Numer. Math. 58, 243–254 (1990)

Numerische Mathematik

Equilibria of Runge-Kutta methods

E. Hairer[1], **A. Iserles**[2], **and J.M. Sanz-Serna**[3]

[1] Section de Mathématiques, Université de Genève, Switzerland
[2] Department of Applied Mathematics and Theoretical Physics, University of Cambridge, UK
[3] Departamento de Matemática Aplicada y Computación, Facultad de Ciencias, Universidad de Valladolid, Spain

Received September 5, 1989/June 28, 1990

Summary. It is known that certain Runge-Kutta methods share the property that, in a constant-step implementation, if a solution trajectory converges to a bounded limit then it must be a fixed point of the underlying differential system. Such methods are called *regular*. In the present paper we provide a recursive test to check whether given method is regular. Moreover, by examining solution trajectories of linear equations, we prove that the order of an s-stage regular method may not exceed $2[(s+2)/2]$ and that the maximal order of regular Runge-Kutta method with an irreducible stability function is 4.

Subject classifications: AMS(MOS): 65L05; CR: G1.7.

1 Introduction

The theme of this paper is the investigation of equilibria of Runge-Kutta methods. We assume herewith that the autonomous initial-value problem

$$\mathbf{y}' = \mathbf{f}(\mathbf{y}), \quad t \geqq t_0; \tag{1.1}$$
$$y(t_0) = \mathbf{y}_0 \in \mathscr{R}^n;$$

where $\mathbf{f}$ is a continuous function, is solved by the Runge-Kutta method

$$\begin{aligned} \xi_1 &= \mathbf{f}(\mathbf{y}_m + h(a_{1,1}\xi_1 + a_{1,2}\xi_2 + \ldots + a_{1,s}\xi_s)), \\ \xi_2 &= \mathbf{f}(\mathbf{y}_m + h(a_{2,1}\xi_1 + a_{2,2}\xi_2 + \ldots + a_{2,s}\xi_s)), \\ &\vdots \\ \xi_s &= \mathbf{f}(\mathbf{y}_m + h(a_{s,1}\xi_1 + a_{s,2}\xi_2 + \ldots + a_{s,s}\xi_s)), \\ \mathbf{y}_{m+1} &= \mathbf{y}_m + h(b_1\xi_1 + b_2\xi_2 + \ldots + b_s\xi_s). \end{aligned} \tag{1.2}$$

Offprint requests to: A. Iserles

Here $\mathbf{y}_m$ approximates the exact solution at $t_0 + mh$, $h > 0$. We denote the method (1.2) in the customary way as

$$\begin{array}{c|c} \mathbf{c} & A \\ \hline & \mathbf{b}^T \end{array} \equiv \begin{array}{c|ccc} c_1 & a_{1,1} & a_{1,2} & \dots & a_{1,s} \\ c_2 & a_{2,1} & a_{2,2} & \dots & a_{2,s} \\ \vdots & \vdots & \vdots & & \vdots \\ c_s & a_{s,1} & a_{s,2} & \dots & a_{s,s} \\ \hline & b_1 & b_2 & \dots & b_s \end{array}, \tag{1.3}$$

where $c_k := \sum_{l=1}^{s} a_{k,l}$, $k = 1, \dots, s$.

Classical error bounds can be used to estimate the difference between the Runge-Kutta approximant and the exact solution of (1.1) in a compact time interval [HNW1]. Unfortunately, this leaves open the question of whether, as $t \to \infty$, the dynamics of (1.1) are correctly modeled by the dynamics of the Runge-Kutta solution.

Subjects of interest in the study of dynamics are the invariant objects of (1.1), e.g. fixed points, limit cycles, homoclinic and heteroclinic orbits and strange attractors. Numerical analysis poses two questions: Firstly, are all the invariant objects of (1.1) reproduced, up to an error inherent in the numerical procedure, by the approximant (1.2). Secondly, does each invariant object of (1.2) have a continuous counterpart in (1.1). The first question has been debated in [BEL1], [BEY1], [ISE1], [ISE2] and [KLL1]. The aim of the present paper is to answer the second question in the particular case of fixed points and for the constant step-size $h > 0$.

Let $\mathscr{F}$ be the set of all the zeros of $\mathbf{f}$. Obviously, $\mathscr{F}$ is precisely the set of all the possible bounded limits of the exact solution $\mathbf{y}(t)$, for all sets of initial values in $\mathscr{R}^n$. Furthermore, let $\mathscr{F}_h^*$ denote the set of all the possible bounded limits of the iterated map $\mathbf{y}_m \mapsto \mathbf{y}_{m+1}$ which is induced by (1.2). Note the dependence on h. It has been proved in [ISE1, ISE2] that $\mathscr{F} \subseteq \mathscr{F}_h^*$ is always valid but that it is possible for $\mathscr{F}_h^* \setminus \mathscr{F}$ to be non-empty. Runge-Kutta schemes significantly differ in that respect from multistep methods, that always obey $\mathscr{F}_h^* \equiv \mathscr{F}$.

