ELSEVIER

Applied Numerical Mathematics 25 (1997) 229–241

APPLIED NUMERICAL MATHEMATICS

Regular Runge–Kutta pairs

Desmond J. Higham [1]

Department of Mathematics, University of Strathclyde, Livingstone Tower, 26 Richmond Street, Glasgow G1 1XH, Scotland, UK

Abstract

Time-stepping methods that guarantee to avoid spurious fixed points are said to be *regular*. For fixed stepsize Runge–Kutta formulas, this concept has been well studied. Here, the theory of regularity is extended to the case of embedded Runge–Kutta pairs used in variable stepsize mode with local error control. First, the limiting case of a zero error tolerance is considered. A recursive regularity test, based on the folding technique of Hairer, Iserles and Sanz-Serna (1990), is developed. It is then shown how regularity at zero tolerance carries through to the case of small tolerances. Finally, the property of regularity for all tolerances is characterized. © 1997 Published by Elsevier Science B.V.

Keywords: Error control; Spurious fixed point; Variable time-stepping

1. Background

An extreme case of a numerical method breaking down arises when a Runge–Kutta (RK) formula computes a spurious steady state; that is, a fixed point which is unrelated to the underlying differential equation. Iserles [6] defined *regular* RK formulas to be those that guarantee to avoid this difficulty whenever the vector field is continuous. A recursive test for regularity was given by Hairer et al. [3], where it was shown, for example, that the only formula that is both explicit and regular is (ignoring redundancies) the forward Euler method. Humphries [5] gave some fundamental results about the existence of spurious solutions for small stepsizes. More recently, Jackiewicz et al. introduced the related concept of *strong regularity* and applied it to RK formulas for both ordinary and delay differential equations [7,8].

The work mentioned above concerns RK methods in fixed stepsize mode. In practice, most high-quality RK software uses embedded RK formula pairs with a variable time-stepping strategy based on local error control. Aves et al. [1] looked at the potential for spurious fixed points in the context of local error control. They showed that for most of the widely-used explicit RK pairs spurious fixed points

[1] The work of this author was supported by the Engineering and Physical Sciences Research Council of the UK under grants GR/H94634 and GR/K80228. E-mail: na.dhigham@na-net.ornl.gov.

SSDI 0168-9274(97)00062-7

are admitted, but are generically unstable, with the instability growing like the inverse of the local error tolerance. However, they also showed that in these cases it is possible to construct differential equations where stable, spurious fixed points exist for arbitrarily small tolerances.

In this work, we perform a systematic study of regularity for RK pairs. Our approach combines ideas from [3], where the concept of a *folding* was introduced, and [1], where the importance of the limiting case of a zero tolerance was identified. Work is currently underway to find characterizations and order barriers for RK pairs with the properties analyzed here.

2. Runge–Kutta pairs

Consider the initial value ordinary differential equation system

$$y'(t) = f\bigl(y(t)\bigr), \qquad y(0) = y_0 \in \mathbb{R}^m. \tag{1}$$

An s-stage embedded RK pair for solving (1) numerically is determined by the coefficients $\{a_{ij}\}_{i,j=1}^{s}$ and $\{c_i, b_i, \widehat{b}_i\}_{i=1}^{s}$, where the weights $\{b_i\}_{i=1}^{s}$ belong to the main formula. Given $y_n \approx y(t_n)$ and a stepsize $h_n > 0$, we compute $y_{n+1} \approx y(t_{n+1})$, with $t_{n+1} = t_n + h_n$, according to

$$\xi_i = f\left(y_n + h_n \sum_{j=1}^{s} a_{ij}\xi_j\right), \tag{2}$$

$$y_{n+1} = y_n + h_n \sum_{i=1}^{s} b_i\xi_i. \tag{3}$$

It is convenient to use

$$z_i := y_n + h_n \sum_{j=1}^{s} a_{ij}\xi_j$$

to denote the arguments at which f is evaluated. The secondary formula, whose weights are $\{\widehat{b}_i\}_{i=1}^{s}$, produces

$$\widehat{y}_{n+1} = y_n + h_n \sum_{i=1}^{s} \widehat{b}_i\xi_i. \tag{4}$$

In a typical time-stepping algorithm, the difference between the main and secondary solution is used to approximate the local error over the step, and is compared with the user-supplied tolerance parameter τ. The *error criterion* is

$$\|\widehat{y}_{n+1} - y_{n+1}\| \leqslant \tau \quad \text{or} \quad \|\widehat{y}_{n+1} - y_{n+1}\| \leqslant \tau h_n, \tag{5}$$

for error-per-step (EPS) or for error-per-unit-step (EPUS) control, respectively. If the error criterion is met, then the step is accepted. Otherwise, the step is rejected and re-taken with a smaller stepsize. After a successful step, the usual formula for the next stepsize is

$$h_{n+1} = h_n\left(\frac{\theta\tau}{\|\widehat{y}_{n+1} - y_{n+1}\|}\right)^{1/q} \quad \text{or} \quad h_{n+1} = h_n\left(\frac{h_n\theta\tau}{\|\widehat{y}_{n+1} - y_{n+1}\|}\right)^{1/(q-1)}, \tag{6}$$

for EPS and EPUS control, respectively. Here q is the largest integer such that $\|\widehat{y}_{n+1} - y_{n+1}\| = \mathrm{O}(h_n^q)$ as $h_n \to 0$, and $\theta \in (0,1)$ is a constant safety factor. Typically, other restrictions such as $0.5 \leqslant h_n/h_{n+1} \leqslant 5$ are imposed, but these are not relevant for our analysis.

Throughout this work, we assume that the RK coefficients satisfy the standard conditions $\sum_{j=1}^{s} a_{ij} = c_i$, for $i = 1, 2, \ldots, s$ and the consistency conditions $\sum_{i=1}^{s} b_i = \sum_{i=1}^{s} \widehat{b}_i = 1$. We also assume that $\boldsymbol{b} \neq \widehat{\boldsymbol{b}}$, so that the pair consists of distinct formulas (although this property is not necessarily shared by a folding of the pair—see Section 3). We let $\boldsymbol{a}_i^{\mathrm{T}}$ denote the ith row of $A = (a_{ij})$, and we suppose that $\boldsymbol{a}_k - \boldsymbol{a}_l \neq 0$ when $k \neq l$, so that no stage is redundant.

