Stepsize Restrictions for Boundedness and Monotonicity of Multistep Methods

W. Hundsdorfer*, A. Mozartova†, M.N. Spijker‡

Abstract

In this paper 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.

2000 Mathematics Subject Classification: 65L06, 65M06, 65M20.
Keywords and Phrases: initial value problems, method of lines, multistep methods, monotonicity, boundedness, positivity, TVD, TVB, strong stability.

1 Introduction

1.1 Monotonicity assumptions

In this paper we consider initial value problems for systems of ordinary differential equations (ODEs) on a vector space $\mathbb{V}$, written as

$$u'(t) = F(u(t)) \quad (t \geq 0)\,, \qquad u(0) = u_0\,, \tag{1.1}$$

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

$$\|v + \tau_0 F(v)\| \;\leq\; \|v\| \qquad \text{for all } v \in \mathbb{V}. \tag{1.2}$$

Assumption (1.2) 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

$$\|u_n\| \;\leq\; \|u_0\| \tag{1.3}$$

for all $n \geq 1$ under the stepsize restriction $\Delta t \leq \tau_0$. For general one-step methods, property (1.3) under a stepsize restriction $\Delta t \leq c\,\tau_0$ is often referred to as *monotonicity* or *strong stability preservation* (SSP).

Useful and well-known examples for (1.2) 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

*CWI, P.O. Box 94079, 1090-GB Amsterdam, The Netherlands (willem.hundsdorfer@cwi.nl).
†CWI, P.O. Box 94079, 1090-GB Amsterdam, The Netherlands (a.mozartova@cwi.nl). Work of this author is supported by a grant from the Netherlands Organisation for Scientific Research NWO.
‡Mathematical Institute, Leiden University, P.O. Box 9512, 2300-RA Leiden, The Netherlands (spijker@math.leidenuniv.nl).

$\|v\|_{\mathrm{TV}} = \sum_{j=1}^{M} |v_{j-1} - v_j|$ (with $v_0 = v_M$), arising from one-dimensional partial differential equations (PDEs), see for instance [4, 15, 18].

Some of the results in this paper will be formulated with sublinear functionals instead of seminorms.[1] This makes it possible to take, for example, maximum principles into consideration as in [23], by requiring that (1.2) holds for the functionals $\|v\|_+ = \max_j v_j$ and $\|v\|_- = -\min_j v_j$. Another example, from [11], 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 $v \in \mathbb{R}^M$.

1.2 Monotonicity and boundedness for linear multistep methods

To solve (1.1) 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

$$u_n = \sum_{j=1}^{k} a_j u_{n-j} + \Delta t \sum_{j=0}^{k} b_j F(u_{n-j}) \tag{1.4}$$

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 paper that

$$\sum_{j=1}^{k} a_j \;=\; 1\,, \qquad \sum_{j=1}^{k} j\, a_j = \sum_{j=0}^{k} b_j\,, \qquad b_0 \geq 0\,. \tag{1.5}$$

The two equalities in (1.5) 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

$$c \;=\; \min_{1 \leq j \leq k} \frac{a_j}{b_j}\,, \tag{1.6}$$

with convention $a/0 = +\infty$ if $a \geq 0$. From (1.2) it can then be shown that

$$\|u_n\| \;\leq\; \max_{0 \leq j < k} \|u_j\| \tag{1.7}$$

for $n \geq k$, under the stepsize restriction $\Delta t \leq c\,\tau_0$; see e.g. [4, 23]. This property can be viewed as an extension of (1.3) for multistep methods with arbitrary starting values.

Results of this type for nonlinear problems were derived originally in [21] with the total variation seminorm, and (1.7) 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 [5] – has become popular. Related work for nonlinear problems was done in [16, 17, 20, 24] for *contractivity*, where one considers $\|\tilde{u}_n - u_n\|$ with differences of two numerical solutions instead of $\|u_n\|$ as in (1.7). Finally we mention that related results on nonnegativity preservation and contractivity or monotonicity for *linear* problems were derived already in [1, 22], again for methods with all $a_j, b_j \geq 0$ and with $\Delta t \leq c\,\tau_0$.

[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.

In order to conclude (1.7) from (1.2) 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 [23].

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 [16], 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 (1.7).

Instead of (1.7), we will consider

$$\|u_n\| \;\leq\; \mu \cdot \max_{0\leq j<k} \|u_j\| \tag{1.8}$$

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 [12, 14] for (1.8) to be valid with arbitrary seminorms under assumption (1.2) 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 paper we will use the general framework of [10] 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 (1.3) 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 [14] for some explicit two-step methods.

1.3 Outline of the paper

To illustrate the relevance of the results we first present in Section 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 some notations are introduced, together with a formulation of the linear multistep method (1.4) that is suited for application of the general boundedness results of [10].

The main results are presented in Section 4. Using the framework of [10], 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 (1.3) – for combinations of linear multistep methods and Runge-Kutta starting procedures.

Section 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 6 the maximal stepsize coefficients γ for either boundedness or monotonicity with some specific starting procedures.

Finally, Section 7 contains some concluding remarks together with comments on multistep schemes that are related to the linear multistep methods (1.4).

