# MONOTONICITY-PRESERVING LINEAR MULTISTEP METHODS

WILLEM HUNDSDORFER *, STEVEN J. RUUTH †, AND RAYMOND J. SPITERI ‡

**Abstract.** In this paper we provide an analysis of monotonicity properties for linear multistep methods. These monotonicity properties include positivity and the diminishing of total variation. We also pay particular attention to related boundedness properties such as the total-variation-bounded (TVB) property. In the analysis the multistep methods are considered in combination with suitable starting procedures. This allows for monotonicity statements for classes of methods which are important and often used in practice but which were thus far not covered by theoretical results.

**Key words.** Multistep schemes, monotonicity, positivity, TVD, TVB, strong stability

**AMS subject classifications.** 65L06, 65M06, 65M20

**1. Introduction.** In this paper we shall be concerned with preservation of certain monotonicity properties for systems of ordinary differential equations (ODEs) in $\mathbb{R}^m$, $m \geq 1$,

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

Specifically we are interested in the discrete preservation of these properties by numerical approximations $w_n \approx w(t_n)$, $t_n = n\Delta t$, generated by linear multistep methods. The multistep methods will be considered in combination with suitable starting procedures. Hence for a given problem (1.1) and step size $\Delta t$, we can regard the sequence $\{w_n\}_{n\geq 1}$ as being determined by the initial value $w_0$ only, just as for the true solution of (1.1).

There are a number of closely related monotonicity concepts. In this paper we shall mainly consider the property

$$\|w_n\| \;\leq\; \|w_0\| \qquad \text{for all } n \geq 1,\, w_0 \in \mathbb{R}^m \,, \tag{1.2}$$

where $\|\cdot\|$ is a given semi-norm, such as the total variation over the components; see for instance (2.1). Related concepts, such as positivity and contractivity, are considered in the next section. Note that for one-step methods, such as Runge–Kutta methods, property (1.2) is equivalent to

$$\|w_n\| \;\leq\; \|w_{n-1}\| \qquad \text{for all } n \geq 1 \text{ with arbitrary } w_0 \in \mathbb{R}^m \,. \tag{1.3}$$

The relevant monotonicity property should hold for the ODE system (1.1) itself of course. In the following we assume that there is a maximal step size $\Delta t_{FE} > 0$, under which (1.3) holds for the forward Euler method,

$$\|v + \Delta t F(v)\| \;\leq\; \|v\| \qquad \text{for all } 0 < \Delta t \leq \Delta t_{FE},\;\; v \in \mathbb{R}^m \,, \tag{1.4}$$

*CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands (Willem.Hundsdorfer@cwi.nl).

†Department of Mathematics, Simon Fraser University, Burnaby, British Columbia, V5A 1S6 Canada (sruuth@sfu.ca). The work of this author was partially supported by a grant from NSERC Canada.

‡Department of Computer Science, Dalhousie University, Halifax, Nova Scotia, B3H 1W5 Canada (spiteri@cs.dal.ca). The work of this author was partially supported by a grant from NSERC Canada.

and we shall determine constants $C_{LM}$ such that the property is valid for a multistep method with suitable starting procedure under the step size restriction

$$\Delta t \;\leq\; C_{LM}\, \Delta t_{FE}\,. \tag{1.5}$$

In our analysis it is crucial to consider the linear multistep method in combination with suitable starting procedures. If a linear $k$-step method is considered with *arbitrary* starting vectors $w_1, \ldots, w_{k-1}$, in addition to $w_0$, then a natural generalization of (1.3) is

$$\|w_n\| \;\leq\; \max_{0\leq j\leq k-1} \|w_j\| \qquad \text{for all } n \geq k,\;\; w_0, w_1, \ldots, w_{k-1} \in \mathbb{R}^m\,. \tag{1.6}$$

This is a common generalization for analysis purposes of multistep methods. However, there is no direct analogy with the analysis of (1.1), where the solution is determined by $w_0$ only. More importantly, it turns out that the insistence on arbitrary starting vectors severely limits the class of methods for which monotonicity can be demonstrated; for example the familiar BDF and Adams methods are then excluded. Consequently, relevant properties of many popular methods used in practice have not been covered yet by theoretical results.

In § 2 some related monotonicity properties are briefly discussed, together with existing results on multistep methods of the type (1.6) that were obtained in [2, 4, 12, 13, 15, 17]. Then in § 3 we analyze the time-step restrictions for (1.2) of both explicit and implicit two-step methods with various starting procedures. Apart from the monotonicity property (1.2) we also consider related boundedness properties $\|w_n\| \leq M\|w_0\|$ with constant $M \geq 1$. In § 4 we extend our analysis to linear multistep methods of higher order. Finally in § 5 we provide some numerical examples to illustrate our results.

## 2. Background.

**2.1. Related monotonicity concepts.** If the ODE system (1.1) is derived from a spatial discretization of a one-dimensional partial differential equation (PDE), then the components $w_j(t)$ of $w(t)$ will approximate the PDE solution $u(x,t)$ at grid points $x = x_j$ or surrounding cells. In that case $w_n = (w_j^n)$ contains the fully discrete method-of-lines approximation to $u(x_j, t_n)$. Consider for vectors $v = (v_j)$ the semi-norm $\|v\| = \mathrm{TV}(v)$ given by

$$\mathrm{TV}(v) = \sum_j |v_j - v_{j-1}|\,. \tag{2.1}$$

We note that this is a semi-norm, and not a norm, because $\mathrm{TV}(v)$ may vanish for $v \neq 0$; viz. $v_j$ constant. For a pure initial-value PDE on an unbounded domain, the index $j$ will run over all integers; but in general, boundary or periodicity conditions will result in a finite-dimensional ODE system. If (1.3) is valid, that is $\mathrm{TV}(w_n) \leq \mathrm{TV}(w_{n-1})$, $n \geq 1$, the scheme is called *total variation diminishing* (TVD). With property (1.2) we have

$$\mathrm{TV}(w_n) \;\leq\; M\ \mathrm{TV}(w_0)\,, \quad n \geq 1\,, \tag{2.2}$$

with constant $M = 1$. A scheme satisfying (2.2) with some $M \geq 1$ is called *total variation bounded* (TVB). Although this is formally weaker than TVD, conservative schemes with this boundedness property are known to converge to the correct entropy solutions for hyperbolic conservation laws; see for instance [5] for details. If (2.2) holds with $M = 1$, no spatial oscillations can be introduced during the time stepping; such spatial oscillations can be regarded as local overshoots and undershoots. Moreover, the scheme will then also be *monotonicity preserving* in the sense that if the initial data $w_j^0$ is monotonically increasing or decreasing in $j$, then this will be preserved over time [14].

An other property related to avoiding undershoots is *positivity* [2], where it is required that

$$w_n \ \geq\ 0 \quad \text{whenever } w_0 \ \geq\ 0\,. \tag{2.3}$$

Here inequalities for vectors are to be interpreted component-wise. The corresponding requirement on the forward Euler method then reads

$$v + \Delta t F(v) \ \geq\ 0 \qquad \text{for all } 0 < \Delta t \leq \Delta t_{\scriptscriptstyle FE},\ \ v \geq 0\,. \tag{2.4}$$

Although we shall mainly focus on the relations (1.2), (1.4), results for positivity with (2.3), (2.4) can be derived in the same way. Positivity is often a natural requirement for general ODE systems, not necessarily semi-discrete PDEs, especially if the components $w_j(t)$ represent physical quantities such as mass or chemical concentrations that must be nonnegative by definition.

Further, related to (1.3) one can consider the *contractivity* property where the difference $\|\tilde{w}_n - w_n\|$ between any two sequences is required to be nonincreasing with increasing $n$ [10, 17]. If we are dealing with a genuine norm, this is a strong stability requirement. Recently [4], methods satisfying (1.3) have also been called *strong stability preserving*, and there is a tradition in the computational gas dynamics literature of referring to TVB, TVD, monotonicity-preservation, and other nonlinear conditions as *nonlinear stability* [11]. However, properties like TVD or positivity are not directly related to the classical numerical concept of stability which deals with growth between two sequences, one of which is viewed as a perturbation of the other. For linear problems we could well associate (1.3) with numerical stability, whereas for general nonlinear problems it may be viewed as a (strong) boundedness property.

