UNIVERSITY OF

SUSSEX

AT BRIGHTON

Centre for Mathematical Analysis and Its Applications

University of Sussex
Falmer, Brighton BN1 9QH
UK

Tel: 01273 678108 Fax: 01273 678097
Email: CMAIA@sussex.ac.uk
Web: http://www.maths.susx.ac.uk/CMAIA

Research Report No: 2000/13

PHASE SPACE ERROR CONTROL FOR DYNAMICAL SYSTEMS II

A.R. Humphries and N. Christodoulou

Phase Space Error Control for Dynamical Systems II

A.R. Humphries[†‡] and N. Christodoulou[†]

14th September 2000

Abstract. We introduce a phase space error control, which is a generalisation of the error control first proposed in [8]. Variable time-stepping algorithms for initial value ordinary differential equations are traditionally designed to solve a problem for a fixed initial condition and over a finite time. It can be shown that these algorithms may perform poorly for long time computations which pass through a small neighbourhood of a fixed point. In this regime there are orbits that are bounded in space but unbounded in time, and the classical error-per-step or error-per-unit-step philosophy may be improved upon. A new error criterion is introduced that essentially bounds the truncation error at each step by a fraction of the solution arc length over the corresponding time interval. This new control can be incorporated within a standard algorithm as an additional constraint at negligible additional computational cost. The new criterion is shown to be admissible, in the sense that it can always be satisfied with non-zero step-sizes. It is shown that this new criterion has a positive effect on the linear stability and dynamical properties, and hence improves behaviour in the neighbourhood of stable fixed points and saddle points. Furthermore spurious fixed points are prevented. Implementation details and numerical results are given.

Key words. Adaptivity, fixed point, long time simulations, stability, stepsize.

AMS subject classifications. 65L06

1 Introduction

We are concerned with variable time-stepping methods for dynamical systems defined by autonomous initial value ordinary differential equations (ODEs)

$$y_t = f(y), \qquad y(0) = y_0 \in \mathbb{R}^m, \tag{1.1}$$

where $f : \mathbb{R}^m \to \mathbb{R}^m$ is assumed to be Lipschitz continuous.

In [8] a new type of phase space based error control was proposed. We will introduce a generalisation of that phase space error control and analyse its dynamical properties, and also address a number of implementation issues not previously fully resolved.

In a dynamical systems context an accurate solution of (1.1) over a given finite time-interval with a particular y_0 is often of little relevance; rather, it is the global behaviour of the system for general values of y_0 in the limit as $t \to \infty$ that is of interest.

When a fixed time-stepping numerical method is used to approximate the flow of (1.1) on or near to a chaotic attractor the error between the numerical approximation and exact solution grows exponentially in time. This leads us to question the meaningfulness of the numerical solution in the limit as $t \to \infty$. This issue has now been studied in detail, and the approach of considering the numerical solution as a discrete dynamical system in its own right, and then

[†]Centre for Mathematical Analysis and its Applications, School of Mathematical Sciences, University of Sussex, Brighton, BN1 9QH, UK. Supported by EPSRC Grant No. GR/M06925.

[‡]A.R. Humphries is grateful to Universidade de Aveiro, Portugal for their hospitality.

comparing the dynamics of this system with the dynamics of (1.1), has been particularly fruitful (see [13] and the references therein).

It is widely accepted that to be efficient an ODE algorithm must be adaptive; that is, the step-size must be varied according to some error measure. In contrast to the fixed step-size case, a dynamical systems oriented theory for variable step-size algorithms is far from complete. Contributions in this area include [3, 5, 6] on behaviour near stable equilibria, [7, 12] on systems with particular nonlinear structures, and [1] on spurious fixed points.

To motivate our work, we mention three areas in which typical adaptive ODE algorithms perform badly. The first is behaviour around a stable fixed point. Hall [5] showed that typical methods fail to capture the correct dynamics in this very simple and important scenario. An illustration of this behaviour was given in [8]. A second area where poor behaviour can arise was identified in [1], where it was shown that almost all adaptive explicit Runge-Kutta methods admit stable *spurious* fixed points for arbitrarily small tolerances.

The third example of poor performance of a typical adaptive ODE algorithm, and perhaps the most important in a dynamical systems context, is near to saddle points. In a chaotic attractor it is often the unstable manifolds of the fixed points which organise the flow on the attractor. The numerical solution will thus only be give a good approximation to the flow on the attractor if it models the unstable manifolds well. But to do this it must produce good approximations to the local unstable manifolds. It was shown in [8], and is illustrated again below in Figure 3 that the typical adaptive ODE algorithm fails to do this. Trajectories of (1.1) which approach a saddle point close to the stable manifold and should pass close to the fixed point before exiting close to the unstable manifold, can result in numerical trajectories which do not pass close to the fixed point and unstable manifold, or which oscillate about the unstable manifold. In this case, we cannot be confident that the numerical solution is giving a good approximation to the attractor or the dynamics on it.

To tackle these issues we introduce a new error control, the principal component of which is to demand that the numerically generated solution $\{y_n\}_{n=0}^{\infty}$ satisfies the phase space θ (PS_θ) constraint

$$\|y_{n+1} - y_n - \Delta t_n[(1-\theta)f(y_n) + \theta f(y_{n+1})]\| \leqslant \varphi \Delta t_n \|(1-\theta)f(y_n) + \theta f(y_{n+1})\|, \tag{1.2}$$

at every step, where $\varphi \in (0,1)$ is a user defined parameter akin to a tolerance, and $\theta \in [0,1]$ is also a parameter to be chosen. This is a generalisation of the PS error control introduced in [8], which corresponded to (1.2) with $\theta = 1/2$. Although the constraint is suitable for application to any numerical method, we will restrict attention in this paper to embedded Runge-Kutta pairs.

We will motivate this error control in Section 2.2, but let us immediately give some numerical examples of a typical adaptive ODE algorithm displaying poor behaviour as outlined above, and how this is remedied by the addition of the PS_θ constraint.

First consider the RK1(2) and DOPRI5(4) methods (defined in Section 2) applied to (1.1) with

$$f(y) = \begin{bmatrix} -5 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \tag{1.3}$$

and $y(0) = [1, 10^{-4}]^T$. A typical adaptive algorithm (as defined in Section 2) produces the dynamics observed in Figure 1.

For the RK1(2) method the numerical solution gives a persistent spurious oscillation, whilst the DOPRI5(4) method converges to a spurious fixed point. Although in both cases the final solution is order of the tolerance from the fixed point, the spurious behaviour persists for arbitrary small tolerances, and it is not possible to force the solution to converge to the fixed point.

If we now apply the RK1(2) method with PS_θ error control we obtain the numerical solution in Figure 2(i), where we see that the numerical solution converges to the true fixed point. In

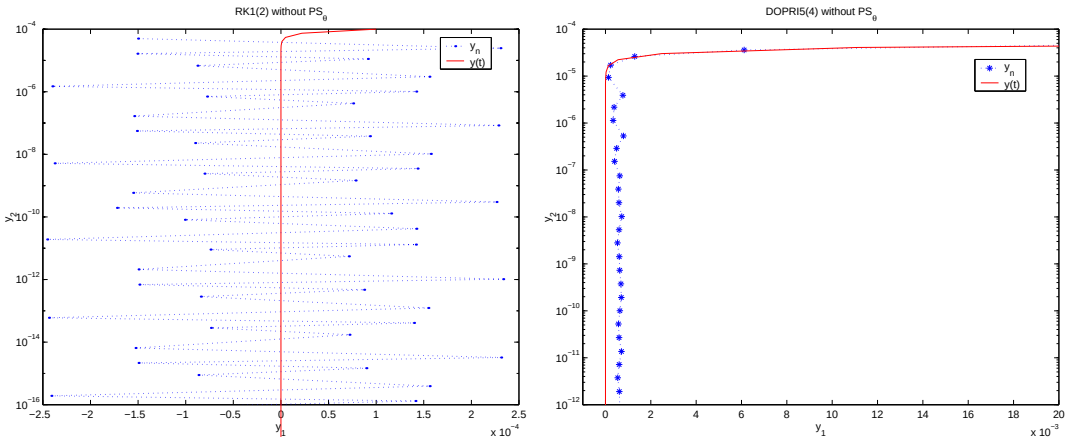

Figure 1: Numerical solutions of a typical adaptive algorithm near a stable fixed point for (i) RK1(2), (ii) DOPRI5(4).

Figure 2(ii) we show the step-sizes used by the two algorithms. The typical adaptive algorithm has many step-size rejections, whilst the PS$_\theta$ algorithm has no rejections and quickly converges to a constant value. Similar behaviour is seen for the DOPRI5(4) method.

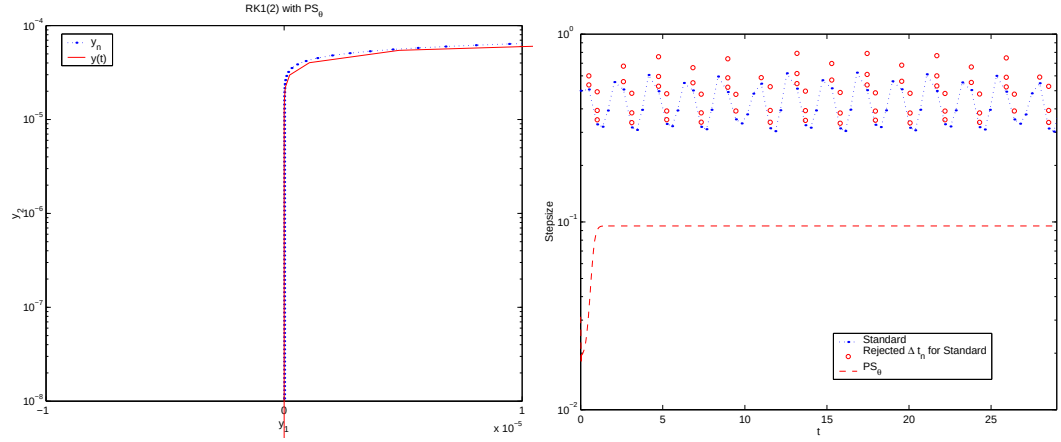

Figure 2: (i) Numerical solution using RK1(2) with PS$_\theta$ error control around a stable fixed point. (ii) Step-sizes used by the typical and PS$_\theta$ augmented algorithms.

Now, we apply the RK2(3) and Fehlberg4(5) methods to (1.1) with

$$f(y) = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \tag{1.4}$$

so that the origin is a saddle point, and take $y(0) = [0.99, 10^{-10}]^T$; very close to the stable manifold.

In one case the numerical solution has a spurious oscillation about the unstable manifold, and although this oscillation decays as the solution moves away from the fixed point, the numerical solution can ultimately end up on either side of the unstable manifold depending on the exact initial condition; thus the property of the unstable manifold of the fixed point acting as a separatrix is lost by the numerical solution. In the other example the numerical solution does not pass as close to the fixed point or the local unstable manifold as it should, and there is also

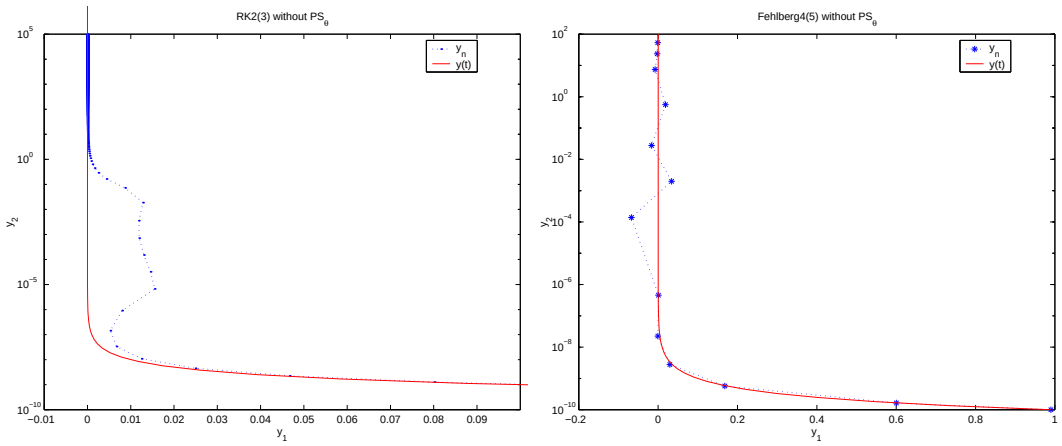

Figure 3: Numerical solutions of a typical adaptive algorithm near a saddle point for (i) RK2(3), (ii) Fehlberg4(5).

a significant phase difference between the exact and numerical solutions. As in the previous example this behaviour persists for arbitrary small tolerances.

If we now apply the RK2(3) method with PS$_\theta$ error control we obtain the numerical solution in Figure 4(i), where we see that the numerical solution follows the exact solution very closely. In Figure 4(ii) we show the step-sizes used by the two algorithms. The PS$_\theta$ algorithm quickly settles to a constant step-size whilst the solution is near the local stable manifold then adjusts to a different constant step-size whilst the solution is near to the local unstable manifold. In contrast the poorer dynamical behaviour of the typical adaptive algorithm results from the large step-sizes that it uses whilst the solution is near to the origin. Note that ultimately as y_2 becomes large the local error estimate determines and reduces the step-size in both algorithms; the different times at which it does so reveals the large phase shift introduced by the typical adaptive algorithm. Similar behaviour is seen for the Fehlberg4(5) method.

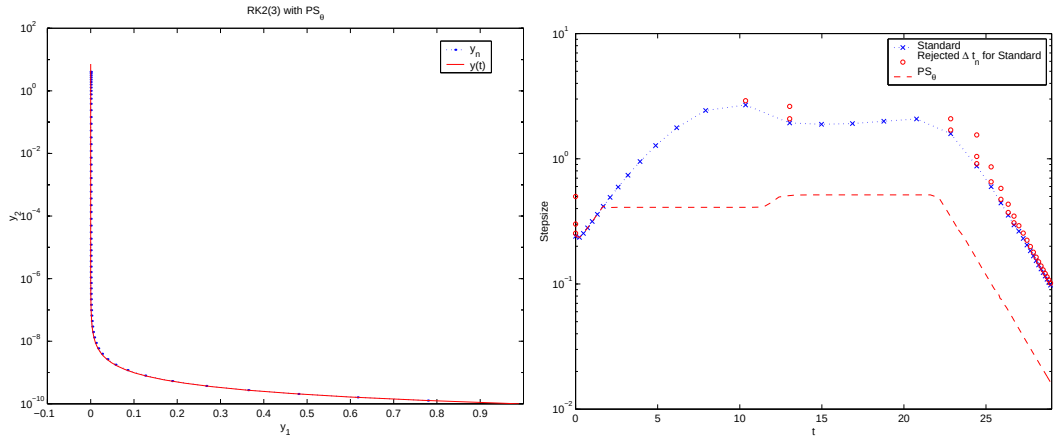

Figure 4: (i) Numerical solution using RK2(3) with PS$_\theta$ error control around a saddle point. (ii) Step-sizes used by the typical and PS$_\theta$ augmented algorithms.

In the next section we outline the traditional error control approach, and in Section 2.2 we motivate the PS$_\theta$ error control.

In Section 3 we show that the PS$_\theta$ error condition (1.2) is admissible in the sense that it can

always be satisfied with step-sizes bounded away from zero. In Section 4 we briefly show that the PS_θ error control prevents spuriosity.

In Section 5 we study the linear stability properties of PS_θ error control in detail. In [8] it was shown that with $\theta = 1/2$ on linear scalar problems the numerical solution displays decay or growth in modulus in accordance with the exact solution. Whilst this is clearly a desirable dynamical property, it does not resolve the problem of oscillatory solutions seen above. To prevent oscillations we require additionally that the linear stability function of the Runge-Kutta method $R(\lambda\Delta t_n) > 0$ for any acceptable step-size Δt_n. Only for certain methods is this achieved with $\theta = 1/2$, but we show that by increasing the value of θ this can be achieved for any method. We also consider the generalisation to the complex case.

