Advances in Computational Mathematics 7 (1997) 361–382

Stepsize selection for tolerance proportionality in explicit Runge–Kutta codes

M. Calvo [a], D.J. Higham [b], J.I. Montijano [a] and L. Rández [a]

[a] *Departamento de Matemática Aplicada, Universidad de Zaragoza, 50009 Zaragoza, Spain*
[b] *Department of Mathematics, University of Strathclyde, Glasgow G1 1XH, Scotland, UK*

Received June 1994; revised December 1996
Communicated by W.H. Enright

The potential for adaptive explicit Runge–Kutta (ERK) codes to produce global errors that decrease linearly as a function of the error tolerance is studied. It is shown that this desirable property may not hold, in general, if the leading term of the locally computed error estimate passes through zero. However, it is also shown that certain methods are insensitive to a vanishing leading term. Moreover, a new stepchanging policy is introduced that, at negligible extra cost, ensures a robust global error behaviour. The results are supported by theoretical and numerical analysis on widely used formulas and test problems. Overall, the modified stepchanging strategy allows a strong guarantee to be attached to the complete numerical process.

Keywords: initial value problems, Runge–Kutta methods, stepsize control

AMS subject classification: 65L05

1. Introduction

What guarantees do we have for the global error of our ordinary differential equation (ODE) initial value software? Standard codes input a tolerance parameter, δ, that is intended to control the accuracy of the computation – reducing δ should lead to a more accurate result. It is clearly important to understand the precise effect of changing δ, and to derive the strongest possible results about the behaviour of the global error as a function of δ. Stetter [13,14] studied this problem and derived conditions under which the global error is asymptotically linear in δ, as $\delta \to 0$. Further work on explicit Runge–Kutta (ERK) methods appears in [9]. The analysis in these references makes use of asymptotic expansions, and, in particular, the results apply only when the leading term in the error estimate is bounded away from zero. In this work we show that the leading term can pass through zero with common methods and problems, and we confirm numerically that this has a deleterious effect on the global error behaviour. On the other hand, we show that certain pairs of ERK formulas have a desirable property that helps to suppress these difficulties. Finally, we show that by altering the stepsize selection policy, adaptive ERK methods can be made insensitive

to the effect of a vanishing leading term in the error estimate. Our philosophy is that widely-used software should satisfy the strongest possible reliability conditions. Hence, we believe that it is extremely worthwhile to develop time-stepping algorithms for which positive results can be proved.

In the remainder of this section we briefly introduce the definitions and results that motivate this work. Further details can be found in [9].

Adaptive ERK methods for an initial value system

$$y'(t) = f\big(t, y(t)\big), \quad y(0) = y_0, \qquad 0 \leqslant t \leqslant t_{\text{end}}, \tag{1.1}$$

where f is assumed to be sufficiently smooth, generate a discrete set of approximations $y_n \approx y(t_n)$ to the solution of (1.1) on a nonuniform grid $\{t_n\}$ in $[0, t_{\text{end}}]$. The stepsizes $h_n := t_n - t_{n-1},\ n = 1, 2, \dots$, are chosen dynamically in an attempt to satisfy the user's accuracy requirement. Moreover, most recent ERK methods are able to produce a continuous approximate solution which is given by a piecewise interpolant to the underlying discrete approximation.

Throughout this paper, we will assume that the basic elements of an adaptive ERK method are a main advancing explicit formula $(t_{n-1}, y_{n-1}) \to (t_n, y_n)$ of order p (perhaps with a continuous extension) together with a locally-based *error estimator* $e_n = e(t_{n-1}, y_{n-1}, h_n)$ which is computed in the step $t_{n-1} \to t_n$ and is used to control the behaviour of the error in the current step and to provide an optimal selection of the next step according to the user supplied tolerance parameter δ. Further we suppose that the error estimator possesses an asymptotic expansion of the form

$$e_n = e(t_{n-1}, y_{n-1}, h_n) = h_n^p \widetilde{\psi}(t_{n-1}, y_{n-1}) + \mathrm{O}\big(h_n^{p+1}\big). \tag{1.2}$$

Let us recall that e_n is used to control the error and to monitor the stepsize in the following way. First, for some given norm $\|\cdot\|$, $\text{est}_n := \|e_n\|$, gives the *error estimate* for the step, which is compared with δ. If $\text{est}_n \leqslant \delta$ then the approximation y_n is accepted, otherwise the step is re-taken with a smaller stepsize. The standard formula for changing stepsize after a successful step is

$$h_{n+1} = \theta \left(\frac{\delta}{\text{est}_n} \right)^{1/p} h_n, \tag{1.3}$$

with $\theta \in (0, 1)$ a constant safety factor. After a rejected step, this formula can be used to give a new stepsize with which to re-take the step, or some other rule can be applied. The precise details of step rejections are not important for our analysis.

Next let us note that many adaptive ERK methods fit into this framework. First, in the case of a pair of formulas of orders p and $p-1$ with extrapolated error-per-step control, the advancing formula is the higher order one and the local error estimator e_n is the difference between the solutions provided by the two formulas. Second, in the case of a pair of formulas of orders p and $q > p$ with an error-per-unit-step control, the advancing formula has order p and the local error estimator is given as h_n^{-1} times the difference between the solutions of the two formulas. Finally, we consider error and stepsize control based on the defect; as described, for example, in [4] and

the references cited therein. If, over each interval $[t_{n-1}, t_n]$, the method provides a continuous extension $\bar{z}_n(t)$ of order p that interpolates the underlying discrete solution, then the defect $\delta_n(t)$ defined by

$$\delta_n(t) = \bar{z}_n'(t) - f\big(t, \bar{z}_n(t)\big), \quad t \in [t_{n-1}, t_n],$$

admits an expansion in powers of $h_n = t_n - t_{n-1}$ with a leading term of order $O(h_n^p)$. For the error estimator we may then take a single sampled value $\delta_n(t_{n-1} + \tau^* h_n)$ with a suitable, fixed choice of $\tau^* \in (0, 1)$.

Next we define tolerance proportionality (TP).

Definition. Suppose that a numerical method produces a discrete approximation $\{t_n, y_n\}$ for any sufficiently small tolerance δ. Then the method is said to exhibit *tolerance proportionality* on (1.1) if there exists an interpolating function $\eta_\delta(t)$, defined for each δ, such that $\eta_\delta(t_n) = y_n$ for all n and

$$\eta_\delta(t) - y(t) = v(t)\delta + g_\delta(t), \quad t \in [0, t_{\text{end}}], \tag{1.4}$$

where

- $v(t)$ is C^1 and independent of δ,
- $g_\delta(t)$ is continuous and $o(\delta)$, and
- $g_\delta'(t)$ is piecewise continuous, with the possible discontinuities occurring at the meshpoints $\{t_n\}$, and is $o(\delta)$.

This definition involves an interpolant $\eta_\delta(t)$. The interpolant is introduced because the discrete solution is defined on a mesh that varies with δ. In words, (1.4) demands that, for small δ, the global error at any fixed point t must be asymptotically linear in δ. Further, the global error $\eta_\delta'(t) - y'(t)$ in the first derivative approximation must also be asymptotically linear in δ. (We mention that a generalisation of this definition to allow for a leading term that behaves like δ^α, with $\alpha \neq 1$, was studied in [10].) For simplicity, in the rest of this work we will not explicitly indicate the dependence of $\eta(t)$ upon δ.

The following result appears in [9, section 2], and is essentially a formalisation of results of Stetter [13,14].

Result 1.1. If an adaptive ERK algorithm as described above is used to solve (1.1), and if

1. the stepsizes satisfy $\max_n\{h_n\} \to 0$ as $\delta \to 0$,
2. $\widetilde{\psi}(t, y(t))$ does not vanish on $[0, t_{\text{end}}]$,

then an interpolant $\eta(t)$ can be defined for which the method exhibits TP.

