NUMERICAL STABILITY

Stability Estimates and Resolvent Conditions
in the Numerical Solution of Initial Value Problems

M.N. Spijker

Lecture Notes

December 1998

NUMERICAL STABILITY

M.N. Spijker

Address of the author
Department of Mathematics and Computer Science
University of Leiden
P.O. Box 9512
2300 RA Leiden
The Netherlands

These lecture notes are in a preliminary form and, in various respects, incomplete. The author will be thankful for criticism and suggestions for improvement.

I am most thankful to my wife Marijke for making the TeX-file of these Lecture Notes!

CONTENTS

1 Partial differential equations and numerical methods

1.1 Diffusion, convection and reaction

We shall use the following notations for the so-called *gradient* and *divergence* operators,

$$\text{grad } f(x,y,z) = \begin{pmatrix} \frac{\partial}{\partial x} f(x,y,z) \\ \frac{\partial}{\partial y} f(x,y,z) \\ \frac{\partial}{\partial z} f(x,y,z) \end{pmatrix},$$

$$\text{div } \begin{pmatrix} f_1(x,y,z) \\ f_2(x,y,z) \\ f_3(x,y,z) \end{pmatrix} = \frac{\partial}{\partial x} f_1(x,y,z) + \frac{\partial}{\partial y} f_2(x,y,z) + \frac{\partial}{\partial z} f_3(x,y,z).$$

We consider a chemical species in a three-dimensional region Ω, and denote its concentration (in mass per unit volume), at the time t and location $p = (x,y,z)$ in Ω, by $u(p,t) = u(x,y,z,t)$.

We denote the rate of generation of the chemical species per unit volume, due to some chemical reaction, by $r(p,t,u(p,t))$. This rate thus depends on the position p, time t and actual concentration $u(p,t)$. The rate can have a positive or negative value, and be zero for a nonreacting species. If there would be no flow of the species through Ω, e.g. if Ω consists of a substance impermeable to the species, the concentration would satisfy the differential equation

$$\frac{\partial}{\partial t} u(p,t) = r(p,t,u(p,t)).$$

But, in many cases of practical interest (fluids or gasses) we have to consider a *flux* through Ω influencing $u(p,t)$. We represent this flux by a vector $F(p,t) \in \mathbb{R}^3$ with components $F_1(p,t)$, $F_2(p,t)$, $F_3(p,t)$. This vector has a direction coinciding with the direction of the flux, and its Euclidean length equals the amount of species flowing per time unit through a unit surface that is perpendicular to the flux (at the position p and time t). In order to determine the effect of the flux F on $\frac{\partial}{\partial t} u(p,t)$ we consider a small cube $\Delta\Omega$ in Ω with center at $p = (x,y,z)$. Let $\Delta\Omega$ be bounded by surfaces perpendicular to the $x-, y-$ and $z-$ axes with distances from p equal to $\frac{\Delta x}{2}$, $\frac{\Delta y}{2}$, $\frac{\Delta z}{2}$, respectively. By simple geometric arguments it is possible to determine the effect of the flux on the increase of the amount of species, per time unit, in the cube. This effect is approximately equal to

$$\Big[F_1(x - \frac{\Delta x}{2}, y, z, t) - F_1(x + \frac{\Delta x}{2}, y, z, t)\Big]\Delta y \Delta z + \Big[F_2(x, y - \frac{\Delta y}{2}, z, t) - F_2(x, y + \frac{\Delta y}{2}, z, t)\Big]\Delta x \Delta z$$

$$+\Big[F_3(x, y, z - \frac{\Delta z}{2}, t) - F_3(x, y, z + \frac{\Delta z}{2}, t)\Big]\Delta x \Delta y,$$

that is to

$$-\text{div } F(p,t) \cdot \Delta x \Delta y \Delta z.$$

Therefore the effect of the flux on the increase of $u(p,t)$, per time unit, equals $-\text{div } F(p,t)$. The total effect of reaction and flux is expressed in

$$\frac{\partial}{\partial t} u(p,t) = -\text{div } F(p,t) + r(p,t,u(p,t)),$$

which we shall call the *conservation of mass equation*.

In general the flux vector $F(p,t)$ can be decomposed,

$$F(p,t) = u(p,t) \cdot V(p,t) + W(p,t).$$

Here the first term in the right-hand member is the *convection flux* of the species. It is caused by a general flow field $V(p,t) \in \mathbb{R}^3$ which is independent of the concentration of the chemical species. The amount of the species transported with the motion $V(p,t)$ is represented by the vector $u(p,t) \cdot V(p,t)$. The second term is called the *diffusion flux*. It is due to the molecular, thermal agitation, and is also present in fluids or gasses at rest. According to *Ficks' law*

$$W(p,t) = -k(p,t) \cdot \text{grad } u(p,t),$$

where $k(p,t)$ is called the *diffusion coefficient* at p and at time t.

Substituting the above expressions in the conservation of mass equation we obtain the partial differential equation

$$\frac{\partial}{\partial t}u(p,t) = \text{div } [k(p,t)\,\text{grad } u(p,t)] - \text{div } [u(p,t)\,V(p,t)] + r(p,t,u(p,t)). \tag{1.1.1}$$

It is called a *transport equation*, or also a *diffusion-convection-reaction equation*, for $u(p,t)$.

In practical situations one may be interested in computing the solution $u(p,t)$ to (1.1.1) for $p \in \Omega,\ t > 0$, under an *initial condition* of the form

$$u(p,0) = f(p) \quad \text{for } p \in \Omega, \tag{1.1.2}$$

with given function $f(p)$. In general, the relations (1.1.1), (1.1.2) are supplemented by conditions for $t > 0$ at the points p belonging to the *boundary* $\partial\Omega$ of Ω. These *boundary conditions* are usually of one of the following three types,

$$u(p,t) = g(p,t), \tag{1.1.3a}$$

$$\frac{\partial}{\partial n}u(p,t) = g(p,t), \tag{1.1.3b}$$

$$\frac{\partial}{\partial n}u(p,t) = c(p,t) \cdot [g(p,t) - u(p,t)]. \tag{1.1.3c}$$

Here $g(p,t)$ and $c(p,t)$ are given functions. Further, n denotes the outward normal to $\partial\Omega$ at p, with length 1, so that $\frac{\partial}{\partial n}u(p,t)$ is equal to the inner product of n and grad $u(p,t)$.

Condition (1.1.3a) is called a *Dirichlet boundary condition*. It occurs if there is free contact between the interior and exterior of Ω at $p \in \partial\Omega$, and the concentration outside Ω is known to be equal to $u_0(p,t)$. If the medium in which the chemical species is solved is the same —outside and inside of $\partial\Omega$ — we have $g(p,t) = u_0(p,t)$.

Condition (1.1.3b) is called a *Von Neumann boundary condition*. E.g. if the boundary $\partial\Omega$ is impermeable to the chemical species, and the normal component $V_n(p,t)$ of $V(p,t)$ at $p \in \partial\Omega$ vanishes, we have (1.1.3b) with $g(p,t) = 0$.

Condition (1.1.3c) is a *mixed condition*. E.g. if the boundary is semipermeable with regard to the chemical species, we may have $c(p,t) > 0$ and $g(p,t)$ as in (1.1.3a).

In cases where $\frac{\partial}{\partial z}u(p,t) = 0$, or $\frac{\partial}{\partial z}u(p,t) = \frac{\partial}{\partial y}u(p,t) = 0$, throughout Ω and for all $t \geq 0$, it may be convenient to consider the concentration as a function only of the variable (x,y,t) or (x,t), respectively. In this way one can arrive e.g. at the following four lower-dimensional versions of the above.

$$\frac{\partial}{\partial t}u(x,t) = \frac{\partial^2}{\partial x^2}u(x,t), \quad u(0,t) = g_0(t), \quad u(1,t) = g_1(t), \tag{1.1.4}$$
$$u(x,0) = f(x), \quad \text{where } 0 \le x \le 1,\ t \ge 0.$$

This is the pure diffusion equation with Dirichlet boundary conditions in 1 dimension.

$$\frac{\partial}{\partial t}u(x,t) = \frac{\partial}{\partial x}[b(x)u(x,t)] + c(x,t), \quad u(0,t) = 0, \tag{1.1.5}$$
$$u(x,0) = f(x), \quad \text{where } 0 \le x \le 1,\ t \ge 0.$$

This is a one-dimensional convection-reaction equation, where the rate $c(x,t)$ of the chemical reaction is independent of the concentration $u(x,t)$.

$$\frac{\partial}{\partial t}u(x,t) = \frac{\partial}{\partial x}[a(x)\frac{\partial}{\partial x}u(x,t)] + \frac{\partial}{\partial x}[b(x)u(x,t)] + c(x)u(x,t) + d(x), \tag{1.1.6}$$
$$u(0,t) = g(t), \quad \frac{\partial}{\partial x}u(1,t) = 0, \quad u(x,0) = f(x), \quad \text{where } 0 \le x \le 1,\ t \ge 0.$$

This is a one-dimensional diffusion-convection-reaction equation. In case $b(0) < 0$, the convection at $x = 0$ is from left to right, and the boundary point $x = 0$ lies *upwind*, sometimes called *upstream*, of the calculation domain. Correspondingly, the condition $u(0,t) = g(t)$ is called an *inflow* boundary condition. In case $b(0) > 0$, the same condition would be called an *outflow* boundary condition.

In the following example we consider the square $\Omega = \{(x,y) :\ 0 \le x \le 1,\ 0 \le y \le 1\}$, with boundary $\partial\Omega = \{(x,y) :\ (x,y) \in \Omega, \text{ and } x(x-1)\,y(y-1) = 0\}$.

$$\frac{\partial}{\partial t}u(x,y,t) = \frac{\partial^2}{\partial x^2}u(x,y,t) + \frac{\partial^2}{\partial y^2}u(x,y,t) \qquad \text{for } (x,y) \in \Omega, \tag{1.1.7}$$
$$u(x,y,t) = g(x,y,t) \qquad \text{for } (x,y) \in \partial\Omega,$$
$$u(x,y,0) = f(x,y) \qquad \text{for } (x,y) \in \Omega,$$
$$\text{where } t \ge 0.$$

This is a pure diffusion equation with Dirichlet boundary conditions in 2 dimensions.

1.2 Semi-discretization by finite difference methods

1.2.1 Basic finite difference approximations

The *finite difference method* can be used for obtaining numerical approximations to the solutions of the partial differential equations reviewed above under given initial-boundary conditions. The method essentially consists in replacing the partial derivatives in the original problem by finite difference quotients so as to obtain a related problem which is easier to solve than the original one. In this section we focus on such replacements for the derivatives with respect to the space variables (x, y and z) only. This process, where derivatives with respect to t are not affected, is called *semi-discretization*.

We start by listing various difference approximations to the derivatives of a function $v(x)$ defined for $0 \le x \le 1$. We assume $v(x)$ to have continuous derivatives of the orders occurring in

the following expressions, and for integer values $p \geq 1$ we denote by $v^{(p)}(x)$ the p-th derivative of $v(x)$. Further, we use the notation

$$\left|v^{(p)}\right|_\infty = \max_{0\leq x\leq 1}\left|v^{(p)}(x)\right|.$$

All arguments of v, occurring below, belong to the interval $[0,1]$, and $\delta = \Delta x$ denotes a positive increment of the variable x. The following six relations are valid:

(1.2.1a) $$v'(x) = \delta^{-1}[v(x+\delta) - v(x)] + R,$$
with $|R| \leq \dfrac{\delta}{2}|v^{(2)}|_\infty$ *(forward difference approximation).*

(1.2.1b) $$v'(x) = \delta^{-1}[v(x) - v(x-\delta)] + R,$$
with $|R| \leq \dfrac{\delta}{2}|v^{(2)}|_\infty$ *(backward difference approximation).*

(1.2.1c) $$v'(x) = (2\delta)^{-1}[v(x+\delta) - v(x-\delta)] + R,$$
with $|R| \leq \dfrac{\delta^2}{6}|v^{(3)}|_\infty$ *(centered approximation).*

(1.2.2a) $$v''(x) = \delta^{-2}[v(x-\delta) - 2v(x) + v(x+\delta)] + R, \quad \text{with } |R| \leq \frac{\delta^2}{12}|v^{(4)}|_\infty.$$

(1.2.2b) $$v''(x) = \frac{1}{12}\delta^{-2}[-v(x-2\delta) + 16v(x-\delta) - 30v(x) + 16v(x+\delta) - v(x+2\delta)] +$$
$$+ R, \quad \text{with } |R| \leq \frac{\delta^4}{90}|v^{(6)}|_\infty.$$

(1.2.2c) $$\frac{1}{12}[v''(x-\delta) + 10v''(x) + v''(x+\delta)] = \delta^{-2}[v(x-\delta) - 2v(x) + v(x+\delta)] + R,$$
with $|R| \leq \dfrac{\delta^4}{240}|v^{(6)}|_\infty$ *(Numerov's approximation).*

The above upper bounds for $|R|$ can be obtained by expanding R in powers of δ, with the use of Taylor's formula applied to the function v. In order to arrive at the bounds in (1.2.2b), (1.2.2c) one has to express the remainder term associated with the Taylor polynomial as an integral involving the 6-th derivative of v. Further, in proving (1.2.2c), Taylor's formula should as well be applied to the function v''.

The left-hand members in the formulas (1.2.1), (1.2.2) can be approximated by the corresponding right-hand members in which the residual R is suppressed.

1.2.2 Finite difference approximations using non-equidistant arguments

It is also possible to establish finite difference approximations with *nonuniformly distributed arguments*. For instance, let $\delta_1 > 0,\ \delta_2 > 0$ and consider the expression

$$\beta \cdot v(x-\delta_1) + \alpha \cdot v(x) + \gamma \cdot v(x+\delta_2),$$

where $\alpha,\ \beta,\ \gamma$ are coefficients which are still to be determined. By Taylor's formula this expression equals

$$(\alpha+\beta+\gamma)v(x) + (\gamma\delta_2 - \beta\delta_1)v'(x) + \left[\gamma(\delta_2)^2 + \beta(\delta_1)^2\right]\frac{v''(x)}{2} - R,$$

where

$$|R| \le \Big[|\gamma|(\delta_2)^3 + |\beta|(\delta_1)^3\Big]\frac{1}{6}|v'''|_\infty.$$

Suppose one wants to approximate $v''(x)$. Then the requirements on $\alpha,\ \beta,\ \gamma$ should be that

$$\alpha+\beta+\gamma = 0, \qquad \gamma\delta_2 - \beta\delta_1 = 0, \qquad \gamma(\delta_2)^2 + \beta(\delta_1)^2 = 2,$$

which yields

$$\alpha = \frac{-2}{\delta_1\delta_2}, \qquad \beta = \frac{2}{\delta_1(\delta_1+\delta_2)}, \qquad \gamma = \frac{2}{\delta_2(\delta_1+\delta_2)}.$$

From the above it can be seen that

$$v''(x) = \frac{2}{\delta_1(\delta_1+\delta_2)}v(x-\delta_1) - \frac{2}{\delta_1\delta_2}v(x) + \frac{2}{\delta_2(\delta_1+\delta_2)}v(x+\delta_2) + R, \tag{1.2.3}$$

$$\text{with} \quad |R| \le \frac{1}{3}\cdot\max(\delta_1,\delta_2)\cdot|v'''|_\infty.$$

1.2.3 Semi-discretization in one space variable

In the process of semi-discretization by finite differences we choose *gridpoints* p in the calculation domain, and we approximate the values $u(p,t)$ by quantities $U(p,t)$ for all gridpoints p.

For instance, consider the approximation of the solution to (1.1.4) on a so-called *nonuniform grid* using formula (1.2.3). We choose $\delta_1 > 0, \delta_2 > 0, \ldots, \delta_{s+1} > 0$ with $\delta_1+\delta_2+\cdots+\delta_{s+1} = 1$, and define gridpoints $p = x_j$ by $x_0 = 0,\ x_{j+1} = x_j + \delta_{j+1}$ (for $j = 0,1,\ldots,s$) . In view of (1.2.3) we define approximations $U_j(t) = U(x_j,t) \approx u(x_j,t)$ by requiring, for $1 \le j \le s$,

$$\frac{d}{dt}U_j(t) = \frac{2}{\delta_j(\delta_j+\delta_{j+1})}U_{j-1}(t) - \frac{2}{\delta_j\delta_{j+1}}U_j(t) + \frac{2}{\delta_{j+1}(\delta_j+\delta_{j+1})}U_{j+1}(t), \qquad U_j(0) = f(x_j),$$

and by putting $U_0(t) = g_0(t),\ U_{s+1}(t) = g_1(t)$. These requirements can be cast in the compact form

$$\frac{d}{dt}U(t) = AU(t) + r(t) \quad \text{for } t \ge 0, \qquad U(0) = u_0. \tag{1.2.4}$$

Here A is a tridiagonal $s \times s$ matrix, $r(t)$ and u_0 are given vectors in the s-dimensional complex vector space $\mathbb{C}^s$, and $U(t) \in \mathbb{C}^s$ has components $U_1(t), U_2(t), \cdots, U_s(t)$. The solution to the system of *ordinary* differential equations in (1.2.4) thus provides an approximation to the solution of the *partial* differential equation in (1.1.4).

The use of variable increments δ_j can be advantageous in cases where the smoothness of the true solution $u(x,t)$ varies significantly when x runs through $[0,1]$. In such cases one may choose δ_j especially small in those parts of $[0,1]$ where $u(x,t)$ is expected to be nonsmooth.

If the smoothness of $u(x,t)$ is not expected to vary strongly with x it may be recommended to use a so-called *uniform grid*, i.e. $x_{j+1} = x_j + \delta$ with $\delta = (s+1)^{-1}$ for $j = 0,1,\ldots,s$. Similarly as in the case of a nonuniform grid, we arrive at a semi-discrete problem of the form (1.2.4). But now we can use simply (1.2.2a) instead of the more general formula (1.2.3). In this case the matrix A is of the simple form

$$A = \delta^{-2}\begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & 1 \\ & & & 1 & -2 \end{bmatrix}.$$

More accurate approximations to the solution of (1.1.4) may be obtained, on a uniform grid, if one replaces the second order derivatives $\frac{\partial^2}{\partial x^2}u(x_j,t)$ by the finite difference expression in (1.2.2b) for $j = 2,3,\ldots,s-1$ and only by the expression from (1.2.2a) for $j = 1,s$. In this way one again arrives at a semi-discrete problem of the form (1.2.4) but now with a pentadiagonal matrix A.

Also Numerov's approximation (1.2.2c) can be used to construct a semi-discrete version of (1.1.4). On a uniform grid, where $x_j = j\cdot\delta,\quad \delta = (s+1)^{-1}$, one arrives, for $1 \le j \le s$, at the requirements

$$\frac{1}{12}[U'_{j-1}(t) + 10U'_j(t) + U'_{j+1}(t)] = \delta^{-2}[U_{j-1}(t) - 2U_j(t) + U_{j+1}(t)], \qquad U_j(0) = f(x_j).$$

By using $U_0(t) = g_0(t),\ U_{s+1}(t) = g_1(t)$ these requirements can be formulated in the compact form

$$A_1\frac{d}{dt}U(t) = A_0U(t) + q(t) \quad \text{ for } t \ge 0, \qquad U(0) = u_0,$$

with known vectors $q(t),\ u_0 \in \mathbb{C}^s$, and tridiagonal $s\times s$ matrices A_0, A_1. Multiplying by $(A_1)^{-1}$ we see that this initial value problem for $U(t)$ is of the form (1.2.4) with $A = (A_1)^{-1}A_0,\quad r(t) = (A_1)^{-1}q(t)$.

1.2.4 Semi-discretization in more than one space variable

In the case of two or three space variables we can proceed similarly as above.

As an illustration we consider the semi-discretization of problem (1.1.7) using the finite difference approximation (1.2.2a). Choose an integer $N \ge 1,\ \delta = (N+1)^{-1}$, and define the gridpoints $p_{jk} = (j\delta, k\delta)$ for $j = 0,1,\ldots,N+1$ and $k = 0,1,\ldots,N+1$. Replacing the partial derivatives $\frac{\partial^2}{\partial x^2}u(p_{jk},t), \frac{\partial^2}{\partial y^2}u(p_{jk},t)$ occurring in (1.1.7) (when $x = j\delta,\ y = k\delta$) by the finite difference approximations

$$\delta^{-2}[u(p_{j-1,k},t) - 2u(p_{jk},t) + u(p_{j+1,k},t)], \qquad \delta^{-2}[u(p_{j,k-1},t) - 2u(p_{jk},t) + u(p_{j,k+1},t)],$$

respectively, one obtains the system of ordinary differential equations

$$\begin{cases} \dfrac{d}{dt}U(p_{jk},t) = \delta^{-2}[U(p_{j-1,k},t) + U(p_{j+1,k},t) + U(p_{j,k-1},t) + U(p_{j,k+1},t) - 4U(p_{jk},t)], \\ \text{where } j = 1,2,\ldots,N \text{ and } k = 1,2,\ldots,N. \end{cases}$$

In the right-hand members of this system we set $U(p_{mn},t) = g(p_{mn},t)$ for all $p_{mn} \in \partial\Omega$ (cf. (1.1.7)). Moreover, we supplement the system by the initial conditions

$$U(p_{jk},0) = f(p_{jk}), \quad \text{where } j = 1,2,\ldots,N \text{ and } k = 1,2,\ldots,N,$$

in conformity with the initial condition in (1.1.7).

For $t \ge 0$ we introduce a vector $U(t) \in \mathbb{C}^s$, with $s = N^2$, the components of which are, in some fixed order, the values $U(p_{jk},t)$ (with $1 \le j \le N,\ 1 \le k \le N$). The above system of N^2 ordinary differential equations, together with the corresponding initial conditions, can again be cast in the form (1.2.4). The components of u_0 are equal to values $f(p_{jk})$, and the components of $r(t)$ depend on values $g(p_{mn},t)$. Further, each row of the matrix A contains at most 5 nonzero entries. The entries on the main diagonal equal $-4/\delta^2$, and the nonzero off-diagonal elements are equal to $1/\delta^2$.

1.2.5 Semi-discretization of diffusion and convection phenomena

1.2.5.1 Approximating diffusion and convection terms

We now consider in some detail various finite difference approximations to the diffusion and convection terms occurring in (1.1.6).

We shall use, for any real ξ, the notations

$$\xi^+ = \begin{cases} \xi & \text{if } \xi > 0, \\ 0 & \text{otherwise} \end{cases} \qquad \text{and} \qquad \xi^- = \begin{cases} \xi & \text{if } \xi < 0, \\ 0 & \text{otherwise} \end{cases} .$$

One easily sees that

$$\xi = \xi^+ + \xi^-, \quad |\xi| = \xi^+ - \xi^-, \quad (-\xi)^+ = -(\xi)^-, \quad (-\xi)^- = -(\xi)^+.$$

In the following we consider an increment $\delta > 0$ and given function $a(x),\ b(x),\ v(x)$. Further, we denote by $\varepsilon(x)$ certain parameter values to be specified below.

We consider the following four finite difference approximations.

(1.2.5a) $\quad (b(x)v(x))' = \{\beta \cdot v(x-\delta) + \alpha \cdot v(x) + \gamma \cdot v(x+\delta)\} + R, \quad \text{with}$

$$\gamma = \gamma(x) = \delta^{-1}\Big[\frac{1}{2}b(x+\delta/2) + \varepsilon(x+\delta/2)\Big([b(x+\delta/2)]^+ - \frac{1}{2}b(x+\delta/2)\Big)\Big],$$

$$\beta = \beta(x) = \delta^{-1}\Big[-\frac{1}{2}b(x-\delta/2) + \varepsilon(x-\delta/2)\Big(\frac{1}{2}b(x-\delta/2) - [b(x-\delta/2)]^-\Big)\Big],$$

$$\alpha = \alpha(x) = -\beta(x) - \gamma(x) + \delta^{-1}[-b(x-\delta/2) + b(x+\delta/2)].$$

(1.2.5b) $\quad (b(x)v(x))' = \{\beta \cdot v(x-\delta) + \alpha \cdot v(x) + \gamma \cdot v(x+\delta)\} + R, \quad \text{with}$

$$\gamma = \gamma(x) = \delta^{-1}\Big[\frac{1}{2}b(x+\delta) + \varepsilon(x+\delta)\Big([b(x+\delta)]^+ - \frac{1}{2}b(x+\delta)\Big)\Big],$$

$$\beta = \beta(x) = \delta^{-1}\Big[-\frac{1}{2}b(x-\delta) + \varepsilon(x-\delta)\Big(\frac{1}{2}b(x-\delta) - [b(x-\delta)]^-\Big)\Big],$$

$$\alpha = \alpha(x) = \delta^{-1}\varepsilon(x)\big([b(x)]^- - [b(x)]^+\big).$$

(1.2.6a) $\quad (a(x)v'(x))' = \{\beta \cdot v(x-\delta) + \alpha \cdot v(x) + \gamma \cdot v(x+\delta)\} + R, \text{ with}$

$$\gamma = \gamma(x) = \delta^{-2}a(x+\delta/2),$$

$$\beta = \beta(x) = \delta^{-2}a(x-\delta/2),$$

$$\alpha = \alpha(x) = -\beta(x) - \gamma(x).$$

(1.2.6b) $\quad (a(x)v'(x))' = \{\beta \cdot v(x-\delta) + \alpha \cdot v(x) + \gamma \cdot v(x+\delta)\} + R, \text{ with}$

$$\gamma = \gamma(x) = \frac{1}{2}\delta^{-2}[a(x) + a(x+\delta)],$$

$$\beta = \beta(x) = \frac{1}{2}\delta^{-2}[a(x-\delta) + a(x)],$$

$$\alpha = \alpha(x) = -\beta(x) - \gamma(x).$$

For $a(x),\ v(x)$ sufficiently smooth the residuals R in (1.2.6a), (1.2.6b) can be seen to be $\mathcal{O}(\delta^2)$ — which is similar to the residual in (1.2.2a). If $\varepsilon(x) \equiv 0$ and $b(x),\ v(x)$ are sufficiently smooth, then

also the residuals R in (1.2.5a), (1.2.5b) can be seen to be $\mathcal{O}(\delta^2)$—which is similar to the residual in (1.2.1c). But, in general, if the ε-values in (1.2.5a) or (1.2.5b) are nonvanishing parameters, then the corresponding residuals are only $\mathcal{O}(\delta)$.