Finally we note that the term *monotonicity* appears in the numerical analysis literature for a variety of related concepts. For example, it is sometimes also used for properties like maximum principles ($\min_j w_j^0 \leq w_j^n \leq \max_j w_j^0$) or comparison principles ($\tilde{w}_0 \leq w_0$ implies $\tilde{w}_n \leq w_n$). In this paper we restrict ourselves to (1.2) and (2.3), but related properties could be studied in a similar way.

**2.2. Monotonicity with arbitrary starting values.** In this paper we mainly consider explicit linear multistep methods

$$w_n \ =\ \sum_{j=1}^{k} \big(a_j w_{n-j} + b_j \Delta t F(w_{n-j})\big)\,, \quad n \geq k\,, \tag{2.5}$$

where starting vectors $w_0, w_1, \cdots, w_{k-1}$ are either given or computed by an appropriate starting procedure. Consistency of the method implies that

$$\sum_{j=1}^{k} a_j \ =\ 1\,. \tag{2.6}$$

Assume for the moment that all $a_j, b_j \geq 0$. By regarding the step (2.5) as a linear combination of scaled forward Euler steps,

$$w_n \;=\; \sum_{j=1}^{k} a_j \left(w_{n-j} + c_j \Delta t F(w_{n-j})\right), \qquad c_j = b_j / a_j \,, \tag{2.7}$$

it easily follows that

$$\|w_n\| \;\leq\; \max_{1 \leq j \leq k} \|w_{n-j}\| \tag{2.8}$$

will hold under (1.4) with the step size restriction

$$\Delta t \;\leq\; K_{\scriptscriptstyle LM} \Delta t_{\scriptscriptstyle FE} \,, \qquad K_{\scriptscriptstyle LM} = \min_{1 \leq j \leq k} \left(\frac{a_j}{b_j}\right) \quad \text{if } a_j, b_j \geq 0 \text{ for all } j \,. \tag{2.9}$$

By convention terms of the form $0/0$ should be omitted in the minimization, and if some coefficient $a_j, b_j$ is negative we leave $K_{\scriptscriptstyle LM}$ undefined. Note that (2.8) can also be formulated equivalently as (1.6).

This result, obtained with scaled forward Euler steps, is due to Shu [15], where it was formulated with total variations. Related results for multistep methods were derived by Bolley & Crouzeix [2] in terms of positivity for linear systems. Contractivity for linear systems was studied by Spijker [17] and Lenferink [12, 13]. The results in [2, 4, 13, 17] also cover implicit methods; we discuss implicit methods in some detail in § 3.

However, these results exclude many schemes that are useful in practice and also may give unnecessary step size restrictions. This is due to the fact that (2.8) should hold with *arbitrary* initial vectors $w_0, w_1, \ldots, w_{k-1}$. As a simple example consider the familiar BDF2 method applied to the trivial problem $w'(t) = 0$. Then

$$w_2 = \frac{4}{3} w_1 - \frac{1}{3} w_0 \,.$$

It is obvious that one cannot have $w_2 \geq 0$ for arbitrary $w_0, w_1 \geq 0$. Likewise it is not always possible to have $\|w_2\| \leq \|w_0\|$ whenever $\|w_1\| \leq \|w_0\|$. On the other hand it is also clear that only the choice $w_1 = w_0$ makes sense for this trivial problem, in which case the inequality $\|w_2\| \leq \|w_0\|$ trivially holds. For this reason we shall analyze the monotonicity properties of multistep schemes with suitable starting procedures. As a result schemes like BDF2 can be included in the theory.

REMARK 2.1. To arrive at (2.9) the assumption $a_j \geq 0$ is necessary to have a convex combination of scaled forward Euler steps. The assumption $b_j \geq 0$ is then needed to ensure that the scaled step sizes $c_j \Delta t$ are nonnegative. As noted in [15, 16], the latter assumption can be avoided for discretizations of the conservation law

$$u_t + f(u)_x = 0 \,,$$

by first applying the discretization in time followed by the spatial discretization (i.e., a transverse-method-of-lines discretization), instead of starting with the semi-discrete system $w' = F(w)$. The only modification to our previous treatment is that if some $b_j < 0$ then $F(w_{n-j})$ in (2.5) should be replaced by $\tilde{F}(w_{n-j})$, where $w' = -\tilde{F}(w)$ is the semi-discretization of

$$u_t - f(u)_x = 0 \,,$$

that is of the equation with reversed time. Its realization in practice is simply a reversal of the upwind direction in the spatial discretization. Along with (1.4), we then also assume

$$\|v - \Delta t \tilde{F}(v)\| \;\le\; \|v\| \qquad \text{for all } 0 < \Delta t \le \Delta t_{FE},\;\; v \in \mathbb{R}^m \,, \tag{2.10}$$

and this counteracts the negativity of $a_j/b_j$ in (2.7). Instead of (2.9) this modification will give the step size restriction

$$\Delta t \;\le\; \tilde{K}_{LM}\,\Delta t_{FE}\,, \qquad \tilde{K}_{LM} = \min_{1\le j\le k}\left(\frac{a_j}{|b_j|}\right) \quad \text{if } a_j \ge 0 \text{ for all } j \tag{2.11}$$

to achieve (2.8). $\diamond$

## 3. Two-Step Methods.

**3.1. Reformulations.** In this section we derive monotonicity results for two-step methods, including some familiar implicit methods [1, 7]. The standard form is written as

$$w_n - b_0\Delta t F_n \;=\; a_1 w_{n-1} + a_2 w_{n-2} + b_1\Delta t F_{n-1} + b_2\Delta t F_{n-2}\,, \qquad n \ge 2\,, \tag{3.1}$$

where $F_{n-j} = F(w_{n-j})$. To obtain precise results, this recursion will be fully written out to include the starting values. Let $\theta \ge 0$ be a parameter to be specified later. Then the two-step recursion can be written in three-step form as

$$\begin{aligned} w_n - b_0\Delta t F_n \;&=\; (a_1-\theta)w_{n-1} + (b_1+\theta b_0)\Delta t F_{n-1} + (a_2+\theta a_1)w_{n-2} \\ &+\; (b_2+\theta b_1)\Delta t F_{n-2} + \theta a_2 w_{n-3} + \theta b_2 \Delta t F_{n-3}\,, \qquad n \ge 3\,. \end{aligned}$$

Continuing this way, by subtracting and adding $\theta^j w_{n-j}$ and using (3.1), we arrive at

$$\begin{aligned} w_n - b_0\Delta t F_n \;&=\; (a_1-\theta)w_{n-1} + (b_1+\theta b_0)\Delta t F_{n-1} \\ &+\; \sum_{j=2}^{n-2} \theta^{j-2}\Big((a_2+\theta a_1-\theta^2)w_{n-j} + (b_2+\theta b_1+\theta^2 b_0)\Delta t F_{n-j}\Big) \\ &+\; \theta^{n-3}\Big((a_2+\theta a_1)w_1 + (b_2+\theta b_1)\Delta t F_1 + \theta a_2 w_0 + \theta b_2 \Delta t F_0\Big)\,. \end{aligned} \tag{3.2}$$

This formula is valid for all $n \ge 3$, with empty sums naturally defined as zero. The reformulation (3.2) will be the basis for our derivations. To bound the last term in (3.2), together with $w_1, w_2$, appropriate starting procedures will be considered. Further we shall determine $\theta$ so as to obtain nonnegative coefficients

$$a_1 - \theta \ge 0\,, \quad a_2 + \theta(a_1-\theta) \ge 0\,, \quad b_1 + \theta b_0 \ge 0\,, \quad b_2 + \theta(b_1+\theta b_0) \ge 0\,, \tag{3.3}$$

with optimal ratio $r(\theta)$ given by

$$r_1(\theta) = \frac{a_1-\theta}{b_1+\theta b_0}\,, \quad r_2(\theta) = \frac{a_2+\theta(a_1-\theta)}{b_2+\theta(b_1+\theta b_0)}\,, \quad r(\theta) = \min\Big(r_1(\theta), r_2(\theta)\Big)\,. \tag{3.4}$$

As before, values 0/0 are ignored when taking the minimum.

**3.2. Explicit second-order two-step methods.** The maximal size of the threshold factor $K_{LM}$ in (2.9) for explicit $k$-step methods of order $p$ has been analyzed by Lenferink [12]. For explicit methods of order $p = 1$ we have $K_{LM} \le 1$, a bound which is already attained by Euler's method. For explicit methods with $k \ge 2$, Lenferink showed that

$$K_{LM} \;\le\; \frac{k-p}{k-1}\,. \tag{3.5}$$

Hence $K_{LM} > 0$ is not possible for second-order explicit two-step methods. By allowing $b_2 < 0$, Shu [15] obtained an explicit two-step method with $p = 2$, $\tilde{K}_{LM} = \frac{1}{2}$. However this result is only applicable to semi-discretizations of conservation laws. Moreover, with $b_1 > 0$, $b_2 < 0$, both $F_j$ and $\tilde{F}_j$ have to be calculated in the process, making the scheme twice as expensive computationally as the standard form (3.1).

