Applied Numerical Mathematics 12 (1993) 315–330
North-Holland

APNUM 407

Error control for initial value problems with discontinuities and delays

Desmond J. Higham

Department of Mathematics and Computer Science, University of Dundee, Dundee, Scotland DD1 4HN, UK

Abstract

Higham, D.J., Error control for initial value problems with discontinuities and delays, Applied Numerical Mathematics 12 (1993) 315–330.

When using software for ordinary differential equation (ODE) initial value problems, it is not unreasonable to expect the global error to decrease linearly with the user-supplied error tolerance. For standard ODEs, conditions on an algorithm that guarantee such "tolerance proportionality" asymptotically (as the error tolerance tends to zero) were derived by Stetter. Here we extend the analysis to cover a certain class of ODEs with low-order derivative discontinuities, and the class of ODEs with constant delays. We show that standard error control techniques will be successful if discontinuities are handled correctly and delay terms are calculated with sufficiently accurate interpolants. It is perhaps surprising that several delay ODE algorithms that have been proposed do not use sufficiently accurate interpolants to guarantee asymptotic proportionality. Our theoretical results are illustrated numerically.

Keywords. Delay ordinary differential equations; discontinuity; global error; interpolation; local error; defect; residual; tolerance proportionality; Runge–Kutta.

1. Introduction

A typical user of ordinary differential equation (ODE) initial value software will define a problem, specify one or more output points, and choose an error tolerance. Hence, despite the fact that the stepsize plays a central role in the design and analysis of ODE solvers, the meshpoints selected by the algorithm are normally transparent to the user. For this reason it is pertinent to ask how the error in the numerical approximation behaves as a function of the error tolerance. This question was addressed by Stetter [18,20], who derived sufficient conditions for an algorithm to exhibit "tolerance proportionality", that is, an approximately linear relationship between error and tolerance. Tolerance proportionality (TP) is widely regarded as an extremely desirable property (indeed it is often, and sometimes erroneously, taken for granted by users). The DETEST package [8], for example, uses a linear least squares fit of error versus tolerance as one criterion for evaluating the performance of an initial value solver.

Correspondence to: D.J. Higham, Department of Mathematics and Computer Science, University of Dundee, Dundee, Scotland DD1 4HN, UK. E-mail: dhigham@uk.ac.dund.mcs.

Further analysis directed at explicit Runge–Kutta (RK) methods with continuous extensions was given in [11]. This work is a sequel to [11] and its aim is to extend the existing analysis to allow for ODEs with low-order derivative discontinuities and ODEs with constant delays.

In the rest of this section, we very briefly outline the results that will be used later; for more details, and numerical examples, see [11].

We consider the solution of the nonstiff initial value problem

$$y'(t) = f(t,\, y(t)), \qquad y(t_0) = y_0 \in \mathbb{R}^N, \quad t_0 \leqslant t \leqslant t_{\text{end}}, \tag{1.1}$$

where the range of integration $[t_0, t_{\text{end}}]$ is finite, using a pth-order RK method. When such a method advances from $y_{n-1} \approx y(t_{n-1})$ to $y_n \approx y(t_n)$ over a step of length $h_n := t_n - t_{n-1}$, the *local error* for the step is defined as

$$\text{le}_n := y_n - z_n(t_n),$$

where the *local solution*, $z_n(t)$, satisfies $z_n'(t) = f(t,\, z_n(t))$ and $z_n(t_{n-1}) = y_{n-1}$. Unless otherwise stated, we assume that the problem (1.1) is sufficiently smooth for the local error expansion

$$\text{le}_n = h_n^{p+1}\psi(y_{n-1},\, t_{n-1}) + \mathrm{O}\big(h_n^{p+2}\big) \tag{1.2}$$

to hold, where the continuous function ψ is independent of h_n. We further assume that a locally-based measure of the error, $\| e(y_{n-1},\, t_{n-1},\, h_n) \|$, is computed in the course of the step, where

$$e(y_{n-1},\, t_{n-1},\, h_n) = h_n^p \tilde{\psi}(y_{n-1},\, t_{n-1}) + \mathrm{O}\big(h_n^{p+1}\big), \tag{1.3}$$

and $\tilde{\psi}$ is continuous and independent of h_n. Typically, $e(y_{n-1},\, t_{n-1},\, h_n)$ will involve the difference of two approximations to $y(t_n)$. Also, the norm $\| \cdot \|$ may incorporate componentwise absolute and relative weights specified by the user.

The step is accepted if $\| e(y_{n-1},\, t_{n-1},\, h_n) \| \leqslant \delta$, where δ is the user-supplied error tolerance, otherwise the step is retaken from t_{n-1} with a smaller stepsize h_n. The usual method for selecting the next stepsize is to take a fixed portion of the asymptotically optimal stepsize, h_{opt}; that is,

$$h_{\text{new}} = \theta h_{\text{opt}}, \quad h_{\text{opt}} = \left(\frac{\delta}{\| e(y_{n-1},\, t_{n-1},\, h_n) \|} \right)^{1/p} h_n, \tag{1.4}$$

with $\theta \in (0,\, 1)$ constant.

We will use $\eta(t)$ to denote any continuous interpolant that takes the value y_n at t_n for $n = 0,\, 1, \ldots$. In particular $\eta_1(t)$ denotes the *ideal* interpolant from [19], which spreads the local error uniformly over each step:

$$\eta_I(t) := z_n(t) + \frac{(t - t_{n-1})}{h_n} \text{le}_n, \quad t \in (t_{n-1},\, t_n]. \tag{1.5}$$

We point out that $\eta_I(t)$ is not necessarily computable, and that $\eta_I'(t)$ generally has a jump discontinuity at each meshpoint t_n.

In the following theorem, which is taken from [11, Theorem 2.1], we use the convention that "piecewise continuous" means continuous except possibly at the meshpoints $\{t_n\}$, and "piece-

wise C^1" means continuous with a first derivative which is continuous except possibly at the meshpoints. We also point out that the theorem does not require the function $\eta(t)$ to be related to a numerical solution—this fact will be exploited in Section 3.

Theorem 1.1. *Given the initial value problem* (1.1) *suppose* $\eta(t)$ *is piecewise* C^1 *and satisfies* $\eta(t_0) = y_0$. *Let* $\varepsilon(t) := \eta(t) - y(t)$ *denote the global error in* $\eta(t)$. *Then the conditions* (A) *and* (B) *below are equivalent*:

(A) $\varepsilon(t) = v(t)\delta + g(t)$, $t_0 \leqslant t \leqslant t_{\text{end}}$, *where* $v(t)$ *is* C^1 *and independent of* δ, *and* $g(t)$ *is piecewise* C^1 *with zeroth and first derivatives of* $o(\delta)$.

(B) $\eta'(t) - f(t, \eta(t)) = \gamma(t)\delta + s(t)$, $t_0 \leqslant t \leqslant t_{\text{end}}$, *where* $\gamma(t)$ *is continuous and independent of* δ, *and* $s(t)$ *is piecewise continuous and* $o(\delta)$.

