Acta Numerica (1993), *pp.* 199–237

Linear stability analysis in the numerical solution of initial value problems

J.L.M. van Dorsselaer
J.F.B.M. Kraaijevanger
M.N. Spijker
Department of Mathematics and Computer Science,
University of Leiden
The Netherlands
E-mail: spijker@rulcri.leidenuniv.nl

This article addresses the general problem of establishing upper bounds for the norms of the nth powers of square matrices. The focus is on upper bounds that grow only moderately (or stay constant) when n, or the order of the matrices, increases. The so-called resolvent condition, occurring in the famous Kreiss matrix theorem, is a classical tool for deriving such bounds.

Recently the classical upper bounds known to be valid under Kreiss's resolvent condition have been improved. Moreover, generalizations of this resolvent condition have been considered so as to widen the range of applications. The main purpose of this article is to review and extend some of these new developments.

The upper bounds for the powers of matrices discussed in this article are intimately connected with the stability analysis of numerical processes for solving initial(-boundary) value problems in ordinary and partial linear differential equations. The article highlights this connection.

The article concludes with numerical illustrations in the solution of a simple initial-boundary value problem for a partial differential equation.

CONTENTS

1. Introduction

1.1. Linear stability analysis

This article deals with step-by-step methods for the numerical solution of linear differential equations. Both initial-boundary value problems in partial differential equations and initial value problems in ordinary differential equations will be included in our considerations.

A crucial question in the step-by-step solution of such problems is whether the method will behave *stably* or not. Here we use the term stability to designate that any numerical errors, introduced at some stage of the calculations, are propagated in a mild fashion – i.e. do not blow up in the subsequent steps of the methods.

Classical tools to assess the stability *a priori*, in the numerical solution of partial differential equations, include *Fourier transformation* and the corresponding famous Von Neumann condition (see the classical work by Richtmyer and Morton (1967)). Further tools of recognized merit for assessing stability, in the solution of ordinary differential equations, comprise the so-called *stability regions* in the complex plane (see e.g. the excellent works by Butcher (1987) and Hairer and Wanner (1991)). During the last 25 years these stability regions have been studied extensively; numerous papers have appeared dealing with the shape and various peculiarities of these regions.

However, these tools are based on the behaviour that the numerical method would have when applied to quite simple test problems. Accordingly, in the case of partial differential equations, Fourier transformation provides a straightforward and reliable stability criterion primarily only for *pure initial value* problems in linear differential equations with *constant* coefficients. Similarly, in the case of ordinary differential equations, stability regions are primarily relevant only to *scalar* equations

$$U'(t) = \lambda U(t) \quad \text{for } t \geq 0, \tag{1.1}$$

with given complex constant λ.

In the pioneering work by F. John (1952) the scope of Fourier transformation had already been widened in that it was used in deriving sufficient conditions for stability in the numerical solution of linear partial differential equations with *variable* coefficients. For subsequent related work, relevant to equations with variable coefficients and to *initial-boundary value* problems, the reader may consult Richtmyer and Morton (1967), Kreiss (1966), Gustafsson, Kreiss and Sundström (1972), Meis and Marcowitz (1981), Thomée (1990), and the references therein.

Clearly, rigorous stability criteria with a wider scope than the simple classical test equations are important – both from a practical and a theoretical point of view. It is equally important to know to what extent stability regions can be relied upon in assessing stability in the numerical solution of

differential equations more general than (1.1). The present article reviews and extends some recent developments which are relevant to these two questions. No essential use will be made of Fourier transformation.

1.2. Stability and power boundedness

In this paper we shall deal with numerical processes of the form

$$u_n = Bu_{n-1} + b_n \quad \text{for } n = 1, 2, 3, \ldots, \tag{1.2}$$

with a given square matrix B of order $s \geq 1$ and given s-dimensional vectors b_n. The s-dimensional vectors u_n are computed sequentially from (1.2) starting from a given vector u_0. Processes of the form (1.2) occur in the numerical solution of linear initial value problems that are essentially more general than the simple classical test problems mentioned earlier. The vectors u_n provide numerical approximations to the true solution of the initial (-boundary) value problem under consideration.

As an illustration of (1.2) we consider the initial-boundary value problem

$$\begin{aligned} &u_t(x,t) = a(x)u_{xx}(x,t) + b(x)u_x(x,t) + c(x)u(x,t) + d(x), \\ &u_x(0,t) = 0, \quad u(1,t) = g(t), \\ &u(x,0) = f(x), \end{aligned} \tag{1.3}$$

where $0 < x < 1$, $t > 0$ and a, b, c, d, f, g are given functions with

$$a(x) > 0, \quad b(x) \geq 0, \quad c(x) \leq 0.$$

We choose $\Delta t = h > 0$, $\Delta x = 1/s$ and consider the approximation of $u(j/s, nh)$ by quantities u_j^n. The following finite difference scheme has been constructed by standard principles (see Richtmyer and Morton (1967)):

$$\begin{aligned} &h^{-1}(u_j^n - u_j^{n-1}) = \\ &\quad s^2 a(j/s)\{\theta(u_{j-1}^n - 2u_j^n + u_{j+1}^n) + (1-\theta)(u_{j-1}^{n-1} - 2u_j^{n-1} + u_{j+1}^{n-1})\} \\ &\quad + sb(j/s)\{\theta(u_{j+1}^n - u_j^n) + (1-\theta)(u_{j+1}^{n-1} - u_j^{n-1})\} \\ &\quad + c(j/s)\{\theta u_j^n + (1-\theta)u_j^{n-1}\} + d(j/s), \\ &u_{-1}^{n-1} = u_1^{n-1}, \quad u_s^{n-1} = g((n-1)h), \\ &u_j^0 = f(j/s), \end{aligned}$$

where $j = 0, 1, \ldots, s-1$ and $n = 1, 2, 3, \ldots$. θ denotes a parameter, with $0 \leq \theta \leq 1$, specifying the finite difference method.

Defining vectors u_n by

$$u_n = \begin{pmatrix} u_0^n \\ u_1^n \\ \vdots \\ u_{s-1}^n \end{pmatrix} \simeq \begin{pmatrix} u(0, nh) \\ u(1/s, nh) \\ \vdots \\ u((s-1)/s, nh) \end{pmatrix},$$

one easily verifies that the u_n satisfy a relation of the form (1.2). Here the matrix B satisfies

$$B = (I + (1-\theta)hA)(I - \theta hA)^{-1}, \tag{1.4}$$

where I denotes the $s \times s$ identity matrix and $A = (\alpha_{jk})$ an $s \times s$ tridiagonal matrix with its (nonzero) entries given by

$$\begin{array}{lcll} \alpha_{j+1,j+1} & = & -2s^2a(j/s) - sb(j/s) + c(j/s) & (0 \le j \le s-1), \\ \alpha_{j+1,j} & = & s^2a(j/s) & (1 \le j \le s-1), \\ \alpha_{j+1,j+2} & = & s^2a(j/s) + sb(j/s) & (1 \le j \le s-2), \\ \alpha_{j+1,j+2} & = & 2s^2a(j/s) + sb(j/s) & (j = 0). \end{array} \tag{1.5}$$

Suppose the numerical calculations based on the general process (1.2) were performed using a perturbed starting vector $\widetilde{u}_0$, instead of u_0. We would then obtain approximations that we denote by $\widetilde{u}_n$. For instance $\widetilde{u}_0$ may stand for a finite-digit representation in a computer of the true u_0, and the $\widetilde{u}_n$ then stand for the numerical approximations obtained in the presence of the rounding error $v_0 = \widetilde{u}_0 - u_0$.

In the stability analysis of (1.2) the crucial question is whether the difference $v_n = \widetilde{u}_n - u_n$ (for $n \ge 1$) can be bounded suitably in terms of the perturbation $v_0 = \widetilde{u}_0 - u_0$. Since

$$v_n = \widetilde{u}_n - u_n = [B\widetilde{u}_{n-1} + b_n] - [Bu_{n-1} + b_n]$$

we have

$$v_n = Bv_{n-1},$$

and consequently

$$v_n = B^n v_0.$$

The last expression makes clear that a central issue in stability analysis is the question of whether given matrices have powers that are uniformly bounded. Accordingly, in the following we focus, for an arbitrary $s \times s$ matrix B, on the stability property

$$\|B^n\| \le M_0 \quad \text{for } n = 0, 1, 2, \ldots, \tag{1.6}$$

where M_0 is a positive constant. For the time being $\|\cdot\|$ stands for the spectral norm, i.e. for the norm induced by the Euclidean vector norm in $\mathbb{C}^s$ (for an $s \times s$ matrix A we have $\|A\| = \max\{|Ax|/|x| : x \in \mathbb{C}^s$ with $x \neq 0\}$, where $|\cdot|$ denotes the Euclidean norm defined by $|x| = \sqrt{x^*x}$ with x^* standing for the Hermitian adjoint of the column vector $x \in \mathbb{C}^s$).

1.3. Power boundedness and the eigenvalue condition

For any given matrix B one can easily deduce from its Jordan canonical form (see, e.g., Horn and Johnson (1990)) a criterion for the existence of

an M_0 with property (1.6). A necessary and sufficient requirement for the existence of such an M_0 is the following *eigenvalue condition*:

> All eigenvalues λ of B have modulus $|\lambda| \leq 1$, and the geometric multiplicity of each eigenvalue with modulus 1 is equal to its algebraic multiplicity. (1.7)

However, in the stability analysis of numerical processes one is often interested in property (1.6) for all B belonging to some infinite family $\mathcal{F}$ of matrices. The crucial question then is whether a single finite M_0 exists such that (1.6) holds simultaneously for all B belonging to $\mathcal{F}$. In this situation, (1.7) may only provide a condition that is necessary (and not sufficient) for such an M_0 to exist.

For instance, in the example of Section 1.2 one can only expect great accuracy in the approximations u_j^n to $u(j/s, nh)$ when Δx (and Δt) become very small. Accordingly one is primarily interested in bounds for B^n that are uniformly valid for all B of the form (1.4), (1.5) with arbitrarily small $\Delta x = 1/s$.

An instructive counterexample, illustrating the fact that criterion (1.7) can be misleading for the case of families $\mathcal{F}$, is provided by the $s \times s$ bidiagonal matrices

$$B = \begin{pmatrix} -1/2 & 3/2 & & 0 \\ & -1/2 & \ddots & \\ & & \ddots & 3/2 \\ 0 & & & -1/2 \end{pmatrix}. \tag{1.8}$$

Matrices of the form (1.8) may be thought of as arising in the numerical solution of the initial-boundary value problem

$$u_t(x,t) = u_x(x,t),$$
$$u(1,t) = 0, \quad u(x,0) = f(x), \quad \text{where } 0 < x < 1,\ t > 0.$$

Consider the finite difference scheme

$$h^{-1}(u_j^n - u_j^{n-1}) = s(u_{j+1}^{n-1} - u_j^{n-1}),$$
$$u_s^{n-1} = 0, \quad u_j^0 = f(j/s).$$

Here $\Delta t = h > 0$, $\Delta x = 1/s$, and u_j^n approximates $u(j/s, nh)$ for $j = 0, 1, \ldots, s-1$ and $n = 1, 2, 3, \ldots$. Clearly, with the choice $hs = 3/2$ this finite difference scheme can be written in the form (1.2) with B as in (1.8).

For each $s \geq 1$ the matrix B defined by (1.8) satisfies the eigenvalue condition (1.7).

Defining the $s \times s$ shift matrix E by

$$E = \begin{pmatrix} 0 & 1 & & 0 \\ & 0 & \ddots & \\ & & \ddots & 1 \\ 0 & & & 0 \end{pmatrix}, \tag{1.9}$$

we have from (1.8) the expression

$$B = -\tfrac{1}{2}I + \tfrac{3}{2}E,$$

so that

$$B^n = \sum_{k=0}^{n} \binom{n}{k} (-\tfrac{1}{2})^{n-k} (\tfrac{3}{2})^k E^k.$$

Defining x to be the s-dimensional vector whose jth component equals $\xi_j = (-1)^j$, and denoting the jth component of $y = B^n x$ by η_j we easily obtain, from the above expression for B^n,

$$|\eta_j| = \sum_{k=0}^{n} \binom{n}{k} (\tfrac{1}{2})^{n-k} (\tfrac{3}{2})^k = 2^n \quad \text{for } 1 \le j \le s-n.$$

For $s > n$ we thus have

$$\left(\sum_{j=1}^{s} |\eta_j|^2 \right)^{1/2} \ge \sqrt{s-n}\, 2^n,$$

and since

$$\left(\sum_{j=1}^{s} |\xi_j|^2 \right)^{1/2} = \sqrt{s},$$

it follows that $\|B^n\| \ge \sqrt{1 - n/s}\, 2^n$. Denoting the $s \times s$ matrix B by B_s we thus have

$$\|(B_{2n})^n\| \ge 2^{n-1/2} \quad \text{for } n = 1, 2, 3, \ldots .$$

Clearly, no M_0 can exist such that (1.6) is valid for all B belonging to $\mathcal{F} = \{B_s : s = 1, 2, 3, \ldots\}$.

The fact that the eigenvalue criterion can be a misleading guide to stability was already known in the 1960s, see e.g. Parter (1962). A related, but stronger, necessary requirement for stability is the so-called *Godunov–Ryabenkii stability condition*, a discussion of which can be found, e.g., in Richtmyer and Morton (1967), Morton (1980) and Thomée (1990). The latter condition is not satisfied in example (1.8).

The earlier counterexample is similar to examples in Richtmyer and Morton (1967), Spijker (1985), Kreiss (1990) and Reddy and Trefethen (1992).

Further examples of instability under the eigenvalue condition (1.7) can be found in Griffiths, Christie and Mitchell (1980), Kraaijevanger, Lenferink and Spijker (1987) and Lenferink and Spijker (1991b). The matrices B in these references have s different eigenvalues λ with $|\lambda| < 1$, and occur in the numerical solution of problems of the form (1.3). See Trefethen (1988) and Reddy and Trefethen (1992) for related counterexamples in spectral methods.