The possible existence of spurious asymptotics is not universal to all Runge-Kutta methods. In line with [ISE2], we say that a method (1.3) is *regular* if $\mathscr{F}_h^* = \mathscr{F}$ for all $h > 0$ and all initial value problems (1.1) – otherwise it is said to be *irregular*. Examples, introduced in [ISE1] and [ISE2], include the regular schemes

$$\begin{array}{c|cc} \frac{1}{2} - \frac{\sqrt{3}}{6} & \frac{1}{4} & \frac{1}{4} - \frac{\sqrt{3}}{6} \\ \frac{1}{2} + \frac{\sqrt{3}}{6} & \frac{1}{4} + \frac{\sqrt{3}}{6} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \tag{1.4}$$

(fourth-order Gauss-Legendre, a Butcher $\mathbf{I}_A$ method [BUT 1]) and

$$
\begin{array}{c|ccc}
0 & 0 & 0 & 0 \\
\frac{1}{2} & \frac{5}{24} & \frac{1}{3} & -\frac{1}{24} \\
1 & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \\
\hline
 & \frac{1}{6} & \frac{2}{3} & \frac{1}{6}
\end{array} \tag{1.5}
$$

(fourth-order Clippinger-Dimsdale [BUT 1], a Lobatto III_A method [EHL 1]), and the irregular

$$
\begin{array}{c|ccc}
\frac{1}{2}-\frac{\sqrt{15}}{10} & \frac{5}{36} & \frac{2}{9}-\frac{\sqrt{15}}{15} & \frac{5}{36}-\frac{\sqrt{15}}{30} \\
\frac{1}{2} & \frac{5}{36}+\frac{\sqrt{15}}{24} & \frac{2}{9} & \frac{5}{36}-\frac{\sqrt{15}}{24} \\
\frac{1}{2}+\frac{\sqrt{15}}{10} & \frac{5}{36}+\frac{\sqrt{15}}{30} & \frac{2}{9}+\frac{\sqrt{15}}{15} & \frac{5}{36} \\
\hline
 & \frac{5}{18} & \frac{4}{9} & \frac{5}{18}
\end{array} \tag{1.6}
$$

(sixth-order Gauss-Legendre, a Butcher $\mathbf{I}_A$ method [BUT 1]) as well as the second-order explicit scheme

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

Furthermore, it has been proved in [ISE 1] that two-stage Runge-Kutta methods of order $p \geqq 2$ are regular if and only if $a_{1,1}+a_{2,2}=\frac{1}{2}$.

The technique used in [ISE 2] to prove irregularity is based on applying a Runge-Kutta method to the logistic equation $y'=\kappa y(1-y)$. In that case an intuitive explanation is provided by means of the underlying bifurcation diagram. Unfortunately, the logistic equation is not powerful enough to study regularity in general methods. In the present paper we adopt a different approach, investigating whether the Runge-Kutta Eqs. (1.2) may, for a specially "tailored" function $\mathbf{f}$, produce a solution (with $\mathbf{y}_{m+1}=\mathbf{y}_m$) such that $\mathbf{f}(\mathbf{y}_m) \neq \mathbf{0}$. Clearly, this is equivalent to irregularity.

In § 2 we introduce our formalism and use it to characterise regular Runge-Kutta methods in terms of a recursive algebraic condition. This provides an easy computational means of checking a method for regularity, as well as an analytic tool.

In § 3 we analyse the stability function of regular Runge-Kutta methods. We prove that its coefficients are related to Bernoulli numbers. This is exploited to produce the order barrier $2[(s+1)/2]$ for a regular s-stage method. The added requirement of A-stability lowers this bound to 4, subject to the stability function being irreducible. Moreover, we derive the explicit from of those regular, maxi-

mal-order methods whose defining matrix A has only real and positive eigenvalues.

Implementation of (1.2) usually involves a mechanism for step selection. Thus, if (1.2) is written as the map

$$\mathbf{y}_{m+1} = \mathbf{G}(\mathbf{y}_m, h), \tag{1.7}$$

a step of the implemented algorithm reads

$$\begin{aligned} \mathbf{y}_{m+1} &= \mathbf{G}(\mathbf{y}_m, h_m), \\ h_{m+1} &= \gamma(\mathbf{y}_m, h_m), \end{aligned} \tag{1.8}$$

where γ represents the underlying step-control technique. The dynamics of (1.8) were studied in [GRI1] and [HAL1]. Constant-step analysis is relevant to (1.8) in the sense that if $(\hat{\mathbf{y}}, \hat{h})$ is a fixed point of (1.8) then $\hat{\mathbf{y}}$ is a fixed point of (1.7) with the constant step $h = \hat{h}$. Consequently, if (1.2) is regular then every variable-step implementation has no spurious fixed points. Of course, it is perfectly possible for (1.7) to possess for some $h > 0$ a spurious fixed point $\hat{\mathbf{y}}$ that is not inherited by (1.8), since $h \neq \gamma(\hat{\mathbf{y}}, h)$. In other words, the step selection mechanism may operate to eliminate spurious fixed points.