Here we consider the monotonicity property (1.2) for schemes (3.1), and optimal threshold factors $C_{LM}$ will be derived for second-order explicit two-step methods combined with suitable starting procedures. The main assumptions on the starting procedures will be

$$\begin{aligned} &\|w_1\| \;\le\; M\,\|w_0\|\,, \quad \|w_2\| = \|a_1w_1 + b_1\Delta t F_1 + a_2 w_0 + b_2 \Delta t F_0\| \;\le\; M\,\|w_0\|\,, \\ &\|(a_2+\theta a_1)w_1 + (b_2+\theta b_1)\Delta t F_1 + \theta a_2 w_0 + \theta b_2 \Delta t F_0\| \;\le\; (a_2+\theta) M\,\|w_0\| \end{aligned} \tag{3.6}$$

for a given step size $\Delta t > 0$ and with $M = 1$. To derive weaker properties, such as (2.2), constants $M > 1$ will also be allowed.

LEMMA 3.1. *Assume (1.4) holds. Let $\theta \ge 0$ satisfy (3.3) and let $r(\theta)$ be given by (3.4) with $b_0 = 0$. Suppose that $\Delta t \le r(\theta)\Delta t_{FE}$ and (3.6) holds with $M \ge 1$. Then*

$$\|w_n\| \;\le\; M\,\|w_0\| \quad \textit{for all } n \ge 1\,. \tag{3.7}$$

*Proof.* From (3.2) and (3.3) we obtain

$$\|w_n\| \le (a_1-\theta)\,\|w_{n-1}\| \,+\, \sum_{j=2}^{n-2} \theta^{j-2}\big(a_2+\theta a_1-\theta^2\big)\,\|w_{n-j}\| \,+\, \theta^{n-3}(a_2+\theta)\,M\,\|w_0\|\,.$$

By assumption, the lemma is valid for $n = 1, 2$. Since we have, in view of (2.6), the relation

$$(a_1-\theta) + \sum_{j=2}^{n-2} \theta^{j-2}\big(a_2+\theta a_1-\theta^2\big) + \theta^{n-3}(a_2+\theta) \;=\; 1\,, \quad n \ge 3,$$

the proof now follows easily by induction. $\square$

To apply this lemma to specific methods we shall determine $\theta$ to obtain an optimal constant

$$C^*_{LM} \;=\; \max\{\, r(\theta) \,:\, \theta \text{ satisfies } (3.3)\,\}\,. \tag{3.8}$$

This will give a step size restriction $\Delta t \le C^*_{LM}\Delta t_{FE}$ which is intrinsic for the specific two-step method. The requirement (3.6) with $M = 1$ may give an additional restriction, say $\Delta t \le C^0_{LM}\Delta t_{FE}$, depending on the starting procedure and the coefficients of the multistep method. For the combined scheme we then obtain the monotonicity property (1.2) under the step size restriction (1.5) with

$$C_{LM} \;=\; \min\{C^0_{LM}, C^*_{LM}\}\,. \tag{3.9}$$

The above derivation will be applied to explicit second-order two-step methods, which constitute a one-parameter family given by (3.1) with $b_0 = 0$ and

$$a_1 = 2-\xi\,,\quad a_2 = \xi - 1\,,\quad b_1 = 1 + \frac{1}{2}\xi\,,\quad b_2 = \frac{1}{2}\xi - 1\,. \tag{3.10}$$

The methods in this class are zero-stable if and only if $0 < \xi \le 2$, and we shall restrict ourselves to these parameter values. Examples of practical interest are the two-step Adams–Bashforth method ($\xi = 1$) and the extrapolated BDF2 method ($\xi = \frac{2}{3}$). With this class of second-order methods it follows, by a straightforward but somewhat tedious calculation, that optimality in (3.8) is attained by setting $b_2 + \theta b_1 = 0$, which gives

$$\theta \;=\; \frac{2-\xi}{2+\xi}\,, \qquad C^*_{LM} \;=\; \frac{2(1+\xi)(2-\xi)}{(2+\xi)^2}\,. \tag{3.11}$$

To obtain a complete bound (3.9) various starting procedures will be considered next.

Remark 3.2. In the remainder of this section we shall focus primarily on condition (3.6) with $M = 1$. For these results all coefficients in the occurring expressions will be required to be nonnegative. Consequently, results on positivity (2.3) with (2.4) can be derived under the same assumptions.

We shall also derive results with $M > 1$. These will only be qualitative. Precise bounds for $M$ can be obtained by using

$$\max_{\Delta t \le C\Delta t_{FE}} \|v + \gamma \Delta t F(v)\| \;\le\; \max\{1, |2\gamma C - 1|\}\,\|v\| \tag{3.12}$$

for arbitrary $C > 0$, $\gamma \in \mathbb{R}$, and $v \in \mathbb{R}^m$. This relation is an obvious consequence of (1.4) if $0 \le \gamma C \le 1$. For values $\gamma C$ outside the interval $[0,1]$, it follows by using in addition the implication $\Delta t_{FE}\|F(v)\| \le 2\,\|v\|$ from (1.4). ◇

**3.2.1. Starting with the forward Euler method.** The natural candidate to compute the starting vector $w_1$ for an explicit two-step method of order $p = 2$ is the forward Euler method

$$w_1 = w_0 + \Delta t F_0\,.$$

Of course, the forward Euler method itself is only first-order accurate; but because it is applied only once, the accuracy of the combined scheme will still be of order two.

With this starting procedure the first condition in (3.6) holds with $M = 1$ for $\Delta t \le \Delta t_{FE}$, of course. The second condition, $\|w_2\| \le M\|w_0\|$, can be written as

$$\|(a_1 - \tilde\theta)w_1 + b_1\Delta t F_1 + (a_2 + \tilde\theta)w_0 + (b_2 + \tilde\theta)\Delta t F_0\| \le M\|w_0\|\,,$$

where an optimal $\tilde\theta$ should be selected. With $M = 1$ it is easily seen that the optimal value is $\tilde\theta = \frac{1}{2}(2-\xi)$, under which the inequality holds for all step sizes

$$\Delta t \;\le\; \frac{2-\xi}{2+\xi}\,\Delta t_{FE}\,. \tag{3.13}$$

With larger step sizes we will have a bound with $M > 1$. We note that this step size restriction for $M = 1$ is more stringent than $\Delta t \le C^*_{LM}\Delta t_{FE}$ for any $\xi \in (0,2)$. Finally, with $\theta$ given by (3.11), the third condition in (3.6) reads

$$\|(a_2+\theta)w_0 + (a_2+\theta a_1+\theta b_2)\Delta t F_0\| \;\le\; (a_2+\theta)M\,\|w_0\|\,,$$

which can be written here as

$$\|w_0 + \frac{1}{2\xi}(3\xi-2)\Delta t F_0\| \;\le\; M\,\|w_0\|\,.$$

Hence $M = 1$ now requires

$$\Delta t \;\le\; \frac{2\xi}{3\xi-2}\,\Delta t_{FE}\,, \qquad \xi \ge \frac{2}{3}\,. \tag{3.14}$$

If either the step size is allowed to be larger or $0 < \xi < \frac{2}{3}$, then we obtain a bound with $M > 1$ (see Remark 3.2), where it should be mentioned that we will have $M \sim \xi^{-1}$ for $\xi \downarrow 0$. We note that for $\xi \ge \frac{2}{3}$ the restriction (3.14) is less stringent than (3.13). The above results can be summarized as follows.

THEOREM 3.3. *Consider the explicit second-order two-step method (3.1), (3.10), and let $w_1$ be computed by the forward Euler method. Then the monotonicity property (1.2) will hold under (1.4) with the restriction*

$$\Delta t \;\le\; \frac{2-\xi}{2+\xi}\,\Delta t_{FE}\,, \qquad \frac{2}{3} \le \xi \le 2\,.$$

*Under (1.4) with the weaker restriction*

$$\Delta t \;\le\; \frac{2(1+\xi)(2-\xi)}{(2+\xi)^2}\,\Delta t_{FE}\,, \qquad 0 < \xi \le 2\,,$$

*the boundedness property (3.7) will hold with $M \ge 1$.*

**3.2.2. A modified two-step starting procedure.** The use of the forward Euler method as starting procedure for the second-order two-step methods (3.1), (3.10) leads to a step size restriction for monotonicity that is more stringent than $\Delta t \le C^*_{LM}\Delta t_{FE}$. Similar restrictions were obtained with standard two-stage Runge–Kutta methods.

