Applied Numerical Mathematics 5 (1989) 71–87
North-Holland

ALGEBRAIC STABILITY AND ERROR PROPAGATION IN RUNGE–KUTTA METHODS

J.F.B.M. KRAAIJEVANGER and M.N. SPIJKER
Department of Mathematics and Computer Science, University of Leiden, 2300 RA Leiden, The Netherlands

This paper is concerned with Runge–Kutta methods for the numerical solution of initial value problems in ordinary differential equations. For these methods we review the fundamental concept of algebraic stability (introduced in 1979 independently by Burrage and Butcher [1] and by Crouzeix [2]). We prove a new theorem implying that algebraic stability is a necessary and sufficient condition for a stable propagation of numerical errors.

Dahlquist and Jeltsch [6] introduced a generalized concept of algebraic stability and proved its equivalence to contractivity under the assumption that the Runge–Kutta methods are not confluent. Our theorem also implies this equivalence when the methods are confluent.

1. Introduction

1.1. Runge–Kutta methods

We deal with the numerical solution of the system of ordinary differential equations

$$\frac{\mathrm{d}}{\mathrm{d}t}U(t) = f(t, U(t)) \tag{1.1}$$

under an initial condition $U(0) = u_0$. Here u_0 is a given vector in the s-dimensional real vector space $\mathbb{R}^s$ and $f: \mathbb{R} \times \mathbb{R}^s \to \mathbb{R}^s$ denotes a given function. Further, $U(t) \in \mathbb{R}^s$ is unknown (for $t > 0$).

The general *Runge–Kutta method* for the approximation of $U(t)$ can be characterized by a so-called *coefficient scheme*

$$S = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mm} \\ b_1 & b_2 & \dots & b_m \end{pmatrix}.$$

Here a_{ij}, b_j are real constants specifying the method.

The method consists in generating approximations $u_1, u_2, u_3, \dots$ obtained by computing for $n = 1, 2, 3, \dots$

$$u_n = u_{n-1} + h_n \sum_{j=1}^{m} b_j f(t_{n-1} + c_j h_n, y_j). \tag{1.2a}$$

0168-9274/89/$3.50

Here y_j are vectors (depending on n) that satisfy the relations

$$y_i = u_{n-1} + h_n \sum_{j=1}^{m} a_{ij} f(t_{n-1} + c_j h_n,\ y_j), \quad i = 1, 2, \ldots, m. \tag{1.2b}$$

In the above $h_n > 0$ denotes the nth *step size* and $t_n = h_1 + h_2 + \cdots + h_n$, $u_n \simeq U(t_n)$. Further $c_j = a_{j1} + a_{j2} + \cdots + a_{jm}$ and m is called the *number of stages*.

In case the computations would start with a slightly perturbed initial vector $\tilde{u}_0$, instead of u_0, we would obtain $\tilde{u}_1$, $\tilde{u}_2$, $\tilde{u}_3, \ldots$ satisfying for $n = 1, 2, 3, \ldots$

$$\tilde{u}_n = \tilde{u}_{n-1} + h_n \sum_{j=1}^{m} b_j f(t_{n-1} + c_j h_n,\ \tilde{y}_j), \tag{1.3a}$$

$$\tilde{y}_i = \tilde{u}_{n-1} + h_n \sum_{j=1}^{m} a_{ij} f(t_{n-1} + c_j h_n,\ \tilde{y}_j), \quad i = 1, 2, \ldots, m. \tag{1.3b}$$

In this paper we are interested in conditions under which the errors $\tilde{u}_n - u_n$ do not grow unduly when n increases.

1.2. Contractivity

Assume

$$\langle f(t, x) - f(t, y),\ x - y \rangle \leqslant 0 \qquad \text{for } t \in \mathbb{R},\quad x, y \in \mathbb{R}^s. \tag{1.4}$$

Here $\langle \cdot, \cdot \rangle$ stands for an arbitrary inner product in $\mathbb{R}^s$, and $|\cdot|$ will denote the corresponding norm. As is well known (see e.g. [4,7]) assumption (1.4) implies that, for any two solutions U and $\tilde{U}$ to (1.1), the norm $|\tilde{U}(t) - U(t)|$ does not increase when t increases. It is natural to require that this property of the differential equation carries over to the numerical process. We thus arrive at the requirement

$$|\tilde{u}_n - u_n| \leqslant |\tilde{u}_{n-1} - u_{n-1}|, \quad n = 1, 2, 3, \ldots. \tag{1.5}$$

Let $\{h_n\}$ be a sequence of step sizes and f and $\langle \cdot, \cdot \rangle$ as above. Following the terminology of [6,7] we call the (method specified by the) coefficient scheme S *contractive with respect to* $\{h_n\}$, f and $\langle \cdot, \cdot \rangle$, if (1.5) holds whenever $\{u_n\}$ and $(\tilde{u}_n\}$ are two sequences satisfying (1.2) and (1.3).

In order to formulate a simple criterion for contractivity the following definitions are needed. We introduce the $m \times m$ matrices

$$B = \operatorname{diag}(b_1, b_2, \ldots, b_m), \qquad A = (a_{ij}),$$

and the column vector

$$b = (b_1, b_2, \ldots, b_m)^{\mathrm{T}}.$$

We call the coefficient scheme S *algebraically stable* if the two matrices

$$B \quad \text{and} \quad BA + A^{\mathrm{T}}B - bb^{\mathrm{T}}$$

are positive-semidefinite. Further, we call a coefficient scheme S with m stages *reducible* if there is a scheme S' with $m' < m$ stages generating the same approximations u_n as S. Conditions on S implying reducibility were given in [6,10] (cf. also [7]). In the following we shall deal with

coefficient schemes S that are *irreducible* in the sense that the conditions in [6] or those in [10] are violated.

The following theorem follows easily from the material presented in [1,2,10].

Theorem 1.1. *Let the coefficient scheme S be irreducible (in the sense of* [10]*). Then algebraic stability is necessary and sufficient for contractivity with respect to all $\{h_n\}$, f and $\langle \cdot, \cdot \rangle$ satisfying* (1.4).

1.3. Looking for stability instead of contractivity

We call the coefficient scheme S *stable with respect to* $\{h_n\}$, f and $\langle \cdot, \cdot \rangle$, if the effect of an arbitrary perturbation $\tilde{u}_0 - u_0$ can be bounded by

$$|\tilde{u}_n - u_n| \leqslant \gamma \cdot |\tilde{u}_0 - u_0|, \quad n = 1, 2, 3, \ldots. \tag{1.6}$$

Here γ stands for any constant independent of n (but possibly depending on $\{h_n\}$, f, $\langle \cdot, \cdot \rangle$, u_0, $\tilde{u}_0$).

Stability is a weaker property than contractivity, since (1.5) implies (1.6) with $\gamma = 1$. However, in many applications of contractivity the property of stability would do as well. For instance, if $\tilde{u}_0$ stands for a finite digit representation (in a computer) of the true u_0 then $\tilde{u}_0 - u_0$ stands for a rounding error. In this situation not only (1.5) but also (1.6) shows that the effect of the rounding error on the subsequent approximations is favourable in that this error does not grow unlimitedly during the calculations. Further, in case of stability with γ only depending on S (and not on $\{h_n\}$, f, $\langle \cdot, \cdot \rangle$, u_0, $\tilde{u}_0$) one can prove so-called B-convergence estimates (see [7,8]). In the literature such estimates have usually been derived by using (1.5), but (1.6) is already sufficient for this purpose (cf. [15]).

From Theorem 1.1 we see that algebraic stability is a sufficient condition for stability with respect to all $\{h_n\}$, f and $\langle \cdot, \cdot \rangle$ satisfying (1.4). But, within the class of schemes S that fail to be algebraically stable there are schemes with other favourable properties (e.g. requiring little work for computing u_n from u_{n-1}). Therefore the following question is of importance.

