MATHEMATICS OF COMPUTATION
Volume 75, Number 254, Pages 655–672
S 0025-5718(05)01794-1
Article electronically published on November 17, 2005

ON MONOTONICITY AND BOUNDEDNESS PROPERTIES OF LINEAR MULTISTEP METHODS

WILLEM HUNDSDORFER AND STEVEN J. RUUTH

Abstract. In this paper an analysis is provided of nonlinear monotonicity and boundedness properties for linear multistep methods. Instead of strict monotonicity for arbitrary starting values we shall focus on generalized monotonicity or boundedness with Runge-Kutta starting procedures. This allows many multistep methods of practical interest to be included in the theory. In a related manner, we also consider contractivity and stability in arbitrary norms.

1. Introduction

Nonlinear monotonicity and boundedness properties are often of importance for the numerical solution of partial differential equations (PDEs) with nonsmooth solutions. This holds in particular for hyperbolic conservation laws, for which specialized spatial discretizations are often used to enforce TVD (total variation diminishing) or TVB (total variation boundedness) properties in one spatial dimension or maximum-norm bounds in more dimensions. Applying such a spatial discretization, one wants of course also to preserve such properties in the time integration of the resulting semidiscrete system.

In this paper we consider initial value problems for systems of ordinary differential equations (ODEs) in $\mathbb{R}^m$, with arbitrary $m \geq 1$,

$$w'(t) = F(w(t)), \quad w(0) = w_0 . \tag{1.1}$$

In our applications these systems will usually arise by spatial discretization of a PDE. Specifically we are interested in the discrete preservation of monotonicity and boundedness properties of numerical approximations $w_n \approx w(t_n)$, $t_n = n\Delta t$, $\Delta t > 0$, generated by linear multistep methods.

In the following it is assumed there is a maximal step size $\Delta t_{FE} > 0$ such that

$$\|v + \Delta t F(v)\| \leq \|v\| \qquad \text{for all } \ 0 < \Delta t \leq \Delta t_{FE}, \ v \in \mathbb{R}^m, \tag{1.2}$$

where $\|\cdot\|$ is a given seminorm, such as the total variation over the components, or a genuine norm, such as the maximum norm. Of course, with the forward Euler method this leads to

$$\|w_n\| \leq \|w_0\| \qquad \text{for all } \ n \geq 1 , \tag{1.3}$$

Received by the editor March 10, 2004 and, in revised form, January 6, 2005.
2000 *Mathematics Subject Classification.* Primary 65L06, 65M06, 65M20.
Key words and phrases. Multistep schemes, monotonicity, boundedness, TVD, TVB, contractivity, stability.
The work of the second author was partially supported by a grant from NSERC Canada.

whenever the step size restriction $\Delta t \leq \Delta t_{FE}$ is satisfied.

In this paper similar properties are studied for linear multistep methods

$$w_n \;=\; \sum_{j=1}^{k} a_j w_{n-j} \;+\; \sum_{j=0}^{k} b_j \Delta t F(w_{n-j})\,, \qquad n \geq k\,. \tag{1.4}$$

In the following the notation $F_{n-j} = F(w_{n-j})$ is used, and it will be assumed throughout that

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

The starting vectors $w_0, w_1, \ldots, w_{k-1}$ are either given or computed by an appropriate starting procedure, and we shall mainly deal with the property

$$\|w_n\| \;\leq\; M\,\|w_0\| \quad \text{for all} \;\; n \geq 1\,. \tag{1.6}$$

This will be referred to as *monotonicity* if $M = 1$ and as *boundedness* if $M > 1$. We shall determine constants C_{LM} such that (1.6) is valid for a multistep method with suitable starting procedure under the step size restriction $\Delta t \leq C_{LM}\Delta t_{FE}$. In our results, the size of M is determined by the coefficients of the multistep method and the specific starting procedure.

Multistep schemes of high order satisfying such boundedness properties have been constructed recently in [13]. In numerical tests these schemes proved to be superior to existing monotone multistep schemes. In this paper we provide the theoretical framework for monotonicity and boundedness properties of these schemes.

The outline of this paper is as follows. In Section 2 we briefly discuss some well-established concepts that will be generalized in this paper. Section 3 contains the main results on monotonicity and boundedness, together with examples of explicit methods with order $p = k$. In Section 4 the results are extended to include perturbations and generalizations of the assumption (1.2). Section 5 contains bounds on maximal step sizes for explicit and implicit multistep methods. Some experimental optimal bounds for classes of explicit methods are discussed in an appendix.

2. Background material

2.1. **Norms.** In this paper $\|\cdot\|$ will be an arbitrary norm, e.g., the maximum norm $\|\cdot\|_\infty$, or a seminorm, e.g., the discrete total variation $\|\cdot\|_{TV}$ over the components. For inner-product norms different results exist. For example, the G-stability property [2, 6] then gives unconditional stability for many implicit second-order schemes, including the trapezoidal rule and the implicit BDF2 scheme.

With general (semi-)norms, like $\|\cdot\|_\infty$ or $\|\cdot\|_{TV}$, much more stringent restrictions on the allowable step sizes arise, even for simple linear systems and implicit methods; see for instance the results in [17] and the experiments in [7, Sect. 5.1]. Such (semi-)norms are mainly relevant for problems with nonsmooth solutions. This is common with hyperbolic conservation laws, and the results in this paper should mainly be regarded with such applications in mind.

2.2. **Contractivity and stability.** The monotonicity and boundedness concepts for sequences of approximations can also be reformulated to deal with the difference of two sequences. Such results will be considered for an ODE system

$$v'(t) = G(v(t))\,, \qquad v(0) = v_0\,, \tag{2.1}$$

where it is assumed that

$$\|\tilde{v} - v + \Delta t\big(G(\tilde{v}) - G(v)\big)\| \;\le\; \|\tilde{v} - v\| \quad \text{for } \tilde{v}, v \in \mathbb{R}^m,\ 0 < \Delta t \le \Delta t_{FE}\,. \tag{2.2}$$

Suppose an appropriate Runge-Kutta starting procedure is used to generate $v_1, \ldots, v_{k-1}$ from the given v_0, and subsequent approximations v_n are computed by the linear multistep method. Along with the sequence $\{v_n\}$ we also consider $\{\tilde{v}_n\}$ starting with a perturbed $\tilde{v}_0$ and possibly a different starting procedure. Let $w_n = \tilde{v}_n - v_n$ and $F_n = G(\tilde{v}_n) - G(v_n)$. For these differences we still have recursion (1.4), and consequently (1.6) then gives *contractivity* if $M = 1$. For $M \ge 1$ we get stability with respect to initial perturbations.

For nonlinear semidiscrete hyperbolic equations, with suitable norm, it may happen that assumption (1.6) is valid whereas (2.2) does not hold. By means of compactness arguments it can then still be possible to prove convergence; see for example [12, Sect. 12.12]. For that reason, property (1.6) is also sometimes referred to as (nonlinear) stability and methods satisfying $\|w_n\| \le \max_{j<n} \|w_j\|$ are nowadays often called strong-stability preserving (SSP). Of course, for linear problems the assumptions (1.3) and (2.2) are equivalent.

2.3. **Arbitrary starting values.** Results concerning contractivity and monotonicity (TVD/SSP) for methods with nonnegative coefficients can be found in [4, 10, 11, 14, 16, 17, 18]. Suppose that all $a_j, b_j \ge 0$, and for such methods let

$$K_{LM} \;=\; \min_{1\le j\le k} \frac{a_j}{b_j}\,, \tag{2.3}$$

with the convention $a/0 = +\infty$ if $a \ge 0$. Then it is easy to show that we have

$$\|w_n\| \;\le\; \max_{0\le j\le k-1} \|w_j\| \qquad \text{for all } \; n \ge k \tag{2.4}$$

under the step size restriction $\Delta t \le K_{LM}\Delta t_{FE}$. This holds for arbitrary starting values for the multistep scheme. However, the methods with nonnegative coefficients form a small class, and the step size requirement $\Delta t \le K_{LM}\Delta t_{FE}$ can be very restrictive. For example, it was shown in [10] that for an explicit k-step method ($k > 1$) of order p we have $K_{LM} \le (k-p)/(k-1)$. The most interesting explicit methods have $p = k$, so then we cannot have $K_{LM} > 0$. For implicit methods of order $p \ge 2$ we have $K_{LM} \le 2$; see [11] and also Section 5.