Condition (A) is a formalization of the concept of tolerance proportionality. For any fixed point $t_0 \leqslant t \leqslant t_{\text{end}}$, it ensures that the global error is asymptotically linear in δ. The condition is strong in the sense that it also requires the global error in $\eta'(t)$ to be asymptotically linear in δ. The equivalent condition (B) provides a more useful characterization from the point of view of analysis.

We will make the following assumptions regarding the numerical solutions:

(1) The stepsizes satisfy $\max_n\{h_n\} = o(1)$ as $\delta \to 0$.

(2) The initial stepsize is chosen so that

$$\| e(y_0, t_0, h_1) \| = \theta^{-p}\delta + o(\delta), \tag{1.6}$$

(3) The function $\| \tilde{\psi}(y(t), t) \|$ from (1.3) is non-vanishing.

Note that assumption (1) implies convergence of the Runge–Kutta solution; that is, $\varepsilon(t) \to 0$ as $\delta \to 0$ (see, for example, [10, Theorem 3.4]). Also, from assumption (3) the error control criterion $\| e(y_{n-1}, t_{n-1}, h_n) \| \leqslant \delta$ implies that a function that is $O(h_n^p)$ is also $O(\delta)$. It is shown in [11] that under assumptions (1), (2), and (3) the ideal interpolant satisfies condition (A), and hence the algorithm exhibits tolerance proportionality.

In recent years it has become common to augment the discrete RK approximation with a computable interpolant, or continuous extension, $q(t) \approx y(t)$, satisfying $q(t_n) = y_n$ and $q'(t_n) = f(t_n, y_n)$. Computable interpolants can be used to provide graphical output, off-meshpoint approximations, and approximate roots of the solution; see, for example, [3,5,6]. It was shown in [11] that such interpolants will not satisfy condition (A) of Theorem 1.1, although they may satisfy a weaker condition. To be more specific, the RK interpolants that have been proposed in the literature can be split into two groups; if l is the largest integer such that, for every fixed $\tau \in [0, 1]$,

$$q(t_{n-1} + \tau h_n) - z_n(t_{n-1} + \tau h_n) = O(h_n^l),$$

then q is *higher order* if $l = p + 1$ and *lower order* if $l = p$. Higher-order interpolants satisfy

$$q(t) - y(t) = v(t)\delta + o(\delta), \quad t_0 \leqslant t \leqslant t_{\text{end}}, \tag{1.7}$$

where $v(t)$ is C^1 and independent of δ, but *not* $q'(t) - y'(t) = v'(t)\delta + o(\delta)$. Hence they preserve the TP in the solution approximation, but not in the first derivative approximation. For lower-order interpolants we have $q(t) - y(t) = O(\delta)$, but the leading term in the global error does not, in general, depend linearly upon δ. This difference in behaviour between the two classes of interpolants plays a key role in our analysis for delay ODEs.

In the next section we look at the effect of overriding the usual stepsize selection mechanism in order to hit an output point exactly or to integrate across a discontinuity. As well as being of interest in their own right, the results of Section 2 are used in Section 3 where error control methods for constant delay ODE systems are analysed. In both sections we test our predictions numerically. We give our conclusions in Section 4.

2. Output points and discontinuities

Suppose that the RK method described above reaches the point t_{n-1} and uses (1.4) to compute the "natural" stepsize h_{new} with which to continue the integration. There are certain circumstances under which a method will artificially restrict the stepsize to $h^* < h_{\text{new}}$ in order to hit the point $t^* := t_{n-1} + h^*$ exactly. This may happen, for example, if t^* has been specified as an output point, or if a low-order derivative of the solution is known to have a discontinuity at t^*. In this case, we have $z_n(t_{n-1}) - y(t_{n-1}) = v(t_{n-1})\delta + o(\delta)$. A standard differential inequality [10, Theorem 10.2] then gives

$$z_n(t) - y(t) = O(\delta) \quad \text{for } t_{n-1} \leqslant t \leqslant t^*.$$

Assuming that f is Lipschitzian, it follows that

$$z_n'(t) - y'(t) = O(\delta) \quad \text{for } t_{n-1} \leqslant t \leqslant t^*,$$

and hence

$$\begin{aligned} z_n(t') - y(t') &= z_n(t_{n-1}) - y(t_{n-1}) + O(h^*\delta) \\ &= v(t_{n-1})\delta + o(\delta). \end{aligned}$$

Since $v(t)$ is C^1, we have

$$z_n(t^*) - y(t^*) = v(t^*)\delta + o(\delta). \tag{2.1}$$

Now the numerical approximation $y^* \approx y(t^*)$ satisfies $y^* - z_n(t^*) = O(h^{*p+1})$, and hence $y^* - z_n(t^*) = o(\delta)$, so that, from (2.1),

$$y^* - y(t^*) = v(t^*)\delta + o(\delta), \tag{2.2}$$

showing that TP in the solution is maintained at t^*.

To examine the first derivative approximation $\eta_1'(t)$ we note from (1.5) that

$$\eta_1'(t^*) - y'(t^*) = z_n'(t^*) + \frac{\text{le}^*}{h^*} - y'(t^*),$$

where le* denotes the local error over the last step, $\text{le}^* := y^* - z_n(t^*)$. Hence

$$\begin{aligned} \eta_1'(t^*) - y'(t^*) &= f(t^*, z_n(t^*)) - f(t^*, y(t^*)) + \frac{\text{le}^*}{h^*} \\ &= f_y(t^*, y(t^*))v(t^*)\delta + \frac{\text{le}^*}{h^*} + o(\delta), \end{aligned} \tag{2.3}$$

using (2.1). Now the quantity le^*/h^* behaves like $O(h^{*p})$ as $h^* \to 0$, and hence will not necessarily be negligible compared with the first term in (2.3). The key point to note is that, as $\delta \to 0$, h^* will follow a decaying sawtooth pattern, changing discontinuously each time a meshpoint coincides with t^*. Hence le^*/h^* will not behave like a linear function of δ, and it

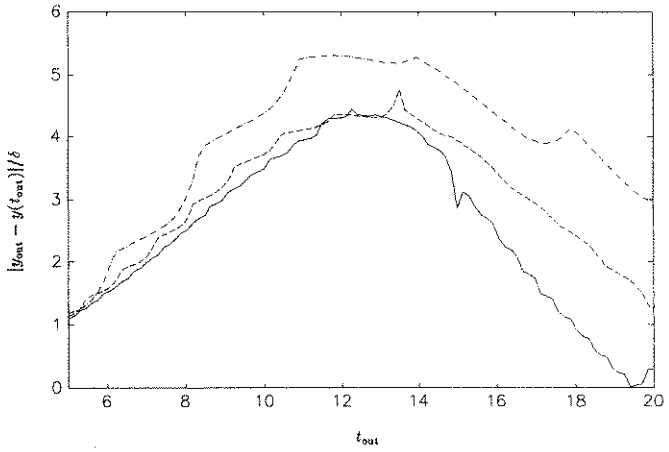

Fig. 1. Logistic equation. (Line type becomes more solid as δ decreases.)

follows that TP in $\eta_1'(t^*)$ cannot be guaranteed. However, the computable interpolants, $q(t)$, for which $q'(t^*) = f(t^*, y^*)$, satisfy

$$q'(t^*) - y'(t^*) = f(t^*, y^*) - f(t^*, y(t^*)) = f_y(t^*, y(t^*))v(t^*)\delta + o(\delta), \qquad (2.4)$$

using (2.2). The function $f_y(t, y(t))v(t)$ is independent of δ, and hence we have a proportionality result for $q'(t^*)$.