> Is it necessary for a coefficient scheme S to be algebraically stable if it is only required to be stable (instead of contractive) with respect to all $\{h_n\}$, f and $\langle \cdot, \cdot \rangle$ satisfying (1.4)?

An analysis of the well-known trapezoidal rule (specified by $m = 2$, $a_{11} = a_{12} = 0$, $a_{21} = a_{22} = b_1 = b_2 = \frac{1}{2}$) suggests that the answer to this question is positive. The corresponding scheme S fails to be algebraically stable. Further, an application of S to the initial value problem

$$\frac{\mathrm{d}}{\mathrm{d}t} U(t) = -20(1+t)^{-1}(U(t) - 1), \qquad U(0) = u_0 = 1$$

reveals severe instability. Choosing $h_n = 2^{n-1}$ we obtain from (1.2) the values $u_n = 1$ whereas (1.3) yields

$$\tilde{u}_n = 1 + (\tilde{u}_0 - 1)\lambda^n \quad \text{with } \lambda = -\tfrac{3}{2}$$

(cf. also [16, pp. 181–182]).

In the present paper we shall prove that the answer to the above question is positive indeed (cf. also [13, Theorem 4.2]).

As an illustration we consider the following example.

$$S = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -\frac{1}{2} & 1 & 0 & 0 \\ -1 & \frac{1}{2} & 1 & 0 \\ 2 & -2 & -1 & 1 \\ \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{pmatrix}. \tag{1.7}$$

This coefficient scheme is diagonally implicit (i.e. $a_{ij} = 0$ for all $i < j$) and therefore allows an efficient numerical solution of the system of equations (1.2b) (cf. e.g. [7]). The corresponding four-stage method has order of accuracy four (see e.g. [16]). Further, the coefficient scheme is irreducible (in the sense of [10]) and fails to be algebraically stable although it is strongly A-stable (i.e. $|(\tilde{u}_n - u_n)/(\tilde{u}_{n-1} - u_{n-1})| < 1$ and $\lim |(\tilde{u}_n - u_n)/(\tilde{u}_{n-1} - u_{n-1})| < 1$ (for $h_n \to \infty$) whenever $f(t, x) \equiv \lambda x$ where λ is a complex scalar with $\mathrm{Re}\ \lambda < 0$). From Theorem 1.1 we only can conclude that there is no contractivity with respect to all $\{h_n\}$, f and $\langle \cdot, \cdot \rangle$ satisfying (1.4). Using the results formulated in this article, it easily follows that the coefficient scheme (1.7) is not even stable with respect to all those $\{h_n\}$, f and $\langle \cdot, \cdot \rangle$.

1.4. Outline of the rest of this paper

In Section 2 we present the main results of this paper in such a fashion that they are readily accessible also for the reader who is not interested in the proofs.

In Section 2.1 we turn to a framework, due to Dahlquist and Jeltsch [6], in which functions f are considered that satisfy a generalized version of the inequality (1.4). In this framework these authors proved a generalized version of algebraic stability to be equivalent to contractivity (see (1.5)). Their proof is valid only for schemes S with $c_i \neq c_j$ (for all $i \neq j$). Our main Theorem 2.4, formulated in Section 2.1, implies that this generalized version of algebraic stability is not only equivalent to contractivity but also to stability (in the sense of (1.6)). Moreover this theorem is valid also for schemes S violating the above requirement $c_i \neq c_j$ (for $i \neq j$).

In Section 2.2 we include into our considerations the important question whether the systems of algebraic equations (1.2b) have unique solutions $y_1, y_2, \ldots, y_m$. Using Theorem 2.4 we give in Corollary 2.8 conditions under which the above mentioned generalized version of algebraic stability nicely becomes equivalent to the simultaneous presence of contractivity and of unique solutions to these systems of equations.

In Section 2.3 we discuss various modified versions of the theorems presented in Sections 2.1 and 2.2.

Finally, Section 3 is of a technical nature. In Section 3.1 we present a series of lemmata needed in Section 3.2. The latter section contains the key result, Theorem 3.8. Our proof of this theorem can be viewed as a nontrivial extension of the proof in [10] of the fact that contractivity implies algebraic stability. At the end of Section 3.2 we give a proof of Theorem 2.4 based on Theorem 3.8.

2. Formulation of the main results

2.1. Error propagation

Let $\langle \cdot, \cdot \rangle$ be a given inner product in $\mathbb{R}^s$ with corresponding norm $|\cdot|$. For arbitrary functions $f: \mathbb{R} \times \mathbb{R}^s \to \mathbb{R}^s$ we define the quantity $\nu[f] \in [-\infty, \infty]$ by

$$\nu[f] = \sup \frac{\langle f(t, x) - f(t, y), x - y \rangle}{|f(t, x) - f(t, y)|^2}.$$

Here the supremum is for all real t and all $x, y \in \mathbb{R}^s$ with $f(t, x) \neq f(t, y)$. Using the convention $\sup \emptyset = -\infty$ we see that assumption (1.4) is equivalent to

$$\nu[f] \leqslant 0.$$

Let S be a given coefficient scheme with A, B and b as in Section 1.2. Following ideas of [6] we focus on situations where, for some real $\rho \in (0, \infty)$,

$$B \text{ and } \left(BA + A^{\mathrm{T}}B - bb^{\mathrm{T}} + \rho^{-1}B\right) \text{ are positive-semidefinite.} \tag{2.1}$$

Definition 2.1. The *radius of algebraic stability* $r(S) \in [0, \infty]$ is given by

$$r(S) = \sup\{\rho \mid \rho = 0 \text{ or } (0 < \rho < \infty \text{ and } (2.1) \text{ holds})\}.$$

Clearly the coefficient scheme S is algebraically stable (as defined in Section 1.2) if and only if

$$r(S) = \infty.$$

The radius $r(S)$ can be computed by using the following lemma (see [6]).

Lemma 2.2. *Let the scheme S be irreducible (in the sense of [6]). Then $r(S) > 0$ if and only if $b_i > 0$ $(i = 1, 2, \ldots, m)$. In this case $r(S) = -\lambda^{-1}$ (if $\lambda < 0$), $r(S) = \infty$ (if $\lambda \geqslant 0$) where λ denotes the smallest eigenvalue of the matrix*

$$B^{-1/2}(BA + A^{\mathrm{T}}B - bb^{\mathrm{T}})B^{-1/2}.$$

In the next theorem we formulate a simple criterion for contractivity. In this theorem we refer, for given $\alpha \in (-\infty, 0]$ and $H \in (0, \infty]$, to the propositions:

(P1) $r(S) \geqslant (-2\alpha)^{-1} H$.

(P2) The scheme S is contractive with respect to all $\{h_n\}$, f and $\langle \cdot, \cdot \rangle$ satisfying $0 < h_n \leqslant H$ and $\nu[f] \leqslant \alpha$.

In case $\alpha = 0$ or $H = \infty$ the right-hand member in (P1) stands for ∞. Adapting the main result in [6] to the above definitions and notations one arrives at the following theorem.

Theorem 2.3. *Let S be a given coefficient scheme and let $\alpha \in (-\infty, 0]$, $H \in (0, \infty]$. Then* (P1) *implies* (P2). *In case $c_i \neq c_j$ (for $i \neq j$) proposition* (P2) *also implies* (P1).

This theorem gives rise to the question whether the restriction to the case $c_i \neq c_j$ (for $i \neq j$) in the second statement of the theorem is essential. Moreover the question arises whether it is

necessary for a scheme S (with $c_i \neq c_j$ (for $i \neq j$)) to satisfy (P1) if it is only required to have the following property (P3) (which is weaker than (P2)).