The commonly known classes of methods, such as the Adams or BDF-type methods, are not included in this theory since some of the coefficients a_j, b_j are negative. However, it was shown in [7] that the boundedness property (1.6) may hold for such methods if the starting values $w_1, \ldots, w_{k-1}$ are generated from w_0 by a consistent starting procedure. For a given multistep method, the constant M in (1.6) will be determined by the starting procedure; see Section 3.3. With special starting procedures and a modified step size restriction we can still have $M = 1$. As we shall see, such boundedness results with starting procedures do apply to many multistep methods of practical interest.

Methods with such monotonicity or boundedness properties and optimal step size restrictions were recently constructed in [13]. In that paper numerical tests showed much improvement in computational efficiency over the class of methods with nonnegative coefficients. An analysis for two-step methods was presented in [7], together with some (partial) results on explicit Adams and BDF methods. This paper provides a general framework to study the monotonicity and boundedness properties of linear k-step methods with starting procedures.

3. Monotonicity and boundedness with starting procedures

To derive monotonicity and boundedness results for linear multistep methods, we begin with a reformulation of the schemes for theoretical purposes. With this reformulation we shall see the influence of the starting procedures on the results for the multistep methods.

3.1. **Reformulations and main results.** Consider the k-step method (1.4) and let $\theta_1, \theta_2, \ldots$ be a bounded sequence of nonnegative parameters. We denote

$$\Theta_j = \prod_{i=1}^{j} \theta_i \quad \text{for } j > 0\,, \qquad \Theta_0 = 1\,, \qquad \Theta_j = 0 \quad \text{for } j < 0\,. \tag{3.1}$$

By subtracting $\theta_1 w_{n-1}$ from the right-hand side of (1.4) and then adding this term but using the recursion, the k-step method is written as an equivalent $(k+1)$-step method with a free parameter. Continuing this way, by subtracting and adding $\Theta_j w_{n-j}$, $j = 2, \ldots, n-k$, substituting w_{n-j} in terms of $w_{n-j-1}, \ldots, w_{n-j-k}$, and collecting terms, it follows that

$$\begin{aligned} w_n \,-\, b_0 \Delta t F_n = & \sum_{j=1}^{n-k} \big(\alpha_j w_{n-j} + \beta_j \Delta t F_{n-j} \big) \\ & + \sum_{j=n-k+1}^{n} \big(\alpha_{n,j}^R w_{n-j} + \beta_{n,j}^R \Delta t F_{n-j} \big)\,, \end{aligned} \tag{3.2a}$$

where the coefficients α_j, β_j are given by

$$\alpha_j = \sum_{i=1}^{k} a_i \Theta_{j-i} \,-\, \Theta_j\,, \qquad \beta_j = \sum_{i=0}^{k} b_i \Theta_{j-i} \tag{3.2b}$$

for all $j \geq 1$, and the coefficients of the remainder term are

$$\alpha_{n,j}^R = \sum_{i=k-n+j}^{k} a_i \Theta_{j-i}\,, \qquad \beta_{n,j}^R = \sum_{i=k-n+j}^{k} b_i \Theta_{j-i} \tag{3.2c}$$

for $n-k+1 \leq j \leq n$. To verify that (3.2) holds for $n \geq k$, first observe that it is valid for $n = k$ (in which case $\alpha_{n,j}^R = a_j$, $\beta_{n,j}^R = b_j$), and then use induction with respect to n together with partial summation. Note that by the construction we still have

$$\sum_{j=1}^{n-k} \alpha_j \;+\; \sum_{j=n-k+1}^{n} \alpha_{n,j}^R \;=\; 1\,, \qquad n \geq k\,, \tag{3.3}$$

in view of the consistency relation in (1.5).

In the following we consider parameter sequences $\{\theta_i\}$ satisfying

$$\theta_j \geq 0 \quad \text{for all } j \geq 1\,, \qquad \theta_j = \theta_* \quad \text{for } j > l \tag{3.4}$$

with some $l \geq 0$. The parameters will be selected such that

$$\alpha_j \geq 0\,, \quad \beta_j \geq 0 \qquad \text{for all } \; j \geq 1\,, \tag{3.5a}$$

and for such a parameter sequence we define

$$\gamma_{LM} \;=\; \min_{j \geq 1} \; \frac{\alpha_j}{\beta_j}\,. \tag{3.5b}$$

The dependence on the choice of the θ_i is omitted in the notation. The optimal value for γ_{LM} over parameter sequences (3.4) will be denoted by C_{LM}. Such optimal values will generally depend on the range for θ_* that will be allowed. The restriction $\theta_j = \theta_*$, $j > l$, was imposed for practical optimization purposes in [13], and it will also be convenient in the analysis; with this restriction the signs of α_j, β_j and the size of the ratios α_j/β_j in (3.5) need only be taken into account for $j \le k + l$.

Theorem 3.1. *Consider a k-step method* (1.4). *Let* γ_{LM} *be given by* (3.5) *with* $\theta_* < 1$. *Assume* $w_1, \ldots, w_{k-1}$ *are computed from* w_0 *by a Runge-Kutta starting procedure. Then there is an* $M \ge 1$, *determined by the starting procedure, such that*

$$\|w_n\| \ \le \ M \, \|w_0\| \qquad \textit{for} \ \ n \ge 1, \ \ \Delta t \le \gamma_{LM} \Delta t_{FE} \, .$$

The proof of this theorem is given in Section 3.3. As we shall see, the assumption $\theta_* < 1$ is related to zero-stability of the multistep method. With regard to the size of M, we note already that in experiments in [13] bounds very close to 1 were found if $w_1, \ldots, w_{k-1}$ are computed from w_0 with standard Runge-Kutta starting procedures. The bound $M = 1$ can sometimes be enforced by selecting special procedures, and, possibly, a modified step size restriction. See Remark 3.6 for additional comments.

In [13] optimal values C_{LM} for the γ_{LM} in (3.5) were found numerically for given step numbers k and order p. For several interesting cases this led to a sequence $\{\theta_i\}$ with $\theta_{l+1} = 0$ for some $l \ge 0$, that is, $\theta_* = 0$. In such a situation another generalization of (2.4) can be formulated.

Theorem 3.2. *Consider a k-step method* (1.4). *Let* γ_{LM} *be given by* (3.5) *where* $\theta_{l+1} = 0$ *for some* $l \ge 0$. *Then*

$$\|w_n\| \ \le \ \max_{0 \le j \le k+l-1} \|w_j\| \qquad \textit{for} \ \ n \ge k + l, \ \ \Delta t \le \gamma_{LM} \Delta t_{FE} \, .$$

Proof. If $\theta_{l+1} = 0$, then also $\alpha_j, \beta_j = 0$ for $j > k + l$. The reformulation (3.2) then reduces to

$$w_n \ - \ b_0 \Delta t F_n \ = \ \sum_{j=1}^{k+l} \big(\, \alpha_j w_{n-j} + \beta_j \Delta t F_{n-j} \, \big) \tag{3.6}$$

for $n \ge k + l$. By simple arguments it follows from (1.2) that

$$\|w_n\| \ \le \ \|w_n - b_0 \Delta t F_n\| \tag{3.7}$$

(see for example [7, p. 614]); this is just unconditional monotonicity of the backward Euler method. The proof now follows directly from (3.6). □

3.2. **Examples.** Optimal values for the γ_{LM} in (3.5), for a given linear multistep method, were denoted as C_{LM} in [13]. Such optimal values are often called *threshold values*. Here we shall distinguish the threshold values C^1_{LM} for $\theta_* \in [0, 1)$ (relevant for Theorem 3.1) and C^0_{LM} for $\theta_* = 0$ (relevant for Theorem 3.2). Mathematically this involves all possible integers $l \ge 0$. Numerical optimal values are found by selecting a fixed, large l, and the resulting optimization is then carried out by using the BARON optimization package [1]; see [13] for details.