Inasmuch as step selection should be preferred whenever possible, constant-step implementations are sometimes used, e.g. when solving large systems of ODEs that arise from semidiscretization of partial differential equations of evolution. Practitioners employing constant steps should be aware of the possibility of spurious equilibria arising from irregular Runge-Kutta methods.

2 Characterisation of regularity

We assume herewith that $\mathbf{b}^T \mathbf{1} = 1$, where $\mathbf{1} := [1, 1, \ldots, 1]^T$. This is only natural, since the condition is necessary for consistency. We recall that $A\mathbf{1} = \mathbf{c}$.

The method (1.3) is said to be *regular* if for all $h > 0$, all positive integers n and all continuous mappings $\mathbf{f}: \mathscr{R}^n \mapsto \mathscr{R}^n$ the Eqs. (1.2), together with $\mathbf{y}_{m+1} = \mathbf{y}_m$, imply that $\mathbf{f}(\mathbf{y}_m) = \mathbf{0}$, the zero vector. In other words, (1.3) may not produce a solution with wrong steady state for any $h > 0$ and any differential system (1.1).

Given a Runge-Kutta method (1.3) and a step-size $h > 0$ we consider the system

$$\begin{aligned} y + h \sum_{j=1}^{s} a_{i,j} \xi_j &= z_i, \qquad i = 1, \ldots, s \\ \sum_{j=1}^{s} b_j \xi_j &= 0, \end{aligned} \tag{2.1}$$

where y, ξ_j and z_i are vectors in $\mathscr{R}^n$.

Lemma 1. *The method* (1.3) *is regular if and only if, for every* $n \geq 1$ *and for every step-size* $h > 0$, *every solution of* (2.1), *which satisfies* $\xi_k = \xi_l$ *whenever* $z_k = z_l$, *admits an index* ν *such that* $\xi_\nu = 0$ *and* $y = z_\nu$.

Proof. This follows from the definition of regularity because we can construct a continuous function f such that

$$f(z_i)=\xi_i, \quad i=1, \ldots, s, \tag{2.2}$$

if and only if the data z_i, ξ_i satisfy $\xi_k=\xi_l$ whenever $z_k=z_l$. □

We denote by $\mathbf{e}_i$ the ith unit vector, whose dimensionality should be transparent from the context. A Runge-Kutta method (1.3) is said to be *essentially one-stage* (*EOS*) if there exists $k\in\{1, \ldots, s\}$ such that

$$\begin{aligned}\mathbf{b}&=\mathbf{e}_k,\\ \mathbf{a}_k&=c_k\,\mathbf{e}_k,\end{aligned} \tag{2.3}$$

where $\mathbf{a}_l^T$ is the lth row of A. Thus, a step $\mathbf{y}_m\mapsto\mathbf{y}_{m+1}$ with the underlying method produces the same result as a step with the one-stage Runge-Kutta

$$\begin{array}{c|c} c_k & a_{k,k} \\ \hline & b_k \end{array}.$$

(The converse is not always correct – it is entirely possible for $\mathbf{f}$ to exist such that the s-stage implicit equations have no solution, whereas the one-step method is soluble.)

Lemma 2. *An EOS Runge-Kutta method is regular.*

Proof. Let $\hat{\mathbf{y}}$ be a fixed point of an EOS method that obeys (2.3). Since, in the terminology of (2.2), $\sum_i b_i\hat{\xi}_i=\mathbf{0}$, we obtain $\hat{\xi}_k=\mathbf{0}$. Thus, (1.2) yields $\mathbf{f}(\hat{\mathbf{y}}+h\sum_j a_{k,j}\hat{\xi}_j)=\mathbf{0}$ and, by virtue of the second relation of (2.3), $\mathbf{f}(\hat{\mathbf{y}})=\mathbf{0}$ concluding the proof of regularity. □

EOS methods are exceptional and, arguably, not very interesting. It is the non-EOS case that merits the greater attention. The following theorem shows that it is possible to reduce the question of regularity of a non-EOS s-stage method to that of an $(s-1)$-stage method:

Theorem 3. *Let the Runge-Kutta method be non-EOS. Then*

(i) *Regularity implies that there exist distinct* $k, l\in\{1, \ldots, s\}$ *such that* $\mathbf{a}_k-\mathbf{a}_l$ *is a scalar multiple of* $\mathbf{b}$;