(P3) The scheme S is stable with respect to all $\{h_n\}$, f and $\langle \cdot, \cdot \rangle$ where $0 < h_n < H$, f is continuous and $\nu[f] < \alpha$.

Both of the above questions are answered by the following theorem constituting the main result of this paper.

Theorem 2.4. *Let $\alpha \in (-\infty, 0]$ and $H \in (0, \infty]$ be given and let the coefficient scheme S be irreducible (in the sense of* [10]*). Then* (P3) *implies* (P1).

A combination of Theorems 2.3 and 2.4 yields:

Corollary 2.5. *Let $\alpha \in (-\infty, 0]$ and $H \in (0, \infty]$ be given and let S be irreducible (in the sense of* [10]*). Then the algebraic condition* (P1)*, the contractivity property* (P2) *and the stability property* (P3) *are equivalent to each other.*

Choosing $\alpha = 0$ and $H = \infty$ the above implies:

Corollary 2.6. *Let S be irreducible (in the sense of* [10]*). Then the following three propositions are equivalent*:

(p1) *S is algebraically stable.*

(p2) *S is contractive with respect to all $\{h_n\}$, f and $\langle \cdot, \cdot \rangle$ satisfying* (1.4).

(p3) *S is stable with respect to all $\{h_n\}$, f and $\langle \cdot, \cdot \rangle$ satisfying* (1.4).

The last corollary answers the question raised in Section 1.3.

2.2. Solvability of the algebraic equations

For given $\alpha \in (-\infty, 0]$ and $H \in (0, \infty]$ one might conjecture that condition (P1) guarantees the existence of a (unique) solution $y_1, y_2, \ldots, y_m$ to (1.2b) provided $0 < h_n \leqslant H$, $\nu[f] \leqslant \alpha$ and f is continuous. But, in [3] a counterexample was constructed showing this conjecture to be false in case $\alpha = 0$ and $H \in (0, \infty]$. Without proof we mention that this construction can be adapted so as to yield a counterexample also when $\alpha < 0$ and $H \in (0, \infty)$. Only in the exceptional case when $\alpha < 0$ and $H = \infty$ it follows from [14] or [11] that the above conjecture is true (provided S is irreducible [6]).

It is clear that contractivity in the absence of solutions to equations (1.2b) makes little sense. Therefore we look for conditions under which (P1) is still equivalent to contractivity accompanied by existence and uniqueness for equations (1.2b).

The following lemma will be helpful.

Lemma 2.7. *Let $\alpha \in (-\infty, 0]$ and $H \in (0, \infty]$ be given and let the coefficient scheme S be irreducible (in the sense of* [6]*). Let $f: \mathbb{R} \times \mathbb{R}^s \to \mathbb{R}^s$ be continuous with $\nu[f] < \alpha$, and $u_{n-1} \in \mathbb{R}^s$, $0 < h_n \leqslant H$. Then condition* (P1) *guarantees that the system of equations* (1.2b) *has a unique solution $y_1, y_2, \ldots, y_m$.*

Proof. From (P1) it follows that $r(S) > 0$. By Lemma 2.2 all b_i are thus positive. An application of [11, Corollary 2.4] (with $h = h_n$, $D = B$ and $\beta = \lambda = 0$) completes the proof. □

Combining Lemma 2.7 with Corollary 2.5 we arrive at the following final corollary.

Corollary 2.8. *Let $\alpha \in (-\infty, 0]$ and $H \in (0, \infty]$ be given and let the scheme S be irreducible (both in the sense of* [6] *and of* [10]). *Then condition* (P1) *is necessary and sufficient for the following property* (P4):

(P4) *The scheme S is contractive with respect to all $\{h_n\}$, f and $\langle \cdot, \cdot \rangle$ where $0 < h_n \leqslant H$, f is continuous and $\nu[f] < \alpha$. Moreover for all such h_n and f the systems* (1.2b) *have unique solutions.*

2.3. Modifications and counterexamples

(1) In [3,6] functions $f: \mathbb{R} \times \mathbb{C}^s \to \mathbb{C}^s$ were considered instead of $f: \mathbb{R} \times \mathbb{R}^s \to \mathbb{R}^s$. In view of the usual isomorphism between $\mathbb{C}$ and $\mathbb{R}^2$ this discrepancy is not essential. Therefore the results of [3,6] used in the Sections 2.1 and 2.2 are applicable indeed. Conversely, our theory presented above could have been formulated as well in a complex framework.

(2) In [6] contractivity was studied also for functions f with $\nu[f] > 0$. We have restricted our considerations to the case $\nu[f] \leqslant 0$ for several reasons. First of all we have not been able to prove a full analogue of Theorem 2.4 with $\alpha > 0$. Further, the presentation in the Sections 2.1 and 2.2 would have become considerably more complicated. Moreover the requirement of contractivity would have become unnatural since no continuous analogue of this property need be present in the differential equation (1.1). Finally we would have been obliged to deal with step size restrictions of type $h_n \geqslant H > 0$ (cf. [6]) which look unrealistic.

(3) Proposition (P3) occurring in Theorem 2.4 concerns possibly *variable step sizes* $h_n \in (0, H)$. Theorem 2.4 does not remain true if proposition (P3) would be weakened by allowing only *constant step sizes* $h_n = h \in (0, H)$ (for $n = 1, 2, 3, \ldots$). A counterexample is provided by the trapezoidal rule. Choosing $\alpha = 0$, $H = \infty$ and $h_n = h \in (0, \infty)$ it can be proved (cf. [5,12]) that (1.6) holds with

$$\gamma = |\tilde{u}_0 - u_0|^{-1} \cdot |\tilde{u}_0 - u_0 + \tfrac{1}{2}h[f(t_0, \tilde{u}_0) - f(t_0, u_0)]|.$$

The weaker version of (P3) thus holds, but (P1) (with $(-2\alpha)^{-1}H = \infty$) is violated since the trapezoidal rule is not algebraically stable.

(4) The above counterexample leaves the question unsettled whether Theorem 2.4 remains true if (P3) is modified by allowing only $h_n = h \in (0, H)$ and simultaneously requiring *stability with γ in* (1.6) *only depending on S* (and not on h, f, $\langle \cdot, \cdot \rangle$, u_0, $\tilde{u}_0$).

We give a counterexample showing that the answer to this question is negative.

Let f be continuous, $\nu[f] \leqslant \alpha = 0$, $H = \infty$, $h_n = h \in (0, \infty)$ and

$$S_0 = \begin{pmatrix} \frac{1}{2} & 0 \\ -\frac{1}{2} & 2 \\ -\frac{1}{2} & \frac{3}{2} \end{pmatrix}.$$

Since $b_1 = -\frac{1}{2}$ there is no algebraic stability so that (P1) is violated. Further S_0 is irreducible (in the sense of [10]).

