ELSEVIER

Applied Numerical Mathematics 24 (1997) 233–246

APPLIED NUMERICAL MATHEMATICS

Numerical stability, resolvent conditions and delay differential equations

M.N. Spijker

Department of Mathematics and Computer Science, University of Leiden, Niels Bohrweg 1, 2333 CA Leiden, Netherlands

Abstract

This paper deals with the problem of estimating a priori the error growth in discretizations of linear initial value problems.

A review is presented of various recent stability estimates which are valid under the Kreiss resolvent condition or under a strengthened version thereof. Moreover, a conjecture is formulated to the effect that errors cannot grow at a faster rate than s^{β}, where $\beta < 1$ and s denotes the order of the matrices under consideration.

Also a weaker version of the Kreiss resolvent condition is discussed. Under that condition, a stability estimate is proved which grows linearly with the order of the matrices under consideration.

The paper concludes by presenting an application of the Kreiss resolvent condition in the error growth analysis for discretizations of delay differential equations. © 1997 Elsevier Science B.V.

1. Introduction

1.1. The purpose of the paper

This paper is concerned with step-by-step methods for the numerical solution of linear initial value problems. It deals with initial value problems for differential equations without delay-term as well as with equations in which a delayed argument of the dependent variable occurs.

A crucial question in the step-by-step solution of such problems is whether the numerical process will behave *stably* or not. Here the term stable is used to designate that any numerical errors introduced at some stage of the calculations are propagated in a mild fashion.

Recently new tools for stability analysis were developed that are related to the so-called *resolvent condition of Kreiss*. These tools can be used in proving stability. The purpose of this paper is to review, extend and apply some of these recent results.

PII S0168-9274(97)00023-8

1.2. Organization of the paper

In Sections 2 and 3 we deal with step-by-step methods specified by $s \times s$ matrices B. We relate the stability analysis of these methods to the problem of establishing suitable upper bounds on the norm $\|B^n\|$ (for $n \geqslant 1$, $s \geqslant 1$). Further, we recall the fact that the eigenvalues of B can be a highly misleading guide to estimating $\|B^n\|$.

Section 4 reviews upper bounds for $\|B^n\|$, valid under the Kreiss resolvent condition; these bounds grow linearly with $n \geqslant 1$ or $s \geqslant 1$.

Section 5 discusses improved upper bounds for $\|B^n\|$, valid under a slightly *stronger* version of the Kreiss condition. A conjecture is presented according to which the norm $\|B^n\|$ grows at most at a rate s^β, with $\beta < 1$, when $s \to \infty$.

Section 6 deals with a *weaker* version of the original Kreiss condition. A theorem is presented which implies that $\|B^n\|$ does not grow faster than linearly with s, under the weaker condition.

Section 7 explores the relevance of the foregoing to the stability analysis of numerical methods for delay differential equations.

2. Stability analysis of linear, numerical processes

We consider the following numerical process:

$$u_n = Bu_{n-1} + b_n \quad (n = 1, 2, 3, \ldots).$$

Here b_n denote given vectors in the s-dimensional complex vectorspace $\mathbb{C}^s$, and B is a given $s \times s$ matrix (with complex entries). Further, $u_n \in \mathbb{C}^s$ are numerical approximations that are computed recursively, for $n = 1, 2, 3, \ldots$, starting from a given $u_0 \in \mathbb{C}^s$.

Processes of the above form arise in a great variety of situations, in particular in the numerical solution of initial value problems for (systems of) linear differential equations—see, e.g., [19] and Sections 3 and 7 of the present paper.

Suppose the numerical computations based on the above process were performed using a slightly perturbed starting vector, say $\widetilde{u}_0$, instead of u_0. For $n \geqslant 1$, we then would obtain approximations $\widetilde{u}_n$, instead of u_n, satisfying

$$\widetilde{u}_n = B\widetilde{u}_{n-1} + b_n \quad (n = 1, 2, 3, \ldots).$$

For instance, $\widetilde{u}_0$ may stand for a finite-digit representation in a computer of the true u_0; the vectors $\widetilde{u}_n$, for $n \geqslant 1$, 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 the above numerical process the crucial question is whether the *propagated error* $v_n = \widetilde{u}_n - u_n$ can be bounded suitably in terms of the initial error $v_0 = \widetilde{u}_0 - u_0$. Therefore, one may be looking for bounds on the propagated error of the form

$$|v_n| \leqslant M|v_0| \quad (n \geqslant 1).$$

Here M stands for a (moderate) constant, and $|x|$ denotes an arbitrary norm for the vectors $x \in \mathbb{C}^s$.