Along with the usual typographical symbol □ to indicate the end of a proof, we will use in this paper also the symbol ◇ to mark the end of examples or remarks.

2 A numerical illustration

To illustrate the relevance of our monotonicity and boundedness concepts, we consider two-step methods of the form

$$u_n = \tfrac{3}{2}u_{n-1} - \tfrac{1}{2}u_{n-2} + \Delta t\,\beta F(u_{n-1}) + \Delta t\,(\tfrac{1}{2} - \beta)F(u_{n-2})\,. \tag{2.1}$$

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 the linear stability regions, as shown in Figure 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 (1.7) 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 (1.3) 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 (1.2) is satisfied for $\tau_0 = \frac{1}{2}\Delta x$ in the maximum norm and the total variation seminorm; see for instance [15, 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.

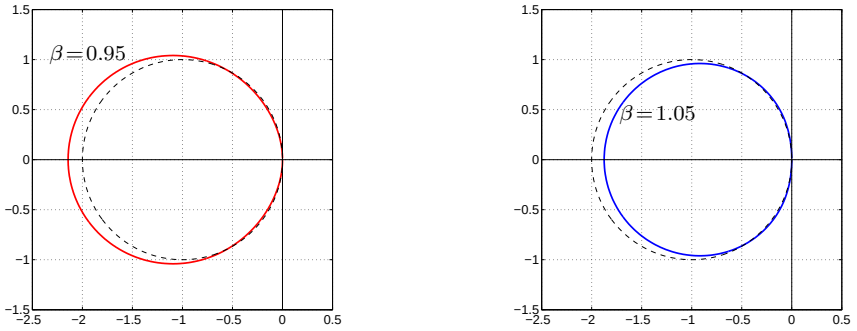

Figure 1: Stability regions of the two-step methods (2.1) with $\beta = 0.95$ (left), $\beta = 1.05$ (right). For comparison, the circle $\{\zeta \in \mathbb{C} : |\zeta + 1| = 1\}$ is displayed by the dashed curve.

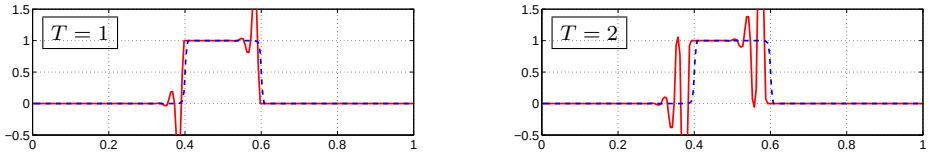

Figure 2: Numerical solutions at $T = 1$ and $T = 2$ for the two-step methods (2.1) with $\beta = 1.05$ (dashed), $\beta = 0.95$ (solid lines).

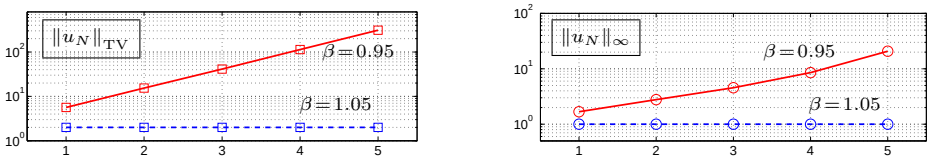

Figure 3: Values of $\|u_N\|_{\mathrm{TV}}$ (left) and $\|u_N\|_\infty$ (right) for $T = 1, 2, \ldots, 5$ and the two-step methods (2.1) with $\beta = 1.05$ (dashed), $\beta = 0.95$ (solid lines).

The numerical solutions for the two schemes are shown in Figure 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, 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 (1.8) is not satisfied here for the scheme with $\beta = 0.95$.

With the results of this paper the different behaviour of these two closely related schemes can be explained. As we will see in Section 6.1, to satisfy the boundedness property (1.8) or the monotonicity property (1.3) with forward Euler starting procedure, the method with $\beta = 1.05$ allows much larger stepsizes than the method with $\beta = 0.95$.

3 Notations and input-output formulations

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,

$$E = \begin{pmatrix} 0 & & & \\ 1 & 0 & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{pmatrix} \in \mathbb{R}^{m \times m}, \tag{3.1}$$

and define

$$A \;=\; \sum_{j=1}^{k} a_j E^j\,, \qquad B \;=\; \sum_{j=0}^{k} b_j E^j\,, \tag{3.2}$$

where $E^0 = I$. These $A, B \in \mathbb{R}^{m\times m}$ are lower triangular Toeplitz matrices containing the coefficients of the method (1.4). For $m \ge k$ 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 $m \ge 1$, we define $J = [e_1, \ldots, e_m, O]$ for $1 \le m < k$, 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 $y_i = \sum_{j=1}^{l} \kappa_{ij} x_j \in \mathbb{V}$ $(1 \le i \le m)$. (In case $\mathbb{V} = \mathbb{R}^M$ with $M \ge 1$, 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 $K \ge 0$ when all entries κ_{ij} of this matrix are nonnegative.

