Boundedness and Monotonicity Properties in Numerical Initial Value Problems

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus
prof. mr. S.C.J.J. Kortmann,
volgens besluit van het college van decanen
in het openbaar te verdedigen op maandag 26 november 2012
om 15:30 uur

door

Anna Motsartova

geboren op 20 oktober 1981, te Chkalovsk, USSR

Promotoren:	prof. dr. W. Hundsdorfer	
	prof. dr. M.N. Spijker	Universiteit Leiden
Manuscriptcommissie:	prof. dr. E. Koelink	
	prof. dr. J. Frank	UVA/CWI
	prof. dr. S. Gottlieb	University of Massachusetts, Dartmouth

The work in this thesis has been carried out at the Centrum Wiskunde & Informatica (CWI). The research was funded by the Netherlands Organisation for Scientific Research (NWO).

ISBN:

To my parents
To my family
To the future

Boundedness and Monotonicity Properties in Numerical Initial Value Problems

Anna Motsartova

Preface

The thesis consists of four chapters preceded by an introduction and followed by a summary. The introduction has been written with the intention to be understandable also for the reader who is not specialized in the field. The chapters are based on papers which were published or submitted for publication in scientific journals. These papers are self-contained, and each of them may be read independently of the others. Details are listed below:

1. Chapter 1 is based on the paper by W. Hundsdorfer, A. Mozartova, M.N. Spijker: *Stepsize conditions for boundedness in numerical initial value problems,* SIAM J. Numer. Anal. 47 (2009), 3797–3819.

2. Chapter 2 is based on the paper by W. Hundsdorfer, A. Mozartova, M.N. Spijker: *Special boundedness properties in numerical initial value problems,* BIT, 51, 4 (2011), 909-936.

3. Chapter 3 is based on the paper by W. Hundsdorfer, A. Mozartova, M.N. Spijker: *Stepsize restrictions for boundedness and monotonicity of multistep methods,* J. Sci. Comput. 50, 2 (2012), 265-286

4. Chapter 4 is entitled *Comparison of Boundedness and Monotonicity Properties of One-Leg and Linear Multistep Methods,* a joint work with W. Hundsdorfer, to be submitted.

The main conclusions obtained in this thesis are listed in the summary at the end of this thesis.

Contents

Introduction

0.1 Monotonicity for time discretizations

From science, engineering, economics and the financial sciences frequently problems arise that need mathematical modelling for their solutions. Examples range in scale from the behaviour of cells in biology to the formation and development of galaxies. For such problems the solutions of the mathematical models quite often can't be achieved without resorting to the methods of numerical mathematics.

Systems of *ordinary differential equations* (ODEs) naturally arise when modelling processes that evolve in time. For example, systems of ODEs often model the motion of a body by its position and velocity; the evolution of chemical and biological species; the change of the temperature of an object in a given environment; and even the dynamics of the price of a stock.

Many interesting systems have solutions with steep gradients. Numerical approximations to such solutions often exhibit strong numerical oscillations, leading to non-physical over- and undershoots. In this thesis we discuss monotonicity properties of time discretizations. The preservation of monotonicity properties is essential for numerical schemes to approximate non-smooth solutions in a qualitatively correct manner, which means to avoid oscillations in the numerical solutions.

Initial value problems. Consider a process that evolves in time. Usually, the state of the process is known at a particular initial moment whereas its evolution has to be determined. One then arrives at an *initial value problem* (IVP) for a system of ODEs.

In this thesis we consider IVPs for general systems of ODEs in a vector space $\mathbb{V}$ for $t \geq 0$ with given initial value u_0, written as

$$\frac{d}{dt}u(t) = F(u(t)), \quad u(0) = u_0, \tag{1}$$

where $F : \mathbb{V} \to \mathbb{V}$ is a given function. The problem then is to find $u(t) \in \mathbb{V}$ for $t > 0$. In numerical applications, $\mathbb{V}$ will be a finite dimensional vector space; typically $\mathbb{V} = \mathbb{R}^M$ or $\mathbb{C}^M$.

Much study has been devoted to the solution of (1). Unfortunately, in most problems that arise in practise a useful analytical expression for the solution cannot be obtained. Therefore, it is common to seek approximate solutions by means of numerical methods.

Approximations u_n to the true solution values $u(t_n)$ at $t_n = n\Delta t$ can be obtained by a time stepping method with a positive time step Δt and $n =$

$1, 2, 3, ...$ We will say that a time stepping method is of *order* p, if for any fixed ODE with a sufficiently often differentiable function F the temporal error satisfies

$$||u(t_n) - u_n|| \leq C \cdot \Delta t^p$$

with some constant $C > 0$ and under a certain norm $||\cdot||$.

The basic assumption on F. Let $||\cdot||$ be a norm, a seminorm or a sublinear functional on $\mathbb{V}$. Recall that $||\cdot|| : \mathbb{V} \to \mathbb{R}$ is called a *sublinear functional* if

$$||\alpha v + \beta w|| \leq \alpha||v|| + \beta||w||$$

for all $\alpha, \beta \geq 0$ and $v, w \in \mathbb{V}$. It is a *seminorm* if we have in addition

$$\|-v\| = \|v\| \geq 0$$

for all $v \in \mathbb{V}$. If it also holds that

$$\|v\| = 0 \text{ only if } v = 0,$$

then $\|\cdot\|$ is a *norm*. In the following we will often consider the *maximum norm* $||v||_\infty = \max(|v_1|, \ldots, |v_M|)$ or the *total variation* (TV) seminorm $||v||_{TV} = \sum_{j=1}^{M} |v_{j-1} - v_j|$ for any vector $v \in \mathbb{R}^M$ with components v_j and $v_0 = v_M$.

In many papers one starts from an assumption about F which, for a given $\tau_0 > 0$, amounts to

$$\|v + \tau_0 F(v)\| \leq \|v\| \qquad \text{for all } v \in \mathbb{V}, \tag{2}$$

see e.g. Gotlieb, Ketcheson & Shu (2011). As we will see shortly, this assumption is relevant to many systems obtained by spatial discretization of a conservation law with suitable (semi-) norms.

It is easy to see that (2) implies $\|v + \Delta t F(v)\| \leq \|v\|$ for all $\Delta t \leq \tau_0$. Consequently, applying the forward Euler method

$$u_n = u_{n-1} + \Delta t F(u_{n-1})$$

for $n \geq 1$ with step size $\Delta t > 0$, we obtain

$$||u_n|| \leq ||u_0|| \qquad \text{for } n \geq 1, \tag{3}$$

under the step size restriction $\Delta t \leq \tau_0$. Property (3) will be referred to as *monotonicity* or *strong stability* of the numerical scheme. The largest τ_0 for which (2) holds can be viewed as the maximal step size giving monotonicity with the forward Euler method.

If F satisfies a Lipschitz condition then the forward Euler method does convergence to the exact solution on any bounded time interval $[0, T]$, see e.g. Hairer, Nørsett & Wanner (1993). In this case we can conclude

$$||u(t)|| \leq ||u(0)|| \qquad \text{for all } t \geq 0, \tag{4}$$

showing monotonicity of the exact ODE solution.

For many ODE's, the basic assumption (2) is also necessary for (4). This is illustrated by the following example.

Example 0.1.1. Consider the scalar, complex test equation $u'(t) = \lambda u(t)$ with $\lambda \in \mathbb{V} = \mathbb{C}$. This equation is known in numerical mathematics as the *Dahlquist test equation*, and the behaviour of numerical methods applied to this equation is often studied. Let $||\cdot|| = |\cdot|$ modulus. Then the basic assumption (2) is valid if and only if

$$|1 + \tau_0 \lambda| \leq 1. \tag{5}$$

Note that the set of λ satisfying (5) is a disc with a centre in $-1/\tau_0$ and radius $1/\tau_0$. On the other hand, since $u(t) = e^{\lambda t} u(0)$, (4) holds if and only if

$$\operatorname{Re} \lambda \leq 0, \tag{6}$$

that is the left half-plane including the imaginary axis. Only for the boundary case where $\operatorname{Re} \lambda = 0$ and $\lambda \neq 0$ we do have (4) but not (2). If $\operatorname{Re} \lambda < 0$ then the basic assumption (2) holds for some $\tau_0 > 0$. Therefore it is seen that the basic assumption is not only sufficient but also necessary if $\operatorname{Re}\lambda < 0$. $\diamond$

The problem with the forward Euler method is that it is only first order accurate in time. We want to consider higher order methods. A method will be called *strong stability preserving* (SSP) if there is a $\gamma > 0$ such that $\|u_n\| \leq \|u_{n-1}\|$ whenever the basic assumption (2) holds and $\Delta t \leq \gamma \tau_0$. The goal is now to specify such stepsize coefficients γ. The maximal stepsize coefficient γ is often called the *monotonicity threshold*.

In the next example we will show how, starting from the basic assumption (2), one can derive monotonicity properties for some other time stepping schemes.

Example 0.1.2. The implicit Euler method

$$u_n = u_{n-1} + \Delta t F(u_n)$$

is monotone under assumption (2) with any given sublinear functional or seminorm $||\cdot||$, without any time step restriction ($\gamma = \infty$). This is easily seen from

$$\Big(1 + \frac{\Delta t}{\tau_0}\Big) u_n = u_{n-1} + \frac{\Delta t}{\tau_0}\Big(u_n + \tau_0 F(u_n)\Big),$$

$$\Big(1 + \frac{\Delta t}{\tau_0}\Big) ||u_n|| \leq ||u_{n-1}|| + \frac{\Delta t}{\tau_0} ||u_n||,$$

showing $||u_n|| \leq ||u_{n-1}||$ for any $\Delta t > 0$.

The implicit trapezoidal rule

$$u_n = u_{n-1} + \tfrac{1}{2}\Delta t F(u_{n-1}) + \tfrac{1}{2}\Delta t F(u_n)$$

has stepsize coefficient $\gamma = 2$. This follows from the fact that the method consists of a forward Euler half-step followed by a backward Euler half-step (with $\frac{1}{2}\Delta t$).

The explicit trapezoidal rule (modified Euler)

$$\bar{u}_n = u_{n-1} + \Delta t F(u_{n-1}), \quad u_n = u_{n-1} + \tfrac{1}{2}\Delta t F(u_{n-1}) + \tfrac{1}{2}\Delta t F(\bar{u}_n)$$

has stepsize coefficients $\gamma = 1$. This becomes more apparent by writing the second stage as

$$u_n = \tfrac{1}{2}u_{n-1} + \tfrac{1}{2}\big(\bar{u}_n + \Delta t F(\bar{u}_n)\big).$$

Later we will see that $\gamma = 2$ for the implicit trapezoidal rule and $\gamma = 1$ for the explicit trapezoidal rule are optimal. ◇

From these simple examples we see that methods have to be rewritten sometimes in a form that is more convenient to make the monotonicity apparent. More importantly, we see that there is no direct relation with the usual linear stability properties of the methods for the test equation $u' = \lambda u$, $\lambda \in \mathbb{C}$. After all, the implicit trapezoidal rule is A-stable, i.e., (3) holds for this test equation with any $\lambda \in \mathbb{C}$, $\operatorname{Re}\lambda \leq 0$ without any restriction on the time step $\Delta t > 0$, whereas its explicit counterpart is only conditionally stable. In fact, the backward Euler method will turn out to be the only well-known method with threshold value $\gamma = \infty$.

0.2 Scalar conservation laws in one spatial dimension

In the examples in this thesis time stepping methods are used for solving ordinary differential equations arising from a spatial discretization of *partial differential equations* (PDEs). ODE problems with non-smooth solutions often come from a spatial discretization of hyperbolic PDEs, which pose particular difficulties for numerical methods because their solutions typically contain discontinuities.

0.2.1 Hyperbolic conservation laws.

A PDE of the form

$$u_t + (f(u))_x = 0, \tag{7}$$

with appropriate initial and boundary conditions, is called a *conservation law*. Here u is a function of x and t, and the subscripts refer to partial derivatives; for example, $u_t = \frac{\partial}{\partial t}u$. In (7), the function u usually represents the density of some quantity and $f(u)$ the *flux*. Conservation laws arise in fluid dynamics and

many other fields. By integrating in x, we see that for any $x_1 < x_2$, the integral of u over $[x_1, x_2]$ changes only because of fluxes through the endpoints:

$$\frac{d}{dt}\int_{x_1}^{x_2} u(x,t)dx = f(u(x_1,t)) - f(u(x_2,t)). \tag{8}$$

A well-known nonlinear example of a conservation law is the *inviscid Burgers equation*,

$$u_t + (\tfrac{1}{2}u^2)_x = 0. \tag{9}$$

This equation appears in studies of gas dynamics and traffic flow, and it serves as a prototype for nonlinear hyperbolic equations and conservation laws in general.

A crucial phenomenon that arises with the Burgers equation and other conservation laws is the formation of *shocks*, which are discontinuities that may appear after a certain finite time and then propagate in a regular manner. Figure 1 shows an example.

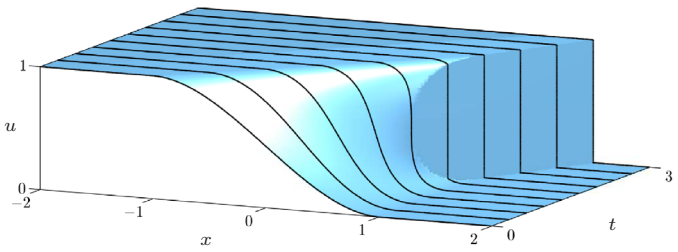

FIGURE 1: Formation of a shock.

Figure 1 is not as straightforward as it looks. It suggests that a shock simply forms and propagates. But (9) is a PDE, defined by derivatives that do not exist for discontinuous functions. A question arises: in what sense do these discontinuous curves satisfy the PDE?

To answer this question we define *weak solutions* by working with the conservation principle (8) rather than the PDE. A weak solution of (9) or (7) is a function $u(x,t)$, not necessarily smooth, that satisfies the underlying integral conservation law

$$\int_{x_1}^{x_2} [u(x,t_2) - u(x,t_1)]dx + \int_{t_1}^{t_2} [f(u(x_2,t)) - f(u(x_1,t))]dt = 0, \tag{10}$$

for all $x_1 < x_2$ and $t_1 < t_2$. Then this is the solution to be approximated by the numerical scheme.

From (8) or (10) one can derive the velocity s of a shock that separates states u_L and u_R on the left and right of a discontinuity. The result is the

Rankine-Hugoniot formula

$$s = \frac{f(u_R) - f(u_L)}{u_R - u_L},$$

see e.g. LeVeque (2002). Hence for the Burgers equation (9) we have $s = \frac{1}{2}(u_L + u_R)$. In Figure 1, the shock has velocity exactly 1/2.

However, weak solutions to (10) may be not unique. In such a situation, we can modify the differential equation slightly by adding a small amount of viscosity, or diffusion, obtaining

$$u_t + (f(u))_x = \varepsilon u_{xx} \tag{11}$$

where $\varepsilon > 0$ is a constant. If ε is very small, then we might expect solutions to (11) to be very close to solutions of (7), which has $\varepsilon = 0$, see e.g. LeVeque (2002). However, the equation (11) is *parabolic* rather than hyperbolic, and it can be proved that for any $\varepsilon > 0$ this equation has a unique solution for all time $t > 0$, and it is smooth. The curves of Figure 1 are what one obtains by taking the limit $\varepsilon \to 0$. The idea of introducing the small parameter ε and looking at the limit $\varepsilon \to 0$ is called the *vanishing-viscosity* approach to define a sensible solution to the hyperbolic equation. This simple idea is the right one, from a physical point of view, in many applications.

To get the solution obtained by the method of vanishing-viscosity, for convex f, one must impose the additional condition that shocks are permitted only if they satisfy

$$f'(u_L) > s > f'(u_R). \tag{12}$$

This is called an *entropy condition*. For a nonconvex flux functions f, the corresponding conditions can be found in LeVeque (2002). For the Burgers equation (9), the entropy condition reads $u_L > s > u_R$.

Often, for conservation laws with discontinuous solutions which are nonoscillating or positive, numerical methods do produce spurious oscillations or negative values, respectively. To avoid this, we need suitable discretizations in space and time.

0.2.2 Spatial discretization

In many applications the ODE system (1) is obtained by spatial discretization of a partial differential equation. As mentioned above, the concept of monotonicity preserving time stepping methods often arises in the numerical solution of hyperbolic partial differential equations with discontinuous solutions.

As an interesting example, consider the one-dimensional scalar conservation law

$$u_t + f(u)_x = 0 \qquad (t > 0\ ,\ 0 < x < 1), \tag{13a}$$

together with a given initial profile and periodic boundary condition

$$u(x, 0) = u_0(x), \quad u(0, t) = u(1, t). \tag{13b}$$

We would like to find a numerical approximation u in which the spatial derivative $f(u)_x$ has been discretized on a grid $\{x_i\}$ in the interval $(0,1)$. The result is then a system of ODEs, called the *semi-discrete system*,

$$u'(t) = F(u(t)). \tag{14}$$

Here $u(t)$ is a vector in $\mathbb{R}^M$ with components $u_i(t)$ approximating the PDE solution at the grid points, $u_i(t) \approx u(x_i, t)$, or approximating the averages over the cells, $u_i(t) \approx \int_{x_i-\frac{1}{2}\Delta x}^{x_i+\frac{1}{2}\Delta x} u(x,t)dx$.

Let $u_{\Delta x}$ be the appropriate restriction of the PDE solution to the grid, either as point values (finite differences) or as cell averages (finite volumes). Then a spatial discretization is said to be of *order* q, in a suitable norm $||\cdot||$ on $\mathbb{R}^M$, if

$$||u'_{\Delta x}(t) - F(u_{\Delta x}(t))|| = O(\Delta x^q). \tag{15}$$

First-order upwind discretization. Suppose that the flux function f is differentiable and $f'(w) \geq 0$ for all $w \in \mathbb{R}$. Using first-order upwind spatial discretization

$$\frac{1}{\Delta x}\big(f(u(x-\Delta x)) - f(u(x))\big) = -f(u(x))_x + O(\Delta x), \tag{16}$$

on a uniform mesh with mesh width $\Delta x = 1/M$ leads to the semi-discrete system

$$u_i'(t) = \frac{1}{\Delta x}\big(f(u_{i-1}(t)) - f(u_i(t))\big), \qquad i = 1, \ldots, M, \tag{17}$$

where $u_i(t)$ approximates $u(x_i, t)$ at the grid point $x_i = i\Delta x$. Due to spatial periodicity we have $u_0(t) = u_M(t)$.

This semi-discrete system fits in the general ODE form (1) with $\mathbb{V} = \mathbb{R}^M$ and the components of $F(v)$ given by $F_i(v) = \Delta x^{-1}(f(v_{i-1}) - f(v_i))$, $i = 1, \ldots, M$, where $v_0 = v_M$.

If $f'(w) \leq 0$ then the semi-discrete system (17) should be replaced by

$$u_i'(t) = \frac{1}{\Delta x}\big(f(u_i(t)) - f(u_{i+1}(t))\big), \qquad i = 1, \ldots, M, \tag{18}$$

where $u_{M+1}(t) = u_1(t)$.

The first-order upwind scheme is very diffusive and often not accurate enough in applications. Since the global space-time discretization error is a sum of the spatial error and the temporal error, to obtain a fully discrete numerical solution of order two, for example, we need both spatial and time discretizations at least of order two. Therefore, to show the relevance of our theory for higher-order time stepping methods we will use in the following higher-order spatial discretizations. On the other hand, common higher-order spatial discretizations produce oscillatory solutions with the possibility of negative values that may be non-physical; for example, for densities or concentrations. Undesirable negative approximations can always be just "cut off" to achieve non-negativity. But

this can destroy the conservation property, leading to incorrect shock speeds, because we are adding mass. Moreover, we do not eliminate under- and overshoots.

In the following we will show for the hyperbolic problem (13) how to achieve a spatial discretization that has better accuracy than the first-order upwind scheme and at the same time positive solutions without under- and overshoots. For this we will use a technique called *limiting*.

Discretization by flux limiting. Consider again a hyperbolic conservation law (13) with f differentiable and $f'(w) \geq 0$ for all $w \in \mathbb{R}$. (This last assumption is only for convenience of the presentation.) The spatial discretization is taken on a uniform grid, with grid points $x_i = i\Delta x$, as

$$u_i'(t) = \frac{1}{\Delta x}\big(f(u_{i-\frac{1}{2}}) - f(u_{i+\frac{1}{2}})\big), \tag{19}$$

where the values $u_{i\pm\frac{1}{2}}(t)$ approximate $u(x_{i\pm\frac{1}{2}}, t)$ at the cell boundaries $x_{i\pm\frac{1}{2}}$. These approximate values, expressed in terms of neighbouring values $u_j(t)$, determine the actual discretization. A spatial discretization in this form (19) is said to be in *flux form* or *conservation form*.

As an example, the so-called third-order upwind-biased scheme is obtained by taking

$$u_{i+\frac{1}{2}} = \tfrac{1}{6}\big(-u_{i-1} + 5u_i + 2u_{i+1}\big) = u_i + \big(\tfrac{1}{3} + \tfrac{1}{6}\theta_i\big)\big(u_{i+1} - u_i\big),$$

where θ_i is the ratio

$$\theta_i = \frac{u_i - u_{i-1}}{u_{i+1} - u_i}, \tag{20}$$

see e.g. Hundsdorfer & Verwer (2003). The resulting scheme can be shown to be of order three when interpreted as a finite volume scheme. This spatial discretization does however introduce some numerical oscillations.

Spatial discretizations in the flux form (19) that will give better accuracy than first-order upwind, but do not give rise to numerical oscillations, can be achieved by modifying the fluxes of higher-order discretizations by the so-called limiting technique. We consider the formula

$$u_{i+\frac{1}{2}} = u_i + \psi(\theta_i)(u_{i+1} - u_i), \tag{21}$$

where ψ is called the *limiter function*. We will choose this limiter function such that $\psi(0) = 0$,

$$0 \leq \psi(\theta) \leq 1 \quad \text{and} \quad 0 \leq \psi(\theta)/\theta \leq 1 \quad \text{for all} \quad \theta \in \mathbb{R}, \tag{22}$$

where $0/0$ is taken as 0.

This property (22) holds with the limiter function

$$\psi(\theta) = \max\big(0, \min(1, \tfrac{1}{3} + \tfrac{1}{6}\theta, \theta)\big) \tag{23}$$

which coincides with the original third-order upwind-biased function $\psi(\theta) = \frac{1}{3} + \frac{1}{6}\theta$ for $\frac{2}{5} \leq \theta \leq 4$. This limiter function was introduced by Koren (1993). Another example of a limiter function satisfying the conditions is

$$\psi(\theta) = \tfrac{1}{2} \cdot \frac{\theta + |\theta|}{1 + |\theta|} \tag{24}$$

introduced by van Leer (1974).

For a smooth solution profile we have $\theta_i \approx 1$, except near extrema. If we take the Koren limiter (23), which gives

$$\psi(\theta) = \tfrac{1}{3} + \tfrac{1}{6}\theta \quad \text{for } \theta \approx 1,$$

then the accuracy of the third-order scheme will be maintained in smooth regions away from extrema. With the van Leer limiter (24), linearization of the limiter function near $\theta = 1$, i.e. replacement of $\psi(\theta)$ by $\psi(1) + \psi'(1)(\theta - 1) = \frac{1}{4} + \frac{1}{4}\theta$, gives the so-called Fromm scheme, which is of order two, again in smooth regions away from extrema.

Remark 0.2.1. The development of a discontinuity from a smooth initial function is a typical property of nonlinear hyperbolic equations. For example, the formation of a shock in fluid flows is described by such equations. For the hyperbolic problem (7), correct discretizations are important to have a correct propagation of the discontinuities. With a semi-discrete scheme (19) in conservation form we can expect that a steep travelling front or shock is computed in the correct location. Lax and Wendroff (1960) proved that if the numerical approximations converge to some function, then this function will in fact be a weak solution of the conservation law, and it will therefore satisfy the Rankine-Hugoniot relation for the shock speed. In LeVeque (1992, Fig.12.1) the importance of the conservation form for the discretization is illustrated for the inviscid Burgers equation (9). ◇

Properties of the semi-discrete system. Consider a spatial discretization in the flux form (19) with (21) and limiter function (23), (24), or $\psi = 0$ (first-order upwind). As before, it is assumed that the flux function f is differentiable, and we also assume that there is an $\alpha > 0$ such that

$$0 \leq f'(w) \leq \alpha \qquad \text{for all } w \in \mathbb{R}. \tag{25}$$

It will be shown that the resulting semi-discrete system of ODEs (14) satisfies the basic assumption (2) in the maximum norm and TV seminorm.

The basic assumption (2). Taking $F_i(u) = \Delta x^{-1}(f(u_{i-\frac{1}{2}}) - f(u_{i+\frac{1}{2}}))$, it is seen from the mean value theorem that

$$F_i(u) = \frac{f'(v_i)}{\Delta x}(u_{i-\frac{1}{2}} - u_{i+\frac{1}{2}}),$$

with v_i some intermediate point between $u_{i-\frac{1}{2}}$ and $u_{i+\frac{1}{2}}$. We also have

$$u_{i-\frac{1}{2}} - u_{i+\frac{1}{2}} = \big(1 - \psi(\theta_{i-1}) + \tfrac{1}{\theta_i}\psi(\theta_i)\big)(u_{i-1} - u_i), \tag{26}$$

with $0 \le (1 - \psi(\theta_{i-1}) + \psi(\theta_i)/\theta_i) \le 2$.

Consequently, using a spatial discretization in the flux form (19) based on (21) with Koren limiter or the van Leer limiter, on a uniform mesh with width $\Delta x = 1/M$, leads to an initial value problem $u' = F(u)$, $u(0) = u_0$, with $\mathbb{V} = \mathbb{R}^M$, for which the components of $F(u)$ can be written as

$$F_i(u) = \beta_i(u)(u_{i-1} - u_i), \quad 0 \le \beta_i(u) \le \frac{2\alpha}{\Delta x} \tag{27}$$

for $i = 1, \ldots, M$, where $u_0 = u_M$. From this observation it easily follows that the basic assumption (2) will be satisfied in the maximum norm with $\tau_0 = \frac{\Delta x}{2\alpha}$. With the same τ_0 it can also be shown quite easily that condition (2) will be satisfied with the TV seminorm.

Note that in this example, the use of first-order upwind spatial discretization gives $F_i(u) = \beta_i(u)(u_{i-1} - u_i)$, with $0 \le \beta_i(u) \le \frac{\alpha}{\Delta x}$ instead of (27). Then condition (2) will be satisfied in the maximum norm and the TV seminorm with $\tau_0 = \frac{\Delta x}{\alpha}$. But then we only have first-order accuracy.

The Lipschitz condition. For the discretizations with a limiter, the values at the cell boundaries can be written as

$$u_{i+\frac{1}{2}} = h(u_{i-1}, u_i, u_{i+1}),$$

where the function $h : \mathbb{R}^3 \to \mathbb{R}$ is given by formula (21). Let us take as an example the Koren limiter (23); for the van Leer limiter (24) the following arguments will be similar. Due to the ratios θ_i in the limiter, it is not obvious that the function h will satisfy a Lipschitz condition, but in fact it does. This result seems not available in the standard literature, and therefore it will be indicated here how to prove it.

For this, first note that

$$h(u_{i-1}, u_i, u_{i+1}) = u_i + \psi(d_{i-1}/d_i) \cdot d_i, \qquad d_i = u_{i+1} - u_i \quad \text{for all } i.$$

The function $g(\xi, \eta) = \psi(\eta/\xi)\,\xi$, from $\mathbb{R}^2$ to $\mathbb{R}$, is continuous on $\mathbb{R}^2$, and it is differentiable in the arguments ξ, η, except for the special cases $\xi = 0$, $\eta = 0$, $\eta = \frac{2}{5}\xi$ and $\eta = 4\xi$. Away from these four lines in $\mathbb{R}^2$, the partial derivatives are uniformly bounded,

$$\left|\frac{\partial}{\partial\xi} g(\xi, \eta)\right| = |\psi(\theta) - \psi'(\theta)\theta| \le 1 + \tfrac{2}{3}, \qquad \left|\frac{\partial}{\partial\eta} g(\xi, \eta)\right| = |\psi'(\theta)| \le 1,$$

with $\theta = \eta/\xi$. It follows that g satisfies a Lipschitz condition on $\mathbb{R}^2$.

Using this, it now seen that the function h satisfies a Lipschitz condition on $\mathbb{R}^3$, from which it is also clear that the function F will satisfy a Lipschitz condition on $\mathbb{R}^M$ for fixed M.

0.3 Monotonicity properties

There are many related monotonicity properties. In this section we will consider two examples of such properties by considering different seminorms or sublinear functionals.

Example 0.3.1. A notable monotonicity property is the so-called *total variation diminishing* (TVD) property:

$$||u_n||_{TV} \leq ||u_{n-1}||_{TV} \qquad \text{for } n \geq 1, \tag{28}$$

which means that the total variation of the numerical solution does not increase. This property frequently appears in the literature on computational fluid dynamics.

Instead of the TVD property, we can consider the more general *total variation boundedness* (TVB) property, for which it is required that a finite μ exists such that, for all $n \geq 1$ and any u_0,

$$\|u_n\|_{TV} \leq \mu \cdot \|u_0\|_{TV}. \tag{29}$$

With nonlinear conservation laws (7) this property is often crucial to obtain convergence towards the physically relevant solution. It was shown in Harten, Hyman & Lax (1976) that if the numerical scheme satisfies the TVB property and the physically relevant solution is identified by the entropy condition (12), then convergence towards the correct solution is guaranteed. ◇

Example 0.3.2. Another related monotonicity property is *preservation of nonnegativity*:

$$u_n \geq 0 \quad \text{ whenever } u_0 \geq 0,$$

where inequalities for vectors in $\mathbb{R}^M$ should be interpreted component-wise. This property is often called *positivity*, which is short but not entirely correct. It is often necessary to ensure a correct physical meaning of approximations. For example, for problems whose solutions are concentrations. Such problems arise frequently when modelling chemical reactions or semidiscretizing PDEs of advection-diffusion type. Solving such problems numerically with nonnegative initial vectors, it is natural to demand nonnegativity of the resulting numerical approximations. For linear systems of ODEs nonnegativity was investigated by Bolley & Crouzeix (1978). Later nonnegativity preservation theory for nonlinear problems was developed by Horvath (1998, 2005).

To illustrate that preservation of nonnegativity can also be cast in our general framework with sublinear functionals, we consider

$$||v||_0 = -\min\{0, v_1, \ldots, v_M\} \qquad \text{for } v = (v_1, v_2, \ldots, v_M)^T \in R^M. \tag{30}$$

It was shown above that flux limiting gives a semi-discrete system for which the basic assumption (2) is satisfied in the maximum norm and TV seminorm. Using similar arguments, it follows that the basic assumption (2) is satisfied with this

functional $||\cdot||_0$ and $\tau_0 = \frac{\Delta x}{2\alpha}$ for any ODE system (27). Here we have $||v||_0 = 0$ if and only if $v \geq 0$, that is all components of v are nonnegative. Consequently, the monotonicity property (3) then implies nonnegativity for these ODE systems.

Therefore from monotonicity for arbitrary sublinear functionals one can conclude nonnegativity. The example also illustrates that apart from (semi-)norms it is useful to consider sublinear functionals, because this leads to such an important property as preservation of nonnegativity. ◇

There are other related monotonicity properties, for example the *maximum principle*

$$\min_j u_{0,j} \leq u_{n,i} \leq \max_j u_{0,j} \quad \text{for all } n \geq 0 \text{ and } 1 \leq i \leq M,$$

where the $u_{n,i}, u_{0,j}$ are the components of the vectors $u_n, u_0 \in \mathbb{R}^M$. This can be associated with the absence of unwanted global overshoots and undershoots, and it can be cast in the general framework (2), (3), by introducing suitable sublinear functionals, see e.g. Spijker (2007).

0.4 Numerical Illustrations

To demonstrate questions which can be solved using monotonicity and boundedness theory, we consider two numerical illustrations.

Two-step Adams-Bashforth (AB2) time discretization. Consider the advection equation

$$u_t + u_x = 0 \qquad \text{for } \; t > 0 \,, \;\; 0 < x < 1 \,, \tag{31}$$

with periodic boundary condition $u(0,t) = u(1,t)$ and an initial block profile:

$$\begin{cases} u(x,0) = 1 & \text{if } \; 0.4 \leq x \leq 0.6; \\ u(x,0) = 0 & \text{otherwise.} \end{cases}$$

The advection equation (31) has the general solution $u(x,t) = u(x-t,0)$, revealing that the initial profile is transported without change of shape along parallel straight characteristic lines. Note that the block function is discontinuous and hence not differentiable. Consequently, the characteristic solution is not a solution of the differential equation in the classical sense. It is a solution of the underlying integral conservation law.

The spatial discretization is taken on a uniform grid with mesh width $\Delta x = 1/M$, where M is the total of grid points, using Koren limiter scheme (see Section 0.2.2 above for the semi-discrete form). This gives us a semi-discrete system of ODEs for which the monotonicity assumption (2) is satisfied for $\tau_0 = 0.5\Delta x$ in the maximum norm and TV seminorm.

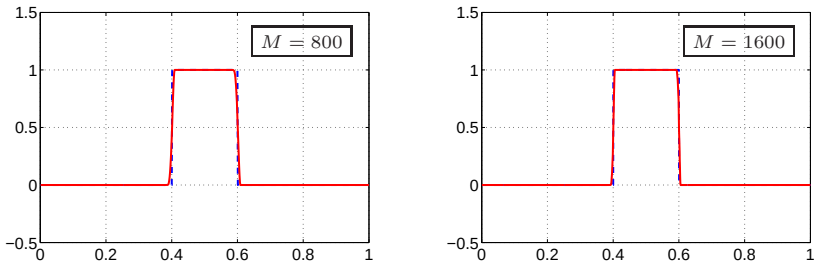

FIGURE 2: The red line is AB2 solutions at $t = 1$ for the linear advection equation with $\Delta t = 0.25\Delta x$. The blue dashed line is an exact PDE solution.

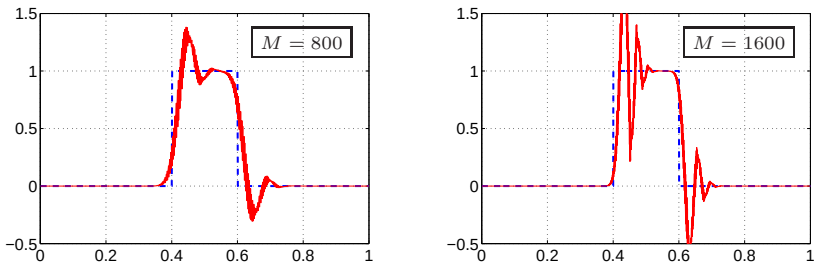

FIGURE 3: The red line is AB2 solutions at $t = 1$ for the linear advection equation with $\Delta t = 0.5\Delta x$. The blue dashed line is an exact PDE solution.

To discretize the resulting nonlinear semi-discrete system of ODEs in time we take the well-known two-step Adams-Bashforth method

$$u_n = u_{n-1} + \frac{3}{2}\Delta t\, F(u_{n-1}) - \frac{1}{2}\Delta t\, F(u_{n-2}). \tag{32}$$

This method has order 2, as defined in Section 0.1. The first approximation u_1 is computed by the forward Euler method: $u_1 = u_0 + \Delta t F(u_0)$.

Numerical solutions are shown in the Figures 2 and 3, with spatial component x horizontally for the output time $t = T$ with $T = 1$. Dashed lines indicate the exact PDE solution. The behaviour of the scheme is seen to be very different from Figure 2 to Figure 3. Whereas for $\Delta t/\Delta x = 0.25$ we get a nice monotonic behaviour, the scheme with $\Delta t/\Delta x = 0.5$ produces large oscillations.

Let us first consider the results with $\Delta t/\Delta x = 0.25$. The scheme gives results close to the exact solution, see Figure 2. In this case the values $||u_N||_{TV}$ are constant equal to 2 for all observed M, and the discrete L_1 errors $||u_N - u(T)||_1$ ($||v||_1 = \sum_i |v_i|$, $N = T/\Delta t$ and $u(T)$ is the exact solution) go to zero for increasing M, see Figure 4. On the other hand, for $\Delta t/\Delta x = 0.5$, in

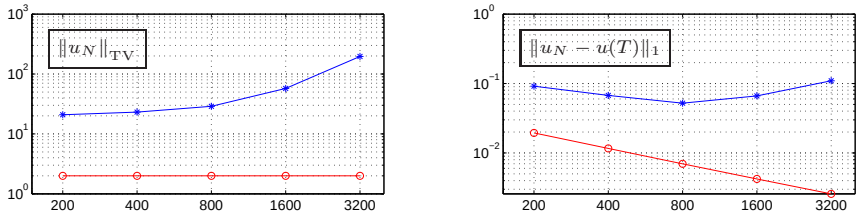

FIGURE 4: Values of $\|u_N\|_{\mathrm{TV}}$ (left picture) and $\|u_N - u(T)\|_1$ (right picture) for $M = 200, 400, 800, 1600, 3200$ and the AB2 method with $\Delta t = 0.5\Delta x$ (blue line, $*$ markers), $\Delta t = 0.25\Delta x$ (red line, $\circ$ markers).

Figure 3, we see that if Δt and Δx are decreased while keeping $\Delta t/\Delta x$ fixed, the oscillations become more and more pronounced. The evolution of the total variation seminorm and L_1 norm is shown in Figure 4 revealing a marked growth for increasing, large values of the dimension M if $\Delta t/\Delta x = 0.5$.

For this well-known second order Adams-Bashforth method the different behaviour according to $\Delta t/\Delta x$ can be explained by the monotonicity and boundedness theory presented in this thesis in Chapters I and III. Furthermore, in Chapter III we will answer the question: what is the largest Δt for the AB2 method such that the behaviour is monotonic with the forward Euler starting procedure $u_1 = u_0 + \Delta t F(u_0)$?

Two-step backward differentiation formulas (BDFs). As a next illustration, we consider the Buckley-Leverett equation

$$u_t + f(u)_x = 0, \qquad f(u) = \frac{cu^2}{cu^2 + (1-u)^2}. \tag{33}$$

This equation provides a simple model for two immiscible fluids in a porous medium and has applications in oil-reservoir simulation. The unknown u here represents the saturation of water in an oil reservoir and lies between 0 and 1. The constant $c > 0$ gives the mobility ratio of the two fluid components.

We consider this problem with $c = 3$, $0 < t \leq \frac{1}{4}$ and $0 \leq x \leq 1$ with inflow condition $u(0,t) = \frac{1}{2}$ and an initial block-function:

$$\begin{cases} u(x,0) = \frac{1}{2} & \text{if } \ x = 0; \\ u(x,0) = 0 & \text{if } \ 0 < x \leq \frac{1}{2}; \\ u(x,0) = 1 & \text{if } \ \frac{1}{2} < x \leq 1. \end{cases}$$

The flux function f is monotonically increasing for $0 \leq u \leq 1$. The solution is shown in Figure 5 with the blue dashed line.

We use a fixed grid with mesh width $\Delta x = 5 \cdot 10^{-3}$. For spatial discretization we use the limited conservative scheme (19) based on (21) with van Leer limiter

(24). With limiting the semi-discrete solution is accurate, only the shocks are slightly diffused, over a few grid cells. Figure 5 shows the initial profile and a time-accurate reference solution which corresponds to the exact solution of the semi-discrete system.

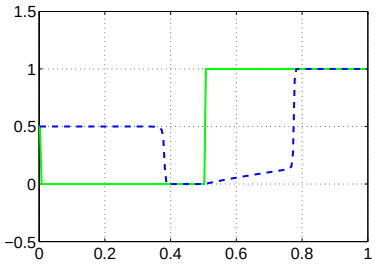

FIGURE 5: The initial profile (green solid line) and a time-accurate semi-discrete solution (blue dashed line) at $t = 1/4$ for the Buckley-Leverett equation.

The resulting semi-discrete system is integrated by the implicit 2-step BDF method

$$u_n = \frac{4}{3}u_{n-1} - \frac{1}{3}u_{n-2} + \frac{2}{3}\Delta t F(u_n) \tag{34}$$

and its explicit counterpart, the extrapolated BDF2 method

$$u_n = \frac{4}{3}u_{n-1} - \frac{1}{3}u_{n-2} + \frac{4}{3}\Delta t F(u_{n-1}) - \frac{2}{3}\Delta t F(u_{n-2}). \tag{35}$$

The first approximation u_1 for (34) is computed by $u_1 = u_0 + \Delta t F(u_1)$ and for (35) by $u_1 = u_0 + \Delta t F(u_0)$. Both methods have order 2. To solve the algebraic system for the implicit method, a Newton-type iteration is used. It is stressed that per step the implicit method is much more costly than the explicit one. Hence the implicit method can only be efficient if it allows much larger time steps than the explicit method. The implicit method is A-stable, G-stable, see e.g. Butcher (2003) or Hairer & Wanner (1996). In a von Neumann analysis, assuming a smooth PDE solution, linearization and constant (frozen) coefficients, this would give unconditional stability, whereas stability of the explicit method implies that the ratio $\Delta t/\Delta x$ can be no more than about 0.5, with the present with the present spatial discretization. But, as we will see below, the two schemes have approximately the same monotonicity threshold.

We use this example to illustrate once more the importance of monotonicity. Due to the monotonicity restriction, the implicit BDF2 method cannot be used with large time steps if undershoots and overshoots are to be avoided. In the Figures 6 and 7 the numerical solutions at time $t = \frac{1}{4}$ are plotted as functions of x with solid lines. Dashed lines indicate a time-accurate semi-discrete solution on the same grid. In Figure 6, where $\Delta t = 0.25\Delta x$, both methods give

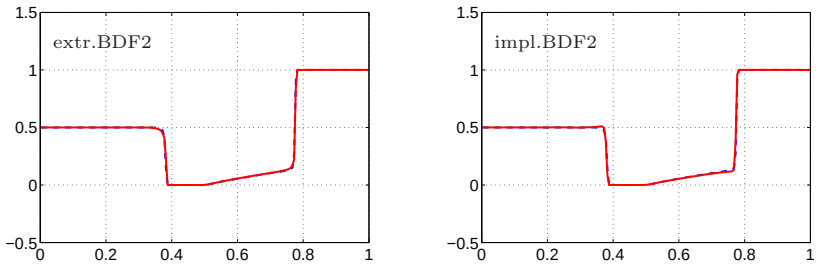

FIGURE 6: BDF2 solutions at $t = 1/4$ for the Buckley-Leverett equation with $\Delta t = 0.25\Delta x$. The dashed line is a time-accurate semi-discrete solution ($\Delta x = 5{\cdot}10^{-3}$).

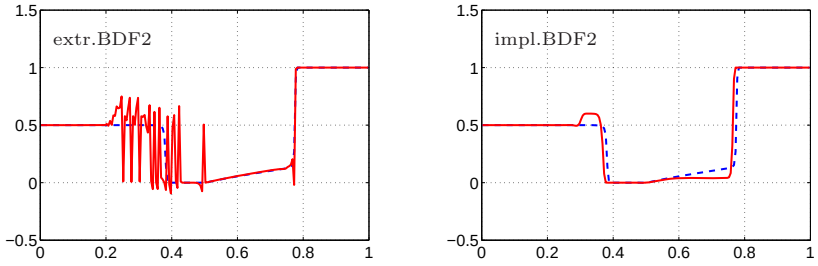

FIGURE 7: BDF2 solutions at $t = 1/4$ for the Buckley-Leverett equation with $\Delta t = 0.5\Delta x$. The dashed line is a time-accurate semi-discrete solution ($\Delta x = 5{\cdot}10^{-3}$).

results close to the exact semi-discrete solution. However, if the time step size is increased to $\Delta t = 0.5\Delta x$ we see from Figure 7 that now the explicit solution becomes unstable, but at the same time the implicit solution becomes very inaccurate: both the shock speed and the shock height are no longer correct. This disappointing qualitative behaviour of the implicit method is due to loss of monotonicity for large step sizes, giving over- and undershoots after the shocks.

0.5 Monotonicity and boundedness properties for RKMs, LMMs and GLMs: reviewing some literature

In this thesis we deal with general time stepping methods of Runge-Kutta or linear multistep type. Such methods, and others, all fit in the framework of general linear methods.

0.5.1 Runge-Kutta methods

The general Runge-Kutta method (RKM), for computing u_n for $n = 1, 2, \ldots$, can be written in the form

$$y_i^{[n]} = u_{n-1} + \Delta t \cdot \sum_{j=1}^{s} a_{ij}\, F(y_j^{[n]}) \quad (1 \le i \le s), \tag{36a}$$

$$u_n = u_{n-1} + \Delta t \cdot \sum_{j=1}^{s} b_j\, F(y_j^{[n]}). \tag{36b}$$

Here a_{ij} and b_j are parameters defining the method. Furthermore, the vectors $y_i^{[n]}$ $(1 \le i \le s)$ are internal approximations used for computing u_n from u_{n-1}, cf. e.g. Butcher (1987) or Hairer, Nørsett & Wanner (1993). If $a_{ij} = 0$ (for $j \ge i$), the method is called *explicit*. Define the $s \times s$ matrix A by $A = (a_{ij})$ and the column vector $b \in \mathbb{R}^s$ by $b = (b_1, b_2, \ldots, b_s)^T$.

For method (36), much attention has been paid in the literature to the monotonicity property

$$\|y_i^{[n]}\| \le \|u_{n-1}\| \qquad (\text{for } 1 \le i \le s), \tag{37a}$$

$$\|u_n\| \le \|u_{n-1}\|. \tag{37b}$$

For classes of RKMs, positive stepsize-coefficients c were determined, such that monotonicity, in the sense of (37), is present for all Δt with

$$0 < \Delta t \le c \cdot \tau_0,$$

as long as the basic assumption (2), associated with the forward Euler method, is satisfied. For explicit RKMs, this was done by rewriting the right-hand members of (36) as convex combinations of forward Euler steps; see e.g. Shu & Osher (1988), Spiteri & Ruuth (2002), and Ruuth (2006). For more general RKMs, stepsize-coefficients were obtained, e.g., in Gottlieb, Shu & Tadmor (2001), Ferracina & Spijker (2004), Higueras (2005) and Spijker (2007, Section 3.2.1).

An important characteristic quantity for Runge-Kutta methods was introduced by Kraaijevanger (1991). Following this author we denote his quantity by $R(A, b)$, and in defining it, we focus on values $\xi \le 0$ for which

$$\begin{gathered}(I - \xi A) \text{ is invertible, } \; A(I - \xi A)^{-1} \ge 0, \quad b^T(I - \xi A)^{-1} \ge 0, \\ (I - \xi A)^{-1} e \ge 0 \quad \text{and} \quad 1 + \xi b^T (I - \xi A)^{-1} e \ge 0.\end{gathered} \tag{38}$$

Here I denotes the identity matrix of order s, and e stands for the column vector in $\mathbb{R}^s$ all whose components are equal to 1. All inequalities for matrices and vectors should be interpreted entry-wise and component-wise, respectively. Then for a given RKM, the number $R(A, b)$ is defined as

$$R(A, b) = \sup\{r : \, r \ge 0 \text{ and } (38) \text{ holds for all } \xi \text{ with } \, -r \le \xi \le 0\}.$$

In case at least one of the inequalities $A \geq 0$, $b \geq 0$ is violated, $R(A,b) = 0$ is defined. For an arbitrary explicit Runge-Kutta method with $s \geq 1$ of order p, it was shown by Kraaijevanger (1991) that

$$R(A,b) \leq s - p + 1 \quad \text{if } 1 \leq p \leq s.$$

Special attention was paid to the problem of determining, for a given RKM, the corresponding maximal stepsize coefficient c. In Higueras (2004, 2005), Ferracina & Spijker (2004, 2005) conditions were derived under which this coefficient equals $R(A,b)$. Earlier results about $R(A,b)$ were extended in Spijker (2007, Section 2.2).

0.5.2 Linear multistep methods

Runge-Kutta methods are sometimes referred to as *one-step* methods, since they evolve the solution from t_{n-1} to t_n without needing to know the solutions at $t_{n-2}, t_{n-3}, \cdots$, etc. There is a broad class of more sophisticated integration methods, known as *linear multistep methods*, which for computing u_n use the previously calculated $u_{n-1}, u_{n-2}, \cdots, u_{n-k}$ (for a k-step method). The main advantages of Runge-Kutta methods are that changes in the stepsize are easy to implement, and, in contrast to multistep methods, we do not have to compute the first few steps by some other (one-step) integration method. The advantage of linear multistep methods is that they require significantly fewer computations per step than Runge-Kutta methods of comparable accuracy.

The general linear k-step method (LMM) can be written in the form

$$u_n = \sum_{j=1}^{k} a_j u_{n-j} + \Delta t \sum_{j=0}^{k} b_j F(u_{n-j}) \tag{39}$$

for $n \geq k$, where the parameters a_j, b_j define the method, e.g. Butcher (1987), Hairer, Nørsett & Wanner (1993). If $b_0 = 0$, the method is called *explicit*. The starting values for this multistep recursion, $u_0, u_1, \ldots, u_{k-1} \in \mathbb{V}$, are supposed to be given, or computed by a Runge-Kutta method. Examples of these methods are the two-step Adams-Bashforth method (32) and the two-step BDF method (34).

If all $a_j, b_j \geq 0$, then from the basic assumption (2) it can be shown that

$$\|u_n\| \;\leq\; \max_{0 \leq j < k} \|u_j\| \tag{40}$$

for $n \geq k$, under the stepsize restriction

$$\Delta t \leq c \cdot \tau_0, \qquad c \;=\; \min_{1 \leq j \leq k} \frac{a_j}{b_j}\,, \tag{41}$$

with the convention $a/0 = +\infty$ if $a \geq 0$; see e.g. Spijker (2007), Gottlieb, Ketcheson & Shu (2009). Property (40) can be viewed as an extension of (3) for multistep methods with arbitrary starting values.

Results of this type for nonlinear problems were derived in Shu (1988) (with $b_0 = 0$), originally in the total variation seminorm. Related results for linear systems were given in Bolley & Crouzeix (1978) for positivity, and in Lenferink (1989), Spijker (1983) for *contractivity*: where one considers $\|\tilde{u}_n - u_n\|$ with differences of two numerical solutions instead of $\|u_n\|$ as in (40). More recently, with arbitrary seminorms or more general *convex functionals*, i.e.

$$||\lambda v + (1-\lambda)w|| \leq \lambda||v|| + (1-\lambda)||w||$$

for $0 \leq \lambda \leq 1$ and $v, w \in \mathbb{V}$, the term SSP (strong stability preserving) – introduced in Gottlieb, Shu & Tadmor (2001) – has become popular. Related work for nonlinear problems was done in Lenferink (1991), Sand (1986), and Vanselov (1983) for contractivity.