In Section 6 we consider implementation details, which are similar to the PS error control of [8], with the exception of step-size selection. In particular we show how the PS_θ error control can be incorporated within a typical adaptive ODE algorithm as an additional constraint. Step-size selection is non-trivial for phase space error controls as they are not based on a simple error estimate, and the step-size selection scheme used in [8] often leads to highly irregular step-size sequences for solutions passing close to fixed points of non-scalar problems. We show why this arises and introduce a new step-size selection scheme which leads to stable step-sizes (with fast linear convergence to a constant value) near fixed points.

In Section 7 we give fuller details of the numerical simulations presented above, and present additional numerical simulations which illustrate and confirm our analysis, as regards the dynamics of the numerical solution and the step-size sequences near to fixed points. Although in principal the PS_θ error control (1.2) can be applied with arbitrary norm, we note that in practice the 2-norm should be preferred to the ∞-norm.

In summary, we have introduced an error control motivated from a geometrical, or phase space, viewpoint. The new control does not influence the numerical solution in most regions of phase space, but improves the performance near fixed points. More precisely, the new control is designed to positively affect the linear stability properties around true fixed points. This enhancement is particularly relevant when the numerical solution is to be driven to a stable fixed point, and more generally when computations take place around (stable or unstable) invariant manifolds. The new control is also proved to prevent spurious fixed points that might otherwise be allowed by the adaptive algorithm.

The PS_θ control analysed here was motivated by a residual test based on the theta method. There are many other geometrically-based controls that could be considered, and analysing the benefits of such controls is clearly of interest. Moreover, we hope to have illustrated that traditional error control algorithms are fundamentally tied to the finite-time/fixed initial value paradigm, and that other approaches can be fruitful for adaptive, long time simulations.

2 Embedded Runge-Kutta Formulae and Error Control

Most of the ideas in this work apply to general variable step-size algorithms. However to state precise results we focus on explicit embedded Runge-Kutta methods. We now describe the main details of a typical adaptive algorithm of the type found in numerical software libraries, and for which we have already presented numerical solutions in the introduction. Further details can be found, for example, in [4, 11].

An embedded Runge-Kutta pair is defined by

$$Y_i \;=\; y_n + \Delta t_n \sum_{j=1}^{i-1} a_{ij} f(Y_j), \qquad 1 \leqslant i \leqslant s. \tag{2.1}$$

$$y_{n+1} = y_n + \Delta t_n \sum_{i=1}^{s} b_i f(Y_i), \tag{2.2}$$

$$\widehat{y}_{n+1} = y_n + \Delta t_n \sum_{i=1}^{s} \widehat{b}_i f(Y_i), \tag{2.3}$$

where Δt_n is the step-size such that $\Delta t_n := t_{n+1} - t_n$ and $t_n = \sum_{j=0}^{n-1} \Delta t_n$. In equation (2.2), y_{n+1} gives an approximation to the solution $y(t_{n+1})$ of (1.1), whereas $\widehat{y}_{n+1}$ is the result of a different Runge-Kutta formula applied at y_n and is used only for step-size control.

The coefficients of the formula pair $\{a_{ij}, b_i, \widehat{b}_i\}$, for $1 \leqslant i \leqslant s$ and $1 \leqslant j \leqslant i-1$, define a particular method, which is usually represented using the Butcher tableau

$$\begin{array}{c|c} & A \\ \hline & b \\ & \widehat{b} \end{array}.$$

The simple RK1(2) method

$$\begin{array}{c|cc} 0 & 0 & \\ \frac{1}{2} & 0 & \\ \hline & 1 & 0 \\ & 0 & 1 \end{array}, \tag{2.4}$$

corresponding to Euler's method with second order error control is a useful test method. We will also use the RK2(3) method (also known as Fehlberg2(3))

$$\begin{array}{c|ccc} 0 & 0 & 0 & \\ 1 & 0 & 0 & \\ \frac{1}{4} & \frac{1}{4} & 0 & \\ \hline & \frac{1}{2} & \frac{1}{2} & 0 \\ & \frac{1}{6} & \frac{1}{6} & \frac{2}{3} \end{array}, \tag{2.5}$$

and the RK2(3)B, Fehlberg4(5), DOPRI5(4) and DOPRI8(7) methods the parameters of which are all stated in [4].

Equations (2.1)-(2.3) are denoted by $RKp(q)$, where p is the order of the method using Y_i and y_{n+1}, and q is the order of the method using Y_i and $\widehat{y}_{n+1}$. The order of the secondary formula $\widehat{y}_{n+1}$, may be higher or lower than that of the main formula y_{n+1}. If $p > q$, then the method is said to be in *extrapolation mode*, whereas if $p < q$, then it is said to be in *non-extrapolation mode*. Thus Fehlberg4(5) is in non-extrapolation mode, and if implemented in extrapolation mode would be denoted Fehlberg5(4).

2.1 Standard Error Control

In typical local error control the difference between the solutions y_n, $\widehat{y}_n$, yields an estimate of the local error which can be used for step-size control. This is usually given by

$$E(y_n, \Delta t_n) = \Delta t_n^{r-1} \| y_{n+1} - \widehat{y}_{n+1} \|, \tag{2.6}$$

with either $r = 1$ (Error-Per-Step (EPS)) or $r = 0$ (Error-Per-Unit-Step (EPUS)). This formula is useful both to control the local error and for time-step selection. An approximation y_{n+1} is regarded as acceptable if

$$E(y_n, \Delta t_n) \leqslant \tau, \tag{2.7}$$

where $\tau > 0$ is some user-defined tolerance.

The constraint (2.7) must be coupled to a step-size selection mechanism. The theory that we develop will be largely independent of this mechanism; thus we will not consider it in detail, but merely note that it is usually based on the formula

$$\Delta t_{n+1} = \gamma \left(\frac{\tau}{E(y_n, \Delta t_n)} \right)^{1/\widetilde{q}} \Delta t_n, \tag{2.8}$$

where $\gamma \in (0,1)$ is a safety factor and $\widetilde{q}$ is the largest integer such that $E(y_n, \Delta t_n) = \mathcal{O}(\Delta t_n^{\widetilde{q}})$; so $\widetilde{q} = \min(p,q)+1$, and thus with $q = p+1$ (non-extrapolation mode) $\widetilde{q} = q$ and with $p = q+1$ (extrapolation mode) $\widetilde{q} = p$.

2.2 Phase Space Error Control

In [8], a new error control, (PS error control), was introduced such that the local error is bounded at each step by an approximation to a fraction of the solution arc-length of the exact solution of (1.1) over the corresponding time interval. This error control can be added into the standard adaptive algorithm as an additional constraint without any significant extra computational cost. It is shown that it has positive effect in a neighbourhood of stable fixed points and furthermore spurious fixed points and period two solutions are prevented.

To bound the local error at each step as a fraction of the solution arc-length the step-size control (2.7) could be augmented by the additional constraint

$$\|\widehat{y}_{n+1} - y_{n+1}\| \leqslant \varphi \|\widehat{y}_{n+1} - y_n\|,$$

where $\|\widehat{y}_{n+1} - y_{n+1}\|$ is an error estimate, $\|\widehat{y}_{n+1} - y_n\|$ approximates the solution arc-length, and $\varphi \in (0,1)$ a user defined parameter specifies the allowable error per step as a fraction of solution arc length. However, it is difficult to analyse this error control since it contains two Runge-Kutta methods, and in practice any analysis would have to be repeated for each pair of methods. Thus, since this error control does not force closeness in some power of the step-size but in an $\mathcal{O}(1)$ sense, $\widehat{y}_{n+1}$ maybe replaced by some other Runge-Kutta formula chosen with respect to linear stability or any other properties. We replace $\widehat{y}_{n+1}$ with

$$y_n + \Delta t_n[(1-\theta)f(y_n) + \theta f(y_{n+1})] \tag{2.9}$$

to obtain the PS_θ error control (1.2);

$$\|y_{n+1} - y_n - \Delta t_n[(1-\theta)f(y_n) + \theta f(y_{n+1})]\| \leqslant \varphi \Delta t_n \|(1-\theta)f(y_n) + \theta f(y_{n+1})\|.$$

This is a generalisation of the PS control of [8] which corresponds to (1.2) with $\theta = 1/2$, and we refer to [8] for further justification/motivation of this error control.

Finally in this section we note the similarity of (2.9) and to the theta method

$$y_{n+1} = y_n + \Delta t_n[(1-\theta)f(y_n) + \theta f(y_{n+1})]. \tag{2.10}$$

The crucial difference is that when (2.9) is inserted into (1.2) the $f(y_{n+1})$ is calculated using the order p Runge-Kutta method from the embedded pair, not the theta method (2.10). Nevertheless setting $\varphi = 0$ in (1.2) would force the numerical solution to be equivalent to that of the theta method, and we will see below that dynamical properties of the theta method are still inherited by the adaptive Runge-Kutta method when $\varphi > 0$. We emphasise that (1.2) is a geometric requirement unrelated to order, and as such can be applied to a RK$p(q)$ method for arbitrary p and q.

3 Admissibility of PS_θ Error Control

In this section we demonstrate that the PS_θ error control (1.2) is admissible when applied to (2.1)-(2.2), in the sense that we can find an infinite solution sequence $\{y_n\}_{n=0}^{\infty}$ such that the error control (1.2) is satisfied at every step. Moreover we show that for this solution sequence it is not possible to have both $\{\Delta t_n\}_{n=0}^{\infty}$ and $\{y_n\}_{n=0}^{\infty}$ bounded; hence we avoid circumstances where $\{y_n\}_{n=0}^{\infty}$ remains bounded but the numerical solution does not progress beyond some *finite* time interval.

The term *admissible* was introduced in [12], where under structural assumptions on f it was proved that the infinite sequence $\{y_n\}_{n=0}^{\infty}$ is bounded with $\sum_{n=0}^{\infty}\Delta t_n$ unbounded for particular Runge-Kutta methods. In this paper will not make any structural assumptions on f, hence will be only able to show that either $\{y_n\}_{n=0}^{\infty}$ or $\sum_{n=0}^{\infty}\Delta t_n$ is unbounded.

We require some notation. Let

$$\mathbb{A} = \max_i \sum_{j=1}^{s} |a_{ij}| \qquad \text{and} \qquad \mathbb{B} = \sum_{i=1}^{s} |b_i|.$$

Note that for consistency of the Runge-Kutta method we require $\mathbb{B} \geqslant 1$. We also require the following lemma.

Lemma 3.1 *Suppose f is Lipschitz on $\mathcal{B} \subseteq \mathbb{R}^m$ with a Lipschitz constant L, $y_n \in \mathcal{B}$ and $\Delta t_n < 1/(L(\mathbb{A}+\theta\mathbb{B}))$. Then any solution of the Runge-Kutta method (2.1)-(2.2) which satisfies $Y_i \in \mathcal{B}$ for all i also satisfies*

$$\begin{aligned} &\|f(Y_i) - \theta f(y_{n+1}) - (1-\theta)f(y_n)\| \\ &\qquad \leqslant \left[\frac{L(\mathbb{A}+\theta\mathbb{B})\Delta t_n}{1 - L(\mathbb{A}+\theta\mathbb{B})\Delta t_n}\right] \|\theta f(y_{n+1}) + (1-\theta)f(y_n)\| \quad \forall i = 1, \ldots, s. \end{aligned} \tag{3.1}$$

Proof. By using the triangle inequality and the Lipschitz continuity we obtain

$$\begin{aligned} \|f(Y_i) - \theta f(y_{n+1}) - (1-\theta)f(y_n)\| &\leqslant L(1-\theta)\|Y_i - y_n\| + L\theta\|Y_i - y_{n+1}\| \\ &\leqslant L\theta\Delta t_n \|\sum_{j=1}^{s} b_j f(Y_j)\| + L\Delta t_n \|\sum_{j=1}^{s} a_{ij} f(Y_j)\| \\ &\leqslant L\Delta t_n(\mathbb{A}+\theta\mathbb{B})\mu, \end{aligned} \tag{3.2}$$

where $\mu = \max_i \|f(Y_i)\|$. Now using the triangle inequality and (3.2)

$$\begin{aligned} \|f(Y_i)\| &\leqslant \|f(Y_i) - \theta f(y_{n+1}) - (1-\theta)f(y_n)\| + \|\theta f(y_{n+1}) + (1-\theta)f(y_n)\| \\ &\leqslant L\Delta t_n(\mathbb{A}+\theta\mathbb{B})\mu + \|\theta f(y_{n+1}) + (1-\theta)f(y_n)\|. \end{aligned}$$

Thus in particular

$$\mu \leqslant L\Delta t_n(\mathbb{A}+\theta\mathbb{B})\mu + \|\theta f(y_{n+1}) + (1-\theta)f(y_n)\|,$$

and provided $\Delta t_n < 1/(L(\mathbb{A}+\theta\mathbb{B}))$ this implies that

$$\mu \leqslant \frac{1}{1 - L(\mathbb{A}+\theta\mathbb{B})\Delta t_n} \|\theta f(y_{n+1}) + (1-\theta)f(y_n)\|.$$

Hence result follows from (3.2). □

We now prove a result for the case where f is globally Lipschitz, and then consider the more general case of f locally Lipschitz. (Recall that f is said to be locally Lipschitz if f satisfies a

Lipschitz condition on every bounded subset $\mathcal{B} \subset \mathbb{R}^m$; where the Lipschitz constant may depend upon $\mathcal{B}$, [9, 13].) We do this because the essence of the proofs of the two results is the same, but the globally Lipschitz case is easier and clearer to follow since it does not require some technicalities that arise in the locally Lipschitz case.

Theorem 3.2 *Let $f : \mathbb{R}^m \to \mathbb{R}^m$ be globally Lipschitz. Then the solution sequence of the ERK formula (2.1)-(2.2) satisfies the PS_θ error control (1.2), for any $\theta \in [0, 1]$, at every step if*

$$\Delta t_n \leqslant \frac{\varphi}{L(\mathbb{A} + \theta\mathbb{B})(\mathbb{B} + \varphi)}. \tag{3.3}$$

Proof. We have

$$\begin{aligned} \|y_{n+1} - y_n - \Delta t_n[\theta f(y_{n+1}) + (1-\theta)f(y_n)]\| &= \Delta t_n \| \sum_{i=1}^{s} b_i (f(Y_i) - \theta f(y_{n+1}) - (1-\theta)f(y_n))\| \\ &\leqslant \Delta t_n \mathbb{B} \max_{1 \leqslant i \leqslant s} \|f(Y_i) - \theta f(y_{n+1}) - (1-\theta)f(y_n)\|. \end{aligned}$$

However, since $\varphi \in (0, 1)$ and by consistency $\mathbb{B} \geqslant 1$, (3.3) implies that

$$\Delta t_n < \frac{1}{L(\mathbb{A} + \theta\mathbb{B})}.$$

Hence Lemma 3.1 implies that

$$\begin{aligned} &\|y_{n+1} - y_n - \Delta t_n[\theta f(y_{n+1}) + (1-\theta)f(y_n)]\| \\ &\qquad \leqslant \Delta t_n \left[\frac{L(\mathbb{A} + \theta\mathbb{B})\mathbb{B}\Delta t_n}{1 - L(\mathbb{A} + \theta\mathbb{B})\Delta t_n} \right] \|\theta f(y_{n+1}) + (1-\theta)f(y_n)\|. \end{aligned}$$

Now from (3.3) we obtain

$$\varphi \geqslant \frac{L(\mathbb{A} + \theta\mathbb{B})\mathbb{B}\Delta t_n}{1 - L(\mathbb{A} + \theta\mathbb{B})\Delta t_n}.$$

Thus the theta error control (1.2) holds, as required. □

The above theorem shows that when f is globally Lipschitz, for any y_n we can find Δt_n and hence y_{n+1} such that the PS_θ error control (1.2) is satisfied. Thus we can always find a solution sequence $\{y_n\}_{n=0}^{\infty}$, when f is globally Lipschitz. Moreover, (3.3) shows that we can choose the solution sequence so that $\{\Delta t_n\}_{n=0}^{\infty}$ is uniformly bounded away from zero, and hence that $\sum_{n=0}^{\infty} \Delta t_n$ is unbounded.