As an example we consider here explicit two- and three-step methods with order $p = k$. We saw already in Section 2 that for such methods nonnegativity of all coefficients a_j, b_j and $K_{LM} > 0$ is not possible.

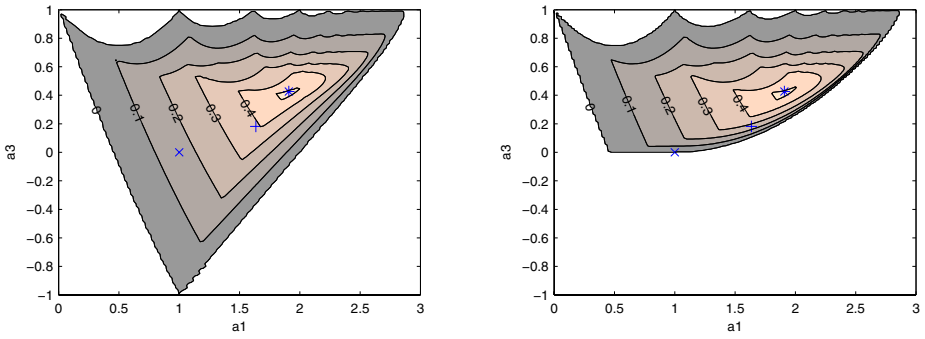

FIGURE 1. Threshold values C^1_{LM} and C^0_{LM} for explicit third-order three-step schemes. Contour levels: $0.0, 0.1, \ldots, 0.5$. Markers: $\times$ for AB3, $+$ for eBDF3 and $*$ for TVB_0(3,3).

For the explicit second-order two-step methods the optimal threshold values C^1_{LM} were obtained in [7] by choosing constant θ_j, which turned out to be optimal for this class of methods. Well-known examples are the two-step Adams-Bashforth method (AB2, $C^1_{LM} = \frac{4}{9}$) and the extrapolated BDF2 scheme (eBDF2, $C^1_{LM} = \frac{5}{8}$). As we shall see, positive threshold values with $\theta_* = 0$ are not possible for this class of methods.

The threshold values for explicit third-order three-step methods, with constraints $\theta_* < 1$ and $\theta_* = 0$, are given in Figure 1. This class of methods forms a two-parameter family, and here we use the coefficients a_1, a_3 as free parameters. Zero-stability of these methods is valid for (a_1, a_3) in a triangle with vertices $(-1, 1)$, $(1, -1)$ and $(3, 1)$. Close to the edge connecting $(1, -1)$ and $(3, 1)$ the methods have large error constants [5] and also the numerical optimizations for C^1_{LM} are not very accurate there.

This class of methods with $p = k = 3$ contains, for instance, the well-known three-step Adams-Bashforth method (AB3, $C^1_{LM} \approx 0.16$) and the three-step extrapolated backward differentiation formula (eBDF3, $C^1_{LM} \approx 0.39$). Also marked in the figure is the optimal method TVB_0(3,3) from [13], which has $C^1_{LM} = C^0_{LM} \approx 0.53$. It is surprising that for many methods in the upper half of the figures there is little difference between C^0_{LM} and C^1_{LM}. In particular, numerical optimization of C^1_{LM} (with $\theta_* \in [0, 1)$) produced the method TVB_0(3,3) for which $\theta_* = 0$.

Some general necessary conditions for having positive thresholds C^1_{LM} and C^0_{LM} will be presented in Section 5. Here we mention that $a_k > 0$ is necessary for having $C^0_{LM} > 0$, as suggested already by Figure 1 for the case $k = 3$.

3.3. **Technical results.** We consider a sequence $\{\theta_i\}$ as in (3.4) with limit point θ_*, such that all α_j, $\beta_j \geq 0$. The resulting γ_{LM} in (3.5) need not necessarily be an optimal value C_{LM}, although for applications that will be the most interesting situation.

First note that if $\theta_{l+1} = 0$ for some $l \geq 0$, then we can take all subsequent θ_j to be zero, because the coefficients in (3.2) will not be affected by these θ_j. Therefore there are effectively two cases: $\theta_* = 0$ and $\theta_* > 0$, and in the latter case we may assume that $\theta_j > 0$ for all $j \geq 1$. Further note that the coefficients α_j, β_j would

grow exponentially for $j > l$ if $\theta_* > 1$. It will be shown below that this cannot happen with a zero-stable scheme.

3.3.1. *Generating polynomials.* To establish a relation between the assumptions in Theorems 3.1, 3.2 and more commonly known properties of linear multistep methods, consider the polynomials

$$\rho(\zeta) = \zeta^k - \sum_{j=1}^{k} a_j \zeta^{k-j}, \qquad \sigma(\zeta) = \sum_{j=0}^{k} b_j \zeta^{k-j}. \tag{3.8}$$

Since $\rho(1) = 0$, according to the consistency relation (1.5), we can write

$$\rho(\zeta) = (\zeta - 1)\hat{\rho}(\zeta) \tag{3.9}$$

with $\hat{\rho}$ a polynomial of degree $k-1$. If $F \equiv 0$ the multistep recursion (1.4) has ρ as its characteristic polynomial. The method is called *zero-stable* if all roots of ρ have modulus at most one and the roots of modulus one are simple. This means that the scheme is stable for $F \equiv 0$ with arbitrary initial values, and this also gives stability for nonstiff problems; see for instance [5]. Because by zero-stability no roots of ρ are outside the unit circle, and $\rho(\theta) > 0$ for large positive θ, it is obvious that zero-stability implies

$$\hat{\rho}(\theta) > 0, \quad \rho(\theta) \geq 0 \quad \text{whenever } \theta \geq 1. \tag{3.10}$$

For any $j \geq k$, the coefficients α_j, β_j can be written in terms of Θ_{j-k} and $\theta_{j-k+1}, \ldots, \theta_j$. If $j \geq k + l$ it is easily seen that

$$\alpha_j = -\Theta_{j-k}\, \rho(\theta_*), \qquad \beta_j = \Theta_{j-k}\, \sigma(\theta_*). \tag{3.11}$$

For a zero-stable method, having $\alpha_j \geq 0$ thus implies $\theta_* \leq 1$. Moreover, we see that $\theta_* = 1$ will give $\alpha_j = 0$. In that case we can still have $\gamma_{LM} > 0$, provided also $\beta_j = 0$, but we shall see below that this case is not very interesting for practical purposes.

If the polynomials ρ and σ do not have a common root, the method is said to be *irreducible* [5]. Reducible methods are not used in practice since the asymptotic properties are the same as for the $(k-1)$-step method that results by dividing out the common factor of ρ, σ. In this paper reducible methods do appear, for example in the proof of Theorem 3.2, but these are only for theoretical purposes, not for actual computations.

Lemma 3.3. *Suppose the method* (1.4) *is irreducible and* $\gamma_{LM} > 0$. *Then*

$$\rho(\theta_*) < 0, \quad \sigma(\theta_*) \geq 0 \quad \textit{and} \quad 0 < \gamma_{LM} \leq -\frac{\rho(\theta_*)}{\sigma(\theta_*)}.$$

If the method is also zero-stable, then $\theta_* < 1$.

Proof. Consider the index $j = k + l$, so that $\Theta_{j-k} \neq 0$ (even if $\theta_* = 0$). If $\gamma_{LM} > 0$, then $\alpha_j, \beta_j \geq 0$ and $\alpha_j = 0$ only if $\beta_j = 0$. But ρ and σ have no common roots, and thus $\alpha_j > 0$. The proof now follows directly from (3.10) and (3.11). □

We note that the upper bound for γ_{LM} in this lemma does not always provide a useful estimate. For example, with the two-step methods of order two, the θ_* was chosen in [7] such that $\sigma(\theta_*) = 0$. Other upper bounds for γ_{LM} (and for the optimal C_{LM}) are given in Section 5.

3.3.2. *Proof of Theorem* 3.1. To prove Theorem 3.1, we start with a technical result with concrete conditions on the starting values. Subsequently, these conditions will be analyzed.