3.2 Formulations with input vectors

In order to apply the theory obtained in [10], we will formulate the multistep scheme (1.4) in terms of input and output vectors. The *output vectors* of the scheme are $y_n = u_{k-1+n}$, $n \ge 1$. The starting values $u_0, u_1, \ldots, u_{k-1}$ will enter the scheme in the first k steps in the combinations

$$x_l \;=\; \sum_{j=l}^{k} a_j u_{k-1+l-j} + \Delta t \sum_{j=l}^{k} b_j F(u_{k-1+l-j}) \qquad \text{for } 1 \le l \le k\,. \tag{3.3}$$

The multistep scheme (1.4) then can be written as

$$y_n \;=\; x_n + \sum_{j=1}^{n-1} a_j y_{n-j} + \Delta t \sum_{j=0}^{n-1} b_j F(y_{n-j}) \qquad \text{for } 1 \le n \le k\,, \tag{3.4a}$$

$$y_n \;=\; \sum_{j=1}^{k} a_j y_{n-j} + \Delta t \sum_{j=0}^{k} b_j F(y_{n-j}) \qquad \text{for } n > k\,, \tag{3.4b}$$

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 (3.4) 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

$$y \;=\; \boldsymbol{J}x + \boldsymbol{A}y + \Delta t\, \boldsymbol{B}\boldsymbol{F}(y)\,. \tag{3.5}$$

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

$$\max_{1\le n\le m} \|y_n\| \;\le\; \mu \cdot \max_{1\le j\le k} \|x_j\| \qquad \begin{array}{l}\text{whenever (1.2) is valid, } \Delta t \le \gamma\,\tau_0\text{, and}\\ x, y \text{ satisfy (3.5)}, \; m \ge 1,\end{array} \tag{3.6}$$

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 (1.2) is satisfied, as

well as all x, y satisfying (3.5) and $m \geq 1$. Therefore γ and μ will *not* depend on a particular problem (1.1) under consideration.

A convenient form to derive results on boundedness is obtained by multiplying relation (3.5) by $(\boldsymbol{I} - \boldsymbol{A} + \gamma \boldsymbol{B})^{-1}$ with $\gamma > 0$. This yields

$$y \;=\; \boldsymbol{R}\, x + \boldsymbol{P}\Big(y + \frac{\Delta t}{\gamma}\boldsymbol{F}(y)\Big)\,, \tag{3.7}$$

where $R = (r_{ij}) \in \mathbb{R}^{m\times k}$ and $P = (p_{ij}) \in \mathbb{R}^{m\times m}$ are given by

$$R \;=\; (I - A + \gamma B)^{-1} J\,, \qquad P \;=\; (I - A + \gamma B)^{-1}\gamma B\,. \tag{3.8}$$

Note that $I - A + \gamma B$ is invertible for any $\gamma > 0$, because $b_0 \geq 0$, and therefore (3.7) is still equivalent to (3.5). 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

$$\mathrm{spr}(|P|) < 1\,, \qquad (I - |P|)^{-1} \;\geq\; 0\,. \tag{3.9}$$

3.3 Application of a general result on boundedness

To obtain boundedness results for the multistep methods we will use a general result from [10]. The connection with the notation used in that paper is established by writing (3.5) in the form

$$y \;=\; \boldsymbol{S}x + \Delta t\, \boldsymbol{T}\boldsymbol{F}(y) \tag{3.10}$$

with $S \in \mathbb{R}^{m\times k}$ and $T \in \mathbb{R}^{m\times m}$ defined by

$$S = (I - A)^{-1} J\,, \qquad T = (I - A)^{-1} B\,. \tag{3.11}$$

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 (1.5) imply that the linear multistep method is exact for first-degree polynomial solutions: if $u_j = \alpha + \beta \cdot j\Delta t$ $(0 \leq j < k)$ and $F(u) \equiv \beta$, then $u_n = \alpha + \beta \cdot n\Delta t$ for all $n \geq k$. Since $y_n = u_{k-1+n}$ $(n \geq 1)$ in (3.10), it follows by varying $\alpha, \beta \in \mathbb{R}$ that

$$e_j^T S \;\neq\; 0 \qquad \text{for all } j\,, \tag{3.12a}$$

$$(e_i - e_j)^T [\, S\; T\,] \;\neq\; 0 \qquad \text{if } \; i \neq j\,, \tag{3.12b}$$

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 2.4 in [10] now yields the following result:

Theorem 3.1 *Consider a linear multistep method (1.4) satisfying (1.5). Then, for any seminorm $\|\cdot\|$ on $\mathbb{V}$, the boundedness property (3.6) is valid provided that*

$$\|(I - |P|)^{-1}|R|\,\|_\infty \;\leq\; \mu \qquad \textit{for all } m\,. \tag{3.13}$$

Moreover, condition (3.13) is necessary for (3.6) to be valid for the class of spaces $\mathbb{V} = \mathbb{R}^M$, $M \geq 1$, with the maximum norm.

In the above result, proving necessity of (3.13) is by far the most difficult part, and for that part the conditions (3.12) 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 (3.7) 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 (1.2) 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 (3.13) implies (3.6).

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 (1.5) is satisfied.

To formulate the results we will use some standard concepts for linear multistep methods, which can be found in [2, 7], 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.

