Eidgenössische
Technische Hochschule
Zürich

Ecole polytechnique fédérale de Zurich
Politecnico federale di Zurigo
Swiss Federal Institute of Technology Zurich

Generalized disks of contractivity for explicit and implicit Runge-Kutta methods

Germund Dahlquist* and Rolf Jeltsch

Research Report No. 2008-20
May 2008

Seminar für Angewandte Mathematik
Eidgenössische Technische Hochschule
CH-8092 Zürich
Switzerland

*Dept. of Numerical Analysis and Computing Science, The Royal Institute of Technology, S-100 44 Stockholm 70, Sweden

Generalized disks of contractivity for explicit and implicit Runge-Kutta methods

Germund Dahlquist[*] and Rolf Jeltsch

Seminar für Angewandte Mathematik
Eidgenössische Technische Hochschule
CH-8092 Zürich
Switzerland

Research Report No. 2008-20   May 2008

Abstract

The A-contractivity of Runge-Kutta methods with respect to an inner-product norm, was investigated thoroughly by Butcher and Burrage (who used the term B-stability). Their theory is here extended to contractivity in a circular region tangential to the imaginary axis at the origin. The largest possible circle is calculated for many known explicit Runge-Kutta methods. As a rule it is considerably smaller than the stability region, and in several cases it degenerates to a point.

[*] Dept. of Numerical Analysis and Computing Science, The Royal Institute of Technology, S-100 44 Stockholm 70, Sweden

1 Introduction

We investigate the contractivity of Runge-Kutta methods when applied to nonlinear differential equations. While stability of a method is concerned with the boundedness of the numerical result, contractivity requests that the difference of any two numerical solutions, computed with the same stepsize, does not grow in a certain norm. For one-step methods and the natural norm, given by the differential equation, both concepts are identical if the differential equation is linear with constant coefficients. In the other cases contractivity is a stronger requirement.

For linear multistep methods contractivity has been introduced by Dahlquist [4], where it was called G-stability. G stands for a positive definite matrix which is method dependent and is used to define a norm in the space of numerical solutions. Nevanlinna and Liniger [10] have treated contractivity of linear multistep methods using method independent norms, such as the maximum norm. Butcher [3] introduced B-stability which is contractivity for non-linear, autonomous contractive differential equations using the natural norm. In [1] similar contractivity concepts have been discussed, namely AN-stability for non-autonomous linear and BN-stability for non-autonomous nonlinear systems. These concepts reduce to A-stability in the linear constant coefficient case and are thus only reasonable for implicit methods. We extend the contractivity concept for Runge-Kutta methods in such a way that explicit methods are included too. We will be using the natural norm in contrast to [2] where an idea similar to Dahlquist's G-stability is introduced. In all these concepts one requests a certain monotonicity condition for the differential equation. In the present article this condition is given in (2.9). Then it is shown that the numerical method when applied to such a differential equation is contractive for either arbitrary or special choices of the stepsize h.

In the remaining part of this section we give an outline of the article. In Section 2 basic notations and definitions are given. In particular the monotonicity condition for the nonlinear differential equations and the concept of contractivity are described. In Section 3 the r-circle contractivity is introduced. If a method is r-circle contractive then the stability region contains the interior or exterior of a disk of radius $|r|$ which is tangential to the imaginary axis at the origin. However, the converse is not true, i.e. there are methods whose stability region contains a disk of radius r with the origin on the boundary which are not r-circle contractive. We then give purely algebraic necessary and sufficient condition in terms of the coefficients for a method to be r-circle contractive. An algorithm is given which enables one to compute r for any given explicit or implicit Runge-Kutta method. It is natural to introduce the concept of reducible methods. An m-stage Runge-Kutta method is reducible if there exists an m'-stage Runge-Kutta method with $m' < m$ and both methods give identical results on any computer which carries out additions of 0 and multiplications by 0 without round-off errors. It is then shown that for irreducible r-circle contractive methods $\frac{1}{r}$ is a continuous function of the coefficients of the method and that this is not the case if one admits reducible methods. Further confluent methods are introduced. A method is called confluent if at least two of the row sums of the coefficient matrix A are equal. It is then shown that to any confluent method, which is r-circle contractive and to any $\varepsilon > 0$ there exists a nonconfluent method which is r'-circle contractive and $|\frac{1}{r} - \frac{1}{r'}| < \varepsilon$. In Section 4 we show that one has numerical contractivity for nonlinear differential equations if the method is r-circle contractive, if h is chosen appropriately. In Section 5 we show that for an explicit r-circle contractive method one has $r \leq m$, where m is the number of stages. This result is sharp. Further if r is negative then the error order $p = 1$ and $r \leq \frac{1}{2c}$ where c is the error constant of the method. Finally, we list r of many of the well known explicit Runge-Kutta methods.

2 The methods and the test equation

For solving initial value problems

$$(2.1) \qquad y'(t) = f\big(t, y(t)\big), \;\; y(0) \text{ given}, \;\; y, f \in \mathbb{R}^s \text{ or } \mathbb{C}^s\,,$$

we consider m stage Runge-Kutta methods. Let $h > 0$ be the stepsize, $t_n = nh$ and y_n is the numerical approximation to the exact solution $y(t_n)$. The numerical solution y_{n+1} at $t_{n+1} = t_n + h$ is computed as

$$(2.2) \qquad y_{n+1} = y_n + h \sum_{j=1}^{m} b_j\, f(t_n + c_j h, Y_j)\,,$$

where

$$(2.3) \qquad Y_i = y_n + h \sum_{j=1}^{m} a_{ij}\, f(t_n + c_j h, Y_j), \quad i = 1, 2, \ldots, m\,.$$