In order to explain why parameter values different from zero may still be useful we focus on the physical meaning of the convection term in (1.1.6). In fact, $\frac{\partial}{\partial x}[b(x)u(x,t)]$ can be interpreted as the rate at which the concentration of a chemical species increases, at the position x and time t, due to the 1-dimensional flow field $V(x) = -b(x)$ (cf. Section 1.1).

For the time being assume $b(x) > 0$. Then the flow is from right to left, and therefore the values $u(x,t')$ with $t' > t$ will be influenced by the so-called upwind values $u(x',t)$ with $x' > x$. If one wants to use a finite difference approximation to $\frac{\partial}{\partial x}[b(x)u(x,t)]$ for determining $u(x,t')$ with $t' > t$, it thus seems quite natural to choose the forward difference approximation (1.2.1a) (with v replaced by $b \cdot u$) and not (1.2.1b) or (1.2.1c). The last two approximations would force $u(x,t')$ with $t' > t$ to depend on concentration values $u(x',t)$, with $x' < x$. But, these values should have no influence on $u(x,t')$. Similarly, for $b(x) < 0$ it is natural to use (1.2.1b) (with v replaced by $b \cdot u$).

Putting $\varepsilon(x-\delta) = \varepsilon(x) = \varepsilon(x+\delta) = 1$ in (1.2.5b), this formula reduces to (1.2.1a) (with v replaced by $b \cdot v$) if $b(x-\delta) > 0,\ b(x) > 0,\ b(x+\delta) > 0$; and it reduces to (1.2.1b) (with v replaced by $b \cdot v$) if $b(x-\delta) < 0,\ b(x) < 0,\ b(x+\delta) < 0$. With all parameter values equal to 1 formula (1.2.5b) is called an *upwind finite difference approximation*. Although, for small $\delta > 0$, its accuracy is lower than of (1.2.5b) with all ε-values equal to 0, it is the more natural approximation from a physical point of view. In practical applications the *upwind parameters* $\varepsilon(x)$ are often chosen such that $0 < \varepsilon(x) < 1$.

Arguments similar to the above apply to the finite difference approximation in (1.2.5a). Again the values $\varepsilon(x-\delta/2)$ and $\varepsilon(x+\delta/2)$ are called upwind parameters.

1.2.5.2 Semi-discretization of problem (1.1.6)

We shall construct a semi-discrete version of (1.1.6). We choose an integer $s \geq 2$, we put $\delta = s^{-1},\ x_\lambda = \lambda \cdot \delta$, and we define $a_\lambda = a(x_\lambda),\ b_\lambda = b(x_\lambda),\ c_\lambda = c(x_\lambda),\ d_\lambda = d(x_\lambda)$. Approximations $U_j(t) \approx u(x_j,t)$ (for $j = 1,2,\ldots,s$) can be obtained by solving

$$\frac{d}{dt}U(t) = AU(t) + r(t) \ \text{ for } t \geq 0, \qquad U(0) = u_0. \tag{1.2.7a}$$

Here $U(t) \in \mathbb{C}^s$ has the components $U_j(t)$ (for $j = 1,2,\ldots,s$), the initial vector $u_0 \in \mathbb{C}^s$ has the components $f(x_j)$ (for $j = 1,2,\ldots,s$), and the vectors $r(t) \in \mathbb{C}^s$ depend on $d_j,\ g(t)$. The $s \times s$ matrix A is of the form

$$A \ = \ A_2 \ + \ A_1 \ + A_0 \ , \tag{1.2.7b}$$

where $A_2,\ A_1,\ A_0$ correspond to the first three terms in the right-hand member of the partial differential equation in (1.1.6).

We define A_0 to be the diagonal matrix

$$A_0 = (\alpha_{jk}) \ \text{ with } \alpha_{jj} = c_j,\ \alpha_{jk} = 0 \quad (\text{for } 1 \leq j \leq s,\ 1 \leq k \leq s,\ j \neq k). \tag{1.2.8}$$

The $s \times s$ matrices $A_1,\ A_2$ are both tridiagonal; they can be constructed by using (1.2.5a), (1.2.6a) or (1.2.5b), (1.2.6b). In all cases we denote the entries of A_1, A_2 by α_{jk} (for $1 \leq j \leq s,\ 1 \leq k \leq s$), and we define

$$\begin{aligned}
&\alpha_{j,j+1} = \gamma(x_j) \quad (1 \leq j \leq s-1),\\
&\alpha_{j,j-1} = \beta(x_j) \quad (2 \leq j \leq s-1), \quad \alpha_{s,s-1} = \beta(x_s) + \gamma(x_s),\\
&\alpha_{j,j} = \alpha(x_j) \qquad (1 \leq j \leq s).
\end{aligned}$$

Here $\alpha(x),\ \beta(x),\ \gamma(x)$ are as in (1.2.5a), (1.2.6a), (1.2.5b) or (1.2.6b). In the above expression for $\alpha_{s,s-1}$ we have included the term $\gamma(x_s)$ so as to express the boundary condition $\frac{\partial}{\partial x}u(1,t)=0$ (see (1.1.6)).

Basing our construction on (1.2.5a), (1.2.6a) we arrive at the following definitions (1.2.9a), (1.2.10a), and basing it on (1.2.5b), (1.2.6b) at the definitions (1.2.9b), (1.2.10b).

(1.2.9a)
$$
\begin{aligned}
&A_1=(\alpha_{jk}) \text{ with , for } j=1,2,\dots,s,\\
&\alpha_{j,j+1}=\delta^{-1}\Big[\frac{1}{2}b_{j+\frac{1}{2}}+\varepsilon_{j+\frac{1}{2}}\Big(b^+_{j+\frac{1}{2}}-\frac{1}{2}b_{j+\frac{1}{2}}\Big)\Big],\\
&\alpha_{j,j-1}=\delta^{-1}\Big[-\frac{1}{2}b_{j-\frac{1}{2}}+\varepsilon_{j-\frac{1}{2}}\Big(\frac{1}{2}b_{j-\frac{1}{2}}-b^-_{j-\frac{1}{2}}\Big)\Big]+\delta_{j,s}\alpha_{s,s+1},\\
&\alpha_{j,j}=-\alpha_{j,j-1}-\alpha_{j,j+1}+\delta_{j,s}\alpha_{s,s+1}+\delta^{-1}[-b_{j-\frac{1}{2}}+b_{j+\frac{1}{2}}].
\end{aligned}
$$

(1.2.9b)
$$
\begin{aligned}
&A_1=(\alpha_{jk}) \text{ with, for } j=1,2,\dots,s,\\
&\alpha_{j,j+1}=\delta^{-1}\Big[\frac{1}{2}b_{j+1}+\varepsilon_{j+1}\Big(b^+_{j+1}-\frac{1}{2}b_{j+1}\Big)\Big],\\
&\alpha_{j,j-1}=\delta^{-1}\Big[-\frac{1}{2}b_{j-1}+\varepsilon_{j-1}\Big(\frac{1}{2}b_{j-1}-b^-_{j-1}\Big)\Big]+\delta_{j,s}\alpha_{s,s+1},\\
&\alpha_{j,j}=\delta^{-1}\varepsilon_j[b^-_j-b^+_j].
\end{aligned}
$$

(1.2.10a)
$$
\begin{aligned}
&A_2=(\alpha_{jk}) \text{ with}\\
&\alpha_{j,j+1}=\delta^{-2}a_{j+\frac{1}{2}} \quad (\text{ for } 1\le j\le s-1),\\
&\alpha_{j,j-1}=\delta^{-2}[a_{j-\frac{1}{2}}+\delta_{j,s}a_{s+\frac{1}{2}}] \quad (\text{ for } 2\le j\le s),\\
&\alpha_{j,j}=-\delta^{-2}[a_{j-\frac{1}{2}}+a_{j+\frac{1}{2}}] \quad (\text{ for } 1\le j\le s).
\end{aligned}
$$

(1.2.10b)
$$
\begin{aligned}
&A_2=(\alpha_{jk}) \text{ with}\\
&\alpha_{j,j+1}=\frac{1}{2}\delta^{-2}[a_j+a_{j+1}] \quad (\text{ for } 1\le j\le s-1),\\
&\alpha_{j,j-1}=\frac{1}{2}\delta^{-2}[a_{j-1}+a_j+\delta_{j,s}(a_s+a_{s+1})] \quad (\text{ for } 2\le j\le s),\\
&\alpha_{j,j}=-\delta^{-2}\Big[\frac{1}{2}a_{j-1}+a_j+\frac{1}{2}a_{j+1}\Big] \quad (\text{ for } 1\le j\le s).
\end{aligned}
$$

In the above four definitions we have used the *Kronecker index* δ_{jk} defined by

$$\delta_{jk}=1 \quad \text{for } j=k, \qquad \text{and} \quad \delta_{jk}=0 \quad \text{for } j\ne k.$$

Further, the values $a_\lambda=a(x_\lambda),\ b_\lambda=b(x_\lambda)$ which occur above with $x_\lambda>1$ should be interpreted as obtained by a mathematical extension of the given functions $a(x),\ b(x)$ defined for $0\le x\le 1$. These values $a_\lambda,\ b_\lambda$ need not to have a relation to any physical quantities outside the computation domain $[0,1]$. In (1.2.9a) and (1.2.9b) the quantities ε_λ are upwind parameters, with $0\le\varepsilon_\lambda\le 1$, and the quantities $\alpha_{s,s+1},\ \alpha_{1,0}$ have been introduced only for notational convenience.

1.3 Notes and remarks

For further discussion of the physical processes considered in Section 1.1, see e.g. Crank (1975), Hirsch (1988), Patankar (1980) and Zlatev (1995). In these references the transport equation (1.1.1) is discussed as well as various boundary conditions like those in (1.3).

There exists a vast literature on finite difference methods. The majority of the material on finite difference approximations in Section 1.2 can be found in one of the numerous books on this subject. We mention, in particular, the classical books Forsythe & Wasow (1960), Richtmyer & Morton (1967) and the more recent works by Hirsch (1988), Meiss & Marcowitz (1981), Roos, Stynes & Tobiska (1996), Shashkov (1996), Strikwerda (1989), Thomas (1995), Thomée (1990).

For the use of formula (1.2.2b), in deriving a semi-discrete version of a (nonlinear) diffusion-reaction problem with Dirichlet boundary conditions, see Stys & Stys (1991).

Upwinding, as discussed in Section 1.2.5, is related to what some authors call the addition of artificial diffusion. For relevant literature, see e.g. Griffiths, Christie & Mitchell (1980), Hirsch (1988), Patankar (1980), Roos, Stynes & Tobiska (1996), Strikwerda (1989).

Semi-discretization of (1.1.1) can be achieved by methods different from those described in Section 1.2, in particular by methods based on *finite volumes, finite elements* or *(pseudo) spectral approximations.* For finite volume methods see e.g. Hirsch (1988), Patankar (1980), Roos, Stynes & Tobiska (1996); for finite element methods see e.g. Hirsch (1988), Oden & Reddy (1976), Quarteroni & Valli (1994), Roos, Stynes & Tobiska (1996), Strang & Fix (1973); and for pseudo spectral methods see e.g. Canuto, Hussaini, Quarteroni & Zang (1988), Fornberg (1996), Gottlieb & Orszag (1977), Quarteroni & Valli (1994).

2 Analysis

2.1 Basic material

2.1.1 Definitions

Let $\varphi(t)$ denote a real function of a real variable t. This function is said to be *isotone* if $\varphi(t_1) \leq \varphi(t_2)$ for all $t_1 \leq t_2$ in the domain of definition of φ. Conversely, if $\varphi(t_1) \geq \varphi(t_2)$ for all such t_1, t_2, then the function is called *antitone*.

Let f denote any scalar function. If the k-th derivative of f exists, it will be denoted by $f^{(k)}$. We define

$$f^{(0)} = f.$$

Let $\gamma = \alpha + \mathrm{i}\beta = \rho\,\mathrm{e}^{\mathrm{i}\theta}$ denote any complex number, with real $\alpha,\ \beta,\ \theta$ and $\rho \geq 0$. Then we write

$$\alpha = \mathrm{Re}\,(\gamma), \qquad \beta = \mathrm{Im}\,(\gamma), \qquad \gamma^* = \alpha - \mathrm{i}\beta,$$

to denote the *real part*, *imaginary part* and the *complex conjugate* of γ, respectively. Further, if $\gamma \neq 0$, we define $\mathrm{Arg}\,(\gamma)$, the *principal value* of the *argument*, to be equal to θ with $-\pi < \theta \leq \pi$. We define $\mathrm{Arg}\,(0) = 0$, so that

$$-\pi < \mathrm{Arg}\,(\gamma) \leq \pi \quad \text{ for all } \gamma \in \mathbb{C}.$$

Let γ be any complex number, and $\rho \geq 0$. We denote the *circle* and *disk* in the complex plane with center γ and radius ρ by

$$\Gamma[\gamma, \rho] = \{\zeta :\ |\zeta - \gamma| = \rho\}, \qquad D[\gamma, \rho] = \{\zeta :\ |\zeta - \gamma| \leq \rho\},$$

respectively.

Let V be an arbitrary subset of the complex plane. The intersection of all convex sets W with $V \subset W \subset \mathbb{C}$ is called the *convex hull* of V, and is denoted by

$$\mathrm{conv}\ V.$$

Note that $\mathrm{conv}\ V$ is the smallest convex set in $\mathbb{C}$ containing the set V. The *interior*, *closure* and *boundary* of V are denoted by

$$\mathrm{int}\ V, \quad \mathrm{cl}\ V, \quad \partial V,$$

respectively. The *distance* from $\zeta \in \mathbb{C}$ to V is

$$\mathrm{d}(\zeta, V) = \ \inf\{|\zeta - \eta| :\ \eta \in V\}.$$

The left *half-plane* is denoted by

$$\mathbb{C}_- = \{\zeta :\ \zeta \in \mathbb{C} \text{ with } \mathrm{Re}\,(\zeta) \leq 0\},$$

and for $\delta \geq 0$ the *wedge* $W(\delta)$ is defined by

$$W(\delta) = \{\zeta :\ \zeta \in \mathbb{C} \text{ satisfies } |\mathrm{Arg}\,(-\zeta)| \leq \delta\}.$$

Whenever m and n are integers with $n < m$, we adopt the conventions

$$\max_{m\le j\le n} \cdots = 0, \qquad \sum_{j=m}^{n} \cdots = 0, \qquad \prod_{j=m}^{n} \cdots = 1,$$

and we put

$$0! = 1, \qquad 0^0 = 1.$$

2.1.2 Theorems

Let $R(\zeta) = P(\zeta)/Q(\zeta)$ be a given rational function, where $P(\zeta)$, $Q(\zeta)$ are polynomials of degrees m and n, respectively. We assume that $P(\zeta)$ and $Q(\zeta)$ have no common zeros.

Suppose ζ_1 is a zero of $Q(\zeta)$ with multiplicity $k \ge 1$. Then ζ_1 is called a *pole of order* k of $R(\zeta)$. We have $Q(\zeta) = (\zeta-\zeta_1)^k Q_1(\zeta)$, with $Q_1(\zeta)$ of degree $(n-k)$ and $Q_1(\zeta_1) \ne 0$. By expanding $P(\zeta)$ and $Q_1(\zeta)$ in powers of $(\zeta-\zeta_1)$ we see that, for $\zeta \ne \zeta_1$ and $|\zeta-\zeta_1|$ sufficiently small,

$$R(\zeta) = \alpha_{-k}(\zeta-\zeta_1)^{-k} + \cdots + \alpha_{-1}(\zeta-\zeta_1)^{-1} + \alpha_0 + \alpha_1(\zeta-\zeta_1) + \alpha_2(\zeta-\zeta_1)^2 + \cdots.$$

The expression $\alpha_{-k}(\zeta-\zeta_1)^{-k} + \cdots + \alpha_{-1}(\zeta-\zeta_1)^{-1}$ is called the *principal part* of $R(\zeta)$ at ζ_1, and α_{-1} is the *residue* of $R(\zeta)$ at ζ_1.

By subtracting from $R(\zeta)$ the principal parts corresponding to *all* different zeros of $Q(\zeta)$ we are left with a function $f(\zeta)$ that is holomorphic on $\mathbb{C}$. There are constants K, ρ_0 such that for all ζ with $|\zeta| \ge \rho_0$ we have the inequality

$$|f(\zeta)| \le K \cdot |\zeta|^{m-n}.$$

Further, for all $\zeta \in \mathbb{C}$ the value $f(\zeta)$ can be represented as the sum of a convergent power series

$$f(\zeta) = \gamma_0 + \gamma_1\zeta + \gamma_2\zeta^2 + \cdots.$$

Consequently, for any $\rho \ge \rho_0$, integration along the positively oriented circle $\Gamma[0,\rho]$ yields

$$|\gamma_k| = \Big|\frac{1}{2\pi i}\oint \frac{f(\zeta)}{\zeta^{k+1}}d\zeta\Big| \le K\cdot\rho^{m-n-k}.$$

By letting $\rho \to \infty$, we see that $\gamma_k = 0$ for all $k > m-n$. We thus arrive at the following.

Theorem 2.1.1 (*Partial fraction decomposition*) *Let $P(\zeta)$, $Q(\zeta)$ be nonzero polynomials of degrees m, n, respectively. Let $\zeta_1, \zeta_2, \ldots, \zeta_q$ be the zeros of $Q(\zeta)$, with multiplicities $k_1, k_2, \ldots, k_q$, respectively. Then, for $\zeta \ne \zeta_1, \ldots, \zeta_q$, we have*

$$\frac{P(\zeta)}{Q(\zeta)} = R_0(\zeta) + R_1(\zeta) + \cdots + R_q(\zeta).$$

Here $R_0(\zeta)$ is a nonzero polynomial of degree $(m-n)$ if $m \ge n$, and $R_0(\zeta) \equiv 0$ if $m < n$, so that we can write

$$R_0(\zeta) = \sum_{l=0}^{m-n} \alpha_{l,0}\,\zeta^l.$$

Further, for $1 \le j \le q$, the quantities $R_j(\zeta)$ are the principal parts of $P(\zeta)/Q(\zeta)$ at ζ_j, so that we can write

$$R_j(\zeta) = \sum_{l=1}^{k_j} \alpha_{-l,j}\,(\zeta-\zeta_j)^{-l} \qquad (\text{ for } 1 \le j \le q).$$

Without proof we state the following important theorem concerning the factorial function $n!$

Theorem 2.1.2 (*Stirlings formula*) *For each integer $n \geq 1$ there is a number θ, with $0 < \theta < 1$, such that*

$$n! = \Big(\frac{n}{\mathrm{e}}\Big)^n \sqrt{2\pi n}\, \exp\Big(\frac{\theta}{12n}\Big).$$

2.2 Estimates for the arc length of curves in the complex plane

2.2.1 The arc length

Let $\alpha < \beta$. Suppose $g(t)$ and $h(t)$ are piecewise continuously differentiable real functions defined for $\alpha \leq t \leq \beta$. The function $f(t) = g(t) + \mathrm{i}h(t)$ defines a curve Γ in the complex plane. The *length* of the curve Γ can be specified by

$$L = \int_\alpha^\beta |f'(t)|dt. \tag{2.2.1}$$

For each $t \in [\alpha, \beta]$ there is a real ω such that $g'(t) = |f'(t)| \cos\omega$, $h'(t) = |f'(t)| \sin\omega$, which implies

$$\int_0^{2\pi} |g'(t)\cos\theta + h'(t)\sin\theta|d\theta = \int_0^{2\pi} |\cos(\omega - \theta)| \cdot |f'(t)|d\theta = 4|f'(t)|.$$

We thus arrive at the representation

$$L = \tfrac{1}{4}\int_0^{2\pi} \Big\{\int_\alpha^\beta |g'(t)\cos\theta + h'(t)\sin\theta|dt\Big\}d\theta.$$

For each θ the quantity

$$L(\theta) = \int_\alpha^\beta \Big|\frac{d}{dt}\big\{g(t)\cos\theta + h(t)\sin\theta\big\}\Big|dt$$

is equal to the length of the projection of the curve Γ onto the straight line passing through the origin with angle θ to the real axis. Therefore, we have proved the following *expression of Cauchy for the length of a curve* in terms of the length of its projections:

$$L = \tfrac{1}{4}\int_0^{2\pi} L(\theta)\, d\theta. \tag{2.2.2}$$

2.2.2 The image of a circle under a rational function

We define a rational function $R(\zeta) = P(\zeta)/Q(\zeta)$ to be of *order* s if $P(\zeta)$, $Q(\zeta)$ are polynomials of degrees not exceeding s. We shall estimate the arc length of the image of a circle $\Gamma[\gamma, \rho]$ under a rational function of order s. This length can be written in the form

$$L = \int_0^{2\pi} \Big|\frac{d}{dt}R(\gamma + \rho\mathrm{e}^{\mathrm{i}t})\Big|dt = \int_{\Gamma[\gamma,\rho]} |R'(\zeta)|\,|d\zeta|.$$

We shall relate the length L to the quantity

$$M = \max\big\{|R(\zeta)| :\ \zeta \text{ lies on } \Gamma[\gamma, \rho]\big\}.$$

The main result of this subsection, Theorem 2.2.2, states that

$$L \;\le\; 2\pi s \cdot M,$$

whenever $R(\zeta)$ is a rational function of order s without poles on $\Gamma[\gamma,\rho]$.

This result is equivalent to the following remarkable fact. Denote the quantities L, M corresponding to the rational function $P(\zeta) = (\zeta-\gamma)^s$ by L_P, M_P, respectively. Since $L_P = 2\pi\ s\cdot\rho^s,\ M_P = \rho^s$, we see that, for all $R(\zeta)$ under consideration, the ratio L/M will never exceed the ratio L_P/M_P.

Our proof of Theorem 2.2.2 will rely on the following lemma.

Lemma 2.2.1 (*On the arc length of the image of a circle under a rational function*) *Let $R(\zeta) = P(\zeta)/Q(\zeta)$ be a rational function of order s without poles on $\Gamma[\gamma,\rho]$. Denote by $M(\theta)$ the maximum value of the projection of $R(\zeta)$ (where ζ runs through $\Gamma[\gamma,\rho]$) on the straight line passing through the origin with angle θ to the real axis, i.e.*

$$M(\theta) = \max\Big\{\mathrm{Re}\left(\mathrm{e}^{-\mathrm{i}\theta}R(\zeta)\right) :\ \zeta \text{ lies on } \Gamma[\gamma,\rho]\Big\}.$$

Then the length L of the image of $\Gamma[\gamma,\rho]$ under the mapping R satisfies

$$L \;\le\; s\int_0^{2\pi} M(\theta)\,d\theta.$$

Proof. 1. With no loss of generality we assume $\gamma = 0$, $\rho = 1$. In view of Cauchy's expression (2.2.2) it is sufficient to prove

$$L(\theta) \le 2s[M(\theta) + M(\theta+\pi)] \quad \text{for } \ 0 \le \theta \le 2\pi, \tag{2.2.3}$$

with

$$f(t) = g(t) + \mathrm{i}h(t) = R(\mathrm{e}^{\mathrm{i}t}), \quad 0 \le t \le 2\pi.$$

2. Let $\theta \in [0, 2\pi]$ be given, and write

$$F(t) = g(t)\cos\theta + h(t)\sin\theta.$$

Since $L(\theta) = \int_0^{2\pi} |F'(t)|\,dt$, we can assume with no loss of generality, that $F'(t)$ does not vanish identically on $[0, 2\pi]$. Therefore

$$L(\theta) \;=\; \sum_{j=1}^{k}\Big|\int_{t_{j-1}}^{t_j} F'(t)\,dt\Big|,$$

where $t_0 = 0 < t_1 < \ldots < t_k = 2\pi$ and $F'(t) \neq 0$ on each open interval (t_{j-1}, t_j).