4.1 Boundedness with respect to the input vectors

First we consider the boundedness property (3.6) with $\mu > 0$ arbitrary, giving boundedness with respect to the input vectors $x_1, \ldots, x_k$ defined by (3.3). 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:

(4.1) $\left\{ \begin{array}{l} \text{there is a } \mu > 0 \text{ such that the boundedness property (3.6) is valid for} \\ \text{all } \mathbb{V} = \mathbb{R}^M,\ M \ge 1, \text{ with maximum norm } \|\cdot\|_\infty\,; \end{array} \right.$

(4.2) $\left\{ \begin{array}{l} \text{there is a } \mu > 0 \text{ such that the boundedness property (3.6) is valid for} \\ \text{any vector space } \mathbb{V} \text{ and seminorm } \|\cdot\|\,. \end{array} \right.$

The next theorem provides necessary and sufficient conditions for these statements. The proof of the theorem will be given in Section 5.

Theorem 4.1 *Consider an irreducible, zero-stable linear multistep method, and let $\gamma > 0$. Then each of the statements (4.1) and (4.2) is equivalent to*

$$-\gamma \in \mathrm{int}(\mathcal{S})\,, \qquad P \ge 0 \quad \textit{(for all } m\textit{)}\,. \tag{4.3}$$

Along with (4.1), (4.2), we also consider the following stronger statement on boundedness for arbitrary nonnegative sublinear functionals:

$$(4.4)\quad \begin{cases} \text{there is a } \mu > 0 \text{ such that the boundedness property (3.6) is valid for} \\ \text{any vector space } \mathbb{V} \text{ and nonnegative sublinear functional } \|\cdot\|\,. \end{cases}$$

Here the restriction to sublinear functionals that are nonnegative has been made to get a similar formulation as for seminorms; see Remark 5.4 below.

Theorem 4.2 *Suppose the linear multistep method is zero-stable, $\gamma > 0$ and*

$$(4.5)\qquad R \geq 0\,, \quad P \geq 0 \quad \textit{(for all } m\textit{)}\,.$$

Then statement (4.4) holds.

Also the proof of this theorem will be given in Section 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). Consequently, for irreducible zero-stable linear two-step methods, each of the statements (4.1), (4.2), (4.4) 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.

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 (3.3). In general, it is more natural to consider boundedness with respect to the starting vectors $u_0, \ldots, u_{k-1}$, as in (1.8). We therefore consider, similar to (3.6), the following boundedness property of the linear multistep scheme (1.4):

$$(4.6)\quad \max_{k \leq n < k+m} ||u_n|| \;\leq\; \tilde{\mu} \cdot \max_{0 \leq j < k} ||u_j|| \qquad \begin{array}{l} \text{whenever (1.2) is valid, } \Delta t \leq \gamma\,\tau_0, \text{ and} \\ \text{(1.4) holds for } k \leq n < k+m,\ m \geq 1. \end{array}$$

If $\|\cdot\|$ is a seminorm, it is easily seen from (1.2) and (3.3) 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 (3.6) holds with stepsize coefficient γ and factor μ, then there is a $\tilde{\mu}$ such that (4.6) holds.

The reverse is also true for seminorms. To see this, first note that (3.4b) is the same as (1.4), only with a shifted index. Therefore (4.6) implies $\max_{k+1 \leq i \leq k+m} \|y_i\| \leq \tilde{\mu} \max_{1 \leq j \leq k} \|y_j\|$ when (1.2) is valid and $\Delta t \leq \gamma \tau_0$. From (3.4a) 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 [14, p. 614]. 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 (3.6) and (4.6) are for seminorms essentially equivalent, in the following sense:

Lemma 4.3 *Suppose $\|\cdot\|$ is a seminorm on a vector space $\mathbb{V}$, and let $\gamma > 0$. Then (3.6) holds with some $\mu > 0$ if and only if (4.6) holds with some $\tilde{\mu} > 0$.*

For sublinear functionals such an equivalence does not hold. As we know from Theorem 4.2, zero-stability and $P, R \geq 0$ is sufficient for having (3.6) 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 [1] with the functional $\|v\|_0 = -\min\{0, v_1, \ldots, v_M\}$ on $\mathbb{R}^M$, it can be shown that to have (4.6) 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 (1.6).

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 (1.6). 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 (3.3),

$$x \;=\; \boldsymbol{R}_0\, u_0 + \boldsymbol{P}_0\Big(w + \frac{\Delta t}{\gamma}\boldsymbol{F}(w)\Big) \tag{4.7}$$

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 4.4 *Let $\|\cdot\|$ be a sublinear functional on a vector space $\mathbb{V}$. Suppose (4.7) holds with $\|w_j\| \leq \|u_0\|$ $(1 \leq j \leq m_0)$, $y \in \mathbb{V}^m$ satisfies (3.5), and*

$$R\,R_0 \geq 0\,, \qquad R\,P_0 \geq 0\,, \qquad P \geq 0\,. \tag{4.8}$$

Then $\|y_i\| \leq \|u_0\|$ for $1 \leq i \leq m$ whenever (1.2) is valid and $\Delta t \leq \gamma\tau_0$.

Proof. From (3.7) 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 $e = R\,R_0 1 + R\,P_0\bar{e} + Pe$. 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 $\|w_j\| \le \|u_0\|$ for $1 \le j \le m_0$ as soon as (1.2) is valid and $\Delta t \le \gamma\tau_0$, 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 [6, 9]. Necessary condition for monotonicity are found in [23]; it can be shown from those results that the condition (4.8) is necessary in Theorem 4.4 under a weak irreducibility condition on the combined scheme.

