BIT Numer Math (2011) 51:909–936
DOI 10.1007/s10543-011-0349-x

BIT

Special boundedness properties in numerical initial value problems

W. Hundsdorfer · A. Mozartova · M.N. Spijker

Received: 18 August 2010 / Accepted: 15 June 2011 / Published online: 21 September 2011

Abstract For Runge-Kutta methods, linear multistep methods and other classes of general linear methods much attention has been paid in the literature to important nonlinear stability properties known as total-variation-diminishing (TVD), strong stability preserving (SSP) and monotonicity. Stepsize conditions guaranteeing these properties were studied by Shu and Osher (J. Comput. Phys. 77:439–471, 1988) and in numerous subsequent papers. Unfortunately, for many useful methods it has turned out that these properties do not hold. For this reason attention has been paid in the recent literature to the related and more general properties called total-variation-bounded (TVB) and boundedness.

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

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

Presented at the BIT50 conference in Lund, Sweden 17–20 June 2010.

Communicated by Gustaf Söderlind.

W. Hundsdorfer · A. Mozartova (✉)
CWI, P.O. Box 94079, 1090-GB Amsterdam, The Netherlands
e-mail: a.mozartova@cwi.nl

W. Hundsdorfer
e-mail: willem.hundsdorfer@cwi.nl

M.N. Spijker
Mathematical Institute, Leiden University, P.O. Box 9512, 2300-RA Leiden, The Netherlands
e-mail: spijker@math.leidenuniv.nl

Keywords Initial value problem · Method of lines (MOL) · Ordinary differential equation (ODE) · General linear method (GLM) · Total-variation-diminishing (TVD) · Strong-stability-preserving (SSP) · Monotonicity · Total-variation-bounded (TVB) · Boundedness

Mathematics Subject Classification (2000) 65L05 · 65L06 · 65L20 · 65M20

1 Introduction

1.1 Bounds for numerical approximations

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

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

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

$$y_i = \sum_{j=1}^{l} s_{ij} x_j + \Delta t \cdot \sum_{j=1}^{m} t_{ij} F_j(y_j) \quad (1 \leq i \leq m). \tag{1.2}$$

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

The process (1.2) is in particular relevant to the important and very large class of general linear methods (GLMs), introduced by Butcher [1], cf. also e.g. Butcher [2, 3], Hairer and Wanner [7], Hairer, Nørsett and Wanner [8]. This class comprises, e.g., all Runge-Kutta methods, linear multistep methods and multistep-multistage variants thereof.

We can represent $N \geq 1$ consecutive steps of any GLM canonically by a process of the generic type (1.2) with $m = N(s+r)$, where s is the number of internal stages and r the number of external stages computed at each step of the GLM. In this situation, the vectors x_j $(1 \leq i \leq l)$ stand for the starting vectors of the GLM, whereas the vectors y_i $(1 \leq i \leq m)$ represent the $N \cdot s$ internal and $N \cdot r$ external stage approximations computed during the N steps. Furthermore, the parameter matrices $S = (s_{ij}) \in \mathbb{R}^{m \times l}$, $T = (t_{ij}) \in \mathbb{R}^{m \times m}$, corresponding to the process (1.2), are determined by the number of steps N as well as by the coefficients of the given GLM. Detailed examples of such representations, as well as alternative representations of actual multistep-multistage methods, can be found in Spijker [22] for $N = 1$ and in Hundsdorfer, Mozartova and Spijker [17] for $N > 1$; cf. also Sect. 4 of the present paper.

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

section, we assume $\|\cdot\|$ to be a *convex functional*, i.e.

$$\|\lambda v + (1-\lambda)w\| \le \lambda\|v\| + (1-\lambda)\|w\| \quad (\text{for } 0 \le \lambda \le 1 \text{ and } v, w \in \mathbb{V}). \tag{1.3}$$

In applications, $\|\cdot\|$ will often be a norm or seminorm, see (2.8) below. But, more general convex functionals are useful as well, notably in connection with discrete maximum principles and preservation of nonnegativity; cf. e.g. Spijker [22] and Sect. 3.4 of the present paper.

