BIT Numerical Mathematics **43:** 363–385, 2003.

ANALYSIS OF ERROR GROWTH VIA STABILITY REGIONS IN NUMERICAL INITIAL VALUE PROBLEMS*

K. J. IN 'T HOUT and M. N. SPIJKER †

Mathematical Institute, Leiden University, Niels Bohrweg 1, 2333 CA Leiden, The Netherlands. email: spijker@math.leidenuniv.nl

Abstract.

This paper concerns the stability analysis of numerical methods for solving time dependent ordinary and partial differential equations. In the literature stability estimates for such methods were derived, under a condition which can be viewed as a transplantation of the Kreiss resolvent condition (from the unit disk to the stability region S of the numerical method). These estimates tell us that errors in the numerical time stepping process cannot grow faster than linearly with $\min\{s, n\}$. Here n denotes the number of time steps, and s stands for the order of the (spatial discretization) matrices involved.

In this paper we address the natural question of whether the above stability estimates can be improved so as to imply an error growth at a slower rate than $\min\{s, n\}$ (when $n \to \infty$, $s \to \infty$). Our results concerning this question are as follows: (a) for all (practical) Runge–Kutta and other one-step formulas, we show that the estimates from the literature are sharp in that error growth at the rate $\min\{s, n\}$ can actually occur, (b) for linear multistep formulas we find that, rather surprisingly, some of the stability estimates can substantially be improved and extended, whereas others are sharp.

The results proved in this paper are also relevant to (suitably scaled spatial discretization) matrices whose ε-pseudo-eigenvalues lie at a distance not exceeding $K\varepsilon$ from the stability region S of the time stepping method, for all $\varepsilon > 0$ and fixed constant K.

AMS subject classification (2000): Primary: 65L20; Secondary: 65L05, 65L06, 65M12, 65M20.

Key words: Initial value problem, discretization, numerical method, Runge–Kutta formula, linear multistep formula, error growth, stability estimate, stability analysis, stability region, resolvent condition, ε-pseudospectrum.

1 Introduction.

1.1 The purpose of the paper

This paper concerns the theoretical analysis of numerical methods for solving *initial value problems*. The focus will be on the behaviour of methods in the

*Received May 2002. Revised November 2002. Communicated by Syvert P. Nørsett.

†The research by the first author has been partially supported by the Netherlands Organization for Scientific Research (NWO).

solution of the system of *ordinary differential equations*

$$U'(t) = AU(t) + f(t) \quad (t \geq 0), \tag{1.1a}$$

under the initial condition

$$U(0) = u_0. \tag{1.1b}$$

Here A denotes a given fixed matrix of order $s \geq 1$. Further, $U(t)$, $f(t)$ and u_0 denote vectors in the s-dimensional complex space $\mathbb{C}^s$, with $U(t)$ unknown and $f(t)$, u_0 prescribed.

Problems of the form (1.1) arise, e.g., in applying semi-discretization (method of lines) to initial-boundary value problems for *partial differential equations*. The dimension s of (1.1) is then related to the accuracy of the semi-discretization and can attain (arbitrarily) large values. The analysis to be given in this paper will encompass problems (1.1), arising in this way.

Current methods, for solving initial value problems, produce in a step-by-step manner numerical approximations corresponding to consecutive discrete values t_n of the time variable t. A crucial question about these methods is whether they behave *stably* or not. Here we refer by the term stable to the situation where any (numerical) error, introduced at some stage of the computations, is propagated in a mild fashion - i.e. its effect does not grow unduly in the subsequent steps of the numerical method.

The *stability region* S of a numerical method is a classical tool for obtaining insight in the actual stability behaviour which is present when a method is applied to problem (1.1). But, this region is defined in terms of the method's behaviour in the solution of a very simple scalar test problem. As a result, for certain problems (1.1) a careless use of stability regions can lead to a completely wrong assessment of stability: for matrices A that are not normal[1], severe instability can manifest itself, although all eigenvalues of A, suitably scaled by the time step h, belong to (the interior of) S; see, e.g., Parter [12], Griffiths *et al.* [5], Morton [11], Spijker [16], Reddy and Trefethen [14].

Clearly, it is an important question of how stability regions should be used in (carefully) proving stability, also for non-normal matrices. A powerful, general approach to this question consists in imposing, on the matrix hA, a *resolvent condition* which can be viewed as a transplantation of the classical Kreiss resolvent condition (from the unit disk to S). Under such a condition on hA, stability estimates can be proved which do not fail for non-normal matrices; see, e.g., Lubich and Nevanlinna [9], Reddy and Trefethen [13, 14], Spijker and Straetemans [19], Toh and Trefethen [21].

The purpose of this paper is to address the natural question of whether the stability estimates, obtainable from the literature under the above resolvent condition on hA, can be improved. We shall bring to light that some of these estimates can be substantially sharpened (and extended), whereas others are best possible.

[1] A matrix A is said to be normal if it commutes with its Hermitian adjoint, i.e. $AA^* = A^*A$. This property is equivalent to A having an orthogonal basis of eigenvectors.

1.2 Outline of the rest of the paper

In Section 2 we state our conclusions, without proof, about stability estimates for Runge–Kutta and other one-step methods applied to problem (1.1).

In Section 2.1, formula (2.4), we formulate the resolvent condition (on the matrix hA) which we shall deal with in our paper. Theorem 2.1 gives a general stability estimate, which was obtained in the literature under condition (2.4).

In Section 2.2 we turn to the question of whether the stability estimate of Theorem 2.1 is sharp (in a sense specified precisely). The main result of this subsection is formulated in Theorem 2.2. This theorem immediately leads to the important conclusion that, for all practical one-step methods, the stability estimate of Theorem 2.1 is sharp indeed.

In Section 3 we state our conclusions, without proof, about stability estimates for linear multistep methods applied to problem (1.1).

In Section 3.1 we outline two stability estimates that were obtained in the literature for linear multistep methods under condition (2.4).

In Section 3.2, Theorem 3.2, we formulate our first main result of Section 3. This theorem can be viewed as an analogue of Theorem 2.1 relevant for the case of linear multistep methods; it gives a stability estimate which substantially extends and improves the two estimates of Section 3.1.

In Section 3.3 we study the question of whether the stability estimate of Theorem 3.2 is sharp. We present two theorems with regard to this question. The first theorem, Theorem 3.3, immediately leads to the interesting conclusion that in a certain non-trivial situation a much stronger stability estimate is possible than the one of Theorem 3.2. The second theorem, Theorem 3.4, can be viewed as an analogue of Theorem 2.2; it directly implies the important result that, in the situation complementary to the one of Theorem 3.3, the stability estimate given by Theorem 3.2 is sharp.

In Section 4 we present the proofs of all theorems of the Sections 2.2, 3.3 that correspond to *lower bounds* (for norms of powers of numerical solution operators). We begin this section by deriving two general lemmas which provide us with matrices A that are especially useful for obtaining suitable lower bounds. Next, by using these matrices A, we prove the Theorems 2.2 and 3.4.

In Section 5 we present the proofs of all theorems of the Sections 3.2, 3.3 that correspond to *upper bounds* (for the norms of the powers of numerical solution operators). We prove Theorem 3.2, by showing that the relevant companion matrix fulfills an appropriate resolvent bound under condition (2.4). Subsequently, we readily prove Theorem 3.3, by using the Jordan canonical form of A.

2 One-step methods: a general stability estimate from the literature and the formulation of our main result.

2.1 A general stability estimate known from the literature

We consider numerical processes of the form

$$u_{n+1} = \varphi(hA)u_n + f_n \quad (n = 0, 1, 2, \ldots). \tag{2.1}$$

Here u_n (for $n \geq 1$) denote vectors in $\mathbb{C}^s$ which are computed successively according to (2.1), starting from a given $u_0 \in \mathbb{C}^s$. Further, $h > 0$ is the *stepsize*, f_n are given vectors in $\mathbb{C}^s$, and A is a complex $s \times s$ matrix. By $\varphi(z)$ we denote a given rational function; we define $\varphi(hA) = P(hA)[Q(hA)]^{-1}$, when $P(z)$, $Q(z)$ are polynomials such that $\varphi(z) \equiv P(z)/Q(z)$ and the matrix $Q(hA)$ is invertible.

Many one-step methods for the numerical solution of ordinary differential equations, like Runge–Kutta methods or Rosenbrock methods (see, e.g., Butcher [2] or Hairer and Wanner [6]), reduce—when applied to (1.1)—to processes of the form (2.1). The vectors u_n then approximate $U(t)$ at the points $t = t_n = nh$ $(n = 1, 2, 3, \ldots)$, and the rational function $\varphi(z)$ satisfies

$$\varphi(0) = \varphi'(0) = 1. \tag{2.2}$$

Suppose the numerical calculations, by process (2.1), were performed starting from an initial vector $\widetilde{u}_0$ deviating (slightly) from the true u_0, e.g., due to rounding off. The numerical approximations, say $\widetilde{u}_n$, obtained that way, will then deviate from the true vectors u_n (for $n \geq 1$). Since both the (perturbed) approximations $\widetilde{u}_n$ and the (unperturbed) u_n satisfy recurrence relation (2.1), the *propagated error* $\widetilde{u}_n - u_n$ is related to the *starting error* $\widetilde{u}_0 - u_0$ via the formula

$$\widetilde{u}_n - u_n = \varphi(hA)^n(\widetilde{u}_0 - u_0) \quad (\text{for } n \geq 1).$$

Accordingly, in analysing the stability of (2.1) one is concerned with establishing (moderate) bounds on $\varphi(hA)^n$ (for $n \geq 1$).