Example 4.5 Consider a two-step method, and let $c_j = a_j - \gamma b_j$ $(j = 1, 2)$. Then

$$x \;=\; \begin{pmatrix} c_2 & c_1 \\ 0 & c_2 \end{pmatrix} \begin{pmatrix} u_0 \\ u_1 \end{pmatrix} \;+\; \gamma \begin{pmatrix} b_2 & b_1 \\ 0 & b_2 \end{pmatrix} \begin{pmatrix} u_0 + \frac{1}{\gamma}\Delta t\, F(u_0) \\ u_1 + \frac{1}{\gamma}\Delta t\, F(u_1) \end{pmatrix}. \tag{4.9}$$

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

$$u_1 = r_0 u_0 + q_0\Big(u_0 + \frac{\Delta t}{\gamma}F(u_0)\Big) + q_1\Big(u_1 + \frac{\Delta t}{\gamma}F(u_1)\Big) \tag{4.10}$$

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 (4.7) with

$$R_0 = \begin{pmatrix} c_2 + c_1 r_0 \\ c_2 r_0 \end{pmatrix}, \qquad P_0 = \begin{pmatrix} c_1 q_0 + \gamma b_2 & c_1 q_1 + \gamma b_1 \\ c_2 q_0 & c_2 q_1 + \gamma b_2 \end{pmatrix}, \tag{4.11}$$

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

$$u_1 = r_0 u_0 + q_0\Big(u_0 + \frac{\Delta t}{\gamma}F(u_0)\Big) + q_1\Big(\bar{u}_1 + \frac{\Delta t}{\gamma}F(\bar{u}_1)\Big) \tag{4.12}$$

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

$$R_0 = \begin{pmatrix} c_2 + c_1 r_0 \\ c_2 r_0 \end{pmatrix}, \qquad P_0 = \begin{pmatrix} c_1 q_0 + \gamma b_2 & c_1 q_1 & \gamma b_1 \\ c_2 q_0 & c_2 q_1 & \gamma b_2 \end{pmatrix}, \tag{4.13}$$

and $w = (u_0, \bar{u}_1, u_1)^T \in \mathbb{V}^3$. ◇

5 Technical derivations and proofs

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

$$(I - A + \gamma B)^{-1} \;=\; \sum_{j\ge 0} \rho_j E^j\,, \tag{5.1}$$

$$P \;=\; (I - A + \gamma B)^{-1}\gamma B \;=\; \sum_{j\ge 0} \pi_j E^j\,, \tag{5.2}$$

with coefficients $\rho_j, \pi_j \in \mathbb{R}$. Note that $R \in \mathbb{R}^{m\times k}$ contains the first r columns of $(I - A + \gamma B)^{-1}$, $r = \min\{k, m\}$.

It is convenient to define $\rho_j = 0$ for $j < 0$. The coefficients ρ_n then satisfy the multistep recursion

$$\rho_n \;=\; \sum_{j=1}^{k} a_j \rho_{n-j} - \gamma \sum_{j=0}^{k} b_j \rho_{n-j} + \delta_{n0} \qquad (n \ge 0)\,, \tag{5.3}$$

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

$$\pi_n \;=\; \gamma \sum_{j=0}^{k} b_j \rho_{n-j} \qquad (n \ge 0)\,. \tag{5.4}$$

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

$$u'(t) \;=\; \lambda\, u(t) \qquad \text{with } \Delta t\, \lambda = -\gamma\,. \tag{5.5}$$

Lemma 5.1 *If $-\gamma \in \mathcal{S}$ then $\max_{0\le n<\infty} |\rho_n| < \infty$. 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 $|\rho_n| \le \kappa\, \theta^n$ for all $n \ge 0$.*

Proof. From (5.3) we see that the coefficients ρ_n are obtained by applying the linear multistep method to (5.5). 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 (5.3) 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 [3, Thm. I.4.2]. Writing the solution of (5.3) in terms of these characteristic roots thus provides the proof. □

Corollary 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 $m \ge k$

$$e_m^T P e \;=\; 1 - (\rho_{m-1}, \ldots, \rho_1, \rho_0)\, v \;=\; 1 - \textstyle\sum_{j=1}^{k} \rho_{m-j} v_j\,.$$

The proof now follows from the previous lemma. □

The recursions (5.3), (5.4) will be used to compute numerically the largest stepsize coefficient γ such that $R \ge 0$ or $P \ge 0$ 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 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 $P \ge 0$ (for all m) with some $\gamma > 0$ requires $b_1 \ge 0$ and $a_1 b_1 + b_2 \ge 0$. These two inequalities were mentioned already in [12], but now it is seen that these are really needed for boundedness. ◇

5.2 Proofs of Theorems 4.1, 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. (3.11). Application of Lemma 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 (4.3) in Theorem 4.1

The following arguments are somewhat similar to those used in the proof of Corollary 3.3 of [10], 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 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 3.1 thus shows that the statements (4.1), (4.2) are valid.

Necessity of (4.3) in Theorem 4.1

To finish the proof of Theorem 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 (4.1).