By subtracting the above recurrence relation satisfied by the vectors u_n from the one satisfied by $\widetilde{u}_n$, one obtains $v_n = \widetilde{u}_n - u_n = (B\widetilde{u}_{n-1} + b_n) - (Bu_{n-1} + b_n) = Bv_{n-1}$. Hence

$$v_n = B^n v_0.$$

For arbitrary $s \times s$ matrices $A = (\alpha_{ij})$, we use the notation

$$\|A\| = \sup |Ax|/|x|,$$

where the supremum is over all vectors $x \neq 0$ in $\mathbb{C}^s$. For instance, if $|x| = |x|_\infty = \max_{1 \leqslant i \leqslant s} |\xi_i|$, for vectors x with components ξ_i, then

$$\|A\| = \|A\|_\infty = \max_{1 \leqslant i \leqslant s} \sum_{j=1}^{s} |\alpha_{ij}|.$$

The best upper bound for $|v_n|$ in terms of $|v_0|$, for arbitrary $v_0 \in \mathbb{C}^s$, thus reads

$$|v_n| \leqslant \|B^n\| \cdot |v_0|.$$

Therefore, the stability analysis of our numerical process amounts to deriving bounds on $\|B^n\|$, e.g., of the form

$$\|B^n\| \leqslant M \quad (n \geqslant 0). \tag{1}$$

In the rest of the paper we focus on the general problem of deriving suitable upper bounds on $\|B^n\|$.

3. An eigenvalue condition

Conditions that are sufficient in order that (1) holds are easily formulated in terms of the *eigenvalues* of the matrix B. Consider, e.g., the following condition:

All eigenvalues λ of B have a modulus $|\lambda| < 1$. (2)

From the Jordan canonical form (cf., e.g., [5]) of the matrix B one sees that (2) implies the existence of a constant M with property (1).

However, from a practical point of view, condition (2) can be a very misleading guide to stability. This fact was noticed, e.g., already by Parter [13].

An instructive example, illustrating that condition (2) is unreliable, is provided by the $s \times s$ bidiagonal matrix

$$B = \begin{pmatrix} -\frac{1}{2} & 0 & \dots & \dots & 0 \\ \frac{3}{2} & -\frac{1}{2} & \ddots & & \vdots \\ 0 & \frac{3}{2} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \frac{3}{2} & -\frac{1}{2} \end{pmatrix}.$$

This matrix may be thought of as arising in the numerical solution of the initial-boundary value problem

$$u_t(x,t) + u_x(x,t) = 0, \qquad u(0,t) = 0, \qquad u(x,0) = f(x),$$

where $0 \leqslant x \leqslant 1$, $t \geqslant 0$ and f denotes a given function. Consider the so-called upwind finite difference scheme

$$\frac{1}{\Delta t}(u_{m,n} - u_{m,n-1}) + \frac{1}{\Delta x}(u_{m,n-1} - u_{m-1,n-1}) = 0,$$
$$u_{0,n-1} = 0, \qquad u_{m,0} = f(m\Delta x)$$

(where $\Delta t > 0$, $\Delta x = 1/s$, $u_{m,n} \simeq u(m\Delta x, n\Delta t)$ for $m = 1,2,\ldots,s$ and $n = 1,2,3,\ldots$). Clearly, the vectors $u_n \in \mathbb{C}^s$, with components $u_{m,n}$ $(1 \leqslant m \leqslant s)$, satisfy $u_n = Bu_{n-1}$, with B as specified above when $\Delta t/\Delta x = 3/2$. Note that the order $s = 1/\Delta x$ of B is related to the accuracy of the finite difference scheme; the order of B has to increase without bound if the discretization error in the finite difference scheme is to approach zero. Therefore let us focus on large values s.

For each $s \geqslant 1$, the eigenvalues of the above matrix B equal $-1/2$, so that (2) is fulfilled. But, an easy calculation shows that

$$\|B^n\|_\infty = 2^n \quad \text{for } n = 1,2,\ldots,s-1.$$

Hence, any M for which (1) holds, satisfies

$$M \geqslant 2^{s-1}.$$

For moderately large values of s, say $s \simeq 100$, we have $M \gtrsim 10^{30}$, so that actually instability manifests itself, although the eigenvalue condition (2) is fulfilled.

The above example shows that, under condition (2), the size of the constant M in (1) is not under control; M can grow at an exponential rate with s. In the following we shall review conditions on B under which the size of $\|B^n\|$ is nicely under control.

4. The classical resolvent condition of Kreiss

Kreiss [7] related property (1) to the condition that

$$\zeta I - B \text{ is invertible, and } \|(\zeta I - B)^{-1}\| \leqslant \frac{L}{|\zeta| - 1}, \quad \text{for all complex } \zeta \text{ with } |\zeta| > 1. \tag{3}$$

Here I denotes the identity matrix, and L is a real constant. Since $(\zeta I - B)^{-1}$ is called the *resolvent* of B at ζ, we shall refer to (3) as the *Kreiss resolvent condition.* (In fact, Kreiss dealt only with the special case of the spectral norm $\|A\| = \|A\|_2$, but this restriction is immaterial to the present discussion.) In many cases of practical interest it is easier to verify (3) than (1).

Property (1) implies that B has a spectral radius $\rho(B) \leqslant 1$. For $|\zeta| > 1$ we thus conclude from (1) that $\rho(\zeta^{-1}B) < 1$ and

$$\begin{aligned}\|(\zeta I - B)^{-1}\| &= |\zeta|^{-1} \cdot \|(I - \zeta^{-1}B)^{-1}\| = |\zeta|^{-1}\|I + \zeta^{-1}B + \zeta^{-2}B^2 + \cdots\| \\ &\leqslant |\zeta|^{-1}(1 + |\zeta|^{-1} + |\zeta|^{-2} + \cdots) \cdot M = M \cdot (|\zeta| - 1)^{-1}.\end{aligned}$$

We see that (1) implies (3) with $L = M$. Kreiss succeeded in proving that, conversely, (3) implies (1) with $M = M_{L,s}$ only depending on L and s.