Let the sequence $\{u_n\}$ be generated by an application of S_0. Defining mappings $G_0(t): \mathbb{R}^s \to \mathbb{R}^s$ such that $u_n = G_0(t_{n-1})u_{n-1}$ we have

$$u_n = G_0(t_{n-1}) \cdot \cdots \cdot G_0(t_1)G_0(t_0)u_0.$$

An easy calculation shows that $G_0(t) = G_2(t - \frac{1}{2}h)G_1(t)$ where $G_2(t)$ and $G_1(t)$ are mappings corresponding to the coefficient schemes

$$S_2 = \begin{pmatrix} 2 \\ \frac{3}{2} \end{pmatrix} \quad \text{and} \quad S_1 = \begin{pmatrix} \frac{1}{2} \\ -\frac{1}{2} \end{pmatrix},$$

respectively. Further $G_1(t + \frac{3}{2}h)G_2(t) = G_3(t)$ where $G_3(t)$ corresponds to the coefficient scheme

$$S_3 = \begin{pmatrix} 2 & 0 \\ \frac{3}{2} & \frac{1}{2} \\ \frac{3}{2} & -\frac{1}{2} \end{pmatrix}.$$

Writing $G_i(\lambda h) = G_{i,\lambda}$ we thus have

$$u_n = G_{2,n-3/2}G_{3,n-5/2} \cdot \cdots \cdot G_{3,3/2}G_{3,1/2}G_{3,-1/2}G_{1,0}u_0.$$

Suppose $v = G_3(t)u$. Then

$$\begin{aligned} v &= u + \tfrac{3}{2}hf(t + 2h,\ y_1) - \tfrac{1}{2}hf(t + 2h,\ y_2), \\ y_1 &= u + 2hf(t + 2h,\ y_1), \\ y_2 &= u + \tfrac{3}{2}hf(t + 2h,\ y_1) + \tfrac{1}{2}hf(t + 2h,\ y_2). \end{aligned}$$

Using $v[f] \leqslant 0$ we obtain

$$\langle y_1 - y_2,\ y_1 - y_2 \rangle = \tfrac{1}{2}h\langle f(t + 2h,\ y_1) - f(t + 2h,\ y_2),\ y_1 - y_2 \rangle \leqslant 0.$$

Hence $y_1 = y_2$ so that $v = G_4(t)u$ where the mapping $G_4(t)$ corresponds to

$$S_4 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}.$$

We thus arrive at the following expression for u_n,

$$u_n = G_{2,n-3/2}G_{4,n-5/2} \cdot \cdots \cdot G_{4,3/2}G_{4,1/2}G_{4,-1/2}G_{1,0}u_0.$$

Since S_2 and S_4 are algebraically stable the corresponding mappings $G_2(t)$ and $G_4(t)$ satisfy a Lipschitz condition on $\mathbb{R}^s$ with Lipschitz constant $L = 1$. The scheme S_1 is not algebraically stable but $G_1(t)$ satisfies a Lipschitz condition with $L = 2$. This can be seen from the fact that $G_1(t)u \equiv \frac{1}{2}(3u - G_5(t)u)$ where $G_5(t)$ corresponds to the algebraically stable scheme

$$S_5 = \begin{pmatrix} \frac{1}{2} \\ 1 \end{pmatrix}.$$

The above expression for u_n, combined with a similar expression for $\tilde{u}_n$, thus shows that

$$|\tilde{u}_n - u_n| \leqslant 2|\tilde{u}_0 - u_0|.$$

We have proved stability of S_0 with $\gamma = 2$ independent of h, f, $\langle \cdot, \cdot \rangle$, u_0, $\tilde{u}_0$.

(5) In our definition of stability (in Section 1.3) the constant γ is allowed to depend on (the Lipschitz constant of) the function f. This dependence causes the property of stability to be of little value in the case of so-called stiff problems (1.1) (i.e. problems with a large Lipschitz constant, cf. [7,8,13,16]). However, the fact that we have adopted this weak version of stability makes our Theorem 2.4 a strong one. Evidently, Theorem 2.4 (as well as Corollaries 2.5 and 2.6) remain valid if property (P3) would refer to a stronger stability concept in which γ is allowed to depend on S (and not on $\{h_n\}$, f, $\langle \cdot, \cdot \rangle$, u_0, $\tilde{u}_0$).

3. The proof of Theorem 2.4

3.1. Technical lemmata

In this section S denotes a given coefficient scheme with m stages.

We shall denote by Z arbitrary diagonal matrices of the form $Z = \operatorname{diag}(z_1, z_2, \ldots, z_m)$ with complex z_j. Further I denotes the identity matrix of order m, and e stands for the column vector in $\mathbb{R}^m$ all of whose components are equal to 1.

In the following the set $\mathcal{D}$ stands for

$$\mathcal{D} = \{ Z \mid \operatorname{Re}(z_j) < 0 \ (1 \leqslant j \leqslant m) \text{ and } (I - AZ) \text{ is nonsingular} \}.$$

For arbitrary $Z \in \mathcal{D}$ we define

$$K(Z) = 1 + b^{\mathrm{T}} Z (I - AZ)^{-1} e$$

and

$$L(Z) = \max\{ -|z_j|^2 / \operatorname{Re}(z_j) \mid 1 \leqslant j \leqslant m \}.$$

The following lemma can easily be proved by using the material in [6].

Lemma 3.1. *Let $\rho \in (0, \infty]$ be given. Assume that, for all $Z \in \mathcal{D}$ with $L(Z) < 2\rho$, we have $|K(Z)| \leqslant 1$. Then $r(S) \geqslant \rho$.*

We further need the following extension, due to [9], of a result already given in [10].

Lemma 3.2 (Hundsdorfer). *Let S be irreducible (in the sense of* [10]*) and $\gamma \geqslant 0$. Then there exist vectors $x = (\xi_1, \xi_2, \ldots, \xi_m)^{\mathrm{T}}$ and $y = (\eta_1, \eta_2, \ldots, \eta_m)^{\mathrm{T}}$ in $\mathbb{R}^m$ such that $y = Ax$ and*

$$\xi_i \neq \xi_j, \quad (\eta_i - \eta_j)/(\xi_i - \xi_j) < -\gamma \qquad \text{for } 1 \leqslant i < j \leqslant m.$$

Proof. In [10, pp. 328–331] an algorithm is presented for the construction of vectors $x = (\xi_1, \xi_2, \ldots, \xi_m)^{\mathrm{T}}$ and $y = (\eta_1, \eta_2, \ldots, \eta_m)^{\mathrm{T}}$ in $\mathbb{R}^m$ with

$$y = Ax, \qquad (\xi_i - \xi_j)(\eta_i - \eta_j) < 0 \quad \text{for } 1 \leqslant i < j \leqslant m \tag{3.1}$$

under the assumption that S is irreducible. This algorithm produces a finite sequence of triples $(x^{(n)}, y^{(n)}, K^{(n)})$ $(n = 0, 1, 2, \ldots, N)$ with

$$x^{(n)} \in \mathbb{R}^m, \qquad y^{(n)} \in \mathbb{R}^m, \qquad y^{(n)} = Ax^{(n)} \quad \text{and} \quad K^{(n)} \subset \{1, 2, \ldots, m\}.$$

The initial triple $(x, y, K) = (x^{(0)}, y^{(0)}, K^{(0)})$ is defined by

$$x = e, \qquad y = Ae, \qquad K = \{1, 2, \ldots, m\},$$

and therefore it obviously has the property

$$(\eta_i - \eta_j)/(\xi_i - \xi_j) < -\gamma \quad \text{for all } i, j \text{ with } \xi_i \neq \xi_j. \tag{3.2}$$

In each step of the algorithm the triple $(x^{(n)}, y^{(n)}, K^{(n)})$ is transformed into the triple $(x^{(n+1)}, y^{(n+1)}, K^{(n+1)})$. Each such transformation depends on a small parameter $\delta = \delta_n \in (0, 1)$ which has to be chosen sufficiently small. Slightly modifying some of the arguments presented in [10] it is easily seen that, for δ_n small enough, property (3.2) remains valid for the transformed triple

$$(x, y, K) = \left(x^{(n+1)}, y^{(n+1)}, K^{(n+1)}\right), \quad n = 0, 1, 2, \ldots, N-1.$$

Consequently property (3.2) also holds for $x = x^{(N)}$ and $y = y^{(N)}$. Since it is shown in [10] that $x = x^{(N)}$ and $y = y^{(N)}$ also satisfy (3.1), the proof of the lemma immediately follows. □

In the following $\mathcal{F}(\alpha)$ denotes the class of all mappings $\psi : \mathbb{C} \to \mathbb{C}$ with

$$\operatorname{Re}\left[(\psi(x) - \psi(y))\overline{(x - y)}\right] \leqslant \alpha |\psi(x) - \psi(y)|^2 \quad \text{for all } x, y \in \mathbb{C}.$$

We present three technical lemmata concerning the class $\mathcal{F}(\alpha)$.