Let $M \geq 1$, and consider the following conditions on the starting values,

$$\|w_j\| \;\leq\; M\,\|w_0\| \qquad \text{for} \;\; j = 1, \ldots, k-1 \tag{3.12}$$

and

$$\Big\| \sum_{i=0}^{k-1} \big(\alpha^R_{n,n-i} w_i + \beta^R_{n,n-i} \Delta t F_i\big) \Big\| \;\leq\; \sum_{i=0}^{k-1} \alpha^R_{n,n-i}\, M\,\|w_0\| \qquad \text{for} \;\; n \geq k\,. \tag{3.13}$$

We note that for a sequence satisfying (3.4) these inequalities need only to be verified for $n = k, k+1, \ldots, 2k+l-1$. The size of the constant M will depend on the starting procedure that is used to generate $w_1, \ldots, w_k$ from w_0.

Lemma 3.4. *Consider method* (1.4) *with* γ_{LM} *given by* (3.5). *Assume* (3.12), (3.13) *with* $M \geq 1$ *and* $\Delta t \leq \gamma_{LM} \Delta t_{FE}$. *Then the boundedness property* (1.6) *holds.*

Proof. From (3.2), (3.5) and (3.13) we obtain

$$\|w_n - b_0 \Delta t F_n\| \;\leq\; \sum_{j=1}^{n-k} \alpha_j\,\|w_{n-j}\| \;+\; \sum_{j=n-k+1}^{n} \alpha^R_{n,j}\, M\,\|w_0\|\,,$$

and by the assumption (3.12) the theorem is valid for $n \leq k-1$. Using (3.3) and (3.7), the proof thus follows directly by induction. □

To study the starting condition (3.13) we may assume that $\theta_* > 0$; otherwise we are in a situation where Theorem 3.2 applies. Let us denote

$$\delta_{n-k+1} \;=\; \sum_{i=0}^{k-1} \alpha^R_{n,n-i}\,, \qquad n \geq k\,. \tag{3.14}$$

Then we want to know that all δ_j $(j \geq 1)$ are positive, or at least nonnegative, in order to see whether (3.13) can be satisfied. For this, first note that

$$\delta_1 \;=\; 1\,, \qquad \delta_{j+1} \;=\; \delta_j - \alpha_j \quad \text{for all } j \geq 1\,. \tag{3.15}$$

This last relation easily follows from (3.3). As a consequence we thus know that the sequence $\{\delta_j\}$ is nonincreasing in j.

For $j \geq k$, the δ_j can be written in terms of Θ_{j-k} and $\theta_{j-k+1}, \ldots, \theta_{j-1}$. Hence for $j \geq k+l$ we have

$$\delta_{j+1} \;=\; \theta_*\,\delta_j\,, \tag{3.16}$$

and in view of (3.11), (3.15) it thus also follows that

$$\alpha_j \;=\; (1-\theta_*)\delta_j\,, \qquad \delta_j \;=\; \Theta_{j-k}\,\hat{\rho}(\theta_*)\,. \tag{3.17}$$

Combining this with Lemma 3.3 and (3.10) directly yields the following result.

Lemma 3.5. *Suppose method* (1.4) *is irreducible, zero-stable and* $\theta_* > 0$, $\gamma_{LM} > 0$. *Then* $\theta_* < 1$ *and* $\delta_j > 0$ *for all* $j \geq 1$.

As observed before, for a sequence (3.4) we get $k+l$ inequalities in (3.13), and the coefficients in the right-hand side are $\delta_1, \ldots, \delta_{k+l}$. If all these $\delta_j > 0$, then condition (3.13) can be fulfilled for any Runge-Kutta starting procedure with $\Delta t \leq \gamma_{LM} \Delta t_{FE}$ for some (sufficiently large) constant $M \geq 1$. This gives the proof of Theorem 3.1.

Remark 3.6. A quantification of M can be given for any specific starting procedure of Runge-Kutta type by using the inequality

$$\max_{\Delta t \le C\Delta t_{FE}} \|v + s\Delta t F(v)\| \;\le\; \max\big(1, |2Cs-1|\big)\,\|v\| \tag{3.18}$$

for $C > 0$, $s \in \mathbb{R}$, and $v \in \mathbb{R}^m$; see also [7, Rem. 3.2]. However such computed bounds for M were found to be much larger than experimental values in numerical tests. We will therefore not elaborate on such estimates.

Furthermore, we note that with an M that is specified in advance, for instance $M = 1$, conditions on the starting procedure and extra conditions on the time step may arise in order to fulfill (3.13). Examples for this can be found in [7]; in numerical tests such additional restrictions were found to be less relevant than the primary time step restriction $\Delta t \le \gamma_{LM}\Delta t_{FE}$ with optimal $\gamma_{LM} = C^1_{LM}$.

Remark 3.7. We can allow $\theta_* = 1$ in Lemma 3.4, but that does not yield results of practical interest. As an example, consider the two-step method

$$w_n \;=\; 2w_{n-1} - w_{n-2} + \Delta t F_{n-1} - \Delta t F_{n-2}\,.$$

This method is not zero-stable, since ρ has double root 1, but taking all $\theta_j = 1$ gives in fact monotonicity with $\gamma_{LM} = 1$ under the starting condition

$$\|w_1 - w_0 - \Delta t F_0\| \;\le\; 0\,,$$

which means of course that w_1 has to be computed by the forward Euler method. Having boundedness or monotonicity for an unstable method may seem contradictory, but it should be realized that the above method is reducible: if w_1 is computed by forward Euler, then the whole sequence $\{w_n\}$ is a forward Euler sequence. Formally the method is second-order consistent, but because of the weak instability it is only first-order convergent.

4. Generalizations

The above results allow various generalizations. Here we discuss the inclusion of perturbations, and the replacement of assumption (1.2) by boundedness assumptions on finite time intervals.

4.1. **Inclusion of perturbations.**

Instead of the multistep recursion (1.4) we can also consider a perturbed version

$$w_n \;-\; b_0\Delta t F_n \;=\; \sum_{j=1}^{k} \big(a_j w_{n-j} + b_j \Delta t F_{n-j}\big) \;+\; d_n\,, \qquad n \ge k\,, \tag{4.1}$$

with perturbations d_n on each step. In the following theorem the influence of these perturbations will be bounded by

$$S \;=\; \sum_{j=0}^{\infty} \Theta_j\,. \tag{4.2}$$

Note that this S will be a finite number for any sequence (3.4) with $\theta_* < 1$.

Theorem 4.1. *Consider method* (1.4) *with* γ_{LM} *given by* (3.5). *Assume the starting conditions* (3.12), (3.13) *are valid with* $M \ge 1$ *and* $\Delta t \le \gamma_{LM}\Delta t_{FE}$. *Then the solution of* (4.1) *can be bounded by*

$$\|w_n\| \;\le\; M\,\|w_0\| \;+\; (n-k+1)S \max_{k\le j\le n} \|d_j\|\,, \qquad n \ge k\,.$$

Proof. The reformulation for (4.1) becomes

$$
\begin{aligned}
w_n \;-\; b_0 \Delta t F_n = & \sum_{j=1}^{n-k} \left(\alpha_j w_{n-j} + \beta_j \Delta t F_{n-j} \right) \\
& + \sum_{j=n-k+1}^{n} \left(\alpha_{n,j}^R w_{n-j} + \beta_{n,j}^R \Delta t F_{n-j} \right) \;+\; \sum_{j=0}^{n-k} \Theta_j \, d_{n-j} \,,
\end{aligned} \tag{4.3}
$$

for $n \geq k$. Under the conditions of Lemma 3.4 we thus obtain

$$
\|w_n\| \;\leq\; \sum_{j=1}^{n-k} \alpha_j \, \|w_{n-j}\| \;+\; \delta_{n-k+1} \, M \, \|w_0\| \;+\; \sum_{j=0}^{n-k} \Theta_j \, \|d_{n-j}\| \,,
$$

where $\alpha_1 + \cdots + \alpha_{n-k} + \delta_{n-k+1} = 1$ for $n > k$, and $\delta_1 = 1$. Hence

$$
\|w_n\| \;\leq\; (1 - \delta_{n-k+1}) \max_{j \leq n-1} \|w_j\| \;+\; \delta_{n-k+1} \, M \, \|w_0\| \;+\; S \max_{j \leq n} \|d_j\| \,.
$$

By induction with respect to $n = k, k+1, \ldots$, the result easily follows. □

A similar result can be derived for differences of two sequences, $w_n = \tilde{v}_n - v_n$, with an equation $v'(t) = G(v(t))$ satisfying (2.2). If we take v_n as an unperturbed multistep result and $\tilde{v}_n = v(t_n)$, then the d_n will represent local truncation errors. For a pth-order method these will be $d_n = \mathcal{O}(\Delta t^{p+1})$, provided the solution is sufficiently smooth. The above result thus gives stability and convergence in general norms such as the maximum norm. This provides a generalization of results in [14, 18] for schemes with nonnegative coefficients, for which we can take $\theta_j \equiv 0$ and $M = S = 1$.