Using the order conditions, it was shown by Lenferink (1989) that the maximal size of the threshold factor c for an explicit k-step method of order p is bounded by

$$\begin{cases} c \leq 1 & \text{if } p = 1, \\ c \leq \frac{k-p}{k-1} & \text{if } p \geq 2. \end{cases} \tag{42}$$

The upper bound $c = 1$ for $p = 1$ is attained by the forward Euler method. Optimal higher-order multistep methods have been constructed by Shu (1988), Lenferink (1989) and Gottlieb, Shu & Tadmor (2001).

In order to conclude (40) from the basic assumption (2) for arbitrary (semi-)norms or sublinear functionals, the condition that all $a_j, b_j \geq 0$ and $\Delta t \leq c\,\tau_0$ is *necessary*. In fact, this condition is already needed if we only consider maximum norms instead of arbitrary (semi-)norms; see Spijker (2007).

Consider again Adams-Bashforth method (32), BDF method (34) and EBDF method (35). As it was shown in Section 0.4, there are Δt such that the methods give accurate approximations to the true solutions. But these methods have some negative coefficients and, therefore, are not covered by the above theory.

So, the methods with nonnegative coefficients form only a small class, excluding the well-known methods of the Adams or BDF-type, and the stepsize requirement $\Delta t \leq c\,\tau_0$ (within this class) can be very restrictive. It is therefore of interest to study the following weaker boundedness property

$$\|u_n\| \;\leq\; \mu \cdot \max_{0 \leq j < k} \|u_j\| \tag{43}$$

for $n \geq k$, under the stepsize restriction $\Delta t \leq \gamma\tau_0$, where the stepsize coefficient $\gamma > 0$ and the factor $\mu \geq 1$ are determined by the multistep method.

Sufficient conditions were derived in Hundsdorfer & Ruuth (2006) and Hundsdorfer, Ruuth & Spiteri (2003) for (43) to be valid with arbitrary seminorms under the basic assumption (2) and $\Delta t \leq \gamma\tau_0$.

0.5.3 General linear methods

We recall that LMMs and RKMs are examples of methods belonging to the important and very large class of *general linear methods* (GLMs), introduced by Butcher (1966), and studied extensively in the literature –see e.g. Butcher (1987, 2003), Hairer, Nørsett & Wanner (1993), Hairer & Wanner (1996), and the references therein.

The general linear method, for solving (1), depends on parameters c_j $(1 \leq j \leq q)$ and parameter matrices $A = (\alpha_{ij}) \in \mathbb{R}^{q \times l}$, $\quad B = (\beta_{ij}) \in \mathbb{R}^{q \times q}$, where $1 \leq l \leq q$. The method can be written in the following form:

$$y_i = \sum_{j=1}^{l} \alpha_{ij} u_j^{[n-1]} + \Delta t \sum_{j=1}^{q} \beta_{ij} F(y_j) \quad (1 \leq i \leq q), \tag{44a}$$

$$u_i^{[n]} = y_{q-l+i} \quad (1 \leq i \leq l). \tag{44b}$$

Here $u_i^{[n-1]}$ are input vectors available at the n-th step of the method, whereas y_i are (intermediate) approximations used for computing the $u_i^{[n]}$ at the new time level $(n = 1, 2, 3, \ldots)$; cf. e.g. Butcher (1966), Butcher (1987, pp. 338).

Obviously, the Runge-Kutta method (36) can be written in the form (44), with $l = 1$, $q = s + 1$, $u_1^{[n]} = u_n \simeq u(n \cdot \Delta t)$ and $\alpha_{i1} = 1$, $\quad \beta_{ij} = a_{ij}$ (for $1 \leq j \leq s$), $\beta_{ij} = 0$ (for $j = s + 1$).

The linear multistep method (39) can be written in the form (44), with $l = k$, $q = k+1$ and $u_i^{[n]} = u_{n-1+i}$ $(1 \leq i \leq k,\ n \geq 0)$, $y_i = u_{n-2+i}$ $(1 \leq i \leq k+1,\ n \geq 1)$, and with $A = \binom{I}{a}$, $B = \binom{O}{b}$, where I denotes the $k \times k$ identity matrix, O the $k \times (k+1)$ zero matrix and $a = (a_k, \ldots, a_1)$, $b = (b_k, \ldots, b_0)$.

For method (44) monotonicity is studied in the form

$$\|y_i\| \leq \max_{1 \leq j \leq l} \|u_j^{[n-1]}\| \quad (1 \leq i \leq m), \tag{45}$$

under the stepsize restriction $\Delta t \leq \gamma \tau_0$. Then the goal is to derive stepsize-coefficients γ with the following important property:

> Condition $0 < \Delta t \leq \gamma \cdot \tau_0$ implies monotonicity, whenever $\mathbb{V}$ is a vector space, $\|\cdot\|$ a convex function on $\mathbb{V}$, and the function $F : \mathbb{V} \to \mathbb{V}$ satisfy the basic assumption (2); (46)

see e.g. Spijker (2007) and the references therein.

For arbitrary GLMs the maximal stepsize-coefficient γ for monotonicity was determined in Spijker (2007). It was shown there that the maximal stepsize-coefficients γ for monotonicity is also relevant to a discrete maximal principle and numerical contractivity of GLMs.

0.6 Scope of the thesis

Chapter I: Stepsize Conditions for Boundedness in Numerical Initial Value Problems

As an important example to motivate the study, we consider the total variation seminorm $\|\cdot\|_{TV}$. Numerical processes, satisfying $\|u_n\|_{TV} \leq \|u_{n-1}\|_{TV}$, play a special role in the solution of hyperbolic conservation laws, cf. e.g. Harten (1983), Shu (1988), Shu & Osher (1988), LeVeque (2002), Hundsdorfer & Verwer (2003). Clearly, the monotonicity property (37) or (40) with $\|\cdot\| = \|\cdot\|_{TV}$ implies a TVB property, in that there is a finite μ such that

$$\|u_n\|_{TV} \leq \mu \cdot \max_{0 \leq j \leq k-1} \|u_j\|_{TV} \quad \text{(for all } n \geq k\text{)}. \tag{47}$$

As discussed in Section 0.3, satisfying (47) is of crucial importance for suitable convergence properties when $\Delta t \to 0$, and constitutes one of the underlying reasons why attention has been paid in the literature to the monotonicity properties (37) or (40), see e.g. LeVeque (2002).

Unfortunately, there are well known RKMs and LMMs, with a record of practical success, for which there exist *no* positive stepsize coefficients γ such that the monotonicity property (37) or (40), respectively, holds whenever $\Delta t \leq \gamma\tau_0$. This was discussed already for LMMs in Section 0.5.2. Therefore one cannot conclude in the way described above that the methods are total-variation bounded. It is therefore worthwhile to study for GLMs directly boundedness properties similar to (43).

Moreover, some special LMMs were found with a positive stepsize coefficient γ such that the boundedness property (43) holds under the basic assumption (2) and $\Delta t \leq \gamma\tau_0$, although the monotonicity property (40) is violated, see Hundsdorfer & Ruuth (2003, 2006), Ruuth & Hundsdorfer (2005). It would be of much interest to know whether similar results are possible for other LMMs, as well as for RKMs and more general GLMs.

In Chapter I we answer these questions. We present a generic framework for deriving best possible stepsize conditions which guarantee boundedness of actual RKMs, LMMs and GLMs. Besides being helpful in finding stepsize conditions that are sufficient for boundedness, the framework leads to necessary conditions as well. The contents of this chapter are equal to W. Hundsdorfer, A. Mozartova, M.N. Spijker: Stepsize conditions for boundedness in numerical initial value problems, SIAM J. Numer. Anal. 47 (2009), 3797–3819.

Chapter II: Special boundedness properties in numerical initial value problems

The fact that for many useful GLMs there exists *no* $\gamma > 0$ such that (45) is present, has led us to study in Chapter I general bounds similar to (43), which are formally weaker than (45) but still useful because they can reveal essential

boundedness properties of the numerical methods, like TVB. These general bounds are relevant in cases where the monotonicity property (46) is violated. But they suffer still from the following two inconveniences: (1) the corresponding stepsize conditions, of type $0 < \Delta t \leq \gamma \cdot \tau_0$, involve complicated conditions on γ which are often difficult to check in practice; (2) the general bounds are relevant to seminorms but not to the wider class of convex functionals.

The question arises of whether some special bounds can be found which improve one or both of these two inconveniences, and which are present in cases where (46) is violated.

Chapter II is essentially addressed to this question. We find special bounds which can still be present in cases where the monotonicity property (45) is violated, and which are the best possible in a definite sense. Moreover, these special bounds are relevant to a class of functionals $||\cdot||$ that is wider than the class of seminorms. Finally, and most importantly in view of applications, the corresponding stepsize conditions $0 < \Delta t \leq \gamma \cdot \tau_0$ involve a condition on γ which is easier to check in practice than the conditions relevant to the general bounds. The contents of this chapter are equal to W. Hundsdorfer, A. Mozartova, M.N. Spijker: Special boundedness properties in numerical initial value problems, BIT 51 (2011), 909-936.

Chapter III: Stepsize Restrictions for Boundedness and Monotonicity of Multistep Methods

As mentioned in Section 0.5.2, the linear multistep methods with nonnegative coefficients form only a small class, excluding many well-known methods. For instance, most explicit k-step methods of order p used in practice have $p = k$, and for such methods we cannot have $c > 0$; c.f. (41), (42). Furthermore, the sufficient conditions which were derived in Hundsdorfer & Ruuth (2006) and Hundsdorfer, Ruuth & Spiteri (2003) for boundedness (43) are not very transparent and not easy to verify for given methods.

The generic framework presented in Chapter I, for deriving best possible stepsize conditions which guarantee boundedness of GLMs, can be used to obtain conditions for boundedness (43) of linear multistep methods. These conditions are not only sufficient but also necessary. Moreover, these conditions are more simple than the sufficient coonditions in Hundsdorfer & Ruuth (2006) and Hundsdorfer, Ruuth & Spiteri (2003).

In practice, the starting values are not arbitrary, of course. From a given u_0, the vectors $u_1, \ldots, u_{k-1}$ can be computed by a Runge-Kutta method. For such combinations of linear multistep methods and Runge-Kutta starting procedures the monotonicity property can be studied in the form (3) under the stepsize restriction $\Delta t \leq \gamma \tau_0$.

Consider the AB2 method (32) and BDF2 method (34). These methods have negative coefficients, therefore we cannot conclude (40) (for arbitrary (semi-)norm or sublinear functional) for these methods. But as seen from the illustrations in Section 0.4, the numerical solutions of these methods can have nice

monotone behaviour with suitable starting procedures, see Figure 2 and 6.

The natural question arises for combinations of linear multistep methods and Runge-Kutta starting procedures of whether stepsize restrictions $\Delta t \leq \gamma \tau_0$ can be established which guarantee the monotonicity property (3). Some partial results in this direction were obtained in Hundsdorfer, Ruuth & Spiteri (2003) for some explicit two-step methods. One may wonder whether these results can be generalized for arbitrary LMMs.

In Chapter III we address these questions. Using the framework of Chapter I, we obtain necessary and sufficient conditions for boundedness. These conditions are relatively transparent and easy to verify numerically for given classes of methods. We will also give conditions that ensure monotonicity – as in (3) – for combinations of linear multistep methods and Runge-Kutta starting procedures. The contents of this chapter are equal to W. Hundsdorfer, A. Mozartova, M.N. Spijker: Stepsize restrictions for boundedness and monotonicity of multistep methods, J. Sci. Comput. 50 (2012), 265-286.

Chapter IV: Comparison of Boundedness and Monotonicity Properties of One-Leg and Linear Multistep Methods

Instead of linear multistep methods, boundedness can be considered for the related class of one-leg methods. With given coefficients a_j, $1 \leq j \leq k$, and b_j, $0 \leq j \leq k$, as for a linear multistep method (39), the corresponding *k-step one-leg method* can be written in the form

$$u_n = \sum_{j=1}^{k} a_j u_{n-j} + \Delta t \beta F(v_n)\,, \qquad v_n = \sum_{j=0}^{k} \hat{b}_j u_{n-j} \tag{48}$$

for $n \geq k$, with a natural scaling for one-leg methods: $\hat{b}_j = b_j/\beta$ and $\beta = \sum_{j=0}^{k} b_j$, $\beta \neq 0$. The starting values $u_0, u_1, \ldots, u_{k-1} \in \mathbb{V}$ are supposed to be given, or computed by a Runge-Kutta method. These methods were originally introduced in Dahlquist (1976) to facilitate the analysis of linear multistep methods. Stability results with inner-product norms for one-leg methods often have a somewhat nicer form than for linear multistep methods; see e.g. Butcher (2003), Hairer & Wanner (1996). A question is whether the one-leg analysis can give results in a nicer form than for the linear multistep methods.

In Chapter IV we study boundedness for one-leg methods, in the sense of (43). Using results for linear multistep methods of Chapter III, it will be shown that the maximal stepsize coefficient for boundedness of a one-leg method is the same as for the associated linear multistep method. Simplification of the analysis is not achieved with one-leg methods.

In view of the close connection between one-leg and linear multistep methods, it is not very surprising that the stepsize coefficients for boundedness are the same. However, it will be also shown that combinations of one-leg methods and Runge-Kutta starting procedures may give different, and possibly larger,

stepsize coefficients for monotonicity than with the linear multistep methods and the same starting procedures. These results will be worked out in detail for the class of explicit two-step methods.

For a more detailed introduction to the topics of this thesis, and for related literature, we refer to the beginning of each chapter.

Chapter 1

Stepsize Conditions for Boundedness in Numerical Initial Value Problems

For Runge-Kutta methods, linear multistep methods and classes of general linear methods much attention has been paid, in the literature, to special nonlinear stability requirements indicated by the terms total variation diminishing, strong stability preserving and monotonicity. Stepsize conditions, guaranteeing these properties, were derived by Shu & Osher (1988) and in numerous subsequent papers. These special stability requirements imply essential boundedness properties for the numerical methods, among which the property of being total variation bounded. Unfortunately, for many well-known methods, the above special requirements are violated, so that one cannot conclude in this way that the methods are (total variation)bounded.

In this chapter, we focus on stepsize conditions for boundedness directly, rather than via the detour of the above special stability properties. We present a generic framework for deriving best possible stepsize conditions which guarantee boundedness of actual RKMs, LMMs and GLMs, thereby generalizing results on the special stability properties mentioned above.

1.1 Introduction

1.1.1 Monotonicity and boundedness

Consider an initial value problem, for a system of ordinary differential equations, of type

$$\frac{d}{dt}u(t) = F(t, u(t)) \quad (t \geq 0), \quad u(0) = u_0. \tag{1.1}$$

In this chapter we study step-by-step-methods for computing numerical approximations u_n to the true solution values $u(n\Delta t)$, where Δt denotes a positive stepsize and $n = 1, 2, 3, ...$

Monotonicity of Runge-Kutta methods

The general Runge-Kutta method (RKM), for computing u_n, can be written in the form

$$v_i^{[n]} = u_{n-1} + \Delta t \sum_{j=1}^{s} a_{ij} F((n-1+c_j)\Delta t, v_j^{[n]}) \quad (1 \le i \le s+1), \tag{1.2a}$$

$$u_n = v_{s+1}^{[n]}. \tag{1.2b}$$

Here a_{ij} and c_j are parameters defining the method, whereas $v_i^{[n]}$ $(1 \le i \le s)$ are intermediate approximations used for computing $u_n = v_{s+1}^{[n]}$ from u_{n-1} $(n = 1, 2, 3, \ldots)$, cf. e.g. Butcher (1987) or Hairer, Nørsett & Wanner (1987). If $a_{ij} = 0$ (for $j \ge i$), the method is called *explicit*.

In the following, $\mathbb{V}$ stands for the vector space on which the differential equation is defined, and $\|\cdot\|$ denotes a seminorm on $\mathbb{V}$ (i.e.: $\|u+v\| \le \|u\|+\|v\|$ and $\|\lambda\, v\| = |\lambda|\,\|v\|$ for all $u, v \in \mathbb{V}$ and real λ). Much attention has been paid in the literature to the property

$$\|v_i^{[n]}\| \le \|u_{n-1}\| \quad (\text{for } 1 \le i \le s+1). \tag{1.3}$$

Clearly, (1.3) implies $\|u_n\| \le \|u_{n-1}\|$. The last inequality, as well as property (1.3), is often referred to by the term *monotonicity* or *strong stability*; it is of particular importance in situations where (1.1) results from (method of lines) semidiscretizations of time-dependent partial differential equations. Choices for $\|\cdot\|$ which occur in that context, include e.g. the *supremum norm* $\|x\| = \|x\|_\infty = \sup_i |\xi_i|$ and the *total variation seminorm* $\|x\| = \|x\|_{TV} = \sum_i |\xi_{i+1} - \xi_i|$ (for vectors x with components ξ_i).

Numerical processes, satisfying $\|u_n\|_{TV} \le \|u_{n-1}\|_{TV}$, play a special role in the solution of hyperbolic conservation laws and are called *total variation diminishing* (TVD), cf. e.g. Harten (1983), Shu (1988), Shu & Osher (1988), LeVeque (2002), Hundsdorfer & Verwer (2003). For such processes there is, trivially, *total variation boundedness* (TVB), in that there is a finite value μ such that, for all $n \ge 1$,

$$\|u_n\|_{TV} \le \mu \cdot \|u_0\|_{TV}. \tag{1.4}$$

Satisfying (1.4) is of crucial importance for suitable convergence properties when $\Delta t \to 0$, and constitutes one of the underlying reasons why attention has been paid in the literature to (1.3), see e.g. LeVeque (2002).

Conditions on Δt which guarantee (1.3) were given in the literature, mainly for autonomous differential equations (i.e. F is independent of t). These conditions apply, however, equally well to general F and we discuss them below for that case. In many papers, one starts from an assumption about F which, for given $\tau_0 > 0$, essentially amounts to

$$\|v + \tau_0\, F(t, v)\| \le \|v\| \quad (\text{for } t \in \mathbb{R},\, v \in \mathbb{V}). \tag{1.5}$$

Assumption (1.5) means that the forward Euler method is monotonic with stepsize τ_0. It can be interpreted as a condition on the manner in which the semidiscretization is performed, in case $\frac{d}{dt}u(t) = F(t, u(t))$ stands for a semidiscrete version of a partial differential equation.

For classes of RKMs, positive *stepsize-coefficients* γ were determined, such that monotonicity, in the sense of (1.3), is present for all Δt with

$$0 < \Delta t \leq \gamma \cdot \tau_0, \tag{1.6}$$

see e.g. Shu & Osher (1988), Gottlieb, Shu & Tadmor (2001), Spiteri & Ruuth (2002), Ferracina & Spijker (2004, 2005), Higueras (2004, 2005), Ruuth (2006), Spijker (2007, Section 3.2.1).

Monotonicity of linear multistep methods

The linear multistep method (LMM), for computing u_n, can be written in the form

$$u_n = \sum_{j=1}^{k} a_j \, u_{n-j} \; + \; \Delta t \cdot \sum_{j=0}^{k} b_j \, F((n-j)\,\Delta t, \, u_{n-j}), \tag{1.7}$$

where the parameters a_j, b_j define the method, $\sum a_j = 1$ - cf. e.g. Butcher (1987), Hairer, Nørsett & Wanner (1993). If $b_0 = 0$, the method is called *explicit*.

For method (1.7), a study was made of monotonicity, in the sense of the inequality

$$\|u_n\| \leq \max_{1 \leq j \leq k} \|u_{n-j}\|. \tag{1.8}$$

For classes of LMMs, positive stepsize-coefficients γ were determined, with the property that (1.5), (1.6) guarantee (1.8), see e.g. Shu (1988), Gottlieb, Shu & Tadmor (2001), Hundsdorfer & Ruuth (2003), Spijker (2007, Section 3.2.2). Clearly, (1.8) with $\|\cdot\| = \|\cdot\|_{TV}$ implies again (trivially) a TVB-property, in that there is a finite μ such that, for all $n \geq k$,

$$\|u_n\|_{TV} \leq \mu \cdot \max_{0 \leq j \leq k-1} \|u_j\|_{TV}. \tag{1.9}$$

Boundedness

Unfortunately, there are well known RKMs and LMMs, with a record of practical success, for which there exist *no positive stepsize-coefficients* γ such that (1.5), (1.6) always imply (1.3) or (1.8), respectively. Examples are the Adams methods and BDFs with $k \geq 2$ as well as the Dormand-Prince formula, cf. e.g. Hairer, Nørsett & Wanner (1993). Moreover, no second order (implicit) RKMs or LMMs exist with $\gamma = \infty$, see e.g. Spijker (1983, Sections 2.2, 3.2). These circumstances suggest that there are situations where monotonicity may be too strong a theoretical demand, and that it is worthwhile to study, along

with monotonicity, also directly the following weaker *boundedness properties* for methods (1.2) and (1.7), respectively:

$$\|v_i^{[n]}\| \leq \mu \cdot \|u_0\| \quad \text{(for } 1 \leq i \leq s+1 \text{ and all } n \geq 1\text{)}, \tag{1.10}$$

$$\|u_n\| \leq \mu \cdot \max_{0 \leq j \leq k-1} \|u_j\| \quad \text{(for all } n \geq k\text{)}. \tag{1.11}$$

Here μ stands for a finite constant (independent of n) which is allowed to be greater than 1. The requirements (1.10), (1.11), with $\|\cdot\| = \|\cdot\|_{TV}$, still imply the TVB-property – which highlights the importance of studying (1.10), (1.11).

Recently – see Hundsdorfer & Ruuth (2003, 2006), Ruuth & Hundsdorfer (2005) – some special LMMs were found with a positive stepsize-coefficient γ such that (1.11) holds under conditions (1.5), (1.6), although (1.8) is violated. The question of whether similar results are possible for other LMMs, as well as for step-by-step methods of a different kind, seems not to have been considered in the literature thus far.

1.1.2 Scope of the chapter

Boundedness of general linear methods

We recall that LMMs and RKMs are examples of methods belonging to the important and very large class of *general linear methods* (GLMs), introduced by Butcher (1966), and studied extensively in the literature – see e.g. Butcher (1987, 2003), Hairer, Nørsett & Wanner (1993), Hairer & Wanner (1996), and the references therein.

In this chapter, we shall consider, for GLMs, boundedness properties, similar to (1.10), (1.11). A generic framework will be presented which facilitates the computation of stepsize-coefficients γ related to such properties.

The theory in the present chapter of the thesis can be viewed as a (nontrivial) extension of an approach to monotonicity of GLMs given earlier in the literature, cf. Spijker (2007). Its usefulness will be illustrated briefly in the present chapter of the thesis, whereas in Chapters 3 and 4 the theory will be applied in a more general analysis for classes of GLMs.

Organization of the chapter

Section 1.2 deals with stepsize-coefficients γ related to explicit bounds for the output vectors of a generic numerical process. Our main theorems, Theorems 1.2.2 and 1.2.4, provide an algebraic criterion in terms of γ, viz. (1.23), for these bounds to be valid in situations of practical relevance.

In Section 1.3, we give results related to Theorems 1.2.2 and 1.2.4. In Section 1.3.1, we apply the theorems so as to obtain simplified conditions for bounding the generic process. We also recover easily a concise criterion for monotonicity obtained earlier in the literature (but derived differently), cf. Spijker (2007). In Section 1.3.2, a lemma is presented which is helpful when applying the main

theorems in the boundedness analysis of actual GLMs. In Section 1.3.3, we illustrate the significance of the general theory shortly, by applying it in resolving the question of boundedness for some concrete numerical methods.

In Section 1.4 we give the proofs of Theorems 1.2.2, 1.2.4.

1.2 Bounds for a generic numerical process

In this section, we shall study bounds for the output vectors of a generic numerical process. We are interested in these bounds, primarily because they facilitate significantly the derivation of actual boundedness results for given GLMs. In Section 1.2.1 we first describe GLMs, whereas in Section 1.2.2 we introduce the generic numerical process and relate it to GLMs. In the Sections 1.2.3 and 1.2.4 we present criteria for the existence of the above mentioned bounds for the generic process.

In all of the following, $\mathbb{V}$ denotes again the vector space on which the differential equation is defined, and $\|\cdot\|$ stands for an arbitrary given seminorm on $\mathbb{V}$.

1.2.1 General linear methods

The general linear method, for solving (1.1), depends on parameters c_j $(1 \leq j \leq q)$ and parameter matrices $A = (\alpha_{ij}) \in \mathbb{R}^{q\times l}$, $B = (\beta_{ij}) \in \mathbb{R}^{q\times q}$, where $1 \leq l \leq q$. The method can be written in the following form:

$$v_i^{[n]} = \sum_{j=1}^{l} \alpha_{ij} u_j^{[n-1]} + \Delta t \sum_{j=1}^{q} \beta_{ij} F((n-1+c_j)\Delta t, v_j^{[n]}) \qquad (1 \leq i \leq q), \quad (1.12a)$$

$$u_i^{[n]} = v_{q-l+i}^{[n]} \qquad (1 \leq i \leq l). \quad (1.12b)$$

Here $u_i^{[n-1]}$ are input vectors available at the n-th step of the method, whereas $v_i^{[n]}$ are (intermediate) approximations used for computing the input vectors $u_i^{[n]}$ for the next step ($n = 1, 2, 3, \ldots$); cf. e.g. Butcher (1966), Butcher (1987, pp. 338).

Obviously, the Runge-Kutta method (1.2) is an example of (1.12), with $l = 1$, $q = s+1$, $u_1^{[n]} = u_n \simeq u(n \cdot \Delta t)$ and $\alpha_{i1} = 1$, $\beta_{ij} = a_{ij}$ (for $1 \leq j \leq s$), $\beta_{ij} = 0$ (for $j = s+1$).

The linear multistep method (1.7) is another example of (1.12), with $l = k$, $q = k+1$ and $u_i^{[n]} = u_{n-1+i}$ $(1 \leq i \leq k,\ n \geq 0)$, $v_i^{[n]} = u_{n-2+i}$ $(1 \leq i \leq k+1,\ n \geq 1)$. Method (1.7) can be written in the form (1.12) with $c_j = j-1$, $A = \binom{I}{a}$, $B = \binom{O}{b}$, where I denotes the $k \times k$ identity matrix, O the $k \times (k+1)$ zero matrix and $a = (a_k, \ldots, a_1)$, $b = (b_k, \ldots, b_0)$.

For completeness, we note that GLMs are often represented differently from (1.12), viz. in a partitioned form with parameters $u_{ij}, v_{ij}, a_{ij}, b_{ij}, c_j$, as follows:

$$Y_i = \sum_{j=1}^{l} u_{ij} y_j^{[n-1]} + \Delta t \sum_{j=1}^{s} a_{ij} F((n-1+c_j)\Delta t, Y_j) \qquad (1 \le i \le s), \qquad (1.13a)$$

$$y_i^{[n]} = \sum_{j=1}^{l} v_{ij} y_j^{[n-1]} + \Delta t \sum_{j=1}^{s} b_{ij} F((n-1+c_j)\Delta t, Y_j) \qquad (1 \le i \le l), \qquad (1.13b)$$

see e.g. Hairer & Wanner (1991, p. 313), Butcher (2003, p. 358). Here s is the number of internal approximations Y_i, and l is again the number of vectors $y_i^{[n]}$ which propagate from step to step. Clearly, (1.13) is formally of type (1.12) with $q = l + s$ and $u_i^{[n]}$, $v_i^{[n]}$ defined, with obvious vector notations, by $u^{[n]} = y^{[n]}$, $v^{[n]} = \binom{Y}{y^{[n]}}$. In this chapter, we aim at bounding simultaneously Y and $y^{[n]}$, in terms of $y^{[0]}$, so that we find it convenient to use a representation of the GLM in which Y and $y^{[n]}$ are lumped together. In the following, we shall thus deal with representation (1.12) rather than (1.13).

Definition 1.2.1. (Boundedness of general linear methods). *We define method (1.12) to be bounded, with constant μ (for given stepsize Δt, vector space $\mathbb{V}$, seminorm $\|.\|$ and function F), if for all $N \ge 1$ we have*

$$\|v_i^{[n]}\| \ \le \ \mu \cdot \max_{1 \le j \le l} \|u_j^{[0]}\| \quad (\textit{for } 1 \le n \le N \textit{ and } 1 \le i \le q), \qquad (1.14)$$

whenever $u_i^{[n-1]}$, $u_i^{[n]}$, $v_i^{[n]} \in \mathbb{V}$ satisfy (1.12) (for $1 \le n \le N$).

Note that (1.14) implies (1.10) or (1.11), respectively, if method (1.12) stands for a RKM or LMM in the way indicated above.

1.2.2 A generic numerical process, with a simple form

For studying boundedness of (1.12), it is convenient to represent in a concise form all relations, involved in specifying $v_i^{[N]}$ (for any given $N \ge 1$). We describe now a standard representation of N consecutive steps of the GLM, to which we will refer in the following as the *canonical representation*. We combine all vectors $v_i^{[n]}$ (with $1 \le i \le q$ and $1 \le n \le N$) into one single vector $y = [y_i] \in \mathbb{V}^m$, where $m = N \cdot q$, and $y_i \in \mathbb{V}$ $(1 \le i \le m)$. Furthermore, we introduce shorthand notations for $u_i^{[0]}$ and $F((n-1+c_i)\Delta t,\, v)$. Defining, for $1 \le i \le l$ and $1 \le j \le q$,

$$x_i = u_i^{[0]}, \ \ y_{(n-1)q+j} = v_j^{[n]}, \ \ F_{(n-1)q+j}(v) = F((n-1+c_j)\Delta t, v), \qquad (1.15)$$

we can rewrite the relations (1.12) (for $n = 1, \ldots N$) in the following form:

$$y_i = \sum_{j=1}^{l} s_{ij}\, x_j \ + \ \Delta t \cdot \sum_{j=1}^{m} t_{ij}\, F_j(y_j) \quad (1 \le i \le m). \qquad (1.16)$$

To specify the coefficient matrices $S = (s_{ij}) \in \mathbb{R}^{m\times l}$, $T = (t_{ij}) \in \mathbb{R}^{m\times m}$, we denote the matrices consisting of the last l rows of $A = (\alpha_{ij})$ and $B = (\beta_{ij})$ by A_0 and B_0, respectively. It can be seen that S is made up of $q \times l$ blocks S_n, and T of $q \times q$ blocks $T_{n,j}$ $(1 \le n \le N,\ 1 \le j \le N)$, where

$$S_n = A\,(A_0)^{n-1}, \tag{1.17a}$$

$$T_{n,j} = O\ \ (j > n), \quad T_{n,n} = B, \quad T_{n,j} = A\,(A_0)^{n-j-1}B_0\ \ (n > j). \tag{1.17b}$$

Furthermore, when $F : \mathbb{R} \times \mathbb{V} \to \mathbb{V}$ satisfies (1.5), then definition (1.15) implies

$$\|v + \tau_0\, F_i(v)\| \ \le\ \|v\| \quad (\text{for } 1 \le\ i \le\ m,\ \text{and } v \in \mathbb{V}). \tag{1.18}$$

For analysing boundedness of (1.12), it is sometimes also handy to use *non-canonical representations*, of N steps of the method, cf. e.g. Section 1.3.3. Such representations share with the canonical representation the form (1.16), with property (1.18), but violate (1.17). Therefore, unless specified otherwise, in the following discussion of (1.16) we shall *not* assume S, T to satisfy (1.17), so that the conclusions, to be obtained about (1.16), can be applied both to canonical and non-canonical representations of method (1.12).

We shall interpret $x_i \in \mathbb{V}$ and $y_i \in \mathbb{V}$ as *input* and *output vectors*, respectively, of the *generic process* (1.16). In the situation (1.16), (1.18), we shall focus on the bound

$$\|y_i\| \ \le\ \mu \cdot \max_{1\le j\le l} \|x_j\| \quad (\text{for } 1 \le i \le m). \tag{1.19}$$

We shall say that *process (1.16) satisfies the bound (1.19)* (for given stepsize Δt, vector space $\mathbb{V}$, seminorm $\|.\|$ and functions $F_i : \mathbb{V} \to \mathbb{V}$), if (1.19) holds whenever x_i and $y_i \in \mathbb{V}$ satisfy (1.16).

Clearly when (1.16) stands, as above, for N versions of (1.12) via the relations (1.15), (1.17), then boundedness of the GLM, defined in Section 1.2.1, corresponds to the situation where process (1.16) satisfies the bound (1.19) - with constant μ independent of $N = 1, 2, 3, \ldots$.

In Sections 1.2.3, 1.2.4, we shall present, without proof, the basic results of the chapter, Theorems 1.2.2, 1.2.4. The theorems give conditions, on the ratio $\Delta t/\tau_0$, in order that process (1.16), with arbitrary parameter matrices $S = (s_{ij})$, $T = (t_{ij})$, satisfies the bound (1.19).

1.2.3 Satisfying the bound (1.19) for arbitrary functions F_i

In this subsection, we shall give our first main result, Theorem 1.2.2. The theorem deals with γ and μ such that the following general and fundamental property is present:

> Condition $0 < \Delta t \le \gamma \cdot \tau_0$ implies that process (1.16) satisfies the bound (1.19), whenever $\mathbb{V}$ is a vector space with seminorm $\|\cdot\|$, and arbitrary functions $F_i : \mathbb{V} \to \mathbb{V}$ satisfy (1.18). (1.20)

Theorem 1.2.2 concerns not only the above property (1.20), but also the following weaker property (1.21), in which the focus is on the *maximum norm*, defined by $\|x\|_\infty = \max_i |\xi_i|$ (for vectors $x \in \mathbb{R}^m$ with components ξ_i).

> Condition $\Delta t = \gamma \cdot \tau_0$ implies that process (1.16) satisfies the bound (1.19), when $\mathbb{V} = \mathbb{R}^m$, $\|\cdot\| = \|\cdot\|_\infty$, and arbitrary $F_i : \mathbb{R}^m \to \mathbb{R}^m$ satisfy (1.18). (1.21)

The theorem below will show that the general property (1.20) is already present as soon as (1.21) is in force. Moreover, the theorem will give an algebraic criterion, in terms of γ, μ, for (1.20), (1.21) to be valid.

In formulating the criterion we need some further notations. For any $m \times k$ matrix $A = (a_{ij})$, we put $\|A\|_\infty = \max_{x \neq 0} \frac{\|Ax\|_\infty}{\|x\|_\infty}$ and we recall the well known formula

$$\|A\|_\infty = \max_i \sum_j |a_{ij}|.$$

We define $|A| = (|a_{ij}|)$, and denote the *spectral radius* of square matrices A by $\mathrm{spr}(A)$.

For values γ such that $I + \gamma T$ is invertible, we introduce the matrices

$$Q = (q_{ij}) = (I + \gamma T)^{-1}, \quad P = (p_{ij}) = Q(\gamma T), \quad R = (r_{ij}) = Q S. \tag{1.22}$$

Our criterion –for properties (1.20), (1.21)– involves the following requirements:

$$I + \gamma T \text{ is invertible,} \tag{1.23a}$$
$$\mathrm{spr}(|P|) < 1, \tag{1.23b}$$
$$\| (I - |P|)^{-1} |R| \|_\infty \leq \mu. \tag{1.23c}$$

Theorem 1.2.2. (Criterion for the bound (1.19), when arbitrary F_i satisfy (1.18)). *Consider process (1.16), with arbitrary coefficient matrices $S = (s_{ij})$ and $T = (t_{ij})$, and let positive τ_0, γ, μ be given. Then condition (1.23) is necessary and sufficient for property (1.20), as well as for (1.21).*

Since property (1.20) is a-priori stronger than (1.21), the essence of the above theorem is that the algebraic condition (1.23) implies the (strong) statement (1.20), whereas already the (weaker) statement (1.21) implies (1.23).

Clearly, when γ satisfies (1.23a), (1.23b), the theorem shows that the smallest μ, for which statements (1.20), (1.21) hold, is equal to

$$\mu = \| (I - |P|)^{-1} |R| \|_\infty. \tag{1.24}$$

In many practical situations, condition (1.23c) is the essential requirement rather than conditions (1.23a) or (1.23b). One easily sees that the last two conditions will be satisfied, with any $\gamma > 0$, if T is lower triangular with nonnegative

diagonal entries. This applies notably to the situation where T is strictly lower triangular, which corresponds to a numerical process that is explicit.

Finally, we note that when, for a given GLM, boundedness (in the sense of Definition 1.2.1) is analysed via the canonical representation, one arrives by Theorem 1.2.2 at requirement (2.12c) *uniformly* for $m = Nq$, with $N = 1, 2, 3, \ldots$ This is in general not easy to verify. More simple conditions and applications will be presented in Section 1.3. When the matrices P, R only depend on N, γ and the coefficients of the underlying GLM (as is the case in the canonical representation), the stepsize-coefficient γ for boundedness only depends on the method, and not on the class of problems under consideration, characterized by τ_0 in (1.5) or (1.18).

1.2.4 Satisfying the bound (1.19) for restricted functions F_i

Our second main result, Theorem 1.2.4 below, deals with important situations not adequately covered by Theorem 1.2.2. It is often *not* natural to allow – as in Theorem 1.2.2 – that all functions F_i are different from each other.

For instance, if in (1.12) we have $c_i = c_j$ for some $i \neq j$, or if the differential equation is autonomous, then N successive applications of (1.12) are represented canonically – via (1.15), (1.17) – by a process 1.16 with $F_i = F_j$ for some, or all, indices $i \neq j$.

Also when $c_i \neq c_j$ (for all $i \neq j$), and the differential equation is *non*-autonomous, it can happen that the canonical representation, obtained via (1.15), (1.17), amounts to a process (1.16) with $F_i = F_j$ for some indices $i \neq j$. According to (1.15), this situation occurs as soon as $n_1 + c_i = n_2 + c_j$ for some n_1, n_2, i, j with $n_1\, q + i \neq n_2\, q + j$. When a general LMM, cf. (1.7), is represented as a GLM as indicated in Section 1.2.1, then $N \geq 2$ applications of the GLM provide an example of this situation.

Below we shall see that, in cases where some of the functions F_i are equal to each other, condition (1.23) can be an unnecessarily strong requirement on γ in order that the stepsize restriction $0 < \Delta t \leq \gamma \cdot \tau_0$ implies the bound (1.19).

In order to describe general situations where some of the functions F_i are equal to each other, we consider index sets $\mathcal{I}_\rho$ with $\mathcal{I}_\rho \subset \{1, \ldots, m\}$ (for $1 \leq \rho \leq r$), and functions $F_i : \mathbb{V} \to \mathbb{V}$ (for $1 \leq i \leq m$), such that

$$\mathcal{I}_1, \ldots, \mathcal{I}_r \text{ are nonempty and mutually disjoint, with } \cup_{\rho=1}^{r} \mathcal{I}_\rho = \{1, \ldots, m\}, \tag{1.25}$$

$$F_i = F_j \text{ whenever } i \text{ and } j \text{ belong to the same index set } \mathcal{I}_\rho. \tag{1.26}$$

Below, we shall deal with the following variant of property (1.20), in which the functions F_i are restricted according to (1.26):

$$\text{Condition } 0 < \Delta t \leq \gamma \cdot \tau_0 \text{ implies that process (1.16) satisfies the bound (1.19), whenever } \mathbb{V} \text{ is a vector space with seminorm } \|\cdot\|, \text{ and functions } F_i : \mathbb{V} \to \mathbb{V} \text{ satisfy (1.18), (1.26).} \tag{1.27}$$

We will see that finding a criterion for (1.27) is more subtle an issue than for (1.20). It will turn out to be convenient to consider, in addition to the above property (1.27), the following weaker version:

> Condition $\Delta t = \gamma \cdot \tau_0$ implies that process (1.16) satisfies the bound (1.19), whenever $\mathbb{V} = \mathbb{R}^m$ with seminorm $\|\cdot\|$, and $F_i : \mathbb{R}^m \to \mathbb{R}^m$ satisfy (1.18), (1.26). (1.28)

Note that, because arbitrary seminorms occur in (1.28), this weaker version is *not* related to the original property (1.27), in the same way as the weaker version (1.21) is related to (1.20). An adaptation of (1.21), for the situation at hand, reads as follows:

> Condition $\Delta t = \gamma \cdot \tau_0$ implies that process (1.16) satisfies the bound (1.19), when $\mathbb{V} = \mathbb{R}^m$, $\|\cdot\| = \|\cdot\|_\infty$, and $F_i : \mathbb{R}^m \to \mathbb{R}^m$ satisfy (1.18), (1.26). (1.29)

By Theorem 1.2.2, condition (1.23) is still *sufficient* in order that (1.27), (1.28), (1.29) hold. But, the following simple Example 1.2.3 shows that the condition is *no longer necessary* – cf. also Section 1.3.3 for a more natural, but less simple, counterexample.

Example 1.2.3. Consider process 1.16 with $l = 1$, $m = 2$ and $s_{i,1} = 1$, $t_{i,1} = 3$, $t_{i,2} = -2$. Suppose (1.25), (1.26) with $r = 1$, $\mathfrak{I}_1 = \{1, 2\}$, i.e. $F_1 = F_2$, and consider $\gamma \geq 1/4$.

One easily sees that requirement (1.23a) is fulfilled, and $\mathrm{spr}(|P|) \geq 1$. Therefore, condition (1.23b) is violated.

On the other hand, the process at hand is nothing but the (backward Euler) method $y_2 = y_1 = x_1 + \Delta t\, F_1(y_1)$, which is of the form (1.16) with $\tilde{l} = \widetilde{m} = 1$ and $\tilde{S} = 1$, $\widetilde{T} = 1$. Condition (1.23) is fulfilled by $\tilde{S}$, $\widetilde{T}$, with $\mu = 1$, for any $\gamma > 0$.

In line with Theorem 1.2.2 (applied with $\tilde{S}$, $\widetilde{T}$), we can conclude that the original process (with $m = 2$) must have property (1.27), with $\mu = 1$, for any $\gamma > 0$, although (1.23) is violated for $\gamma \geq 1/4$. ◇

In the following, we will see that violation of condition (1.23) while (1.27) is valid – as in the above example – is a phenomenon related to reducibility of the generic process (1.16). We will deal below with two irreducibility assumptions under which (1.23) cannot be violated.

In formulating these assumptions, we denote the i-th row and j-th column of any matrix A by $A(i,:)$ and $A(:,j)$, respectively. By $\widehat{T} = (\hat{t}_{ij})$ we denote the matrix defined by

$$\hat{t}_{ij} = t_{ij} \quad (\text{if } S(j,:) \neq 0), \qquad \hat{t}_{ij} = 0 \quad (\text{if } S(j,:) = 0).$$

By $[\,S\;T\,]$ and $[\,S\;\widehat{T}\,]$ we denote the $m \times (l+m)$ matrices whose first l columns equal those of S, and last m columns equal those of T and $\widehat{T}$, respectively.

We will use the *irreducibility assumption*

$$[S\ T](i,:) \neq [S\ T](j,:) \quad \text{(if } i \neq j \text{ are in the same } \mathcal{I}_\rho \text{ and } T(:,i) \neq 0,\ T(:,j) \neq 0\text{)}, \tag{1.30}$$

as well as the slightly stronger assumption

$$[S\ \widehat{T}](i,:) \neq [S\ \widehat{T}](j,:) \quad \text{(if } i \neq j \text{ are in the same } \mathcal{I}_\rho \text{ and } T(:,i) \neq 0,\ T(:,j) \neq 0\text{)}. \tag{1.31}$$

Clearly, if $r < m$ and there is *no* irreducibility in the sense of (1.30), then process (1.16) –with F_i satisfying (1.26)– is equivalent to a process (1.16) with a smaller value of m.

Theorem 1.2.4. (Criterion for the bound (1.19), when F_i satisfy (1.18), (1.26)). *Consider process (1.16), with arbitrary coefficient matrices $S = (s_{ij})$ and $T = (t_{ij})$. Let positive τ_0, γ, μ be given, and assume (1.25).*

(i) Assume irreducibility in the sense of (1.30). Then condition (1.23) is necessary and sufficient for property (1.27), as well as for (1.28).

(ii) Assume irreducibility in the sense of (1.31). Then condition (1.23) is necessary and sufficient for property (1.27), as well as for (1.29).

The above statement (i) shows that, under the irreducibility assumption (1.30), property (1.28) implies the algebraic property (1.23). On the other hand, statement (ii) reveals that under the stronger irreducibility assumption (1.31), already the weaker property (1.29) implies (1.23). The natural question thus arises of whether statements (i), (ii) can be combined and strengthened into the following proposition:

(iii) Assume irreducibility in the sense of (1.30). Then condition (1.23) is necessary and sufficient for property (1.27), as well as for (1.29).

The following counterexample answers the above question in the negative: statement (iii) is in general *not* true!

Example 1.2.5. Consider process 1.16 with $l = 1$, $m = 3$ and $s_{1,1} = 0$, $s_{2,1} = s_{3,1} = 1$, $t_{i,1} = i$, $t_{1,2} = t_{1,3} = 0$, $t_{2,2} = t_{3,2} = 3$, $t_{2,3} = t_{3,3} = -2$. Suppose (1.25), (1.26) with $r = 1$, $\mathcal{I}_1 = \{1, 2, 3\}$, i.e. $F_1 = F_2 = F_3$, and consider $\gamma = 1/4$.

The irreducibility assumption (1.30) is fulfilled. Furthermore, one easily sees that (1.23a) is fulfilled, but $\mathrm{spr}(|P|) = 1$. Therefore, condition (1.23) is violated.

On the other hand, for $\Delta t = \tau_0/4$ and $\mathbb{V}$, $\|\cdot\|$, F_i as in (1.29), it can be seen that $\|y_1\| = \|\Delta t\, F(y_1)\| = 0$, $\|y_2\| = \|y_3\| \leq \|x_1\|$. With $\mu = 1$, we thus have property (1.29). ◇

Theorem 1.2.2 can formally be viewed as a special case of Theorem 1.2.4 –the latter theorem, with $r = m$ and the trivial index sets $\mathcal{I}_\rho = \{\rho\}$, implies the former. We have formulated Theorem 1.2.2 separately in view of its importance

and simplicity: it does not need (1.25), (1.26) nor (1.30), (1.31). Moreover, by formulating first Theorem 1.2.2 explicitly, we could show in a natural way, via Example 1.2.3, that some additional (irreducibility) assumption is needed in order that condition (1.23) is the appropriate criterion when some F_i are equal.

1.3 Results related to the main theorems

1.3.1 Alternative conditions for properties (1.20), (1.27)

In this section we study process (1.16) with arbitrary coefficient matrices $S = (s_{ij})$ and $T = (t_{ij})$. We shall give conditions, for properties (1.20) and (1.27), which are in general simpler and easier to check than (1.23). In deriving these conditions, we shall use a lemma about condition (1.33b) which will be presented first in Section 1.3.1.

The same notations will be used as in Section 1.2, notably (1.22), and any inequalities between matrices or vectors should be understood entry-wise or component-wise, respectively.

Background regarding condition (1.23b)

The following lemma, about condition (1.23b), will be used in Sections 1.3, 1.4.

Lemma 1.3.1. (Interpretations of (1.23b). *Assume (1.23a). Then each of the following three requirements is equivalent to (1.23b).*
(i) $I - |P|$ *is invertible, with* $(I - |P|)^{-1} \geq 0$*;*
(ii) $I - |P|$ *is invertible, and* $\mathrm{spr}(|P|) \leq 1$*;*
(iii) There exist no real scalar λ *and vector* $\varphi \in \mathbb{R}^m$ *with:*

$$(\lambda I - |P|)\,\varphi = 0, \quad \varphi \neq 0, \quad \varphi \geq 0, \quad \lambda \geq 1. \tag{1.32}$$

Proof. One easily sees that (1.23b) implies each of the properties (i), (ii), (iii). Conversely, applying the Perron-Frobenius theory as presented e.g. in Horn & Johnson (1988, p. 503), it follows that (1.23b) is implied by (ii) as well as by (iii).

We shall complete the proof of the lemma by assuming (i) and proving (iii). Suppose, (iii) does *not* hold, i.e. there are λ, φ satisfying (1.32). Then $0 \geq -\varphi = (I - |P|)^{-1}\{(\lambda - 1)\,\varphi\} \geq 0$, so that $\varphi = 0$, which contradicts (1.32). □

Simplified conditions for properties (1.20), (1.27), with arbitrary μ

The following neat condition on γ will turn out to be quite useful:

$$I + \gamma T \text{ is invertible,} \tag{1.33a}$$
$$P \geq 0, \tag{1.33b}$$
$$R \geq 0. \tag{1.33c}$$

Assume (1.33a) and (1.33b) are fulfilled. We then see from Lemma 1.3.1 and the formula

$$I - P \;=\; Q \;=\; (I + \gamma\, T)^{-1} \tag{1.34}$$

(which follows from (1.22)), that condition (1.23b) is equivalent to: $\mathrm{spr}(P) \;\leq\; 1$.

For matrices $S,\ T$ satisfying (1.33), (1.23b), we have

$$\|(I - |P|)^{-1}|R|\|_\infty = \|(I-P)^{-1}R\|_\infty = \|Q^{-1}Q\,S\|_\infty = \|S\|_\infty.$$

For such matrices we have also $S = (I-P)^{-1}\,R$, with $(I-P)^{-1} \geq 0$, so that $S \geq 0$ and $\|(I-|P|)^{-1}|R|\|_\infty = \max_i \sum_j s_{ij}$.

Consequently, under assumption (1.33), the conditions (1.23b), (1.23c) are equivalent to

$$\mathrm{spr}(\,P\,) \leq 1 \qquad \text{and} \qquad \sum_j s_{i\,j} \;\leq\; \mu \quad (1 \leq i \leq m). \tag{1.35}$$

In view of this equivalency, we have the following useful corollary to Theorems 1.2.2, 1.2.4:

Corollary 1.3.2. (Criterion for properties (1.20), (1.27), when $P \geq 0$, $R \geq 0$). *Let arbitrary matrices $S = (s_{ij})$, $T = (t_{ij})$ and positive values $\tau_0,\ \gamma,\ \mu$ be given, such that (1.33) is fulfilled. Then the following two statements are valid.*
(i) Condition (1.35) is necessary and sufficient for property (1.20).
(ii) Assume (1.25), (1.30). Then (1.35) is necessary and sufficient for property (1.27).