As an alternative starting procedure that can be used for semi-discrete conservation laws following Remark 2.1, we compute $w_1$ with the forward Euler method but use for the second step a modified scheme,

$$w_1 = w_0 + \Delta t F_0\,, \;\; w_2 = a_1 w_1 + a_2 w_0 + b_1\Delta t F_1 + \alpha b_2 \Delta t F_0 + (1-\alpha) b_2 \Delta t \tilde F_0\,, \tag{3.15}$$

where $\tilde F_0 = \tilde F(w_0)$ and $\alpha \in [0,1]$ is to be determined later. We assume $\tilde F$ satisfies (2.10). We note that because $\tilde F$ is evaluated only once (for $w_0$), the computational costs will not increase significantly.

Consider the optimal $\theta$ value (3.11), for which $b_2 + \theta b_1 = 0$. With the modified second step, it follows that (3.2) with $b_0 = 0$ will change accordingly to

$$\begin{aligned} w_n = (a_1-\theta)w_{n-1} + b_1\Delta t F_{n-1} + \sum_{j=2}^{n-1}\theta^{j-2}(a_2+\theta a_1-\theta^2)w_{n-j} \\ + \theta^{n-2}\Big(\theta w_1 + a_2 w_0 + \alpha b_2 \Delta t F_0 + (1-\alpha) b_2\Delta t\tilde F_0\Big)\,. \end{aligned}$$

If the last term can be bounded by $\theta^{n-2}(a_2+\theta)\|w_0\|$ and if $\|w_1\| \le \|w_0\|$, the result of Lemma 3.1 will remain valid with $M=1$. With the forward Euler approximation $w_1$, we thus get the requirement

$$\|(a_2+\theta)w_0 + (\alpha b_2+\theta)\Delta t F_0 + (1-\alpha)b_2 \Delta t \tilde{F}_0\| \;\le\; (a_2+\theta)\|w_0\| \tag{3.16}$$

for $\Delta t \le C^0_{LM}\Delta t_{FE}$, where an optimal $C^0_{LM} \in (0,1]$ will be selected by a favourable choice of the parameter $\alpha$.

The contribution of $F_0$ in this inequality is minimized by taking $\alpha = -\theta/b_2 = 1/b_1$. By using (2.10) it then follows that $\|w_1\| \le \|w_0\|$ and (3.16) will be satisfied for $\Delta t \le C^0_{LM}\Delta t_{FE}$ with

$$C^0_{LM} = \min\left\{\, 1\,, \frac{2\xi}{2-\xi}\right\}.$$

Taking $C_{LM} = \min\{C^0_{LM}, C^*_{LM}\}$, we can summarize this result as follows.

THEOREM 3.4. *Consider the explicit second-order two-step method (3.1), (3.10) for $n \ge 3$, and let $w_1, w_2$ be computed by (3.15) with $\alpha = 1/b_1$. Then the monotonicity property (1.2) will hold under (1.4) and (2.10) with the step size restriction*

$$\Delta t \;\le\; C_{LM}\,\Delta t_{FE}\,, \qquad C_{LM} \;=\; \begin{cases} \dfrac{2\xi}{2-\xi} & \text{if} \quad 0<\xi<\dfrac{1}{\frac12+\sqrt{2}}\,, \\[2ex] \dfrac{2(1+\xi)(2-\xi)}{(2+\xi)^2} & \text{if} \quad \dfrac{1}{\frac12+\sqrt{2}} \le \xi \le 2\,. \end{cases}$$

If the step size restriction in this theorem for $\xi < 1/(\frac12+\sqrt{2})$ is not satisfied, but still $\Delta t \le C^*_{LM}\Delta t_{FE}$, then we will have, as with other starting procedures, the boundedness property (3.7) with $M>1$.

A somewhat related result was obtained in [9] for the extrapolated BDF2 method ($\xi = \frac23$) in the so-called one-leg form. For that particular method it was demonstrated that (1.2) holds for all $\Delta t \le C^*_{LM}\Delta t_{FE}$ provided that an appropriate two-stage Runge–Kutta starting method is used. Also it was observed in [9] that this implies boundedness of $\|w_n\|$ for the standard multistep form (3.1) of the extrapolated BDF2 method if a special starting procedure is used. From the above we see that the boundedness property holds for any starting procedure and all $0<\xi\le 2$.

**3.3. Implicit second-order two-step methods.** In this section we consider the implicit two-step methods of order 2. These methods form a two-parameter family with coefficients

$$a_1 = 2-\xi\,,\;\; a_2 = \xi-1\,,\;\; b_0=\eta\,,\;\; b_1 = 1+\frac12\xi-2\eta\,,\;\; b_2 = \eta+\frac12\xi-1\,. \tag{3.17}$$

As for the explicit methods (3.10) we need $0<\xi\le 2$ for zero-stability. The methods are $A$-stable if and only if in addition $\eta \ge \frac12$. If $\eta = \frac12$ these methods are reducible to the trapezoidal rule, in the sense that if $w_1$ is calculated by the trapezoidal rule then the whole sequence $\{w_n\}$ will satisfy the trapezoidal rule recurrence; see [3, 7]. Two interesting sub-classes in (3.17) are $\xi = \frac23$, giving BDF2-type methods, and $\xi=1$, giving implicit 2-step Adams methods.

In order to deal with implicit terms in (3.1), we shall use, in addition to (1.4),

$$\|v\| \;\le\; \|v - \Delta t F(v)\| \qquad \text{for all } \Delta t>0,\;\; v\in\mathbb{R}^m\,. \tag{3.18}$$

This can be interpreted as a condition on the implicit Euler method: $\|v_1\| \le \|v_0\|$ if $v_1 = v_0 + \Delta t F(v_1)$. It might appear that (3.18) should be imposed as an additional assumption to (1.4), but it is in fact a simple consequence. From $v_1 = v_0 + \Delta t F(v_1)$ it follows that

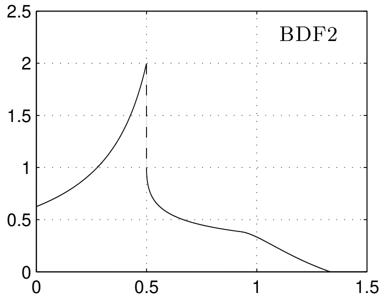
$$\Big(1 + \frac{\Delta t}{\Delta t_{FE}}\Big) v_1 \;=\; v_0 + \frac{\Delta t}{\Delta t_{FE}} \Big( v_1 + \Delta t_{FE} F(v_1) \Big) \,,$$

$$\Big(1 + \frac{\Delta t}{\Delta t_{FE}}\Big) \|v_1\| \;\le\; \|v_0\| + \frac{\Delta t}{\Delta t_{FE}} \|v_1\| \,,$$

and hence $\|v_1\| \le \|v_0\|$ for any $\Delta t > 0$. Thus under (1.4), the implicit Euler method gives the monotonicity property (1.2) without any step size restriction. However, this is only a first-order method and for practical applications higher accuracy is often required. In the following we therefore concentrate on the class of second-order methods (3.17).

It was shown by Lenferink [13], in terms of contractivity for linear systems, that the threshold value $K_{LM}$ in (2.9) is bounded by 2 for all 2-step methods of order $p > 1$. The optimal $K_{LM} = 2$ is attained by the trapezoidal rule. In view of the results for explicit methods, one might hope that such severe restrictions can be circumvented in our formulation (1.2) with suitable starting procedures. Using (3.18) we can follow the derivation of Lemma 3.1, just as for explicit methods. Depending on the starting procedure, according to (3.6), this will give monotonicity (1.2) or boundedness (3.7) for $\Delta t \le r(\theta)\Delta t_{FE}$. We now consider the factors $C^*_{LM}$ that are obtained by optimal values for $\theta$ in (3.8).