(ii) *Suppose that* $\mathbf{a}_k-\mathbf{a}_l$ *is a scalar multiple of* $\mathbf{b}$ *and reorder the Runge-Kutta tableau so that* $k\mapsto 1$ *and* $l\mapsto s$. *Define a new* $(s-1)$*-stage Runge-Kutta method*

$$\begin{array}{c|c} \mathbf{c}^* & A^* \\ \hline & \mathbf{b}^{*T} \end{array} \tag{2.4}$$

by specifying

$$\begin{aligned} A^*&=[I_{s-1}, \mathbf{0}]\,A\begin{bmatrix} I_{s-1} \\ \mathbf{e}_1^T \end{bmatrix};\\ \mathbf{c}^*&=[I_{s-1}, \mathbf{0}]\,\mathbf{c};\\ \mathbf{b}^{*T}&=\mathbf{b}^T\begin{bmatrix} I_{s-1} \\ \mathbf{e}_1^T \end{bmatrix}, \end{aligned} \tag{2.5}$$

where I_{s-1} is the $(s-1)\times(s-1)$ identity matrix. Note that $\mathbf{c}^*=A^*\mathbf{1}_{s-1}$ and $\mathbf{b}^{*T}\mathbf{1}_{s-1}=1$, where $\mathbf{1}_{s-1}=[1,\ldots,1]^T\in\mathscr{R}^{s-1}$. We say that (2.5) is a *folding* of (1.3). Then (1.3) is regular if and only if (2.5) is regular.

Proof. Assuming that (1.3) is regular, we stipulate first that $\mathbf{b}\neq\mathbf{e}_k$ for all $k=1,\ldots,s$. Thus, there exists an open orthant

$$\mathscr{K}:=\{[\xi_1,\ldots,\xi_s]^T\in\mathscr{R}^s:(-1)^{\sigma(l)}\xi_l>0,\ l=1,\ldots,s\},$$

where $\sigma(l)=\pm 1$, $l=1,\ldots,s$, such that $\mathscr{J}:=\mathscr{K}\cap\mathbf{b}^\perp\neq\emptyset$, where

$$\mathbf{b}^\perp=\{[\xi_1,\ldots,\xi_s]^T\in\mathscr{R}^s:\sum_j b_j\xi_j=0\}$$

is the orthogonal complement of $\mathbf{b}$ in $\mathscr{R}^s$. Thus, $\mathscr{J}$ is a cone – a convex set, closed under the map $[\xi_1,\ldots,\xi_s]\mapsto[\lambda\xi_1,\ldots,\lambda\xi_s]$ for all $\lambda>0$. We contend that

$$A\mathscr{J}\subset\bigcup_{\substack{k,l\\k\neq l}}\{\mathbf{z}\in\mathscr{R}^s:z_k=z_l\}. \tag{2.6}$$

For suppose that (2.6) is false. Then there exists $[\zeta_1,\ldots,\zeta_s]^T\in\mathscr{J}$ such that the s-vector $\mathbf{z}:=A\zeta$ has pairwise distinct components. Since $\zeta\in\mathbf{b}^\perp$, it is true that $b_1\zeta_1+\ldots+b_s\zeta_s=0$. Thus, the set $\{\zeta_1,\ldots,\zeta_s,0,z_1,\ldots,z_s\}$ is admissible (for $n=1$). Moreover since $\zeta\in\mathscr{K}$, it follows by Lemma 1 that the method is irregular, contradicting our assumption.

Since (2.6) is valid, convexity of $A\mathscr{J}$ implies that there exist distinct k and l such that

$$A\mathscr{J}\subset\{\mathbf{z}\in\mathscr{R}^s:z_k=z_l\}. \tag{2.7}$$

Thus, $A\mathscr{J}\subset(\mathbf{e}_k-\mathbf{e}_l)^\perp$. Consequently, $\mathscr{J}$ is contained in the orthogonal complement of $A^T(\mathbf{e}_k-\mathbf{e}_l)$. Since that orthogonal complement is a subspace of $\mathscr{R}^s$, we have

$$\operatorname{span}\mathscr{J}\subset(A^T(\mathbf{e}_k-\mathbf{e}_l))^\perp.$$

But $\operatorname{span}\mathscr{J}=\mathbf{b}^\perp$, hence

$$\mathbf{b}^\perp\subset(A^T(\mathbf{e}_k-\mathbf{e}_l))^\perp\equiv(\mathbf{a}_k-\mathbf{a}_l)^\perp$$

and $\mathbf{a}_k-\mathbf{a}_l$ is, indeed, a scalar multiple of $\mathbf{b}\neq\mathbf{0}$.

