant to the required accuracy.

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

Other authors have considered techniques for approximating the breaking points (see, *e.g.*, Feldstein and Neves (1984) and Hauber (1997)). The algorithm of Feldstein and Neves (1984) checks at all steps whether

$$d_\zeta(t_n + \theta h_{n+1}) = \alpha\big(t_n + \theta h_{n+1}, \eta(t_n + \theta h_{n+1})\big) - \zeta \tag{9.8}$$

changes sign for some $\zeta \in \mathcal{B}$. In such a case, the zero of d_ζ is computed and the new breaking point is inserted in $\mathcal{B}$. Note that the use of η to compute the new breaking point is not reliable in general. In fact, in the current integration interval the solution would not be smooth and consequently $\eta(t)$ would be a bad approximation of the solution. A modification of this idea was considered by Hauber (1997), who proposed extrapolating the continuous output of the preceding step, *i.e.*, to replace (9.8) by

$$d_\zeta(t_{n-1} + \theta h_n) = \alpha\big(t_{n-1} + \theta h_n, \hat{\eta}(t_{n-1} + \theta h_n)\big) - \zeta, \quad \theta > 1.$$

Although this idea allows us to overcome the problem due to the lack of smoothness of the solution, using the collocation polynomial computed in the interval $[t_{n-1}, t_n]$ for $t > t_n$ may also determine an inaccurate computation. The work of Enright and Hayashi (1997) also uses this kind of extrapolation to cross breaking points.

The basic idea presented here is related to the fact that, in the algorithm which computes the RK step, the step-size is not fixed but variable. This allows for a more accurate computation of breaking points and for an improvement of the convergence theorem, as illustrated below by Theorem 9.1.

9.2. Convergence and accuracy of breaking points

This section is devoted to illustrating the theoretical aspects associated with the solution of (9.4)–(9.5).

Accuracy of the computed breaking points

Concerning the coupling of the Runge–Kutta equations and the equation for the breaking point (9.4)–(9.5), the following error bound is obtained (see Guglielmi and Hairer (2008) for the proof).

Theorem 9.1. Let $y(t)$ be the solution of (8.1) and let ζ^* and ξ^* be exact breaking points of the problem such that $\alpha\big(\xi^*, y(\xi^*)\big) = \zeta^*$. Furthermore, let ζ be an approximation of ζ^*. If

$$\frac{\mathrm{d}}{\mathrm{d}t}\big(\alpha(t, y(t))\big)\big|_{t=\xi^*} \neq 0, \tag{9.9}$$

then the computed breaking point $\xi = t_n + h_{n+1}$, obtained by solving (9.4)–(9.5), satisfies the error estimate

$$|\xi - \xi^*| \leq C\big(\|y_{n+1} - y(t_{n+1})\| + |\zeta - \zeta^*|\big)$$

for some constant $C > 0$.

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

This means that the breaking points are computed to the same order as the numerical solution at grid points.

Although the order is the same as that obtained using an extrapolation of the dense output computed in the previous interval (see (9.3)), the error constant is expected to be much smaller. Usually, in fact, stiffly accurate collocation methods such as Radau-IIA methods exhibit an error at mesh points which is much smaller than the uniform error in the integration interval. Therefore, making breaking points to coincide with mesh points should improve accuracy, as is actually confirmed by several numerical experiments.

Convergence

By means of Theorem 9.1 above, it is easy to refine Theorem 8.8 and to avoid the assumption that the mesh contains all the exact breaking points where the solution $y(t)$ is not at least C^p-continuous.

Theorem 9.2. (Convergence) Consider the DDE

$$y'(t) = f\big(t, y(t), y\big(\alpha(t, y(t))\big)\big), \quad t_0 \le t \le t_f,$$
$$y(t) = \phi(t), \quad t \le t_0,$$

with simple breaking points (*i.e.*, (9.9) holds). Assume that the hypotheses of Theorem 8.8 hold except that, instead of the exact breaking points, those obtained by solving (9.4)–(9.5) are included in the mesh.

If the underlying collocation method has discrete order p (and uniform order $q = \nu$), then the DDE method (8.24)–(8.27) has discrete global order and uniform global order $q' = \min\{p, \nu + 1\}$.

Table 9.1. Numerical results for equation (9.10), where FE stands for the number of function evaluations, ERR for the error at the final point and ERRBP for an average error in the computation of the breaking points.

	Radar5: old version		Radar5: new version		
$-\log$ (tol)	FE	ERR	FE	ERR	ERRBP
2	94	$0.41\,10^{-1}$	97	$0.13\,10^{-3}$	$0.55\,10^{-4}$
4	146	$0.55\,10^{-3}$	147	$0.14\,10^{-5}$	$0.63\,10^{-6}$
6	247	$0.40\,10^{-3}$	198	$0.32\,10^{-7}$	$0.13\,10^{-7}$
8	443	$0.15\,10^{-5}$	276	$0.60\,10^{-9}$	$0.25\,10^{-9}$
10	733	$0.85\,10^{-7}$	490	$0.52\,10^{-10}$	$0.21\,10^{-10}$
12	1622	$0.85\,10^{-9}$	932	$0.46\,10^{-12}$	$0.20\,10^{-12}$

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

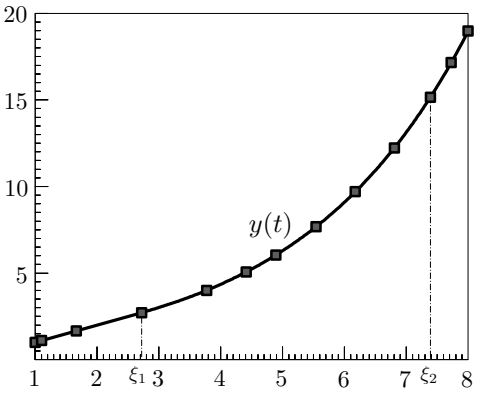


Figure 9.1. The solution of (9.10) and its numerical approximation.

Example 9.1. Consider the equation (see Neves (1975) and Paul (1994))

$$\begin{aligned} y'(t) &= \frac{y(t)\, y(\log(y(t)))}{t}, \quad 1 \le t \le 8, \\ y(t) &= 1, \quad t \le 1. \end{aligned} \tag{9.10}$$

The exact solution is

$$y(t) = \begin{cases} 1, & t < \xi_0^*, \\ t, & \xi_0^* \le t < \xi_1^*, \\ \mathrm{e}^{t/\mathrm{e}}, & \xi_1^* \le t < \xi_2^*, \\ \dots & \dots \ \dots \ \dots \end{cases}$$

and the breaking points are $\xi_0^* = 1$ (of order 0), $\xi_1^* = \mathrm{e}$ (of order 1), $\xi_2^* = \mathrm{e}^2$ (of order 2), *etc.*

Table 9.1 shows the numerical integration of (9.10) by the code Radar5. In version 1.1 of the code the breaking points were not computed explicitly, but only implicitly, through the error control, and the step-size was simply driven by error estimates. In the current version, 2.1, an explicit computation of the breaking points is implemented, according to the algorithm previously described, which solves (9.4)–(9.5).

In Figure 9.1 we show the exact solution and its numerical approximation, obtained with relative and absolute error tolerances per step, $R_{\mathrm{tol}} = A_{\mathrm{tol}} = 5 \cdot 10^{-5}$. The computed breaking points are

$$\begin{aligned} \xi_1 &= 2.71828358623074\cdots, \\ \xi_2 &= 7.38905155206748\cdots. \end{aligned}$$

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

The first numerical breaking point is essentially exact because the solution is linear in the first interval. The second numerical breaking point has an error $|\xi_2 - \xi_2^*| = 4.5469 \cdot 10^{-6}$. The number of accepted steps is 13 and the number of rejected steps is 2. The relative error at the final point is $\mathrm{ERR} \approx 1.001 \cdot 10^{-4}$.

10. Neutral problems with state-dependent delays

Now we consider neutral problems of the form (8.2). Without loss of generality, we focus our attention on systems in autonomous form

$$\begin{aligned} y'(t) &= f\big(y(t),\, y\big(\alpha(y(t))\big),\, y'\big(\alpha(y(t))\big)\big), \quad t_0 \le t \le t_f, \\ y(t) &= \phi(t), \quad t \le t_0, \end{aligned} \tag{10.1}$$

where, as usual, $\alpha(y(t))$ denotes the deviated argument.

In general, at the initial point t_0 the right-hand derivative

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

is different from the left-hand derivative $\phi'(t_0)$, *i.e.*, it does not satisfy the splicing condition (4.6) assumed in Theorem 4.5. This irregularity at t_0 is propagated by the deviated argument $\alpha(y(t))$ to further breaking points, where the first derivative of the solution is not continuous.

Due to such jump discontinuities in the first derivative of the solution, because of a breaking point, problem (10.1) has to be considered as a discontinuous differential equation (see, *e.g.*, Filippov (1964, 1988)). Moreover, the delay being state-dependent, existence and uniqueness of a classical solution are no longer assured, independently of the regularity of f. Therefore, the solution might either terminate, or even bifurcate, in the presence of a breaking point. This leads us in a natural way to consider *weak* (or *generalized*) solutions, which may allow the integrator to prolong the solution beyond those breaking points where the classical solution ceases to exist. To this end, we consider some possible *regularizations*, and define weak solutions to be the limits of the solutions of the regularized problems as the regularization parameters tend to zero.