We remark that [9] includes the additional requirement that the initial stepsize h_1 be chosen so that

$$\text{est}_1 = \theta^p \delta + o(\delta).$$

However, this condition can be removed – our proof of theorem 4.1 shows how this can be done.

Let us now consider the assumptions 1 and 2. The first assumption requires the maximum stepsize to decrease to zero with δ. This scenario is perfectly reasonable – as the user asks for more accuracy, the algorithm is forced to take smaller steps. However, it must be stated as an assumption, since it is possible to construct examples where the error criterion $\text{est}_n \leqslant \delta$ holds on all steps but $\max_n\{h_n\} \nrightarrow 0$ as $\delta \to 0$. Aves et al. [1] show how such behaviour can occur in the presence of *spurious fixed points*, but in general these examples are unstable and will not be seen in practice. Overall, assumption 1 is unlikely to be violated.

The second assumption concerns the behaviour of the function $\widetilde{\psi}$, which, in general, depends upon both the problem (1.1) and the ERK coefficients. The algorithm does not monitor $\widetilde{\psi}(t, y(t))$ during the course of an integration, and there is no guarantee that assumption 2 will hold. The proof of result 1.1 in [9] shows that $v(t)$ in (1.4) is the solution to the linear (variational) problem

$$v'(t) - f_y\big(t, y(t)\big)v(t) = \theta^p \psi\big(t, y(t)\big) \big/ \big\|\widetilde{\psi}\big(t, y(t)\big)\big\|, \quad v(0) = 0. \tag{1.5}$$

Here the function ψ appears in the leading term of the local error expansion and is defined as follows. The *local solution*, $z_n(t)$, over a step from (t_{n-1}, y_{n-1}) to (t_n, y_n), is defined as

$$z_n'(t) = f\big(t, z_n(t)\big), \qquad z_n(t_{n-1}) = y_{n-1},$$

and the *local error* expansion in powers of h_n satisfies

$$\text{le}_n := y_n - z_n(t_n) = h_n^{p+1}\psi(t_{n-1}, y_{n-1}) + \text{O}\big(h_n^{p+2}\big). \tag{1.6}$$

It is clear from (1.5) that the analysis generally breaks down if assumption 2 does not hold. We mention that the related analysis of Shampine [12, assumption (5.6), p. 101] also requires a non-vanishing leading term in the error estimate.

In the next section we show that assumption 2 is necessary in general to ensure that an ERK code provides a reliable tolerance proportionality for the problem (1.1): we find examples with commonly used ERK algorithms and test problems where $\widetilde{\psi}(t, y(t))$ passes through zero, and we show numerically that the TP property is then lost. In section 3 we show that certain, special formulas exist where $\widetilde{\psi}(t, y(t)) = 0$ is less likely to cause difficulties. The fourth section introduces a new, modified stepsize strategy for which assumption 2 can be relaxed.

2. Examples of breakdown

In this section we present some examples of methods and IVPs where the assumptions in result 1.1 are not satisfied and TP is observed not to hold in practice.

First, let us recall a difficulty which arises in the stepsize estimators of some ERK methods when applied to equations of the form $y'(t) = f(t)$, where f is an

arbitrary function. As noted by Shampine [11], there are ERK methods such that for these particular equations the local error estimator (1.2) vanishes and according to (1.3) arbitrary stepsizes are allowed. This breakdown in the stepsize policy may introduce larger errors than desired in the numerical computation. This fact was analyzed in detail by Shampine [11] for Fehlberg's (7,8) pair [6], and a fix was devised.

Later, Verner [15] proposed new families of ERK pairs of several orders in which the local error estimators do not vanish identically. This was accomplished by choosing the coefficients of the RK formulas so that the local truncation errors of the corresponding pairs are different for every problem. In this way (unless all elementary differentials of the problem vanish identically) the above mentioned difficulty is absent in Verner's pairs. Moreover, Verner's approach has been followed for most of the ERK pairs that have been constructed in recent years.

In this work, we show that there are standard ERK pairs and widely-used test problems for which either the main formula or the local error estimate do not reflect exactly the orders O(h^{p+1}) and O(h^p) of (1.6) and (1.2). This behaviour (which will not be detected by a typical code) can degrade the tolerance proportionality.

If $\psi(t, y(t))$ in (1.6) is identically zero then $v(t)$ in (1.5) is zero, and hence the rate of decrease of the global error in (1.4) is faster than linear. A more dangerous situation, from the tolerance proportionality point of view, arises when $\widetilde{\psi}(t, y(t))$ vanishes at one or more isolated points in the integration interval. In the remainder of this paper a value $t = t^*$ such that $\widetilde{\psi}(t^*, y(t^*)) = 0$ and $\psi(t^*, y(t^*)) \neq 0$ will be called a TP-singular point. Clearly, for a given ERK method, the existence and location of TP-singular points depend on the IVP and cannot be easily detected in advance.

We now present a simple example in which TP-singular points can be easily identified analytically, and we demonstrate numerically that the TP property is lost.

Consider the A4 problem of the DETEST [5] set of problems, usually known as the logistic equation,

$$y' = \frac{y}{4}\left(1 - \frac{y}{20}\right), \quad y(0) = 1, \qquad t \in [0, 20], \tag{2.1}$$

whose analytical solution is

$$y(t) = \frac{20}{1 + 19\exp(-t/4)}\,. \tag{2.2}$$

As numerical method we take a second order ERK pair given by the advancing formula

$$y_n = y_{n-1} + h_n f_{n,1}, \tag{2.3}$$

and the error estimate

$$e_n = h_n(f_{n,0} - f_{n,1}), \tag{2.4}$$

where

$$f_{n,0} = f(t_{n-1}, y_{n-1}), \qquad f_{n,1} = f\left(t_{n-1} + \frac{h_n}{2}, y_{n-1} + \frac{h_n}{2} f_{n,0}\right).$$

Here the error estimate comes from differencing Euler's method and the second order formula (2.3). We will refer to this pair as RK2(1)a.

Standard theory (see, for example, [2]) shows that for the method (2.3)–(2.4) on an autonomous scalar equation, the expansions (1.2) and (1.6) have the form

$$e_n = -\frac{h_n^2}{2}\big[f'(f)(y_{n-1})\big] + \mathrm{O}\big(h_n^3\big),$$
$$\mathrm{le}_n = -\frac{h_n^3}{24}\big[\big(f''(f,f) + 4f'\big(f'(f)\big)\big)(y_{n-1})\big] + \mathrm{O}\big(h_n^4\big).$$

In particular, for the logistic equation (2.1) the functions in the leading terms are

$$\widetilde{\psi}(y) = -\frac{(10-y)y(20-y)}{6400}, \tag{2.5}$$
$$\psi(y) = \frac{y(20-y)(9y^2-180y+800)}{1024000}. \tag{2.6}$$

To show the existence of a TP-singular point, note that the solution (2.2) of the logistic equation increases monotonically in the interval $[0, 20]$ from $y(0) = 1$ to $y(20) \in (10, 20)$. Therefore, in view of the expression for $\widetilde{\psi}(y)$ in (2.5), there is precisely one TP-singular point given by the solution of $y(t) = 10$; that is, $t^* = 4\ln(19) = 11.78\ldots$. Note that for any $\mu > 0$, $\widetilde{\psi}(y(t)) \neq 0$ for $t \in [0, t^* - \mu]$ and therefore assumption 2 in result 1.1 holds. Consequently, if assumption 1 is also satisfied then TP follows over this subinterval.

In order to see the effect of t^*, several experiments were performed. First, we approximated the leading function $\widetilde{\psi}(t, y(t))$ of the local error estimate using only the information provided by the numerical integration. For a given tolerance δ, we monitored successively the sequence of stepsizes and the corresponding gridpoints t_{n-1}, $n = 1, 2, \ldots$, as well as the quotients

$$\frac{|e_n|}{h_n^2} = \big|\widetilde{\psi}(t_{n-1}, y_{n-1}) + \mathrm{O}(h_n)\big|,$$

that approximate asymptotically the modulus of the function $|\widetilde{\psi}(t, y(t))|$ at the gridpoints. Figure 1 plots these values (using linear interpolation) for $\delta = 10^{-3}$, 10^{-4}, 10^{-5}, and it can be seen that they reflect very accurately the existence of the TP-singular point predicted by the theory.