We always request

$$(2.4a) \qquad \sum_{j=1}^{m} b_j = 1$$

and

$$(2.4b) \qquad c_i = \sum_{j=1}^{m} a_{ij}\;.$$

Observe that by (2.4a) the method has an error order of at least one. (2.4b) is not necessary for a method to be convergent, see [11]. However, it is convenient in notation to have (2.4b) and practically all known methods satisfy (2.4b). Moreover, the extension of the present results to methods without (2.4b) is trivial. If the matrix $A = \{a_{ij}\}$ is strictly lower triangular then the method is called explicit otherwise implicit. We call a method *nonconfluent* if all c_i are distinct and *confluent* otherwise. For compactness in notation we introduce the vectors $Y, F_n(Y) \in \mathbb{R}^{ms}$ or $\mathbb{C}^{ms}$ and $\mathbb{1} \in \mathbb{R}^m$ defined by

$$(2.5) \qquad Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_m \end{pmatrix}, \quad F_n(Y) := \begin{pmatrix} f(t_n + c_1 h, Y_1) \\ f(t_n + c_2 h, Y_2) \\ \vdots \\ f(t_n + c_m h, Y_m) \end{pmatrix}, \quad \mathbb{1} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}.$$

With $b^T = (b_1, b_2, \ldots, b_m)$ one can write (2.2) as

$$(2.6) \qquad y_{n+1} = y_n + hb^T \otimes I_s F_n(Y)$$

and (2.3) takes the form

$$(2.7) \qquad Y = \mathbb{1} \otimes y_n + hA \otimes I_s F_n(Y)\,.$$

Here $\otimes$ denotes the Kronecker-product, see [5,p. 116] and I_s is the $s \times s$ identity matrix.

The aim of this article is to show for Runge-Kutta methods that for any two numerical solutions $\{y_n\}_{n=0,1,\ldots}$, $\{z_n\}_{n=0,1,\ldots}$ which are computed with the same h one has

$$(2.8) \qquad \|y_{n+1} - z_{n+1}\| \leq \|y_n - z_n\|, \quad n = 0, 1, \ldots\ .$$

We assume here that $\|u\| := \langle u, u\rangle^{\frac{1}{2}}$ where $\langle\cdot,\cdot\rangle$ is an innerproduct defined on $\mathbb{R}^s$ or $\mathbb{C}^s$. Note that in contrast to G-stability [4] and the nonlinear stability in [2] the norm does not depend on the method used but only on the differential equation treated. We talk of numerical contractivity if (2.8) is satisfied. The main purpose of this article is to show numerical contractivity. To do this we need to impose conditions on the differential equations and on the methods. The condition on the method is the r-circle contractivity which is treated in Section 3. For the differential equation we request the monotonicity condition

$$(2.9) \qquad Re\big\langle f(t,y) - f(t,z), y - z\big\rangle \leq -\alpha\, \|f(t,y) - f(t,z)\|^2 \quad \forall y, z \in \mathbb{R}^s \text{ or } \mathbb{C}^s\,.$$

In Section 4 we shall show that if α, r and the stepsize h satisfy the inequality (4.2) then one has numerical contractivity. To clarify the condition (2.9) we observe that for a linear equation $y' = \lambda y$ condition (2.9) becomes

$$Re(1+\alpha\lambda)/\lambda \geq 0\,. \tag{2.10}$$

Thus if we introduce the generalized disks

$$D(r) = \begin{cases} \{\lambda \in \mathbb{C} \mid \ \ |\lambda + r| \leq r\} & \text{if } r > 0 \\ \{\lambda \in \overline{\mathbb{C}} \mid \ \ Re\,\lambda \leq 0\} & \text{if } r = \infty \\ \{\lambda \in \overline{\mathbb{C}} \mid \ \ |\lambda + r| \geq -r\} & \text{if } r < 0 \end{cases} \tag{2.11}$$

then (2.10) is equivalent to $\lambda \in D(\frac{1}{2\alpha})$. If $\alpha \geq 0$ then (2.9) implies that for two solutions $y(t)$ and $z(t)$ of (2.1) one has

$$\frac{d}{dt}\,\|y(t) - z(t)\|^2 \leq 0 \text{ for all } t\,.$$

Further observe that α is not invariant against scaling. Let $y(t)$ be a solution of (2.1) and define $z(t) := y(\tau t)$. Then $z(t)$ is a solution of the scaled system

$$z'(t) = g(t,z)$$

where

$$g(t,z) := \tau f(t\tau, y)\,.$$

If (2.1) satisfies (2.9) with $\alpha = \alpha_f$ then g satisfies (2.9) with $\alpha = \alpha_g = \tau\alpha_f$. Moreover (2.9) with $\alpha > 0$ implies that f is Lipschitz-continuous with $\frac{1}{\alpha}$ as Lipschitz-constant for one has

$$\begin{aligned} \|f(t,y) - f(t,z)\|\,\|y - z\| &\geq Re\big\langle -f(t,y) + f(t,z), y - z\big\rangle \\ &\geq \alpha\|f(t,y) - f(t,z)\|^2\,. \end{aligned}$$

Here we have used Schwarz's inequality and (2.9).

3 The *r*-circle contractivity

In this section we define r-circle contractivity. In order to motivate this definition we consider the scalar test equation

$$y' = \lambda(t)y(t), \ \ \lambda(t) \in \mathbb{C}\,. \tag{3.1}$$

If one applies (2.6), (2.7) to (3.1) the numbers

$$\zeta_i = h\lambda(t_n + hc_i), \quad i = 1, 2, \ldots, m \tag{3.2}$$

and $\zeta = (\zeta_1, \zeta_2, \ldots, \zeta_m)^T$ are needed. Assume that (3.1) satisfies the monotonicity condition (2.9) then $\zeta_i \in D(r)$ with $r = h/2\alpha$. If the c_i are distinct then one can choose any m complex numbers $\zeta_i \in D(r)$ and find a smooth $\lambda(t)$ such that (3.2) holds. Applying (2.6), (2.7) to (3.1) leads to

$$y_{n+1} = K(\zeta) y_n \tag{3.3}$$

where

$$K(\zeta) = 1 + b^T Z (I_m - AZ)^{-1} \mathbb{1} \tag{3.4}$$

with

$$Z = \operatorname{diag}(\zeta_1, \zeta_2, \ldots, \zeta_m)\,, \tag{3.5}$$

see [1]. Clearly we have numerical contractivity if $|K(\zeta)| \leq 1$. This leads to the