To complete the proof of (i) we need to consider the case of $\mathbf{b}$ being a coordinate vector, $\mathbf{e}_1$, say. Since the method is non-EOS, $\mathbf{a}_1$ may not be a multiple of $\mathbf{b}$. Thus, the set $\mathscr{J}:=\{\xi\in\mathscr{R}^s:\mathbf{a}_1^T\xi>0,\ \xi_1=0\}$ is non-empty: in fact, its projection to the $(s-1)$-dimensional space spanned by $\xi_2,\ldots,\xi_s$ is a hyperplane. We can now choose an orthant $\mathscr{K}$ in the $(s-1)$-dimensional space (keeping $\xi_1=0$) such that $\mathscr{J}:=\mathscr{K}\cap\mathscr{J}\neq 0$ and $\mathscr{J}$ is a cone. The inclusion (2.7) is seen again to hold: otherwise there exists $[\zeta_1,\ldots,\zeta_s]^T\in\mathscr{J}$ such that $\mathbf{z}:=A\zeta$ has pairwise distinct components. Since $\mathbf{b}=\mathbf{e}_1$, we have $b_1\zeta_1+\ldots+b_s\zeta_s=b_1\zeta_1=0$. Thus, the set $\{\zeta_1,\ldots,\zeta_s,0,z_1,\ldots,z_s\}$ is admissible (for $n=1$). Moreover, since $\zeta_i\neq 0$ for $2\leqq i\leqq s$ and $z_1=\mathbf{a}_1^T\zeta>0$, Lemma 1 implies that the method is irregular, a contradiction. This establishes (2.7) and the remainder of the proof is identical to the case of $\mathbf{b}$ not being a coordinate vector.

Finally, we assume that $a_1 - a_s$ is a scalar multiple of b. Then, every solution of (2.1) also solves

$$y + h \sum_{j=1}^{s-1} a^*_{ij} \xi_j = z_i, \qquad i = 1, \ldots, s-1 \tag{2.8}$$

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

Further, if $\xi_1, \ldots, \xi_{s-1}$ is a solution of (2.8) then, by adding $\xi_s = \xi_1$, we get a solution of (2.1) which satisfies $z_s = z_1$. This and Lemma 1 imply that the original Runge-Kutta method is regular if and only if (2.4) is so. □

Corollary 4. *Let* (1.3) *be a consistent explicit scheme* (i.e. $a_{k,l} = 0$ *for all* $1 \leqq l \leqq k \leqq s$). *Then it is regular if and only if it produces the same solution sequence as the first-order forward Euler method* $\mathbf{y}_{m+1} = \mathbf{y}_m + h\mathbf{f}(\mathbf{y}_m)$.

Proof. If the method is EOS (and this includes the case $s = 1$) then explicitness implies that $\mathbf{a}_k = \mathbf{0}$ in (2.3) and everything reduces to forward Euler. Thus, let us assume that the method is non-EOS. In particular $s \geqq 2$ and, by Theorem 3 there exist distinct k and l such that $\mathbf{a}_k - \mathbf{a}_l = (c_k - c_l)\,\mathbf{b}$. There are two cases: if $\mathbf{a}_k = \mathbf{a}_l$ then the stages $\boldsymbol{\xi}_k$ and $\boldsymbol{\xi}_l$ are identical. Thus, we can remove the kth stage, say, replacing b_l by $b_k + b_l$. Otherwise $c_k \neq c_l$ and, since the method is explicit,

$$b_s = \frac{a_{k,s} - a_{l,s}}{c_k - c_l} = 0.$$

Thus, the sth stage does not contribute to the solution sequence and can be omitted.

In either case the method is equivalent to an $(s-1)$-stage explicit method. We continue by induction, progressively reducing the number of stages until encountering an EOS method which, as we have already seen, is equivalent to forward Euler. □

Note that irregularity of (1.3) means *either* that the first condition of Theorem 3 fails (i.e. the weights are not a multiple of a difference between two rows of the Runge-Kutta matrix) *or* that, if it satisfied, the folded method is irregular. This yields the following recursive terminating algorithm for determination of regularity:

The regularity test

(1) Set $\bar{s} := s$, $\bar{A} := A$, $\bar{\mathbf{b}} := \mathbf{b}$, $\bar{\mathbf{c}} := \mathbf{c}$.
(2) Check if there exists $k \in \{1, \ldots, \bar{s}\}$ such that

$$\bar{b}_j = \delta_{k,j}, \qquad j = 1, \ldots, \bar{s},$$

$$\bar{a}_{k,j} = \bar{c}_k \delta_{k,j}, \qquad j = 1, \ldots, \bar{s},$$

where $\delta_{k,j}$ is Kronecker's delta. If so (and, in particular if $\bar{s} = 1$) go to (5).

(3) Check whether there exist $1 \leqq k < l \leqq \bar{s}$ such that

$$\bar{a}_{k,j} - \bar{a}_{l,j} = (\bar{c}_k - \bar{c}_l)\,\bar{b}_j, \quad j = 1, \ldots, \bar{s}.$$

If no such k and l exist then go to (6).