For the generic process (1.2), as well as for special instances thereof, much attention has been paid in the literature to the derivation of suitable upper bounds for $\|y_i\|$, in terms of the input vectors x_j, under the basic assumption that for given $\tau_0 > 0$

$$\|v + \tau_0 F_i(v)\| \le \|v\| \quad (\text{for } 1 \le i \le m, \text{ and } v \in \mathbb{V}); \tag{1.4}$$

cf. e.g. Ferracina and Spijker [4], Gottlieb, Ketcheson and Shu [5], Gottlieb, Shu, and Tadmor [6], Higueras [9, 10], Hundsdorfer and Ruuth [13, 14], Hundsdorfer, Ruuth and Spiteri [15], Shu and Osher [20], Spijker [22].

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

$$s_{i1} + s_{i2} + \cdots + s_{il} = 1 \quad (1 \le i \le m). \tag{1.5}$$

In case the process (1.2) stands for *just one step* ($N = 1$) of a GLM, this condition corresponds to preconsistency, cf. Spijker [22], Butcher [3]. Henceforth we will refer to (1.5) as the *preconsistency condition* for process (1.2).

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

$$\|y_i\| \le \max_{1 \le j \le l} \|x_j\| \quad (\text{for } 1 \le i \le m) \tag{1.6}$$

has received much attention. The process has been called *monotonic* or *strongly stable* (for given stepsize Δt, vector space $\mathbb{V}$, functional $\|\cdot\|$ and functions $F_i : \mathbb{V} \to \mathbb{V}$) if the last bound holds whenever x_i and y_i satisfy (1.2). Algebraic characterizations were derived of stepsize-coefficients γ with the following important property:

> Condition $0 < \Delta t \le \gamma \cdot \tau_0$ implies monotonicity, whenever $\mathbb{V}$ is a vector space, $\|\cdot\|$ a convex function on $\mathbb{V}$, and the functions $F_i : \mathbb{V} \to \mathbb{V}$ satisfy the basic assumption (1.4); (1.7)

see e.g. Spijker [22] and the references therein.

Unfortunately, for many useful GLMs there exists *no* $\gamma > 0$ such that the above property is present, when one step of the method ($N = 1$) is represented as a preconsistent process of the form (1.2); some examples are given in Sect. 4 of this paper. Furthermore, in important situations, processes of generic type (1.2) arise which even fail to satisfy the preconsistency condition (1.5). Sometimes, deeper insight into a given GLM can be gained by representing $N > 1$ consecutive steps of the method as such a generic process; cf. Sect. 4.

These difficulties have led various authors to study bounds for $\|y_i\|$ that differ from the monotonicity bound (1.6) by a factor $\mu \geq 1$, i.e.

$$\|y_i\| \leq \mu \cdot \max_{1 \leq j \leq l} \|x_j\| \quad (\text{for } 1 \leq i \leq m). \tag{1.8}$$

Such general bounds are formally weaker than (1.6) but still useful because they can reveal essential boundedness properties of the numerical methods under consideration, like the property of being *total-variation bounded*—for this important concept see e.g. LeVeque [18]. Stepsize conditions corresponding to general bounds (1.8) were derived, e.g., in Ruuth and Hundsdorfer [19], Hundsdorfer, Mozartova and Spijker [17].

The general bounds obtained thus far in the literature are relevant in cases where the monotonicity property (1.7) is violated or even the preconsistency condition (1.5) is not in force. On the other hand, these bounds suffer still from the following two inconveniences: (1) the corresponding stepsize conditions, of type $0 < \Delta t \leq \gamma \cdot \tau_0$, involve complicated conditions on γ which are often difficult to check in practice; (2) the general bounds are relevant to seminorms but not to any wider class of functionals satisfying (1.3).

1.2 Scope of the paper

The main purpose of the present paper is to establish stepsize conditions guaranteeing special bounds for the generic process (1.2), thereby circumventing the two inconveniences just mentioned above. We shall find special bounds which can still be present in cases where the monotonicity property (1.7) or the preconsistency condition (1.5) is violated, and which are the best possible in a definite sense. Moreover, these special bounds are relevant to a class of functionals $\|\cdot\|$ that is wider than the class of seminorms. Finally, and most importantly in view of applications, the corresponding stepsize conditions $0 < \Delta t \leq \gamma \cdot \tau_0$ involve a condition on γ which is easier to check in practice than the conditions relevant to the general bounds given in the literature.