Any application of method (1.4) 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 (1.2) with $\tau_0 = 1$, $\mathbb{V} = \mathbb{R}^2$ and $\|\cdot\| = \|\cdot\|_\infty$. Using Lemma 4.3, it thus follows that property (3.6) implies $\gamma\cdot\mathcal{D} \subset \mathcal{S}$. Therefore, if $\gamma > 0$, then $-\gamma \in \mathrm{int}(\mathcal{S})$ is certainly necessary for (4.1).

Assuming $-\gamma \in \mathrm{int}(\mathcal{S})$, it remains to show that $P \geq 0$ is necessary for (3.13). 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 5.2 that $\sum_{j=0}^{\infty}\pi_j = 1$. We can write (3.13) 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, (3.13) 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

$$(I-D)^{-1}e_i \;\leq\; \tilde{\mu}\, e \qquad (\text{for all } m \geq l+1 \text{ and } 1 \leq i \leq l)\,. \tag{5.6}$$

The bounds (5.6) are related to stability of the recursion

$$\eta_n = \sum_{j=0}^{l} \delta_j \eta_{n-j} \qquad (\text{for } n \geq l) \tag{5.7}$$

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, (5.6) implies stability of the recursion (5.7). 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 (3.13) is *not* satisfied. According to Theorem 3.1, also (4.1) is then not satisfied, which completes the proof of Theorem 4.1.

Sufficiency of (4.5) in Theorem 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 (3.9), we also have have $\mathrm{spr}(|P|) < 1$. Assuming (1.2) and $\Delta t \leq \gamma\tau_0$, it follows from Theorem 3.9 in [11] that

$$\|y_i\| \leq \mu_i \cdot \max_{1 \leq j \leq k} \|x_j\| \qquad (1 \leq i \leq m) \tag{5.8}$$

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 (3.6) then follows.

Remark 5.4 Replacement of the μ_i in (5.8) 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 (3.6). Theorem 4.2 has therefore been formulated for nonnegative sublinear functionals only.

Necessary and sufficient conditions for boundedness with the form (5.8) for arbitrary sublinear functionals are given in [11]. 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. ◇

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 (5.3), (5.4) for the coefficients ρ_n and π_n in terms of the roots of the characteristic polynomial of the recursion (5.3). 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

$$c_1 \geq 0\,, \qquad c_1^2 + 4c_2 \geq 0\,. \tag{5.9}$$

We note that under condition (5.9) 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 (5.9) holds together with

$$b_0 c_1 + b_1 \geq 0\,, \qquad b_0(c_1^2 + c_2) + b_1 c_1 + b_2 \geq 0\,, \qquad b_0\theta^2 + b_1\theta + b_2 \geq 0\,, \tag{5.10}$$

where $\theta = \frac{1}{2}c_1 + \frac{1}{2}\sqrt{c_1^2+4c_2}$. The first two inequalities in (5.10) just mean that $\pi_1, \pi_2 \geq 0$.

Remark 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

$$(I - A + \gamma B)^{-1} = (I - (1-\gamma)E)^{-1}(I + a_2 E)^{-1}\,.$$

We see from (5.9) 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. ◇

5.4 Remark on the construction in [12, 14]

Multiplication of (3.5) 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 [14] for $k = 2$ and in [12] 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 [14] 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 $P, R \geq 0$.

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 (4.3) and (4.8), 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 (5.3), (5.4) 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 (5.3) 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 $\operatorname{sgn}(\pi_n) = \operatorname{sgn}(c \sum_{j=0}^{k} b_j \theta^{-j})$ 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 (4.8) amount to $\sum_{j=1}^{k} v_j \rho_{n-j} \geq 0$ 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 $m \geq 1$.

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.

In Figure 4 (left panel) the maximal values of γ are displayed for which $P \geq 0$. As noted in Section 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 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 $P \geq 0$.

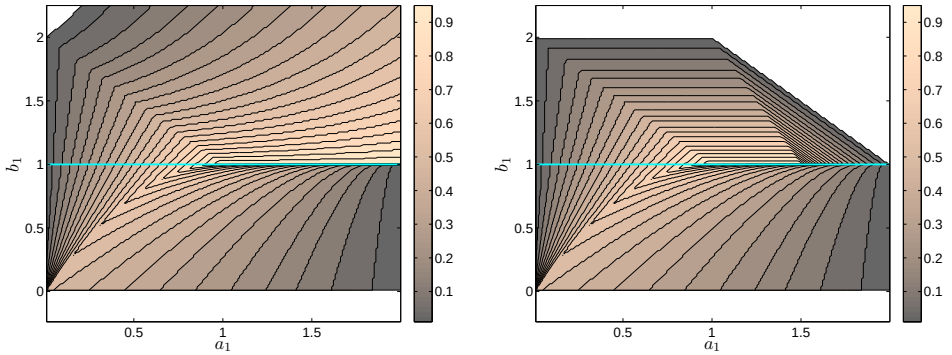

Figure 4: Explicit two-step methods of order one, with parameters $a_1 \in [0, 2)$ horizontally and $b_1 \in [-\frac{1}{4}, \frac{9}{4}]$, $b_1 \neq 1$, vertically. Left panel: threshold $\gamma > 0$ for boundedness. Right panel: threshold $\gamma > 0$ for monotonicity with forward Euler starting procedure. Contour levels at $j/20$, $j = 0, 1, \ldots$; for the 'white' areas there is no positive γ.