We introduce the notations

$$a = \min_{0\le t\le 2\pi} F(t), \qquad b = \max_{0\le t\le 2\pi} F(t),$$

and we denote the range of values $F(t)$ obtained when t runs through the interval (t_{j-1}, t_j) by $F(t_{j-1}, t_j)$. Defining

$$\varphi_j(x) = 1 \text{ for } x \in F(t_{j-1}, t_j) \ \text{ and } \ \varphi_j(x) = 0 \text{ for } x \in [a,b]\backslash F(t_{j-1}, t_j),$$

we have

$$L(\theta) = \sum_{j=1}^{k} \int_a^b \varphi_j(x)\,dx = \int_a^b \sum_{j=1}^{k} \varphi_j(x)\,dx.$$

Below we shall prove

$$\sum_{j=1}^{k} \varphi_j(x) \le 2s \qquad \text{for } a \le x \le b. \tag{2.2.4}$$

Applying (2.2.4) we obtain

$$L(\theta) \le 2s(b-a) = 2s\left[\max_{0\le t\le 2\pi} \mathrm{Re}\big(\mathrm{e}^{-\mathrm{i}\theta}R(\mathrm{e}^{\mathrm{i}t})\big) - \min_{0\le t\le 2\pi} \mathrm{Re}\big(\mathrm{e}^{-\mathrm{i}\theta}R(\mathrm{e}^{\mathrm{i}t})\big)\right]$$

which proves (2.2.3).

3. In order to prove (2.2.4) we assume $x \in [a,b]$ to be given, and we write $\zeta = \mathrm{e}^{\mathrm{i}t}$ for $0 \le t \le 2\pi$. A straightforward calculation shows that

$$F(t) = x \tag{2.2.5}$$

is equivalent to

$$\mathrm{e}^{-\mathrm{i}\theta}P(\zeta)[Q(\zeta)]^* + \mathrm{e}^{\mathrm{i}\theta}[P(\zeta)]^*Q(\zeta) - 2xQ(\zeta)[Q(\zeta)]^* = 0$$

(where γ^* denotes the complex conjugate of γ). Multiplying the latter equality by ζ^s we arrive, in view of $\zeta^* = \zeta^{-1}$, at a relation

$$p(\zeta) = 0,$$

where $p(\zeta)$ is a polynomial of a degree not exceeding $2s$. Moreover $p(\zeta)$ does not vanish identically (since $F'(t)$ does not). Therefore, there exist at most $2s$ different values $t \in (0, 2\pi)$ with (2.2.5). This implies that $x \in F(t_{j-1}, t_j)$ for at most $2s$ different values j, so that (2.2.4) is fulfilled. □

Theorem 2.2.2 (*On the arc length of the image of a circle under a rational function*) *Let $R(\zeta)$ be a rational function of order s without poles on $\Gamma[\gamma,\rho]$. Then the length L of the image of $\Gamma[\gamma,\rho]$ under the mapping R satisfies*

$$L \;\le\; 2\pi s \cdot \max\{|R(\zeta)| :\; \zeta \textit{ lies on } \Gamma[\gamma,\rho]\}.$$

Proof. Immediate from Lemma 2.2.1. □

2.3 Notes and remarks

The partial fraction decomposition given in Section 2.1.2, as well as elementary properties of holomorphic functions, can be found in most of the numerous handbooks on complex analysis, e.g. Ahlfors (1966), Henrici (1974), Rudin (1974).

For Stirling's formula, given in Section 2.1.2, see e.g. Henrici (1974); for Cauchy's derivation of formula (2.2.2), see Cauchy (1888).

In the special case where $R(\zeta)$ is a rational function of order s, with all of its (possible) poles at $\zeta = \gamma$, Theorem 2.2.2 is an easy consequence of *Bernstein's inequality* (see e.g. Edwards (1967)). This inequality tells us that, for any such $R(\zeta)$, we have

$$\rho \cdot \max_{|\zeta-\gamma|=\rho} |R'(\zeta)| \;\leq\; s \cdot \max_{|\zeta-\gamma|=\rho} |R(\zeta)|.$$

In view of

$$L \;\leq\; 2\pi\rho \cdot \max_{|\zeta-\gamma|=\rho} |R'(\zeta)|,$$

one arrives immediately at the inequality given by Theorem 2.2.2.

For the case of general rational fuctions $R(\zeta)$, Theorem 2.2.2 was proved in Spijker (1991). The arguments used above in the proof of Lemma 2.2.1 are similar to those in that reference.

A beautiful extension of Theorem 2.2.2, dealing with the image of a circle under the mapping $R(\zeta)$ on the Riemann sphere, was given by Wegert & Trefethen (1994). These authors gave a short proof of their extension by using a general formula, for the arc length of curves lying on a sphere, due to H. Poincaré.

3 Linear algebra

3.1 The Jordan canonical form

The set of all $s \times s$ matrices $A = (\alpha_{jk})$, with complex entries α_{jk}, is denoted by $\mathbb{C}^{s,s}$. With the usual addition $A + B$ and multiplication $\gamma \cdot A$, for $A, B \in \mathbb{C}^{s,s}$ and $\gamma \in \mathbb{C}$, the set $\mathbb{C}^{s,s}$ becomes a complex vector space of dimension s^2.

The $s \times s$ identity matrix and zero matrix are denoted by

$$I \text{ and } O,$$

respectively. Further, for all $s \times s$ matrices A we define

$$A^0 = I.$$

For any square matrix A we denote the set of its eigenvalues by $\sigma[A]$. This set is called the *spectrum* of A. The *spectral radius* is defined by

$$\rho(A) = \max\{|\lambda| : \ \lambda \in \sigma[A]\}.$$

Note that A is *regular*, i.e. A^{-1} exists, if and only if $0 \notin \sigma[A]$.

Without proof we state the following fundamental theorem.

Theorem 3.1.1 (*Jordan canonical form*) *Let A be a given $s \times s$ matrix. Then there exist a regular $T \in \mathbb{C}^{s,s}$ and block-diagonal $J = \text{diag}\,(J_1, J_2, \ldots, J_r) \in \mathbb{C}^{s,s}$ with*

$$A = TJT^{-1}.$$

Here the blocks J_k are bidiagonal matrices of order s_k,

$$J_k = \begin{bmatrix} \lambda_k & 1 & & & 0 \\ & \lambda_k & 1 & & \\ & & \ddots & \ddots & \\ & & & \ddots & 1 \\ 0 & & & & \lambda_k \end{bmatrix}, \quad \textit{with} \quad \lambda_k \in \sigma[A] \ \textit{ for } k = 1, 2, \ldots, r.$$

The Jordan matrix J is unique up to permutations of the Jordan blocks J_k.

Since the order of J_k is s_k we see that, for $\lambda \in \sigma[A]$, the sum

$$\sum_{\lambda_k = \lambda} s_k$$

is equal to the so-called *algebraic multiplicity* of λ, i.e. the multiplicity of λ as a zero of the characteristic polynomial of A. Clearly

$$s_1 + s_2 + \cdots + s_r = s.$$

We denote the identity matrix of order s_k by I_k, and define E_k by

$$J_k = \lambda_k I_k + E_k.$$

Further we define $s \times s$ matrices $P_k, \ R_k$ by

$$P_k = T \text{ diag } (O, \ldots, O, I_k, O, \ldots, O)\ T^{-1}, \qquad R_k = T \text{ diag } (O, \ldots, O, E_k, O, \ldots, O)\ T^{-1}.$$

From the above one easily obtains the following theorem (in which δ_{jk} is the Kronecker index already introduced in Subsection 1.2.5.2).

Theorem 3.1.2 (*Spectral representation theorem*) *Let A be a given $s \times s$ matrix. Then there is an integer r, with $1 \le r \le s$, such that A can be represented in the form*

$$A = \sum_{k=1}^{r} (\lambda_k P_k + R_k).$$

Here $\lambda_k \in \sigma[A]$, $P_j P_k = \delta_{jk} P_j$, $P_k R_k = R_k P_k = R_k$ for $1 \le k \le r$, $1 \le j \le r$, and $P_1 + P_2 + \cdots + P_r = I$. Further, for $1 \le k \le r$, we have $P_k \ne O$, and there are integers $s_k \ge 1$ such that $(R_k)^m = O$ if and only if $m \ge s_k$. The integers s_k satisfy $s_1 + s_2 + \cdots + s_r = s$.

For $\zeta \notin \sigma[A]$ we shall often deal with the matrix

$$(\zeta I - A)^{-1}.$$

It is called the *resolvent* of A at ζ. From the spectral representation of A one easily obtains the representation of the resolvent as given in the following theorem.

Theorem 3.1.3 (*Representation for the resolvent*) *Let $A \in \mathbb{C}^{s,s}$ and $\zeta \notin \sigma[A]$. Then*

$$(\zeta I - A)^{-1} = \sum_{k=1}^{r} \Big[(\zeta - \lambda_k)^{-1} P_k + (\zeta - \lambda_k)^{-2} R_k + \cdots + (\zeta - \lambda_k)^{-s_k} (R_k)^{s_k - 1} \Big],$$

where λ_k, P_k, R_k, s_k are as in the spectral representation theorem.

We denote the complex conjugate of any complex number γ by γ^*, and recall that the *Hermitian adjoint* A^* of $A = (\alpha_{jk}) \in \mathbb{C}^{s,s}$ is the $s \times s$ matrix whose entry in the j-th row and k-th column equals α_{kj}^* (for $1 \le j \le s$, $1 \le k \le s$). The matrix A is said to be *normal* if $AA^* = A^*A$, and *unitary* if $AA^* = I$.

Without proof we state the following important theorem about normal matrices.

Theorem 3.1.4 (*Jordan canonical form of normal matrices*) *Let A be a given $s \times s$ matrix. Then A is normal if and only if a Jordan decomposition $A = TJT^{-1}$ exists with T unitary and all blocks J_k of order 1.*

3.2 Norms and ε-pseudo-eigenvalues

With $\mathbb{C}^s$ we denote the vector space of all column vectors

$$x = \begin{pmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_s \end{pmatrix}, \quad \text{with complex } \xi_1, \xi_2, \ldots, \xi_s.$$

We recall that the *Hermitian adjoint* x^* of x is the row vector given by

$$x^* = (\xi_1^*, \xi_2^*, \ldots, \xi_s^*).$$

Arbitrary norms on $\mathbb{C}^s$ will be denoted by $|\cdot|$, and the so-called *Hölder norms* by

$$|x|_p = \left\{ \sum_{j=1}^{s} |\xi_j|^p \right\}^{1/p} \qquad \text{for } \ 1 \le p < \infty,$$

$$|x|_p = \max_{1 \le j \le s} |x_j| \qquad \text{for } \ p = \infty.$$

We thus have $|x|_2 = \sqrt{x^* x}$, also called the *Euclidean norm* of x.

By a *linear functional* on $\mathbb{C}^s$ we mean a linear mapping from $\mathbb{C}^s$ to $\mathbb{C}$. Such a mapping can be represented by a row vector with s complex components. Without proof we state the following lemma, which is useful in various practical situations.

Lemma 3.2.1 (*On linear functionals*) *Let $|\cdot|$ be an arbitrary norm on $\mathbb{C}^s$, and $y \in \mathbb{C}^s$. Then there exists a linear functional F on $\mathbb{C}^s$ such that*

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

Let $|\cdot|$ be an arbitrary vector norm on $\mathbb{C}^s$. For any $s \times s$ matrix A we write

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

the supremum being over all $x \in \mathbb{C}^s$, $x \neq 0$. The function $\|\cdot\|$ is easily seen to be a norm on $\mathbb{C}^{s,s}$, the so-called *matrix norm induced* by the vector norm $|\cdot|$. We denote the matrix norms, induced by the Hölder norms, by $\|\cdot\|_p$. We have

$$\|A\|_\infty = \max_{1 \le j \le s} \sum_{k=1}^{s} |\alpha_{jk}|, \quad \|A\|_1 = \|A^*\|_\infty, \quad \|A\|_2 = \sqrt{\rho(A^* A)}.$$

The norm $\|A\|_2$ is also called the *spectral norm* of A. Clearly, for unitary A we have $\|A\|_2 = 1$.

In the rest of this Section 3.2, $\|\cdot\|$ will stand for a matrix norm induced by an *arbitrary* vector norm $|\cdot|$ on $\mathbb{C}^s$.

Since $\|\cdot\|$ is a norm on the vector space $\mathbb{C}^{s,s}$, we can measure the 'distance' between two $s \times s$ matrices A and B by the quantity $\|A - B\|$. Accordingly, for given $A_k \in \mathbb{C}^{s,s}$, we shall write $\lim_{k\to\infty} A_k = B$ and $\sum_{k=0}^{\infty} A_k = B$ to denote the situation where $B \in \mathbb{C}^{s,s}$ is such that $\lim_{k\to\infty} \|B - A_k\| = 0$ or $\lim_{n\to\infty} \|B - \sum_{k=0}^{n} A_k\| = 0$, respectively. We note that the limit concept thus defined in $\mathbb{C}^{s,s}$ is independent of the underlying norm $\|\cdot\|$. This is a consequence of the fact that for any two norms, say $\|\cdot\|$ and $\|\cdot\|'$, there is a constant $\alpha > 0$ such that $\|A\|' \le \alpha \cdot \|A\|$ (for all $A \in \mathbb{C}^{s,s}$).

It is easily seen that, for any $A, B \in \mathbb{C}^{s,s}$, we have

$$\|AB\| \le \|A\| \, \|B\|, \quad \text{and} \quad \|I\| = 1.$$

From Theorem 3.1.2 (or from the arguments to be given in Example 3.3.7) one can obtain the *spectral radius formula*

$$\lim_{n\to\infty} \|A^n\|^{1/n} = \rho(A).$$

In the following definition we denote by ε an arbitrary, nonnegative real constant.

Definition 3.2.2 λ is an *ε-pseudo-eigenvalue* of $A \in \mathbb{C}^{s,s}$ if
(i) there is an $E \in \mathbb{C}^{s,s}$ with $\|E\| \leq \varepsilon$ such that $\lambda \in \sigma[A+E]$.
The set of all ε-pseudo-eigenvalues of A is called the *ε-pseudospectrum*, and is denoted by $\sigma_\varepsilon[A]$. □

Note that in general the answer to the question of whether λ is an ε-pseudo-eigenvalue of A not ony depends on ε and A, but also on the vector norm in $\mathbb{C}^s$ by which the matrix norm $\|\cdot\|$ is induced. Clearly, for $\varepsilon = 0$, the ε-pseudospectrum of A is equal to the spectrum $\sigma[A]$.

The concept of an ε-pseudo-eigenvalue can be related to the following properties.
(ii) There is an $s \times s$ matrix E with $\|E\| = \varepsilon$ such that $\lambda \in \sigma[A+E]$.
(iii) There is an $s \times s$ matrix U with $\|U\| = 1$ and $\|(A-\lambda I)U\| \leq \varepsilon$.
(iv) There is a vector $u \in \mathbb{C}^s$ with $|u| = 1$ and $|(A-\lambda I)u| \leq \varepsilon$.
(v) $A - \lambda I$ is singular, or $A - \lambda I$ is regular with $\|(A-\lambda I)^{-1}\|^{-1} \leq \varepsilon$.

Theorem 3.2.3 (*Characterizations of ε-pseudo-eigenvalues*) *For given $\varepsilon > 0$, $s \geq 2$, $A \in \mathbb{C}^{s,s}$ the properties (i)–(v) are equivalent to each other.*

3.3 The Dunford-Taylor integral

Let $\varphi_{jk}(\zeta)$ be holomorphic complex functions on an open subset Ω of the complex plane, where $1 \leq j \leq s$, $1 \leq k \leq s$. We define the mapping Φ, from Ω into $\mathbb{C}^{s,s}$, by

$$\Phi(\zeta) = (\varphi_{jk}(\zeta)) \quad \text{for} \quad \zeta \in \Omega.$$

For any rectifiable curve Γ in Ω we define the *integral of* $\Phi(\zeta)$ *along* Γ by

$$\int_\Gamma \Phi(\zeta)d\zeta \;=\; \Big(\int_\Gamma \varphi_{jk}(\zeta)d\zeta\Big) \;\in\; \mathbb{C}^{s,s}.$$

Let $\|\cdot\|$ be any norm on $\mathbb{C}^{s,s}$. Then the inequality

$$\left\|\int_\Gamma \Phi(\zeta)d\zeta\right\| \leq \int_\Gamma \|\Phi(\zeta)\|\;|d\zeta|$$

holds. It is a consequence of the fact that the norm of a sum does not exceed the sum of the norms of the individual terms, and that the integral of $\Phi(\zeta)$ along Γ can be seen to be a limit of Riemann sums.

Let A be a given $s \times s$ matrix, and $f(\zeta)$ holomorphic on Ω. Under the assumption that $\sigma[A] \subset \Omega$ we define $f(A)$ by the so-called *Dunford-Taylor integral*.

$$f(A) = \frac{1}{2\pi \mathrm{i}} \int_\Gamma f(\zeta)(\zeta I - A)^{-1}d\zeta.$$

Here Γ consists of a finite number of rectifiable simple closed curves Γ_k with positive orientation. The interiors Ω_k of Γ_k are assumed to be mutually disjoint and to be such that

$$\sigma[A] \;\subset\; \bigcup_k \Omega_k, \qquad \bigcup_k (\Omega_k \cup \Gamma_k) \;\subset\; \Omega.$$

We note that $f(A)$ does not depend on Γ as long as Γ satisfies these conditions. This follows by applying Cauchy's integral theorem to the entries $\varphi_{jk}(\zeta)$ of $\Phi(\zeta) = f(\zeta)(\zeta I - A)^{-1}$ and using the above definition of the integral along Γ.

Without proof we state the following basic theorem about the Dunford-Taylor integral.

Theorem 3.3.1 *Let $A \in \mathbb{C}^{s,s}$ and $f(\zeta)$, $g(\zeta)$ holomorphic on an open set Ω with $\sigma[A] \subset \Omega$. Then*

(a) $f(\zeta \cdot I) = f(\zeta) \cdot I \quad$ *for all* $\zeta \in \Omega$,
(b) $(\alpha \cdot f + \beta \cdot g)(A) = \alpha \cdot f(A) + \beta \cdot g(A) \quad$ *for all* $\alpha, \beta \in \mathbb{C}$,
(c) $(f \cdot g)(A) = f(A) \cdot g(A)$,
(d) $\sigma[f(A)] = f(\sigma[A])$.

Assume, additionally: $B \in \mathbb{C}^{s,s}$ with $AB = BA$; Ω is connected and, for each $\lambda \in \sigma[A]$, the closed disk with center λ and radius $\rho(B)$ belongs to Ω. Then $\sigma[A+B] \subset \Omega$ and

(e) $$f(A+B) = \sum_{k=0}^{\infty} \frac{1}{k!} f^{(k)}(A) B^k.$$

Note that the addition and multiplications in the left-hand members of (b), (c) involve scalar functions, whereas those in the right-hand members $s \times s$ matrices. The right-hand member of (d) stands for $\{f(\lambda) : \ \lambda \in \sigma[A]\}$.

Example 3.3.2 Let $R(\zeta) = P(\zeta)/Q(\zeta)$ be a rational function and A an $s \times s$ matrix. We shall say that $R(A)$ *exists* if $Q(\lambda) \neq 0$ for all $\lambda \in \sigma[A]$. In this case the matrix $Q(A)$ is regular by (d), and from (c) it can be deduced that

$$R(A) = P(A)\big[Q(A)\big]^{-1} = \big[Q(A)\big]^{-1} P(A).$$

Using (b), (c) it can also be shown that the partial fraction decomposition of $R(\zeta)$ (see Theorem 2.1.1) applies to $R(A)$. If $R(A)$ exists, we have

$$\begin{cases} R(A) = R_0(A) + R_1(A) + \cdots + R_q(A), \\ R_0(A) = \displaystyle\sum_{l=0}^{m-n} \alpha_{l,0} A^l, \ R_j(A) = \sum_{l=1}^{k_j} \alpha_{-l,j} (A - \zeta_j I)^{-l} \ \ (\text{for } 1 \leq j \leq q), \end{cases}$$

where the coefficients $\alpha_{l,j}$ and the integers q, k_j are as in Theorem 2.1.1. □

Example 3.3.3 Let $f(\zeta) = \mathrm{e}^{\zeta}$. Writing $\mathrm{e}^A = \exp(A) = f(A)$, it follows from (e) that

$$\mathrm{e}^A = \sum_{k=0}^{\infty} \frac{1}{k!} A^k.$$ □

Further, for real t and positively oriented circle $\Gamma = \Gamma[0, \rho]$ with ρ sufficiently large, we have

$$\mathrm{e}^{tA} = \frac{1}{2\pi i} \int_{\Gamma} \mathrm{e}^{\xi} (\xi I - tA)^{-1} d\xi = \frac{1}{2\pi i} \int_{\Gamma} \mathrm{e}^{t\zeta} (\zeta I - A)^{-1} d\zeta.$$

Using the last of these two integral representations it can easily be seen that, when $h \to 0$, the matrix $h^{-1}\big[\mathrm{e}^{(t+h)A} - \mathrm{e}^{tA}\big]$ tends to $A\mathrm{e}^{tA}$. We thus have

$$\frac{d}{dt}\big(\mathrm{e}^{tA}\big) \ = \ A\mathrm{e}^{tA}.$$

Example 3.3.4 Let A, B be commuting $s \times s$ matrices. From (e) we have

$$(A+B)^n = \sum_{k=0}^{n} \binom{n}{k} A^k B^{n-k}$$

□

We state the following neat theorem without proof.

Theorem 3.3.5 *Let $f(\zeta)$ be holomorphic on an open set Ω containing the spectrum of the matrix A. Let $g(\xi)$ be holomorphic on $f(\Omega)$, and define $h(\zeta) = g(f(\zeta))$ (for $\zeta \in \Omega$). Then $h(A) = g(f(A))$.*

By substituting the representation for the resolvent, given in Theorem 3.1.3, in the Dunford-Taylor integral one arrives after an easy calculation at the following theorem.

Theorem 3.3.6 (*Spectral representation for $f(A)$*) *Let A be a given $s \times s$ matrix, and $f(\zeta)$ holomorphic on an open set Ω containing $\sigma[A]$. Then*

$$(3.3.1)\quad f(A) = \sum_{k=1}^{r} \Big[f(\lambda_k)P_k + f'(\lambda_k)R_k + \frac{1}{2!} f^{(2)}(\lambda_k)(R_k)^2 + \cdots + \frac{1}{(s_k-1)!} f^{(s_k-1)}(\lambda_k)(R_k)^{s_k-1} \Big].$$

Here λ_k, P_k, R_k, s_k are as in Theorem 3.1.2 (Spectral representation theorem).

Using the notations of Theorem 3.1.1 (Jordan canonical form), we can rewrite formula (3.3.1) as

$$(3.3.2a)\qquad f(A) = TDT^{-1},$$

where

$$(3.3.2b)\qquad D = \operatorname{diag}(D_1, D_2, \ldots, D_r)$$

is a block-diagonal matrix composed of matrices D_k of order s_k. Using the definition of E_k given in Section 3.1, we see that D_k allows the elegant representation

$$(3.3.3c)\qquad D_k = \sum_{j=0}^{\infty} \frac{f^{(j)}(\lambda_k)}{j!} (E_k)^j.$$

¿From this representation it follows that D_k is an upper triangular matrix of the form

$$(3.3.3d)\qquad D_k = \begin{bmatrix} f(\lambda_k) & \frac{f^{(1)}(\lambda_k)}{1!} & \frac{f^{(2)}(\lambda_k)}{2!} & \cdots & \frac{f^{(s_k-1)}(\lambda_k)}{(s_k-1)!} \\ 0 & f(\lambda_k) & \frac{f^{(1)}(\lambda_k)}{1!} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \frac{f^{(2)}(\lambda_k)}{2!} \\ \vdots & & \ddots & \ddots & \frac{f^{(1)}(\lambda_k)}{1!} \\ 0 & \cdots & \cdots & 0 & f(\lambda_k) \end{bmatrix}.$$

Example 3.3.7. We illustrate Theorem 3.3.6 by using it in a short proof of the spectral radius formula given in Section 3.2. For any integer $n \geq 1$ and $f(\zeta) = \zeta^n$, we have by Theorem 3.3.1(c)

$$A^n = f(A),$$

and by Theorem 3.3.6

$$f(A) = \sum_{k=1}^{r} \Big[(\lambda_k)^n P_k + \frac{n}{1!} (\lambda_k)^{n-1} R_k + \frac{n(n-1)}{2!} (\lambda_k)^{n-2} (R_k)^2 + \cdots + \binom{n}{s_k - 1} (\lambda_k)^{n-s_k+1} (R_k)^{s_k-1} \Big].$$

In view of the last two equalities there is a constant γ such that for $n = 1, 2, 3, \ldots$

$$\|A^n\| \ \leq \ \gamma \, n^p [\rho(A)]^n, \qquad \text{with} \ \ p = \max_k (s_k - 1).$$

Since $\|A^n\| \geq \rho(A^n) = [\rho(A)]^n$ there follows

$$\rho(A) \ \leq \ \|A^n\|^{1/n} \ \leq \ (\gamma \, n^p)^{1/n} \rho(A),$$

from which we obtain the spectral radius formula by letting $n \to \infty$.