To summarise, in the case where the final stepsize is reduced to hit t^* exactly, TP in the solution approximation is retained at all points, TP in $\eta_1'(t)$ is lost at t^*, and TP in $q'(t)$ is introduced at t^*. A rather surprising consequence is that, for $t_0 < t \leqslant t^*$,

- $\eta_1'(t)$ gives TP at *all* t except $t = t^*$,
- $q'(t)$ gives TP at *no* t except $t = t^*$.

We illustrate these phenomena using the logistic equation (problem A4, unscaled, from DETEST [8])

$$y'(t) = \tfrac{1}{4}y(t)\left(1 - \tfrac{1}{20}y(t)\right), \qquad y(0) = 1, \qquad (2.5)$$

which has solution $y(t) = 20/(1 + 19\exp(-\frac{1}{4}t))$. We implemented the fourth- and fifth-order pair RK5(4)7FM of Dormand and Prince [4] in extrapolated error-per-step mode; that is, with the fifth-order formula advancing the solution and the difference between the fourth- and fifth-order values giving the error estimate. Mixed relative–absolute weights were used in the error measure. The code was made to reduce the final stepsize, if necessary, so as to hit the output point $t^* = t_{out}$ exactly. The problem was solved repeatedly using 100 equally spaced values of t^* in $[5, 20]$, and after each integration we recorded the norm of the global error in y^*, $f(t^*, y^*)$, and $\eta_1'(t^*)$. Since $\eta_1'(t^*)$ is not computable in general, we used the approximation $f(t^*, \tilde{u}(t^*)) + \widetilde{le}^*/h^*$. Here $\tilde{u}(t^*)$ is the result of a step from $\{t_{n-1}, y_{n-1}\}$ of length h^* using an eighth-order RK formula, and $\widetilde{le}^* = y^* - \tilde{u}(t^*)$. The tests were performed for error tolerances of $\delta = 10^{-5}, 10^{-7}, 10^{-9}$. The results are plotted in Figs. 1–3. (In these, and all subsequent figures, discrete values are joined by straight lines for clarity, and the line type changes from dashed/dotted to dashed to solid as the tolerance decreases.)

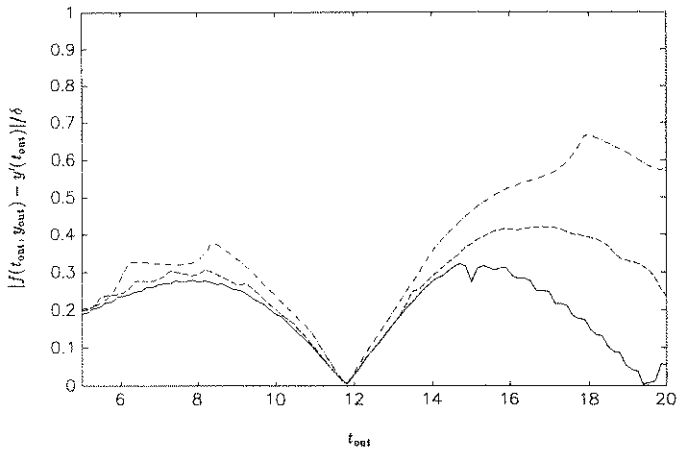

Fig. 2. Logistic equation.

We see from Fig. 1 that the global error to tolerance ratio for y^* appears to be converging to a discernible limit function as δ decreases. (The limit function is not the same as that in [11] for the same method and test problem, since here we are using mixed, rather than absolute, weights in the error measure.) The TP behaviour of $f(t^*, y^*)$, given in Fig. 2, is also reasonably good. For $\eta_1'(t^*)$, however, the ratio does not settle down to a limit. Comparing Figs. 2 and 3 we see that the $\eta_1'(t^*)$ ratios seem to correspond to those for $f(t^*, y^*)$ with "random" oscillations added. This is what we would expect from equations (2.3) and (2.4); the nonsmooth le^*/h^* term in (2.3) is clearly making its presence felt.

Next, we must consider what happens when the integration is *re-started* from the point t^*. This is essentially the same as applying the method to a new initial value problem, except that the initial value y^* is not exact, but satisfies (2.2). It can be shown that Theorem 1.1 extends to the case where the initial value has an error that is asymptotically linear in δ—a more general version of this result is proved in the next section. It follows that TP is maintained after crossing t^*, and, by induction, when a finite number of discontinuities are encountered.

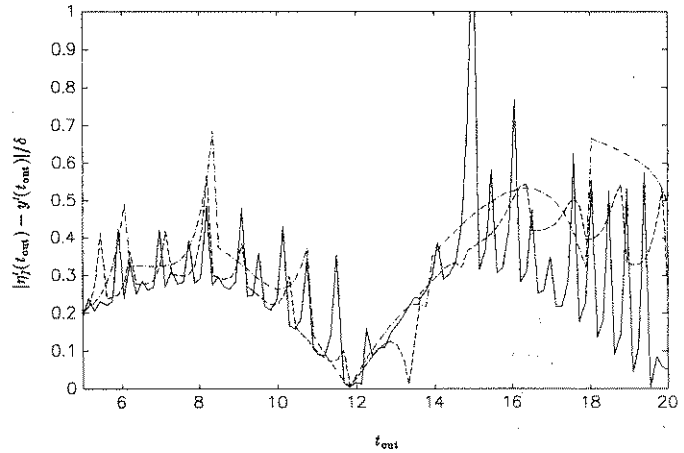

Fig. 3. Logistic equation.

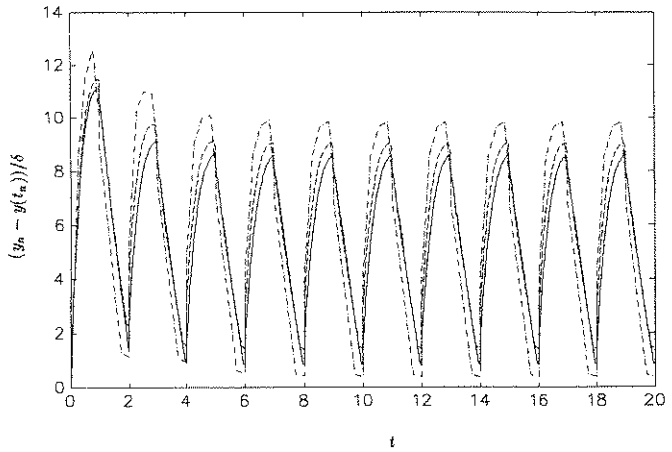

Fig. 4. Problem F2.