We now consider the case where f is locally Lipschitz. We require the following lemma.

Lemma 3.3 *Suppose f is Lipschitz with Lipschitz constant L on $\mathcal{N}(\mathcal{B}, \varepsilon)$, where $\mathcal{B} \subset \mathbb{R}^m$, $\varepsilon > 0$ and*

$$\mathcal{N}(\mathcal{B}, \varepsilon) = \Big\{ x \in \mathbb{R}^m : dist(x, \mathcal{B}) < \varepsilon \Big\}, \tag{3.4}$$

and let

$$\mathcal{M} = \sup_{y \in \mathcal{N}(\mathcal{B}, \varepsilon)} \|f(y)\| < \infty. \tag{3.5}$$

If

$$\Delta t_n < \min\left(\frac{\varepsilon}{\mathbb{A}\mathcal{M}}, \frac{1}{L\mathbb{A}} \right) \tag{3.6}$$

then for any $y_n \in \mathcal{B}$ the unique solution of (2.1)-(2.2) satisfies

$$Y_i \in \mathcal{B}(y_n, \varepsilon) \subset \mathcal{N}(\mathcal{B}, \varepsilon) \ \ \forall i = 1, \ldots, s,$$

where

$$\mathcal{B}(y_n, \varepsilon) = \Big\{ x \in \mathbb{R}^m : \|x - y_n\| < \varepsilon \Big\}.$$

Proof. See Lemma 4.2.4 in [8]. □

Theorem 3.4 *Let $f : \mathbb{R}^m \to \mathbb{R}^m$ be locally Lipschitz. Then for any bounded set $\mathcal{B}$ and any $y_n \in \mathcal{B} \subset \mathbb{R}^m$ there exists $\widehat{\Delta t} = \widehat{\Delta t}(\mathcal{B}) > 0$ such that y_{n+1} in the explicit Runge-Kutta formula (2.1)-(2.2) satisfies the PS_θ condition (1.2) for all $\Delta t_n \in (0, \widehat{\Delta t}(\mathcal{B}))$, for any $\theta \in [0, 1]$.*

Proof. Choose $\varepsilon > 0$ and let $\mathcal{N}(\mathcal{B}, \varepsilon)$ and $\mathcal{M}$ be defined by equations (3.4) and (3.5) respectively, and let L be a Lipschitz constant for f on $\mathcal{N}(\mathcal{B}, \varepsilon)$. Define

$$\widehat{\Delta t} = \min\left(\frac{\varepsilon}{\mathbb{A}\mathcal{M}}, \frac{\varphi}{L(\mathbb{A} + \theta\mathbb{B})(\mathbb{B} + \varphi)}, \frac{\varepsilon}{\mathbb{B}\mathcal{M}}\right).$$

Since $\varphi \in (0, 1)$ and $\mathbb{B} \geqslant 1$ then

$$\frac{\varphi}{L(\mathbb{A} + \theta\mathbb{B})(\mathbb{B} + \varphi)} < \frac{1}{L\mathbb{A}}.$$

Hence Lemma 3.3 shows that $Y_i \in \mathcal{B}(y_n, \varepsilon)$ for all i, and since $\Delta t_n < \varepsilon/(\mathbb{B}\mathcal{M})$ we also conclude that $y_{n+1} \in \mathcal{B}(y_n, \varepsilon)$.

Now follow the proof of Theorem 3.2 applying (3.1) from Lemma 3.1 with $\mathcal{B} = \mathcal{B}(y_n, \varepsilon)$ to derive the result. □

4 Prevention of Spuriosity

We show that the PS_θ error control (1.2) does not allow the numerical solution to have spurious fixed points. Moreover when $\theta = 1/2$ it does not allow period two solutions. Note that the following result is independent of the method used to generate the solution sequence $\{y_n\}_{n=0}^{\infty}$ or the step-size sequence $\{\Delta t_n\}_{n=0}^{\infty}$.

Theorem 4.1 *An algorithm that satisfies the PS_θ control (1.2) for any $\theta \in [0, 1]$ does not allow spurious fixed points. Moreover if $\theta = 1/2$ the algorithm does not allow period two solutions.*

Proof. If $y_{n+1} = y_n = y^*$ in (1.2) then

$$(1 - \varphi)\Delta t_n \|f(y^*)\| \leqslant 0. \tag{4.1}$$

from which it follows that $f(y^*) = 0$ and hence that the method does not admit spurious fixed points.

If $\theta = 1/2$ then the method reduces to the PS control and the result follows from Theorem 5.1 of [8]. □

Although the possibility of spurious period two solutions cannot be eliminated entirely when $\theta \neq 1/2$ we will show in the next section that they can be prevented near to fixed points of linear problems for $\theta \neq 1/2$, and so should not arise near to hyperbolic fixed points of nonlinear problems.

5 Linear Stability Analysis

Consider the explicit Runge-Kutta method (2.1)-(2.3) applied to the linear scalar test problem

$$y_t = \lambda y, \qquad y(0) = y_0 \in \mathbb{R}, \tag{5.1}$$

where either $\lambda \in \mathbb{R}$ or $\lambda \in \mathbb{C}$. Then the numerical solution evolves according to

$$y_{n+1} = R(z_n) y_n, \tag{5.2}$$

where $z_n = \lambda \Delta t_n$ with Δt_n determined by the particular time-stepping strategy in use, and $R(\cdot)$ is the *linear stability function* of the method, which for an explicit s-stage method is given by

$$R(z) = \sum_{i=0}^{s} c_i z^i, \tag{5.3}$$

where by consistency $c_0 = c_1 = 1$. Recall also that the linear stability region, $\mathcal{S}$, of a Runge-Kutta method is given by $\mathcal{S} = \{z \in \mathbb{C} : |R(z)| \leqslant 1\}$.

In the next two sections we investigate the behaviour of adaptive explicit Runge-Kutta methods under PS_θ error control (1.2) when applied to this test problem for $\lambda \in \mathbb{R}$ and $\lambda \in \mathbb{C}$, respectively. These problems are relevant following the argument in [8] that in a general a single real or complex conjugate pair of eigenvalues will dominate the dynamical behaviour near to the stable and unstable manifolds of a fixed point.

Note that with (5.2) the PS_θ condition (1.2) becomes

$$\Big|R(z_n) - 1 - z_n[\theta R(z_n) + (1-\theta)]\Big| \leqslant \varphi \Big| z_n[\theta R(z_n) + (1-\theta)]\Big|. \tag{5.4}$$

5.1 Real λ

The following lemma will be useful.

Lemma 5.1 *Suppose the adaptive ERK method (2.1)-(2.2) is applied to the linear scalar test problem (5.1) with $\lambda \in \mathbb{R}$, and suppose that the PS_θ error control (1.2) is satisfied with $\theta \in [0,1]$. Let*

$$p = R(z_n) - 1, \tag{5.5}$$
$$q = -z_n[\theta R(z_n) + (1-\theta)]. \tag{5.6}$$

Then $pq < 0$, and furthermore if $p < 0$ then

$$q \geqslant \frac{1}{1+\varphi}[1 - R(z_n)]. \tag{5.7}$$

Proof. If $pq \geqslant 0$ then p and q have the same sign. Therefore

$$\begin{aligned} |R(z_n) - 1 - z_n[\theta R(z_n) + (1-\theta)]| &= |R(z_n) - 1| + |z_n[\theta R(z_n) + (1-\theta)]| \\ &> \varphi |z_n[\theta R(z_n) + (1-\theta)]| \end{aligned}$$

which contradicts (5.4). Thus $pq < 0$.

Now assume that $p < 0$ (equivalently $R(z_n) < 1$) and that

$$0 < q < \frac{1}{1+\varphi}(1 - R(z_n)). \tag{5.8}$$

Then

$$|R(z_n) - 1 - z_n[\theta R(z_n) + (1-\theta)]| \geqslant (1 - R(z_n)) - |q| > \frac{\varphi}{1+\varphi}(1 - R(z_n)), \tag{5.9}$$

by (5.8). But (5.8) also implies that

$$1 - R(z_n) > (1+\varphi)\Big|z_n[\theta R(z_n) + (1-\theta)]\Big|,$$

and combining this with (5.9) we obtain

$$\Big|R(z_n) - 1 - z_n[\theta R(z_n) + (1-\theta)]\Big| > \varphi\Big|z_n[\theta R(z_n) + (1-\theta)]\Big|,$$

which contradicts the PS_θ error control (5.4), and hence establishes (5.7). □

In the following theorem we show that the PS_θ error control (1.2) preserves the stability of (5.1), in the sense that with suitable parameters $|R(z_n)| < 1$ when $\lambda < 0$ and $|R(z_n)| > 1$ when $\lambda > 0$.

Theorem 5.2 *Suppose the adaptive ERK method (2.1)-(2.2) is applied to the linear scalar test problem (5.1) with $\lambda \in \mathbb{R}$, subject to the PS_θ error control (1.2) with $\theta \in [0,1]$ and $\varphi \in (0,1)$.*

(a) If $\lambda < 0$ then

(I) for $\theta \in (0, 1/2)$, $R(z_n) \in (-(1-\theta)/\theta, 1)$ for all $\varphi \in (0,1)$. Moreover let $\mathcal{L} < 0$ be such that $R(z) \notin [-(1-\theta)/\theta, -1]$ for all $z \leqslant \mathcal{L}$. Then $|R(z_n)| < 1$

(i) for all $\varphi \in (0, -1 - \frac{2}{\mathcal{L}(1-2\theta)})$ if $\theta \in (\frac{1}{2} + \frac{1}{\mathcal{L}}, \frac{1}{2} + \frac{1}{2\mathcal{L}})$,

(ii) for all $\varphi \in (0,1)$ if $\theta \in [\frac{1}{2} + \frac{1}{2\mathcal{L}}, \frac{1}{2})$,

(II) for $\theta \in [1/2, 1]$, $|R(z_n)| < 1$ for all $\varphi \in (0,1)$.

(b) If $\lambda > 0$ then $|R(z_n)| > 1$

(i) for all $\varphi \in (0,1)$ if $\theta \in [0, 1/2]$,

(ii) for all $\varphi \in (0,1)$ and all $\theta \in [0,1]$ if $\mathcal{M} \leqslant 1/2$,

(iii) for all $\varphi \in (0,1)$ if $\theta \in [0, 1/(2\mathcal{M})]$, and $1/2 < \mathcal{M} < 1$,

(iv) for all $\varphi \in (0, -1 + 1/(\mathcal{M}\theta))$ if $\theta \in (1/2, 1]$, and $1/2 \leqslant \mathcal{M} \leqslant 1$,

(v) for all $\varphi \in (0, -1 + 1/(\mathcal{M}\theta))$ if $\theta \in (1/2, 1/\mathcal{M}]$, and $1 < \mathcal{M} < 2$,

where $\mathcal{M}$ is the smallest number such that $|R(z)| > 1$ for all $z > \mathcal{M}$.

Proof. (a) Consider $\lambda < 0$ and suppose that $|R(z_n)| \geqslant 1$.

If $R(z_n) \geqslant 1$ then by inspection p, q have the same sign, since $z_n < 0$. But by Lemma 5.1 this cannot occur.

Now, if $R(z_n) \leqslant -1$ then $p \leqslant -2$. If also $\theta \in [1/2, 1]$, then $q \leqslant -z_n(1-2\theta) \leqslant 0$ since $z_n < 0$ and again we get a contradiction by Lemma 5.1.

It only remains to consider the case where $R(z_n) \leqslant -1$, $\theta \in (0, 1/2)$ and hence by Lemma 5.1

$$q \geqslant \frac{1}{1+\varphi}[1 - R(z_n)].$$

But using (5.6) and rearranging we obtain

$$R(z_n) \geqslant \frac{1 + z_n(1-\theta)(1+\varphi)}{1 - z_n\theta(1+\varphi)} = -\frac{1-\theta}{\theta} + \frac{1}{\theta(1 - z_n\theta(1+\varphi))} > -\frac{1-\theta}{\theta}.$$

But $R(z_n) \leqslant -1$ by assumption. So to avoid a contradiction we require

$$-1 \geqslant \frac{1 + z_n(1-\theta)(1+\varphi)}{1 - z_n\theta(1+\varphi)}.$$

Rearranging this, and also noting that $z_n < 0$ we obtain

$$z_n \leqslant -\frac{2}{(1+\varphi)(1-2\theta)} < 0.$$

It follows from (5.3) that for an explicit consistent method $|R(z)| \to \infty$ as $|z| \to \infty$. Thus there exists $\mathcal{L} < 0$ such that $R(z) \notin [-(1-\theta)/\theta, -1]$ for all $z < \mathcal{L}$. Thus it is not possible for $R(z_n) \leqslant -1$ if

$$-\frac{2}{(1+\varphi)(1-2\theta)} \leqslant \mathcal{L},$$

or equivalently

$$\varphi < -1 - \frac{2}{\mathcal{L}(1-2\theta)}.$$

Now $-1 - \frac{2}{\mathcal{L}(1-2\theta)} > 0$ if and only if $\theta > \frac{1}{2} + \frac{1}{\mathcal{L}}$, and $-1 - \frac{2}{\mathcal{L}(1-2\theta)} \geqslant 1$ if and only if $\theta \geqslant \frac{1}{2} + \frac{1}{2\mathcal{L}}$. Statement (a) of the theorem follows.

(b) Now consider $\lambda > 0$ and suppose that $|R(z_n)| \leqslant 1$, which implies that $p \leqslant 0$.

If $\theta \in [0, 1/2]$, then $1 - 2\theta \leqslant \theta R(z_n) + 1 - \theta \leqslant 1$ and thus since $z_n > 0$ we obtain $q \leqslant 0$. Thus p and q have the same sign which cannot occur by Lemma 5.1.

Now suppose that, $|R(z_n)| \leqslant 1$ and $\theta \in (\frac{1}{2}, 1]$. By Lemma 5.1 we must have

$$q \geqslant \frac{1}{1+\varphi}[1 - R(z_n)].$$

But using (5.6) and $z_n > 0$ we obtain

$$R(z_n)(1 - z_n\theta(1+\varphi)) \geqslant 1 + z_n(1-\theta)(1+\varphi) > 0. \tag{5.10}$$

Recall that we need $q > 0$ to avoid a contradiction, but $q = -z_n[\theta R(z_n) + (1-\theta)]$ and $z_n > 0$. So for $q > 0$ we require $\theta R(z_n) + (1-\theta) < 0$ which implies that

$$R(z_n) < -\frac{1-\theta}{\theta} \leqslant 0,$$

for $\theta \in (1/2, 1]$. But substituting $R(z_n) < 0$ in (5.10) implies that $1 - z_n\theta(1+\varphi) < 0$, and thus

$$z_n > \frac{1}{\theta(1+\varphi)} > 0.$$

But, for an explicit method the domain of absolute stability $\mathcal{S}$ is bounded. So there exists $\mathcal{M} \geqslant 0$ such that $|R(z)| > 1$ for all $z > \mathcal{M} \geqslant 0$. Now if

$$\frac{1}{\theta(1+\varphi)} \geqslant \mathcal{M} \tag{5.11}$$

we derive a contradiction since we have shown that for $|R(z_n)| \leqslant 1$ we need $z_n > \frac{1}{\theta(1+\varphi)} \geqslant \mathcal{M} \geqslant z_n$. Rearranging (5.11) we obtain that

$$\varphi < -1 + \frac{1}{\theta\mathcal{M}}, \tag{5.12}$$

guarantees $|R(z_n)| > 1$. Now $-1 + \frac{1}{\theta\mathcal{M}} > 0$ if and only if $\theta \leqslant 1/\mathcal{M}$, and $-1 + \frac{1}{\theta\mathcal{M}} \geqslant 1$ if and only if $\theta \geqslant 1/(2\mathcal{M})$. This completes the proof. □

Remark For many common methods, including all methods with $c_i \geqslant 0$ for all i, the stability function (5.3) satisfies $R(z) > 1$ for all $z > 0$. For such methods Theorem 5.2(b) is trivial, and so Theorem 5.2(a) is the more significant result.