As in Baker and Paul (2006) and Bellen and Guglielmi (2009), we give the following definition where, with respect to Definition 1.1, the value of y' is assigned at any point of the integration interval. As before, we let $\mathcal{B}$ denote the set of breaking points.

Definition 10.1. We say that a function $y(t)$ is a solution to problem (10.1) in $[t_0, t_f]$ if:

(i) it is continuous on $[t_0, t_f]$;

(ii) it is continuously differentiable in $[t_0, t_f] \setminus \mathcal{B}$;

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

(iii) it satisfies (10.1) in $[t_0, t_f] \setminus \mathcal{B}$;

(iv) at those breaking points $\xi \in \mathcal{B}$ where (10.1) is not satisfied, we have

$$\lim_{t \searrow \xi} y'(t) = f\big(y(\xi), y\big(\alpha(y(\xi))\big), z\big),$$

where

$$z = \lim_{t \searrow \xi} y'\big(\alpha(y(t))\big).$$

Thus $y'(t)$ is the usual two-sided derivative for all $t \in [t_0, t_f]$, except for the breaking points ξ where (10.1) is not satisfied. At such points we take it to be the one-sided right derivative $\lim_{t \searrow \xi} y'(t)$.

In general, for neutral equations there is no smoothing effect. Therefore, breaking points of order zero may be propagated throughout the integration interval.

Since Theorem 4.5 is not applicable, in order to study the existence of a solution, we make the assumption that the set $\mathcal{B}$ is finite.

10.1. Neutral problems as discontinuous differential equations

Let $\xi > \zeta$ be a breaking point of order zero, that is,

$$\alpha(y(\xi)) = \zeta,$$

where ζ is a previous breaking point, the ancestor of ξ, where the derivative of the solution has a jump discontinuity.

Let

$$\begin{aligned} x^+(s) &= y(s) \quad \text{for} \quad s \geq \zeta, \\ x^-(s) &= y(s) \quad \text{for} \quad s < \zeta, \end{aligned} \tag{10.2}$$

and let $x'^+(s)$ and $x'^-(s)$ be the corresponding derivatives.

Since we assumed that the set $\mathcal{B}$ is finite, they are defined and smooth in a suitable neighbourhood of ζ. Then we can locally write problem (10.1) in the form

$$y'(t) = h(y(t)) = \begin{cases} h^+(y(t)) & \text{if} \quad \alpha(y(t)) > \zeta, \\ h^-(y(t)) & \text{if} \quad \alpha(y(t)) < \zeta, \\ h^+(y(\xi)) & \text{if} \quad t = \xi \quad \text{and} \quad \alpha(y(t)) \nearrow \zeta, \\ h^-(y(\xi)) & \text{if} \quad t = \xi \quad \text{and} \quad \alpha(y(t)) \searrow \zeta, \end{cases} \tag{10.3}$$

where

$$\begin{aligned} h^+(y(t)) &= f\big(y(t),\, x^+\big(\alpha(y(t))\big),\, x'^+\big(\alpha(y(t))\big)\big), \\ h^-(y(t)) &= f\big(y(t),\, x^-\big(\alpha(y(t))\big),\, x'^-\big(\alpha(y(t))\big)\big). \end{aligned} \tag{10.4}$$

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

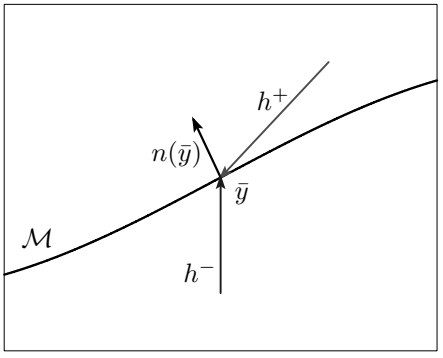


Figure 10.1. The vector fields h^+ and h^- and the normal n to the manifold oriented towards the region $\{y \mid g(y) > 0\}$.

Note that (10.3) is a differential equation with discontinuous right-hand side. In fact, the discontinuity occurs at at $t = \zeta$, where

$$x'^{+}\big(\alpha(y(\xi))\big) \neq x'^{-}\big(\alpha(y(\xi))\big).$$

On the contrary, the solution is continuous, *i.e.*, $x^{+}\big(\alpha(y(\xi))\big) = x^{-}\big(\alpha(y(\xi))\big)$.

Let us introduce the so-called *switching function*

$$g(y(t)) = \alpha(y(t)) - \zeta, \tag{10.5}$$

whose zeros identify the instants when the right-hand side of (10.4) switches from h^+ to h^- or *vice versa*.

If we introduce the manifold

$$\mathcal{M} = \{y \mid g(y) = 0\},$$

which separates the two regions where the vector field of the differential equation is smooth, we have the situation illustrated in Figure 10.1.

Consider $\bar{y} \in \mathcal{M}$, that is, $g(\bar{y}) = 0$, and assume that $\bar{t}$ is such that $y(\bar{t}) = \bar{y}$. Then let ∇ denote the gradient with respect to y and set

$$n(\bar{y}) = \frac{\nabla g(\bar{y})}{\|\nabla g(\bar{y})\|} \quad \text{if} \quad \nabla g(\bar{y}) \neq 0.$$

Finally, consider the quantities

$$\big\langle n(\bar{y}), h^+(\bar{y})\big\rangle \quad \text{and} \quad \big\langle n(\bar{y}), h^-(\bar{y})\big\rangle. \tag{10.6}$$

If the conditions

$$\begin{aligned} \big\langle n(\bar{y}), h^+(\bar{y})\big\rangle &< 0, \\ \big\langle n(\bar{y}), h^-(\bar{y})\big\rangle &> 0, \end{aligned} \tag{10.7}$$

occur, then the vector fields h^+ and h^- have a normal direction with respect

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

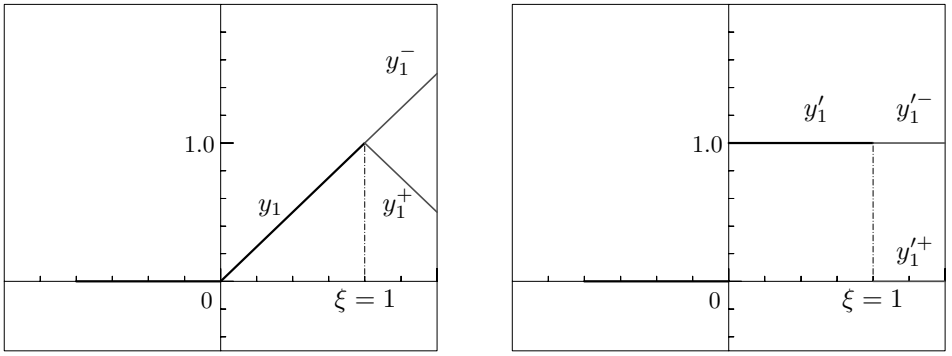


Figure 10.2. The solution of (10.9) terminates at $\xi = 1$.

to the manifold at $\bar{y}$ which is oriented towards the manifold itself. This means that the classical solution to (10.3) ceases to exist. This situation is illustrated in Figure 10.1.

On the contrary, in the two cases when

$$\big\langle n(\bar{y}), h^+(\bar{y})\big\rangle \cdot \big\langle n(\bar{y}), h^-(\bar{y})\big\rangle > 0, \tag{10.8}$$

a unique classical solution keeps on existing in a right neighbourhood of $\bar{t}$.

Finally, if $\big\langle n(\bar{y}), h^+(\bar{y})\big\rangle > 0$, $\big\langle n(\bar{y}), h^-(\bar{y})\big\rangle = 0$ and $\big\langle n(y), h^-(y)\big\rangle \le 0$ in a neighbourhood of $\bar{y}$ (or in the specular case $\big\langle n(\bar{y}), h^-(\bar{y})\big\rangle < 0$, $\big\langle n(\bar{y}), h^+(\bar{y})\big\rangle = 0$ and $\big\langle n(y), h^+(y)\big\rangle \ge 0$ in a neighbourhood of $\bar{y}$) two solutions are admissible, so that uniqueness is lost.

Example 10.1. Let us consider the system

$$\begin{aligned} y_1'(t) &= 1 - 2\, y_1'(y_1(t) - 1), \\ y_2'(t) &= 2 - \frac{1}{2} y_2'(y_1(t) - 1), \end{aligned} \tag{10.9}$$

with initial data $y_1(t) = y_2(t) \equiv 0$ for $t \le 0$.

The solution exists until $t = 1$ and is given by $y_1(t) = t, y_2(t) = 2\,t$. Then it terminates at $t = \xi = 1$.

We have

$$\bar{y} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad \nabla g(\bar{y}) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad h^+(\bar{y}) = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \quad h^-(\bar{y}) = \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$$

We see that conditions (10.7) are satisfied. In fact,

$$\begin{aligned} \big\langle \nabla g(\bar{y}), h^+(\bar{y})\big\rangle &= -1, \\ \big\langle \nabla g(\bar{y}), h^-(\bar{y})\big\rangle &= 1. \end{aligned}$$

This implies termination of the classical solution.