(4) Set

$$\begin{aligned}
&\bar{s} := \bar{s} - 1;\\
&\bar{b}_k := \bar{b}_k + \bar{b}_l;\\
&\bar{b}_j := \bar{b}_{j+1}, \quad j = l, \ldots, \bar{s};\\
&\bar{c}_i := \bar{c}_{i+1}, \quad i = l, \ldots, \bar{s};\\
&\bar{a}_{i,k} := \bar{a}_{i,k} + \bar{a}_{i,l}, \quad i = 1, \ldots, l-1;\\
&\bar{a}_{i,k} := \bar{a}_{i+1,k} + \bar{a}_{i+1,l}, \quad i = l, \ldots, \bar{s};\\
&\bar{a}_{i,j} := \bar{a}_{i+1,j}, \quad i = l, \ldots, \bar{s}, \quad j = 1, \ldots, l-1,\\
&\bar{a}_{i,j} := \bar{a}_{i+1,j+1}, \quad i,j = l, \ldots, \bar{s}
\end{aligned}$$

and go to (2).

(5) The method is regular. Terminate.

(6) The method is irregular. Terminate.

The dimensionality of a function $\mathbf{f}$ that demonstrates irregularity is of interest. The proofs of Lemma 1 and Theorem 3 show that if condition (i) fails then we can choose a scalar f to interpolate $f(z_i) = \xi_i$, $i = 1, \ldots, s$. Likewise, if (i) is valid but the folded method is irregular then it is seen from the proof that both methods (the original one and the folded) produce a spurious equilibrium for the same problem (1.1). Thus, by iterating the folding (like in the regularity test) until condition (i) fails (as it must, sooner or latter, otherwise we will reach an EOS method, which, by Lemma 2, must be regular) we see that, again, we can choose a scalar f. Moreover, in both cases we may take a polynomial f of degree not exceeding s. Thus, irregularity means that even fairly "simple" systems (1.1) may give rise to wrong equilibria.

An inherent shortcoming of the regularity test is its non-constructivity: It is easy to use it to verify whether a specific method (1.3) is regular, but far more complicated, because of its recursive nature, to answer questions regarding the highest order of a regular Runge-Kutta method of given number of stages, implicitness structure etc. In a forthcoming paper by K. Burrage [BUR 1] the regularity test will serve as a point of departure for the derivation of a non-recursive regularity condition.

3 The stability function of regular Runge-Kutta methods

The test of § 2 provides the means to check whether any given Runge-Kutta method is regular. In the present section we derive some necessary conditions for regularity. They will be obtained by applying the method to the linear test equation $y' = \lambda y$. For this problem the numerical solution is given by $y_{m+1} = R(h\lambda)\, y_m$, where

$$R(z) = \frac{P(z)}{Q(z)} = 1 + z\mathbf{b}^T(I - zA)^{-1}\mathbf{1} \tag{3.1}$$

is the *stability function* of the method. Throughout this section we shall assume that $R(z)$ is an *irreducible* rational function, so that the poles of $R(z)$ are exactly the zeros of $\det(I-zA)$. Moreover, we let $q:=\max\{\deg P, \deg Q\}\leqq s$.

Theorem 5. *If a Runge-Kutta method is regular and of order $p\geqq 1$, then*

$$R(z)=\frac{1+(1+\alpha_1)z+\alpha_2 z^2+\alpha_3 z^3+\ldots+\alpha_q z^q}{1+\alpha_1 z+\alpha_2 z^2+\alpha_3 z^3+\ldots+\alpha_q z^q}. \tag{3.2}$$

Proof. Regularity implies that $R(z)\neq 1$ whenever $z\neq 0$. Consequently, $P(z)\neq Q(z)$ for $z\neq 0$. This together with the order assumption implies that $P(z)=Q(z)+z$. □

Theorem 6. *The stability function* (3.2) *satisfies $R(z)=e^z+O(z^{p+1})$ iff*

$$\alpha_i=\frac{B_i}{i!} \quad \text{for } i=1,\ldots,p-1,$$

where B_i are the Bernoulli numbers (recall that $B_{2j+1}=0$ for $j=1,2,\ldots$ and that $B_0=1$, $B_1=-1/2$, $B_2=1/6$, $B_4=-1/30$, $B_6=1/42$, etc.).

Proof. The assumption $P(z)=Q(z)+z$ inserted into $P(z)/Q(z)=e^z+O(z^{p+1})$ implies

$$Q(z)=\frac{z}{e^z-1}+O(z^p)$$

and the result follows from the definition of the Bernoulli numbers

$$\frac{z}{e^z-1}=\sum_{i\geqq 0} B_i\frac{z^i}{i!}. \quad \square$$

It has been proved in [ISE2] that a two-stage Runge-Kutta method of order $p\geqq 2$ is regular if and only if it is equivalent to a method with $\operatorname{tr} A=1/2$. Herewith we provide a short proof of this statement.