In order to study stability in a (reliable) quantitative framework, we assume that $|\cdot|$ is an arbitrary given norm on $\mathbb{C}^s$. We denote the corresponding induced matrix norm by $\|\cdot\|$, i.e. for all $s \times s$ matrices B we define

$$\|B\| = \max\{|Bx|/|x| : x \in \mathbb{C}^s \text{ with } x \neq 0\}. \tag{2.3}$$

We define the *stability region* of any rational function $\varphi(z)$ by

$$S = \{z : z \in \mathbb{C} \text{ with } |\varphi(z)| \leq 1\}.$$

In the following we shall deal with bounds on $\|\varphi(hA)^n\|$ under the assumption that

$$(zI - hA) \text{ is invertible, and } \|(zI - hA)^{-1}\| \leq \frac{K}{\mathrm{d}(z, S)} \text{ for all } z \in \mathbb{C} \setminus S. \tag{2.4}$$

Here I denotes the $s \times s$ identity matrix, K is a positive constant and $\mathrm{d}(z, S)$ denotes the distance from z to the set S,

$$\mathrm{d}(z, S) = \inf\{|z - x| : x \in S\}.$$

One usually calls $(zI - hA)^{-1}$ the *resolvent* of hA at z, and we shall refer to (2.4) as the *Kreiss resolvent condition on hA with respect to S, with constant K*. This terminology is used because condition (2.4), with S equal to the unit disk $|z| \leq 1$ and with matrix norm $\|\cdot\|$ equal to the spectral norm, reduces to

one of the equivalent conditions occurring in the famous Kreiss matrix theorem (see, e.g., Richtmyer and Morton [15]).

For the sake of completeness we mention that (2.4) can be reformulated in terms of the so-called *ε-pseudospectrum* $\Lambda_\varepsilon(hA) = \{\lambda : \lambda$ is an eigenvalue of some $s \times s$ matrix B with $\|B - hA\| \le \varepsilon\}$: condition (2.4) is equivalent to the requirement that, for all $\varepsilon > 0$, the set $\Lambda_\varepsilon(hA)$ lies within the set $\{z : \mathrm{d}(z, S) \le K\varepsilon\}$ (see Reddy and Trefethen [13, 14], Dorsselaer *et al.* [4]).

Under condition (2.4) (and various additional conditions regarding $\varphi(z)$) several authors proved upper bounds on $\|\varphi(hA)^n\|$ which grow mildly (linearly) with n or s; see Dorsselaer *et al.* [4], Lenferink and Spijker [8], Lubich and Nevanlinna [9], Reddy and Trefethen [14], Spijker and Straetemans [19], Toh and Trefethen [21]. The following is a known general result:

THEOREM 2.1. *Let $\varphi(z)$ be an arbitrary given rational function. Then there is a constant γ, depending only on φ, such that*

(2.5) $\|\varphi(hA)^n\| \le \gamma \cdot K \cdot \min\{s, n\}$ *(whenever $n \ge 1$, $s \ge 1$, $h > 0$, $K \ge 1$, $\|\cdot\|$ is any norm of the form* (2.3)*, and A is an arbitrary $s \times s$ matrix satisfying* (2.4)*).*

For examples and illustrations to the stability estimate (2.5) we refer to the literature just mentioned.

2.2 New conclusions about sharpness

In the stability analysis of process (2.1), one is interested in bounds on $\|\varphi(hA)^n\|$ particularly for large values of n and s. Accordingly, we shall say that the stability estimate (2.5) is *sharp* if the upper bound $\gamma \cdot K \cdot \min\{s, n\}$ cannot be replaced by any function, say $F(K, s, n)$, growing more slowly than in proportion to $\min\{s, n\}$ (for all fixed K and $n \to \infty$, $s \to \infty$). Clearly, it is an important question of whether (2.5) is sharp in this sense.

Below we shall address the sharpness question for arbitrary given rational functions $\varphi(z)$. But, first we shall focus on two special cases in which $\varphi(z)$ has a very simple structure. Throughout the paper, we shall use the notation

$$\|B\|_\infty = \max\{|Bx|_\infty / |x|_\infty : x \in \mathbb{C}^m \text{ with } x \ne 0\},$$

for any $m \times m$ matrix B. Here $|\cdot|_\infty$ denotes the maximum norm, i.e., $|x|_\infty = \max_i |\xi_i|$ (for any $x \in \mathbb{C}^m$ with components $\xi_1, \xi_2, \ldots, \xi_m$).

Case (i). Assume $\varphi(z) \equiv z$.

Let $\|\cdot\| = \|\cdot\|_\infty$. Then there exist constants $c > 0$, $K \ge 1$, $h > 0$ and $s \times s$ matrices $A_{s,n}$ (for $n \ge 1$, $s \ge 1$) which satisfy at the same time the following two conditions:

(2.6a) Property (2.4) holds, with $A = A_{s,n}$ (for $n \ge 1$, $s \ge 1$);

(2.6b) $\|\varphi(hA)^n\| \ge c \cdot \min\{s, n\}$, with $A = A_{s,n}$ (for $n \ge 1$, $s \ge 1$).

A proof of the existence of such c, K, h, $A_{s,n}$ can be found in Borovykh and Spijker [1]; for closely related material see Kraaijevanger [7]. We conclude that, for the case $\varphi(z) \equiv z$, the estimate (2.5) is sharp in the sense specified above.

One might conjecture that conclusions, similar to the above for $\varphi(z) \equiv z$, can be reached for all rational functions $\varphi(z)$. But, the following case shows that this is not true.

Case (ii). Assume $\varphi(z)$ is constant.
In this case, where $\varphi(z)$ equals some constant value α, it is convenient to distinguish between the situation where $|\alpha| \leq 1$ and where $|\alpha| > 1$. If $|\alpha| \leq 1$, then $\|\varphi(hA)^n\| = |\alpha^n| \leq 1$, so that the upper bound $\gamma \cdot K \cdot \min\{s, n\}$ in (2.5) can be replaced, e.g., by

$$F(K, s, n) \equiv 1. \tag{2.7}$$

If $|\alpha| > 1$, then the corresponding stability region S is empty, so that (2.4) cannot be fulfilled. Therefore, the choice (2.7) is still formally correct. In conclusion, for any value α, the stability estimate in Theorem 2.1 is far from sharp.

The following theorem constitutes our main result about one-step methods; it will be proved in Section 4.

THEOREM 2.2. *Let $\varphi(z)$ be an arbitrary rational function which is not constant. Let $\|\cdot\| = \|\cdot\|_\infty$. Then there exist constants $c > 0$, $K \geq 1$, $h > 0$ and $s \times s$ matrices $A_{s,n}$ (for $n \geq 1$, $s \geq 1$) satisfying at the same time* (2.6*a*) *and* (2.6*b*).

In view of our above discussion of Case (ii) and of (2.2), Theorem 2.2 immediately leads to the following two interesting conclusions.

CONCLUSION 2.3. *Let $\varphi(z)$ be an arbitrary given rational function. Then the estimate* (2.5) *in Theorem* 2.1 *is sharp if and only if $\varphi(z)$ is not constant.*

CONCLUSION 2.4. *Suppose process* (2.1) *originates from applying to* (1.1) *some one-step method for ordinary differential equations, e.g., a Runge–Kutta or Rosenbrock method. Then the estimate* (2.5) *of Theorem* 2.1 *is sharp.*

3 Multistep methods: stability estimates from the literature and the formulation of our main results.

3.1 Stability estimates from the literature

We consider numerical processes of the form

$$\sum_{i=0}^{k} \alpha_i u_{n+i} = hA \sum_{i=0}^{k} \beta_i u_{n+i} + f_n \quad (\text{for } n \geq 0). \tag{3.1}$$

Here k is a fixed positive integer, and u_n (for $n \geq k$) denote vectors in $\mathbb{C}^s$ computed successively from starting vectors $u_0, u_1, \ldots, u_{k-1}$. Further, h, A and f_n

have a similar meaning as in Section 2.1, and α_i, β_i are real constants specifying the numerical process. We introduce the polynomials

$$\rho(\zeta) = \sum_{i=0}^{k} \alpha_i \zeta^i \quad \text{and} \quad \sigma(\zeta) = \sum_{i=0}^{k} \beta_i \zeta^i,$$

and in all of the following we make the (standard) assumptions

$$\alpha_k = 1, \ |\alpha_0| + |\beta_0| > 0, \ \rho(\zeta) \text{ and } \sigma(\zeta) \text{ have no common root.}$$

Linear multistep methods as well as *one-leg methods* (see, e.g., Butcher [2] or Hairer and Wanner [6]) reduce–when applied in the solution of (1.1)—to processes of the form (3.1).