Theorem 5.2 is unsatisfactory for simulation of dynamical systems. Although it guarantees that for a linear scalar problem the fixed point of the numerical solution will be stable (unstable) when the fixed point of the underlying dynamical system is stable (unstable), it allows the possibility of $R(z_n) < 0$. As was argued in the introduction, this is very undesirable as it allows spurious oscillations to be introduced to the numerical simulation. Thus we seek stronger conditions on the parameters to ensure that $R(z_n) > 0$.

Theorem 5.3 *Suppose the adaptive ERK method (2.1)-(2.2) is applied to the linear scalar test problem (5.1) with $\lambda \in \mathbb{R}$, subject to the PS_θ error control (1.2) with $\theta \in [0, 1]$ and $\varphi \in (0, 1)$.*

(a) If $\lambda < 0$ then $0 < R(z_n) < 1$

(i) for all $\varphi \in (0, -1 - \frac{1}{\mathcal{L}^*(1-\theta)})$ for $\theta \in (1 - \frac{1}{\mathcal{L}^*}, 1 - \frac{1}{2\mathcal{L}^*})$,

(ii) for all $\varphi \in (0, 1)$ for $\theta \in [1 - \frac{1}{2\mathcal{L}^*}, 1]$,

where $\mathcal{L}^* < 0$ is such that $R(z) \notin [-(1-\theta)/\theta, 0]$ for all $z \leqslant \mathcal{L}^*$.

(b) If $\lambda > 0$ then $R(z_n) > 1$

(i) for all $\varphi \in (0, 1)$ if $\theta = 0$.

If, furthermore there exists $\mathcal{M}^* \geqslant 0$ such that $R(z) > 1$ for all $z \geqslant \mathcal{M}^*$ then $R(z_n) > 1$

(ii) for all $\varphi \in (0, 1)$ and all $\theta \in [0, 1]$ if $\mathcal{M}^* \leqslant 1/2$,

(iii) for all $\varphi \in (0, 1)$ if $\theta \in [0, 1/(2\mathcal{M}^*)]$, and $\mathcal{M}^* > 1/2$,

(iv) for all $\varphi \in (0, -1 + 1/(\mathcal{M}^*\theta))$ if $\theta \in (1/(2\mathcal{M}^*), 1]$, and $1/2 < \mathcal{M}^* \leqslant 1$,

(v) for all $\varphi \in (0, -1 + 1/(\mathcal{M}^*\theta))$ if $\theta \in (1/(2\mathcal{M}^*), 1/\mathcal{M}^*)$, and $\mathcal{M}^* > 1$.

Proof. (a) Noting that $\mathcal{L}^* \leqslant \mathcal{L}$, it follows from Theorem 5.2 that $|R(z_n)| < 1$ under the conditions given in (a), and it only remains to prove that $R(z_n) > 0$. Let p and q again be defined by (5.5) and (5.6) respectively.

If $R(z_n) \leqslant 0$ then $p \leqslant -1$. Hence by Lemma 5.1 if the PS_θ control (1.2) is satisfied with $R(z_n) \leqslant 0$ and $z_n < 0$ then

$$q \geqslant \frac{1}{1+\varphi}[1 - R(z_n)]$$

which implies that

$$R(z_n) \geqslant \frac{1 + z_n(1-\theta)(1+\varphi)}{1 - z_n\theta(1+\varphi)} > -\frac{(1-\theta)}{\theta}.$$

Note that, since by assumption $R(z_n) \leqslant 0$, this gives an immediate contradiction for $\theta = 1$. For $\theta \in [0, 1)$ to avoid a contradiction we require

$$0 \geqslant R(z_n) \geqslant \frac{1 + z_n(1-\theta)(1+\varphi)}{1 - z_n\theta(1+\varphi)},$$

and thus since $z_n < 0$,

$$0 \geqslant 1 + z_n(1-\theta)(1+\varphi),$$

or on rearranging,

$$z_n \leqslant -\frac{1}{(1+\varphi)(1-\theta)} < 0.$$

It follows from (5.3) that for an explicit consistent method $|R(z)| \to \infty$ as $|z| \to \infty$. Thus there exists $\mathcal{L}^* < 0$ such that $R(z_n) \notin [-(1-\theta)/\theta, 0]$ for all $z_n < \mathcal{L}^*$. Thus it is not possible for $R(z_n) \leqslant 0$ if

$$-\frac{1}{(1+\varphi)(1-\theta)} \leqslant \mathcal{L}^*,$$

or equivalently

$$\varphi < -1 - \frac{1}{\mathcal{L}^*(1-\theta)}.$$

Now $-1 - \frac{1}{\mathcal{L}^*(1-\theta)} > 0$ if and only if $\theta > 1 + \frac{1}{\mathcal{L}^*}$, and $-1 - \frac{1}{\mathcal{L}^*(1-\theta)} \geqslant 1$ if and only if $\theta \geqslant 1 + \frac{1}{2\mathcal{L}^*}$. Statement (a) of the theorem follows.
(b) Consider $\lambda > 0$, and note that since $\mathcal{M}^* \geqslant \mathcal{M}$, it follows from Theorem 5.2 that $|R(z_n)| > 1$ under the conditions given in (b), and hence it only remains to prove that $R(z_n) > 0$.

If $R(z_n) < -1$ then $p \leqslant -2$. Hence by Lemma 5.1

$$q \geqslant \frac{1}{1+\varphi}[1 - R(z_n)] \geqslant 1$$

which implies

$$R(z_n)(1 - z_n\theta(1+\varphi)) \geqslant 1 + z_n(1-\theta)(1+\varphi) > 0. \tag{5.13}$$

But substituting $R(z_n) < -1$ in (5.13) implies that $1 - z_n\theta(1+\varphi) < 0$ (which proves the $\theta = 0$ case), and thus for $\theta \neq 0$,

$$z_n > \frac{1}{\theta(1+\varphi)} > 0.$$

Now if the stability function (5.3) has $c_s > 0$ then there exists $\mathcal{M}^* \geqslant 0$ such that $R(z) > 1$ for all $z > \mathcal{M}^* \geqslant 0$. Now if

$$\frac{1}{\theta(1+\varphi)} \geqslant \mathcal{M}^* \tag{5.14}$$

we derive a contradiction since we have shown that for $R(z_n) < -1$ we need $z_n > \frac{1}{\theta(1+\varphi)} \geqslant \mathcal{M}^* \geqslant z_n$. Rearranging (5.14) we obtain that

$$\varphi < -1 + \frac{1}{\theta\mathcal{M}^*}, \tag{5.15}$$

guarantees $R(z_n) > 1$. Now $-1 + \frac{1}{\theta\mathcal{M}^*} > 0$ if and only if $\theta \leqslant 1/\mathcal{M}^*$, and $-1 + \frac{1}{\theta\mathcal{M}^*} \geqslant 1$ if and only if $\theta \geqslant 1/(2\mathcal{M}^*)$. This completes the proof. □

Remark Theorem 5.3 shows that if $\lambda < 0$ and the PS_θ error control (1.2) is satisfied for any $\varphi \in (0,1)$ then provided $\theta \leqslant 1$ is sufficiently large then $0 < R(z_n) < 1$, and so the numerical approximation to the solution of (5.1) converges monotonically to the fixed point, just like the exact solution to (5.1). For many methods, including all methods with $c_i \geqslant 0$ for all i, the stability function (5.3) satisfies $R(z) > 1$ for all $z > 0$, and so for $\lambda > 0$ we have $R(z_n) > 1$ and hence monotonic divergence of the numerical solution from the fixed point, irrespective of the time-stepping strategy. For methods with $c_s > 0$, but $R(z) \leqslant 1$ for some $z > 0$, the theorem shows that if $\lambda > 0$ and the PS_θ error control (1.2) is satisfied for any $\varphi \in (0,1)$ then provided

$\theta \geqslant 0$ is sufficiently small then $R(z_n) > 1$, and so the numerical approximation to the solution of (5.1) diverges monotonically from the fixed point. However, for methods with $c_s < 0$ in the stability function $R(z)$ (5.3), including the DOPRI8(7) method, the theorem only guarantees that $R(z_n) > 1$ for $\lambda > 0$ if $\theta = 0$.

We denote the stability function of the two-stage theta method (2.10) by $R_\theta(z)$ so that

$$R_\theta(z) = \frac{1 + (1-\theta)z}{1 - \theta z}, \tag{5.16}$$

and now show how the numerical solution to (5.1) generated by a PS_θ controlled method is geometrically related to that of the two-stage theta method.

Theorem 5.4 *Suppose the adaptive ERK method (2.1)-(2.2) is applied to the linear scalar test problem (5.1) with $\lambda \in \mathbb{R}$, subject to the PS_θ error control (1.2) with $\theta \in [0,1]$ and $\varphi \in (0,1)$.*

(i) If $\lambda < 0$ and (θ, φ) satisfy one of Theorem 5.3(a)(i)-(ii) then

$$R_\theta((1+\varphi)z_n) \leqslant R(z_n) \leqslant R_\theta((1-\varphi)z_n).$$

(ii) If $\lambda > 0$ and (θ, φ) satisfy one of Theorem 5.3(b)(i)-(v) then $0 < z_n < 1/[\theta(1-\varphi)]$ and

$$R_\theta((1-\varphi)z_n) \leqslant R(z_n).$$

If furthermore, $0 < z_n < 1/[\theta(1+\varphi)]$, then

$$R_\theta((1-\varphi)z_n) \leqslant R(z_n) \leqslant R_\theta((1+\varphi)z_n).$$

Proof. (i) By Theorem 5.3(a), $0 < R(z_n) < 1$. Hence by Lemma 5.1

$$-z_n[\theta R(z_n) + (1-\theta)] \geqslant \frac{1}{1+\varphi}[1 - R(z_n)]$$

which implies that

$$R(z_n) \geqslant \frac{1 + z_n(1-\theta)(1+\varphi)}{1 - z_n\theta(1+\varphi)} = R_\theta(z_n).$$

Now, suppose that the right-hand inequality fails so that $R_\theta((1-\varphi)z_n) < R(z_n)$. Using (5.16) and rearranging (noting that $1 - \theta(1-\varphi)z_n > 0$), we have

$$R(z_n) - 1 - [\theta R(z_n) + (1-\theta)]z_n > -\varphi z_n[\theta R(z_n) + (1-\theta)].$$

But by Theorem 5.3(a), $0 < R(z_n) < 1$ if $\lambda < 0$. Thus $1 - \theta < \theta R(z_n) + (1-\theta) < 1$, and the term on the right-hand side is positive. Therefore

$$|R(z_n) - 1 - [\theta R(z_n) + (1-\theta)]z_n| > \varphi |z_n[\theta R(z_n) + (1-\theta)]|,$$

which contradicts (5.4), thus the right-hand side inequality holds.
(ii) Now consider $\lambda > 0$. Theorem 5.3(b) shows $R(z_n) > 1$. Suppose that

$$R(z_n)[1 - \theta(1-\varphi)z_n] < 1 + (1-\theta)(1-\varphi)z_n,$$

then

$$R(z_n) - 1 - z_n[\theta R(z_n) + (1-\theta)] < -\varphi z_n[\theta R(z_n) + (1-\theta)].$$

But since the right-hand side is negative, this implies that

$$|R(z_n) - 1 - z_n[\theta R(z_n) + (1-\theta)]| > \varphi |z_n[\theta R(z_n) + (1-\theta)]|$$

which contradicts (5.4) and hence we have

$$R(z_n)[1 - \theta(1-\varphi)z_n] \geqslant 1 + (1-\theta)(1-\varphi)z_n. \tag{5.17}$$

Since the right-hand side of (5.17) is positive, so is the left-hand side, which implies that

$$z_n < \frac{1}{\theta(1-\varphi)}.$$

Now dividing (5.17) by $1 - \theta(1-\varphi)z_n$ (which we have just shown to be positive) gives

$$R(z_n) \geqslant \frac{1 + (1-\theta)(1-\varphi)z_n}{1 - \theta(1-\varphi)z_n} = R_\theta((1-\varphi)\lambda z_n),$$

as required. Now suppose further that $0 < z_n < 1/[\theta(1+\varphi)]$. If

$$R(z_n) > \frac{1 + (1-\theta)(1+\varphi)z_n}{1 - \theta(1+\varphi)z_n},$$

then

$$R(z_n) - 1 - z_n[\theta R(z_n) + (1-\theta)] > \varphi z_n[\theta R(z_n) + (1-\theta)].$$

But $R(z_n) > 1$ implies that the right-hand side is positive and hence

$$|R(z_n) - 1 - z_n[\theta R(z_n) + (1-\theta)]| > \varphi |z_n[\theta R(z_n) + (1-\theta)]|$$

which contradicts (5.4). Hence we must have

$$R(z_n) \leqslant \frac{1 + (1-\theta)(1+\varphi)z_n}{1 - \theta(1+\varphi)z_n} = R_\theta((1+\varphi)z_n. \qquad \square$$

Whilst the preceding theorems give an indication of the behaviour of PS_θ error control for $z \in \mathbb{R}$, the behaviour can be much better for particular methods. In particular, the Fehlberg 2(3) method (in non-extrapolation mode) and the DOPRI5(4) method (in extrapolation mode) have the properties that

(i) $R(z) > 0$ for all $z \in \mathbb{R}$,

(ii) $R(z) > 1$ for all $z > 0$.

Thus for these methods under PS_θ error control (1.2) applied to (5.1) we have

$$\left.\begin{array}{c} 0 < R(z_n) < 1 \text{ if } \lambda < 0, \\ 1 < R(z_n) \text{ if } \lambda > 0, \end{array}\right\} \forall \varphi \in (0,1),\ \forall \theta \in [0,1], \tag{5.18}$$

making them particularly good candidates to use with this error control.

For general methods with $\lambda < 0$, Theorem 5.3(a)(i-ii) implies the existence of $0 \leqslant \theta^- < \theta^+ < 1$ such that $0 < R(z_n) < 1$ if $\lambda < 0$,

(i) for $\varphi \in (0, \varphi^*(\theta))$ for $\theta \in (\theta^-, \theta^+)$,

(ii) for all $\varphi \in (0,1)$ for $\theta \in [\theta^+, 1]$.

The values of θ^-, θ^+ and $\varphi^*(\theta)$ are easy to calculate using matlab [10]. Let $z^* < 0$ be such that $R(z^*) = 0$, then since by Theorem 5.2 $R(z_n) < 1$, we only need to find the values of θ, φ at which $R(z^*) = 0$ and z^* satisfies (5.4) with equality, so that

$$\varphi^* = \frac{|-1 - z^*(1-\theta)|}{|z^*(1-\theta)|}. \tag{5.19}$$

The values of θ^- and θ^+ correspond to $\varphi^* = 0$ and $\varphi^* = 1$ respectively in (5.19), and hence $\theta^- = 1 + 1/z^*$ and $\theta^+ = 1 + 1/(2z^*)$, where of course z^* depends on the method (2.2). The resulting values for some popular methods are given in Table 1. The corresponding values of φ^* for $\theta \in [0, 1]$ are plotted in Figure 5 for several of the methods; graphs are similar for other methods.