We illustrate this behaviour using problem F2 from DETEST, which is also used in [7],

$$y'(t) = \begin{cases} 55 - 1.5y(t) & \text{for } \lfloor t \rfloor \text{ even}, \\ 55 - 0.5y(t) & \text{for } \lfloor t \rfloor \text{ odd}, \end{cases} \quad 0 \leqslant t \leqslant 20,$$

$$y(0) = 110.$$

(Here $\lfloor t \rfloor$ denotes the integer part of t.) We see that there are discontinuities at $t^* \in \{i\}_{i=1}^{20}$. The problem was solved with the RK algorithm described earlier, except that the stepsize selection was altered so that rather than including the points of discontinuity in the mesh, we crossed them with stepsizes of $o(\delta)$. (This was done in attempt to model the more realistic situation where the discontinuities are not known exactly [7].) Figure 4 records the global error to tolerance ratios at the meshpoints for $\delta = 10^{-5}$, 10^{-7}, 10^{-9}. In Fig. 5 we present the corresponding picture when the standard stepsize selection strategy was not changed. In the former case the ratios appear to be approaching a limit, whereas in the latter case the errors are much larger and do not settle down.

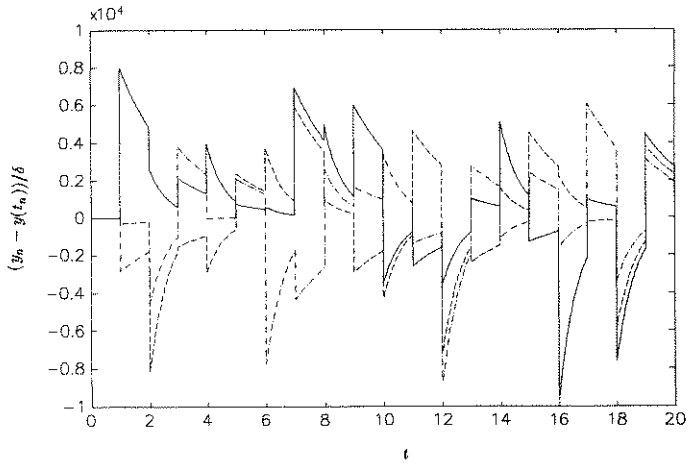

Fig. 5. Problem F2.

3. ODEs with constant delays

One of the simplest examples of a delay differential equation is given by

$$y'(t) = y(t-1), \quad t \geqslant 0, \tag{3.1}$$
$$y(t) = 1, \quad t \in [-1, 0].$$

Here we see that $y(t) = t + 1$ over $[0, 1]$, and in general $y(t)$ is a polynomial of degree i over $[i-1, i]$. Also, notice that $y'(t)$ has a jump discontinuity at $t = 0$ and this discontinuity is propagated in such a way that $y^{(i)}(t)$ is discontinuous at $t = i - 1$.

The general problem considered here is a system of ODEs with k constant delays, which we write as

$$y'(t) = F(t, y(t), y(t-\tau_1), y(t-\tau_2), \ldots, y(t-\tau_k)) \in \mathbb{R}^N, \quad 0 \leqslant t \leqslant t_{\text{end}}, \tag{3.2}$$
$$y(t) = \Phi(t), \quad t \in [-\tau_k, 0].$$

We assume that the delays are ordered so that $0 < \tau_1 < \tau_2 < \cdots < \tau_k$, and that the initial function, $\Phi(t)$, has $p + 1$ continuous derivatives. As we noted in the example (3.1), if $\Phi(t)$ does not match $y(t)$ smoothly at $t = 0$, then derivative discontinuities will be propagated throughout the solution. It can be shown that $y'(t)$ is generally discontinuous at $t = 0$ and that a discontinuity in $y^{(i)}(t)$ at $t = t^*$ leads to a possible discontinuity in $y^{(i+1)}(t)$ at $t = t^* + \tau_j$, for $j = 1, 2, \ldots, k$. (For an analysis of the location and order of discontinuities in more general classes of delay ODEs, see [14,21,22].)

We assume that F in (3.2) is a smooth function of each of its arguments, and in particular that if $r(t)$ is a given function with $p + 1$ continuous derivatives, then the standard initial value problem (IVP) $y'(t) = F(t, y(t), r(t-\tau_1), r(t-\tau_2), \ldots, r(t-\tau_k))$, $y(0) = y_0$, is sufficiently smooth for the expansions (1.2) and (1.3) to hold for any initial value y_0.

Now due to the discontinuity propagation in (3.2), it follows that there exist a finite number of points $\{\hat{t}_i\}_{i=1}^m$ such that $0 < \hat{t}_1 < \hat{t}_2 < \cdots < \hat{t}_m$ and $y(t)$ has $p + 1$ continuous derivatives over each subinterval $(\hat{t}_i, \hat{t}_{i+1})$, and also over $(\hat{t}_m, \infty)$. Moreover, the discontinuity points $\hat{t}_i$ can be computed a priori.

The most natural approach for solving (3.2) numerically is to use an interpolation procedure to approximate the retarded values, $y(t-\tau_i)$, and then to apply a standard ODE method to the resulting IVP (see [1,10,12,13,15–17] for examples). Here, we assume that an explicit RK method is used, with error control and stepsize selection as described in Section 1, and with a corresponding interpolant. In other words, we apply the RK method to the ODE

$$y^{q\prime}(t) = F(t, y^q(t), q(t-\tau_1), q(t-\tau_2), \ldots, q(t-\tau_k)), \quad 0 \leqslant t \leqslant t_{\text{end}}, \tag{3.3}$$
$$y^q(0) = \Phi(0),$$

where $q(t) := \Phi(t)$ for $t \in [-\tau_k, 0]$, and for $t > 0$, $q(t)$ denotes either a higher- or lower-order interpolant to the discrete approximation, as described in Section 1. (The superscript q emphasises that y^q depends upon q, and therefore upon the error tolerance δ.) Note that since we are concerned with an $h_n \to 0$ analysis, we may assume that on a general step from t_{n-1} to t_n, the retarded values needed by the RK scheme lie to the left of t_{n-1}, and hence interpolation (rather than extrapolation) can be used. We will suppose that the discontinuity points $\hat{t}_i$ are located a priori and incorporated into the mesh. Our aim is to examine what conditions on the interpolation process are necessary/sufficient to guarantee tolerance proportionality.