In Sect. 2 of this paper, we review and extend bounds and monotonicity results for the generic process (1.2), as given thus far in the existing literature. In the Sects. 2.1, 2.2, we give a brief review of known monotonicity results for the generic process (1.2), thereby focusing on a classical simple condition on the stepsize-coefficient γ. Moreover, we consider a property which is a-priori more refined than pure monotonicity and we characterize in Theorem 2.4 stepsize conditions guaranteeing this property. In Sect. 2.3, we specify two generalizations of the bound $\|y_i\| \leq \max_{1 \leq j \leq l} \|x_j\|$ which are relevant to generic processes which need not be preconsistent. Theorem 2.5 characterizes stepsize conditions guaranteeing these generalizations.

Section 3 contains the main theoretical findings of the paper. In Sect. 3.1, we formulate explicitly, for the generic process (1.2), the special bounds mentioned above (for $\|y_i\|$ in terms of $\|x_j\|$), and mention three features which distinguish them from more general standard bounds (1.8). In Sect. 3.2, we study, in the situation of these special bounds, the characterizations provided by Theorem 2.5. We find simplified versions of these characterizations, viz. (3.4)–(3.7). In Sect. 3.3, we study the special

bounds for the case of seminorms $\|\cdot\|$; we find that these bounds are the best possible in the sense specified by Theorem 3.4. The main theorem of Sect. 3.3, Theorem 3.5, gives simplified criteria for stepsize conditions guaranteeing the special bounds. Section 3.4 deals with the special bounds for the case of a natural class of functionals—the so-called sublinear functionals—which is essentially larger than the class of seminorms. Theorem 3.8 reveals the surprising fact that the special bounds are the only bounds which make sense in the context of general sublinear functionals. The main theorem of Sect. 3.4, Theorem 3.9, gives among other things a mild condition under which the classical simple condition on γ, reviewed in Sect. 2, characterizes stepsize conditions guaranteeing the special bounds for sublinear functionals.

In Sect. 4 we illustrate the significance of the special boundedness theory by applying it to some concrete numerical methods. For most of these methods, the monotonicity results, as given in the literature, see e.g. [5, 22], are *not* (directly) applicable. Moreover, the boundedness theory, as given e.g. in [17] would lead to very complicated conditions. In Sect. 4.2 we study the two-step Adams-Bashforth method. When writing one step of the method in a standard fashion as a generic process of type (1.2), there is no $\gamma > 0$ such that the monotonicity property (1.7) is present. But, by writing $N \geq 1$ steps of the method judiciously in the generic form (1.2), it turns out that Theorems 3.5, 3.9 yield conclusions which can nicely be interpreted in terms of boundedness and nonnegativity preservation of the method. In Sect. 4.3 we analyze a large class of k-step methods, containing both predictor-corrector methods and hybrid multistep methods. The monotonicity results, known from the literature, are not valid for many popular schemes of this class. By applying Theorem 3.9, we will show that for many methods of practical interest relevant boundedness properties are valid.

2 Reviewing and extending results from the literature

2.1 Preliminaries

Let I stand for the identity matrix of order m, and let $S = (s_{ij})$, $T = (t_{ij})$ denote the coefficient matrices corresponding to the generic process (1.2). Similarly as in [17, 22], we introduce the matrices

$$P = (p_{ij}) = (I + \gamma T)^{-1}(\gamma T), \qquad R = (r_{ij}) = (I + \gamma T)^{-1} S. \tag{2.1}$$

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

When working with P and R, the invertibility of $I + \gamma T$ will be implicitly assumed. Actually, to study boundedness properties this assumption can be made without loss of generality. To see this, we formulate the following lemma, which is an analogue of a result from [22, Lemma 4.2]. The proof of this lemma is compact, so we repeat it here.

Lemma 2.1 (Invertibility of $I + \gamma T$) *Let $\tau_0 > 0$, $\gamma > 0$ be given and $\Delta t = \gamma \cdot \tau_0$. Let $\mathbb{V} = \mathbb{R}$, $\|\cdot\| = |\cdot|$ and assume μ is a constant such that the general bound (1.8) holds whenever $F_i : \mathbb{V} \to \mathbb{V}$ fulfill the basic assumption (1.4) and $y_i, x_j \in \mathbb{V}$ satisfy (1.2). Then $I + \gamma T$ is invertible.*