The following corollary to Theorem 1.2.2 is useful in cases where (1.33a), (1.33b) hold, but (1.33c) is violated. It can be applied when constants $\varrho_j,\ \sigma,\ \tau$ are available such that the matrices $R = (r_{ij}),\ T = (t_{ij}),\ P = (p_{ij})$ satisfy

$$\mathrm{spr}(\,P\,) \leq 1 \quad \text{and} \quad \sum_k |r_{jk}| \leq \varrho_j, \quad \sum_j \varrho_j \leq \sigma, \quad \max_{i,\,j} |t_{ij}| \;\leq\; \tau. \tag{1.36}$$

Corollary 1.3.3. (Condition for property (1.20), when $P \geq 0$). *Let arbitrary matrices $S = (s_{ij})$, $T = (t_{ij})$ and positive values $\tau_0,\ \gamma$ be given, such that (1.33a), (1.33b) are fulfilled. Then condition (1.36) guarantees property (1.20) with*

$$\mu \;\;=\;\; \max_j \varrho_j \;+\; \gamma \cdot \max_i \textstyle\sum_j |t_{ij}|\, \varrho_j \;\;\leq\;\; (1 + \gamma\,\tau)\,\sigma.$$

Proof. In view of Theorem 1.2.2, it is sufficient to prove (1.23b), (1.23c) for the above μ. Condition (1.23b) follows, from Lemma 1.3.1 and (1.34), as above. Furthermore, condition (1.23c) is fulfilled, because $\|(I-|P|)^{-1}\,|R|\|_\infty = \|(I-P)^{-1}\,|R|\|_\infty = \|(I+\gamma\,T)\,|R|\|_\infty \leq \|R\|_\infty + \gamma\,\|T\;|R|\|_\infty \leq \max_j \varrho_j + \gamma \cdot \max_i \sum_j |t_{ij}|\,\varrho_j$. □

Simplified criterion for properties (1.20) and (1.27), with $\mu = 1$

Throughout this subsection we assume that $\mu = 1$ and the matrix $S = (s_{ij})$ satisfies

$$s_{i\,1} + s_{i\,2} + \cdots + s_{i\,l} = 1 \quad (1 \le i \le m). \tag{1.37}$$

Assumption (1.37) is e.g. fulfilled when (1.16) stands for the canonical representation of N steps of a method (1.12) with coefficients α_{ij} satisfying

$$\alpha_{i\,1} + \alpha_{i\,2} + \cdots + \alpha_{i\,l} = 1 \quad (1 \le i \le q) \tag{1.38}$$

- this follows easily from (1.17a). GLMs are often represented with coefficients α_{ij} such that (1.38) is in force, cf. e.g. the examples in Section 1.3.3.

We shall find that condition (1.33) is the appropriate criterion for properties (1.20), (1.27), by proving the equivalence of (1.33) and (1.23) (with $\mu = 1$). In our proof we shall use the notation E_k to denote the $k \times 1$ matrix with all entries equal to 1.

First, assume (1.33). In order to prove (1.23b), (1.23c) (with $\mu = 1$), we note that $P\,E_m = P\,S\,E_l = (I - Q)\,S\,E_l = E_m - R\,E_l \le E_m$. It follows that $\|P\|_\infty \le 1$, so that $\mathrm{spr}(P) \le 1$. Hence, (1.35) is in force, which in Section 1.3.1 was proved to be equivalent to (1.23b), (1.23c).

Conversely, assume (1.23) (with $\mu = 1$). We have $E_m = S\,E_l = (I - P)^{-1}\,R\,E_l \le (I - |P|)^{-1}\,|R|\,E_l \le E_m$. Hence $(I - |P|)^{-1}\,|R|\,E_l = E_m$, which implies $|P|\,E_m + |R|\,E_l = E_m = P\,E_m + R\,E_l$. Therefore, $(|P| - P)\,E_m + (|R| - R)\,E_l = 0$, so that $P = |P| \ge 0$, $R = |R| \ge 0$, i.e. (1.33).

In view of the equivalency of (1.33) and (1.23), the Theorems 1.2.2, 1.2.4 yield the following corollary, which is closely related to a monotonicity result formulated earlier in the literature (but derived differently), cf. Spijker (2007).

Corollary 1.3.4. (Criterion for properties (1.20), (1.27), with $\mu = 1$). *Let arbitrary matrices $S = (s_{ij})$, $T = (t_{ij})$ and positive τ_0, γ be given. Assume (1.37). Then the following two statements are valid.*
(i) Condition (1.33) is necessary and sufficient for property (1.20) with $\mu = 1$.
(ii) Assume (1.25), (1.30). Then (1.33) is necessary and sufficient for (1.27) with $\mu = 1$.

1.3.2 The matrices T, P and R, for the canonical representation of GLMs

By representing N steps of method (1.12) in the form (1.16) canonically – cf. (1.15), (1.17) – and a subsequent application of one of the Theorems 1.2.2, 1.2.4 or Corollaries 1.3.2, 1.3.3, 1.3.4, one can obtain conditions for boundedness of the GLM. Because such conditions involve the corresponding T, P and R – cf. (1.22) – we shall study these matrices, in the subsequent Lemma 1.3.5. The lemma will be applied in Section 1.3.3.

From (1.17), (1.22), we see that the matrices S, T, P, Q, R, respectively, corresponding to the canonical representation of N steps of (1.12) reduce, for $N = 1$, simply to:

$$A = (\alpha_{i\,j}), \quad B = (\beta_{i\,j}), \quad L = (I + \gamma\, B)^{-1}, \quad K = L(\gamma B), \quad M = L\, A. \tag{1.39}$$

The following lemma relates (conditions on) T, P, R for *any* $N \geq 1$, directly to the simple matrices (1.39). We denote by K_0, M_0 the matrices consisting of the last l rows of K and M, respectively. Note that M_0 equals the $l \times l$ stability matrix $M(z)$ of the GLM at the point $z = -\gamma$, cf. e.g. Butcher (2003, p. 381).

Lemma 1.3.5. (On the matrices T, P, R of the canonical representation). *For given $\gamma > 0$, $\mu > 0$ and integer $N \geq 1$, the following statements are valid.*
(i) Matrix T satisfies (1.23a) if and only if $I + \gamma\, B$ is invertible.
(ii) If (1.23a) holds, then matrix P satisfies (1.23b) if and only if $\mathrm{spr}(|K|) < 1$.
(iii) If (1.23a) holds, then R is made up of $q \times l$ blocks R_n, and P of $q \times q$ blocks $P_{n,j}$, where $1 \leq n \leq N$, $1 \leq j \leq N$ and

$$R_n = M(M_0)^{n-1},\ P_{n,j} = 0\,(j > n),\ P_{n,n} = K,\ P_{n,j} = M(M_0)^{n-j-1}K_0\,(n > j).$$

Proof. Part (i) follows from (1.17b), and (ii) follows from the expressions for $P_{n,j}$ given in (iii).

To analyse the blocks R_n, we rewrite $(I + \gamma\, T)\, R = S$ in terms of these blocks, using (1.17): $\gamma \sum_{j=1}^{n-1} A\,(A_0)^{n-j-1}\, B_0\, R_j + (I + \gamma\, B)\, R_n = A\,(A_0)^{n-1}$ $(n \geq 1)$. To give this relation a more convenient form, we introduce the $l \times q$ matrix $H = [\,O\ I\,]$, composed of the $l \times (q - l))$ zero matrix O and the $l \times l$ identity matrix I. Clearly, $A_0 = H\, A$, $B_0 = H\, B$, $K_0 = H\, K$, $M_0 = H\, M$. We put $\bar{A} = A\, H$, $\bar{M} = M\, H$, so that

$$\gamma \sum_{j=1}^{n-1} \bar{A}^{\,n-j} B\, R_j \ + \ (I + \gamma\, B)\, R_n \ = \ \bar{A}^{\,n-1} A \ \ (n \geq 1).$$

We modify this relation, by premultiplying it with $\bar{A}$ and replacing n by $n - 1$. Subtracting this modified equality from the original one, we obtain $(I + \gamma\, B)\, R_n = \bar{A}\, R_{n-1}$, so that $R_n = \bar{M}\, R_{n-1}$ $(n \geq 2)$. Hence $R_n = (\bar{M})^{n-1} R_1 = (\bar{M})^{n-1} M = M\,(M_0)^{n-1}$ $(n \geq 1)$.

To complete the proof, we conclude from $(I + \gamma\, T)\, P = \gamma\, T$ and (1.17b), that P has a block Toeplitz structure, with $q \times q$ blocks $P_{n,j} = P_{n-j+1}$ where $P_k = 0$ $(k \leq 0)$, $P_1 = K$. Similarly as above we find $\gamma \sum_{j=1}^{k-1} \bar{A}^{\,k-j} B\, P_j + (I + \gamma\, B)\, P_k = \gamma\, \bar{A}^{\,k-1} B$ $(k \geq 1)$ and $(I + \gamma\, B)\, P_k = \bar{A}\, P_{k-1}$, so that $P_k = (\bar{M})^{k-1} K = M\, M_0^{k-2} K_0$ $(k \geq 2)$. □

1.3.3 Examples of actual boundedness results obtainable from the theory

This section only serves to make evident the practical relevance of the generic process (1.16) and the applicability of the above theory to the boundedness anal-

ysis of given GLMs, see Definition 1.2.1. Accordingly, below we will illustrate the theory by applying it just to a few actual numerical methods.

For ease of presentation, and also to illustrate (1.25), (1.26) and Theorem 1.2.4 with $r < m$, we deal throughout this section with autonomous problems –i.e. F in (1.1) is independent of t, and (1.5) reduces to

$$\|v + \tau_0 F(v)\| \leq \|v\| \quad (\text{for } v \in \mathbb{V}). \tag{1.40}$$

Below we shall study boundedness of various methods, by looking for stepsize coefficients γ and constants μ such that

Condition $0 < \Delta t \leq \gamma \cdot \tau_0$ implies boundedness with constant μ, cf. Definition 1.2.1, whenever $\mathbb{V}$ is a vector space with seminorm $\|\cdot\|$ and $F : \mathbb{V} \to \mathbb{V}$ satisfies (1.40). (1.41)

Clearly, when (1.41) holds with $\mu = 1$, then γ is a stepsize coefficient for *monotonicity*.

Two explicit RKMs

Following Gottlieb & Shu (1998), we consider two explicit RKMs (1.2), with $s = 2$, the nonzero coefficients of which are given by (1.42) and (1.43), respectively:

$$a_{21} = 1, \;\; a_{31} = a_{32} = 1/2, \tag{1.42}$$

$$a_{21} = -20, \;\; a_{31} = 41/40, \;\; a_{32} = -1/40. \tag{1.43}$$

Both methods are of second order and yield identical numerical approximations when applied to linear autonomous problems.The first method is monotonic ((1.41) with $\mu = 1$) with stepsize-coefficient $\gamma = 1$, whereas for method (1.43) there exists *no* positive stepsize-coefficient γ for monotonicity, cf. e.g. the paper just mentioned and Ferracina & Spijker (2004) or Higueras(2004).

To analyse for both methods the boundedness property (1.41) (with arbitrary $\mu \geq 1$), we represent the methods as GLMs (1.12) with coefficient matrices A, B –as indicated in Section 1.2.1– and consider the corresponding canonical representation of $N \geq 1$ steps, cf. (1.15), (1.16), (1.17). Because F is independent of t, we have properties (1.25), (1.26) with $r = 1$, $\mathfrak{I}_1 = \{1, \cdots, m\}$. From (1.17) one sees that (1.30) and (1.31) are fulfilled, so that Theorem 1.2.4 can be applied. It follows that property (1.41) is present if and only if condition (1.23) is fulfilled (for all $N \geq 1$). From Lemma 1.3.5 we see that conditions (1.23a), (1.23b) are fulfilled, with any $\gamma > 0$, for both methods. In order to express the dependence of (1.23c) on N, we put $\mu_N = ||(I - |P|)^{-1} |R|||_\infty$.

For method (1.42), when $N \geq 1$, it is possible to find by a computation based on Lemma 1.3.5 that

$$\mu_N = 1 \quad (\text{for } 0 < \gamma \leq 1), \quad \mu_N = (1 + 2\,\gamma\,(\gamma - 1))^N \quad (\text{for } \gamma \geq 1).$$

Hence, for *any* given $\mu \geq 1$, the largest stepsize-coefficient γ, for which method (1.42) has the boundedness property (1.41), is equal to $\gamma = 1$.

For method (1.43), a similar computation yields $\mu_N = (1+\frac{\gamma}{20}+\gamma^2)^{N-1}\,(1+40\,\gamma)$ for $N \geq 1$ and $0 < \gamma \leq 2$). From this expression we can conclude that there exists *no* positive γ for which method (1.43) has the boundedness property (1.41) with *any* $\mu \geq 1$.

We think these conclusions, about methods (1.42), (1.43), nicely supplement and confirm the discussion of the methods, as presented in Gottlieb & Shu (1998): method (1.42) is superior to (1.43) not only regarding monotonicity, but also with respect to boundedness.

We have not displayed the details of the computations leading to the above expressions for μ_N, because we want to keep the size of the chapter within reasonable limits.

A two-stage RKM depending on a parameter θ

We shall give an example showing that the canonical representation of N steps of an (irreducible) RKM can fail to satisfy the irreducibility condition (1.30), with the result that Theorem 1.2.4 does *not* yield a necessary condition for boundedness. The example will also provide an instance of a *non-canonical* representation yielding a boundedness result that is *not* obtainable via the canonical representation. Finally, it will show, unlike the examples in Section 1.3.3, that the restrictions on γ for *boundedness* of RKMs can be less severe than for *monotonicity*.

We consider the two-stage RKM, given by (1.2) with $s = 2$, $a_{1,1} = a_{1,2} = 0$, $a_{2,1} = a_{3,1} = 1-\theta$, $a_{2,2} = a_{3,2} = \theta$, with real parameter θ. We write the method concisely as (1.12) with $l = 1$, $q = 2$, $A = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, $B = \begin{pmatrix} 0 & 0 \\ 1-\theta & \theta \end{pmatrix}$, and consider the corresponding canonical representation (1.16) of N consecutive steps of the method. We see from Lemma 1.3.5 that (1.23a), (1.23b) hold, if and only if $1 + 2\,\gamma\,\theta > 0$. Assuming this inequality to be fulfilled, it is possible to find by a computation using Lemma 1.3.5, that $\mu_N = ||(I - |P|)^{-1}\,|R|||_\infty$ equals $\mu_N = \lambda^N$ $(N \geq 1)$, where

$$\lambda \;=\; \frac{|1+\gamma\,(\theta-1)| + \gamma\,|\theta-1|}{1+\gamma\,\theta-\gamma\,|\theta|} \;\geq 1.$$

We see that $\lambda = 1$, if and only if

$$0 \;\leq\; \theta \;\leq\; 1, \quad \gamma\,(1-\theta) \;\leq\; 1. \tag{1.44}$$

This does *not* allow us to conclude via Theorem 1.2.4 –with $r = 1$, $\mathfrak{I}_1 = \{1, \cdots, m\}$ as in Section 1.3.3– that condition (1.44) is *necessary* for *boundedness* (property (1.41) with any fixed $\mu \geq 1$), because the irreducibility condition (1.30) on $[\,S\;T\,]$ is violated for $N \geq 2$.

On the other hand, Theorem 1.2.4 can be applied – with $r = 1$, $\mathfrak{I}_1 = \{1, \cdots, m\}$ –to the canonical representation for $N = 1$, because $[\,S\;T\,] = [\,A\;B\,]$ satisfies (1.30). Since $\mu_1 = \lambda$, condition (1.44) is *necessary and sufficient* for

monotonicity ((1.41) with $\mu = 1$); this follows also e.g. from Corollary 1.3.4, from Ferracina & Spijker (2004) or Higueras (2004).

To prove that boundedness is possible under a weaker condition than (1.44), we represent N steps of the method $-$*not* canonically$-$ by (1.16) with $l = 1$, $m = N$, $s_{n,1} = 1$, $t_{nj} = 0$ $(j > n)$, $t_{nj} = \theta$ $(j = n)$, $t_{nj} = 1$ $(j < n)$ and $y_n = u_n$, $x_1 = u_0 + \Delta t\,(1-\theta)\,F(u_0)$. Since $[\,S\;T\,]$ now satisfies (1.30) (with $r = 1$, $I = \{1, \ldots, m\}$), we can apply e.g. Corollary 1.3.4 to the situation at hand. A computation shows that (1.33) holds if and only if $0 \le \theta$, $\gamma\,(1-\theta) \le 1$. Hence, for any $\theta > 1$, $\gamma > 0$, the conditions (1.6), (1.40) imply that

$$\|u_n\| \le \|x_1\| = \left\| \left(1 + \tfrac{(\theta-1)\,\Delta t}{\tau_0}\right) u_0 \,-\, \tfrac{(\theta-1)\,\Delta t}{\tau_0}\,(u_0 + \tau_0\,F(u_0)) \right\| \le \mu\,\|u_0\|,$$

with $\mu = 1 + 2\,(\theta-1)\,\gamma$.

In conclusion, for $\theta > 1$, there exists no positive stepsize coefficient for monotonicity, whereas any $\gamma > 0$ is a stepsize coefficient corresponding to the boundedness property (1.41), with $\mu = 1 + 2\,(\theta-1)\,\gamma$.

One-leg Adams-Bashforth method

We consider the so-called one-leg version of the second order Adams-Bashforth method,

$$u_n \;=\; u_{n-1} \,+\, \Delta t\,F\big(\tfrac{3}{2}\,u_{n-1} - \tfrac{1}{2}\,u_{n-2}\big), \tag{1.45}$$

cf. e.g. Butcher (1987), Hairer, Nørsett & Wanner (1987), Hairer & Wanner (1996). This method is *not monotonic*, in that there exists *no* positive γ with the property that (1.40), (1.45), (1.6) always imply $\|u_n\| \le \max\{\|u_{n-1}\|, \|u_{n-2}\|\}$. This follows e.g. directly from Spijker (1983, Theorem 3.3).

We will see that, in spite of the above negative result, there exist positive γ and μ such that

$$\|u_n\| \le \mu \cdot \max\{\|u_0\|, \|u_1\|\} \quad (\text{for } 0 < \Delta t < \gamma \cdot \tau_0, \text{ and all } n \ge 2), \tag{1.46}$$

as soon as (1.40) and (1.45) (for $n \ge 2$) are in force.

Below we shall prove this boundedness result by rewriting method (1.45) as a GLM, and applying Corollary 1.3.3 in combination with Lemma 1.3.5 to the canonical representation, cf. (1.15), (1.16), (1.17).

We introduce, for $n \ge 1$, the vectors $v_1^{[n]} = -\frac{1}{2}u_{n-1} + \frac{3}{2}u_n$, $v_2^{[n]} = u_n$, $v_3^{[n]} = u_{n+1}$ and $u_1^{[n-1]} = u_{n-1}$, $u_2^{[n-1]} = u_n$, so that (1.45) is equivalent to the GLM (1.12), with

$$q = 3,\;\; l = 2 \quad \text{and} \quad A = \begin{pmatrix} -\frac{1}{2} & \frac{3}{2} \\ 0 & 1 \\ 0 & 1 \end{pmatrix},\; B = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

Clearly, if this GLM satisfies (1.41) with positive γ, μ, then method (1.45) has the boundedness property mentioned above, cf. (1.46).

In order to apply Corollary 1.3.3 to the canonical representation of the GLM, we have to check conditions (1.33a), (1.33b) and (1.36). Because B is strictly lower triangular, we see directly from Lemma 1.3.5 (i) that (1.33a) is fulfilled for any $\gamma > 0$.

To analyse (1.33b) we consider, for any $\gamma > 0$, the expressions for the blocks $P_{n,j}$ given by Lemma 1.3.5 (iii). One easily sees that $P_{n,j} \geq 0$ ($j \geq n$). Furthermore, it can be seen that $P_{n,j} \geq 0$ (for $j = n-1$ and $j = n-2$) if and only if $\gamma \leq 4/9$. From now on we assume $\gamma = 4/9$. In the analysis of $P_{n,j}$ with $j \leq n-3$, via Lemma 1.3.5 (iii), it is convenient to use the following representation for the powers of M_0:

$$(M_0)^k = \begin{pmatrix} x_{k-1} & y_{k-1} \\ x_k & y_k \end{pmatrix},$$

where $x_{k+1} = \frac{1}{3}\, x_k + \frac{2}{9}\, x_{k-1}$, $x_0 = 0$, $x_1 = \frac{2}{9}$ and $y_{k+1} = \frac{1}{3}\, y_k + \frac{2}{9}\, y_{k-1}$, $y_0 = 1$, $y_1 = \frac{1}{3}$ (for $k \geq 1$). Substituting this representation (with $k = n-j-1$) in the expression for $P_{n,j}$ of Lemma 1.3.5 (iii), it can be seen that $P_{n,j} \geq 0$ (for $j \leq n-3$), which proves (1.33b).

The first inequality in (1.36) is fulfilled −with $\mathrm{spr}(P) = 0$− because the blocks $P_{n,n}$ are strictly lower triangular. A computation, using the above representation for $(M_0)^k$, shows that the remaining inequalities in (1.36) are fulfilled as well, with $\varrho_j = 2$ (for $j = 1$), $\varrho_j = 3^{-n}[2^n - (-1)^n]$ (for $j = 3n-2$, $n \geq 2$), $\varrho_j = 3^{-n-1}[2^{n+2} - (-1)^n]$ (for $j = 3n-1$, $n \geq 1$), $\varrho_j = 3^{-n-2}[2^{n+3} + (-1)^n]$ (for $j = 3n$, $n \geq 1$) and $\sigma = 31/4$, $\tau = 3/2$. The upperbound $(1 + \gamma\,\tau)\,\sigma$ of Corollary 1.3.3 thus amounts to $155/12$, from which we conclude that method (1.45) has the boundedness property (1.46), with $\gamma = 4/9$ and $\mu = 155/12 \simeq 12.9$.

A smaller value for μ can be obtained by a straightforward −but slightly longer− computation of the expression $\mu = \max_j \varrho_j + \gamma \cdot \max_i \sum_j |t_{ij}|\, \varrho_j$, see Corollary 1.3.3. In this way one can arrive at a similar conclusion as above, but with $\gamma = 4/9$ and the better value $\mu = 31/9 \simeq 3.4$.

We note, for completeness, that the above results could not have been obtained by a similar application of Corollary 1.3.2, instead of Corollary 1.3.3, because condition (1.33c) is violated, in the situation at hand, for all $N \geq 1$ and $\gamma > 0$.

A two-stage GLM

Our last example illustrates that conclusions about boundedness can sometimes be reached by a rather *short calculation*. We consider the second order method for solving (1.1) (with $F(t,v) = F(v)$),

$$u_1^{[n]} = -u_1^{[n-1]} + 2\, u_2^{[n-1]}, \tag{1.47a}$$

$$u_2^{[n]} = u_2^{[n-1]} + \Delta t \cdot F(u_1^{[n]}), \tag{1.47b}$$

where $u_1^{[n-1]} \simeq u((n-1/2)\,\Delta t)$ and $u_2^{[n-1]} \simeq u(n\,\Delta t)$ $(n = 1, 2, 3, \ldots)$. We write the method as (1.12), with $l = q = 2$, $A = \begin{pmatrix} -1 & 2 \\ 0 & 1 \end{pmatrix}$, $B = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$, and consider the corresponding canonical representation (1.16), cf. (1.15), (1.17). Because the matrix $[\,S\;T\,]$ satisfies the irreducibility condition (1.30) with $r = 1$, $\mathfrak{I}_1 = \{1, \ldots, m\}$, we can apply Theorem 1.2.4 in the situation at hand.

Let any $\gamma > 0$ be given. From Lemma 1.3.5 we see easily that the corresponding matrices T, P satisfy conditions (1.23a), (1.23b). By Theorem 1.2.4, the boundedness property (1.41) thus holds, for any given μ, if and only if $\mu_N = \|(I - |P|)^{-1}\,|R|\|_\infty$ is such that $\sup\{\mu_N : N \geq 1\} \leq \mu$.

Because $(I - |P|)^{-1}\,|R| \geq |R|$, we see from Lemma 1.3.5 (iii) that $\mu_N \geq \|R\|_\infty \geq \|M^N\|_\infty$, with M as in (1.39). From the expression $M = \begin{pmatrix} -1 & 2 \\ \gamma & 1-2\gamma \end{pmatrix}$, it follows that $\mathrm{spr}(M) = \gamma + \sqrt{1+\gamma^2} > 1$, so that $\mu_N \to \infty$ for $N \to \infty$.

We conclude that *there is no boundedness*, in the sense of (1.41), for any positive γ and μ.

1.4 Proof of Theorems 1.2.2, 1.2.4

Because Theorem 1.2.2 follows from Theorem 1.2.4 by choosing in the latter theorem the trivial index sets $\mathfrak{I}_\rho = \{\rho\}$ (for $1 \leq \rho \leq m = r$), it is enough to prove below Theorem 1.2.4.

The sufficiency of condition (1.23), in parts (i) and (ii) of Theorem 1.2.4, is a direct consequence of Proposition 1.4.2, to be given in Section 1.4.1, and the fact that (1.20) implies the three properties (1.27), (1.28) and (1.29) (for any index sets $\mathfrak{I}_\rho$ as in (1.25)).

The necessity of condition (1.23), in Theorem 1.2.4, follows directly from Proposition 1.4.6, to be given in Section 1.4.2, and the fact that property (1.27) implies both (1.28) and (1.29).

1.4.1 Sufficiency of condition (1.23)

In the following, we shall write (1.16) and similar relations more concisely, by using the following notations relevant to the vector space $\mathbb{V}$. For any integer $k \geq 1$ and vectors $x_1, \ldots, x_k \in \mathbb{V}$, we denote the vector in $\mathbb{V}^k$ with components x_i by

$$x = [x_i] = \begin{pmatrix} x_1 \\ \vdots \\ x_k \end{pmatrix} \in \mathbb{V}^k.$$

Furthermore, we denote with a bold-face letter the linear operators from $\mathbb{V}^k$ to $\mathbb{V}^m$ determined in a natural way by $m \times k$ matrices: for any matrix $A = (a_{ij}) \in \mathbb{R}^{m\times k}$ and $x = [x_i] \in \mathbb{V}^k$ we define $\boldsymbol{A}(x) = y$, where $y = [y_i] \in \mathbb{V}^m$ is given by $y_i = \sum_{j=1}^k a_{ij}\,x_j$ $(1 \leq i \leq m)$.

We combine the vectors x_i and y_i, occurring in (1.16), into vectors $x = [x_i] \in \mathbb{V}^l$ and $y = [y_i] \in \mathbb{V}^m$, respectively. Furthermore, for given functions $F_i : \mathbb{V} \to \mathbb{V}$ $(1 \le i \le m)$, we define a function $\boldsymbol{F}$, from $\mathbb{V}^m$ to $\mathbb{V}^m$, by $\boldsymbol{F}(y) = [F_i(y_i)] \in \mathbb{V}^m$ for $y = [y_i] \in \mathbb{V}^m$. With these notations, the relations (1.16) can be written as an equality in $\mathbb{V}^m$:

$$y \;=\; \boldsymbol{S}x \;+\; \Delta t \cdot\; \boldsymbol{T}\,\boldsymbol{F}(y). \tag{1.48}$$

The subsequent lemma is a variant to Spijker (2007, Lemma 4.1). It will be useful, in the present section for proving Proposition 1.4.2, and later on for proving Proposition 1.4.6. We shall use the notations 1.22, and relate 1.48 – with F_i satisfying (1.18), (1.26) – to the conditions

$$y = \boldsymbol{R}\,x \;+\; \boldsymbol{P}\,z, \quad \text{with} \quad \|z_i\| \le \|y_i\| \;\; (1 \le i \le m), \tag{1.49a}$$

$$y_i \ne y_j \;\; \text{whenever } z_i \ne z_j \text{ and } i, j \text{ belong to the same index set } \mathfrak{I}_\rho. \tag{1.49b}$$

Lemma 1.4.1. (Reformulation of (1.48) with F_i satisfying (1.18), (1.26)). *Let $\tau_0 > 0$, $\gamma > 0$, $I + \gamma T$ invertible, and assume (1.25). Let $x = [x_i] \in \mathbb{V}^l$ and $y = [y_i] \in \mathbb{V}^m$ be given. Then the following three statements are equivalent:*

The vectors x, y satisfy (1.48) for some Δt with $0 < \Delta t \le \gamma \cdot \tau_0$ and some functions $F_i : \mathbb{V} \to \mathbb{V}$ satisfying (1.18), (1.26); (1.50)

The vectors x, y satisfy (1.48) with $\Delta t = \gamma \cdot \tau_0$ and some functions $F_i : \mathbb{V} \to \mathbb{V}$ satisfying (1.18), (1.26); (1.51)

There exists a vector $z = [z_i] \in \mathbb{V}^m$ such that (1.49) holds. (1.52)

Proof. Assume (1.50). In order to prove (1.51), we define $\theta = \Delta t/(\gamma\,\tau_0)$ and $\tilde{F}_i = \theta \cdot F_i$, so that x, y satisfy (1.48) also with $\Delta t = \gamma \cdot \tau_0$ and F_i replaced by $\tilde{F}_i$. Clearly, $\tilde{F}_i = \tilde{F}_j$ for i, j in the same index set, and $\|v + \tau_0\,\tilde{F}_i(v)\| = \|(1-\theta)\,v + \theta\,[v + \tau_0\,F_i(v)]\| \le \|v\|$. This implies (1.51).

Assume (1.51). In order to prove (1.52), we rewrite (1.48) as

$$(I + \gamma\,\mathbf{T})\,y = \;\mathbf{S}\,x + \gamma\,\mathbf{T}\,[y + \tau_0\,\mathbf{F}(y)],$$

from which we see that x, y satisfy (1.49a) with $z = [z_i] = y + \tau_0\,\mathbf{F}(y)$. Furthermore, when $z_i \ne z_j$ and i, j belong to the same index set $\mathfrak{I}_\rho$, we have $y_i + \tau_0\,F_i(y_i) \ne y_j + \tau_0\,F_i(y_j)$, which implies (1.49b). Hence, (1.52) holds.

Assume (1.52). We shall prove (1.50). For $i \in \mathfrak{I}_\rho$ we define $F_i(v) = (1/\tau_0)\,(z_k - y_k)$ (if $v = y_k,\ k \in \mathfrak{I}_\rho$) and $F_i(v) = 0$ (otherwise). In view of (1.49b) this is a proper definition, and $F_i = F_j$ for i, j in the same index set, i.e. (1.26). Furthermore, we see that x, y satisfy (1.48) with $\Delta t = \gamma \cdot \tau_0$. Finally, for $i \in \mathfrak{I}_\rho$, we have $\|v + \tau_0\,F_i(v)\| = \|z_k\| \le \|v\|$ (if $v = y_k,\ k \in \mathfrak{I}_\rho$) and $\|v + \tau_0\,F_i(v)\| = \|v\|$ (otherwise), so that (1.18) is fulfilled. This completes the proof of (1.50). □

Proposition 1.4.2. (Sufficiency of condition (1.23) for property (1.20)). *Let $\tau_0 > 0$ be given, and assume γ, μ are positive constants such that (1.23) holds. Then process (1.16) has the boundedness property (1.20).*

Proof. Assume condition (1.23) is fulfilled, and consider x_i, y_i satisfying (1.16), in the situation where (1.18) holds and $0 < \Delta t \leq \gamma \cdot \tau_0$. Applying Lemma 1.4.1 (with the trivial index sets $\mathfrak{I}_\rho = \{\rho\}$, $1 \leq \rho \leq r = m$), we have (1.49a), from which we obtain

$$[\,\|y_i\|\,] \;\leq\; [\,\|r_i\|\,] + |P|\,[\,\|z_i\|\,] \;\leq\; [\,\|r_i\|\,] + |P|\,[\,\|y_i\|\,], \text{ with } [\,r_i\,] = \boldsymbol{R}\,x.$$

Consequently, $(I - |P|)\,[\,\|y_i\|\,] \leq [\,\|r_i\|\,]$. By Lemma 1.3.1, the matrix $I - |P|$ is invertible with $(I - |P|)^{-1} \geq 0$. Therefore, $[\,\|y_i\|\,] \;\leq\; (I - |P|)^{-1}\,[\,\|r_i\|\,]$, which implies

$$[\,\|y_i\|\,] \;\leq\; (I - |P|)^{-1}\,|R|\,[\,\|x_i\|\,]. \tag{1.53}$$

An application of (1.23c) shows that the components in the right-hand member of the last inequality do not exceed $\mu \cdot (\max_j \|x_j\|)$, which completes the proof of Proposition 1.4.2. □

1.4.2 Necessity of condition (1.23)

In this section we shall prove the necessity of condition (1.23) for properties (1.28) and (1.29), under the irreducibility assumptions (1.30) and (1.31), respectively. We assume throughout the section that τ_0, γ, μ are given positive constants and, unless stated otherwise, that $\mathfrak{I}_\rho$ are arbitrary given index sets of type (1.25).

Formulation of Proposition 1.4.6

To demonstrate the necessity of condition (1.23) we will formulate Proposition 1.4.6. To prove this proposition we will need three lemmas, the first of which is

Lemma 1.4.3. (Invertibility of $\boldsymbol{I + \gamma\,T}$). *Property (1.28), as well as property (1.29), implies that the matrix $I + \gamma\,T$ is invertible.*

Proof. Assume (1.28) or (1.29). Let $\eta = [\eta_i] \in \mathbb{R}^m$ with $(I + \gamma\,T)\,\eta = 0$. We shall prove $\eta = 0$.

We define $F_i(v) = -(1/\tau_0)\,v$ (for all $v \in \mathbb{V} = \mathbb{R}^m$), so that (1.26), (1.18) are fulfilled with $\|\cdot\| = \|\cdot\|_\infty$. We see that (1.16) is satisfied, with $\Delta t = \gamma \cdot \tau_0$, by the vectors $x_i = 0$ $(1 \leq i \leq l)$ and $y_i = \eta_i\,e_1$ $(1 \leq i \leq m)$, where e_1 is the first unit vector in $\mathbb{V} = \mathbb{R}^m$.

By (1.28) or (1.29), there follows $|\eta_i| = \|y_i\|_\infty \leq \mu \cdot \max_j \|x_j\|_\infty = 0$, so that $\eta = 0$. □

In proving that property (1.28) implies (1.23), we shall make use of vectors $\xi = [\xi_j] \in \mathbb{R}^l$ and $\eta = [\eta_j]$, $\zeta = [\zeta_j] \in \mathbb{R}^m$ satisfying the following condition:

$$\eta = R\,\xi + P\,\zeta, \text{ with } \eta_j \neq \eta_k \text{ (for all } j \neq k \text{ belonging to the same index set } \mathcal{I}_\rho). \tag{1.54}$$

Furthermore, in proving that the (weaker) property (1.29) implies (1.23), we shall use vectors $\xi = [\xi_j] \in \mathbb{R}^l$ and $\eta = [\eta_j]$, $\zeta = [\zeta_j] \in \mathbb{R}^m$ satisfying the subsequent (stronger) condition:

$$\eta = R\,\xi + P\,\zeta, \text{ with } |\zeta_j| \leq |\eta_j| \text{ (for } 1 \leq j \leq m), \text{ and } \eta_j \neq \eta_k \text{ (for all } j \neq k \text{ belonging to the same index set } \mathcal{I}_\rho). \tag{1.55}$$

We shall see that vectors ξ, η, ζ exist satisfying conditions (1.54) and (1.55), respectively, if the following (simplified) versions of assumptions (1.30) and (1.31) are fulfilled:

$$[S\,T\,](i,:) \neq [S\,T\,](j,:) \quad \text{(if } i \neq j \text{ belong to the same index set } \mathcal{I}_\rho), \tag{1.56}$$

$$[S\,\widehat{T}\,](i,:) \neq [S\,\widehat{T}\,](j,:) \quad \text{(if } i \neq j \text{ belong to the same index set } \mathcal{I}_\rho). \tag{1.57}$$

Our proof of Proposition 1.4.6 needs also the following two lemmas.

Lemma 1.4.4. (Relevance of (1.54), (1.55) for condition (1.23)).
(i) Assume (1.28), and suppose ξ, η, ζ satisfy (1.54). Then (1.23) is fulfilled.
(ii) Assume (1.29), and suppose ξ, η, ζ satisfy (1.55). Then (1.23) is fulfilled.

Lemma 1.4.5. (Conditions for (1.54), (1.55)). *Let $I + \gamma\,T$ be invertible. Then the following two implications are valid.*
(i) Assumption (1.56) implies the existence of ξ, η, ζ satisfying (1.54).
(ii) Assumption (1.57) implies the existence of ξ, η, ζ satisfying (1.55).

Since the proof of these two lemmas is rather long, it will be given separately in Section 1.4.2.

Proposition 1.4.6. (Necessity of condition (1.23) for properties (1.28), (1.29)).
(i) Assume (1.28) and irreducibility in the sense of (1.30). Then (1.23) holds.
(ii) Assume (1.29) and irreducibility in the sense of (1.31). Then (1.23) holds.

Proof. Because of Lemma 1.4.3, we can assume that $I + \gamma\,T$ is invertible.

In order to prove Part (i) of the proposition, we assume (1.28), (1.30). We denote by I^0 the set of all indices i, with $1 \leq i \leq m$ and $T(:,i) = 0$.

First, assume there are *no* index sets $\mathcal{I}_\rho$ containing a pair of indices $i \neq j$ with $j \in I^0$. Conditions (1.30) and (1.56) are then equivalent. Hence, combining Lemma 1.4.5 (i) and Lemma 1.4.4 (i), we obtain (1.23).

Next, assume there do exist sets $\mathcal{I}_\rho$ containing indices $i \neq j$ where $j \in I^0$. We note that the functions F_j, with $j \in I^0$, *do not enter actually* in the basic

relations (1.16). Accordingly, it is immaterial for these relations whether or not a given function F_j, with $j \in I^0$, is equal to any F_i with $i \neq j$. Therefore, we can refine the given partition $\mathfrak{I}_1 \cup \cdots \cup \mathfrak{I}_r = \{1, \cdots, m\}$ into one with regard to which properties (1.28) and (1.56) hold: the refined partition is obtained, from the original one, by creating new separate index sets for all indices $j \in I^0$ belonging to an (old) index set $\mathfrak{I}_\rho$ with at least two different indices.

From (the original) property (1.28) one sees that (1.28) is still present with regard to the new, refined partition. Moreover, the original property (1.30) implies that (1.56) is valid with regard to the new index sets. Therefore, we arrive at (1.23), again by combining Lemma 1.4.5 (i) and 1.4.4 (i) (in the situation of the new partition).

To prove Part (ii) of the proposition, assume (1.29), (1.31), and define I^0 as above.

First assume there are *no* sets $\mathfrak{I}_\rho$ containing indices $i \neq j$ where $j \in I^0$. Conditions (1.31) and (1.57) are then equivalent. Hence, Lemmas 1.4.5 (ii) and 1.4.4 (ii) yield (1.23).

Next assume there do exist sets $\mathfrak{I}_\rho$ with indices $i \neq j$ where $j \in I^0$. Using the above refined partition, similarly as in the proof of Part (i), we arrive again at (1.23) by combining Lemma 1.4.5 (ii) and 1.4.4 (ii). □

Proof of the Lemmas 1.4.4, 1.4.5

The sole purpose of the present section is to prove Lemmas 1.4.4, 1.4.5. Throughout the section we assume, with no loss of generality, that $I + \gamma T$ is invertible. We shall use the notation

$$\operatorname{sgn}(\alpha) = 1 \text{ (for } \alpha \geq 0), \qquad \operatorname{sgn}(\alpha) = -1 \text{ (for } \alpha < 0).$$

Proof of Lemma 1.4.4

The proof of this lemma is divided into several parts.

Part 1a. Assume (1.28), and let ξ, η, ζ satisfy (1.54). We shall prove (1.23b) via Lemma 1.3.1, by assuming that λ and φ satisfy (1.32), and deducing a contradiction from that assumption.

We shall prove $\varphi = 0$, by using special vectors $x = [x_j] \in \mathbb{V}^l$ and $y = [y_j]$, $z = [z_j] \in \mathbb{V}^m$, where x_j, y_j, $z_j \in \mathbb{V} = \mathbb{R}^m$ have components x_{ij}, y_{ij}, z_{ij}, respectively. We define, for $1 \leq i \leq m$, $1 \leq j \leq m$, $1 \leq k \leq l$,

$$x_{ik} = 0, \quad z_{ij} = \operatorname{sgn}(p_{ij})\, \varphi_j, \quad y_{ij} = \sum_{k=1}^{l} r_{jk}\, x_{ik} + \sum_{k=1}^{m} p_{jk}\, z_{ik}.$$

We have $y = \boldsymbol{R}x + \boldsymbol{P}z$, and because $y_{jj} = \sum_{k=1}^{m} |p_{jk}|\, \varphi_k = \lambda \varphi_j$, there follows

$$\|z_j\|_\infty = |z_{ij}| = \varphi_j \leq y_{jj} = \|y_j\|_\infty \quad (1 \leq i \leq m,\, 1 \leq j \leq m). \tag{1.58}$$

First, suppose $y_j \neq y_k$ for all $j \neq k$ belonging to the same index set $\mathfrak{I}_\rho$. Then x, y, z satisfy (1.49), with $\|\cdot\| = \|\cdot\|_\infty$, so that, by Lemma 1.4.1, the

vectors x, y satisfy (1.51) with $\mathbb{V} = \mathbb{R}^m$, $\|\cdot\| = \|\cdot\|_\infty$. By property (1.28) and (1.58), there follows $\|\varphi\|_\infty \le \max_j \|y_j\|_\infty \le \mu \cdot \max_k \|x_k\|_\infty = 0$. Hence $\varphi = 0$, which contradicts (1.32) and thus proves (1.23b).

Next, suppose $y_q = y_s$ for two indices $q < s$ belonging to the same set $\mathfrak{I}_\rho$. In this situation, we modify (only) the q-th component of all x_j, y_j, z_j into $\tilde{x}_{qj} = \xi_j$, $\tilde{y}_{qj} = \eta_j$, $\tilde{z}_{qj} = \zeta_j$, and we denote the resulting vectors by $\tilde{x}_j$, $\tilde{y}_j$, $\tilde{z}_j$, respectively. The vectors $\tilde{x} = [\tilde{x}_j]$, $\tilde{y} = [\tilde{y}_j]$, $\tilde{z} = [\tilde{z}_j]$ satisfy the following variant of condition (1.49):

$$\tilde{y} = \boldsymbol{R}\,\tilde{x} + \boldsymbol{P}\,\tilde{z}, \quad \tilde{y}_j \ne \tilde{y}_k \ \text{ (for all } j \ne k \text{ in the same index set).} \tag{1.59}$$

In order that $\tilde{x}$, $\tilde{y}$, $\tilde{z}$ actually fulfill (1.49), we define the special seminorm

$$\|\psi\| = \max\{|\psi_i| : i \ne q\} \quad (\text{for all } \psi = [\psi_i] \in \mathbb{V} = \mathbb{R}^m).$$

Because y_j, z_j satisfy (1.58), we have

$$\|\tilde{z}_j\| = \|z_j\|_\infty \le \|y_j\|_\infty = \|\tilde{y}_j\| \ \text{ (for } 1 \le j \le m) \tag{1.60}$$

(where $\|y_j\|_\infty = \|\tilde{y}_j\|$, with $j = q$, follows from: $\|\tilde{y}_q\| = \|\tilde{y}_s\| = \|y_s\|_\infty = \|y_q\|_\infty$).

Clearly, with the above special seminorm in $\mathbb{V}$, the vectors $\tilde{x}$, $\tilde{y}$, $\tilde{z}$ fulfill (1.49), so that $\tilde{x}$, $\tilde{y}$ satisfy (1.51). Using property (1.28) and the last equality in (1.60), we find $\max_j \|y_j\|_\infty = \max_j \|\tilde{y}_j\| \le \mu \cdot \max_k \|\tilde{x}_k\| = 0$. In view of (1.58), it follows that $\varphi = 0$, which proves (1.23b).

Part 1b. Assuming (1.28), (1.54), we shall prove (1.23c). We have $\|(I - |P|)^{-1}|\,R\,|\|_\infty = \|\varphi\|_\infty$, with $\varphi = [\varphi_i] \in \mathbb{R}^m$, where the values $\varphi_i \ge 0$ satisfy the linear equations

$$\varphi_j = \sum_{k=1}^{l} |r_{jk}| + \sum_{k=1}^{m} |p_{jk}|\varphi_k \quad (1 \le j \le m).$$

Condition (1.23c) is thus equivalent to

$$\|\varphi\|_\infty \le \mu. \tag{1.61}$$

We shall prove this inequality, using again some special vectors $x = [x_j] \in \mathbb{V}^l$ and $y = [y_j]$, $z = [z_j] \in \mathbb{V}^m$, where x_j, y_j, $z_j \in \mathbb{V} = \mathbb{R}^m$ have components x_{ij}, y_{ij}, z_{ij}. In view of the linear equations satisfied by $\varphi_1, \ldots, \varphi_m$, we define now

$$x_{ik} = \operatorname{sgn}(r_{ik}), \quad z_{ij} = \operatorname{sgn}(p_{ij})\,\varphi_j, \quad y_{ij} = \sum_{k=1}^{l} r_{jk}\,x_{ik} + \sum_{k=1}^{m} p_{jk}\,z_{ik}.$$

Clearly $y = \boldsymbol{R}\,x + \boldsymbol{P}\,z$, and because $y_{jj} = \sum_{k=1}^{l} |r_{jk}| + \sum_{k=1}^{m} |p_{jk}|\,\varphi_k = \varphi_j$, the relations (1.58) are again fulfilled.

First, suppose $y_j \neq y_k$ for all $j \neq k$ belonging to the same index set $\mathfrak{I}_\rho$. Then x, y, z satisfy (1.49), with $\|\cdot\| = \|\cdot\|_\infty$, so that, by Lemma 1.4.1, the vectors x, y satisfy (1.51) with $\mathbb{V} = \mathbb{R}^m$, $\|\cdot\| = \|\cdot\|_\infty$. By property (1.28) and (1.58), there follows $\|\varphi\|_\infty \leq \max_j \|y_j\|_\infty \leq \mu \cdot \max_k \|x_k\|_\infty = \mu$, which implies (1.61).

Next, suppose $y_q = y_s$, where $q < s$ belong to the same set $\mathfrak{I}_\rho$. We modify the q-th component of x_j, y_j, z_j as above in Part 1a of the proof. The resulting vectors $\tilde{x} = [\tilde{x}_j]$, $\tilde{y} = [\tilde{y}_j]$, $\tilde{z} = [\tilde{z}_j]$ satisfy again (1.59), and –in view of (1.58)– they satisfy also (1.60).

Consequently, $\tilde{x}$, $\tilde{y}$, $\tilde{z}$ fulfill condition (1.49), so that $\tilde{x}$, $\tilde{y}$ satisfy (1.51) with the special seminorm defined above. Using property (1.28) and the last equality in (1.60), we find $\max_j \|y_j\|_\infty = \max_j \|\tilde{y}_j\| \leq \mu \cdot \max_k \|\tilde{x}_k\| = \mu$, which proves again (1.61).

Part 2a. Assume (1.29) and (1.55). We shall again prove (1.23b) via Lemma 1.3.1.

Denote by λ and $\varphi = [\varphi_i]$, $x = [x_j] = [[x_{ij}]]$, $y = [y_j] = [[y_{ij}]]$, $z = [z_j] = [[z_{ij}]]$ the same scalar and vectors as in Part 1a of the proof, so that (1.58) is again in force.

First, suppose $y_j \neq y_k$ for all $j \neq k$ belonging to the same index set $\mathfrak{I}_\rho$. Similarly as in Part 1a, we arrive at $\varphi = 0$, which proves (1.23b).

Next, suppose $y_q = y_s$ where $q < s$ belong to the same set $\mathfrak{I}_\rho$. Define $\tilde{x}_j$, $\tilde{y}_j$, $\tilde{z}_j$ as in Part 1a, but now with ξ, η, ζ satisfying (1.55). We have again (1.59), (1.60), and therefore

$$\|\tilde{z}_j\|_\infty = \max\{\|\tilde{z}_j\|, |\zeta_j|\} \leq \max\{\|\tilde{y}_j\|, |\eta_j|\} = \|\tilde{y}_j\|_\infty. \tag{1.62}$$

Hence, $\tilde{x}_j$, $\tilde{y}_j$, $\tilde{z}_j$ satisfy (1.49) with $\|\cdot\| = \|\cdot\|_\infty$. Via Lemma 1.4.1 and property (1.29) we obtain $\|\tilde{y}_j\|_\infty \leq \mu \cdot \|\xi\|_\infty$, and in view of (1.58), (1.60) there follows $\|\varphi\|_\infty \leq \mu \cdot \|\xi\|_\infty$.

By suitable scaling of ξ, η, ζ, with property (1.55), we can achieve that $\|\xi\|_\infty$ is arbitrarily close to zero. Hence, $\varphi = 0$, which proves (1.23b).

Part 2b. Assuming (1.29), (1.55), we shall prove (1.23c).

The beginning of the proof runs as in Part 1b above, using (1.55) instead of (1.54). We arrive again at (1.23c), via (1.61), if $y_j \neq y_k$ for all $j \neq k$ belonging to the same set $\mathfrak{I}_\rho$.

If $y_q = y_s$, for some $q < s$ belonging to the same $\mathfrak{I}_\rho$, we proceed as in Part 2a above, and introduce $\tilde{x}_j$, $\tilde{y}_j$, $\tilde{z}_j$ satisfying (1.49) with $\|\cdot\| = \|\cdot\|_\infty$. From Lemma 1.4.1 and property (1.29) it follows that $\|\tilde{y}_j\|_\infty \leq \mu \cdot \max_k\{1, \|\xi\|_\infty\}$, and in view of (1.58), (1.60) we obtain $\|\varphi\|_\infty \leq \mu \cdot \max_k\{1, \|\xi\|_\infty\}$.

By arranging that $\|\xi\|_\infty < 1$, we obtain (1.61) and therefore also (1.23c). □

Proof of Lemma 1.4.5

Part 1. For given $\xi = [\xi_i] \in \mathbb{R}^l$ and $\lambda = [\lambda_i] \in \mathbb{R}^m$, one can define $\eta = [\eta_i]$, $\zeta = [\zeta_i] \in \mathbb{R}^m$ by

$$\eta_i = \sum_k s_{ik}\,\xi_k + \sum_k t_{ik}\,\lambda_k, \quad \zeta_i = \eta_i + \lambda_i/\gamma \quad (1 \le i \le m). \tag{1.63}$$

The definition is easily seen to imply

$$\eta = R\,\xi + P\,\zeta. \tag{1.64}$$

This simple implication will be used, several times, below.

Assuming (1.56), one can see that ξ_i, λ_i exist, such that η_i, defined by (1.63), satisfy

$$\eta_i \ne \eta_j \quad (\text{for any } i \ne j \text{ in the same index set } \mathcal{I}_\rho). \tag{1.65}$$

Because (1.63) implies (1.64), it follows that ξ, η, ζ exist satisfying (1.54).