Remark 4.2. We can compare such stability-convergence results with classical estimates based on a Lipschitz condition, as found in [5], for example. For this, note that (2.2) implies

$$
\|G(\tilde{v}) - G(v)\| \;\leq\; L \, \|\tilde{v} - v\|
$$

with $L = 2/\Delta t_{FE}$. The standard stability results will involve bounds with $\exp(Lt_n)$. If such a Lipschitz condition is valid for a hyperbolic PDE, we will have $\Delta t_{FE} \sim \Delta x$, where Δx is the mesh width in space, and estimates with $\exp(Lt_n)$ are then completely useless. Our results, on the other hand, lead to reasonable stability bounds under a CFL restriction on $\Delta t / \Delta x$, with constants M and S that are independent of the mesh width Δx.

4.2. **Generalized boundedness assumptions.** In semidiscretizations of scalar conservation laws the monotonicity assumption (1.2) can be valid if a so-called TVD-limiter is used. Such limiters do not distinguish between genuine extrema and numerically induced extrema caused by oscillations. Consequently numerical diffusion must be added locally near genuine extrema to maintain the TVD property, leading to significant errors. To reduce this dissipation (at the cost of potentially introducing small oscillations) more relaxed limiters are often used such as the TVB-limiter of [15].

To generalize our results to these systems and others exhibiting growth, we consider the assumption

$$
\|v + \Delta t F(v)\| \;\leq\; (1 + c\Delta t) \, \|v\| \;+\; \kappa \Delta t \tag{4.4}
$$

for arbitrary $v \in \mathbb{R}^m$ and $0 < \Delta t \le \Delta t_{FE}$, where $c, \kappa \ge 0$. For the forward Euler method $w_n = w_{n-1} + \Delta t F_{n-1}$ with $\Delta t \le \Delta t_{FE}$ it then easily follows that

$$\|w_n\| \;\le\; \mathrm{e}^{c\,t_n}\,\|w_0\| \;+\; \frac{1}{c}\left(\mathrm{e}^{c\,t_n} - 1\right)\kappa\,, \qquad n \ge 1\,,$$

with the convention $\frac{1}{c}(\mathrm{e}^{ct} - 1) = t$ in the case $c = 0$. This gives boundedness on finite time intervals $[0, T]$. Here we derive a similar result for multistep methods. For simplicity we consider (1.4) without perturbations. The generalization (4.4) was recently considered in [3] for boundedness results with Runge-Kutta methods. We also remark that the TVB-limiters of Shu [15] can now be included by choosing $\kappa > 0$.

Theorem 4.3. *Consider method* (1.4) *with* $\gamma_{LM} > 0$ *given by* (3.5). *Assume the starting conditions* (3.12), (3.13) *are satisfied with* $M \ge 1$ *and* $\Delta t \le \gamma_{LM}\Delta t_{FE}$. *For implicit methods, assume also* $\Delta t \le \Delta t^*$ *where* $b_0 c \Delta t^* < 1$. *Then there are* $M^* \ge 1$, $c^*, \kappa^* \ge 0$ *such that*

$$\|w_n\| \;\le\; \mathrm{e}^{c^* t_{n-k+1}} M^*\,\|w_0\| \;+\; \frac{1}{c^*}\left(\mathrm{e}^{c^* t_{n-k+1}} - 1\right)\kappa^*\,, \qquad n \ge k\,.$$

For explicit methods we can take $M^* = M$, $c^* = c/\gamma_{LM}$ *and* $\kappa^* = \kappa/\gamma_{LM}$. *For implicit methods the* M^*, c^*, κ^* *are determined by* $M, c, \kappa, \gamma_{LM}$ *and* Δt^*.

Proof. Let $v_n = w_n - b_0 \Delta t F_n$ and denote $c' = c/\gamma_{LM}$, $\kappa' = \kappa/\gamma_{LM}$. By the reformulation (3.2a) we then obtain

$$\begin{aligned}\|v_n\| &\le \sum_{j=1}^{n-k} \|\alpha_j w_{n-j} + \beta_j \Delta t F_{n-j}\| \;+\; \delta_{n-k+1} M \|w_0\| \\ &\le \sum_{j=1}^{n-k} \alpha_j \big((1 + c'\Delta t)\|w_{n-j}\| + \kappa' \Delta t\big) \;+\; \delta_{n-k+1} M \|w_0\|\end{aligned}$$

for all $n \ge k$. Since $\sum_{j=1}^{n-k} \alpha_j + \delta_{n-k+1} = 1$, it follows that

$$\begin{aligned}\|v_n\| &\le (1 - \delta_{n-k+1})(1 + c'\Delta t) \max_{j<n} \|w_j\| \\ &\qquad + (1 - \delta_{n-k+1})\kappa' \Delta t \;+\; \delta_{n-k+1} M \|w_0\|\,.\end{aligned} \tag{4.5}$$

Let us first consider explicit methods, where $v_n = w_n$. Consider the induction assumption

$$\|w_j\| \;\le\; \mathrm{e}^{c' t_{j-k+1}} M \|w_0\| \;+\; \frac{1}{c'}\big(\mathrm{e}^{c' t_{j-k+1}} - 1\big)\,\kappa'\,, \tag{4.6}$$

which is valid for $j = k - 1$. Assuming it to hold for $j = k, \ldots, n-1$, we obtain

$$\begin{aligned}\|w_n\| &\le (1 - \delta_{n-k+1})\Big(\mathrm{e}^{c' t_{n-k+1}} M \|w_0\| + \mathrm{e}^{c'\Delta t}\,\frac{1}{c'}\big(e^{c' t_{n-k}} - 1\big)\kappa'\Big) \\ &\quad + (1 - \delta_{n-k+1})\kappa' \Delta t \;+\; \delta_{n-k+1} M \|w_0\|\,,\end{aligned}$$

and consequently

$$\|w_n\| \;\le\; \mathrm{e}^{c' t_{n-k+1}} M \|w_0\| \;+\; \mathrm{e}^{c'\Delta t}\,\frac{1}{c'}\big(\mathrm{e}^{c' t_{n-k}} - 1\big)\kappa' \;+\; \kappa' \Delta t\,,$$

from which it follows that (4.6) also holds for $j = n$.