It must be remarked that, at least theoretically, the existence of a TP-singular point could cause the algorithm for stepsize changing to breakdown at a gridpoint very close to the singular point. However, in this case, such a situation never occurred due to the fact that the higher order terms of e_n do not vanish simultaneously. The overall effect is a rapid increase in the size of the steps when the numerical solution goes through the isolated point $t = t^*$.

Figure 2 plots the stepsizes for $\delta = 10^{-3}$, 10^{-4}, 10^{-5}, 10^{-6} and it can be seen that the largest stepsizes occur precisely in the vicinity of the TP-singular point. (As

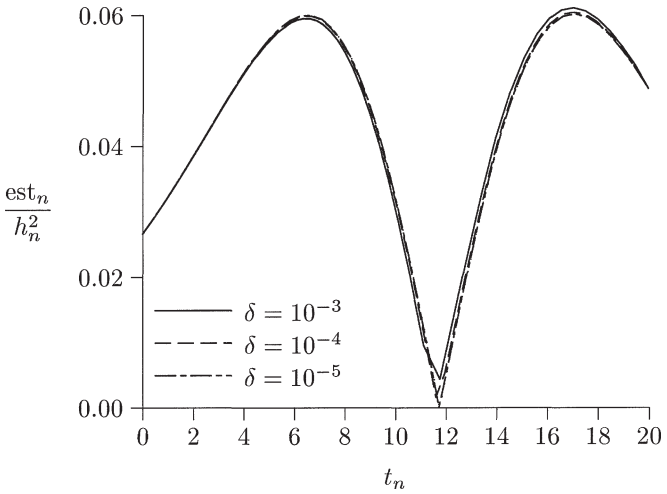

Figure 1. A4 problem, RK2(1)a.

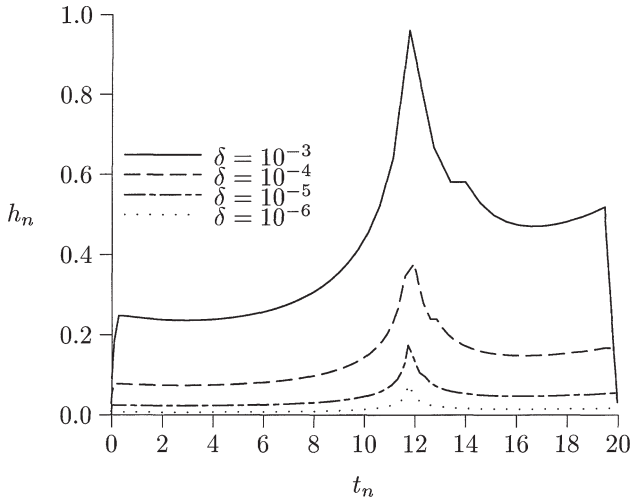

Figure 2. A4 problem, RK2(1)a.

is standard in ERK codes, we placed an upper bound on the ratio between the sizes of two consecutive steps.)

Finally, to study to what extent the TP property is satisfied we computed, for several values of δ, the linear interpolant obtained with the points

$$\left(t_n, \frac{y_n - y(t_n)}{\delta}\right). \tag{2.7}$$

If TP holds in the whole interval then as $\delta \to 0$ these curves should tend to the fixed curve $v(t)$ from (1.4). Figure 3 plots these curves for $\delta = 10^{-4}, \ldots, 10^{-10}$ and it is

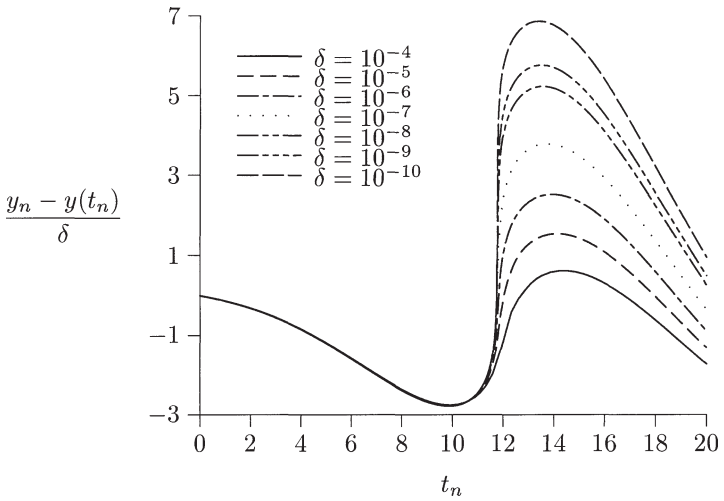

Figure 3. A4 problem, RK2(1)a.

clear that in the neighbourhood of the singular TP-point the curves do not show a convergent behaviour, in contrast with the behaviour in the first part of the integration interval.

The computations performed in example 1 were repeated for other ERK methods and problems. In particular for the A3 problem of the DETEST set and the pair (2.3)–(2.4) there are also several TP-points that can be easily calculated and the TP property does not hold.

Next we study the existence of TP-singular points in the numerical solution of the A4 problem with the well known pair RK5(4)7FM of RK formulas of orders 5 and 4 due to Dormand and Prince [3]. By using a symbolic package it is found that for the quadratic function $f(y)$ of (2.1), the expansions of e_n and le_n are given by

$$e_n = h_n^6 \psi(y_{n-1}) + \mathrm{O}\big(h_n^7\big), \qquad \mathrm{le}_n = h_n^5 \widetilde{\psi}(y_{n-1}) + \mathrm{O}\big(h_n^6\big),$$

where

$$\psi(y) = \frac{y(y-20)(y-10)(2y^4 - 80y^3 + 1355y^2 - 11100y + 36000)}{106168320000000}$$

and

$$\widetilde{\psi}(y) = -\frac{y(y-20)(7673y^4 - 306920y^3 + 4898300y^2 - 36582000y + 104760000)}{2654208000000000}.$$

It is easily seen that in the range $1 < y < 20$, the function $\widetilde{\psi}(y)$ vanishes at the values

$$y_1 = 7.918499387, \qquad y_2 = 12.08150061,$$

therefore in view of (2.2) we have the TP-points

$$t_1^* = 10.08786115, \qquad t_2^* = 13.46765068.$$

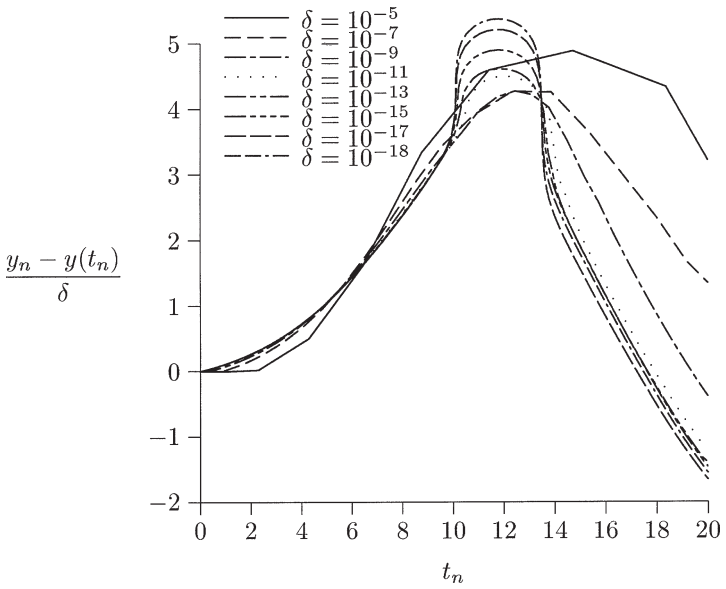

Figure 4. A4 problem, DOPRI5(4).