In the context of fixed stepsizes, the following definition of a spurious fixed point is used.

Definition 2.1. For a single RK formula $\{A, \boldsymbol{b}, \boldsymbol{c}\}$ with a given, fixed stepsize ($h_n \equiv h > 0$), by a *spurious fixed point* we mean a point y with $f(y) \neq 0$ such that $y_n \equiv y$ is a constant solution of the map defined by (2) and (3).

For variable time-stepping algorithms, both y_n and h_n may vary with n. To make it clear whether we are discussing fixed points with respect to y_n or to (h_n, y_n), we make the following definitions.

Definition 2.2. For a RK pair $\{A, \boldsymbol{b}, \widehat{\boldsymbol{b}}, \boldsymbol{c}\}$ in variable stepsize mode with a given tolerance τ, by a *global spurious fixed point* we mean a pair (h, y) with $f(y) \neq 0$ such that $y_n \equiv y$, $h_n \equiv h > 0$ is a constant solution of the map defined by (2)–(4) and (6). (Note that in this case the error criterion (5) is automatically satisfied.)

Definition 2.3. For a RK pair $\{A, \boldsymbol{b}, \widehat{\boldsymbol{b}}, \boldsymbol{c}\}$ in variable stepsize mode with a given tolerance τ, by a *local spurious fixed point* we mean a pair (h, y) with $f(y) \neq 0$ such that $y_{n+1} = y_n = y$, $h_n = h > 0$ is a solution to (2)–(3) and the error criterion (5) is satisfied.

We emphasize that the definition of a *local* spurious fixed point refers to behaviour over a single step, and if a method avoids local spurious fixed points then it will automatically avoid global spurious fixed points.

3. Regularity at zero tolerance

Our approach in this work is to assume that the error tolerance τ, rather than the stepsize, is small. It is clear from (2)–(5) that a global or local spurious fixed point of the algorithm must have $\sum_{i=1}^{s} b_i \xi_i = 0$ and $\sum_{i=1}^{s} \widehat{b}_i \xi_i = \mathrm{O}(\tau)$. Hence, for small tolerances, as mentioned in [1], a fixed point of the algorithm must be a fixed point of the main formula and within $\mathrm{O}(\tau)$ of being a fixed point of the secondary formula. In this section, therefore, we concentrate on the limiting case of $\tau = 0$. Section 4 shows how these results extend to the case of small τ. We point out that Lemmas 3.1, 3.2, 3.4 and 3.5 below do not require $\boldsymbol{b}$ and $\widehat{\boldsymbol{b}}$ to be distinct. This allows us to accommodate the folding process that is introduced later in the section.

We begin with the following definition.

Definition 3.1. A RK pair $\{A, \boldsymbol{b}, \widehat{\boldsymbol{b}}, \boldsymbol{c}\}$ is *regular at* $\tau = 0$, denoted R_0, if for every problem (1) where f is continuous and every stepsize $h > 0$ the individual RK formulas $\{A, \boldsymbol{b}, \boldsymbol{c}\}$ and $\{A, \widehat{\boldsymbol{b}}, \boldsymbol{c}\}$, in fixed stepsize mode, do not simultaneously admit a spurious fixed point.

We emphasize that R_0 is a weaker demand than asking for at least one of the individual formulas to be regular. For example, consider the second and third order explicit pair of Fehlberg [2]

$$\begin{array}{|c} A \\ \hline \boldsymbol{b}^{\mathrm{T}} \\ \hline \widehat{\boldsymbol{b}}^{\mathrm{T}} \end{array} = \begin{array}{|ccc} 1 & & \\ \frac{1}{4} & \frac{1}{4} & \\ \hline \frac{1}{6} & \frac{1}{6} & \frac{4}{6} \\ \hline \frac{1}{2} & \frac{1}{2} & \end{array}.$$

The individual formulas are not regular. (This follows from [3, Corollary 4], where it is proved that any explicit RK formula with order greater than one cannot be regular.) However, any fixed point that is shared by the formulas must satisfy

$$\tfrac{1}{6}\xi_1 + \tfrac{1}{6}\xi_2 + \tfrac{4}{6}\xi_3 = 0 \text{ and } \tfrac{1}{2}\xi_1 + \tfrac{1}{2}\xi_2 = 0 \quad \Longrightarrow \quad \xi_3 = 0.$$

Also,

$$z_3 = y + \tfrac{1}{4}\xi_1 + \tfrac{1}{4}\xi_2 = y.$$

So, $f(y) = f(z_3) = \xi_3 = 0$. Hence, any simultaneous fixed point is not spurious. This shows that the pair is R_0. By contrast, the well-known fourth and fifth order pairs DOPRI(5,4) of Dormand and Prince and RKF45 of Fehlberg are not R_0—this follows as a by-product of the analysis in [1, Section 5].

In general, from (2)–(4), the system of equations that must be satisfied by a simultaneous fixed point is

$$y + h\sum_{j=1}^{s} a_{ij}\xi_j = z_i, \quad 1 \leqslant i \leqslant s, \tag{7}$$

$$\sum_{j=1}^{s} b_j\xi_j = 0, \tag{8}$$

$$\sum_{j=1}^{s} \widehat{b}_j\xi_j = 0. \tag{9}$$

Note that the values ξ_i must correspond to $f(z_i)$ for some continuous function f; hence we cannot have $\xi_k \neq \xi_l$ when $z_k = z_l$. This motivates the following definition.

Definition 3.2. A solution $(h, y, \boldsymbol{z}, \boldsymbol{\xi})$ to (7)–(9) is *valid* if $h > 0$ and $z_k = z_l \Rightarrow \xi_k = \xi_l$.

We may now state a simple characterization of regularity at $\tau = 0$.

Lemma 3.1. *A RK pair is* R_0 *if and only if for every valid solution to* (7)–(9) *there exists an index* ν *such that* $y = z_\nu$ *and* $\xi_\nu = 0$.