Next, consider implicit methods. We have

$$\|w_j\| \;\le\; \frac{1}{1-b_0c\Delta t}\,\|v_j\| \;+\; \frac{b_0\kappa\Delta t}{1-b_0c\Delta t}\,. \tag{4.7}$$

This relation easily follows from

$$\Big(1+\frac{b_0\Delta t}{\Delta t_{FE}}\Big)w_j \;=\; v_j \;+\; \frac{b_0\Delta t}{\Delta t_{FE}}\big(w_j+\Delta t_{FE}F(w_j)\big)\,,$$

$$\Big(1+\frac{b_0\Delta t}{\Delta t_{FE}}\Big)\|w_j\| \;\le\; \|v_j\| \;+\; \frac{b_0\Delta t}{\Delta t_{FE}}\big((1+c\Delta t_{FE})\|w_j\|+\kappa\Delta t_{FE}\big)\,.$$

Combining (4.5) and (4.7) gives

$$\begin{aligned}\|w_n\| \le{}& (1-\delta_{n-k+1})\,\frac{1+c'\Delta t}{1-b_0c\Delta t}\,\max_{j<n}\|w_j\| \\ &+\;(1-\delta_{n-k+1})\,\frac{\kappa'\Delta t}{1-b_0c\Delta t}\;+\;\delta_{n-k+1}\,\frac{M\|w_0\|}{1-b_0c\Delta t}\;+\;\frac{b_0\kappa\Delta t}{1-b_0c\Delta t}\,.\end{aligned}$$

Taking $M^*=M/(1-b_0c\Delta t^*)$, we can select $c^*\ge c$, $\kappa^*\ge\kappa$ such that

$$\|w_n\| \;\le\; (1-\delta_{n-k+1})\,\mathrm{e}^{c^*\Delta t}\max_{j<n}\|w_j\| \;+\; \kappa^*\Delta t \;+\; \delta_{n-k+1}M^*\|w_0\|\,,$$

which leads as before to the desired estimate. □

5. Upper bounds for the threshold values

In this section we consider some additional points related to the maximal values C_{LM} for the γ_{LM} in (3.5) with parameter sequences $\{\theta_j\}$ satisfying (3.4). As in Section 3.3, we shall distinguish the thresholds C^0_{LM} with $\theta_*=0$ and C^1_{LM} with $\theta_*\in[0,1)$. Of course, we always have $C^0_{LM}\le C^1_{LM}$.

5.1. **Stability regions.** The basic equation for linear stability considerations is the scalar complex test equation $w'(t)=\lambda w(t)$. This can also be converted to an equivalent system in $\mathbb{R}^2$ to remain formally within the class of real equations (1.1). The *stability region* $\mathcal{S}$ consists of those $z=\Delta t\lambda\in\mathbb{C}$ for which the multistep scheme will be stable for arbitrary starting values. We can bound C^1_{LM} in terms of the largest disc $\mathcal{D}_r=\{z\in\mathbb{C}:|z+r|\le r\}$ fitting in the stability region.

For the test equation $w'(t)=\lambda w(t)$, the monotonicity assumption (1.2) will hold provided $z=\Delta t\lambda\in\mathcal{D}_1$. If $\theta_*<1$, then we know that the starting conditions (3.12), (3.13) can be satisfied for any set of starting values by adjusting M, showing stability for $\Delta t\le C^1_{LM}\Delta t_{FE}$ of the multistep recursion, and thus $C^1_{LM}z\in\mathcal{S}$. Consequently, $\mathcal{D}_r\subset\mathcal{S}$ for $r=C^1_{LM}$.

This implies for example that no $C^1_{LM}>0$ exists for the explicit two-step midpoint (leap-frog) method or the Nyström methods; see also [7, Rem. 4.3].

It was shown in [9] that $\mathcal{D}_r\subset\mathcal{S}$ implies $r\le1$ for explicit methods, with equality $r=1$ only for the forward Euler method. The same thus holds for C^1_{LM}.

Note that this general upper bound $C^1_{LM}\le1$ is the same as the upper bound $K_{LM}\le1$ for explicit methods with nonnegative coefficients. However, whereas $K_{LM}>0$ does not hold for most methods of practical interest, the class of methods with $C^1_{LM}>0$ is much larger and it does include many useful methods.

5.2. **Positive threshold values.** Application of Lemma 3.3 with $\theta_* = 0$ shows that

$$C^0_{LM} > 0 \qquad \Longrightarrow \qquad a_k > 0\,, \quad b_k \geq 0\,, \tag{5.1}$$

and if $a_k > 0$, $b_k \geq 0$, then $C^0_{LM} \leq a_k/b_k$. For zero-stable methods with order $p = k$ this necessary condition for $C^0_{LM} > 0$ cannot hold if $k = 2$, see [7], and the numerical optimizations in [13] indicate that this is also the case with $k = 4$, 6. For $k = 3$, 5, on the other hand, these numerical optimizations did produce schemes with $\theta_* = 0$ when trying to optimize C^1_{LM} for a given step number k and order p, leading for instance to the $\mathrm{TVB}_0(3,3)$ scheme discussed in Section 3.3.

The upper bound for C^1_{LM} obtained from Lemma 3.3 with $\theta \in [0,1)$ does in general not provide a useful estimate. For explicit methods the condition $\mathcal{D}_r \subset \mathcal{S}$ for $r = C^1_{LM}$ often gives a much better bound, though usually not sharp, while for implicit A-stable methods this does not yield a useful bound. Here we give some simple but useful upper bounds based on the first few α_j, β_j.

With explicit methods we have $\alpha_1 = a_1 - \theta_1$, $\beta_1 = b_1$ and $\beta_2 = b_2 + b_1\theta_1$. To have $\beta_2 \geq 0$ we need $\theta_1 \geq -b_2/b_1$, and therefore

$$C^1_{LM} \leq \frac{\alpha_1}{\beta_1} \;\leq\; \frac{a_1 + b_2/b_1}{b_1} \;=\; \frac{1}{b_1^2}\bigl(a_1b_1 + b_2\bigr)\,. \tag{5.2}$$

This was used in [7] to guarantee the optimality of the threshold values C^1_{LM} found with constant θ_j for explicit second-order two-step methods. As a consequence of (5.2) we have for explicit methods the necessary condition

$$C^1_{LM} > 0\,, \quad b_0 = 0 \qquad \Longrightarrow \qquad a_1 > 0\,, \quad b_1 \geq 0 \quad a_1b_1 + b_2 > 0\,. \tag{5.3}$$

This result was used in [7, 13] to show that there is no positive threshold value for the explicit Adams methods with $k \geq 4$ and the extrapolated BDF schemes with $k = 6$. In the contour plot for C^1_{LM} in Figure 1, with $k = p = 3$, the lower-left (nearly triangular-shaped) region roughly coincides with the region where $a_1b_1 + b_2 \leq 0$.

For implicit methods we have $\alpha_1 = a_1 - \theta_1$, $\beta_1 = b_1 + \theta_1 b_0$. Since $b_0 \geq 0$ we then have the necessary condition

$$C^1_{LM} > 0 \qquad \Longrightarrow \qquad a_1 > 0\,, \quad b_1 + a_1b_0 \geq 0\,. \tag{5.4}$$

An example will be seen in Figure 2 below.

5.3. **Implicit methods.** For the construction of optimal methods in [13] only explicit methods were considered. The reason was that with implicit methods threshold values are found that are not much larger than with explicit methods. From a practical point of view this means that implicit methods do not allow large time steps if monotonicity properties are crucial. An exception is the backward Euler method with $K_{LM} = \infty$; see, e.g., formula (3.7). In this section upper bounds for C^1_{LM} will be derived for methods of order two or larger.

5.3.1. *Example.* As an illustration, we show in Figure 2 the threshold values with $\theta_* < 1$ for implicit second-order two-step methods. These methods form a two-parameter family, and we can take a_1, b_0 as free parameters. The methods are zero-stable for $0 \leq a_1 < 2$ and A-stable if we also have $b_0 \geq \frac{1}{2}$. Interesting cases are, for example, $a_1 = 1$ and $a_1 = \frac{4}{3}$, giving the two-step Adams and BDF2-type methods, respectively. The standard implicit BDF2 method corresponds to $a_1 = \frac{4}{3}$, $b_0 = \frac{2}{3}$, and the (third-order) Adams-Moulton method has $a_1 = 1$, $b_0 = \frac{5}{12}$.

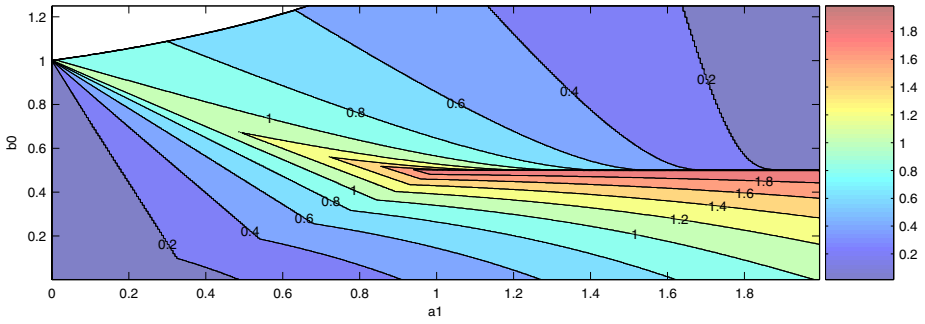

FIGURE 2. Threshold values C^1_{LM} for second-order two-step methods.