Definition 3.1. *A Runge-Kutta method is called* ***r-circle contractive*** *if $D(r)$ is the largest generalized disk with $r \neq 0$ and*

$$|K(\zeta)| \leq 1 \textit{ for all } \zeta \in D^m(r)\,. \tag{3.6}$$

A method is called ***circle contractive*** *if (3.6) holds from some $r \neq 0$.*

Note that for a confluent method applied to (3.1) one never has $\zeta_i \neq \zeta_j$ if $c_i = c_j$. Nevertheless we request (3.6). One reason for this is, as we shall see at the end of this section, that with the present definition $\frac{1}{r}$ is a continuous function of the coefficients a_{ij} and b_j if the method is irreducible. Clearly $D(r) \subset S$, where S is the stability region of the method, given by

$$S = \left\{ \mu \in \overline{\mathbb{C}} \,\middle|\, K(\mu \mathbb{1}| \leq 1 \right\}.$$

Let

$$Q = BA + A^T B - bb^T = \left(q_{ij}\right)_{i=1,\, j=1}^{m \quad m}\,, \tag{3.7}$$

where

$$B = \operatorname{diag}(b_1, b_2, \ldots, b_m)\,. \tag{3.8}$$

Theorem 3.2. *A Runge-Kutta method is r-circle contractive if and only if*

$$b_j \geq 0 \text{ for } j = 1, 2, \ldots, m \tag{3.9}$$

and $\rho = -\frac{1}{r}$ is the largest number such that

$$w^T Q w \geq \rho w^T B w \text{ for all } w \in \mathbb{R}^m\,. \tag{3.10}$$

Proof. Corollary 4.3 states that (3.9) and (3.10) with an arbitrary ρ' imply (3.6) with $r' = -\frac{1}{\rho'}$ if $\rho' \neq 0$ and $r' = \infty$ if $\rho' = 0$. We further need the converse result, namely

Lemma 3.3. *Assume (3.6) holds for some $r' \neq 0$, r' may be infinite. Then (3.9) and (3.10) hold for $\rho' = -\frac{1}{r'}$ if r' is finite and $\rho' = 0$ otherwise.*

From this lemma and Corollary 4.3 follows immediately the theorem with $\rho = -\frac{1}{r}$. □

To show Lemma 3.3 we need the following lemma of Burrage and Butcher [1].

Lemma 3.4. *Let Z be such that $I_m - AZ$ is nonsingular and let*

$$u = (I_m - AZ)^{-1}\,\mathbb{1}\,. \tag{3.11}$$

Then

$$|K(\zeta)|^2 - 1 = 2\sum_{i=1}^{m} b_i\,|u_i|^2\,Re\,\zeta_i - \sum_{i,j=1}^{m} q_{ij}\,\overline{\zeta}_i\,\overline{u}_i\,\zeta_j\,u_j\,. \tag{3.12}$$

Proof of Lemma 3.3. Assume that for r' one has (3.6), that is

$$|K(\zeta)|^2 - 1 \leq 0 \text{ for all } \zeta \in D^m(r')\,. \tag{3.13}$$

To prove $b_j \geq 0$, assume on the contrary that $b_i < 0$ for some i. Choose $\zeta_j = 0$ for $j \neq i$ and $\zeta_i = -\varepsilon$. For $\varepsilon > 0$ sufficiently small one has $\zeta \in D^m(r')$. By (3.11),

$$u_j = 1 + \psi_j(\varepsilon), \text{ where } |\psi_j(\varepsilon)| \to 0 \text{ as } \varepsilon \to 0, j = 1, 2, \ldots, m\,. \tag{3.14}$$

The right hand side of (3.12) becomes

$$-2b_i\varepsilon + \varepsilon k(\varepsilon) \tag{3.15}$$

with $|k(\varepsilon)| \to 0$ as $\varepsilon \to 0$. (3.15) is positive for ε sufficiently small. This contradicts (3.13).

In order to show that $w^T(Q + \frac{1}{r'}\ B)w$ is nonnegative we assume the contrary. Let $w = (w_1, w_2, \dots, w_m)^T \in \mathbb{R}^m$ be such that

$$\sum_{i,j=1}^{m} q_{ij}\, w_i w_j + \frac{1}{r'} \sum_{i=1}^{m} b_i w_i^2 < 0\,. \tag{3.16}$$

Let $\varphi_j = \frac{w_j}{r'}$ and $\zeta_j = -r' + r' e^{i\varphi_j \varepsilon} = i w_j\, \varepsilon - \frac{w_j^2}{2r'}\, \varepsilon^2 + O(\varepsilon^3)$. By construction $\zeta = (\zeta_1, \zeta_2, \dots, \zeta_m) \in D^m(r')$ for all ε. Since $\zeta_j \to 0$ as $\varepsilon \to 0$ (3.14) holds again. We substitute ζ_j in the right hand side of (3.12) and find

$$|K(\zeta)|^2 - 1 = \Big(-\frac{1}{r'} \sum_{i=1}^{m} b_i w_i^2 - \sum_{i,j=1}^{m} q_{ij}\, w_i w_j \Big)\, \varepsilon^2 + \varepsilon^2 k_1(\varepsilon) \tag{3.17}$$

with $|k_1(\varepsilon)| \to 0$ as $\varepsilon \to 0$. Hence (3.17) gives a contradiction to (3.13) for ε sufficiently small. Thus (3.10) holds for $\rho' = -\frac{1}{r'}$. □

Remark 3.5. From Theorem 3.2 follows easily that an algebraically stable method in the sense of Burrage and Butcher [1] is r-circle contractive with a non-positive r.