As in the above example we have plotted in figure 4 the quotients (2.7) for this problem and several values of δ. We see that the TP property is maintained at the beginning of the integration interval, but after crossing the first TP-singular point t_1^*, the curves do not show convergent behaviour for $\delta \to 0$. The same behaviour is observed after crossing the second TP-singular point t_2^*.

3. Insensitivity to a vanishing leading term

When TP holds, the variational equation (1.5) defines the function $v(t)$ which in view of (1.4) can be regarded as the "asymptotic proportionality function". In the examples of the previous section, one manifestation of the breakdown is that the right-hand side of the variational equation (1.5) becomes unbounded (at the TP-singular points). We notice, however, that it is possible for the ratio

$$\frac{\psi(t, y(t))}{\|\widetilde{\psi}(t, y(t))\|} \tag{3.1}$$

appearing in (1.5) to be a regular function, even in the presence of TP-singular points, since $\psi(t, y(t))$ and $\|\widetilde{\psi}(t, y(t))\|$ may vanish simultaneously. This is certainly true for the non-extrapolated error-per-unit-step schemes mentioned in section 1, since in this case $\psi(t, y(t)) = \widetilde{\psi}(t, y(t))$. In this section we show that ERK pairs exist for which the ratio (3.1) is always regular in the more widely-used extrapolated error-per-step mode. This implies that a bounded solution of (1.5) exists for all differential equations.

For convenience, we assume throughout this section that the differential system (1.1) is written in autonomous form. This can, of course, be ensured by, if necessary, adding a new component $y_0(t) \equiv t$ to the vector $y(t)$ that satisfies the trivial equation $\mathrm{d}y_0/\,\mathrm{d}t = 1$ together with $y_0(0) = 0$.

As a fist example we consider the second order method given by the advancing formula

$$y_n = y_{n-1} + \frac{h_n}{4}(f_{n,0} + 3f_{n,1}),$$

and the error estimate

$$e_n = \frac{3h_n}{4}(f_{n,1} - f_{n,0}),$$

where

$$f_{n,0} = f(y_{n-1}), \qquad f_{n,1} = f\left(y_{n-1} + \frac{2}{3}h_n f_{n,0}\right).$$

We refer to this method as RK2(1)b. A simple calculation shows that

$$e(y;h) = -\frac{h^2}{2}f'(f)(y) + \mathrm{O}\big(h^3\big),$$
$$\mathrm{le}(y;h) = -\frac{h^3}{6}f'\big(f'(f)\big)(y) + \mathrm{O}\big(h^4\big),$$

and hence

$$\widetilde{\psi}(y) = -\frac{1}{2}f'(f)(y), \qquad \psi(y) = -\frac{1}{6}f'\big(f'(f)\big)(y).$$

Since f' is a linear operator and $\psi(y) = (1/3)f'(\widetilde{\psi}(y))$, the function (3.1) is regular for all f.

To check numerically the TP of this method we have applied it to the IVPs considered in section 2. First, we consider the logistic equation A4 in which $f(y) = y(20 - y)/80$ and

$$\widetilde{\psi}(y) = -\frac{(10-y)y(20-y)}{6400},$$
$$\psi(y) = -\frac{(10-y)^2 y(20-y)}{12 \times 40^3}.$$

Then (3.1) becomes $-|10 - y|/120$ where $y = y(t)$ is the exact solution of the IVP given by (2.2). The function $v(t)$ is the solution of the variational equation

$$v' = \frac{10 - y(t)}{40}\,v - \frac{|10 - y(t)|}{120}, \quad v(0) = 0, \qquad t \in [0, 20]. \tag{3.2}$$

Taking into account that

$$\big|10 - y(t)\big| = \begin{cases} 10 - y(t) & \text{if } t \leqslant t^* = 4\ln(19), \\ -(10 - y(t)) & \text{if } t \geqslant t^*, \end{cases}$$

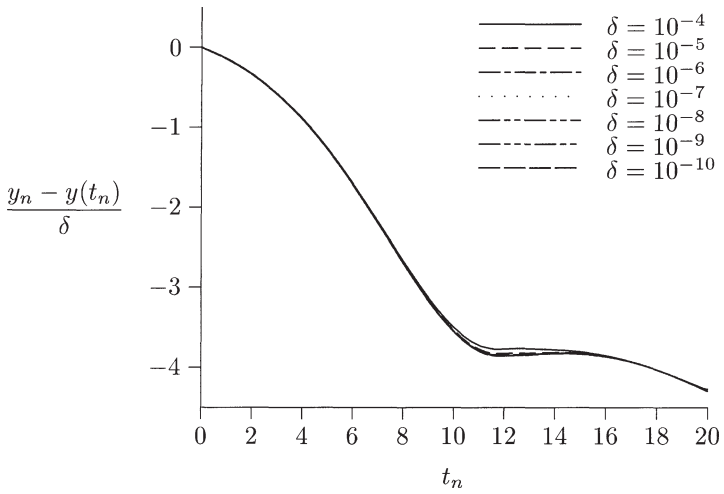

Figure 5. A4 problem, RK2(1)b.

where $y(t)$ is given by (2.2), the general solution of (3.2) is

$$v(t) = \begin{cases} (-80/57)y'(t) + 1/3 & \text{if } t \leqslant t^*, \\ (-248/285)y'(t) - 1/3 & \text{if } t \geqslant t^*. \end{cases} \tag{3.3}$$

In figure 5 we plot for $\delta = 10^{-4}, \ldots, 10^{-10}$ the piecewise linear functions that interpolate the data points of (2.7). We see that for $\delta \to 0$ they converge to $v(t)$, which has an horizontal tangent at the TP-singular point $t^* = 4\ln(19) \approx 11.78$.

Finally, we have constructed a third order ERK method with four stages where the coefficients have been chosen with the aim that (3.1) is a regular function. The Butcher array of its coefficients is

0				
1/3	1/3			
2/3	−1	5/3		
8/9	52/81	−20/81	40/81	
b	35/320	144/320	60/320	81/320
$\widehat{b}$	1/4	0	3/4	0

Here the b_i are the coefficients of the third order formula and the $\widetilde{b}_i$ the coefficients of a second order auxiliary formula which is used to control the local error in the usual way, that is, $e_n = \widetilde{y}_{n+1} - y_{n+1}$. This pair will be denoted RK3(2)b.

After some calculation it is found that for this method

$$|\widetilde{\psi}(y)| = \frac{1}{6}|f'(f'(f))(y)|,$$
$$\psi(y) = \frac{5}{72}\,\frac{1}{4!}\,f'(f'(f'(f)))(y),$$

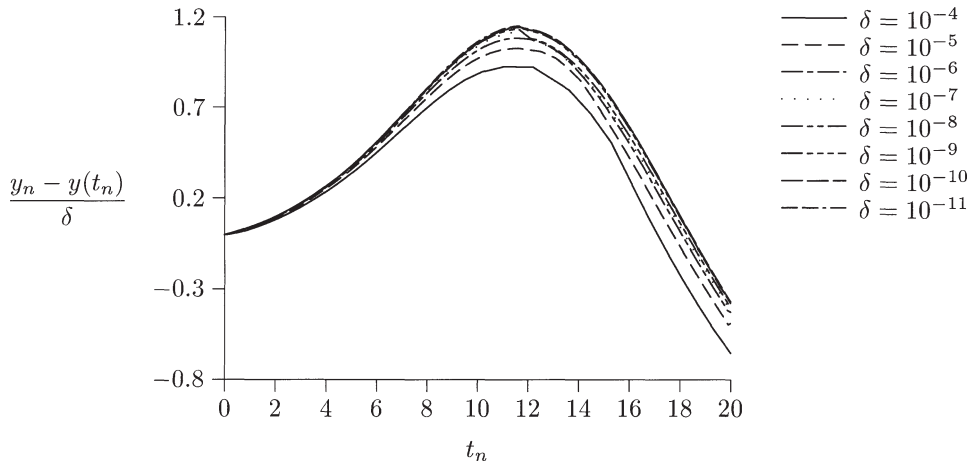

Figure 6. A4 problem, RK3(2)b.

and therefore (3.1) is a regular function at TP-singular points.