We note that the C^1_{LM} values given here are somewhat larger for $b_0 \geq 1$ than the values presented in [7], where constant θ_j were used. In the quadrangle defined by the inequalities $0 \leq a_1 \leq 1$, $\frac{1}{2}a_1 \leq b_0 \leq 1 - \frac{1}{4}a_1$ we have nonnegative coefficients, and for most of that region the value of $K_{LM} = \min_j(a_j/b_j)$ is close to the displayed C^1_{LM}. Furthermore, it should be noted that $b_1 = 2 - \frac{1}{2}a_1 - 2b_0$ due to second-order consistency. Combining this with (5.4), it is seen that a positive threshold cannot be obtained for $b_0 > (2-\frac{1}{2}a_1)/(2-a_1)$, corresponding to the region in the upper-left corner in Figure 2.

The maximal values $C^1_{LM} = 2$ are found for $b_0 = \frac{1}{2}$ and $a_1 \geq 1$. For any fixed $a_1 \in [1, 2]$ we see the following behaviour: if the parameter b_0 is increased, starting with $b_0 = 0$, we first get an increase of C^1_{LM}, up to the value 2 for $b_0 = \frac{1}{2}$, but after that there is a decrease of C^1_{LM}. It will be shown below that this behaviour is quite general for implicit methods of order $p \geq 2$.

5.3.2. *Upper bound for K_{LM}.* To derive general upper bounds for C^1_{LM} we first study the optimal values K_{LM} for methods with nonnegative coefficients.

Consider an implicit k-step method of order $p \geq 2$ with all $a_j, b_j \geq 0$. In the following results we shall only use the order-two conditions. Together with $\sum_{j=0}^{k-1} a_{k-j} = 1$, see (1.5), these order conditions are

$$\sum_{j=0}^{k-1} \left(j^q a_{k-j} + q\, j^{q-1} b_{k-j} \right) \;=\; k^q - q k^{q-1} b_0\,, \quad q = 1,\, 2\,.$$

Let $c_j = a_j - K b_j$ and assume $c_j \geq 0$ for $j = 1, \ldots, k-1$, that is, $K \leq K_{LM}$. In terms of these coefficients, the order conditions can also be written as

$$\sum_{j=0}^{k-1} \left(c_{k-j} + K b_{k-j} \right) \quad = \quad 1\,, \tag{5.5a}$$

$$\sum_{j=0}^{k-1} \left(j\, c_{k-j} + (Kj+1) b_{k-j} \right) \quad = \quad k - b_0\,, \tag{5.5b}$$

$$\sum_{j=0}^{k-1} \left(j^2 c_{k-j} + (Kj^2 + 2j) b_{k-j} \right) \quad = \quad k(k - 2b_0)\,. \tag{5.5c}$$

By taking a linear combination of these relations, multiplying (5.5a) by λ and (5.5b) by μ, with λ, μ chosen such that

$$\lambda + \mu(k - b_0) - k(k - 2b_0) \;=\; 0\,,$$

it is seen that

$$\sum_{j=0}^{k-1} \big(\lambda + \mu j - j^2\big)\, c_{k-j} \;=\; -\sum_{j=0}^{k-1} \big(K(\lambda + \mu j - j^2) + (\mu - 2j)\big)\, b_{k-j}\,.$$

Let $s = \pm 1$. Depending on b_0, we shall select below suitable $\lambda, \mu \in \mathbb{R}$ such that

$$s(\lambda + \mu j - j^2) \;\geq\; 0\,, \qquad j = 0, 1, \ldots, k-1\,.$$

Since all $c_{k-j}, b_{k-j} \geq 0$ it then follows that

$$s\big(K(\lambda + \mu j - j^2) + (\mu - 2j)\big) \;\leq\; 0$$

for some index j. For both cases $s = +1$ and $s = -1$ we thus obtain

$$K \;\leq\; \max_{0 \leq j \leq k-1} \varphi(j)\,, \qquad \varphi(j) \;=\; \frac{2j - \mu}{\lambda + \mu j - j^2}\,. \tag{5.6}$$

First consider $b_0 \leq \frac{1}{2}$. Take $\lambda = 0$, $\mu = k(k - 2b_0)/(k - b_0)$. Then the function φ will attain its maximum in (5.6) for $j = k - 1$. Hence we get the following upper bound for K_{LM}:

$$K_{LM} \;\leq\; \frac{k^2 - 2k + 2b_0}{(k-1)((1-b_0)k - b_0)}\,.$$

This is monotonically increasing in b_0; if $b_0 = 0$ its value is $(k-2)/(k-1)$ and if $b_0 = \frac{1}{2}$ the value equals 2. If we allow k to be arbitrarily large we get the upper bound

$$K_{LM} \;\leq\; 1/(1 - b_0) \quad \text{for} \;\; b_0 \leq \tfrac{1}{2}\,. \tag{5.7}$$

In fact this bound can be shown to hold for any first-order method with $b_0 \leq 1$.

Next we consider $b_0 > \frac{1}{2}$, and we now take $\mu = 2(k-1)$, $\lambda = -k(k-2) - 2b_0$. Then

$$\varphi(j) \;=\; \frac{2i}{i^2 + 2b_0 - 1}\,, \qquad i = k - 1 - j\,.$$

Here it is easily seen that

$$K_{LM} \;\leq\; \begin{cases} 1/b_0 & \text{if} \;\; \frac{1}{2} < b_0 \leq 1\,, \\ 1/\sqrt{2b_0 - 1} & \text{if} \;\; 1 \leq b_0\,. \end{cases} \tag{5.8}$$

Hence the optimal threshold value is $K_{LM} = 2$, which is achieved by the trapezoidal rule. This was already stated in [11, p. 186], and in that reference also bounds on K_{LM} can be found for higher-order implicit methods, partly obtained by numerical optimizations.

5.3.3. *Upper bound for C^1_{LM}.* The above bounds (5.7), (5.8) for the thresholds K_{LM} with arbitrary step number k lead to the following result.

Theorem 5.1. *For any irreducible, zero-stable, implicit linear multistep method of order $p \geq 2$ we have*

$$C^1_{LM} \;\leq\; \begin{cases} 1/(1 - b_0) & \textit{if} \;\; 0 \leq b_0 \leq \frac{1}{2}\,, \\ 1/b_0 & \textit{if} \;\; \frac{1}{2} \leq b_0 \leq 1\,, \\ 1/\sqrt{2b_0 - 1} & \textit{if} \;\; 1 \leq b_0\,. \end{cases} \tag{5.9}$$

Proof. Let us denote the right-hand side of (5.9) by $U(b_0)$. Along with method (1.4) and the reformulation (3.2a) with $\theta_* < 1$, we also consider formula (3.2a) without the remainder terms,

$$w_n \; - \; b_0 \Delta t F_n \; = \; \sum_{j=1}^{\kappa} \left(\alpha_j w_{n-j} + \beta_j \Delta t F_{n-j} \right), \tag{5.10}$$

where $\kappa = n - k$ and $\alpha_j, \beta_j \geq 0$. The omitted k remainder terms have magnitude $\epsilon = \theta_*^{\kappa}$. Therefore the truncated formula (5.10) is a linear κ-step method for which the order-two conditions will be satisfied within $\mathcal{O}(\epsilon)$ accuracy; that is, (5.5) is valid in terms of the coefficients α_j, β_j and step number κ (instead of a_j, b_j and k) if we modify the right-hand sides by adding an $\mathcal{O}(\epsilon)$ term. Now we can repeat the arguments of Section 5.3.2 for this truncated method to obtain

$$\min_{1 \leq j \leq \kappa} \frac{\alpha_j}{\beta_j} \; \leq \; U(b_0) + \mathcal{O}(\epsilon) \, .$$

By taking κ sufficiently large, it is thus seen that the above upper bounds for K_{LM} with arbitrarily large step numbers k also apply to the threshold C^1_{LM} of the original method (1.4). □

For practical applications the most important fact is that large threshold values are not possible. Numerical illustrations of the strong oscillations that can occur with standard implicit methods for the advection test equation $u_t + u_x = 0$, with TVD-limiters in the spatial discretization, can be found in [8, Sect. III.1]. Explicit methods are therefore preferable if monotonicity properties are crucial. For applications with very stiff terms, for instance convection-reaction with stiff reactions, some form of splitting or an implicit-explicit approach may of course be more beneficial if the difficulties with monotonicity arise from the nonstiff (or mildly stiff) parts of the equation that allow explicit treatment.