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

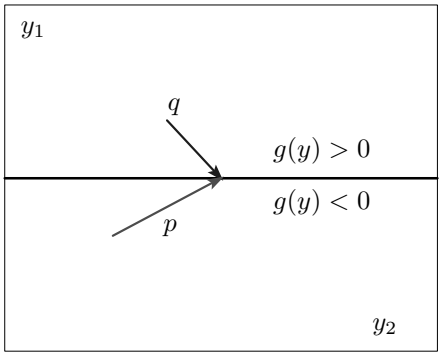


Figure 10.3. Problem (10.9) reformulated as a discontinuous differential equation.

In other words, as illustrated in Figure 10.2, we can explain termination by considering, for $t \in [1, 1+\delta]$ with δ sufficiently small, the pair of differential equations

$$\begin{aligned} y'^{+}(t) &= h^{+}(y^{+}(t)), \\ y'^{-}(t) &= h^{-}(y^{-}(t)), \end{aligned} \tag{10.10}$$

where

$$h^{+}\big(y^{+}(t)\big) = q = \begin{pmatrix} -1 \\ 1 \end{pmatrix} \quad \text{and} \quad h^{-}\big(y^{-}(t)\big) = p = \begin{pmatrix} 1 \\ 2 \end{pmatrix}. \tag{10.11}$$

We have that $y^{+}(t)$ is a local solution of the considered neutral problem (10.9) if $y_1^{+}(t) - 1 > 0$ for $t \in (1, 1+\delta]$ and that $y^{-}(t)$ is a local solution of (10.9) if $y_1^{-}(t) - 1 < 0$ for $t \in (1, 1+\delta]$. Since $y_1^{+}(t) = 2 - t$ and $y_1^{-}(t) = t$, none of the previous conditions is fulfilled (see Figure 10.2).

Since $y'(y_1(s) - 1) = H(y_1(s) - 1)$ until $y_1(s) \le 2$, H being the Heaviside function, the problem (10.9) can be reformulated, at least locally, as the discontinuous differential equation

$$y'(t) = \begin{cases} p & \text{if } \ g(y) < 0, \\ q & \text{if } \ g(y) > 0, \end{cases} \tag{10.12}$$

where $g(y) = y_1 - 1$ (see Figure 10.3).

This is the case of a system of ODEs with a vector field which is discontinuous on a linear manifold $\mathcal{M} = \{y \mid g(y) = 0\} = \{y \mid y_1 = 1\}$ of codimension 1. Furthermore, the vector field is constant on the two half-spaces separated by the manifold.

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

10.2. Regularization by a time average of the discontinuous vector field

Let ξ be a termination point. Following Fusco and Guglielmi (2009), we consider for $t > \xi$ the regularized problem

$$y'^{\varepsilon}(t) = \frac{1}{\varepsilon}\int_{t-\varepsilon}^{t} \big[H\big(g(y^{\varepsilon}(s))\big)h^{+}(y^{\varepsilon}(s)) + \big(1 - H\big(g(y^{\varepsilon}(s))\big)\big)h^{-}(y^{\varepsilon}(s))\big]\,\mathrm{d}s, \tag{10.13}$$

where H denotes the Heaviside function and $y^{\varepsilon}(t) = \phi(t)$, $t \le \xi$, ϕ being a C^1-function. Our goal is that of studying the existence of a solution for sufficiently small $\varepsilon > 0$ and that of investigating the limit behaviour of such solutions as $\varepsilon \longrightarrow 0^{+}$.

Theorem 10.1. (Existence) Let g, h^{-}, h^{+} be smooth functions. Then there exists $\varepsilon_0 > 0$ such that $\forall \varepsilon \in (0, \varepsilon_0)$, there exist $T > 0$ and $C > 0$, independent of ε, such that the problem (10.13) has a C^1-solution $y^{\varepsilon} : [\xi, \xi + T] \to \mathbb{R}^d$ with

- $|g(y^{\varepsilon}(t))| \le C\varepsilon,\ t \in [\xi, \xi + T]$.

Moreover, there exists a C^1-function y^0 such that

- $g(y^0(t)) \equiv 0$;
- $\lim\limits_{\varepsilon \to 0} \|y^{\varepsilon} - y^0\|_{C^0[\xi, \xi+T]} = 0$;
- $y'^0(t) = \mu(t)h^{+}(y^0(t)) + (1 - \mu(t))h^{-}(y^0(t)) \in \mathcal{T}_{y^0}\mathcal{M}$ with $\mu(t) \in [0, 1]$, $t \in [0, T]$;

where $\mathcal{M} = \{y \mid g(y) = 0\}$ is the manifold delimiting the two smooth regions of the vector fields and $\mathcal{T}_{y^0}\mathcal{M}$ is the linear manifold tangent to $\mathcal{M}$ at $y^0(t)$.

According to Theorem 10.1, the dynamics of the limit solution y^0 takes place in the manifold $\mathcal{M}$.

Then it is natural to define a *weak solution* of the problem (10.1) after a termination point as the solution of the *limit problem*

$$\begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} y'(t) \\ \mu'(t) \end{pmatrix} = \begin{pmatrix} h(t, y(t), \mu(t)) \\ \zeta - \alpha(y(t)) \end{pmatrix}, \tag{10.14}$$

I being the identity matrix and

$$\begin{aligned} h(t, y(t), \mu(t)) = {} & \mu(t)\, f(t, y(t), y(\zeta), x'^{+}(\zeta)) \\ & + (1 - \mu(t))\, f(t, y(t), y(\zeta), x'^{-}(\zeta)), \end{aligned}$$

with consistent initial data $(y(\xi), \mu(\xi))$ at $t = \xi$.

This system includes the constraint $\alpha(y(t)) = \zeta$ and, hence, is a differential-algebraic equation of index 2. It replaces the original problem (10.1) for $t \ge \xi$ until a classical solution is recovered.

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

10.3. Neutral problems as implicit delay equations

In view of the numerical integration of (10.1), by introducing a new variable $z(t) = y'(t)$, we rewrite it as the equivalent implicit system

$$\begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} y'(t) \\ z'(t) \end{pmatrix} = \begin{pmatrix} z(t) \\ -z(t) + f\big(y(t), y\big(\alpha(y(t))\big), z\big(\alpha(y(t))\big)\big) \end{pmatrix}, \tag{10.15}$$

with the initial conditions $y(t) = \phi(t)$, $z(t) = \phi'(t)$, $t \le t_0$, where I denotes the identity matrix.

Note that (10.15) falls in the general class of implicit problems

$$\begin{aligned} M\,u'(t) &= f\big(u(t), u\big(\alpha(t, u(t))\big)\big), \quad t_0 \le t \le t_f, \\ u(t) &= \psi(t), \quad t \le t_0, \end{aligned}$$

where the $d \times d$ matrix M is constant and possibly singular, which will be studied in Section 11.

Regularization by a singular perturbation

Following Bellen and Guglielmi (2009), as a regularization of (10.15) we consider the singularly perturbed problem

$$\begin{pmatrix} I & 0 \\ 0 & \varepsilon I \end{pmatrix} \begin{pmatrix} y'_\varepsilon(t) \\ z'_\varepsilon(t) \end{pmatrix} = \begin{pmatrix} z_\varepsilon(t) \\ -z_\varepsilon(t) + f\big(y_\varepsilon(t), y_\varepsilon\big(\alpha(y_\varepsilon(t))\big), z_\varepsilon\big(\alpha(y_\varepsilon(t))\big)\big) \end{pmatrix}, \tag{10.16}$$

which coincides with (10.15) for $\varepsilon = 0$.

Under standard assumptions on f, problem (10.16) admits a solution on a bounded interval for any fixed $\varepsilon > 0$. If the initial datum $(y_\varepsilon, z_\varepsilon) = (\phi, \psi)$ is continuous, then the corresponding solution is also continuous.

Although a theoretical analysis of the limit of the solution of (10.16) as $\varepsilon \to 0$ is still missing, the numerical experiments provided by Bellen and Guglielmi (2009) suggest that a limit solution exists.

Example 10.2. We consider problem (10.9) again, *i.e.*,

$$\begin{aligned} y'_1(t) &= 1 - 2\,y'_1(y_1(t) - 1), \\ y'_2(t) &= 2 - \frac{1}{2} y'_2(y_1(t) - 1), \end{aligned}$$

with initial data $y_1(t) = y_2(t) \equiv 0$ for $t \le 0$, $y_1(t) = t$, $y_2(t) = 2t$ for $0 \le t \le 1$. We have seen that $\xi = 1$ is a termination point.