Determination of the optimal factors $C^*_{LM}$ in analytical form is cumbersome, even if we restrict ourselves to sub-classes such as $\xi = \frac{2}{3}$ and $\xi = 1$. On the other hand, numerically it is easy to compute the optimal $\theta$ values in (3.8). The corresponding threshold values $C^*_{LM}$ are given in Figure 3.1 for $\xi = \frac{2}{3}, 1$ as function of $\eta$. We note that $C^*_{LM} = \frac{1}{2}$ for the familiar implicit BDF2 method ($\xi = \eta = \frac{2}{3}$).



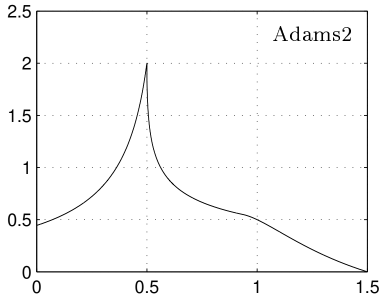




FIG. 3.1. *Threshold values $C^*_{LM}$ versus $\eta \in [0, 1.5]$, with $\xi = \frac{2}{3}$ (left) and $\xi = 1$ (right).*

The results are rather disappointing. The largest numbers $C^*_{LM} = 2$ are found for the values $\eta = \frac{1}{2}$, and numerical verification shows that the same also holds with other choices of $\xi \in (0, 2]$. With $\eta = \frac{1}{2}$ the function $r_2$ in (3.4) has a removable singularity, which is related to the reducibility of the method, and this is the reason why the curves for $\eta < \frac{1}{2}$ and $\eta > \frac{1}{2}$ are very different, even having a discontinuity for the case $\xi = \frac{2}{3}$.

Of course, for a complete bound suitable starting procedures, such as the implicit Euler method, should also be taken into account. However, the main result is that we obtain restrictions that are hardly better than those for explicit methods, and such restrictions have been confirmed in numerical experiments [8]. For practical purposes this means that the implicit schemes are not competitive with the explicit ones if monotonicity properties like (1.2) or (2.3) are crucial in an application. For this reason we shall restrict ourselves in the following section to explicit methods.

REMARK 3.5. For the class of BDF2-type methods, threshold values for monotonicity were calculated analytically in [8] for linear, constant-coefficient problems $w'(t) = Aw(t)$. The curve in Figure 3.1 with $\xi = \frac{2}{3}$ almost coincides with the linear result for $\eta \lesssim 0.9$, but for larger $\eta$ an extra condition sets in due to nonlinearity. For linear systems the restrictions in (3.3) can be relaxed by allowing negative values for $b_0 + \theta b_1$ and $b_2 + \theta(b_0 + \theta b_1)$.

The essential difference between linear and nonlinear results is most easily illustrated by the simple one-step method

$$w_n - \eta \Delta t F_n = w_{n-1} + (1-\eta)\Delta t F_{n-1}, \tag{3.19}$$

with parameter $\eta \geq 0$, whose stability function is given by

$$R(z) = (1-\eta z)^{-1}(1 + (1-\eta)z)\,.$$

Let $\gamma$ be the largest number such that $R$ and all its derivatives are nonnegative on $[-\gamma, 0]$. It has been shown in [2, 17] (in terms of positivity and contractivity) that the monotonicity property (1.2) will hold under (1.4) for linear systems $w'(t) = Aw(t)$ provided that $\Delta t \leq \gamma \Delta t_{FE}$. Thus for linear problems we get the restriction

$$\Delta t \;\leq\; \gamma\, \Delta t_{FE}\,, \qquad \gamma = \begin{cases} (1-\eta)^{-1} & \text{if } \eta < 1\,, \\ \infty & \text{if } \eta \geq 1\,. \end{cases}$$

On the other hand, for nonlinear problems the optimal condition is seen to be

$$\Delta t \;\leq\; C\, \Delta t_{FE}\,, \qquad C = \begin{cases} (1-\eta)^{-1} & \text{if } \eta < 1\,, \\ \infty & \text{if } \eta = 1\,, \\ 0 & \text{if } \eta > 1\,. \end{cases}$$

Note that for $\eta > 1$ the coefficient in front of $F_{n-1}$ becomes negative. For linear problems this can be counteracted by the implicit term, but for general nonlinear problems we need this coefficient to be nonnegative. ◇

## 4. Higher-Order Methods.

This section contains derivations of boundedness results for various important higher-order explicit linear multistep methods: the extrapolated BDF and explicit Adams methods of order three or greater. To study the boundedness property (3.7), with $M \geq 1$, it is not necessary to specify the starting schemes: although the value of $M$ may vary according to the choice of starting procedure, the boundedness property itself is independent of this choice.

### 4.1. Reformulations.

We begin with a reformulation for the explicit multistep schemes. This is similar to formula (3.2) for 2-step schemes, but to obtain proper step size restrictions different $\theta_j$ will be used in the various stages. To keep the presentation concise we give the reformulation here in detail only for 3-step schemes. Consider (2.5) with $k = 3$. Then by subtracting and adding $\theta_j w_{n-j}$, $j = 1, 2, \ldots, n-3$, substituting

$w_{n-j}$ in terms of $w_{n-j-1}, \ldots, w_{n-j-3}$, and collecting terms, it follows that $w_n$ can be expressed as

$$w_n = \sum_{j=1}^{n-3} \big(\alpha_j w_{n-j} + \beta_j \Delta t F_{n-j}\big) + \sum_{i=0}^{2} \big(\rho_{i,n} w_i + \sigma_{i,n} \Delta t F_i\big)\,, \tag{4.1}$$

where the coefficients $\alpha_j$, $\beta_j$ are given by

$$\alpha_1 = a_1 - \theta_1\,, \quad \alpha_2 = a_2 + a_1\theta_1 - \theta_1\theta_2\,, \quad \alpha_3 = a_3 + a_2\theta_1 + a_1\theta_1\theta_2 - \theta_1\theta_2\theta_3\,,$$

$$\alpha_j = \Big(\prod_{k=1}^{j-3} \theta_k\Big)\Big(a_3 + a_2\theta_{j-2} + a_1\theta_{j-2}\theta_{j-1} - \theta_{j-2}\theta_{j-1}\theta_j\Big)\,, \quad j \geq 4\,,$$

$$\beta_1 = b_1\,, \quad \beta_2 = b_2 + b_1\theta_1\,, \quad \beta_3 = b_3 + b_2\theta_1 + b_1\theta_1\theta_2\,,$$

$$\beta_j = \Big(\prod_{k=1}^{j-3} \theta_k\Big)\Big(b_3 + b_2\theta_{j-2} + b_1\theta_{j-2}\theta_{j-1}\Big)\,, \quad j \geq 4\,.$$

We shall take $\theta_i \geq 0$ such that

$$\alpha_j \geq 0\,, \quad \beta_j \geq 0 \qquad \text{for all } j \geq 1\,, \tag{4.2}$$

and we define

$$C^*_{LM} \;=\; \max_{\{\theta_i\}_{i\geq 1}} \; \min_{j \geq 1} \; \frac{\alpha_j}{\beta_j}\,. \tag{4.3}$$

Then it follows, similar to Lemma 3.1, that the boundedness property (3.7) will hold with $M \geq 1$ under the step size restriction $\Delta t \leq C^*_{LM} \Delta t_{FE}$. To obtain results on monotonicity (1.2), that is, $M = 1$, it is also necessary to study the coefficients $\rho_{i,n}$, $\sigma_{i,n}$ of the remainder term in (4.1) and to include specific starting procedures.

For $k$-step methods with $k \geq 4$ we can proceed similarly. In the above reformulation (4.1) we get the same expressions for $\alpha_1$, $\alpha_2$, $\alpha_3$ and $\beta_1$, $\beta_2$, $\beta_3$; the other $\alpha_j$, $\beta_j$ will then involve more terms.