This important result of Kreiss has often been used with great success in the stability analysis of numerical processes, see, e.g., [3,4,12,14].

In the following we focus on cases where the dimension s is large. Therefore, it is important to understand in which way $M_{L,s}$ depends on s. The original proof of Kreiss does not provide a sharp

value for $M_{L,s}$. Improved values for $M_{L,s}$ were obtained successively by Morton [11], Tadmor [17], LeVeque and Trefethen [8] and Spijker [15]; see [21] for a nice historical survey. Eventually, the following theorem was obtained (for its proof see, e.g., [19] and Lemma 5b in Section 6 below).

Theorem 1. *The resolvent condition* (3) *implies that*

$$\|B^n\| \leqslant 4 \cdot L \cdot n \quad (n \geqslant 1,\ s \geqslant 1), \tag{4a}$$

$$\|B^n\| \leqslant \mathrm{e} \cdot L \cdot s \quad (n \geqslant 1,\ s \geqslant 1), \tag{4b}$$

We see that under condition (3) the size of $\|B^n\|$ is rather well under control. There is *no* strong (exponential) instability. Under condition (3) the propagated error increases at most linearly with the number of steps n, and with the dimension s.

In the following we review recent modifications of Theorem 1.

5. A strong version of the Kreiss resolvent condition

The estimate (4) is essentially best possible when the general resolvent condition (3) is in force, see [19]. The question poses itself of whether still better stability estimates can be established under conditions that are slightly stronger than (3) and fulfilled in cases of practical interest. This question was dealt with, e.g., in [10,18]. In the subsequent we review some conclusions pertinent to this question obtained by Spijker and Straetemans [16].

We deal with the following, stronger version of (3):

$$\zeta I - B \text{ is invertible, and } \|(\zeta I - B)^{-1}\| \leqslant \frac{L}{\mathrm{d}(\zeta, W)}, \quad \text{for all } \zeta \in \mathbb{C}\backslash W. \tag{5}$$

Here W is an arbitrary set with the following three properties.

(i) W is a closed subset of the unit disk $\{\zeta\colon\ |\zeta| \leqslant 1\}$.

(ii) The intersection of W and the unit circle $\{\zeta\colon\ |\zeta| = 1\}$ consists only of a finite number of points.

(iii) At each of these points of intersection the contact between W and the unit circle is of order at most q.

In property (iii) we denote by q an arbitrary constant, with $0 \leqslant q < \infty$. The property amounts to the requirement that positive β_0, β_1 exist such that $1 - |\xi| \geqslant \beta_1 |\xi - \zeta|^{1+q}$ whenever ζ is a point of intersection and $\xi \in W$, $|\xi - \zeta| \leqslant \beta_0$ (cf. [16]). Further, in (5) we denote by $\mathrm{d}(\zeta, W)$ the distance from ζ to W.

Under condition (5) we shall look for improvements over (4) of the form

$$\|B^n\| \leqslant \gamma \cdot L \cdot n^{\alpha} \quad (n \geqslant 1,\ s \geqslant 1), \tag{6a}$$

and

$$\|B^n\| \leqslant \gamma \cdot L \cdot s^{\beta} \quad (n \geqslant 1,\ s \geqslant 1). \tag{6b}$$

Here γ denotes an arbitrary constant (only depending on W and q), and α, β are constants (only depending on q) with $0 \leqslant \alpha < 1$, $0 \leqslant \beta < 1$.

In [16], a counterexample is constructed which implies certain lower bounds for the above exponents α, β. The counterexample involves the norm $\|A\| = \|A\|_\infty$, and is relevant to any given fixed $q \geqslant 0$. It deals with a set W satisfying (i)–(iii), with fixed constants $L > 1$, $c > 0$, and with the exponent

$$p = 1 - \frac{1}{1+q}.$$

The counterexample consists in a sequence $\{B_s\}$ of $s \times s$ matrices $(s = 1, 2, 3, \ldots)$ satisfying (5) as well as

$$\|(B_s)^n\| \geqslant c \cdot n^p = c \cdot s^p \quad (\text{for } n = s = 1, 2, 3, \ldots).$$

According to this counterexample, the general condition (5) may imply (6a) or (6b) only when

$$\alpha \geqslant 1 - \frac{1}{1+q} \quad \text{and} \quad \beta \geqslant 1 - \frac{1}{1+q}.$$

The following theorem shows that our lower bound for the exponent α is best possible.

Theorem 2. *Let $q \geqslant 0$, and let W satisfy* (i)–(iii). *Then there exists a constant γ (only depending on W, q) such that* (5) *implies* (6a) *with $\alpha = 1 - 1/(1+q)$.*

For the proof of the theorem we refer to [16]. Note that for $q = 0$ we have $\alpha = 0$, in which case the estimate (6a) corresponds to what sometimes is called *strong stability*. In this case the norm $|v_n|$ of the propagated error does not grow at all, with n or s.