The regularized problem of the form (10.13) is

$$y'^{\varepsilon}(t) = \frac{1}{\varepsilon} \int_{t-\varepsilon}^{t} \big[H\big(g(y^\varepsilon(s))\big)\, q + \big(1 - H\big(g(y^\varepsilon(s))\big)\big)\, p\big]\, \mathrm{d}s, \tag{10.17}$$

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

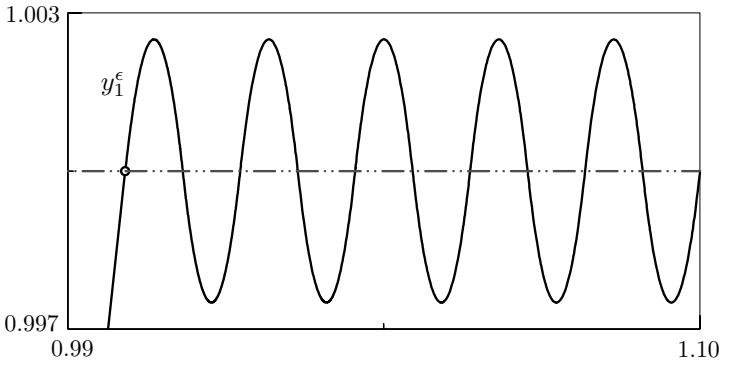


Figure 10.4. The solution component y_1^ε for $\varepsilon = 10^{-2}$.

where the vector fields p and q are constant and the corresponding manifold $\mathcal{M}$ is linear.

The solution of (10.17) can be found explicitly. The first component y_1^ε is periodic of period 2ε and continuously differentiable for $t \geq 1$ (see Figure 10.4). It is given by

$$y_1^\varepsilon(t) = \begin{cases} 1 + (t-1) - \frac{(t-1)^2}{\varepsilon}, & 1 \leq t \leq 1+\varepsilon, \\ 1 - (t-1-\varepsilon) + \frac{(t-1-\varepsilon)^2}{\varepsilon}, & 1+\varepsilon \leq t \leq 1+2\varepsilon, \end{cases}$$

in the interval $[1, 1+2\varepsilon]$ and is repeated periodically for $t \geq 1+2\varepsilon$. The second component y_2^ε is given by the sum of a periodic function and of a linear function

$$y_2^\varepsilon(t) = \frac{3}{2}\, t + \frac{1}{2}\, y_1^\varepsilon(t), \quad t \geq 1.$$

Note that, for all $t \geq 1$, the solution remains ε-close to the manifold

$$\mathcal{M} = \{y \mid y_1 = 1\}.$$

We also observe that y_1 and y_2 converge in the C^0-topology as $\varepsilon \to 0$, that is, there exist

$$y_1^0(t) = \lim_{\varepsilon \to 0} y_1^\varepsilon(t) = 1, \tag{10.18}$$

$$y_2^0(t) = \lim_{\varepsilon \to 0} y_2^\varepsilon(t) = \frac{3}{2}\, t + \frac{1}{2}. \tag{10.19}$$

Hence, y_1^0 and y_2^0 naturally represent the weak solution to the original problem (10.9).

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

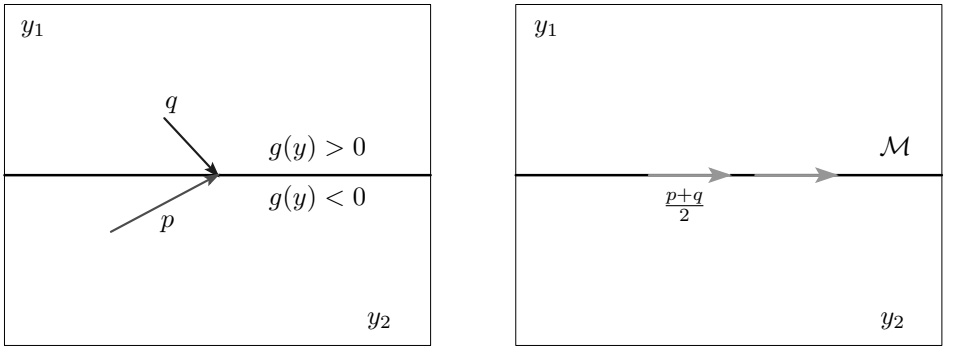


Figure 10.5. The limit problem associated to (10.17).

By (10.14), for $t \geq 1$ we get

$$\begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} y'(t) \\ \mu'(t) \end{pmatrix} = \begin{pmatrix} \mu(t)\, q + (1 - \mu(t))\, p \\ y_1(t) - 1 \end{pmatrix}. \tag{10.20}$$

The consistent initial value for μ is $\mu(1) = \frac{1}{2}$.

By differentiating twice the constraint $y_1(t) \equiv 1$, we get

$$\mu'(t)(p_1 - q_1) = 0 \implies \mu'(t) = 0,$$

which yields

$$\mu(t) = \frac{1}{2}.$$

As a consequence, for $t \geq 1$ the weak solution satisfies the equation

$$y'(t) = \frac{p}{2} + \frac{q}{2} = \begin{pmatrix} 0 \\ 3/2 \end{pmatrix}. \tag{10.21}$$

According to Theorem 10.1, the right-hand side of (10.20) gives the unique convex combination of p and q which lies on the manifold $\mathcal{M}$ (see Figure 10.5). This agrees with the definition of generalized solution of the discontinuous problem given in Filippov (1964, 1988).

Now, we consider the second proposed regularization (10.16), *i.e.*,

$$\begin{aligned} y_1'^{\varepsilon}(t) &= z_1^{\varepsilon}(t), \\ y_2'^{\varepsilon}(t) &= z_2^{\varepsilon}(t), \\ \varepsilon z_1'^{\varepsilon}(t) &= 1 - 2\, z_1^{\varepsilon}\big(y_1^{\varepsilon}(t) - 1\big) - z_1^{\varepsilon}(t), \\ \varepsilon z_2'^{\varepsilon}(t) &= 2 - \frac{1}{2} z_2^{\varepsilon}\big(y_1^{\varepsilon}(t) - 1\big) - z_2^{\varepsilon}(t), \end{aligned} \tag{10.22}$$

with initial data $y_1^{\varepsilon}(t) = y_2^{\varepsilon}(t) = z_1^{\varepsilon}(t) = z_2^{\varepsilon}(t) \equiv 0$ for $t \leq 0$.

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

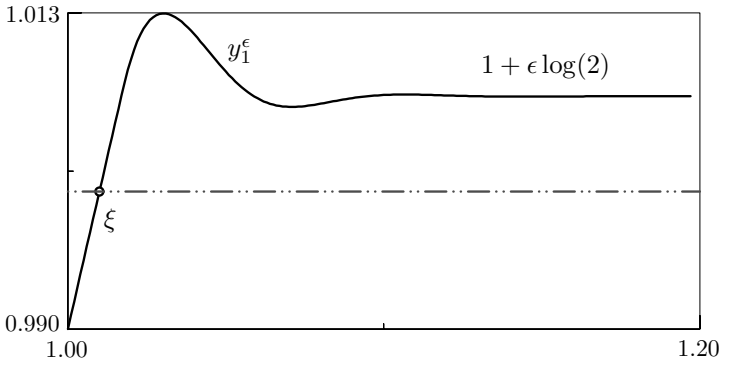


Figure 10.6. The first component of the solution of (10.22) in a neighbourhood of $\xi = 1$ (computed for $\varepsilon = 1/100$).

The first component y_1^ε of the solution has the behaviour shown in Figure 10.6.[1] In particular, it is still ε-close to the manifold $\mathcal{M}$ and its limit as $\varepsilon \to 0$ still coincides with (10.18). In any case, y_1^ε does not exhibit the oscillations of the previous regularizaton (compare to Figure 10.4). From a numerical point of view, the oscillations whose wavelength is of order ε constitute a challenging difficulty.

10.4. The numerics of weak solutions

Concerning the numerical approximation of weak solutions, we have two options: one is that of integrating, when required, the limit problem (10.14) and the other is that of approximating it by integrating the regularized problems (10.13) or (10.16).

Alternating the integration of (10.14) and (10.1) implies a repeated event detection and a possible change of problem whenever (10.7) occurs, as well as when the weak solution is such that

$$\big\langle n(y(t^*)), h^+(y(t^*))\big\rangle \cdot \big\langle n(y(t^*)), h^-(y(t^*))\big\rangle = 0$$

at some instant t^*. On the contrary, the numerical integration of the regularized problems does not require this. Nevertheless, due to the high oscillations, the numerical integration of (10.13) appears quite expensive, whereas integrating the singularly perturbed problem (10.16) seems to be cheaper. In any case, step-size restrictions are to be expected whenever some component of the z-variables (the solution derivatives) has a jump. Indeed, such jumps correspond to steep transitions of the same component of the z_ε variables.

[1] The asymptotic value $1 + \varepsilon \log(2)$ of y_1^ε was obtained using an asymptotic expansion of the solution, by E. Hairer, to whom we are grateful.