Lemma 3.3. *Let $\psi \in \mathcal{F}(\alpha)$ with $\alpha < 0$. Then ψ satisfies a Lipschitz condition,*

$$|\psi(x) - \psi(y)| \leqslant (-1/\alpha)|x - y| \quad \textit{for all } x, y \in \mathbb{C}.$$

Proof. Use $\operatorname{Re}[(\psi(x) - \psi(y))\overline{(x - y)}] \geqslant -|\psi(x) - \psi(y)| \cdot |x - y|$ (cf. also [6]). □

Lemma 3.4. *Let $\psi_1, \psi_2 \in \mathcal{F}(\alpha)$ with $\alpha \leqslant 0$. Then, for any $\lambda \in [0, 1]$, also*

$$\psi = \lambda\psi_1 + (1 - \lambda)\psi_2 \in \mathcal{F}(\alpha).$$

Proof. Let $x, y \in \mathbb{C}$ and put $z = x - y$ and $w_i = \psi_i(x) - \psi_i(y)$ $(i = 1, 2)$. Then we have

$$\begin{aligned}
&\operatorname{Re}\left[(\psi(x) - \psi(y))\overline{(x - y)}\right] - \alpha|\psi(x) - \psi(y)|^2 \\
&\quad = \lambda \operatorname{Re}(w_1\bar{z}) + (1 - \lambda) \operatorname{Re}(w_2\bar{z}) \\
&\qquad - \alpha\left[\lambda^2 |w_1|^2 + (1 - \lambda)^2 |w_2|^2 + 2\lambda(1 - \lambda) \operatorname{Re}(w_1\bar{w}_2)\right] \\
&\quad \leqslant \alpha\lambda(1 - \lambda)\left[|w_1|^2 + |w_2|^2 - 2 \operatorname{Re}(w_1\bar{w}_2)\right] \\
&\quad = \alpha\lambda(1 - \lambda)|w_1 - w_2|^2 \leqslant 0. \qquad \square
\end{aligned}$$

Lemma 3.5. *Let $V \subset \mathbb{C}$ and $\psi_0 : V \to \mathbb{C}$ be such that*

$$\operatorname{Re}\left[(\psi_0(x) - \psi_0(y))\overline{(x - y)}\right] \leqslant \alpha|\psi_0(x) - \psi_0(y)|^2 \quad \textit{for all } x, y \in V$$

for some $\alpha < 0$. Then there exists an extension $\psi \in \mathcal{F}(\alpha)$ of ψ_0.

Proof. Note that $\mathbb{C}$ can be seen as a real Hilbert space with inner product $\langle x, y\rangle = \mathrm{Re}(x\bar{y})$. The corresponding norm $\|x\| = \langle x, x\rangle^{1/2}$ is equal to the modulus $|x|$. Hence the assumption on ψ_0 can be written as

$$\langle \psi_0(x) - \psi_0(y), x - y\rangle \leqslant \alpha \|\psi_0(x) - \psi_0(y)\|^2 \quad \text{for all } x, y \in V,$$

and the proof of the lemma immediately follows from [17, Theorem 4.3]. □

The following two lemmata will be essential in the proof of Theorem 3.8.

Lemma 3.6. *Let S be irreducible (in the sense of* [10]*), and $Z \in \mathcal{D}$. Let $\varepsilon > 0$, $L = L(Z) + \varepsilon$ and $q_0, \tilde{q}_0 \in \mathbb{C}$ with $q_0 \neq \tilde{q}_0$. Then a real $\delta > 0$ and a mapping $\psi \in \mathcal{F}(-1/L)$ exist such that for all complex q and $\tilde{q}$ with $|q - q_0| \leqslant \delta$ and $|\tilde{q} - \tilde{q}_0| \leqslant \delta$ there are p_i and $\tilde{p}_i$ satisfying*

$$p_i = q + \sum_{j=1}^{m} a_{ij}\psi(p_j), \quad \tilde{p}_i = \tilde{q} + \sum_{j=1}^{m} a_{ij}\psi(\tilde{p}_j), \qquad 1 \leqslant i \leqslant m. \tag{3.3}$$

Further, the corresponding quantities

$$q_1 = q + \sum_{j=1}^{m} b_j\psi(p_j) \quad \text{and} \quad \tilde{q}_1 = \tilde{q} + \sum_{j=1}^{m} b_j\psi(\tilde{p}_j)$$

satisfy

$$|\tilde{q}_1 - q_1| \geqslant (|K(Z)| - \varepsilon) \cdot |\tilde{q}_0 - q_0|. \tag{3.4}$$

Proof. (1) Let x and y be as in Lemma 3.2 with $\gamma = 1/L$. From the definition of $K(Z)$ we have

$$K(Z) = 1 + \sum_{j=1}^{m} b_j z_j s_j, \qquad s_i = 1 + \sum_{j=1}^{m} a_{ij} z_j s_j, \quad 1 \leqslant i \leqslant m.$$

Let δ and t be positive parameters to be specified below and define $p_{0,i}$ and $\tilde{p}_{0,i}$ by

$$p_{0,i} = t\eta_i + q_0, \quad \tilde{p}_{0,i} = p_{0,i} + (\tilde{q}_0 - q_0)s_i, \qquad 1 \leqslant i \leqslant m.$$

Next we introduce complex disks B_i and $\tilde{B}_i$ by

$$B_i = \{\zeta \mid \zeta \in \mathbb{C}, |\zeta - p_{0,i}| \leqslant \delta\}, \quad \tilde{B}_i = \{\zeta \mid \zeta \in \mathbb{C}, |\zeta - \tilde{p}_{0,i}| \leqslant \delta\}, \qquad 1 \leqslant i \leqslant m.$$

Note that $B_i = \tilde{B}_i$ if $s_i = 0$. We assume that δ is small enough and t large enough to imply

$$B_i \cap B_j = \tilde{B}_i \cap \tilde{B}_j = B_i \cap \tilde{B}_j = \emptyset \quad \text{for all } i, j \text{ with } i \neq j$$

and

$$B_i \cap \tilde{B}_i = \emptyset \quad \text{for all } i \text{ with } s_i \neq 0.$$

On the set $V = \bigcup_{i=1}^{m}(B_i \cup \tilde{B}_i)$ we define a mapping $\psi_0 : V \to \mathbb{C}$ as follows. For all $q, \tilde{q} \in \mathbb{C}$ with $|q - q_0| \leqslant \delta$ and $|\tilde{q} - \tilde{q}_0| \leqslant \delta$ we define

$$p_i = p_{0,i} + (q - q_0), \quad \psi_0(p_i) = t\xi_i, \qquad 1 \leqslant i \leqslant m, \tag{3.5a}$$

$$\tilde{p}_i = \tilde{p}_{0,i} + (\tilde{q} - \tilde{q}_0), \quad \psi_0(\tilde{p}_i) = t\xi_i + (\tilde{q}_0 - q_0)z_i s_i, \qquad 1 \leqslant i \leqslant m. \tag{3.5b}$$

A straightforward calculation shows that

$$\mathrm{Re}\big[(\psi_0(x) - \psi_0(y))\overline{(x - y)}\big] \leqslant (-1/L)\,|\psi_0(x) - \psi_0(y)|^2 \quad \text{for all } x, y \in V$$

provided δ is small enough and t large enough. In view of Lemma 3.5 there exists an extension $\psi \in \mathcal{F}(-1/L)$ of ψ_0.