Figure 6 plots the piecewise linear functions that interpolate the data points of (2.7) for $\delta = 10^{-4}, \ldots, 10^{-11}$ and the A4 problem. Our numerical tests imply that ERK pairs for which (3.1) is always regular perform well in the presence of TP-singular points. Asking for (3.1) to be regular, however, is a very strong requirement that places many constraints on the coefficients of the elementary differentials which appear in the leading terms of the local truncation errors of the pair of formulas. Furthermore, although the condition guarantees a bounded solution for (1.5), it does not automatically ensure TP. It is possible to state additional assumptions under which TP holds, but we do not pursue this approach here because the derivation of high order ERK methods with these requirements turns out to be a very difficult task. In the next section we advocate an alternative stepsize selection scheme that is designed to work with any ERK pair.

4. New stepsize selection policy

4.1. Description

In this section we show how the stepsize selection formula (1.3) can be altered to avoid the difficulties encountered in section 2. Since the standard formula works well when $\widetilde{\psi}(t, y(t)) \neq 0$, we wish to modify the stepchanging policy only when the numerical solution is "close" to a TP-singular point. Our overall aim is to produce, at little computational cost, an algorithm for which a strong guarantee (theorem 4.1) can be established.

Thus, we advocate replacing the stepsize formula (1.3) by

$$h_{n+1} = \theta \left(\frac{\delta}{\text{estmax}_n} \right)^{1/p} h_n, \tag{4.1}$$

where $\text{estmax}_n = \max\{\text{est}_n, \overline{\text{est}}_n\}$. Here,

$$\overline{\text{est}}_n = h_n^p \min\{\text{estint}_n, \text{estabs}\}, \tag{4.2}$$

with

$$\text{estint}_n = \kappa \frac{1}{t_n} \sum_{i=1}^{n} \frac{\text{est}_i}{h_i^{p-1}}, \tag{4.3}$$

and κ and estabs are small constants. Appropriate values for κ and estabs depend upon the ERK coefficients. Details are given in subsection 4.4. The error criterion remains the same – a step is accepted if $\text{est}_n \leqslant \delta$ and rejected otherwise.

The motivation behind (4.1)–(4.3) is that when the error estimate est_n becomes "small", it is replaced by $\overline{\text{est}}_n$ in the stepchanging formula. The two quantities estint_n and estabs represent relative and absolute thresholds. Since $\text{est}_i/h_i^p \approx \|\widetilde{\psi}(t_{i-1}, y(t_{i-1}))\|$, the sum

$$\sum_{i=1}^{n} \frac{\text{est}_i}{h_i^{p-1}} \approx \sum_{i=1}^{n} \big\|\widetilde{\psi}\big(t_{i-1}, y(t_{i-1})\big)\big\| h_i$$

represents a Riemann sum that approximates the integral of $\|\widetilde{\psi}(t, y(t))\|$ over the current range of integration. Since $\kappa \ll 1$ we will usually have $\text{est}_n > h_n^p \text{estint}_n$, but if the leading term in est_n becomes relatively small, then $\overline{\text{est}}_n$ will dominate. The absolute threshold, estabs, is introduced to account for the case where the error estimate changes by several orders of magnitude, but does not become close to zero. Our numerical tests revealed that this case can arise in practice; in particular on the D5 problem from DETEST.

4.2. Analysis

To investigate the TP behaviour of the new stepsize selection scheme we will make use of a result in [9], which shows that a sufficient condition for TP is that the local error on each step satisfies

$$\text{le}_n = \gamma(t_n) h_n \delta + \text{o}(h_n \delta), \tag{4.4}$$

where $\gamma(t)$ is continuous and independent of δ.

We make the following assumptions.

1. The stepsizes satisfy $\max_n\{h_n\} \to 0$ as $\delta \to 0$.
2. $\widetilde{\psi}(0, y(0)) \neq 0$.

Comparing these with the assumptions in result 1.1 we see that the second assumption has been relaxed, so that $\widetilde{\psi}(t, y(t))$ is only required to be nonzero at $t = 0$. We will discuss this further in the next subsection.

We note that assumption 1 implies that the numerical solution converges; $\max_n \|y_n - y(t_n)\| \to 0$ as $\delta \to 0$, see, for example, [8]. We also mention that

no extra assumption is made about the way that the initial stepsize, h_1, is chosen. We require only that $h_1 \to 0$ as $\delta \to 0$, and $\text{est}_1 \leqslant \delta$.

We now prove a number of intermediate results which ultimately show that (4.4) holds, for $n > 1$.

Lemma 4.1. There exists a constant $\alpha > 0$ (independent of δ) such that, for all sufficiently small δ,

$$\alpha \leqslant \frac{\min_{n>1}\{h_n\}}{\max_{n>1}\{h_n\}}\,. \tag{4.5}$$

Proof. From assumption 2, there exist an $\varepsilon > 0$ and $0 < \widehat{t} < t_{\text{end}}$ such that $\|\widetilde{\psi}(t, y(t))\| > \varepsilon$ for $0 \leqslant t \leqslant \widehat{t}$. The standard error estimate satisfies (1.2) and hence

$$\begin{aligned} \text{est}_n &= \big\|\widetilde{\psi}(t_{n-1}, y_{n-1})\big\| h_n^p + \mathrm{O}\big(h_n^{p+1}\big) \\ &= \big\|\widetilde{\psi}\big(t_{n-1}, y(t_{n-1})\big)\big\| h_n^p + \mathrm{o}\big(h_n^p\big). \end{aligned} \tag{4.6}$$

Hence, for all sufficiently small δ,

$$\text{est}_n \geqslant \frac{\varepsilon}{2}\, h_n^p, \quad \text{for all } 0 \leqslant t_n \leqslant \widehat{t}. \tag{4.7}$$

Now consider the Riemann sum estimate, estint_n. Given any meshpoint $t_n > \widehat{t}$, let t_N be a meshpoint such that $\widehat{t}/2 < t_N < \widehat{t}$. Then

$$\text{estint}_n = \kappa\, \frac{1}{t_n} \sum_{i=1}^{n} \frac{\text{est}_i}{h_i^p}\, h_i \geqslant \kappa\, \frac{1}{t_{\text{end}}} \sum_{i=1}^{N} \frac{\text{est}_i}{h_i^p}\, h_i \geqslant \kappa\, \frac{1}{t_{\text{end}}} \sum_{i=1}^{N} \frac{\varepsilon}{2}\, h_i, \tag{4.8}$$

using (4.7). The right-hand side of this inequality satisfies

$$\kappa\, \frac{1}{t_{\text{end}}} \sum_{i=1}^{N} \frac{\varepsilon}{2}\, h_i = \kappa \frac{1}{t_{\text{end}}}\, \frac{\varepsilon}{2}\, t_N > \kappa\, \frac{\varepsilon}{4}\, \frac{\widehat{t}}{t_{\text{end}}}\,. \tag{4.9}$$

Hence, there exists a constant $\gamma > 0$ such that

$$\overline{\text{estint}}_n \geqslant \gamma, \quad \text{for all } \widehat{t} < t_n \leqslant t_{\text{end}}. \tag{4.10}$$

Taking $\alpha_1 = \min\{\varepsilon/2, \gamma, \text{estabs}\}$, it follows from (4.7) and (4.10) that, for sufficiently small δ,

$$\text{estmax}_n := \max\big\{\text{est}_n, \overline{\text{est}}_n\big\} \geqslant \alpha_1 h_n^p. \tag{4.11}$$