Appendix A. Optimizations and optimal methods

In [13] optimizations of the threshold values were performed over various classes of explicit k-step schemes with order p, using the Baron optimization package. As indicated in Section 3.2, finding the threshold values C^1_{LM} and C^0_{LM} for any given method involves mathematically all possible integers $l \geq 0$; numerical optimal values are found by selecting a fixed, sufficiently large l.

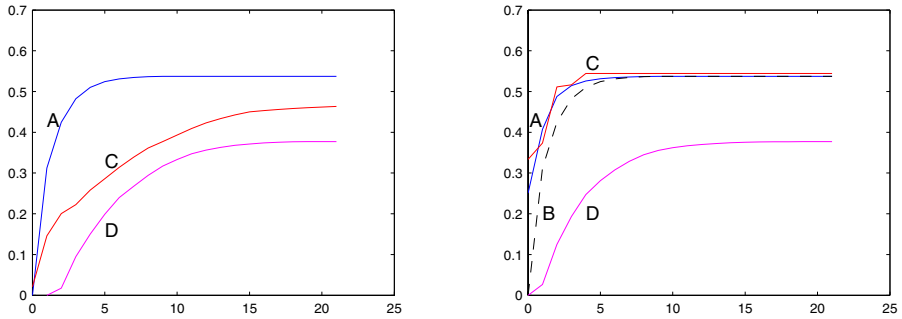

Figure 3. Optimal values γ_{LM} for $l = 0, 1, \ldots, 21$, with $\theta_* = 0$ or $\theta_* \in [0, 1)$, for explicit methods with given k, p.

To illustrate this procedure, we consider here optimizations of the values γ_{LM} for fixed integers l, with either $\theta_* = 0$ or $\theta_* \in [0, 1)$, over some classes of explicit methods with given step number k and order p.

In Figure 3 the optimal values are plotted for several choices of (k, p) with integers $l = 0, 1, \ldots$ on the horizontal axis. One sees that the values for increasing l quickly level out to optimal threshold values.

In these plots, $l = 0$ also is included, meaning that all θ_j equal θ_*. If $\theta_* = 0$ the optimal γ_{LM} values then correspond of course with the optimal K_{LM} over these classes of methods. For $(k, p) = (5, 4)$ a small value $K_{LM} \approx 0.02$ is possible. For the other choices of (k, p) there is no positive K_{LM}; see also [4, 13].

Furthermore, we note that for $p = 4$, $k = 4, 5$, the case $\theta_* \in [0, 1)$ yields optimal values that are actually achieved by methods with $\theta_* = 1$, but these methods are not zero-stable (double root 1 for the ρ-polynomials). Also nearby methods with θ_* slightly less than 1 cannot be recommended; these methods have large error constants. For this reason the optimization for $(k, p) = (4, 4)$ was performed in [13] with $\theta_* \in [0, 0.7]$, leading to the TVB(4,4) method in Table 3.2 of that paper.

Optimizations of this kind yielded a number of schemes in [13] with step number k up to 7 and order $p = k$ or $p = k - 1$. The schemes with $\theta_* = 0$ were denoted as $\mathrm{TVB}_0(k, p)$ and for these schemes the result of Theorem 3.2 is valid. For the other $\mathrm{TVB}(k, p)$ schemes of [13] the boundedness result of Theorem 3.1 applies.

References

[1] M. Tawarmalani, N.V. Sahinidis, *Convexification and Global Optimization in Continuous and Mixed-Integer Nonlinear Programming: Theory, Algorithms, Software, and Applications*. Nonconvex Optimization and Its Applications 65, Kluwer, 2002. MR1961018 (2004a:90001)

[2] G. Dahlquist, *Error analysis for a class of methods for stiff nonlinear initial value problems*, Procs. Dundee Conf. 1975, Lecture Notes in Math. 506, G.A. Watson (ed.), Springer, 1976, pp. 60-74. MR0448898 (56:7203)

[3] L. Ferracina, M.N. Spijker, *Stepsize restrictions for total-variation-boundedness in general Runge-Kutta procedures*. Appl. Numer. Math. 53 (2005), pp. 265–279. MR2128526

[4] S. Gottlieb, C.-W. Shu and E. Tadmor, *Strong stability preserving high-order time discretization methods*, SIAM Review 42 (2001), pp. 89-112. MR1854647 (2002f:65132)

[5] E. Hairer, S.P. Nørsett and G. Wanner, *Solving Ordinary Differential Equations I – Nonstiff Problems*, Second edition, Springer Series in Comput. Math. 8, Springer, 1993. MR1227985 (94c:65005)

[6] E. Hairer, G. Wanner, *Solving Ordinary Differential Equations II – Stiff and Differential-Algebraic Problems*, Second edition, Springer Series in Comput. Math. 14, Springer, 1996. MR1439506 (97m:65007

[7] W. Hundsdorfer, S.J. Ruuth and R.J. Spiteri, *Monotonicity-preserving linear multistep methods*, SIAM J. Numer. Anal. 41 (2003), pp. 605-623. MR2004190 (2004g:65093)

[8] W. Hundsdorfer, J.G. Verwer, *Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations*. Springer Series in Comput. Math. 33, Springer, 2003. MR2002152 (2004g:65001)

[9] R. Jeltsch, O. Nevanlinna, *Stability of explicit time discretizations for solving initial value problems*, Numer. Math. 37 (1981), pp. 61-91. MR0615892 (82g:65042)

[10] H.W.J. Lenferink, *Contractivity preserving explicit linear multistep methods*, Numer. Math. 55 (1989), pp. 213-223. MR0987386 (90f:65058)

[11] H.W.J. Lenferink, *Contractivity preserving implicit linear multistep methods*, Math. Comp. 56 (1991), pp. 177-199. MR1052098 (91i:65129)

[12] R.J. LeVeque, *Finite Volume Methods for Hyperbolic Problems*. Cambridge Texts in Applied Mathematics, Cambridge University Press, 2002. MR1925043 (2003h:65001)

[13] S.J. Ruuth, W. Hundsdorfer, *High-order linear multistep methods with general monotonicity and boundedness properties*. To appear in J. Comp. Phys., 2005.

[14] J. Sand, *Circle contractive linear multistep methods*, BIT 26 (1986), pp. 114-122. MR0833836 (87h:65124)
[15] C.-W. Shu, *TVB uniformly high-order schemes for conservation laws*, Math. Comp. 49 (1987), pp. 105-121. MR0890256 (89b:65208)
[16] C.-W. Shu, *Total-variation-diminishing time discretizations*, SIAM J. Sci. Stat. Comp. 9 (1988), pp. 1073-1084. MR0963855 (90a:65196)
[17] M.N. Spijker, *Contractivity in the numerical solution of initial value problems*, Numer. Math. 42 (1983), pp. 271-290. MR0723625 (85b:65067)
[18] R. Vanselov, *Nonlinear stability behaviour of linear multistep methods*, BIT 23 (1983), pp. 388-396. MR0705005 (84k:65090)

CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands
E-mail address: willem.hundsdorfer@cwi.nl

Department of Mathematics, Simon Fraser University, Burnaby, British Columbia, V5A 1S6 Canada
E-mail address: sruuth@sfu.ca