Method	θ^-	θ^+
RK1(2)	0	0.5
RK2(3)B	0.3746	0.6873
Fehlberg4(5)	0.5138	0.7569
Fehlberg5(4)	0.5760	0.7880
DOPRI8(7)	0.7285	0.8643

Table 1: Values of θ^- and θ^+ for some popular methods

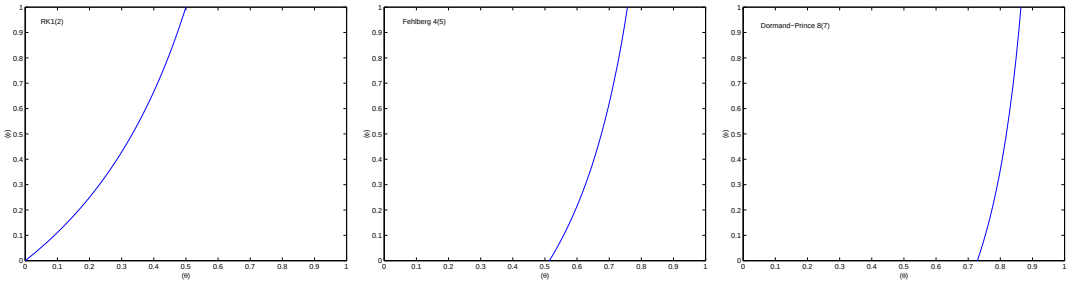

Figure 5: The maximum value φ^* as a function of λ, such that $0 < R(z_n) < 1$ when (5.4) is satisfied for $\lambda < 0$ with $\varphi \in (0, \varphi^*)$ for (i) RK1(2) (ii) Fehlberg4(5) (iii) DOPRI8(7).

Note that with the exception of DOPRI8(7) all of these methods have $R(z) > 1$ for all $z > 0$ and so satisfy $R(z_n) > 1$ when $\lambda > 0$ regardless of the error control used.

5.2 Complex λ

We now indicate how far the results above can be extended to the case of complex λ. It will be useful to define the acceptable region, $\mathcal{Q}(\varphi, \theta) \in \mathbb{C}$, for any $\theta \in [0, 1]$ and any $\varphi \in (0, 1)$ to be the set of points in $\mathbb{C}$ which satisfy (5.4). Note that for a consistent method

$$R(z) - 1 - z[\theta R(z) + (1-\theta)] = \mathcal{O}(z^2),$$

$$z[\theta R(z) + (1-\theta)] = z + \mathcal{O}(z^2),$$

and that both expressions are polynomials of degree $s + 1$. Thus by (5.4) for any $\varphi > 0$, the acceptable region $\mathcal{Q}(\varphi, \theta)$ contains a neighbourhood of the origin. Also, since the terms of degree $s + 1$ have different coefficients for $\varphi \in (0, 1)$ the acceptable region $\mathcal{Q}(\varphi, \theta)$ is bounded.

Since as noted in [8], for explicit methods the stability region intersects the imaginary axis in a neighbourhood of the origin only at the origin, it follows that there will exist points close to the origin and the imaginary axis which satisfy (5.4) but for which either $Re(\lambda) < 0$ and $|R(\lambda\Delta t_n)| > 1$ or $Re(\lambda) > 0$ and $|R(\lambda\Delta t_n)| < 1$. As noted in [8], the concept of $A(\alpha)$-stability can be generalized to this situation, although we will not pursue this further here.

It follows trivially from (5.4) that the acceptable region $\mathcal{Q}(\varphi, \theta)$ is monotonically increasing in φ; that is $\mathcal{Q}(\varphi_1, \theta) \subseteq \mathcal{Q}(\varphi_2, \theta)$ if $\varphi_1 \leqslant \varphi_2$. In order to gain some insight we investigate $\mathcal{Q}(0, \theta)$ which is given by solving

$$R(z) - 1 - z[\theta R(z) + (1 - \theta)] = 0. \tag{5.20}$$

For an explicit s-stage method (5.20) is a polynomial of degree $s + 1$ and hence has $s + 1$ roots. For every consistent method at least two of these roots are at the origin, and at least three roots are at the origin for method of at least second order if $\theta = 1/2$. The location of the other roots influences the acceptable region.

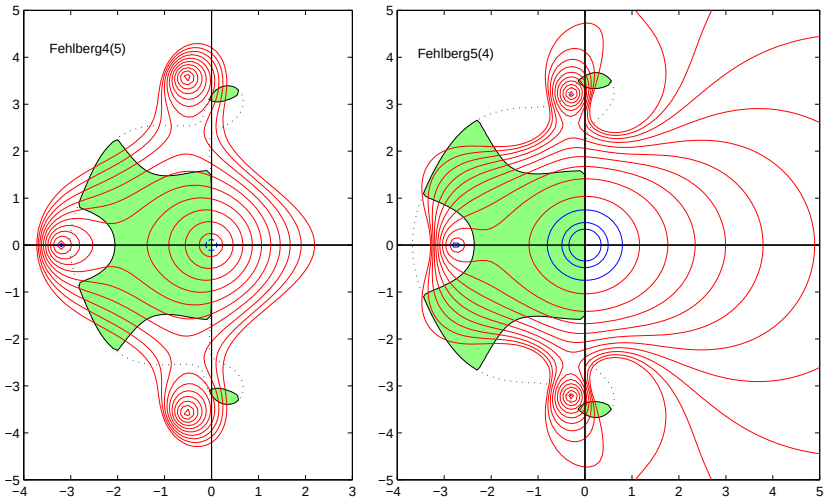

Figure 6: The regions of absolute stability (with the subset on which $Re(R(z)) \geqslant 0$ shaded) and the acceptable regions $\mathcal{Q}(\varphi, \theta)$ for $\varphi = 0.01, 0.02, 0.05, 0.1, 0.2, 0.3, \ldots, 1$, for (i) the Fehlberg4(5) method with $\theta = 0.1$, and (ii) the Fehlberg5(4) method with $\theta = 0.5$.

In Figure 6, the acceptable regions $\mathcal{Q}(\varphi, \theta)$ for the Fehlberg 4(5) method and its extrapolation version Fehlberg 5(4) are shown for various values of φ. That $\mathcal{Q}(\varphi, \theta)$ is monotonically increasing in φ with its form for small φ dependent on its behaviour for $\varphi = 0$ is clearly apparent from the plots. Note that for the Fehlberg4(5) method with $\theta = 0.1$ there are three distinct non-zero roots of (5.20), one of which is real, and all of which fall in the left half-plane outside the stability region of the method. This means that if we apply PS_θ error control to this method with $\theta = 0.1$ and a negative eigenvalue, however close to zero we take φ we cannot ensure that the numerical solution mimics the contraction of (5.1). By Theorem 5.2(a)(I)(i) we can ensure contraction of the numerical solution for small φ in the case of a real negative eigenvalue by taking θ sufficiently large. This result can also be seen to extend to the complex case, since for $0 < \varphi \ll 1$, all points of $\mathcal{Q}(\varphi, \theta)$ are close to points of $\mathcal{Q}(0, \theta)$, where the latter set is given by (5.20), or on rearranging,

$$R(z) = \frac{1 + (1 - \theta)z}{1 - z\theta} = R_\theta(z). \tag{5.21}$$

Since the region of absolute stability $\mathcal{S}_\theta$ of the theta method (2.10) contains the entire left half complex plane for any $\theta \geqslant 1/2$, it follows that for $\theta \geqslant 1/2$, if $z \in \mathcal{Q}(0, \theta)$ with $Re(z) < 0$ then

$z \in \mathcal{S}_\theta$, hence $|R(z)| = |R_\theta(z)| < 1$ and so $z \in \mathcal{S}$ the region of absolute stability of the method (2.2). Thus for all φ sufficiently small and all θ sufficiently large $\mathcal{Q}_-(\varphi,\theta) \subseteq \mathcal{S}_-$ where $\mathcal{Q}_-(\varphi,\theta)$ and $\mathcal{S}_-$ are the subsets of $\mathcal{Q}(\varphi,\theta)$ and $\mathcal{S}$ in the left half complex plane.

In Figure 6(ii), we indeed see that $\mathcal{Q}(0,\theta) \subseteq \mathcal{S}$ for the Fehlberg 5(4) method with $\theta = 0.5$, so that $\mathcal{Q}_-(\varphi,1/2) \subseteq \mathcal{S}_-$ for φ sufficiently small. However we would like a stronger result. There are five points on the boundary of $\mathcal{S}$ for the Fehlberg 5(4) method with $R(z) = -1$, corresponding to spurious period two solutions; however by Theorem 4.1 these points are excluded from $\mathcal{Q}(\varphi,1/2)$. But as $\theta < \theta^-$ (see Table 1) there do exist points in $\mathcal{Q}(\varphi,1/2)$ on the negative real axis with $R(z) < 0$. We have already noted in Theorem 5.3 and its preamble that $R(z_n) < 0$ results in spatial oscillations which lead to poor approximations to dynamical behaviour. In the complex case, the $\arg(R(z_n))$ corresponds to a rotation, with $\arg(R(z_n)) = \pi$ again corresponding to a spurious oscillation, and $\arg(R(z_n))$ close to π giving a perturbed oscillation, and chaotic looking dynamics. All such spuriosity can be avoided by choosing parameters such that for $z \in \mathcal{Q}(\varphi,\theta)$ with $Re(z) < 0$, not only $|R(z)| < 1$ but also $Re(R(z)) \geqslant 0$; which corresponds to $|\arg(R(z))| \leqslant \pi/2$ so that we require the numerical solution to perform at least four steps when approximating a periodic orbit. This is achievable, since it follows from (5.21) that for $z \in \mathcal{Q}(0,\theta)$,

$$Re(R(z)) \geqslant 0, \quad \text{if} \quad \left| z - \frac{1-2\theta}{2\theta(1-\theta)} \right| \leqslant \frac{1}{2\theta(1-\theta)}. \tag{5.22}$$

In particular for $\theta = 1$, $Re(R(z)) > 0$ for all z such that $Re(z) < 1$. Thus for all φ sufficiently small and all θ sufficiently large $\mathcal{Q}_-(\varphi,\theta) \subseteq \mathcal{S}_-^+$ where $\mathcal{S}_-^+$ is the subset of $\mathcal{S}$ in the left half complex plane on which $Re(R(z)) \geqslant 0$.

Note that for both cases shown in Figure 6 there are points of $\mathcal{Q}(0,\theta)$ close to the imaginary axis, and to the boundary of $\mathcal{S}$, and that for both methods we could get incorrect stability of the numerical solution for near imaginary eigenvalues. However λ imaginary corresponds to a non-hyperbolic fixed point, and for λ close to imaginary the fixed point is close to being non-hyperbolic. This suggests that PS_θ error control might not perform too well near non-hyperbolic fixed points. However, since standard time-stepping performs poorly even for hyperbolic fixed points, in this paper we will content ourselves with trying to derive a method which performs well near hyperbolic fixed points, and will not address further the behaviour of the method with near imaginary eigenvalues.

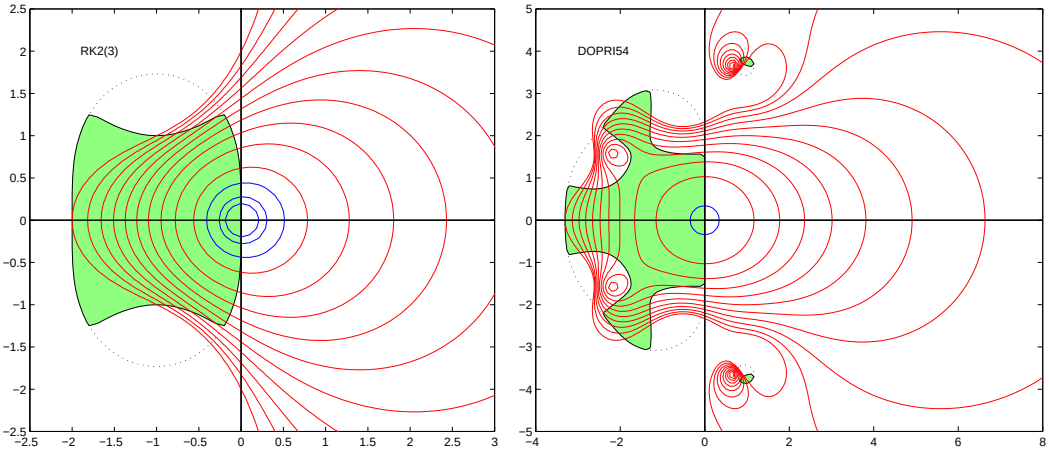

Figure 7: The regions of absolute stability (with the subset on which $Re(R(z)) \geqslant 0$ shaded) and the acceptable regions $\mathcal{Q}(\varphi,\theta)$ for $\varphi = 0.01, 0.02, 0.05, 0.1, 0.2, 0.3, \ldots, 1$, for (i) the Fehlberg2(3) method and (ii) the DOPRI5(4) method, both with $\theta = 0.5$.

In Figure 7, the acceptable regions $\mathcal{Q}(\varphi,\theta)$ for the Fehlberg 2(3) and DOPRI5(4) methods are shown for various values of φ with $\theta = 1/2$. Recall that the Fehlberg2(3) and DOPRI5(4) methods satisfy (5.18). We consider $\theta = 0.5$, since then Theorem 4.1 guarantees that the methods admit no spurious two solutions. We see from Figure 7(i) that the Fehlberg2(3) method behaves excellently with

1. $|R(z_n)| > 1$ for all $z_n \in \mathcal{Q}(\varphi, 1/2)$ with $Re(z_n) > 0$ for all $\varphi \in (0,1)$
2. $|R(z_n)| < 1$ and $Re(R(z_n)) > 0$ for all $z_n \in \mathcal{Q}(\varphi, 1/2)$ with $Re(z_n) < 0$ except near to the imaginary axis, for all $\varphi \in (0, 0.5)$.

The DOPRI5(4) method does not perform quite so well in that there are two points $z, \overline{z} \in \mathcal{Q}(0, 1/2)$ with $Re(R(z)) < 0$. However unless $\arg(\lambda)$ is exactly equal to $\arg(z)$ or $\arg(\overline{z})$, by choosing φ sufficiently small we can ensure that $Re(R(z_n)) > 0$, if we wish to. Alternatively following (5.22) we can increase the value of θ to ensure that $Re(R(z)) > 0$ for all $z \in \mathcal{Q}(0,\theta)$.

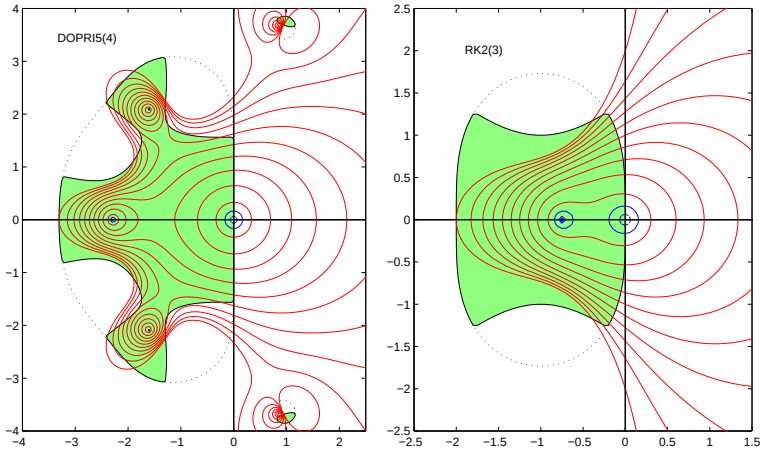

Figure 8: The regions of absolute stability (with the subset on which $Re(R(z)) \geqslant 0$ shaded) and the acceptable regions $\mathcal{Q}(\varphi,\theta)$ for $\varphi = 0.01, 0.02, 0.05, 0.1, 0.2, 0.3, \ldots, 1$, for (i) the DOPRI5(4) method and (ii) the Fehlberg2(3) method, both with $\theta = 0.8$.