By introducing the vectors $U_n = (u_{n+k-1}^{\mathrm{T}}, \ldots, u_{n+1}^{\mathrm{T}}, u_n^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{C}^{sk}$, the numerical process (3.1) can be rewritten in the compact form

$$U_{n+1} = \Phi(hA)U_n + F_n \quad (n = 0, 1, 2, \ldots).$$

Here $\Phi(hA)$ denotes the companion matrix of order sk defined by

$$\Phi(hA) = \begin{pmatrix} \varphi_{k-1}(hA) & \ldots & \varphi_1(hA) & \varphi_0(hA) \\ I & & & \\ & \ddots & & \\ & & I & O \end{pmatrix},$$

with

$$\varphi_i(hA) = (-\alpha_i I + \beta_i hA)(\alpha_k I - \beta_k hA)^{-1} \quad (\text{for } 0 \le i \le k-1);$$

further all components of $F_n \in \mathbb{C}^{sk}$ are zero with the exception of the first s ones, which are equal to those of $(\alpha_k I - \beta_k hA)^{-1} f_n$. Clearly, in the stability analysis of (3.1), the derivation of (moderate) bounds on $\Phi(hA)^n$ (for $n \ge 1$) is of crucial importance.

Similarly as in Section 2.1, we denote by $|\cdot|$ an arbitrary norm in $\mathbb{C}^s$, and by $\|\cdot\|$ the corresponding induced matrix norm (see (2.3)). We define a norm $|\cdot|$ in $\mathbb{C}^{sk}$ by

$$|x| = \max_i |x_i| \quad \text{for } x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{pmatrix} \text{ with } x_i \in \mathbb{C}^s. \tag{3.2a}$$

Further, for any $sk \times sk$ matrix $B = (B_{ij})$, composed of $s \times s$ blocks B_{ij} (where $1 \le i, j \le k$), we define

$$\|B\| = \|(B_{ij})\| = \max\{|Bx|/|x| : x \in \mathbb{C}^{sk} \text{ with } x \ne 0\}. \tag{3.2b}$$

It is easily seen that, for such block matrices $B = (B_{ij})$,

$$\max_{i,j} \|B_{ij}\| \le \|B\| \le \max_{i} \sum_{j} \|B_{ij}\|. \tag{3.2c}$$

The *stability region* S of process (3.1) is a subset of $\overline{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$ defined by

$$S = \{z : z \in \overline{\mathbb{C}} \text{ with } \sup_{n \ge 1} \|\Phi(z)^n\| < \infty\}.$$

Here $\|\cdot\|$ is defined according to (3.2a,b), with $s = 1$. Further, we use the conventions $a/0 = \infty$ (for $a \in \overline{\mathbb{C}}$, $a \ne 0$), $a/0 = 0$ (for $a = 0$). In order to formulate a useful characterization of S, it is convenient to use the following terminology: any polynomial $p(\zeta)$ satisfies the *root condition* if all its roots have a modulus not exceeding 1, and any roots with modulus 1 are simple. It is easily seen that, for any $z \in \overline{\mathbb{C}}$,

(3.3a) $z \in S \cap \mathbb{C} \iff \alpha_k - z\beta_k \ne 0$ and $\rho(\zeta) - z\sigma(\zeta)$ satisfies the root condition;

(3.3b) $z = \infty \in S \iff \beta_k \ne 0$ and $\sigma(\zeta)$ satisfies the root condition.

Below we shall shortly indicate two stability estimates obtainable from the literature.

In describing the first estimate, we shall refer to the following conditions:

(3.4a) S is bounded;

(3.4b) $\sigma(\zeta) \ne 0$ and $\rho'(\zeta)\sigma(\zeta) - \rho(\zeta)\sigma'(\zeta) \ne 0$ (whenever $|\zeta| = 1$).

The subsequent theorem is a neat counterpart of Theorem 2.1, valid for processes (3.1) satisfying (3.4):

THEOREM 3.1. *Assume the numerical process* (3.1) *satisfies condition* (3.4). *Then there is a constant* γ, *depending only on the coefficients* α_i, β_i, *such that*

(3.5) $\|\Phi(hA)^n\| \le \gamma \cdot K \cdot \min\{s, n\}$ *(whenever* $n \ge 1$, $s \ge 1$, $h > 0$, $K \ge 1$, $\|\cdot\|$ *is any norm as in* (3.2*a*, *b*), *and* A *is an arbitrary* $s \times s$ *matrix satisfying* (2.4)*).*

The above theorem is essentially due to Reddy and Trefethen [13, 14]. These authors use in their papers a terminology and framework somewhat different from ours; but in their 1992 paper they still make assumption (3.4) and they use arguments which are easily seen to imply the theorem stated above. (On the other hand, their arguments do *not* imply (3.5) when (3.4) is violated.)

Clearly, the restriction (3.4) in the above theorem is irksome; e.g., linear multistep methods with an unbounded stability region (which are important for stiff problems) are not included. Lubich and Nevanlinna [9] made the important observation that certain stability estimates can still be established *without*

assuming (3.4). In order to sketch their estimates, we introduce $s \times s$ matrices $B_{ij}(n)$ and $r_n(hA)$ (for $n \geq 0$), with

$$\Phi(hA)^n = \begin{pmatrix} B_{11}(n) & \dots & B_{1k}(n) \\ \cdot & & \cdot \\ \cdot & & \cdot \\ B_{k1}(n) & \dots & B_{kk}(n) \end{pmatrix}, \quad r_n(hA) = B_{11}(n)(\alpha_k I - \beta_k hA)^{-1}.$$

In the paper just mentioned (pp. 311, 312) an upper bound of the following form is derived:

$$\|r_n(hA)\| \leq \gamma \cdot K \cdot \min\{s, n+1\}.$$

Such a bound is directly relevant to estimating the effect on u_n of (perturbations in) the vectors f_n (see (3.1)). It can also be used indirectly, in bounding $\|\Phi(hA)^n\|$, because

$$B_{ij}(n) = \sum_{m=1}^{j} r_{n+m-i}(hA)(\alpha_{k+m-j} I - \beta_{k+m-j} hA),$$

where we use the convention $r_m(hA) = O$ (for $m < 0$). By using (3.2c), it can easily be seen that the above upper bound for $\|r_n(hA)\|$ implies the stability estimate

$$\|\Phi(hA)^n\| \leq \delta(hA) \cdot \gamma K \cdot \min\{s, n\}, \tag{3.6a}$$

where

$$\delta(hA) = k(k+1) \sum_{i=1}^{k} \|\alpha_i I - \beta_i hA\|. \tag{3.6b}$$

Unfortunately, in the numerical solution of stiff problems (for which typically $\|hA\| >> 1$) the size of the factor $\delta(hA)$ is not under control.

In the next Section 3.2 we shall extend (3.5) to a class of general processes of the form (3.1) (not necessarily satisfying (3.4)). In Section 3.3 we shall bring to light in which cases the estimate (3.5) is sharp.

For the sake of completeness, we note that Reddy and Trefethen [13, 14] related the resolvent condition (2.4) also to a strong stability estimate of the form $\|\Phi(hA)^n\| \leq c$ $(n \geq 1)$. Under condition (3.4), they proved that this strong estimate implies (2.4), with a ratio K/c only depending on the coefficients α_i, β_i. In the following, we shall not deal further with this implication, or possible variants thereof - because we feel its study is not needed for answering the main question raised in Section 1.1 and we want to keep the present paper sufficiently concise.

3.2 A new stability result

Our first main result for multistep methods is given in the following theorem; it will be proved in Section 5.

THEOREM 3.2. *Assume the stability region S of process* (3.1) *is closed in* $\overline{\mathbb{C}}$. *Then there is a constant* γ, *depending only on the coefficients* α_i, β_i, *such that* (3.5) *is valid.*

For most numerical processes (3.1) of practical interest the condition that S is closed in $\overline{\mathbb{C}}$ is fulfilled; see, e.g., Dahlquist, Mingyou, and LeVeque [3] and Hairer and Wanner [6, Sect. V.7]. Moreover, it can readily be seen that S is closed in $\overline{\mathbb{C}}$ whenever (3.4) holds. Accordingly, Theorem 3.2 constitutes a far-reaching generalization of Theorem 3.1.

Furthermore, with an eye to processes (3.1) violating (3.4), it is important to note that the stability estimate given by Theorem 3.2 amounts to an essential improvement over (3.6).

3.3 New conclusions about sharpness

In this subsection we consider the question of whether the stability estimate (3.5) is sharp (in a similar sense as explained at the start of Section 2.2). The following theorem, which will be proved in Section 5, leads to a first interesting conclusion regarding this question.