Proof. The proof is entirely analogous to that of [3, Lemma 1], but is included here for completeness.

"*Only if*". First, suppose the pair is R_0. Then, given any valid solution it is possible to construct a function (for example, by componentwise polynomial interpolation) such that $f(z_i) = \xi_i$. If $y = z_\nu$ but $\xi_\nu \neq 0$ then we have $f(y) \neq 0$. Alternatively, if $y \neq z_i$ for all i then we can add the extra interpolation condition $f(y) = 1$. Hence, in order for the solution to be non-spurious we must have $y = z_\nu$ and $\xi_\nu = 0$, for some ν.

"*If*". This follows immediately, since any fixed point is forced to satisfy $f(y) = f(z_\nu) = \xi_\nu = 0$. □

It is clear from the proof of Lemma 3.1 that in order to investigate whether a pair is R_0, it is necessary and sufficient to deal with scalar problems ($m = 1$). So henceforth, for convenience, we will assume that $m = 1$.

We now make a definition that will play a similar role to that of essentially one step (EOS) in [3].

Definition 3.3. A RK pair is *observably regular at* $\tau = 0$, denoted OR_0, if there exists an index $1 \leqslant p \leqslant s$ such that $\boldsymbol{e}_p \in \operatorname{span}\{\boldsymbol{b}, \widehat{\boldsymbol{b}}\}$ and $\boldsymbol{a}_p \in \operatorname{span}\{\boldsymbol{b}, \widehat{\boldsymbol{b}}\}$. (Here $\boldsymbol{e}_p$ denotes the pth column of the identity matrix.)

There is an immediate result.

Lemma 3.2. *A RK pair that is* OR_0 *is also* R_0.

Proof. Suppose the pair is OR_0. Given any fixed point, it follows from (8) and (9) that $\boldsymbol{e}_p^{\mathrm{T}}\boldsymbol{\xi} = 0$ and $\boldsymbol{a}_p^{\mathrm{T}}\boldsymbol{\xi} = 0$; that is, from (7), $y = z_p$. Hence, from Lemma 3.1, the pair is R_0. □

To proceed with the analysis, we will use the following linear algebra result.

Lemma 3.3. *For any integer $s > 2$, suppose we are given a pair of vectors $\boldsymbol{q}, \boldsymbol{r} \in \mathbb{R}^s$ and a finite set of nonzero vectors $V = \{\boldsymbol{v}^{[i]}\}_{i=1}^N$ with $\boldsymbol{v}^{[i]} \in \mathbb{R}^s$. Then if $\boldsymbol{v}^{[i]} \notin \operatorname{span}\{\boldsymbol{q}, \boldsymbol{r}\}$ for all $1 \leqslant i \leqslant N$, there exists a vector $\boldsymbol{u} \in \mathbb{R}^s$ such that*

$$\boldsymbol{u}^{\mathrm{T}}\boldsymbol{q} = \boldsymbol{u}^{\mathrm{T}}\boldsymbol{r} = 0 \quad \text{and} \quad \boldsymbol{u}^{\mathrm{T}}\boldsymbol{v}^{[i]} \neq 0, \quad 1 \leqslant i \leqslant N.$$

Proof. Suppose that $\boldsymbol{q}$ and $\boldsymbol{r}$ are linearly independent. Then there exists an orthonormal basis $\{\boldsymbol{c}^{[j]}\}_{j=1}^s$ for $\mathbb{R}^s$ such that $\operatorname{span}\{\boldsymbol{c}^{[1]}, \boldsymbol{c}^{[2]}\} = \operatorname{span}\{\boldsymbol{q}, \boldsymbol{r}\}$. Each $\boldsymbol{v}^{[i]}$ has an expansion $\boldsymbol{v}^{[i]} = \sum_{j=1}^s \beta_j^i \boldsymbol{c}^{[j]}$, where $\|[\beta_3^i, \beta_4^i, \ldots, \beta_s^i]^{\mathrm{T}}\| \neq 0$.

Now let $\boldsymbol{u}$ have an expansion of the form $\boldsymbol{u} = \sum_{j=3}^s \gamma_j \boldsymbol{c}^{[j]}$. Clearly $\boldsymbol{u}^{\mathrm{T}}\boldsymbol{q} = \boldsymbol{u}^{\mathrm{T}}\boldsymbol{r} = 0$. The remaining conditions are

$$[\gamma_3, \gamma_4, \ldots, \gamma_s] \begin{bmatrix} \beta_3^i \\ \beta_4^i \\ \vdots \\ \beta_s^i \end{bmatrix} \neq 0, \quad 1 \leqslant i \leqslant N.$$

That is, we are left with the problem of finding a vector in $\mathbb{R}^{s-2}$ that is not orthogonal to any of the vectors $[\beta_3^i, \beta_4^i, \ldots, \beta_s^i]^{\mathrm{T}}$ in $\mathbb{R}^{s-2}$. A simple inductive argument shows that such a vector exists.

A similar argument works when $\boldsymbol{q}$ and $\boldsymbol{r}$ are linearly dependent. □

Using this lemma, we can prove the following result.

Lemma 3.4. *Suppose that $s > 1$ and the RK pair is not* OR$_0$. *If the pair is* R$_0$ *then $\boldsymbol{a}_k - \boldsymbol{a}_l \in \operatorname{span}\{\boldsymbol{b}, \widehat{\boldsymbol{b}}\}$ for some $k \neq l$.*

Proof. Suppose that the pair is R$_0$ and $\boldsymbol{a}_k - \boldsymbol{a}_l \notin \operatorname{span}\{\boldsymbol{b}, \widehat{\boldsymbol{b}}\}$ for all $k \neq l$. We must show that the pair is OR$_0$.