We can obtain a corresponding upper bound. Let $K = \max_{[0,t_{\text{end}}]} \|\widetilde{\psi}(t, y(t))\|$. It follows from (4.6) that, for sufficiently small δ,

$$\text{est}_n \leqslant 2K h_n^p. \tag{4.12}$$

Also, using (4.12),

$$\overline{\text{est}}_n \leqslant h_n^p \text{estint}_n = \kappa h_n^p \frac{1}{t_n} \sum_{i=1}^{n} \frac{\text{est}_i}{h_i^p} h_i \leqslant \kappa h_n^p \frac{1}{t_n} \sum_{i=1}^{n} 2Kh_i \leqslant \kappa 2K h_n^p. \tag{4.13}$$

Hence, from (4.12) and (4.13) there exists a constant $\alpha_2 > 0$ such that, for sufficiently small δ,

$$\text{estmax}_n := \max\bigl\{\text{est}_n, \overline{\text{est}}_n\bigr\} \leqslant \alpha_2 h_n^p. \tag{4.14}$$

Now, in the stepsize formula (4.1) it follows from (4.11) and (4.14) that, for sufficiently small δ,

$$\frac{\min_{n>1}\{h_n\}}{\max_{n>1}\{h_n\}} \geqslant \left(\frac{\alpha_1}{\alpha_2}\right)^{1/p} =: \alpha, \tag{4.15}$$

where α is a constant. □

Lemma 4.2. A quantity that is $O(h_n)$ as $\delta \to 0$ must also be $O(\delta^{1/p})$.

Proof. The error criterion ensures that $\text{est}_n \leqslant \delta$. Hence, from (4.7),

$$\frac{\varepsilon}{2} h_n^p \leqslant \text{est}_n \leqslant \delta, \quad \text{for all } 0 \leqslant t_n \leqslant \widehat{t}, \tag{4.16}$$

and so,

$$\min\{h_n\} \leqslant \left(\frac{2}{\varepsilon}\right)^{1/p} \delta^{1/p}. \tag{4.17}$$

Hence, using lemma 4.1,

$$\max\{h_n\} \leqslant \frac{1}{\alpha} \min\{h_n\} \leqslant \frac{1}{\alpha}\left(\frac{2}{\varepsilon}\right)^{1/p} \delta^{1/p}, \tag{4.18}$$

and the result follows. □

Lemma 4.3. The Riemann sum error estimate behaves like

$$\text{estint}_n = \Psi(t_{n-1}) + o(1), \tag{4.19}$$

where

$$\Psi(t) = \begin{cases} \kappa t^{-1} \int_0^t \|\widetilde{\psi}(\mu, y(\mu))\| \, \mathrm{d}\mu & \text{if } t > 0, \\ \kappa \|\widetilde{\psi}(0, y(0))\| & \text{if } t = 0, \end{cases} \tag{4.20}$$

and hence

$$\overline{\text{est}}_n = h_n^p \min\bigl\{\Psi(t_{n-1}), \text{estabs}\bigr\} + o\bigl(h_n^p\bigr).$$

Proof. The standard estimate satisfies (4.6) and hence

$$\text{estint}_n = \kappa \frac{1}{t_n} \sum_{i=1}^{n} \frac{\text{est}_i}{h_i^p} h_i = \kappa \frac{1}{t_n} \sum_{i=1}^{n} \big(\|\widetilde{\psi}\big(t_{i-1}, y(t_{i-1})\big)\| h_i + \text{o}(h_i) \big). \tag{4.21}$$

Now,

$$\sum_{i=1}^{n} \|\widetilde{\psi}\big(t_{i-1}, y(t_{i-1})\big)\| h_i = \int_0^{t_{n-1}} \|\widetilde{\psi}\big(\mu, y(\mu)\big)\| \, \mathrm{d}\mu + \text{o}(1). \tag{4.22}$$

Also, from lemma 4.1,

$$\sum_{i=1}^{n} \text{o}(h_i) \leqslant n \text{o}\big(\min_n \{h_n\} \big) = n \min_n \{h_n\} \, \text{o}(1) \leqslant t_n \, \text{o}(1) = \text{o}(1). \tag{4.23}$$

Using (4.22) and (4.23) in (4.21) gives the result. □

Note that $\Psi(t, y(t)) > 0$ for $t > 0$ since, by assumption, $\|\widetilde{\psi}(t, y(t))\| \geqslant \varepsilon$ on some interval $[0, \widehat{t}\,]$.

Lemma 4.4. For sufficiently small δ there are no step rejections when formula (4.1) is used. Also, for $n > 1$, the quantity estmax_n satisfies

$$\text{estmax}_n = \theta^p \delta + \text{o}(\delta). \tag{4.24}$$

Proof. We will show that using the stepsize formula (without imposing the error criterion) leads to (4.24). Since $\text{est}_n \leqslant \text{estmax}_n$ and $\theta < 1$ it will follow that, for sufficiently small δ, the error criterion is automatically satisfied, and no rejections occur.

On a general step, we know from (4.6) and (4.19) that

$$\text{estmax}_n = C(t_{n-1}) h_n^p + \text{o}\big(h_n^p\big), \tag{4.25}$$

where

$$C(t) = \max\big\{ \|\widetilde{\psi}\big(t, y(t)\big)\|, \min\big\{ \Psi(t), \text{estabs} \big\} \big\}.$$

Now, in (4.1), we have

$$h_{n+1}^p = \theta^p \frac{\delta}{h_n^p (C(t_{n-1}) + \text{o}(1))} h_n^p = \theta^p \frac{\delta}{C(t_{n-1})} + \text{o}(\delta). \tag{4.26}$$

Hence, on the next step, using lemma 4.2,

$$\text{estmax}_{n+1} = C(t_n) h_{n+1}^p + \text{o}\big(h_{n+1}^p\big) = \frac{C(t_n)}{C(t_{n-1})} \theta^p \delta + \text{o}(\delta).$$

Now the continuity of $C(t)$ gives the result. □

Lemma 4.5. The local error for $n > 1$ satisfies

$$\mathrm{le}_n = \frac{\psi(t_n, y(t_n))\theta^p}{C(t_n)}\, h_n\delta + \mathrm{o}(h_n\delta), \tag{4.27}$$

where

$$C(t) = \max\{\|\widetilde{\psi}(t, y(t))\|, \min\{\Psi(t), \mathrm{estabs}\}\}.$$

Proof. From (4.25),

$$\frac{\mathrm{estmax}_n}{C(t_{n-1})} = h_n^p + \mathrm{o}\big(h_n^p\big). \tag{4.28}$$

Hence, in (1.6),

$$\mathrm{le}_n = \psi(t_{n-1}, y_{n-1})\,\frac{\mathrm{estmax}_n}{C(t_{n-1})}\, h_n + \mathrm{o}\big(h_n^{p+1}\big). \tag{4.29}$$

Now lemma 4.4 gives

$$\mathrm{le} = \frac{\psi(t_{n-1}, y_{n-1})\theta^p\delta}{C(t_{n-1})}\, h_n + \mathrm{o}(h_n\delta) + \mathrm{o}\big(h_n^{p+1}\big). \tag{4.30}$$

Using lemma 4.2 and the continuity of ψ and C we find

$$\mathrm{le}_n = \frac{\psi(t_n, y(t_n))\theta^p\delta}{C(t_n)}\, h_n + \mathrm{o}(h_n\delta). \tag{4.31}$$

□

Theorem 4.1. Under the assumptions 1 and 2 stated at the start of this subsection, the ERK algorithm with stepchanging formula (4.1) exhibits TP.