(2) Let $q, \tilde{q} \in \mathbb{C}$ be given with $|q - q_0| \leqslant \delta$ and $|\tilde{q} - \tilde{q}_0| \leqslant \delta$. With the definition of p_i and $\tilde{p}_i$ in (3.5) it is easily verified that condition (3.3) is fulfilled. Defining q_1 and $\tilde{q}_1$ as stated in the lemma we finally have

$$\tilde{q}_1 - q_1 = K(Z) \cdot (\tilde{q}_0 - q_0) + (\tilde{q} - \tilde{q}_0) - (q - q_0).$$

Consequently, (3.4) holds provided δ is chosen so small that also $2\delta \leqslant \varepsilon |\tilde{q}_0 - q_0|$. □

Lemma 3.7. *Let positive real numbers θ and L, a complex number q_0^*, and for $j = 1, 2, \ldots, m$ a mapping $\phi_j^* : \mathbb{C} \to \mathbb{C}$ be given with*

$$|\phi_j^*(x) - \phi_j^*(y)| \leqslant L|x - y| \quad \text{for all } x, y \in \mathbb{C}.$$

If

$$\theta L \max_i \sum_{j=1}^{m} |a_{ij}| \leqslant \tfrac{1}{2},$$

then the system of equations

$$x_i = q_0^* + \theta \sum_{j=1}^{m} a_{ij} \phi_j^*(x_j), \quad 1 \leqslant i \leqslant m$$

has a unique solution $(x_1, x_2, \ldots, x_m)^{\mathrm{T}} \in \mathbb{C}^m$ and the corresponding quantity

$$q_1^* = q_0^* + \theta \sum_{j=1}^{m} b_j \phi_j^*(x_j)$$

satisfies

$$|q_1^* - q_0^*| \leqslant 2\theta\mu[\,|b_1| + |b_2| + \cdots + |b_m|\,]$$

with $\mu = \max_j |\phi_j^*(q_0^*)|$.

Proof. The above system of equations can be rewritten as $X = G(X)$ where $X = (x_1, x_2, \ldots, x_m)^{\mathrm{T}}$ and $G : \mathbb{C}^m \to \mathbb{C}^m$ is defined as

$$G_i(X) = q_0^* + \theta \sum_{j=1}^{m} a_{ij} \phi_j^*(x_j), \quad 1 \leqslant i \leqslant m.$$

Introducing the maximum norm $|X|_\infty = \max_j |x_j|$, it is easily seen that G is a contraction,

$$|G(X) - G(Y)|_\infty \leqslant \tfrac{1}{2} |X - Y|_\infty \quad \text{for all } X, Y \in \mathbb{C}^m.$$

In view of the contraction mapping theorem it follows that G has a unique fixed point $X = (x_1, x_2, \ldots, x_m)^{\mathrm{T}}$. Note that for $i = 1, 2, \ldots, m$

$$\begin{aligned} |\phi_i^*(x_i)| &\leqslant |\phi_i^*(q_0^*)| + L|x_i - q_0^*| \\ &\leqslant \mu + \theta L \sum_{j=1}^{m} |a_{ij} \phi_j^*(x_j)| \leqslant \mu + \tfrac{1}{2} \max_j |\phi_j^*(x_j)|. \end{aligned}$$

Hence

$$|\phi_j^*(x_j)| \leqslant 2\mu, \quad 1 \leqslant j \leqslant m,$$

and the above inequality for $|q_1^* - q_0^*|$ immediately follows. □

3.2. Proving the main result

Theorem 3.8. *Let S be a given irreducible* [10] *Runge–Kutta scheme, and $Z \in \mathscr{D}$. Let $0 < \varepsilon < |K(Z)|$, $L = L(Z) + \varepsilon$ and $w_0, \tilde{w}_0 \in \mathbb{C}$ with $w_0 \neq \tilde{w}_0$. Then a sequence $\theta_1, \theta_2, \theta_3, \ldots$ and a continuous function $\phi: \mathbb{R} \times \mathbb{C} \to \mathbb{C}$ exist with*

$$0 < \theta_n \leqslant 1, \quad \phi(\tau, \cdot) \in \mathscr{F}(-1/L) \qquad \text{for all } n \geqslant 1, \quad \tau \in \mathbb{R}$$

such that there are complex $v_j^{(n)}$, $\tilde{v}_j^{(n)}$, w_n and $\tilde{w}_n$ satisfying for all $n \geqslant 1$

$$v_i^{(n)} = w_{n-1} + \theta_n \sum_{j=1}^{m} a_{ij} \phi\left(\tau_{n-1} + c_j\theta_n, v_j^{(n)}\right), \quad 1 \leqslant i \leqslant m, \tag{3.6a}$$

$$\tilde{v}_i^{(n)} = \tilde{w}_{n-1} + \theta_n \sum_{j=1}^{m} a_{ij} \phi\left(\tau_{n-1} + c_j\theta_n, \tilde{v}_j^{(n)}\right), \quad 1 \leqslant i \leqslant m, \tag{3.6b}$$

$$w_n = w_{n-1} + \theta_n \sum_{j=1}^{m} b_j \phi\left(\tau_{n-1} + c_j\theta_n, v_j^{(n)}\right), \tag{3.7a}$$

$$\tilde{w}_n = \tilde{w}_{n-1} + \theta_n \sum_{j=1}^{m} b_j \phi\left(\tau_{n-1} + c_j\theta_n, \tilde{v}_j^{(n)}\right), \tag{3.7b}$$

and for all even $n \geqslant 2$

$$|\tilde{w}_n - w_n| \geqslant (|K(Z)| - \varepsilon)^{n/2} \cdot |\tilde{w}_0 - w_0|. \tag{3.8}$$

Here we use the notation $\tau_{n-1} = \theta_1 + \theta_2 + \cdots + \theta_{n-1}$.

Proof. (1) Let S, Z, ε, L, w_0 and $\tilde{w}_0$ be as in the assumptions of the theorem. We define $\rho = (|K(Z)| - \varepsilon)^{1/2}$. For arbitrary integers $k \geqslant 1$ we will write T_k for the set

$$\{\tau \mid \tau = \tau_{k-1} + c_j\theta_k, 1 \leqslant j \leqslant m\}.$$

Further we put $T_\infty = \bigcup_{k=1}^{\infty} T_k$. In part (2) of the proof we will construct a sequence of mappings $\phi_0 \subset \phi_2 \subset \phi_4 \subset \phi_6 \subset \cdots$, a sequence of step sizes $\theta_1, \theta_2, \theta_3, \ldots$ (with $\theta_k = 1$ if k is even, and $0 < \theta_k \leqslant 1$ if k is odd) and complex numbers w_k, $\tilde{w}_k$, $v_j^{(k)}$, $\tilde{v}_j^{(k)}$ $(1 \leqslant j \leqslant m,\ k \geqslant 1)$ such that for $n = 0, 2, 4, 6, \ldots$ we have the properties (3.9)–(3.12):

$$\phi_n: \left(\bigcup_{k=1}^{n} T_k\right) \times \mathbb{C} \to \mathbb{C} \text{ satisfies } \phi_n(\tau, \cdot) \in \mathscr{F}(-1/L) \text{ for all } \tau \in \bigcup_{k=1}^{n} T_k, \tag{3.9}$$

$$v_i^{(k)} = w_{k-1} + \theta_k \sum_{j=1}^{m} a_{ij} \phi_n\left(\tau_{k-1} + c_j\theta_k, v_j^{(k)}\right), \quad 1 \leqslant i \leqslant m, \quad 1 \leqslant k \leqslant n, \tag{3.10a}$$

$$\tilde{v}_i^{(k)} = \tilde{w}_{k-1} + \theta_k \sum_{j=1}^{m} a_{ij} \phi_n\left(\tau_{k-1} + c_j\theta_k, \tilde{v}_j^{(k)}\right), \quad 1 \leqslant i \leqslant m, \quad 1 \leqslant k \leqslant n, \tag{3.10b}$$

$$w_k = w_{k-1} + \theta_k \sum_{j=1}^{m} b_j \phi_n\left(\tau_{k-1} + c_j\theta_k,\ v_j^{(k)}\right), \qquad 1 \leqslant k \leqslant n, \tag{3.11a}$$

$$\tilde{w}_k = \tilde{w}_{k-1} + \theta_k \sum_{j=1}^{m} b_j \phi_n\left(\tau_{k-1} + c_j\theta_k,\ \tilde{v}_j^{(k)}\right), \qquad 1 \leqslant k \leqslant n, \tag{3.11b}$$

$$|\tilde{w}_k - w_k| \geqslant \rho^k |\tilde{w}_0 - w_0|, \qquad k = 0, 2, 4, \ldots, n. \tag{3.12}$$