First, we prove the result when $s > 2$. Let

$$V := \{\boldsymbol{a}_k - \boldsymbol{a}_l\}_{k \neq l} \cup \{\boldsymbol{a}_i\colon\ \boldsymbol{a}_i \notin \operatorname{span}\{\boldsymbol{b}, \widehat{\boldsymbol{b}}\}\} \cup \{\boldsymbol{e}_i\colon\ \boldsymbol{e}_i \notin \operatorname{span}\{\boldsymbol{b}, \widehat{\boldsymbol{b}}\}\}.$$

Then taking $\boldsymbol{q} = \boldsymbol{b}$ and $\boldsymbol{r} = \widehat{\boldsymbol{b}}$ in Lemma 3.3 it follows that there exists a $\boldsymbol{u}$ such that

$$\boldsymbol{u}^{\mathrm{T}}\boldsymbol{b} = 0, \tag{10}$$

$$\boldsymbol{u}^{\mathrm{T}}\widehat{\boldsymbol{b}} = 0, \tag{11}$$

$$\boldsymbol{u}^{\mathrm{T}}(\boldsymbol{a}_k - \boldsymbol{a}_l) \neq 0, \quad k \neq l, \tag{12}$$

$$\boldsymbol{u}^{\mathrm{T}}\boldsymbol{a}_i \neq 0, \qquad \boldsymbol{a}_i \notin \operatorname{span}\{\boldsymbol{b}, \widehat{\boldsymbol{b}}\}, \tag{13}$$

$$\boldsymbol{u}^{\mathrm{T}}\boldsymbol{e}_i \neq 0, \qquad \boldsymbol{e}_i \notin \operatorname{span}\{\boldsymbol{b}, \widehat{\boldsymbol{b}}\}. \tag{14}$$

Hence, in (7)–(9) we may set $y = h = 1$ (arbitrarily) and $\boldsymbol{\xi} = \boldsymbol{u}$, thereby defining $\boldsymbol{z}$.

It follows from (12) that $z_k \neq z_l$ when $k \neq l$. Hence the solution is valid. Lemma 3.1 then shows that there exists an index ν such that $y = z_\nu$ and $\xi_\nu = 0$; that is, $\boldsymbol{u}^{\mathrm{T}}\boldsymbol{a}_\nu = 0$ and $\boldsymbol{u}^{\mathrm{T}}\boldsymbol{e}_\nu = 0$. So, from (13) and (14) we must have $\boldsymbol{a}_\nu \in \operatorname{span}\{\boldsymbol{b}, \widehat{\boldsymbol{b}}\}$ and $\boldsymbol{e}_\nu \in \operatorname{span}\{\boldsymbol{b}, \widehat{\boldsymbol{b}}\}$, showing that the pair is OR$_0$.

When $s = 2$ and $\boldsymbol{b} \neq \widehat{\boldsymbol{b}}$ the pair is OR$_0$. (This follows because $\operatorname{span}\{\boldsymbol{b}, \widehat{\boldsymbol{b}}\} = \mathbb{R}^2$.) We also consider the case where $s = 2$ and $\boldsymbol{b} = \widehat{\boldsymbol{b}}$, since this may arise when we construct foldings (see below). In this case, taking $\boldsymbol{u} \neq 0$ orthogonal to $\boldsymbol{b}$, we have $\boldsymbol{u}^{\mathrm{T}}\boldsymbol{b} = \boldsymbol{u}^{\mathrm{T}}\widehat{\boldsymbol{b}} = 0$ and $(\boldsymbol{a}_1 - \boldsymbol{a}_2)^{\mathrm{T}}\boldsymbol{u} \neq 0$. Setting $y = h = 1$ (arbitrarily) and $\boldsymbol{\xi} = \boldsymbol{u}$, thereby defining $\boldsymbol{z}$, we have a valid solution of (7)–(9) with $z_1 \neq z_2$. From Lemma 3.1 we must have $y = z_\nu$ and $\xi_\nu = 0$ for some index ν. This means that $\boldsymbol{u}^{\mathrm{T}}\boldsymbol{a}_\nu = 0 = \boldsymbol{u}^{\mathrm{T}}\boldsymbol{e}_\nu$ or, equivalently, $\boldsymbol{a}_\nu \in \operatorname{span}\{\boldsymbol{b}, \widehat{\boldsymbol{b}}\}$ and $\boldsymbol{e}_\nu \in \operatorname{span}\{\boldsymbol{b}, \widehat{\boldsymbol{b}}\}$, showing that the pair is OR$_0$. □

We remark that this result is analogous to [3, Theorem 3(i)], although the technique of proof is different. Furthermore, the idea from [3, Theorem 3(ii)] of using *foldings* to generate a recursive test for regularity can be applied, as we now show.

Definition 3.4. Given an s-stage ($s > 1$) RK pair $\{A, \boldsymbol{b}, \widehat{\boldsymbol{b}}, \boldsymbol{c}\}$, the *folding* of this pair is the $s - 1$ stage pair $\{A^*, \boldsymbol{b}^*, \widehat{\boldsymbol{b}}^*, \boldsymbol{c}^*\}$ defined by

$$\begin{array}{c|c} c^* & A^* \\ \hline & \boldsymbol{b}^{*\mathrm{T}} \\ \hline & \widehat{\boldsymbol{b}}^{*\mathrm{T}} \end{array} = \begin{array}{c|ccccc} c_1 & a_{11}+a_{1s} & a_{12} & \cdots & \cdots & a_{1,s-1} \\ c_2 & a_{21}+a_{2s} & a_{22} & \cdots & \cdots & a_{2,s-1} \\ \vdots & \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & \vdots & & \ddots & \vdots \\ c_{s-1} & a_{s-1,1}+a_{s-1,s} & a_{s-1,2} & \cdots & \cdots & a_{s-1,s-1} \\ \hline & b_1+b_s & b_2 & \cdots & \cdots & b_{s-1} \\ \hline & \widehat{b}_1+\widehat{b}_s & \widehat{b}_2 & \cdots & \cdots & \widehat{b}_{s-1} \end{array}.$$

Lemma 3.5. *With $s > 1$, suppose that $\boldsymbol{a}_k - \boldsymbol{a}_l \in \operatorname{span}\{\boldsymbol{b}, \widehat{\boldsymbol{b}}\}$ for some $k \neq l$. Reorder the rows of the RK tableau so that $k \mapsto 1$ and $l \mapsto s$. The folding of this pair is* R_0 *if and only if the original pair is* R_0.