Part 2. Assuming (1.57), we shall determine scalars ε, μ_k, ξ_k, with

$$0 \le \varepsilon\,\mu_k \le 2\,\gamma, \tag{1.66}$$

such that the system of equations

$$\eta_i = \sum_k s_{ik}\,\xi_k - \varepsilon \sum_k t_{ik}\,\mu_k\,\eta_k \quad (1 \le i \le m) \tag{1.67}$$

has a solution $\eta = [\eta_i]$ satisfying (1.65). Using the implication (1.63) $\Rightarrow$ (1.64) (with $\lambda_i = -\varepsilon\,\mu_i\,\eta_i$), one sees that such scalars ε, μ_k, ξ_k lead to (1.55) (with $\zeta_i = (1 - \frac{\varepsilon\,\mu_i}{\gamma})\,\eta_i$).

To find ε, μ_k, ξ_k with the above properties, consider first any fixed μ_k, ξ_k, and note that the corresponding system (1.67) has a solution $\eta_i = \eta_i(\varepsilon)$, for $\varepsilon > 0$ small enough, with

$$\eta_i(\varepsilon) = \sigma_i - \varepsilon\,\tau_i + \mathcal{O}(\varepsilon^2) \ \ (\text{for } \varepsilon \downarrow 0), \ \ \sigma_i = \sum_k s_{ik}\,\xi_k, \ \ \tau_i = \sum_k t_{ik}\,\sigma_k\,\mu_k. \tag{1.68}$$

Aiming at (1.65) (with $\eta_i = \eta_i(\varepsilon)$), we are lead by (1.68) to fix ξ_k such that

$$\sigma_i \ne \sigma_j \ (\text{for } S(i,:) \ne S(j,:)), \quad \sigma_i \ne 0 \ (\text{for } S(i,:) \ne 0).$$

Below we shall specify μ_k, in terms of values ϱ_k which are determined such that $\operatorname{sgn}(\varrho_k) = \operatorname{sgn}(\sigma_k)$ (for $1 \le k \le m$) and $\sum_k \hat{t}_{ik}\,\varrho_k \ne \sum_k \hat{t}_{jk}\,\varrho_k$ (for $\hat{T}(i,:) \ne \hat{T}(j,:)$). We define $\mu_k = \varrho_k/\sigma_k$ (if $\sigma_k \ne 0$) and $\mu_k = 0$ (if $\sigma_k = 0$). It follows that

$$\mu_k \ge 0 \ (\text{for } 1 \le k \le m) \quad \text{and} \quad \tau_i \ne \tau_j \ (\text{for } \hat{T}(i,:) \ne \hat{T}(j,:)).$$

Because of (1.57), the values σ_i, τ_i corresponding to ξ_k, μ_k thus specified, satisfy

$$(\sigma_i,\,\tau_i) \ne (\sigma_j,\,\tau_j) \quad (\text{for any } i \ne j \text{ in the same index set } \mathcal{I}_\rho).$$

Combining these inequalities with (1.68), it follows that (1.65) (with $\eta_i = \eta_i(\varepsilon)$) and (1.66) hold for sufficiently small $\varepsilon > 0$. Hence ε, μ_k, ξ_k exist with the properties stated above. □

Chapter 2

Special boundedness properties in numerical initial value problems

For Runge-Kutta methods, linear multistep methods and other classes of general linear methods much attention has been paid in the literature to important nonlinear stability properties known as total variation diminishing (TVD), strong stability preserving (SSP) and monotonicity. Stepsize conditions guaranteeing these properties were studied by Shu & Osher (1988) and in numerous subsequent papers. Unfortunately, for many useful methods it has turned out that these properties do not hold. For this reason attention has been paid in the recent literature to the related and more general properties called total variation bounded (TVB) and boundedness.

In the present chapter we focus on stepsize conditions guaranteeing boundedness properties of a special type. These boundedness properties are optimal, and distinguish themselves also from earlier boundedness results by being relevant to sublinear functionals, discrete maximum principles and preservation of nonnegativity. Moreover, the corresponding stepsize conditions are more easily verified in practical situations than the conditions for general boundedness given thus far in the literature.

The theoretical results are illustrated by application to the two-step Adams-Bashforth method and a class of two-stage multistep methods.

2.1 Introduction

2.1.1 Bounds for numerical approximations

In this chapter we deal with the numerical solution of initial value problems of the form

$$\frac{d}{dt}u(t) = F(t, u(t)) \quad (t \geq 0), \quad u(0) = u_0. \tag{2.1}$$

We shall study a wide class of numerical methods for solving such problems; thereby basing our study on the analysis of an abstract generic numerical process

of the type

$$y_i = \sum_{j=1}^{l} s_{ij}\, x_j \;+\; \Delta t \cdot \sum_{j=1}^{m} t_{ij}\, F_j(y_j) \quad (1 \le i \le m). \tag{2.2}$$

Here $\Delta t > 0$ denotes the stepsize, the vectors x_j $(1 \le j \le l)$ are the input vectors of the process, and y_i $(1 \le i \le m)$ the output vectors. In applications to concrete numerical methods, the output vectors usually stand for approximations to the exact solution $u(t)$ of the differential equation at certain time levels $\bar{t}_i$, that is, $y_i \approx u(\bar{t}_i)$ $(1 \le i \le m)$, and $F_i(y_i) = F(\bar{t}_i, y_i)$.

The process (2.2) is in particular relevant to the important and very large class of general linear methods (GLMs), introduced by Butcher (1966), cf. also e.g. Butcher (1987), (2003), Hairer & Wanner (1996), Hairer, Nørsett & Wanner (1993). This class comprises, e.g., all Runge-Kutta methods, linear multistep methods and multistep-multistage variants thereof.

We can represent $N \ge 1$ consecutive steps of any GLM canonically by a process of the generic type (2.2) with $m = N(s + r)$, where s is the number of internal stages and r the number of external stages computed at each step of the GLM. In this situation, the vectors x_j $(1 \le i \le l)$ stand for the starting vectors of the GLM, whereas the vectors y_i $(1 \le i \le m)$ represent the $N \cdot s$ internal and $N \cdot r$ external stage approximations computed during the N steps. Furthermore, the parameter matrices $S = (s_{ij}) \in \mathbb{R}^{m \times l}$, $T = (t_{ij}) \in \mathbb{R}^{m \times m}$, corresponding to the process (2.2), are determined by the number of steps N as well as by the coefficients of the given GLM. Detailed examples of such representations, as well as alternative representations of actual multistep-multistage methods, can be found in Spijker (2007) for $N = 1$ and in Chapter 1 for $N > 1$; cf. also Section 2.4 of this chapter.

We denote by $\mathbb{V}$ the vector space on which the differential equation is defined, and by $||\cdot||$ a real functional on $\mathbb{V}$, i.e. $||v|| \in \mathbb{R}$ for all $v \in \mathbb{V}$. In the rest of the present section, we assume $||\cdot||$ to be a *convex functional*, i.e.

$$\|\lambda v + (1-\lambda)w\| \;\le\; \lambda\,\|v\| + (1-\lambda)\,\|w\| \qquad (\text{for } 0 \le \lambda \le 1 \text{ and } v, w \in \mathbb{V}). \tag{2.3}$$

In applications, $||\cdot||$ will often be a norm or seminorm, see (2.16) below. But, more general convex functionals are useful as well, notably in connection with discrete maximum principles and preservation of nonnegativity; cf. e.g. Spijker (2007) and Section 2.3.4 of the present chapter.

For the generic process (2.2), as well as for special instances thereof, much attention has been paid in the literature to the derivation of suitable upper bounds for $||y_i||$, in terms of the input vectors x_j, under the basic assumption that for given $\tau_0 > 0$

$$\|v + \tau_0\, F_i(v)\| \;\le\; \|v\| \quad (\text{for } 1 \le i \le m, \text{ and } v \in \mathbb{V}); \tag{2.4}$$

cf. e.g. Ferracina & Spijker (2004), Gottlieb, Ketcheson & Shu (2009), Gottlieb, Shu, & Tadmor (2001), Higueras (2004), (2005), Hundsdorfer & Ruuth (2003),

(2006), Hundsdorfer, Ruuth & Spiteri (2003), Shu & Osher (1988), Spijker (2007).

In most papers, the focus has been on the situation where the coefficients of the generic process satisfy the condition

$$s_{i\,1} + s_{i\,2} + \cdots + s_{i\,l} = 1 \quad (1 \leq i \leq m). \tag{2.5}$$

In case the process (2.2) stands for *just one step* ($N = 1$) of a GLM, this condition corresponds to preconsistency, cf. Spijker (2007), Butcher (2003). Henceforth we will refer to (2.5) as the *preconsistency condition* for process (2.2).

For preconsistent generic processes representing one step of a GLM, the bound

$$\|y_i\| \ \leq \ \max_{1\leq j\leq l} \|x_j\| \quad (\text{for } 1 \leq i \leq m) \tag{2.6}$$

has received much attention. The process has been called *monotonic* or *strongly stable* (for given stepsize Δt, vector space $\mathbb{V}$, functional $\|\cdot\|$ and functions $F_i : \mathbb{V} \to \mathbb{V}$) if the last bound holds whenever x_i and y_i satisfy (2.2). Algebraic characterizations were derived of stepsize-coefficients γ with the following important property:

> Condition $0 < \Delta t \leq \gamma \cdot \tau_0$ implies monotonicity, whenever $\mathbb{V}$ is a vector space, $\|\cdot\|$ a convex function on $\mathbb{V}$, and the functions $F_i : \mathbb{V} \to \mathbb{V}$ satisfy the basic assumption (2.4); (2.7)

see e.g. Spijker (2007) and the references therein.

Unfortunately, for many useful GLMs there exists *no* $\gamma > 0$ such that the above property is present, when one step of the method ($N = 1$) is represented as a preconsistent process of the form (2.2); some examples are given in Section 2.4 of this chapter. Furthermore, in important situations, processes of generic type (2.2) arise which even fail to satisfy the preconsistency condition (2.5). Sometimes, deeper insight into a given GLM can be gained by representing $N > 1$ consecutive steps of the method as such a generic process; cf. Section 2.4.

These difficulties have led various authors to study bounds for $\|y_i\|$ that differ from the monotonicity bound (2.6) by a factor $\mu \geq 1$, i.e.

$$\|y_i\| \ \leq \ \mu \cdot \max_{1\leq j\leq l} \|x_j\| \quad (\text{for } 1 \leq i \leq m). \tag{2.8}$$

Such general bounds are formally weaker than (2.6) but still useful because they can reveal essential boundedness properties of the numerical methods under consideration, like the property of being *total variation bounded* - for this important concept see e.g. LeVeque (2002). Stepsize conditions corresponding to general bounds (2.8) were derived, e.g., in Ruuth & Hundsdorfer (2005), Chapter 1 of this thesis.

The general bounds obtained thus far in the literature are relevant in cases where the monotonicity property (2.7) is violated or even the preconsistency condition (2.5) is not in force. On the other hand, these bounds suffer still from

the following two inconveniences: (1) the corresponding stepsize conditions, of type $0 < \Delta t \leq \gamma \cdot \tau_0$, involve complicated conditions on γ which are often difficult to check in practice; (2) the general bounds are relevant to seminorms but not to any wider class of functionals satisfying (2.3).

2.1.2 Scope of the chapter

The main purpose of the present chapter is to establish stepsize conditions guaranteeing special bounds for the generic process (2.2), thereby circumventing the two inconveniences just mentioned above. We shall find special bounds which can still be present in cases where the monotonicity property (2.7) or the preconsistency condition (2.5) is violated, and which are the best possible in a definite sense. Moreover, these special bounds are relevant to a class of functionals $||\cdot||$ that is wider than the class of seminorms. Finally, and most importantly in view of applications, the corresponding stepsize conditions $0 < \Delta t \leq \gamma \cdot \tau_0$ involve a condition on γ which is easier to check in practice than the conditions relevant to the general bounds given in the literature.

In Section 2.2 of this chapter, we review and extend bounds and monotonicity results for the generic process (2.2), as given thus far in the existing literature. In the Sections 2.2.1, 2.2.2, we give a brief review of known monotonicity results for the generic process (2.2), thereby focussing on a classical simple condition on the stepsize-coefficient γ. Moreover, we consider a property which is a-priori more refined than pure monotonicity and we characterize in Theorem 2.2.4 stepsize conditions guaranteeing this property. In Section 2.2.3, we specify two generalizations of the bound $\|y_i\| \leq \max_{1 \leq j \leq l} \|x_j\|$ which are relevant to generic processes which need not be preconsistent. Theorem 2.2.5 characterizes stepsize conditions guaranteeing these generalizations.

Section 2.3 contains the main theoretical findings of the chapter. In Section 2.3.1, we formulate explicitly, for the generic process (2.2), the special bounds mentioned above (for $||y_i||$ in terms of $||x_j||$), and mention three features which distinguish them from more general standard bounds (2.8). In Section 2.3.2, we study, in the situation of these special bounds, the characterizations provided by Theorem 2.2.5. We find simplified versions of these characterizations, viz. (2.25)-(2.28). In Section 2.3.3, we study the special bounds for the case of seminorms $||\cdot||$; we find that these bounds are the best possible in the sense specified by Theorem 2.3.4. The main theorem of Section 2.3.3, Theorem 2.3.5, gives simplified criteria for stepsize conditions guaranteeing the special bounds. Section 2.3.4 deals with the special bounds for the case of a natural class of functionals – the so-called sublinear functionals – which is essentially larger than the class of seminorms. Theorem 2.3.8 reveals the surprising fact that the special bounds are the only bounds which make sense in the context of general sublinear functionals. The main theorem of Section 2.3.4, Theorem 2.3.9, gives among other things a mild condition under which the classical simple condition on γ, reviewed in Section 2.2, characterizes stepsize conditions guaranteeing the special bounds for sublinear functionals.

In Section 2.4 we illustrate the significance of the special boundedness theory by applying it to some concrete numerical methods. For most of these methods, the monotonicity results, as given in the literature, see e.g. Gottlieb, Ketcheson & Shu (2009), Spijker (2007), are *not* (directly) applicable. Moreover, the boundedness theory, as given e.g. in Chapter 1 would lead to very complicated conditions. In Section 2.4.2 we study the two-step Adams-Bashforth method. When writing one step of the method in a standard fashion as a generic process of type (2.2), there is no $\gamma > 0$ such that the monotonicity property (2.7) is present. But, by writing $N \geq 1$ steps of the method judiciously in the generic form (2.2), it turns out that Theorems 2.3.5, 2.3.9 yield conclusions which can nicely be interpreted in terms of boundedness and nonnegativity preservation of the method. In Section 2.4.3 we analyse a large class of k-step methods, containing both predictor-corrector methods and hybrid multistep methods. The monotonicity results, known from the literature, are not valid for many popular schemes of this class. By applying Theorem 2.3.9, we will show that for many methods of practical interest relevant boundedness properties are valid.

2.2 Reviewing and extending results from the literature

2.2.1 Preliminaries

Let I stand for the identity matrix of order m, and let $S = (s_{ij})$, $T = (t_{ij})$ denote the coefficient matrices corresponding to the generic process (2.2). Similarly as in Chapter 1 of this thesis, Spijker (2007), we introduce the matrices

$$P = (p_{ij}) = (I + \gamma T)^{-1}(\gamma T), \qquad R = (r_{ij}) = (I + \gamma T)^{-1} S. \tag{2.9}$$

These matrices depend explicitly on γ, and they are defined if γ is such that $I + \gamma T$ is invertible.

When working with P and R, the invertibility of $I + \gamma T$ will be implicitly assumed. Actually, to study boundedness properties this assumption can be made without loss of generality. To see this, we formulate the following lemma, which is an analogue of a result from Spijker (2007, Lemma 4.2). The proof of this lemma is compact, so we repeat it here.

Lemma 2.2.1 (Invertibility of $I + \gamma T$)**.** *Let $\tau_0 > 0$, $\gamma > 0$ be given and $\Delta t = \gamma \cdot \tau_0$. Let $\mathbb{V} = \mathbb{R}$, $\|\cdot\| = |\cdot|$ and assume μ is a constant such that the general bound (2.8) holds whenever $F_i : \mathbb{V} \to \mathbb{V}$ fulfil the basic assumption (2.4) and $y_i, x_j \in \mathbb{V}$ satisfy (2.2). Then $I + \gamma T$ is invertible.*

Proof. Let $\eta = [\eta_i] \in \mathbb{R}^m$ such that $(I + \gamma T)\,\eta = 0$. We shall prove $\eta = 0$.

We define $F_i(v) = -(1/\tau_0)\, v$ (for all $v \in \mathbb{V}$), so that the basic assumption (2.4) is fulfilled with $\|\cdot\| = |\cdot|$. Clearly, (2.2) is satisfied, with $\Delta t = \gamma \cdot \tau_0$, by the vectors $x_i = 0$ $(1 \leq i \leq l)$ and $y_i = \eta_i$ $(1 \leq i \leq m)$. By applying (2.8), there follows $|\eta_i| = |y_i| \leq \mu \cdot \max_j |x_j| = 0$, therefore $\eta = 0$. □

In the following, we shall frequently deal with values γ satisfying the condition that

$$(I + \gamma\, T)^{-1}(\gamma T) \geq 0, \qquad (I + \gamma\, T)^{-1} S \geq 0.$$

These inequalities – as well as any other inequalities between matrices appearing below – should be interpreted entry-wise. The above condition can evidently be rewritten, less explicitly but more simply, as

$$P \geq 0, \qquad R \geq 0. \tag{2.10}$$

This form can more easily be compared (than the more explicit form) with a series of conditions on γ to be studied in the rest of the chapter. In view of the essential use of the above condition made (directly or indirectly) in the existing literature on monotonicity, we will refer to it as the *classical condition on* γ.

2.2.2 Monotonicity with arbitrary convex functionals $||\cdot||$

We shall recall briefly some concepts and results from the literature which are related to the monotonicity property (2.7). The next two theorems follow directly from Spijker (2007, Theorems 2.2, 2.4).

Theorem 2.2.2 (Criterion for monotonicity with arbitrary convex functional $||\cdot||$)**.** *Consider a generic process (2.2) satisfying the preconsistency condition (2.5). Let $\gamma > 0$ be given. Then the monotonicity property (2.7) is present, if and only if γ satisfies the classical condition (2.10).*

In the following, we use, for any given matrix $A = (a_{ij})$, the notation $\mathrm{Inc}(A)$ to denote the *incidence matrix* of A, given by

$$\mathrm{Inc}(A) = (\hat{a}_{ij}), \quad \text{where } \hat{a}_{ij} = 1 \ (\text{if } a_{ij} \neq 0), \ \ \hat{a}_{ij} = 0 \ (\text{if } a_{ij} = 0).$$

Theorem 2.2.3 (Conditions on S, T)**.** *Let the preconsistency condition (2.5) be fulfilled. Then there is a $\gamma > 0$ satisfying the classical condition (2.10), if and only if $S \geq 0$, $T \geq 0$, $\mathrm{Inc}(TS) \leq \mathrm{Inc}(S)$ and $\mathrm{Inc}(T^2) \leq \mathrm{Inc}(T)$.*

Clearly, for given matrices S, T, it is rather easy, by applying Theorems 2.2.2 and 2.2.3, to see whether there is a positive stepsize-coefficient γ such that the monotonicity property (2.7) is present.

For preconsistent processes, the classical condition (2.10) will be proved to imply an interesting variant of the standard monotonicity bound $\|y_i\| \leq \max_{1\leq j\leq l} \|x_j\|$. The variant is as follows:

$$\|y_i\| \ \leq \sum_{j=1}^{l} |s_{ij}|\, \|x_j\| \quad (1 \leq i \leq m). \tag{2.11}$$

Note that, when all s_{ij} are nonnegative, the last bound is of particular interest because it is more refined and gives, in general, more information than the

standard monotonicity bound. Clearly, all s_{ij} are nonnegative as soon as the monotonicity property (2.7) is present for some $\gamma > 0$; cf. Theorems 2.2.2, 2.2.3.

We shall say that *process (2.2) satisfies the bound (2.11)* (for given stepsize Δt, vector space $\mathbb{V}$, functional $\|\cdot\|$ and functions $F_i : \mathbb{V} \to \mathbb{V}$), if (2.11) holds whenever x_i and $y_i \in \mathbb{V}$ satisfy (2.2). The following (refined) property is an obvious variant of the standard monotonicity property (2.7):

> Condition $0 < \Delta t \leq \gamma \cdot \tau_0$ implies the bound (2.11), whenever $\mathbb{V}$ is a vector space, $\|\cdot\|$ a convex functional on $\mathbb{V}$, and the functions $F_i : \mathbb{V} \to \mathbb{V}$ satisfy the basic assumption (2.4). (2.12)

The following theorem shows that this property is present under the same conditions as the standard monotonicity property (2.7).

Theorem 2.2.4 (Criterion for property (2.12))**.** *Consider a generic process (2.2) satisfying the preconsistency condition (2.5). Let $\gamma > 0$ be given. Then property (2.12) is present, if and only if γ satisfies the classical condition (2.10).*

Proof. 1. Let the basic assumption (2.4) be fulfilled, and let $0 < \Delta t \leq \gamma \cdot \tau_0$, where γ satisfies the classical condition (2.10). We denote by E_k the $k \times 1$ matrix with all entries equal to 1. Note that, since $R = (I - P)S$ and $SE_l = E_m$, we have $RE_l + PE_m = E_m$, i.e. $\sum_{j=1}^{l} r_{ij} + \sum_{j=1}^{m} p_{ij} = 1$.

We rewrite process (2.2), using the notations (2.9), in the form

$$y_i = \sum_{j=1}^{l} r_{ij}\, x_j \;+\; \sum_{j=1}^{m} p_{ij}\, \big(y_j + \theta\tau_0 F_j(y_j)\big) \quad (1 \leq i \leq m), \quad \theta = \frac{\Delta t}{\gamma\tau_0}.$$

We denote the column vector in $\mathbb{R}^l$ with components $||x_i||$ by $[||x_i||]$, and we use a similar notation with regard to y_i and $F_i(y_i)$. Using the convexity property of the functional $\|\cdot\|$, there follows $[||y_i||] \leq R[||x_j||] + P[||y_i + \theta\tau_0 F_i(y_i)||]$. Because $P \geq 0$, we have $P[||y_i + \theta\tau_0 F_i(y_i)||] = P[||\theta(y_i + \tau_0 F_i(y_i)) + (1-\theta)y_i||] \leq P[||y_i||]$, so that

$$[||y_i||] \leq (I + \gamma T)^{-1} S[||x_j||] + (I - (I + \gamma T)^{-1})[||y_i||], \tag{2.13}$$

i.e. $(I + \gamma T)^{-1}[||y_i||] \leq (I + \gamma T)^{-1} S[||x_j||]$. In view of Theorem 2.2.3, the matrices S and $I + \gamma T$ are nonnegative, so that the bound (2.11) is in force. Property (2.12) has thus been proved.

2. Conversely, assume property (2.12) is present. We shall use the notation

$$\operatorname{sgn}(\alpha) = 1 \;\; (\text{for } \alpha \geq 0), \quad \operatorname{sgn}(\alpha) = -1 \;\; (\text{for } \alpha < 0).$$

Applying property (2.12) in the special situation where $\mathbb{V} = \mathbb{R}$, $||v|| = v$, $F_i = 0$, $x_j = \operatorname{sgn}(s_{i_0 j})$, we see from the corresponding bound (2.11) that $\sum_j |s_{i_0 j}| \leq \sum_j |s_{i_0 j}| \operatorname{sgn}(s_{i_0 j})$, so that $s_{i_0 j} \geq 0$. Hence, all $s_{ij} \geq 0$.

In the general situation, the bound (2.11) thus implies, for $1 \leq i \leq m$,

$$\|y_i\| \;\leq\; \sum_n s_{in}\|x_n\| \;\leq\; \big(\sum_n s_{in}\big)\, \max_j \|x_j\| \;=\; \max_j \|x_j\|.$$

It follows that property (2.12) implies the standard monotonicity property (2.7) and – by Theorem 2.2.2 – also the classical condition (2.10). □

2.2.3 General bounds with seminorms $||\cdot||$

With an eye to cases where the preconsistency condition (2.5) or the monotonicity property (2.7) (with $\gamma > 0$) is violated, we shall review and extend, in this section, some results from the literature about bounds which are more general than those considered above. We shall focus on the general bounds

$$\|y_i\| \;\leq\; \mu_i \cdot \max_{1\leq j\leq l} \|x_j\| \quad (\text{for } 1 \leq i \leq m), \tag{2.14}$$

$$\|y_i\| \leq \sum_{j=1}^{l} \mu_{ij} \, \|x_j\| \quad (\text{for } 1 \leq i \leq m), \tag{2.15}$$

where for the time being μ_i and μ_{ij} denote arbitrary coefficients. Clearly, when $\mu_i = 1$, $\mu_{ij} = |s_{ij}|$, these bounds reduce to the bounds (2.6) and (2.11), respectively.

In this section, we shall deal with the situation where $||\cdot||$ is a *seminorm*, i.e.

$$||v + w|| \leq ||v|| + ||w|| \;\text{ and }\; ||\lambda\, v|| = |\lambda|\, ||v|| \tag{2.16}$$

for all real λ and v, $w \in \mathbb{V}$. We shall say that *process (2.2) satisfies the bound (2.14) or (2.15)* (for given stepsize Δt, vector space $\mathbb{V}$, seminorm $\|.\|$ and functions $F_i : \mathbb{V} \to \mathbb{V}$), if (2.14) or (2.15), respectively, holds whenever x_i and $y_i \in \mathbb{V}$ satisfy (2.2). Below we shall focus on stepsize-coefficients γ which are related to the above two general bounds by means of the following two properties:

Condition $0 < \Delta t \leq \gamma \cdot \tau_0$ implies that process (2.2) satisfies the bound (2.14), whenever $\mathbb{V}$ is a vector space, $\|\cdot\|$ a seminorm on $\mathbb{V}$, and the functions $F_i : \mathbb{V} \to \mathbb{V}$ satisfy the basic assumption (2.4), (2.17)

Condition $0 < \Delta t \leq \gamma \cdot \tau_0$ implies that process (2.2) satisfies the bound (2.15), whenever $\mathbb{V}$ is a vector space, $\|\cdot\|$ a seminorm on $\mathbb{V}$, and the functions $F_i : \mathbb{V} \to \mathbb{V}$ satisfy the basic assumption (2.4). (2.18)

In formulating conditions on γ for these properties, we need some notations. For any matrix $A = (a_{ij})$, we define the matrix $|A|$ by $|A| = (|a_{ij}|)$. For square matrices A, we denote the *spectral radius* by $\mathrm{spr}(A)$. Furthermore we introduce the $m \times 1$ matrix

$$(\mu_i) = (\mu_1, \mu_2, \cdots, \mu_m)^T$$

and the $m \times l$ matrix

$$(\mu_{ij}) = \begin{pmatrix} \mu_{11} & \cdots & \mu_{1l} \\ \vdots & & \vdots \\ \mu_{m1} & \cdots & \mu_{ml} \end{pmatrix}. \tag{2.19}$$

We shall relate properties (2.17) and (2.18), respectively, to the following two requirements:

$$\mathrm{spr}(|P|) < 1 \quad \text{and} \quad (I - |P|)^{-1} |R|\, E_l \leq (\mu_i), \tag{2.20}$$

$$\mathrm{spr}(|P|) < 1 \quad \text{and} \quad (I - |P|)^{-1} |R| \leq (\mu_{ij}). \tag{2.21}$$

Note that, for given coefficient matrices S, T, these requirements amount to conditions on γ - cf. the definition (2.9) of P, R.

The following theorem is a variant of a result given earlier in Chapter 1. In fact, when all μ_i are equal to each other, part (I) of the theorem is an immediate corollary to Theorem 1.2.2 in Chapter 1 just mentioned.

Theorem 2.2.5 (Criteria for the properties (2.17), (2.18))**.** *Consider an arbitrary generic process (2.2). Let $\gamma > 0$ and arbitrary μ_i, μ_{ij} be given. Then the following two propositions are valid:*

(I) Property (2.17) is present, if and only if γ is such that condition (2.20) is fulfilled.

(II) Property (2.18) is present, if and only if γ is such that condition (2.21) is fulfilled.

Proof. The conditions (2.20), (2.21) imply (2.17) and (2.18), respectively, by similar arguments as used in part 1 of the proof of Theorem 2.2.4. Using the arguments of the mentioned proof and (2.16), we get now $[||y_i||] \leq |R|[||x_j||] + |P|[||y_i||]$ instead of (2.13). There follows $[||y_i||] \leq (I - |P|)^{-1} |R|[||x_j||]$. By (2.21) we arrive at property (2.18). Since $(I-|P|)^{-1}|R|[||x_j||] \leq (I-|P|)^{-1}|R|E_l \cdot \max_k ||x_k||]$, by (2.20) we arrive at property (2.17).

The necessity of the conditions (2.20) and (2.21) can be proved by almost the same arguments as already given in Chapter 1 (Section 1.4.2). □

Theorem 2.2.5 has a wider scope, certainly, than the theorems of Section 2.2.2, in that μ_i and μ_{ij} are arbitrary coefficients and the preconsistency condition (2.5) is not needed.

On the other hand, it is in general much more difficult to see whether the conditions (2.20), (2.21) are fulfilled than to check the classical condition (2.10). Moreover, unlike the theorems in Section 2.2.2, Theorem 2.2.5 is only relevant to seminorms (and e.g. not to certain convex functionals arising in connection with discrete maximum principles and preservation of nonnegativity). These obvious weaknesses of Theorem 2.2.5 are among the reasons for dealing in the following section with bounds of a very special form.

2.3 Bounds of a special form

2.3.1 Special choices for μ_i, μ_{ij}

Below we shall focus on the bounds of the preceding subsection in the case where

$$\mu_i = \textstyle\sum_j |s_{ij}| \qquad \text{and} \qquad \mu_{ij} = |s_{ij}|, \tag{2.22}$$

so that the general bounds (2.14), (2.15), respectively, take the special form

$$\|y_i\| \;\le\; \Big(\sum_{j=1}^{l} |s_{ij}|\Big) \cdot \max_{1\le j\le l} \|x_j\| \qquad (1 \le i \le m), \tag{2.23}$$

$$\|y_i\| \;\le\; \sum_{j=1}^{l} |s_{ij}|\, \|x_j\| \qquad (1 \le i \le m). \tag{2.24}$$

Below we list three important features by which these special bounds distinguish themselves from the general bounds considered in the preceding subsection.

First of all, property (2.17) with $\mu_i = \sum_j |s_{ij}|$, as well as property (2.18) with $\mu_{ij} = |s_{ij}|$, can be interpreted as an extension, to *all* F_i (satisfying the basic assumption(2.4)), of a bound which is trivially fulfilled when $F_i(v) \equiv 0$. In fact, in the subsequent Theorem 2.3.4, we shall see that the above special bounds (2.23), (2.24) are the *best possible*, in the sense that, for any $\gamma > 0$, the general boundedness properties (2.17), (2.18) *cannot* be valid with coefficients smaller than (2.22).

Secondly, as will be seen in Theorem 2.3.8 below, the equalities $\mu_i = \sum_j |s_{ij}|$ are necessary in order that any bound (2.14) holds for a natural class of functionals $||\cdot||$ that is larger than the class of seminorms. Similarly, the equalities $\mu_{ij} = |s_{ij}|$ must be fulfilled in order that any bound (2.15) holds for this larger class.

Finally and most importantly in view of applications, the above criteria (2.20) and (2.21), respectively, will turn out to reduce to much simpler forms when $\mu_i = \sum_j |s_{ij}|$ or $\mu_{ij} = |s_{ij}|$.

2.3.2 Simplified conditions when $\mu_i = \sum_j |s_{ij}|$ and $\mu_{ij} = |s_{ij}|$

In this section we shall analyse and simplify the above general conditions (2.20), (2.21) in the special situations $\mu_i = \sum_j |s_{ij}|$ and $\mu_{ij} = |s_{ij}|$. Our first result is as follows:

Lemma 2.3.1 (Conditions (2.20), (2.21) with $\mu_i = \sum_j |s_{ij}|$ and $\mu_{ij} = |s_{ij}|$)**.** *Condition (2.20) with $\mu_i = \sum_j |s_{ij}|$ is equivalent to (2.21) with $\mu_{ij} = |s_{ij}|$.*

Proof. To prove the lemma, we assume $\mathrm{spr}(|P|) < 1$ and $\mu_i = \sum_j |s_{ij}|$, $\mu_{ij} = |s_{ij}|$.

Suppose condition (2.20) is fulfilled. Since $(\mu_i) = |S|E_l = |(I-P)^{-1}R|E_l \le (I-|P|)^{-1}|R|E_l$, condition (2.20) is equivalent to $|S|E_l = (I-|P|)^{-1}|R|E_l$, which can be rewritten as $|S|E_l = |P||S|E_l + |R|E_l$. Because of the last equality and $|S| = |R+PS| \le |R| + |P||S|$, it follows that

$$|S| = |P||S| + |R|.$$

Hence, $(I-|P|)^{-1}|R| = |S| = (\mu_{ij})$, which implies (2.21).

Conversely, (2.21) implies $(I-|P|)^{-1}|R|E_l \le |S|E_l$, i.e. (2.20). □

Below, we shall specify situations in which the general conditions (2.20) and (2.21) can be simplified to one of the subsequent four requirements:

$$\operatorname{spr}(|P|) < 1 \quad \text{and} \quad |P\,S| = |P|\,|S| \le |S|, \;\; |R| \le |S|; \tag{2.25}$$

$$\operatorname{spr}(|P|) < 1 \quad \text{and} \quad P\,S = |P|\,S, \quad R \ge 0; \tag{2.26}$$

$$\operatorname{spr}(P) < 1 \quad \text{and} \quad P \ge 0, \quad R \ge 0; \tag{2.27}$$

$$P \ge 0, \quad R \ge 0, \quad S \ge 0. \tag{2.28}$$

Lemma 2.3.2 (Simplifications of (2.20), (2.21) with $\mu_i = \sum_j |s_{ij}|$, $\mu_{ij} = |s_{ij}|$)**.**
(I) Condition (2.20) as well as condition (2.21), with the choice (2.22), is equivalent to (2.25).
(II) If $S \ge 0$, then condition (2.25) is equivalent to (2.26).
(III) If S has no row equal to zero, then the three conditions (2.26), (2.27) and (2.28) are equivalent to each other.

Proof. (I) In view of Lemma 2.3.1, it is enough to show that condition (2.21) with $\mu_{ij} = |s_{ij}|$ is equivalent to (2.25).

From the proof of Lemma 2.3.1 it is evident that condition (2.21), with $\mu_{ij} = |s_{ij}|$, is equivalent to

$$\operatorname{spr}(|P|) < 1 \;\text{ and }\; |S| = |P||S| + |R|. \tag{2.29}$$

The last equality implies $|P|\,|S| = |P\,S|$, because $|S| = |PS+R| \le |PS|+|R| \le |P||S| + |R|$. Furthermore, because $S = PS + R$, we have

$$|S| = |PS| + |R|$$

as soon as $|PS| \le |S|$ and $|R| \le |S|$. It follows that condition (2.29) is equivalent to (2.25).

(II) Assume $S \ge 0$. In order to prove the equivalence of (2.25) and (2.26), assume $\operatorname{spr}(|P|) < 1$.

Suppose (2.25) is fulfilled. Since $R = S - PS$ and $|S| = |S - PS| + |PS|$, we have

$$|R| + |PS| = S = R + PS \le R + |PS| = R + |P|S,$$

which implies $R \ge 0$ and $P\,S = |P|\,S$. Therefore we have (2.26).

Conversely, from (2.26) and $S = R + PS$ we have

$$|P||S| + |R| = |S| = |PS + R| \le |PS| + |R| \le |P||S| + |R|.$$

Hence, (2.26) implies (2.25).

(III) Assume S has no row equal to zero. We shall prove successively that (2.26) $\Rightarrow$ (2.27) $\Rightarrow$ (2.28) $\Rightarrow$ (2.27) $\Rightarrow$ (2.26).

Assume (2.26). Since $(I-|P|)S \ge 0$, we have $S = (I-|P|)^{-1}(I-|P|)S \ge 0$. Denoting by σ_i the entries of SE_l, we have $\sigma_i = \sum_j s_{ij} > 0$ (for $1 \le i \le m$).

Since $(|P|-P)S = 0$, we have $(|P|-P)SE_l = 0$ and thus $\sum_j (|p_{ij}| - p_{ij})\sigma_j = 0$. Hence, $P \geq 0$, and therefore we have (2.27).

Furthermore, (2.27) implies that $S = (I-P)^{-1}R = (I+P+P^2+\ldots)R \geq 0$, so that (2.27) implies (2.28).

In order to prove that property (2.28) leads to (2.27), it is enough to show that $\mathrm{spr}(P) < 1$. Introducing $D = \mathrm{Diag}(\sigma_1, \ldots, \sigma_m)$ with $\sigma_i = \sum_j s_{ij}$, we have

$$D^{-1}PDE_m = D^{-1}PSE_l = D^{-1}(S-R)E_l \leq D^{-1}SE_l = E_m.$$

It follows that $\mathrm{spr}(P) = \mathrm{spr}(D^{-1}PD) \leq 1$. Since $P = I - (I+\gamma T)^{-1} \geq 0$ has no eigenvalue 1, we conclude from the Perron-Frobenius theory (see e.g. Horn (1988, p. 503)) that $\mathrm{spr}(P) < 1$.

It is easy to see that (2.27) leads to (2.26). □

Remark 2.3.3. Let $\gamma > 0$. Then condition (2.27) is equivalent to

$$P \geq 0, \quad R \geq 0, \quad T \geq 0. \tag{2.30}$$

In order to show this, first assume (2.27). Then $I + \gamma T = (I-P)^{-1} = I + P + P^2 + \ldots \geq I$, which yields (2.30).

Next suppose that (2.30) is fulfilled. Applying the Perron-Frobenius theory as presented e.g. in Horn (1988, p. 503), it follows that there is a vector $x \in \mathbb{R}^m$ with $0 \leq x \neq 0$, such that $Px = \lambda x$ where $\lambda = \mathrm{spr}(P)$. Clearly, $(I+\gamma T)^{-1}x = (I-P)x = (1-\lambda)x$, and therefore

$$x = (1-\lambda)(I+\gamma T)x.$$

Because $Tx \geq 0$, the assumption that $\lambda \geq 1$, would lead to:

$$0 \leq x = (1-\lambda)x + \gamma(1-\lambda)Tx \leq (1-\lambda)x \leq 0.$$

This would imply $x = 0$, which is a contradiction; therefore $\mathrm{spr}(P) < 1$. □

2.3.3 Special bounds with seminorms $||\cdot||$

Clearly, with the choice (2.22), the general properties (2.17), (2.18), respectively, reduce to

> Condition $0 < \Delta t \leq \gamma \cdot \tau_0$ implies that process (2.2) satisfies the special bound (2.23), whenever $\mathbb{V}$ is a vector space, $\|\cdot\|$ a seminorm on $\mathbb{V}$, and the functions $F_i : \mathbb{V} \to \mathbb{V}$ satisfy the basic assumption (2.4). (2.31)

> Condition $0 < \Delta t \leq \gamma \cdot \tau_0$ implies that process (2.2) satisfies the special bound (2.24), whenever $\mathbb{V}$ is a vector space, $\|\cdot\|$ a seminorm on $\mathbb{V}$, and the functions $F_i : \mathbb{V} \to \mathbb{V}$ satisfy the basic assumption (2.4). (2.32)

In this section we shall analyse these two special properties, and arrive at relatively simple conditions on γ for the properties to be present.

But, we shall first present Theorem 2.3.4, which shows a crucial feature of the statements (2.31), (2.32): the theorem tells us that the estimates (2.23) and (2.24) occurring in these statements are the best possible, in the sense that, for any $\gamma > 0$, the more general properties (2.17), (2.18) cannot be valid with smaller choices for μ_i and μ_{ij} than (2.22). We have

Theorem 2.3.4 (Lower bounds for μ_i and μ_{ij})**.**
(I) If $\gamma > 0$ and μ_i are such that property (2.17) holds, then $\mu_i \geq \sum_j |s_{ij}|$ (for $1 \leq i \leq m$).
(II) If $\gamma > 0$ and μ_{ij} are such that property (2.18) holds, then $\mu_{ij} \geq |s_{ij}|$ (for $1 \leq i \leq m$, $1 \leq j \leq l$).

Proof. In order to prove statement (I), assume property (2.17) is valid with $\gamma > 0$ and $\mu_{i_0} < \sum_j |s_{i_0 j}|$ for some index i_0. Then, in the situation where $\mathbb{V} = \mathbb{R}$, $||v|| = |v|$, $F_i = 0$ and $x_j = \mathrm{sgn}(s_{i_0 j})$, we have

$$\left\|\sum_j s_{i_0 j} x_j\right\| \leq \mu_{i_0} \cdot \max_{1 \leq j \leq l} \|x_j\| < \sum_j |s_{i_0 j}| = \left\|\sum_j s_{i_0 j} x_j\right\|.$$

This yields a contradiction, so that (I) must be true.

To prove statement (II), assume property (2.18) is present with $\gamma > 0$ and $\mu_{i_0 j_0} < |s_{i_0 j_0}|$ for some pair (i_0, j_0). Then, applying this property to the situation where $\mathbb{V} = \mathbb{R}$, $\|v\| = |v|$, $F_i = 0$, $x_j = \mathrm{sgn}(s_{i_0 j})$ (for $j = j_0$) and $x_j = 0$ (for $j \neq j_0$), we arrive at

$$\|s_{i_0 j_0} x_{j_0}\| \leq \mu_{i_0 j_0} \|x_{j_0}\| < |s_{i_0 j_0}| = \|s_{i_0 j_0} x_{j_0}\|.$$

This yields again a contradiction, so that statement (II) must be true. □

Our main result about the special boundedness properties (2.31), (2.32) will be formulated in Theorem 2.3.5. The theorem shows that criteria for these properties are possible which are in general much simpler than the criteria, given in Section 2.2.3, for the more general boundedness properties (2.17), (2.18).

Theorem 2.3.5 (Simplified criteria for the special properties (2.31) and (2.32))**.** *Consider an arbitrary generic process (2.2), and let $\gamma > 0$. Then the following propositions are valid:*
(I) Condition (2.25) is necessary and sufficient for property (2.31) as well as for property (2.32).
(II) If $S \geq 0$, then condition (2.26) is necessary and sufficient for property (2.31) as well as for property (2.32).
(III) If $S \geq 0$ has no row equal to zero, then the classical condition (2.10) is necessary and sufficient for property (2.31) as well as for (2.32).

Proof. Part (I) follows from a combination of Theorem 2.2.5 and Lemma 2.3.2.

Part (II) follows from part (I) and Lemma 2.3.2.

In order to prove statement (III), assume $S \geq 0$ has no row equal to zero. Combining part (II) of Theorem 2.3.5 and part (III) of Lemma 2.3.2, it follows that property (2.31) as well as (2.32) is equivalent to condition (2.28). Because $S \geq 0$, the last condition is equivalent to the classical condition (2.10). □

Property (2.32) is a-priori stronger than (2.31). Therefore the essence of the above theorem is that conditions (2.25), (2.26) and (2.10), under the appropriate assumptions on S, imply the strong statement (2.32), whereas already the weaker statement (2.31), under the same assumptions on S, implies conditions (2.25), (2.26) and (2.10).

2.3.4 Special bounds with general sublinear functionals $||\cdot||$

In this section we shall deal with bounds for $\|y_i\|$, where the functional $||\cdot||$ is *not* necessarily a seminorm. The following two examples provide some motivation for dealing with such bounds.

Example 2.3.6. Consider the functionals $\|v\| = \|v\|_+$ and $\|v\| = \|v\|_-$ defined by

$$\|v\|_+ = \max_i v_i\,, \quad \|v\|_- = -\min_i v_i \tag{2.33}$$

for $v = (v_1, v_2, \ldots, v_M)^T \in \mathbb{V} = \mathbb{R}^M$. These two functionals are no seminorms. But, they are highly relevant to *discrete maximum principles* for actual numerical processes, cf. Hundsdorfer & Verwer (2003, p. 118), Spijker (2007, p. 1235). ◇

Example 2.3.7. Another useful functional which fails to be a seminorm, is given by

$$\|v\|_0 = -\min\{0, v_1, \ldots, v_M\} \tag{2.34}$$

for $v = (v_1, v_2, \ldots, v_M)^T \in \mathbb{V} = \mathbb{R}^M$. For this non-negative functional we have $||v||_0 = 0$ if and only if $v \geq 0$, where this inequality is to be interpreted component-wise. One sees that any boundedness property $\max_i \|y_i\|_0 \leq \mu \cdot \max_j \|x_j\|_0$ implies the *preservation-of-nonnegativity* property: $y_i \geq 0$ (for $1 \leq i \leq m$) whenever all $x_j \geq 0$. For the practical relevance of this property, e.g. in the numerical solution of reaction-diffusion-convection equations, one may consult e.g. Hundsdorfer & Verwer (2003). ◇

Since the above functionals $\|v\|_+$, $\|v\|_-$ and $\|v\|_0$ violate the seminorm condition (2.16), the material of Sections 2.2.3 and 2.3.3 does *not* apply. It is therefore natural to look for versions of Theorems 2.2.5, 2.3.4 and 2.3.5 which are relevant to classes of functionals that are larger than the one specified by (2.16). Below we shall focus on functionals $||\cdot||$ which are only required to be *sublinear*, i.e.

$$||\alpha v + \beta w|| \leq \alpha||v|| + \beta||w|| \qquad \text{(for all } \alpha,\, \beta \geq 0 \text{ and } v,\, w \in \mathbb{V}). \tag{2.35}$$

Note that this requirement is equivalent to $||v+w|| \leq ||v||+||w||$, $||\lambda\, v|| = \lambda\, ||v||$ (for all $\lambda \geq 0$ and $v,\, w \in \mathbb{V}$). One easily sees that the three functionals in the above examples are sublinear.

In line with the above, we shall study the question for which values $\gamma > 0$ the process (2.2) has either of the following two general boundedness properties:

Condition $0 < \Delta t \leq \gamma \cdot \tau_0$ implies the bound (2.14), whenever $\mathbb{V}$ is a vector space, $\|\cdot\|$ a sublinear functional on $\mathbb{V}$, and the functions $F_i : \mathbb{V} \to \mathbb{V}$ satisfy the basic assumption (2.4). (2.36)

Condition $0 < \Delta t \leq \gamma \cdot \tau_0$ implies the bound (2.15), whenever $\mathbb{V}$ is a vector space, $\|\cdot\|$ a sublinear functional on $\mathbb{V}$, and the functions $F_i : \mathbb{V} \to \mathbb{V}$ satisfy the basic assumption (2.4). (2.37)

The following theorem may be viewed as a variant of Theorem 2.3.4 tuned to sublinear functionals. It shows, somewhat surprisingly, that we loose nothing by focusing on bounds with the coefficients (2.22).

Theorem 2.3.8 (Expressions for μ_i and μ_{ij})**.**
(I) If $\gamma > 0$ and μ_i are such that property (2.36) is present, then $\mu_i = \sum_j |s_{ij}|$ $(1 \leq i \leq m)$ and $S \geq 0$.
(II) If $\gamma > 0$ and μ_{ij} are such that property (2.37) is present, then $\mu_{ij} = |s_{ij}|$ $(1 \leq i \leq m,\ 1 \leq j \leq l)$ and $S \geq 0$.

Proof. (I) It follows from Theorem 2.3.4 that

$$\textstyle\sum_j |s_{ij}| \leq \mu_i \ \ (\text{for } 1 \leq i \leq m). \tag{2.38}$$

Applying property (2.36) to the situation where $\mathbb{V} = \mathbb{R}$, $||v|| = v$, $F_i(v) \equiv 0$, and choosing successively all $x_j = 1$ and all $x_j = -1$, we find $\sum_j s_{ij} \leq \mu_i$ and $(-\sum_j s_{ij}) \leq (-\mu_i)$, respectively. Hence

$$\textstyle\mu_i = \sum_j s_{ij} \ \ (\text{for } 1 \leq i \leq m).$$

Combining this equality and (2.38), we arrive at proposition (I).

(II) It follows from Theorem 2.3.4 that

$$\textstyle\sum_j |s_{ij}| \leq \sum_j \mu_{ij} \ \ (\text{for } 1 \leq i \leq m). \tag{2.39}$$

Applying property (2.37) to the situation where $\mathbb{V} = \mathbb{R}$, $||v|| = v$, $F_i(v) \equiv 0$, we conclude that $\sum_j s_{ij} x_j = y_i \leq \sum_j \mu_{ij} x_j$ $(1 \leq i \leq m)$, for all real values x_j. This implies

$$\mu_{ij} = s_{ij} \ \ (\text{for } 1 \leq i \leq m,\ 1 \leq j \leq l).$$

Combining this equality and (2.39), we arrive at proposition (II). □

Theorem 2.3.8 shows that the special bounds (2.23), (2.24), respectively, are the only bounds of type (2.14), (2.15) which make sense in the context of general sublinear functionals $\|\cdot\|$. Accordingly, we shall focus on the following special versions of the general properties (2.36) and (2.37), respectively:

Condition $0 < \Delta t \leq \gamma \cdot \tau_0$ implies that process (2.2) satisfies the special bound (2.23), whenever $\mathbb{V}$ is a vector space, $\|\cdot\|$ a sublinear functional on $\mathbb{V}$, and the functions $F_i : \mathbb{V} \to \mathbb{V}$ satisfy the basic assumption (2.4), (2.40)

Condition $0 < \Delta t \leq \gamma \cdot \tau_0$ implies that process (2.2) satisfies the special bound (2.24), whenever $\mathbb{V}$ is a vector space, $\|\cdot\|$ a sublinear functional on $\mathbb{V}$, and the functions $F_i : \mathbb{V} \to \mathbb{V}$ satisfy the basic assumption (2.4). (2.41)

Our main result about these two properties has been formulated in Theorem 2.3.9. The theorem can be regarded as a neat version of Theorem 2.3.5, parts (I) and (III), adapted to sublinear functionals.

Theorem 2.3.9 (Criteria for the properties (2.40) and (2.41))**.** *Consider an arbitrary generic process (2.2), and let $\gamma > 0$. Then the following propositions are valid:*

(I) Condition (2.27) is necessary and sufficient for property (2.40) as well as for property (2.41).

(II) If $S \geq 0$ has no row equal to zero, then the classical condition (2.10) is necessary and sufficient for property (2.40) as well as for property (2.41).

Proof. (I) We prove necessity and sufficiency of (2.27) separately.

1 (Sufficiency). It is easy to see that property (2.41) implies (2.40). Therefore, it is enough to prove that condition (2.27) implies (2.41). The last implication can be proved by almost the same arguments as used in part 1 of the proof of Theorem 2.2.4 in Section 2.2. Note that again the inequalities $I+\gamma T \geq 0$ and $S \geq 0$ are needed, which follow now from: $I+\gamma T = (I-P)^{-1} = I + P + P^2 + \ldots \geq 0$ and $S = (I-P)^{-1}R \geq 0$.