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

We define $F_i(v) = -(1/\tau_0)v$ (for all $v \in \mathbb{V}$), so that the basic assumption (1.4) is fulfilled with $\|\cdot\| = |\cdot|$. Clearly, (1.2) is satisfied, with $\Delta t = \gamma \cdot \tau_0$, by the vectors $x_i = 0$ $(1 \le i \le l)$ and $y_i = \eta_i$ $(1 \le i \le m)$. By applying (1.8), there follows $|\eta_i| = |y_i| \le \mu \cdot \max_j |x_j| = 0$, therefore $\eta = 0$. □

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

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

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

$$P \ge 0, \qquad R \ge 0. \tag{2.2}$$

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

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

We shall recall briefly some concepts and results from the literature which are related to the monotonicity property (1.7). The next two theorems follow directly from [22, Theorems 2.2, 2.4].

Theorem 2.2 (Criterion for monotonicity with arbitrary convex functional $\|\cdot\|$) *Consider a generic process* (1.2) *satisfying the preconsistency condition* (1.5). *Let* $\gamma > 0$ *be given. Then the monotonicity property* (1.7) *is present, if and only if* γ *satisfies the classical condition* (2.2).

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

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

Theorem 2.3 (Conditions on S, T) *Let the preconsistency condition* (1.5) *be fulfilled. Then there is a* $\gamma > 0$ *satisfying the classical condition* (2.2), *if and only if* $S \ge 0$, $T \ge 0$, $\mathrm{Inc}(TS) \le \mathrm{Inc}(S)$ *and* $\mathrm{Inc}(T^2) \le \mathrm{Inc}(T)$.

Clearly, for given matrices S, T, it is rather easy, by applying Theorems 2.2 and 2.3, to see whether there is a positive stepsize-coefficient γ such that the monotonicity property (1.7) is present.

For preconsistent processes, the classical condition (2.2) will be proved to imply an interesting variant of the standard monotonicity bound $\|y_i\| \le \max_{1 \le j \le l} \|x_j\|$.

The variant is as follows:

$$\|y_i\| \le \sum_{j=1}^{l} |s_{ij}| \|x_j\| \quad (1 \le i \le m). \tag{2.3}$$

Note that, when all s_{ij} are nonnegative, the last bound is of particular interest because it is more refined and gives, in general, more information than the standard monotonicity bound. Clearly, all s_{ij} are nonnegative as soon as the monotonicity property (1.7) is present for some $\gamma > 0$; cf. Theorems 2.2, 2.3.

We shall say that *process* (1.2) *satisfies the bound* (2.3) (for given stepsize Δt, vector space $\mathbb{V}$, functional $\|\cdot\|$ and functions $F_i : \mathbb{V} \to \mathbb{V}$), if (2.3) holds whenever x_i and $y_i \in \mathbb{V}$ satisfy (1.2). The following (refined) property is an obvious variant of the standard monotonicity property (1.7):

> Condition $0 < \Delta t \le \gamma \cdot \tau_0$ implies the bound (2.3), whenever $\mathbb{V}$ is a vector space, $\|\cdot\|$ a convex functional on $\mathbb{V}$, and the functions $F_i : \mathbb{V} \to \mathbb{V}$ satisfy the basic assumption (1.4). (2.4)

The following theorem shows that this property is present under the same conditions as the standard monotonicity property (1.7).

Theorem 2.4 (Criterion for property (2.4)) *Consider a generic process* (1.2) *satisfying the preconsistency condition* (1.5). *Let* $\gamma > 0$ *be given. Then property* (2.4) *is present, if and only if* γ *satisfies the classical condition* (2.2).

Proof 1. Let the basic assumption (1.4) be fulfilled, and let $0 < \Delta t \le \gamma \cdot \tau_0$, where γ satisfies the classical condition (2.2). We denote by E_k the $k \times 1$ matrix with all entries equal to 1. Note that, since $R = (I - P)S$ and $SE_l = E_m$, we have $RE_l + PE_m = E_m$, i.e. $\sum_{j=1}^{l} r_{ij} + \sum_{j=1}^{m} p_{ij} = 1$.

We rewrite process (1.2), using the notations (2.1), in the form

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

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

$$[\|y_i\|] \le (I + \gamma T)^{-1} S[\|x_j\|] + (I - (I + \gamma T)^{-1})[\|y_i\|], \tag{2.5}$$

i.e. $(I + \gamma T)^{-1}[\|y_i\|] \le (I + \gamma T)^{-1} S[\|x_j\|]$. In view of Theorem 2.3, the matrices S and $I + \gamma T$ are nonnegative, so that the bound (2.3) is in force. Property (2.4) has thus been proved.