We conclude this subsection with the remark that in some special cases the eigenvalue criterion can be reliable. For normal matrices B (i.e. $B^*B = BB^*$ with B^* denoting the Hermitian adjoint of B) the stability estimate (1.6) is valid with $M_0 = 1$ as soon as all eigenvalues of B have a modulus not exceeding 1 (see, e.g., Horn and Johnson (1990)). But, in general, one has to look for conditions different from (1.7).

1.4. Power boundedness and the resolvent condition

The famous *Kreiss matrix theorem* (see, e.g., Kreiss (1962) and Richtmyer and Morton (1967)) relates (1.6) to conditions on B which are more reliable than (1.7). One of these conditions involves the so-called *resolvent* $(\zeta I - B)^{-1}$ of B, and reads as follows:

$$\begin{aligned} &\zeta I - B \text{ is invertible and } \|(\zeta I - B)^{-1}\| \le M_1(|\zeta| - 1)^{-1} \\ &\text{for all complex numbers } \zeta \notin D. \end{aligned} \tag{1.10}$$

Here M_1 is a positive constant, I the $s \times s$ identity matrix and

$$D = \{\zeta : \zeta \in \mathbb{C} \text{ and } |\zeta| \le 1\}$$

the closed *unit disk* in the complex plane.

If (1.6) is satisfied, then all eigenvalues of B lie in D, so that for all $\zeta \notin D$ the matrix $\zeta I - B$ is invertible and

$$\|(\zeta I - B)^{-1}\| = \left\| \sum_{k=0}^{\infty} \zeta^{-k-1} B^k \right\| \le \sum_{k=0}^{\infty} |\zeta|^{-k-1} M_0 = M_0(|\zeta| - 1)^{-1}.$$

Hence (1.6) implies (1.10) with $M_1 = M_0$. The Kreiss matrix theorem asserts that, conversely, (1.10) implies (1.6) with M_0 depending only on M_1 and the dimension s, but otherwise independent of the matrix B.

The Kreiss matrix theorem has often been used in the stability analysis of numerical methods for solving initial value problems for partial differential equations. In the classical situation the matrices B are obtained by Fourier transformation of the numerical solution operators, and they stand essentially for the so-called amplification matrices (see, e.g., Richtmyer and Morton (1967)). These matrices are of a *fixed* finite order s. On the other hand, the implication of (1.6) by (1.10) can also be used without Fourier transformation, with B standing for the numerical solution operator in (1.2).

In this situation we may be dealing with a family of matrices B of finite – but not *uniformly* bounded – orders s. Therefore, of particular interest is the dependence of the stability constant M_0 in (1.6) on the dimension s (see also Tadmor (1981)).

Various authors (see, e.g., Morton (1964), Miller and Strang (1966), Tadmor (1981), LeVeque and Trefethen (1984) and Spijker (1991)) have studied the size of (the optimal) M_0 as a function of M_1 and s, and recently some open problems in this field were solved. Moreover, the implication of (1.6) by (1.10) as previously discussed has recently been generalized in several directions. More general norms than the spectral norm have been dealt with and the resolvent condition (1.10) has been adapted to domains different from the unit disk D. In the latter case the matrices B in (1.6) and (1.10) are not the same, but are related to each other by a given (rational) transformation.

1.5. Scope of the rest of the article

In the rest of this article we review and extend some of the recent results just mentioned, and illustrate them in the numerical solution of initial (-boundary) value problems.

In Section 2 we still deal with resolvent condition (1.10) with respect to the unit disk D, but we consider general norms on the vector space of all $s \times s$ matrices. In this situation we focus on the best upper bounds for $\|B^n\|$ that are possible under condition (1.10).

In Section 3.1 we relate estimates like (1.6) more explicitly to the stability analysis of numerical methods for the solution of ordinary and partial differential equations. In Section 3.2 we show that in this analysis it is useful to consider resolvent conditions with respect to regions $V \subset \mathbb{C}$ that are different from the unit disk D. The focus will be on regions V that are contained in the stability regions corresponding to the numerical methods under consideration. We give a review of stability estimates from the literature based on resolvent conditions with respect to such V. Section 3.3 provides various comments on these estimates. We confine our considerations throughout to so-called one-step methods. For related stability results pertinent to (linear) multistep methods we refer to Crouzeix (1987), Grigorieff (1991), Lubich (1991), Lubich and Nevanlinna (1991) and Reddy and Trefethen (1990, 1992).

Section 4 deals with various concepts and problems that are related to (generalized) resolvent conditions. In Section 4.1 the resolvent condition is related to the concept of ϵ-pseudospectra recently used by Trefethen and others (see e.g. Trefethen (1992) and Reddy and Trefethen (1990, 1992)). In Section 4.2 it is related to the so-called M-numerical range introduced by Lenferink and Spijker (1990). Part of the material presented here is used in some proofs given in Section 2. In Section 4.3 we consider the problem of

bounding the exponential function of a matrix under the assumption that the matrix satisfies a resolvent condition (with respect to the complex left half plane).

In Section 5 we focus on the range of applications of the stability results reviewed in Section 3. Moreover, a numerical illustration is presented involving the solution of a partial differential equation.

2. Stability estimates under resolvent conditions with respect to the unit disk

2.1. The classical situation for arbitrary $M_1 \geq 1$

We start by reviewing classical upper bounds for $\|B^n\|$ that were derived from (1.10), with $\|\cdot\|$ standing for the spectral norm.

As already mentioned in the introduction, the Kreiss matrix theorem asserts, for the spectral norm, that resolvent condition (1.10) implies power boundedness (1.6) with a stability constant M_0 depending only on M_1 and the dimension s. According to Tadmor (1981), the original proof by Kreiss (1962) yields an upper bound $\|B^n\| \leq M_0$ with

$$M_0 \simeq (M_1)^{s^s},$$

which is far from sharp. After successive improvements by various authors (Morton, 1964; Miller and Strang, 1966), it was Tadmor (1981) who succeeded in proving a bound that is linear in s,

$$\|B^n\| \leq 32e\pi^{-1}sM_1.$$

LeVeque and Trefethen (1984) lowered this upper bound to $2esM_1$, and conjectured that the latter bound can be improved further to

$$\|B^n\| \leq esM_1 \quad \text{for } n = 0, 1, 2, \ldots . \tag{2.1}$$

Moreover, these authors showed by means of a counterexample that the factor e in (2.1) cannot be replaced by any smaller constant if the upper bound is to be valid for arbitrary factors M_1 in (1.10) and arbitrarily large integers s.

Smith (1985) proved a result which, combined with the arguments of LeVeque and Trefethen (1984), leads to the improved upper bound $\|B^n\| \leq \pi^{-1}(\pi+2)esM_1$, which is an improvement over the upper bound $2esM_1$ but still weaker than conjecture (2.1). The conjecture was finally proved to be true by Spijker (1991) (see also Wegert and Trefethen (1992)).

In addition to the upper bound (2.1), which is linear in s and independent of n, it is possible to derive an upper bound from (1.10) that is linear in n and independent of s. By the Cauchy integral formula (see, e.g., Conway

(1985)) we have

$$B^n = \frac{1}{2\pi \mathrm{i}} \int_\Gamma \zeta^n (\zeta I - B)^{-1} \, \mathrm{d}\zeta, \tag{2.2}$$

where the contour of integration Γ is any positively oriented circle $|\zeta| = 1+\epsilon$ with $\epsilon > 0$. Choosing $\epsilon = 1/n$ it readily follows from (1.10) and (2.2) that

$$\|B^n\| \leq (1+1/n)^n (n+1) M_1 \leq e(n+1)M_1 \quad \text{for } n = 1, 2, 3, \ldots \tag{2.3}$$

(see also Reddy and Trefethen (1990) and Lubich and Nevanlinna (1991)).

In the next subsection we will discuss a generalization of the upper bounds (2.1), (2.3) to norms different from the spectral norm. We will also investigate the sharpness of these bounds in the general case.

2.2. Stability estimates for arbitrary $M_1 \geq 1$ and arbitrary norms

In this subsection we consider a generalization of the upper bounds (2.1), (2.3) to the case where $\|\cdot\|$ is an *arbitrary norm* on $\mathbb{C}^{s,s}$, the vector space of all complex $s \times s$ matrices. If the norm is submultiplicative (i.e. $\|AB\| \leq \|A\|\|B\|$ for all $A, B \in \mathbb{C}^{s,s}$) the norm is called a *matrix norm*. Norms for which $\|I\| = 1$ are called *unital*.

Theorem 2.1 Let $s \geq 1$, $B \in \mathbb{C}^{s,s}$ and $\|\cdot\|$ denote an arbitrary norm on the vector space $\mathbb{C}^{s,s}$.

(a) If (1.6) holds for some M_0, then (1.10) holds with $M_1 = M_0$;
(b) If (1.10) holds for some M_1, then

$$\|B^n\| \leq (1+1/n)^n \min(s, n+1) M_1 \quad \text{for } n = 1, 2, 3, \ldots . \tag{2.4}$$

Proof. 1. The proof of (a) is the same as the proof in Section 1.4 for the spectral norm. Since the proof of (2.3) as given in Section 2.1 also remains valid for arbitrary norms, the proof of (b) is complete if we can show that

$$\|B^n\| \leq (1+1/n)^n s M_1 \quad \text{for } n = 1, 2, 3, \ldots . \tag{2.5}$$

In order to prove (2.5) we now consider arbitrary but fixed $n \geq 1$ and B satisfying (1.10).

2. A well known corollary to the Hahn–Banach theorem (see, e.g., Chapter 3 in Rudin (1973), or Chapter 5 in Horn and Johnson (1990)) states that, corresponding to any normed vector space X and vector $y \in X$, there exists a linear transformation $F : X \to \mathbb{C}$ with

$$F(y) = \|y\| \quad \text{and} \quad |F(x)| \leq \|x\| \quad \text{for all } x \in X.$$

Applying this result with $X = \mathbb{C}^{s,s}, y = B^n$ we see that a linear $F : \mathbb{C}^{s,s} \to \mathbb{C}$ exists with

$$|F(A)| \leq \|A\| \quad \text{for all } s \times s \text{ matrices } A, \tag{2.6}$$

$$F(B^n) = \|B^n\|. \tag{2.7}$$

Combination of (2.7) and (2.2) yields

$$\|B^n\| = \frac{1}{2\pi\,\mathrm{i}} \int_\Gamma \zeta^n R(\zeta)\,\mathrm{d}\zeta,$$

where Γ is the positively oriented circle $|\zeta| = 1 + 1/n$ and R is the rational function defined by $R(\zeta) = F((\zeta I - B)^{-1})$. Integration by parts gives

$$\|B^n\| = \frac{-1}{2\pi\,\mathrm{i}(n+1)} \int_\Gamma \zeta^{n+1} R'(\zeta)\,\mathrm{d}\zeta \le \frac{1}{2\pi n}(1+1/n)^n \int_\Gamma |R'(\zeta)|\,\mathrm{d}\zeta|. \tag{2.8}$$

3. Let E_{jk} stand for the $s \times s$ matrix with entry in the jth row and kth column equal to 1, and all other entries 0. Denoting the entries of the matrix $(\zeta I - B)^{-1}$ by $r_{jk}(\zeta)$ we thus have

$$(\zeta I - B)^{-1} = \sum_{j,k} r_{jk}(\zeta) E_{jk},$$

and therefore also

$$R(\zeta) = \sum_{j,k} r_{jk}(\zeta) F(E_{jk}).$$

We define a rational function to be of order s if its numerator and denominator are polynomials of a degree not exceeding s. By Cramer's rule, the $r_{jk}(\zeta)$ are rational functions of order s with the same denominator. Hence $R(\zeta)$ is also of order s.

It was proved by Spijker (1991) that, for any rational function $R(\zeta)$ which has no poles on the circle Γ and is of order s, the following inequality holds:

$$\int_\Gamma |R'(\zeta)|\,\mathrm{d}\zeta| \le 2\pi s \max_\Gamma |R(\zeta)|. \tag{2.9}$$

The proof of (2.5) now easily follows by a combination of (2.8), (2.9), (2.6) and (1.10). □

We remark that this proof of (2.5) is essentially based on ideas taken from LeVeque and Trefethen (1984) and Lenferink and Spijker (1991a). For an interesting discussion and generalization of inequality (2.9) we refer to Wegert and Trefethen (1992).

In the following theorem we focus on the sharpness of the bound (2.4) in the case $n = s - 1$.

Theorem 2.2 Let $s \ge 2$ and an arbitrary norm $\|\cdot\|$ on $\mathbb{C}^{s,s}$ be given. Then

$$\sup\{\|B^{s-1}\|/M_1(B) :\ B \in \mathbb{C}^{s,s},\ M_1(B) < \infty\} = \left(1 + \frac{1}{s-1}\right)^{s-1} s, \tag{2.10}$$

where $M_1(B)$ denotes the smallest M_1 such that (1.10) holds (we define $M_1(B) = \infty$ if (1.10) is not fulfilled for any M_1).

Proof. Define $B \in \mathbb{C}^{s,s}$ by $B = \gamma E$, where $\gamma > 0$ is large and the $s \times s$ matrix E is defined by (1.9). We have

$$\begin{aligned} M_1(B) &= \sup_{|\zeta|>1} (|\zeta| - 1)\|(\zeta I - B)^{-1}\| = \sup_{|\zeta|>1} \frac{|\zeta| - 1}{|\zeta|} \left\| \sum_{j=0}^{s-1} \left(\frac{\gamma}{\zeta} E\right)^j \right\| \\ &\leq \sum_{j=0}^{s-1} \mu_j \gamma^j \|E^j\| = \mu_{s-1} \gamma^{s-1} \|E^{s-1}\| \left(1 + \mathcal{O}\left(\gamma^{-1}\right)\right), \end{aligned}$$

where

$$\mu_j = \sup_{|\zeta|>1} (|\zeta| - 1)|\zeta|^{-j-1} = \max_{0 \leq x \leq 1} (1 - x)x^j = j^j (j+1)^{-j-1},$$

so that

$$\begin{aligned} \|B^{s-1}\|/M_1(B) &\geq 1/\mu_{s-1} + \mathcal{O}\left(\gamma^{-1}\right) \\ &= \left(1 + \frac{1}{s-1}\right)^{s-1} s + \mathcal{O}\left(\gamma^{-1}\right) \quad (\text{as } \gamma \to \infty). \end{aligned}$$

It follows that the left-hand member of (2.10) is not smaller than the right-hand member. In view of (2.4) the proof is complete. □

Corollary 2.3 For each $s \geq 1$, let a norm $\|\cdot\| = \|\cdot\|^{(s)}$ be given on $\mathbb{C}^{s,s}$. Then there exist matrices $B_s \in \mathbb{C}^{s,s}$ for $s = 1, 2, 3, \ldots$, such that $M_1(B_s) < \infty$ and

$$\|(B_s)^{s-1}\|^{(s)} \sim esM_1(B_s) \quad (\text{as } s \to \infty), \tag{2.11}$$

where $M_1(B_s)$ has the same meaning as in Theorem 2.2.