Theorems 1 and 2, together with the above lower bounds for α and β, provide some evidence for the following conjecture.

Conjecture 3. Let $q \geqslant 0$, and let W satisfy (i)–(iii). Then there exists a constant γ (only depending on W, q) such that (5) implies (6b) with $\beta = 1 - 1/(1+q)$.

6. A weak version of the Kreiss resolvent condition

In applications of Theorems 1 and 2 one has to check whether (3) or (5) is satisfied. In certain cases this may be difficult, and one may even not succeed in establishing (3). Therefore, an important issue is the question of whether the stability estimate (4), which follows by virtue of Theorem 1 from (3), is still valid under a resolvent condition that is easier to prove than (3).

We consider the following variant to the Kreiss resolvent condition (3):

$$\begin{aligned} &(\zeta I - B) \text{ is invertible } (\text{for } |\zeta| > 1), \text{ and} \\ &\|(\zeta I - B)^{-1}\| \leqslant L \cdot \frac{|\zeta|^{\delta \cdot s}}{|\zeta| - 1} \quad (\text{for } 1 < |\zeta| < \rho). \end{aligned} \tag{7}$$

Here $\delta \geqslant 0$, $\rho > 1$ denote constants, and s is the order of the matrix B. Clearly, (7) is a weaker version of (3) and therefore easier to prove.

The following theorem shows that, under condition (7), one can partly reach the same conclusion as in Theorem 1.

Theorem 4. *Let $\delta \geqslant 0$, $\rho > 1$, and B of order s. Then condition* (7) *implies*

(a) $\|B^n\| \leqslant \gamma \cdot 4L \cdot (n + \delta s) \quad (n \geqslant 1,\ s \geqslant 1)$,

(b) $\|B^n\| \leqslant \gamma \cdot \mathrm{e}L \cdot (1 + \delta)s \quad (n \geqslant 1,\ s \geqslant 1)$,

where $\gamma = (\sqrt{\rho} - 1)^{-1}(\sqrt{\rho} + 1)$.

The theorem follows immediately from a combination of the following Lemmas 5a and 5b. The first lemma relates (7) to the following resolvent condition:

$$(\zeta I - B) \text{ is invertible, and } \|(\zeta I - B)^{-1}\| \leqslant K \cdot \frac{|\zeta|^{\delta \cdot s}}{|\zeta| - 1} \quad \text{for all complex } \zeta \text{ with } |\zeta| > 1. \tag{7'}$$

Lemma 5a. *Let $\delta \geqslant 0$, $\rho > 1$. Then* (7) *implies* (7′) *with* $K = (\sqrt{\rho} + 1)(\sqrt{\rho} - 1)^{-1} \cdot L$.

Proof. For $|\zeta| \geqslant \rho$ we have the Dunford–Taylor representation

$$(\zeta I - B)^{-1} = (2\pi \mathrm{i})^{-1} \int (\zeta - z)^{-1}(zI - B)^{-1}\,\mathrm{d}z,$$

where the integration is along any positively oriented circle $|z| = \sigma$, with $1 < \sigma < \rho$. From this representation we obtain, in view of (7),

$$\|(\zeta I - B)^{-1}\| \leqslant (2\pi)^{-1} \int |\zeta - z|^{-1} \cdot L \cdot \frac{|z|^{\delta s}}{|z| - 1}|\mathrm{d}z| \leqslant L|\zeta|^{\delta s} \frac{\sigma}{(\sigma - 1)(|\zeta| - \sigma)}.$$

Choosing $\sigma = \sqrt{\rho}$ we obtain, for $|\zeta| \geqslant \rho$,

$$\|(\zeta I - B)^{-1}\| \leqslant \frac{\sigma}{\sigma - 1}\left(1 + \frac{\sigma - 1}{|\zeta| - \sigma}\right) \cdot L \frac{|\zeta|^{\delta s}}{|\zeta| - 1} \leqslant \frac{\sigma + 1}{\sigma - 1} \cdot L \cdot \frac{|\zeta|^{\delta s}}{|\zeta| - 1},$$

which proves (7′) with $K = (\sqrt{\rho} + 1)(\sqrt{\rho} - 1)^{-1} \cdot L$. □

Lemma 5b. *Let $\delta \geqslant 0$. Then the resolvent condition* (7′) *implies that*

(a) $\|B^n\| \leqslant 4K \cdot (n + \delta s) \quad (n \geqslant 1,\ s \geqslant 1)$,

(b) $\|B^n\| \leqslant \mathrm{e}K \cdot (1 + \delta)s \quad (n \geqslant 1,\ s \geqslant 1)$.

Proof. (1) Let the $s \times s$ matrix B satisfy (7′), and let $n \geqslant 1$. According to a well known corollary to the Hahn–Banach theorem (see, e.g., [5, Chapter 5]), there exists a linear mapping F, from the vector space of all $s \times s$ matrices to $\mathbb{C}$, such that $|F(A)| \leqslant \|A\|$ (for all $s \times s$ matrices A) and $F(B^n) = \|B^n\|$. For any $\varepsilon > 0$ we have

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

where Γ denotes the positively oriented circle $|\zeta| = 1 + \varepsilon$. Consequently,

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

where $R(\zeta)$ is the rational function defined by $R(\zeta) = F((\zeta I - B)^{-1})$.