We note immediately that the first smooth subinterval will be $(0, \tau_1)$ and that on this subinterval we are, in effect, solving the standard IVP

$$\begin{aligned} &y'(t) = F(t,\ y(t),\ \Phi(t-\tau_1),\ \Phi(t-\tau_2),\ldots,\Phi(t-\tau_k)),\\ &y(0) = \Phi(0). \end{aligned} \tag{3.4}$$

Since this ODE does not depend upon δ, the results mentioned in Section 1 apply directly, and in particular we conclude that higher-order interpolants will satisfy (1.7) over $(0, \hat{t}_1)$ while lower-order interpolants give $q(t) - y(t) = O(\delta)$, but do not give (1.7) in general.

Now suppose that we re-start at $\hat{t}_1 (= \tau_1)$. To proceed with the analysis we define the local solution over a general step from t_{n-1} to t_n by

$$\begin{aligned} &z_n^{q\prime}(t) = F(t,\ z_n^q(t),\ q(t-\tau_1),\ q(t-\tau_2),\ldots,q(t-\tau_k)),\\ &z_n^q(t_{n-1}) = y_{n-1} \end{aligned}, \tag{3.5}$$

and the local error at t_n by

$$\mathrm{le}_n^q = y_n - z_n^q(t_n).$$

The corresponding ideal interpolant can then be defined by

$$\eta_I^q(t) := z_n^q(t) + \frac{(t-t_{n-1})}{h_n}\mathrm{le}_n^q, \quad t \in (t_{n-1},\ t_n].$$

Our approach is to examine the global error $\eta_I^q(t) - y(t)$ over $(\hat{t}_1, \hat{t}_2)$ by splitting it into two components, $y^q(t) - y(t)$ and $\eta_I^q(t) - y^q(t)$. First we look at $y^q(t) - y(t)$, and show that with higher-order interpolation if we regard $y^q(t)$ as an approximation to $y(t)$ then condition (B) of Theorem 1.1, and hence also condition (A), is satisfied.

Using $f^y(t,\ r(t))$ to denote $F(t,\ r(t),\ y(t-\tau_1),\ y(t-\tau_2),\ldots,y(t-\tau_k))$, for a given function $r(t)$, we have

$$\begin{aligned} y^{q\prime}(t) - f^y(t,\ y^q(t)) = {} & F(t,\ y^q(t),\ q(t-\tau_1),\ q(t-\tau_2),\ldots,q(t-\tau_k))\\ & - F(t,\ y^q(t),\ y(t-\tau_1),\ y(t-\tau_2),\ldots,y(t-\tau_k)). \end{aligned} \tag{3.6}$$

Hence $y^{q\prime}(t) - f^y(t,\ y^q(t)) = O(\max_i \| q(t-\tau_i) - y(t-\tau_i)\|)$. Since we solved a standard IVP (3.4) over the first subinterval, we know that $O(\max_i \| q(t-\tau_i) - y(t-\tau_i)\|) = O(\delta)$ for either higher- or lower-order interpolants. Using this in (3.6) it follows from a standard differential inequality (see, for example, [10, Theorem 10.2, p. 56]) that $\| y^q(t) - y(t)\| = O(\delta)$. Further, writing $y^q(t) = y(t) + (y^q(t) - y(t))$ and $q(t-\tau_i) = y(t-\tau_i) + (q(t-\tau_i) - y(t-\tau_i))$ in (3.6) and expanding, we find that

$$\begin{aligned} &y^{q\prime}(t) - f^y(t,\ y^q(t))\\ &\quad = \sum_{i=1}^{k} \frac{\partial F}{\partial z_i}(t,\ y(t),\ y(t-\tau_1),\ldots,y(t-\tau_k))(q(t-\tau_i) - y(t-\tau_i))\\ &\qquad + O\Big(\max_i \| q(t-\tau_i) - y(t-\tau_i)\|^2\Big)\\ &\qquad + O\Big(\| y(t) - y^q(t)\| \max_i \| q(t-\tau_i) - y(t-\tau_i)\|\Big)\\ &\qquad + O\big(\| y(t) - y^q(t)\|^2\big), \end{aligned}$$

where $\partial F/\partial z_i$ denotes the partial derivative of $F(t, y, z_1, z_2, \ldots, z_k)$ with respect to z_i. It follows that

$$y^{q\prime}(t) - f^y(t, y^q(t)) = \sum_{i=1}^{k} \frac{\partial F}{\partial z_i}(t, y(t), y(t-\tau_1), \ldots, y(t-\tau_k))(q(t-\tau_i) - y(t-\tau_i)) + O(\delta^2). \quad (3.7)$$

Now, with a higher-order interpolant we have $q(t-\tau_i) - y(t-\tau_i) = v_i(t)\delta + o(\delta)$, where $v_i(t) := v(t-\tau_i)$ is continuous and independent of δ, and hence from (3.7)

$$y^{q\prime}(t) - f^y(t, y^q(t)) = \Gamma(t)\delta + o(\delta),$$

where $\Gamma(t)$ is continuous and independent of δ, and the $o(\delta)$ remainder is clearly continuous. We may thus apply the equivalence result of Theorem 1.1 to deduce that

$$y^q(t) - y(t) = V(t)\delta + G(t), \quad (3.8)$$

where $V(t)$ is C^1 and independent of δ, and $G(t)$ is piecewise C^1 with zeroth and first derivatives of $o(\delta)$.

For the lower-order interpolant, however, we know that, in general, $q(t-\tau_i) - y(t-\tau_i)$ does not behave linearly (asymptotically) as a function of δ, and hence (3.8) does not hold in general.

Next we show that the error control method causes the ideal interpolant $\eta_I^q(t)$ to give TP relative to the "approximate" true solution $y^q(t)$. To do this we generalise Theorem 1.1 to allow for the fact that $y^q(t)$ depends upon δ.

Theorem 3.1. *Recall that $y(t)$ is the solution to* (3.2), *and let $y^q(t)$ be the solution to* (3.3). *Let $\eta(t)$ be a piecewise C^1 approximation to $y^q(t)$, and let $\varepsilon(t) := \eta(t) - y^q(t)$ denote the corresponding error. Suppose that*

$$\varepsilon(\hat{t}_1) = K\delta + o(\delta), \quad (3.9)$$

where K is a constant vector. Then the following conditions are equivalent:

(A) *$\varepsilon(t) = v(t)\delta + g(t)$, $t \in (\hat{t}_1, \hat{t}_2)$, where $v(t)$ is C^1 and independent of δ, and $g(t)$ is piecewise C^1 with zeroth and first derivatives of* $o(\delta)$.

(B) *$\eta'(t) - F(t, \eta(t), q(t-\tau_1), q(t-\tau_2), \ldots, q(t-\tau_k)) = \gamma(t)\delta + s(t)$, $t \in (\hat{t}_1, \hat{t}_2)$, where $\gamma(t)$ is continuous and independent of δ, and $s(t)$ is piecewise continuous and* $o(\delta)$.