Proof. Lemma 4.5 is the essential result, showing that the local error has the required form (4.4) for every $n > 1$. If the lemma also applied for $n = 1$ then TP would follow automatically. However, it is unrealistic to assume that (4.4) holds on the first step, and we show below that the behaviour on the first step is not important, provided that the error criterion is satisfied. Loosely, on the first step the global and local errors are identical, and the global error that is introduced has a negligible o(δ) effect. More formally, since $\widetilde{\psi}(0, y(0)) \neq 0$ (by assumption), the error criterion $\mathrm{est}_1 \leqslant \delta$ and the expansion (1.2) imply that $h_1 = \mathrm{O}(\delta^{1/p})$ so that

$$y_1 - y(t_1) = \mathrm{O}\big(h_1^{p+1}\big) = \mathrm{O}\big(\delta^{(p+1)/p}\big) = \mathrm{o}(\delta). \tag{4.32}$$

Now let $x(t)$ be the function on $[t_1, t_{\mathrm{end}}]$ satisfying

$$x'(t) = f\big(t, x(t)\big), \quad x(t_1) = y_1.$$

Note that $x(t)$ depends upon δ, via t_1. However, the analysis in section 2 of [9] still applies to give TP from lemma 4.5; there exists an interpolant $\eta(t)$ and a C^1 function $v(t)$ independent of δ such that

$$\eta(t) - x(t) = v(t)\delta + \mathrm{o}(\delta), \tag{4.33}$$

for any $t \in [t_1, t_{\text{end}}]$. Now, using (4.32) it follows from a standard differential inequality (see, for example, [8, section I.10]), that

$$x(t) - y(t) = \mathrm{o}(\delta),$$

for any $t \in [t_1, t_{\text{end}}]$. Combining this with (4.33) gives $\eta(t) - y(t) = v(t)\delta + \mathrm{o}(\delta)$ for any $t \in [t_1, t_{\text{end}}]$. Since $t_1 \to 0$ as $\delta \to 0$ and $\eta(0) = y_0$, it follows that

$$\eta(t) - y(t) = v(t)\delta + \mathrm{o}(\delta), \quad \text{for any } t \in [0, t_{\text{end}}].$$

To see that a similar result holds for the first derivative, combine the expansions

$$\eta'(t) - x'(t) = v'(t)\delta + \mathrm{o}(\delta)$$

and

$$x'(t) - y'(t) = f\big(t, x(t)\big) - f\big(t, y(t)\big) = \mathrm{o}(\delta). \qquad \square$$

4.3. Discussion

The analysis above shows that the new stepchanging policy (4.1) requires $\widetilde{\psi}(t, y(t)) \neq 0$ only at $t = 0$, rather than over the whole interval $[0, t_{\text{end}}]$. In fact, it is possible to further modify the policy so that $\widetilde{\psi}(0, y(0)) \neq 0$ is not needed. This could be done, for example, by forcing $h_1 = \mathrm{O}(\delta^{1/p})$ and adding εh_n^p, where $\varepsilon > 0$ is constant, to the definition of $\overline{\text{est}}_n$ in (4.2). However, with this approach there is a danger that a poor choice of ε can affect the normal behaviour of the stepsize formula away from TP-singular points. Further, we feel that choosing the initial stepsize so that the integration is started "on scale" is a separate issue that is important in its own right. Gladwell et al. [7], for example, have derived a sophisticated algorithm for computing h_1. Hence, overall, we regard the case $\widetilde{\psi}(0, y(0)) = 0$ as a difficulty that must be addressed by the initial stepsize formula, rather than by the general stepchanging formula. If the initial phase of the integration is reliable, in the sense that for some fixed $t^* > 0$, where $\widetilde{\psi}(t^*, y(t^*)) \neq 0$, we have

$$\eta(t^*) - y(t^*) = K\delta + \mathrm{o}(\delta),$$

then the analysis above is readily adapted to show that TP is maintained over $[t^*, t_{\text{end}}]$.

4.4. Numerical tests and conclusions

We now describe some numerical tests with the new stepchanging technique.

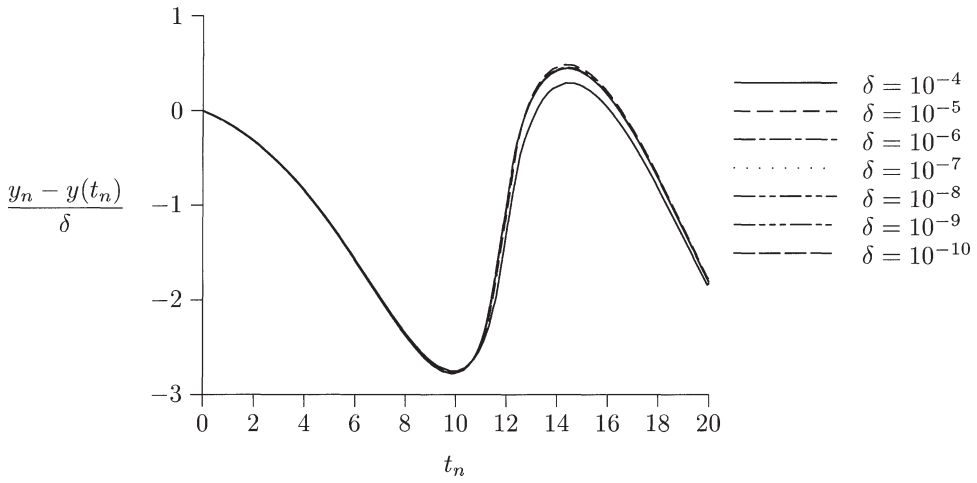

Figure 7. A4 problem, RK2(1)a.

In all cases the initial stepsize was computed as

$$h_1 = \left(\frac{\delta}{\max\{\|f(t_0, y_0)\|, 10^{-p}\}} \right)^{1/p}.$$

For efficiency, the parameters κ and estabs in the stepchanging strategy must be fine-tuned for a particular ERK pair. Loosely, if the parameters are too large, then an unnecessary switch from est_n to $\overline{\text{est}}_n$ may be made in (4.1); that is, a TP-singular point may be falsely signalled. This causes the code to choose smaller stepsizes than necessary, which may reduce the efficiency (although TP is, of course, maintained). Conversely, if the parameters are too small, then TP-singular points may be missed, except at very stringent tolerances. After detailed testing on a range of problems, we chose the following values:

	κ	estabs
RK2(1a)	0.2	4.0×10^{-2}
DOPRI5(4)	0.5	2.5×10^{-5}

In figure 7 we plot the quotients (2.7) for the RK2(1)a pair on the A4 problem. We see that the curves obtained for $\delta \to 0$ converge clearly to a limit, which is the graph of the "asymptotic proportionality function" $v(t)$, confirming that TP holds.

Figure 8 gives the corresponding results for the RK5(4)7FM pair of Dormand and Prince, also called DOPRI5(4) [3,8], on the same A4 problem. As remarked above, two TP-singular points exist, but TP also holds.

Next, we consider the cost of the new stepchanging policy. The cost-per-step involved in replacing (1.3) by (4.1) is clearly negligible. On problems where there are no TP-singular points, the change will usually not affect the numerical solution. On those where TP-singular points exist, we have found that the stepsizes are only modified in the neighbourhood of TP-singular points. Figures 9 and 10 show standard

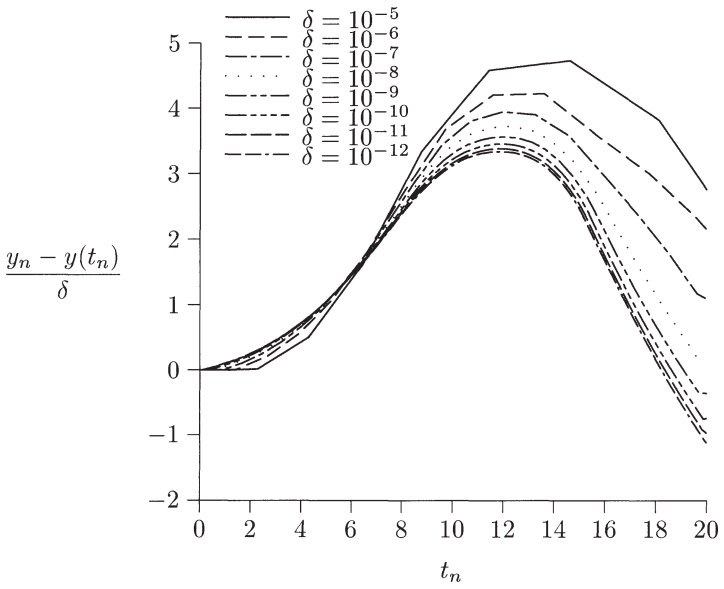

Figure 8. A4 problem, DOPRI5(4).