THEOREM 3.3. *Assume the stability region S of process* (3.1) *consists of a finite number of points. Then there is a constant* γ, *depending only on the* α_i, β_i, *such that*

$$\|\Phi(hA)^n\| \leq \gamma \cdot K$$

whenever $n \geq 1$, $s \geq 1$, $h > 0$, $K \geq 1$, $\|\cdot\|$ *is any norm as in* (3.2a, b), *and A is an arbitrary $s \times s$ matrix satisfying* (2.4).

This theorem shows that, whenever the set S is finite, a much stronger stability estimate is valid than (3.5). (For natural examples of multistep methods with finite S; see, e.g., Hairer and Wanner [6, Sect. V.1].)

The problem arises whether in addition to S being finite there exist further situations in which (3.5) can essentially be improved. The subsequent theorem, to be proved in Section 4, will enable us to settle this problem.

THEOREM 3.4. *Assume the stability region S of process* (3.1) *is closed in* $\overline{\mathbb{C}}$ *and does not consist of a finite number of points. Let* $\|\cdot\| = \|\cdot\|_\infty$. *Then there exist constants* $c > 0$, $K \geq 1$, $h > 0$ *and* $s \times s$ *matrices* $A_{s,n}$ *(for* $n \geq 1$, $s \geq 1$*) satisfying at the same time the following two conditions:*

(3.7a) *Property (2.4) holds, with* $A = A_{s,n}$ *(for* $n \geq 1$, $s \geq 1$*);*

(3.7b) $\|\Phi(hA)^n\| \geq c \cdot \min\{s, n\}$, *with* $A = A_{s,n}$ *(for* $n \geq 1$, $s \geq 1$*).*

The above theorem can be viewed as a convenient analogue of Theorem 2.2 for the case of multistep processes.

By combining the Theorems 3.3 and 3.4, we immediately arrive at the following, neat conclusion concerning the sharpness of the stability estimate in Theorem 3.2.

CONCLUSION 3.5. *Let* (3.1) *be an arbitrary given process such that* S *is closed in* $\overline{\mathbb{C}}$. *Then the estimate* (3.5) *in Theorem* 3.2 *is sharp if and only if* S *does not consist of a finite number of points.*

4 Proving lower bounds.

4.1 Constructing general matrices $A_{s,n}$

In the present Section 4.1 we shall first focus on $s \times s$ matrices A with two properties which can be viewed as generalizations of (2.6a) and (2.6b), respectively. These general properties will be specified in Lemma 4.1. Next, by combining this lemma and results from the literature, we shall easily conclude that $s \times s$ matrices $A_{s,n}$ exist having properties very similar to those in the Theorems 2.2 and 3.4. This conclusion will be formulated concisely in Lemma 4.2. The latter lemma will be essential in proving the two theorems in the Sections 4.3 and 4.4, respectively.

We assume that C, D and T are $s \times s$ matrices, with T invertible, such that

$$C = TDT^{-1}, \quad D = \operatorname{diag}(\delta_1, \delta_2, \ldots, \delta_s); \tag{4.1a}$$

$$\delta_j = \delta_0 \cdot \mathrm{e}^{\mathrm{i}j\theta} \quad (1 \le j \le s), \quad |\delta_0| = 1, \quad \theta > 0; \tag{4.1b}$$

$$0 \le (s-1)\theta \le \pi. \tag{4.1c}$$

We use the notation

$$E_j = \operatorname{diag}(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_s), \text{ with } \varepsilon_i = 1 \text{ (for } i = j) \text{ and } \varepsilon_i = 0 \text{ (for } i \ne j),$$

and we define the projection matrices

$$P_j = TE_jT^{-1} \quad (\text{for } 1 \le j \le s). \tag{4.2a}$$

By $\|\cdot\|$ we denote an arbitrary matrix norm of the form (2.3), and we assume that β is a constant with

$$\|P_1 + P_2 + \ldots + P_j\| \le \beta \quad (\text{for } 1 \le j \le s). \tag{4.2b}$$

Matrices C having essentially the properties (4.1), (4.2) were used, e.g., in McCarthy and Schwartz [10], Kraaijevanger [7], Borovykh and Spijker [1], Spijker *et al.* [20]. In these papers, actual norms $\|\cdot\|$ and matrices C were specified for which $K = \sup_{|\zeta|>1}(|\zeta| - 1)\|(\zeta I - C)^{-1}\|$ is of "moderate size" whereas $\|C^n\|$ is "large". The size of K was estimated by using the important inequality $K \le 1 + \beta\pi$ (which is essentially due to McCarthy and Schwartz [10]).

In our Lemma 4.1, we shall assume that C is an arbitrary matrix with the properties (4.1), (4.2). Further, we shall assume that $f(z)$ is a complex function such that, for some $z_0 \in \mathbb{C}$ and $r > 0$,

$$f(z) \text{ is holomorphic for } |z - z_0| < r, \text{ and } f'(z_0) \ne 0; \tag{4.3a}$$

(4.3b) $\zeta_0 = f(z_0)$ has a modulus $|\zeta_0| = 1$.

We shall use the notation

(4.3c) $S_f = \{z : |z - z_0| < r \text{ and } |f(z)| = 1\}$.

LEMMA 4.1. *Let $\|\cdot\|$ be an arbitrary matrix norm of the form* (2.3)*, and assume* (4.1)–(4.3)*. Then for each integer $n \geq 1$ there exists an $s \times s$ matrix A with the following two properties:*

(4.4a) *$(zI - A)$ is invertible, and $\|(zI - A)^{-1}\| \leq \dfrac{\gamma_1 \beta}{\mathrm{d}(z, S_f)}$ for all $z \in \mathbb{C} \setminus S_f$;*

(4.4b) $\|f(A)^n\| \geq \gamma_0 \|C^n\|$.

Here γ_0, γ_1 are positive constants only depending on f, z_0 and r (and not on n or any of the quantities occurring in (4.1), (4.2) *).*

In order to keep the present subsection sufficiently concise and transparant, we postpone the proof of the above lemma to Section 4.2. We shall use Lemma 4.1 in proving the following conclusion:

LEMMA 4.2. *Assume* (4.3)*. Let $\|\cdot\| = \|\cdot\|_\infty$. Then there exist positive constants γ_0, K and $s \times s$ matrices $A_{s,n}$ such that, for all $n \geq 1$ and $s \geq 1$,*

(4.5a)
$(zI - A_{s,n})$ is invertible, and $\|(zI - A_{s,n})^{-1}\| \leq \dfrac{K}{\mathrm{d}(z, S_f)}$ for all $z \in \mathbb{C} \setminus S_f$;

(4.5b) $\|f(A_{s,n})^n\| \geq \gamma_0 \cdot \min\{s, n\}$.

Here γ_0 and K only depend on f, z_0 and r (and not on n or s).

PROOF. 1. We assume first that $1 \leq s \leq n$. We define the $s \times s$ matrix C by (4.1), with $\delta_j = \mathrm{e}^{\mathrm{i}j\theta}$, $\theta = \pi/n$ and with $T = (t_{ij})$ where $t_{ij} = 1$ (for $j \leq i$) and $t_{ij} = 0$ (for $j > i$). By Kraaijevanger [7], and also by Borovykh and Spijker [1, pp. 166, 167], it was shown that the corresponding matrices P_j, defined by (4.2a), satisfy (4.2b) with

$$\|\cdot\| = \|\cdot\|_\infty \quad \text{and} \quad \beta = 1.$$

Further, by a direct computation (similar to computations in the above two papers) one can see that

$$\|C^n\|_\infty = 2s - 1.$$

We apply Lemma 4.1 to the situation at hand, and denote the corresponding $s \times s$ matrix A by $A_{s,n}$. Since (4.4a) holds with $\gamma_1 \beta = \gamma_1$, we arrive at (4.5a) with $K = \gamma_1$. Further, $\|f(A_{s,n})^n\|_\infty \geq \gamma_0(2s - 1) \geq \gamma_0 s = \gamma_0 \cdot \min\{s, n\}$, so that (4.5b) holds.

2. Next, we assume that $1 \le n < s$. We define the $s \times s$ matrix

$$A_{s,n} = \begin{pmatrix} A_{n,n} & O \\ O & z_0 I \end{pmatrix},$$

where $A_{n,n}$ was specified above in Part 1 of the proof.

By using that $(zI - A_{n,n})$ is invertible and $\|(zI - A_{n,n})^{-1}\|_\infty \le \gamma_1/\mathrm{d}(z, S_f)$ (for all $z \in \mathbb{C} \setminus S_f$), one can conclude that $A_{s,n}$ satisfies (4.5a) with $K = \max\{\gamma_1, 1\} = \gamma_1$.