3.4 The logarithmic norm

Throughout this section $|\cdot|$ denotes an arbitrary norm on $\mathbb{C}^s$, and $\|\cdot\|$ stands for the corresponding induced matrix norm.

Consider the initial value problem

$$U'(t) = A\,U(t), \quad U(0) = v,$$

where v is a given vector in $\mathbb{C}^s$ with norm $|v| = 1$. ¿From the relation $\frac{d}{dt}(\mathrm{e}^{tA}) = A\mathrm{e}^{tA}$ (obtained in Example 3.3.3), the solution $U(t) \in \mathbb{C}^s$ can be seen to be equal to $U(t) = e^{tA}v$ (for $t \in \mathbb{R}$). The logarithmic norm of A, to be defined below, is useful in estimating the norm of $\mathrm{e}^{tA}v$.

For real t with $t \neq 0$ and $v, w \in \mathbb{C}^s$ we define the difference quotient

$$m(v, w; t) = t^{-1}[|v + tw| - |v|].$$

The following simple lemma is quite useful. It can be proved by elementary considerations.

Lemma 3.4.1 *For fixed vectors v, w the quantity $m(v, w; t)$ is isotone with respect to $t \in \mathbb{R} \setminus \{0\}$. Further $-|w| \leq m(v, w; t) \leq |w|$.*

In view of this lemma we can define the (finite) quantities

$$m_-(v, w) = \lim_{t \to 0-} m(v, w; t), \qquad m_+(v, w) = \lim_{t \to 0+} m(v, w; t).$$

For $A \in \mathbb{C}^{s,s}$ we introduce the notations

$$m_-(A) = \sup_{|v|=1} m_-(v, Av), \qquad m_+(A) = \sup_{|v|=1} m_+(v, Av).$$

Further, for $A \in \mathbb{C}^{s,s}$ and $t \neq 0$, we define

$$\mu(A; t) = \sup_{|v|=1} m(v, Av; t).$$

Using the definition of $\|\cdot\|$, in terms of the norm $|\cdot|$ on $\mathbb{C}^s$, it can be seen that

$$\mu(A; t) = t^{-1} \Big[\big\| (I + tA)^{-1} \big\|^{-1} - 1 \Big] \quad \text{for} \ \ t < 0, \qquad |t|\rho(A) < 1,$$

and

$$\mu(A; t) = t^{-1} \Big[\big\| I + tA \big\| - 1 \Big] \quad \text{for} \ \ t > 0.$$

Theorem 3.4.2 (*Representations of the logarithmic norm*) *We have*

$$\lim_{t \to 0-} \mu(A; t) = m_-(A) = m_+(A) = \lim_{t \to 0+} \mu(A; t).$$

Any of the four quantities in the theorem is called the *logarithmic norm* of A, and will be denoted by

$$\mu(A).$$

Theorem 3.4.3 (*Characterization of the logarithmic norm*) *The logarithmic norm of A is equal to the smallest constant ω such that*

$$\left\|\mathrm{e}^{tA}\right\| \le \mathrm{e}^{t\omega} \qquad \text{for all } \ t > 0.$$

Proof. 1. From elementary numerical analysis it is known that the explicit Euler method, for the numerical solution of the initial value problem $U'(t) = AU(t)$, $U(0) = v$, is convergent. Hence $(I + \frac{t}{n}A)^n v$ tends to $\mathrm{e}^{tA}v$ as $n \to \infty$. Therefore $(I + \frac{t}{n}A)^n$ tends to e^{tA} (for $n \to \infty$). We have, for $t > 0$,

$$\left\|(I + \tfrac{t}{n}A)^n\right\| \le \{1 + \tfrac{t}{n}\mu(A; \tfrac{t}{n})\}^n = \{1 + \tfrac{t}{n}[\mu(A) + \varepsilon_n]\}^n,$$

with $\varepsilon_n \to 0$ for $n \to \infty$. Consequently $\|\mathrm{e}^{tA}\| \le \mathrm{e}^{t\mu(A)}$ for $t > 0$.

2. Suppose ω is such that $\|\mathrm{e}^{tA}\| \le \mathrm{e}^{t\omega}$ (for all $t > 0$).

Let $v \in \mathbb{C}^s$ with $|v| = 1$, and let $U(t)$ denote the solution to the initial value problem

$$U'(t) = AU(t), \qquad U(0) = v.$$

For $t > 0$ we have

$$m(v, Av; t) = t^{-1}\big[|v + tAv| - 1\big] = t^{-1}\big[|U(t)| + \varepsilon(t) - 1\big],$$

where

$$\lim_{t \to 0+} t^{-1} \cdot \varepsilon(t) \ = \ 0.$$

Hence

$$m_+(v, Av) \ = \ \lim_{t \to 0+} m\,(v, Av; t) \ \le \ \lim_{t \to 0+} \big[t^{-1}(\mathrm{e}^{t\omega} - 1) + t^{-1}\varepsilon(t)\big] \ = \ \omega.$$

Since $\mu(A) = m_+(A)$, it follows that $\mu(A) \le \omega$. □

Theorem 3.4.4 (*Properties of the logarithmic norm*) *Let $A, B \in \mathbb{C}^{s,s}$ and $\gamma \in \mathbb{C}$. Then the following relations are valid.*

(i) $-\|A\| \le \mu(A) \le \|A\|$,

(ii) $\mu(I) = 1, \quad \mu(O) = 0, \quad \mu(A + \gamma \cdot I) = \mu(A) + \mathrm{Re}\ \gamma$,

(iii) $\mu(\gamma \cdot A) = \gamma \cdot \mu(A)$ *provided* $\gamma \ge 0$ *(positive homogeneity)*,

(iv) $\mu(A + B) \le \mu(A) + \mu(B)$ *(sub-additivity)*,

(v) $|\mu(A) - \mu(B)| \le \|A - B\|$ *(continuity)*,

(vi) $\max\{\mathrm{Re}\ \lambda : \ \lambda \in \sigma[A]\} \le \mu(A)$.

We denote the logarithmic norm of A corresponding to the norm $|\cdot|_p$ on $\mathbb{C}^s$ by

$$\mu_p(A).$$

Theorem 3.4.5 (*Expressions for $\mu_p(A)$*) *For arbitrary $A \in \mathbb{C}^{s,s}$ we have*

$$\mu_\infty(A) = \max_j \ \big(\mathrm{Re}\ (\alpha_{jj}) + \sum_{k \ne j} |\alpha_{jk}|\big), \qquad \mu_1(A) = \mu_\infty(A^*),$$

$$\mu_2(A) = \max_{|v|=1} \ \mathrm{Re}\ (v^* A v) = \max\{\ \lambda : \ \lambda \in \sigma[\tfrac{1}{2}(A + A^*)]\}.$$

For normal matrices A we have

$$\mu_2(A) = \max\{\ \mathrm{Re}\ \lambda : \ \lambda \in \sigma[A]\}.$$

Definition 3.4.6 The matrix $A \in \mathbb{C}^{s,s}$ is said to satisfy a *circle condition*, with respect to the disk $D[\gamma, \rho]$, if

$$\|A - \gamma \cdot I\| \leq \rho.$$

□

In some of the following it will be important to check whether A satisfies a circle condition, and to interprete such a condition in terms of $\|A\|$ and $\mu(A)$. The following theorem is useful in this context.

Theorem 3.4.7 (*On circle conditions*) *Let $A \in \mathbb{C}^{s,s}$ and $\alpha > \omega$.*

a) *If $\|A + \frac{1}{2}(\alpha - \omega)I\| \leq \frac{1}{2}(\alpha + \omega)$ then $\|A\| \leq \alpha, \quad \mu(A) \leq \omega$.*

b) *Assume all diagonal elements α_{jj} of $A = (\alpha_{jk})$ are real, and $p = 1$ or $p = \infty$. Then*

$$\|A + \tfrac{1}{2}(\alpha - \omega)I\|_p \leq \tfrac{1}{2}(\alpha + \omega) \text{ if and only if } \|A\|_p \leq \alpha, \quad \mu_p(A) \leq \omega.$$

Proof. a) Since

$$\|A\| - \tfrac{1}{2}(\alpha - \omega) \leq \|A + \tfrac{1}{2}(\alpha - \omega)I\| \leq \tfrac{1}{2}(\alpha + \omega),$$

we have $\|A\| \leq \alpha$.

Writing $\gamma = \frac{1}{2}(\alpha - \omega)$ we obtain

$$\mu(A) = \mu(A + \gamma I) - \gamma \leq \|A + \gamma I\| - \gamma \leq (\gamma + \omega) - \gamma,$$

and therefore $\mu(A) \leq \omega$.

b) Since for any matrix A we have $\|A\|_1 = \|A^*\|_\infty, \ \mu_1(A) = \mu_\infty(A^*)$, it is sufficient to prove the statement for $p = \infty$.

Let $\|A\|_\infty \leq \alpha$, $\mu_\infty(A) \leq \omega$. For $1 \leq j \leq s$ we introduce

$$\sigma_j = \sum_{k \neq j} |\alpha_{jk}|.$$

From the expressions, given in the above for $\|A\|_\infty$, $\mu_\infty(A)$, we see that

$$|\alpha_{jj}| + \sigma_j \leq \alpha, \qquad \alpha_{jj} + \sigma_j \leq \omega.$$

In order to prove $\|(A + \frac{1}{2}(\alpha - \omega)I\|_\infty \leq \frac{1}{2}(\alpha + \omega)$ we only have to show that

$$|\alpha_{jj} + \tfrac{1}{2}(\alpha - \omega)| + \sigma_j \leq \tfrac{1}{2}(\alpha + \omega)$$

for $1 \leq j \leq s$.

Fix j, and define θ by

$$\begin{aligned} \theta = 0 \quad &\text{if} \quad \alpha_{jj} + \tfrac{1}{2}(\alpha - \omega) \geq 0, \\ \theta = 1 \quad &\text{if} \quad \alpha_{jj} + \tfrac{1}{2}(\alpha - \omega) < 0. \end{aligned}$$

We have

$$\begin{aligned} |\alpha_{jj} + \tfrac{1}{2}(\alpha - \omega)| + \sigma_j &= \theta\{-\alpha_{jj} - \tfrac{1}{2}(\alpha - \omega) + \sigma_j\} + (1 - \theta)\{\alpha_{jj} + \tfrac{1}{2}(\alpha - \omega) + \sigma_j\} \\ &\leq \theta\{|\alpha_{jj}| - \tfrac{1}{2}(\alpha - \omega) + \sigma_j\} + (1 - \theta)\{\alpha_{jj} + \tfrac{1}{2}(\alpha - \omega) + \sigma_j\} \\ &\leq \theta\{\alpha - \tfrac{1}{2}(\alpha - \omega)\} + (1 - \theta)\{\omega + \tfrac{1}{2}(\alpha - \omega)\} \\ &= \tfrac{1}{2}(\alpha + \omega). \end{aligned}$$

□

3.5 The M-numerical range

The *classical numerical range* of a given matrix $A \in \mathbb{C}^{s,s}$ is defined as the set of complex numbers

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

It is known to be a convex set containing all eigenvalues of A. Below we consider a generalization which will be useful in the following sections.

Let $|\cdot|$ be an arbitrary norm on $\mathbb{C}^s$, and $M \geq 1$. For a given matrix $A \in \mathbb{C}^{s,s}$ we define the disk $D[\gamma, \rho]$ to be *suitable* if

$$\left\|(A - \gamma I)^k\right\| \leq M\rho^k \quad \text{for} \ \ k = 1, 2, 3, \ldots. \tag{3.5.1}$$

Using the spectral radius formula (Section 3.2) we see that, for any suitable disk $D[\gamma, \rho]$, the inclusion $\sigma[A - \gamma I] \subset D[0, \rho]$ is valid. This implies that

$$\sigma[A] \subset D[\gamma, \rho]. \tag{3.5.2}$$

In view of (3.5.2), the best enclosure of $\sigma[A]$ obtainable from relations of the form (3.5.1) equals the intersection of all suitable disks. This is a motivation for the following definition.

Definition 3.5.1 The *M-numerical range* of A with respect to the norm $|\cdot|$ is the set $\tau[A]$ given by

$$\tau[A] = \bigcap D[\gamma, \rho],$$

where the intersection is over all disks that are suitable. If we want to express the dependence of $\tau[A]$ on M, we write

$$\tau[A] = \tau[A, M],$$

and the numerical ranges corresponding to the Hölder norms $|\cdot|_p$ are denoted by

$$\tau_p[A, M]. \qquad \square$$

Theorem 3.5.2 (*Basic properties of the M-numerical range*) *For arbitrary norm $|\cdot|$ and constant $M \geq 1$ the following holds:*

(i) $\tau[A, M]$ *is a closed, bounded, convex subset of the complex plane,*
(ii) $\tau[\zeta_0 I + \zeta_1 A, M] = \zeta_0 + \zeta_1 \cdot \tau[A, M]$ *for all* $\zeta_0, \ \zeta_1 \in \mathbb{C}$,
(iii) $\tau[A, M_2] \subset \tau[A, M_1]$ *for* $1 \leq M_1 \leq M_2$,
(iv) $\text{conv } \sigma[A] \subset \tau[A, M]$,
(v) $\bigcap\limits_{M \geq 1} \tau[A, M] = \text{conv } \sigma[A]$.

Proof. The first four properties follow easily from the above definition. In order to prove property (v), choose any $\gamma_0 \in \mathbb{C}$, $\rho_0 > 0$ such that conv $\sigma[A]$ lies in the interior of the disk $D[\gamma_0, \rho_0]$. Since $\rho(A - \gamma_0 I) < \rho_0$ the spectral radius formula (Section 3.2) tells us that

$$\lim_{k \to \infty} \left\|(A - \gamma_0 I)^k\right\|^{\frac{1}{k}} < \rho_0.$$

Therefore, there is an M_0 with

$$\left\|(A - \gamma_0 I)^k\right\| \leq M_0(\rho_0)^k \quad \text{for} \ \ k = 1, 2, 3, \ldots,$$

so that $\tau[A, M_0] \subset D[\gamma_0, \rho_0]$.

Since conv $\sigma[A]$ is equal to the intersection of all disks $D[\gamma_0, \rho_0]$ of the above type, we see that conv $\sigma[A]$ must be equal to the intersection of all sets $\tau[A, M]$ with $M \geq 1$, i.e. (v). $\square$

Let W be an arbitrary convex subset of $\mathbb{C}$, and let $\zeta \in \mathbb{C}$. The *distance* from ζ to W is defined by

$$\mathrm{d}(\zeta, W) = \inf\,\{|\zeta - \xi| :\ \xi \in W\}.$$

If ξ belongs to the *boundary* ∂W of W and

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

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

We shall deal with the following four conditions on $A \in \mathbb{C}^{s,s}$ with respect to $W \subset \mathbb{C}$.

(I) $\tau[A, M] \subset W$.

(II) $\zeta I - A$ is regular and $\big\|(\zeta I - A)^{-k}\big\| \le M \cdot \big[\mathrm{d}(\zeta, W)\big]^{-k}$ for all $\zeta \in \mathbb{C} \setminus W$ and $k = 1, 2, 3, \ldots$.

(III) $\big\|\exp\,[t\mathrm{e}^{-\mathrm{i}\theta}(A - \xi I)]\big\| \le M$ for all $t \ge 0$, $\xi \in \partial W$, and normal directions θ to W at ξ.

(IV) There is a vector norm $|\cdot|_0$ on $\mathbb{C}^s$ such that the corresponding 1-numerical range satisfies $\tau_0[A, 1] \subset W$ and $|x| \le |x|_0 \le M \cdot |x|$ (for all $x \in \mathbb{C}^s$).

Theorem 3.5.3 (*Main theorem on the M-numerical range*) *Let $|\cdot|$ be an arbitrary norm on $\mathbb{C}^s$, and $A \in \mathbb{C}^{s,s}$. Let $M \ge 1$, and W an arbitrary nonempty, closed, and convex subset of $\mathbb{C}$. Let $\tau[A, M]$ be the M-numerical range of A with respect to $|\cdot|$. Then the conditions (I)–(IV) are equivalent to each other.*

Clearly, $\tau[A, M]$ is the smallest nonempty, closed and convex set $W \subset \mathbb{C}$ with property (I). Therefore, the above theorem reveals *three new characterizations* of the M-numerical range. We see that $\tau[A, M]$ equals the smallest nonempty, closed and convex set $W \subset \mathbb{C}$ with property (II), and the same holds with regard to the properties (III) and (IV).

We note that, for $M = 1$, further characterizations of the set $\tau[A, M]$ are possible. One of these is easily obtained by applying the characterization of the logarithmic norm (see Theorem 3.4.3 with $\omega = 0$) to the norm-inequality in (III). We see that, for $M = 1$ and $|\cdot|$, A, W as in the above theorem, the conditions (I)–(IV) become equivalent to the condition

(V) $\mu\big(\mathrm{e}^{-\mathrm{i}\theta}A\big) \le \mathrm{Re}(\mathrm{e}^{-\mathrm{i}\theta}\xi)$ for all $\xi \in \partial W$, and normal directions θ to W at ξ.

Theorem 3.5.4 (*Expressions for $\tau_p[A, 1]$*) *For arbitrary $A = (\alpha_{jk}) \in \mathbb{C}^{s,s}$ let $A' = (\alpha'_{jk})$ with $\alpha'_{jk} = \alpha_{kj}$, and let*

$$D_j = D[\alpha_{jj}, \sigma_j], \qquad \sigma_j = \sum_{k \ne j} |\alpha_{jk}| \qquad \text{for } \ j = 1, 2, \ldots, s.$$

We have

$$\tau_\infty[A, 1] = \mathrm{conv} \bigcup_{j=1}^{s} D_j, \qquad \tau_1[A, 1] = \tau_\infty[A', 1],$$

$$\tau_2[A, 1] = \{x^* A x :\ x \in \mathbb{C}^s \text{ with } \ x^* x = 1\}.$$

For normal A one has

$$\tau_2[A, 1] = \mathrm{conv}\ \sigma[A].$$

Using the above expressions for $\tau_p[A,1]$ with $p = 1, \infty$ one can establish a simple relation between circle conditions and the set $\tau_p[A,1]$. We have the following theorem.

Theorem 3.5.5 (*Relation between circle conditions and $\tau_p[A,1]$ for $p = 1, \infty$*) *Let $D[\gamma,\rho]$ be an arbitrary disk, $A = (\alpha_{jk}) \in \mathbb{C}^{s,s}$ and $p = 1$ or ∞. Then A satisfies a circle condition with respect to $D[\gamma,\rho]$ and $\|\cdot\|_p$, if and only if $\tau_p[A,1] \subset D[\gamma,\rho]$.*

Proof. Since $\|A\|_1 = \|A'\|_\infty$, $\tau_1[A,1] = \tau_\infty[A',1]$, where A' is as in Theorem 3.5.4, we only have to consider the case $p = \infty$.

If A satisfies a circle condition with respect to $D[\gamma,\rho]$, we have $\tau_\infty[A,1] \subset D[\gamma,\rho]$ by the definition of the 1-numerical range.

Conversely, assume $\tau_\infty[A,1] \subset D[\gamma,\rho]$. In view of Theorem 3.5.4, and with the notations of that theorem, there follows

$$D[\alpha_{jj},\sigma_j] \ \subset \ D[\gamma,\rho],$$

and consequently $|\alpha_{jj} - \gamma| + \sigma_j \le \rho$ (for $j = 1,2,\dots,s$). Hence $\|A - \gamma I\|_\infty \le \rho$. □

3.6 Linear algebra concepts and semi-discretization

We shall illustrate some of the above concepts and theorems by applying them to the matrices A originating in the process of semi-discretization described in Section 1.2.5.

Consider problem (1.1.6) and assume

$$a(x) > 0, \quad b(x) \le 0, \quad c(x) \le 0, \quad \text{and } \ b(x) \text{ is antitone .} \tag{3.6.1}$$

We shall deal with the matrix A defined by (1.2.7b), (1.2.8), (1.2.9a), (1.2.10a). Assume, with regard to (1.2.9a), that

$$\text{all upwind parameters satisfy } \ \varepsilon_\lambda = 1. \tag{3.6.2}$$

In this situation the matrix $A = (\alpha_{jk})$ is tridiagonal; we have

$$\begin{aligned}
&\alpha_{j,j+1} = \delta^{-2} a_{j+\frac12} \quad (\text{for } 1 \le j \le s-1),\\
&\alpha_{j,j-1} = \delta^{-2} a_{j-\frac12} - \delta^{-1} b_{j-\frac12} \quad (\text{for } 2 \le j \le s-1),\\
&\alpha_{j,j-1} = \delta^{-2}\big[a_{s-\frac12} + a_{s+\frac12}\big] - \delta^{-1} b_{s-\frac12} \quad (\text{for } j = s),\\
&\alpha_{j,j} = -\delta^{-2}\big[a_{j-\frac12} + a_{j+\frac12}\big] + \delta^{-1} b_{j+\frac12} + c_j \quad (\text{for } 1 \le j \le s).
\end{aligned} \tag{3.6.3}$$

From (3.6.1) we have, for $2 \le j \le s$, the inequalities

$$\alpha_{jj} + \sum_{k \ne j} \big|\alpha_{jk}\big| \ = \ \delta^{-1}\big[b_{j+\frac12} - b_{j-\frac12}\big] + c_j \le 0,$$

and

$$\alpha_{11} + \sum_{k \ne 1} \big|\alpha_{1k}\big| = -\delta^{-2} a_{\frac12} + \delta^{-1} b_{\frac32} + c_j \le 0.$$

Therefore, by Theorem 3.4.5,

$$\mu_\infty(A) \le 0. \tag{3.6.4}$$

By virtue of Theorem 3.4.3 there follows

$$\|\exp\,(tA)\|_\infty \le 1 \quad \text{for all } \; t > 0.$$

The last conclusion can be interpreted as a *stability result for the semi-discrete problem* (1.2.7a). Suppose $\widetilde{U}\,(t)$ is a solution to (1.2.7a) with u_0 replaced by $\widetilde{u}_0$. Then

$$\widetilde{U}\,(t) - U(t) = \exp\,(tA)(\widetilde{u}_0 - u_0),$$

and since the above matrix norm of exp (tA) does not exceed 1, we have

$$\left|\widetilde{U}\,(t) - U(t)\right|_\infty \le |\,\widetilde{u}_0 - u_0|_\infty\ .$$

This means that any perturbation in the initial vector u_0 is not amplified, when the time t increases and differences between solutions are measured in the maximum norm $|\cdot|_\infty$. This stability result for (1.2.7a) is a semi-discrete analogue of a stability property for (1.1.6) which is plausible from (3.6.1) when viewing $a(x)$, $b(x)$, $c(x)$ as corresponding to diffusion, convection and reaction, respectively.

An essential point in the above derivation of (3.6.4) is the property of A that all of its off-diagonal entries are nonnegative. If (3.6.2) is not fulfilled, then this property need not be present, and accordingly (3.6.4) can be violated. This explains once more why upwind approximations are advantageous.

In the situation (3.6.2) one can estimate $\mu_\infty(A)$, similarly as above, also for arbitrary $a(x) > 0$ and $b(x)$, $c(x)$ not necessarily satisfying (3.6.1). In this case (3.6.4) is replaced by an inequality

$$\mu_\infty(A) \le \omega \tag{3.6.5}$$

with a constant ω, depending only on the functions $b(x)$, $c(x)$.

The above estimates for $\mu_\infty(A)$ can still be established for values ε_λ somewhat smaller than 1. In fact, if

$$\varepsilon_\lambda \ge 1 - \frac{2a_\lambda}{\delta|b_\lambda|} \quad \text{for } \; \lambda = \frac{3}{2}, \frac{5}{2}, \ldots, \frac{2s+1}{2}, \tag{3.6.6}$$

then all off-diagonal entries of A are still nonnegative, and the same estimates are valid as for the fully upwind approximations defined by (3.6.2). The ratio

$$P_\lambda = \frac{\delta|b_\lambda|}{a_\lambda}$$

is called the *(mesh-)Péclet number* at the location $x_\lambda = \lambda \cdot \delta$. If diffusion dominates convection, we have small Péclet numbers, and (3.6.6) may be fulfilled with $\varepsilon_\lambda \equiv 0$. But, for strongly convection dominated problems the Péclet number is large, and (3.6.6) forces ε_λ to be close to 1.

We turn again to the situation (3.6.1), (3.6.2). From the general expression for $\|A\|_\infty$ (cf. Section 3.2) we see that the tridiagonal matrix A defined by (3.6.3) satisfies

$$\|A\|_\infty \le \alpha, \tag{3.6.7}$$

with

$$\alpha = 4\delta^{-2}|a|_\infty + 2\delta^{-1}|b|_\infty + |c|_\infty. \tag{3.6.8}$$

By applying Theorem 3.4.7 (part b), with $\omega = 0$ and $p = \infty$, it follows that A satisfies the circle condition

$$\|A + \frac{\alpha}{2}I\|_\infty \le \frac{\alpha}{2}. \tag{3.6.9}$$

This inequality implies that the 1-numerical range $\tau_\infty[A, 1]$ is contained in the disk $D[-\frac{\alpha}{2}, \frac{\alpha}{2}]$. More refined information about the shape of $\tau_\infty[A, 1]$ can be obtained by using Theorem 3.5.4.