In Figure 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 (2.1) 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 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 $\beta = 0.95$.

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 5 (left panel). These values correspond to those found earlier in [12, Fig. 2]. We now see from Theorem 4.1 that – somewhat surprisingly – the latter values, which were obtained by ad-hoc arguments, are not only sufficient but also necessary for boundedness.

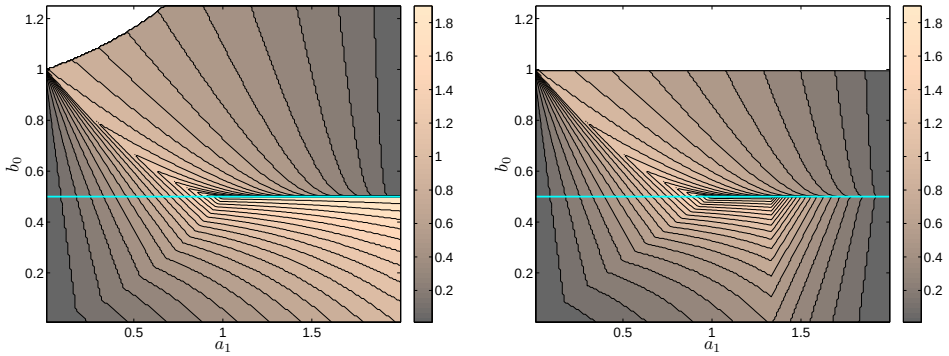

Figure 5: Implicit two-step methods of order two, with parameters $a_1 \in [0, 2)$ horizontally and $b_0 \in [0, \frac{5}{4}]$, $b_0 \neq \frac{1}{2}$, vertically. Left panel: thresholds $\gamma > 0$ for boundedness. Right panel: thresholds $\gamma > 0$ for monotonicity with the θ-method, $\theta = b_0$, 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 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.

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 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 [12, Fig. 1]. The left panel of the figure shows the maximal γ for which $P \geq 0$ and $-\gamma \in \text{int}(\mathcal{S})$.

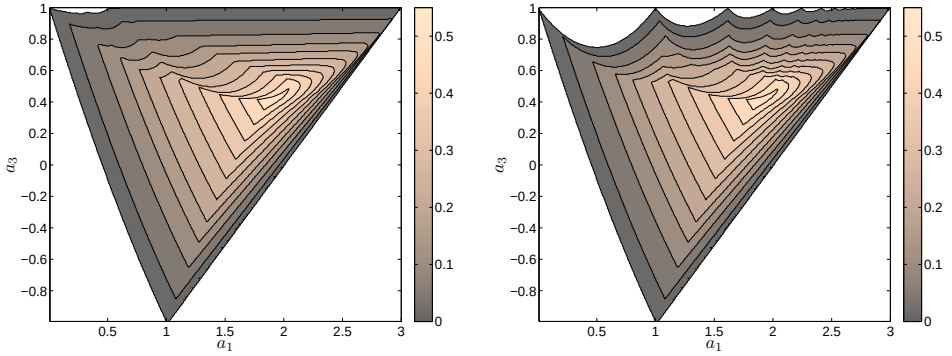

Figure 6: Explicit three-step methods of order three, with parameters $a_1 \in [0, 3]$ horizontally and $a_3 \in [-1, 1]$ vertically. Left panel: threshold $\gamma > 0$ for boundedness, that is, $P \geq 0$ and $-\gamma \in \text{int}(\mathcal{S})$. Right panel: maximal $\gamma > 0$ such that $P \geq 0$ 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 [19], 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 [12, 19] in a new and wider perspective.

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 [19], given in [13] 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 7 (left panel) the maximal value of γ is shown such that the methods are zero-stable, $-\gamma \in \text{int}(\mathcal{S})$ and $P \geq 0$. The right panel shows the error constants (defined as in [7, Sect. III.2]) for the zero-stable methods.

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 [19] 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 5.4, while keeping the error constants at a moderate size.

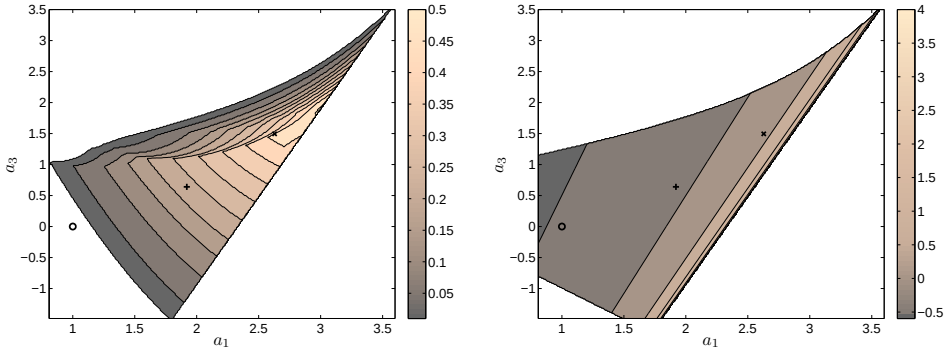

Figure 7: Explicit four-step methods of order four, with parameters described above. Left panel: threshold $\gamma > 0$ for boundedness. Contour levels at $j/20$, $j = 0, 1, \ldots$; for the 'white' areas there is no positive γ. Right panel: $\log_{10}$ of the absolute error constants for zero-stable methods. Markers: $\circ$ for AB4, $+$ for EBDF4 and $\times$ for TVB(4,4).