Further, $\|f(A_{s,n})^n\|_\infty \ge \|f(A_{n,n})^n\|_\infty \ge \gamma_0 n = \gamma_0 \cdot \min\{s, n\}$, so that (4.5b) still holds. This completes the proof. □

REMARK 4.1. In the above, no other application of Lemma 4.1 is made than in proving Lemma 4.2. But, other applications are possible as well. One could combine Lemma 4.1, e.g., with the material in Spijker *et al.* [20], so as to arrive at interesting variants of Lemma 4.2 and Theorem 2.2 which are relevant to the spectral norm $\|\cdot\|_2$ (rather than the norm $\|\cdot\|_\infty$). However, the corresponding lower bounds for $\|\varphi(hA_{s,n})^n\|_2$ would *not* be good enough for proving the sharpness (if $\varphi(z)$ is not constant) of Theorem 2.1. For that reason we do not go further into this application here.

4.2 Proof of Lemma 4.1

Part 1

In order to prove Lemma 4.1, we assume that $\|\cdot\|$ is an arbitrary norm of the form (2.3) and that (4.1)–(4.3) hold.

We define

$$\mu_j = \zeta_0 \mathrm{e}^{\mathrm{i}(j-1)\theta} \quad (\text{for } 1 \le j \le s).$$

Clearly, $\mu_1 = \zeta_0$ and $\mu_j = \mu_1 \delta_j/\delta_1$. Therefore, in order to arrive at (4.4b) (with $\gamma_0 = 1$), one could think of choosing A such that $f(A) = (\mu_1/\delta_1)\, C$, i.e. $A = \sum_{j=1}^s \lambda_j P_j$ with eigenvalues λ_j satisfying

$$|\lambda_j - z_0| < r \quad \text{and} \quad f(\lambda_j) = \mu_j. \tag{4.6}$$

However, the existence of λ_j satisfying (4.6) can be guaranteed only for those μ_j which lie sufficiently close to μ_1. In fact, by the inverse function theorem for holomorphic functions, there exist open sets X and Y, with

$$z_0 \in X \subset \{z : |z - z_0| < r\}, \quad \zeta_0 \in Y,$$

such that f maps X in a 1–1 fashion onto Y and such that the inverse $g = f^{-1}$ is holomorphic on Y with $g'(\zeta) \ne 0$ (for $\zeta \in Y$). We choose $\tau > 0$ so small that $\zeta_0 \mathrm{e}^{\mathrm{i}t} \in Y$ (for $0 \le t \le \tau$), and we denote by q the largest integer with

$$1 \le q \le s \quad \text{and} \quad (q-1)\theta \le \tau.$$

We see that $\lambda_j = g(\mu_j)$ satisfies (4.6) for $j = 1, 2, \ldots, q$. Consequently, for any integers m and p satisfying

$$1 \le p \le q, \quad m \ge 0, \quad m + p \le s, \tag{4.7a}$$

we can define

$$A = \sum_{j=1}^{p} \lambda_j P_{m+j}. \tag{4.7b}$$

Clearly

$$f(A) = \sum_{j=1}^{p} \mu_j P_{m+j}. \tag{4.7c}$$

Below we shall prove (4.4a), (4.4b) for some A of the form (4.7b).

Part 2

With an eye to (4.4b), we want to relate C^n to matrices of the form

$$f(A)^n = \sum_{j=1}^{p} (\mu_j)^n P_{m+j}.$$

For this purpose we define the integer $k \geq 0$ by

$$kq \leq s - 1 < (k+1)q. \tag{4.8}$$

We have

$$C^n = \left[\sum_{l=0}^{k-1} \Big(\sum_{j=lq+1}^{lq+q} P_j \Big) + \sum_{j=kq+1}^{s} P_j \right] C^n.$$

Since $P_{m+j} C^n = (\delta_{m+j})^n P_{m+j} = \big(\delta_1 \mathrm{e}^{\mathrm{i}m\theta}/\mu_1\big)^n \cdot (\mu_j)^n P_{m+j}$, it follows that

$$\|C^n\| \leq \sum_{l=0}^{k-1} \Big\| \sum_{j=1}^{q} (\mu_j)^n P_{lq+j} \Big\| + \Big\| \sum_{j=1}^{s-kq} (\mu_j)^n P_{kq+j} \Big\|.$$

Among the $k+1$ terms in the right-hand member of the last inequality there must be a term which is greater than or equal to $\|C^n\|/(k+1)$. Consequently, there exist m, p satisfying (4.7a) such that, for A defined by (4.7b),

$$\|f(A)^n\| \geq \|C^n\|/(k+1).$$

First, suppose $k \geq 1$. In view of (4.8) we then have $q \leq s-1$, and in view of the definition of q we arrive at the inequality $q\theta > \tau$. Together with (4.1c) this leads to $k \leq (s-1)\theta/(q\theta) < \pi/\tau$. Applying the above lower bound for $\|f(A)^n\|$, we conclude that (4.4b) is valid, with $\gamma_0 = 1/(1+\pi/\tau)$. It is evident that (4.4b) holds with the same value γ_0 in case $k = 0$.

Part 3

Let A be any matrix of the form (4.7b), with m, p as in (4.7a). In view of our conclusion in Part 2 above, it only remains to show that (4.4a) is fulfilled, with some γ_1 only depending on the quantities in (4.3).

Let $z \in \mathbb{C} \setminus S_f$. The matrix $zI - A$ is invertible, with

$$(zI - A)^{-1} = \sum_{j=1}^{p} (z - \lambda_j)^{-1} P_{m+j}.$$

Defining $R_j = \sum_{i=1}^{j} P_{m+i}$, we thus have

$$(zI - A)^{-1} = \frac{1}{z - \lambda_p} R_p + \sum_{j=1}^{p-1} \left(\frac{1}{z - \lambda_j} - \frac{1}{z - \lambda_{j+1}} \right) R_j.$$

From (4.2b) we see that $\|R_j\| \leq 2\beta$, so that

$$\|(zI - A)^{-1}\| \leq 2\beta \cdot [1 + \alpha(z)] \cdot \frac{1}{\mathrm{d}(z, S_f)},$$

where

$$\alpha(z) = \mathrm{d}(z, S_f) \sum_{j=1}^{p-1} \left| \frac{1}{z - g(\mu_j)} - \frac{1}{z - g(\mu_{j+1})} \right|.$$

It remains to prove that $\alpha(z)$ can be bounded uniformly for $z \in \mathbb{C} \setminus S_f$.

The following lemma is tailor-made for bounding $\alpha(z)$.

LEMMA 4.3. *Let $W \subset \mathbb{C}$, and $F : [0, \tau] \to W$. Assume F is continuously differentiable with $F'(t) \neq 0$ and $F(t) \neq F(t')$ for all $t, t' \in [0, \tau]$, $t' \neq t$. Then*

$$\sup_z \mathrm{d}(z, W) \int_0^\tau \left| \frac{\mathrm{d}}{\mathrm{d}t} \left(\frac{1}{z - F(t)} \right) \right| \mathrm{d}t < \infty,$$

where the supremum is over all $z \in \mathbb{C} \setminus W$.

This lemma can be viewed as a generalization of Lemma 5.2 in Spijker and Straetemans [18]. The proof of the present Lemma 4.3 runs along the same lines as the proof of the lemma in that paper; for this reason we omit it here.

Defining $F(t) = g(\zeta_0 \mathrm{e}^{\mathrm{i}t})$, we have

$$\begin{aligned} \alpha(z) &= \mathrm{d}(z, S_f) \sum_{j=1}^{p-1} \left| \frac{1}{z - F((j-1)\theta)} - \frac{1}{z - F(j\theta)} \right| \\ &\leq \mathrm{d}(z, S_f) \int_0^\tau \left| \frac{\mathrm{d}}{\mathrm{d}t} \left(\frac{1}{z - F(t)} \right) \right| \mathrm{d}t. \end{aligned}$$

An application of Lemma 4.3, with $W = S_f$, completes the proof of Lemma 4.1. □

4.3 Proof of Theorem 2.2

The proof of Theorem 2.2 is very short thanks to Lemma 4.2.

Let $\varphi(z)$ be as assumed in Theorem 2.2. Then there exist complex ζ_0 and z_0 such that $|\zeta_0| = 1$, $\varphi(z_0) = \zeta_0$, $\varphi'(z_0) \neq 0$. We choose $r > 0$ so small that

$\varphi(z)$ has no poles for $|z - z_0| < r$, and we define $f(z) = \varphi(z)$ (for $|z - z_0| < r$). Now $f(z)$ satisfies (4.3a,b), and the corresponding set S_f, defined in (4.3c), is contained in the stability region S of $\varphi(z)$. Clearly, $\mathbb{C} \setminus S \subset \mathbb{C} \setminus S_f$ and $\mathrm{d}(z, S_f) \geq \mathrm{d}(z, S)$ (for $z \in \mathbb{C} \setminus S$).

By applying Lemma 4.2, and using the last inclusion and inequality, we easily arrive at the conclusion stated in Theorem 2.2. □

4.4 Proof of Theorem 3.4

Part 1

The proof of Theorem 3.4 consists of two parts. In this first part, we shall prove the existence of a complex function $f(z)$ and of $z_0 \in \mathbb{C}$, $r > 0$ with the following properties: (4.3a,b) hold, the set S_f given by (4.3c) is contained in the stability region S of process (3.1), and

$$\rho(f(z)) - z\sigma(f(z)) = 0 \quad \text{whenever } |z - z_0| < r. \tag{4.9}$$

Here $\rho(\zeta)$, $\sigma(\zeta)$ denote the polynomials defined in Section 3.1. Henceforth, we shall assume that S satisfies the assumptions of Theorem 3.4.