Proof. Immediate from Theorem 2.2. □

This corollary was proved by LeVeque and Trefethen (1984) for the spectral norm. Our proof of Theorem 2.2 is essentially based on ideas taken from that paper.

In view of (2.4), the estimate (2.1) is valid for general norms $\|\cdot\|$ on $\mathbb{C}^{s,s}$. By virtue of Corollary 2.3, this general version of (2.1) is sharp in the sense of (2.11). However, it should be emphasized that this does not resolve the sharpness question for given *fixed* M_1, since $M_1(B_s)$ in (2.11) may depend on s. In the next two subsections we will focus on the situation where M_1 is a given fixed number.

2.3. About the best stability estimates for $M_1 = 1$

In the special situation where the resolvent condition (1.10) holds with $M_1 = 1$, the upper bound (2.4) can be improved in various ways. First we concentrate on arbitrary matrix norms on $\mathbb{C}^{s,s}$, and at the end of this subsection we focus on matrix norms $\|\cdot\|_p$ induced by the *pth Hölder norm on* $\mathbb{C}^s$ (for $p = 1, 2, \infty$).

Theorem 2.4 Let $s \geq 1$, $B \in \mathbb{C}^{s,s}$ and $\|\cdot\|$ denote an arbitrary matrix norm on $\mathbb{C}^{s,s}$. If (1.10) holds with $M_1 = 1$, then

$$\|B^n\| \leq n! n^{-n} \mathrm{e}^n \leq \sqrt{2\pi(n+1)} \quad \text{for } n = 1, 2, 3, \ldots . \tag{2.12}$$

Proof. Property (1.10) with $M_1 = 1$ implies that

$$\|\mathrm{e}^{zB}\| \leq \mathrm{e}^{|z|} \quad \text{for all } z \in \mathbb{C}.$$

This can be seen from Theorem 4.7, to be presented in Section 4, or more directly by using

$$\mathrm{e}^{zB} = \lim_{k\to\infty} \left[I - \frac{z}{k}B\right]^{-k} = \lim_{k\to\infty} \left(\frac{k}{z}\right)^k \left[\frac{k}{z}I - B\right]^{-k}$$

and applying (1.10) with $\zeta = k/z$.

We have

$$B^n = \frac{n!}{2\pi\,\mathrm{i}} \int_\Gamma z^{-n-1} \mathrm{e}^{zB}\,\mathrm{d}z,$$

where Γ is the positively oriented circle with radius n and centre 0. Therefore $\|B^n\| \leq n! n^{-n} e^n$. From Stirling's formula

$$n! = (n/\mathrm{e})^n \sqrt{2\pi n} \exp[\theta_n (12n)^{-1}] \quad \text{with } 0 < \theta_n < 1$$

(see, e.g., Abramowitz and Stegun (1965)), we finally obtain

$$\|B^n\| \leq \sqrt{2\pi n} \exp[(12n)^{-1}] \leq \sqrt{2\pi(n+1)}.$$

□

This proof of (2.12) is essentially based on ideas taken from Bonsall and Duncan (1980) (see also Bonsall and Duncan (1971)). Another proof can be given along the lines of Lubich and Nevanlinna (1991) (Theorem 2.1) or McCarthy (1992).

The next theorem shows that the upper bound for $\|B^n\|$ in (2.12) is sharp. For the elegant proof, which is beyond the scope of this paper, we refer to McCarthy (1992).

Theorem 2.5 Let $s \geq 2$ be given. Then there exists a vector norm on $\mathbb{C}^s$ such that the $s \times s$ shift matrix E, defined by (1.9), has the following two properties with respect to the corresponding induced matrix norm $\|\cdot\|$:

(a) E satisfies the resolvent condition (1.10) with $M_1 = 1$;
(b) $\|E^n\| = n!e^n n^{-n} \geq \sqrt{2\pi n}$ for $n = 1, 2, \ldots, s-1$.

According to the following theorem the stability estimate (2.12) can be substantially improved for the case of some important matrix norms.

Theorem 2.6 Let $s \geq 1$, $Q \in \mathbb{C}^{s,s}$ invertible, and $p = 1, 2$ or ∞. Let the norm $\|\cdot\|$ on $\mathbb{C}^{s,s}$ be defined by $\|A\| = \|QAQ^{-1}\|_p$ (for all $A \in \mathbb{C}^{s,s}$). Then (1.10) with $M_1 = 1$ implies (1.6) with $M_0 = 1$ (if $p = 1$ or ∞) or $M_0 = 2$ (if $p = 2$).

Proof. Since the result for general invertible Q easily follows from the result for $Q = I$, it is sufficient to consider the latter case only.

Let $p = \infty$. Suppose $B = (\beta_{jk})$ satisfies (1.10) with $\|\cdot\| = \|\cdot\|_\infty$, $M_1 = 1$. Clearly (4.11) holds with $W = D$, $M = 1$. By Theorem 4.7 relation (4.10) holds as well. In view of the expression for $\tau_1[B]$ (with $\|\cdot\| = \|\cdot\|_\infty$) given at the end of Section 4.2, we conclude that each disk with its centre at β_{jj} and radius $\rho_j = \sum_{k \neq j} |\beta_{jk}|$ lies in the unit disk D. Consequently, $|\beta_{jj}| + \rho_j \leq 1$, and

$$\|B\|_\infty = \max_{1 \leq j \leq s} (|\beta_{jj}| + \rho_j) \leq 1,$$

so that (1.6) holds with $M_0 = 1$.

For $p = 1$ the proof follows from the result for $p = \infty$ and the fact that $\|A\|_1 = \|A^{\mathrm{T}}\|_\infty$ for all $A \in \mathbb{C}^{s,s}$ (where A^{T} denotes the transpose of A).

For $p = 2$ the value $M_0 = 2$ was stated, e.g., in Reddy and Trefethen (1992) and McCarthy (1992). The proof runs as follows. It can be seen by a straightforward calculation (or directly from the material in Bonsall and Duncan (1980) or Lenferink and Spijker (1990)) that the numerical range $\{x^*Bx : x \in \mathbb{C}^s \text{ with } x^*x = 1\}$ is contained in the unit disk D. The proof continues by applying Berger's inequality (see, e.g., Pearcy (1966), Richtmyer and Morton (1967 p. 89), Bonsall and Duncan (1980) or Horn and Johnson (1990)). This inequality reads

$$r(A^n) \leq [r(A)]^n \quad \text{for } n = 1, 2, 3, \ldots,$$

where A is any $s \times s$ matrix, and $r(A)$ denotes the so-called numerical radius of A defined by

$$r(A) = \max \{|x^*Ax| : x \in \mathbb{C}^s \text{ with } x^*x = 1\}.$$

Since $r(B) \leq 1$, there follows

$$r(B^n) \leq 1.$$

By splitting B^n into a sum $B^n = A_1 + \mathrm{i}A_2$ with Hermitian A_1, A_2, and by noting that for any Hermitian A the relation $\|A\|_2 = r(A)$ is valid, we

finally obtain

$$\|B^n\| \le \|A_1\| + \|A_2\| = r(A_1) + r(A_2) \le 2r(B^n) \le 2.$$

□

2.4. About the best stability estimates for fixed $M_1 > 1$

Theorem 2.1 shows that if resolvent condition (1.10) is satisfied with fixed M_1, then $\|B^n\|$ can grow at most linearly with n or s. Corollary 2.3 reveals that the corresponding upper bound is sharp – if we allow M_1 to be variable.

For the special case $M_1 = 1$, however, this linear growth with n or s is too pessimistic, as can be seen from Theorems 2.4 and 2.6 in the previous subsection.

For fixed values $M_1 > 1$ the question also arises as to whether the upper bound (2.4) can be improved. We do not know of any positive results in this direction. In the following we shall therefore present negative results only – in the form of lower bounds for $\|B^n\|$.

A negative result we have seen already is Theorem 2.5, which is also relevant for any fixed $M_1 > 1$. It shows that $\|B^n\|$ may grow at the rate $\sqrt{n}$ or $\sqrt{s}$.

The following two theorems show what growth rates can be achieved for the three important matrix norms $\|\cdot\|_p$, $p = 1, 2, \infty$.

Theorem 2.7 Let $p = 1$ or $p = \infty$. Then there exist $C > 0$ and $M > 1$ such that

$$\sup_{s,B} \|B^n\|_p \ge C\sqrt{n} \quad \text{for } n = 0, 1, 2, \ldots,$$

where the supremum is over all integers $s \ge 1$ and all matrices $B \in \mathbb{C}^{s,s}$ satisfying the resolvent condition (1.10) with $M_1 = M$ and $\|\cdot\| = \|\cdot\|_p$.

Proof. The proof for the case $p = \infty$ easily follows from a straightforward adaptation of Example 2.2 in Lubich and Nevanlinna (1991) to the finite-dimensional case.

More precisely, let ϕ denote a Möbius transformation that maps the unit disk onto itself and is not just a rotation (such ϕ exist, see, e.g., Henrici (1974)). We define $B_s = \phi(E_s)$, where E_s stands for the $s \times s$ matrix E defined by (1.9). From the material in Lubich and Nevanlinna (1991) it follows that every B_s satisfies the resolvent condition (1.10) with $\|\cdot\| = \|\cdot\|_\infty$ and a constant M_1 independent of s, and that

$$\lim_{s\to\infty} \|B_s^n\|_\infty \ge C\sqrt{n} \quad \text{for } n = 0, 1, 2, \ldots,$$

where C is a positive constant. This proves the theorem for the case $p = \infty$.

For $p = 1$ the result is obtained by noting that $\|A\|_1 = \|A^{\mathrm{T}}\|_\infty$ for all $A \in \mathbb{C}^{s,s}$. □

Theorem 2.8 Let $M > \pi + 1$ be given. Then there exist a constant $C > 0$ and matrices $B_s \in \mathbb{C}^{s,s}$ for $s = 2, 4, 6, \ldots$, such that all B_s satisfy (1.10) with $M_1 = M$, $\|\cdot\| = \|\cdot\|_2$, and

$$\|(B_s)^{s/2}\|_2 \geq C(\log s)^{1/2}/\log\log s.$$

Proof. It was shown by McCarthy and Schwartz (1965) that for each $M > \pi + 1$ there exist a constant $\gamma > 0$ and $s \times s$ matrices $E_{s,j}$ (for all even positive s and $j = 1, 2, \ldots, s$) with the following properties:

$$(E_{s,j})^2 = E_{s,j} \neq 0, \quad E_{s,j}E_{s,k} = 0 \ (j \neq k), \quad \sum_{j=1}^{s} E_{s,j} = I, \tag{2.13}$$

$$\Big\| \sum_{j \text{ odd}} E_{s,j} \Big\|_2 \geq \gamma(\log s)^{1/2}/\log\log s, \tag{2.14}$$

$$\begin{aligned} &B_s = \textstyle\sum_{j=1}^{s} e^{2\pi ij/s} E_{s,j} \text{ satisfies (1.10) with} \\ &M_1 = M \text{ and } \|\cdot\| = \|\cdot\|_2. \end{aligned} \tag{2.15}$$

For even s we have

$$(B_s)^{s/2} = \sum_{j=1}^{s} (-1)^j E_{s,j} = I - 2 \sum_{j \text{ odd}} E_{s,j}.$$

In view of (2.14) this implies

$$\|(B_s)^{s/2}\|_2 \geq 2\gamma(\log s)^{1/2}/\log\log s - 1 \quad \text{for } s = 2, 4, 6, \ldots.$$

Since all $(B_s)^{s/2} \neq 0$ there exists a constant C with the property stated in the theorem. □

For additional interesting examples for the matrix norms $\|\cdot\|_p$ with $p = 2$, ∞ we refer to McCarthy (1992).

We also mention that after completion of the present article new results related to this were found by Kraaijevanger (1992) for the matrix norm $\|\cdot\|_\infty$.

3. Stability estimates under resolvent conditions with respect to general regions V

3.1. Linear stability analysis and stability regions

Consider an initial value problem for a system of s *ordinary differential equations* of the form

$$\begin{aligned} U'(t) &= AU(t) + b(t) \quad (t \geq 0), \\ U(0) &= u_0. \end{aligned} \tag{3.1}$$

Here A is a given constant $s \times s$ matrix, and u_0, $b(t)$ are given vectors in $\mathbb{C}^s$. The vector $U(t) \in \mathbb{C}^s$ is unknown for $t > 0$.

In this section we analyse the stability of numerical processes for approximating $U(t)$. This analysis will also be relevant to classes of numerical processes for solving *partial differential equations*.