In Figure 8(i) the acceptable region $\mathcal{Q}(\varphi,\theta)$ for DOPRI5(4) method is shown for various values of φ with $\theta = 0.8$. Note that $Re(R(z)) > 0$ for all $z \in \mathcal{Q}(0,\theta)$, however there is now a $z \in \mathcal{Q}(0,\theta)$ with $z < 0$ (and real). The same can be seen to occur in In Figure 7(ii) for the Fehlberg2(3) method. Real non-zero values of $z \in \mathcal{Q}(0,\theta)$ always occur for θ near to $1/2$ for methods of order 2 and above, since then, as noted already, (5.20) has three roots at zero for $\theta = 1/2$ and two roots at zero for $\theta \neq 1/2$. Real non-zero $z \in \mathcal{Q}(0,\theta)$ is undesirable for two reasons. Firstly, we would like to be able to force reductions in the step-size by reducing φ, so that we can achieve arbitrary accuracy of the numerical solution. From Figure 7 we see that this cannot be guaranteed for $\lambda < 0$ in this case, since there is a $z \in \mathcal{Q}(0,\theta)$ with $z < 0$ which satisfies (5.4) for all $\varphi \in (0,1)$. Nevertheless, this value of z is below the linear stability limit, so the method will still drive the solution to the fixed point, thus performing better than the standard error control. Secondly, in this case

$$\frac{\Big|R(z) - 1 - z[\theta R(z) + (1-\theta)]\Big|}{\Big|z[\theta R(z) + (1-\theta)]\Big|}, \tag{5.23}$$

is not monotonically increasing as $z < 0$ becomes larger in modulus. Thus, when we seek to set up an algorithm, in Section 6, to drive the step-size to a limit $\Delta t_n \to \Delta t^*$ such that (5.4) is

satisfied with the ratio (5.23) equal to $\chi\varphi$ for some safety factor $\chi < 1$, there will be multiple possible values of Δt^*. Thus, in general, PS_θ error control (1.2) should not be applied with $\theta \approx 1/2$ but $\theta \neq 1/2$, or when there exists real $z \in \mathcal{Q}(0,\theta)$ with $z < 0$.

In Figure 9 the acceptable regions $\mathcal{Q}(\varphi,\theta)$ for the methods in Table 1 are shown for various values of φ with $\theta = \theta^+$. Recall that θ^+ is defined so that the boundary of $\mathcal{Q}(1,\theta^+)$ crosses the negative real axis at the point at which $R(z) = 0$. With the exception of the RK1(2) method, φ sufficiently small ensures that if $z \in \mathcal{Q}(\varphi,\theta^+)$ with z not close to the imaginary axis then not only is $z \in \mathcal{S}$ but also $Re(R(z)) > 0$. Thus with these parameters these methods do not admit spurious oscillations near to fixed points.

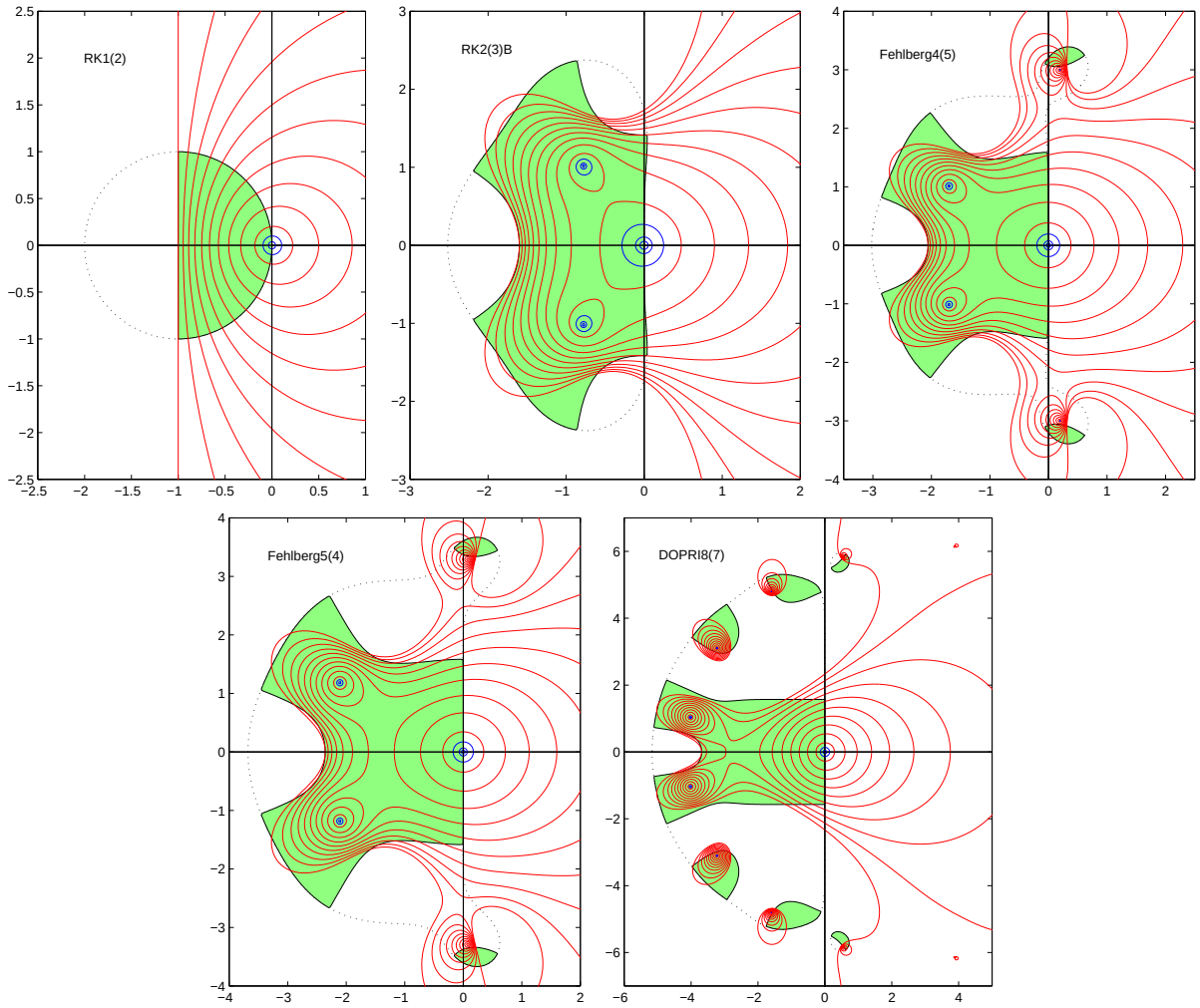

Figure 9: The regions of absolute stability (with the subset on which $Re(R(z)) \geqslant 0$ shaded) and the acceptable regions $\mathcal{Q}(\varphi,\theta)$ for $\varphi = 0.01, 0.02, 0.05, 0.1, 0.2, 0.3, \ldots, 1$, for the methods in Table 1 with $\theta = \theta^+$.

We note finally that the choice of $\theta = \theta^+$ is not arbitrary. These methods do not perform as well for θ significantly smaller than θ^+ for two reasons. Firstly if $\theta < \theta^-$ then there exist points of $\mathcal{Q}(0,\theta)$ on the negative real axis either outside of $\mathcal{S}$ or inside $\mathcal{S}$ with $R(z) < 0$, as in Figure 6, resulting in spurious oscillatory solutions. Even if $\theta \in (\theta^-,\theta^+)$ then there can exist $z \in \mathcal{Q}(0,\theta)$ with $z < 0$ and real, similar to the case of Figure 8 and undesirable for the same reasons.

6 Algorithm

In this section we show how to incorporate the PS_θ constraint into a practical variable time-stepping algorithm, by making a few changes to a traditional time-stepping algorithm. We now outline a strategy which is similar to that using for the PS error control in [8], only differing in perhaps its most important aspect, the step-size selection. Note that choosing a new step-size when the PS_θ constraint is violated (or is close to being violated), is non-trivial as we do not have a simple error estimate to use.

It is convenient to use the following common representation of a Runge-Kutta method,

$$k_i = f(y_n + \Delta t_n \sum_{j=1}^{i-1} a_{ij} k_j), \quad 1 \leqslant i \leqslant s, \tag{6.24}$$

$$y_{n+1} = y_n + \Delta t_n \sum_{i=1}^{s} b_i k_i, \tag{6.25}$$

with

$$E(y_n, \Delta t_n) = \Delta t_n^r \| \sum_{i=1}^{s} e_i k_i \| \tag{6.26}$$

where $r = 0$ or 1 for EPUS or EPS control, respectively. This is identical to (2.1)-(2.3) and (2.6)-(2.6) noting that $k_i = f(Y_i)$ for $i = 1, \ldots, s$ and letting $e_i = b_i - \widehat{b}_i$.

Some care is needed when implementing PS_θ control in finite precision arithmetic. Although Theorem 3.4 shows that there is always an acceptable step-size, since both the right and left-hand sides of (1.2) tend to zero as $\Delta t \to 0$, in practice rounding errors could cause the rejection of what is otherwise an acceptable step-size. To avoid unnecessary cancellation, we implement (1.2) in the equivalent form

$$\|(b_1 + \theta - 1) f(y_n) - \theta f(y_{n+1}) + \sum_{i=2}^{s} b_i k_i\| \leqslant \varphi \|\theta f(y_{n+1}) + (1-\theta) f(y_n)\|. \tag{6.27}$$

The basic algorithm for solving (1.1) over $0 \leqslant t \leqslant T$ can be summarised as follows.

Algorithm 6.1 (PS_θ)

set $n = 0$, $y_0 = y(0)$, $t_0 = 0$, $k_1 = f(y_0)$ *and choose* Δt_0
while $t_n < T$
  compute k_i, $i = 2, ..., s$ *from (6.24)*
  $y_{new} = y_n + \Delta t_n \sum_{i=1}^{s} b_i k_i$
  $f_{new} = f(y_{new})$
  $E(y_{new}, \Delta t_n) = \Delta t_n^r \| \sum_{i=1}^{s} e_i k_i \|$
  $T_l = \|(b_1 + \theta - 1) k_1 - \theta f_{new} + \sum_{i=2}^{s} b_i k_i\|$
  $T_r = \|\theta f_{new} + (1-\theta) k_1\|$
  if $E(y_{new}, \Delta t_n) \leqslant \tau$ ***and*** $T_l \leqslant \varphi T_r$
    $y_{n+1} = y_{new}$
    $k_1 = f_{new}$
    $t_{n+1} = t_n + \Delta t_n$
    compute Δt_{new} *and set* $\Delta t_{n+1} = \Delta t_{new}$
    increment n *to* $n+1$
  else
    compute Δt_{new} *and set* $\Delta t_n = \Delta t_{new}$
  end
end

We now describe in detail the strategy for computing Δt_{new}. It is common to include a *maximum step-size ratio*, $\alpha > 1$, in a code. A typical choice is $\alpha = 5$. Consecutive step-sizes must satisfy $\Delta t_{n+1} \leqslant \alpha \Delta t_n$; this restricts the relative increase of the step-size over each step. It is also common to impose a maximum step-size, $\Delta t_{\max}$, so that $\Delta t_n \leqslant \Delta t_{\max}$ for all n. Thus, using the standard formula (2.8) we calculate

$$\Delta t_{\text{est}} = \gamma \left(\frac{\tau}{E(y_n, \Delta t_n)} \right)^{1/q} \Delta t_n, \tag{6.28}$$

and set

$$\Delta t_{\text{new}} = \min\{\Delta t_{\text{est}}, \alpha \Delta t_n, \Delta t_{\max}, T - t_n\}. \tag{6.29}$$

In our new step-size selection strategy, we allow α to change on each step in order to take account of the extra constraint (6.27). Recall that our overall aim is to depart from the step-size that would be predicted by the local error based formula only when the phase space error is significant. Hence, letting $r := T_l/T_r$, we set $\alpha = \alpha_1$ if $r < \beta_{\min}$, where α_1 is the maximum step-size ratio used by the traditional strategy and $\beta_{\min}$ is a small parameter, such as $\varphi/10$. In this way, we expect the new strategy to be invisible away from fixed points.

If the constraint $r \leqslant \varphi$ is violated then we force the step-size to be halved; that is, we set $\alpha = 1/2$.

To avoid excessive rejections it is important to have a safety factor. Hence, we introduce a parameter $\beta_{\max}$ (say, $\beta_{\max} = \varphi/2$) such that we set $\alpha < 1$ if $r > \beta_{\max}$, so that we force a step-size reduction at the next step if we are close to rejecting. We set $\alpha = 1$ if $r = \beta_{\max}$ and if $r < \beta_{\max}$ we set $\alpha > 1$ so that step-size increases are only allowed if $r < \beta_{\max}$.

Specifically we set

$$\alpha(r) := \begin{cases} \alpha_1 & \text{when } r \leqslant \beta_{\min}, \\ \alpha_2(r) := p_1 r^2 + p_2 r + p_3 & \text{when } \beta_{\min} \leqslant r \leqslant \beta_{\max}, \\ \alpha_3(r) := q_1 r^2 + q_2 r + q_3 & \text{when } \beta_{\max} \leqslant r \leqslant \varphi, \\ 1/2 & \text{when } r \geqslant \varphi, \end{cases} \tag{6.30}$$

where the coefficients of the functions $\alpha_2(r)$ and $\alpha_3(r)$ are chosen so that

$$\begin{array}{lll} \alpha_2(\beta_{min}) = \alpha_1, & & \alpha_3(\beta_{max}) = 1, \\ \alpha_2(\beta_{max}) = 1, & \text{and} & \alpha_3(\varphi) = \frac{1}{2}, \\ \alpha_2'(\beta_{max}) = -\frac{1}{\beta_{max}\kappa}, & & \alpha_3'(\beta_{max}) = -\frac{1}{\beta_{max}\kappa}, \end{array} \tag{6.31}$$

where the parameter κ is a strictly positive integer, dependent on the method is used, defined later in Table 2. With this dynamic choice of α, equation (6.29) defines the new stepsize selection process.

Thus for r between $\beta_{\min}$ and $\beta_{\max}$, α is a quadratic function which decreases from α_1 to 1, and for r between $\beta_{\max}$ and φ, α is a quadratic function which decreases from 1 to 1/2. We will show below that α is strictly monotonic decreasing for $r \in [\beta_{\min}, \varphi]$ if $\beta_{\max} \geqslant \varphi/(1+\kappa)$.

In [8] simple linear functions were used for $\alpha_2(r)$ and $\alpha_3(r)$. However this does not work well for orbits on (or near) the stable manifold of a fixed point y^*, as in this case the solution will tend to the fixed point, and ideally we would expect the step-size to also tend to a limit. Near to the fixed point y^* the step-size will be fixed if $r = \beta_{\max}$ so that $\alpha(r) = 1$, but we will see below that stability of this fixed point of the step-size will depend on the value of $\alpha'(\beta_{\max})$. For the linear functions used in [8] $\alpha'(r)$ is discontinuous at $r = \beta_{\max}$. In contrast with the quadratic functions defined by (6.30),(6.31) continuity of $\alpha'(r)$ is ensured and $\alpha'(\beta_{\max})$ can be set to a suitable value.

One final point must be made about Algorithm 6.1. In the case where the numerical solution is driven to a fixed point, both T_l and T_r tend to zero. Hence, to avoid division by zero errors, let δ be the machine precision and compute r as follows.

if $T_r > \delta$
  $r := T_l/T_r$
else
  if $T_l \leqslant \delta$
    $r := \beta_{\max}$
  else
    $r := \varphi$
  end
end

Thus a decrease in the step-size is forced if T_r is small but T_l is not, and the same step-size is kept if both T_r and T_l are small.