(2) Using the above integral representation for $\|B^n\|$, in combination with the inequality

$$\big|R(\zeta)\big| \leqslant K \cdot \frac{|\zeta|^{\delta \cdot s}}{|\zeta| - 1} \quad (\text{for } |\zeta| > 1),$$

we arrive at

$$\big\|B^n\big\| \leqslant K \cdot \varepsilon^{-1} \cdot (1 + \varepsilon)^{n+\delta s+1}.$$

Choosing $\varepsilon = (n + \delta s)^{-1}$, there follows,

$$\big\|B^n\big\| \leqslant \big[1 + (n + \delta s)^{-1}\big]^{n+\delta s} K \cdot (n + \delta s + 1).$$

From this inequality we obtain the upper bound

$$\big\|B^n\big\| \leqslant 4K \cdot (n + \delta s),$$

which proves part (a) of the lemma.

We also obtain the upper bound

$$\big\|B^n\big\| \leqslant \mathrm{e}K \cdot (n + \delta s + 1).$$

(3) Following [8], we perform a partial integration in the above integral representation for $\|B^n\|$ so as to get

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

There follows

$$\big\|B^n\big\| \leqslant \big[2\pi(n + 1)\big]^{-1}(1 + \varepsilon)^{n+1} \cdot \Lambda,$$

where

$$\Lambda = \int_{\Gamma} \big|R'(\zeta)\big||\mathrm{d}\zeta|.$$

It is easily verified that $R(\zeta)$ is a rational function, with no poles on the circle Γ, and with degrees of its numerator and denominator not exceeding s (see, e.g., [19, p. 209]). For rational functions with these properties, the integral Λ can be estimated, according to [15,21], in the following way:

$$\Lambda \leqslant 2\pi s \cdot \max_{\Gamma} \big|R(\zeta)\big|.$$

Along the curve Γ our function $R(\zeta)$ satisfies the inequality $|R(\zeta)| \leqslant K \cdot \varepsilon^{-1}(1 + \varepsilon)^{\delta s}$. Combining this inequality with the above estimate of Λ and our last upper bound for $\|B^n\|$, we obtain

$$\big\|B^n\big\| \leqslant K \cdot s(n + 1)^{-1} \cdot \varepsilon^{-1} \cdot (1 + \varepsilon)^{n+\delta \cdot s+1}.$$

Choosing $\varepsilon = (n + \delta s)^{-1}$, there follows

$$\big\|B^n\big\| \leqslant \mathrm{e}K \cdot \frac{(n + \delta s + 1)s}{n + 1}.$$

(4) Combining the last inequality with the upper bound for $\|B^n\|$ obtained at the end of part (2) of the proof, it follows that

$$\|B^n\| \leqslant \mathrm{e}K \cdot (n + \delta s + 1) \cdot \min\left\{1, \frac{s}{n+1}\right\}.$$

Hence

$$\|B^n\| \leqslant \mathrm{e}K(1+\delta) \cdot s.$$

This completes the proof of the lemma. □

Remark 6. For $\delta = 0$, the upper bounds for $\|B^n\|$ in Theorem 4 tend to $4Ln$ and $\mathrm{e}Ls$, respectively, when $\rho \to \infty$. This implies that Theorem 1 can be viewed as a corollary to Theorem 4.

7. Numerical stability in the solution of delay differential equations

7.1. A linear, scalar test equation

We shall shortly explore the applicability of the foregoing in the numerical solution of the initial value problem

$$Z'(t) = f\big(Z(t), Z(t-\tau)\big) \quad (t \geqslant 0), \qquad Z(t) = g(t) \quad (t \leqslant 0).$$

Here f, g are given functions, $\tau > 0$ is a fixed delay, and $Z(t)$ is unknown (for $t > 0$).

We shall deal with the following, well known, version of the *trapezoidal rule*:

$$z_n = z_{n-1} + \frac{h}{2}\big[f(z_n, z_{n-s+1}) + f(z_{n-1}, z_{n-s})\big] \quad (n \geqslant 1). \tag{8}$$

Here s is an integer with $s \geqslant 2$, and $h = \tau/(s-1)$ denotes the *stepsize*. Further, z_n denote approximations to $Z(t)$ at the *gridpoints* $t = t_n = nh$. Defining $z_i = g(t_i)$ $(i = 0, -1, \ldots, 1-s)$, one may compute, for $n = 1, 2, 3, \ldots$, approximations z_n by successive applications of (8).

Following Barwell [1], many authors studied the stability of numerical methods, for the above initial value problem, by analyzing the methods when applied to the following, linear test problem:

$$Z'(t) = \lambda Z(t) + \mu Z(t-\tau) \quad (t \geqslant 0), \qquad Z(t) = g(t) \quad (t \leqslant 0).$$

Here λ, μ denote fixed, complex coefficients, and $g(t), Z(t) \in \mathbb{C}$.