Proof. Suppose $\boldsymbol{a}_k - \boldsymbol{a}_l \in \operatorname{span}\{\boldsymbol{b}, \widehat{\boldsymbol{b}}\}$ for some $k \neq l$, and reorder the rows of the RK tableau so that $k \mapsto 1$ and $l \mapsto s$. Let $\{A^*, \boldsymbol{b}^*, \widehat{\boldsymbol{b}}^*, \boldsymbol{c}^*\}$ denote the folding of this pair and consider the systems

$$y + h\sum_{j=1}^{s} a_{ij}\xi_j = z_i, \quad 1 \leqslant i \leqslant s, \tag{15}$$

$$\sum_{j=1}^{s} b_j\xi_j = 0, \tag{16}$$

$$\sum_{j=1}^{s} \widehat{b}_j\xi_j = 0, \tag{17}$$

and

$$y + h\sum_{j=1}^{s-1} a^*_{ij}\xi_j = z_i, \quad 1 \leqslant i \leqslant s-1, \tag{18}$$

$$\sum_{j=1}^{s-1} b^*_j\xi_j = 0, \tag{19}$$

$$\sum_{j=1}^{s-1} \widehat{b}^*_j\xi_j = 0. \tag{20}$$

Since $\boldsymbol{a}_1 - \boldsymbol{a}_s \in \operatorname{span}\{\boldsymbol{b}, \widehat{\boldsymbol{b}}\}$, any valid solution $(h, y, \boldsymbol{z}, \boldsymbol{\xi})$ to (15)–(17) must have $z_1 = z_s$ and hence $\xi_1 = \xi_s$. It follows that $(h, y, \{z_i\}_{i=1}^{s-1}, \{\xi_i\}_{i=1}^{s-1})$ solves (18)–(20). Conversely, given a valid solution to (18)–(20), by taking $\xi_s = \xi_1$ and $z_s = z_1$ we obtain a valid solution to (15)–(17). This shows that the original pair is R_0 if and only if the folding is R_0. □

From Lemmas 3.2, 3.4 and 3.5 we can construct the following regularity test.

Recursive test for regularity at $\tau = 0$
Start with the original RK pair.

1. If the current pair is OR_0, then stop: the original pair is R_0.
2. If $\boldsymbol{a}_k - \boldsymbol{a}_l \in \operatorname{span}\{\boldsymbol{b}, \widehat{\boldsymbol{b}}\}$ for some $k \neq l$ then reorder the rows of the RK tableau so that $k \mapsto 1$ and $l \mapsto s$ and apply the recursive test to the folding. Otherwise stop: the original pair is not R_0.

The test is guaranteed to terminate, since if we repeatedly fold down to the level $s = 1$, the current pair is OR_0.

4. Regularity for nonzero tolerances

Our next aim is to move from the limiting case of zero tolerance to the practically-interesting small tolerance regime. Roughly, the Implicit Function Theorem tells us that a spurious solution at $\tau = 0$ can be extended to a spurious solution for small τ. However, care must be taken in deriving results of this type since the validity condition $z_k = z_l \Rightarrow \xi_k = \xi_l$ must not be violated.

As for the $\tau = 0$ case discussed in the previous section, the regularity issue is completely determined by behaviour on scalar problems; so we assume $m = 1$.

We begin with a lemma that gives a canonical form for spurious fixed points.

Lemma 4.1. *If a RK pair admits a global spurious fixed point $(h, y, \boldsymbol{z}, \boldsymbol{\xi})$ for some $\tau \geqslant 0$, then for the same τ, h, y it admits a global spurious fixed point $(h, y, \widehat{\boldsymbol{z}}, \widehat{\boldsymbol{\xi}})$, where $\widehat{z}_k \neq \widehat{z}_l$ for $k \neq l$, unless $\boldsymbol{a}_k - \boldsymbol{a}_l \in \operatorname{span}\{\boldsymbol{b}, \widehat{\boldsymbol{b}}\}$. The same result holds for local spurious fixed points.*

Proof. First we consider global spurious fixed points. We have

$$y + h\boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{\xi} = z_i, \quad 1 \leqslant i \leqslant s, \tag{21}$$

$$\boldsymbol{b}^{\mathrm{T}}\boldsymbol{\xi} = 0, \tag{22}$$

$$\widehat{\boldsymbol{b}}^{\mathrm{T}}\boldsymbol{\xi} = \alpha\tau, \tag{23}$$

where $\alpha = \pm\theta/h$ for EPS control and $\alpha = \pm\theta$ for EPUS control. Since the solution is spurious, we have

$$y \neq z_i \text{ for all } i, \quad \text{or} \quad y = z_v,\ \xi_v \neq 0. \tag{24}$$

If $z_k = z_l$ and $\boldsymbol{a}_k - \boldsymbol{a}_l \notin \operatorname{span}\{\boldsymbol{b}, \widehat{\boldsymbol{b}}\}$ for some $k \neq l$, then $\exists \boldsymbol{v}$ such that $\widehat{\boldsymbol{b}}^{\mathrm{T}}\boldsymbol{v} = \boldsymbol{b}^{\mathrm{T}}\boldsymbol{v} = 0$ and $(\boldsymbol{a}_k - \boldsymbol{a}_l)^{\mathrm{T}}\boldsymbol{v} = 1$. Perturbing $\boldsymbol{\xi}$ to $\widehat{\boldsymbol{\xi}} := \boldsymbol{\xi} + \varepsilon\boldsymbol{v}$, for some $\varepsilon \neq 0$, produces a solution $(h, y, \widehat{\boldsymbol{z}}, \widehat{\boldsymbol{\xi}}) := (h, y, \{z_i + \varepsilon h\boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{v}\}_{i=1}^{s}, \boldsymbol{\xi} + \varepsilon\boldsymbol{v})$. We can choose ε small enough so that $z_i \neq z_j \Rightarrow \widehat{z}_i \neq \widehat{z}_j$ and (24) remains true for the new solution. Now $\widehat{z}_k - \widehat{z}_l = \varepsilon h(\boldsymbol{a}_k - \boldsymbol{a}_l)^{\mathrm{T}}\boldsymbol{v} = h\varepsilon \neq 0$. Continuing this approach, if necessary, we can always reduce the solution to the required form.