2. Conversely, assume property (2.4) is present. We shall use the notation

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

Applying property (2.4) in the special situation where $\mathbb{V} = \mathbb{R}$, $\|v\| = v$, $F_i = 0$, $x_j = \mathrm{sgn}(s_{i_0 j})$, we see from the corresponding bound (2.3) that $\sum_j |s_{i_0 j}| \le \sum_j |s_{i_0 j}| \mathrm{sgn}(s_{i_0 j})$, so that $s_{i_0 j} \ge 0$. Hence, all $s_{ij} \ge 0$.

In the general situation, the bound (2.3) thus implies, for $1 \le i \le m$,

$$\|y_i\| \le \sum_n s_{in} \|x_n\| \le \left(\sum_n s_{in} \right) \max_j \|x_j\| = \max_j \|x_j\|.$$

It follows that property (2.4) implies the standard monotonicity property (1.7) and—by Theorem 2.2—also the classical condition (2.2). □

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

With an eye to cases where the preconsistency condition (1.5) or the monotonicity property (1.7) (with $\gamma > 0$) is violated, we shall review and extend, in this section, some results from the literature about bounds which are more general than those considered above. We shall focus on the general bounds

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

$$\|y_i\| \le \sum_{j=1}^{l} \mu_{ij} \|x_j\| \quad (\text{for } 1 \le i \le m), \tag{2.7}$$

where for the time being μ_i and μ_{ij} denote arbitrary coefficients. Clearly, when $\mu_i = 1$, $\mu_{ij} = |s_{ij}|$, these bounds reduce to the bounds (1.6) and (2.3), respectively.

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

$$\|v + w\| \le \|v\| + \|w\| \quad \text{and} \quad \|\lambda v\| = |\lambda| \|v\| \quad (\text{for all real } \lambda \text{ and } v, w \in \mathbb{V}). \tag{2.8}$$

We shall say that *process* (1.2) *satisfies the bound* (2.6) *or* (2.7) (for given stepsize Δt, vector space $\mathbb{V}$, seminorm $\|\cdot\|$ and functions $F_i : \mathbb{V} \to \mathbb{V}$), if (2.6) or (2.7), respectively, holds whenever x_i and $y_i \in \mathbb{V}$ satisfy (1.2). Below we shall focus on stepsize-coefficients γ which are related to the above two general bounds by means of the following two properties:

> Condition $0 < \Delta t \le \gamma \cdot \tau_0$ implies that process (1.2) satisfies the bound (2.6), whenever $\mathbb{V}$ is a vector space, $\|\cdot\|$ a seminorm on $\mathbb{V}$, and the functions $F_i : \mathbb{V} \to \mathbb{V}$ satisfy the basic assumption (1.4), (2.9)

> Condition $0 < \Delta t \le \gamma \cdot \tau_0$ implies that process (1.2) satisfies the bound (2.7), whenever $\mathbb{V}$ is a vector space, $\|\cdot\|$ a seminorm on $\mathbb{V}$, and the functions $F_i : \mathbb{V} \to \mathbb{V}$ satisfy the basic assumption (1.4). (2.10)

In formulating conditions on γ for these properties, we need some notations. For any matrix $A = (a_{ij})$, we define the matrix $|A|$ by $|A| = (|a_{ij}|)$. For square matrices

A, we denote the *spectral radius* by $\mathrm{spr}(A)$. Furthermore we introduce the $m \times 1$ matrix

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

and the $m \times l$ matrix

$$(\mu_{ij}) = \begin{pmatrix} \mu_{11} & \cdots & \mu_{1l} \\ \vdots & & \vdots \\ \mu_{m1} & \cdots & \mu_{ml} \end{pmatrix}. \tag{2.11}$$

We shall relate properties (2.9) and (2.10), respectively, to the following two requirements:

$$\mathrm{spr}(|P|) < 1 \quad \text{and} \quad (I - |P|)^{-1}|R|E_l \le (\mu_i), \tag{2.12}$$

$$\mathrm{spr}(|P|) < 1 \quad \text{and} \quad (I - |P|)^{-1}|R| \le (\mu_{ij}). \tag{2.13}$$

Note that, for given coefficient matrices S, T, these requirements amount to conditions on γ—cf. the definition (2.1) of P, R.