In order to describe the situation where some of the b_j are equal to zero it is convenient to introduce the

Definition 3.6. *An m-stage Runge-Kutta method is called reducible if there exist two sets S and T such that $S \neq \emptyset$, $S \cap T = \emptyset$, $S \cup T = \{1, 2, \dots, m\}$ and*

$$b_k = 0 \quad \textit{if} \ \ k \in S \tag{3.18}$$

$$a_{jk} = 0 \quad \textit{if} \ \ j \in T \ \textit{and} \ k \in S\,. \tag{3.19}$$

The method is called ***irreducible*** *if it is not reducible.*

This definition says that the stages with index in S don't have an influence on the final outcome of the integration provided multiplications by 0 and additions of 0 are performed exactly. It the method is reducible it is equivalent to the m'-stage Runge-Kutta method which consists of the stages with index in T only. Hence m' is the number of elements in T and $m' < m$.

We study now Theorem 3.2 for r-circle contractive methods with some $b_k = 0$. Let S and T be such that $S \cup T = \{1, 2, \dots, m\}$ and

$$b_k = 0 \quad \text{for} \ \ k \in S \tag{3.20}$$

$$b_j > 0 \quad \text{for} \ \ j \in T\,. \tag{3.21}$$

By (3.7), $q_{kk} = 0$ for $k \in S$. Hence for $Q - \rho B$ to be nonnegative definite it is necessary that

$$q_{kj} = 0, \quad j = 1, 2, \ldots, m \text{ for all } k \in S\,. \tag{3.22}$$

Since $q_{kj} = a_{jk} b_j$ when $b_k = 0$ one finds that (3.22) is satisfied if and only if

$$a_{jk} = 0 \text{ whenever } j \in T \text{ and } k \in S\,. \tag{3.23}$$

Thus (3.20), (3.21) and (3.23) imply that the method is reducible. We have therefore shown the

Corollary 3.7. *An irreducible Runge-Kutta method is r-circle contractive if and only if*

$$b_j > 0 \quad \textit{for} \quad j = 1, 2, \ldots, m \tag{3.24}$$

and $\rho = -\frac{1}{r}$ is the largest number such that

$$w^T Q w \geq \rho w^T B w \quad \textit{for all} \quad w \in \mathbb{R}^m\,. \tag{3.25}$$

Let $\mathcal{A}$ be the set of all irreducible circle contractive Runge-Kutta methods. Hence, by Corollary 3.7, a Runge-Kutta method is in $\mathcal{A}$ if and only if all b_j are positive. The methods in $\mathcal{A}$ are the ones which interest us. If a method is not in $\mathcal{A}$ it is either not circle contractive or it is reducible and after deleting the irrelevant stages one may have a member of $\mathcal{A}$. In the following we shall compute r of a given method in $\mathcal{A}$. Since all b_j are positive it follows that $B^{1/2} = \operatorname{diag}(b_1^{1/2}, b_2^{1/2}, \ldots, b_m^{1/2})$ is nonsingular. Using the transformation $B^{1/2} w = v$ reduces (3.25) to

$$v^T B^{-\frac{1}{2}} Q B^{-\frac{1}{2}} v \geq \rho v^T v \quad \text{for all} \quad v \in \mathbb{R}^m\,. \tag{3.26}$$

Let $\nu_1, \nu_2, \ldots, \nu_m$ be the eigenvalues of the real and symmetric matrix $B^{-1/2} Q B^{-1/2}$. Hence the largest ρ for which (3.26) holds is $\rho_{\min} = \min_{i=1,\ldots,m} \nu_i$ and thus by Corollary 3.7 one has

$$r = \begin{cases} \infty & \text{if } \min\limits_{i=1,\ldots,m} \nu_i = 0 \\ -\dfrac{1}{\min\limits_{i=1,\ldots,m} \nu_i} & \text{otherwise}\,. \end{cases} \tag{3.27}$$

Clearly the set $\mathcal{A}$ is open and $\rho_{\min} = \min_{i=1,\ldots,m} \nu_i = \frac{1}{r}$ is a continuous function of the coefficients a_{ij} and b_j of the methods. However, if some of the b_j tend to zero the following possibilities can occur. Either the limiting method is no longer r-circle contractive, see for example Heun's method in Section 6, or else it must become reducible. In the latter case r may depend continuously on b_j or not as the following example shows.

Example 3.8. Let

$$A = \begin{pmatrix} 0 & 0 \\ 0 & \alpha \end{pmatrix}$$

$$b^T = (1-\varepsilon, \varepsilon)\,.$$

Clearly (2.4) and (3.9) are satisfied for all $\varepsilon \in [0,1]$. For $\varepsilon \in (0,1)$ we find

$$B^{-\frac{1}{2}} Q B^{-\frac{1}{2}} = \begin{pmatrix} \varepsilon - 1 & -\varepsilon^{\frac{1}{2}}(1-\varepsilon)^{\frac{1}{2}} \\ -\varepsilon^{\frac{1}{2}}(1-\varepsilon)^{\frac{1}{2}} & -\varepsilon + 2\alpha \end{pmatrix}\,.$$

The eigenvalues are

$$\nu_{1,2} = \frac{1}{2}\left(2\alpha - 1 \pm \sqrt{(2\alpha+1)^2 - 8\alpha\varepsilon}\right).$$

Hence

$$\lim_{\varepsilon \to 0^+} \rho_{\min}(c) = \begin{cases} -1 & \text{if } \alpha \geq -\dfrac{1}{2} \\ 2\alpha & \text{if } \alpha < -\dfrac{1}{2}\,. \end{cases}$$

However, if $\varepsilon = 0$ then the method is reducible and can be reduced to Euler's method with

$$A = (0)$$

$$b^T = (1)$$

and

$$\rho_{\min}(0) = -1\,.$$

Hence one has a discontinuity on $\partial\mathcal{A}$ if $\alpha < -\frac{1}{2}$, and if $\alpha \geq -\frac{1}{2}$ then $\rho_{\min}(\varepsilon)$ is continuous in $[0,1]$.