To elucidate this relevance, we assume an initial-boundary value problem to be given for a linear partial differential equation with variable coefficients in the differential operator (which depend on the space variable x but not on the time variable t). Applying the method of *semi-discretization*, where discretization is applied to the space variable x only, one arrives at an initial value problem for a large system of the form (3.1). In this case the matrix A, the inhomogeneous term $b(t)$, and the vector u_0 are determined by the original initial-boundary value problem and by the process of semi-discretization. The solution $U(t)$ to (3.1) then provides an approximation to the solution of the original initial-boundary value problem. For an example we refer to problem (1.3); by replacing the derivatives with respect to x in (1.3) by the same *finite difference* quotients as referred toin Section 1.2, one arrives at an initial value problem (3.1) with the tridiagonal $s \times s$ matrix $A = (\alpha_{jk})$ given by (1.5). We note that problems (3.1) arise not only when the semi-discretization relies on the introduction of finite differences, but also when it is based on *spectral* approximations (see Gottlieb and Orszag (1977) and Canuto, Hussaini, Quarteroni and Zang (1988)) or on the *finite element* method (see, e.g., Oden and Reddy (1976) and Strang and Fix (1973)).

Many step-by-step methods for the numerical solution of ordinary differential equations, like Runge–Kutta methods or Rosenbrock methods (see Butcher (1987) and Hairer and Wanner (1991)), reduce – when applied to (3.1) – to processes of the form

$$u_n = \varphi(hA)u_{n-1} + b_n \quad \text{for } n = 1, 2, 3, \ldots . \tag{3.2}$$

Here $\varphi(\zeta) = P(\zeta)/Q(\zeta)$ is a rational function, depending only on the underlying step-by-step method. $P(\zeta), Q(\zeta)$ are polynomials, without common zeros, such that $\varphi(0) = \varphi'(0) = 1$. Further, $h = \Delta t > 0$ denotes the *step-size*, and we define $\varphi(hA) = P(hA)Q(hA)^{-1}$ when $Q(hA)$ is invertible. The vectors $b_n \in \mathbb{C}^s$ are related to $b(t)$, and $u_n \simeq U(nh)$ are calculated successively from (3.2). It is worth noting that many numerical processes in partial differential equations which are *not* constructed with the process of semi-discretization in mind are still of the form (3.2), and can *a posteriori* be conceived as relying on semi-discretization. For instance, it follows from (1.4) that the process constructed in Section 1.2 can be written in the form (3.2) with

$$\varphi(\zeta) = (1 + (1-\theta)\zeta)(1-\theta\zeta)^{-1}$$

and $A = (\alpha_{jk})$ satisfying (1.5).

Since (3.2) is a special case of (1.2), the stability analysis of (3.2) amounts

to investigating the growth of matrices B^n with

$$B = \varphi(hA).$$

In this analysis it is useful to introduce the *stability region* S, defined by

$$S = \{\zeta : \zeta \in \mathbb{C} \text{ with } Q(\zeta) \neq 0 \text{ and } |\varphi(\zeta)| \leq 1\}. \tag{3.3}$$

Consider the following requirement on hA with regard to S,

$$\begin{aligned} &\sigma[hA] \subset S, \text{ and for each } \zeta \in \partial S \text{ which is a zero of the} \\ &\text{minimal polynomial of } hA \text{ with multiplicity } m > 1, \\ &\text{the derivatives } \varphi^{(j)}(\zeta) \text{ vanish for } j = 1, 2, \ldots, m-1. \end{aligned} \tag{3.4}$$

Here $\sigma[hA]$ denotes the *spectrum* (set of eigenvalues) of hA, and ∂S the *boundary* of S. For the concept of minimal polynomial see, e.g., Horn and Johnson (1990). The spectral mapping theorem (see Conway (1985) or Rudin (1973)) states that, if $Q(\zeta) \neq 0$ for all $\zeta \in \sigma[hA]$, then

$$\sigma[\varphi(hA)] = \{\varphi(\zeta) : \zeta \in \sigma[hA]\}.$$

Hence, the condition $\sigma[hA] \subset S$ in (3.4) is equivalent to the condition $\sigma[B] \subset D$ in (1.7) with $B = \varphi(hA)$. Further, from the Jordan canonical form of hA it can be deduced that the condition regarding $\zeta \in \partial S$ in (3.4) is equivalent to the condition on the geometric multiplicities in (1.7). Consequently, (3.4) is equivalent to (1.7). It follows that (3.4) is a necessary and sufficient condition in order that a finite M_0 exists with stability property (1.6) for $B = \varphi(hA)$.

We note that most functions $\varphi(\zeta)$ of practical interest have nonvanishing derivatives $\varphi'(\zeta)$ on the whole of ∂S. In this case (3.4) simply reduces to $\sigma[hA] \subset S$ and the condition that all $\zeta \in \partial S \cap \sigma[hA]$ are zeros of the minimal polynomial of hA with multiplicity 1.

In general (3.4) has similar advantages (it is relatively simple to verify, and reliable for normal matrices) and disadvantages (quite unreliable for families of matrices that are not normal) as the eigenvalue condition (1.7). In the rest of this section we adapt (3.4) to conditions on hA that reliably predict stability – also for nonnormal matrices and norms $\|\cdot\|$ on $\mathbb{C}^{s,s}$ different from the spectral norm. An advantage of these conditions on hA over a resolvent condition on $B = \varphi(hA)$ (as dealt with in Section 2) lies in the circumstance that, in general, hA has a simpler structure than B, and that knowledge available about S can be exploited.

3.2. Reviewing stability estimates from the literature

In the literature various stability results can be found, which are essentially based on the use of resolvent conditions of the form

$$\begin{aligned} &\zeta I - hA \text{ is invertible and } \|(\zeta I - hA)^{-1}\| \leq M_1 \, \mathrm{d}(\zeta, V)^{-1} \\ &\text{for all complex numbers } \zeta \notin V. \end{aligned} \tag{3.5}$$

Here, V is a closed subset of the stability region S (see (3.3)), M_1 is a constant, $\|\cdot\|$ denotes a norm on $\mathbb{C}^{s,s}$ and $\mathrm{d}(\zeta, V) = \min\{|\zeta - \eta| : \eta \in V\}$ is the distance from ζ to V. Under additional assumptions, to be stated below, it is shown in the literature that (3.5) implies a stability estimate

$$\|\varphi(hA)^n\| \leq M_1 g(n, s) \quad \text{for } n = 1, 2, 3, \ldots, \tag{3.6}$$

where the function g *only depends on* φ *and* V (and not on h, A, M_1 or $\|\cdot\|$).

In the following we list some of these stability results. We assume throughout that (3.5) is satisfied with closed $V \subset S$ and a norm $\|\cdot\|$ on $\mathbb{C}^{s,s}$. In each separate case we formulate the relevant additional assumptions and the resulting function g.

For any $W \subset \mathbb{C}$ we denote by ∂W the *boundary* of W, and write $\mathbb{C}^- = \{\zeta : \zeta \in \mathbb{C} \text{ with } \mathrm{Re}\,\zeta \leq 0\}$.

1 In Lenferink and Spijker (1991a) (Theorem 2.2) estimate (3.6) is proved with $g(n, s) \equiv \gamma s$ where γ depends only on φ and V. The additional assumptions are: V is bounded and convex; $\varphi'(\zeta) \neq 0$ on $\partial V \cap \partial S$; and ∂V lies on an algebraic curve.

2 In Lenferink and Spijker (1991b) (Lemma 3.3) estimate (3.6) is proved with $g(n, s) \equiv \gamma n$ where γ depends only on φ and V. The additional assumptions are: V is bounded and convex; and $\|\cdot\|$ is induced by a vector norm on $\mathbb{C}^s$.

3 In Reddy and Trefethen (1992) (Theorem 7.1) estimate (3.6) is proved with $g(n, s) \equiv \gamma \min(n, s)$ where γ depends only on φ. The additional assumptions are: $V = S$, S is bounded; $\varphi'(\zeta) \neq 0$ on ∂S; and $\|\cdot\|$ is a weighted spectral norm (i.e. $\|B\| = \|QBQ^{-1}\|_2$ for all $B \in \mathbb{C}^{s,s}$, where Q is an invertible matrix).

4 In Lubich and Nevanlinna (1991) (Theorem 3.1) estimate (3.6) is proved with $g(n, s) \equiv \gamma \min(n, s)$ where γ depends only on φ. The additional assumptions are: $V = \mathbb{C}^-$ and $\|\cdot\|$ is induced by a vector norm on $\mathbb{C}^s$.

5 From the material in the important paper by Brenner and Thomée (1979) it follows that (3.6) holds with $g(n, s) \equiv \gamma\sqrt{n}$ where γ depends only on φ. The additional assumptions are: $V = \mathbb{C}^-$, $M_1 = 1$ and $\|\cdot\|$ is induced by a vector norm on $\mathbb{C}^s$.

6 For $\delta \geq 0$ the *wedge* $W(\delta)$ is defined by $W(\delta) = \{\zeta : \zeta = 0 \text{ or } |\arg\zeta - \pi| \leq \delta\}$. In Lenferink and Spijker (1991b) (Lemma 3.1) estimate (3.6) is proved with $g(n, s) \equiv \gamma$ where γ depends only on φ and V. The additional assumptions are: V is a bounded convex subset of $W(\alpha)$, where $0 \leq \alpha < \pi/2$, $V \subset \mathrm{int}(S) \cup \{0\}$; and $\|\cdot\|$ is induced by a vector norm on $\mathbb{C}^s$.

7 In Crouzeix, Larsson, Piskarev and Thomée (1991) (Theorem 5) estimate (3.6) is proved with $g(n, s) \equiv \gamma$ where γ depends only on φ and

V. The additional assumptions, slightly adapted in order to fit in our framework, are: $V = W(\alpha)$, $S \supset W(\beta)$, $0 \leq \alpha < \beta \leq \pi/2$ and $\|\cdot\|$ is induced by a vector norm on $\mathbb{C}^s$. For related material see Palencia (1991, 1992) and Lubich and Nevanlinna (1991).

8 For $\rho > 0$ the *disk* $D(\rho)$ is defined by $D(\rho) = \{\zeta : \zeta \in \mathbb{C} \text{ and } |\zeta+\rho| \leq \rho\}$. In Lubich and Nevanlinna (1991) (Theorem 3.4) estimate (3.6) is proved with $g(n,s) \equiv \gamma\sqrt{1+nr_0}$ where γ depends only on φ. The additional assumptions are: $r_0 > 0$, $V = D(r_0)$, $S \supset \mathbb{C}^-$ and $\|\cdot\|$ is induced by a vector norm on $\mathbb{C}^s$. (The assumption $S \supset \mathbb{C}^-$ can be relaxed, see Lubich and Nevanlinna (1991).)

9 The quantity $r = \sup\{\rho : \rho > 0 \text{ and } D(\rho) \subset S\}$ is called the *stability radius* of the step-by-step method (3.2) (see, e.g., Kraaijevanger *et al.* (1987)). In Lenferink and Spijker (1991b) (Sections 2.3 and 2.4) it was noted that, for $0 < r < \infty$, estimate (3.6) holds with $g(n,s) \equiv \gamma\sqrt{n}$ where γ only depends on φ. The additional assumptions are: $M_1 = 1$, $\|\cdot\| = \|\cdot\|_\infty$ and $V = D(r)$. Next consider $r \in (0,\infty]$ and $0 < r_0 < r$. If (3.5) holds with $V = D(r_0)$, then, again under the assumptions $M_1 = 1$, $\|\cdot\| = \|\cdot\|_\infty$, inequality (3.6) even holds with $g(n,s) \equiv \gamma$, where γ depends only on φ and r_0 (see Kraaijevanger *et al.* (1987) and Lenferink and Spijker (1991b)).

10 In Brenner and Thomée (1979) and Lubich and Nevanlinna (1991) more refined estimates of the form (3.6) were derived for functions φ satisfying special conditions. For example, from Lubich and Nevanlinna (1991) (Theorem 3.2) it follows that, in the situation of point 4, an estimate (3.6) with $g(n,s) \equiv \gamma \min(n^\alpha, s)$, $\alpha < 1$, is possible for functions φ with $|\varphi(\zeta)|$ not identically 1 on the imaginary axis. We refer to Brenner and Thomée (1979) and Lubich and Nevanlinna (1991) for more details.

3.3. Various comments on stability estimates from the literature

Remark 3.1 Results 1, 2, 6 and 8 in the last subsection were proved by using integral representations of the form

$$\varphi(hA)^n = \frac{1}{2\pi\,\mathrm{i}} \int_\Gamma \varphi(\zeta)^n (\zeta I - hA)^{-1}\,\mathrm{d}\zeta,$$

where Γ is a proper curve in the complex plane surrounding V, and by estimating the integral (see, e.g., the proof of Theorem 2.1). Results 5, 7 and 10 were proved by using related, but different, integral representations for $\varphi(hA)^n$.

Results 3 and 4 were obtained by first proving that resolvent condition (3.5) for hA implies a resolvent condition (1.10) for $B = \varphi(hA)$ (with a different constant M_1) and then applying (a version of) Theorem 2.1 to this matrix B.

Finally, the proof of Result 9 relies on an expansion of $\varphi(hA)^n$ in a power series

$$\varphi(hA)^n = \gamma_0 I + \gamma_1(hA + \rho I) + \gamma_2(hA + \rho I)^2 + \cdots \tag{3.7}$$

with $\rho = r$ or r_0, and on bounding the terms of the series using the fact that the resolvent condition (3.5) (with $M_1 = 1$, $\|\cdot\| = \|\cdot\|_\infty$ and $V = D(\rho)$) implies a so-called *circle condition* $\|hA + \rho I\|_\infty \leq \rho$. The latter implication, which is in fact an equivalence, follows immediately from Theorem 2.6 (with $B = \rho^{-1}(hA + \rho I)$), and was stated in Lenferink and Spijker (1991b) (Section 2.4). In Kraaijevanger *et al.* (1987), Nevanlinna (1984) and Spijker (1985) this circle condition was combined with (3.7) to yield the desired stability bounds.

Remark 3.2 We note that Results 2, 3, 4, 6, 7 and 8, although formulated in Kraaijevanger *et al.* (1987), Nevanlinna (1984) and Spijker (1985) for special norms, are valid as well for arbitrary norms $\|\cdot\|$ on $\mathbb{C}^{s,s}$. This can be seen by a straightforward adaptation of the proofs in Kraaijevanger *et al.* (1987), Nevanlinna (1984) and Spijker (1985).