The following theorem is a variant of a result given earlier in the literature, see [17]. In fact, when all μ_i are equal to each other, part (I) of the theorem is an immediate corollary to Theorem 2.2 in the paper just mentioned.

Theorem 2.5 (Criteria for the properties (2.9), (2.10)) *Consider an arbitrary generic process* (1.2). *Let* $\gamma > 0$ *and arbitrary* μ_i, μ_{ij} *be given. Then the following two propositions are valid*:

(I) *Property* (2.9) *is present, if and only if* γ *is such that condition* (2.12) *is fulfilled.*
(II) *Property* (2.10) *is present, if and only if* γ *is such that condition* (2.13) *is fulfilled.*

Proof The conditions (2.12), (2.13) imply (2.9) and (2.10), respectively, by similar arguments as used in part 1 of the proof of Theorem 2.4. Using the arguments of the mentioned proof and (2.8), we get now $[\|y_i\|] \le |R|[\|x_j\|] + |P|[\|y_i\|]$ instead of (2.5). There follows $[\|y_i\|] \le (I - |P|)^{-1}|R|[\|x_j\|]$. By (2.13) we arrive at property (2.10). Since $(I - |P|)^{-1}|R|[\|x_j\|] \le (I - |P|)^{-1}|R|E_l \cdot \max_k \|x_k\|]$, by (2.12) we arrive at property (2.9).

The necessity of the conditions (2.12) and (2.13) can be proved by almost the same arguments as already given in [17, Sect. 4.2]. □

Theorem 2.5 has a wider scope, certainly, than the theorems of Sect. 2.2, in that μ_i and μ_{ij} are arbitrary coefficients and the preconsistency condition (1.5) is not needed.

On the other hand, it is in general much more difficult to see whether the conditions (2.12), (2.13) are fulfilled than to check the classical condition (2.2). Moreover, unlike the theorems in Sect. 2.2, Theorem 2.5 is only relevant to seminorms (and e.g. not to certain convex functionals arising in connection with discrete maximum principles and preservation of nonnegativity). These obvious weaknesses of Theorem 2.5 are among the reasons for dealing in the following section with bounds of a very special form.

3 Bounds of a special form

3.1 Special choices for μ_i, μ_{ij}

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

$$\mu_i = \sum_j |s_{ij}| \quad \text{and} \quad \mu_{ij} = |s_{ij}|, \tag{3.1}$$

so that the general bounds (2.6), (2.7), respectively, take the special form

$$\|y_i\| \le \left(\sum_{j=1}^{l} |s_{ij}| \right) \cdot \max_{1 \le j \le l} \|x_j\| \quad (1 \le i \le m), \tag{3.2}$$

$$\|y_i\| \le \sum_{j=1}^{l} |s_{ij}| \|x_j\| \quad (1 \le i \le m). \tag{3.3}$$

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

First of all, property (2.9) with $\mu_i = \sum_j |s_{ij}|$, as well as property (2.10) with $\mu_{ij} = |s_{ij}|$, can be interpreted as an extension, to *all* F_i (satisfying the basic assumption(1.4)), of a bound which is trivially fulfilled when $F_i(v) \equiv 0$. In fact, in the subsequent Theorem 3.4, we shall see that the above special bounds (3.2), (3.3) are the *best possible*, in the sense that, for any $\gamma > 0$, the general boundedness properties (2.9), (2.10) *cannot* be valid with coefficients smaller than (3.1).

Secondly, as will be seen in Theorem 3.8 below, the equalities $\mu_i = \sum_j |s_{ij}|$ are necessary in order that any bound (2.6) holds for a natural class of functionals $\|\cdot\|$ that is larger than the class of seminorms. Similarly, the equalities $\mu_{ij} = |s_{ij}|$ must be fulfilled in order that any bound (2.7) holds for this larger class.

Finally and most importantly in view of applications, the above criteria (2.12) and (2.13), respectively, will turn out to reduce to much simpler forms when $\mu_i = \sum_j |s_{ij}|$ or $\mu_{ij} = |s_{ij}|$.