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

10.5. Numerical integration of neutral problems in implicit form

In order to integrate (10.15), it is natural to consider the method

$$\begin{pmatrix} Y^i_{n+1} - y_n \\ 0 \end{pmatrix} = h_{n+1} \begin{pmatrix} \sum_{j=1}^{\nu} a_{ij} Z^j_{n+1} \\ \sum_{j=1}^{\nu} a_{ij} \big(-Z^j_{n+1} + f\big(Y^j_{n+1}, \tilde{Y}^j_{n+1}, \tilde{Z}^j_{n+1}\big)\big) \end{pmatrix}, \quad \text{for } i = 1, \dots, \nu, \tag{10.23}$$

where

$$\tilde{Y}^j_{n+1} = \begin{cases} \phi(\alpha^j_{n+1}) & \text{if } \alpha^j_{n+1} < t_0, \\ \eta(\alpha^j_{n+1}) & \text{if } t_m \le \alpha^j_{n+1} < t_{m+1}, \end{cases}$$

for some $m \le n$ and

$$\tilde{Z}^j_{n+1} = \begin{cases} \phi'(\alpha^j_{n+1}) & \text{if } \alpha^j_{n+1} < t_0, \\ \lambda(\alpha^j_{n+1}) & \text{if } t_m \le \alpha^j_{n+1} < t_{m+1}, \end{cases}$$

with $\alpha^j_{n+1} = \alpha(Y^j_{n+1})$.

The continuous approximation $\eta(t)$ is still given by (8.32) and, for the continuous approximation $\lambda(t)$, one can consider two options. The first is chosen if t_m is not a computed breaking point and is given by

$$\lambda(t_m + \theta h_{m+1}) = \sum_{i=0}^{\nu} \ell_i(\theta) Z^i_{m+1}, \quad \theta \in [0,1], \tag{10.24}$$

where $Z^0_{m+1} = z_m$. The second option is chosen when t_m is a computed breaking point and is given by

$$\lambda(t_m + \theta h_{m+1}) = \sum_{i=1}^{\nu} \hat{\ell}_i(\theta) Z^i_{m+1}, \quad \theta \in [0,1], \tag{10.25}$$

where $\hat{\ell}_i(\theta)$, $i = 1, \dots, \nu$, are the Lagrange polynomials of degree $\nu - 1$ involving the collocation abscissae $c_1, \dots, c_\nu$ only.

Observe that the choice (10.24)–(10.25) also provides a generally discontinuous approximation of the solution derivative $z(t)$ at the computed breaking point t_m, according to the fact that, in general, the solution $y(t)$ is only C^0-continuous at breaking points.

In the case where $\alpha^j_{n+1} \in (t_n, t_{n+1}]$, *i.e.*, when the corresponding delay is smaller than the current step-size, $\eta(\alpha^j_{n+1})$ and $\lambda(\alpha^j_{n+1})$ are not known *a priori*, but only implicitly through the current stage values which are still to be computed.

For a non-singular coefficient matrix A (this is the case, for example, for

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

Gauss and Radau IIA methods), (10.23) gives

$$Z^j_{n+1} = f\big(Y^j_{n+1}, \tilde{Y}^j_{n+1}, \tilde{Z}^j_{n+1}\big), \quad j = 1, \dots, \nu,$$

whereas the first row remains

$$Y^i_{n+1} = y_n + h_{n+1} \sum_{j=1}^{\nu} a_{ij} Z^j_{n+1}, \quad i = 1, \dots, \nu.$$

In the case of overlapping, for some j we may have

$$\tilde{Y}^j_{n+1} = \eta(t_n + \theta^j_{n+1} h_{n+1}) = \sum_{i=0}^{\nu} \ell_i(\theta^j_{n+1}) Y^i_{n+1}$$

and

$$\tilde{Z}^j_{n+1} = \lambda(t_n + \theta^j_{n+1} h_{n+1}) = \sum_{i=0}^{\nu} \ell_i(\theta^j_{n+1}) Z^i_{n+1},$$

where $\theta^j_{n+1} = \alpha^j_{n+1}/h_{n+1}$. Therefore, all the approximated delayed terms $\tilde{Y}^j_{n+1}$ and $\tilde{Z}^j_{n+1}$ may be written in terms of Y^j_{n+1} and Z^j_{n+1}. It turns out that these are the same values provided by the approach described in Section 8 with the option (8.5) for the neutral equations in the form (10.1). Consequently, the two approaches are equivalent and the method converges according to Theorem 8.9, which holds under the crucial assumption that exact breaking points are included in the mesh Δ.

10.6. Checking existence and uniqueness numerically

Whenever the solution ceases to exist, a code which has not been designed to check termination typically stops the integration after the step-size has been reduced to a minimal value. This is certainly inconvenient, since the cause of this arrest would remain unclear. Hence it is important to check the possible termination numerically.

For $\bar{y} = y(\xi) \in \mathbb{R}^d$ such that $g(\bar{y}) = \alpha(\bar{y}) - \zeta = 0$, we must compute the sign of scalar products (10.6), or equivalently of

$$\big\langle \nabla g(\bar{y}), h^+(\bar{y}) \big\rangle = \sum_{i=1}^{d} \frac{\partial \alpha}{\partial y_i}(y(\xi))\, f_i(y(\xi), y(\zeta), x'^+(\zeta)), \tag{10.26}$$

$$\big\langle \nabla g(\bar{y}), h^-(\bar{y}) \big\rangle = \sum_{i=1}^{d} \frac{\partial \alpha}{\partial y_i}(y(\xi))\, f_i(y(\xi), y(\zeta), x'^-(\zeta)), \tag{10.27}$$

where f_i denotes the ith component of f.

The idea for a numerical investigation is based on the observation that, for a point $\bar{y} \in \mathcal{M}$, we can approximate (10.6) in the following way. According to Hairer and Wanner (1996) and Guglielmi and Hairer (2008), by

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

considering a first-order approximation of $g(\bar{y} + \delta\, h^{\pm}(\bar{y}))$, we get

$$g(\bar{y} + \delta\, h^{\pm}(\bar{y})) = g(\bar{y}) + \delta \left\langle \nabla g(\bar{y}), h^{\pm}(\bar{y}) \right\rangle + \mathcal{O}(\delta^2).$$

For a small $\delta > 0$, exploiting the property $g(\bar{y}) = 0$ yields

$$\left\langle \nabla g(\bar{y}), h^{+}(\bar{y}) \right\rangle \approx \frac{1}{\delta}\, g(\bar{y} + \delta\, h^{+}(\bar{y})), \tag{10.28}$$

$$\left\langle \nabla g(\bar{y}), h^{-}(\bar{y}) \right\rangle \approx \frac{1}{\delta}\, g(\bar{y} + \delta\, h^{-}(\bar{y})). \tag{10.29}$$

Note that this corresponds to applying a step of the Euler method to the pair of problems

$$y'(t) = h^{+}(y(t)) \quad \text{and} \quad y'(t) = h^{-}(y(t))$$

with step-size δ.

Let $t_n = \xi^*$ (approximating ξ) and $t_m = \zeta^*$ (approximating ζ) be a numerical breaking point and its ancestor, respectively. Then let

$$\lambda^{-}(t) = \lambda(t), \quad t \in [t_{m-1}, t_m),$$

$$\lambda^{+}(t) = \lambda(t), \quad t \in [t_m, t_{m+1}),$$

be the polynomial extensions of the derivative of the solution on the right-hand and left-hand side of the breaking point t_m, respectively. Such polynomials are clearly well defined in a whole neighbourhood of t_m. Observe that, in general, we expect that $\lambda^{+}(t_m) \neq \lambda^{-}(t_m)$.

Now, in order to proceed, it is sufficient to replace $x'^{+}(s)$ and $x'^{-}(s)$ by $\lambda^{+}(s)$ and $\lambda^{-}(s)$ in (10.1) (see Guglielmi and Hairer (2008)). Then, with

$$y_n^{+} = y_n + \delta f(y_n, y_m, \lambda^{+}(t_m)) \quad \text{and} \quad y_n^{-} = y_n + \delta f(y_n, y_m, \lambda^{-}(t_m)),$$

by using (10.4)–(10.5) and (10.28)–(10.29), at y_n we obtain

$$\begin{aligned} \left\langle \nabla g(y_n), h^{+}(y_n) \right\rangle &\approx a_\delta^{+} = \frac{\alpha(t_n + \delta, y_n^{+}) - t_m}{\delta}, \\ \left\langle \nabla g(y_n), h^{-}(y_n) \right\rangle &\approx a_\delta^{-} = \frac{\alpha(t_n + \delta, y_n^{-}) - t_m}{\delta}. \end{aligned} \tag{10.30}$$

If $a_\delta^{+} \cdot a_\delta^{-} > 0$, so that the solution continues to exist, the integration proceeds with the right-hand limit of $z(t)$ at t_n. On the contrary, if $a_\delta^{+} < 0$ and $a_\delta^{-} > 0$, the solution ceases to exist at t_n. Finally, if $a_\delta^{+} \approx 0$ or $a_\delta^{-} \approx 0$, a further analysis could determine whether the solution bifurcates at t_n.

10.7. Accuracy and breaking points

As in the non-neutral case, we want to overcome the need to include the exact breaking points in the mesh Δ. Moreover, with the use of the differential-algebraic formulation (10.15), we also intend to control the error in the z variable, that is, in the derivative of the solution y.