Further, it is easy to see that Result 9 is also valid for norms $\|\cdot\|$ defined by $\|B\| = \|QBQ^{-1}\|_p$ (for all $B \in \mathbb{C}^{s,s}$), where Q is an invertible matrix and $p = 1$ or ∞.

Remark 3.3 In all of the above, the resolvent condition (3.5) occurs as a *sufficient condition* for stability estimates of the form (3.6). Reddy and Trefethen (1992) (Theorem 7.1) succeeded in showing (for the weighted spectral norm, see Result 3) that (3.5) is also a *necessary condition* for stability. In fact, they showed – for any matrix hA belonging to a specific family $\mathcal{F}$ defined in their paper – that, in general, strong stability (i.e. $\|\varphi(hA)^n\| \leq M_0$ for all $n \geq 0$) implies the resolvent condition (3.5) with $V = S$ and $M_1 = \gamma M_0$. Here γ depends only on φ and $\mathcal{F}$.

Remark 3.4 Modifications of Results 3, 5 and 9 can be proved if we relax slightly the assumption $V \subset S$ for the set V in the resolvent condition (3.5). This can be useful in applications (see Section 5).

(a) Let S be bounded and $\varphi'(\zeta) \neq 0$ on ∂S. Further, let $\beta > 0$ and $h > 0$ be given. Suppose that the resolvent condition (3.5) is (only) satisfied with respect to the set $V = S + \beta h D$ (but not necessarily with respect to the smaller set $V = S$ itself).
It follows from Reddy and Trefethen (1992) (Theorem 8.2) that there exist positive constants $\gamma_1, \gamma_2, \gamma_3$ (only depending on φ) such that these assumptions imply the stability estimate

$$\|\varphi(hA)^n\| \leq M_1\gamma_1 e^{\gamma_2\beta nh} \min(n, s) \quad \text{for } n = 1, 2, 3, \ldots \tag{3.8}$$

whenever $\beta h \leq \gamma_3$. This was proved in Reddy and Trefethen (1992) for the weighted spectral norm (defined in Result 3). The proof in that paper can be adapted in a straightforward way to arbitrary norms on $\mathbb{C}^{s,s}$.

(b) Let $r \leq \infty$ have the same meaning as in Result 9, and let $M_1 = 1$, $\|\cdot\| = \|\cdot\|_\infty$. Further, let $0 < r_0 < \infty$, $r_0 \leq r$ and $\beta > 0$ be given. Then there exists a constant $h_0 > 0$ such that φ is analytic on $W = D(r_0) + \beta h_0 D = \{\zeta : \zeta \in \mathbb{C} \text{ and } |\zeta + r_0| \leq r_0 + \beta h_0\}$. Suppose that $0 < h \leq h_0$ and (3.5) is (only) satisfied with $V = D(r_0) + \beta h D$. These assumptions imply the stability estimate

$$\|\varphi(hA)^n\| \leq \gamma_1 e^{\gamma_2 \beta n h} \sqrt{n} \quad \text{for } n = 1, 2, 3, \ldots, \tag{3.9}$$

where the constants γ_1, γ_2 depend only on φ, r_0 and βh_0 (and not on h, n, s or A). The proof is again based on the expansion (3.7) (with $\rho = r_0$), and can be given in two steps. First we apply Theorem 2.6 (with $B = (r_0 + \beta h)^{-1}(hA + r_0 I)$) to obtain $\|hA + r_0 I\| \leq r_0 + \beta h$ and then use (3.7) and estimates for the $|\gamma_k|$ (see Spijker (1985), Corollary 4.3) to derive (3.9). Further, it is easy to see that this result is also valid for norms $\|\cdot\|$ defined by $\|B\| = \|QBQ^{-1}\|_p$ (for all $B \in \mathbb{C}^{s,s}$), where Q is an invertible matrix and $p = 1$ or ∞.

(c) An estimate of the form (3.9) can also be proved if we replace the condition $V = \mathbb{C}^-$ in Result 5 by $V = \mathbb{C}^- + \beta h D$. We refer to Brenner and Thomée (1979) (Theorem 1) for more details.

Remark 3.5 Some of the arguments recently used in Kreiss (1990) and Kreiss and Wu (1992) are closely related to the above, and can be interpreted as yielding a result of the form (3.6). The assumptions on φ which are made in Kreiss (1990) and Kreiss and Wu (1992) in order to derive stability estimates comprise:

$$\text{The half disk } \{\zeta : \operatorname{Re}\zeta \leq 0,\ |\zeta| \leq R_1\} \text{ is contained in } S, \tag{3.10}$$

$$\begin{aligned} &\varphi \text{ is a polynomial which does not transform any} \\ &\text{two different points } \zeta \text{ with } \operatorname{Re}\zeta = 0, |\zeta| < R_1 \\ &\text{into one and the same image point } z \text{ with } |z| = 1. \end{aligned} \tag{3.11}$$

Here R_1 is a given positive constant. Assume the $s \times s$ matrix hA satisfies

$$\|hA\| \leq R < R_1, \tag{3.12}$$

$$\|(\zeta I - hA)^{-1}\| \leq K_1 (\operatorname{Re}\zeta)^{-1} \quad \text{for all } \zeta \in \mathbb{C} \text{ with } \operatorname{Re}\zeta > 0. \tag{3.13}$$

Although the setting in Kreiss (1990) and Kreiss and Wu (1992) is different in appearance from the one we use here, Theorem 3.2 in Kreiss and Wu (1992) essentially states that (3.10)–(3.13) imply

$$\|(e^{\zeta} I - \varphi(hA))^{-1}\| \leq K_0 (\operatorname{Re}\zeta)^{-1} \quad \text{for all } \zeta \text{ with } \operatorname{Re}\zeta > 0. \tag{3.14}$$

We now show that this conclusion is related to the stability results described earlier. First of all, the assumptions (3.12), (3.13) imply our resolvent condition (3.5) with $V = \{\zeta : \operatorname{Re}\zeta \leq 0,\ |\zeta| \leq R\}$ and $M_1 = \sqrt{2}K_1$. Further, (3.14) can be proved to be equivalent to a resolvent condition of the form (1.10) with $B = \varphi(hA)$. Therefore, by Theorem 2.1, (3.14) implies a result of the form (3.6).

The stability estimates which are focused on in Kreiss (1990) and Kreiss and Wu (1992) are pertinent to l_2 norms, and essentially different from (1.6) or (3.6). In fact, the estimates (3.6) are relevant to stability with respect to perturbations in the initial value u_0 of process (3.2), whereas the estimates in Kreiss (1990) and Kreiss and Wu (1992) are relevant to *stability with respect to perturbations in the vectors* b_n of (3.2). In Kreiss (1990) and Kreiss and Wu (1992) this stability concept, referred to as *stability in a generalized sense*, is argued to be equivalent to an inequality of the form (3.14) (see Kreiss and Wu (1992) (Theorem 3.1)). Moreover, an analogous concept (of stability in a generalized sense) for the continuous problem (3.1) is stated to be equivalent to a resolvent condition of the form (3.13).

4. Various related concepts and problems

4.1. ϵ-pseudospectra

The useful concept of ϵ-pseudospectra has been introduced and studied by Landau (1975), Varah (1979), Reddy and Trefethen (1990, 1992), Trefethen (1992) and others. The focus in these papers is on the (weighted) spectral norm. The main purpose of this subsection is to extend the notion of ϵ-pseudospectra to the situation of general matrix norms, and to relate it to the resolvent condition (3.5).

Let $\|\cdot\|$ denote an arbitrary matrix norm on $\mathbb{C}^{s,s}$. Let B be an $s \times s$ matrix and $\epsilon > 0$. Consider for a given complex number λ the situation where

$$\text{there exists an } s \times s \text{ matrix } E \text{ with } \|E\| \leq \epsilon \text{ such that } \lambda \in \sigma[B+E]. \quad (4.1)$$

Analogously to Reddy and Trefethen (1990, 1992), Reichel and Trefethen (1992) and Trefethen (1992) we give the following definition.

Definition 4.1 The set of all complex numbers λ satisfying (4.1) is called the *ϵ-pseudospectrum* of B and is denoted by $\sigma_\epsilon[B]$.

We emphasize that – unlike the spectrum $\sigma[B]$ – the pseudospectrum $\sigma_\epsilon[B]$ depends on the norm $\|\cdot\|$.

The concept of an ϵ-pseudospectrum can be related to the following properties:

$$\begin{aligned}&\text{There exists an } s \times s \text{ matrix } E \text{ with } \|E\| = \epsilon\\ &\text{such that } \lambda \in \sigma[B+E];\end{aligned} \quad (4.2)$$

There exists an $s \times s$ matrix U with $\|U\| = 1$
such that $\|(B - \lambda I)U\| \leq \epsilon$; (4.3)

$B - \lambda I$ is singular,
or $B - \lambda I$ is invertible with $\|(B - \lambda I)^{-1}\| \geq \epsilon^{-1}$. (4.4)

We have

Theorem 4.2 (a) Let $\|\cdot\|$ be a matrix norm on $\mathbb{C}^{s,s}$. Then (4.1) and (4.2) are equivalent (provided $s \geq 2$). Moreover (4.2) implies (4.3), and (4.3) implies (4.4). If $\|I\| = 1$ then (4.3) and (4.4) are equivalent.

(b) Let $\|\cdot\|$ be induced by a vector norm $|\cdot|$ on $\mathbb{C}^s$, $s \geq 2$. Then properties (4.1) – (4.4) are equivalent to each other. Moreover, they are equivalent to the requirement that

there exists a vector $u \in \mathbb{C}^s$ with $|u| = 1$ such that $|(B - \lambda I)u| \leq \epsilon$. (4.5)

Proof. (a) First we prove the equivalence of (4.1) and (4.2). The implication of (4.1) by (4.2) is trivial. To prove the reverse implication we assume there exists a matrix E with $\|E\| < \epsilon$ and a vector $u \neq 0$ such that $(B + E - \lambda I)u = 0$. When $s \geq 2$ we can choose a matrix C with $C \neq 0$ and $Cu = 0$. Define the matrix $E(t) = E + tC$ for $t \geq 0$. There exists a positive t_0 such that $\|E(t_0)\| = \epsilon$ and $\lambda \in \sigma[B + E(t_0)]$, which proves (4.2).

Assume (4.2). Define $V = [u, 0, 0, \ldots, 0] \in \mathbb{C}^{s,s}$ where $u \in \mathbb{C}^s$ is an eigenvector of $B + E$ corresponding to the eigenvalue λ. Defining $U = \|V\|^{-1}V$ we arrive at $\|U\| = 1$ and $(B + E)U = \lambda U$. Hence $\|(B - \lambda I)U\| = \|EU\| \leq \epsilon$, which proves (4.3).

Assume (4.3). For invertible $B - \lambda I$ we get with $E = (B - \lambda I)U$ the relation $\|E\| \leq \epsilon$, and therefore $1 = \|(B - \lambda I)^{-1}E\| \leq \|(B - \lambda I)^{-1}\|\epsilon$, which proves (4.4).

Assume (4.4) and $\|I\| = 1$. If $B - \lambda I$ is singular then (4.3) holds with $U = [u, 0, 0, \ldots, 0]$, where $u \in \mathbb{C}^s$ is in the null space of $B - \lambda I$ and is chosen such that $\|U\| = 1$. If $B - \lambda I$ is invertible then (4.3) holds with $U = \|(B - \lambda I)^{-1}\|^{-1}(B - \lambda I)^{-1}$.

(b) Assume (4.3). Choosing $v \in \mathbb{C}^s$ with $|v| = 1$, $|Uv| = 1$, we have $|(B - \lambda I)Uv| \leq \epsilon$. With $u = Uv$ we arrive at (4.5).

Assume (4.5). Taking $X = \mathbb{C}^s$ and $y = u$ in the corollary to the Hahn–Banach theorem formulated in the proof of Theorem 2.1, we see that there exists a linear transformation $F : \mathbb{C}^s \to \mathbb{C}$ with

$$F(u) = 1 \quad \text{and} \quad |F(x)| \leq |x| \quad \text{for all } x \in \mathbb{C}^s.$$

Defining the matrix E by

$$Ex = -F(x)(B - \lambda I)u \quad \text{for all } x \in \mathbb{C}^s,$$

it follows that $Eu = -(B - \lambda I)u$ and $\|E\| = |(B - \lambda I)u| \leq \epsilon$, which proves (4.1).

In view of (a) the proof is complete. □

Remark 4.3 (a) In part (a) of Theorem 4.2 the assumption $\|I\| = 1$ is essential for the equivalence of (4.3) and (4.4). This can be seen as follows. Let $s \geq 1$, $\epsilon = \|I\|^{-1}$, $\lambda = 0$ and $B = I$. Then (4.4) is always satisfied but (4.3) holds if and only if $\|I\| = 1$.

(b) For arbitrary matrix norms, (4.1) can be a stronger condition than (4.3), even if $\|I\| = 1$. This can be seen from the following example. On $\mathbb{C}^{2,2}$ we define a matrix norm by

$$\|A\| = \max\{\|A\|_1, \|A\|_\infty\} \quad \text{for all } A = (a_{ij}) \in \mathbb{C}^{2,2}$$

(see, e.g., Horn and Johnson (1990, p. 308)) and choose

$$\lambda = 0, \quad \epsilon = \tfrac{1}{2} \quad \text{and} \quad B = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

One easily verifies that condition (4.3) is satisfied by taking

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

But, a straightforward calculation reveals that $B + E$ is invertible for all $s \times s$ matrices E with $\|E\| \leq 1/2$, showing that Condition (4.1) is violated.

Remark 4.4 If $\|\cdot\|$ is the spectral norm, then Conditions (4.1)–(4.5) are all equivalent to the requirement that $B - \lambda I$ has a singular value σ with $\sigma \leq \epsilon$ (see Reddy and Trefethen (1990), Trefethen (1992) and Varah (1979)).

Following Reddy and Trefethen (1990, 1992) we now formulate a theorem which shows that the resolvent condition (3.5) can be nicely interpreted in terms of the ϵ-pseudospectra of the matrix hA.