2 (Necessity). For proving the necessity it is enough to show that (2.40) implies (2.27). To prove this implication, we (only) assume (2.40) to hold in the situation where

$$\mathbb{V} = \mathbb{R}^m, \quad ||v|| = \max_k v^{[k]} \quad (\text{for } v \in \mathbb{V} \text{ with components } v^{[k]} \ (1 \leq k \leq m)).$$

We define functions $F_j : \mathbb{V} \to \mathbb{V}$ by

$$F_j(v) = \tau_0^{-1}(-y_j + z_j) \quad (\text{for } v = y_j), \quad F_j(v) = 0 \quad (\text{otherwise}),$$

where y_j, z_j are vectors in $\mathbb{V}$ - to be specified below - satisfying

$$\|z_j\| \leq \|y_j\| \quad (1 \leq j \leq m)). \tag{2.42}$$

Clearly the functions F_j defined in this fashion satisfy the basic assumption (2.4).

We consider the matrices $P = (p_{ij})$, $R = (r_{ij})$ (cf. (2.9)) and define the components of $x_j, z_j \in \mathbb{V}$ by $x_j^{[k]} = -1$ (if $r_{kj} < 0$), $x_j^{[k]} = 0$ (otherwise), and $z_j^{[k]} = -1$ (if $p_{kj} < 0$), $z_j^{[k]} = 0$ (otherwise). We define the vectors $y_i \in \mathbb{V}$ by $y_i = \sum_{j=1}^{l} r_{ij}x_j + \sum_{j=1}^{m} p_{ij}z_j$ $(1 \leq i \leq m)$. A short calculation shows that x_i, y_i satisfy the relations (2.2) with the functions F_j as defined above and $\Delta t = \gamma\tau_0$.

We denote by ρ_i the sum of the absolute values of the negative entries in the i-th row of R, and by π_i the sum of the absolute values of the negative entries in the i-th row of P. By the definition of y_i, we have $||y_i|| \geq y_i^{[i]} = \rho_i + \pi_i$ $(1 \leq$

$i \leq m$). Because $||z_i|| \leq 0$, the inequalities (2.42) are in force, so that the basic assumption (2.4) is valid.

Applying property (2.40) to the situation at hand, there follows

$$\rho_i + \pi_i \leq ||y_i|| \leq \left(\sum_j |s_{ij}|\right) \cdot \max_j ||x_j|| \leq 0 \qquad (1 \leq i \leq m),$$

which proves $P \geq 0,\ R \geq 0$. The remaining inequality, $\mathrm{spr}(P) < 1$, follows e.g. by applying Theorem 2.3.5, part (I).

(II) Let the classical condition (2.10) be fulfilled. Then (2.28) holds as well. So, by Lemma 2.3.2, part(III), condition (2.27) is fulfilled. From part (I) (of Theorem 2.3.9) we conclude that (2.40) and (2.41) hold.

Conversely, assume property (2.40) or (2.41). By Theorem 2.3.5, part (III), we arrive at (2.10). □

Since property (2.41) is a-priori stronger than (2.40), the essence of the above theorem is that conditions (2.27), (2.10) (under the appropriate assumptions on S) imply the strong statement (2.41), whereas already the weaker statement (2.40) implies (2.27) and (2.10)(under the same assumptions on S).

2.3.5 Various natural questions

In this section we ask and answer five natural questions about possible simplifications or extensions of Lemma 2.3.2 and Theorems 2.3.5, 2.3.9. For each of these questions we will provide counterexamples.

Question 2.3.10. Because all of the conditions (2.26), (2.27), (2.28) and (2.10) are more simple in appearance than condition (2.25), the question arises of whether the last condition can be replaced by any of the first four conditions in Lemma 2.3.2 (part (I)) or in Theorem 2.3.5 (part(I)).

To answer this question, consider the generic process (2.2) with $l = 2$, $m = 1$ and $s_{11} = -2$, $s_{12} = 1$, $t_{11} = 1$. Let $\gamma > 0$. It is easy to see that condition (2.25) is fulfilled. Hence, the properties (2.31) and (2.32) are present. But, we do *not* have $R \geq 0$, so that the conditions (2.26), (2.27), (2.28) and (2.10) are violated. Therefore, none of the last four conditions can replace condition (2.25) in Lemma 2.3.2 (part (I)) or in Theorem 2.3.5 (part(I)). □

Question 2.3.11. Because the conditions (2.27), (2.28) and (2.10) are more simple than (2.26), the question arises of whether condition (2.26) can be replaced by one of the first three conditions, in Lemma 2.3.2 (part (II)) or in Theorem 2.3.5 (part(II)).

The following counterexample proves that such a replacement is *not* possible. Consider process (2.2) with $l = m = 1$ and $s_{11} = 0$, $t_{11} = -1$. Let $\gamma = 1/4$. One easily sees that condition (2.26) is fulfilled, so that the properties (2.31) and (2.32) are present. But, we do not have $P \geq 0$, so that (2.27), (2.28) and (2.10) are violated. Therefore, none of the last three conditions can replace condition (2.26) in Lemma 2.3.2 (part (II)) or in Theorem 2.3.5 (part(II)). □

Question 2.3.12. Because the classical condition (2.10) is more simple than (2.27), the question arises of whether condition (2.27) can be replaced by (2.10) in Theorem 2.3.9 (part(I)).

The following counterexample proves that such replacement is *not* possible. Consider process (2.2) with $l = m = 1$ and $s_{11} = 0$, $t_{11} = -1$. Let $\gamma = 2$. One easily sees that condition (2.27) is violated, so that the properties (2.40) and (2.41) are not present. But (2.10) is fulfilled. Therefore, condition (2.10) cannot replace (2.27) in Theorem 2.3.9 (part(I)). □

Question 2.3.13. One may ask whether the condition $S \geq 0$ can be omitted in Theorem 2.3.5 (part(III)) or in Theorem 2.3.9 (part(II)).

To answer this question, consider process (2.2) with $l = m = 1$ and $s_{11} = -1$, $t_{11} = -1$. Let $\gamma = 2$. It is easy to see that condition (2.10) is fulfilled, but *not* (2.25) or (2.27). Hence, all of the special boundedness properties (2.31), (2.32), (2.40) and (2.41) are not present. Therefore, the condition $S \geq 0$ cannot be omitted in Theorem 2.3.5 (part(III)) or in Theorem 2.3.9 (part(II)). □

Question 2.3.14. Finally, we consider the question of whether the condition of S having no row equal to zero, can be omitted in Theorem 2.3.9 (part(II)). A negative answer to this question easily follows from the counterexample used above in resolving Question 2.3.12. □

2.4 Applications of the theory

2.4.1 Preliminaries

Below we shall illustrate the preceding theory by applying it to some well-known numerical methods. In these applications, we will restrict ourselves, for ease of representation, to autonomous problems, i.e. F in the initial value problem (2.1) is independent of t. Accordingly, in the generic process (2.2), we assume $F_j = F$, and the basic assumption (2.4) takes the form

$$\|v + \tau_0 F(v)\| \leq \|v\| \quad (\text{for all } v \in \mathbb{V}). \tag{2.43}$$

In Section 2.4.2 we shall deal with the two-step ($k = 2$) Adams-Bashforth LMM and in Section 2.4.3 with a class of k-step 2-stage methods. All of these methods generate vectors $u_n \in \mathbb{V}$ (for $n \geq k$) from starting vectors $u_0, \ldots, u_{k-1} \in \mathbb{V}$, where $u_n \approx u(n \cdot \Delta t)$ and k is fixed. We call a k-step method *bounded with factor* μ (for given stepsize Δt, vector space $\mathbb{V}$, functional $||\cdot||$ and function F) if

$$\|u_n\| \leq \mu \cdot \max_{0 \leq j \leq k-1} \|u_j\| \quad (k \leq n \leq k - 1 + N), \tag{2.44}$$

whenever $N \geq 1$ and $u_n \in \mathbb{V}$ ($k \leq n \leq k - 1 + N$) are generated from any $u_0, \ldots, u_{k-1} \in \mathbb{V}$ by N successive applications of the method. Boundedness with factor $\mu = 1$ will be referred to as *monotonicity* of the method.

We recall that boundedness and monotonicity with the so-called total variation seminorm (defined by $||x|| = ||x||_{TV} = \sum_i |\xi_{i+1} - \xi_i|$ for vectors x with components ξ_i) correspond to the important concepts *total variation bounded* and *total variation diminishing*, respectively, cf. e.g. Hundsdorfer & Verwer (2003), LeVeque (2002).

In the following we shall focus on the situation where the functional $||\cdot||$ is a seminorm. We shall consider stepsize-coefficients $\gamma > 0$ and factors μ such that

$$\begin{array}{l}\text{Condition } 0 < \Delta t \le \gamma \cdot \tau_0 \text{ implies boundedness with factor } \mu,\\ \text{whenever } \mathbb{V} \text{ is a vector space with seminorm } \|\cdot\|, \text{ and } F : \mathbb{V} \to \mathbb{V}\\ \text{satisfies the basic assumption (2.43).}\end{array} \tag{2.45}$$

In case γ, μ satisfy (2.45), we will say that γ is a *stepsize-coefficient for boundedness of the method with factor* μ; in case γ satisfies (2.45) with $\mu = 1$, we will call it a *stepsize-coefficient for monotonicity*. Below we shall look for stepsize-coefficients γ with property (2.45) by considering representations (2.2) of N consecutive steps of the method under consideration.

2.4.2 The two-step Adams-Bashforth method

The well-known 2-step Adams-Bashforth method reads

$$u_n = u_{n-1} + \Delta t\Big[\tfrac{3}{2}F(u_{n-1}) - \tfrac{1}{2}F(u_{n-2})\Big]; \tag{2.46}$$

it yields numerical approximations $u_n \approx u(n\Delta t)$ $(n = 2, 3, \ldots)$, starting from u_0 and $u_1 \approx u(\Delta t)$. In this section we shall look at the relevance of Theorems 2.2.2, 2.2.4, 2.3.5, 2.3.9 to this method, thereby representing N consecutive numerical steps in two different ways as a process of type (2.2).

In order to describe our first (and most natural) representation, we put $l = 2$, $m = N + 2$ and $x_1 = u_0$, $x_2 = u_1$, $y_i = u_{i-1}$ $(1 \le i \le m)$. Clearly, the equalities (2.46) hold for $2 \le n \le N + 1$ if and only if

$$\begin{array}{l} y_1 = x_1,\\ y_2 = x_2,\\ y_i = x_2 - \frac{1}{2}\Delta t\, F(y_1) + \Delta t \sum_{j=2}^{i-2} F(y_j) + \frac{3}{2}\Delta t\, F(y_{i-1}) \quad (3 \le i \le m)\,. \end{array} \tag{2.47}$$

These relations are equivalent to the relations in (2.2), with coefficients s_{ij}, t_{ij} defined by:

$$(s_{ij}) = S = \begin{pmatrix} 1 & 0\\ 0 & 1\\ \vdots & \vdots\\ 0 & 1 \end{pmatrix}, \qquad (t_{ij}) = T = \begin{pmatrix} 0 & & & & & \\ 0 & 0 & & & & \\ -\frac{1}{2} & \frac{3}{2} & 0 & & & \\ -\frac{1}{2} & 1 & \frac{3}{2} & 0 & & \\ \vdots & \vdots & \ddots & \ddots & \ddots & \\ -\frac{1}{2} & 1 & \cdots & 1 & \frac{3}{2} & 0 \end{pmatrix}.$$

With these definitions, the equalities (2.46) (for $2 \leq n \leq N+1$) thus hold if and only if (2.2) is fulfilled.

For the matrix T at hand, we see that $I + \gamma T$ is invertible for all $\gamma > 0$. Furthermore, because the preconsistency condition (2.5) is fulfilled, one might hope to be able to prove the monotonicity property (2.7) and its variant (2.12), for some $\gamma > 0$, by applying Theorems 2.2.2 and 2.2.4. If this were possible, such a γ would be a stepsize-coefficient for monotonicity in the sense specified in Section 2.4.1.

However, a short calculation shows that the matrix $P = (I+\gamma T)^{-1}(\gamma T)$ has a negative entry (for any $\gamma > 0$ and all $N \geq 1$), so that we cannot conclude, by applying the Theorems 2.2.2, 2.2.4, that there is $\gamma > 0$ for which the properties (2.7), (2.12) hold. Similarly, Theorems 2.3.5, 2.3.9 cannot be applied here so as to arrive at the boundedness property (2.45) with positive γ. The following negative statement can be proved, e.g. by applying the material in Spijker (1983, Theorem 3.3):

Proposition 2.4.1. *For the two-step Adams-Bashforth method (2.46) there exists no positive stepsize-coefficient γ for monotonicity.*

In spite of this statement, we will see below that a positive stepsize-coefficient for *boundedness* can be determined by applying Theorem 2.3.5 and representing the equalities (2.46) (for $2 \leq n \leq N+1$) in the generic form (2.2) with less obvious matrices S, T than used above.

We consider the representation in the generic form (2.2), with $l = 2$, $m = N$, $y_i = u_{i+1}$ $(1 \leq i \leq m)$ and input vectors

$$x_1 \,=\, u_1 \,+\, \tfrac{3}{2}\Delta t\, F(u_1) \,-\, \tfrac{1}{2}\Delta t\, F(u_0)\,, \qquad x_2 \,=\, -\tfrac{1}{2}\Delta t\, F(u_1)\,.$$

Clearly, the equalities (2.46) (for $2 \leq n \leq N+1$) amount to

$$\begin{aligned} &y_1 = x_1, \\ &y_i \,=\, x_1 \,+\, x_2 \,+\, \Delta t \textstyle\sum_{j=1}^{i-2} F(y_j) \,+\, \frac{3}{2}\Delta t\, F(y_{i-1}) \qquad (2 \leq i \leq m)\,. \end{aligned} \tag{2.48}$$

The first N steps of the Adams-Bashforth method can thus be represented by the generic process (2.2), with $l = 2$, $m = N$ and

$$(s_{ij}) = S = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{pmatrix}, \qquad (t_{ij}) = T = \begin{pmatrix} 0 & & & & \\ \frac{3}{2} & 0 & & & \\ 1 & \frac{3}{2} & 0 & & \\ \vdots & \ddots & \ddots & \ddots & \\ 1 & \cdots & 1 & \frac{3}{2} & 0 \end{pmatrix}. \tag{2.49}$$

Note that this matrix S violates the preconsistency condition (2.5), so that the monotonicity theory of Section 2.2.2 is not relevant here. But, the special boundedness theory of Section 2.3 still applies.

To be able to apply Theorem 2.3.5, we shall determine expressions for P and R corresponding to S, T just defined. A short calculation shows that

$$(I+\gamma T)^{-1} = \begin{pmatrix} q_0 & & & \\ q_1 & q_0 & & \\ \vdots & \ddots & \ddots & \\ q_{m-1} & \cdots & q_1 & q_0 \end{pmatrix},$$

where $q_0 = 1$, $q_1 = -\frac{3}{2}\gamma$ and $q_i = (1-\frac{3}{2}\gamma)q_{i-1} + \frac{1}{2}\gamma\, q_{i-2}$ for $i \geq 2$. It follows that

$$R = \begin{pmatrix} r_0 & 0 \\ r_1 & r_0 \\ \vdots & \vdots \\ r_{m-1} & r_{m-2} \end{pmatrix}, \quad P = -\begin{pmatrix} 0 & & & \\ q_1 & 0 & & \\ \vdots & \ddots & \ddots & \\ q_{m-1} & \cdots & q_1 & 0 \end{pmatrix},$$

where $r_i = q_0 + q_1 + \cdots + q_i$. Using the recurrence relation satisfied by q_i, one finds for $0 < \gamma \leq \frac{4}{9}$ and $i \geq 1$ that $q_i \leq 0$ and $\gamma \cdot r_i = -[(1-\gamma)q_i + \frac{\gamma}{2}q_{i-1}] \geq 0$. Hence, the classical condition (2.10) is fulfilled for any $\gamma \in (0, \frac{4}{9}]$. In the rest of this section we assume $\gamma = \frac{4}{9}$.

From proposition (III) of Theorem 2.3.5, we conclude that the generic process (2.2) (with coefficients given by (2.49)) has the special boundedness property (2.32). Using this property and the definition of x_1, x_2 in force, it follows that condition $0 < \Delta t \leq \gamma \cdot \tau_0$ implies:

$$\|u_n\| \leq \|u_1 + \tfrac{3}{2}\Delta t\, F(u_1) - \tfrac{1}{2}\Delta t\, F(u_0)\| + \| - \tfrac{1}{2}\Delta t\, F(u_1)\| \tag{2.50}$$

for $2 \leq n \leq N+1$, whenever u_n is generated by applying the Adams-Bashforth method under the basic assumption (2.43). Here $||\cdot||$ stands for an arbitrary seminorm on the vector space $\mathbb{V}$.

For $0 < \Delta t \leq \gamma \cdot \tau_0$ and any seminorm $||\cdot||$, we have

$$\|\Delta t F(v)\| = (\Delta t/\tau_0)\| - v + (v + \tau_0 F(v))\| \;\leq\; 2\gamma\|v\|,$$

which can be seen to imply

$$\|u_1 + \tfrac{3}{2}\Delta t\, F(u_1) - \tfrac{1}{2}\Delta t\, F(u_0)\| \;\leq\; \|u_1\| + \gamma\|u_0\|;$$

hence,

$$\|u_1 + \tfrac{3}{2}\Delta t F(u_1) - \tfrac{1}{2}\Delta t F(u_0)\| + \| - \tfrac{1}{2}\Delta t F(u_1)\| \;\leq\; (1+\gamma)\|u_1\| + \gamma\|u_0\|. \tag{2.51}$$

Combining this inequality with the above bound for $\|u_n\|$, we arrive at the following:

Proposition 2.4.2. *For the two-step Adams-Bashforth method (2.46), the step-size condition $0 < \Delta t \leq \frac{4}{9}\tau_0$ implies boundedness with factor $\mu = 17/9$, whenever $\mathbb{V}$ is a vector space with seminorm $\|\cdot\|$, and $F : \mathbb{V} \to \mathbb{V}$ satisfies the basic assumption (2.43).*

By applying Theorem 2.3.9, instead of Theorem 2.3.5, we find similarly as above that the bound (2.50) is valid under the basic assumption (2.43), when $||\cdot||$ is an arbitrary *sublinear functional* on the vector space $\mathbb{V}$. But, in the general situation of sublinear functionals, we *cannot* derive similarly as above that the inequality (2.51) is valid.

To give a simple illustration of the estimate (2.50), with a sublinear functional $||\cdot||$ which is no seminorm, we consider $\mathbb{V} = \mathbb{R}^M$ with functional $||\cdot|| = ||\cdot||_0$ (given by (2.34)). Applying Theorem 2.3.9 to the situation at hand, and defining $v \geq 0$ by nonnegativity of all components of $v \in \mathbb{V}$, yields:

Proposition 2.4.3. *Consider the two-step Adams-Bashforth method (2.46) in the situation where* $\mathbb{V} = \mathbb{R}^M$ *and* $||\cdot|| = ||\cdot||_0$ *(see (2.34)). Assume* $F : \mathbb{V} \to \mathbb{V}$ *satisfies the basic assumption (2.43). Then the stepsize condition* $0 < \Delta t \leq \frac{4}{9}\tau_0$ *implies that*

$$u_n \geq 0 \quad (2 \leq n \leq N+1),$$

whenever u_n *is obtained from* $u_0,\ u_1$ *with* $u_1 + \frac{3}{2}\Delta t\, F(u_1) \geq \frac{1}{2}\Delta t\, F(u_0)$ *and* $F(u_1) \leq 0$.

We note that there exists *no* positive stepsize-coefficient γ, such that the inequalities $u_n \geq 0$ are valid for $0 < \Delta t \leq \gamma \cdot \tau_0$, under the more natural assumption that

$$u_0 \geq 0,\ \ u_1 \geq 0 \qquad \text{and} \qquad v + \tau_0\, F(v) \geq 0 \quad (\text{for all } v \in \mathbb{R}^M \text{ with } v \geq 0).$$

This can be seen, for example, by considering $\mathbb{V} = \mathbb{R}$, $F(v) \equiv v$ and $u_0 = 1$, $u_1 = 0$.

2.4.3 Predictor-corrector methods and hybrid multistep methods

Notations

Using an explicit linear multistep method (LMM), with coefficients $\hat{a}_j$, $\hat{b}_j$, as a predictor for an implicit LMM, with coefficients a_j, b_j, results in a numerical process of type

$$v_n = \sum_{j=1}^{k} \hat{a}_j u_{n-j} + \Delta t \sum_{j=1}^{k} \hat{b}_j F(u_{n-j}), \tag{2.52a}$$

$$u_n = \sum_{j=1}^{k} a_j u_{n-j} + \Delta t \sum_{j=1}^{k} b_j F(u_{n-j}) + \Delta t\, b_0 F(v_n), \tag{2.52b}$$

where $k \geq 1$ is fixed and $n = k, k+1, \ldots$, cf. e.g. Butcher (2003), Hairer, Nørsett & Wanner (1993), Huang (2009). The starting values for this method are $u_0, u_1, \ldots, u_{k-1} \in \mathbb{V}$.

Throughout this section we assume $b_0 > 0$, $\sum_{j=1}^{k} \hat{a}_j = 1$, $\sum_{j=1}^{k} a_j = 1$, as well as zero-stability, i.e. all roots of the equation $\xi^k = \sum_{j=1}^{k} a_j \xi^{k-j}$ have a modulus $|\xi| \le 1$, and the roots with $|\xi| = 1$ are simple.

Methods of type (2.52) are called predictor-corrector methods if u_n and v_n, respectively, are final and tentative approximations to the solution at $t_n = n\Delta t$. If a predictor (2.52a) corresponds to a method with order of accuracy k, and a corrector (2.52b) to a method with order $k+1$, then the predictor-corrector method (2.52) has order $k+1$. The most popular schemes of this type are obtained by combining the explicit Adams-Bashforth and implicit Adams-Moulton methods, cf. the literature mentioned above.

The formulas (2.52) can also stand for so-called hybrid multistep methods, also known as modified linear multistep methods, where v_n approximates the solution at a point $\bar{t}_n = (n-\kappa)\Delta t$, with an extra parameter $\kappa \neq 0$; cf. the above literature.

We shall represent $N \ge 1$ steps of the general method (2.52) as a process of type (2.2), were $y = [y_i] \in \mathbb{V}^m$, $m = 2N$, with

$$y_i = u_{k-1+i}\,, \quad y_{N+i} = v_{k-1+i} \qquad \text{for } 1 \le i \le N\,. \tag{2.53}$$

For the input vector we take $x = [x_j] \in \mathbb{V}^l$, $l = 2k$, defined by

$$x_i = \sum_{j=i}^{k} a_j u_{k-1+i-j} + \Delta t \sum_{j=i}^{k} b_j F(u_{k-1+i-j}) \qquad (1 \le i \le k), \tag{2.54a}$$

$$x_{i+k} = \sum_{j=i}^{k} \hat{a}_j u_{k-1+i-j} + \Delta t \sum_{j=i}^{k} \hat{b}_j F(u_{k-1+i-j}) \qquad (1 \le i \le k). \tag{2.54b}$$

To write the relations (2.52), (2.53) specifying $y_1, y_2, \ldots, y_m$ in a compact way, we give the following definitions. For any $m \times r$ matrix $S = (s_{ij})$ we denote by the boldface symbol $\boldsymbol{S}$ the corresponding linear map from $\mathbb{V}^r$ to $\mathbb{V}^m$, that is, $y = \boldsymbol{S}x$ if $y_i = \sum_{j=1}^{r} s_{ij} x_j \in \mathbb{V}$ $(1 \le i \le m)$. Let I be the $N \times N$ identity matrix. Let $J_0 \in \mathbb{R}^{N\times k}$ be the matrix that consists of either the first N rows of the $k \times k$ identity matrix (when $1 \le N < k$), or the first k columns of I (when $N \ge k$). Furthermore, let $A_0 \in \mathbb{R}^{N\times N}$ be the lower triangular Toeplitz matrix with diagonal entries 0, entries a_j on the j-th lower diagonal $(1 \le j \le \min\{k, N-1\})$ and with the remaining entries 0 again. The matrices $B_0, \hat{A}_0, \hat{B}_0 \in \mathbb{R}^{N\times N}$ are defined likewise with coefficients $b_j, \hat{a}_j, \hat{b}_j$ $(1 \le j \le \min\{k, N-1\})$, respectively (the coefficient b_0 does not enter into the matrix B_0).

It is easy to see that the relations (2.52) (for $k \le n \le k-1+N$) are equivalent to

$$y = \boldsymbol{J}x + \boldsymbol{A}y + \Delta t \boldsymbol{B}\boldsymbol{F}(y)\,, \tag{2.55}$$

where $\boldsymbol{F}(y) = [F(y_j)] \in \mathbb{V}^m$, and $J \in \mathbb{R}^{m\times l}$, A, $B \in \mathbb{R}^{m\times m}$ are given by

$$J = \begin{pmatrix} J_0 & 0 \\ 0 & J_0 \end{pmatrix}, \quad A = \begin{pmatrix} A_0 & O \\ \hat{A}_0 & O \end{pmatrix}, \quad B = \begin{pmatrix} B_0 & b_0 I \\ \hat{B}_0 & O \end{pmatrix}. \tag{2.56}$$

The generic form (2.2) is thus obtained with coefficient matrices $(s_{ij}) = S = (I - A)^{-1}J$ and $(t_{ij}) = T = (I - A)^{-1}B$.

Monotonicity for predictor-corrector methods and hybrid multistep methods

Let us first take a brief look at standard monotonicity with respect to the starting vectors $u_0, \ldots, u_{k-1}$. For this, it is convenient to introduce $\check{a}_j = a_j - \gamma b_0 \hat{a}_j$ and $\check{b}_j = b_j - \gamma b_0 \hat{b}_j$ (for $j = 1, \ldots, k$). The relations (2.52) imply that

$$u_n = \sum_{j=1}^{k} \check{a}_j u_{n-j} + \Delta t \sum_{j=1}^{k} \check{b}_j F(u_{n-j}) + \gamma b_0 \big(v_n + \frac{\Delta t}{\gamma} F(v_n)\big) .$$

By combining this equality with (2.52a), we arrive at the following theorem; see also e.g. Gottlieb, Shu & Tadmor (2001), Huang (2009), Spijker (2007).

Theorem 2.4.4. *Consider method (2.52) with $n = k, k+1, \ldots, k-1+N$. Let $\|\cdot\|$ be a convex functional on the vector space $\mathbb{V}$, and assume $F : \mathbb{V} \to \mathbb{V}$ satisfies the basic assumption (2.43). Let $\gamma > 0$ be such that*

$$\hat{a}_j \geq \gamma \hat{b}_j \geq 0 , \quad \check{a}_j \geq \gamma \check{b}_j \geq 0 \qquad (j = 1, \ldots, k). \tag{2.57}$$

Then the stepsize restriction $0 < \Delta t \leq \gamma \cdot \tau_0$ implies that

$$\|u_n\| \leq \max_{0 \leq j \leq k-1} \|u_j\| \quad (k \leq n \leq k-1+N). \tag{2.58}$$

Note that, under a weak irreducibility assumption, condition (2.57) is not only sufficient but also necessary for the above bound (2.58), see Spijker (2007).

However, the methods (2.52) with coefficients satisfying condition (2.57) (with $\gamma > 0$) form a small class, excluding popular schemes, for instance obtained by combining explicit and implicit Adams-type methods as indicated above. Furthermore, in view of results for LMMs of Ruuth & Hundsdorfer (2005), one can expect that the stepsize requirement $\Delta t \leq \gamma \cdot \tau_0$ (with γ such that (2.57) holds) may be unnecessarily restrictive if γ is only required to be a stepsize-coefficient for boundedness (in the sense of Section 2.4.1).

Below we apply the theory of Section 2.3 in an analysis of the methods (2.52) which is also relevant in cases where condition (2.57) is violated.

Special bounds for predictor-corrector methods and hybrid multistep methods

Below we shall look for stepsize-coefficients for boundedness using the representation of (2.52) in the generic form (2.2) with the matrices S, T specified in Section 2.4.3.

For this T, the matrix $I + \gamma T$ is invertible for all $\gamma > 0$. To prove this, we consider the alternative ordering

$$y_{2i-1} = v_{k-1+i}, \quad y_{2i} = u_{k-1+i} \quad (1 \leq i \leq N), \tag{2.59}$$

which yields a representation of type (2.55) with strictly lower triangular matrices, say, $\underline{A}$, $\underline{B}$. The corresponding matrix $\underline{T} = (I - \underline{A})^{-1}\underline{B}$ is also strictly lower triangular. With our original ordering, viz. (2.53), we thus have a matrix $T = V\underline{T}V^{-1}$, where V is a permutation matrix, and therefore $I + \gamma T$ is invertible. To derive boundedness results it will be convenient to use the original ordering (2.53).

Substituting the expressions for S and T (given at the end of Section 2.4.3) into the definition (2.9) of P and R, we arrive at

$$R \;=\; KJ\,, \qquad P \;=\; \gamma KB\,, \qquad K \;=\; (I - A + \gamma B)^{-1}\,. \tag{2.60}$$

Because $P = V\underline{P}V^{-1}$, with $\underline{P} = \gamma\underline{T}(I + \gamma\underline{T})^{-1}$ and $\mathrm{spr}(\underline{P}) = 0$, we have also $\mathrm{spr}(P) = 0$.

Let $\check{K}_0 = (I - \check{A}_0 + \gamma\check{B}_0)^{-1}$, $\check{A}_0 = A_0 - \gamma b_0\hat{A}_0$, $\check{B}_0 = B_0 - \gamma b_0\hat{B}_0$. It can be seen that

$$K = \begin{pmatrix} I - A_0 + \gamma B_0 & \gamma b_0 I \\ -\hat{A}_0 + \gamma\hat{B}_0 & I \end{pmatrix}^{-1} = \begin{pmatrix} \check{K}_0 & -\gamma b_0\check{K}_0 \\ (\hat{A}_0 - \gamma\hat{B}_0)\check{K}_0 & (I - A_0 + \gamma B_0)\check{K}_0 \end{pmatrix}.$$

This gives

$$R \;=\; \begin{pmatrix} \check{K}_0 J_0 & -\gamma b_0\,\check{K}_0 J_0 \\ (\hat{A}_0 - \gamma\hat{B}_0)\,\check{K}_0 J_0 & (I - A_0 + \gamma B_0)\,\check{K}_0 J_0 \end{pmatrix}. \tag{2.61}$$

Using the fact that lower triangular Toeplitz matrices commute, it is found that

$$P \;=\; \gamma \begin{pmatrix} (B_0 - \gamma b_0\hat{B}_0)\,\check{K}_0 & b_0\,\check{K}_0 \\ \big((I - A_0)\hat{B}_0 + \hat{A}_0 B_0\big)\,\check{K}_0 & b_0(\hat{A}_0 - \gamma\hat{B}_0)\,\check{K}_0 \end{pmatrix}. \tag{2.62}$$

We have

$$S \;=\; \begin{pmatrix} (I - A_0)^{-1}J_0 & O \\ \hat{A}_0(I - A_0)^{-1}J_0 & J_0 \end{pmatrix}.$$

By considering the upper-right blocks of R, P, S and PS, $|P|S$ it can be seen that none of conditions (2.25)–(2.28) is fulfilled (for any $\gamma > 0$ and all $N \geq 1$). Hence, Theorem 2.3.5 cannot be applied here directly so as to arrive at property (2.45) with positive γ. However, we shall see below that a positive *stepsize-coefficient for boundedness* can be found by modifying the matrix S and applying Theorem 2.3.9.

Let

$$\tilde{x}_i = x_i - \gamma b_0 x_{i+k}\,, \quad \tilde{x}_{i+k} = x_{i+k} \qquad \text{for } \; i = 1, \ldots, k\,. \tag{2.63}$$

Then $x = V\,\tilde{x}$ with $V = \begin{pmatrix} I & \gamma b_0 I \\ O & I \end{pmatrix}$. Below we shall deal with process (2.55) written in the equivalent form

$$y = \tilde{\boldsymbol{S}}\tilde{x} + \Delta t\boldsymbol{T}\boldsymbol{F}(y)\,, \tag{2.64}$$

where $\tilde{S} = (\tilde{s}_{ij}) = (I - A)^{-1}JV = SV$. Defining $\tilde{R} = (I + \gamma T)^{-1}\tilde{S}$ (cf. (2.9)) we get in view of (2.60)

$$\tilde{R} \;=\; K\,J\,V \;=\; \begin{pmatrix} \check{K}_0 J_0 & O \\ (\hat{A}_0 - \gamma \hat{B}_0)\,\check{K}_0 J_0 & J_0 \end{pmatrix}. \tag{2.65}$$

We now have $\tilde{R} \geq 0$ (for all $N \geq 1$) whenever

$$P \geq 0 \qquad \text{(for all } N \geq 1\text{)}. \tag{2.66}$$

This leads directly to the following result.

Lemma 2.4.5. *Consider N consecutive steps of method (2.52) written in the form (2.64). Let $||\cdot||$ be a sublinear functional on the vector space $\mathbb{V}$. Assume $F : \mathbb{V} \to \mathbb{V}$ satisfies the basic assumption (2.43) and $\gamma > 0$ is such that (2.66) holds. Then the stepsize restriction $0 < \Delta t \leq \gamma \cdot \tau_0$ implies that the output vectors y_i defined by (2.53) satisfy*

$$\|y_i\| \leq \tilde{\mu}_i \cdot \max_{1 \leq j \leq l} \|\tilde{x}_j\| \qquad (1 \leq i \leq 2N)\,,$$

with $\tilde{\mu}_i = \sum_j |\tilde{s}_{ij}|$.

Proof. To prove this lemma, we apply part (I) of Theorem 2.3.9 with S replaced by $\tilde{S}$. □

Consider $\tilde{\mu} = \max_i \tilde{\mu}_i = \|\tilde{S}\|_\infty$. Using the definition (2.56) and the expression (2.65), there follows after a little calculation that

$$\tilde{S} \;=\; \begin{pmatrix} I & \gamma b_0 I \\ \hat{A}_0 & I - \check{A}_0 \end{pmatrix} \begin{pmatrix} S_0 & 0 \\ 0 & S_0 \end{pmatrix},$$

with $S_0 = (I - A_0)^{-1}J_0$. We find that $\tilde{\mu} \leq \|(I - A_0)^{-1}J_0\|_\infty \cdot \max\big\{1 + \gamma b_0, 1 + \sum_{j=1}^{k} \big(|\hat{a}_j| + |\check{a}_j|\big)\big\}$. Due to the assumption of zero-stability we have $\sup_{N \geq 1} \|S_0\|_\infty < \infty$, so that $\tilde{\mu}$ can be bounded, uniformly with respect to N.

Consider $\gamma > 0$ such that (2.66) holds and let $0 < \Delta t \leq \gamma \cdot \tau_0$. Then from Lemma 2.4.5 and (2.54), (2.63), it follows that

$$\|u_n\| \leq \tilde{\mu} \cdot \max\Big\{ \sum_{j=1}^{k} \big(|\check{a}_j - \gamma \check{b}_j| + |\gamma \check{b}_j|\big), \sum_{j=1}^{k} \big(|\hat{a}_j - \gamma \hat{b}_j| + |\gamma \hat{b}_j|\big) \Big\} \cdot \max_{0 \leq j \leq k-1} \|u_j\|$$

for $k \leq n \leq k - 1 + N$, whenever u_n is generated from $u_0, \ldots, u_{k-1} \in \mathbb{V}$ by applying method (2.52) under the basic assumption (2.43), where $||\cdot||$ is a seminorm on the vector space $\mathbb{V}$. Thus we arrive at the following theorem.

Theorem 2.4.6. *Assume $\gamma > 0$ is such that $P \geq 0$ (for all $N \geq 1$). Then γ is a stepsize-coefficient for boundedness of the method (2.52) (in the sense of Section 2.4.1).*

Results for third order explicit two-step methods of the form (2.52)

In this section we study method (2.52) with $k = 2$, $u_n \approx u(n\Delta t)$, $v_n \approx u((n-\kappa)\Delta t)$. Requiring order $p = 3$ leaves 3 free parameters a_1, $\hat{a}_1$, κ and the remaining coefficients can be computed by the formulas: $a_2 = 1 - a_1$, $b_0 = (4 + a_1)/(6(1-\kappa)(2-\kappa))$, $b_1 = (8 - 12\kappa - (4 - 3\kappa)a_1)/(6(1-\kappa))$, $b_2 = (4 - (5 - 3\kappa)a_1)/(6(2-\kappa))$, $\hat{a}_2 = 1 - \hat{a}_1$, $\hat{b}_1 = 2 - \frac{\hat{a}_1}{2} - 2\kappa + \frac{\kappa^2}{2}$, $\hat{b}_2 = -\frac{\hat{a}_1}{2} + \kappa - \frac{\kappa^2}{2}$. The method is zero-stable if and only if $a_1 \in [0, 2)$.

For these methods we will compute the maximal values of γ such that $P \geq 0$ for $N = 1, \ldots, 1000$; it was verified that with larger N the results did not differ anymore noticeably.

First we study the methods with $\kappa = 0$, corresponding to the classical two-step predictor-corrector methods. The result is shown in the left panel of Figure 2.1. We note that there are no methods in this class for which the monotonicity condition (2.57) holds with $\gamma > 0$. The displayed values of γ for boundedness with these predictor-corrector methods are rather low; the maximal value is approximately 0.36, corresponding to $a_1 \approx 0.765$, $\hat{a}_1 \approx 1.673$.

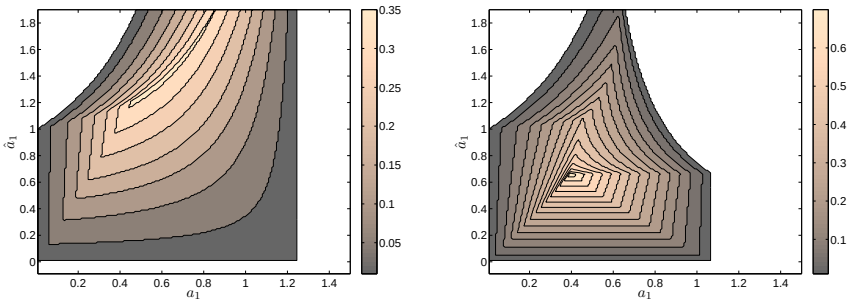

FIGURE 2.1: Maximal values $\gamma > 0$ such that $P \geq 0$ for the methods (2.52) with $k = 2$ of order $p = 3$, with parameters $a_1 \in [0, 1.5]$ horizontally and $\hat{a}_1 \in [-0.1, 1.95]$ vertically. Left panel: standard predictor-corrector methods, $\kappa = 0$. Right panel: hybrid methods with $\kappa = 1 - \frac{1}{3}\sqrt{3}$. Contour levels at $j/20$, $j = 0, 1, \ldots$; for the 'white' areas, there is no positive γ.

A numerical search revealed that larger values of γ can be found by allowing $\kappa \neq 0$. The right panel of Figure 2.1 shows the values of γ with $\kappa = 1 - \frac{1}{3}\sqrt{3}$. The largest $\gamma \approx 0.73$ is found with $a_1 \approx 0.392$, $\hat{a}_1 \approx 0.667$ and this γ is optimal within the whole class (2.52) with $k = 2$, $p = 3$.

Rather surprisingly, this method coincides with the method found in Spijker (2007, Section 3.2.3) which is optimal with respect to the monotonicity condition (2.57). The latter method corresponds to $a_1 = 6\sqrt{3} - 10$, $\hat{a}_1 = \frac{2}{3}$. These parameters coincide (up to four decimal digits) with the values for a_1, $\hat{a}_1$ obtained numerically by our search using condition (2.66), corresponding to the right panel in Figure 2.1. In fact, if $\hat{a}_1 \leq \frac{2}{3}$ the monotonicity condition (2.57) seems to give the same γ as the boundedness condition (2.66). If $\hat{a}_1 > \frac{2}{3}$

then the method has some negative coefficient, so then there is no positive γ for monotonicity with arbitrary starting values. But, as shown by Figure 2.1, for such $\hat{a}_1$ we can still have positive stepsize-coefficients γ for boundedness.

Chapter 3

Stepsize Restrictions for Boundedness and Monotonicity of Multistep Methods

In this chapter nonlinear monotonicity and boundedness properties are analyzed for linear multistep methods. We focus on methods which satisfy a weaker boundedness condition than strict monotonicity for arbitrary starting values. In this way, many linear multistep methods of practical interest are included in the theory. Moreover, it will be shown that for such methods monotonicity can still be valid with suitable Runge-Kutta starting procedures. Restrictions on the stepsizes are derived that are not only sufficient but also necessary for these boundedness and monotonicity properties.

3.1 Introduction

3.1.1 Monotonicity assumptions

In this chapter we consider initial value problems for systems of ordinary differential equations (ODEs) on a vector space $\mathbb{V}$, written as

$$u'(t) = F(u(t)) \quad (t \geq 0)\,, \qquad u(0) = u_0\,, \tag{3.1}$$

with $F : \mathbb{V} \to \mathbb{V}$ and $u_0 \in \mathbb{V}$ given. Let $\|\cdot\|$ be a norm or seminorm on $\mathbb{V}$. In the following it is assumed that there is a constant $\tau_0 > 0$ such that

$$\|v + \tau_0 F(v)\| \;\leq\; \|v\| \qquad \text{for all } v \in \mathbb{V}. \tag{3.2}$$

Assumption (3.2) implies $\|v + \Delta t\, F(v)\| \leq \|v\|$ for all $\Delta t \in (0, \tau_0]$. Consequently, when applying the forward Euler method $u_n = u_{n-1} + \Delta t\, F(u_{n-1})$, $n \geq 1$, with stepsize $\Delta t > 0$ to compute approximations $u_n \approx u(t_n)$ at $t_n = n\Delta t$, we have

$$\|u_n\| \;\leq\; \|u_0\| \tag{3.3}$$

for all $n \geq 1$ under the stepsize restriction $\Delta t \leq \tau_0$. For general one-step methods, property (3.3) under a stepsize restriction $\Delta t \leq c\,\tau_0$, with some constant $c > 0$, is often referred to as *monotonicity* or *strong stability preservation* (SSP).

Useful and well-known examples for (3.2) involve $v = (v_1, \ldots, v_M)^T \in \mathbb{V} = \mathbb{R}^M$ with the maximum norm $\|v\|_\infty = \max_{1\leq j\leq M} |v_j|$ or the total variation seminorm $\|v\|_{\mathrm{TV}} = \sum_{j=1}^M |v_{j-1} - v_j|$ (with $v_0 = v_M$), arising from one-dimensional partial differential equations (PDEs), see for instance Gottlieb, Ketcheson & Shu (2009), Hundsdorfer & Verwer (2003), LeVeque (2002).

Some of the results in this chapter will be formulated with sublinear functionals instead of seminorms.[1] This makes it possible to take, for example, maximum principles into consideration as in Spijker (2007), by requiring that (3.2) holds for the functionals $\|v\|_+ = \max_j v_j$ and $\|v\|_- = -\min_j v_j$. Another example, from Chapter 2, is $\|v\|_0 = -\min\{0, v_1, \ldots, v_M\}$, by which preservation of nonnegativity can be included in the theory. We note that this last sublinear functional is nonnegative, that is, $\|v\| \geq 0$ for all $v \in \mathbb{R}^M$.

3.1.2 Monotonicity and boundedness for linear multistep methods

To solve (3.1) numerically we consider multistep methods. We will be primarily concerned with linear k-step methods, where the approximations $u_n \approx u(t_n)$ at the points $t_n = n\Delta t$ are computed by

$$u_n = \sum_{j=1}^{k} a_j u_{n-j} + \Delta t \sum_{j=0}^{k} b_j F(u_{n-j}) \tag{3.4}$$

for $n \geq k$. The starting values for this multistep recursion, $u_0, u_1, \ldots, u_{k-1} \in \mathbb{V}$, are supposed to be given, or computed by a Runge-Kutta method.

It will be assumed throughout this chapter that

$$\sum_{j=1}^{k} a_j \;=\; 1\,, \qquad \sum_{j=1}^{k} j\,a_j \;=\; \sum_{j=0}^{k} b_j\,, \qquad b_0 \geq 0\,. \tag{3.5}$$

The two equalities in (3.5) are the conditions for consistency of order one. The assumption $b_0 \geq 0$ will be convenient; it holds for all well-known implicit methods, and, of course, also for any explicit method.

Suppose that all $a_j, b_j \geq 0$, and for such a method let

$$c \;=\; \min_{1\leq j\leq k} \frac{a_j}{b_j}\,, \tag{3.6}$$

with convention $a/0 = +\infty$ if $a \geq 0$. From (3.2) it can then be shown that

$$\|u_n\| \;\leq\; \max_{0\leq j<k} \|u_j\| \tag{3.7}$$

[1] Recall that $\varphi : \mathbb{V} \to \mathbb{R}$ is called a sublinear functional if $\varphi(v+w) \leq \varphi(v) + \varphi(w)$ and $\varphi(cv) = c\varphi(v)$ for all real $c \geq 0$ and $v, w \in \mathbb{V}$. It is a seminorm if we have in addition $\varphi(-v) = \varphi(v) \geq 0$ for all $v \in \mathbb{V}$. If it also holds that $\varphi(v) = 0$ only if $v = 0$, then φ is a norm.

for $n \geq k$, under the stepsize restriction $\Delta t \leq c\,\tau_0$; see e.g. Gottlieb, Ketcheson & Shu (2009), Spijker (2007). This property can be viewed as an extension of (3.3) for multistep methods with arbitrary starting values.

Results of this type for nonlinear problems were derived originally in Shu (1988) with the total variation seminorm, and (3.7) with this seminorm is known as the TVD (total variation diminishing) property. More recently, with arbitrary seminorms or more general convex functionals, the term SSP (strong stability preserving) – introduced in Gottlieb, Shu & Tadmor (2001) – has become popular. Related work for nonlinear problems was done in Lenferink (1991), Sand (1986), Vanselov (1983) for *contractivity*, where one considers $\|\tilde{u}_n - u_n\|$ with differences of two numerical solutions instead of $\|u_n\|$ as in (3.7). Finally we mention that related results on nonnegativity preservation and contractivity or monotonicity for *linear* problems can be found e.g. in Bolley & Crouzeix (1978), Ketcheson (2009), Lenferin (1989) and Spijker (1983), again primarily for methods with all $a_j, b_j \geq 0$ and with $\Delta t \leq c\,\tau_0$.

In order to conclude (3.7) from (3.2) for arbitrary (semi-)norms or sublinear functionals, the condition that all $a_j, b_j \geq 0$ and $\Delta t \leq c\,\tau_0$ is *necessary*. In fact, this condition is already needed if we only consider maximum norms instead of arbitrary (semi-)norms; see Spijker (2007).

The methods with nonnegative coefficients form only a small class, excluding the well-known methods of the Adams or BDF-type, and the stepsize requirement $\Delta t \leq c\,\tau_0$ (within this class) can be very restrictive. For instance, as shown in Lenferink (1989), for an explicit k-step method ($k > 1$) of order p we have $c \leq (k-p)/(k-1)$. Most explicit methods used in practice have $p = k$, and for such methods we cannot have $c > 0$. It is therefore of interest to study properties that are more relaxed than (3.7).

Instead of (3.7), we will consider

$$\|u_n\| \;\leq\; \mu \cdot \max_{0 \leq j < k} \|u_j\| \tag{3.8}$$

for $n \geq k$, under the stepsize restriction $\Delta t \leq \gamma\tau_0$, where the stepsize coefficient $\gamma > 0$ and the factor $\mu \geq 1$ are determined by the multistep method. With the total variation seminorm this is known as the TVB (total variation boundedness) property.

Sufficient conditions were derived in Hundsdorfer & Ruuth (2006), Hundsdorfer, Ruuth & Spiteri (2003) for (3.8) to be valid with arbitrary seminorms under assumption (3.2) and $\Delta t \leq \gamma\tau_0$. The sufficient conditions of those papers are not very transparent and not easy to verify for given methods. In the present chapter we will use the general framework of Chapter 1 to obtain more simple conditions for boundedness, and these conditions are not only sufficient but also necessary.

In practice, the starting values are not arbitrary, of course. From a given u_0, the vectors $u_1, \ldots, u_{k-1}$ can be computed by a Runge-Kutta method. For such combinations of linear multistep methods and Runge-Kutta starting procedures we will study the monotonicity property (3.3) under a stepsize restriction

$\Delta t \leq \gamma\tau_0$. By writing the total scheme in a special Runge-Kutta form we will obtain sharp stepsize conditions for this type of monotonicity. This gives a generalization of earlier, partial results in this direction obtained in Hundsdorfer, Ruuth & Spiteri (2003) for some explicit two-step methods.

3.1.3 Outline of the chapter

To illustrate the relevance of the results we first present in Section 3.2 a numerical example with two simple two-step methods applied to a semi-discrete advection equation. The coefficients a_j, b_j of the two methods are close to each other, but the behaviour of the methods with respect to boundedness and monotonicity turns out to be very different.

In Section 3.3 some notations are introduced, together with a formulation of the linear multistep method (3.4) that is suited for application of the general boundedness results of Chapter 1.

The main results are presented in Section 3.4. Using the framework of Chapter 1, we will obtain necessary and sufficient conditions for boundedness. These conditions are relatively transparent and easy to verify numerically for given classes of methods. We will also give conditions that ensure monotonicity –as in (3.3)– for combinations of linear multistep methods and Runge-Kutta starting procedures.

Section 3.5 contains some technical derivations and the proofs of the main theorems on boundedness. We will see that, for all methods of practical interest, the stepsize coefficients γ for boundedness are completely determined by particular properties of the method when applied to the test equation $u'(t) = \lambda\, u(t)$ with $\Delta t\, \lambda = -\gamma$.

For some classes of methods, with two free parameters, we will present and discuss in Section 3.6 the maximal stepsize coefficients γ for either boundedness or monotonicity with some specific starting procedures.

Finally, Section 3.7 contains a few concluding remarks, putting our results in a somewhat wider perspective and addressing briefly the possibility of analoguous results for variants of the linear multistep methods (3.4).

Along with the usual typographical symbol $\Box$ to indicate the end of a proof, we will use in this chapter also the symbol $\diamond$ to mark the end of examples or remarks.