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

In line with (9.4)–(9.5), we couple the RK equations (10.23) to the equation for the breaking point ξ (with ancestor ζ) whenever the presence of a breaking point has been detected, so as to obtain the system

$$0 = \alpha\big(\eta(t_n + h_{n+1})\big) - \zeta, \tag{10.31}$$

$$\begin{pmatrix} Y_{n+1}^i - y_n \\ 0 \end{pmatrix} = h_{n+1} \begin{pmatrix} \sum_{j=1}^{\nu} a_{ij} Z_{n+1}^j \\ \sum_{j=1}^{\nu} a_{ij}\big(-Z_{n+1}^j + f\big(Y_{n+1}^j, \tilde{Y}_{n+1}^j, \tilde{Z}_{n+1}^j\big)\big) \end{pmatrix}, \quad \text{for } \ i = 1, \dots, \nu, \tag{10.32}$$

for the unknowns $Y_{n+1}^1, \dots, Y_{n+1}^\nu$, $Z_{n+1}^1, \dots, Z_{n+1}^\nu$ and h_{n+1}.

We still assume that the breaking points are simple, *i.e.*,

$$\frac{\mathrm{d}}{\mathrm{d}t}\big(\alpha(y(t))\big)\big|_{t=\xi} \neq 0, \tag{10.33}$$

and we get the following analogous result to Theorem 9.2 (see Guglielmi and Hairer (2008)).

Theorem 10.2. Consider a smooth problem (10.1) with simple breaking points (*i.e.*, (10.33) holds) and with non-vanishing delay satisfying the hypothesis (H_1) (see Section 9) and assume that, instead of the exact breaking points, those obtained by solving (10.31)–(10.32) are inserted into the mesh.

If the underlying collocation method has discrete order p (and uniform order $q = \nu$), then the resulting method for the NDDE (10.1) still has discrete global order and uniform global order q', where $q' = \min\{p, \nu + 1\}$.

If one considers the ν-stage Radau IIA methods, whose classical order is $p = 2\nu - 1$, with interpolants $\eta(t)$ and $\lambda(t)$ of uniform order $q = \nu$, the NDDE method converges with global uniform order $q' = \nu + 1$ for any $\nu \geq 2$. Such order results hold for any step-size, including the case of overlapping. The 3-stage method is used, for example, in the code Radar5.

Concerning the accuracy of the z variable, we cannot obtain any uniform estimate if we do not use a simple trick. As a matter of fact, for problems with state-dependent delays it is not possible to obtain uniform bounds to the global error of z. In fact, if a mesh point t_n is a numerically computed breaking point, the corresponding exact breaking point is slightly different in general. If the solution derivative has a jump discontinuity at this point, here the global error might be large independently of h.

It is possible to bypass this difficulty by comparing the solution derivative and its numerical approximation at slightly different times. In particular, Guglielmi and Hairer (2008) proved that

$$\lambda(t) - z(s) \ = \ \mathcal{O}(h^{q'}),$$

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

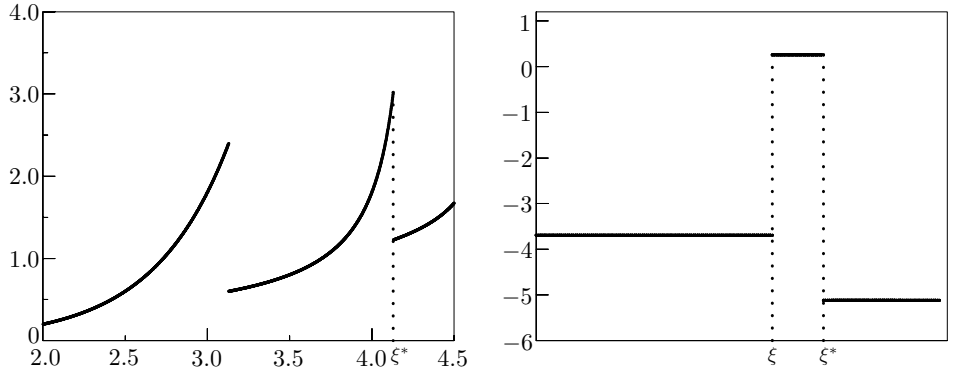


Figure 10.7. The solution component z of equation (10.34) and its numerical approximation (*left*); logarithm of the error of the numerical approximation of z in a neighbourhood of the breaking point ξ (*right*).

where $s = s(t)$ is a suitable smooth function which satisfies $s = t + \mathcal{O}(h^{q'})$ and $h = \max_{n\geq 1} h_n$.

Example 10.3. Let us consider the equation

$$\begin{aligned} y'(t) &= y'(y(t)) + \frac{1}{5}y(t), \quad 2 \leq t \leq 5, \\ \phi(t) &= (t-1)^2, \quad t \leq 2. \end{aligned} \tag{10.34}$$

Figure 10.7 shows that the derivative of the solution is completely inaccurate in a neighbourhood of the breaking point $\xi^* = 4.130469677\cdots$ of amplitude proportional to the error tolerance (which, in the specific case, is $R_{\text{tol}} = A_{\text{tol}} = 10^{-4}$). In fact, the computed breaking point is $\xi = 4.130454\cdots$.

11. Implicit problems with state-dependent delays

As anticipated in the previous section, here we face the study of implicit systems of DDEs of the general form

$$\begin{aligned} M\,u'(t) &= f\big(u(t), u\big(\alpha(t, u(t))\big)\big), \quad t_0 \leq t \leq t_f, \\ u(t) &= \psi(t), \quad t \leq t_0, \end{aligned} \tag{11.1}$$

where the $d \times d$ matrix M is constant and possibly singular.

Note that, for the sake of simplicity, we consider the autonomous case with a single deviated argument.

Besides NDDEs, this class of problems also includes singularly perturbed problems and, since we allow M to be singular, a variety of delay differential-algebraic equations (see, *e.g.*, the models in Shampine and Gahinet (2006)).

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

11.1. The numerical scheme

Consider an *implicit collocation method*

$$M(U^i_{n+1} - u_n) = h_{n+1} \sum_{j=1}^{\nu} a_{ij} f\big(U^j_{n+1}, \tilde{U}^j_{n+1}\big), \quad i = 1, \dots, \nu, \tag{11.2}$$

$$u_{n+1} = u_n + h_{n+1} \sum_{i=1}^{\nu} b_i f\big(U^i_{n+1}, \tilde{U}^i_{n+1}\big),$$

with ν distinct abscissae $c_1, \dots, c_\nu$ such that $c_i \neq 0$, $i = 1, \dots, \nu$, where $\tilde{U}^i_{n+1}$ is an approximation to $u(\alpha(u(t_n + c_i h_{n+1})))$ defined by

$$\tilde{U}^i_{n+1} = \begin{cases} \phi(\alpha^i_{n+1}) & \text{if } \alpha^i_{n+1} < t_0, \\ \eta(\alpha^i_{n+1}) & \text{if } \alpha^i_{n+1} \geq t_0, \end{cases} \tag{11.3}$$

where, in turn,

$$\alpha^i_{n+1} = \alpha(U^i_{n+1}).$$

As in (8.32), the continuous approximate solution is given by

$$\eta(t_n + \theta h_{n+1}) = \sum_{i=0}^{\nu} \ell_i(\theta) U^i_{n+1}, \quad \theta \in [0, 1), \tag{11.4}$$

where $U^0_{n+1} = u_n$ and $c_0 = 0$.

If the mesh point t_n is a computed breaking point, in the step $[t_n, t_{n+1}]$ the polynomial (11.4) can optionally be replaced by

$$\eta(t_n + \theta h_{n+1}) = \sum_{i=1}^{\nu} \ell_i(\theta) U^i_{n+1}, \quad \theta \in [0, 1],$$

which interpolates the internal stage values only, but not u_n (see Guglielmi and Hairer (2001)). The use of this option is important in the presence of a jump discontinuity in some component of the solution, since it permits to have also a discontinuity in the continuous approximation of the solution. Hence, in general, we have $\eta(t_n) \neq u_n$.

As usual, for overlapping, that is, when $\alpha^j_{n+1} \in (t_n, t_{n+1}]$ for some j, the term $\eta\big(\alpha^j_{n+1}\big)$ is not known *a priori*, but only implicitly through the unknown stage values $U^1_{n+1}, \dots, U^\nu_{n+1}$.

11.2. Computing the breaking points

As for the neutral case examined in the previous section, the computation of the breaking points is based on the coupling of the system of the Runge–Kutta equations (11.2) and of the equation for the breaking point

$$0 = \alpha\big(t_n + h_{n+1}, \eta(t_n + h_{n+1})\big) - \zeta. \tag{11.5}$$

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

Generalizing what we have explained in Section 10.7, in this case it is possible to prove the following result (which is stated in Guglielmi and Hairer (2008)).

Theorem 11.1. Consider a smooth problem (11.1) with simple breaking points, *i.e.*, such that

$$\frac{\mathrm{d}}{\mathrm{d}t}\big(\alpha(u(t))\big)\big|_{t=\xi^*} \neq 0,$$

and with non-vanishing delay satisfying the hypothesis (H_1) (see Section 9). Moreover, assume that the uniform global error of the RK method (11.2) is of size $\mathcal{O}(h^r)$ ($h = \max_{n\geq 1} h_n$) if the exact breaking points of order $\leq r$ are inserted into the mesh.

Then, if, instead of the exact breaking points, those computed by solving (11.2)–(11.5) are used, the global error of the resulting method satisfies

$$\max_{t_0\leq t\leq t_f} |u(s) - \eta(t)| = \mathcal{O}(h^r), \tag{11.6}$$

where $s = s(t)$ is a suitable smooth function such that $s(t) = t + \mathcal{O}(h^r)$.