Theorem 4.5 Let the norm $\|\cdot\|$ on $\mathbb{C}^{s,s}$ be induced by a vector norm on $\mathbb{C}^s$, and let V, h, A, M_1 be as in Section 3.2. Then the resolvent condition (3.5) is equivalent to the requirement that for all $\epsilon > 0$ the set $\sigma_\epsilon[hA]$ is contained in $V + M_1 \epsilon D = \{\zeta : \zeta = \xi + \eta \text{ with } \xi \in V, |\eta| \leq M_1 \epsilon\}$.

The theorem can be proved in a straightforward way by using the fact that, according to Theorem 4.2(b), Properties (4.1) and (4.4) with $B = hA$ are equivalent in the situation of the theorem.

Following the ideas of Trefethen (1992), the concept of ϵ-pseudospectra can also be used to *determine numerically* regions V and constants M_1 such that (3.5) holds. In order to explain how this can be done we assume $\|\cdot\|$

to be induced by a vector norm on $\mathbb{C}^s$, write $B = hA$, choose a fixed $\epsilon > 0$, denote the boundary of $\sigma_\epsilon[B]$ by Γ_ϵ and its length by $|\Gamma_\epsilon|$. The set

$$V = \sigma_\epsilon[B] \tag{4.6}$$

can be determined numerically, e.g., by checking for a large set of complex numbers λ whether (4.4) is satisfied. A corresponding constant M_1 can be computed from the formula

$$M_1 = |\Gamma_\epsilon|(2\pi\epsilon)^{-1}. \tag{4.7}$$

In order to establish (4.7) we note that for $\zeta \notin V$ we have

$$(\zeta I - B)^{-1} = \frac{1}{2\pi \mathrm{i}} \int_{\Gamma_\epsilon} (\zeta - \lambda)^{-1}(\lambda I - B)^{-1}\, \mathrm{d}\lambda$$

and therefore

$$\|(\zeta I - B)^{-1}\| \leq \frac{|\Gamma_\epsilon|}{2\pi} \max_{\lambda \in \Gamma_\epsilon} |(\zeta - \lambda)^{-1}| \epsilon^{-1} = M_1\, \mathrm{d}(\zeta, V)^{-1}.$$

It should be noted that both V and M_1 depend on ϵ so that it may pay to evaluate (4.6) and (4.7) for various values of ϵ.

We refer to Trefethen (1992) for closely related considerations and many further interesting applications of ϵ-pseudospectra.

4.2. The M-numerical range

When applying the stability results discussed in Sections 3.2 and 3.3, one may want to *prove* rigorously resolvent conditions of the form (3.5). In the following we show that the concept of the M-numerical range, introduced by Lenferink and Spijker (1990), can be helpful. The M-numerical range, to be defined below, can be viewed as a generalization of the *classical numerical range* (for an $s \times s$ matrix B),

$$\{x^* B x : x \in \mathbb{C}^s \text{ with } x^* x = 1\}.$$

The resolvent condition (3.5) will be seen to be satisfied when V contains the M_1-numerical range of hA.

Let $\|\cdot\|$ be a matrix norm on $\mathbb{C}^{s,s}$, and M a constant with $M \geq \|I\|$. Assume B is a given $s \times s$ matrix. We focus on disks

$$D[\gamma, \rho] = \{\zeta : \zeta \in \mathbb{C} \text{ with } |\zeta - \gamma| \leq \rho\}$$

with arbitrary $\gamma \in \mathbb{C}$, $\rho \geq 0$ such that

$$\|(B - \gamma I)^k\| \leq M\rho^k \quad \text{for } k = 1, 2, 3, \ldots. \tag{4.8}$$

Definition 4.6 The *M-numerical range of B with respect to the norm* $\|\cdot\|$ is the set $\tau_M[B]$ defined by

$$\tau_M[B] = \bigcap D[\gamma, \rho], \tag{4.9}$$

where the intersection is over all disks $D[\gamma, \rho]$ with property (4.8).

Let W be a nonempty, closed and convex subset of $\mathbb{C}$. If ξ belongs to the boundary ∂W of W and

$$\mathrm{Re}\{e^{-i\theta}(\zeta - \xi)\} \leq 0 \quad \text{for all } \zeta \in W,$$

where θ is a real constant, then θ is called a *normal direction* to W at ξ.

In order to formulate a basic theorem about the M-numerical range we consider the following four conditions on B:

$$\tau_M[B] \subset W, \tag{4.10}$$

$$\zeta I - B \text{ is invertible and } \|(\zeta I - B)^{-k}\| \leq M \cdot \mathrm{d}(\zeta, W)^{-k} \text{ for all } \zeta \notin W \text{ and } k = 1, 2, 3, \ldots, \tag{4.11}$$

$$\|\exp[te^{-i\theta}(B - \xi I)]\| \leq M \text{ for all } t \geq 0, \xi \in \partial W \text{ and normal directions } \theta \text{ to } W \text{ at } \xi, \tag{4.12}$$

$$\text{there is a unital matrix norm } \|\cdot\|' \text{ on } \mathbb{C}^{s,s} \text{ with corresponding 1-numerical range } \tau_1'[B] \subset W \text{ and } M^{-1}\|A\| \leq \|A\|' \leq M\|A\| \text{ (for all } s \times s \text{ matrices } A). \tag{4.13}$$

The following theorem was proved by Lenferink and Spijker (1990).

Theorem 4.7 Properties (4.10)–(4.13) are equivalent to each other.

Clearly, $\tau_M[B]$ is the smallest nonempty, closed and convex set $W \subset \mathbb{C}$ with property (4.10). Therefore, Theorem 4.7 reveals three new characterizations of the M-numerical range. We see that $\tau_M[B]$ equals the smallest nonempty, closed and convex set $W \subset \mathbb{C}$ with property (4.11), and the same holds with regard to properties (4.12) and (4.13).

It is clear that (4.11) is fulfilled for any set W with

$$\tau_M[B] \subset W \subset \mathbb{C}.$$

In view of Definition 4.6 we thus can make the two following observations.

(I) If V is any closed subset of $\mathbb{C}$ with $\tau_M[hA] \subset V$, then (3.5) is fulfilled with $M_1 = M$.

(II) In order to construct a set V as in (I) we only have to determine a finite number of pairs γ_j, ρ_j such that $B = hA$ satisfies (4.8) for all $\gamma = \gamma_j$, $\rho = \rho_j$. Clearly the set $V = \bigcap_j D[\gamma_j, \rho_j]$ is as required.

In Lenferink and Spijker (1990) (Theorem 3.1) and Lenferink and Spijker

(1991b) stability estimates were derived essentially along the lines of the observations (I), (II).

We finally note that for $M = 1$ the set (4.9) coincides with the so-called *algebra numerical range* (see, e.g., Bonsall and Duncan (1980) and Lenferink and Spijker (1990)). In this case some simple expressions for $\tau_1[B]$ are known. For $\|\cdot\| = \|\cdot\|_p$ with $p = 1, 2, \infty$ these expressions are as follows.

- Let $\|\cdot\| = \|\cdot\|_\infty$. Then $\tau_1[B]$ is equal to the convex hull of the union of the Gerschgorin disks $D[\gamma_j, \rho_j]$ defined by $\gamma_j = \beta_{jj}$ and $\rho_j = \sum_{k \neq j} |\beta_{jk}|$, where β_{jk} denote the entries of B (see, e.g., Lenferink and Spijker (1990) (Section 3.1.1)).
- Let $\|\cdot\| = \|\cdot\|_1$. Then $\tau_1[B]$ is easily seen to be equal to the 1-numerical range of B^{T} with respect to $\|\cdot\|_\infty$.
- Let $\|\cdot\| = \|\cdot\|_2$. Then $\tau_1[B]$ equals the classical numerical range $\{x^* B x : x \in \mathbb{C}^s \text{ with } x^* x = 1\}$ – see the papers mentioned earlier.

4.3. Bounds on the exponential function of a matrix

In Section 3 we focused on stability of the time stepping process (3.2). In this subsection we shall investigate stability of the underlying initial value problem (3.1) itself.

Suppose the initial value u_0 in (3.1) is replaced by a slightly perturbed vector $\widetilde{u}_0$ and $\widetilde{U}(t)$ is the solution to (3.1) with initial value $\widetilde{u}_0$. In analogy to Section 1.2, (3.1) is said to be *stable* if a small perturbation $v_0 = \widetilde{u}_0 - u_0$ always yields errors $V(t) = \widetilde{U}(t) - U(t)$ (for $t > 0$) that are also small. Therefore, the stability analysis of the initial value problem (3.1) amounts to bounding $V(t)$ (for $t > 0$) suitably in terms of v_0. Since $V(t) = \mathrm{e}^{tA} v_0$ we consider the stability property

$$\|\mathrm{e}^{tA}\| \leq M_0 \quad \text{for all } t \geq 0, \tag{4.14}$$

where M_0 is a positive constant and $\|\cdot\|$ a norm on $\mathbb{C}^{s,s}$.

By using the Jordan canonical form of A it can be easily seen that there exists a finite M_0 with the stability property (4.14) if and only if the following eigenvalue condition is satisfied:

$$\begin{aligned} &\text{All eigenvalues } \lambda \text{ of } A \text{ have a real part } \mathrm{Re}\,\lambda \leq 0, \\ &\text{and the geometric multiplicity of each eigenvalue } \lambda \\ &\text{with } \mathrm{Re}\,\lambda = 0 \text{ is equal to its algebraic multiplicity.} \end{aligned} \tag{4.15}$$

Similar to the situation for the eigenvalue conditions (1.7) and (3.4), eigenvalue condition (4.15) can be reliable (e.g. for normal matrices) or misleading (for families of matrices that are not normal). A reliable criterion for the stability property (4.14), in general situations, can be based on the resolvent

of A, and reads

$$\begin{aligned}&\zeta I - A \text{ is invertible and } \|(\zeta I - A)^{-1}\| \le M_1 (\operatorname{Re}\zeta)^{-1} \\ &\text{for all } \zeta \text{ with } \operatorname{Re}\zeta > 0.\end{aligned} \tag{4.16}$$

In the following we shall discuss the relation between the stability property (4.14) and the resolvent condition (4.16).

By using the formula

$$(\zeta I - A)^{-1} = \int_0^\infty \mathrm{e}^{-\zeta t} \mathrm{e}^{tA}\, \mathrm{d}t \quad \text{for all } \zeta \text{ with } \operatorname{Re}\zeta > 0$$

one can easily see that (4.14) implies (4.16) with $M_1 = M_0$. Conversely, (4.16) implies (4.14) with M_0 depending only on M_1 and the dimension s, but otherwise independent of A. Various authors have studied the size of the optimal M_0 as a function of M_1 and s for the spectral norm or other special norms (see Miller (1968), Laptev (1975), Gottlieb and Orszag (1977) and LeVeque and Trefethen (1984)). The following theorem sharpens and generalizes their results to the case of arbitrary norms on $\mathbb{C}^{s,s}$.

Theorem 4.8 Let $s \ge 1$, $A \in \mathbb{C}^{s,s}$ and $\|\cdot\|$ denote an arbitrary norm on $\mathbb{C}^{s,s}$. If (4.16) holds for some M_1, then

$$\|\mathrm{e}^{tA}\| \le \mathrm{e} s M_1 \quad \text{for all } t \ge 0. \tag{4.17}$$

Proof. The proof is analogous to that of Theorem 2.1, and is based on the representation of e^{tA} for $t > 0$ as

$$\mathrm{e}^{tA} = \frac{1}{2\pi \mathrm{i}} \int_{\operatorname{Re}\zeta = t^{-1}} \mathrm{e}^{t\zeta} (\zeta I - A)^{-1}\, \mathrm{d}\zeta$$

(see also LeVeque and Trefethen (1984)). □

The sharpness of the bound (4.17) is considered in the following theorem, which generalizes a result by LeVeque and Trefethen (1984) for the spectral norm to the case of arbitrary norms on $\mathbb{C}^{s,s}$.

Theorem 4.9 Let $s \ge 2$ and an arbitrary norm $\|\cdot\|$ on $\mathbb{C}^{s,s}$ be given. Then we have for all $t > 0$

$$\begin{aligned}&\sup\{\|\mathrm{e}^{tA}\|/M_1(A) : A \in \mathbb{C}^{s,s}, M_1(A) < \infty\} \ge \\ &\qquad \frac{s^s}{(s-1)!} \mathrm{e}^{-(s-1)} > \mathrm{e}(2\pi)^{-1/2}(s-1)^{1/2},\end{aligned} \tag{4.18}$$

where $M_1(A)$ denotes the smallest M_1 such that (4.16) holds (we define $M_1(A) = \infty$ if (4.16) is not fulfilled for any M_1).

Proof. Let $t > 0$ be given. Define $A \in \mathbb{C}^{s,s}$ by $A = -\alpha I + \gamma E$ where $\alpha > 0$ will be specified later, $\gamma > 0$ is large and E is the matrix defined by (1.9). After some calculations similar to those in the proof of Theorem 2.2

we obtain the relations

$$\begin{aligned} M_1(A) &\leq s^{-s}(s-1)^{s-1}\alpha^{1-s}\gamma^{s-1}\|E^{s-1}\|\left(1+\mathcal{O}\left(\gamma^{-1}\right)\right), \\ \|\mathrm{e}^{tA}\| &= \mathrm{e}^{-\alpha t}\frac{t^{s-1}}{(s-1)!}\gamma^{s-1}\|E^{s-1}\|\left(1+\mathcal{O}\left(\gamma^{-1}\right)\right) \end{aligned}$$

and hence

$$\|\mathrm{e}^{tA}\|/M_1(A) \geq \mathrm{e}^{-\alpha t}(\alpha t)^{s-1}s^s(s-1)^{1-s}/(s-1)! + \mathcal{O}\left(\gamma^{-1}\right) \quad (\text{as } \gamma \to \infty).$$

If we choose $\alpha = (s-1)/t$, the right-hand side of the inequality tends to $s^s\mathrm{e}^{-(s-1)}/\,(s-1)!$ as $\gamma \to \infty$, which is strictly larger than $\mathrm{e}(2\pi)^{-1/2}(s-1)^{1/2}$ by Stirling's formula (see e.g. the proof of Theorem 2.4). □

Note that the upper bound $\|\mathrm{e}^{tA}\|/M_1(A) \leq es$ of Theorem 4.8 and the lower bound (4.18) of Theorem 4.9 differ by a factor $\sim \sqrt{2\pi s}$. This is a less satisfactory situation than in Section 2.2, where the upper bound $\|B^n\|/M_1(B) \leq es$ was shown to be essentially sharp. Further, Theorem 4.9 does not shed any light on the sharpness question for *fixed* constants M_1, since arbitrarily large $M_1(A)$ are allowed in (4.18).