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

In this section we shall analyze and simplify the above general conditions (2.12), (2.13) in the special situations $\mu_i = \sum_j |s_{ij}|$ and $\mu_{ij} = |s_{ij}|$. Our first result is as follows:

Lemma 3.1 (Conditions (2.12), (2.13) with $\mu_i = \sum_j |s_{ij}|$ and $\mu_{ij} = |s_{ij}|$) *Condition* (2.12) *with* $\mu_i = \sum_j |s_{ij}|$ *is equivalent to* (2.13) *with* $\mu_{ij} = |s_{ij}|$.

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

Suppose condition (2.12) is fulfilled. Since $(\mu_i) = |S|E_l = |(I - P)^{-1}R|E_l \le (I - |P|)^{-1}|R|E_l$, condition (2.12) is equivalent to $|S|E_l = (I - |P|)^{-1}|R|E_l$, which

can be rewritten as $|S|E_l = |P||S|E_l + |R|E_l$. Because of the last equality and $|S| = |R + PS| \le |R| + |P||S|$, it follows that

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

Hence, $(I - |P|)^{-1}|R| = |S| = (\mu_{ij})$, which implies (2.13).

Conversely, (2.13) implies $(I - |P|)^{-1}|R|E_l \le |S|E_l$, i.e. (2.12). □

Below, we shall specify situations in which the general conditions (2.12) and (2.13) can be simplified to one of the subsequent four requirements:

$$\mathrm{spr}(|P|) < 1 \quad \text{and} \quad |PS| = |P||S| \le |S|, \ |R| \le |S|; \tag{3.4}$$

$$\mathrm{spr}(|P|) < 1 \quad \text{and} \quad PS = |P|S, \ R \ge 0; \tag{3.5}$$

$$\mathrm{spr}(P) < 1 \quad \text{and} \quad P \ge 0, \ R \ge 0; \tag{3.6}$$

$$P \ge 0, \qquad R \ge 0, \qquad S \ge 0. \tag{3.7}$$

Lemma 3.2 (Simplifications of (2.12), (2.13) with $\mu_i = \sum_j |s_{ij}|$ and $\mu_{ij} = |s_{ij}|$)

(I) *Condition* (2.12) *as well as condition* (2.13), *with the choice* (3.1), *is equivalent to* (3.4).

(II) *If* $S \ge 0$, *then condition* (3.4) *is equivalent to* (3.5).

(III) *If* S *has no row equal to zero, then the three conditions* (3.5), (3.6) *and* (3.7) *are equivalent to each other.*

Proof (I) In view of Lemma 3.1, it is enough to show that condition (2.13) with $\mu_{ij} = |s_{ij}|$ is equivalent to (3.4).

From the proof of Lemma 3.1 it is evident that condition (2.13), with $\mu_{ij} = |s_{ij}|$, is equivalent to

$$\mathrm{spr}(|P|) < 1 \quad \text{and} \quad |S| = |P||S| + |R|. \tag{3.8}$$

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

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

as soon as $|PS| \le |S|$ and $|R| \le |S|$. It follows that condition (3.8) is equivalent to (3.4).

(II) Assume $S \ge 0$. In order to prove the equivalence of (3.4) and (3.5), assume $\mathrm{spr}(|P|) < 1$.

Suppose (3.4) is fulfilled. Since $R = S - PS$ and $|S| = |S - PS| + |PS|$, we have

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

which implies $R \ge 0$ and $PS = |P|S$. Therefore we have (3.5).

Conversely, from (3.5) and $S = R + PS$ we have

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

Hence, (3.5) implies (3.4).

(III) Assume S has no row equal to zero. We shall prove successively that (3.5) $\Rightarrow$ (3.6) $\Rightarrow$ (3.7) $\Rightarrow$ (3.6) $\Rightarrow$ (3.5).

Assume (3.5). Since $(I - |P|)S \ge 0$, we have $S = (I - |P|)^{-1}(I - |P|)S \ge 0$. Denoting by σ_i the entries of SE_l, we have $\sigma_i = \sum_j s_{ij} > 0$ (for $1 \le i \le m$). Since $(|P| - P)S = 0$, we have $(|P| - P)SE_l = 0$ and thus $\sum_j (|p_{ij}| - p_{ij})\sigma_j = 0$. Hence, $P \ge 0$, and therefore we have (3.6).

Furthermore, (3.6) implies that $S = (I - P)^{-1}R = (I + P + P^2 + \cdots)R \ge 0$, so that (3.6) implies (3.7).