In the case of a local spurious fixed point (23) changes to $\widehat{\boldsymbol{b}}^{\mathrm{T}}\boldsymbol{\xi} = \widehat{\alpha}\tau$, where $|\widehat{\alpha}| \leqslant 1/h$ for EPS control and $|\widehat{\alpha}| \leqslant 1$ for EPUS control. The same technique of proof can be used. □

We now show how a lack of regularity when $\tau = 0$ extends to a lack of regularity for small τ.

Theorem 4.1. *If a RK pair is not R_0 then there exist constants τ^*, K, h, y, C (depending only on the RK pair) such that given any $0 < \tau \leqslant \tau^*$ there exists a continuous function f (depending upon τ)*

with a global Lipschitz constant bounded above by K for which the RK pair admits a global spurious fixed point (h, y), *on the ODE* (1) *with*

$$|f(y)| \geqslant C > 0.$$

Proof. Suppose the pair is not R_0. Then there exists a spurious solution at $\tau = 0$, with $f(y) = \widehat{C} \neq 0$, characterized by $(h, y, \boldsymbol{z}, \boldsymbol{\xi})$. Let $\xi_{\max} := \max_i\{|\xi_i|\}$. By Lemma 4.1 we can assume that $z_k \neq z_l$ for $k \neq l$, unless $\boldsymbol{a}_k - \boldsymbol{a}_l \in \operatorname{span}\{\boldsymbol{b}, \widehat{\boldsymbol{b}}\}$.

Case 1. Suppose all z_i are distinct. Now $\exists \boldsymbol{v}$ such that $\boldsymbol{b}^{\mathrm{T}}\boldsymbol{v} = 0$ and $\widehat{\boldsymbol{b}}^{\mathrm{T}}\boldsymbol{v} = 1$. Perturbing $\boldsymbol{\xi}$ to $\widehat{\boldsymbol{\xi}} := \boldsymbol{\xi} + \varepsilon\boldsymbol{v}$, for some $\varepsilon > 0$, produces a solution

$$(h, y, \widehat{\boldsymbol{z}}, \widehat{\boldsymbol{\xi}}) := \left(h, y, \{z_i + \varepsilon h \boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{v}\}_{i=1}^{s}, \boldsymbol{\xi} + \varepsilon\boldsymbol{v}\right).$$

We may choose $\tau = h\varepsilon/\theta$ for EPS control and $\tau = \varepsilon/\theta$ for EPUS control so that, for this τ, the new solution satisfies the error criterion (5) and gives $h_{n+1} = h_n$ in (6).

Now, given a set of real numbers $\{q_i\}_{i=1}^{n}$, define the quantity mindist by

$$\operatorname{mindist}\left(\{q_i\}_{i=1}^{n}\right) := \min_{i \neq j} |q_i - q_j|.$$

Case 1(a). Suppose $y \neq z_i$ for all $1 \leqslant i \leqslant s$. Then

$$\beta := \operatorname{mindist}\left(\{z_i\}_{i=1}^{s}, y\right) > 0.$$

Choosing ε such that $|\varepsilon h \boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{v}| \leqslant \beta/4$ for all i ensures that

$$\operatorname{mindist}\left(\{\widehat{z}_i\}_{i=1}^{s}, y\right) \geqslant \beta/2.$$

This guarantees that the new solution is valid. By reducing ε, if necessary, we can also ensure that $\max_i |\widehat{\xi}_i| \leqslant 2\xi_{\max}$. We can take f to be any function satisfying $f(\widehat{z}_i) = \widehat{\xi}_i$ for $1 \leqslant i \leqslant s$ and $f(y) = \widehat{C}$. Choosing, for simplicity, a piecewise linear interpolant, we find that the Lipschitz constant is bounded above by the maximum of

$$\max_{i \neq j} \frac{|\widehat{\xi}_i - \widehat{\xi}_j|}{|\widehat{z}_i - \widehat{z}_j|} \quad \text{and} \quad \max_{i} \frac{|\widehat{C} - \widehat{\xi}_i|}{|\widehat{z}_i - y|}.$$

The maximum of these two quantities is bounded by

$$K := 2\max\{4\xi_{\max}, C + 2\xi_{\max}\}/\beta,$$

where $C = |\widehat{C}|$.

Case 1(b). Suppose $y = z_\nu$, $\widehat{C} = \xi_\nu \neq 0$. The proof proceeds in a similar way to that of Case 1(a). Let $C = |\xi_\nu|/2$ and $\beta := \operatorname{mindist}(\{z_i\}_{i=1}^{s}) > 0$. Choose ε small enough so that $|\varepsilon h \boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{v}| \leqslant \beta/4$ for all i, giving $\operatorname{mindist}(\{\widehat{z}_i\}_{i=1}^{s}) \geqslant \beta/2$ and $\min_{i \neq \nu} |\widehat{z}_i - y| \geqslant 3\beta/4$ and, by further reduction of ε if necessary, so that $|\widehat{\xi}_\nu| \geqslant C$ and $\max_i |\widehat{\xi}_i| \leqslant 2\xi_{\max}$. The perturbed solution is valid (since the $\widehat{z}_i$ are distinct) and we may set $f(y) = \widehat{\xi}_\nu$. Choosing f to be a piecewise linear interpolant gives a Lipschitz constant bounded above by the maximum of

$$\max_{i \neq j} \frac{|\widehat{\xi}_i - \widehat{\xi}_j|}{|\widehat{z}_i - \widehat{z}_j|} \leqslant \frac{4\xi_{\max}}{\beta/2} = 8\frac{\xi_{\max}}{\beta} \quad \text{and} \quad \max_{i \neq \nu} \frac{|\widehat{\xi}_i - f(y)|}{|\widehat{z}_i - y|} \leqslant \frac{4\xi_{\max}}{3\beta/4} = \frac{16}{3}\frac{\xi_{\max}}{\beta}.$$