3.2 A numerical illustration

To illustrate the relevance of our monotonicity and boundedness concepts, we consider two-step methods of the form

$$u_n = \tfrac{3}{2}u_{n-1} - \tfrac{1}{2}u_{n-2} + \Delta t\, \beta F(u_{n-1}) + \Delta t\, (\tfrac{1}{2} - \beta)F(u_{n-2})\,. \qquad (3.9)$$

We take two methods within this class: $\beta = 0.95$ and $\beta = 1.05$. Both methods have order one. Moreover the error constants are very similar, and so are

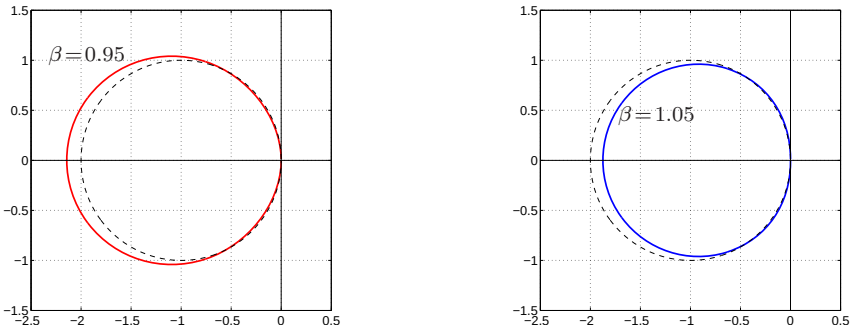

FIGURE 3.1: Stability regions of the two-step methods (3.9) with $\beta = 0.95$ (left), $\beta = 1.05$ (right). For comparison, the circle $\{\zeta \in \mathbb{C} : |\zeta + 1| = 1\}$ is displayed by the dashed curve.

the linear stability regions, as shown in Figure 3.1. However, as we will see shortly, these two methods have a very different monotonicity and boundedness behaviour.

Note that for both methods we have $a_2 < 0$ and $b_2 < 0$. Therefore the monotonicity property (3.7) with arbitrary starting vectors and seminorms does not apply. Instead of an arbitrary u_1 we consider the forward Euler starting procedure $u_1 = u_0 + \Delta t F(u_0)$. The combination of the two-step methods with forward Euler may give a scheme for which the monotonicity property (3.3) is valid.

Monotonicity and boundedness properties are of importance for problems with non-smooth solutions. Such ODE problems often arise from conservation laws with shocks or advection dominated PDEs with steep gradients, after suitable spatial discretization.

A simple illustration is provided by the one-dimensional linear advection equation

$$\frac{\partial}{\partial t}u(x,t) + \frac{\partial}{\partial x}u(x,t) = 0 \qquad \text{for } \; t > 0 \;\text{ and }\; 0 < x < 1$$

with periodic boundary conditions. The initial profile is chosen as a block-function: $u(x,0) = 1$ if $0.4 \le x \le 0.6$, and $u(x,0) = 0$ otherwise. The spatial discretization is taken on a uniform grid with mesh width $\Delta x = 1/200$, using a standard flux-limited scheme −the so-called Koren limiter− giving a semi-discrete system of ODEs for which the monotonicity assumption (3.2) is satisfied for $\tau_0 = \frac{1}{2}\Delta x$ in the maximum norm and the total variation seminorm; see for instance Hundsdorfer & Verwer (2003, Sect. III.1).

Subsequently, the resulting nonlinear semi-discrete system is integrated in time with the above two methods and Courant number $\Delta t/\Delta x$ equal to 1/3. The first approximation u_1 is computed by the forward Euler method.

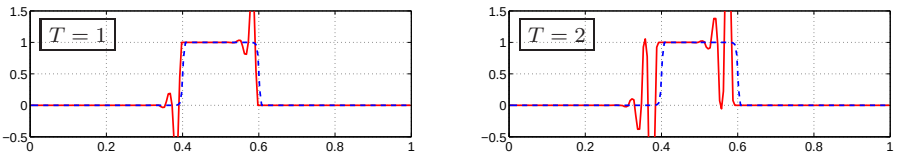

FIGURE 3.2: Numerical solutions at $T = 1$ and $T = 2$ for the two-step methods (3.9) with $\beta = 1.05$ (dashed), $\beta = 0.95$ (solid lines).

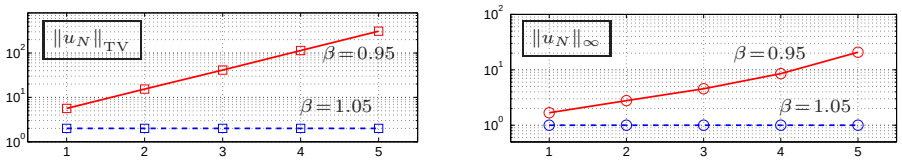

FIGURE 3.3: Values of $\|u_N\|_{\mathrm{TV}}$ (left) and $\|u_N\|_\infty$ (right) for $T = 1, 2, \ldots, 5$ and the two-step methods (3.9) with $\beta = 1.05$ (dashed), $\beta = 0.95$ (solid lines).

The numerical solutions for the two schemes are shown in Figure 3.2, with spatial component x horizontally, for the output times $t = T$ with $T = 1, 2$. The behaviour of the two schemes is seen to be very different. Whereas we get a nice monotonic behaviour for $\beta = 1.05$, the scheme with $\beta = 0.95$ produces large oscillations.

The oscillations with $\beta = 0.95$ become more and more pronounced for increasing time. The evolution of the total variation and maximum norm of u_N ($N = T/\Delta t$) is shown in Figure 3.3, revealing an exponential growth. On the other hand, for the scheme with $\beta = 1.05$ these values are constant: $\|u_N\|_{\mathrm{TV}} = 2$, $\|u_N\|_\infty = 1$. A similar behaviour can also be observed if T is held fixed, say $T = 1$, and the $\Delta t, \Delta x$ are decreased while keeping the Courant number $\Delta t/\Delta x$ fixed. Apparently the boundedness property (3.8) is not satisfied here for the scheme with $\beta = 0.95$.

With the results of this chapter the different behaviour of these two closely related schemes can be explained. As we will see in Section 3.6.1, to satisfy the boundedness property (3.8) or the monotonicity property (3.3) with forward Euler starting procedure, the method with $\beta = 1.05$ allows much larger stepsizes than the method with $\beta = 0.95$.

3.3 Notations and input-output formulations

3.3.1 Some notations

For any given $m \geq 1$ we will denote by $e_1, e_2, \ldots, e_m$ the unit basis vectors in $\mathbb{R}^m$, that is, the j-th element of e_i equals one if $i = j$ and zero otherwise. Furthermore, $e = e_1 + e_2 + \cdots + e_m$ is the vector in $\mathbb{R}^m$ with all components equal to one. The $m \times m$ identity matrix is denoted by I. If it is necessary to specify the dimension we will use the notations $e_j^{[m]}$, $e^{[m]}$ and $I^{[m]}$ for these unit vectors and the identity matrix I.

Let $E = [e_2, \ldots, e_m, 0]$ be the $m \times m$ backward shift matrix,

$$E = \begin{pmatrix} 0 & & & \\ 1 & 0 & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{pmatrix} \in \mathbb{R}^{m \times m}, \tag{3.10}$$

and define

$$A \;=\; \sum_{j=1}^{k} a_j E^j\,, \qquad B \;=\; \sum_{j=0}^{k} b_j E^j\,, \tag{3.11}$$

where $E^0 = I$. These $A, B \in \mathbb{R}^{m \times m}$ are lower triangular Toeplitz matrices containing the coefficients of the method (3.4). For $m \geq k$ we also introduce $J = [e_1, \ldots, e_k] \in \mathbb{R}^{m \times k}$, containing the first k columns of the identity matrix I. To make the notations fitting for any $m \geq 1$, we define $J = [e_1, \ldots, e_m, O]$ for $1 \leq m < k$, with O being the $m \times (k - m)$ zero matrix.

For any $m \times l$ matrix $K = (\kappa_{ij})$ we denote by the boldface symbol $\boldsymbol{K}$ the associated linear mapping from $\mathbb{V}^l$ to $\mathbb{V}^m$, that is, $y = \boldsymbol{K}x$ for $y = [y_i] \in \mathbb{V}^m$, $x = [x_i] \in \mathbb{V}^l$ if $y_i = \sum_{j=1}^{l} \kappa_{ij} x_j \in \mathbb{V}$ $(1 \leq i \leq m)$. (In case $\mathbb{V} = \mathbb{R}^M$ with $M \geq 1$, then $\boldsymbol{K}$ is the Kronecker product of K with $I^{[M]}$.) Furthermore, the $m \times l$ matrix with entries $|\kappa_{ij}|$ will be denoted by $|K|$, and we define $\|K\|_\infty = \max_i \sum_j |\kappa_{ij}|$.

Inequalities for vectors or matrices are to be understood component-wise. In particular, we will use the notation $K \geq 0$ when all entries κ_{ij} of this matrix are nonnegative.

3.3.2 Formulations with input vectors

In order to apply the theory obtained in Chapter 1, we will formulate the multistep scheme (3.4) in terms of input and output vectors. The *output vectors* of the scheme are $y_n = u_{k-1+n}$, $n \geq 1$. The starting values $u_0, u_1, \ldots, u_{k-1}$ will enter the scheme in the first k steps in the combinations

$$x_l \;=\; \sum_{j=l}^{k} a_j u_{k-1+l-j} + \Delta t \sum_{j=l}^{k} b_j F(u_{k-1+l-j}) \qquad (1 \leq l \leq k)\,. \tag{3.12}$$

The multistep scheme (3.4) then can be written as

$$y_n \;=\; x_n + \sum_{j=1}^{n-1} a_j y_{n-j} + \Delta t \sum_{j=0}^{n-1} b_j F(y_{n-j}) \qquad (1 \le n \le k)\,, \tag{3.13a}$$

$$y_n \;=\; \sum_{j=1}^{k} a_j y_{n-j} + \Delta t \sum_{j=0}^{k} b_j F(y_{n-j}) \qquad (n > k)\,, \tag{3.13b}$$

where the starting values are contained within the source terms in the first k steps. We will refer to the vectors $x_1, \ldots, x_k \in \mathbb{V}$ as the *input vectors* for the scheme.

To obtain a convenient notation, we consider m steps of the multistep scheme, $m \ge 1$, leading to (3.13) with $n = 1, 2, \ldots, m$. Let $y = [y_i] \in \mathbb{V}^m$, $x = [x_i] \in \mathbb{V}^k$, and define $\boldsymbol{F}(y) = [F(y_i)] \in \mathbb{V}^m$. We can now write the resulting scheme in a compact way as

$$y \;=\; \boldsymbol{J}x + \boldsymbol{A}y + \Delta t\, \boldsymbol{B}\boldsymbol{F}(y)\,. \tag{3.14}$$

To study boundedness, the number of steps m is allowed to be arbitrarily large. Consider, for given vector space $\mathbb{V}$ and seminorm $\|\cdot\|$, the boundedness property

$$\max_{1\le n\le m} \|y_n\| \le \mu \cdot \max_{1\le j\le k} \|x_j\| \quad \begin{array}{l}\text{whenever (3.2) is valid, } \Delta t \le \gamma\tau_0,\\ \text{and } x, y \text{ satisfy (3.14), } m \ge 1,\end{array} \tag{3.15}$$

with a stepsize coefficient $\gamma > 0$ and boundedness factor $\mu \ge 1$. Note that this property involves all $F : \mathbb{V} \to \mathbb{V}$ for which the monotonicity assumption (3.2) is satisfied, as well as all x, y satisfying (3.14) and $m \ge 1$. Therefore γ and μ will *not* depend on a particular problem (3.1) under consideration.

A convenient form to derive results on boundedness is obtained by multiplying relation (3.14) by $(\boldsymbol{I} - \boldsymbol{A} + \gamma\boldsymbol{B})^{-1}$ with $\gamma > 0$. This yields

$$y \;=\; \boldsymbol{R}x + \boldsymbol{P}\Big(y + \frac{\Delta t}{\gamma}\boldsymbol{F}(y)\Big)\,, \tag{3.16}$$

where $R = (r_{ij}) \in \mathbb{R}^{m\times k}$ and $P = (p_{ij}) \in \mathbb{R}^{m\times m}$ are given by

$$R \;=\; (I - A + \gamma B)^{-1} J\,, \qquad P \;=\; (I - A + \gamma B)^{-1}\gamma B\,. \tag{3.17}$$

Note that $I - A + \gamma B$ is invertible for any $\gamma > 0$, because $b_0 \ge 0$, and therefore (3.16) is still equivalent to (3.14). The matrix P is again a lower triangular Toeplitz matrix, and it has the entry $\pi_0 = \gamma b_0/(1+\gamma b_0) \in [0, 1)$ on the diagonal. The spectral radius $\mathrm{spr}(|P|)$ of the matrix $|P| = (|p_{ij}|)$ also equals π_0, and because this is less than one it follows that $(I - |P|)^{-1} = \sum_{j=0}^{\infty} |P|^j$. We thus have

$$\mathrm{spr}(|P|) < 1\,, \qquad (I - |P|)^{-1} \;\ge\; 0\,. \tag{3.18}$$

3.3.3 Application of a general result on boundedness

To obtain boundedness results for the multistep methods we will use a general result from Chapter 1. The connection with the notation used in that chapter is established by writing (3.14) in the form

$$y \;=\; \boldsymbol{S}x + \Delta t\, \boldsymbol{T}\boldsymbol{F}(y) \tag{3.19}$$

with $S \in \mathbb{R}^{m\times k}$ and $T \in \mathbb{R}^{m\times m}$ defined by

$$S = (I-A)^{-1}J\,, \qquad T = (I-A)^{-1}B\,. \tag{3.20}$$

We note that the matrix $I+\gamma T = (I-A)^{-1}(I-A+\gamma B)$ is invertible for $\gamma > 0$, and $R = (I+\gamma T)^{-1}S$, $P = (I+\gamma T)^{-1}\gamma T$. Furthermore, the consistency conditions in (3.5) imply that the linear multistep method is exact for first-degree polynomial solutions: if $u_j = \alpha + \beta\cdot j\Delta t$ $(0 \le j < k)$ and $F(u) \equiv \beta$, then $u_n = \alpha + \beta\cdot n\Delta t$ for all $n \ge k$. Since $y_n = u_{k-1+n}$ $(n \ge 1)$ in (3.19), it follows by varying $\alpha, \beta \in \mathbb{R}$ that

$$e_j^T S \;\neq\; 0 \qquad \text{for all } j\,, \tag{3.21a}$$

$$(e_i - e_j)^T [\,S\;T\,] \;\neq\; 0 \qquad \text{if } \; i \neq j\,, \tag{3.21b}$$

where $[S\;T]$ is the $m\times(k+m)$ matrix whose first k columns equal those of S and whose last m columns are equal to those of T. Application of Theorem 1.2.4, part (ii) in Chapter 1 now yields the following result:

Theorem 3.3.1. *Consider a linear multistep method (3.4) satisfying (3.5). Then, for any seminorm $\|\cdot\|$ on $\mathbb{V}$, the boundedness property (3.15) is valid provided that*

$$\|(I-|P|)^{-1}|R|\,\|_\infty \;\le\; \mu \qquad \textit{for all } m\,. \tag{3.22}$$

Moreover, condition (3.22) is necessary for (3.15) to be valid for the class of spaces $\mathbb{V} = \mathbb{R}^M$, $M \ge 1$, with the maximum norm.

In the above result, proving necessity of (3.22) is by far the most difficult part, and for that part the conditions (3.21) are relevant. Showing sufficiency is much easier, and we will repeat the main arguments here. For this purpose, note that for any seminorm $\|\cdot\|$, relation (3.16) implies

$$\|y_i\| \;\le\; \sum_{j=1}^{k} |r_{ij}|\,\|x_j\| + \sum_{j=1}^{m} |p_{ij}|\,\|y_j\| \qquad (1 \le i \le m)$$

whenever (3.2) is satisfied and $\Delta t \le \gamma\tau_0$. Setting $\eta = (\eta_i) \in \mathbb{R}^m$, $\xi = (\xi_j) \in \mathbb{R}^k$ with $\eta_i = \|y_i\|$ and $\xi_j = \|x_j\|$, we thus obtain

$$\eta \;\le\; |R|\,\xi + |P|\,\eta\,,$$

where $|R| = (|r_{ij}|)$, $|P| = (|p_{ij}|)$. Since $(I-|P|)^{-1} \ge 0$, it follows that

$$\eta \;\le\; (I-|P|)^{-1}|R|\,\xi\,,$$

from which it is seen directly that (3.22) implies (3.15).

3.4 Boundedness and monotonicity results

In this section conditions are given for boundedness and monotonicity of linear multistep methods. It will always be assumed that (3.5) is satisfied.

To formulate the results we will use some standard concepts for linear multistep methods, which can be found in Butcher (2003), Hairer, Nørsett & Wanner (1993), for example. The *stability region* of the linear multistep method is denoted by $\mathcal{S}$, and its interior by $\mathrm{int}(\mathcal{S})$. If $0 \in \mathcal{S}$ the method is said to be *zero-stable*. The method is called *irreducible* if the generating polynomials $\rho(\zeta) = \zeta^k - \sum_{j=1}^{k} a_j \zeta^{k-j}$ and $\sigma(\zeta) = \sum_{j=0}^{k} b_j \zeta^{k-j}$ have no common factor.

3.4.1 Boundedness with respect to the input vectors

First we consider the boundedness property (3.15) with $\mu > 0$ arbitrary, giving boundedness with respect to the input vectors $x_1, \ldots, x_k$ defined by (3.12). As we will see, this can be linked to some linear stability properties of the method and non-negativity of the matrices P, R. It is important to note that these $m \times m$ matrices depend explicitly on γ, and we are interested in m arbitrarily large.

For a given linear multistep method and given $\gamma > 0$ we consider the following two statements:

$$\begin{cases} \text{there is a } \mu > 0 \text{ such that the boundedness property (3.15)} \\ \text{is valid for all } \mathbb{V} = \mathbb{R}^M,\ M \geq 1, \text{ with maximum norm } \|\cdot\|_\infty\,; \end{cases} \tag{3.23}$$

$$\begin{cases} \text{there is a } \mu > 0 \text{ such that the boundedness property (3.15)} \\ \text{is valid for any vector space } \mathbb{V} \text{ and seminorm } \|\cdot\|\,. \end{cases} \tag{3.24}$$

The next theorem provides necessary and sufficient conditions for these statements. The proof of the theorem will be given in Section 3.5.

Theorem 3.4.1. *Consider an irreducible, zero-stable linear multistep method, and let $\gamma > 0$. Then each of the statements (3.23) and (3.24) is equivalent to*

$$-\gamma \in \mathrm{int}(\mathcal{S})\,, \qquad P \geq 0 \quad \textit{(for all } m\textit{)}\,. \tag{3.25}$$

Along with (3.23), (3.24), we also consider the following stronger statement on boundedness for arbitrary nonnegative sublinear functionals:

$$\begin{cases} \text{there is a } \mu > 0 \text{ such that the boundedness property (3.15)} \\ \text{is valid for any vector space } \mathbb{V} \text{ and nonnegative sublinear} \\ \text{functional } \|\cdot\|\,. \end{cases} \tag{3.26}$$

Here the restriction to sublinear functionals that are nonnegative has been made to get a similar formulation as for seminorms; see Remark 3.5.4 below.

Theorem 3.4.2. *Suppose the linear multistep method is zero-stable, $\gamma > 0$ and*

$$R \geq 0\,, \quad P \geq 0 \quad \textit{(for all } m\textit{)}\,. \tag{3.27}$$

Then statement (3.26) holds.

Also the proof of this theorem will be given in Section 3.5. In that section we will also see that if $k = 2$ and the method is irreducible, then $P \geq 0$ (for all m) implies $R \geq 0$ (for all m). We also have: (3.26) $\Rightarrow$ (3.24) $\Rightarrow$ (3.23) $\Rightarrow$ $P \geq 0$, where the last implication follows from Theorem 3.4.1. Consequently, for irreducible zero-stable linear two-step methods, each of the statements (3.23), (3.24), (3.26) is valid with stepsize coefficient $\gamma > 0$ if and only if $P \geq 0$ (for all m).

In the above results, zero-stability has been assumed in advance. It is clear, by considering $F \equiv 0$, that this is also a necessary condition for the relevant boundedness properties.

3.4.2 Boundedness with respect to the starting vectors

The above results provide criteria for boundedness with respect to the input vectors $x_1, \ldots, x_k$ defined in (3.12). In general, it is more natural to consider boundedness with respect to the starting vectors $u_0, \ldots, u_{k-1}$, as in (3.8). We therefore consider, similar to (3.15), the following boundedness property of the linear multistep scheme (3.4):

$$\max_{k \leq n < k+m} ||u_n|| \leq \tilde{\mu} \cdot \max_{0 \leq j < k} ||u_j|| \quad \begin{array}{l} \text{whenever (3.2) is valid, } \Delta t \leq \gamma\, \tau_0, \\ \text{and (3.4) holds for } k \leq n < k+m, \\ m \geq 1. \end{array} \tag{3.28}$$

If $\|\cdot\|$ is a seminorm, it is easily seen from (3.2) and (3.12) that

$$\|x_i\| \;\leq\; \sum_{j=1}^{k} \big(|a_j - \gamma b_j| + \gamma |b_j|\big) \cdot \max_{0 \leq l < k} \|u_l\|$$

for $i = 1, \ldots, k$. Consequently, if (3.15) holds with stepsize coefficient γ and factor μ, then there is a $\tilde{\mu}$ such that (3.28) holds.

The reverse is also true for seminorms. To see this, first note that (3.13b) is the same as (3.4), only with a shifted index. Therefore property (3.28) implies $\max_{k+1 \leq i \leq k+m} \|y_i\| \leq \tilde{\mu} \max_{1 \leq j \leq k} \|y_j\|$ when (3.2) is valid and $\Delta t \leq \gamma \tau_0$. From (3.13a) we see that

$$\|y_n\| \;\leq\; \|y_n - \Delta t b_0 F(y_n)\| \;\leq\; \|x_n\| + \sum_{j=1}^{n-1} \big(|a_j - \gamma b_j| + \gamma |b_j|\big) \|y_{n-j}\|$$

for $1 \leq n \leq k$. Here the first inequality follows by monotonicity of the backward Euler method for any stepsize; see for instance Hundsdorfer, Ruuth & Spiteri (2003). By induction with respect to n it is now seen that there are $\nu_1, \nu_2, \ldots, \nu_k$, only depending on the coefficients a_j, b_j and γ, such that

$$\|y_n\| \;\leq\; \nu_n \cdot \max_{1 \leq j \leq n} \|x_j\| \qquad (1 \leq n \leq k)\,.$$

It follows from the above that the boundedness properties (3.15) and (3.28) are for seminorms essentially equivalent, in the following sense:

Lemma 3.4.3. *Suppose $\|\cdot\|$ is a seminorm on a vector space $\mathbb{V}$, and let $\gamma > 0$. Then (3.15) holds with some $\mu > 0$ if and only if (3.28) holds with some $\tilde{\mu} > 0$.*

For sublinear functionals such an equivalence does not hold. As we know from Theorem 3.4.2, zero-stability and $P, R \geq 0$ is sufficient for having (3.15) with nonnegative sublinear functionals, and we will see in later examples that this is satisfied with $\gamma > 0$ for many methods, including methods with some negative coefficients a_j, b_j. On the other hand, by combining results on nonnegativity preservation as given in Bolley & Crouzeix (1978) with the functional $\|v\|_0 = -\min\{0, v_1, \ldots, v_M\}$ on $\mathbb{R}^M$, it can be shown that to have (3.28) with $\gamma > 0$ for all nonnegative sublinear functionals we need all $a_j, b_j \geq 0$ and $\gamma \leq c$ with $c > 0$ given by (3.6).

3.4.3 Monotonicity with starting procedures

For methods with nonnegative coefficients a_j and b_j we know that monotonicity is valid with respect to arbitrary starting values $u_0, u_1, \ldots, u_{k-1}$, with stepsize coefficient $\gamma \leq c$ given by (3.6). As mentioned before, this only applies to a small class of methods, and usually only under severe stepsize restrictions. Most popular methods used in practice have some negative coefficients. For such methods it is useful to consider specific starting procedures to compute $u_1, \ldots, u_{k-1}$ from u_0. For a given stepsize, this provides an input vector x determined by u_0. For suitable starting procedures we may still have monotonicity with respect to u_0, even if the multistep methods has some negative coefficients.

Assume that a Runge-Kutta type starting procedure is used, producing a vector $w = [w_j] \in \mathbb{V}^{m_0}$ such that $u_i = w_{j_i}$ for $i = 0, 1, \ldots, k-1$; the remaining w_j are internal stage vectors of the starting procedure. For given $\gamma > 0$ we write, using (3.12),

$$x \;=\; \boldsymbol{R}_0\, u_0 + \boldsymbol{P}_0\Big(w + \frac{\Delta t}{\gamma}\boldsymbol{F}(w)\Big) \tag{3.29}$$

with matrices $P_0 \in \mathbb{R}^{k\times m_0}$ and $R_0 \in \mathbb{R}^{k\times 1}$ determined by the starting procedure and the coefficients of the linear multistep method. Examples are given below.

Theorem 3.4.4. *Let $\|\cdot\|$ be a sublinear functional on a vector space $\mathbb{V}$. Suppose (3.29) holds with $\|w_j\| \leq \|u_0\|$ $(1 \leq j \leq m_0)$, $y \in \mathbb{V}^m$ satisfies (3.14), and*

$$R\,R_0 \geq 0\,, \qquad R\,P_0 \geq 0\,, \qquad P \geq 0\,. \tag{3.30}$$

Then $\|y_i\| \leq \|u_0\|$ for $1 \leq i \leq m$ whenever (3.2) is valid and $\Delta t \leq \gamma\tau_0$.

Proof. From (3.16) we obtain

$$y \;=\; \boldsymbol{R}\,\boldsymbol{R}_0 u_0 + \boldsymbol{R}\,\boldsymbol{P}_0\Big(w + \frac{\Delta t}{\gamma}\boldsymbol{F}(w)\Big) + \boldsymbol{P}\Big(y + \frac{\Delta t}{\gamma}\boldsymbol{F}(y)\Big)\,.$$

Setting $\eta = (\eta_i) \in \mathbb{R}^m$, $\eta_i = \|y_i\|$, it follows that

$$\eta \;\leq\; \big(R\,R_0 + R\,P_0\,\bar{e}\big)\|u_0\| \;+\; P\,\eta\,,$$

with unit vector $\bar{e} = e^{[m_0]} \in \mathbb{R}^{m_0}$. For the special case $F \equiv 0$, all w_j, y_i will be equal to u_0, from which it is seen that $e = R\,R_0 1 + R\,P_0 \bar{e} + P e$. Consequently

$$(I - P)\,\eta \;\leq\; (I - P)\,e \cdot \|u_0\|\,,$$

and since $(I - P)^{-1} \geq 0$ we obtain $\eta \leq e \cdot \|u_0\|$. □

A standard starting procedure consists of taking $k-1$ steps with a given s-stage Runge-Kutta method with stepsize Δt. In order to guarantee that $\|w_j\| \leq \|u_0\|$ for $1 \leq j \leq m_0$ as soon as (3.2) is valid and $\Delta t \leq \gamma \tau_0$, the Runge-Kutta method itself should be monotonic/SSP with stepsize coefficient γ.

Any Runge-Kutta starting procedure combined with m steps of the linear multistep method can be written together as one step of a 'big' Runge-Kutta method with $m_0 + m$ stages. The above result could therefore –in principle– also have been derived from the results in Ferracina & Spijker (2005), Higueras (2005). Necessary condition for monotonicity are found in Spijker (2007); it can be shown from those results that the condition (3.30) is necessary in Theorem 3.4.4 under a weak irreducibility condition on the combined scheme.

The following example shows how the matrices R_0, P_0 are obtained for some simple methods.

Example 3.4.5. Consider a two-step method, and let $c_j = a_j - \gamma b_j$ $(j = 1, 2)$. Then

$$x \;=\; \begin{pmatrix} c_2 & c_1 \\ 0 & c_2 \end{pmatrix} \begin{pmatrix} u_0 \\ u_1 \end{pmatrix} \;+\; \gamma \begin{pmatrix} b_2 & b_1 \\ 0 & b_2 \end{pmatrix} \begin{pmatrix} u_0 + \frac{1}{\gamma}\Delta t\, F(u_0) \\ u_1 + \frac{1}{\gamma}\Delta t\, F(u_1) \end{pmatrix}. \tag{3.31}$$

Suppose u_1 is computed by the θ-method, $u_1 = u_0 + \Delta t(1-\theta)F(u_0) + \Delta t\theta F(u_1)$. This can be written as

$$u_1 = r_0 u_0 + q_0\Big(u_0 + \frac{\Delta t}{\gamma} F(u_0)\Big) + q_1\Big(u_1 + \frac{\Delta t}{\gamma} F(u_1)\Big) \tag{3.32}$$

with $r_0 = (1+\theta\gamma)^{-1}(1-(1-\theta)\gamma)$, $q_0 = (1+\theta\gamma)^{-1}(1-\theta)\gamma$, and $q_1 = (1+\theta\gamma)^{-1}\theta\gamma$. This leads to (3.29) with

$$R_0 = \begin{pmatrix} c_2 + c_1 r_0 \\ c_2 r_0 \end{pmatrix}, \qquad P_0 = \begin{pmatrix} c_1 q_0 + \gamma b_2 & c_1 q_1 + \gamma b_1 \\ c_2 q_0 & c_2 q_1 + \gamma b_2 \end{pmatrix}, \tag{3.33}$$

and $w = (u_0, u_1)^T \in \mathbb{V}^2$. Of course, if the multistep method is explicit we will take $\theta = 0$, in which case $r_0 = 1 - \gamma$, $q_0 = \gamma$ and $q_1 = 0$.

Another natural starting procedure for explicit methods is the explicit trapezoidal rule (also known as the modified Euler method)

$$\bar{u}_1 = u_0 + \Delta t\, F(u_0)\,, \qquad u_1 = u_0 + \tfrac{1}{2}\Delta t\, F(u_0) + \tfrac{1}{2}\Delta t\, F(\bar{u}_1)\,.$$

Here we get

$$u_1 = r_0 u_0 + q_0\Big(u_0 + \frac{\Delta t}{\gamma} F(u_0)\Big) + q_1\Big(\bar{u}_1 + \frac{\Delta t}{\gamma} F(\bar{u}_1)\Big) \tag{3.34}$$

with $r_0 = 1 - \gamma + \frac{1}{2}\gamma^2$, $q_0 = \frac{1}{2}\gamma(1-\gamma)$ and $q_1 = \frac{1}{2}\gamma$. This gives

$$R_0 = \begin{pmatrix} c_2 + c_1 r_0 \\ c_2 r_0 \end{pmatrix}, \qquad P_0 = \begin{pmatrix} c_1 q_0 + \gamma b_2 & c_1 q_1 & \gamma b_1 \\ c_2 q_0 & c_2 q_1 & \gamma b_2 \end{pmatrix}, \tag{3.35}$$

and $w = (u_0, \bar{u}_1, u_1)^T \in \mathbb{V}^3$. ◇

3.5 Technical derivations and proofs

3.5.1 Recursions for the coefficients of P and R

We first take a closer look at the lower triangular $m \times m$ Toeplitz matrices

$$(I - A + \gamma B)^{-1} \;=\; \sum_{j \geq 0} \rho_j E^j \,, \tag{3.36}$$

$$P \;=\; (I - A + \gamma B)^{-1} \gamma B \;=\; \sum_{j \geq 0} \pi_j E^j \,, \tag{3.37}$$

with coefficients $\rho_j, \pi_j \in \mathbb{R}$. Note that the first r columns of $(I - A + \gamma B)^{-1}$, $r = \min\{k, m\}$, appear in $R \in \mathbb{R}^{m \times k}$.

It is convenient to define $\rho_j = 0$ for $j < 0$. The coefficients ρ_n then satisfy the multistep recursion

$$\rho_n \;=\; \sum_{j=1}^{k} a_j \rho_{n-j} - \gamma \sum_{j=0}^{k} b_j \rho_{n-j} + \delta_{n0} \qquad (n \geq 0)\,, \tag{3.38}$$

with Kronecker delta symbol δ_{n0} (whose value equals one if $n = 0$ and zero otherwise). In terms of these ρ_n, the coefficients π_n are given by

$$\pi_n \;=\; \gamma \sum_{j=0}^{k} b_j \rho_{n-j} \qquad (n \geq 0)\,. \tag{3.39}$$

This gives a direct link between these coefficients ρ_n, π_n and the behaviour of the linear multistep method applied to the scalar equation

$$u'(t) \;=\; \lambda\, u(t) \qquad \text{with } \Delta t\, \lambda = -\gamma\,. \tag{3.40}$$

Lemma 3.5.1. *If $-\gamma \in \mathcal{S}$ then $\max_{0 \leq n < \infty} |\rho_n| < \infty$. Furthermore, if the method is irreducible and $-\gamma \in \mathrm{int}(\mathcal{S})$, then there is a $\kappa > 0$ and $\theta \in (0,1)$ such that $|\rho_n| \leq \kappa\, \theta^n$ for all $n \geq 0$.*

Proof. From (3.38) we see that the coefficients ρ_n are obtained by applying the linear multistep method to (3.40). If $-\gamma \in \mathcal{S}$ this recursion is stable, and therefore the $|\rho_n|$ are bounded uniformly in n.

The characteristic roots of the recursion (3.38) are given by algebraic functions of γ. If the method is irreducible these functions are not (locally) constant. It follows that for any $-\gamma \in \mathrm{int}(\mathcal{S})$ there is a $\theta \in (0,1)$ such that the maximum modulus of the characteristic roots is less than θ; see Crouzeix & Raviart (1980, Thm. I.4.2). Writing the solution of (3.38) in terms of these characteristic roots thus provides the proof. □

Corollary 3.5.2. *Suppose the method is irreducible and* $-\gamma \in \mathrm{int}(\mathcal{S})$. *Then* $\sum_{j=0}^{\infty} \pi_j = 1$.

Proof. We have $\sum_{j=0}^{m-1} \pi_j = e_m^T P e = e_m^T \big(I - (I - A + \gamma B)^{-1}(I - A)\big) e$. Let $v = (I - A)e$. Then only the first k components v_j are nonzero. Consequently we obtain for $m \geq k$

$$e_m^T P e \;=\; 1 - (\rho_{m-1}, \ldots, \rho_1, \rho_0)\, v \;=\; 1 - \textstyle\sum_{j=1}^{k} \rho_{m-j} v_j \,.$$

The proof now follows from the previous lemma. □

The recursions (3.38), (3.39) will be used to compute numerically the largest stepsize coefficient γ such that $R \geq 0$ or $P \geq 0$ with large m. Necessary conditions for these inequalities can be obtained by computing the first few coefficients ρ_j and π_j by hand.

Example 3.5.3. For explicit methods we have

$$\rho_0 = 1\,, \quad \rho_1 = a_1 - \gamma b_1\,, \quad \rho_2 = a_1^2 + a_2 - \gamma(2a_1 b_1 + b_2) + \gamma^2 b_1^2\,,$$
$$\pi_0 = 0\,, \qquad \pi_1 = \gamma b_1\,, \qquad \pi_2 = \gamma(a_1 b_1 + b_2) - \gamma^2 b_1^2\,.$$

It is clear that the inequality $P \geq 0$ (for all m) with some $\gamma > 0$ requires $b_1 \geq 0$ and $a_1 b_1 + b_2 \geq 0$. These two inequalities were mentioned already in Hundsdorfer & Ruuth (2006), but now it is seen that these are really needed for boundedness. ◇

3.5.2 Proofs of Theorems 3.4.1, 3.4.2

Along with R and P, we will use in this section the $m \times m$ Toeplitz matrices $(I - A)^{-1} = \sum_{j \geq 0} \sigma_j E^j$ and $T = (I - A)^{-1} B = \sum_{j \geq 0} \tau_j E^j$, with entries $\sigma_j, \tau_j \in \mathbb{R}$ on the j-th lower diagonal, and we write $S = (I - A)^{-1} J$, cf. (3.20). Application of Lemma 3.5.1 with $\gamma = 0$ shows that if the method is zero-stable, then there is an $\alpha_0 > 0$ such that $|\sigma_j| \leq \alpha_0$ for all $j \geq 0$.

Sufficiency of (3.25) in Theorem 3.4.1

The following arguments are somewhat similar to those used in the proof of Corollary 3.3 of Chapter 1, although the notations are not completely matching.

Assume the linear multistep method is irreducible and zero-stable, $-\gamma \in \mathrm{int}(\mathcal{S})$ and $P \geq 0$. Setting $\beta_0 = \sum_{j=0}^{k} |b_j|$, it follows that $|\tau_j| \leq \alpha_0\beta_0$ for all $j \geq 0$. Lemma 3.5.1 shows that there is an $\alpha_1 > 0$ such that $\sum_{j=0}^{\infty} |\rho_j| \leq \alpha_1$. Since $P \geq 0$, we have

$$(I - |P|)^{-1}|R| = (I - P)^{-1}|R| = (I + \gamma T)|R|\,,$$

and consequently $\|(I - |P|)^{-1}|R|\|_\infty \leq (1 + \gamma\alpha_0\beta_0 k)\alpha_1$. Application of Theorem 3.3.1 thus shows that the statements (3.23), (3.24) are valid.

Necessity of (3.25) in Theorem 3.4.1

To finish the proof of Theorem 3.4.1 it has to be shown that for an irreducible, zero-stable method the conditions $P \geq 0$ and $-\gamma \in \mathrm{int}(\mathcal{S})$ are necessary for (3.23).

Any application of method (3.4) to the scalar, complex test equation $u'(t) = \lambda\, u(t)$ with $\lambda = \alpha + i\beta$ and real α, β, can be reformulated as an application to $u'(t) = F(u(t))$ in $\mathbb{V} = \mathbb{R}^2$ with $F(v) = (\alpha v_1 - \beta v_2, \beta v_1 + \alpha v_2)$ for $v = (v_1, v_2) \in \mathbb{V}$. Choosing $\lambda \in \mathcal{D} = \{\alpha + i\beta : -2 \leq \alpha \leq 0, |\beta| \leq \min(2 + \alpha, -\alpha)\}$, we have (3.2) with $\tau_0 = 1$, $\mathbb{V} = \mathbb{R}^2$ and $\|\cdot\| = \|\cdot\|_\infty$. Using Lemma 3.4.3, it thus follows that property (3.15) implies $\gamma \cdot \mathcal{D} \subset \mathcal{S}$. Therefore, if $\gamma > 0$, then $-\gamma \in \mathrm{int}(\mathcal{S})$ is certainly necessary for (3.23).

Assuming $-\gamma \in \mathrm{int}(\mathcal{S})$, it remains to show that $P \geq 0$ is necessary for (3.22). Let us write as before $P = \sum_{j\geq 0} \pi_j E^j$ with coefficients $\pi_j \in \mathbb{R}$. Because $-\gamma \in \mathrm{int}(\mathcal{S})$ we know by Corollary 3.5.2 that $\sum_{j=0}^{\infty} \pi_j = 1$. We can write (3.22) as

$$(I - |P|)^{-1}|R|\,\bar{e} \;\leq\; \mu\, e \qquad (\text{for all } m)\,,$$

where $\bar{e} = e^{[k]} \in \mathbb{R}^k$ and $e = e^{[m]} \in \mathbb{R}^m$.

Suppose some π_j are negative. Then there is an $l \geq 1$ with $\sum_{j=0}^{l} |\pi_j| > 1$. Consider now $m > l$, and let

$$D = \sum_{j=0}^{l} \delta_j E^j \qquad \text{with } \; \delta_j = |\pi_j| \;\text{ for }\; 0 \leq j \leq l\,.$$

We have $|R|\,\bar{e} \geq \left(e_1^T |R|\,\bar{e}\right) e_1 = (1 + \gamma b_0)^{-1} e_1$. Furthermore

$$(I - |P|)^{-1} - (I - D)^{-1} \;=\; (I - |P|)^{-1}(|P| - D)(I - D)^{-1} \;\geq\; 0\,,$$

and therefore $(I - |P|)^{-1}e_1 \geq (I - D)^{-1}e_1$. Consequently, (3.22) implies $(I - D)^{-1}e_1 \leq \tilde{\mu} e$ for all $m \geq l + 1$ with $\tilde{\mu} = (1 + \gamma b_0)\mu$. Note that $(I - D)^{-1}$ is again a lower triangular Toeplitz matrix, and therefore we also have

$$(I - D)^{-1}e_i \;\leq\; \tilde{\mu}\, e \qquad (\text{for all } m \geq l + 1 \text{ and } 1 \leq i \leq l)\,. \tag{3.41}$$

The bounds (3.41) are related to stability of the recursion

$$\eta_n = \sum_{j=0}^{l} \delta_j \eta_{n-j} \qquad (\text{for } \; n \geq l) \tag{3.42}$$

with starting values $\eta_0, \ldots, \eta_{l-1} \in \mathbb{R}$. For given $\eta_0, \ldots, \eta_{l-1}$ the solution for m steps of this recursion can be written as $(I-D)^{-1}\xi$ where $\xi = \sum_{i=1}^{l} \xi_i e_i \in \mathbb{R}^m$ collects the starting values in the form of source terms in the first l steps. Therefore, (3.41) implies stability of the recursion (3.42). However, this l-step recursion has characteristic polynomial

$$d(\zeta) \; = \; \zeta^l - \sum_{j=0}^{l} \delta_j \zeta^{l-j} \, .$$

Since $\delta_0 = \gamma b_0/(1+\gamma b_0)$ and $\sum_{j=0}^{l} \delta_j > 1$, we have $d(1) < 0$ but $d(\zeta) > 0$ for $\zeta \gg 1$. Hence there is a root larger than one, which contradicts stability of the recursion.

Consequently, having some negative entries in P implies that (3.22) is *not* satisfied. According to Theorem 3.3.1, also (3.23) is then not satisfied, which completes the proof of Theorem 3.4.1.

Sufficiency of (3.27) in Theorem 3.4.2

Let $\|\cdot\|$ be an arbitrary sublinear functional. If $P, R \geq 0$ then $S = (I-P)^{-1}R \geq 0$. Moreover, according to (3.18), we also have have $\mathrm{spr}(|P|) < 1$. Assuming (3.2) and $\Delta t \leq \gamma \tau_0$, it follows from Theorem 3.9 in Chapter 2 that

$$\|y_i\| \; \leq \mu_i \cdot \max_{1 \leq j \leq k} \|x_j\| \qquad (1 \leq i \leq m) \tag{3.43}$$

with $\mu_i = \sum_{j=1}^{k} \sigma_{i-j}$, where $\sigma_l = 0$ if $l < 0$. If the method is zero-stable, then $\mu = \sup_{1 \leq i < \infty} \mu_i < \infty$. For nonnegative sublinear functionals the property (3.15) then follows.

Remark 3.5.4. Replacement of the μ_i in (3.43) by $\mu = \sup_i \mu_i$ is not allowed for arbitrary sublinear functionals. Boundedness properties for arbitrary sublinear functionals should therefore not be expressed with (3.15). Theorem 3.4.2 has therefore been formulated for nonnegative sublinear functionals only.

Necessary and sufficient conditions for boundedness with the form (3.43) for arbitrary sublinear functionals are given in Chapter 2. However, as noted before, this will not lead to results in terms of the natural starting values $u_0, \ldots, u_{k-1}$, and therefore this will not be pursued here. $\diamond$

3.5.3 Conditions for $R \geq 0$ and $P \geq 0$ with two-step methods

For the case $k = 2$ we can formulate necessary and sufficient conditions for having $R \geq 0$ or $P \geq 0$ (for all $m \geq 1$) by writing down explicitly the solutions

of the recurrence relations (3.38), (3.39) for the coefficients ρ_n and π_n in terms of the roots of the characteristic polynomial of the recursion (3.38). The derivations are rather technical and not very revealing. Therefore we only present the results here, without the full derivation.

So, assume $k = 2$, and let $c_j = (1+\gamma b_0)^{-1}(a_j - \gamma b_j)$ for $j = 1, 2$. Setting $\rho_i = 0$ for $i < 0$ and $\rho_0 = (1 + \gamma b_0)^{-1}$, the coefficients ρ_n are given by the recursion $\rho_n = c_1\rho_{n-1} + c_2\rho_{n-2}$ for $n \geq 1$. Furthermore $\pi_n = \gamma b_0\rho_n + \gamma b_1\rho_{n-1} + \gamma b_2\rho_{n-2}$ for $n \geq 0$. These coefficients also satisfy the recursion $\pi_n = c_1\pi_{n-1} + c_2\pi_{n-2}$ for $n \geq 3$.

By solving the recursion in terms of the characteristic roots $\theta_\pm = \frac{1}{2}c_1 \pm \frac{1}{2}\sqrt{c_1^2 + 4c_2}$, thereby considering the cases of real or complex characteristic roots separately, it follows by some computations that $R \geq 0$ (for all m) if and only if

$$c_1 \geq 0\,, \qquad c_1^2 + 4c_2 \geq 0\,. \tag{3.44}$$

We note that under condition (3.44) the characteristic roots are real and $\theta_+ \geq |\theta_-|$.

The conditions for $P \geq 0$ can be studied in a similar way. For irreducible methods it can then be shown – by rather tedious calculations – that we have $P \geq 0$ (for all m) if and only if (3.44) holds together with

$$b_0c_1+b_1 \geq 0\,, \qquad b_0(c_1^2+c_2)+b_1c_1+b_2 \geq 0\,, \qquad b_0\theta^2+b_1\theta+b_2 \geq 0\,, \tag{3.45}$$

where $\theta = \frac{1}{2}c_1 + \frac{1}{2}\sqrt{c_1^2 + 4c_2}$. The first two inequalities in (3.45) just mean that $\pi_1, \pi_2 \geq 0$.

Remark 3.5.5. For any irreducible linear two-step method it is seen from the above that $R \geq 0$ is a necessary condition for $P \geq 0$ (for all m). To show that irreducibility is essential for this, consider an explicit two-step method with $a_1 + a_2 = 1$, $b_0 = 0$, $b_1 = 1$ and $b_2 = a_2$. Here we find that $\rho(\zeta) = (\zeta - 1)(\zeta + a_2)$ and $\sigma(\zeta) = \zeta + a_2$, so $\zeta = -a_2$ is a common root of the ρ and σ polynomials.

We have

$$(I - A + \gamma B)^{-1} = (I - (1 - \gamma)E)^{-1}(I + a_2E)^{-1}\,.$$

We see from (3.44) that $R \geq 0$ iff $\gamma \leq a_1 = 1 - a_2$. However, when calculating P the common factor drops out, resulting in

$$P = (I - (1 - \gamma)E)^{-1}\gamma E\,,$$

and therefore $P \geq 0$ iff $\gamma \leq 1$. Consequently, if $a_1 < 1$, then $P \geq 0$ does not imply $R \geq 0$ for these reducible methods. ◇

3.5.4 Remark on the construction in Hundsdorfer & Ruuth (2006) and Hundsdorfer, Ruuth & Spiteri (2003)

Multiplication of (3.14) with a Toeplitz matrix $K = \sum_{j\geq 0} \kappa_j E^j$ gives

$$y \;=\; \tilde{\boldsymbol{R}}\, x + \big(\tilde{\boldsymbol{P}} - \gamma\tilde{\boldsymbol{Q}}\big)\, y + \gamma\tilde{\boldsymbol{Q}}\,\big(y + \frac{\Delta t}{\gamma}\boldsymbol{F}(y)\big)\,,$$

where $\tilde{R} = KJ$, $\tilde{P} = I - K(I - A)$ and $\tilde{Q} = KB$. Taking $\kappa_0 = (1 + \gamma b_0)^{-1}$, we have $\mathrm{spr}(\tilde{P}) = |1 - \kappa_0| < 1$. If $K \geq 0$ is such that $\tilde{P} \geq \gamma \tilde{Q} \geq 0$, then we obtain as before

$$\eta \;\leq\; (I - \tilde{P})^{-1} \tilde{R}\, e \cdot \max_i \|x_i\| \;=\; (I - A)^{-1} J\, e \cdot \max_i \|x_i\|\,,$$

where $\eta = (\eta_i) \in \mathbb{R}^m$ with $\eta_i = \|y_i\|$.

Basically –in somewhat disguised form– this is what was used in Hundsdorfer, Ruuth & Spiteri (2003) for $k = 2$ and in Hundsdorfer & Ruuth (2006) for $k > 2$. In those papers, for a given integer l, chosen sufficiently large, the sequence $\{\kappa_j\}$ was taken to be geometric after index l, that is, $\kappa_{j+1}/\kappa_j = \theta$ for $j \geq l$. Subsequently, $\kappa_1, \ldots, \kappa_l, \theta \geq 0$ were determined (by an optimization code) to yield an optimal γ such that $\tilde{P} \geq \gamma \tilde{Q} \geq 0$. In fact, for $k = 2$ the whole sequence was taken in Hundsdorfer, Ruuth & Spiteri (2003) to be geometric, $\kappa_j = \kappa_0 \theta^j$, $j \geq 0$.

The present approach is more elegant. Moreover, it has a wider scope in that it gives conditions that are not only sufficient but also necessary for boundedness. It is remarkable that for many interesting methods the maximal values for γ seem to be the same. In this respect, note that if we take $K = (I - A + \gamma B)^{-1}$ then $K \geq 0$, $\tilde{P} \geq \gamma \tilde{Q} \geq 0$ is equivalent to $P, R \geq 0$.

3.6 Examples

For some families of methods, with two free parameters, we will display in contour plots the maximal values of γ such that we have boundedness with arbitrary input vectors (for seminorms) or monotonicity with starting procedures (for sublinear functionals), using (3.25) and (3.30), respectively. These maximal stepsize coefficients will be called *threshold values*.

The main criterion for boundedness is $P \geq 0$ for all $m \geq 1$. To verify this criterion, we compute the coefficients π_j from (3.38), (3.39) for $1 \leq j \leq m$ with a finite m, and check whether these coefficients are nonnegative. It is not a-priori clear how large this m should be taken in order to conclude that *all* π_j are nonnegative. The figures in this section were made with $m = 1000$, and it was verified that with a larger m the results did not differ anymore visually. For most methods a much smaller m would have been sufficient. Numerical inspection shows that in the generic case the recursion (3.38) has one dominant characteristic root $\theta \in \mathbb{R}$, giving asymptotically $\rho_n = c\,\theta^n(1 + \mathcal{O}(\kappa^n))$ for large n, with $c, \kappa \in \mathbb{R}$, $|\kappa| < 1$, and then $\mathrm{sgn}(\pi_n) = \mathrm{sgn}(c \sum_{j=0}^{k} b_j \theta^{-j})$ is constant for n large enough, provided θ is positive.

The threshold values for monotonicity with starting procedures can be obtained in a similar way: the first two inequalities in (3.30) amount to the inequality $\sum_{j=1}^{k} v_j \rho_{n-j} \geq 0$ for all $n \geq 1$ where $v = (v_1, \ldots, v_k)^T$ is any column of R_0 or P_0.

In the following, we will simply write $P \geq 0$ and $R \geq 0$ if the relevant inequality holds for all $m \geq 1$.