Note that the set C of confluent methods in $\mathcal{A}$ is a surface in $\mathcal{A}$ of lower dimension. Thus by continuity of $\frac{1}{r}$ as a function of a_{ij} and b_j any confluent r-circle contractive method in $\mathcal{A}$ can be approximated by a non confluent r-circle contractive method such that $\frac{1}{r}$ is as close to $\frac{1}{r'}$ as one wishes. This property would **not** hold if we would have replaced (3.6) by

$$|K(\zeta)| \leq 1 \text{ for all } \zeta \in D^m(r) \cap V \tag{3.28}$$

where

$$V = \{\zeta \in \mathbb{C}^m \,|\, \zeta_i = \zeta_j \text{ whenever } c_i = c_j\}$$

as we can see in the following

Example 3.9. Consider the classical 3-stage Nyström method of order 3 given by

$$A = \begin{pmatrix} 0 & 0 & 0 \\ \frac{2}{3} & 0 & 0 \\ 0 & \frac{2}{3} & 0 \end{pmatrix} ,$$

$$b^T = \left(\begin{array}{ccc} \frac{1}{4} & \frac{3}{8} & \frac{3}{8} \end{array} \right) ,$$

see [9], p. 48. If one computes r using the above algorithm one obtains $r \approx 0.92668857$. If we would have used (3.28) instead of (3.6) in the definition of r-circle contractivity one would have found $r_c = 3$. However, for $a_{31} = \varepsilon$ sufficiently small (3.6) and (3.28) are identical. Thus using (3.28) instead of (3.6) would have resulted in an r which does not depend continuously on the coefficients of the method. This is one reason for choosing (3.6) rather than (3.28). The main reason, however, is the Theorem 4.1 of the next section.

4 Nonlinear contractivity

Theorem 4.1. *Assume the differential equation satisfies the monotonicity condition (2.9) and the Runge-Kutta method is r-circle contractive. Then two numerical solutions y_n and z_n computed using the same stepsize $h > 0$ satisfy*

$$\|y_{n+1} - z_{n+1}\| \le \|y_n - z_n\| \text{ for } n = 0, 1, 2, \ldots \tag{4.1}$$

provided

$$\begin{cases} \dfrac{h}{r} \le 2\alpha & \text{if } r \ne \infty \\ \alpha \ge 0 \text{ and } h \text{ arbitrary} & \text{if } r = \infty\,. \end{cases} \tag{4.2}$$

Proof. First we observe that it is enough to show (4.1) for $n = 0$ only. Subtracting from (2.6) the corresponding equation for the solution $\{z_n\}_{n=0,1,\ldots}$ gives

$$x_1 = x_0 + hb^T \otimes I_s F \tag{4.3}$$

where we have used the abbreviations

$$x_0 = y_0 - z_0, \;\; x_1 = y_1 - z_1, \;\; F = F_0(Y) - F_0(Z)$$

and $Z \in \mathbb{R}^{ms}$ or $\mathbb{C}^{ms}$ is given by

$$Z = \begin{pmatrix} Z_1 \\ Z_2 \\ \dots \\ Z_m \end{pmatrix} . \tag{4.4}$$

In a similar fashion one obtains from (2.7) the equation

$$X = \mathbb{1} \otimes x_0 + hA \otimes I_s F \, , \tag{4.5}$$

where $X = Y - Z$. It is enough to show that

$$\|x_1\|^2 - \|x_0\|^2 \leq 0 \, . \tag{4.6}$$

Substituting (4.3) in (4.6) leads to

$$\|x_1\|^2 - \|x_0\|^2 = h2Re\big\langle x_0, b^T \otimes I_s F \big\rangle + h^2 \|b^T \otimes I_s F\|^2 \, . \tag{4.7}$$

The first term on the right hand side can be simplified if we introduce the following product $[\ ,\ ]$ in $\mathbb{R}^{ms}$ or $\mathbb{C}^{ms}$. Let $U, V \in \mathbb{R}^{ms}$ or $\mathbb{C}^{ms}$ be given by

$$U = \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_m \end{pmatrix} , \quad V = \begin{pmatrix} V_1 \\ V_2 \\ \vdots \\ V_m \end{pmatrix}$$

where $U_i, V_i \in \mathbb{R}^s$ or $\mathbb{C}^s$. Then

$$[U, V] = \sum_{j=1}^{m} b_i \big\langle U_j, V_j \big\rangle \, . \tag{4.8}$$

Hence

$$\big\langle x_0, b^T \otimes I_s F \big\rangle = [\mathbb{1} \otimes x_0, F] \, . \tag{4.9}$$

In order to show (4.6) we need an upper bound for $Re[\mathbb{1} \otimes x_0, F]$. The following lemma is an easy consequence of (2.9) and the definition (4.8).

Lemma 4.2. *Assume $b_j \geq 0$ for $j = 1, 2, \dots, m$ and that the monotonicity condition (2.9) holds. Then*

$$Re[F, X + \alpha F] \leq 0 \, . \tag{4.10}$$

Eliminating X from (4.10) using (4.5) leads to

$$Re[F, \mathbb{1} \otimes x_0] \leq -h \, Re\Big[F, A \otimes I_s F + \frac{\alpha}{h} \, F\Big] \, . \tag{4.11}$$

Using (4.9) and (4.11) in (4.7) gives

$$\|x_1\|^2 - \|x_0\|^2 \leq -h^2 \, Re \, P(F) \tag{4.12}$$

where

$$P(F) = 2\Big[\Big(A \otimes I_s + \frac{\alpha}{h} \, I_{ms}\Big) \, F, F\Big] - \|b^T \otimes I_s F\|^2 \, . \tag{4.13}$$