For the special situation where (4.16) holds with $M_1 = 1$, the upper bound (4.17) can be improved. This is the content of the following theorem, which is a well-known result in semigroup theory (see, e.g., Pazy (1983) or Theorem 4.7 above).

Theorem 4.10 Let $M_1 = 1$ and $\|\cdot\|$ be a matrix norm. Then (4.16) implies (4.14) with $M_0 = 1$.

In the remainder of this section we will answer the question whether – in addition to the upper bound (4.17) – there exists an upper bound depending only on t and M_1. This would be analogous to the situation in Section 2, where $\|B^n\|$ was not only bounded by esM_1, but also by $\mathrm{e}(n+1)M_1$ (see Theorem 2.1). Clearly, this question is equivalent to the existence of a function g such that

$$\|\mathrm{e}^{tA}\| \leq g(t, M_1) \quad \text{for all } t \geq 0,$$

whenever resolvent condition (4.16) is fulfilled. The nonexistence of such a function g is proved in Theorem 4.11.

Theorem 4.11 The matrices

$$A_s = \begin{pmatrix} -1 & -2 & \cdots & -2 \\ & -1 & \ddots & \vdots \\ & & \ddots & -2 \\ 0 & & & -1 \end{pmatrix} \in \mathbb{C}^{s,s}, \quad s \geq 1,$$

satisfy the resolvent condition (4.16) with $\|\cdot\| = \|\cdot\|_\infty$ and $M_1 = 2$. Moreover we have

$$\lim_{s\to\infty} \|e^{tA_s}\|_\infty = \infty \quad \text{for all } t > 0. \tag{4.19}$$

Proof. For $A = A_s$ we have $A = -(I+E)(I-E)^{-1}$, where E is the matrix defined by (1.9). Hence we arrive for $\zeta \in \mathbb{C}$ with $\mathrm{Re}\,\zeta > 0$ at

$$(\zeta I - A)^{-1} = \frac{1}{\zeta+1}\left\{ I - \frac{2}{\zeta+1}\sum_{j=1}^{s-1}\left(\frac{\zeta-1}{\zeta+1}\right)^{j-1} E^j \right\},$$

from which we obtain

$$\begin{aligned}(\mathrm{Re}\,\zeta)\|(\zeta I - A)^{-1}\|_\infty &\le \frac{\mathrm{Re}\,\zeta}{|\zeta+1|}\left\{1 + \frac{2}{|\zeta+1|}\sum_{j=1}^{\infty}\left|\frac{\zeta-1}{\zeta+1}\right|^{j-1}\right\} \\ &= (\mathrm{Re}\,\zeta)|\zeta+1|^{-1}\{1 + 2(|\zeta+1| - |\zeta-1|)^{-1}\} \le 2,\end{aligned}$$

implying (4.16) with constant $M_1 = 2$.

In order to prove (4.19) we fix $t > 0$ and define the complex function f by $f(\zeta) = \exp[-t(1+\zeta)(1-\zeta)^{-1}]$ (for all $\zeta \neq 1$). The function f is analytic on $\mathbb{C} \setminus \{1\}$ and can therefore be represented on the open unit disk by a power series

$$f(\zeta) = \sum_{n=0}^{\infty} a_n(t)\zeta^n.$$

Since $f(e^{i\theta}) = \exp[-it/\tan(\frac{1}{2}\theta)]$ (for small positive θ), we see that the limit $\lim_{\theta\to 0} f(e^{i\theta})$ does not exist, implying that

$$\sum_{n=0}^{\infty} |a_n(t)| = \infty. \tag{4.20}$$

The proof of (4.19) is completed by combining (4.20) with

$$e^{tA} = f(E) = \sum_{n=0}^{s-1} a_n(t)E^n, \quad \|e^{tA}\|_\infty = \sum_{n=0}^{s-1} |a_n(t)|.$$

□

We remark that after completion of the present paper new results related to this were found by Kraaijevanger (1992) for the maximum norm $\|\cdot\|_\infty$.

5. Applications and examples

5.1. Range of applications

It is clear from Sections 1.2 and 3.1 that the stability estimates discussed in Sections 2 and 3 are relevant to numerical processes for solving linear

differential equations which are essentially more general than the classical test problems mentioned in Section 1.1. The results of Sections 2 and 3 have a potential for clarifying actual stability problems in cases where Fourier transformation techniques are unlikely to be successful. Such cases comprise linear differential equations with nonsmooth variable coefficients, spectral methods, and finite difference or finite element methods with highly irregular geometries.

Still, at first sight, most of the stability results in Sections 2 and 3 may be considered to be quite weak in that the upper bounds for $\|B^n\|$ do not remain bounded as $n \to \infty$ or $s \to \infty$. However, in computational practice troublesome instability usually manifests itself by an exponential growth of the error. Evidently such growth is not possible when the upper bounds of Sections 2 and 3 are in force – these upper bounds grow at the rate of some power of s or n. In fact, various authors have allowed such polynomial growth in their definition of stability – e.g. Strang (1960), Forsythe and Wasow (1960) and Gottlieb and Orszag (1977).

In Section 1.2 we indicated that bounds on $\|B^n\|$ are useful when analysing the propagation of *rounding errors* $v_0 = \tilde{u}_0 - u_0$. But the stability estimates of Sections 2 and 3 are also relevant to the question of how fast the so-called *global discretization errors*

$$d_n = U(nh) - u_n \tag{5.1}$$

approach zero when $h = \Delta t \to 0$. Here $U(t)$, u_n satisfy (3.1) and (3.2), respectively. We define the *local discretization error* e_n by $e_n = h^{-1} r_n$, where r_n denotes the residual in the right-hand member of (3.2) when u_n and u_{n-1} in that formula are replaced by $U(nh)$ and $U((n-1)h)$, respectively. Writing $B = \varphi(hA)$ we then have $d_n = B d_{n-1} + h e_n$ and therefore

$$d_n = h \sum_{j=1}^{n} B^{n-j} e_j. \tag{5.2}$$

From this representation it is evident that the stability estimates from Sections 2 and 3, in combination with bounds on the local discretization errors, can be used to derive bounds on the errors (5.1). We note that the same holds true when in the numerical solution of a given partial differential equation, with solution $u(x,t)$, one replaces the vector $U(nh)$ in (5.1) by a suitable projection in $\mathbb{C}^s$ of the true $u(x,t)$. Of course e_n should then be defined accordingly.

If $nh = t > 0$ is fixed, and the bounds on $\|B^n\|$ grow with some power of n (or s), then a straightforward application of (5.2) yields bounds on the global errors that are of a *lower* order than the local discretization errors. But, Strang (1960) has already shown that, even in the presence of such polynomial growth, it may be possible to establish bounds on the global

discretization errors that are of the *same* order as the local errors – provided the problem itself is sufficiently smooth.

For subsequent extensions of Strang's result to linear and nonlinear problems, see, e.g., Strang (1964), Spijker (1972), Brenner and Thomée (1979), Thomée (1990), and the references therein. In the case of nonlinear problems the basic assumptions in these papers include the requirement that a linearization of the actual numerical process is stable (in the sense that polynomial growth is allowed). Therefore the stability analysis of linear processes (as in the present paper) may contribute to the stability analysis of numerical processes for nonlinear differential equations, see also López-Marcos and Sanz-Serna (1988).

We finally comment on the relevance of the bounds on $\|\exp(tA)\|$ obtained in Section 4.3. Similar to the situation for $\|\varphi(hA)^n\|$ discussed above, these bounds are not only relevant for studying the effect of initial perturbations $v_0 = \widetilde{u}_0 - u_0$ (such as rounding errors) on the solution $U(t)$ of initial value problem (3.1), but also for analysing the global discretization error

$$d(t) = \widetilde{U}(t) - U(t)$$

when (3.1) is obtained by semi-discretization of a partial differential equation. Here $\widetilde{U}(t)$ denotes a suitable projection in $\mathbb{C}^s$ of the solution to the partial differential equation. Defining the corresponding local discretization error $e(t)$ to be the residual appearing in the right-hand side of the differential equation in (3.1) when $U(t)$ is replaced by $\widetilde{U}(t)$, we readily obtain $d'(t) = Ad(t) + e(t)$, so that

$$d(t) = \int_0^t \exp((t-\tau)A)e(\tau)\,\mathrm{d}\tau.$$

From this representation, which is a continuous analogue of (5.2), one can derive bounds on the global errors $d(t)$ by combining bounds on the local errors $e(t)$ and the bounds on $\|\exp(tA)\|$ obtained in Section 4.3.

5.2. Examples pertinent to the theory of Section 3

In order to illustrate some of the preceding notions and theorems we consider the simple initial-boundary value problem

$$\begin{aligned} &u_t(x,t) = (a(x)u(x,t))_x + g(x,t), \\ &u(x,0) = f(x), \quad u(1,t) = 0, \quad \text{where } 0 < x < 1,\ t > 0. \end{aligned} \tag{5.3}$$

Here a, g, f denote given functions with $a(x) \geq 0$. The values $u(x,t)$ are considered unknown for $0 \leq x < 1$, $t > 0$.

We select an integer $s \geq 1$ and define $\Delta x = 1/s$. Approximating $(au)_x$ by the forward difference quotient (see, e.g., Richtmyer and Morton (1967))

$$(a(x)u(x,t))_x \simeq (\Delta x)^{-1}\{a(x+\Delta x)u(x+\Delta x,t) - a(x)u(x,t)\},$$

problem (5.3) is transformed into a semi-discrete problem of the form (3.1) with $A = (\alpha_{jk})$, where

$$\begin{cases} \alpha_{jj} &= -sa((j-1)/s) & (j = 1, 2, \ldots, s), \\ \alpha_{j,j+1} &= sa(j/s) & (j = 1, 2, \ldots, s-1), \\ \alpha_{jk} &= 0 & \text{otherwise.} \end{cases}$$

Further,

$$\begin{aligned} b(t) &= (g(0,t), g(1/s,t), \ldots, g((s-1)/s,t))^{\mathrm{T}}, \\ u_0 &= (f(0), f(1/s), \ldots, f((s-1)/s))^{\mathrm{T}}, \end{aligned}$$

and the jth component $U_j(t)$ of the solution $U(t)$ to (3.1) approximates the solution $u(x,t)$ to (5.3) at $(x,t) = ((j-1)/s, t)$ (for $j = 1, 2, \ldots, s$).

In the following we focus on conditions that guarantee the stability of the fully discrete numerical process (3.2). We will derive upper bounds for $\|\varphi(hA)^n\|_p$ in the cases $p = 1$ and $p = \infty$. For simplicity we assume that the ratio $\mu = h/\Delta x$ is fixed. Further we introduce the constants

$$\alpha = \max_{0 \le x \le 1} a(x), \quad \beta = \max_{0 \le x \le 1} a'(x).$$

Case 1: $p = 1$. For the norm $\|\cdot\|_1$ one easily verifies that the matrix hA satisfies

$$\|hA + \alpha\mu I\|_1 \le \alpha\mu.$$

Applying part (a) of Theorem 2.1 to the matrix $B = I + (\alpha\mu)^{-1}hA$ we see that hA satisfies the resolvent condition (3.5) with

$$M_1 = 1 \quad \text{and} \quad V = \{\zeta : |\zeta + \alpha\mu| \le \alpha\mu\}.$$

Suppose that $\alpha\mu \le r$, where r is the stability radius, which was defined in Section 3.2 (Result 9) to be the radius of the largest disk in the complex left half-plane which is tangent to the imaginary axis at the origin and lies in the stability region S (defined by (3.3)). Then it follows from Remark 3.2 and the material in Section 3.2 that

$$\|\varphi(hA)^n\|_1 \le \gamma \min(s, \sqrt{n}) \quad \text{for } n = 1, 2, 3, \ldots,$$

where γ depends only on φ.

Under the more stringent condition $\alpha\mu < r$ it follows from Result 9 (with $r_0 = \alpha\mu$) and Remark 3.2 that we even have

$$\|\varphi(hA)^n\|_1 \le \gamma \quad \text{for } n = 1, 2, 3, \ldots,$$

where γ depends only on φ and $\alpha\mu$. *Case 2:* $p = \infty$. For the norm $\|\cdot\|_\infty$ one easily verifies that the matrix hA satisfies

$$\|hA + \alpha\mu I\|_\infty \le \alpha\mu + \beta h.$$

When $\beta \le 0$ we can proceed as in Case 1. In the following we assume that $\beta > 0$.

An application of part (a) of Theorem 2.1 to the matrix $B = (\alpha\mu + \beta h)^{-1}(hA + \alpha\mu I)$ shows that hA satisfies the resolvent condition (3.5) with

$$M_1 = 1 \quad \text{and} \quad V = \{\zeta : |\zeta + \alpha\mu| \leq \alpha\mu + \beta h\}.$$

Suppose that $\alpha\mu \leq r$ and $r < \infty$. Let the stability region S be bounded and $\varphi'(\zeta) \neq 0$ on ∂S. Then it follows from Remark 3.4 (parts (a) and (b)) that we have

$$\|\varphi(hA)^n\|_\infty \leq \gamma_1 e^{\gamma_2 \beta n h} \min(s, \sqrt{n}) \quad \text{for} \quad n = 1, 2, 3, \ldots$$

whenever $\beta h \leq \gamma_3$. Here $\gamma_1, \gamma_2, \gamma_3 > 0$ only depend on φ.