3.6.1 Explicit linear two-step methods of order one

Consider the class of explicit two-step methods of order (at least) one. With this class of methods we can take a_1, b_1 as free parameters, and set $a_2 = 1 - a_1$, $b_2 = 2 - a_1 - b_1$. The methods are zero-stable for $0 \leq a_1 < 2$. In case $b_1 = 2 - \frac{1}{2}a_1$ the order is two. The methods with $b_1 = 1$ or $a_1 = 2$ are reducible.

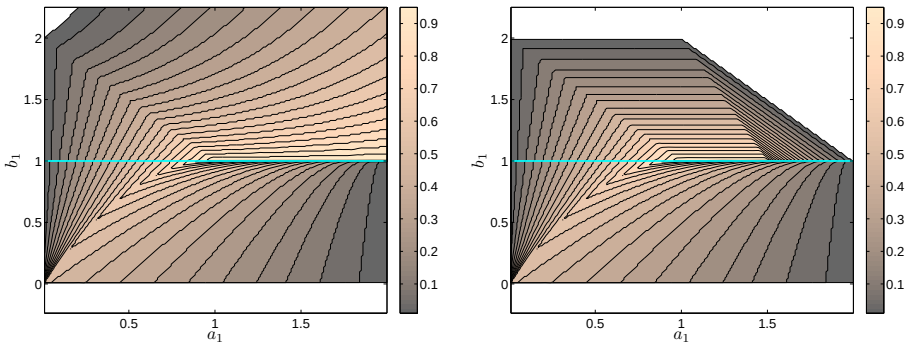

FIGURE 3.4: Explicit two-step methods of order one, with parameters $a_1 \in [0, 2)$ horizontally and $b_1 \in [-\frac{1}{4}, \frac{9}{4}]$, $b_1 \neq 1$, vertically. Left panel: threshold $\gamma > 0$ for boundedness. Right panel: threshold $\gamma > 0$ for monotonicity with forward Euler starting procedure. Contour levels at $j/20$, $j = 0, 1, \ldots$; for the 'white' areas there is no positive γ.

In Figure 3.4 (left panel) the maximal values of γ are displayed for which $P \geq 0$. As noted in Section 3.4.1, for the irreducible two-step methods these values of γ correspond to the threshold values for boundedness. For the 'white' areas in the contour plot there is no positive γ. We already know from Example 3.5.3 that if $b_1 < 0$ or $a_1 + b_1 - a_1 b_1 > 2$, then there is no $\gamma > 0$ for which $P \geq 0$.

In Figure 3.4 (right panel), the maximal values of γ are shown for which we have monotonicity with the forward Euler starting procedure. Note that $b_1 = 1$ is a special (reducible) case: starting with forward Euler, the whole scheme reduces to an application of the forward Euler method, so then we have monotonicity with $\gamma = 1$.

The methods (3.9) correspond to $a_1 = \frac{3}{2}$ and $b_1 = \beta$. It is now clear why $\beta = 0.95$ gave a much worse behaviour than $\beta = 1.05$ in the numerical example of Section 3.2. The maximal stepsize coefficient for boundedness is $\gamma \approx 0.35$ if $\beta = 0.95$ and $\gamma \approx 0.93$ if $\beta = 1.05$. With forward Euler start the maximal stepsize coefficient for monotonicity is $\gamma \approx 0.35$ if $\beta = 0.95$, and it is $\gamma \approx 0.82$ if $\beta = 1.05$. Therefore, the method with $\beta = 1.05$ allows much larger stepsizes for boundedness and monotonicity than the method with $\beta = 0.95$.

3.6.2 Implicit linear two-step methods of order two

Likewise we can consider the implicit two-step methods of order (at least) two, with free parameters a_1 and b_0. The remaining coefficients are then determined by $a_2 = 1 - a_1$, $b_1 = 2 - \frac{1}{2}a_1 - 2b_0$ and $b_2 = -\frac{1}{2}a_1 + b_0$. Again, the methods are zero-stable if $a_1 \in [0, 2)$, and they are A-stable if we also have $b_0 \geq \frac{1}{2}$. In case $b_0 = \frac{1}{3} + \frac{1}{12}a_1$ the order is three. The methods with $b_0 = \frac{1}{2}$ are reducible (to the trapezoidal rule).

The threshold values for boundedness are displayed in Figure 3.5 (left panel). These values correspond to those found earlier in Hundsdorfer & Ruuth (2006, Fig. 2). We now see from Theorem 3.4.1 that −somewhat surprisingly− the latter values, which were obtained by ad-hoc arguments, are not only sufficient but also necessary for boundedness.

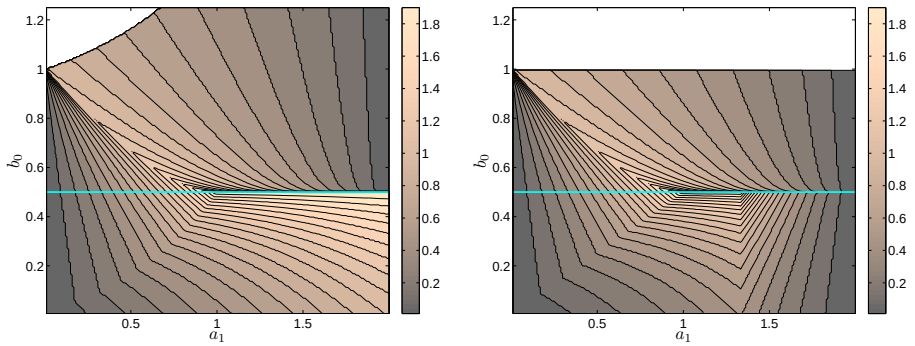

FIGURE 3.5: Implicit two-step methods of order two, with parameters $a_1 \in [0, 2)$ horizontally and $b_0 \in [0, \frac{5}{4}]$, $b_0 \neq \frac{1}{2}$, vertically. Left panel: thresholds $\gamma > 0$ for boundedness. Right panel: thresholds $\gamma > 0$ for monotonicity with the θ-method, $\theta = b_0$, as starting procedure. Contour levels at $j/10$, $j = 0, 1, \ldots$; for the 'white' areas there is no positive γ.

For the starting procedure we consider the θ-method, with $\theta = 1$ (backward Euler) or $\theta = b_0$. One might think that the monotonicity properties would be optimal with $\theta = 1$. That turns out not to be the case. In Figure 3.5 (right panel) the monotonicity thresholds are plotted for $\theta = b_0$. For $\theta = 1$ these thresholds become zero in the lower-right part ($b_0 \leq \frac{1}{2}a_1$) of the parameter plane; this is due to lack of monotonicity after one application of the two-step method.

3.6.3 Explicit linear three-step methods of order three

The class of explicit three-step methods of order three can be described with a_1, a_3 as free parameters, and then $a_2 = 1 - a_1 - a_3$, $b_1 = \frac{1}{12}(28 - 5a_1 - a_3)$, $b_2 = -\frac{8}{12}(1 + a_1 - a_3)$, $b_3 = \frac{1}{12}(4 + a_1 + 5a_3)$. Inspection shows that these

methods are zero-stable for (a_1, a_3) inside the triangle with vertices $(-1, 1)$, $(1, -1)$ and $(3, 1)$. Well-known examples in this class are the three-step Adams-Bashforth method, with $a_1 = 1$, $a_3 = 0$, and the extrapolated BDF3 method, with $a_1 = \frac{18}{11}$, $a_3 = \frac{2}{11}$.

In Figure 3.6 (right panel) the maximal value of γ is shown such that both $P \geq 0$ and $R \geq 0$. This corresponds to the values found Hundsdorfer, Ruuth (2006, Fig. 1). The left panel of the figure shows the maximal γ for which $P \geq 0$ and $-\gamma \in \text{int}(\mathcal{S})$.

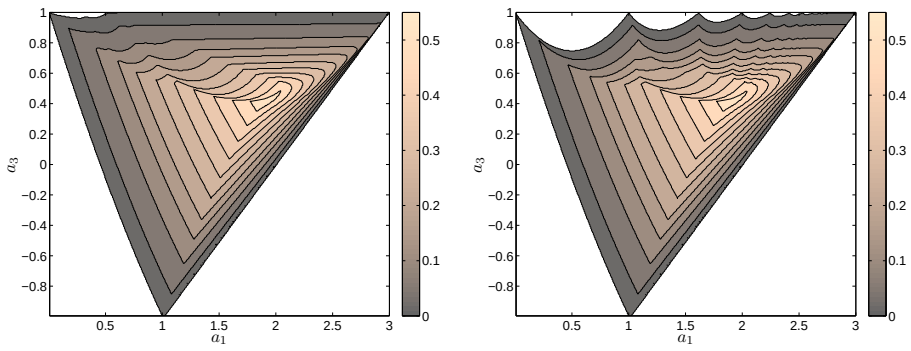

FIGURE 3.6: Explicit three-step methods of order three, with parameters $a_1 \in [0, 3]$ horizontally and $a_3 \in [-1, 1]$ vertically. Left panel: threshold $\gamma > 0$ for boundedness, that is, $P \geq 0$ and $-\gamma \in \text{int}(\mathcal{S})$. Right panel: maximal $\gamma > 0$ such that $P \geq 0$ and $R \geq 0$. Contour levels at $j/20$, $j = 0, 1, \ldots$; for the 'white' areas there is no positive γ.

It is seen that for many of the methods with $a_3 > 0.5$ the maximal γ for which $P \geq 0$ is slightly lager than for $P, R \geq 0$. For $a_3 < 0.5$ there is very little difference in the two pictures. In particular, the method obtained by optimization in Ruuth & Hundsdorfer (2005), with $a_1 \approx 1.91$ and $a_3 \approx 0.43$, is still optimal with respect to the threshold value, with $\gamma \approx 0.53$. Once again, these results put the earlier findings of Hundsdorfer & Ruuth (2006), Ruuth & Hundsdorfer (2005) in a new and wider perspective.

3.6.4 Explicit linear four-step methods of order four

For the class of explicit four-step methods of order four, the order conditions read $a_4 = 1 - (a_1 + a_2 + a_3)$, $b_4 = -\frac{1}{24}(9a_1 + 8a_2 + 9a_3)$, $b_3 = \frac{1}{6}(\frac{5}{2}a_1 + 2a_2 + \frac{9}{2}a_3 + 16a_4 - 18b_4)$, $b_2 = \frac{1}{2}(-a_1 + 3a_3 + 8a_4 - 4b_3 - 6b_4)$, $b_1 = a_1 + 2a_2 + 3a_3 + 4a_4 - (b_2 + b_3 + b_4)$. This still leaves three free parameters a_1, a_2, a_3, which makes visualization difficult.

We therefore consider a plane that contains three important schemes within this class: the explicit four-step Adams-Bashforth method (AB4), the extrapolated BDF4 scheme (EBDF4) and the method TVB(4,4) from Ruuth & Hunds-

dorfer (2005), given in Hundsdorfer & Ruuth (2007) with rational coefficients. Now two degrees of freedom remain. We take a_1, a_3 as free parameters, and set $a_2 = \frac{76772}{68211}(1 - a_1) - \frac{43115}{68211}a_3$.

In Figure 3.7 (left panel) the maximal value of γ is shown such that the methods are zero-stable, $-\gamma \in \mathrm{int}(\mathcal{S})$ and $P \geq 0$. The right panel shows the error constants (defined as in Hairer, Nørsett & Wanner (2003, Sect. III.2) for the zero-stable methods).

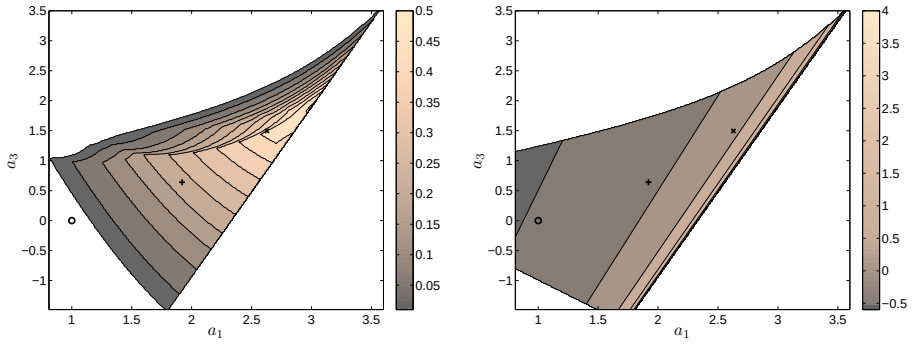

FIGURE 3.7: Explicit four-step methods of order four, with parameters described above. Left panel: threshold $\gamma > 0$ for boundedness. Contour levels at $j/20$, $j = 0, 1, \ldots$; for the 'white' areas there is no positive γ. Right panel: $\log_{10}$ of the absolute error constants for zero-stable methods. Markers: $\circ$ for AB4, $+$ for EBDF4 and $\times$ for TVB(4,4).

It is seen that the threshold value γ for boundedness is relatively large for the method TVB(4,4), with $a_1 \approx 2.63$ and $a_3 \approx 1.49$. This method was derived in Ruuth & Hundsdorfer (2005) by numerical optimization of γ within the class of explicit four-step methods of order four, based on the sufficient condition for boundedness discussed in Section 3.5.4, while keeping the error constants at a moderate size.

It is clear from the figure that the threshold value for boundedness can be slightly increased by taking (a_1, a_3) closer to $(3, 2)$. But then the error constant becomes much larger. Therefore the conclusion of Ruuth & Hundsdorfer (2005) still stands: the TVB(4,4) scheme gives a good compromise between a moderate error constant 2.38 and a relatively large stepsize coefficient $\gamma \approx 0.45$.

3.7 Concluding remarks

Based on the general framework of Chapter 1, we have obtained in this chapter stepsize restrictions for linear multistep methods that are necessary and sufficient for boundedness with maximum norms or arbitrary seminorms (Theorem 3.4.1). This puts the previously found, more complicated sufficient condi-

tions of Hundsdorfer & Ruuth (2006), Hundsdorfer, Ruuth & Spiteri (2003) in a better and wider perspective.

Moreover, it is now also seen that the essential condition for boundedness, $P \geq 0$, arises as a natural condition for monotonicity of linear multistep methods with Runge-Kutta starting procedures (Theorem 3.4.4). Optimizing the starting procedures for given classes of multistep methods is part of our ongoing research.

Instead of linear multistep methods, boundedness can be considered for the related class of one-leg methods. These methods were originally introduced in Dahlquist (1975) to facilitate the analysis of linear multistep methods. Stability results with inner-product norms for one-leg methods often have a somewhat nicer form than for linear multistep methods; see e.g. Butcher (2003), Hairer & Wanner (1996). We have found that the maximal stepsize coefficient for boundedness (with general seminorms) of a one-leg method is the same as for the associated linear multistep method, but simplification of the theory is *not* achieved in this way.

In the same way one can study the important class of predictor-corrector methods. However, for such methods the matrices P and R do become much more complicated than for linear multistep methods. Instead of simple Toeplitz matrices we then have to work with block matrices where the blocks have a Toeplitz structure. Sufficient conditions for boundedness are presented in Chapter 2.

Chapter 4

Comparison of Boundedness and Monotonicity Properties of One-Leg and Linear Multistep Methods

One-leg multistep methods have some advantage over linear multistep methods with respect to storage of the past results. In this chapter boundedness and monotonicity properties with arbitrary (semi-)norms or convex functionals are analyzed for such multistep methods. The maximal stepsize coefficient for boundedness and monotonicity of a one-leg method is the same as for the associated linear multistep method when arbitrary starting values are considered. It will be shown, however, that combinations of one-leg methods and Runge-Kutta starting procedures may give very different, and possibly larger, stepsize coefficients for monotonicity than the linear multistep methods with the same starting procedures.

4.1 Introduction

4.1.1 The ODE systems and basic assumptions

We consider general systems of ordinary differential equations (ODEs) in a vector space $\mathbb{V}$ on the time interval $[0, \infty)$ with given initial value, written as

$$u'(t) = F(u(t)), \quad u(0) = u_0, \tag{4.1}$$

with $u_0 \in \mathbb{V}$ and $F : \mathbb{V} \to \mathbb{V}$. In the following we will make the following basic assumption: there is a constant $\tau_0 > 0$ such that

$$\|v + \tau_0 F(v)\| \leq \|v\| \qquad \text{for all } v \in \mathbb{V}, \tag{4.2}$$

where $\|\cdot\|$ denotes a norm, a seminorm, or a convex functional on $\mathbb{V}$.

It is easy to see that (4.2) implies $\|v + \Delta t\, F(v)\| \le \|v\|$ for all $\Delta t \in (0, \tau_0]$. Consequently, applying the forward Euler method $u_n = u_{n-1} + \Delta t\, F(u_{n-1})$, $n \ge 1$, with stepsize $\Delta t > 0$ to compute approximations $u_n \approx u(t_n)$ at $t_n = n\Delta t$, we obtain $\|u_n\| \le \|u_0\|$ for $n \ge 1$ under the stepsize restriction $\Delta t \le \tau_0$. For general one-step methods, this property under a stepsize restriction $\Delta t \le \gamma\, \tau_0$ is often referred to as *monotonicity* or *strong stability preservation* (SSP). In this chapter we shall study similar properties for multistep methods.

In applications, the vector space $\mathbb{V}$ usually is the $\mathbb{R}^M$. Useful norms are for example the maximum norm $\|v\|_\infty = \max_{1\le j\le M} |v_j|$ and the sum norm $\|v\|_1 = \sum_{j=1}^M |v_j|$ for $v = (v_j) \in \mathbb{R}^M$. Norms that are generated by an inner product, such as the Euclidian norm $\|v\|_2 = (\sum_{j=1}^M |v_j|^2)^{1/2}$, are not focussed on in this chapter; for such norms other boundedness results exist, under more relaxed stepsize conditions, related to G-stability, see Dahlquist (1975) or Hairer & Wanner (1996), for example.

To include related properties, such as maximum principles (as in Spijker (2007)) and positivity preservation (as in Chapter 2), it can be useful to consider more general functionals. Recall that $\varphi : \mathbb{V} \to \mathbb{R}$ is a *convex functional* on $\mathbb{V}$ if

$$\varphi(\lambda v + (1-\lambda)w) \;\le\; \lambda\,\varphi(v) + (1-\lambda)\,\varphi(w) \qquad (\text{for } 0 \le \lambda \le 1 \text{ and } v, w \in \mathbb{V}).$$

It is called a *nonnegative sublinear functional* if $\varphi(v+w) \le \varphi(v) + \varphi(w)$ and $\varphi(cv) = c\varphi(v) \ge 0$ for all real $c \ge 0$ and $v, w \in \mathbb{V}$. It is a *seminorm* if we have in addition $\varphi(-v) = \varphi(v) \ge 0$ for all $v \in \mathbb{V}$. Finally, if it also holds that $\varphi(v) = 0$ only if $v = 0$, then φ is a *norm*.

4.1.2 Linear multistep and one-leg methods

In the following we will consider one-leg and linear multistep methods for finding the approximations $u_n \approx u(t_n)$ at the step points $t_n = n\Delta t$, $n \ge 1$. It is supposed that starting vectors $u_0, u_1, \ldots, u_{k-1} \in \mathbb{V}$ are known.

A *linear multistep method* applied to (4.1) reads

$$u_n = \sum_{j=1}^{k} a_j u_{n-j} + \Delta t \sum_{j=0}^{k} b_j F(u_{n-j}) \tag{4.3}$$

for $n \ge k$. The parameters a_j, b_j and $k \in \mathbb{N}$ define the method. Along with this linear multistep method, we also consider the corresponding *k-step one-leg method*

$$u_n = \sum_{j=1}^{k} a_j u_{n-j} + \Delta t \beta F(v_n)\,, \qquad v_n = \sum_{j=0}^{k} \hat{b}_j u_{n-j} \tag{4.4}$$

for $n \ge k$, where $\hat{b}_j = b_j/\beta$ and $\beta = \sum_{j=0}^k b_j \ne 0$. If $b_0 = 0$ these multistep methods are called *explicit*, and if $b_0 \ne 0$ they are called *implicit*.

It will be assumed throughout this chapter that

$$\sum_{j=1}^{k} a_j \;=\; 1\,, \qquad \sum_{j=1}^{k} j\,a_j = \sum_{j=0}^{k} b_j > 0\,, \qquad b_0 \geq 0\,. \tag{4.5a}$$

Here, the two equalities give the conditions for consistency of order one. Having $\sum_{j=0}^{k} b_j > 0$ is then necessary for zero-stability of the methods. The assumption $b_0 \geq 0$ will be convenient in this chapter; it holds for all well-known methods. Furthermore, for the generating polynomials $\rho(\zeta) = \zeta^k - \sum_{j=1}^{k} a_j \zeta^{k-j}$ and $\sigma(\zeta) = \sum_{j=0}^{k} b_j \zeta^{k-j}$ it will be assumed that

$$\rho(\zeta) \text{ and } \sigma(\zeta) \text{ have no common factor}\,. \tag{4.5b}$$

Methods that do not satisfy this last condition are said to be reducible (in the sense of Dahlquist), and these are essentially equivalent to a $(k-1)$-step method.

One-leg methods were introduced by Dahlquist (1975), originally only to facilitate the analysis of linear multistep methods. Subsequently, it was realized that one-leg methods might be useful on their own, not just as an analysis tool. It is known that the conditions for consistency of order p are the same if $p = 1, 2$, but for larger p the one-leg method has to satisfy more order conditions than the corresponding linear multistep method; cf. Hairer & Wanner (1996), for instance.

On the other hand, one-leg methods have an advantage over the corresponding linear multistep methods with respect to storage, which is very important for large-scale problems when function evaluations of F are expensive. If, for example, $a_k, b_k \neq 0$, then for a step (4.3) with the linear multistep method we need storage of the vectors $u_{n-1}, \ldots, u_{n-k}$ and $F(u_{n-2}), \ldots, F(u_{n-k})$, together with an evaluation of $F(u_{n-1})$. For a step (4.4) with the one-leg method only storage of $u_{n-1}, \ldots, u_{n-k}$ is needed, together with evaluation of $F(v_n)$.

4.1.3 Scope of the chapter

In this chapter we will first consider the property

$$\|u_n\| \;\leq\; \mu \cdot \max_{0 \leq j < k} \|u_j\| \qquad \text{for all } \; n \geq k \;\text{ and }\; 0 < \Delta t \leq \gamma\,\tau_0\,, \tag{4.6}$$

where the factor $\mu \geq 1$ and the stepsize coefficient $\gamma \geq 0$ are determined by the multistep method. If (4.6) holds with $\gamma > 0$ and $\mu = 1$ whenever the basic assumption (4.2) is satisfied, then this property will be called *monotonicity*. For many interesting methods, this property (4.6) will only hold with some $\mu > 1$, in which case we refer to it as *boundedness*.

It is known, see e.g. Gottlieb, Ketcheson & Shu (2011), Spijker (2007), that the condition for monotonicity for either the linear multistep method (4.3) or the one-leg method (4.4) reads

$$a_j \;\geq\; \gamma\, b_j \;\geq\; 0 \qquad (1 \leq j \leq k)\,. \tag{4.7}$$

This requires that all coefficients of the method are non-negative, which severely restricts the class of methods. It is therefore of interest to study more relaxed properties.

The boundedness property (4.6), with $\mu \geq 1$, has been studied for linear multistep methods. Sufficient stepsize conditions $\Delta t \leq \gamma \tau_0$ were derived in Hundsdorfer, Ruuth & Spiteri (2003), Hundsdorfer & Ruuth (2006) for having (4.6) with arbitrary seminorms under the basic assumption (4.2). Using the general framework of Chapter 1, more simple conditions were found in Chapter 3, and these more simple conditions were shown be sufficient and necessary.

In (4.6) the starting values $u_1, \ldots, u_{k-1}$ are arbitrary. In practice these starting values will be computed from the given initial value u_0, for instance by a Runge-Kutta method. For such combinations of multistep methods and Runge-Kutta starting procedures the following monotonicity property

$$\|u_n\| \ \leq \ \|u_0\| \qquad \text{for all} \ \ n \geq 1 \ \ \text{and} \ \ 0 < \Delta t \leq \gamma \, \tau_0 \, , \tag{4.8}$$

can still be valid, even if the multistep method itself is not monotone, but only bounded for arbitrary starting values, that is, (4.6) is valid with $\mu > 1$, not with $\mu = 1$.

For some combinations of linear multistep methods and Runge-Kutta starting procedures, the monotonicity property (4.8) was studied in Chapter 3, where conditions were derived with arbitrary seminorms and nonnegative sublinear functionals. Earlier, for some two-step methods sufficient conditions with arbitrary seminorms were found in Hundsdorfer, Ruuth & Spiteri (2003).

We will first consider the boundedness property (4.6) with arbitrary staring vectors. It will be seen that the maximal stepsize coefficient for boundedness of a one-leg method is the same as for the associated linear multistep method. In view of the close connection between one-leg and linear multistep methods, this result is not surprising.

Next, we will give conditions for having monotonicity (4.8) for multistep methods with a starting procedure. Then different conditions will arise for the one-leg and linear multistep methods.

A detailed study of these conditions for the class of explicit two-step methods will reveal that combinations of these one-leg methods with natural Runge-Kutta starting procedures can give monotonicity with much larger stepsizes than for the linear multistep methods with the same starting procedures.

4.2 General framework

In this chapter we will use general results on monotonicity of Spijker (2007) and on boundedness of Chapter 1. We will apply the results to the multistep methods (4.3) and (4.4), but our framework does allow general multistep methods with any number of internal stages. The notation used in Spijker (2007) and Chapter 1 will be followed as far as possible.

For any given $m \geq 1$, let $e_1, e_2, \ldots, e_m$ stand for the unit basis vectors in $\mathbb{R}^m$, that is, the j-th element of e_i equals one if $i = j$ and zero otherwise. The $m \times m$ identity matrix is denoted by I. Furthermore, $e = e_1 + e_2 + \cdots + e_m$ is the vector in $\mathbb{R}^m$ with all components equal to one. If it is necessary to specify the dimension we will denote these unit vectors by $e^{[m]}$; usually the proper dimension will be clear from the context.

For any $m \times l$ matrix $K = (\kappa_{ij})$ we will denote by the boldface symbol $\boldsymbol{K}$ the associated linear mapping from $\mathbb{V}^l$ to $\mathbb{V}^m$, that is, $\eta = \boldsymbol{K}\xi$ for $\eta = [\eta_i] \in \mathbb{V}^m$, $\xi = [\xi_i] \in \mathbb{V}^l$ if $\eta_i = \sum_{j=1}^{l} \kappa_{ij}\xi_j \in \mathbb{V}$ $(1 \leq i \leq m)$. Inequalities for vectors or matrices are to be understood component-wise. In particular, we will use the notation $K \geq 0$ when all entries κ_{ij} of this matrix are non-negative.

Application of a multistep method with a fixed number of steps leads to a process of the generic form

$$y_i \;=\; \sum_{j=1}^{k} s_{ij} x_j \;+\; \Delta t \sum_{j=1}^{m} t_{ij} F(y_j) \qquad (1 \leq i \leq m)\,, \tag{4.9}$$

producing the output vectors $y_1, y_2, \ldots, y_m$ from the input data $x_1, \ldots, x_k$ in $\mathbb{V}$. Typically, the set of output vectors will contain approximations u_n, $n \geq k$, whereas the input vectors x_j will consist of linear combinations of the starting vectors $u_0, u_1, \ldots, u_{k-1}$ and their function values $F(u_0), F(u_1), \ldots, F(u_{k-1})$.

Let $y = [y_i] \in \mathbb{V}^m$, $x = [x_i] \in \mathbb{V}^k$, and denote $\boldsymbol{F}(y) = [F(y_i)] \in \mathbb{V}^m$. The coefficient matrices for the process (4.9) are $S = (s_{ij}) \in \mathbb{R}^{m \times k}$ and $T = (t_{ij}) \in \mathbb{R}^{m \times m}$. With the above notations, the generic process (4.9) can be written in a compact way as

$$y \;=\; \boldsymbol{S}\, x + \Delta t \boldsymbol{T}\boldsymbol{F}(y)\,. \tag{4.10}$$

Let $[S\;T]$ be the $m \times (k+m)$ matrix whose first k columns equal those of S and whose last m columns are equal to those of T. As we will see, the generic processes that are generated by the multistep methods will be such that all rows of S are not zero and all rows of $[S\;T]$ are different from each other. With unit basis vectors $e_i \in \mathbb{R}^m$, $1 \leq i \leq m$, this can be expressed as

$$e_j^T S \;\neq\; 0 \qquad \text{for all } j\,, \tag{4.11a}$$

$$e_i^T [\,S\;T\,] \;\neq\; e_j^T [\,S\;T\,] \qquad \text{if } \; i \neq j\,. \tag{4.11b}$$

It is obvious that two identical rows in $[S\,T]$ lead to two output vectors y_i and y_j, $i \neq j$, with $y_i = y_j$ for any function F and arbitrary input vectors x_l. This was called reducibility in Spijker (2007). In this chapter we will refer to such a scheme as *reducibile* (in the sense of Spijker), and a scheme for which all rows of $[S\;T]$ are different from each other is called *irreducible* (in the sense of Spijker).

Boundedness for arbitrary starting vectors

If $\gamma > 0$ is such that $I + \gamma T$ is not singular, we can write the process also in the form

$$y \;=\; \boldsymbol{R}x + \boldsymbol{P}\Big(y + \frac{\Delta t}{\gamma}\boldsymbol{F}(y)\Big)\,, \tag{4.12}$$

where $R \in \mathbb{R}^{m\times k}$ and $P \in \mathbb{R}^{m\times m}$ are given by

$$R \;=\; (I+\gamma T)^{-1}S\,, \qquad P \;=\; (I+\gamma T)^{-1}\gamma T\,. \tag{4.13}$$

The number of steps with the multistep methods will be arbitrary, so the number m will allowed to be arbitrarily large as well. Consider, for given vector space $\mathbb{V}$ and seminorm $\|\cdot\|$, the boundedness property

$$\left\{\begin{array}{ll} \max\limits_{1\le i\le m} ||y_i|| \;\le\; \mu \cdot \max\limits_{1\le j\le k} ||x_j|| & \text{whenever (4.2) is valid, } \Delta t \le \gamma\,\tau_0, \\ & \text{and } x, y \text{ satisfy (4.10)}, \; m \ge 1, \end{array}\right. \tag{4.14}$$

with a stepsize coefficient $\gamma > 0$ and boundedness factor $\mu \ge 1$. Note that this bound holds uniformly for all initial value problems (4.1) under the basic assumption (4.2) with given $\tau_0 > 0$.

For any $m \times m$ matrix $K = (\kappa_{ij})$, let $\mathrm{spr}(K)$ be the spectral radius of K, and let $\|K\|_\infty = \max_i \sum_j |\kappa_{ij}|$ stand for the induced maximum norm of K. If $K = (\kappa_{ij})$ is an $m \times l$ matrix, then the matrix with entries $|\kappa_{ij}|$ is denoted by $|K|$, From Proposition 1.4.2 and Theorem 1.2.4 in Chapter 1 we have the following result:

Theorem 4.2.1. *Assume $I + \gamma T$ is not singular, and $spr(|P|) < 1$. Then, for any vector space $\mathbb{V}$ with seminorm $\|\cdot\|$, the boundedness property (4.14) is valid provided that*

$$\|(I - |P|)^{-1}|R|\,\|_\infty \;\le\; \mu \qquad \textit{for all } m\,. \tag{4.15}$$

Moreover, if (4.11) holds, then the condition (4.15) is necessary for (4.14) to be valid for the class of spaces $\mathbb{V} = \mathbb{R}^M$, $M \ge 1$, with the maximum norm.

In this theorem, the assumptions that $I+\gamma T$ is not singular and $\mathrm{spr}(|P|) < 1$ could be included into (4.15). However, for the multistep methods considered in this chapter these assumptions will hold trivially.

Furthermore, we note that boundedness as in (4.14), that is, boundedness with respect to the input vectors x_j, can be considered for functionals that are more general than seminorms. However this does not lead to boundedness results with respect to the starting vectors $u_0, \ldots, u_{k-1}$, as in (4.6), unless additional constraints on the methods are imposed. For example, as pointed out in Chapter 3, for linear multistep methods that would lead again to the very strict conditions (4.7).

Monotonicity with starting procedures

Instead of arbitrary starting vectors $u_0, u_1, \ldots, u_{k-1}$ for the multistep methods, we will consider Runge-Kutta starting procedures to generate these vectors from u_0. Assume this starting procedure produces a vector $w = [w_i] \in \mathbb{V}^{m_0}$, $m_0 \geq k$, where $u_j = w_{i_j}$ for $j = 0, 1, \ldots, k-1$ and the remaining w_i are internal stage vectors of the starting procedure.

The whole starting procedure, which may consist of several steps of a Runge-Kutta method, can be conveniently written as a single step

$$w = \boldsymbol{e_0} u_0 + \Delta t \boldsymbol{K_0} \boldsymbol{F}(w)\,, \tag{4.16}$$

where $e_0 = e^{[m_0]} = (1, \ldots, 1)^T \in \mathbb{R}^{m_0}$, and $K_0 \in \mathbb{R}^{m_0 \times m_0}$ is the coefficient matrix of this Runge-Kutta starting procedure. As is well known, see e.g. Spijker (2007), the conditions of Kraaijevanger (1991)

$$(I + \gamma K_0)^{-1} e_0 \geq 0\,, \qquad (I + \gamma K_0)^{-1} \gamma K_0 \geq 0 \tag{4.17}$$

guarantee that the starting procedure itself is monotone with stepsize coefficient γ, that is, $\|w_j\| \leq \|u_0\|$ $(1 \leq j \leq m_0)$ whenever (4.2) is valid, $\Delta t \leq \gamma\, \tau_0$, for any vector space $\mathbb{V}$ and convex functional $\|\cdot\|$.

The above Runge-Kutta starting procedure will give an input vector of the form

$$x = \boldsymbol{S_0} u_0 + \Delta t \boldsymbol{T_0} \boldsymbol{F}(w) \tag{4.18}$$

with $S_0 \in \mathbb{R}^{k \times 1}$, $T_0 \in \mathbb{R}^{k \times m_0}$. The total scheme, consisting of the multistep method and starting procedure can therefore be written as

$$\begin{cases} w = \boldsymbol{e_0}\, u_0 + \Delta t \boldsymbol{K_0} F(w)\,, \\ y = \boldsymbol{S} \boldsymbol{S_0} u_0 + \Delta t \boldsymbol{S}\, \boldsymbol{T_0} \boldsymbol{F}(w) + \Delta t \boldsymbol{T} \boldsymbol{F}(y)\,. \end{cases} \tag{4.19}$$

For the multistep methods, with sum of a_j equal to one, the output vectors y_i will be consistent approximations to $u(t_n)$ for some $n \geq 0$. By considering $F \equiv 0$ it then follows that

$$S\, S_0 = e\,, \tag{4.20}$$

where $e = (1, 1, \ldots, 1)^T \in \mathbb{R}^m$. Therefore, for any fixed m, the total scheme (4.19) is then just an $(m_0{+}m)$-stage Runge-Kutta method, with an $(m_0{+}m) \times (m_0{+}m)$ coefficient matrix

$$K = \begin{pmatrix} K_0 & O \\ S\, T_0 & T \end{pmatrix}. \tag{4.21}$$

To obtain monotonicity results we substitute $\gamma(v + \frac{\Delta t}{\gamma} \boldsymbol{F}(v)) - \gamma v$ for the terms $\Delta t \boldsymbol{F}(v)$ in (4.19). This gives, after a little manipulation,

$$\begin{cases} w = (\boldsymbol{I} + \gamma \boldsymbol{K_0})^{-1} \boldsymbol{e_0}\, u_0 + (\boldsymbol{I} + \gamma \boldsymbol{K_0})^{-1} \gamma \boldsymbol{K_0} \Big(w + \frac{\Delta t}{\gamma} \boldsymbol{F}(w) \Big)\,, \\ y = \boldsymbol{R}\, \boldsymbol{R_0} u_0 + \boldsymbol{R}\, \boldsymbol{P_0} \Big(w + \frac{\Delta t}{\gamma} \boldsymbol{F}(w) \Big) + \boldsymbol{P} \Big(y + \frac{\Delta t}{\gamma} \boldsymbol{F}(y) \Big)\,, \end{cases} \tag{4.22}$$

with matrices R, P as before and

$$R_0 = S_0 - \gamma T_0 (I + \gamma K_0)^{-1} e_0 \,, \qquad P_0 = \gamma T_0 (I + \gamma K_0)^{-1} \,. \tag{4.23}$$

These expressions arise in a natural way by writing $x = \boldsymbol{R_0} u_0 + \boldsymbol{P_0}\big(w + \frac{\Delta t}{\gamma} \boldsymbol{F}(w)\big)$, together with relation (4.12).

We will consider, for m arbitrarily large, and a given vector space $\mathbb{V}$ with convex functional $\|\cdot\|$, the following monotonicity property with stepsize coefficient $\gamma > 0$,

$$\left\{ \begin{array}{ll} \max\limits_{1 \le n \le m} ||y_n|| \le ||u_0|| & \text{whenever (4.2) is valid, } \Delta t \le \gamma \tau_0 \text{, and} \\ & x, y \text{ satisfy (4.10),(4.16), (4.18), } m \ge 1. \end{array} \right. \tag{4.24}$$

As we will see next, this type of monotonicity of the multistep methods with starting procedures will hold under the condition

$$R\, R_0 \ge 0 \,, \quad R\, P_0 \ge 0 \,, \quad P \ge 0 \qquad (\text{for all } m \ge 1) \,, \tag{4.25}$$

where R, P are defined by (4.13). The following result is similar to Theorem 3.4.4 of Chapter 3, where sufficiency of condition (4.25) was proven for nonnegative sublinear functionals.

Theorem 4.2.2. *Assume $I + \gamma T$ is not singular, $spr(|P|) < 1$, and the starting procedure is such that (4.17) holds. Let $\|\cdot\|$ be a convex functional on a vector space $\mathbb{V}$. Then (4.25) implies the monotonicity property (4.24). Moreover, if all rows of the matrix K in (4.21) are different from each other, then (4.25) is also necessary for this monotonicity property to hold for the class of spaces $\mathbb{V} = \mathbb{R}^M$, $M \ge 1$, with the maximum norm.*

Proof. Assume (4.17). Let $\eta = (\eta_i) \in \mathbb{R}^m$ with $\eta_i = \|y_i\|$. Since we have $\|w_j + \frac{\Delta t}{\gamma} F(w_j)\| \le \|w_j\| \le \|u_0\|$ for $1 \le j \le m_0$, it follows from the second equality in (4.22) that

$$\eta \;\le\; R\, R_0 \cdot \|u_0\| \;+\; R\, P_0\, e_0 \cdot \|u_0\| \;+\; P\, \eta \,.$$

In case $F \equiv 0$, all vectors w_j, y_i will be equal to u_0, from which it is seen that $e = R\, R_0 1 + R\, P_0 e_0 + P e$. Therefore,

$$(I - P)\eta \;\le\; (I - P)\, e \cdot \|u_0\| \,.$$

Since $\mathrm{spr}(P) < 1$, we have $(I - P)^{-1} = \sum_{j \ge 0} P^j \ge 0$, and therefore $\eta \le e \cdot \|u_0\|$, that is, $\|y_i\| \le \|u_0\|$ for $1 \le i \le m$.

If all rows of K are different from each other, the necessity of (4.25) follows from Spijker (2007), by considering (4.22) for fixed m as a step of a Runge-Kutta method with coefficient matrix K. □

Because (4.22) has the form of a Runge-Kutta method, sufficiency of (4.25) could −in principle− also have been derived from the results in Ferracina & Spijker (2005), Higueras (2005). If the coefficient matrix K in (4.21) has some identical rows, this Runge-Kutta method is reducible (in the sense of Spijker).

4.3 Formulations of the multistep methods

In order to apply the above results on boundedness and monotonicity we will formulate the multistep methods (4.3) and (4.4) in terms of input and output vectors, similar as in Chapter 3.

As before, $e_1, e_2, \ldots, e_m$ stand for the unit basis vectors in $\mathbb{R}^m$, and I is the $m \times m$ identity matrix. Further, $E = [e_2, \ldots, e_m, 0]$ will denote the $m \times m$ backward shift matrix, that is, all entries of E are zero except the entries on the first lower diagonal, which are 1.

Let $A, B \in \mathbb{R}^{m\times m}$ be defined by

$$A \;=\; \sum_{j=1}^{k} a_j E^j\,, \qquad B \;=\; \sum_{j=0}^{k} b_j E^j\,, \tag{4.26}$$

where $E^0 = I$. These are lower triangular Toeplitz matrices, with coefficients a_j, b_j on the j-th lower diagonal. For $m \geq k$ we also introduce $J = [e_1, \ldots, e_k] \in \mathbb{R}^{m\times k}$, containing the first k columns of the identity matrix I. To make the notations fitting for any $m \geq 1$, we define $J = [e_1, \ldots, e_m, O]$ for $1 \leq m < k$, with O being the $m \times (k-m)$ zero matrix. Finally, $A_0, B_0 \in \mathbb{R}^{k\times k}$ are given by

$$A_0 = \begin{pmatrix} a_k & \cdots & a_2 & a_1 \\ & a_k & & a_2 \\ & & \ddots & \vdots \\ & & & a_k \end{pmatrix}, \qquad B_0 = \begin{pmatrix} b_k & \cdots & b_2 & b_1 \\ & b_k & & b_2 \\ & & \ddots & \vdots \\ & & & b_k \end{pmatrix}. \tag{4.27}$$

4.3.1 Formulations of linear multistep methods with input vectors

The output vectors of the linear multistep scheme (4.3) are $y_n = u_{k-1+n}$, $n \geq 1$. The starting values $u_0, u_1, \ldots, u_{k-1}$ will enter the scheme in the first k steps in the combinations

$$x_l \;=\; \sum_{j=l}^{k} a_j u_{k-1+l-j} + \Delta t \sum_{j=l}^{k} b_j F(u_{k-1+l-j}) \qquad \text{for } 1 \leq l \leq k\,. \tag{4.28}$$

The multistep scheme (4.3) then can be written as

$$y_n \;=\; x_n + \sum_{j=1}^{n-1} a_j y_{n-j} + \Delta t \sum_{j=0}^{n-1} b_j F(y_{n-j}) \qquad \text{for } 1 \leq n \leq k\,, \tag{4.29a}$$

$$y_n \;=\; \sum_{j=1}^{k} a_j y_{n-j} + \Delta t \sum_{j=0}^{k} b_j F(y_{n-j}) \qquad \text{for } n > k\,, \tag{4.29b}$$

where the starting values are contained within the source terms in the first k steps. The vectors $x_1, \ldots, x_k \in \mathbb{V}$ are the input vectors for the scheme.

Consider m steps of the multistep scheme, $m \geq k$, leading to (4.29) with $n = 1, 2, \ldots, m$. The resulting scheme can be written as

$$y \;=\; \boldsymbol{J}x + \boldsymbol{A}y + \Delta t\, \boldsymbol{BF}(y)\,. \tag{4.30}$$

Clearly this is of the form (4.10) with

$$S = (I - A)^{-1}J\,, \qquad T = (I - A)^{-1}B\,, \tag{4.31}$$

which gives (4.12) with

$$R \;=\; (I - A + \gamma B)^{-1}J\,, \qquad P \;=\; (I - A + \gamma B)^{-1}\gamma B\,. \tag{4.32}$$

If we consider the problem (4.1) with $F \equiv \beta$ and solution $u(t) = \alpha + \beta t$, then exact starting values $u_j = u(t_j)$ $(0 \leq j < k)$ will give $u_n = u(t_n)$ (for all $n \geq k$) because of consistency of the methods. From this it is easily seen that (4.11) holds, and therefore the scheme is irreducible (in the sense of Spijker). It should be remarked that this is not directly related to the Dahlquist irreducibility condition (4.5b) for the multistep methods.

The matrix $I - A + \gamma B$ is invertible for any $\gamma > 0$, because $b_0 \geq 0$. The matrix P is again a lower triangular Toeplitz matrix, and it has the entry $\pi_0 = \gamma b_0/(1 + \gamma b_0) \in [0, 1)$ on the main diagonal. The spectral radius $\mathrm{spr}(|P|)$ of the matrix $|P|$ is therefore less than one.

The coefficients of the matrices R and P are easily found recursively. Let $\rho_j = 0$ for $j < 0$. If we set $(I - A + \gamma B)^{-1} = \sum_{n\geq 0} \rho_n E^n$ and $P = \sum_{n\geq 0} \pi_n E^n$, then these Toeplitz coefficients ρ_n, π_n are given by $\rho_0 = 1/(1 + \gamma b_0)$ and

$$\rho_n \;=\; \sum_{j=1}^{k} a_j \rho_{n-j} - \gamma \sum_{j=0}^{k} b_j \rho_{n-j} \qquad \text{for } n \geq 1\,, \tag{4.33a}$$

$$\pi_n \;=\; \gamma \sum_{j=0}^{k} b_j \rho_{n-j} \qquad \text{for } n \geq 0\,. \tag{4.33b}$$

An inequality of the type $Rv \geq 0$ for all $m \geq 1$, with a vector $v = (v_1, \ldots, v_k)^T$, is now equivalent to having $\sum_{j=1}^{k} v_j \rho_{n-j} \geq 0$ for all $n \geq 1$.

4.3.2 Formulation of one-leg methods with input vectors

To derive results for one-leg methods, we will proceed in a similar way, using an input-output formulation. To distinguish the arising vectors and associated matrices for the one-leg methods from those of the linear multistep methods, we will use the upper bar symbol for the one-leg vectors and matrices. In particular, the matrices S, T, R, P will be as in (4.31) and (4.32) for the linear multistep methods, and the corresponding matrices for the one-leg methods will be denoted by $\bar{S}$, $\bar{T}$, $\bar{R}$ and $\bar{P}$. Likewise, in the generic form (4.9), (4.10) the dimensions m, k will now read $\bar{m}$ and $\bar{k}$.

Consider m steps of the one-leg method (4.4), and let $\bar{y} = [\bar{y}_i] \in \mathbb{V}^{\bar{m}}$, $\bar{m} = 2m$, with

$$\bar{y}_i = u_{k-1+i}\,, \quad \bar{y}_{m+i} = v_{k-1+i} \qquad \text{for } 1 \le i \le m\,. \tag{4.34}$$

As input we have $\bar{x} = [\bar{x}_j] \in \mathbb{V}^{\bar{k}}$, $\bar{k} = 2k$, with

$$\bar{x}_i = \sum_{j=i}^{k} a_j u_{k-1+i-j}\,, \quad \bar{x}_{i+k} = \sum_{j=i}^{k} \hat{b}_j u_{k-1+i-j} \qquad i = 1, \ldots, k\,. \tag{4.35}$$

Let $J \in \mathbb{R}^{m\times k}$ and $A, B \in \mathbb{R}^{m\times m}$ be as before. Then the m steps of the one-leg method can be written as

$$\bar{y} = \bar{\boldsymbol{J}}\bar{x} + \bar{\boldsymbol{A}}\bar{y} + \Delta t \bar{\boldsymbol{B}}\boldsymbol{F}(\bar{y})\,, \tag{4.36}$$

where $\bar{J} \in \mathbb{R}^{\bar{m}\times\bar{k}}$ and $\bar{A}, \bar{B} \in \mathbb{R}^{\bar{m}\times\bar{m}}$ are given by

$$\bar{J} = \begin{pmatrix} J & 0 \\ 0 & J \end{pmatrix}, \quad \bar{A} = \begin{pmatrix} A & O \\ \frac{1}{\beta}B & O \end{pmatrix}, \quad \bar{B} = \begin{pmatrix} O & \beta I \\ O & O \end{pmatrix}, \tag{4.37}$$

with zero matrices $O \in \mathbb{R}^{m\times m}$ and $0 \in \mathbb{R}^{m\times k}$. We can rewrite (4.36) in the following form, comparable to (4.10),

$$\bar{y} \;=\; \bar{\boldsymbol{S}}\bar{x} + \Delta t\, \bar{\boldsymbol{T}}\boldsymbol{F}(\bar{y}) \tag{4.38}$$

with $\bar{S} \in \mathbb{R}^{\bar{m}\times\bar{k}}$ and $\bar{T} \in \mathbb{R}^{\bar{m}\times\bar{m}}$ defined by

$$\bar{S} = (\bar{I} - \bar{A})^{-1}\bar{J}\,, \qquad \bar{T} = (\bar{I} - \bar{A})^{-1}\bar{B}\,, \tag{4.39}$$

with $\bar{m} \times \bar{m}$ identity matrix $\bar{I}$. Working out these matrices, in terms of $S = (I - A)^{-1}J$ and $T = (I - A)^{-1}B$, gives

$$\bar{S} \;=\; \begin{pmatrix} S & 0 \\ \frac{1}{\beta}BS & J \end{pmatrix}, \qquad \bar{T} \;=\; \begin{pmatrix} O & \beta(I - A)^{-1} \\ O & T \end{pmatrix}. \tag{4.40}$$

This can be further rewritten, for example with $BS = TJ$.

If there is only one index j, $0 \le j \le k$, with $b_j \neq 0$, then the one-leg method is the same as the linear multistep method. For genuine one-leg methods, with $b_j \neq 0$ for at least two indices j, it will now be shown that the scheme is irreducible (in the sense of Spijker).

Lemma 4.3.1. *Suppose $b_j \neq 0$ for two or more indices $0 \le j \le k$. Then all rows of $[\bar{S}\ \bar{T}]$ are different from each other.*

Proof. The u_n, v_n are consistent approximations to $u(t_n)$, $u(\bar{t}_n)$, respectively, with $\bar{t}_n = \sum_{j=0}^{k} \hat{b}_j t_{n-j}$. Using the same arguments as for the linear multistep methods, it follows that the first m rows of $[\bar{S}\ \bar{T}]$ are different from each other, and the same holds for the last m rows.

It remains to show that none of the first m rows can be equal to any of the last m rows. For this it is sufficient to show that the $m \times (m+k)$ matrices

$$C_1 \;=\; \begin{pmatrix} 0 & \beta(I-A)^{-1} \end{pmatrix} \qquad \text{and} \qquad C_2 \;=\; \begin{pmatrix} J & T \end{pmatrix}$$

have no common rows. Because of the entry J in C_2, it is clear that the first k rows of C_2 cannot coincide with any of the rows of C_1.

Since the lower triangular Toeplitz matrices B and $I-A$ commute, we also have $T = B(I-A)^{-1}$. Equal rows of $\beta(I-A)^{-1}$ and T can therefore only happen if $\beta e_i^T (I-A)^{-1} = e_j^T B(I-A)^{-1}$ with $i \geq j$. But then $\beta e_i^T = e_j^T B$, which implies that either $i, j \leq k$ or that only the coefficient b_{i-j} is not zero.