3.7 Notes and Remarks

A proof of Theorem 3.1.1 (Jordan canonical form) and Theorem 3.1.4 (Jordan canonical form of normal matrices) can be found e.g. in Horn & Johnson (1990). The representation Theorems 3.1.2, 3.1.3 can be derived directly, without using Theorem 3.1.1, by using complex integration theory, see e.g. Kato (1976).

For further discussion of norms in $\mathbb{C}^s$ and $\mathbb{C}^{s,s}$ we refer to Horn & Johnson (1990). Lemma 3.2.1 (Linear functionals) can be viewed as a corollary to the so-called Hahn-Banach theorem from functional analysis, cf. Rudin (1973). For a direct derivation of Lemma 3.2.1, in the context of the finite dimensional space $\mathbb{C}^s$, see e.g. Horn & Johnson (1990).

The spectral radius formula (Section 3.2) is a special case of a general result valid in complex Banach algebras, cf. Rudin (1973). For further discussion of the spectral radius formula, for matrices $A \in \mathbb{C}^{s,s}$, we refer to Horn & Johnson (1990).

The concept of the ε-pseudospectrum was introduced and studied a.o. by Landau (1975), Reddy & Trefethen (1990, 1992), Varah (1979). Very interesting properties of the ε-pseudospectrum, as well as a wealth of applications, were discovered by L.N. Trefethen, see Trefethen (1996). The focus in the works just mentioned is on the (weighted) spectral norms $\|A\|_2$ for $A \in \mathbb{C}^{s,s}$. The ε-pseudospectrum for the case of arbitrary norms $\|A\|$ in $\mathbb{C}^{s,s}$ was discussed in Dorsselaer, Kraaijevanger & Spijker (1993). Theorem 3.2.3 (Characterizations of the ε-pseudo-eigenvalues) has been taken from that paper.

For additional reading about the Dunford-Taylor integral, see Conway (1985), Dowson (1978), Dunford & Schwartz (1958), Kato (1976). A direct and neat proof of the Theorems 3.3.1, 3.3.5, valid for the general case of linear operators in a Banach space, is given in Dowson(1978). We note that the parts (a)–(d) of Theorem 3.3.1 can also be proved easily by using Theorem 3.3.6. For further discussion of the latter theorem, see e.g. Dunford & Schwartz (1958), Kato (1976).

The logarithmic norm was introduced independently by Dahlquist (1959) and Lozinskij (1958). Subsequently, various properties of the logarithmic norm were established by Desoer & Haneda (1972) and Ström (1975). The last equality in Theorem 3.4.2 $(m_+(A) = \lim_{t\to 0+} \mu(A;t))$ and the Theorems 3.4.3, 3.4.4, 3.4.5 can be found, with complete proofs, in the above four references. Lemma 3.4.1 has been taken from Martin (1976, p.37). A proof of the second equality in Theorem 3.4.2 $(m_-(A) = m_+(A))$ can also be found in Martin (1976, p.246) — even for the general situation of operators acting on a Banach space. The fact that $\lim_{t\to 0-} \mu(A;t) = \lim_{t\to 0+} \mu(A;t)$ was proved in Dorsselaer & Spijker (1994). Theorem 3.4.7 (Circle conditions) has been taken from Spijker (1985).

There exists a vast literature on the classical numerical range and closely related issues, see Horn & Johnson (1994). For $M = 1$, the M-numerical range (Definition 3.5.1) can be seen to be equal to the so-called algebra numerical range, well known in some parts of functional analysis, see Bonsall & Duncan (1973, p.42) and (1980, p.35). The M-numerical range, for general $M \geq 1$, was introduced in Lenferink & Spijker (1990). In that paper Theorem 3.5.2 and part of Theorem 3.5.3 (equivalence of (I), (II) and (III)) was proved. The equivalence of (I) and (IV) was obtained in Spijker (1993). The expressions for $\tau_p[A, 1]$, given in Theorem 3.5.4, have been known for a long time, cf. Lenferink & Spijker (1990), Spijker (1993) for corresponding historical remarks.

4 The problem of stability in the numerical solution of differential equations

4.1 Linear stability analysis

We shall deal 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 method.

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 for stability. Further tools of recognized merit for assessing stability, in the solution of ordinary differential equations, comprise so-called *stability regions* in the complex plane. Since the mid sixties these stability regions have been studied extensively; numerous papers have appeared dealing with the shape and various peculiarities of these regions.

However, the above 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 certain numerical methods applied to *pure initial value* problems in linear differential equations with *constant* coefficients. In many cases of practical interest, Fourier transformation is *not* relevant to analysing stability: e.g. for pseudo-spectral methods applied to initial-boundary value problems, for finite difference or finite element methods related to highly irregular grids, and for methods applied to equations with strongly varying coefficients. Similarly, in the case of ordinary differential equations, stability regions are primarily relevant only to numerical methods when applied to *scalar* equations

$$U'(t) = \lambda U(t) \quad \text{for} \quad t \geq 0, \tag{4.1.1}$$

with given complex constant λ.

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 (4.1.1). In the following we shall discuss various theories which are relevant to these two questions. No essential use will be made of Fourier transformation.

4.2 Stability and power boundedness

In the following we shall deal with numerical processes of the form

$$u_n = Bu_{n-1} + r_n \qquad \text{for} \quad n = 1, 2, 3, \ldots, \tag{4.2.1}$$

with a given square matrix B of order $s \geq 1$ and given vectors $r_n \in \mathbb{C}^s$. The s-dimensional vectors u_n are computed sequentially from (4.2.1) starting from a given vector $u_0 \in \mathbb{C}^s$.

Processes of the form (4.2.1) occur in the numerical solution of linear initial value problems that are essentially more general than the simple classical test problems mentioned above. The vectors u_n provide numerical approximations to the true solution of the initial(-boundary) value problem under consideration.

As an illustration to (4.2.1) we consider again the one-dimensional diffusion-convection-reaction problem (1.1.6) of Section 1.1,

$$\begin{aligned}(4.2.2)\qquad &\frac{\partial}{\partial t}u(x,t)=\frac{\partial}{\partial x}[a(x)\frac{\partial}{\partial x}u(x,t)]+\frac{\partial}{\partial x}[b(x)u(x,t)]+c(x)u(x,t)+d(x),\\ &u(0,t)=g(t),\quad \frac{\partial}{\partial x}u(1,t)=0,\quad u(x,0)=f(x),\quad \text{where } 0\le x\le 1,\ t\ge 0.\end{aligned}$$

We choose positive increments $\Delta t=h$, $\Delta x=\delta=1/s$ and consider the approximation of $u(x,t)$ at $x=x_j=j\delta$, $t=nh$. We denote these approximations by u_j^n. The following finite difference scheme has been constructed by applying the 2-stage θ–Runge–Kutta method ($a_{11}=a_{12}=0$, $a_{21}=b_1=1-\theta$, $a_{22}=b_2=\theta$) to (1.2.7), (1.2.8), (1.2.9a) (with all $\varepsilon_\lambda=1$), (1.2.10a).

$$\begin{aligned}h^{-1}(u_j^n-u_j^{n-1})=&\delta^{-2}\left\{a_{j+1/2}\left[\theta u_{j+1}^n+(1-\theta)u_{j+1}^{n-1}\right]-(a_{j+1/2}+a_{j-1/2})\left[\theta u_j^n+\right.\right.\\ &\left.+\ (1-\theta)u_j^{n-1}\right]+a_{j-1/2}\left[\theta u_{j-1}^n+(1-\theta)u_{j-1}^{n-1}\right]\Big\}\\ &+\delta^{-1}\left\{b_{j+1/2}\left[\theta u_j^n+(1-\theta)u_j^{n-1}\right]-b_{j-1/2}\left[\theta u_{j-1}^n+(1-\theta)u_{j-1}^{n-1}\right]\right\}+\\ &+c_j\left[\theta u_j^n+(1-\theta)u_j^{n-1}\right]+d_j,\\ u_0^{n-1}=g((n-1)h),\quad &u_{s+1}^{n-1}=u_{s-1}^{n-1},\qquad u_j^0=f(x_j),\end{aligned}$$

where $j=1,2,\dots,s$ and $n=1,2,3,\dots$. In the above θ denotes a parameter, with $0\le\theta\le 1$, specifying the numerical process, and $x_\lambda=\lambda\delta$, $a_\lambda=a(x_\lambda)$, $b_\lambda=b(x_\lambda)$, $c_\lambda=c(x_\lambda)$, $d_\lambda=d(x_\lambda)$.

Defining vectors u_n by

$$u_n=\begin{pmatrix}u_1^n\\ u_2^n\\ \vdots\\ u_s^n\end{pmatrix}\simeq\begin{pmatrix}u(x_1,nh)\\ u(x_2,nh)\\ \vdots\\ u(x_s,nh)\end{pmatrix}$$

one easily verifies that the u_n satisfy a relation of the form (4.2.1). Here

$$(4.2.3)\qquad B=(I+(1-\theta)hA)(I-\theta hA)^{-1},$$

where $A=(\alpha_{jk})$ is an $s\times s$ tridiagonal matrix with its (nonzero) entries given by

$$\begin{aligned}(4.2.4)\qquad &\alpha_{j,j+1}=\delta^{-2}a_{j+\frac12}\quad(\text{for } 1\le j\le s-1),\\ &\alpha_{j,j-1}=\delta^{-2}a_{j-\frac12}-\delta^{-1}b_{j-\frac12}\quad(\text{for } 2\le j\le s-1),\\ &\alpha_{j,j-1}=\delta^{-2}\left[a_{s-\frac12}+a_{s+\frac12}\right]-\delta^{-1}b_{s-\frac12}\quad(\text{for } j=s),\\ &\alpha_{j,j}=-\delta^{-2}\left[a_{j-\frac12}+a_{j+\frac12}\right]+\delta^{-1}b_{j+\frac12}+c_j\quad(\text{for } 1\le j\le s).\end{aligned}$$

Clearly this matrix A equals the matrix given by (1.2.7b), (1.2.8), (1.2.9a) (with all $\varepsilon_\lambda=1$), (1.2.10a), and it coincides with (3.6.3).

Suppose the numerical calculations based on the general process (4.2.1) 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 (4.2.1) a crucial question is whether the difference $v_n = \widetilde{u}_n - u_n$ (for $n \geq 1$) can be bounded suitably in terms of the perturbation $v_0 = \widetilde{u}_0 - u_0$. Since

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

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 our 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

$$\|B^n\| \leq M_0 \qquad \text{for} \quad n = 0, 1, 2, \ldots, \tag{4.2.5}$$

where M_0 is a positive constant. Throughout this Chapter 4 we denote by $\|\cdot\|$ the spectral norm (see Section 3.2).

4.3 Power boundedness and the eigenvalue criterion

For any given matrix B one can easily deduce a criterion for the existence of an M_0 with property (4.2.5). In view of the Theorems 3.3.1, 3.3.6 we have

$$B^n = \sum_{k=1}^{r} \left[(\lambda_k)^n P_k + \binom{n}{1} (\lambda_k)^{n-1} R_k + \binom{n}{2} (\lambda_k)^{n-2} (R_k)^2 + \cdots + \binom{n}{s_k - 1} (\lambda_k)^{n-s_k+1} (R_k)^{s_k-1} \right]$$

(cf. Example 3.3.7). Here λ_k, P_k and R_k are related to the Jordan decomposition of our matrix B similarly as they were related to the decomposition of A in Section 3.1.

From this representation for B^n we easily conclude that $\|B^n\|$ remains bounded when $n \to \infty$ if and only if

$$|\lambda_k| \leq 1, \quad \text{and } s_k = 1 \quad \text{when } |\lambda_k| = 1.$$

We thus arrive at the following theorem.

Theorem 4.3.1 (*The eigenvalue criterion*) *For a given matrix B there exists a constant M_0 with property (4.2.5) if and only if*

$$\begin{cases} \textit{all eigenvalues } \lambda \textit{ of } B \textit{ have a modulus } |\lambda| \leq 1, \textit{ and any Jordan} \\ \textit{block corresponding to an eigenvalue } \lambda \textit{ with } |\lambda| = 1 \textit{ has order 1 .} \end{cases} \tag{4.3.1}$$

However, in the stability analysis of numerical processes one is often interested in property (4.2.5) for all B belonging to some infinite family $\mathcal{F}$ of matrices. The crucial question then is of whether a single finite M_0 exists such that (4.2.5) holds simultaneously for all B belonging to $\mathcal{F}$. In this situation, (4.3.1) 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 4.2 one can only expect great accuracy in the approximations u_j^n to $u(x_j, nh)$ when δ (and h) become very small. Accordingly one is primarily interested in bounds for B^n that are valid uniformly for all B of the form (4.2.3), (4.2.4) with arbitrarily small $\delta = 1/s$.

An instructive counterexample, illustrating the fact that the criterion (4.3.1) can be misleading for the case of families $\mathcal{F}$, is provided by the $s \times s$ bidiagonal matrices

$$B = \begin{pmatrix} -1/2 & 0 & 0 & \cdots & 0 \\ 3/2 & -1/2 & \ddots & \ddots & \vdots \\ \vdots & 3/2 & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 3/2 & -1/2 \end{pmatrix}. \tag{4.3.2}$$

Matrices of the form (4.3.2) may be thought of as arising in the numerical solution of the initial-boundary value problem

$$\frac{\partial}{\partial t}u(x,t) = -\frac{\partial}{\partial x}u(x,t), \quad u(0,t) = 0, \quad u(x,0) = f(x), \quad \text{where } 0 \le x \le 1,\ t \ge 0.$$

This problem is the same as (1.1.5) with $b(x) = -1$, $c(x,t) = 0$. Application of the (forward) Euler method to a fully upwind semi-discrete approximation, based on (1.2.1b), yields the numerical process

$$h^{-1}(u_j^n - u_j^{n-1}) = \delta^{-1}(u_{j-1}^{n-1} - u_j^{n-1}),$$
$$u_0^{n-1} = 0, \qquad u_j^0 = f(j/s).$$

Here $\Delta t = h > 0$, $\Delta x = \delta = 1/s$, and u_j^n approximates $u(j\delta, nh)$ for $j = 1, 2, \ldots, s$ and $n = 1, 2, 3, \ldots$. Clearly, with the choice $hs = 3/2$ this numerical process can be written in the form (4.2.1) with a matrix B as in (4.3.2).

For each $s \ge 1$ the matrix (4.3.2) satisfies the eigenvalue condition (4.3.1).

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

$$E = \begin{pmatrix} 0 & \cdots & \cdots & \cdots & 0 \\ 1 & \ddots & & & \vdots \\ 0 & 1 & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & 1 & 0 \end{pmatrix}, \tag{4.3.3}$$

we have from (4.3.2) the expression

$$B = -\frac{1}{2}I + \frac{3}{2}E.$$

Consequently (cf. Example 3.3.4),

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

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

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

For $s > n$ we thus have

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

Since

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

the spectral norm of B^n satisfies $\|B^n\| \ge \sqrt{1 - n/s} \cdot 2^n$. Denoting the $s \times s$ matrix B by B_s we thus have

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

Clearly, no M_0 can exist such that (4.2.5) is valid for all B belonging to $\mathcal{F} = \{B_s : s = 1, 2, 3, \ldots\}$.

It should be noted that in some special cases the eigenvalue criterion can be reliable. For normal matrices B we have, with the notation of Theorem 3.1.4 applied to the matrix B,

$$B^n = (TJT^{-1})^n = T \operatorname{diag}(\lambda_1^n, \lambda_2^n, \ldots, \lambda_s^n) T^{-1},$$

so that (4.3.1) implies $\|B^n\| \le \|T\| \cdot \|T^{-1}\|$. Since $T^{-1} = T^*$ we have $\|T\| = \|T^{-1}\| = 1$, and we arrive at the following theorem.

Theorem 4.3.2 (*The eigenvalue criterion for normal matrices*) *Let the matrix B be normal, and $\|\cdot\|$ denote the spectral norm. Then the following three statements are equivalent to each other.*

(i) *There is an M_0 such that (4.2.5) is fulfilled.*

(ii) *$\|B^n\| \le 1$ for $n = 1, 2, 3, \ldots$.*

(iii) *All eigenvalues λ of B have a modulus $|\lambda| \le 1$.*

But, in general the matrices B in (4.2.1) are not normal, and one has to look for conditions different from (4.3.1). In the following chapters we shall deal with conditions which are still reliable for the case of non normal matrices.

4.4 Notes and remarks

For a discussion of Fourier transformations and the corresponding Von Neumann condition for analysing stability we refer to the classical work Richtmyer & Morton (1967) and to Strikwerda (1989).

Already in the pioneering work by F. John (1952) the scope of Fourier transformation was widened in that it was used in deriving sufficient conditions for stability in the numerical solution of linear partial differential equations with mildly *variable* coefficients. For subsequent related work, relevant to equations with variable coefficients and also to *initial-boundary value* problems, the reader may consult Gustafsson, Kreiss & Sundström (1972), Kreiss (1966), Meis & Marcowitz (1981), Richtmyer & Morton (1967), Strikwerda (1989), Thomée (1990) and the references therein.

Stability regions, related to numerical methods for ordinary differential equations, are discussed extensively in the excellent works by Butcher (1987) and Hairer & Wanner (1996).

The fact that the eigenvalue criterion (4.3.1) can be a misleading guide to stability was already known in the sixties, 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 Morton (1980), Richtmyer & Morton (1967), Thomée (1990). The latter condition is not satisfied in example (4.3.2).

The counterexample (4.3.2) is similar to examples in Kreiss (1990), Reddy & Trefethen (1992), Richtmyer & Morton (1967), Spijker (1985). Further examples of instability under the eigenvalue condition (4.3.1) can be found in Griffiths, Christie & Mitchell (1980), Kraaijevanger, Lenferink & Spijker (1987), Lenferink & 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 (4.2.2). See Reddy & Trefethen (1992), Trefethen (1988) for related counterexamples in spectral methods.

5 Stability estimates under resolvent conditions on the numerical solution operator B

5.1 Power boundedness and the Kreiss resolvent condition

By $\|\cdot\|$ we denote, in this Section 5.1, the spectral norm.

In the early sixties H. -O. Kreiss established an important theorem, nowadays called the *Kreiss matrix theorem*, dealing with the uniform boundedness of $\|B^n\|$. Here $n = 1, 2, 3, \ldots$ and B belongs to a given, possibly infinite, family $\mathcal{F}$ of $s \times s$ matrices (with fixed $s \geq 1$). The theorem gives a series of conditions which are equivalent to the existence of a constant M_0 such that

$$\|B^n\| \leq M_0 \qquad \text{for } n = 0, 1, 2, \ldots, \tag{5.1.1}$$

simultaneously for all $B \in \mathcal{F}$.

One of the conditions in the Kreiss theorem involves the resolvent $(\zeta I - B)^{-1}$ of B, and amounts to the requirement that

$$\begin{aligned} &\zeta I - B \text{ is invertible, and } \|(\zeta I - B)^{-1}\| \leq M_1 \cdot (|\zeta| - 1)^{-1}, \\ &\text{for all complex numbers } \zeta \in \mathbb{C} \setminus D. \end{aligned} \tag{5.1.2}$$

Here M_1 is a positive constant, and D denotes the closed unit disk $\{\zeta : \zeta \in \mathbb{C} \text{ with } |\zeta| \leq 1\}$. We shall refer to (5.1.2) as the *Kreiss resolvent condition*. In many cases of practical interest it is easier to verify (5.1.2) than (5.1.1).

Let B be a given $s \times s$ matrix satisfying (5.1.1). Then, by Theorem 4.3.1 (Eigenvalue criterion), all eigenvalues of B lie in D. For $\zeta \in \mathbb{C} \setminus D$, the matrix $\zeta I - B$ is thus invertible and, in view of Theorem 3.3.1,

$$(\zeta I - B)^{-1} = \sum_{n=0}^{\infty} \zeta^{-n-1} B^n.$$

It follows that

$$\|(\zeta I - B)^{-1}\| \leq \sum_{n=0}^{\infty} |\zeta|^{-n-1} \cdot \|B^n\| \leq M_0 \sum_{n=0}^{\infty} |\zeta|^{-n-1} = M_0 \cdot (|\zeta| - 1)^{-1}.$$

We see that (5.1.1) implies (5.1.2) with $M_1 = M_0$.

The Kreiss matrix theorem asserts that, conversely, (5.1.2) implies (5.1.1) with M_0 depending only on M_1 and on the dimension s, but otherwise independent of the matrix B.

The Kreiss 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, considered in the sixties, the matrices B are obtained by Fourier transformation of the numerical solution operators, and they stand essentially for the so-called amplification matrices. These matrices are of a *fixed* finite order s. On the other hand, the implication of (5.1.1) by (5.1.2) can also be used without Fourier transformation, with B standing for the numerical solution operator in (4.2.1). In this situation we may be dealing with a family of matrices B of finite – but *not uniformly* bounded – order s. Therefore, of particular interest is the dependence of the stability constant M_0 in (5.1.1) on the dimension s.

Since the work of Kreiss many authors studied the size of (the optimal) M_0 as a function of M_1 and s, and eventually in the ninetees some open problems in this field were settled. Moreover, the implication of (5.1.1) by (5.1.2) as discussed above was generalized in several directions. More general norms than the spectral norm were dealt with and the resolvent condition (5.1.2) was adapted to domains different from the unit disk D. In the following we shall discuss some of the results just mentioned as well as closely related ones.

5.2 Stability estimates for arbitrary $M_1 \geq 1$ and arbitrary norms

In this section we consider the relation between (5.1.1) and (5.1.2) for the case where $\|\cdot\|$ is a matrix norm on $\mathbb{C}^{s,s}$ induced by an arbitrary vector norm $|\cdot|$ on $\mathbb{C}^s$. Throughout this section $\|\cdot\|$ denotes such a matrix norm.

Theorem 5.2.1 *Let $s \geq 1$, $B \in \mathbb{C}^{s,s}$.*
(a) If (5.1.1) holds for some M_0, then (5.1.2) holds with $M_1 = M_0$;
(b) If (5.1.2) holds for some M_1, then

$$\|B^n\| \leq (1+1/n)^n \, \min\{n+1, s\} \, M_1 \leq \mathrm{e} \cdot \min\{n+1, s\} \, M_1 \quad \text{for} \;\; n = 1, 2, 3, \ldots . \tag{5.2.1}$$

Proof. 1. The proof of (a) is the same as the proof in Section 5.1 for the spectral norm.

In order to prove (b) we consider an arbitrary but fixed $n \geq 1$ and B satisfying (5.1.2). In view of Theorem 3.3.1(c) we can express the n-th power B^n as a Dunford-Taylor integral

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

where Γ is any positively oriented circle $|\zeta| = 1 + \varepsilon$ with $\varepsilon > 0$. Taking norms and applying (5.1.2) we obtain

$$\|B^n\| \, \leq \, \frac{1}{2\pi} \int_\Gamma (1+\varepsilon)^n M_1 \varepsilon^{-1} |d\zeta| = (1 + \varepsilon^{-1})(1+\varepsilon)^n M_1.$$

By choosing $\varepsilon = 1/n$ there follows

$$\|B^n\| \, \leq \, (1+1/n)^n \, (n+1) M_1.$$

Clearly, the proof of (b) is complete if we can show

$$\|B^n\| \leq (1+1/n)^n s M_1 \quad \text{for} \;\; n = 1, 2, 3, \ldots \; . \tag{5.2.2}$$

2. We can regard $\mathbb{C}^{s,s}$ as the space $\mathbb{C}^t$, with $t = s^2$, and the norm $\|\cdot\|$ as a vector norm on $\mathbb{C}^t$. Applying Lemma 3.2.1 (Lineair functionals), with $y = B^n$ and s replaced by t, we see that a linear $F : \mathbb{C}^{s,s} \to \mathbb{C}$ exists with

$$|F(A)| \leq \|A\| \quad \text{for all} \;\; s \times s \text{ matrices } A, \tag{5.2.3}$$

$$F(B^n) = \|B^n\|. \tag{5.2.4}$$

A combination of (5.2.4) and the above integral representation for B^n yields

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

where $R(\zeta) = F\big((\zeta I - B)^{-1}\big)$. Integrating by parts we obtain, with $\varepsilon = 1/n$,

$$\|B^n\| = \frac{-1}{2\pi \mathrm{i}(n+1)} \int_\Gamma \zeta^{n+1} R'(\zeta) d\zeta \, \leq \, \frac{1}{2\pi n} (1+1/n)^n \int_\Gamma |R'(\zeta)| \, |d\zeta|. \tag{5.2.5}$$

3. Let E_{jk} stand for the $s \times s$ matrix with entry in the j-th row and k-th 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 recall that a rational function is 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, and they have the same denominator. Hence $R(\zeta)$ is also a rational function of order s.