Let Γ denote the boundary (in $\overline{\mathbb{C}}$) of S and consider the function $g(\zeta)$, with values in $\overline{\mathbb{C}}$, given by

$$g(\zeta) = \frac{\rho(\zeta)}{\sigma(\zeta)} \quad (\text{for } \zeta \in \mathbb{C}).$$

The assumptions on S imply that Γ does not consist of a finite number of points. Further, it can be seen that $\Gamma \subset \{g(\mathrm{e}^{\mathrm{i}t}) : t \in [0, 2\pi]\}$. We remark that $g(\mathrm{e}^{\mathrm{i}t})$ $(0 \leq t \leq 2\pi)$ is the so-called *root locus curve* of (3.1), see, e.g., Hairer and Wanner [6]. From the above, it directly follows that there exists a sequence of mutually different points on the unit circle, say $\mathrm{e}^{\mathrm{i}t_n}$, with real t_n $(n = 1, 2, 3, \ldots)$, such that

$$z_n := g(\mathrm{e}^{\mathrm{i}t_n}) \in \Gamma \quad \text{for } n \geq 1.$$

The set $\{\mathrm{e}^{\mathrm{i}t_n}\}_{n=1}^{\infty}$ has an accumulation point, which we denote by $\mathrm{e}^{\mathrm{i}t_0}$. Without loss of generality we can assume that either $t_n \uparrow t_0$ or $t_n \downarrow t_0$ (for $n \to \infty$). For ease of presentation we assume the latter limit behaviour to be present. We define

$$z_0 = g(\zeta_0), \quad \zeta_0 = \mathrm{e}^{\mathrm{i}t_0}.$$

Since S is closed, we have $z_n \in S$ for $n \geq 0$.

In the following, we shall prove that there exists a value $\tau > 0$ such that

$$g(\mathrm{e}^{\mathrm{i}t}) \in S \quad \text{whenever } \ t \in [t_0, t_0 + \tau). \tag{4.10}$$

After having shown (4.10) we shall easily establish the existence of a function $f(z)$ as mentioned above.

Assume that $z_0 \in \mathbb{C}$. (We shall shortly discuss the case $z_0 = \infty$ at the end of this part.) Since $z_0 \in S \cap \mathbb{C}$, we have that $\alpha_k - z_0\beta_k \neq 0$ and the polynomial $\rho(\zeta) - z_0\sigma(\zeta)$ satisfies the root condition, see (3.3a). Denote by $\zeta_1, \zeta_2, \ldots, \zeta_k$ the roots of $\rho(\zeta) - z_0\sigma(\zeta) = 0$, ordered such that $\zeta_1 = \zeta_0$, $|\zeta_i| = 1$ (for $1 \leq i \leq l$) and

$|\zeta_i| < 1$ (for $l < i \leq k$) with certain integer $l \geq 1$. The roots ζ_i for $i = 1, 2, \ldots, l$ are all simple. Hence, by virtue of a well known result from complex function theory, there is an open neighborhood $X \subset \mathbb{C}$ of z_0 and for each $i = 1, 2, \ldots, l$ a complex function $f_i(z)$ with the following properties:

(4.11a) $f_i(z)$ is holomorphic for $z \in X$;

(4.11b) $f_i(z_0) = \zeta_i$;

(4.11c) $\rho(f_i(z)) - z\sigma(f_i(z)) = 0$ for $z \in X$.

Furthermore, we can assume that X is so small that

(4.12) if $z \in X$, then: $z \in S \iff |f_i(z)| \leq 1$ for $1 \leq i \leq l$.

In order to arrive at (4.10), for some $\tau > 0$, we choose $\delta \in (0, 1]$ such that

$$g(\mathrm{e}^{\mathrm{i}u}) \in X \quad \text{whenever } u \in \mathbb{C},\ |u - t_0| < \delta.$$

Let $i \in \{1, 2, \ldots, l\}$. With an eye to (4.12), we consider the complex function $E_i(u)$ given by

$$E_i(u) = f_i(g(\mathrm{e}^{\mathrm{i}u})) \quad \text{for } u \in \mathbb{C},\ |u - t_0| < \delta.$$

In view of (4.11a), the function $E_i(u)$ is holomorphic (for $|u - t_0| < \delta$). Let integer $N \geq 1$ be such that $|t_n - t_0| < \delta$ for $n \geq N$. Clearly, if $n \geq N$, then $z_n = g(\mathrm{e}^{\mathrm{i}t_n}) \in X$ and $E_i(t_n) = f_i(z_n)$. Next, application of (4.12) leads to the bound

$$|E_i(t_n)| \leq 1 \quad \text{for } n \geq N. \tag{4.13}$$

We will show that there exists a $\tau > 0$ such that

$$|E_i(t)| \leq 1 \quad \text{whenever } t \in \mathbb{R},\quad t_0 \leq t < t_0 + \tau. \tag{4.14}$$

From the fact that $E_i(u)$ is holomorphic, it readily follows that $|E_i(t)|^2$ can be expressed as a convergent series,

$$|E_i(t)|^2 = \sum_{j=0}^{\infty} e_j\,(t - t_0)^j \quad \text{for } t \in \mathbb{R},\ |t - t_0| < \delta$$

with certain real coefficients e_j ($j = 0, 1, 2, \ldots$). In particular, $e_0 = |E_i(t_0)|^2 = |f_i(z_0)|^2 = |\zeta_i|^2 = 1$, cf. (4.11b). If $e_j = 0$ for $j \geq 1$, then we immediately have (4.14), with $\tau = \delta$. Assume next that there is an integer $J \geq 1$ such that $e_J \neq 0$ and $e_j = 0$ for $1 \leq j \leq J - 1$. We have $|E_i(t)|^2 = 1 + e_J\,(t - t_0)^J\,(1 + \mathcal{O}(t - t_0))$ (for $t \to t_0$). Upon invoking (4.13), and using that $t_n \downarrow t_0$ ($n \to \infty$), we easily obtain from our last expression for $|E_i(t)|^2$ that $e_J \leq 0$ and, subsequently, that there is a $\tau \in (0, \delta]$ such that (4.14) is fulfilled. We conclude, in view of (4.12),

that there exists a $\tau \in (0, \delta]$ such that (4.10) holds. Replacing t_0 and τ by $t_0 + \tau/2$ and $\tau/2$, respectively, we can assume that

$$g(\mathrm{e}^{\mathrm{i}t}) \in S \quad \text{for} \quad t_0 - \tau < t < t_0 + \tau. \tag{4.15}$$

We define the function $f(z)$ by

$$f(z) = f_1(z) \quad \text{for} \quad |z - z_0| < r$$

with a suitable $r > 0$. Consider the disk $Y = \{\zeta : |\zeta - \zeta_0| < \varepsilon\}$ where $\varepsilon = |\mathrm{e}^{\mathrm{i}(t_0+\tau)} - \zeta_0| \in (0, 1)$. Then we let $r > 0$ be such that

$$z \in X \quad \text{and} \quad f_1(z) \in Y \quad \text{whenever } |z - z_0| < r.$$

From the above it readily follows that $f(z)$ possesses the properties as formulated at the beginning of this part. Firstly, from (4.11a,b) and the fact that $f_1(z_0)$ is a simple root, we immediately obtain (4.3a,b). Next, by (4.11c), we directly have (4.9). Finally, using (4.11c), it is easily seen that the set S_f given by (4.3c) satisfies $S_f \subset \{g(\mathrm{e}^{\mathrm{i}t}) : t \in (t_0 - \tau, t_0 + \tau)\}$. In view of (4.15), we thus have $S_f \subset S$. This completes the proof if $z_0 \in \mathbb{C}$.

The proof in the case $z_0 = \infty$ runs along the same lines as above; we prove again that there exists a $\tau \in (0, \delta]$ such that (4.10) holds. For the analysis, it is now convenient to use the complex variable $w = z^{-1}$ instead of z. In this way, we can easily arrange for the domain of $f(z)$ to be contained in $\mathbb{C}$.

Part 2

In this second part, we shall apply Lemma 4.2 and subsequently complete the proof of Theorem 3.4.

Let $f(z)$, z_0 and r be as defined in Part 1 above and $\|\cdot\| = \|\cdot\|_\infty$. We consider an application of Lemma 4.2, where we take radius $r/2$ instead of r (this will turn out to be useful later on). Thus, there exist positive constants γ_0, K and $s \times s$ matrices $A_{s,n}$ such that, for all $n \geq 1$ and $s \geq 1$, conditions (4.5a′) and (4.5b) hold, with

(4.5a′)

$$(zI - A_{s,n}) \text{ is invertible, and } \|(zI - A_{s,n})^{-1}\| \leq \frac{K}{\mathrm{d}(z, S'_f)} \quad \text{for all} \quad z \in \mathbb{C} \setminus S'_f \ ,$$

where $S'_f = \{z : |z - z_0| < r/2 \text{ and } |f(z)| = 1\}$.