The proof is based on an error estimate for the computed breaking points which is analogous to that proved for explicit problems (see Theorem 9.1) and on the classical convergence proof. The only additional difficulty lies in the fact that, for the discontinuous components, it is necessary to align the computed and the exact breaking points in order to obtain a significant error bound. Such an alignment is based on the error estimate $|\xi - \xi^*| = \mathcal{O}(h^r)$ between the computed and the exact breaking point. As a consequence, the global error $|u(t) - \eta(t)|$ is still of size $\mathcal{O}(h^r)$ in $[t_0, t_f]$, except for all the small intervals of the type $[\xi, \xi^*]$, whose size is $\mathcal{O}(h^r)$. This is consistent with what we have shown in Figure 10.7 for Example 10.3.

11.3. Solving the RK equations

In this section we focus on the solution of the RK equations in the particular case when overlapping occurs. The RK system (11.2) has the form

$$F^i_{n+1}\big(U^1_{n+1}, \ldots, U^\nu_{n+1}, \tilde{U}^1_{n+1}, \ldots, \tilde{U}^\nu_{n+1}\big) = 0, \quad i = 1, \ldots, \nu, \tag{11.7}$$

for the unknowns $U^1_{n+1}, \ldots, U^\nu_{n+1}$, where

$$F^i_{n+1} = M(U^i_{n+1} - u_n) - h_{n+1} \sum_{j=1}^{\nu} a_{ij} f\big(U^j_{n+1}, \tilde{U}^j_{n+1}\big).$$

We recall that

$$\tilde{U}^j_{n+1} = \eta(\alpha^j_{n+1}) \quad \text{if} \quad \alpha^j_{n+1} > t_0.$$

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

We are interested in solving (11.7) by means of a suitable iterative Newton process. For sake of conciseness, we omit the dependence of F^i_{n+1}, U^j_{n+1}, $\tilde{U}^j_{n+1}$ and α^j_{n+1} on n. Moreover, we denote by $f(u,\tilde{u})$ the function on the right-hand side of (11.1).

In order to obtain an accurate computation of the derivatives of F^i, we consider the approximation

$$\frac{\partial F^i}{\partial U^k} \approx M\,\delta_{ik} - h_{n+1}\,(a_{ik}\,D^k + \hat{D}^k), \tag{11.8}$$

where δ_{ik} is the Kronecker delta and

$$D^k = \frac{\partial f}{\partial u}(U^k,\tilde{U}^k) + \frac{\partial f}{\partial \tilde{u}}(U^k,\tilde{U}^k)\,\eta'(\alpha^k)\frac{\partial \alpha}{\partial u}(U_k),$$
$$\hat{D}^k = \sum_{j=1}^{\nu} a_{ij}\,\frac{\partial f}{\partial \tilde{u}}(U^k,\tilde{U}^k)\,\frac{\partial \tilde{U}^j}{\partial U^k}.$$

Note that the term $\frac{\partial \tilde{U}^j}{\partial U^k} = 0$ if the deviated argument $\alpha^j \le t_n$. More precisely, since

$$\eta(t_n + \theta h_{n+1}) = \sum_{k=0}^{\nu} \ell_k(\theta) U^k, \quad \theta \in [0,1),$$

in the current interval, we get

$$\frac{\partial \tilde{U}^j}{\partial U^k} = \mathcal{U}_{jk}\,I_d,$$

where I_d denotes the $d \times d$ identity matrix and

$$\mathcal{U}^{jk} = \begin{cases} \ell_k(\theta_j) & \text{if } \theta_j > 0, \\ 0 & \text{otherwise,} \end{cases} \tag{11.9}$$

with

$$\theta_j = \big(\alpha^j - t_n\big)/h_{n+1}.$$

In order to reduce computation, we simplify (11.8) by approximating all derivatives of the functions f and α by

$$\frac{\partial \alpha}{\partial u}(U^k) \approx \frac{\partial \alpha}{\partial u}(u_n), \tag{11.10}$$

$$\frac{\partial f}{\partial u\,\partial \tilde{u}}(U^k,\tilde{U}^k) \approx \frac{\partial f}{\partial u\,\partial \tilde{u}}(u_n,\tilde{u}_n), \tag{11.11}$$

where $\tilde{u}_n = \eta(\alpha^0)$ and $\alpha^0 = \alpha(u_n)$. As a consequence, we expect the Newton process to be linearly convergent.

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

11.4. General form of the Jacobian

In order to solve (11.7) by means of a Newton process, consider the Jacobian

$$J = I \otimes M - h_{n+1}\, A \otimes \left(\frac{\partial f}{\partial u} + \frac{\partial f}{\partial \tilde{u}} \eta'(\alpha^0) \frac{\partial \alpha}{\partial u} \right) - h_{n+1}\, A \cdot \mathcal{U} \otimes \frac{\partial f}{\partial \tilde{u}}. \quad (11.12)$$

In general, J has a full structure and does not admit any transformation allowing for a reduction in the cost of its LU-factorization.

Structure of the Jacobian in the case without overlapping
However, when no overlapping occurs, that is, when $\alpha(U^j) < t_n$ for all $j = 1, \ldots, \nu$, we have $\mathcal{U} = \mathrm{O}$ and, therefore,

$$J = J_0 = I \otimes M - h_{n+1}\, A \otimes B_0, \quad (11.13)$$

where

$$B_0 = \frac{\partial f}{\partial u} + \frac{\partial f}{\partial \tilde{u}} \eta'(\alpha^0) \frac{\partial \alpha}{\partial u}.$$

Then, as in the ODE case (see, *e.g.*, Butcher (1976)), if the RK matrix A is invertible, the matrix J_0 can be pre-multiplied by $(h_{n+1}\, A)^{-1} \otimes I_d$. In such a case it is useful to transform A^{-1}, so as to obtain a block-diagonal matrix D

$$T^{-1}\, A^{-1}\, T = D$$

(see, *e.g.*, Hairer and Wanner (1996)). By introducing the transformed variables $W = (T^{-1} \otimes I_d)\, U$, we obtain an equivalent Newton iteration with Jacobian

$$\hat{J}_0 = h_{n+1}^{-1}\, D \otimes M - I \otimes B_0. \quad (11.14)$$

Such a matrix has block-diagonal structure and, thus, the computational cost for its factorization is much cheaper than that of J.

11.5. Preserving the tensor structure of the Jacobian

Unfortunately, in the case of overlapping the previous transformation is not possible. However, an analogous transformation of the Jacobian J to block-diagonal structure is possible if we approximate the matrix $\mathcal{U}$ by

$$\mathcal{U} \approx \gamma\, I_\nu \quad \text{for an optimal} \quad \gamma \in \mathbb{R},$$

where I_ν stands for the $\nu \times \nu$ identity matrix.

The iteration with no overlapping corresponds to having $\gamma = 0$. In general, simply approximating J by J_0 may prevent the Newton iteration from convergence or make it very slow (see the examples by Castleton and Grimm (1973) and by Waltman (1978) studied in Guglielmi (2005)).

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

A better choice consists in choosing an optimal $\gamma \in \mathbb{R}$, according to some approximation criteria. The idea is then to make use of an inexact Jacobian which can be block-diagonalized and, if the corresponding inexact Newton process does not converge, either to reduce the step-size or to switch to the exact iteration, that is, to make use of (11.12).

We adopt the optimization criterion

$$\gamma^* \longrightarrow \min_{\gamma \in \mathbb{R}} \|\mathcal{U} - \gamma\, I_\nu\|_F^2, \tag{11.15}$$

where $\|\cdot\|_F$ is the Frobenius norm. This choice leads to an explicit formula for the optimal parameter γ^* and has been supported by several numerical experiments.

In the special case $\alpha(y(t)) \equiv t$ (no delay case), we have $\alpha(U^j) = t_n + c_j h_{n+1}$, $j = 1, \ldots, \nu$. Hence, $\mathcal{U} = I_\nu$ and, consequently, $\gamma^* = 1$, which can also be considered a good approximation of the optimal parameter for those cases when the step-size is much larger than the delay.

Example 11.1. Let us consider the Radau IIA 2-stage method, whose tableau is given by

$$\begin{array}{c|cc} \frac{1}{3} & \frac{5}{12} & -\frac{1}{12} \\ 1 & \frac{3}{4} & \frac{1}{4} \\ \hline & \frac{3}{4} & \frac{1}{4} \end{array}\ .$$

The jth row of the 2×2 matrix $\mathcal{U}$ is zero if $\alpha(U^j) < t_n$, $j = 1, 2$. We have that

$$\mathcal{U} = \begin{pmatrix} H(\theta_1) & 0 \\ 0 & H(\theta_2) \end{pmatrix} \cdot \begin{pmatrix} \ell_{11} & \ell_{12} \\ \ell_{21} & \ell_{22} \end{pmatrix},$$

where $\ell_{jk} = \ell_k(\theta_j)$ and $H(\cdot)$ is the unit Heaviside function.