We define the mapping $\phi_\infty : T_\infty \times \mathbb{C} \to \mathbb{C}$ as $\phi_\infty = \phi_0 \cup \phi_2 \cup \phi_4 \cup \cdots$. As $\theta_k = 1$ (for all even $k \geqslant 2$) we have $T_\infty = \{\tau_1', \tau_2', \tau_3', \ldots\}$ with $\tau_1' < \tau_2' < \tau_3' < \cdots$ and $\lim_{k\to\infty} \tau_k' = \infty$. Extending ϕ_∞ to a mapping $\phi : \mathbb{R} \times \mathbb{C} \to \mathbb{C}$ by linear interpolation and by $\phi(\tau, x) = \phi_\infty(\tau_1', x)$ (for $\tau < \tau_1'$ and $x \in \mathbb{C}$), it follows from (3.9) and Lemmata 3.3 and 3.4 that ϕ is a continuous function with $\phi(\tau, \cdot) \in \mathcal{F}(-1/L)$ (for all $\tau \in \mathbb{R}$). Further, in view of (3.10)–(3.12) we have the relations (3.6)–(3.8).

(2) In this part we construct a sequence of mappings $\phi_0 \subset \phi_2 \subset \phi_4 \subset \cdots$, a sequence of step sizes $\theta_1, \theta_2, \theta_3, \ldots$ and complex numbers w_k, $\tilde{w}_k$, $v_j^{(k)}$, $\tilde{v}_j^{(k)}$ $(1 \leqslant j \leqslant m,\ k \geqslant 1)$ such that for $n = 0, 2, 4, \ldots$ the relations (3.9)–(3.12) are valid. We define ϕ_0 to be the function with empty domain. Obviously, with this definition the relations (3.9)–(3.12) are fulfilled with $n = 0$.

Next, suppose that an even integer $N \geqslant 0$ is given and that a mapping ϕ_N, a sequence of step sizes $\theta_1, \theta_2, \ldots, \theta_N$ and complex numbers w_k, $\tilde{w}_k$, $v_j^{(k)}$, $\tilde{v}_j^{(k)}$ $(1 \leqslant j \leqslant m,\ 1 \leqslant k \leqslant N)$ are already defined such that (3.9)–(3.12) hold with $n = N$. In the following we will prove that there exist an extension ϕ_{N+2} of ϕ_N, step sizes θ_{N+1} and θ_{N+2} and complex numbers w_k, $\tilde{w}_k$, $v_j^{(k)}$, $\tilde{v}_j^{(k)}$ $(1 \leqslant j \leqslant m,\ N+1 \leqslant k \leqslant N+2)$ such that (3.9)–(3.12) also hold with $n = N+2$.

Let $\delta > 0$ and $\psi \in \mathcal{F}(-1/L)$ be given by Lemma 3.6 (with $q_0 = w_N$ and $\tilde{q}_0 = \tilde{w}_N$). We define $\theta_{N+1} = \theta$ and $\theta_{N+2} = 1$, where $\theta \in (0, 1]$ will be specified below. Further we define the mapping

$$\phi_{N+2} : \left(\bigcup_{k=1}^{N+2} T_k\right) \times \mathbb{C} \to \mathbb{C}$$

by

$$\phi_{N+2}(\tau, x) = \begin{cases} \phi_N(\tau, x), & \text{if } \tau \in \bigcup_{k=1}^{N} T_k \text{ and } x \in \mathbb{C}, \\ \psi(x), & \text{if } \tau \in (T_{N+1} \cup T_{N+2}) \setminus \left(\bigcup_{k=1}^{N} T_k\right) \text{ and } x \in \mathbb{C}. \end{cases}$$

Clearly ϕ_{N+2} is an extension of ϕ_N satisfying (3.9) with $n = N+2$. Choosing θ so small that

$$\theta L \max_i \sum_{j=1}^{m} |a_{ij}| \leqslant \tfrac{1}{2}$$

and

$$2\theta\left[\sum_{j=1}^{m} |b_j|\right] \cdot \max\left\{ |\phi_{N+2}(\tau, x)| \,|\, \tau \in \bigcup_{k=1}^{N+2} T_k,\ x \in \{w_N, \tilde{w}_N\} \right\} \leqslant \delta,$$

it follows from Lemma 3.3 (with $\alpha = -1/L$) and Lemma 3.7 (with $\phi_j^* = \phi_{N+2}(\tau_N + c_j\theta, \cdot)$ and

$q_0^* = w_N, \tilde{w}_N)$ that there exist complex numbers w_{N+1}, $\tilde{w}_{N+1}$, $v_j^{(N+1)}$, $\tilde{v}_j^{(N+1)}$ $(1 \leqslant j \leqslant m)$ such that

$$v_i^{(N+1)} = w_N + \theta_{N+1} \sum_{j=1}^{m} a_{ij} \phi_{N+2}\left(\tau_N + c_j \theta_{N+1},\ v_j^{(N+1)}\right), \quad 1 \leqslant i \leqslant m, \tag{3.13a}$$

$$\tilde{v}_i^{(N+1)} = \tilde{w}_N + \theta_{N+1} \sum_{j=1}^{m} a_{ij} \phi_{N+2}\left(\tau_N + c_j \theta_{N+1},\ \tilde{v}_j^{(N+1)}\right), \quad 1 \leqslant i \leqslant m, \tag{3.13b}$$

$$w_{N+1} = w_N + \theta_{N+1} \sum_{j=1}^{m} b_j \phi_{N+2}\left(\tau_N + c_j \theta_{N+1},\ v_j^{(N+1)}\right), \tag{3.14a}$$

$$\tilde{w}_{N+1} = \tilde{w}_N + \theta_{N+1} \sum_{j=1}^{m} b_j \phi_{N+2}\left(\tau_N + c_j \theta_{N+1},\ \tilde{v}_j^{(N+1)}\right), \tag{3.14b}$$

$$|w_{N+1} - w_N| \leqslant \delta, \qquad |\tilde{w}_{N+1} - \tilde{w}_N| \leqslant \delta. \tag{3.15}$$

For θ small enough the sets T_{N+2} and $\bigcup_{k=1}^{N} T_k$ are disjoint, and therefore

$$\phi_{N+2}(\tau, x) = \psi(x) \qquad \text{for } \tau \in T_{N+2}, \quad x \in \mathbb{C}.$$