Case 2. Suppose $z_k = z_l$ for some $k \neq l$, where $\boldsymbol{a}_k - \boldsymbol{a}_l \in \operatorname{span}\{\boldsymbol{b}, \widehat{\boldsymbol{b}}\}$. Swap rows of the RK tableau so that $k \mapsto 1$ and $l \mapsto s$ and form the folding of this pair. Arguing along the lines of the proof of Lemma 3.5, we see that $(h, y, \{z_i\}_{i=1}^{s-1}, \{\xi_i\}_{i=1}^{s-1})$ is a valid spurious solution of the folding at $\tau = 0$. We now repeat the current proof for the folding. Note that from Lemma 3.5 the folding is not R_0. If Case 1 is used we are done, since the same $f(y)$ can be used for the unfolded method. Otherwise fold again, and so on. Eventually we must exit via Case 1 (otherwise the number of stages would reach $s = 1$, violating the condition that the pair is not R_0). □

Next, we seek a converse of Theorem 4.1 showing a positive consequence of the R_0 property when τ is small. Note, however, that it is not true that R_0 completely eliminates spurious fixed points for small tolerances. To illustrate this, we construct an example using ideas from [4]. Suppose we solve the ODE $y' = -y$ with an explicit two-stage pair, using a second order main formula and a first order secondary formula. Since $s = 2$ and $\operatorname{span}\{\boldsymbol{b}, \widehat{\boldsymbol{b}}\} = \mathbb{R}^2$, such a pair is OR_0 and hence R_0. However, it is easy to verify that global "spurious" fixed points exist, for any given τ, of the form

$$\text{for EPS:} \quad (h, y) = (2, \pm\theta\tau/2), \quad \text{so } |f(y)| = \theta\tau/2,$$
$$\text{for EPUS:} \quad (h, y) = (2, \pm\theta\tau), \quad \text{so } |f(y)| = \theta\tau.$$

In this example the spuriosity is innocuous in the sense that $f(y) = \mathrm{O}(\tau)$. We now show that this effect is generic.

We begin with a result for the case $s = 2$, and then generalize to an arbitrary R_0 pair.

Lemma 4.2. *Suppose the number of stages $s = 2$. Given $\tau > 0$ and $f : \mathbb{R} \to \mathbb{R}$ for which the pair admits a local spurious fixed point (h, y) for the tolerance τ, let L be the Lipschitz constant for f in a region containing y and the stage values z_1, z_2. Then for EPS control and EPUS control, respectively, we have*

$$|f(y)| \leqslant C\tau(1 + hL)/h \quad \text{and} \quad |f(y)| \leqslant C\tau(1 + hL),$$

where C depends only on the RK pair.

Proof. The fixed point satisfies

$$y + h\boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{\xi} = z_i, \quad i = 1, 2, \tag{25}$$
$$\boldsymbol{b}^{\mathrm{T}}\boldsymbol{\xi} = 0, \tag{26}$$
$$|\widehat{\boldsymbol{b}}^{\mathrm{T}}\boldsymbol{\xi}| \leqslant \alpha\tau, \tag{27}$$

where $\alpha = \theta/h$ for EPS control and $\alpha = \theta$ for EPUS control. Since

$$B := \begin{bmatrix} \boldsymbol{b}^{\mathrm{T}} \\ \widehat{\boldsymbol{b}}^{\mathrm{T}} \end{bmatrix} \in \mathbb{R}^{2\times 2}$$

is nonsingular, it follows from (26)–(27) that

$$\|\boldsymbol{\xi}\|_\infty \leqslant \|B^{-1}\|_\infty \alpha\tau.$$

Hence, using (25),

$$|y - z_1| \leqslant h\|\boldsymbol{a}_1\|_1\|\boldsymbol{\xi}\|_\infty \leqslant h\|\boldsymbol{a}_1\|_1\|B^{-1}\|_\infty \alpha\tau.$$

So,

$$\left|f(y)\right| \leqslant \left|f(y) - f(z_1)\right| + \left|f(z_1)\right| \leqslant Lh\|\boldsymbol{a}_1\|_1\left\|B^{-1}\right\|_\infty \alpha\tau + \left\|B^{-1}\right\|_\infty \alpha\tau,$$

giving the required result. □

Theorem 4.2. *Suppose a RK pair is* R_0. *Suppose that (for either EPS or EPUS control) for some tolerance* τ *and some function* $f:\mathbb{R}\to\mathbb{R}$ *the pair admits a local spurious fixed point* (h, y), *which we assume to be in the canonical form described in Lemma* 4.1. *Suppose further that*

$$\min\left\{\min_{z_i\neq z_j}|z_i - z_j|, \min_{z_i\neq y}|z_i - y|\right\} \geqslant \beta > 0$$

and that f *has a Lipschitz constant bounded above by* L *in a region containing* y *and the stage values* $\{z_i\}_{i=1}^s$. *Then there exist constants* $K = K(h, L)$ *and* $\tau^* = \tau^*(h, \beta)$ *such that*

$$\left|f(y)\right| \leqslant K\tau, \quad \textit{whenever } 0 < \tau \leqslant \tau^*.$$

Proof. We consider the EPUS case. A similar proof works for EPS control. Suppose the pair is R_0 and admits a spurious fixed point of the form stated in the theorem.

Since $\boldsymbol{b}$ and $\widehat{\boldsymbol{b}}$ are linearly independent, there exists a vector $\boldsymbol{v}$ (depending only on the RK pair) such that $\boldsymbol{b}^{\mathrm{T}}\boldsymbol{v} = 0$ and $\widehat{\boldsymbol{b}}^{\mathrm{T}}\boldsymbol{v} = 1$. Let $D := \max_i |\boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{v}|$ and $E = \|\boldsymbol{v}\|_\infty$.

We have

$$\begin{aligned} &y + h\boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{\xi} = z_i, \quad 1 \leqslant i \leqslant s,\\ &\boldsymbol{b}^{\mathrm{T}}\boldsymbol{\xi} = 0,\\ &\widehat{\boldsymbol{b}}^{\mathrm{T}}\boldsymbol{\xi} = \gamma\tau, \end{aligned}$$

where $|\gamma| \leqslant \theta$.