Theorem 7. *A regular Runge-Kutta method of order $p\geqq 2$ satisfies*

$$a_{1,1}+a_{2,2}+\ldots+a_{s,s}=\tfrac{1}{2}. \tag{3.3}$$

For $s=2$ and $p\geqq 2$ condition (3.3) *is sufficient for regularity.*

Proof. The necessity of (3.3) follows from Theorem 6, because $a_{11}+\ldots+a_{ss}=\alpha_1$. For the proof of sufficiency (in the case $s=2$) we use the regularity test of § 2. We assume $c_1\neq c_2$ (otherwise the method would be reducible) so that

$$b_1=\frac{c_2-\frac{1}{2}}{c_2-c_1}, \qquad b_2=\frac{\frac{1}{2}-c_1}{c_2-c_1}.$$

The assumption (3.3) then yields

$$a_{21}-a_{11}=c_2-(a_{11}+a_{22})=c_2-\tfrac{1}{2}=(c_2-c_1)\,b_1$$
$$a_{22}-a_{12}=(a_{11}+a_{22})-c_1=\tfrac{1}{2}-c_1=(c_2-c_1)\,b_2$$

which proves the regularity of the method. □

For many classical implicit Runge-Kutta methods (e.g. Butcher's methods of type I, II and III [BUT 1], or the A-stable methods of Ehle [EHL 1], Axelsson [AXE 1] and Chipman [CHI 1]; see [DEV 1] for a collection of these methods) the stability function is a Padé approximant to the exponential function. The following result shows that most of these methods can not be regular.

Theorem 8. *The only Padé approximants which are of the form* (3.2) *are*

$$R_{1/0}(z)=1+z, \qquad R_{0/1}(z)=\frac{1}{1-z},$$
$$R_{1/1}(z)=\frac{1+\frac{z}{2}}{1-\frac{z}{2}}, \qquad R_{2/2}(z)=\frac{1+\frac{z}{2}+\frac{z^2}{12}}{1-\frac{z}{2}+\frac{z^2}{12}}.$$

Proof. This follows from Theorem 6 and the requirement $p=\deg P+\deg Q$. □

Apart from one-stage methods only the two-stage Gauss method and the 3-stage Lobatto IIIA method are regular. (Note that $q=2$ for the last two methods.) The coefficients of these methods have been given in the introduction.

Theorem 9. *The order p of an s-stage regular Runge-Kutta method satisfies*

$$\begin{aligned} p&\leqq s+2 \quad \textit{if } s \textit{ is even},\\ p&\leqq s+1 \quad \textit{if } s \textit{ is odd}. \end{aligned} \tag{3.4}$$

Proof. This follows immediately from Theorem 6 because $\alpha_j=0$ for $j>q$ and since $q\leqq s$. The different order barriers for s even and s odd are due to the fact that the odd Bernoulli numbers vanish, except B_1. □

Remark. The order barrier (3.4) also holds if we allow the stability function to be reducible. In this situation it can be written as

$$R(z)=\frac{P(z)}{Q(z)}\cdot\frac{S(z)}{S(z)}$$

where $P(0)=Q(0)=S(0)=1$, $\det(I-zA)=Q(z)\cdot S(z)$ and the polynomials P and Q are relatively prime. As in the proof of Theorem 5 we conclude that

$$P(z)=Q(z)+zT(z) \tag{3.5}$$

where $T(z)$ divides $\det(I-zA)$ and hence also $S(z)$. If we insert (3.5) into the order condition $P(z)=Q(z)\cdot e^z+O(z^{p+1})$ we obtain

$$\frac{Q(z)}{T(z)}=\frac{z}{e^z-1}+O(z^p).$$

Since $\deg Q + \deg T \leqq s$, one can deduce from the Padé tableau for the function $z/(e^z-1)$ (see [PER 1]) that p is bounded by (3.4).

For $s \leqq 3$ the order barrier (3.4) is optimal. For $s \geqq 4$ it is not clear whether (3.4) can be improved or not. So far, no regular Runge-Kutta method of order $p > 4$ is known to the authors. For Runge-Kutta methods whose defining matrix A has only real and positive eigenvalues we have the following result:

Theorem 10. *The maximal order of a regular Runge-Kutta method with* $\det(I - zA) = (1-\gamma_1 z)\cdot\ldots\cdot(1-\gamma_s z)$ *and* $\gamma_j > 0$ *is 3.*

Proof. This is a consequence of Theorem 6, because $\alpha_3 = -\sum_{i<j<k} \gamma_i \gamma_j \gamma_k < 0$. □

Example. A general 3-rd order SDIRK method with $a_{ii} = 1/6$ (c.f. condition (3.3)) is given by

$$\begin{array}{c|ccc} 1/6 & 1/6 & & \\ c_2 & c_2 - 1/6 & 1/6 & \\ c_3 & c_3 - \alpha - 1/6 & \alpha & 1/6 \\ \hline & b_1 & b_2 & b_3 \end{array} \tag{3.6}$$

where

$$b_1 = \frac{\frac{1}{3}c_3 - \frac{1}{4}}{(c_3 - c_2)(c_2 - \frac{1}{6})}, \qquad b_2 = \frac{\frac{1}{4} - \frac{1}{3}c_2}{(c_3 - c_2)(c_3 - \frac{1}{6})}, \qquad b_1 = 1 - b_2 - b_3,$$

and

$$\alpha = \frac{(c_3 - c_2)(c_3 - \frac{1}{6})}{36(c_2 - \frac{1}{6})(\frac{1}{4} - \frac{1}{3}c_3)}.$$