Observe that $P(F)$ is quadratic form in F and it remains to show that its real part is nonnegative. Let $G \in \mathbb{R}^{ms}$ or $\mathbb{C}^{ms}$ be written as

$$G = \begin{pmatrix} G_1 \\ G_2 \\ \vdots \\ G_m \end{pmatrix}$$

where $G_i \in \mathbb{R}^s$ or $\mathbb{C}^s$. Hence

$$\begin{aligned} Re \, P(G) &= \sum_{j=1}^{m} b_j \Big(\Big\langle \sum_{i=1}^{m} a_{ji} \, G_i, G_j \Big\rangle + \Big\langle G_j, \sum_{i=1}^{m} a_{ji} \, G_i \Big\rangle \Big) \\ &\quad + 2 \, \frac{\alpha}{h} \, \sum_{j=1}^{m} b_j \langle G_j, G_j \rangle - \Big\langle \sum_{i=1}^{m} b_i \, G_i, \sum_{j=1}^{m} b_j \, G_j \Big\rangle \\ &= \sum_{i=1}^{m} \sum_{j=1}^{m} q_{ij} \, \langle G_i, G_j \rangle + 2 \, \frac{\alpha}{h} \, \sum_{i=1}^{m} b_i \, \langle G_i, G_i \rangle \, . \end{aligned}$$

Thus by (3.10) $Re \, P(G)$ is nonnegative if $-\frac{2\alpha}{h} \leq \rho = -\frac{1}{r}$ if $r \neq \infty$. If $r = \infty$ then α has to be nonnegative and h is arbitrary. This completes the proof of Theorem 4.1. □

Observe that (4.2) covers two totally different situations. If α is nonnegative then the differential equation is contractive and one should either use a method with $r < 0$ or $r = \infty$ with h arbitrary or $r > 0$ with $h \leq 2\alpha r$ in order to obtain a contractive numerical scheme. However, if α is negative the differential equation is no longer contractive and hence one no longer wants a contractive scheme. Thus (4.2) should be violated, that is one either wants that $r \in (0, \infty]$ and then h is arbitrary or $r < 0$ and $h \leq 2\alpha r$. Collecting these results we see that one should always choose $h \leq 2\alpha r$ if $\alpha r \in (0, \infty)$

while for $\alpha r \notin (0, \infty)$ the contractivity or noncontractivity does not give any restriction on h. Clearly accuracy will give restrictions on the choice of the stepsize as well as the solvability of (2.7) for Y. This is, however, not the concern of the present article.

Corollary 4.3. *Assume that*

$$b_j \geq 0 \textit{ for } j = 1, 2, \ldots, m\,, \tag{4.14}$$

and

$$w^T Q w \geq \rho' w^T B w \textit{ for all } w \in \mathbb{R}^m\,. \tag{4.15}$$

Then one has

$$|K(\zeta)| \leq 1 \textit{ for all } \zeta \in D^m(r') \tag{4.16}$$

where

$$r' = \begin{cases} \infty & \textit{if } \rho' = 0 \\ -\dfrac{1}{\rho'} & \textit{otherwise}\,. \end{cases}$$

Proof. Let $\zeta = (\zeta_1, \zeta_2, \ldots, \zeta_m)^T \in D^m(r')$, $Z = \mathrm{diag}(\zeta_1, \ldots, \zeta_m)$. Let $h = 1$, $s = 1$ and $F = ZX$ where $X \in \mathbb{C}^m$. Then (4.3) and (4.5) become

$$x_1 = x_0 + b^T Z X \tag{4.17}$$

and

$$X = x_0 \mathbb{1} + AZX\,. \tag{4.18}$$

Thus

$$x_1 = K(\zeta)\, x_0 \tag{4.19}$$

and we have proved the corollary provided (4.6) holds. This is, however, shown exactly in the same way as in the proof of Theorem 4.1. Just observe that $\zeta \in D^m(r')$ implies $Re[ZX, X + \alpha ZX] \leq 0$ for $\alpha = \frac{1}{2r'}$ and that (4.15) with $\rho' = -\frac{1}{r'}$ implies $Re\, P(G) \geq 0$. □

5 Methods with optimal r and examples

Given an r-circle contractive Runge-Kutta method. Let $D(r_s)$ be the largest generalized disk of form (2.11) in the stability region S. Then one has $D(r) \subset D(r_s)$. The following two examples show that $D(r)$ may be a proper subset of $D(r_s)$.

Example 5.1. The θ-method is given by

$$A = \begin{pmatrix} 0 & 0 \\ \theta & 1-\theta \end{pmatrix}$$

$$b^T = \begin{pmatrix} \theta & 1-\theta \end{pmatrix}$$

or

$$y_{n+1} = y_n + h\big(\theta f(t_n, y_n) + (1-\theta)\, f(t_{n+1}, y_{n+1})\big)\,.$$

For $\theta = 0$ it is reducible and can be reduced to the implicit Euler method with $r(0) = -1$. For $\theta = 1$ it is reducible too and the reduced method is the explicit Euler method with $r(1) = 1$. For $0 \in (0,1)$ one finds $r(\theta) = \frac{1}{\theta}$. In particular, for the trapezoidal rule, where $\theta = \frac{1}{2}$, it follows that $r(\frac{1}{2}) = 2$. This result is in agreement with the fact that the trapezoidal rule is not B-stable, see [12]. To compute the stability region we observe that $K(\mu\mathbb{1}) = (1+\mu\theta)/(1-(1-\theta)\mu)$. Hence $S = D\big(r_s(\theta)\big)$ with $r_s(\theta) = 1/(2\theta - 1)$. Therefore one has

$$\begin{array}{ll} r(0) = -1 = r_s(0) & \text{implicit Euler,} \\ \left.\begin{array}{l} D\big(r(\theta)\big) \text{ is a proper subset} \\ \text{of } D\big(r_s(\theta)\big) \end{array}\right\} & \text{for } 0 < \theta < 1 \\ r(1) = 1 = r_s(1) & \text{explicit Euler.} \end{array}$$

It is, however, known [4] that the θ-method has a simple relation to the one-leg methods,

$$z_{n+1} = z_n + hf\big\{\theta t_n + (1-\theta)\, t_{n+1},\, \theta z_n + (1-\theta)\, z_{n+1}\big\}$$

which is A-contractive for $\theta \le \frac{1}{2}$. This indicates that the disks of contractivity calculated in this paper can be larger if a different norm is used. The ideas of Burrage and Butcher [2] are interesting in this context.