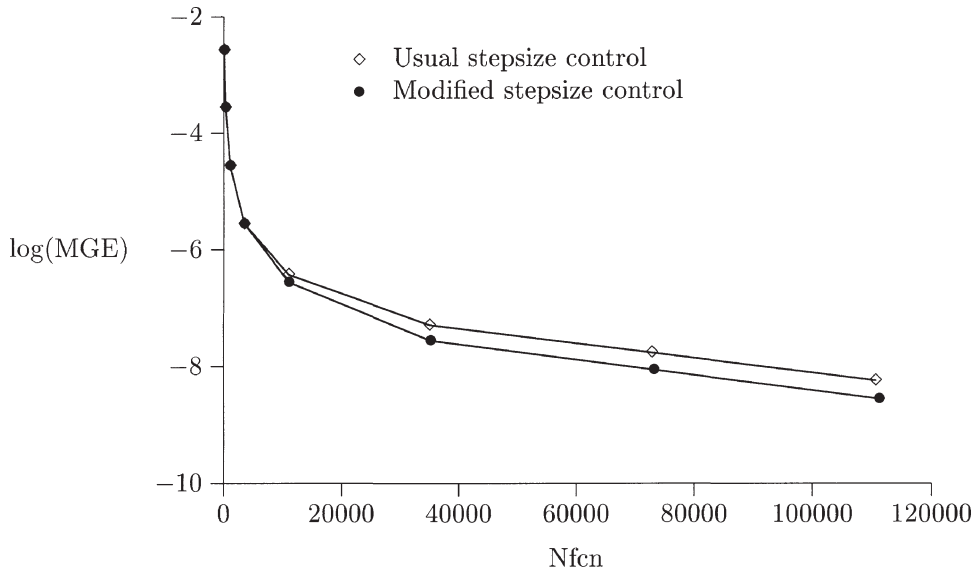

Figure 9. A4 problem, RK2(1)a.

efficiency plots for the two methods on the two problems mentioned above. Here, MGE is the maximum global error over all steps, and NFCN is the number of function evaluations used. In these figures the cost of both stepsize techniques is very similar.

For the sake of brevity, we have presented results for the A4 problem only. A similar study has been carried out for the A3 problem in which the number of

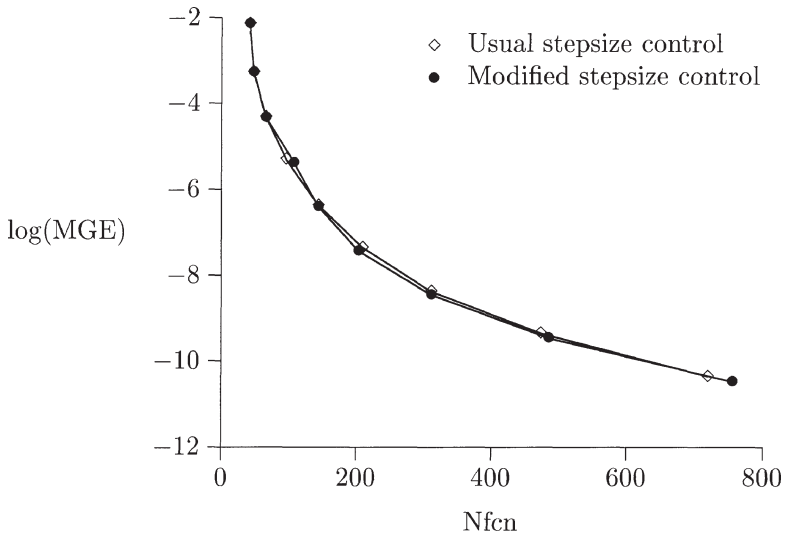

Figure 10. A4 problem, DOPRI5(4).

singular TP points is larger than in the A4 problem and the conclusions are essentially the same. Further numerical tests on higher order ERK methods, in particular the Prince–Dormand 8(7) (DOPRI8 [8]) pair, indicated the existence of TP-singular points on standard test problems.

To conclude, we emphasise that the main aim in this work was to show that the guarantees associated with some adaptive ODE algorithms can be strengthened, at little cost, by modifying the stepchanging process. The modification can be applied to a range of standard algorithms.

Acknowledgements

The research of DJH was supported by a grant from the Science and Engineering Research Council of the UK. DJH also thanks the University of Zaragoza for funding travel and accommodation during this collaboration. The research of the other authors was partially supported by a grant from the Comisión Interministerial de Ciencia y Tecnología of Spain.

References

[1] M.A. Aves, D.F. Griffiths and D.J. Higham, Does error control suppress spuriosity?, SIAM J. Numer. Anal., to appear.
[2] J.C. Butcher, *The Numerical Analysis of Ordinary Differential Equations* (Wiley, New York, 1987).
[3] J.R. Dormand and P.J. Prince, A family of embedded Runge–Kutta formulae, J. Comput. Appl. Math. 6 (1980) 19–26.
[4] W.H. Enright, The relative efficiency of alternative defect control schemes for high-order continuous Runge–Kutta formulas, SIAM J. Numer. Anal. 30(5) (1993) 1419–1445.

[5] W.H. Enright and J.D. Pryce, Two FORTRAN packages for assessing initial value methods, ACM Trans. Math. Software 13 (1987) 1–27.
[6] E. Fehlberg, Classical fifth, sixth, seventh and eighth order Runge–Kutta formulas with stepsize control, NASA Technical Report R-287 (1968).
[7] I. Gladwell, L.F. Shampine and R.W. Brankin, Automatic selection of the initial stepsize for an ODE solver, J. Comput. Appl. Math. 18 (1987) 175–192.
[8] E. Hairer, S.P. Nørsett and G. Wanner, *Solving Ordinary Differential Equations I, Nonstiff Problems* (Springer, 2nd ed., 1993).
[9] D.J. Higham, Global error versus tolerance for explicit Runge–Kutta methods, IMA J. Numer. Anal. 11 (1991) 457–480.
[10] D.J. Higham, The tolerance proportionality of adaptive ODE solvers, J. Comput. Appl. Math. 45 (1993) 227–236.
[11] L.F. Shampine, Quadrature and Runge–Kutta formulas, Appl. Math. Comput. 2 (1976) 161–171.
[12] L.F. Shampine, The stepsizes used by one-step codes for ODEs, Appl. Numer. Math. 1 (1985) 95–106.
[13] H.J. Stetter, Considerations concerning a theory for ODE-solvers, in: *Numerical Treatment of Differential Equations*, eds. R. Burlisch, R. Grigorieff and J. Schröder (Springer, Berlin, 1978) pp. 188–200; also: Lecture Notes in Mathematics 631 (1976).
[14] H.J. Stetter, Tolerance proportionality in ODE-codes, in: *Proceedings 2nd Conference on Numerical Treatment of Ordinary Differential Equations*, ed. R. März, Humboldt University, Berlin (1980) pp. 109–123; also in: *Working Papers for the 1979 SIGNUM Meeting on Numerical Ordinary Differential Equations*, ed. R.D. Skeel, Department of Computer Science, University of Illinois at Urbana–Champaign.
[15] J.H. Verner, Explicit Runge–Kutta methods with estimates of the local truncation error, SIAM J. Numer. Anal. 15(4) (1978) 772–790.