Proof. The proof is based on the proof of [11, Theorem 2.1]. We introduce a third condition, (C), and then prove that (A) $\Rightarrow$ (B), (B) $\Rightarrow$ (C), and (C) $\Rightarrow$ (A).

(C) $\varepsilon'(t) - F_y(t, y^q(t), q(t-\tau_1), q(t-\tau_2), \ldots, q(t-\tau_k))\varepsilon(t) = \gamma(t)\delta + u(t)$, where $\gamma(t)$ is the function appearing in condition (B), and $u(t)$ is piecewise continuous and $o(\delta) + O(\varepsilon(t)^2)$. (Here $F_y(t, y, z_1, z_2, \ldots, z_k)$ denotes the partial derivative of $F(t, y, z_1, z_2, \ldots, z_k)$ with respect to y.)

(A) $\Rightarrow$ (B): Writing $\eta(t) = y^q(t) + \varepsilon(t)$ we have

$$\eta'(t) - F(t, \eta(t), q(t-\tau_1), \ldots, q(t-\tau_k)) = \varepsilon'(t) - F_y(t, y^q(t), q(t-\tau_1), \ldots, q(t-\tau_k))\varepsilon(t) + w(t),$$

where $w(t) = O(\varepsilon(t)^2)$. Hence, since $q(t-\tau_i) - y(t-\tau_i) = O(\delta)$ and $y^q(t) - y(t) = O(\delta)$,

$$\begin{aligned}&\eta'(t) - F(t, \eta(t), q(t-\tau_1), \ldots, q(t-\tau_k))\\ &\quad = \varepsilon'(t) - F_y(t, y(t), y(t-\tau_1), \ldots, y(t-\tau_k))\varepsilon(t) + \bar{w}(t),\end{aligned}$$

where $\bar{w}(t) = O(\varepsilon(t)\delta + \varepsilon(t)^2)$. Finally, using (A),

$$\begin{aligned}&\eta'(t) - F(t, \eta(t), q(t-\tau_1), \ldots, q(t-\tau_k))\\ &\quad = \delta\big[v'(t) - F_y(t, y(t), y(t-\tau_1), \ldots, y(t-\tau_k))v(t)\big]\\ &\qquad + g'(t) - F_y(t, y(t), y(t-\tau_1), \ldots, y(t-\tau_k))g(t) + \bar{w}(t),\end{aligned}$$

which has the required form.

(B) $\Rightarrow$ (C): This follows as in the proof of [11, Theorem 2.1].

(C) $\Rightarrow$ (A): Let $v(t)$ denote the unique solution to the linear initial value problem

$$\begin{aligned}&v'(t) - F_y(t, y(t), y(t-\tau_1), y(t-\tau_2), \ldots, y(t-\tau_k))v(t) = \gamma(t),\\ &v(\hat{t}_1) = K.\end{aligned} \tag{3.10}$$

From (C), we have

$$\varepsilon'(t) - F_y(t, y(t), y(t-\tau_1), y(t-\tau_2), \ldots, y(t-\tau_k))\varepsilon(t) = \gamma(t)\delta + \bar{u}(t), \tag{3.11}$$

where $\bar{u}(t)$ is piecewise continuous and $o(\delta) + O(\varepsilon(t)^2)$. From (3.10) and (3.11), $\varepsilon(t) - \delta v(t)$ satisfies

$$(\varepsilon(t) - \delta v(t))' - F_y(t, y(t), y(t-\tau_1), \ldots, y(t-\tau_k))(\varepsilon(t) - \delta v(t)) = \bar{u}(t).$$

Standard theory (see, for example, [2, p. 86]) shows that this initial value problem has a solution of the form

$$\varepsilon(t) - \delta v(t) = Y(t)\left[\varepsilon(\hat{t}_1) - \delta v(\hat{t}_1) + \int_{\hat{t}_1}^{t} Y^{-1}(\mu)\bar{u}(\mu)\,\mathrm{d}\mu\right], \tag{3.12}$$

where the fundamental solution matrix $Y(t)$ is defined by

$$\begin{aligned}&Y'(t) = F_y(t, y(t), y(t-\tau_1), y(t-\tau_2), \ldots, y(t-\tau_k))Y(t),\\ &Y(\hat{t}_1) = I.\end{aligned}$$

Note that $Y(t)$ is independent of δ, and that $\varepsilon(\hat{t}_1) - \delta v(\hat{t}_1) = \varepsilon(\hat{t}_1) - \delta K = o(\delta)$. It follows that

$$\varepsilon(t) - \delta v(t) = g(t),$$

where $g(t)$ is $o(\delta) + O(\varepsilon(t)^2)$ and continuous, and $g'(t)$ is $o(\delta) + O(\varepsilon(t)^2)$ and piecewise continuous, giving the desired result. □

Now the "approximate" problem $y^{q\prime}(t) = F(t, y^q(t), q(t-\tau_1), q(t-\tau_2), \ldots, q(t-\tau_k))$ is the one that the RK method is actually being asked to solve. We would like to apply the standard ODE analysis in [11] in order to conclude that the error control causes condition (B)

to hold. There is, however, one complication—the (higher- or lower-order) interpolant $q(t)$ is typically only a C^1 function, and hence the approximate problem is not smooth enough for (1.2) and (1.3) to hold. We can sidestep this difficulty by noting that the RK process samples F at a discrete set of points. For each δ, we could replace $q(t)$ by a smoother function that interpolates $q(t)$ at these discrete points and the numerical solution would remain unchanged. Hence we may "pretend" that $q(t)$ is globally C^{p+1}. It follows that condition (B) in Theorem 3.1 is satisfied, allowing us to deduce the desired result.

Corollary 3.2. *Suppose that we solve* (3.2) *over* $[\hat{t}_1, \hat{t}_2]$ *in the manner described above, using either a higher- or lower-order interpolant. Then, provided that given* $y(t)$ *for* $t \leqslant \hat{t}_1$, $\|\tilde{\psi}(y(t), t)\| \neq 0$ *over* $[\hat{t}_1, \hat{t}_2]$ *in* (1.3), *the ideal interpolant satisfies*

$$\eta_I^q(t) - y^q(t) = \tilde{V}(t)\delta + \tilde{G}(t), \tag{3.13}$$

where $\tilde{V}(t)$ *is continuous and independent of* δ, $\tilde{V}(t) \in C^1(\hat{t}_1, \hat{t}_2)$, *and* $\tilde{G}(t)$ *is piecewise* C^1 *with zeroth and first derivatives of* $o(\delta)$.

When a higher-order interpolant is used we may thus combine (3.8) and (3.13) to give

$$\eta_I^q(t) - y(t) = \overline{V}(t)\delta + \overline{G}(t), \tag{3.14}$$

where $\overline{V}(t)$ is continuous and independent of δ, $\overline{V}(t) \in C^1(\hat{t}_1, \hat{t}_2)$, and $\overline{G}(t)$ is piecewise C^1 with zeroth and first derivatives of $o(\delta)$. On the other hand, with a lower-order interpolant we see that since (3.8) does not hold, in general the leading term in $\eta_I^q(t) - y(t)$ will not be linear.