Applying method (8) to the test equation one arrives at the recurrence relation

$$z_n = a \cdot z_{n-1} + b \cdot z_{n-s+1} + b \cdot z_{n-s} \quad (n \geqslant 1),$$

where

$$a = \frac{2+x}{2-x}, \quad b = \frac{y}{2-x} \quad \text{and} \quad x = h\lambda, \quad y = h\mu.$$

Defining $u_n \in \mathbb{C}^s$ and the $s \times s$ companion matrix B by

$$u_n = \begin{pmatrix} z_n \\ z_{n-1} \\ \vdots \\ z_{n-s+2} \\ z_{n-s+1} \end{pmatrix}, \qquad B = \begin{pmatrix} a & 0 & 0 & \dots & 0 & b & b \\ 1 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & \dots & 1 & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 1 & 0 \end{pmatrix},$$

the above recurrence relation is equivalent to

$$u_n = Bu_{n-1} \quad (n \geqslant 1).$$

Clearly, B depends (only) on x, y and s so that we may write $B = B_s(x, y)$.

Following standard practice, in dealing with the above test problem, we consider the *stability region*

$$S = \{(x, y)\text{: for each } s \geqslant 2, \text{the matrix } B_s(x, y) \text{ satisfies (2)}\}.$$

(Various authors, e.g., Zennaro [22], would refer to this set as the P-stability region of method (8).) The above region S was studied by various authors among which [2,9,20,22]. It is known that the interior and closure of S equal

$$\text{int(S)} = \{(x, y)\text{: } \operatorname{Re}[x] < -|y|\},$$
$$\text{cl}(S) = \{(x, y)\text{: } \operatorname{Re}[x] \leqslant -|y|\},$$

respectively.

But, we noticed already in Section 3 that condition (2), occurring in our definition of S, can be a misleading guide to stability—as long as "stability" is interpreted in the sense of Sections 1 and 2. With regard to error propagation the crucial question is of whether $\|B^n\|$ can be bounded suitably, and *not* of whether, for each individual fixed s, one has the asymptotic property $\lim_{n\to\infty} \|B^n\| = 0$.

In the following subsections we focus on estimating $\|B^n\|_\infty$ for $(x, y) \in \text{cl}(S)$ and $n \geqslant 1$, $s \geqslant 2$.

7.2. *A positive stability result, derived by using resolvents*

In the following we deal with a fixed, given element $(x, y) \in \text{cl}(S)$, so that

$$\operatorname{Re}[x] \leqslant -|y|.$$

We shall study the corresponding matrix $B = B_s(x, y)$ for $s \geqslant 2$.

By a short, and rather rough, analysis of the resolvent $(\zeta I - B)^{-1}$ it can be shown that condition (7) is satisfied here, with $\delta = 1$, $\rho = 2$, $\|\cdot\| = \|\cdot\|_\infty$ and L only depending on the given (x, y). By virtue of Theorem 4, we thus have the stability estimate

$$\|B^n\|_\infty \leqslant c \cdot s \quad (n \geqslant 1,\ s \geqslant 2),$$

with a constant c only depending on (x, y).

By putting a bit more effort in the analysis of the resolvent, it can be shown that (7) holds in fact with $\delta = 0$. We have

Lemma 7. *Corresponding to any given $x, y \in \mathbb{C}$ with $\operatorname{Re}[x] \leqslant -|y|$ there is a constant L such that, for all $s \geqslant 2$, the matrix $B = B_s(x, y)$ satisfies condition (7) with $\delta = 0$, $\rho = 2$, $\|\cdot\| = \|\cdot\|_\infty$.*

Proof. (1) Let $x, y \in \mathbb{C}$ be given with $\mathrm{Re}\,[x] \leqslant -|y|$. The corresponding polynomial $P(\zeta) = \det(\zeta I - B)$ equals

$$P(\zeta) = (\zeta - a)\zeta^{s-1} - (\zeta + 1)b.$$

We shall analyze the ratio between $(\zeta + 1)b$ and $(\zeta - a)$ for $|\zeta| > 1$.

Clearly, any ζ with $|\zeta| > 1$ can be written in the form

$$\zeta = \frac{2+z}{2-z}, \quad \text{with } \mathrm{Re}\,[z] > 0,\ z \neq 2.$$

An easy calculation shows that

$$(\zeta - a) = \frac{4(z-x)}{(2-z)(2-x)} \quad \text{and} \quad (\zeta + 1)b = \frac{4y}{(2-z)(2-x)}.$$

We thus have

$$(\zeta + 1)b/(\zeta - a) = \frac{y}{z-x},$$

and since

$$\left|\frac{y}{z-x}\right| \leqslant \frac{-\mathrm{Re}\,[x]}{\mathrm{Re}\,[z] - \mathrm{Re}\,[x]} < 1$$

there follows

$$\left|(\zeta + 1)b/(\zeta - a)\right| < 1 \quad (\text{for } |\zeta| > 1). \tag{9a}$$

This inequality implies that

$$\left|P(\zeta)\right| \geqslant |\zeta - a| \cdot |\zeta|^{s-1} - \left|(\zeta + 1)b\right| \geqslant \left(|\zeta| - |a|\right)\left(|\zeta|^{s-1} - 1\right) \quad (\text{for } |\zeta| > 1). \tag{9b}$$

(2) Since $\mathrm{Re}\,[x] \leqslant 0$, we have $|a| \leqslant 1$. From (9b) we thus see that $(\zeta I - B)$ is invertible whenever $|\zeta| > 1$.