Consequently, if the matrices C_1 and C_2 have common rows, then there is only one index j, $0 \leq j \leq k$, with $b_j \neq 0$. □

As for the linear multistep methods, it follows from consistency of the one-leg method that $\bar{S}$ has no zero rows. The conditions (4.11) are therefore fulfilled with the matrices $\bar{S}, \bar{T}$ instead of S, T, provided the one-leg method is not a linear multistep method.

Since $I + \gamma T$ is invertible for any $\gamma > 0$, the same holds for $I + \gamma\bar{T}$. From (4.36) we therefore obtain the transformed form, comparable to (4.12),

$$\bar{y} \;=\; \bar{\boldsymbol{R}}\bar{x} + \bar{\boldsymbol{P}}\Big(\bar{y} + \frac{\Delta t}{\gamma}\boldsymbol{F}(\bar{y})\Big)\,, \tag{4.41}$$

where $\bar{R} \in \mathbb{R}^{\bar{m}\times\bar{k}}$ and $\bar{P} \in \mathbb{R}^{\bar{m}\times\bar{m}}$ are given by

$$\bar{R} \;=\; (\bar{I} - \bar{A} + \gamma\bar{B})^{-1}\bar{J}\,, \qquad \bar{P} \;=\; (\bar{I} - \bar{A} + \gamma\bar{B})^{-1}\gamma\bar{B}\,. \tag{4.42}$$

These matrices $\bar{R}$ and $\bar{P}$ in (4.42) for the one-leg method will be expressed in terms of the $m \times m$ matrices R and P for the linear multistep method, as given by (4.32). Let us here denote $L = (I - A + \gamma B)^{-1}$. Then it is found that

$$(\bar{I} - \bar{A} + \gamma\bar{B})^{-1} \;=\; \begin{pmatrix} I-A & \beta\gamma I \\ -\frac{1}{\beta}B & I \end{pmatrix}^{-1} \;=\; \begin{pmatrix} L & -\beta\gamma L \\ \frac{1}{\beta}BL & (I-A)L \end{pmatrix}.$$

The blocks consist of products of Toeplitz matrices that commute. Since $LJ = R$, it follows that

$$\bar{R} \;=\; \begin{pmatrix} R & -\beta\gamma R \\ \frac{1}{\beta}BR & (I-A)R \end{pmatrix}, \qquad \bar{P} \;=\; \begin{pmatrix} O & \beta\gamma L \\ O & P \end{pmatrix}. \tag{4.43}$$

Note that since $\mathrm{spr}(|P|) < 1$, we also have $\mathrm{spr}(|\bar{P}|) < 1$. Furthermore we see that $\bar{P} \geq 0$ iff $P \geq 0$ and $R \geq 0$.

4.4 Boundedness for arbitrary starting vectors

In this section conditions are given for boundedness of the multistep methods (4.3) and (4.4). It will always be assumed that (4.5) is satisfied. Furthermore, in this section, boundedness is understood in the the sense of property (4.14) for any seminorm, with some $\mu \geq 1$, and with y_i, x_j, m, k replaced by $\bar{y}_i, \bar{x}_j, \bar{m}$ and $\bar{k}$, respectively, for the one-leg methods.

To formulate the results we will use some standard linear stability concepts for multistep methods, as given in Butcher (2003) Hairer, Nørsett & Wanner (1993), Hairer & Wanner (1996) for instance. We denote the *stability region* of the methods by $\mathcal{S}$, and its interior by $\mathrm{int}(\mathcal{S})$. If $0 \in \mathcal{S}$ the method is said to be *zero-stable*.

It was shown in Chapter 3 that for a zero-stable linear multistep method satisfying (4.5), the condition

$$-\gamma \in \mathrm{int}(\mathcal{S}), \qquad P \geq 0 \quad (\text{for all } m). \tag{4.44}$$

is necessary and sufficient for the boundedness property (4.14) to hold with some $\mu \geq 1$.

As we will see, the same result is valid for the one-leg methods. This can be shown using relations between a linear multistep method and the corresponding one-leg method, as given in Dahlquist (1975) or Hairer & Wanner (1996). It is also possible to prove this from Theorem 4.2.1, which will be done here.

For this, we consider the matrix

$$\bar{M} \;=\; (\bar{I} - |\bar{P}|)^{-1}|\bar{R}| \;=\; \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix}. \tag{4.45}$$

Using the fact that $(I - A)R = (I - P)J$ and $BR = \frac{1}{\gamma}PJ$, it follows by some calculations that the blocks $M_{ij} \in \mathbb{R}^{m \times k}$ can be written as

$$\begin{cases} M_{11} \;=\; (I - |P|)^{-1}|R|\,, \\ M_{12} \;=\; \beta\gamma\left((I - |P|)^{-1}|I - P| + I\right)|R|\,, \\ M_{21} \;=\; \frac{1}{\beta\gamma}\,(I - |P|)^{-1}|P|\,J\,, \\ M_{22} \;=\; (I - |P|)^{-1}|I - P|\,J\,. \end{cases} \tag{4.46}$$

According to Theorem 4.2.1, boundedness of the one-leg method is equivalent to having a bound $\|\bar{M}\|_\infty \leq \mu$ uniformly for $m \geq 1$. By considering the M_{11} block, we therefore see that boundedness of the one-leg method implies boundedness of the linear multistep method.

On the other hand, suppose the linear multistep method is bounded. Then we know that $P \geq 0$. Zero-stability implies that $\|S\|_\infty$ is bounded uniformly in m. Therefore, the maximum norms of the matrices R and

$$(I - |P|)^{-1}J = (I - A + \gamma B)S$$

are also bounded uniformly in m. Using the relations $(I-P)^{-1}P = (I-P)^{-1}-I$ and $(I-P)^{-1}|I-P| \leq (I-P)^{-1}(I+P)$, it follows that the maximum norms of all the blocks M_{ij} in (4.46) are bounded uniformly in m.

In conclusion, we have the following result on boundedness for our multistep methods (4.3) and (4.4).

Theorem 4.4.1. *Consider a one-leg or linear multistep method, satisfying (4.5). Assume the method is zero-stable, and let $\gamma > 0$. Then there is a $\mu \geq 1$ such that the boundedness property (4.14) is valid for any vectorspace $\mathbb{V}$ and seminorm $\|\cdot\|$ if and only if condition (4.44) holds.*

The above result, with equal stepsize coefficients for a one-leg method and its linear multistep counterpart, is hardly surprising, given the close connection between one-leg methods and linear multistep methods. We will see, however, that the allowable stepsizes for one-leg and linear multistep methods can be very different if we require monotonicity with starting procedures.

The boundedness property (4.14) is expressed in terms of the input vectors x_j. However, with seminorms this is easily translated into boundedness with respect to the starting values $u_0, \ldots, u_{k-1}$ as in (4.6), and likewise reversely; see also Section 3.4.2 of Chapter 3.

Remark 4.4.2. As observed before, we have $\bar{P} \geq 0$ iff $P \geq 0$ and $R \geq 0$. This is slightly stronger than having only $P \geq 0$. As a consequence, boundedness of the one-leg method does *not* require $\bar{P} \geq 0$. $\diamond$

4.5 Monotonicity with starting procedures

4.5.1 Linear multistep methods with starting procedures

Consider the Runge-Kutta starting procedure (4.16), producing the vector $w = [w_i] \in \mathbb{V}^{m_0}$. Let $J_0 \in \mathbb{R}^{k \times m_0}$ be the matrix, with columns $e_1, \ldots, e_k \in \mathbb{R}^k$ interceded by zero columns, which selects those components w_i that correspond to a starting value u_j of the multistep method, that is,

$$\boldsymbol{J_0} w = (u_0, \ldots, u_{k-1})^T\,, \qquad \boldsymbol{J_0 F}(w) = (F(u_0), \ldots, F(u_{k-1}))^T\,.$$

Further, let A_0, B_0 be as in (4.27). Then it follows from (4.28) that

$$x \;=\; \boldsymbol{A_0 J_0} w + \Delta t \boldsymbol{B_0 J_0 F}(w)\,. \tag{4.47}$$

This gives the representation $x = \boldsymbol{S_0} u_0 + \Delta t \boldsymbol{T_0 F}(w)$, as in (4.18), with matrices $S_0 \in \mathbb{R}^{k \times 1}$, $T_0 \in \mathbb{R}^{k \times m_0}$ given by

$$S_0 = A_0 J_0 e_0\,, \qquad T_0 = A_0 J_0 K_0 + B_0 J_0\,. \tag{4.48}$$

To obtain monotonicity results, the scheme is now written in the form (4.22) with matrices R, P given by (4.32) and with $R_0 \in \mathbb{R}^{k \times 1}$, $P_0 \in \mathbb{R}^{k \times m_0}$ given by

$$R_0 = A_0 J_0 e_0 - P_0 e_0\,, \qquad P_0 = \gamma(A_0 J_0 K_0 + B_0 J_0)(I + \gamma K_0)^{-1}\,. \tag{4.49}$$

These matrices R_0, P_0 can be further rewritten as

$$R_0 = (A_0 - \gamma B_0) J_0 (I + \gamma K_0)^{-1} e_0 \,, \tag{4.50a}$$

$$P_0 = (A_0 - \gamma B_0) J_0 (I + \gamma K_0)^{-1} \gamma K_0 + \gamma B_0 J_0 \,. \tag{4.50b}$$

Consequently, the conditions (4.25) for monotonicity of the total scheme are:

$$\begin{cases} P \,\geq\, 0 \,, \\ R\,(A_0 - \gamma B_0) J_0 (I + \gamma K_0)^{-1} e_0 \,\geq\, 0 \,, \\ R\,\big((A_0 - \gamma B_0) J_0 (I + \gamma K_0)^{-1} \gamma K_0 + \gamma B_0 J_0\big) \,\geq\, 0 \,. \end{cases} \tag{4.51}$$

The first inequality, $P \geq 0$, is the essential condition for boundedness of the linear multistep method.

4.5.2 One-leg methods with starting procedures

As for the linear multistep methods, we now consider the formulas that are obtained if a Runge-Kutta starting procedure is used for a one-leg method. It is assumed, as before, that this starting procedure is of the form $w = \boldsymbol{e_0} u_0 + \Delta t \boldsymbol{K_0} \boldsymbol{F}(w)$ and $\boldsymbol{J_0} w = (u_0, \ldots, u_{k-1})^T$. From (4.35) it is then seen that

$$\bar{x} = \begin{pmatrix} \boldsymbol{A_0} \\ \frac{1}{\beta} \boldsymbol{B_0} \end{pmatrix} \boldsymbol{J_0} w \;=\; \begin{pmatrix} \boldsymbol{A_0} \\ \frac{1}{\beta} \boldsymbol{B_0} \end{pmatrix} \boldsymbol{J_0} \big(\boldsymbol{e_0} u_0 + \Delta t \boldsymbol{K_0} \boldsymbol{F}(w) \big) \,.$$

This can be written as

$$\bar{x} \;=\; \bar{\boldsymbol{S}}_{\boldsymbol{0}} u_0 + \Delta t \bar{\boldsymbol{T}}_{\boldsymbol{0}} \boldsymbol{F}(w) \,, \tag{4.52}$$

with

$$\bar{S}_0 = \begin{pmatrix} A_0 \\ \frac{1}{\beta} B_0 \end{pmatrix} J_0 e_0 \,, \qquad \bar{T}_0 \;=\; \begin{pmatrix} A_0 \\ \frac{1}{\beta} B_0 \end{pmatrix} J_0 K_0 \,. \tag{4.53}$$

As before, for any fixed number of steps m, the total scheme is an $(m_0{+}\bar{m})$-stage Runge-Kutta method, $\bar{m} = 2m$, with an $(m_0{+}\bar{m}) \times (m_0{+}\bar{m})$ coefficient matrix

$$\bar{K} \;=\; \begin{pmatrix} K_0 & O \\ \bar{S}\,\bar{T}_0 & \bar{T} \end{pmatrix} . \tag{4.54}$$

To derive monotonicity results we substitute, as before, $\gamma(v + \frac{\Delta t}{\gamma} \boldsymbol{F}(v)) - \gamma v$ for all terms $\Delta t \boldsymbol{F}(v)$. This gives, as in (4.22), the total scheme in the form

$$\begin{cases} w = (I + \gamma \boldsymbol{K_0})^{-1} \boldsymbol{e_0}\, u_0 + (I + \gamma \boldsymbol{K_0})^{-1} \gamma \boldsymbol{K_0} \Big(w + \dfrac{\Delta t}{\gamma} \boldsymbol{F}(w) \Big) \,, \\ \bar{y} = \bar{\boldsymbol{R}}\, \bar{\boldsymbol{R}}_{\boldsymbol{0}} u_0 + \bar{\boldsymbol{R}}\, \bar{\boldsymbol{P}}_{\boldsymbol{0}} \Big(w + \dfrac{\Delta t}{\gamma} \boldsymbol{F}(w) \Big) + \bar{\boldsymbol{P}} \Big(\bar{y} + \dfrac{\Delta t}{\gamma} \boldsymbol{F}(\bar{y}) \Big) \,, \end{cases}$$

with matrices $\bar{R}, \bar{P}$ given by (4.43) and

$$\bar{R}_0 = \bar{S}_0 - \bar{P}_0 e_0\,, \qquad \bar{P}_0 = \gamma \bar{T}_0 (I + \gamma K_0)^{-1}\,. \tag{4.55}$$

Inserting the expressions (4.53) into (4.55) gives

$$\bar{R}_0 = \begin{pmatrix} A_0 \\ \frac{1}{\beta} B_0 \end{pmatrix} J_0 (I + \gamma K_0)^{-1} e_0\,, \quad \bar{P}_0 = \begin{pmatrix} A_0 \\ \frac{1}{\beta} B_0 \end{pmatrix} J_0 (I + \gamma K_0)^{-1} \gamma K_0\,. \tag{4.56}$$

To compare the occurring matrices in the monotonicity requirement for the one-leg method with those of the linear multistep method, we note that

$$\bar{R} \begin{pmatrix} A_0 \\ \frac{1}{\beta} B_0 \end{pmatrix} = \begin{pmatrix} R(A_0 - \gamma B_0) \\ \frac{1}{\beta}(BRA_0 + (I - A) R B_0) \end{pmatrix} == \begin{pmatrix} R(A_0 - \gamma B_0) \\ \frac{1}{\beta\gamma}(PJ(A_0 - \gamma B_0) + \gamma J B_0) \end{pmatrix}.$$

The conditions for monotonicity of the total scheme, $\bar{R}\bar{R}_0 \geq 0$, $\bar{R}\bar{P}_0 \geq 0$ and $\bar{P} \geq 0$, therefore read

$$\begin{cases} P \geq 0\,, \qquad R \geq 0\,, \\ R\,(A_0 - \gamma B_0) J_0 (I + \gamma K_0)^{-1} e_0 \geq 0\,, \\ R\,(A_0 - \gamma B_0) J_0 (I + \gamma K_0)^{-1} \gamma K_0 \geq 0\,, \\ \big(PJ(A_0 - \gamma B_0) + \gamma J B_0\big)\, J_0 (I + \gamma K_0)^{-1} e_0 \geq 0\,, \\ \big(PJ(A_0 - \gamma B_0) + \gamma J B_0\big)\, J_0 (I + \gamma K_0)^{-1} \gamma K_0 \geq 0\,. \end{cases} \tag{4.57}$$

Although the one-leg methods give the same stepsize coefficients for boundedness as the corresponding linear multistep methods, this is no longer so if monotonicity with starting procedures is considered.

4.6 Application for explicit two-step methods

As an application of the general formulas derived in the previous section, we will give here detailed results for explicit two-step methods of order one. With this class of methods we can take a_1, b_1 as free parameters, and set $a_2 = 1 - a_1$, $b_2 = 2 - a_1 - b_1$. The methods have order two if $b_1 = 2 - \frac{1}{2} a_1$, and they are zero-stable if $0 \leq a_1 < 2$. The methods with $b_1 = 1$ or $a_1 = 2$ do not satisfy the Dahlquist irreducibility condition (4.5b). Furthermore, if $b_1 = 0$ or $b_2 = 0$, then the one-leg methods coincide with the corresponding linear multistep methods. It will therefore be assumed for the one-leg methods that $b_1 \neq 0$ and $b_2 \neq 0$.

For this family of methods, with free parameters a_1, b_1, we will display in contour plots the maximal values of γ such that we have monotonicity or boundedness with arbitrary starting vectors (for seminorms) or monotonicity with

starting procedures (for convex functionals). These maximal stepsize coefficients will be called *threshold values*.

In these plots, $b_1 = 1$ is a special case: starting with forward Euler, the whole linear two-step scheme reduces to an application of the forward Euler method, so then we have monotonicity with $\gamma = 1$. Furthermore, for the one-leg methods $a_1 + b_1 = 2$ is also a special case: we then have $b_2 = 0$, so the one-leg method is then a linear multistep method, written in a reducible form.

In Figure 4.1 (left picture), the maximal values of γ are shown for which we have monotonicity with arbitrary starting vectors. These values are obtained from condition (4.7). For the 'white' areas in the contour plot, there is no positive γ.

The threshold values for boundedness are shown in Figure 4.1 (right picture). These thresholds, which are the same for the one-leg and linear multistep methods, coincide with those in Figure 4.4 of Chapter 3. We see that for boundedness the area of nonzero thresholds is much larger than for monotonicity and it includes many interesting methods, for example the second order methods with $b_1 = 2 - \frac{1}{2}a_1$.

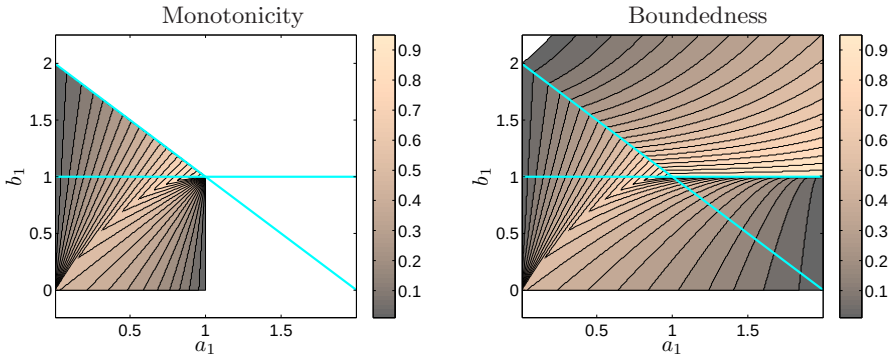

FIGURE 4.1: Explicit two-step methods of order one, with parameters $a_1 \in [0, 2)$ horizontally and $b_1 \in [-\frac{1}{4}, \frac{9}{4}]$, $b_1 \neq 1$, vertically. The contour plot shows the optimal $\gamma > 0$ for monotonicity (left picture) and for boundedness (right picture). The contour levels are at $j/20$, $j = 0, 1, \ldots$; for the 'white' areas, there is no positive γ.

The theshold values for boundedness have been found numerically, verifying the condition $P \geq 0$ with $m = 1000$. Inspection with larger m showed that the results do not differ anymore visually; in fact, for most methods a much smaller value of m would have been sufficient. This condition for boundedness, as well as the conditions for monotonicity –such as (4.51) and the related conditions (4.57) for the one-leg methods– were verified in the same way, using the recursive formulas (4.33) for the coefficients of the relevant Toeplitz matrices.

4.6.1 Starting procedure: the explicit Euler method

Consider explicit two-step methods, and suppose u_1 is computed by the forward Euler method,

$$u_1 = u_0 + \Delta t F(u_0)\,. \tag{4.58}$$

This is of the form (4.16) with $m_0 = 2$, $w = (w_1, w_2)^T = (u_0, u_1) \in \mathbb{V}^2$ and we get

$$K_0 = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \qquad J_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{4.59}$$

For the linear multistep methods we then obtain from (4.50),

$$R_0 = \begin{pmatrix} c_2+(1-\gamma)c_1 \\ (1-\gamma)c_2 \end{pmatrix}, \qquad P_0 = \begin{pmatrix} \gamma c_1+\gamma b_2 & \gamma b_1 \\ \gamma c_2 & \gamma b_2 \end{pmatrix}, \tag{4.60}$$

with $c_j = a_j - \gamma b_j$ for $j = 1, 2$. For the one-leg methods this Euler starting procedure leads to (4.56) with

$$\bar{R}_0 = \begin{pmatrix} a_2+(1-\gamma)a_1 \\ (1-\gamma)a_2 \\ \hat{b}_2+(1-\gamma)\hat{b}_1 \\ (1-\gamma)\hat{b}_2 \end{pmatrix}, \qquad \bar{P}_0 = \begin{pmatrix} \gamma a_1 & 0 \\ \gamma a_2 & 0 \\ \gamma\hat{b}_1 & 0 \\ \gamma\hat{b}_2 & 0 \end{pmatrix}. \tag{4.61}$$

The total schemes with the linear two-step methods are irreducible (in the sense of Spijker) because all u_n are consistent approximations to $u(t)$ at different time levels. The combinations of the two-step one-leg methods and the forward Euler starting procedure are also irreducible (in the sense of Spijker) if $b_j \neq 0$ for $j = 1, 2$. To show this we consider such a total scheme, written as one step of a big Runge-Kutta method with coefficient matrix

$$\bar{K} = \begin{pmatrix} K_0 & O \\ \bar{S}\bar{T}_0 & \bar{T} \end{pmatrix}$$

where $\bar{S} \in \mathbb{R}^{\bar{m}\times 4}$, $\bar{T}_0 \in \mathbb{R}^{4\times 2}$ and $\bar{T} \in \mathbb{R}^{\bar{m}\times\bar{m}}$ are given by

$$\bar{S} = \begin{pmatrix} (I-A)^{-1}J & 0 \\ \frac{1}{\beta}B(I-A)^{-1}J & J \end{pmatrix}, \quad \bar{T}_0 = \begin{pmatrix} a_1 & 0 \\ a_2 & 0 \\ \hat{b}_1 & 0 \\ \hat{b}_2 & 0 \end{pmatrix}, \quad \bar{T} = \begin{pmatrix} O & \beta(I-A)^{-1} \\ O & B(I-A)^{-1} \end{pmatrix}.$$

It is clear that the matrix $\beta(I-A)^{-1}$ has no zero row. The first row of the lower triangular Toeplitz matrix $B(I-A)^{-1}$ is the only zero row of that matrix. The first row of $(\frac{1}{\beta}B(I-A)^{-1}J\ J)\bar{T}_0$ is $(\hat{b}_1\ 0)$, and since $b_1 \neq 0$, $b_2 \neq 0$, it is seen

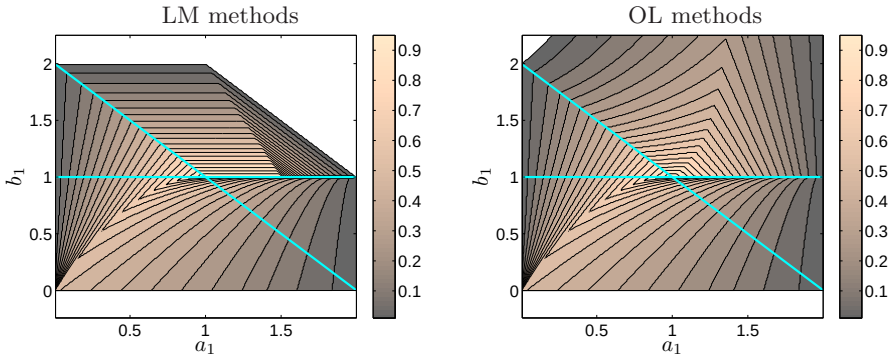

FIGURE 4.2: Optimal $\gamma > 0$ for monotonicity of the explicit two-step methods with explicit Euler starting method. Left picture: linear multistep methods; right picture: one-leg methods. Explanations as in Figure 4.1

that this cannot be equal to any row of K_0. Therefore the coefficient matrix $\bar{K}$ of the total scheme has no equal rows.

In Figure 4.2 the maximal values of γ are shown for which we have monotonicity with the forward Euler starting procedure for the explicit linear two-step methods (left picture) and the explicit one-leg methods (right picture). From this figure we conclude that the monotonicity properties with forward Euler starting procedure are better for these explicit one-leg methods than for the corresponding linear multistep methods.

4.6.2 An example on the relevance of irreducibility

The general conditions (4.25) for monotonicity are always sufficient. If the total scheme is reducible (in the sense of Spijker), that is, if some rows of the matrix K in (4.21) are equal, then these conditions (4.25) may not be necessary. As an illustration for this, consider the linear two-step methods of order one with the following starting procedures:

$$w_1 = u_0\,, \qquad w_2 = u_0 + \Delta t F(w_1)\,, \tag{4.62a}$$

with $u_1 = w_2$, and

$$w_1 = u_0\,, \qquad w_2 = u_0\,, \qquad w_3 = u_0 + \frac{\Delta t}{2}(F(w_1) + F(w_2))\,, \tag{4.62b}$$

with $u_1 = w_3$. Both these starting procedures are just the explicit Euler method, but the second form is reducible, with a redundant second stage. Even though the starting schemes are essentially equivalent and the corresponding input vectors x_j are the same, the matrices R_0 and P_0 will differ, leading to different results with (4.25).

The largest $\gamma \geq 0$ for which (4.25) holds is displayed in Figure 4.3. It is clear that, in comparison to (4.62a), these values γ are often smaller for the reducible procedure (4.62b). Note that for the irreducible case (4.62a) the values γ are necessary and sufficient; in the reducible case these values are only sufficient.

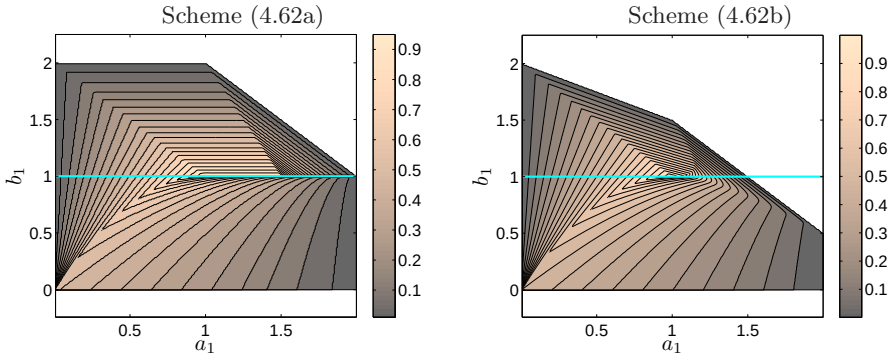

FIGURE 4.3: Largest $\gamma > 0$ for the monotonicity conditions (4.25) with the linear two-step methods and irreducible or reducible Euler start. Left picture: irreducible form (4.62a); right picture: reducible form (4.62b). Explanations as in Figure 4.1

4.6.3 Starting procedure: the explicit trapezoidal rule

Now suppose that u_1 is computed by the explicit trapezoidal rule, also known as the modified Euler method,

$$v_1 = u_0 + \Delta t F(u_0)\,, \quad u_1 = u_0 + \tfrac{1}{2}\Delta t F(u_0) + \tfrac{1}{2}\Delta t F(v_1)\,. \tag{4.63}$$

This fits in our general form with $m_0 = 3$, $w = (w_1, w_2, w_3)^T = (u_0, v_1, u_1)^T$ and

$$K_0 = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}, \qquad J_0 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{4.64}$$

Here we have $\|w_j\| \leq \|u_0\|$ $(1 \leq j \leq m_0)$, whenever (4.2) is valid and $\Delta t \leq \tau_0$.

For the linear multistep method this gives, as in Chapter 3,

$$R_0 = \begin{pmatrix} c_2 + c_1 r_0 \\ c_2 r_0 \end{pmatrix}, \qquad P_0 = \begin{pmatrix} c_1 q_0 + \gamma b_2 & c_1 q_1 & \gamma b_1 \\ c_2 q_0 & c_2 q_1 & \gamma b_2 \end{pmatrix}, \tag{4.65}$$

with $c_j = a_j - \gamma b_j$ for $j = 1, 2$, and $r_0 = 1 - \gamma + \frac{1}{2}\gamma^2$, $q_0 = \frac{1}{2}\gamma(1-\gamma)$, $q_1 = \frac{1}{2}\gamma$.

For the one-leg methods we obtain, with the same r_0, q_0, q_1, the formulas

$$\bar{R}_0 = \begin{pmatrix} a_2 + a_1 r_0 \\ a_2 r_0 \\ \hat{b}_2 + \hat{b}_1 r_0 \\ \hat{b}_2 r_0 \end{pmatrix}, \qquad \bar{P}_0 = \begin{pmatrix} a_1 q_0 & a_1 q_1 & 0 \\ a_2 q_0 & a_2 q_1 & 0 \\ \hat{b}_1 q_0 & \hat{b}_1 q_1 & 0 \\ \hat{b}_2 q_0 & \hat{b}_2 q_1 & 0 \end{pmatrix}. \tag{4.66}$$

In the same way as with the explicit Euler starting procedure, it can be verified that the total schemes are irreducible (in the sense of Spijker), under the assumption $b_j \neq 0$ $(j = 1, 2)$ for the one-leg methods.

The maximal values of $\gamma > 0$ for monotonicity with this explicit trapezoidal rule starting procedure are shown in Figure 4.4; in the left picture for the linear two-step methods, and in the right picture for the one-leg methods.

For the linear multistep methods monotonicity with the explicit trapezoidal rule as starting procedure leads to monotonicity thresholds that are less than or equal to those with the forward Euler method. In fact the results are quite close – but not identical – to those with the reducible procedure (4.62b).

The one-leg methods with the explicit trapezoidal rule as starting method give here almost the same thresholds as with the forward Euler method, except for a parameter region with $a_1, b_1 > 1$. There the thresholds are somewhat improved with the explicit trapezoidal rule. This is in marked contrast to the situation for the corresponding linear multistep methods, where the thresholds deteriorate with this starting method in comparison with the explicit Euler start.

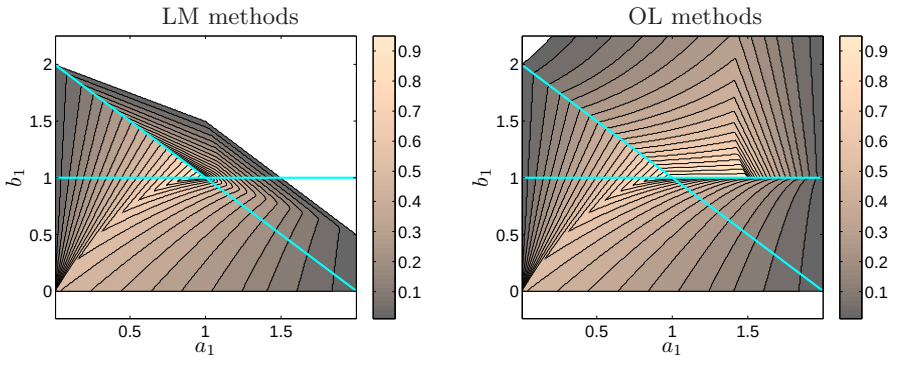

FIGURE 4.4: Optimal $\gamma > 0$ for monotonicity of the explicit two-step methods with with the explicit trapezoidal rule as starting method. Left picture: linear multistep methods; right picture: one-leg methods. Explanations as in Figure 4.1

4.6.4 Explicit two-step methods of order two

The most interesting explicit two-step methods are of course those with order two, $b_1 = 2 - \frac{1}{2}a_1$. This is a one-parameter family with a_1 as free parameter. Here a more clear picture is provided in Figure 4.5, where thresholds are plotted for boundedness and monotonicity with forward Euler and the explicit trapezoidal rule as starting methods for the linear two-step methods and the corresponding one-leg methods.

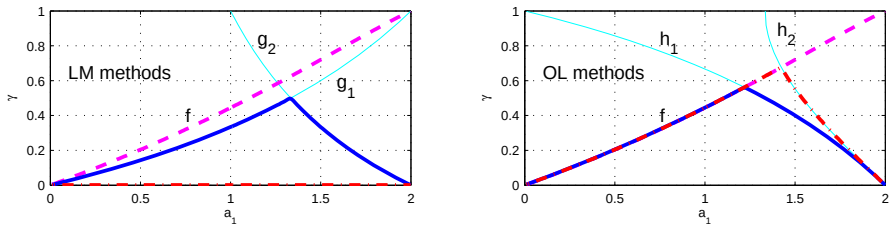

FIGURE 4.5: Explicit two-step methods of order two, with parameter $a_1 \in [0, 2)$. Vertical axis: thresholds for boundedness or monotonicity. Left picture: linear multistep methods; right picture: one-leg methods. Dashed line: boundedness; solid lines: monotonicity with forward Euler start; dash-dotted lines: monotonicity with explicit trapezoidal rule as starting method (with $\gamma = 0$ for the linear two-step methods).

The curves in these figures describing the thresholds are actually quite simple. The condition for boundedness is $\gamma \le f(a_1)$ with $f(z) = 2z(3-z)/(4-z)^2$; this value was shown in Hundsdorfer, Ruuth & Spiteri (2003) to be sufficient and from the requirement $\pi_2 = \gamma(a_1 b_1 + a_2) - \gamma^2 b_1^2 \ge 0$, it directly follows that it is also necessary.

For the linear multistep methods with forward Euler start it can be shown, in the same way as in Pham Thi, Hundsdorfer & Sommeijer (2006), that a sufficient condition for monotonicity is given by $\gamma \le \min\{g_1(a_1), g_2(a_1)\}$ with $g_1(z) = z/(4-z)$, $g_2(z) = (2-z)/z$; see also results in Hundsdorfer, Ruuth & Spiteri (2003) for $a_1 \le \frac{4}{3}$. We see in the figure that these sufficient conditions for monotonicity are also necessary. With an explicit trapezoidal rule start there is no positive threshold for monotonicity; this can be shown by considering the Runge-Kutta method that arises with $m = 1$, giving the coefficient matrix

$$K = \begin{pmatrix} 0 & & & \\ 1 & 0 & & \\ \frac{1}{2} & \frac{1}{2} & 0 & \\ \beta_1 & \beta_2 & \beta_3 & 0 \end{pmatrix} \tag{4.67}$$

with $\beta_1 = b_2 + \frac{1}{2}a_1$, $\beta_2 = \frac{1}{2}a_1$ and $\beta_3 = b_1$ being the weights of this explicit

3-stage method. The conditions for order two imply that $\beta_1 = 0$, and using Thoerem 4.2 of Kraaijevanger (1991) it can be shown that this Runge-Kutta method does not have a positive stepsize coefficient for monotonicity.

For the one-leg methods with forward Euler starting procedure we have $\gamma \leq \min\{f(a_1), h_1(a_1)\}$ as sufficient condition with $h_1(z) = 2(2-z)/(4-z)$. This follows from the results in Pham Thi et al. (2006) on positivity. With an explicit trapezoidal rule start the sufficient condition becomes $\gamma \leq \min\{f(a_1), h_2(a_1)\}$ with $h_2(z) = 1 - \sqrt{(3z-4)/(4-z)}$. This sufficient condition can be derived in a similar way as in Pham Thi et al. (2006). Again, we see from Figure 4.5 that these sufficient conditions for monotonicity are necessary as well. This necessity can be proven by considering the Runge-Kutta methods that arise with $m = 1$ and $m = 2$, that is, after one or two steps of the one-leg method with these starting procedures.

4.7 Concluding remarks

In view of the reduced storage requirements, compared to linear multistep methods, one-leg methods are interesting for large-scale computations. In this chapter results have been presented for having boundedness with arbitrary starting values, as well as for monotonicity with Runge-Kutta starting procedures.

It was seen that the stepsize restriction $\Delta t \leq \gamma\tau_0$ for the boundedness property (4.6) with seminorms is the same for a one-leg method as for the associated linear multistep method. However, differences between one-leg methods and linear multistep methods arise when we consider the monotonicity property (4.8) with starting procedures.

For explicit two-step methods it is seen that the monotonicity properties with standard starting procedures are better for the one-leg methods than for the linear multistep methods. However, no general conclusions are to be drawn from this. For the implicit two-step methods of order two it was found that the requirements for monotonicity, starting with backward Euler or with the θ-method, $\theta = b_0$, were not always better for the one-leg methods than for the corresponding linear multistep methods. These implicit methods turn out to have thresholds less than or equal to two. Since this is not very much larger than for the explicit methods, the implicit methods are not recommended if monotonicity is important, and therefore the results for these implicit methods have not been discussed here in detail.

Numerical tests were performed for scalar conservation laws $u_t + f(u)_x = 0$ in one spatial dimension with $f(u) = u$ (linear advection) and $f(u) = \frac{1}{2}u^2$ (Burgers equation). Spatial discretization was done with a van Leer type flux-limited scheme, for which it is known that the resulting ODE system satisfies the basic assumption (4.2) with τ_0 proportional to the meshwidth in space Δx. However, no significant difference was found in these tests between the explicit two-step one-leg methods and the corresponding linear multistep methods with the various starting procedures, even though the theoretical properties of the

one-leg methods are more favourable than those of the linear multistep methods.

Therefore, the practical relevance of the differences between one-leg and linear multistep methods found in this chapter are not yet fully established. On the other hand, for practical computations the theoretical findings do give a foundation for the explicit two-step one-leg methods with explicit trapezoidal start that is more solid than for the linear multistep methods.

Bibliography

[1] Bolley C. & Crouzeix M. (1978). *Conservation de la positivité lors de la discrétisation des problèmes d'évolution paraboliques*, RAIRO Anal. Numer. 12, 237–245.

[2] Butcher J.C. (1966). *On the convergence of numerical solutions to ordinary differential equations*. Math. Comp. 20, 1-10.

[3] Butcher J.C. (1987). *The numerical analysis of ordinary differential equations*. Wiley.

[4] Butcher J.C. (2003). *Numerical methods for ordinary differential equations*. Wiley.

[5] Crouzeix M. & Raviart P.-A. (1980). *Approximation des Problèmes d'Évolution*. Lecture Notes, University Rennes.

[6] Dahlquist G. (1975). *Error analysis for a class of methods for stiff nonlinear initial value problems*, Procs. Dundee Conf., Lecture Notes in Math. 506, G.A. Watson (ed.), Springer, 60–74.

[7] Ferracina L. & Spijker M.N. (2004). *Stepsize restrictions for the total-variation-diminishing property in general Runge-Kutta methods*, SIAM J. Numer. Anal. 42, 1073–1093.

[8] Ferracina L. & Spijker M.N. (2005). *An extension and analysis of the Shu-Osher representation of Runge-Kutta methods*, Math. Comp. 74, 201–219.

[9] Gottlieb S., Ketcheson D.I. & Shu C.-W. (2009). *High order strong stability preserving time discretizations*, J. Sci. Comput. 38, 251–289.

[10] Gottlieb S., Ketcheson D.I. & Shu C.-W. (2011). *Strong stability preserving Runge-Kutta and multistep time discretizations*, World Scientific Publishing Co. Pte. Ltd.

[11] Gottlieb S. & Shu C.-W. (1998). *Total-variation-diminishing Runge-Kutta schemes*, Math. Comp. 67 73–85.

[12] Gottlieb S., Shu C.-W. & Tadmor E. (2001). *Strong stability preserving high-order time discretization methods*, SIAM Review 43, 89–112.

[13] Hairer E., Nørsett S.P. & Wanner G. (1993). *Solving ordinary differential equations I – Nonstiff problems*. Second edition, Springer Series Comput. Math. 8, Springer.

[14] Hairer E. & Wanner G. (1996). *Solving ordinary differential equations II – Stiff and differential-algebraic problems*. Second edition, Springer Series in Comput. Math. 14, Springer.

[15] Harten E. (1983). *High resolution schemes for hyperbolic conservation laws*. J. Comput. Phys. 49, 357–393.

[16] Harten E., Hyman J.M. & Lax P.D. (1976). *On finite-difference approximations and entropy conditions for shocks*. with appendix by B. Keyfitz. Comput. Pure Appl. Math. 29, 297–322.

[17] Higueras I. (2004). *On strong stability preserving time discretization methods*, J. Sci. Comput. 21, 193–223.

[18] Higueras I. (2005). *Representations of Runge-Kutta methods and strong stability preserving methods*, SIAM J. Numer. Anal. 43, 924–948.

[19] Horn R.A. & Johnson C.R. (1998). *Matrix analysis*, Cambridge University Press, Cambridge.

[20] Horvath Z. (1998). *Positivity of Runge-Kutta and diagonally split Runge-Kutta methods*, Appl. Numer. Math. 28, 309–326.

[21] Horvath Z. (2005). *On the positivity step size threshold of RungeKutta methods*, Appl. Numer. Math. 53, 341–356.

[22] Huang C. (2009). *Strong stability preserving hybrid methods*. Appl. Num. Meth. 59 891–904.

[23] Hundsdorfer W. & Ruuth S.J. (2003). *Monotonicity for time discretizations*, Dundee Conference Report NA/217 2003, D. F. Griffiths and G. A. Watson, eds., University of Dundee, Dundee, UK, 85–94.

[24] Hundsdorfer W. & Ruuth S.J. (2006). *On monotonicity and boundedness properties of linear multistep methods*, Math. Comp. 75, 655–672.

[25] Hundsdorfer W. & Ruuth S.J. (2007). *IMEX extensions of linear multistep methods with general monotonicity and boundedness properties*, J. Comput. Phys. 225, 2016–2042.

[26] Hundsdorfer W., Ruuth S.J. & Spiteri R.J. (2003). *Monotonicity-preserving linear multistep methods*, SIAM J. Numer. Anal. 41, 605–623.

[27] Hundsdorfer W. & Verwer J.G. (2003). *Numerical solution of time-dependent advection-diffusion-reaction equations*. Springer Series in Comput. Math. 33, Springer.

[28] Ketcheson D.I. (2009). *Computation of optimal monotonicity preserving general linear methods*, Math. Comp. 78, 1497–1513.

[29] Koren B. (1993). *A robust upwind discretization method for advection, diffusion and source terms*, in Vreugdenhil C.B. & Koren B., Numerical methods for advection-diffusion problems, Braunschweig: Vieweg, 117

[30] Kraaijevanger J.F.B.M. (1991). *Contractivity of Runge-Kutta methods*, BIT 31, 482–528.

[31] Lax P.D. & Wendroff B. (1960). *Systems of conservation laws*, Comm.Pure Appl.Math., 13, 217–237.

[32] Lenferink H.W.J. (1989). *Contractivity preserving explicit linear multistep methods*, Numer. Math. 55, 213–223.

[33] Lenferink H.W.J. (1991). *Contractivity preserving implicit linear multistep methods*, Math. Comp. 56, 177–199.

[34] LeVeque R.J. (1992). *Numerical methods for conservation laws*, Lecture Notes in Mathematics, ETH Zurich, Birkhauser Verlag, Basel.

[35] LeVeque R.J. (2002). *Finite volume methods for hyperbolic problems*, Cambridge Texts in Applied Mathematics, Cambridge University Press.

[36] Pham Thi N.N., Hundsdorfer W. & Sommeijer B.P. (2006). *Positivity for explicit two-step methods in linear multistep and one-leg form* BIT 46, 875–882

[37] Ruuth S.J. (2006). *Global optimization of explicit strong-stability-preserving Runge-Kutta methods*, Math. Comp. 75, 183–207.

[38] Ruuth S.J. & Hundsdorfer W. (2005). *High-order linear multistep methods with general monotonicity and boundedness properties*, J. Comput. Phys. 209, 226–248

[39] Sand J. (1986). *Circle contractive linear multistep methods*, BIT 26, 114–122.

[40] Shu C.-W. (1988). *Total-variation-diminishing time discretizations*, SIAM J. Sci. Stat. Comp. 9, 1073–1084.

[41] Shu C.-W. & Osher S. (1988). *Efficient implementation of essentially nonoscillatory shock-capturing schemes*, J. Comput. Phys. 77, 439–471

[42] Spijker M.N. (1983). *Contractivity in the numerical solution of initial value problems*, Numer. Math. 42, 271–290.

[43] Spijker M.N. (2007). *Stepsize restrictions for general monotonicity in numerical initial value problems*, SIAM J. Numer. Anal. 45, 1226–1245.

[44] Spiteri R.J. & Ruuth S.J. (2002). *A new class of optimal high-order strong-stability-preserving time discretization methodss*, SIAM J. Numer. Anal. 40, 469–491.

[45] Vanselov R. (1983). *Nonlinear stability behaviour of linear multistep methods*, BIT 23, 388–396.

[46] Van Leer B. (1974). *Towards the ultimate conservative difference scheme II. Monotonicity and conservation combined in a second order scheme*, J. Comp. Phys. 14, 361–370.

Summary

This thesis deals with the numerical solution of systems of ordinary differential equations (ODEs), and in particular ODEs that are obtained by spatial discretization of conservation laws. For such problems, steep gradients or discontinuities in the solutions are common. To ensure convergence of the numerical approximations towards the physically relevant exact solutions, it is important that the numerical methods posses certain boundedness or monotonicity properties. In this thesis we focus on such properties for the time discretizations.

For the time discretization we consider a wide class of methods, the so-called general linear methods. Well-known examples of such methods are Runge-Kutta methods and linear multistep methods. It is known from the literature that some classes of Runge-Kutta methods and linear multistep methods allow a representation, the so-called Shu-Osher form, from which monotonicity properties are easily deduced under suitable assumptions on the systems of ODEs and appropriate conditions on the stepsize.

There are, however, many methods that seem to perform well in practice, but for which there are no appropriate Shu-Osher forms. For this reason we study boundedness properties that are weaker than the monotonicity properties from the literature.

The thesis contains an introduction and four chapters. These four chapters contain material from papers that have been published or submitted for publication in scientific journals. The introduction is written with the intention to be understandable for the reader who is not a specialist in the field.

Chapter I presents a generic framework for deriving the largest stepsizes for which boundedness is still guaranteed. These results can be applied to numerical methods that are not included in the standard monotonicity theory from the literature.

However, these conditions on the stepsizes are not always easy to determine for concrete general linear methods. In Chaper II we therefore study special bounds that allow easier verification, and which are still applicable to cases where the standard monotonicity theory does not hold. These special bounds are relevant for a class of functionals that is wider than the class of seminorms, allowing statements on preservation of non-negativity, for example.

In Chapter III we obtain necessary and sufficient conditions for boundedness of linear multistep methods. These conditions are relatively transparant and easy to verify for given classes of methods. In this chapter, also conditions are obtained that ensure monotonicity for combinations of linear multistep methods and Runge-Kutta starting procedures.

Finally, in Chapter IV boundedness and monotonicity properties are studied for the so-called one-leg multistep methods. It is found that the maximal stepsizes for boundedeness of a one-leg method are the same as for the corresponding linear multistep method, but the conditions for monotonicity with Runge-Kutta starting procedures can lead to very different stepsizes.

Samenvatting

Dit proefschrift behandelt het numerieke oplossen van systemen van gewone differentiaalvergelijkingen, met name systemen die worden verkregen door ruimtelijke discretisatie van behoudswetten. De oplossingen van dergelijke problemen hebben vaak zeer steile hellingen of zelfs discontinuïteiten. Om zeker te zijn van convergentie van de numerieke benaderingen naar de fysisch relevante oplossingen, is het belangrijk dat de numerieke methoden bepaalde begrensdheids- of monotoniciteitseigenschappen bezitten.

Voor de tijdsdiscretisatie bekijken we een brede klasse van numerieke methoden, de zogenaamde 'general linear methods'. Bekende voorbeelden van zulke methoden zijn Runge-Kutta methoden en lineaire meerstapsmethoden. Het is bekend uit de literatuur dat sommige klassen van Runge-Kutta en lineaire meerstapsmethoden in een geschikte vorm geschreven kunnen worden, de zogenaamde Shu-Osher vorm, waaruit monotoniciteitseigenschappen gemakkelijk afgeleid kunnen worden, onder geschikte aannamen voor de systemen van gewone differentiaalvergelijkingen en passende voorwaarden op de stapgrootte.

Er zijn echter vele methoden die goed werken in de praktijk, maar waarvoor er geen geschikte Shu-Osher vorm bestaat. Om deze reden bestuderen we begrensdheidseigenschappen die zwakker zijn dan de monotoniciteitseigenschappen uit de literatuur.

Het proefschrift bevat een inleiding en vier hoofdstukken. Deze vier hoofdstukken bevatten materiaal uit artikelen die gepubliceerd zijn of aangeboden voor publicatie in wetenschappelijke tijdschriften. De inleiding is geschreven met de intentie om begrijpelijk te zijn voor lezers die geen specialist in het vakgebied zijn.

Hoofdstuk I geeft een algemeen kader voor het afleiden van de maximale stapgrootten waarvoor begrensdheid nog gegarandeerd is. Deze resultaten kunnen worden toegepast op numerieke methoden waarvoor de standaard monotoniciteitstheorie uit de literatuur niet van toepassing is.

Deze voorwaarden op de stapgrootten zijn echter niet altijd eenvoudig te bepalen voor concrete numerieke methoden. In hoofdstuk II bestuderen we daarom speciale vormen van begrensdheid die eenvoudiger na te gaan zijn, en die nog steeds van toepassing zijn op gevallen waar de standaard monotoniciteitstheorie niet geldig is. Deze vormen van speciale begrensheid zijn relevant voor een klasse van functionalen die breder is dan de klasse van semi-normen, waardoor, bijvoorbeeld, uitspraken over behoud van niet-negativiteit gedaan kunnen worden.

In hoofdstuk III worden noodzakelijke en voldoende voorwaarden afgeleid

voor begrensdheid van lineaire meerstapsmethoden. Deze voorwaarden zijn relatief transparant en gemakkelijk te controleren voor gegeven klassen van methoden. In dit hoofdstuk worden ook voorwaarden verkregen die monotoniciteit geven voor combinaties van lineaire meerstapsmethoden en Runge-Kutta startprocedures.

Ten slotte, in hoofdstuk IV worden begrensdheids- en monotoniciteitseigenschappen bestudeerd voor de zogenaamde 'one-leg' meerstapsmethoden. Het blijkt dat de maximale stapgrootten voor begrensdheid van een one-leg methode dezelfde zijn als voor de bijbehorende lineaire meerstapsmethode, maar de voorwaarde voor monotoniciteit met een Runge-Kutta startprocedure kan zeer verschillende stapgrootten opleveren.

Curriculum Vitae

Anna Motsartova was born in Chkalovsk, USSR, on October 20, 1981. After completing her secondary school studies, in 1999, she started her studies in mathematics at the Novosibirsk State University, Novosibirsk, Russia. She graduated in 2004. From 2004 to 2006 she was part-time member of the numerical analysis research group at the Institute of Computational Mathematics and Mathematical Geophysics (ICMMG) in Novosibirsk, Russia. From 2007 to 2011 she was employed as Assistent in Opleiding (AiO) at Centrum Wiskunde & Informatica (CWI). During that period she did her PhD research under the supervision of Prof.dr W. Hundsdorfer and Prof.dr. M.N. Spijker that is described in this thesis.