**4.2. Boundedness and TVB.** First we give the step size restrictions for boundedness and the related TVB property for the third-order extrapolated BDF3 scheme

$$w_n = \frac{18}{11} w_{n-1} - \frac{9}{11} w_{n-2} + \frac{2}{11} w_{n-3} + \frac{18}{11} \Delta t F_{n-1} - \frac{18}{11} \Delta t F_{n-2} + \frac{6}{11} \Delta t F_{n-3} \tag{4.4}$$

and the fourth-order extrapolated BDF4 scheme

$$\begin{aligned} w_n \;=\; & \frac{48}{25} w_{n-1} - \frac{36}{25} w_{n-2} + \frac{16}{25} w_{n-3} - \frac{3}{25} w_{n-4} \\ & + \frac{48}{25} \Delta t F_{n-1} - \frac{72}{25} \Delta t F_{n-2} + \frac{48}{25} \Delta t F_{n-3} - \frac{12}{25} \Delta t F_{n-4}\,. \end{aligned} \tag{4.5}$$

THEOREM 4.1. *Assume (1.4) holds. The extrapolated BDF3 (4.4) scheme satisfies the boundedness property (3.7) with* $M \geq 1$ *provided* $\Delta t \leq \frac{7}{18} \Delta t_{FE}$. *For the extrapolated BDF4 (4.5) scheme the boundedness property will hold if* $\Delta t \leq \frac{7}{32} \Delta t_{FE}$. *These values* $\frac{7}{18}, \frac{7}{32}$ *are optimal within (4.2), (4.3).*

*Proof.* Consider (4.4). We first maximize $\alpha_1/\beta_1$ over the constraint $\beta_2 \geq 0$ to get $\theta_1 = 1$. This also maximizes $\alpha_2/\beta_2$; so next we maximize $\alpha_3/\beta_3$ over the constraints

$\alpha_2 \geq 0, \beta_3 \geq 0$ to get $\theta_2 = \frac{2}{3}$. Maximizing $\alpha_4/\beta_4$ over the constraints $\alpha_3 \geq 0, \beta_4 \geq 0$ gives $\theta_3 = \frac{1}{2}$. We can now set the remaining $\theta_j = \frac{1}{2}$, $j \geq 4$ because this choice is admissible, in the sense of (4.2), and does not contribute to the step size restriction; indeed $\frac{1}{2}$ is the value that minimizes the factor $(b_3 + b_2\theta_{j-2} + b_1\theta_{j-2}\theta_{j-1})$ in $\beta_j$, $j \geq 4$. Since

$$\min_{j\geq 1} \frac{\alpha_j}{\beta_j} = \min\left\{\frac{\alpha_1}{\beta_1}, \frac{\alpha_2}{\beta_2}, \frac{\alpha_3}{\beta_3}, \frac{\alpha_4}{\beta_4}, \frac{\alpha_5}{\beta_5}\right\} = \frac{\alpha_1}{\beta_1} = \frac{7}{18},$$

and we first optimized over $\alpha_1/\beta_1$, we see that $C^*_{LM} = \frac{7}{18}$.

The result for the extrapolated BDF4 scheme follows in a similar manner, except that an admissible value for $\theta_3$ is more difficult to find; for this we used a numerical search. We remark that in both cases the constant $M$ will depend on the choice of starting procedure used. □

Another popular class of methods is formed by the explicit $k$-step Adams methods with order $p = k$. The third-order method is

$$w_n = w_{n-1} + \frac{23}{12}\Delta t F(w_{n-1}) - \frac{16}{12}\Delta t F(w_{n-2}) + \frac{5}{12}\Delta t F(w_{n-3}), \tag{4.6}$$

and the coefficients $a_j, b_j$ for the higher-order methods can be found in [6], for example. For these methods the results are less favourable than for the extrapolated BDF schemes.

THEOREM 4.2. *Assume (1.4) holds. The explicit 3-step Adams method satisfies the boundedness property (3.7) with $M \geq 1$ provided $\Delta t \leq \frac{84}{529}\Delta t_{FE}$. For the explicit Adams methods with $k \geq 4$ no positive $C^*_{LM}$ value in (4.3) exists.*

*Proof.* To have $\beta_2 \geq 0$ we need $\theta_1 \geq -b_2/b_1$, and consequently

$$\frac{\alpha_1}{\beta_1} \leq \frac{1 + b_2/b_1}{b_1} = \frac{1}{b_1^2}(b_1 + b_2).$$

If $k = 3$ we have $b_1 = \frac{23}{12}$ and $b_2 = -\frac{16}{12}$, leading to $C^*_{LM} \leq \frac{84}{529}$. Moreover, it follows by some simple calculations that this upper bound is attained by taking all $\theta_i = \frac{16}{23}$.

To show that we cannot have $C^*_{LM} > 0$ if $k \geq 4$, note that the $k$-step explicit Adams method may be written as

$$w_n = w_{n-1} + \Delta t \sum_{j=0}^{k-1} \gamma_j \nabla^j F_{n-1},$$

where $\nabla^j$ represent the usual backward differences and the $\gamma_j$ are positive constants given in [6, Sect. III.1]. A straightforward calculation for $k \geq 4$ shows that

$$b_1 = \sum_{j=0}^{k-1} \gamma_j = \frac{55}{24} + \sum_{j=4}^{k-1} \gamma_j, \qquad b_2 = -\sum_{j=0}^{k-1} j\gamma_j = -\frac{59}{24} - \sum_{j=4}^{k-1} j\gamma_j.$$

Therefore $b_1 + b_2 \leq \frac{-4}{24} < 0$, implying that $\alpha_1/\beta_1 < 0$. Hence the scheme does not possess a positive threshold value $C^*_{LM}$. □

REMARK 4.3. Following the same lines, it is also straightforward to show that none of the explicit Nyström methods [6] has a positive threshold value $C^*_{LM}$. ◇

The generation of monotonicity results for high-order multistep schemes such as extrapolated BDF3 by means of optimized strong-stability-preserving Runge–Kutta starting procedures [4, 18] is part of our current research.

## 5. Numerical Illustrations.

**5.1. Linear positivity test.** As a first numerical test we consider the positivity property (2.3) for the linear advection problem $u_t + u_x = 0$, $0 \le x \le 1$, with inflow boundary condition $u(0,t) = 0$ and initial mass $u(x,0)$ concentrated at the inflow boundary. The semi-discrete system is obtained with first-order upwind discretization in space and constant mesh width $\Delta x = 1/m$. The resulting linear ODE system in $\mathbb{R}^m$ is

$$w'(t) = Aw(t)\,, \quad A = \frac{1}{\Delta x}\begin{pmatrix} -1 & & & \\ 1 & -1 & & \\ & \ddots & \ddots & \\ & & 1 & -1 \end{pmatrix}, \quad w_0 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{5.1}$$

The dimension of the system is taken as $m = 100$. For this system we determined experimentally the largest Courant number $\nu = \Delta t/\Delta x$ for which $w_n \ge 0$ is maintained up to $n = 1000$. We note that with the forward Euler method this will hold up to $\nu = 1$. Further we note that, by changing $w_j(t)$ in (5.1) into $1-w_j(t)$, identical results can be obtained with the condition $\|w_n\|_\infty \le \|w_0\|_\infty$.

First we consider the class of explicit two-step methods (3.10) with parameter values $\xi = j/20$, $j = 0, 1, \dots, 40$. Along with the forward Euler method and the modified two-step procedure (3.15), we also consider the exact starting value $w_1 = \exp(\Delta t A) w_0$. The results are plotted in Figure 5.1, in combination with the theoretical values $C^*_{LM}$ from (3.11).

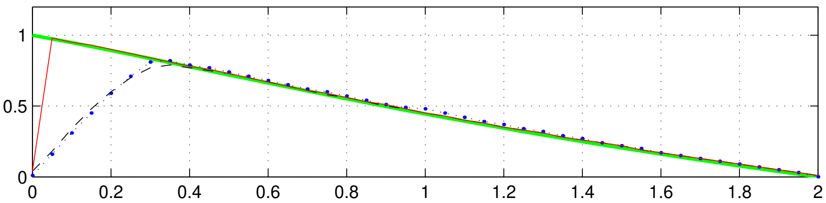



FIG. 5.1. *Positivity test for the explicit two-step methods (3.10). Courant numbers versus $\xi \in [0,2]$, starting with exact solution values [dots], forward Euler [solid line], and the modified two-step procedure (3.15) [dashed line]. The thick gray line is the $C^*_{LM}$-curve from (3.11).*

The influence of the starting values as given in the Theorems 3.3, 3.4 does not show up accurately in Figure 5.1. We note however that the test problem here is linear, whereas the theoretical results were obtained for nonlinear problems.

In a similar manner the behaviour of the implicit two-step Adams and BDF-type methods has been tested. The results are shown in Figure 5.2. The starting value $w_1$ was computed with the implicit Euler method and with method (3.19), where the parameter $\eta$ is the same as in (3.17); taking an exact starting value for $w_1$ did give results close to the latter starting procedure. For $\xi > \frac{1}{2}$ the results with implicit Euler and (3.19) also almost coincide.