Example 5.2. The most general two stage second order explicit Runge-Kutta method is characterized by

$$A = \begin{pmatrix} 0 & 0 \\ \dfrac{1}{2\alpha} & 0 \end{pmatrix}, \quad b^T = (1-\alpha, \alpha), \quad \alpha \neq 0\,,$$

see [6], p. 121. If $\alpha = 1$ the method is reducible and thus not circle contractive. However, for $\alpha \in (0,1)$ one finds by an easy calculation that

$$r(\alpha) = 2\Big/\left(1 + \sqrt{\frac{1}{\alpha(1-\alpha)} - 3}\right), \quad \alpha \in (0,1)\,. \tag{5.1}$$

Here $r(\alpha)$ denotes truly on α, and $r(\alpha) < 1$ for $\alpha \neq \frac{1}{2}$. This is in contrast to the stability region S which is independent of α. In fact $S = \{\mu \in \mathbb{C} \,\big|\, |1 + \mu + \mu^2| \leq 1\}$ and thus

$$r_s(\alpha) = 1 \text{ for all } \alpha \neq 0\,. \tag{5.2}$$

It is well-known that S is bounded for explicit methods. Hence r is positive for explicit circle contractive methods. How large can r actually be?

Theorem 5.3. *Assume an explicit m-stage Runge-Kutta method is r-circle contractive. Then*

$$r \leq m\,. \tag{5.3}$$

Moreover equality is only attained if

$$K(\mu\mathbb{1}) = \left(1 + \frac{\mu}{m}\right)^m, \tag{5.4}$$

which implies that the error order is one. The method with

$$b_i = \frac{1}{m} \qquad i = 1, 2, \ldots, m\,.$$

$$a_{ij} = \begin{cases} 0 & \textit{for } i \leq j \\ \dfrac{1}{m} & \textit{for } i > j \end{cases} \tag{5.5}$$

attains equality in (5.3).

Proof. In [8] it is shown that $r_s \leq m$ with $r_s = m$ if and only if (5.4) holds. Thus from $r \leq r_s$ follows (5.3) and (5.4). If a Runge-Kutta method has error order p, then $K(\mu\mathbb{1}) - e^\mu = O(\mu^{p+1})$. For the special $K(\mu\mathbb{1})$ of (5.4) we find $e^\mu - K(\mu\mathbb{1}) = -\frac{1}{2m}\,\mu^2 + O(\mu^3)$ and thus by (2.4) $p = 1$. An easy calculation shows that $B^{-\frac{1}{2}}QB^{-\frac{1}{2}} = -\frac{1}{m}\,I_m$ for the method given by (5.5). Thus by (3.27) one has $r = m$ and equality in (5.3) holds. □

Let us now consider the same problem for implicit methods. Burrage and Butcher [1] have investigated algebraic stability and shown that there are implicit m-stage Runge-Kutta methods of order $2m$, $2m-1$ and $2m-2$ which are algebraically stable, that is r is nonpositive. The following theorem gives a relation between the size of a negative r and the accuracy of the method.

Theorem 5.4. *Assume the Runge-Kutta method is r-circle contractive with $r < 0$ and*

$$K(\mu\mathbb{1}) - e^{\mu} = -c\mu^{p+1} + O(\mu^{p+2})\,. \tag{5.6}$$

Then

$$p = 1, \;\; c < 0$$

and

$$r \leq \frac{1}{2c}\,.$$

Proof. Let R be the radius of curvature of ∂S at $\mu = 0$. Since $D(r) \subset S$ one has that $0 \leq R \leq -r$. It remains to show that

$$R = \begin{cases} \infty & \text{if } p > 1 \\ -\dfrac{1}{2c} & \text{if } p = 1\,. \end{cases} \tag{5.7}$$

Let ∂S be given in a neighborhood of 0 by the equation $\mu = \xi(t) + it$. $\xi(t)$ is implicitly defined by $|K((\xi(t) + it)\mathbb{1})|^2 = 1$. Using (5.6) we find

$$\begin{aligned} 1 = e^{2\xi(t)}&\Big(1 + c(\xi(t) + it)^{p+1} + O(\xi(t) + it)^{p+2}\Big) \\ &\Big(1 + c(\xi(t) - it)^{p+1} + O(\xi(t) - it)^{p+2}\Big)\,. \end{aligned} \tag{5.8}$$

Implicit differentiation of (5.8) gives

$$\xi'(0) = 0$$

$$\xi''(0) = \begin{cases} 0 & \text{if } p > 1 \\ -2c & \text{if } p = 1 \end{cases}$$

and hence (5.7) follows immediately. □

Observe that for implicit r-circle contractive methods with an nonpositive r the absolute value of r increases as the accuracy increases. In the discussion of Theorem 4.1 we notice that a method with r negative is undesirable. Theorem 5.4 shows that most practical r-circle contractive methods have a nonnegative r. However, there are methods with $r < 0$, for example the implicit Euler method, see Example 5.1.

6 Calculation of r for some explicit methods

We omit the algebraically stable methods given in [1] and restrict ourselves to the explicit methods listed in [9].

All *second order two stage* methods are contained in Example 5.2.

Third order formulas. Observe that for all these formulas one has $r_s \sim 1.25$.

Classic form

$$A = \begin{pmatrix} 0 & 0 & 0 \\ \dfrac{1}{2} & 0 & 0 \\ -1 & 2 & 0 \end{pmatrix}$$

$$b^T = \begin{pmatrix} \dfrac{1}{6} & \dfrac{2}{3} & \dfrac{1}{6} \end{pmatrix} \qquad r = 0.5$$

Nyström form

$$A = \begin{pmatrix} 0 & 0 & 0 \\ \frac{2}{3} & 0 & 0 \\ 0 & \frac{2}{3} & 0 \end{pmatrix}$$

$$b^T = \begin{pmatrix} \frac{1}{4} & \frac{3}{8} & \frac{3}{8} \end{pmatrix} \qquad r \approx 0.927$$

Heun form

$$A = \begin{pmatrix} 0 & 0 & 0 \\ \frac{1}{3} & 0 & 0 \\ 0 & \frac{2}{3} & 0 \end{pmatrix}$$

$$b^T = \begin{pmatrix} \frac{1}{4} & 0 & \frac{3}{4} \end{pmatrix}$$

This method has $b_2 = 0$ and is irreducible. Thus it is not circle contractive.