Now on a general step from t_{n-1} to t_n in the integration over $[\hat{t}_1, \hat{t}_2]$ we have $\eta_I^q(t) - z_n^q(t) = O(h_n^{p+1})$ and, for a higher-order interpolant, $q(t) - z_n^q(t) = O(h_n^{p+1})$. Hence $q(t) - \eta_I^q(t) = O(h_n^{p+1})$, so that $q(t) - \eta_I^q(t) = o(\delta)$ and, using (3.14),

$$q(t) - y(t) = \overline{V}(t)\delta + o(\delta). \tag{3.15}$$

This shows that a higher-order interpolant maintains TP in the $y(t)$ approximation across $[\hat{t}_1, \hat{t}_2]$. By induction, (3.14) and (3.15) remain true when a finite number of smooth subintervals are crossed. The induction is valid provided that the tail of backvalues never crosses into the current subinterval; that is, $\hat{t}_{i+1} - \hat{t}_i \leqslant \tau_1$. There are two cases where this condition does not hold. First, if the coupling in (3.2) is weak, we may be able to take smooth subintervals with width bigger than τ_1. Second, the integration may proceed into the final smooth region $(\hat{t}_m, \infty)$. We will show how to deal with the second case. (The first case can be handled similarly.) Given any fixed point $t > \hat{t}_m$, let t_{N_1} be the furthermost meshpoint such that $t_{N_1} - \hat{t}_m \leqslant \tau_1$, and in general let t_{N_r} be the furthermost meshpoint such that $t_{N_r} - t_{N_{r-1}} \leqslant \tau_1$. In this manner the range $[\hat{t}_m, t]$ can be divided into a finite number of subintervals of width $\leqslant \tau_1$. Now we can inductively obtain (3.14) and (3.15) over each subinterval, so that eventually

$$q(t) - y(t) = \overline{V}(t)\delta + o(\delta) \tag{3.16}$$

at the given point t.

To verify the analysis, we implemented a three-stage, second- and third-order RK pair of Fehlberg [10, p. 170] in extrapolated error-per-step mode, so that $p = 3$. Two alternatives were used for the interpolant q. First, a piecewise quadratic interpolant defined over $[t_{i-1}, t_i]$ by $q(t_{i-1}) = y_{i-1}$, $q'(t_{i-1}) = f(t_{i-1}, y_{i-1})$, and $q(t_i) = y_i$ was implemented. Here $q(t)$ has local order $l = 3 = p$, so we have a lower-order interpolation scheme. The resulting method will be

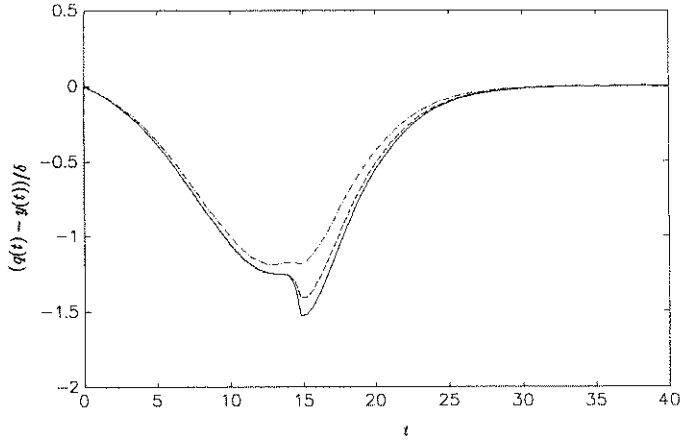

Fig. 6. Delay logistic equation: P3L4 method.

denoted P3L3. Second, we used the standard two-point cubic Hermite interpolant which satisfies $q(t_{i-1}) = y_{i-1}$, $q'(t_{i-1}) = f(t_{i-1}, y_{i-1})$, $q(t_i) = y_i$, and $q'(t_i) = f(t_i, y_i)$. In this case the local order of q is $l = 4 = p + 1$, and the interpolant is of higher order. This method will be denoted P3L4. We also implemented a $p = 4$, $l = 4$ method, which we refer to as P4L4, consisting of the third- and fourth-order RK pair from [6], in extrapolated error-per-step mode, along with the cubic interpolant above. We mention that the $p = 4$, $l = 4$ combination has proved to be a popular choice [12,13].

The algorithms were tested on the logistic equation

$$y'(t) = \tfrac{1}{4}y(t)\big(1 - \tfrac{1}{20}y(t-1)\big), \quad t \geqslant 0,$$
$$y(t) = 1, \quad t \in [-1, 0],$$

which is a delayed analogue of (2.5). Equations of this type arise in the study of population dynamics [10, p. 292]. The global error to tolerance behaviour of the three methods for

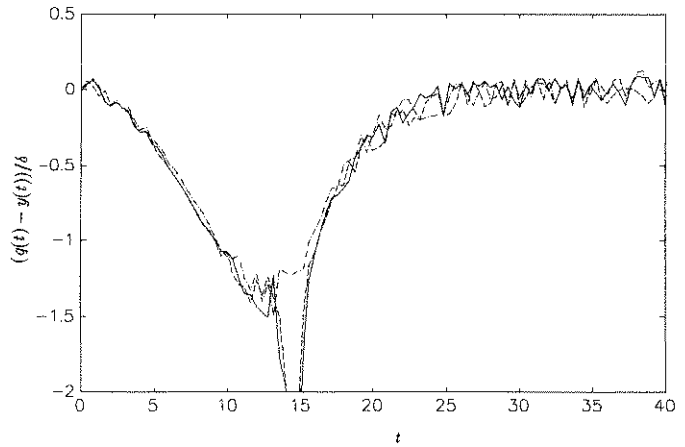

Fig. 7. Delay logistic equation: P3L3 method.

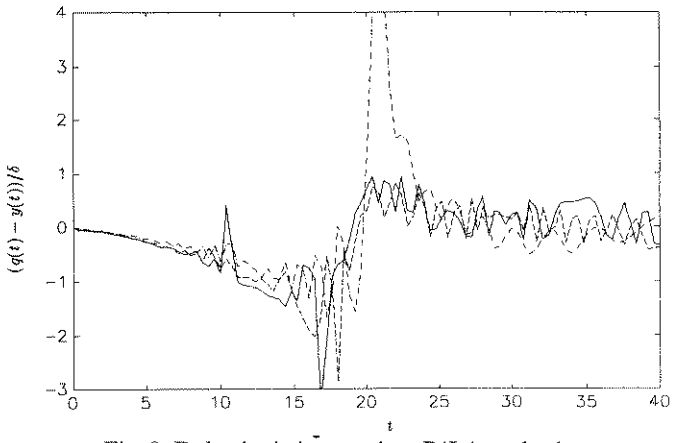

Fig. 8. Delay logistic equation: P4L4 method.