In view of Lemma 3.6 (with $q = w_{N+1}$ and $\tilde{q} = \tilde{w}_{N+1}$) we thus have proved the existence of complex numbers w_{N+2}, $\tilde{w}_{N+2}$, $v_j^{(N+2)}$, $\tilde{v}_j^{(N+2)}$ $(1 \leqslant j \leqslant m)$ with

$$v_i^{(N+2)} = w_{N+1} + \theta_{N+2} \sum_{j=1}^{m} a_{ij} \phi_{N+2}\left(\tau_{N+1} + c_j \theta_{N+2},\ v_j^{(N+2)}\right), \quad 1 \leqslant i \leqslant m, \tag{3.16a}$$

$$\tilde{v}_i^{(N+2)} = \tilde{w}_{N+1} + \theta_{N+2} \sum_{j=1}^{m} a_{ij} \phi_{N+2}\left(\tau_{N+1} + c_j \theta_{N+2},\ \tilde{v}_j^{(N+2)}\right), \quad 1 \leqslant i \leqslant m, \tag{3.16b}$$

$$w_{N+2} = w_{N+1} + \theta_{N+2} \sum_{j=1}^{m} b_j \phi_{N+2}\left(\tau_{N+1} + c_j \theta_{N+2},\ v_j^{(N+2)}\right), \tag{3.17a}$$

$$\tilde{w}_{N+2} = \tilde{w}_{N+1} + \theta_{N+2} \sum_{j=1}^{m} b_j \phi_{N+2}\left(\tau_{N+1} + c_j \theta_{N+2},\ \tilde{v}_j^{(N+2)}\right), \tag{3.17b}$$

$$|\tilde{w}_{N+2} - w_{N+2}| \geqslant \rho^2 |\tilde{w}_N - w_N|. \tag{3.18}$$

Clearly we have proved (3.9)–(3.12) with $n = N + 2$. □

Proof of Theorem 2.4. Let $\alpha \in (-\infty, 0]$ and $H \in (0, \infty]$ be given. Assume that proposition (P3) of Section 2.1 is valid for an irreducible [10] scheme S. We shall prove $r(S) \geqslant (-2\alpha)^{-1} H$.

Suppose $r(S) < (-2\alpha)^{-1} H$. Then, in view of Lemma 3.1, there is a diagonal matrix $Z \in \mathcal{D}$ with $L(Z) < (-\alpha)^{-1} H$ and $|K(Z)| > 1$. We apply Theorem 3.8 with

$$0 < \varepsilon < \min\left(-L(Z) + (-\alpha)^{-1} H,\ |K(Z)| - 1\right),$$

$$L = L(Z) + \varepsilon, \qquad w_0 = 0, \qquad \tilde{w}_0 = 1.$$

Let the sequence $\theta_1, \theta_2, \theta_3, \ldots$ and the function ϕ be as in Theorem 3.8, and let $\rho = (|K(Z)| - \varepsilon)^{1/2}$. Note that

$$L < (-\alpha)^{-1} H, \tag{3.19}$$

$$\rho > 1. \tag{3.20}$$

Let h be arbitrary with $0 < h < H$. We define step sizes h_n and a continuous function $f: \mathbb{R} \times \mathbb{R}^2 \to \mathbb{R}^2$ by

$$h_n = \theta_n h, \quad n = 1, 2, 3, \ldots$$

and

$$f(t, x) = h^{-1}(\operatorname{Re} \phi(t/h, \xi_1 + i\xi_2), \operatorname{Im} \phi(t/h, \xi_1 + i\xi_2))^T$$

$$\text{for } t \in \mathbb{R}, \qquad x = (\xi_1, \xi_2)^T \in \mathbb{R}^2.$$

Clearly $0 < h_n < H$ $(n = 1, 2, 3, \ldots)$. Further, by equipping $\mathbb{R}^2$ with the standard inner product $\langle x, y \rangle = x^T y$ and corresponding Euclidean norm $|x|_2 = (x^T x)^{1/2}$, it easily follows from $\phi(\tau, \cdot) \in \mathcal{F}(-1/L)$ (cf. Theorem 3.8) that

$$\nu[f] \leqslant -h/L. \tag{3.21}$$

Using the usual isomorphy between $\mathbf{C}$ and $\mathbb{R}^2$ it follows from Theorem 3.8 (cf. (3.6), (3.7)) that there exist $y_j^{(n)}$, $\tilde{y}_j^{(n)}$ $(1 \leqslant j \leqslant m,\ n \geqslant 1)$ and u_n, $\tilde{u}_n$ $(n \geqslant 0)$ in $\mathbb{R}^2$ with $u_0 = (0, 0)^T$, $\tilde{u}_0 = (1, 0)^T$ such that relations (1.2) and (1.3) are valid for $n \geqslant 1$. Moreover we obtain from Theorem 3.8 (cf. (3.8))

$$|\tilde{u}_n - u_n|_2 \geqslant \rho^n |\tilde{u}_0 - u_0|_2, \quad n = 2, 4, 6, \ldots.$$

In view of (P3) we conclude that $\nu[f] \geqslant \alpha$, so (3.21) yields $\alpha \leqslant -h/L$. As the latter inequality is valid for arbitrary $h \in (0, H)$ we have $\alpha \leqslant -H/L$, which is a contradiction in view of (3.19). We thus have proved $r(S) \geqslant (-2\alpha)^{-1} H$. □

References

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

[2] M. Crouzeix, Sur la B-stabilité des méthodes de Runge-Kutta, *Numer. Math. 32* (1979) 75–82.

[3] M. Crouzeix, W.H. Hundsdorfer and M.N. Spijker, On the existence of solutions to the algebraic equations in implicit Runge-Kutta methods, *BIT 23* (1983) 84–91.

[4] G. Dahlquist, Stability and error bounds in the numerical integration of ordinary differential equations, *Trans. Roy. Inst. Technol. Stockholm 130* (1959).

[5] G. Dahlquist, Error analysis for a class of methods for stiff non-linear initial value problems, in: G.A. Watson, ed., *Numerical Analysis*, Lecture Notes in Mathematics 506 (Springer, Berlin, 1976) 60–72.

[6] G. Dahlquist and R. Jeltsch, Generalized disks of contractivity for explicit and implicit Runge-Kutta methods, Rept. TRITA-NA-7906, Department of Computer Science, Royal Institute of Technology, Stockholm (1979).

[7] K. Dekker and J.G. Verwer, *Stability of Runge-Kutta Methods for Stiff Nonlinear Differential Equations* (North-Holland, Amsterdam, 1984).

[8] R. Frank, J. Schneid and C.W. Ueberhuber, The concept of B-convergence, *SIAM J. Numer. Anal. 18* (1981) 753–780.

[9] W.H. Hundsdorfer, Private communication (1984).

[10] W.H. Hundsdorfer and M.N. Spijker, A note on B-stability of Runge-Kutta methods, *Numer. Math. 36* (1981) 319–331.
[11] W.H. Hundsdorfer and M.N. Spijker, On the algebraic equations in implicit Runge-Kutta methods, *SIAM J. Numer. Anal. 24* (1987) 583–594.
[12] J.F.B.M. Kraaijevanger, B-convergence of the implicit midpoint rule and the trapezoidal rule, *BIT 25* (1985) 652–666.
[13] M.N. Spijker, Stability in the numerical solution of stiff initial value problems, *Nieuw Archief voor Wiskunde XXX* (1982) 264–276.
[14] M.N. Spijker, Feasibility and contractivity in implicit Runge–Kutta methods, *J. Comput. Appl. Math. 12–13* (1985) 563–578.
[15] M.N. Spijker, Monotonicity and boundedness in implicit Runge-Kutta methods, *Numer. Math. 50* (1986) 97–109.
[16] H.J. Stetter, *Analysis of Discretization Methods for Ordinary Differential Equations* (Springer, Berlin, 1973).
[17] P.P. Wakker, Extending monotone and non-expansive mappings by optimization, *Cahiers du C.E.R.O. 27* (1985) 141–149.