The Courant numbers in Figure 5.2 are close to the theoretical bound $C^*_{LM}$ in Figure 3.1 for $\eta \lesssim 0.9$. In particular the different behaviour for $\eta < \frac{1}{2}$ and $\eta > \frac{1}{2}$ shows up very clearly. Quantitatively, only the results with the BDF-type methods with $\eta \le \frac{1}{2}$ and the implicit Euler method as starting procedure are somewhat more favourable than the bound $C^*_{LM}$. The difference between the curves in Figure 3.1 for the larger $\eta$

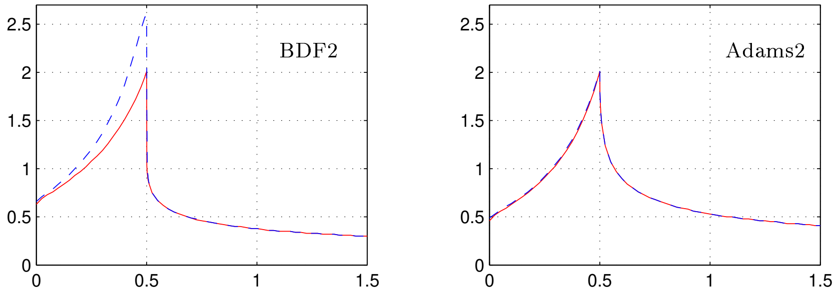



FIG. 5.2. *Positivity test for the implicit two-step methods (3.17) with $\xi = \frac{2}{3}$ [left] and $\xi = 1$ [right]. Courant numbers versus $\eta \in [0, \frac{3}{2}]$, starting with implicit Euler [dashed lines] and method (3.19) [solid lines].*

values is due to the fact that the values $C^*_{LM}$ were obtained for nonlinear problems; see Remark 3.5. As noted previously, the rather small Courant numbers allowed with the implicit methods in practice mean that these implicit second-order two-step methods are not competitive with the explicit ones for problems where monotonicity is crucial.

In Table 5.1 the experimental positivity results are presented for the $k$-step extrapolated BDF schemes (eBDF$k$) and the $k$-step explicit Adams methods, which are also known as the Adams–Bashforth methods (AB$k$), $k = 3, 4$. Here we also list the theoretical bounds on the Courant numbers for these methods that were obtained in § 4. The experimental bounds were found with exact starting values and with high-order Runge–Kutta starting procedures, giving approximately the same values.

TABLE 5.1
*Positivity test for higher-order methods. Experimental Courant numbers and theoretical bounds.*

| | eBDF3 | AB3 | eBDF4 | AB4 |
|---|---|---|---|---|
| Theoretical | $\frac{7}{18} \approx 0.39$ | $\frac{84}{529} \approx 0.16$ | $\frac{7}{32} \approx 0.22$ | 0 |
| Experimental | 0.43 | 0.23 | 0.30 | 0.11 |

**5.2. Nonlinear accuracy test.** To compare the explicit linear multistep methods for a nonlinear example, we consider the Burgers equation

$$u_t + (u^2)_x = 0\,, \quad 0 \le x \le 1\,,\ 0 \le t \le \frac{1}{4},$$

with periodic boundary conditions and initial profile $u(x, 0)$ given by the block function which equals 0 for $x \in (0, \frac{1}{2}]$ and 1 for $x \in (\frac{1}{2}, 1]$. For increasing time the solution $u(x, t)$ consists of a shock at $x = t$ and a rarefaction wave between $\frac{1}{2} \le x \le \frac{1}{2} + 2t$; see Figure 5.3.

Spatial discretization is performed with the flux-limited scheme of van Leer [19], which combines a second-order upwind-biased discretization (in smooth solution regions) with first-order upwind fluxes; see also [14, p. 180] and [8]. For this test, with $u \in [0, 1]$, it can be shown that the forward Euler method is TVD and positive for Courant numbers $\nu = 2\Delta t/\Delta x \le \frac{1}{2}$. However to achieve a reasonable accuracy the Courant number should be taken significantly smaller than $\frac{1}{2}$, because otherwise the

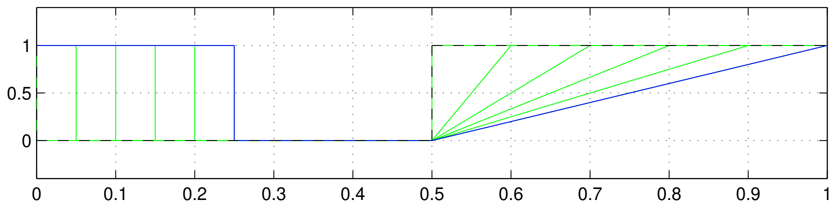



FIG. 5.3. *Solution of Burgers' equation for $0 \leq x \leq 1$ at $t = 0$ [dashed] and $t = \frac{1}{4}$ [solid line]. The light gray lines indicate the time evolution.*

rarefaction wave suffers from compression due to linear instability of the forward Euler method with the second-order discretization. For Courant numbers in the range $[\frac{1}{2}, 1]$ the forward Euler method is no longer strictly TVD, but the oscillations are quite small. This can be understood heuristically by the observation that with the first-order upwind discretization the forward Euler method is TVD up to $\nu = 1$, and in non-smooth regions, where monotonicity matters most, the flux-limited scheme becomes close to first-order upwinding.

The same observations apply to the multistep methods used in this test; the theoretical limits for monotonicity can be nearly doubled without introducing large temporal errors. Still, the theoretical predictions based on the threshold values $C^*_{LM}$ show up when compared with forward Euler. The choice of starting procedures did have only minor significance; for the results presented here the first step was taken with the forward Euler method. In this test the discrete $L_1$-errors

$$\Delta x \sum_{j=1}^{m} |u(x_j, t_n) - w_j^n| , \qquad m\,\Delta x = 1,$$

were measured for different Courant numbers in the range $[0, 1]$ at time $t_n = \frac{1}{4}$. The test was performed on a fixed grid with mesh width $\Delta x = 10^{-2}$. The results for various second-order 2-step methods (3.10) are shown in Figure 5.4.

The methods in the top picture (a) of Figure 5.4 are the extrapolated BDF2 scheme (eBDF2, $\xi = \frac{2}{3}$), the 2-step Adams–Bashforth method (AB2, $\xi = 1$), and the second-order modified 2-step method (Sh2) of Shu [15],

$$w_n = \frac{4}{5} w_{n-1} + \frac{1}{5} w_{n-2} + \frac{8}{5} \Delta t F(w_{n-1}) - \frac{2}{5} \Delta t \tilde{F}(w_{n-2}) , \tag{5.2}$$

which is the modified form of (3.10), $\xi = \frac{6}{5}$, with threshold factor $\tilde{K}_{LM} = \frac{1}{2}$; see Remark 2.1. This scheme is more expensive in CPU time and in this test it does not perform as well as the other two, of which the extrapolated BDF2 scheme has a slight advantage over the explicit 2-step Adams method.

In the bottom picture (b) of Figure 5.4 the results are given for the methods (3.10) with $\xi = \frac{1}{5}, \frac{6}{5}, \frac{9}{5}$. For comparison results for the forward Euler method are also included. As predicted by the bound $C^*_{LM}$ of (3.11), the method with $\xi = \frac{9}{5}$ can only be used with small Courant numbers. The method with $\xi = \frac{1}{5}$ does provide results for larger Courant numbers but its accuracy deteriorates for large $\nu$. The results with $\xi = \frac{6}{5}$ are intermediate, where it should be noted that this method is competitive with the more expensive modified method (5.2) which is based on the same parameter choice.