In view of Part 1 we have $S'_f \subset S$. Using this inclusion, it directly follows from (4.5a′) that condition (3.7a) of Theorem 3.4 is fulfilled, with $h = 1$ (cf. also Section 4.3). Consequently, it remains to show that there exists a constant $c > 0$ such that the lower bound (3.7b) of Theorem 3.4 is valid.

Let $n \geq 1$, $s \geq 1$ and write $A = A_{s,n}$. Let the holomorphic function $F(z)$ be defined by $F(z) = \rho(f(z)) - z\sigma(f(z))$ for $|z - z_0| < r$. Condition (4.5a′) implies that the spectrum of the matrix A is contained in the disk $|z - z_0| < r/2$. In view of (4.9), we thus have $F(A) = O$, i.e.

$$\sum_{i=0}^{k} (\alpha_i I - \beta_i A) f(A)^i = O.$$

Consequently, for any given $u_0 \in \mathbb{C}^s$, the vectors u_m, defined by $u_m = f(A)^m u_0$, satisfy the recurrence relation

$$\sum_{i=0}^{k} \alpha_i u_{m+i} = A \sum_{i=0}^{k} \beta_i u_{m+i} \quad (\text{for } m \geq 0).$$

The matrix $\alpha_k I - \beta_k A$ is invertible (since the spectrum of A lies in S) and, for $m \geq 0$, it follows that $U_m = \Phi(A)^m U_0$, where $U_m = (u_{m+k-1}^{\mathrm{T}}, \ldots, u_{m+1}^{\mathrm{T}}, u_m^{\mathrm{T}})^{\mathrm{T}}$ (see Section 3.1). By considering arbitrary $u_0 \in \mathbb{C}^s$, we easily arrive at the lower bound

$$\|\Phi(A)^n\| \;\geq\; \left(\max_{0 \leq i \leq k-1} \|f(A)\|^i \right)^{-1} \cdot \|f(A)^n\|. \tag{4.16}$$

Since

$$f(A) = \frac{1}{2\pi \mathrm{i}} \int f(z)(zI - A)^{-1}\, \mathrm{d}z,$$

where the integral is along the positively oriented curve $|z - z_0| = 3r/4$, we obtain

$$\|f(A)\| \leq \frac{3}{4} r \cdot M_0 \cdot M_1.$$

Here

$$M_0 = \max_{|z - z_0| = 3r/4} |f(z)| < \infty, \quad M_1 = \max_{|z - z_0| = 3r/4} \|(zI - A)^{-1}\| \leq \frac{K}{r/4}\ ,$$

where the last inequality holds by (4.5a′). Hence,

$$\|f(A)\| \leq 3KM_0. \tag{4.17}$$

By a combination of (4.5b), (4.16) and (4.17), we (immediately) find that there exists a constant $c > 0$, independent of n and s, such that condition (3.7b) holds (with $h = 1$). This completes the proof of Theorem 3.4. □

5 Proving upper bounds.

5.1 Proof of Theorem 3.2

Let $\|B\|$ denote a norm for matrices B of order m, which is induced by any vector norm in $\mathbb{C}^m$; and suppose $\tau > 1$, $L > 0$ are given constants. Assume B is an $m \times m$ matrix, with all eigenvalues in the closed unit disk, such that $\|(\zeta I - B)^{-1}\| \leq L(|\zeta| - 1)^{-1}$ (for all $\zeta \in \mathbb{C}$ with $1 < |\zeta| < \tau$). This assumption is well known to imply the upper bound $\|B^n\| \leq \gamma L \min\{m, n\}$ (for $n \geq 1$), with γ only depending on τ (see, e.g., Spijker [17]). Therefore, Theorem 3.2 is an immediate consequence of the following lemma.

LEMMA 5.1. *Assume the stability region S of process* (3.1) *is closed in* $\overline{\mathbb{C}}$. *Then there exist constants c and $\tau > 1$, only depending on the α_i, β_i, such that $(\zeta I - \Phi(hA))$ is invertible (for $|\zeta| > 1$) and*

$$\|(\zeta I - \Phi(hA))^{-1}\| \leq \frac{cK}{|\zeta| - 1} \quad (\text{for } 1 < |\zeta| < \tau) \tag{5.1}$$

whenever $h > 0$, $K \geq 1$, $\|\cdot\|$ is any norm given by (3.2a, b) and A is an arbitrary $s \times s$ matrix satisfying (2.4).

In the remainder of this subsection we shall prove Lemma 5.1.

Assume henceforth that S is closed in $\overline{\mathbb{C}}$ and that h, K, $\|\cdot\|$, A are given as in Lemma 5.1. Consider the $k \times k$ matrix-valued function $R(\zeta, z)$, with entries $r_{lm}(\zeta, z)$, defined by

$$R(\zeta, z) = (r_{lm}(\zeta, z)) = (\zeta I - \Phi(z))^{-1}$$

for $\zeta \in \mathbb{C}$, $z \in \mathbb{C}$ such that $(\zeta I - \Phi(z))$ is invertible. Since the spectrum of hA is contained in S, the matrix $(\zeta I - \Phi(hA))$ is invertible and

$$(\zeta I - \Phi(hA))^{-1} = R(\zeta, hA) = (r_{lm}(\zeta, hA)) \qquad \text{for } |\zeta| > 1.$$

By a compactness argument, and using (3.2c), one easily sees that the following lemma is valid.

LEMMA 5.2. *Assume that to each l and m (with $1 \leq l \leq k$, $1 \leq m \leq k$) and each ζ_0 (with $|\zeta_0| = 1$) there exist positive constants c_0, ε_0, only depending on the α_i, β_i, with*

$$\|r_{lm}(\zeta, hA)\| \leq \frac{c_0 K}{|\zeta| - 1} \quad \text{for } |\zeta - \zeta_0| < \varepsilon_0, \quad |\zeta| > 1. \tag{5.2}$$

Then there are constants c and $\tau > 1$, only depending on the α_i, β_i, such that (5.1) holds.

Below, the resolvent bound (5.1) will be proved by showing that the assumption of Lemma 5.2 is fulfilled. We shall make use of the following known result (see Dahlquist, Mingyou and LeVeque [3] or Hairer and Wanner [6, Sect. V.7]).

LEMMA 5.3. *There is a constant c_1 (independent of l, m, ζ, z) such that*

$$|r_{lm}(\zeta, z)| \leq \frac{c_1}{|\zeta| - 1} \quad \text{for } 1 \leq l \leq k, \quad 1 \leq m \leq k, \ |\zeta| > 1, \quad z \in S.$$

It can be verified, in a straightforward manner, that each $r_{lm}(\zeta, z)$ is of the form

$$r_{lm}(\zeta, z) = \frac{p_{lm}(\zeta) - q_{lm}(\zeta) z}{\rho(\zeta) - \sigma(\zeta) z} \tag{5.3}$$

with certain polynomials $p_{lm}(\zeta)$, $q_{lm}(\zeta)$. Consequently, we have

$$r_{lm}(\zeta, z) = a_{lm}(\zeta) + b_{lm}(\zeta, z) \tag{5.4a}$$

with

$$a_{lm}(\zeta) = \frac{q_{lm}(\zeta)}{\sigma(\zeta)}, \quad b_{lm}(\zeta, z) = \frac{p_{lm}(\zeta) - q_{lm}(\zeta) g(\zeta)}{\rho(\zeta) - \sigma(\zeta) z}, \quad g(\zeta) = \frac{\rho(\zeta)}{\sigma(\zeta)}. \tag{5.4b}$$

In order to prove that the assumption of Lemma 5.2 is fulfilled, we assume l, m and ζ_0 to be given with $|\zeta_0| = 1$. We shall treat two cases separately.

Case 1. $\sigma(\zeta_0) \neq 0$ *or* $\sigma'(\zeta_0) \neq 0$. Assume ζ is such that $|\zeta| > 1$, $\sigma(\zeta) \neq 0$. Then, according to (5.4), we have

$$r_{lm}(\zeta, hA) = a_{lm}(\zeta)I + \frac{p_{lm}(\zeta) - q_{lm}(\zeta)g(\zeta)}{\sigma(\zeta)}(g(\zeta)I - hA)^{-1}.$$

Applying (2.4), we obtain

$$\begin{aligned}\|r_{lm}(\zeta, hA)\| &\leq |a_{lm}(\zeta)| + \frac{|p_{lm}(\zeta) - q_{lm}(\zeta)g(\zeta)|}{|\sigma(\zeta)| \cdot |g(\zeta) - x(\zeta)|} \cdot K \\ &= |a_{lm}(\zeta)| + |b_{lm}(\zeta, x(\zeta))| \cdot K \\ &\leq \Big(2|a_{lm}(\zeta)| + |r_{lm}(\zeta, x(\zeta))|\Big) \cdot K,\end{aligned}$$

where $x(\zeta) \in \mathbb{C} \cap S$ is such that $\mathrm{d}(g(\zeta), S) = |g(\zeta) - x(\zeta)|$. Consequently, in view of Lemma 5.3, we get

$$\|r_{lm}(\zeta, hA)\| \leq \left(2\frac{|q_{lm}(\zeta)|}{|\sigma(\zeta)|} + \frac{c_1}{|\zeta| - 1}\right) \cdot K.$$

Since $\sigma(\zeta_0) \neq 0$, or $\sigma(\zeta_0) = 0$ and $\sigma'(\zeta_0) \neq 0$, there is an $\varepsilon_0 > 0$ such that

$$\sup\left\{\frac{|\zeta| - 1}{|\sigma(\zeta)|} : \ \ |\zeta - \zeta_0| < \varepsilon_0 \text{ and } |\zeta| > 1\right\} < \infty.$$

It follows that a finite c_0 exists with property (5.2).