Let ζ be given with

$$|\zeta| = 1 + \varepsilon > 1,$$

and let $u = (\xi_1, \xi_2, \ldots, \xi_s)^{\mathrm{T}}$, $v = (\eta_1, \eta_2, \ldots, \eta_s)^{\mathrm{T}}$ be vectors in $\mathbb{C}^s$ such that

$$(\zeta I - B)u = v.$$

By a straightforward calculation it can be verified that

$$\begin{aligned}
P(\zeta)\cdot\xi_1 &= \zeta^{s-1}\eta_1 + b(\zeta+1)\left(\eta_2 + \zeta\eta_3 + \cdots + \zeta^{s-2}\eta_s\right) - b\zeta^{s-1}\eta_s,\\
P(\zeta)\cdot\xi_j &= \zeta^{s-j}\eta_1 + \zeta^{s-j}(\zeta - a)\left(\eta_2 + \zeta\eta_3 + \cdots + \zeta^{j-2}\eta_j\right)\\
&\quad + b(\zeta+1)\left(\eta_{j+1} + \zeta\eta_{j+2} + \cdots + \zeta^{s-j-1}\eta_s\right) - b\zeta^{s-j}\eta_s \quad \text{for } 2 \leqslant j \leqslant s-2,\\
P(\zeta)\cdot\xi_{s-1} &= \zeta\eta_1 + \zeta(\zeta - a)\left(\eta_2 + \zeta\eta_3 + \cdots + \zeta^{s-3}\eta_{s-1}\right) + b(\zeta+1)\eta_s - b\zeta\eta_s,\\
P(\zeta)\cdot\xi_s &= \eta_1 + (\zeta - a)\left(\eta_2 + \zeta\eta_3 + \cdots + \zeta^{s-2}\eta_s\right) - b\eta_s.
\end{aligned}$$

(3) From the above relations we obtain, for $|v|_\infty \leqslant 1$ and $1 \leqslant j \leqslant s$,

$$\left|P(\zeta)\xi_j\right| \leqslant |\zeta|^{s-1} + |\zeta - a|\frac{|\zeta|^{s-1} - 1}{\varepsilon} + |b|\left\{|\zeta|^{s-1} + |\zeta + 1|\frac{|\zeta|^{s-1} - 1}{\varepsilon}\right\}.$$

In view of (9b),

$$\left\|(\zeta I - B)^{-1}\right\|_\infty \leqslant \frac{|\zeta - a| + |(\zeta+1)b|}{|\zeta - a||\zeta|^{s-1} - |(\zeta+1)b|} \cdot \frac{|\zeta|^{s-1} - 1}{\varepsilon} + \frac{2|\zeta|^{s-1}}{|P(\zeta)|}.$$

Using (9a), there follows

$$\left\|(\zeta I - B)^{-1}\right\|_\infty \leqslant 2 \cdot \left(\frac{1}{\varepsilon} + \frac{|\zeta|^{s-1}}{|P(\zeta)|}\right).$$

Assume $b \neq 0$. Then $\operatorname{Re}[x] \leqslant -|y| = -|(2-x)b| < 0$, and therefore $|a| < 1$. From (9b) we obtain

$$\frac{|\zeta|^{s-1}}{|P(\zeta)|} \leqslant \frac{|\zeta|^{s-1}}{(1-|a|)(|\zeta|^{s-1} - 1)} \leqslant \frac{|\zeta|}{(1-|a|)\varepsilon} \leqslant \frac{2}{(1-|a|)\varepsilon},$$

provided $0 < \varepsilon \leqslant 1$. We thus arrive at the resolvent inequality

$$\left\|(\zeta I - B)^{-1}\right\|_\infty \leqslant L/(|\zeta| - 1) \quad (\text{for } 1 < |\zeta| \leqslant 2),$$

with $L = 2 + 4(1-|a|)^{-1}$.

For $b = 0$, we have $|\zeta|^{s-1}/|P(\zeta|) = |\zeta - a|^{-1} \leqslant \varepsilon^{-1}$, so that the above resolvent inequality follows with $L = 4$. □

Combining Lemma 7 and Theorem 4 we arrive at

Conclusion 8. *Corresponding to any pair $(x, y) \in \operatorname{cl}(S)$ there is a constant c such that the matrix $B = B_s(x, y)$ satisfies*

$$\left\|B^n\right\|_\infty \leqslant c \cdot \min(s, n) \quad (n \geqslant 1,\ s \geqslant 2).$$

Here c only depends on x, y and not on n, s.

7.3. A negative stability result, derived by using resolvents

The question poses itself of whether a variant to Lemma 7 exists, in which the constant L is independent of $x, y \in \mathbb{C}$ with $\operatorname{Re}[x] \leqslant -|y|$. The following lemma shows that such a variant is not possible.

Lemma 9. *Let $s \geqslant 2$ be a given, fixed integer. Then there exist no $\delta \geqslant 0$, $\rho > 1$, L such that $B = B_s(x, y)$ satisfies the resolvent condition (7) with $\|\cdot\| = \|\cdot\|_\infty$, uniformly for all $x, y \in \mathbb{C}$ with $\operatorname{Re}[x] < -|y|$.*

Proof. (1) Assume (7) holds, uniformly for $x, y \in \mathbb{C}$ with $\operatorname{Re}[x] < -|y|$.