6.1 Step-size Selection

In this section we give more details and analysis of the step-size selection function $\alpha(r)$ (6.30). It is useful to define parameters $0 < \psi < \chi < 1$ such that

$$\beta_{min} = \psi\varphi, \tag{6.32}$$
$$\beta_{max} = \chi\varphi. \tag{6.33}$$

This together with (6.31) implies

$$p_1 = \frac{\chi\kappa(\alpha_1 - 1) + \psi - \chi}{\chi\varphi^2\kappa(\chi - \psi)^2}, \qquad q_1 = \frac{2 - \chi(2+\kappa)}{2\chi\kappa\varphi^2(1-\chi)^2},$$
$$p_2 = \frac{(\chi^2 - \psi^2) - 2\chi^2\kappa(\alpha_1 - 1)}{\chi\kappa\varphi(\chi - \psi)^2}, \quad \text{and} \quad q_2 = \frac{\chi^2\kappa - (1-\chi^2)}{\chi\varphi\kappa(1-\chi)^2}, \tag{6.34}$$
$$p_3 = \frac{\kappa+1}{\kappa} + \frac{\chi^2\kappa(\alpha_1 - 1) - \chi(\chi - \psi)}{\kappa(\chi - \psi)^2}, \qquad q_3 = \frac{\kappa+1}{\kappa} + \frac{2\chi - \chi^2(2+\kappa)}{2\kappa(1-\chi)^2}.$$

In Figure 10 examples of $\alpha(r)$ for $\kappa = 1$, 2 are plotted. Note that in both cases $\alpha(r)$ appears to be monotonically decreasing for $r \in [\beta_{\min}, \varphi]$. This is a highly desirable property, as the larger r is the closer the PS_θ control (1.2) is to being violated, and hence the less we would want to increase the step-size.

A little algebra shows that $\alpha(r)$ is strictly monotonic decreasing for $r \in [\beta_{\min}, \beta_{\max}] = [\psi\varphi, \chi\varphi]$ provided $\alpha_1 > 1 + (\chi - \psi)/(2\chi\kappa)$ which is satisfied for any $\kappa \geqslant 1$ and any $0 < \psi < \chi < 1$ if $\alpha_1 \geqslant 3/2$. Since α_1 is usually taken to be 5 this presents no problem. However for $\alpha(r)$ to be strictly monotonic decreasing for $r \in [\beta_{\max}, \varphi] = [\chi\varphi, \varphi]$ we require $\chi \geqslant 1/(1+\kappa)$. Thus to ensure monotonicity we must not choose χ too small, but the choice of $\chi = 1/2$ is suitable for all $\kappa \geqslant 1$.

If the solution of (1.1) is tending to a fixed point, then by Theorem 5.2 we expect the numerical solution to also be driven to the fixed point. In a "good" time-stepping strategy the step-size Δt_n should also tend to a constant value, and we will now show how the algorithm given above is set up to achieve this.

Close to a hyperbolic fixed point "most" solutions on the stable manifold will approach the fixed point in the direction of the eigenvector of the Jacobian matrix corresponding to the eigenvalue with negative real part smallest in modulus. This is modelled by the linear problem (5.1), and we again consider PS_θ error control applied to this problem. Note that for the problem (5.1) the PS_θ constraint (1.2) becomes (5.4), which depends only on $z_n = \lambda\Delta t_n$

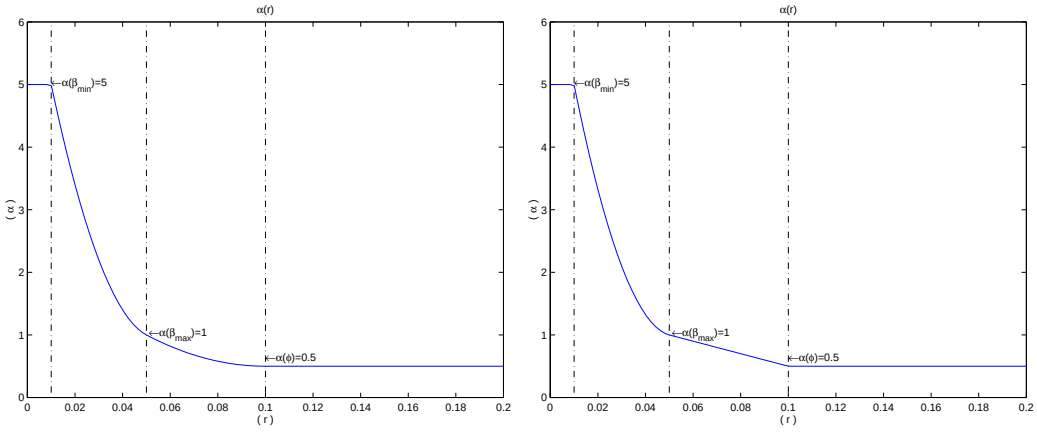

Figure 10: Graph of the function $\alpha(r)$, given by (6.30) for $\kappa = 1$, 2, with $\psi = 0.1$, $\chi = 0.5$, and $\varphi = 0.1$.

and is independent of the spatial position y_n. Thus for this problem we can consider the sequence $\{z_n\}_{n\geqslant 0}$ to determine whether it tends to a constant value, independently of the spatial behaviour. Since the standard time-stepping strategy drives the step-size Δt_n to the linear stability limit in the neighbourhood of a fixed point and the numerical solution remains bounded away from the fixed point (see Hall [5]), for the numerical solution to be driven to the fixed point by our algorithm we require that in a neighbourhood of the fixed point at every step the step-sizes are chosen in (6.29) according to the dynamic maximum step-size ratio $\Delta t_{\text{new}} = \alpha \Delta t_n$ where α defined by (6.30). By Theorem 5.2 this will ensure that the solution is driven to the fixed point for suitable choice of φ and θ.

Thus in the neighbourhood of the fixed point the evolution of the step-size, will be determined by

$$z_{n+1} = \alpha(r(z_n))z_n := \mathcal{F}(z_n), \tag{6.35}$$

where $z_n = \lambda \Delta t_n$, provided $r(z_n) < \varphi$ where the ratio r is given by

$$r(z) = \frac{|R(z) - 1 - z[\theta R(z) + (1-\theta)]|}{|z[\theta R(z) + (1-\theta)]|}. \tag{6.36}$$

By (6.30),(6.31),(6.33) this iteration has a fixed point $z^* = \mathcal{F}(z^*)$ at z^* such that $r(z^*) = \chi\varphi = \beta_{\max}$ and hence $\alpha(r(z^*)) = \alpha(\chi\varphi) = \alpha(\beta_{\max}) = 1$. For this iteration to be stable we require that $|\mathcal{F}'(z^*)| < 1$, with quadratic convergence if $\mathcal{F}'(z^*) = 0$. We now show how to achieve convergence of this iteration for small φ with quadratic convergence in the limit as $\varphi \to 0$.

Since $\mathcal{F}(z_n)$ is given by (6.35),

$$\begin{aligned} \mathcal{F}'(z^*) &= \alpha'(r(z^*))r'(z^*)z^* + \alpha(r(z^*)) \\ &= \alpha'(r(z^*))r'(z^*)z^* + 1. \end{aligned} \tag{6.37}$$

Since the stability function $R(z)$ is given by equation (5.3), in (6.36) we have $R(z_n) - 1 - z_n[\theta R(z_n) + (1-\theta)] = \mathcal{O}(z^k)$, $k \geqslant 2$, and $z_n[\theta R(z_n) + (1-\theta)] = \mathcal{O}(z)$, and thus $r(z) = \mathcal{O}(z)$. We define

$$r_1(z) := z^{-1}(R(z) - 1 - z[\theta R(z) + (1-\theta)]), \tag{6.38}$$

$$r_2(z) := sign(r_1(z))|[\theta R(z) + (1-\theta)]|, \tag{6.39}$$

so that $r(z) = r_1(z)/r_2(z)$, and

$$r'(z) = \frac{r_1'(z)r_2(z) - r_2'(z)r_1(z)}{r_2(z)^2}. \tag{6.40}$$

But $r(z^*) = \chi\varphi$, hence $r_2(z^*) = r_1(z^*)/(\chi\varphi)$, and so (6.40) implies

$$r'(z^*) = \frac{\chi\varphi[r_1'(z^*) - \chi\varphi r_2'(z^*)]}{r_1(z^*)} = \chi\varphi\frac{r_1'(z^*)}{r_1(z^*)} + \mathcal{O}(\chi^2\varphi^2). \tag{6.41}$$

Now for an explicit Runge-Kutta method (5.3) and (6.38) imply that

$$r_1(z) = z^{-1}[(c_2 - \theta)z^2 + \sum_{i=3}^{s+1}(c_i - \theta c_{i-1})z^i] = (c_2 - \theta)z + \sum_{i=2}^{s}(c_{i+1} - \theta c_i)z^i.$$

Let z^κ be the first term with non-zero coefficient then

$$r_1(z) = (c_{\kappa+1} - \theta c_\kappa)z^\kappa + \mathcal{O}(z^{\kappa+1}), \tag{6.42}$$

where $1 \leqslant \kappa \leqslant s$ and c_j's are the coefficients of $R(z)$. Then

$$r_1'(z) = \kappa(c_{\kappa+1} - \theta c_\kappa)z^{\kappa-1} + \mathcal{O}(z^\kappa). \tag{6.43}$$

Thus

$$z\frac{r_1'(z)}{r_1(z)} = z\frac{\kappa(c_{\kappa+1} - \theta c_\kappa)z^{\kappa-1} + \mathcal{O}(z^\kappa)}{(c_{\kappa+1} - \theta c_\kappa)z^\kappa + \mathcal{O}(z^{\kappa+1})} = \kappa + \mathcal{O}(z).$$

Now, $r(z) = \mathcal{O}(z)$, implies that $\mathcal{O}(z^*) = r(z^*) = \chi\varphi$, and thus by (6.41)

$$z^*r'(z^*) = \chi\varphi z^*\frac{r_1'(z^*)}{r_1(z^*)} + \mathcal{O}(z^*\chi^2\varphi^2) = \chi\varphi\kappa + \mathcal{O}(z^*\chi\varphi) = \chi\varphi\kappa + \mathcal{O}(\chi^2\varphi^2). \tag{6.44}$$

Finally (6.37) implies that

$$\mathcal{F}'(z^*) = 1 + \alpha'(\chi\varphi)[\chi\varphi\kappa + \mathcal{O}(\chi^2\varphi^2)] = \mathcal{O}(\chi\varphi)$$

provided $\alpha'(\chi\varphi) = -1/(\chi\varphi\kappa)$, or equivalently $\alpha'(\beta_{\max}) = -1/(\beta_{\max}\kappa)$ as specified in (6.31). Thus with the choice of parameters given we expect the step-size to converge to a constant value in the neighbourhood of a fixed point, with quadratic convergence in the limit as $\varphi \to 0$.

The integer parameter κ in (6.31) follows from (6.42) and depends only on θ and the parameters c_i of the stability function $R(z)$ (5.3) of the method (2.2), where the dependence is given by Table 2.

Condition	κ
$c_2 \neq \theta$	1
$c_2 = \theta$, $c_3 \neq \theta^2$	2
$c_2 = \theta$, $c_3 = \theta^2$, $c_4 \neq \theta^3$	3
$\vdots$	$\vdots$
$c_{i+1} = \theta^i$, $i = 1, \ldots, q-1$ and $c_{q+1} \neq \theta^q$	q

Table 2: Determining $\kappa \in \{1, 2, \ldots, s\}$ for equation (6.31) in terms of the coefficients of the stability function.

Note from Table 2 that there are three main cases.

(i) Method (2.2) of order $p = 1$ and $c_2 \neq \theta$ implies $\kappa = 1$,

(ii) Method (2.2) of order $p \geqslant 2$ and $\theta \neq 1/2$ implies $\kappa = 1$,

(iii) Method (2.2) of order $p \geqslant 3$ and $\theta = 1/2$ implies $\kappa = 2$.

Note that values of $\kappa > 2$ can only arise with first and second order methods. For example $\theta = 1/2$ and $R(z) = 1 + z + z^2/2 + z^3/4$ implies that $\kappa = 3$.

We now consider two examples, where $\mathcal{F}(z^*)$ is computed without approximation to confirm the convergence of the step-size to a constant value as was suggested by the approximate analysis above.

6.2 PS_θ control for RK1(2)

Consider the RK1(2) method (2.4) in non-extrapolation mode, applied to the scalar linear problem (5.1) with $\lambda < 0$. This method has stability function

$$R(z_n) = 1 + z_n, \quad z_n = \lambda \Delta t_n,$$

and from Table 2 we have $\kappa = 1$. Now (6.36) becomes

$$r(z) = \frac{|\theta z|}{|1 + \theta z|}, \tag{6.45}$$

and since the fixed point $z^* < 0$ of the iteration (6.35) satisfies $r(z^*) = \chi\varphi$ we have

$$|\theta z^*| = \chi\varphi |1 + \theta z^*|. \tag{6.46}$$

For any $\theta \neq 0$ there are two cases to consider depending on whether $\theta z^* = \pm\chi\varphi(1 + \theta z^*)$.

(i) If $z^* < -1/\theta$ then (6.46) implies $-\theta z^* = -\chi\varphi(1 + \theta z^*)$ and hence since $\chi\varphi \in (0, 1)$, we have $0 > z^* = \chi\varphi/\theta(1 - \chi\varphi) > 0$, a contradiction.

(ii) Thus $-1/\theta < z^* < 0$ and (6.46) implies $-\theta z^* = \chi\varphi(1 + \theta z^*)$ and hence

$$z^* = -\frac{\chi\varphi}{\theta(1 + \chi\varphi)}. \tag{6.47}$$

Thus since $\chi\varphi \in (0, 1)$ we have $z^* \in (-1/(2\theta), 0)$.

Now for $z^* \in (-1/(2\theta), 0)$ equation (6.45) implies that

$$r(z^*) = \frac{|\theta z^*|}{|1 + \theta z^*|} = \frac{-\theta z^*}{1 + \theta z^*}.$$

Hence using (6.47),

$$z^* r'(z^*) = -\frac{\theta z^*}{(1 + \theta z^*)^2} = \chi\varphi(1 + \chi\varphi).$$

Thus since $\alpha'(\chi\varphi) = -1/(\chi\varphi)$ equation (6.37) implies

$$\mathcal{F}'(z^*) = 1 - \frac{1}{\chi\varphi}\chi\varphi(1 + \chi\varphi) = -\chi\varphi.$$

So if $\varphi = 0$ then $\mathcal{F}'(z^*) = 0$ and in limit of small φ we obtain quadratic convergence of the step-size to z^* in (6.35). Moreover $-1 < -\chi < \mathcal{F}'(z^*) < 0$ for all χ, $\varphi \in (0, 1)$ and so z_n converges to z^* for any $\chi \in (0, 1)$ and any $\varphi \in (0, 1)$.

6.3 $\mathbf{PS}_\theta$ control for RK2(3)

Consider the RK2(3) method (2.5) in non-extrapolation mode, applied to the scalar linear problem (5.1) with $\lambda < 0$. This method has stability function

$$R(z) = 1 + z + \frac{z^2}{2}, \quad z = \lambda \Delta t_n,$$

and from Table 2 we have $\kappa = 2$ for $\theta = 1/2$ and $\kappa = 1$ otherwise. Consider first the case where $\theta = 1/2$ recommended in Section 5.2. Then (6.36) becomes

$$r(z) = \frac{|-z^3/2|}{|z^3/2 + z^2 + 2z|} = \frac{z^2}{z^2 + 2z + 4} \tag{6.48}$$

and since the fixed point $z^* < 0$ of the iteration (6.35) satisfies $r(z^*) = \chi\varphi$ we have

$$\chi\varphi(z^{*2} + 2z^* + 4) = z^{*2},$$

and hence

$$z^* = \frac{\chi\varphi - \sqrt{4\chi\varphi - 3\chi^2\varphi^2}}{1 - \chi\varphi}.$$

It follows from (6.48) that

$$r'(z) = \frac{2z(z+4)}{(z^2 + 2z + 4)^2},$$

Thus since $\alpha'(\chi\varphi) = -1/(2\chi\varphi)$ equation (6.37) implies

$$\mathcal{F}'(z^*) = 1 - \frac{1}{2\chi\varphi} \frac{2z^{*2}(z^* + 4)}{(z^{*2} + 2z^* + 4)^2} = \frac{\chi\varphi(z^* + 1)}{z^*}.$$

The graph of $\mathcal{F}'(z^*)$ against $\chi\varphi$ for $\theta = 1/2$ is given in Figure 11.