Case 1(a). Suppose all z_i are distinct and $y \neq z_i$ for all $1 \leqslant i \leqslant s$. Perturb $\boldsymbol{\xi}$ to $\widehat{\boldsymbol{\xi}} = \boldsymbol{\xi} - \gamma\tau\boldsymbol{v}$, to give a solution $(h, y, \widehat{\boldsymbol{z}}, \widehat{\boldsymbol{\xi}})$ at zero tolerance, where $\widehat{z}_i := z_i - h\gamma\tau\boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{v}$. If

$$hD|\gamma|\tau \leqslant \beta/4, \tag{28}$$

then the set $\{\{\widehat{z}_i\}_{i=1}^s, y\}$ has distinct elements. This means that we have constructed a valid spurious solution at zero tolerance, contradicting the R_0 assumption. Note that (28) is implied by

$$\tau \leqslant \frac{\beta}{4hD\theta} =: \tau^*.$$

Hence, Case 1(a) cannot arise for $\tau \leqslant \tau^*$.

Case 1(b). Suppose all z_i are distinct and $y = z_\nu$ (so $f(y) = \xi_\nu$). The same perturbation used above gives a valid solution at zero tolerance with distinct $\widehat{z}_i$. Since the pair is R_0, this solution cannot be spurious, so we must have $y = \widehat{z}_\nu$ and $\widehat{\xi}_\nu = 0$. But, by construction,

$$\left|\widehat{\xi}_\nu - \xi_\nu\right| \leqslant |\gamma||v_\nu|\tau \leqslant \theta E\tau.$$

Hence, we have $|f(y)| \leqslant \theta E\tau$.

Case 2. Since the solution is in canonical form, in the remaining case there exists a pair of indices $k \neq l$ such that $z_k = z_l$, where $\boldsymbol{a}_k - \boldsymbol{a}_l \in \operatorname{span}\{\boldsymbol{b}, \widehat{\boldsymbol{b}}\}$. Swap rows of the RK tableau so that $k \mapsto 1$

and $l \mapsto s$ and form the folding of this pair. Following the arguments in the proof of Lemma 3.5, we see that a valid fixed point $(h, y, \boldsymbol{z}, \boldsymbol{\xi})$ exists for this τ if and only if the folding has a valid fixed point $(h, y, \{z_i\}_{i=1}^{s-1}, \{\xi_i\}_{i=1}^{s-1})$ for this τ. Note that from Lemma 3.5 the folding is R_0. We now repeat the current proof for the folding. If Case 1(b) arises we are done, otherwise fold again, and so on. Eventually, if Case 1(b) never arises, we reach the level $s = 2$. If we obtain a folding with $s = 2$ stages where $\boldsymbol{b}$ and $\widehat{\boldsymbol{b}}$ are distinct, then an application of Lemma 4.2 gives the result. Otherwise, if $\boldsymbol{b} = \widehat{\boldsymbol{b}}$, we have a spurious fixed point of both individual formulas, contradicting the R_0 assumption. □

So far we have been concerned with small tolerances. It is reasonable to ask which RK pairs can guarantee to avoid spurious fixed points for all $\tau > 0$. Theorem 4.3 below gives a negative answer to this question—only the trivial case of a regular main formula gives rise to a formula pair that is guaranteed never to have a spurious fixed point. In this very demanding sense, standard error control techniques can never regularise a RK formula.

Theorem 4.3. *A RK pair never admits a local spurious fixed point for any continuous function f and any $\tau > 0$ if and only if the main formula (in fixed stepsize mode) never admits a spurious fixed point for any continuous f. The same result is true for global spurious fixed points.*

Proof. The 'if' is trivial. Now suppose that the main formula (in fixed stepsize mode) admits a spurious fixed point. Then we have $(h, y, \boldsymbol{z}, \boldsymbol{\xi})$ such that $y + h\boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{\xi} = z_i$ for $1 \leqslant i \leqslant s$ and $\boldsymbol{b}^{\mathrm{T}}\boldsymbol{\xi} = 0$. If $\widehat{\boldsymbol{b}}^{\mathrm{T}}\boldsymbol{\xi} = \delta \neq 0$ then take $\tau = h|\delta|/\theta$ for EPS control and $\tau = |\delta|/\theta$ for EPUS control to give a global (and hence local) spurious fixed point. Otherwise, if $\widehat{\boldsymbol{b}}^{\mathrm{T}}\boldsymbol{\xi} = 0$, then the pair is not R_0 and we may apply Theorem 4.1. □

Acknowledgements

Most of the work presented here was done during a visit to the numerical analysis group at the University of Geneva. I am grateful to Ernst Hairer and Richard Wain for feedback on this material. I also thank a referee for streamlining the original proof of Lemma 3.4.

References

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

[2] E. Fehlberg, Klassiche Runge–Kutta-Formeln vierter und niedriger Ordnung mit Schrittweiten-Kontrolle und ihre Andwendung auf Wärmeleitungsprobleme, *Computing* 6 (1970) 61–71.

[3] E. Hairer, A. Iserles and J.M. Sanz-Serna, Equilibria of Runge–Kutta methods, *Numer. Math.* 58 (1990) 243–254.

[4] G. Hall, Equilibrium states of Runge–Kutta schemes, *ACM Trans. Math. Software* 11 (1985) 289–301.

[5] A.R. Humphries, Spurious solutions of numerical methods for initial value problems, *IMA J. Numer. Anal.* 13 (1993) 263–290.

[6] A. Iserles, Stability and dynamics of numerical methods for nonlinear ordinary differential equations, *IMA J. Numer. Anal.* 10 (1990) 1–30.

[7] Z. Jackiewicz, R. Vermiglio and M. Zennaro, Regularity properties of Runge–Kutta methods for ordinary differential equations, *Appl. Numer. Math.* 22 (1996) 251–262.

[8] Z. Jackiewicz, R. Vermiglio and M. Zennaro, Regularity properties of Runge–Kutta methods for delay differential equations, *Appl. Numer. Math.* 24 (1997) 263–278.