Ralston's optimum third-order form

$$A = \begin{pmatrix} 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 \\ 0 & \frac{3}{4} & 0 \end{pmatrix}$$

$$b^T = \frac{1}{9}\left(\; 2, \quad 3, \quad 4 \; \right) \qquad r \approx 0.899$$

Kuntzmann's optimum third-order form

$$A = \begin{pmatrix} 0 & 0 & 0 \\ 0.4648162 & 0 & 0 \\ -0.0581020 & 0.8256939 & 0 \end{pmatrix}$$

$$b^T = \begin{pmatrix} 0.2071768 & 0.3585646 & 0.4342585 \end{pmatrix} \qquad r \approx 0.847$$

Fourth order formulas. Observe that for all these formulas one has $r_s \sim 1.4$.

Classical form

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

$$b^T = \begin{pmatrix} \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{pmatrix} \qquad r = 1$$

Kutta form

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & 0 & 0 \\ -\frac{1}{3} & 1 & 0 & 0 \\ 1 & -1 & 1 & 0 \end{pmatrix}$$

$$b^T = \begin{pmatrix} \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{1}{8} \end{pmatrix} \qquad r \approx 0.464$$

Gill form

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 \\ \frac{\sqrt{2}-1}{2} & \frac{2-\sqrt{2}}{2} & 0 & 0 \\ 0 & \frac{-\sqrt{2}}{2} & 1+\frac{\sqrt{2}}{2} & 0 \end{pmatrix}$$

$$b^T = \begin{pmatrix} \frac{1}{6} & \frac{2-\sqrt{2}}{6} & \frac{2+\sqrt{2}}{6} & \frac{1}{6} \end{pmatrix} \qquad r \approx 0.586$$

Kuntzman optimum fourth order form

$$A = \frac{1}{220} \begin{pmatrix} 0 & 0 & 0 & 0 \\ 88 & 0 & 0 & 0 \\ -33 & 165 & 0 & 0 \\ 95 & -75 & 200 & 0 \end{pmatrix}$$

$$b^T = \frac{1}{360} \begin{pmatrix} 55, & 125, & 125, & 55 \end{pmatrix} \qquad r \approx 0.698$$

Ralston's optimum fourth order form given in [9], p. 58 is not circle contractive since $b_2 \sim -0.55198066 < 0$.

The fifth order six stage formulas by Nyström, Luther, Butcher, Sarafayan, Fehlberg and Lawson listed in [9], p. 50-54 are all irreducible and have $b_2 = 0$. Hence these are not circle contractive. The same is true for Shanks almost fifth order 5-stage and the almost sixth order 6-stage methods given in [9]. The sixth order methods of Huta and Butcher [9], p. 55 are not circle contractive since at least one b_j is nonpositive. None of the methods which are used for estimating the local truncation error and are listed in [9], p. 68-76 are circle contractive. Among these formulas one finds methods of Merson, Scraton, Sarafayan, Butcher, Fehlberg, Grabunov and Shakhov.

Finally, we want to point out that we do not claim that the circles calculated here are the true contractivity regions. For the so-called one-leg methods, see e.g. [4] one of the authors has recently shown that the contractivity regions are indeed circles, but we don't yet know if this is true for Runge-Kutta methods.

We also remind of the remark, made in Example 5.1 of Section 5, that one may find different, perhaps larger, contractivity regions with respect to other norms. Therefore, our values of r must not be considered as a final verdict in the comparison of methods. We have found sufficient conditions rather than necessary. Hyman [7] has reported some interesting empirical evidence of the shortcomings of the linear stability theory as a guide-line for the behaviour of Runge-Kutta methods on non-linear problems. We have not yet had the opportunity to study his results from our point of view.

Acknowledgement

We would like to thank Olavi Nevanlinna for stimulating discussions. Thanks are also due to Gene Golub for providing the excellent working conditions during our stay at Stanford University.

References

[1] K. Burrage and J.C. Butcher. *Stability criteria for implicit Runge-Kutta methods.* SIAM J. Numer. Anal., **16**(1):46-57, 1979.

[2] K. Burrage and J.C. Butcher. *Nonlinear stability of a general class of differential equation methods.* BIT **20**(2):185-203, 1980.

[3] J.C. Butcher. *A stability property of implicit Runge-Kutta methods.* BIT **15**:358-361, 1975.

[4] G. Dahlquist. *Error analysis for a class of methods of stiff non-linear initial value problems.* Numer. Anal., Proc. Dundee Conf. 1975, Lect. Notes Math. 506, 60-72, 1976.

[5] W. Gröbner. *Matrizenrechnung.* B.I. Hochschultaschenbücher, 103/103a. Bibliograpisches Institut, Mannheim.

[6] P. Henrici. *Discrete variable methods in ordinary differential equations.* Wiley, New York, 1962.

[7] M. Hyman. (Private communication), 1978.

[8] R. Jeltsch and O. Nevanlinna. *Largest disk of stability of explicit Runge-Kutta methods.* BIT **18**:500-502, 1978.

[9] L. Lapidus and J.H. Seinfeld. *Numerical solution of ordinary differential equations.* Academic Press, New York, 1971.

[10] O. Nevanlinna and W. Liniger. *Contractive methods for stiff differential equations. Part I*: BIT **18**:457-474, 1978. *Part II*: BIT **19**:53-72, 1979.

[11] J. Oliver. *A curiosity of low-order explicit Runge-Kutta methods.* Math. Comput., **29**(132):1032-1036, 1975.

[12] G. Wanner. *A short proof on nonlinear A-stability.* BIT **16**:226-227, 1976.