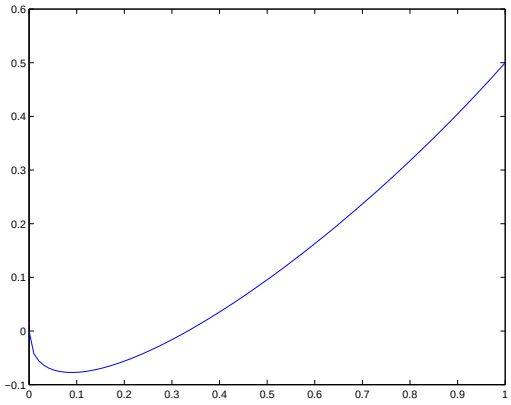

Figure 11: Graph of $\mathcal{F}'(z^*)$ against $\chi\varphi$ for the RK2(3) method with $\theta = 1/2$.

Thus $|\mathcal{F}'(z^*)| < 1$ for all $0 < \chi\varphi < 1$, with quadratic convergence in the limit as $\varphi \to 0$.

If $\theta \neq 1/2$ which implies $\kappa = 1$ the situation is not so simple. For $\theta < 1/2$ a similar argument to that above gives a unique negative value of z^* and then $\mathcal{F}'(z^*)$ can be computed as a function of both θ and $\chi\varphi$, the result of which is given in Figure 12(i). We see that convergence will occur for any $\theta < 1/2$ and $0 < \chi\varphi < 1$, except for $(\theta, \chi\varphi)$ close to $(0.5, 0)$, with $\mathcal{F}'(z^*)$ very close to zero for θ and $\chi\varphi$ both small, indicating rapid convergence of the step-size in this case.

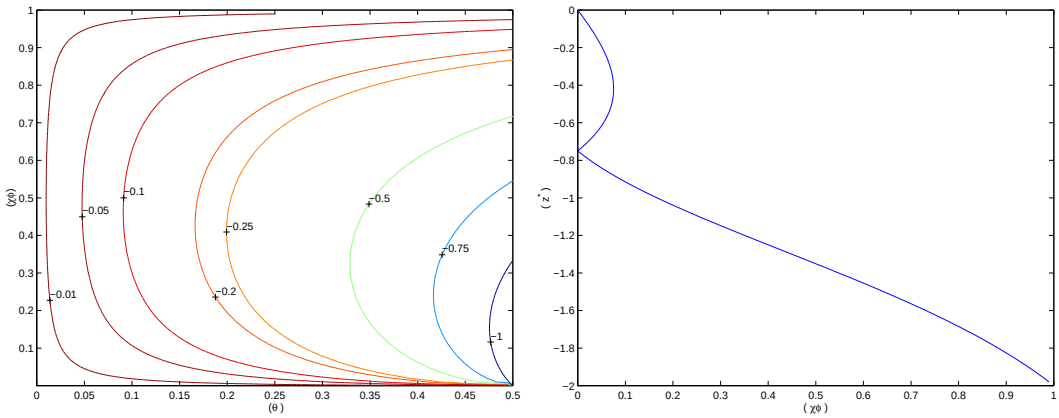

Figure 12: (i) Contours of $\mathcal{F}'(z^*)$ for RK2(3) with $0 < \theta \leqslant 1/2$ and $\chi\varphi \in [0, 1]$. (ii) Fixed point solutions $z^* < 0$ of (6.35) against $\chi\varphi$ for RK2(3) with $\theta = 0.8$.

For $\theta \in (1/2, 1]$ if

$$\chi\varphi < \frac{\theta(2\theta - 3) + \sqrt{2\theta}}{2\theta(2 - \theta)} \tag{6.49}$$

there exist multiple fixed points $z^* < 0$ of the step-size. This can be seen from examining the contours in Figure 8(ii). In Figure 12(ii) we plot these fixed points against $\chi\varphi$ for $\theta = 0.8$. We wish to force the algorithm to converge to the step-size z^* which tends to 0 as $\chi\varphi \to 0$, so that we can attain arbitrary accuracy by decreasing φ sufficiently. However this fixed point only exists for $\chi\varphi$ sufficiently small, which imposes an upper bound on $\chi\varphi$ which is very restrictive for θ close to $1/2$, so we do not advocate implementing the method like this. Nevertheless it can be shown that when (6.49) is satisfied, the step-size selection scheme in Section 6.1 ensures that the required fixed point z^* is stable, whilst the other fixed points are not. See [2] for more details.

6.4 Choice of Parameters

To summarize, the parameter $\varphi \in (0, 1)$, is user defined, and acts akin to a tolerance with smaller values giving more accurate solutions. We suggest

$$\begin{aligned} \psi &= 0.1, \quad \beta_{\min} = \psi\varphi, \\ \chi &= 0.5, \quad \beta_{\max} = \chi\varphi, \end{aligned}$$

although from the analysis above other values are also possible.

To complete the implementation, method dependent values of θ and κ are required. We suggest $\theta = \theta^+$ (though $\theta = 1$ might also be useful) leading to the values in Table 3 for well-known methods.

7 Illustration of Numerical Tests

In this section we illustrate the performance of the PS_θ method on a number of numerical test problems. Extensive testing has been done with the RK1(2), RK2(3), DOPRI5(4), DOPRI8(7) and Fehlberg4(5) methods and conclusions shown here have been found to be valid in general.

We compare the standard adaptive algorithm, as described in Section 2, with the same algorithm augmented by PS_θ control as described in Section 6. In all the examples presented EPUS control is used with parameter values $\varphi = 0.1$, $\psi = 0.1$, $\chi = 0.5$, $\alpha_1 = 5$, $\delta = 10^{-15}$.

Method	θ	κ
RK 1(2)	0.5	1
RK 2(1)	0.5	2
RK 2(3)	0.5	2
RK 2(3)B	0.6873	1
Fehlberg 4(5)	0.7569	1
Fehlberg 5(4)	0.7880	1
Dormand-Price 5(4)	0.5	2
Dormand-Price 8(7)	0.8643	1

Table 3: Suggested values of κ and θ for common methods.

Similar results are obtained for EPS control. The values of κ and θ for each method are as given in Table 3. The two-norm is used in all examples. Local error tolerances of $\tau = 10^{-2}$ or $\tau = 10^{-3}$ were used. These are larger than would be used in practice, but we emphasise that poor dynamic behaviour of the standard adaptive algorithm persists for arbitrary small tolerances.

First consider the DOPRI8(7) method applied to the scalar linear test problem (5.1) with $\lambda = -10$, for $t \in [0, 30]$ with $\tau = 10^{-2}$. Figure 13(i) shows the numerical solution for the standard adaptive algorithm and for the PS_θ method. The standard adaptive solution remains at $\mathcal{O}(\tau)$ from the fixed point, whilst the PS_θ solution is driven to the fixed point. Figure 13(ii) shows the step-sizes used by both methods. The standard adaptive algorithm drives the step-size to the linear stability limit of the method, whilst for the PS_θ method, the step-size tends to a constant value below the linear stability limit. Similar behaviour is seen with other methods with the PS_θ method driving the solution to the fixed point and the step-size to a constant value below the linear stability limit in each case, whilst the standard adaptive algorithm has step-sizes which tend to or oscillate about the linear stability limit and numerical solutions tending to spurious fixed points (if $R(z) = 1$ at the linear stability limit of the method) or period two solutions (if $R(z) = -1$).

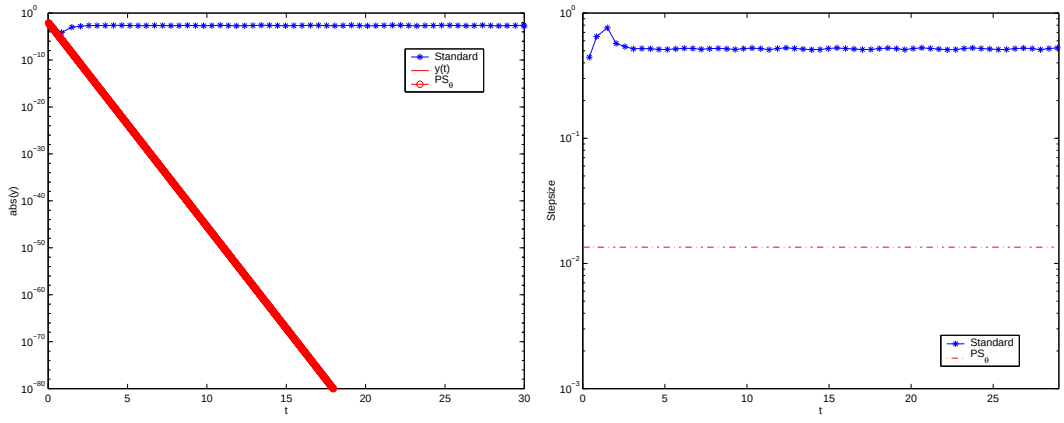

Figure 13: DOPRI8(7) around a scalar stable fixed point $\dot{y} = -10y$, $y_0 = 10^{-2}$, $\Delta t_0 = 0.4$ and $\tau = 1e-2$. (i) Solutions using standard and PS_θ methods. (ii) Step-sizes for each method

Now consider the RK1(2) and DOPRI5(4) methods applied to (1.1) with f defined by (1.3). Using EPUS with $\tau = 10^{-3}$ Figures 1 and 2 in the introduction show that behaviour similar to the previous example results, with the standard algorithm resulting in spurious behaviour

whilst the PS_θ method drives the solution to the fixed point, with a smooth step-size sequence.

Equation (1.1) with f defined by (1.4) illustrates a saddle point. For this example we took $\tau = 10^{-2}$ with the other method parameters for both the standard and PS_θ methods as above. Figure 3 shows that the standard adaptive algorithm either results in oscillations about the unstable manifold (if $R(z) = -1$ at the linear stability limit) or solutions which do not oscillate but fail to pass close to the local unstable manifold (if $R(z) = 1$). In contrast the PS_θ method gives a solution, Figure 4, which closely follows the exact solution.

In Figure 14 we illustrate the improvement in the new step-size selection mechanism of the PS_θ method over that of the PS method of [8], by applying the RK2(3) method to the previous problem. The PS_θ method uses different constant step-sizes near to the local stable and unstable manifolds with a smooth transition between them. In contrast the step-size sequence for the original PS method is unstable near to the stable manifold and transition between the stable and unstable manifold.

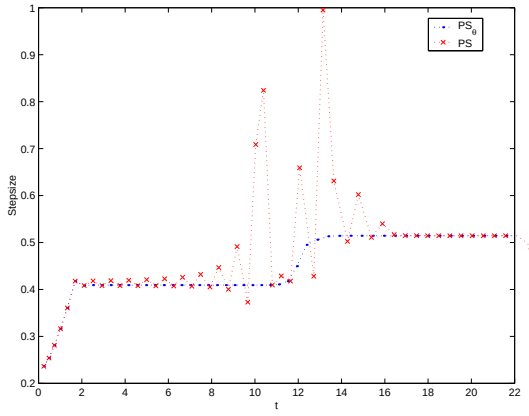

Figure 14: Comparing with step-size sequences of the RK2(3) method with PS and PS_θ control applied to (1.1) with f defined by (1.4).

Next we illustrate the behaviour around a stable fixed point with non-real eigenvalues. We apply the DOPRI8(7) method with EPUS control to (1.1) with

$$f(y) = \begin{bmatrix} -7.947 & 4.668 & 3.0229 & -0.345 \\ -1.278 & -5.527 & -3.639 & -3.533 \\ -0.832 & -2.305 & -4.526 & -4.049 \\ -5.359 & -0.502 & 0.153 & -6 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}. \tag{7.50}$$

$f(y)$ has eigenvalues $-10 \pm 5i$ and $-2 \pm i$. We took $y(0) = [1, 1, 1, 1]^T$, $\Delta t_0 = 0.5$, and $\tau = 1e-3$. Figure 15 gives the solution norm and step-sizes. Just as for the problems with real eigenvalues we see that PS_θ control has the effect of driving the solution towards equilibrium.

Finally we note that use of the 2-norm is not arbitrary. If the ∞-norm is used step-size instabilities arise in problems with saddle points, where there is a transition between step-sizes near the local stable and unstable manifolds, and in problems with complex eigenvalues.

References

[1] M.A. Aves, D.F. Griffiths, and D.J. Higham. Does error control suppress spuriosity? *SIAM J. Num. Anal.*, 34:756–778, 1997.

[2] N. Christodoulou, 2001. University of Sussex, DPhil Thesis, In preparation.

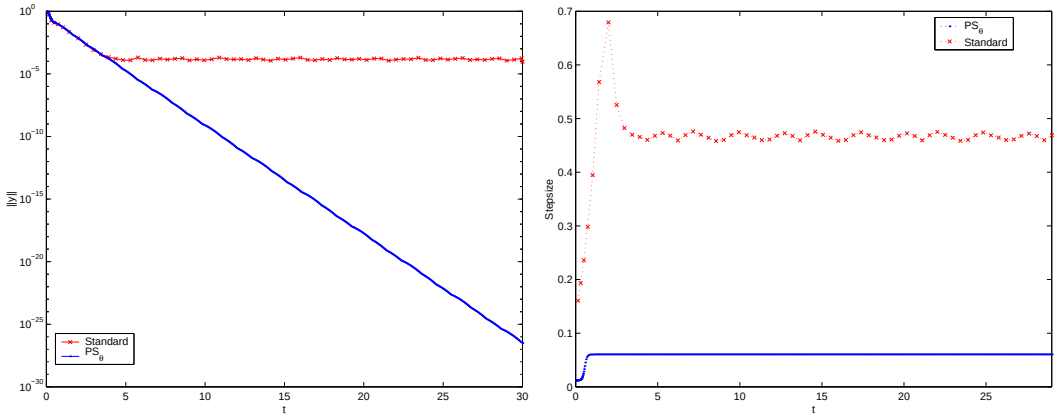

Figure 15: DOPRI8(7) around a stable fixed point with non-real eigenvalues.

[3] D.F. Griffiths. The dynamics of some linear mutlistep methods. In D.F. Griffiths and G.A. Watson, editors, *Proceedings of the 1987 Dundee Conference on Numerical Analysis*, pages 115–134. Pitman Research Notes in Mathematics, 1988.

[4] E. Hairer, S.P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I - Nonstiff Problems*, volume I. Springer-Verlag, Berlin Heidelberg, 1987. First Edition.

[5] G. Hall. Equilibrium states of Runge-Kutta scemes I. *ACM Trans on Math. Software*, 11:289–301, 1985.

[6] G. Hall. Equilibrium states of Runge-Kutta scemes II. *ACM Trans on Math. Software*, 12:183–192, 1986.

[7] D. J. Higham and A. M. Stuart. Analysis of the dynamics of local error control via a piecewise continuous residual. *BIT*, 38:44–57, 1998.

[8] D.J. Higham, A R. Humphries, and R.J. Wain. Phase space error control for dynamical systems. *SIAM J. Sci. Comp.*, 21:2275–2294, 2000.

[9] A.R. Humphries. Spurious solutions of numerical methods for initial value problems. *IMA J. Num. Anal.*, 13:263–290, 1993.

[10] The Math Works, Inc. *MATLAB User's Guide*. Natick, Massachusetts, 1992.

[11] L.F. Shampine. *Numerical Solution of Ordinary Differential Equations*. Chapman and Hall, 1994.

[12] A.M. Stuart and A.R. Humphries. The essential stability of local error control for dynamical systems. *SIAM J. Num. Anal.*, 32:1940–1971, 1995.

[13] A.M. Stuart and A.R. Humphries. *Dynamical Systems and Numerical Analysis*. Cambridge University Press, Cambridge, 1996.