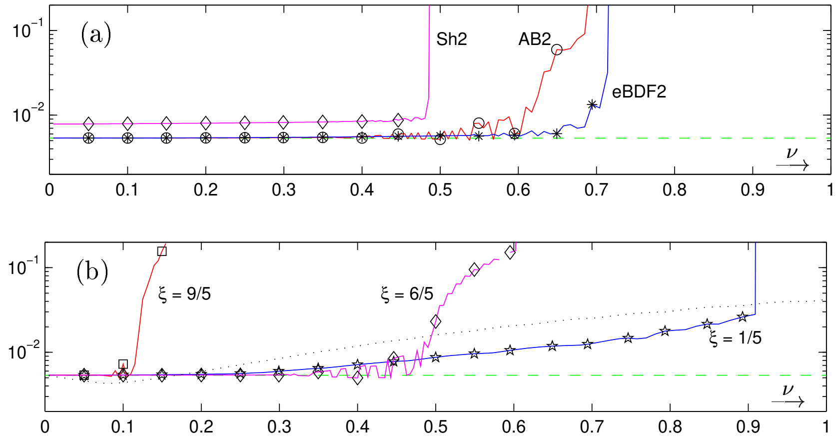



FIG. 5.4. *Burgers' equation, $L_1$-errors versus Courant numbers $\nu$ for explicit two-step methods (3.10). (a) eBDF2 [$\xi = \frac{2}{3}$], AB2 [$\xi = 1$], and Sh2 [$\tilde{\xi} = \frac{6}{5}$]; (b) $\xi = \frac{1}{5}$, $\xi = \frac{6}{5}$, and $\xi = \frac{9}{5}$, together with forward Euler results [dotted line]. The light dashed horizontal line indicates the spatial error.*

In Figure 5.4 we also indicate the spatial errors of the flux-limited van Leer discretization with $\Delta x = 10^{-2}$; that is, the $L_1$ difference between semi-discrete solution and PDE solution at $t = \frac{1}{4}$ on this spatial grid. The modified scheme (5.2) gives larger errors for $\nu \to 0$. This is due to the use of $\tilde{F}$ which introduces some extra numerical dissipation, in particular at the bottom and top of the rarefaction wave. With most of the methods the $L_1$-errors show oscillations as a function of $\nu$ before becoming unbounded. This is an onset of instability, due to spatial oscillations at the top of the shock or rarefaction wave.

Method (3.10) with $\xi = \frac{1}{5}$ could be used with relatively large Courant numbers $\nu$ without becoming unstable, but for the larger values $\nu$ the results are no longer accurate, due to compression of the rarefaction wave. With the forward Euler method this compression is much more pronounced. Time-stepping methods with high order will be mostly beneficial for smooth solutions. The present test is primarily intended to show the relevance of monotonicity. This should also be kept in mind with the results for the third-order methods below.

The fact that the starting procedures did not matter significantly in this test is somewhat more surprising than with the previous linear example. In the derivation of our theoretical results for nonlinear problems no relation at all was assumed between terms like $F(w_n)$ and $F(w_{n-1})$. For grid points $x_j$ adjacent to the shock the spatial discretization becomes close to the first-order upwind scheme and elsewhere we will have $F_j(w_n) = F_j(w_{n-1}) + \mathcal{O}(\Delta t)$. It is not clear, however, how such arguments could be used in a rigorous mathematical fashion.

In the same way some 3-step methods were tested. The results are shown in Figure 5.5. Here we selected the extrapolated BDF3 scheme (eBDF3) and the 3-step Adams–Bashforth method (AB3). Also included are the results for the second-order 3-step method

$$w_n = \frac{3}{4}w_{n-1} + \frac{1}{4}w_{n-3} + \frac{3}{2}\Delta t F(w_{n-1}) \tag{5.3}$$

of Shu [15] with threshold value $K_{LM} = \frac{1}{2}$, which is optimal among the 3-step methods of order 2; see also Lenferink [12]. In the figure this method is indicated as Sh2,3.

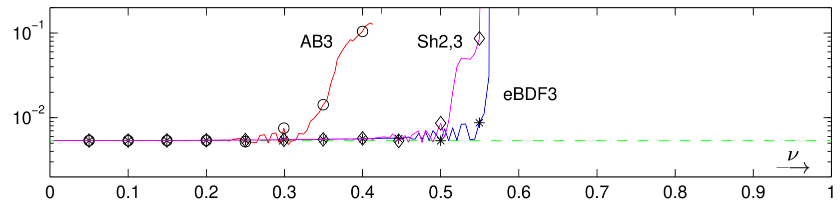



FIG. 5.5. *Burgers' equation, $L_1$-errors versus Courant numbers $\nu$ for the explicit three-step methods eBDF3, AB3, and Sh2,3. The light dashed horizontal line indicates the spatial error.*

Since this is a second-order method, comparison with AB2 or eBDF2 is actually more appropriate. As expected from the theoretical bounds, the eBDF3 scheme does perform better than the AB3 method. Also with these 3-step methods, starting procedures turned out not to be very influential. Here a standard 2-stage second-order Runge–Kutta method was used.

In summary we can say that the theoretical step-size restrictions of § 3 for the monotonicity property (1.2) are probably somewhat pessimistic, but the step-size restrictions under which the more general boundedness property (3.7) could be proved give a good indication for the applicability of the various methods.

**6. Summary and Conclusions.** We have shown that inclusion of starting procedures in multistep schemes allows for statements on monotonicity (1.2) and boundedness (3.7) with classes of methods that are important in practice (such as the Adams and BDF-type methods). We find that the standard second-order two-step methods AB2 (second-order Adams-Bashforth)

$$w_n = w_{n-1} + \frac{3}{2}\Delta t F(w_{n-1}) - \frac{1}{2}\Delta t F(w_{n-2})\,,$$

and eBDF2 (second-order extrapolated BDF)

$$w_n = \frac{4}{3}w_{n-1} - \frac{1}{3}w_{n-2} + \frac{4}{3}\Delta t F(w_{n-1}) - \frac{2}{3}\Delta t F(w_{n-2})\,,$$

have more relaxed time-stepping restrictions than schemes with positive coefficients when monotonicity-preservation is required. Similarly, the well-known higher-order schemes AB3 (4.6), eBDF3 (4.4) and eBDF4 (4.5) have more relaxed time-stepping restrictions than schemes with positive coefficients when related boundedness properties are required. Numerical tests confirm the results of these theoretical studies: these standard methods performed better than specially constructed methods with positive coefficients. We particularly recommend the well-known extrapolated BDF schemes.

Finally we have found that implicit second-order two-step schemes are not competitive with the explicit ones when monotonicity properties are crucial, generalizing one of the results in [8] for the two-step BDF-type schemes. The restrictions on the step sizes for monotonicity and boundedness with the implicit schemes were shown to be only marginally better than for the explicit schemes, and in practical computations this will not be enough to justify the increase in computational work with the implicit schemes.

REFERENCES


[1] U.M. Ascher, S.J. Ruuth and B. Wetton, *Implicit-explicit methods for time-dependent PDE's*. SIAM J. Numer. Anal. 32 (1995), pp. 797-823.

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

[3] G. Dahlquist, *Error analysis for a class of methods for stiff nonlinear initial value problems*. Procs. Dundee Conference 1975, Lectures Notes in Mathematics 506, G.A. Watson (ed.), Springer-Verlag, Berlin, 1976, pp. 60-74.

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

[5] A. Harten, *On a class of high resolution total-variation-stable finite difference schemes*, SIAM J. Numer. Anal. 21 (1984), pp. 1-23.

[6] E. Hairer, S.P. Nørsett and G. Wanner, *Solving Ordinary Differential Equations I – Nonstiff Problems*. Second edition, Springer Series in Computational Mathematics 8, Springer Verlag, 1993.

[7] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations II – Stiff and Differential-Algebraic Problems*. Second edition, Springer Series in Computational Mathematics 14, Springer Verlag, 1996.

[8] W. Hundsdorfer, *Partially implicit BDF2 blends for convection dominated flows*. SIAM J. Numer. Anal. 38 (2001), pp. 1763-1783.

[9] W. Hundsdorfer and J. Jaffré, *Implicit-explicit time stepping with spatial discontinuous finite elements*; to appear Appl. Num. Math.

[10] J.F.B.M. Kraaijevanger, *Contractivity of Runge-Kutta methods*. BIT 31 (1991), pp. 482-528.

[11] C.B. Laney, *Computational Gasdynamics*. Cambridge University Press, 1998.

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

[13] ———, *Contractivity preserving implicit linear multistep methods*. Math. Comp. 56 (1991), pp. 177-199.

[14] R.J. LeVeque, *Numerical Methods for Conservation Laws*. Lectures in Mathematics, ETH Zürich, Birkhäuser Verlag, 1992.

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

[16] C.-W. Shu and S. Osher, *Efficient implementation of essentially non-oscillatory shock-capturing schemes*. J. Comput. Phys. 77 (1988), pp. 439-471.

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

[18] R.J. Spiteri and S.J. Ruuth, *A new class of optimal high-order strong-stability-preserving time-stepping schemes*. SIAM J. Numer. Anal., 40 (2002) pp. 469-491.

[19] B. van Leer, *Towards the ultimate conservative difference scheme II. Monotonicity and conservation combined in a second order scheme*. J. Comput. Phys. 14 (1974), pp. 361-370.