It is 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 [19] still stands: the TVB(4,4) scheme gives a good compromise between a moderate error constant 2.38 and a relatively large stepsize coefficient $\gamma \approx 0.45$.

7 Concluding remarks

Based on the general framework of [10], we have obtained in this paper stepsize restrictions for linear multistep methods that are necessary and sufficient for boundedness with maximum norms or arbitrary seminorms (Theorem 4.1). This puts the previously found, more complicated sufficient conditions of [12, 14] 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 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 to facilitate the analysis of linear multistep methods. Stability results for one-leg methods often have a somewhat nicer form than for linear multistep methods. It can be shown that the maximal stepsize coefficient for boundedness 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 rather complicated. 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 [11].

References

[1] C. Bolley, M. Crouzeix, *Conservation de la positivité lors de la discrétisation des problèmes d'évolution paraboliques*, RAIRO Anal. Numer. 12 (1978), 237–245.

[2] J.C. Butcher, *Numerical Methods for Ordinary Differential Equations*. Wiley, 2003.

[3] M. Crouzeix, P.-A. Raviart, *Approximation des Problèmes d'Évolution*. Lecture Notes, University Rennes, 1980.

[4] S. Gottlieb, D.I. Ketcheson, C.-W. Shu, *High order strong stability preserving high-order time discretizations*, J. Sci. Comput. 38 (2009), 251–289.

[5] S. Gottlieb, C.-W. Shu, E. Tadmor, *Strong stability preserving high-order time discretization methods*, SIAM Review 43 (2001), 89–112.

[6] L. Ferracina, M.N. Spijker, *An extension and analysis of the Shu-Osher representation of Runge-Kutta methods*, Math. Comp. 74 (2005), 201–219.

[7] E. Hairer, S.P. Nørsett, G. Wanner, *Solving Ordinary Differential Equations I – Nonstiff Problems*. Second edition, Springer Series Comput. Math. 8, Springer, 1993.

[8] E. Hairer, G. Wanner, *Solving Ordinary Differential Equations II – Stiff and Differential-Algebraic Problems*. Second edition, Springer Series in Comput. Math. 14, Springer, 1996.

[9] I. Higueras, *Representations of Runge-Kutta methods and strong stability preserving methods*, SIAM J. Numer. Anal. 43 (2005), 924–948.

[10] W. Hundsdorfer, A. Mozartova, M.N. Spijker, *Stepsize conditions for boundedness in numerical initial value problems*, SIAM J. Numer. Anal. 47 (2009), 3797–3819.

[11] W. Hundsdorfer, A. Mozartova, M.N. Spijker, *Special boundedness properties in numerical initial value problems*. To appear, 2010.

[12] W. Hundsdorfer, S.J. Ruuth, *On monotonicity and boundedness properties of linear multistep methods*, Math. Comp. 75 (2006), 655–672.

[13] W. Hundsdorfer, S.J. Ruuth, *IMEX extensions of linear multistep methods with general monotonicity and boundedness properties*, J. Comput. Phys. 225 (2007), 2016–2042.

[14] W. Hundsdorfer, S.J. Ruuth, R.J. Spiteri, *Monotonicity-preserving linear multistep methods*, SIAM J. Numer. Anal. 41 (2003), 605–623.

[15] W. Hundsdorfer, J.G. Verwer, *Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations*. Springer Series in Comput. Math. 33, Springer, 2003.

[16] H.W.J. Lenferink, *Contractivity preserving explicit linear multistep methods*, Numer. Math. 55 (1989), 213–223.

[17] H.W.J. Lenferink, *Contractivity preserving implicit linear multistep methods*, Math. Comp. 56 (1991), 177–199.

[18] R.J. LeVeque, *Finite Volume Methods for Hyperbolic Problems*. Cambridge Texts in Applied Mathematics, Cambridge University Press, 2002.

[19] S.J. Ruuth, W. Hundsdorfer, *High-order linear multistep methods with general monotonicity and boundedness properties*, J. Comput. Phys. 209 (2005), 226–248

[20] J. Sand, *Circle contractive linear multistep methods*, BIT 26 (1986), 114–122.

[21] C.-W. Shu, *Total-variation-diminishing time discretizations*, SIAM J. Sci. Stat. Comp. 9 (1988), 1073–1084.

[22] M.N. Spijker, *Contractivity in the numerical solution of initial value problems*, Numer. Math. 42 (1983), 271–290.

[23] M.N. Spijker, *Stepsize restrictions for general monotonicity in numerical initial value problems*, SIAM J. Numer. Anal. 45 (2007), 1226–1245.

[24] R. Vanselov, *Nonlinear stability behaviour of linear multistep methods*, BIT 23 (1983), 388–396.