Since the function to minimize in (11.15) is quadratic with respect to γ, the minimizer

$$\gamma^* = \frac{-9\, H(\theta_1)\,(-1+\theta_1)\,\theta_1 + H(\theta_2)\,\theta_2\,(-1+3\,\theta_2)}{4}$$

is a global one.

Synthesis of the inexact Newton process

With the previous procedure we obtain an approximation of the Jacobian (11.12) given by

$$J \approx J_{\gamma^*} = I \otimes M - h_{n+1}\, A \otimes B_{\gamma^*}, \tag{11.16}$$

where

$$B_\gamma = \frac{\partial f}{\partial u} + \frac{\partial f}{\partial \tilde{u}} \left(\eta'(\alpha^0) \frac{\partial \alpha}{\partial u} + \gamma I_\nu \right).$$

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

By making use of the same transformation used to obtain $\hat{J}_0$ (see (11.14)), we get

$$\hat{J}_{\gamma^*} = (h_{n+1})^{-1} D \otimes M - I \otimes B_{\gamma^*}, \tag{11.17}$$

which has the same block-diagonal structure as $\hat{J}_0$.

The experimental results obtained on the examples from the test set by Paul (1994) and on the test problems included in the code Radar5 have shown that the use of the inexact Jacobian J_{γ^*} allows us to obtain a more efficient integration of problems with vanishing or small delays since, in most cases, the use of the exact Jacobian J can be avoided.

REFERENCES

C. T. H. Baker (1996), Numerical analysis of Volterra functional and integral equations, in *The State of the Art in Numerical Analysis* (I. S. Duff and G. A. Watson, eds), Clarendon Press, Oxford.

C. T. H. Baker (2000), 'Retarded differential equations', *J. Comput. Appl. Math.* **125**, 309–335.

C. T. H. Baker and C. A. H. Paul (2006), 'Discontinuous solutions of neutral delay differential equations', *Appl. Numer. Math.* **56**, 284–304.

C. T. H. Baker, C. A. H. Paul and D. R. Willé (1995*a*), 'Issues in the numerical solution of evolutionary delay differential equations', *Adv. Comput. Math.* **3**, 171–196.

C. T. H. Baker, C. A. H. Paul and D. R. Willé (1995*b*), A bibliography on the numerical solution of delay differential equations. NA Report 269, Department of Mathematics, University of Manchester.

A. Bellen (1985), Constrained mesh methods for functional differential equations, in *Delay Equations, Approximation and Application*, Vol. 74 of *Internat. Ser. Numer. Math.*, pp. 52–70.

A. Bellen and N. Guglielmi (2009), 'Solving neutral delay differential equations with state dependent delays', *J. Comput. Appl. Math.*, in press.

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

R. Bellman and K. L. Cooke (1963), *Differential-Difference Equations*, Academic Press.

H. Brunner (2004), *Collocation Methods for Volterra Integral and Related Functional Differential Equations*, Cambridge University Press, Cambridge.

J. C. Butcher (1976), 'On the implementation of implicit Runge–Kutta methods', *BIT* **6**, 237–240.

R. N. Castleton and L. J. Grimm (1973), 'A first order method for differential equations of neutral type', *Math. Comput.* **27**, 571–577.

K. L. Cooke and J. Wiener (1984), 'Retarded differential equations with piecewise constant delays', *J. Math. Anal. Appl.* **99**, 265–297.

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

C. W. Cryer (1972), Numerical methods for functional differential equations, in *Delay and Functional Differential Equations and their Applications* (K. Schmitt, ed.), Academic Press, New York, pp. 17–101.

C. W. Cryer and L. Tavernini (1972), 'The numerical solution of Volterra functional differential equations by Euler's method', *SIAM J. Numer. Anal.* **9**, 105–129.

O. Diekmann, S. A. van Gils, S. M. Verduyn Lunel and H. O. Walther (1995), *Delay Equations: Functional-, Complex-, and Nonlinear Analysis*, AMS series, Springer, Berlin.

R. D. Driver (1977), *Ordinary and Delay Differential Equations*, Springer, Berlin.

L. E. El'sgol'ts and S. B. Norkin (1973), *Introduction to the Theory and Application of Differential Equations with Deviating Arguments*, Academic Press, New York.

W. H. Enright and H. Hayashi (1997), 'A delay differential equation solver based on a continuous Runge–Kutta method with defect control', *Numer. Algorithms* **16**, 349–364.

W. H. Enright, K. R. Jackson, S. P. Nørsett and P. G. Thomsen (1988), 'Effective solution of discontinuous IVPs using a Runge–Kutta formula pair with interpolants', *Appl. Math. Comput.* **27**, 313–335.

A. Feldstein (1964), Discretization methods for retarded ordinary differential equation. PhD Thesis, Department of Mathematics, UCLA, Los Angeles.

A. Feldstein and K. W. Neves (1984), 'High order methods for state-dependent delay differential equations with nonsmooth solutions', *SIAM J. Numer. Anal.* **21**, 844–863.

A. Feldstein, K. W. Neves and S. Thompson (2006), 'Sharpness results for state dependent delay differential equations: An overview', *Appl. Numer. Math.* **56**, 472–487.

A. F. Filippov (1964), 'Differential equations with discontinuous right-hand sides', *Trans. Amer. Math. Soc.* **42**, 199–231.

A. F. Filippov (1988), *Differential Equations with Discontinuous Righthand Sides*, Vol. 18 of *Mathematics and its Applications* (Soviet Series), Kluwer Academic, Dordrecht (translated from the Russian).

G. Fusco and N. Guglielmi (2009), A regularization for discontinuous differential equations with application to state-dependent delay differential equations of neutral type. In preparation.

N. Guglielmi (2005), 'On the Newton iteration in the application of collocation methods to implicit delay equations', *Appl. Numer. Math.* **53**, 281–297.

N. Guglielmi and E. Hairer (2001), 'Implementing Radau IIA methods for stiff delay differential equations', *Computing* **67**, 1–12.

N. Guglielmi and E. Hairer (2008), 'Computing breaking points in implicit delay differential equations', *Adv. Comput. Math.* **29**, 229–247.

E. Hairer and G. Wanner (1996), *Solving Ordinary Differential Equations II: Stiff and Differential Algebraic Problems*, Springer Series in Computational Mathematics, Springer, Berlin.

J. K. Hale (1977), *Theory of Functional Differential Equations*, Springer, New York.

J. K. Hale and S. M. Verduyn Lunel (1993), *Introduction to Functional Differential Equations*, Applied Mathematical Sciences, Springer, New York.

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press

R. Hauber (1997), 'Numerical treatment of retarded differential-algebraic equations by collocation methods', *Adv. Comput. Math.* **7**, 573–592.
V. Kolmanovskii and A. Myshkis (1992), *Applied Theory of Functional Differential Equations*, Kluwer, Dordrecht.
V. Kolmanovskii and V. Nosov (1986), *Stability of Functional Differential Equations*, Academic Press, London.
J. Kuang and Y. Cong (2005), *Stability of Numerical Methods for Delay Differential Equations*, Science Press, Beijing.
Y. Kuang (1993), *Delay Differential Equations with Applications in Population Dynamics*, Academic Press, Boston.
S. Maset (2009), Theoretical and numerical analysis of retarded functional differential equations. In preparation.
S. Maset, L. Torelli and R. Vermiglio (2005), 'Runge–Kutta methods for retarded functional differential equations', *Math. Models Methods Appl. Sci.* **15**, 1203–1251.
G. Meinardus and G. Nürnberger (1985), Approximation theory and numerical methods for delay differential equations, in *Delay Equations, Approximation and Application*, Vol. 74 of *Internat. Ser. Numer. Math.*, pp. 13–40.
K. W. Neves (1975), 'Automatic integration of functional differential equations: An approach', *ACM Trans. Math. Software* **1**, 357–368.
C. A. H. Paul (1994), A test set of functional differential equation. NA Report 243, Department of Mathematics, University of Manchester.
L. F. Shampine and P. Gahinet (2006), 'Delay-differential-algebraic equations in control theory', *Appl. Numer. Math.* **56**, 574–588.
L. F. Shampine and S. Thompson (2000), 'Event location for ordinary differential equations', *Comput. Math. Appl.* **39**, 43–54.
L. Tavernini (1971), 'One-step methods for the numerical solution of Volterra functional differential equations', *SIAM J. Numer. Anal.* **4**, 786–795.
P. Waltman (1978), A threshold model of antigen-stimulated antibody production, in *Theoretical Immunology*, Vol. 8 of *Immunology* series, Dekker, New York, pp. 437–453.
W. Wang and S. Li (2004), 'Stability analysis of nonlinear delay differential equations of neutral type', *Math. Numer. Sin.* **26**, 303–314.
D. R. Willé and C. T. H. Baker (1992), 'The tracking of derivative discontinuities in systems of delay differential equations', *Appl. Numer. Math.* **9**, 299–222.
M. Zennaro (1995), Delay differential equations: Theory and numerics, in *Theory and Numerics of Ordinary and Partial Differential Equations* (M. Ainsworth, J. Levesley, W. A. Light and M. Marletta, eds), Clarendon Press, Oxford, pp. 291–333.

https://doi.org/10.1017/S0962492906390010 Published online by Cambridge University Press