By Theorem 2.2.2 we have

$$\int_\Gamma |R'(\zeta)|\,|d\zeta| \le 2\pi s \max_\Gamma |R(\zeta)|. \tag{5.2.6}$$

The proof of (5.2.2) now easily follows by a combination of (5.2.5), (5.2.6), (5.2.3) and (5.1.2). □

In the following theorem we focus on the sharpness of the bound (5.2.1) in the case $n = s-1$.

Theorem 5.2.2 *Let $s \ge 2$. Then*

$$\sup\{\|B^{s-1}\|/M_1(B) :\ B \in \mathbb{C}^{s,s},\ M_1(B) < \infty\} = \big(1 + \frac{1}{s-1}\big)^{s-1} s, \tag{5.2.7}$$

where $M_1(B)$ denotes the smallest M_1 such that (5.1.2) holds (we define $M_1(B) = \infty$ if (5.1.2) 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 (4.3.3). We have

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

where

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

so that

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

It follows that the left-hand member of (5.2.7) is not smaller than the right-hand member. In view of (5.2.1) the proof is complete. □

Corollary 5.2.3 *For each $s \ge 1$, let an induced matrix norm $\|\cdot\|^{(s)}$ 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*

$$\|(B_s)^{s-1}\|^{(s)} \sim \mathrm{e} s M_1(B_s) \quad (\text{as } s \to \infty), \tag{5.2.8}$$

where $M_1(B_s)$ has the same meaning as in Theorem 5.2.2 (with $\|\cdot\| = \|\cdot\|^{(s)}$). □

Proof. Immediate from Theorem 5.2.2. □

In view of (5.2.1), the estimate

$$\|B^n\| \le \mathrm{e}\, s M_1 \quad \text{for } n = 0, 1, 2, \ldots \tag{5.2.9}$$

is valid for general induced matrix norms on $\mathbb{C}^{s,s}$. By virtue of Corollary 5.2.3, this estimate is sharp in the sense of (5.2.8). However, it should be emphasized that this does not resolve the sharpness question for given *fixed* M_1, since $M_1(B_s)$ in (5.2.8) may depend on s. In the next two sections we will focus on the situation where M_1 is a given fixed number.

5.3 Improved stability estimates for $M_1 = 1$

5.3.1 The case of arbitrary norms

In the special situation where the resolvent condition (5.1.2) holds with $M_1 = 1$, the upper bound (5.2.1) can be improved in various ways. First we concentrate on arbitrary induced norms on $\mathbb{C}^{s,s}$.

Theorem 5.3.1 *Let $s \ge 1$, $B \in \mathbb{C}^{s,s}$ and $\|\cdot\|$ denote an arbitrary induced matrix norm on $\mathbb{C}^{s,s}$. If (5.1.2) holds with $M_1 = 1$, then*

$$\|B^n\| \;\le\; n!\,(\mathrm{e}/n)^n \;\le\; \sqrt{2\pi(n+1)} \qquad \text{for } n = 1, 2, \ldots\ . \tag{5.3.1}$$

Proof. Property (5.1.2) with $M_1 = 1$ implies that condition (II)(of Section 3.5) is fulfilled by $A = B$ with $M = 1$, $W = \{\zeta : |\zeta| \le 1\}$. By virtue of Theorem 3.5.3 this A must satisfy the corresponding condition (III) as well, i.e.

$$\left\|\exp\left[t\,\mathrm{e}^{-\mathrm{i}\theta}(B - \mathrm{e}^{\mathrm{i}\theta} I)\right]\right\| \le M = 1 \quad \text{for all } \ t \ge 0 \ \text{ and real } \theta.$$

Consequently,

$$\|\exp\,(\zeta B)\| \le \exp\,(|\zeta|) \quad \text{for all complex } \zeta.$$

From the power series expansion for $\exp(A)$ (Example 3.3.3 in Section 3.3) with $A = \zeta B$, it can be seen that

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

where Γ is the positively oriented circle with radius n and center 0. Therefore $\|B^n\| \le n!\, n^{-n}\mathrm{e}^n$. In view of Stirling's formula (Theorem 2.1.2) we have

$$n! \le (n/\mathrm{e})^n \sqrt{2\pi n}\, \exp\,[(12n)^{-1}],$$

from which it can be deduced that $n!\,(e/n)^n \le \sqrt{2\pi(n+1)}$. □

As the next theorem shows, the upperbound for $\|B^n\|$ in (5.3.1) is not unnecessarily pessimistic.

Theorem 5.3.2 *Let $s \ge 2$, and let $B = (\beta_{ij})$ denote the $s \times s$ matrix with*

$$\beta_{ij} = 1 \ (\text{for } j = i - 1), \quad \text{and} \quad \beta_{ij} = 0 \ (\text{for } j \ne i - 1).$$

Then there is a vector norm in $\mathbb{C}^s$ such that, for the corresponding induced matrix norm $\|\cdot\|$ on $\mathbb{C}^{s,s}$, we have

(i) $\|(\zeta I - B)^{-1}\| \;\le\; (|\zeta| - 1)^{-1}$ *for all $\zeta \in \mathbb{C}$ with $|\zeta| > 1$;*

(ii) $\|B^n\| \;=\; n!\,(\mathrm{e}/n)^n \;\ge\; \sqrt{2\pi \cdot n}$ *for $n = 1, 2, \ldots, s - 1$.*

Proof. 1. For any column vector x in $\mathbb{C}^s$, we denote its components by $x_0, x_1, \ldots, x_{s-1}$, and we introduce the polynomial

$$x(t) = x_0 + x_1 t + \cdots + x_{s-1} t^{s-1}.$$

We define

$$|x| = \inf \left\{ \sum_{k=1}^{m} |\alpha_k| \, \mathrm{e}^{|\lambda_k|} \ : \ m \geq 1, \ \alpha_k \in \mathbb{C}, \ \lambda_k \in \mathbb{C} \ \text{ with } \ \sum_{k=1}^{m} \alpha_k \, \mathrm{e}^{\lambda_k B} = x(B) \right\}.$$

Note that the relation

$$\sum_{k=1}^{m} \alpha_k \, \mathrm{e}^{\lambda_k B} = x(B),$$

which occurs in our last definition, is valid if and only if $\alpha_1, \alpha_2, \ldots, \alpha_m$ satisfy the following system of linear equations:

$$\sum_{k=1}^{m} (\lambda_k)^i \cdot \alpha_k = i! \, x_i \qquad (0 \leq i \leq s-1).$$

From this we see that, for any given $x \in \mathbb{C}^s$ and $m \geq s$, it is possible to find λ_k, α_k such that the above relation is valid. Consequently, $|x|$ is a well defined real number.

In part 2 of the proof we show that $|x|$ is a norm for the vectors $x \in \mathbb{C}^s$.

2. a) Let $x, y \in \mathbb{C}^s$, and $\sum\limits_k \alpha_k \, \mathrm{e}^{\lambda_k B} = x(B)$, $\sum\limits_l \beta_l \, \mathrm{e}^{\mu_l B} = y(B)$. We have

$$\sum_k \alpha_k \, \mathrm{e}^{\lambda_k B} + \sum_l \beta_l \, \mathrm{e}^{\mu_l B} = (x+y)(B),$$

and therefore

$$|x+y| \ \leq \ \sum_k |\alpha_k| \, \mathrm{e}^{|\lambda_k|} + \sum_l |\beta_l| \, \mathrm{e}^{|\mu_l|}.$$

This implies that

$$|x+y| \ \leq \ |x| + |y|.$$

b) For $x \in \mathbb{C}^s, \lambda \in \mathbb{C}$ and any $\alpha_k, \lambda_k \in \mathbb{C}$ with $\sum\limits_k \alpha_k \, \mathrm{e}^{\lambda_k B} = x(B)$,

we have $\sum\limits_k (\lambda \alpha_k) \, \mathrm{e}^{\lambda_k B} = (\lambda x)(B)$, so that

$$|\lambda x| \ \leq \ \sum_k |\lambda \alpha_k| \, \mathrm{e}^{|\lambda_k|} \ = \ |\lambda| \cdot \sum_k |\alpha_k| \, \mathrm{e}^{|\lambda_k|}.$$

This implies that $|\lambda x| \leq |\lambda| \cdot |x|$. For $\lambda \neq 0$ we thus have $|x| = |\lambda^{-1}(\lambda x)| \leq |\lambda^{-1}| \cdot |\lambda x|$, and therefore $|\lambda x| \geq |\lambda| \cdot |x|$, which implies

$$|\lambda x| = |\lambda| \cdot |x|.$$

The last equality is also valid for $\lambda = 0$.

c) Suppose $|x| = 0$. For each $\varepsilon > 0$ there are α_k, λ_k with

$$\sum_k \alpha_k \, \mathrm{e}^{\lambda_k B} = x(B) \quad \text{and} \quad \sum_k |\alpha_k| \, \mathrm{e}^{|\lambda_k|} \ < \ \varepsilon.$$

For all such α_k, λ_k and $0 \leq i \leq s-1$ there follows

$$|x_i| = \left| \sum_k \alpha_k (\lambda_k)^i / i! \right| \ \leq \ \sum_k |\alpha_k| \frac{|\lambda_k|^i}{i!} \ \leq \ \sum_k |\alpha_k| \, \mathrm{e}^{|\lambda_k|} < \varepsilon.$$

Consequently, $|x_i| = 0$, and therefore $x = 0$.

d) We have proved that $|x|$ is a norm for $x \in \mathbb{C}^s$, and we denote the corresponding matrix norm on $\mathbb{C}^{s,s}$ by $\|\cdot\|$.

3. In order to prove (i) we consider arbitrary $x, y \in \mathbb{C}^s$ and $\zeta \in \mathbb{C}$ with

$$y = \mathrm{e}^{\zeta B} x.$$

We have

$$y_i = \sum_{j=0}^{i} \frac{\zeta^j}{j!}\, x_{i-j} \qquad (i = 0, 1, \ldots, s-1),$$

and therefore also

$$y(B) = \mathrm{e}^{\zeta B} x(B).$$

For any $\alpha_k, \lambda_k \in \mathbb{C}$ with $\sum_k \alpha_k \,\mathrm{e}^{\lambda_k B} = x(B)$ there follows $y(B) = \sum_k \alpha_k \,\mathrm{e}^{\zeta B}\,\mathrm{e}^{\lambda_k B}$. In view of Theorem 3.3.1(c) we obtain

$$y(B) = \sum_k \alpha_k \,\mathrm{e}^{(\zeta+\lambda_k)B}.$$

Consequently

$$|y| \;\le\; \sum_k |\alpha_k|\,\mathrm{e}^{|\zeta+\lambda_k|} \;\le\; \mathrm{e}^{|\zeta|} \sum_k |\alpha_k|\,\mathrm{e}^{|\lambda_k|}.$$

It follows that $|y| \le \mathrm{e}^{|\zeta|}|x|$. Hence

$$\|\mathrm{e}^{\zeta B}\| \le \mathrm{e}^{|\zeta|} \quad \text{for all } \zeta \in \mathbb{C}.$$

The matrix B thus satisfies condition III of Theorem 3.5.3 (with $A = B$, and with $M = 1$, $W = \{\zeta : |\zeta| \le 1\}$). By virtue of that theorem, condition II is satisfied as well. This proves (i).

4. Let $v \in \mathbb{C}^s$, with components $v_0 = 1$, $v_i = 0$ $(1 \le i \le s-1)$. Let $1 \le n \le s-1$ and define the vector $w \in \mathbb{C}^s$ by

$$w \;=\; B^n v.$$

We shall show that

$$|v| \le 1,$$

and

$$|w| \ge n!\,(\mathrm{e}/n)^n.$$

In view of the Theorems 5.3.1 and 2.1.2, the last two inequalities prove (ii).

The first of these inequalities follows from the fact that $\sum_{k=1}^{m} \alpha_k \,\mathrm{e}^{\lambda_k B} = v(B)$ holds with $m = 1$, $\alpha_1 = 1$, $\lambda_1 = 0$.

In order to prove the second inequality we consider arbitrary α_k, λ_k such that $\sum_k \alpha_k \,\mathrm{e}^{\lambda_k B} = w(B)$. Since the components of w are given by

$$w_i = 0 \;(0 \le i \le s-1,\; i \ne n), \qquad w_n = 1,$$

we have $w(B) = B^n$, and there follows

$$1 = \left| \sum_k \alpha_k \,\frac{(\lambda_k)^n}{n!} \right| \;\le\; \sum_k \sigma_k \,|\alpha_k|\,\mathrm{e}^{|\lambda_k|},$$

where

$$\sigma_k \;=\; \frac{|\lambda_k|^n\,\mathrm{e}^{-|\lambda_k|}}{n!} \;\le\; \frac{n^n\,\mathrm{e}^{-n}}{n!}.$$

Consequently,

$$n!\,(\mathrm{e}/n)^n \;\le\; \sum_k |\alpha_k|\,\mathrm{e}^{|\lambda_k|},$$

and therefore $|w| \ge n!\,(\mathrm{e}/n)^n$. □

5.3.2 Improvements of (5.3.1) for the case of special norms

According to the following theorem the stability estimate (5.3.1) can be improved substantially for the case of some important matrix norms.

Theorem 5.3.3 *Let $s \ge 1$, $Q \in \mathbb{C}^{s,s}$ 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 (5.1.2) with $M_1 = 1$ implies that*

$$\|B^n\| \;\le\; M_0 \qquad (\text{ for } n = 0, 1, 2, \ldots),$$

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 (5.1.2) with $\|\cdot\| = \|\cdot\|_\infty$, $M_1 = 1$. Clearly (II) (Section 3.5) holds with $W = \{\zeta : |\zeta| \le 1\}$, $M = 1$. By Theorem 3.5.3 we have $\tau_\infty[B, 1] \subset W$. In view of Theorem 3.5.5 the matrix B satisfies a circle condition with respect to the unit disk, which means that $\|B\|_\infty \le 1$. Therefore (5.1.1) holds with $M_0 = 1$.

For $p = 1$ the proof follows from the result for $p = \infty$ and the fact that $\|A\|_1 = \|A^*\|_\infty$ for all $A \in \mathbb{C}^{s,s}$.

For $p = 2$ the proof runs as follows. Similarly as above it can be seen that (5.1.2) implies $\tau_2[B, 1] \subset W = \{\zeta : |\zeta| \le 1\}$. In view of Theorem 3.5.4 we thus have

$$\{x^*Bx :\; x \in \mathbb{C}^s \text{ with } x^*x = 1\} \;\subset\; W.$$

The proof continues by applying Berger's inequality. This inequality reads

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

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 $r(B) \le 1$, there follows

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

We split B^n into a sum $B^n = A_1 + \mathrm{i}A_2$ with Hermitian matrices $A_1 = \frac{1}{2}(A + A^*)$ and $A_2 = \frac{1}{2\mathrm{i}}(A - A^*)$. By noting that for any Hermitian A the relation $\|A\|_2 = r(A)$ is valid, we finally obtain

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

5.4 The best stability estimates for fixed $M_1 > 1$

Theorem 5.2.1 shows that if the resolvent condition (5.1.2) is satisfied with a fixed M_1, then $||B^n||$ can grow at most linearly with n or s. Corollary 5.2.3 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 the Theorems 5.3.1 and 5.3.3 in the previous section.

For other fixed values $M_1 > 1$, also, the question arises of whether the upper bound (5.2.1) can be improved. According to Theorem 5.3.2, a growth of $\|B^n\|$ at the rate $\sqrt{n}$ or $\sqrt{s}$ can occur if arbitrary norms are considered. But, that theorem is not relevant to the important norms $\|\cdot\|_p$, with $p = 1, 2, \infty$. In the following we present two results for these norms. The first result shows, for $p = 1, \infty$, that growth at the rate n and s can occur, whereas the second result establishes, for $p = 2$, a growth which is almost at the rate $\sqrt{\log n}$ and $\sqrt{\log s}$.

Theorem 5.4.1 *Let $M_1 \geq \pi + 1$, and $p = 1$ or $p = \infty$. Let s be a given integer, with $s \geq 1$. Then there is an $s \times s$ matrix B such that (5.1.2) is satisfied with $\|\cdot\| = \|\cdot\|_p$, and*

$$\|B^n\|_p \;\geq\; n \qquad \text{for } \; n = 1, 2, \ldots, s.$$

Theorem 5.4.2 *Let $M_1 > \pi + 1$ 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 (5.1.2) with $\|\cdot\| = \|\cdot\|_2$, and*

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

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

$$(E_{s,j})^2 = E_{s,j} \neq O, \quad E_{s,j}E_{s,k} = O \;(j \neq k), \quad \sum_{j=1}^{s} E_{s,j} = I; \tag{5.4.1}$$

$$\Big\| \sum_{j \;\; odd} E_{s,j} \Big\|_2 \geq \gamma(\log s)^{1/2}/\log\log s; \tag{5.4.2}$$

$$B_s = \sum_{j=1}^{s} \mathrm{e}^{2\pi \mathrm{i} j/s} E_{s,j} \text{ satisfies (5.1.2)} \; . \tag{5.4.3}$$

For even s we have

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

In view of (5.4.2) this implies

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

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

5.5 Notes and remarks

The Kreiss matrix theorem was originally published in Kreiss (1962). Subsequent elaborations and applications to so-called amplification matrices, originating from Fourier transformation, can be found e.g. in Richtmyer & Morton (1967), Strikwerda (1989).

As already mentioned in Section 5.1, the Kreiss matrix theorem asserts, for the spectral norm, that the resolvent condition (5.1.2) implies power boundedness (5.1.1) 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 (cf. Morton (1964) and Miller & Strang (1966)), it was Tadmor (1981) who succeeded in proving a bound that is linear in s,

$$\|B^n\| \leq 32\mathrm{e}\pi^{-1} s M_1.$$

LeVeque & Trefethen (1984) lowered this upper bound to $2\mathrm{e}sM_1$, and conjectured that the latter bound can be improved further to (5.2.9) (with $\|\cdot\|$ still standing for the spectral norm). Moreover, these authors showed by means of a counterexample that the factor e in (5.2.9) cannot be replaced by any smaller constant if the upper bound should be valid for arbitrary factors M_1 in (5.1.2) and arbitrarily large dimensions s.

Smith (1985) proved a result which, combined with the arguments of LeVeque & Trefethen (1984), leads to the bound $\|B^n\| \leq \pi^{-1}(\pi+2)\mathrm{e}sM_1$, which is an improvement over the upper bound $2\mathrm{e}sM_1$ but still weaker than (5.2.9). Eventually (5.2.9) was proved to be true (in Spijker (1991)). For an interesting historical survey see Wegert & Trefethen (1994).

The proof in Section 5.2 of (5.2.1) has been taken from Dorsselaer, Kraaijevanger & Spijker (1993), and is partly based on arguments used earlier by Lenferink & Spijker (1991a,b), Lubich & Nevanlinna (1991), Reddy & Trefethen (1990). The crucial idea to use in the proof a relation of the form (5.2.5) and to bound the integral $\int_\Gamma |R'(\zeta)|\,|d\zeta|$ in terms of $\max_\Gamma |R(\zeta)|$ was used earlier by LeVeque & Trefethen (1984).

Corollary 5.2.3 was proved by LeVeque & Trefethen (1984) for the spectral norm. The proof of Theorem 5.2.2 has been taken from Dorsselaer, Kraaijevanger & Spijker (1993).

The proof of (5.3.1) in Section 5.3.1 is essentially based on ideas taken from Bonsall & Duncan (1980)(see also Bonsall & Duncan (1971)). Another proof can be given along the lines of Lubich & Nevanlinna (1991) (Theorem 2.1).

Our proof of Theorem 5.3.2 strongly relies on ideas taken from Crabb (1970).

The value $M_0 = 2$ (for $p = 2$) in Theorem 5.3.3 has been known already for some time and was stated e.g. in Reddy & Trefethen (1992). For the inequality of Berger, used in the proof of Theorem 5.3.3, we refer to Bonsall & Duncan (1980), Horn & Johnson (1990), Pearcy (1966), Richtmyer & Morton (1967, p.89).

Theorem 5.4.1 follows easily from a clever counterexample presented in Kraaijevanger (1994).

Theorem 5.4.2 and its proof have been taken from Dorsselaer, Kraaijevanger & Spijker (1993).

6 Stability estimates under resolvent conditions on hA

6.1 Linear stability analysis and stability regions

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

$$\begin{cases} U'(t) &= AU(t) + r(t) \qquad (t \geq 0), \\ U(0) &= u_0. \end{cases} \tag{6.1.1}$$

Here A is a given constant $s \times s$ matrix, and u_0, $r(t)$ 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 be relevant also 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 (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 can arrive at an initial value problem for a large system of the form (6.1.1). In this case the matrix A, the inhomogeneous term $r(t)$, 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 (6.1.1) then provides an approximation to the solution of the original initial-boundary value problem. For examples we refer to the Sections 1.2, 3.6, 4.2, where semi-discretization by *finite-difference* methods is dealt with .

Many step-by-step methods for the numerical solution of ordinary differential equations, like Runge-Kutta methods or Rosenbrock methods reduce — when applied to (6.1.1) — to processes of the form

$$u_n = \varphi(hA)u_{n-1} + r_n \quad \text{for } n = 1, 2, 3, \ldots \ . \tag{6.1.2}$$

Here $\varphi(z) = P(z)/Q(z)$ is a rational function, depending only on the underlying step-by-step method, and $P(z)$, $Q(z)$ are polynomials, without common zeros, such that $\varphi(0) = \varphi'(0) = 1$. Further, $h = \Delta t > 0$ is the stepsize, and $\varphi(hA) = P(hA)[Q(hA)]^{-1}$ (when $Q(h\lambda) \neq 0$ for all $\lambda \in \sigma[A]$, cf. Example 3.3.2). The vectors $r_n \in \mathbb{C}^s$ are related to $r(t)$, and $u_n \simeq U(nh)$ are computed successively from (6.1.2). 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 (6.1.2), and can a posteriori be conceived as relying on semi-discretization.

An example of (6.1.2) is provided by the, fully discrete, numerical process constructed in Section 4.2 for the solution of (4.2.2). From (4.2.3) we see that this process is of the form (6.1.2) with

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

and the tridiagonal matrix $A = (\alpha_{jk})$ given by (4.2.4).

Since (6.1.2) is a special case of (4.2.1) the stability analysis of (6.1.2) amounts, as explained at the end of Section 4.2, to investigating the growth of the matrices B^n specified by

$$B = \varphi(hA). \tag{6.1.3}$$

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

$$S = \{z : \ z \in \mathbb{C} \text{ with } Q(z) \neq 0 \text{ and } |\varphi(z)| \leq 1\}. \tag{6.1.4}$$

The following theorem is a variant to Theorem 4.3.1. It is easier to apply than the latter, since it gives a stability criterion in terms of $\varphi(z)$ and hA rather than in terms of the (more complicated) matrix B itself.

Theorem 6.1.1 (*The eigenvalue criterion*) *Let $||\cdot||$ be an arbitrary induced norm on $\mathbb{C}^{s,s}$, and $B = \varphi(hA)$. Then there is an M_0 such that (5.1.1) holds, if and only if*

$$\text{(6.1.5)} \qquad \begin{cases} \text{all eigenvalues of } hA \text{ belong to } S; \text{ and whenever } J_k \text{ is a Jordan block} \\ \text{(with order } s_k > 1) \text{ of } hA \text{ corresponding to an eigenvalue } h\lambda_k \in \partial S, \\ \text{then the derivatives } \varphi^{(j)}(h\lambda_k) \text{ vanish for } j = 1, 2, \ldots, s_k - 1. \end{cases}$$

Proof. Applying Theorem 3.3.6 (with f, A replaced by φ, hA) we see that

$$B^n = \varphi(hA)^n = \sum_{k=1}^{r} (T_k)^n,$$

where

$$T_k = \varphi(h\lambda_k)P_k + Q_k, \quad Q_k = \sum_{j=1}^{s_k-1} \frac{\varphi^{(j)}(h\lambda_k)}{j!}(R_k)^j, \quad \lambda_k \in \sigma[A], \quad s_k \geq 1.$$

There is an M_0 with (5.1.1) if and only if, for each k, the powers $(T_k)^n$ remain bounded for $n \to \infty$.

Let k be given. From the above expressions for T_k, Q_k we see that

$$(T_k)^n = [\varphi(h\lambda_k)P_k + Q_k]^n = \sum_{p=0}^{m_k} \binom{n}{p} \varphi(h\lambda_k)^{n-p}\, Q_k^p\, P_k,$$

where $m_k = \min(n, s_k - 1)$. Moreover,

$$Q_k = \mathrm{O} \quad \text{if and only if} \quad \varphi^{(j)}(h\lambda_k) = 0 \quad \text{for} \quad 1 \leq j \leq s_k - 1.$$

The proof of the theorem is completed by noting that the following four implications are valid.