$\delta = 10^{-4}, 10^{-6}, 10^{-8}$ is plotted in Figs. 6–8. Here the global error in $q(t)$ was computed at 101 equally spaced points in the range [0, 40]. The numerical solution with $\delta = 10^{-10}$ was taken to be the "true" solution in each case. We see in Fig. 6 that the ratios for the P3L4 method behave smoothly and appear to approach a limit function. For P3L3, in Fig. 7, although the behaviour is similar, the ratios do not seem to converge to a limit, but rather oscillate about a fixed curve. This illustrates the potential difference in behaviour between higher- and lower-order interpolants that our analysis predicted. The P4L4 method in Fig. 8 also exhibits nonlinear variation of global error to tolerance.

The methods were also tested on a disease propagation model [10, p. 295]. Here, all three methods displayed good tolerance proportionality—for P3L3 and P4L4 the nonlinear effects, although $O(\delta)$, were not sufficiently large to be visible. (Similar behaviour on certain ODEs was observed in [11].)

We mention that Bellen and Zennaro [23] investigated higher- and lower-order interpolation in a slightly different context. Those authors also found that higher-order interpolants give significant advantages.

4. Discussion

The main conclusion of this work is that when a pth-order Runge–Kutta formula is used to solve a constant delay system of ODEs, higher-order (locally $O(h_n^{p+1})$) interpolation is necessary and sufficient to guarantee asymptotic tolerance proportionality. It is perhaps surprising, therefore, that several of the algorithms that have been put forward in the literature use lower-order (locally $O(h_n^p)$) interpolants (see, for example, [10,12,13]). A possible explanation for this is that lower-order interpolation can be justified via a classical fixed stepsize analysis. Oppelstrup [16] and Roth [17] state that with a fixed stepsize and lower-order interpolant, the global error behaves like $O(h^p)$, which is, of course, the best order that can be achieved. For the variable stepsize case analysed here, this corresponds to the fact that lower-order interpolants allow the global error to be *bounded* linearly with δ. To convert the bound into an equality, higher-order interpolation must be performed.

Acknowledgements

This work began when I was at the University of Toronto, under the support of the Information Technology Research Centre of Ontario and the Natural Sciences and Engineering Research Council of Canada, and was completed at the University of Dundee under the support of the University of Dundee Research Initiatives Fund. I thank Nick Higham for commenting on this manuscript.

References

[1] H. Arndt, Numerical solution of retarded initial value problems with local and global error and stepsize control, *Numer. Math.* 43 (1984) 343–360.
[2] U.M. Ascher, R.M.M. Mattheij and R.D. Russell, *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations* (Prentice-Hall, Englewood Cliffs, NJ, 1988).
[3] J.R. Cash, A block 6(4) Runge–Kutta formula for nonstiff initial value problems, *ACM Trans. Math. Software* 15 (1989) 15–28.
[4] J.R. Dormand and P.J. Prince, A family of embedded Runge–Kutta formulae, *J. Comput. Appl. Math.* 6 (1980) 19–26.
[5] J.R. Dormand and P.J. Prince, Runge–Kutta triples, *Comput. Math. Appl.* 12 (1986) 1007–1017.
[6] W.H. Enright, K.R. Jackson, S.P. Nørsett and P.G. Thomsen, Interpolants for Runge–Kutta formulas, *ACM Trans. Math. Software* 12 (1986) 193–218.
[7] W.H. Enright, K.R. Jackson, S.P. Nørsett and P.G. Thomsen, Effective solution of discontinuous IVPs using a Runge–Kutta formula pair with interpolants, *Appl. Math. Comput.* 27 (1988) 313–335.
[8] W.H. Enright and J.D. Pryce, Two FORTRAN packages for assessing initial value methods, *ACM Trans. Math. Software* 13 (1987) 1–27.
[9] C.W. Gear and O. Østerby, Solving ordinary differential equations with discontinuities, *ACM Trans. Math. Software* 10 (1984) 23–44.
[10] E. Hairer, S.P. Nørsett and G. Wanner, *Solving Ordinary Differential Equations I* (Springer, Berlin, 1987).
[11] D.J. Higham, Global error versus tolerance for explicit Runge–Kutta methods, *IMA J. Numer. Anal.* 11 (1991) 457–480.
[12] K.W. Neves, Automatic integration of functional differential equations: an approach, *ACM Trans. Math. Software* 1 (1975) 357–368.
[13] K.W. Neves, Control of interpolatory error in retarded differential equations, *ACM Trans. Math. Software* 7 (1981) 421–444.
[14] K.W. Neves and A. Feldstein, Characterization of jump discontinuities for state dependent delay differential equations, *J. Math. Anal. Appl.* 56 (1976) 689–707.
[15] H.J. Oberle and H.J. Pesch, Numerical treatment of delay differential equations by Hermite interpolation, *Numer. Math.* 37 (1981) 235–255.
[16] J. Oppelstrup, The RKFHB4 method for delay-differential equations, in: R. Burlisch, R.D. Grigorieff and J. Schröder, eds., *Numerical Treatment of Differential Equations: Proceedings Oberwolfach, 1976*, Lecture Notes in Mathematics 631 (Springer, Berlin, 1978) 133–146.
[17] M.G. Roth, Difference methods for stiff delay differential equations, Ph.D. Thesis, Tech. Report UIUCDCS-R-80-1012, Department of Computer Science, University of Illinois at Urbana-Champagne, IL (1980).
[18] H.J. Stetter, Considerations concerning a theory for ODE-solvers, in: R. Burlisch, R.D. Grigorieff and J. Schröder, eds., *Numerical Treatment of Differential Equations: Proceedings Oberwolfach, 1976*, Lecture Notes in Mathematics 631 (Springer, Berlin, 1978) 188–200.
[19] H.J. Stetter, Interpolation and error estimates in Adams PC-codes, *SIAM J. Numer. Anal.* 16 (1979) 311–323.
[20] H.J. Stetter, Tolerance proportionality in ODE-codes, in: R. März, ed., *Proceedings Second Conference on Numerical Treatment of Ordinary Differential Equations*, Seminarberichte No. 32, Humboldt University, Berlin (1980) 109–123; also in: R.D. Skeel, ed., *Working Papers for the 1979 SIGNUM Meeting on Numerical Ordinary*

Differential Equations, Department of Computer Science, University of Illinois at Urbana-Champagne, IL (1979).
[21] D.R. Willé and C.T.H. Baker, The propagation of derivative discontinuities in systems of delay-differential equations, Numerical Analysis Report 160, University of Manchester, Manchester, England (1988).
[22] D.R. Willé and C.T.H. Baker, The tracking of derivative discontinuities in systems of delay-differential equations, Numerical Analysis Report 185, University of Manchester, Manchester, England (1990).
[23] A. Bellen and M. Zennaro, Numerical solution of delay differential equations by uniform corrections to an implicit Runge–Kutta method, *Numer. Math.* 47 (1985) 301–316.