We consider arbitrary x, y, ζ with $\operatorname{Re}[x] < -|y|$, $1 < |\zeta| < \rho$, and we use the notations of the proof of Lemma 7 (parts 1 and 2).

Let

$$B = B_s(x, y), \qquad (\zeta I - B)u = v,$$

with $\eta_1 = 1$, $\eta_i = 0$ $(i \neq 1)$. Since $\xi_1 = \zeta^{s-1}/P(\zeta)$, we have

$$|\zeta|^{s-1}/|P(\zeta)| \leqslant \left\|(\zeta I - B)^{-1}\right\|_\infty \leqslant L \cdot |\zeta|^{\delta \cdot s}/(|\zeta| - 1). \tag{10}$$

(2) We now focus on real x, y satisfying

$$x < -1, \qquad y = (-1)^s(1+x),$$

with corresponding polynomial

$$P(\zeta) = P(\zeta, x) = \left(\zeta - \frac{2+x}{2-x}\right)\zeta^{s-1} - (\zeta+1)\frac{1+x}{2-x}(-1)^s.$$

For $x \to -\infty$, this polynomial tends to

$$P^*(\zeta) = (\zeta+1)\left(\zeta^{s-1} + (-1)^s\right).$$

Since (10) holds, with $P(\zeta) = P(\zeta, x)$ and $x < -1$, we see that

$$|\zeta|^{s-1}/|P^*(\zeta)| \leqslant L \cdot |\zeta|^{\delta \cdot s}/(|\zeta| - 1).$$

From this inequality we arrive at a contradiction, by considering real $\zeta < -1$ with $\zeta \to -1$. □

In Section 4 we noted that (1) implies (3), with $L = M$. In view of Lemma 9, there exists no L such that $B = B_s(x, y)$ satisfies (3), uniformly for $(x, y) \in \text{int}(S)$. Therefore, we have

Conclusion 10. *Let $s \geqslant 2$ be given. Then there exists no constant c (only depending on s) such that $B = B_s(x, y)$ satisfies*

$$\|B^n\|_\infty \leqslant c \quad (n \geqslant 1),$$

uniformly for $(x, y) \in \text{int}(S)$.

From the above it is clear that Conclusion 8 cannot be sharpened in that the constant c in that conclusion would become independent of $(x, y) \in \text{cl}(S)$.

References

[1] V.K. Barwell, Special stability problems for functional differential equations, *BIT* 15 (1975) 130–135.

[2] M. Calvo and T. Grande, On the asymptotic stability of θ-methods for delay differential equations, *Numer. Math.* 54 (1988) 257–269.

[3] G. Dahlquist, H. Mingyou and R. LeVeque, On the uniform power-boundedness of a family of matrices and the applications to one-leg and linear multistep methods, *Numer. Math.* 42 (1983) 1–13.

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

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

[6] J.F.B.M. Kraaijevanger, Two counterexamples related to the Kreiss matrix theorem, *BIT* 34 (1994) 113–119.

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

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

[9] M.Z. Liu and M.N. Spijker, The stability of the θ-methods in the numerical solution of delay differential equations, *IMA J. Numer. Anal.* 10 (1990) 31–48.

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

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

[12] C. Palencia, Stability of rational multistep approximations of holomorphic semigroups, *Math. Comp.* 64 (1995) 591–599.
[13] S.V. Parter, Stability, convergence, and pseudo-stability of finite-difference equations for an overdetermined problem, *Numer. Math.* 4 (1962) 277–292.
[14] R.D. Richtmyer and K.W. Morton, *Difference Methods for Initial Value Problems* (Wiley, New York, 2nd ed., 1967).
[15] M.N. Spijker, On a conjecture by LeVeque and Trefethen related to the Kreiss matrix theorem, *BIT* 31 (1991) 551–555.
[16] M.N. Spijker and F.A.J. Straetemans, Stability estimates for families of matrices of nonuniformly bounded order, *Linear Algebra Appl.* 239 (1996) 77–102.
[17] E. Tadmor, The equivalence of L_2-stability, the resolvent condition, and strict H-stability, *Linear Algebra Appl.* 41 (1981) 151–159.
[18] E. Tadmor, The resolvent condition and uniform powerboundedness, *Linear Algebra Appl.* 80 (1986) 250–252.
[19] J.L.M. van Dorsselaer, J.F.B.M. Kraaijevanger and M.N. Spijker, Linear stability analysis in the numerical solution of initial value problems, *Acta Numerica* (1993) 199–237.
[20] D.S. Watanabe and M.G. Roth, The stability of difference formulas for delay differential equations, *SIAM J. Numer. Anal.* 22 (1985) 132–145.
[21] E. Wegert and L.N. Trefethen, From the Buffon needle problem to the Kreiss matrix theorem, *Amer. Math. Monthly* 101 (1994) 132–139.
[22] M. Zennaro, P-stability properties of Runge–Kutta methods for delay differential equations, *Numer. Math.* 49 (1986) 305–318.