(i) $|\varphi(h\lambda_k)| < 1 \;\Rightarrow\; (T_k)^n$ remains bounded for $n \to \infty$;

(ii) $|\varphi(h\lambda_k)| = 1$ and $Q_k = \mathrm{O} \;\Rightarrow\; (T_k)^n$ remains bounded for $n \to \infty$;

(iii) $|\varphi(h\lambda_k)| = 1$ and $Q_k \neq \mathrm{O} \;\Rightarrow\; (T_k)^n$ does not stay bounded for $n \to \infty$;

(iv) $|\varphi(h\lambda_k)| > 1 \;\Rightarrow\; (T_k)^n$ does not stay bounded for $n \to \infty$. □

Many functions $\varphi(z)$ of practical interest have non-vanishing derivatives $\varphi'(z)$ on the whole of ∂S. In this case (6.1.5) simply reduces to $\sigma[hA] \subset S$ and the condition that the Jordan blocks J_k of hA have order 1 whenever they correspond to an eigenvalue $h\lambda_k \in \partial S$.

Clearly, condition (6.1.5) always implies that

$$\text{(6.1.6)} \qquad \sigma[hA] \subset S.$$

Moreover, from Theorem 3.1.4 we see that (6.1.6) is equivalent to (6.1.5) if the matrix A is normal.

The analogue of Theorem 4.3.2, in the situation where $B = \varphi(hA)$, is as follows.

Theorem 6.1.2 (*The eigenvalue criterion for normal matrices*) *Let $\|\cdot\|$ denote the spectral norm, and assume $B = \varphi(hA)$ where A is normal. Then the following three statements are equivalent to each other.*

(i) *There is an M_0 such that (5.1.1) holds.*

(ii) $\|B^n\| \leq 1$ *for* $n = 1, 2, 3, \ldots$.

(iii) $\sigma[hA] \subset S$.

Proof. Suppose (5.1.1) holds for some constant M_0. Then, in view of Theorem 6.1.1, condition (6.1.6) is fulfilled.

Suppose (6.1.6) holds. By Theorem 3.1.4, all Jordan blocks of hA have an order 1, and Theorem 3.3.6 thus yields the representation

$$B = \varphi(hA) = \sum_{k=1}^{r} \varphi(h\lambda_k)P_k = T \text{ diag } [\varphi(h\lambda_1), \varphi(h\lambda_2), \ldots, \varphi(h\lambda_s)]T^{-1}.$$

Consequently $\|B^n\| \leq \|T\| \cdot \|T^{-1}\| = 1$ as in the proof of Theorem 4.3.2. □

In general (6.1.5) has similar advantages and disadvantages as the eigenvalue condition (4.3.1). It is relatively simple to verify, and reliable in the situation of normal matrices, for which it reduces to (6.1.6). But, it is unreliable for (families of) matrices that are not normal.

In the rest of this chapter we adapt (6.1.5) 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 Chapter 5) lies in the circumstance that, in general, hA has simpler a structure than B, and that knowledge available about S can be exploited.

The conditions on hA we shall deal with, are all of the following general form,

(6.1.7) $(zI - hA)$ is regular and $\|(zI - hA)^{-1}\| \leq K \cdot \mathrm{d}(z, V)^{-1}$ for all complex numbers $z \in \mathbb{C} \setminus V$.

Here K is a constant, V a closed subset of $\mathbb{C}$, $\|\cdot\|$ denotes an arbitrary induced norm on $\mathbb{C}^{s,s}$, and $\mathrm{d}(z, V)$ is the distance from z to V.

In case V is bounded, we easily see from (6.1.7), by letting $z \to \infty$, that

(6.1.8*a*) $$K \geq 1.$$

In the following we shall always assume that this inequality is fulfilled.

We further note that if

(6.1.8*b*) V is a closed subset of S,

condition (6.1.7) implies (6.1.5). We shall focus on the situation where (6.1.8b) is fulfilled.

In the following sections, condition (6.1.7) will be seen to imply stability estimates of the form

(6.1.9) $$\|\varphi(hA)^n\| \leq K \cdot \Phi(n, s) \quad \text{for } n = 1, 2, 3, \ldots,$$

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

6.2 Stability estimates which grow linearly with n and s

6.2.1. Arbitrary subsets V of the stability region

The following general theorem is valid.

Theorem 6.2.1 *There is a constant γ such that the stability estimate (6.1.9) holds, with*

$$\Phi(n, s) = \gamma \cdot \min\{n, s\},$$

whenever K, V satisfy (6.1.8) and $hA \in \mathbb{C}^{s,s}$ satisfies condition (6.1.7). Here γ only depends on the rational function $\varphi(z)$ (and not on V, $n \geq 1$, $s \geq 1$, $hA \in \mathbb{C}^{s,s}$, $\|\cdot\|$ or K).

This theorem will be proved below for the special case where

(6.2.1a) $\varphi'(z) \neq 0$ on ∂S,
and
(6.2.1b) S is bounded.

The proof will consist of two steps. First, it will be shown that B, given by (6.1.3), satisfies a resolvent condition (5.1.2) (with the same norm $\|\cdot\|$ as appearing in (6.1.7), (6.1.9)). A subsequent application of Theorem 5.2.1 (part (b)) will lead to (6.1.9) with $\Phi(n,s)$ as specified in Theorem 6.2.1.

In proving that B satisfies (5.1.2) we shall make use of two lemmas, the first of which reads as follows.

Lemma 6.2.2 *Let $\rho > 1$, $M > 0$, $B \in \mathbb{C}^{s,s}$ and $\|\cdot\|$ an induced norm on $\mathbb{C}^{s,s}$. Assume*

(6.2.2a) *$(\zeta I - B)$ is regular for all ζ with $|\zeta| > 1$, and*

(6.2.2b) *$\|(\zeta I - B)^{-1}\| \leq M \cdot (|\zeta| - 1)^{-1}$ for all ζ with $1 < |\zeta| < \rho$.*

Then (5.1.2) is satisfied with

$$M_1 = \frac{\sqrt{\rho}+1}{\sqrt{\rho}-1} \cdot M.$$

Proof. For any ζ with $|\zeta| \geq \rho$ we can write $(\zeta I - B)^{-1}$ as a Dunford-Taylor integral,

$$(\zeta I - B)^{-1} = \frac{1}{2\pi \mathrm{i}} \int_\Gamma f(z)(zI - B)^{-1} dz,$$

where $f(z) = (\zeta - z)^{-1}$ and Γ is any positively oriented circle $\Gamma[0, \sigma]$ with $1 < \sigma < \rho$. Consequently, for such ζ, we obtain

$$\|(\zeta I - B)^{-1}\| \;\leq\; (2\pi)^{-1} \int_\Gamma |\zeta - z|^{-1} \cdot M \cdot (|z| - 1)^{-1} |dz| \;\leq\; \frac{M \cdot \sigma}{(|\zeta| - \sigma)(\sigma - 1)}.$$

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

$$(|\zeta| - 1)\, \|(\zeta I - B)^{-1}\| \;\leq\; \frac{M \cdot \sigma}{\sigma - 1}\Big(1 + \frac{\sigma - 1}{|\zeta| - \sigma}\Big) \;\leq\; \frac{\sigma + 1}{\sigma - 1} \cdot M.$$

Since $M_1 \geq M$ the proof is complete. □

For $\alpha > 0$, $\beta > 0$ we shall use, in the following, the notations

$$D_\alpha = \{\zeta :\ \zeta \in \mathbb{C} \text{ with } 1 < |\zeta| \leq 1 + \alpha\},$$
$$S_\beta = \{z :\ z \in \mathbb{C} \text{ with } 0 < \mathrm{d}(z, S) \leq \beta\}.$$

For given $\zeta \in \mathbb{C}$ we shall deal with the rational function

(6.2.3a) $$\psi_\zeta(z) = [\zeta - \varphi(z)]^{-1}.$$

It is easily verified that the order of $\psi_\zeta(z)$ does not exceed the order, say r, of $\varphi(z)$. Therefore the number of different poles of $\psi_\zeta(z)$, denoted by $q(\zeta)$, satisfies

(6.2.3*b*) $$q(\zeta) \le r.$$

We denote the poles of $\psi_\zeta(z)$, and their corresponding orders, by

(6.2.3*c*) $$z_j(\zeta) \quad \text{and} \quad k_j(\zeta) \quad (\text{ for } \ j = 1, 2, \dots, q(\zeta)),$$

respectively. Finally, we use the notation

(6.2.3*d*) $$\varphi(\infty) = \lim_{z\to\infty} \varphi(z).$$

Our second lemma only concerns $\varphi(z)$ (and not the matrix B given by (6.1.3)):

Lemma 6.2.3 *Let $\varphi(z)$ satisfy (6.2.1a,b). Then there are positive α, β with the following properties.*

(i) $|\varphi(\infty)| > 1 + \alpha$;

(ii) *$\varphi(z)$ is regular, and $\varphi'(z) \neq 0$ at all $z \in S_\beta$;*

(iii) *if $\zeta \in D_\alpha$, then $z_j(\zeta) \in S_\beta$ (for $j = 1, 2, \dots, q(\zeta)$);*

(iv) *if $\zeta \in D_\alpha$, then* $$\psi_\zeta(z) \ = \ [\zeta - \varphi(\infty)]^{-1} - \sum_{j=1}^{q(\zeta)} \big[\varphi'(z_j(\zeta))\big]^{-1} \big(z - z_j(\zeta)\big)^{-1};$$

(v) *if $\zeta \in D_\alpha$, then $|\zeta| - 1 \le \mu_1 \cdot \mathrm{d}(z_j(\zeta), S)$ (for $j = 1, 2, \cdots, q(\zeta)$), with $\mu_1 = \max\{|\varphi'(z)| : \ z \in \mathrm{cl}\ (S_\beta)\} < \infty$.*

Proof. 1. Since S is bounded, $\varphi(\infty)$ satisfies $1 \le |\varphi(\infty)| \le \infty$. In case $|\varphi(\infty)| > 1$, it is clear that (i) holds for all $\alpha > 0$ sufficiently small.

Suppose $\varphi(\infty) = \eta$ with $|\eta| = 1$. Then we can write

$$P(z) = \alpha_r z^r + \cdots + \alpha_1 z + \alpha_0, \qquad Q(z) = \beta_r z^r + \cdots + \beta_1 z + \beta_0,$$

where $\alpha_r = \eta \cdot \beta_r \neq 0$. Therefore, for $w \in \mathbb{C}$ with $|w| > 0$ sufficiently small, we have a representation for $\varphi(1/w)$ of the form

$$\varphi(1/w) = \eta \cdot f(w) \ \text{ with } \ f(w) = \frac{1 + \gamma_1 w + \cdots + \gamma_r w^r}{1 + \delta_1 w + \cdots + \delta_r w^r}.$$

Since

$$f(w) = 1 + \varepsilon_1 \cdot w^k + \mathcal{O}(w^2) \quad (\text{ for } w \to 0),$$

with $\varepsilon_1 \neq 0$, $k \ge 1$, it follows that there exist w, arbitrarily close to 0, with $|f(w)| < 1$. This means that $\varphi(1/w)$ assumes values with modulus less than 1 for w arbitrarily close to 0. This contradicts (6.2.1b).

We conclude that (i) holds for all $\alpha > 0$ that are sufficiently small.

2. Denote the set of all poles of $\varphi(z)$ and zeros of $\varphi'(z)$ by T. By (6.1.4), (6.2.1a) the sets ∂S and T have no points in common. Since ∂S is closed and T finite, there exists a $\beta > 0$ such that all points in T have a distance to ∂S greater than β.

Let $z \in S_\beta$. Since $\mathrm{d}(z, \partial S) = \mathrm{d}(z, S)$ it follows that $0 < \mathrm{d}(z, \partial S) \le \beta$. Consequently z does not belong to T. Part (ii) has been proved.

3. We define

$$\sigma = \inf\ \{|\varphi(z)| : \ z \in \mathbb{C} \text{ with } \mathrm{d}(z, S) > \beta\}.$$

First we assume $\sigma > 1$. We choose α satisfying (i) with $0 < \alpha < \sigma - 1$. Let ζ be given, with $\zeta \in D_\alpha$. We then have $1 < |\zeta| < \sigma$. The definition of σ and (6.1.4) imply that all z with $\varphi(z) = \zeta$ must belong to S_β. Hence $z_j(\zeta) \in S_\beta$ for $j = 1, 2, \ldots, q$.

Next we assume $\sigma \leq 1$, which will be shown to lead to a contradiction. The definition of σ implies that there is a sequence of complex numbers $y_1, y_2, y_3, \ldots$ with

$$d(y_k, S) > \beta \quad \text{and} \quad |\varphi(y_k)| \leq 1 + 1/k \quad (\text{ for } \ k = 1, 2, 3, \ldots).$$

From (i) we see that positive numbers ρ and ε exist such that $|\varphi(z)| \geq 1 + \varepsilon$ for all complex z not belonging to the disk $D[0, \rho]$. Hence

$$y_k \in D[0, \rho] \quad \text{for all } k \text{ that are sufficiently large .}$$

Since $D[0, \rho]$ is compact, there is a convergent subsequence of $\{y_k\}$, say $\{y_{k(j)}\}$ with limit y_∞.

Since $\mathrm{d}(y_{k(j)}, S) > \beta$, we have $\mathrm{d}(y_\infty, S) \geq \beta$. But, since $\varphi(y_{k(j)})$ tends to $\varphi(y_\infty)$ (for $j \to \infty$), we also have $|\varphi(y_\infty)| \leq 1$. The last inequality means that $\mathrm{d}(y_\infty, S) = 0$, and we have a contradiction.

4. Let $\zeta \in D_\alpha$. In view of (ii), (iii) we see that at all poles $z_j(\zeta)$ we have $\varphi(z_j(\zeta)) \neq \infty$ and $\varphi'(z_j(\zeta)) \neq 0$. From (6.2.3a) it follows that the principal part of $\psi_\zeta(z)$ at $z_j(\zeta)$ equals

$$-\big[\varphi'(z_j(\zeta)\big]^{-1}\big(z - z_j(\zeta)\big)^{-1}$$

Subtracting from $\psi_\zeta(z)$ all its principal parts we obtain a function that we denote by $\omega_\zeta(z)$. In view of Theorem 2.1.1 the function $\omega_\zeta(z)$ is a polynomial in the variable z. Since $\psi_\zeta(z)$ tends to $[\zeta - \varphi(\infty)]^{-1}$ when $z \to \infty$, we can conclude that

$$\lim_{z \to \infty} \omega_\zeta(z) = \big[\zeta - \varphi(\infty)\big]^{-1}.$$

Clearly, $\omega_\zeta(z)$ is a polynomial of degree zero, and (iv) thus holds.

5. Let $\zeta \in D_\alpha$ and $y \in \partial S$. We have

$$|\zeta| - 1 \ \leq \ |\varphi(z_j(\zeta))| - |\varphi(y)| \ \leq \ |\varphi(z_j(\zeta)) - \varphi(y)|.$$

Denote the straight line segment connecting $z_j(\zeta)$ to $y \in \partial S$ with $|z_j(\zeta) - y| = \mathrm{d}(z_j(\zeta), S)$ by L. Since L is contained in $\mathrm{cl}\,(S_\beta)$ we have

$$|\varphi(z_j(\zeta)) - \varphi(y)| \ \leq \ \mu_1 \cdot |z_j(\zeta) - y| \ = \ \mu_1 \cdot d(z_j(\zeta), S),$$

with μ_1 as in statement (v). This completes the proof of the lemma. □

We now complete the proof of Theorem 6.2.1, making use of the Lemmas 6.2.2 and 6.2.3.

Proof of Theorem 6.2.1 for the case (6.2.1a,b). Assume (6.2.1), (6.1.7), (6.1.8). The matrix $B = \varphi(hA)$ exists (in the sense of Example 3.3.2) since $\sigma[hA] \subset V \subset S$ and the denominator $Q(z)$ of $\varphi(z)$ does not vanish on S. Moreover $(\zeta I - B)$ is regular for all ζ with $|\zeta| > 1$. This follows from the fact that $\sigma[B] = \varphi(\sigma[hA]) \subset \{\zeta : \ |\zeta| \leq 1\}$ (by Theorem 3.3.1 (d) and (6.1.4)).

Choose α, β according to Lemma 6.2.3. Let $\zeta \in D_\alpha$. In view of Lemma 6.2.3 (iv) and Example 3.3.2 we have

$$\big(\zeta I - B\big)^{-1} \ = \ \big[\zeta - \varphi(\infty)\big]^{-1} I - \sum_{j=1}^{q(\zeta)} \big[\varphi'(z_j(\zeta))\big]^{-1}\big(hA - z_j(\zeta)I\big)^{-1}.$$

By taking norms we obtain

$$\big\|\big(\zeta I - B\big)^{-1}\big\| \;\le\; \frac{\alpha}{|\varphi(\infty)| - (1+\alpha)} \cdot \big(|\zeta| - 1\big)^{-1} + \frac{1}{\mu_0} \sum_{j=1}^{q(\zeta)} \big\|\big(z_j(\zeta) I - hA\big)^{-1}\big\|$$

where

$$\mu_0 = \min\,\{|\varphi'(z)| :\; z \in \text{ cl }(S_\beta)\}$$

is positive by Lemma 6.2.3 (ii) and (6.2.1a).

The assumptions (6.1.7), (6.1.8b), in combination with Lemma 6.2.3 (v), yield

$$\big\|\big(z_j(\zeta) I - hA\big)^{-1}\big\| \le K \cdot d\big(z_j(\zeta), V\big)^{-1} \le K \cdot d\big(z_j(\zeta), S\big)^{-1} \le K\mu_1 \cdot \big(|\zeta| - 1\big)^{-1}.$$

In view of (6.2.3b) we thus obtain

$$\big\|\big(\zeta I - B\big)^{-1}\big\| \;\le\; M \cdot \big(|\zeta| - 1\big)^{-1},$$

with

$$M \;=\; \frac{\alpha}{|\varphi(\infty)| - (1+\alpha)} \;+\; rK \cdot \frac{\mu_1}{\mu_0}.$$

Clearly, the assumptions of Lemma 6.2.2 are fulfilled with $\rho = 1 + \alpha$. Therefore, our matrix B satisfies (5.1.2) with

$$M_1 = \frac{\sqrt{1+\alpha} + 1}{\sqrt{1+\alpha} - 1} \left[\big(|\varphi(\infty)| - 1 - \alpha \big)^{-1} \alpha \;+\; r\mu_1\mu_0^{-1} \cdot K \right].$$

In view of (6.1.8a), the matrix B satisfies (5.1.2) with

$$M_1 = \gamma_1 \cdot K,$$

where γ_1 only depends on the rational function $\varphi(z)$.

An application of Theorem 5.2.1 (b) yields

$$\|B^n\| \;\le\; (1 + 1/n)^n \min\{n+1, s\}\gamma_1 K \;\le\; 2\mathrm{e}\gamma_1 \cdot K \cdot \min\{n, s\}.$$

Theorem 6.2.1 has thus been proved, under the assumption (6.2.1a,b), with $\gamma = 2\mathrm{e}\gamma_1$. □

6.2.2 An example in which V is a disk and S is bounded

Let $\|\cdot\|$ denote an arbitrary induced norm on $\mathbb{C}^{s,s}$, and suppose $A \in \mathbb{C}^{s,s}$ satisfies the circle condition

$$\|A - \gamma I\| \le \rho. \tag{6.2.4}$$

In applications of Theorem 6.2.1 the following lemma will be helpful.

Lemma 6.2.4 *Assume (6.2.4). Then, for any $h > 0$, the matrix hA satisfies the resolvent condition (6.1.7) with*

$$V = D[h\gamma, h\rho] \quad \text{and} \quad K = 1.$$

Proof. From (6.2.4) it follows that

$$\|hA - h\gamma I\| \leq h\rho.$$

Denoting the 1-numerical range of hA by $\tau[hA]$ (see Definition 3.5.1) we thus have

$$\tau[hA] \ \subset \ D[h\gamma, h\rho].$$

An application of Theorem 3.5.3 (with hA replacing A, and with $M = 1,\ W = D[h\gamma, h\rho]$) completes the proof. □

We shall illustrate Theorem 6.2.1 in the numerical solution of the diffusion-convection-reaction problem (4.2.2). We define the tridiagonal $s \times s$ matrix A by (4.2.4). With regard to the coefficients $a(x)$, $b(x)$, $c(x)$ we make the assumption (3.6.1).

In Section 3.6 we showed that the matrix A satisfies a circle condition (6.2.4) with $\|\cdot\| = \|\cdot\|_\infty$ and

$$\gamma = -\frac{\alpha}{2}, \quad \rho = \frac{\alpha}{2}, \quad \alpha = 4\delta^{-2}|a|_\infty + 2\delta^{-1}|b|_\infty + |c|_\infty.$$

By virtue of Lemma 6.2.4 the matrix hA satisfies the condition (6.1.7) with

(6.2.5)
$$\|\cdot\| = \|\cdot\|_\infty, \quad K = 1 \ \text{ and}$$
$$V = D[-r_0, r_0], \ \text{where} \ \ r_0 = \frac{2h}{\delta^2}|a|_\infty + \frac{h}{\delta}|b|_\infty + \frac{h}{2}|c|_\infty.$$

We consider the, fully discrete, process of Section 4.2 with a matrix B given by (4.2.3), (4.2.4). This amounts to (6.1.2) with $\varphi(z) = (1 + (1-\theta)z)\,(1 - \theta z)^{-1}$. We consider a fixed θ with $0 \leq \theta < \frac{1}{2}$. The corresponding stability region $S = S_\theta$ is given by the formula

$$S_\theta = D[-r, r] \ \text{ with } \ r = \frac{1}{1-2\theta} \quad \text{ for } \ 0 \leq \theta < 1/2.$$

Clearly, all of the assumptions of Theorem 6.2.1 will be fulfilled as soon as $r_0 \leq r$, i.e.

(6.2.6)
$$\frac{2h}{\delta^2}|a|_\infty + \frac{h}{\delta}|b|_\infty + \frac{h}{2}|c|_\infty \leq \frac{1}{1-2\theta}.$$

Under this condition we thus have stability, (6.1.8) with Φ as in Theorem 6.2.1.

As a numerical illustration we consider problem (4.2.2) with

$$a(x) \equiv 1, \qquad b(x) \equiv -10^3, \qquad c(z) \equiv 0,$$

and the corresponding numerical process with

$$\theta = \frac{1}{4}, \qquad \delta = \frac{1}{s} = \frac{1}{20}\ .$$

In this situation the sufficient condition for stability (6.2.6) amounts to the stepsize restriction

$$h \leq h_0 = \frac{1}{10\,400} \simeq 9.62 \times 10^{-5}.$$

Computer experiments show that with a stepsize $h = 9 \times 10^{-5}$ one has

$$\|\varphi(hA)^n\| \leq 2.4 \quad \text{ for } \ n = 0, 1, 2, \ldots.$$

The last inequality expresses a stable behaviour, which is amply in agreement with Theorem 6.2.1.

It is instructive to compare the above stepsize restriction based on Theorem 6.2.1 with a naive use of the eigenvalue condition as formulated in Theorem 6.1.1. In fact, for $a(x)$, $b(x)$, $c(x)$ and δ as specified above all eigenvalues λ of A can be shown to be different from each other and to satisfy

$$-26\,514 < \lambda < 0.$$

For $\theta = 1/4$ we have $S = D[-2, 2]$, so that the eigenvalue condition (6.1.5) is satisfied for all $h > 0$ with

$$h \leq h_1 = \frac{4}{26\,514} \simeq 15.1 \times 10^{-5}.$$

Computer experiments show that with a stepsize $h = 15 \times 10^{-5}$ one has

$$\max_{n \geq 0} \|\varphi(hA)^n\| \simeq 2.7 \times 10^{12}.$$

Hence from a practical point of view there is mere instability with this h, although $h < h_1$ and (6.1.5) is fulfilled.

6.2.3 An example in which $V = \mathbb{C}_- \subset S$

In the example of Section 6.2.2 we have seen that, in the situation (6.2.5), condition (6.1.8b) boils down to a stepsize restriction of the form $h \leq h_0$, with finite h_0. This is related to the fact that the stability region S, dealt with in the example of Section 6.2.2, is bounded. In cases where S is unbounded it may be possible to establish stability estimates which are valid for *all* $h > 0$.

In order to illustrate this point, we consider the tridiagonal matrix A given by (4.2.4) and assume (3.6.1) to be fulfilled. In Section 6.2.2 we have seen that hA satisfies the resolvent condition (6.1.7) with the specifications (6.2.5). Since $D[-r_0, r_0] \subset \mathbb{C}_-$ it follows that hA also satisfies condition (6.1.7) with

$$\|\cdot\| = \|\cdot\|_\infty, \quad K = 1 \quad \text{and} \quad V = \mathbb{C}_- \ .$$

We again consider the fully discrete process of Section 4.2 with B specified by (4.2.3), (4.2.4). But, now we consider a fixed θ with $\frac{1}{2} \leq \theta \leq 1$. This amounts to (6.1.2) with a rational function $\varphi(z)$ the stability region S of which satisfies

$$\mathbb{C}_- \ \subset \ S.$$

An application of Theorem 6.2.1 shows that the process of Section 4.2, with $\frac{1}{2} \leq \theta \leq 1$, is stable in the sense that

$$||B^n||_\infty \ \leq \ \gamma \cdot \min\{n, s\} \quad \text{for all} \ \ n \geq 1, \ s \geq 1.$$

Here γ only depends on θ, and the result is valid for *any* $h > 0$.