Case 2. $\sigma(\zeta_0) = \sigma'(\zeta_0) = 0$. In view of (3.3b), we have $\infty \notin S$. Since S is closed in $\overline{\mathbb{C}}$, we conclude that S is a bounded subset of $\mathbb{C}$. We shall use the formula

$$r_{lm}(\zeta, hA) = \frac{1}{2\pi\mathrm{i}} \int r_{lm}(\zeta, z)(zI - hA)^{-1}\,\mathrm{d}z,$$

where the integration is over a positively oriented circle with center 0 and radius R. We choose R sufficiently large so as to ensure that $R - 1 > \max\{|x| : \ x \in S\}$. Since $\sigma(\zeta_0) = 0$ and $\rho(\zeta_0) \neq 0$, formula (5.3) yields that we can choose $\varepsilon_0 \in (0, 1)$ so small that

$$M = \sup\{|r_{lm}(\zeta, z)| : \ |\zeta - \zeta_0| < \varepsilon_0, \ |z| \leq R\} < \infty.$$

With these choices for R and ε_0, the above integral formula is valid whenever $|\zeta - \zeta_0| < \varepsilon_0$. We thus find $\|r_{lm}(\zeta, hA)\| \leq RMK \leq RMK/(|\zeta| - 1)$ whenever $|\zeta - \zeta_0| < \varepsilon_0, \ |\zeta| > 1$, which proves (5.2) with $c_0 = R\,M$.

5.2 Proof of Theorem 3.3

Assume that S consists of a finite number of points and that h, K, $\|\cdot\|$, A are given as in Theorem 3.3.

Let $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_m\}$ denote the set of all distinct eigenvalues of hA. From the Jordan canonical form of hA, we have a representation

$$hA = \sum_{j=1}^{m} [\lambda_j P_j + R_j], \tag{5.5}$$

where P_j, R_j are $s \times s$ matrices with $P_1 + P_2 + \ldots + P_m = I$, $P_i P_j = O$ $(i \neq j)$, $P_j P_j = P_j$, $R_j P_j = P_j R_j = R_j$ and $(R_j)^s = O$. We shall prove that

$$R_i = O \quad \text{and} \quad \|P_i\| \leq K \quad (1 \leq i \leq m). \tag{5.6}$$

One easily sees that the resolvent of hA, for $z \in \mathbb{C} \setminus \Lambda$, can be expressed as follows:

$$(zI - hA)^{-1} = \sum_{j=1}^{m} (z - \lambda_j)^{-1} [P_j + (z - \lambda_j)^{-1} R_j + \ldots + (z - \lambda_j)^{-s+1} (R_j)^{s-1}].$$

From (2.4) it follows that $\Lambda \subset S$. Consequently, for z sufficiently close to any given λ_i, we have $\mathrm{d}(z, S) = |z - \lambda_i|$ and $\|(z - \lambda_i)(zI - hA)^{-1}\| \leq K$. Inserting the above expression for the resolvent into the left-hand member of the last inequality, we obtain

$$\|P_i + F_i(z) + \sum_{j \neq i} (z - \lambda_i)(z - \lambda_j)^{-1} [P_j + F_j(z)]\| \leq K,$$

where $F_l(z) \equiv (z - \lambda_l)^{-1} R_l + \ldots + (z - \lambda_l)^{-s+1} (R_l)^{s-1}$ (for $1 \leq l \leq m$). By letting $z \to \lambda_i$, we see that (5.6) must be valid.

We are now ready to prove the existence of a constant γ as in Theorem 3.3. Using (5.5), (5.6) and denoting the Kronecker product by $\otimes$, we have

$$\|\Phi(hA)^n\| = \|\textstyle\sum_{j=1}^{m} \Phi(\lambda_j)^n \otimes P_j\| \leq \sum_{j=1}^{m} \|\Phi(\lambda_j)^n\|_\infty \cdot \|P_j\|.$$

Therefore, $\|\Phi(hA)^n\| \leq \gamma K$, with $\gamma = \sup_{n \geq 1} \sum_{z \in S} \|\Phi(z)^n\|_\infty < \infty$. This completes the proof of Theorem 3.3. □

Acknowledgement.

The authors remember, with pleasure, useful discussions with Dr. N. Borovykh about some of the topics in this paper.

REFERENCES

1. N. Borovykh and M. N. Spijker, *The sharpness of stability estimates corresponding to a general resolvent condition*, Linear Algebra Applic., 311 (2000), pp. 161–175.

2. J. C. Butcher, *The Numerical Analysis of Ordinary Differential Equations*, John Wiley, Chichester, 1987.
3. G. Dahlquist, H. Mingyou, and R. LeVeque, *On the uniform power-boundedness of a family of matrices and the applications to one-leg and linear multistep methods*, Numer. Math., 42 (1983), pp. 1–13.
4. J. L. M. van Dorsselaer, J. F. B. M. Kraaijevanger, M. N. Spijker, *Linear stability analysis in the numerical solution of initial value problems*, Acta Numerica, 1993 (1993), pp. 199–237.
5. D. F. Griffiths, I. Christie, and A. R. Mitchell, *Analysis of error growth for explicit difference schemes in conduction-convection problems*, Int. J. Numer. Meth. Engin., 15 (1980), pp. 1075–1081.
6. E. Hairer and G. Wanner, *Solving Ordinary Differential Equations*, Vol. II, 2nd ed., Springer-Verlag, Berlin, 1996.
7. J. F. B. M. Kraaijevanger, *Two counterexamples related to the Kreiss matrix theorem*, BIT, 34 (1994), pp. 113–119.
8. H. W. J. Lenferink and M. N. Spijker, *On the use of stability regions in the numerical analysis of initial value problems*, Math. Comput., 57 (1991), pp. 221–237.
9. C. Lubich and O. Nevanlinna, *On resolvent conditions and stability estimates*, BIT, 31 (1991), pp. 293–313.
10. C. A. McCarthy and J. Schwartz, *On the norm of a finite Boolean algebra of projections, and applications to theorems of Kreiss and Morton*, Comm. Pure Appl. Math., 18 (1965), pp. 191–201.
11. K. W. Morton, *Stability of finite difference approximations to a diffusion-convection equation*, Int. J. Numer. Meth. Engin., 15 (1980), pp. 677–683.
12. S. V. Parter, *Stability, convergence, and pseudo-stability of finite-difference equations for an over-determined problem*, Numer. Math., 4 (1962), pp. 277–292.
13. S. C. Reddy and L. N. Trefethen, *Lax-stability of fully discrete spectral methods via stability regions and pseudo-eigenvalues*, Comp. Meth. Appl. Mech. Engin., 80 (1990), pp. 147–164.
14. S. C. Reddy and L. N. Trefethen, *Stability of the method of lines*, Numer. Math. 62 (1992), pp. 235–267.
15. R. D. Richtmyer and K. W. Morton, *Difference Methods for Initial-Value Problems*, 2nd ed., John Wiley, New York, 1967.
16. M. N. Spijker, *Stepsize restrictions for stability of one-step methods in the numerical solution of initial value problems*, Math. Comput., 45 (1985), pp. 377–392.
17. M. N. Spijker, *Numerical stability, resolvent conditions and delay differential equations*, Appl. Numer. Math., 24 (1997), pp. 233–246.
18. M. N. Spijker and F. A. J. Straetemans, *Stability estimates for families of matrices of nonuniformly bounded order*, Linear Algebra Appl., 239 (1996), pp. 77–102.
19. M. N. Spijker and F. A. J. Straetemans, *Error growth analysis via stability regions for discretizations of initial value problems*, BIT, 37 (1997), pp. 442–464.
20. M. N. Spijker, S. Tracogna, and B. D. Welfert, *About the sharpness of the stability estimates in the Kreiss matrix theorem*, Math. Comp., 72 (2003), pp. 697–713.
21. K. C. Toh and L. N. Trefethen, *The Kreiss matrix theorem on a general complex domain*, SIAM J. Matrix Anal. Appl., 21 (1999), pp. 145–165.