In order to find regular methods of this type we consider a folding of the third row with the first one. This leads to the conditions

$$\alpha = (c_3 - \tfrac{1}{6})\, b_2, \qquad \tfrac{1}{6} = (c_3 - \tfrac{1}{6})\, b_3,$$

both of which are equivalent to $c_2 + c_3 = \frac{3}{2}$. The folded method can be seen to be regular under the same condition. Consequently, the method (3.6) is regular, provided that $c_2 + c_3 = \frac{3}{2}$.

Finally, we present a stability barrier for regular Runge-Kutta methods. We recall that a method is called *$A(\alpha)$-stable*, if the stability function satisfies

$$|R(z)| \leqq 1 \quad \text{for } |\pi - \arg z| \leqq \alpha. \tag{3.7}$$

It is called *A-stable*, if it satisfies (3.7) with $\alpha = \pi/2$.

Theorem 11. *If the stability function* (3.2) *with* $\alpha_q \neq 0$ *is $A(\alpha)$-stable, then*

$$\alpha \leqq \frac{\pi}{2(q-1)}.$$

Proof. Since $P(z) = Q(z) + z$, the condition $|R(z)| \leqq 1$ (or $|Q(z)+z|^2 \leqq |Q(z)|^2$) is equivalent to

$$2\,\mathrm{Re}(\bar{z}Q(z)) \leqq -|z|^2.$$

On the ray $z = re^{it}$ with $r \to \infty$, $|\pi - t| \leqq \alpha$, we obtain

$$2\alpha_q r^{q+1} \cos((q-1)t) + O(r^q) \leqq -r^2.$$

yielding $\alpha_q \cos((q-1)t) \leqq 0$. Since $A(\alpha)$-stability implies $A(0)$-stability, and that, in turn, implies that $(-1)^q \alpha_q > 0$ (otherwise the stability function has a negative pole), we have $(-1)^{q-1} \cos((q-1)t) \geqq 0$ and our assertion follows. □

Corollary 12. *An A-stable regular Runge-Kutta method with an irreducible stability function obeys $q \leqq 2$, hence its order satisfies $p \leqq 4$.* □

References

[AXE1] Axelsson, O.: A class of A-stable methods. BIT **9**, 185–199 (1969)

[BEL1] Beyn, W.-J., Lorenz, J.: Center manifolds of dynamical systems under discretization. Numer. Funct. Anal. Optimization **9**, 381–414 (1987)

[BEY1] Beyn, W.-J.: On invariant closed curves for one-step methods. Numer. Math. **51**, 103–122 (1987)

[BUR1] Burrage, K.: (to appear)

[BUT1] Butcher, J.C.: Implicit Runge-Kutta processes. Math. Comput. **18**, 50–64 (1964)

[CHI1] Chipman, F.H.: A-stable Runge-Kutta processes. BIT **11**, 384–388 (1971)

[DEV1] Dekker, K., Verwer, J.G.: Stability of Runge-Kutta methods for stiff nonlinear equations, 1st Ed. Amsterdam: North-Holland 1984

[EHL1] Ehle, B.L.: On Padé approximations to the exponential function and A-stable methods for the numerical solution of initial value problems. Ph.D. dissertation, Univ. of Waterloo 1969

[GRI1] Griffiths, D.F.: The dynamics of some linear multistep methods with step-size control. In: Griffiths, D.F., Watson, G.A. (eds.) Numerical analysis 1987, pp. 115–134. Harlow: Pitman 1988

[HAL1] Hall, G.: Equilibrium states of Runge-Kutta codes. ACM Trans. Math. Software **20**, 289–301 (1985)

[HNW1] Hairer, E., Nørsett, S.P., Wanner, G.: Solving ordinary differential equations I: nonstiff problems, 1st Ed. Berlin Heidelberg New York: Springer 1987

[ISE1] Iserles, A.: Stability and dynamics of numerical methods for nonlinear ordinary differential equations. IMA J. Num. Anal. **10**, 1–30 (1990)

[ISE2] Iserles, A.: Nonlinear stability and asymptotics of O.D.E. solvers. In: Agarwal, R.P. (ed.) International Conference on Numerical Mathematics. Basel: Birkhäuser 1989

[KLL1] Kloeden, P.E., Lorenz, J.: Stable attracting sets in dynamical systems and in their one-step discretizations. SIAM J. Numer. Anal. **23**, 986–995 (1986)

[PER1] Perron, O.: Die Lehre von den Kettenbrüchen. 3rd ed. Stuttgart: Teubner 1954

[STE1] Stetter, H.J.: Analysis of discretization methods for ordinary differential equations, 1st Ed. Berlin Heidelberg New York: Springer 1973