6.3 Stability estimates which grow slower than linearly with n

Under conditions on V which are (slightly) stronger than (6.1.8b) variants to Theorem 6.2.1 exist in which

$$\Phi(n, s) \ = \ \gamma \cdot \min\{n^\alpha, s\}$$

with $\alpha < 1$. Below we formulate such a variant with $\alpha = 0$.

We consider the situation where

(6.3.1a) $$0 \leq \theta < \theta' < \pi/2,$$

(6.3.1b) $$W(\theta') \ \subset \ S,$$

(6.3.1c) $$V = W(\theta).$$

The following interesting theorem is valid.

Theorem 6.3.1 *Assume (6.3.1). Then there is a constant γ such that (6.1.7) implies (6.1.9) with $\Phi(n,s) \equiv \gamma$. Here γ only depends on $\varphi(z)$ and V (i.e. on θ).*

6.4 Resolvent conditions and the M-numerical range of hA

6.4.1 The construction of a set V as in the resolvent condition (6.1.7) by using M-numerical ranges

Let $||\cdot||$ be an arbitrary induced norm on $\mathbb{C}^{s,s}$, and $K \geq 1$. Suppose $A \in \mathbb{C}^{s,s}$, $h > 0$ and

$$W = \tau[hA, K]$$

(cf. Section 3.5). From Theorem 3.5.3 we see that

$$\begin{cases} zI - hA \text{ is regular and } \|(zI - hA)^{-1}\| \leq K \cdot [d(z,W)]^{-1} \\ \text{for all } z \notin W. \end{cases}$$

Therefore, we can make the following two observations.

(I) If V is any set with

$$\tau[hA, K] \subset V \subset \mathbb{C},$$

then hA satisfies the resolvent condition (6.1.7).

(II) In order to construct a set V as in observation (I) we only have to determine a finite number of pairs γ_j, ρ_j such that

$$\|(hA - \gamma_j I)^k\| \leq K(\rho_j)^k \quad \text{for} \quad k = 1, 2, 3, \ldots \ .$$

In view of Definition 3.5.1 the set

$$V = \bigcap_j D[\gamma_j, \rho_j]$$

is as required.

6.4.2 An illustration in the numerical solution of the pure diffusion equation

We consider an initial value problem of the form (6.1.1) originating from a semi-discretization of the pure diffusion problem(1.1.4). We assume the semi-discretization to be based on the finite difference formula (1.2.2a). The $s \times s$ matrix $A = (\alpha_{jk})$ in (6.1.1) thus equals the tridiagonal matrix A displayed in Subsection 1.2.3, i.e.

$$\begin{cases} \alpha_{jj} = -2/\delta^2, \quad \alpha_{jk} = 1/\delta^2 \text{ for } |j-k| = 1, \quad \alpha_{jk} = 0 \text{ for } |j-k| > 1, \\ \text{and } \delta = (s+1)^{-1}. \end{cases} \tag{6.4.1}$$

Without proof we state the following result on the matrix A.

Lemma 6.4.1 *Let α be given with $0 < \alpha < \pi/2$. Then there exist constants $K = K_\alpha \geq 1$ and $R = R_\alpha > 0$ such that, for all $s \geq 1$, the $s \times s$ matrix $A = (\alpha_{jk})$ given by (6.4.1) satisfies*

$$\left\| \left(I + \mathrm{e}^{\mathrm{i}(\frac{\pi}{2} - \alpha)} R^{-1} \delta^2 A\right)^k \right\|_\infty \ \leq \ K \quad \text{for} \quad k = 1, 2, 3, \ldots \ .$$

From this lemma we easily see that, for $k = 1, 2, 3, \ldots$,

$$\left\| \left(hA - \gamma_1 I \right)^k \right\|_\infty \ \leq \ K_\alpha(\rho_1)^k \text{ with } \gamma_1 = -\frac{h}{\delta^2} R_\alpha \mathrm{e}^{\mathrm{i}(\alpha - \frac{\pi}{2})}, \quad \rho_1 = |\gamma_1|,$$

and

$$\left\| \left(hA - \gamma_2 I \right)^k \right\|_\infty \ \leq \ K_\alpha(\rho_2)^k \text{ with } \gamma_2 = (\gamma_1)^*, \quad \rho_2 = |\gamma_2|.$$

In view of the above observations (I), (II) we deduce that, for $0 < \alpha < \pi/2$, the matrix hA satisfies (6.1.7) with

$$K = K_\alpha \ \text{ and } \ V = V_\alpha = D[\gamma_1, \rho_1] \bigcap D[\gamma_2, \rho_2].$$

We note that the set V_α satisfies

$$V_\alpha \ \subset \ W(\alpha).$$

This inclusion implies the important fact that for any $\alpha \in (0, \pi/2)$ there is a constant $K = K_\alpha$ such that the matrix given by (6.4.1) satisfies (6.1.7) with $V = W(\alpha)$. Here K_α is independent of $s \geq 1$ and $h > 0$. Combining this fact with Theorem 6.3.1 we arrive at the following theorem.

Theorem 6.4.2 *Let $\varphi(z)$ be such that, for some $\beta \in (0, \pi/2)$, the wedge $W(\beta)$ is contained in the stability region S. Let the $s \times s$ matrix $A = (\alpha_{jk})$ be defined by (6.4.1). Then there is a constant γ such that $\|\varphi(hA)^n\|_\infty \leq \gamma$. Here γ only depends on $\varphi(z)$ (and not on $n \geq 1$, $s \geq 1$ or $h > 0$).*

6.4.3 An illustration in the numerical solution of a diffusion-convection-reaction problem

In the following we illustrate the relevance of observation (I) of Subsection 6.4.1 with $K = 1$ and $\| \cdot \| = \| \cdot \|_\infty$.

We consider the initial-boundary value problem

$$\begin{aligned} (6.4.2) \quad & \frac{\partial}{\partial t} u(x,t) = \frac{\partial^2}{\partial x^2} u(x,t) - 200 \frac{\partial}{\partial x} u(x,t) - 137\,000 \cdot x \cdot u(x,t), \\ & u(0,t) = g_0(t), \quad u(1,t) = g_1(t), \quad u(x,0) = f(x), \quad \text{where } \ 0 \leq x \leq 1 \ \text{ and } \ t \geq 0. \end{aligned}$$

Here $g_0(t)$, $g_1(t)$, $f(x)$ are given functions, and $u(x,t)$ is unknown. We apply the method of semi-discretization using the finite difference formulas (1.2.1c), (1.2.2a). In this way (6.4.2) is transformed into a system of ordinary differential equations of the form (6.1.1) with a tridiagonal $s \times s$ matrix $A = (\alpha_{jk})$ for which

$$(6.4.3) \qquad \begin{cases} \alpha_{j,j-1} = \delta^{-2} + 100\,\delta^{-1} & \text{for } \ 2 \leq j \leq s, \\ \alpha_{j,j} = -2\delta^{-2} - 137\,000 \cdot j \cdot \delta & \text{for } \ 1 \leq j \leq s, \\ \alpha_{j,j+1} = \delta^{-2} - 100\,\delta^{-1} & \text{for } \ 1 \leq j \leq s-1, \end{cases}$$

with $\delta = (s+1)^{-1}$.

With this matrix A we first consider the numerical process (6.1.2) where

$$\varphi(z) = 1 + z + 0.5\, z^2 + 0.0625\, z^3.$$

The stability region S corresponding to this function $\varphi(z)$ contains the real interval $[-6, 0]$.

Let $s = 99$, so that $\delta = 10^{-2}$. Now all eigenvalues λ of A are different from each other, and real with $-157\,000 < \lambda < -20\,000$. Therefore the eigenvalue condition (6.1.5) is fulfilled whenever the stepsize $h > 0$ satisfies

$$h \leq \frac{6}{157\,000} \simeq 3.82 \times 10^{-5}.$$

We choose $h = 3.4 \times 10^{-5}$ so that (6.1.5) is amply fulfilled. But, straightforward numerical experiments show that

$$\max_{n \geq 1} \|\varphi(hA)^n\|_\infty \ > \ 3 \times 10^{11} \quad (\text{ for } \ s = 99,\ h = 3.4 \times 10^{-5}). \tag{6.4.4}$$

This inequality once more illustrates the fact that Theorem 6.1.1 can be an unreliable guide to stability.

Next we consider, with the same matrix A, the numerical process (6.1.2) where

$$\varphi(z) = 1 + z + 0.5\, z^2 + 0.0645\, z^3.$$

For $\sigma \geq 0$, $\rho \geq 0$ we introduce the subset $V(\sigma, \rho)$ of the complex plane defined by

$$V(\sigma, \rho) = \{z :\ z = x + y \text{ with } x \in \mathbb{R},\ y \in \mathbb{C},\ -\sigma - \rho \leq x \leq -\rho,\ |y| \leq \rho\}.$$

With $\sigma_0 = 4.67$ and $\rho_0 = 0.68$ the set $V(\sigma_0, \rho_0)$ is contained in the stability region of $\varphi(z)$, i.e.

$$V(\sigma_0, \rho_0) \ \subset \ S.$$

Using Theorem 3.5.4 we easily see from the expressions (6.4.3) that, for any $\delta > 0$ with $\delta \leq 10^{-2}$, the 1-numerical range of A satisfies

$$\tau_\infty[A, 1] \ \subset \ V\,(137\,000\,,\ 2\delta^{-2}).$$

In view of Theorem 3.5.2 (part (iii)) we thus arrive at

$$\tau_\infty[hA, 1] \ \subset \ V\,(137\,000\,h\,,\ 2h\delta^{-2}).$$

so that observation (I) of Subsection 6.4.1 applies to the situation at hand.

By virtue of Theorem 6.2.1 we can conclude that

$$\|\varphi(hA)^n\| \leq \gamma \cdot \min\{n, s\} \ \text{ for all } \ n \geq 1, \quad s \geq 99 \ \text{ and } h \text{ with} \tag{6.4.5}$$
$$0 < h \leq \min\Big\{\frac{\sigma_0}{137\,000},\ \frac{\delta^2}{2} \cdot \rho_0\Big\}.$$

Here γ is a constant independent of n, s, h.

Let $s = 99$. Since now $\delta = 10^{-2}$ and therefore

$$\min\Big\{\frac{\sigma_0}{137\,000}, \frac{\delta^2}{2} \cdot \rho_0\Big\} = 3.4 \times 10^{-5},$$

the stepsize restriction in (6.4.5) is fulfilled when

$$h = 3.4 \times 10^{-5}.$$

With this stepsize we thus expect a mild error propagation in accordance with the stability estimate in (6.4.5). Straightforward numerical experiments show that actually

$$\|\varphi(hA)^n\|_\infty \leq 3/2 \quad (\text{ for all } n \geq 1,\ s = 99,\ h = 3.4 \times 10^{-5}) \tag{6.4.6}$$

which is in agreement with (6.4.5).

The striking difference between (6.4.4) and (6.4.6) is closely related to the fact that the stability region S corresponding to (6.4.4) does *not* enclose the set $\tau_\infty[hA, 1]$. In fact, this S is so 'small' that it does not contain any complex number z with $\mathrm{Re}(z) = -4$ and $\mathrm{Im}(z) \neq 0$.

6.5 Resolvent conditions and the ε-pseudospectra of hA

In the above we have seen that the resolvent condition (6.1.7) is a handy tool for deriving stability estimates of the general form (6.1.9). Further, we have seen that circle conditions and M-numerical ranges can provide *sufficient* conditions in order that (6.1.7) is fulfilled. But up to now an easy interpretation of (6.1.7) has been missing.

We formulate a theorem which shows that (6.1.7) can nicely be interpreted in terms of the ε-pseudospectra of the matrix hA.

Theorem 6.5.1 *Let V be a closed subset of $\mathbb{C}$. Then the resolvent condition (6.1.7) is equivalent to the requirement that, for each $\varepsilon > 0$, the set $\sigma_\varepsilon[hA]$ is contained in*

$$\{z:\ z \in \mathbb{C} \text{ with } \mathrm{d}(z,V) \le K\cdot\varepsilon\} = \{z:\ z = x+y \text{ with } x\in V,\ |y| \le K\varepsilon\}.$$

This theorem can be proved in a straightforward way by using the fact that, according to Theorem 3.2.3, the statements (i) and (v) of Section 3.2 (with $B = hA$) are equivalent.

We note that the concept of ε-pseudospectra can also be used, in principle, to determine numerically regions V and constants K such that (6.1.7) holds. In order to explain how this may be done we write $B = hA$, choose a fixed $\varepsilon > 0$ and denote the boundary of $\sigma_\varepsilon[B]$ by Γ_ε. The set

$$V = \sigma_\varepsilon[B] \tag{6.5.1}$$

can be determined numerically, e.g. by checking for a large set of complex numbers λ whether (v) (Section 3.2) is satisfied. A corresponding constant K can be computed from the formula

$$K = |\Gamma_\varepsilon|\,(2\pi\varepsilon)^{-1}, \tag{6.5.2}$$

where $|\Gamma_\varepsilon|$ denotes the length of Γ_ε.

In order to establish (6.5.2) we note that for $z \notin V$ we have

$$(zI-B)^{-1} = \frac{1}{2\pi\mathrm{i}}\int_{\Gamma_\varepsilon}(z-\lambda)^{-1}(\lambda I-B)^{-1}d\lambda$$

and therefore

$$\left\|(zI-B)^{-1}\right\| \ \le\ \frac{|\Gamma_\varepsilon|}{2\pi}\ \max_{\lambda\in\Gamma_\varepsilon}\ |(z-\lambda)^{-1}|\varepsilon^{-1}\ =\ K\,\mathrm{d}\,(z,V)^{-1}.$$

It is clear that both V and K depend on ε. Therefore, it may pay to evaluate (6.5.1) and (6.5.2) for various values of $\varepsilon > 0$.

6.6 Notes and remarks

We note that problems of the form (6.1.1) can originate form partial differential equations in cases where the process of semi-discretization does not rely on the introduction of finite differences, but instead on e.g. finite volumes, finite elements or (pseudo) spectral approximations (cf. Section 1.3).

Step-by-step methods for the numerical solution of initial value problems in ordinary differential equations are discussed e.g. in Butcher (1987), Hairer & Wanner (1996). Runge-Kutta methods as well as Rosenbrock methods are discussed extensively in these works.

Condition (6.1.5) was stated earlier in Dorsselaer, Kraaijevanger & Spijker (1993).

The proof of Theorem 6.2.1 for the special case (6.2.1), as presented in Section 6.2.1, is essentially due to Reddy & Trefethen (1992). The value for M_1, given in Lemma 6.2.2, has been taken from Spijker (1996).

Lubich & Nevanlinna (1991) were the first to prove a version of Theorem 6.2.1 in which condition (6.2.1b) is not required; they proved (6.1.9) with $\Phi(n,s) = \gamma \cdot \min\{n,s\}$ for the case where (6.1.7) holds with $V = \mathbb{C}_- \subset S$. A proof of Theorem 6.2.1, for the general case, can be found in Spijker & Straetemans (1996b).

The (numerical) example in Section 6.2.2, with $\varphi(z) = (1+(1-\theta)z)\,(1-\theta z)^{-1},\ \theta = 1/4$, has been taken from Kraaijevanger, Lenferink & Spijker (1987).

Interesting stability estimates were derived, independently of each other, by Palencia (1993, 1995) and Crouzeix, Larsson, Piskarev & Thomée (1993). These estimates, adapted to fit in our terminology, yield Theorem 6.3.1. For further variants to Theorem 6.2.1 with $\Phi(n,s) = \gamma \cdot \min\{n^\alpha, s\},\ 0 \le \alpha < 1$, we refer to Spijker & Straetemans (1996b).

Part of the material in the Sections 6.4.1, 6.4.2 has been taken from Lenferink & Spijker (1990), Spijker (1993). The proof of Lemma 6.4.1, as given in the first of these two references, relies on stability estimates of Thomée (1965). The example in Section 6.4.3 is due to Lenferink & Spijker (1991b).

Material, closely related to Section 6.5, can be found in Dorsselaer, Kraaijevanger & Spijker (1993). Theorem 6.5.1 was formulated earlier in Reddy & Trefethen (1990, 1992).

REFERENCES

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

Ahlfors L. (1966): *Complex analysis*, McGraw-Hill (New York).

Bonsall F.F., Duncan J. (1971): *Numerical ranges of operators on normed spaces and of elements of normed algebras*, Cambridge University Press (Cambridge).

Bonsall F.F., Duncan J. (1973): *Numerical ranges II*, Cambridge University Press (New York).

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

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

Butcher J.C. (1987): *The numerical analysis of ordinary differential equations*, John Wiley (Chichester).

Canuto C., Hussaini M.Y., Quarteroni A., Zang T.A. (1988): *Spectral methods in fluid dynamics*, Springer (New York).

Cauchy A. (1888): *Oeuvres complètes d'Augustin Cauchy*, I re Série, Tome VI, Gauthier-Villars (Paris).

Conway J.B. (1985): *A course in functional analysis*, Springer (New York).

Crabb M.J. (1970): 'The power inequality on normed spaces', *Proceed. Edinb. Mathem. Soc.* **17** (Series II), 237–240.

Crank J. (1975): *The mathematics of diffusion*, Second edition, Clarendon Press (Oxford).

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

Crouzeix M., Larsson S., Piskarev S., Thomée V. (1993): 'The stability of rational approximations of analytic semigroups', *BIT* **33**, 74–84.

Dahlquist G. (1959): *Stability and error bounds in the numerical integration of ordinary differential equations*, Trans. Roy. Inst. Techn. Nr 130 (Stockholm).

Desoer C., Haneda H. (1972): 'The measure of a matrix as a tool to analyze computer algorithms for circuit analysis', *IEEE Trans. Circuit Theory* **19**, 480–486.

Dorsselaer J.L.M. van, Kraaijevanger J.F.B.M., Spijker M.N. (1993): 'Linear stability analysis in the numerical solution of initial value problems', *Acta Numerica* **1993**, 199–237.

Dorsselaer J.L.M. van, Spijker M.N. (1994): 'The error committed by stopping the Newton iteration in the numerical solution of stiff initial value problems', *IMA Journ. Numer.Anal.* **14**, 183–209.

Dowson H.R. (1978): *Spectral theory of linear operators*, Academic Press (London).

Dunford N., Schwartz J.T. (1958): *Linear operators, Part I*, Interscience Publ. (New York).

Edwards R.E. (1967): *Fourier series*, Vol I, Holt, Rinehart and Winston (New York).

Fornberg B. (1996): *A practical guide to pseudospectral methods*, Cambridge University Press (Cambridge).

Forsythe G.E., Wasow W.R. (1960): *Finite difference methods for partial differential equations*, John Wiley (New York).

Gottlieb D., Orszag S.A. (1977): *Numerical analysis of spectral methods*, Soc. Ind. Appl. Math. (Philadelphia).

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

Grigorieff R.D. (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.

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

Hairer E., Wanner G. (1996): *Solving ordinary differential equations*, Vol. II, 2nd edition, Springer (Berlin).

Henrici P. (1974): *Applied and computational complex analysis*, Vol. I, John Wiley (New York).

Hirsch C. (1988): *Numerical computation of internal and external flows*, Vol. I: Fundamentals and numerical discretization, John Wiley (Chichester etc.).

Horn R.A., Johnson C.R. (1990): *Matrix analysis*, Cambridge University Press (Cambridge).

Horn R.A., Johnson C.R. (1994): *Topics in matrix analysis*, Cambridge University Press (Cambridge).

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

Kato T. (1976): *Perturbation theory for linear operators*, 2nd ed., Springer (Berlin).

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

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

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

Kreiss H.-O. (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.

Kreiss H.-O. (1990): 'Well posed hyperbolic initial boundary value problems and stable difference approximations', in *Proceedings of the Third International Conference on Hyperbolic Problems*, Uppsala, Sweden.

Kreiss H.-O., Wu L. (1993): 'On the stability definition of difference approximations for the initial boundary value problem', *Appl. Numer. Mathem.* **12**, 213–227.

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

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

Lenferink H.W.J., Spijker M.N. (1988): 'The relevance of stability regions in the numerical solution of initial value problems', in *Numerical Treatment of Differential Equations* (K. Strehmel, ed.), Teubner (Leipzig), 95–103.

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

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

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

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

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

Lozinskij S.M. (1958): 'Error estimate for numerical integration of ordinary differential equations, Part I', *Izv. Vysš. Učebn. Zaved. Matematika* **6**, 52–90, no 6.

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

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

Martin R.H. (1976): *Nonlinear operators and differential equations in Banach spaces*, John Wiley (New York).

McCarthy C.A. (1971): 'A strong resolvent condition does not imply powerboundedness', Unpublished Note, Departm. Mathem., Univ. Minnesota, Minneapolis.

McCarthy C.A. (1972): 'A strong resolvent condition', Unpublished Note, Departm. Mathem., Univ. Minnesota, Minneapolis.

McCarthy C.A., Schwartz J. (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.

Meiss T., Marcowitz U. (1981): *Numerical solution of partial differential equations*, Springer (New York).

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

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

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

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

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

Nevanlinna O. (1993): *Convergence of iterations for linear equations*, Birkhäuser (Basel).

Nevanlinna O. (1996): 'Remarks on the growth of the resolvent operators for power bounded operators', to appear in Banach Center Publications.

Oden J.T., Reddy J.N. (1976): *An introduction to the mathematical theory of finite elements*, John Wiley (New York).

Palencia C. (1993): 'A stability result for sectorial operators in Banach spaces', *SIAM J. Numer. Anal.* **30**, 1373–1384.

Palencia C. (1995): 'Stability of rational multistep approximations of holomorphic semigroups', *Math. Comp.* **64**, 591–599.

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

Patankar S.V. (1980): *Numerical heat transfer and fluid flow*, Hemisphere Publishing Corporation (New York).

Pazy A. (1983): *Semigroups of linear operators and applications to partial differential equations*, Springer (New York).

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

Quarteroni A., Valli A. (1994): *Numerical approximation of partial differential equations*, Springer (Berlin).

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

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

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

Richtmyer R.D., Morton K.W. (1967): *Difference methods for initial-value problems*, 2nd Ed., John Wiley (New York).

Ritt R.K. (1953): 'A condition that $\lim_{n\to\infty} n^{-1}T^n = 0$', *Proc. Amer. Math. Soc.* **4**, 898–899.

Roos H-G., Stynes M., Tobiska L. (1996): *Numerical methods for singularly perturbed differential equations*, Springer (Berlin).

Le Roux M.-N. (1979): ' Semidiscretization in time for parabolic problems', *Math. Comp.* **33**, 919–931.

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

Rudin W. (1974): *Real and complex analysis*, McGraw-Hill (New York).

Shashkov M. (1996): *Conservative finite-difference methods on general grids.* CRC Press (Boca Raton).

Shields A.L. (1978): 'On Möbius bounded operators', *Acta Sci. Math.* **40**, 371–374.

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

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

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

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

Spijker M.N. (1993): 'Numerical ranges and stability estimates', *Appl. Numer. Math.* **13**, 241–249

Spijker M.N. (1996): 'Numerical stability, resolvent conditions and delay differential equations', To appear in APNUM.

Spijker M.N., Straetemans F.A.J. (1996a): 'Stability estimates for families of matrices of nonuniformly bounded order', *Linear Algebra Appl.* **239**, 77–102.

Spijker M.N., Straetemans F.A.J. (1996b): 'Error growth analysis, via stability regions, for discretizations of initial value problems', To appear in BIT.

Spijker M.N., Straetemans F.A.J. (1996c): 'A note on the order of contact between sets in the complex plane', Submitted for publication.

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

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

Strang G., Fix G.J. (1973): *An analysis of the finite element method*, Prentice-Hall (Englewood Cliffs).

Strikwerda J.C. (1989): *Finite difference schemes and partial differential equations*, Wadsworth (Belmont).

Strikwerda J.C., Wade B.A. (1991): 'Cesaro means and the Kreiss matrix theorem', *Linear Algebra Appl.* **145**, 89–106.

Ström T. (1975): 'On logarithmic norms', *SIAM J. Numer. Anal.* **12**, 741–753.

Stys T., Stys K. (1991): '$\mathcal{O}(h^4)$ locally overconvergent semidiscrete scheme for the equation $u_t = u_{xx} + f(t,x,u)$', *J. Comp. Appl. Math.* **34**, 221–231.

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

Tadmor E. (1986): 'The resolvent condition and uniform power boundedness', *Linear Algebra Appl.* **80**, 250–252.

Thomas J.W. (1995): *Numerical partial differential equations: finite difference methods*, Springer (New York).

Thomée V. (1965): 'Stability of difference schemes in the maximum-norm', *J. Differential Equations* **1**, 273–292.

Thomée V. (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.

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

Trefethen L.N. (1996): *Spectra and pseudospectra: The behaviour of non-normal matrices and operators*, book in preparation.

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

Wegert E., Trefethen L.N. (1994): 'From the Buffon needle problem to the Kreiss matrix theorem', *Amer. Math. Monthly* **101**, 132–139.

Zlatev Z. (1995): *Computer treatment of large air pollution models*, Kluwer Academ. Publish. (Dordrecht).