In case $r = \infty$ we can apply the general result mentioned in Remark 3.4 (part (c)) so as to obtain a similar stability estimate.

Further illustrations of the theory of Section 3 can be found, e.g., in Lenferink and Spijker (1991b) and Reddy and Trefethen (1992). In Kraaijevanger *et al.* (1987) an example was presented pertinent to problem (1.3).

5.3. Numerical illustrations

In order to give a numerical illustration of the material of Section 5.2 we consider the classical fourth-order Runge–Kutta method (see, e.g., Butcher (1987)). Applying this method to the semi-discrete problem (3.1) as specified in Section 5.2, one arrives at a fully discrete process (3.2) with

$$\varphi(\zeta) = 1 + \zeta + \frac{\zeta^2}{2!} + \frac{\zeta^3}{3!} + \frac{\zeta^4}{4!}. \tag{5.4}$$

The corresponding stability radius r is equal to

$$r = 1.393 \tag{5.5}$$

(rounded to four decimal places). For later use we note that it follows from the definition of r that

$$\text{the interval } [-2r, 0] \text{ is contained in } S. \tag{5.6}$$

We consider the matrix A, defined in Section 5.2, with three different choices for the function $a(x)$, viz.

$$a_1(x) = 1, \quad a_2(x) = 1 - x^{10}, \quad a_3(x) = 1 - x.$$

Using the notations of the preceding subsection, we have for all of these functions that

$$\alpha = 1, \quad \beta \leq 0.$$

For given s and function $a(x)$ we shall measure the stability of the corresponding numerical process by the quantity

$$c(\mu, a) = \sup_{n \geq 0} \|\varphi(hA)^n\|_\infty.$$

Table 1. *Values of* $c(\mu, a_i)$ *for* $s = 80$

h	μ	$c(\mu, a_1)$	$c(\mu, a_2)$	$c(\mu, a_3)$
0.0125	1	1	1	1
0.0150	1.2	1.12	1.12	1.08
0.0175	1.4	2.26	2.26	1.30
0.0200	1.6	3.47×10^{9}	2.58×10^{7}	4.20
0.0225	1.8	4.93×10^{19}	7.56×10^{15}	1.73×10^{2}
0.0250	2.0	9.06×10^{50}	2.68×10^{25}	4.17×10^{4}

For $s = 80$ we have listed some values of $c(\mu, a_i)$ in Table 1. In the table we see good stability up to $\mu = 1.4$. This is perfectly in agreement with the conditions of Section 5.2 since, in view of (5.5), the requirement

$$\alpha\mu \leq r$$

amounts to $\mu \leq 1.393$. For $\mu > 1.4$ we see large values in the table, indicating strong instability. It is worth noting that for all $\mu \leq 2.0$ requirement (3.4) is still fulfilled, since for these μ we have

$$\sigma[hA] \subset S \setminus \partial S.$$

This inclusion follows from (5.5) and (5.6) and the fact that, for our functions a_i,

$$\sigma[hA] \subset [-\mu, 0).$$

The numerical results thus confirm the reliability of the stability criteria discussed in Section 5.2, and the failing of the eigenvalue condition (3.4).

For further numerical illustrations related to the material of Sections 2 and 3 we refer to Trefethen (1988), Lenferink and Spijker (1991b) and Reddy and Trefethen (1992). For a numerical illustration pertinent to problem (1.3) see Kraaijevanger *et al.* (1987).

Acknowledgements

The authors wish to thank L.N. Trefethen and S.C. Reddy for many stimulating discussions on the topic of this paper. They are also indebted to L.N. Trefethen and A. Iserles for editorial comments on a preliminary version of this paper, and to J. Groeneweg for performing the computations displayed in Section 5.3. This research has been supported by the Netherlands Organization for Scientific Research (N.W.O.). This research has been made possible by the award of a fellowship of the Royal Netherlands Academy of Arts and Sciences (K.N.A.W.) to Kraaijevanger.

REFERENCES

M. Abramowitz and I.A. Stegun (1965), *Handbook of Mathematical Functions*, Dover (New York).

F.F. Bonsall and J. Duncan (1971), *Numerical Ranges of Operators on Normed Spaces and of Elements of Normed Algebras*, Cambridge University Press (Cambridge).

F.F. Bonsall and J. Duncan (1980), 'Numerical ranges', in *Studies in Functional Analysis* (R.G. Bartle, ed.), The Mathematical Association of America, 1–49.

Ph. Brenner and V. Thomée (1979), 'On rational approximations of semigroups', *SIAM J. Numer. Anal.* **16**, 683–694.

J.C. Butcher (1987), *The Numerical Analysis of Ordinary Differential Equations*, John Wiley (Chichester).

C. Canuto, M.Y. Hussaini, A. Quarteroni and T.A. Zang (1988), *Spectral Methods in Fluid Dynamics*, Springer (New York).

J.B. Conway (1985), *A Course in Functional Analysis*, Springer (New York).

M. Crouzeix (1987), 'On multistep approximation of semigroups in Banach spaces', *J. Comput. Appl. Math.* **20**, 25–35.

M. Crouzeix, S. Larsson, S. Piskarev and V. Thomée (1991), 'The stability of rational approximations of analytic semigroups', Technical Report 1991:28, Chalmers University of Technology & the University of Göteborg (Göteborg).

G.E. Forsythe and W.R. Wasow (1960), *Finite Difference Methods for Partial Differential Equations*, John Wiley (New York).

D. Gottlieb and S.A. Orszag (1977), *Numerical Analysis of Spectral Methods*, Soc. Ind. Appl. Math. (Philadelphia).

D.F. Griffiths, I. Christie and A.R. Mitchell (1980), 'Analysis of error growth for explicit difference schemes in conduction-convection problems', *Int. J. Numer. Meth. Engrg* **15**, 1075–1081.

R.D. Grigorieff (1991), 'Time discretization of semigroups by the variable two-step BDF method', in *Numerical Treatment of Differential Equations* (K. Strehmel, ed.), Teubner (Leipzig), 204–216.

B. Gustafsson, H.-O. Kreiss and A. Sundström (1972), 'Stability theory of difference approximations for mixed initial boundary value problems. II', *Math. Comput.* **26**, 649–686.

E. Hairer and G. Wanner (1991), *Solving Ordinary Differential Equations*, Vol. II, Springer (Berlin).

P. Henrici (1974), *Applied and Computational Complex Analysis*, Vol. 1, John Wiley (New York).

R.A. Horn and C.R. Johnson (1990), *Matrix Analysis*, Cambridge University Press (Cambridge).

F. John (1952), 'On integration of parabolic equations by difference methods', *Comm. Pure Appl. Math.* **5**, 155–211.

J.F.B.M. Kraaijevanger (1992), 'Two counterexamples related to the Kreiss matrix theorem', submitted to *BIT*.

J.F.B.M. Kraaijevanger, H.W.J. Lenferink and M.N. Spijker (1987), 'Stepsize restrictions for stability in the numerical solution of ordinary and partial differential equations', *J. Comput. Appl. Math.* **20**, 67–81.

H.-O. Kreiss (1962), 'Über die Stabilitätsdefinition für Differenzengleichungen die partielle Differentialgleichungen approximieren', *BIT* **2**, 153–181.

H.-O. Kreiss (1966), 'Difference approximations for the initial-boundary value problem for hyperbolic differential equations', in *Numerical Solution of Nonlinear Differential Equations* (D. Greenspan, ed.), John Wiley (New York), 141–166.

H.-O. Kreiss (1990), 'Well posed hyperbolic initial boundary value problems and stable difference approximations', in *Proc. Third Int. Conf. on Hyperbolic Problems, Uppsala, Sweden.*

H.-O. Kreiss and L. Wu (1992), 'On the stability definition of difference approximations for the initial boundary value problem', to appear in *Comm. Pure Appl. Math.*

H.J. Landau (1975), 'On Szegő's eigenvalue distribution theorem and non-Hermitian kernels', *J. d'Analyse Math.* **28**, 335–357.

G.I. Laptev (1975), 'Conditions for the uniform well-posedness of the Cauchy problem for systems of equations', *Sov. Math. Dokl.* **16**, 65–69.

H.W.J. Lenferink and M.N. Spijker (1990), 'A generalization of the numerical range of a matrix', *Linear Algebra Appl.* **140**, 251–266.

H.W.J. Lenferink and M.N. Spijker (1991a), 'On a generalization of the resolvent condition in the Kreiss matrix theorem', *Math. Comput.* **57**, 211–220.

H.W.J. Lenferink and M.N. Spijker (1991b), 'On the use of stability regions in the numerical analysis of initial value problems', *Math. Comput.* **57**, 221–237.

R.J. LeVeque and L.N. Trefethen (1984), 'On the resolvent condition in the Kreiss matrix theorem', *BIT* **24**, 584–591.

J.C. López-Marcos and J.M. Sanz-Serna (1988), 'Stability and convergence in numerical analysis III: linear investigation of nonlinear stability', *IMA J. Numer. Anal.* **8**, 71–84.

C. Lubich (1991), 'On the convergence of multistep methods for nonlinear stiff differential equations', *Numer. Math.* **58**, 839–853.

C. Lubich and O. Nevanlinna (1991), 'On resolvent conditions and stability estimates', *BIT* **31**, 293–313.

C.A. McCarthy (1992), 'A strong resolvent condition need not imply power-boundedness', to appear in *J. Math. Anal. Appl.*

C.A. McCarthy and J. Schwartz (1965), 'On the norm of a finite boolean algebra of projections, and applications to theorems of Kreiss and Morton', *Comm. Pure Appl. Math.* **18**, 191–201.

T. Meis and U. Marcowitz (1981), *Numerical Solution of Partial Differential Equations*, Springer (New York).

J. Miller (1968), 'On the resolvent of a linear operator associated with a well-posed Cauchy problem', *Math. Comput.* **22**, 541–548.

J. Miller and G. Strang (1966), 'Matrix theorems for partial differential and difference equations', *Math. Scand.* **18**, 113–123.

K.W. Morton (1964), 'On a matrix theorem due to H.O. Kreiss', *Comm. Pure Appl. Math.* **17**, 375–379.

K.W. Morton (1980), 'Stability of finite difference approximations to a diffusion-convection equation', *Int. J. Num. Meth. Engrg* **15**, 677–683.

O. Nevanlinna (1984), 'Remarks on time discretization of contraction semigroups', Report HTKK–MAT–A225, Helsinki Univ. Techn. (Helsinki).

J.T. Oden and J.N. Reddy (1976), *An Introduction to the Mathematical Theory of Finite Elements*, John Wiley (New York).

S.V. Parter (1962), 'Stability, convergence, and pseudo-stability of finite-difference equations for an over-determined problem', *Numer. Math.* **4**, 277–292.

A. Pazy (1983), *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer (New York).

C. Pearcy (1966), 'An elementary proof of the power inequality for the numerical radius', *Michigan Math. J.* **13**, 289–291.

S.C. Reddy and L.N. Trefethen (1990), 'Lax-stability of fully discrete spectral methods via stability regions and pseudo-eigenvalues', *Comput. Meth. Appl. Mech. Engrg* **80**, 147–164.

S.C. Reddy and L.N. Trefethen (1992), 'Stability of the method of lines', *Numer. Math.* **62**, 235–267.

L. Reichel and L.N. Trefethen (1992), 'Eigenvalues and pseudo-eigenvalues of Toeplitz matrices', *Linear Algebra Appl.* **162–164**, 153–185.

R.D. Richtmyer and K.W. Morton (1967), *Difference Methods for Initial-value Problems*, 2nd Ed., John Wiley (New York).

W. Rudin (1973), *Functional Analysis*, McGraw-Hill (New York).

J.C. Smith (1985), 'An inequality for rational functions', *Amer. Math. Monthly* **92**, 740–741.

M.N. Spijker (1972), 'Equivalence theorems for nonlinear finite-difference methods', *Springer Lecture Notes in Mathematics*, Vol. **267**, Springer (Berlin), 233–264.

M.N. Spijker (1985), 'Stepsize restrictions for stability of one-step methods in the numerical solution of initial value problems', *Math. Comput.* **45**, 377–392.

M.N. Spijker (1991), 'On a conjecture by LeVeque and Trefethen related to the Kreiss matrix theorem', *BIT* **31**, 551–555.

W.G. Strang (1960), 'Difference methods for mixed boundary-value problems', *Duke Math. J.* **27**, 221–231.

G. Strang (1964), 'Accurate partial difference methods II. Non–linear problems', *Numer. Math.* **6**, 37–46.

G. Strang and G.J. Fix (1973), *An Analysis of the Finite Element Method*, Prentice-Hall (Englewood Cliffs).

E. Tadmor (1981), 'The equivalence of L_2-stability, the resolvent condition, and strict H-stability', *Linear Algebra Appl.* **41**, 151–159.

V. Thomée (1990), 'Finite difference methods for linear parabolic equations', in *Handbook of Numerical Analysis I* (P.G. Ciarlet and J.L. Lions, eds), North-Holland (Amsterdam), 5–196.

L.N. Trefethen (1988), 'Lax-stability vs. eigenvalue stability of spectral methods', in *Numerical Methods for Fluid Dynamics III* (K.W. Morton and M.J. Baines, eds), Clarendon Press (Oxford), 237–253.

L.N. Trefethen (1992), *Spectra and Pseudospectra*, book in preparation.

J.M. Varah (1979), 'On the separation of two matrices', *SIAM J. Numer. Anal.* **16**, 216–222.

E. Wegert and L.N. Trefethen (1992), 'From the Buffon needle problem to the Kreiss matrix theorem', to appear in *Amer. Math. Monthly*.

Published by the Press Syndicate of the University of Cambridge
The Pitt Building, Trumpington Street, Cambridge CB2 1RP
40 West 20th Street, New York, NY 10011-4211, USA
10 Stamford Road, Oakleigh, Melbourne 3166, Australia

First published 1993

Printed in Canada

Library of Congress cataloging in publication data available

A catalogue record for this book is available from the British Library

ISBN 0-521-443563 hardback