In order to prove that property (3.7) leads to (3.6), it is enough to show that $\mathrm{spr}(P) < 1$. Introducing $D = \mathrm{Diag}(\sigma_1, \ldots, \sigma_m)$ with $\sigma_i = \sum_j s_{ij}$, we have

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

It follows that $\mathrm{spr}(P) = \mathrm{spr}(D^{-1}PD) \le 1$. Since $P = I - (I + \gamma T)^{-1} \ge 0$ has no eigenvalue 1, we conclude from the Perron-Frobenius theory (see e.g. [11, p. 503]) that $\mathrm{spr}(P) < 1$.

It is easy to see that (3.6) leads to (3.5). □

Remark 3.3 Let $\gamma > 0$. Then condition (3.6) is equivalent to

$$P \ge 0, \qquad R \ge 0, \qquad T \ge 0. \tag{3.9}$$

In order to show this, first assume (3.6). Then $I + \gamma T = (I - P)^{-1} = I + P + P^2 + \cdots \ge I$, which yields (3.9).

Next suppose that (3.9) is fulfilled. Applying the Perron-Frobenius theory as presented e.g. in [11, p. 503], it follows that there is a vector $x \in \mathbb{R}^m$ with $0 \le x \ne 0$, such that $Px = \lambda x$ where $\lambda = \mathrm{spr}(P)$. Clearly, $(I + \gamma T)^{-1}x = (I - P)x = (1 - \lambda)x$, and therefore

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

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

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

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

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

Clearly, with the choice (3.1), the general properties (2.9), (2.10), respectively, reduce to

> Condition $0 < \Delta t \le \gamma \cdot \tau_0$ implies that process (1.2) satisfies the special bound (3.2), whenever $\mathbb{V}$ is a vector space, $\|\cdot\|$ a seminorm on $\mathbb{V}$, and the functions $F_i : \mathbb{V} \to \mathbb{V}$ satisfy the basic assumption (1.4). (3.10)

Condition $0 < \Delta t \leq \gamma \cdot \tau_0$ implies that process (1.2) satisfies the special bound (3.3), whenever $\mathbb{V}$ is a vector space, $\|\cdot\|$ a seminorm on $\mathbb{V}$, and the functions $F_i : \mathbb{V} \to \mathbb{V}$ satisfy the basic assumption (1.4). (3.11)

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

But, we shall first present Theorem 3.4, which shows a crucial feature of the statements (3.10), (3.11): the theorem tells us that the estimates (3.2) and (3.3) occurring in these statements are the best possible, in the sense that, for any $\gamma > 0$, the more general properties (2.9), (2.10) cannot be valid with smaller choices for μ_i and μ_{ij} than (3.1). We have

Theorem 3.4 (Lower bounds for μ_i and μ_{ij})

(I) *If $\gamma > 0$ and μ_i are such that property (2.9) holds, then $\mu_i \geq \sum_j |s_{ij}|$ (for $1 \leq i \leq m$).*
(II) *If $\gamma > 0$ and μ_{ij} are such that property (2.10) holds, then $\mu_{ij} \geq |s_{ij}|$ (for $1 \leq i \leq m$, $1 \leq j \leq l$).*

Proof In order to prove statement (I), assume property (2.9) is valid with $\gamma > 0$ and $\mu_{i_0} < \sum_j |s_{i_0 j}|$ for some index i_0. Then, in the situation where $\mathbb{V} = \mathbb{R}$, $\|v\| = |v|$, $F_i = 0$ and $x_j = \mathrm{sgn}(s_{i_0 j})$, we have

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

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

To prove statement (II), assume property (2.10) is present with $\gamma > 0$ and $\mu_{i_0 j_0} < |s_{i_0 j_0}|$ for some pair (i_0, j_0). Then, applying this property to the situation where $\mathbb{V} = \mathbb{R}$, $\|v\| = |v|$, $F_i = 0$, $x_j = \mathrm{sgn}(s_{i_0 j})$ (for $j = j_0$) and $x_j = 0$ (for $j \neq j_0$), we arrive at

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

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

Our main result about the special boundedness properties (3.10), (3.11) will be formulated in Theorem 3.5. The theorem shows that criteria for these properties are possible which are in general much simpler than the criteria, given in Sect. 2.3, for the more general boundedness properties (2.9), (2.10).

Theorem 3.5 (Simplified criteria for the special properties (3.10) and (3.11)) *Consider an arbitrary generic process (1.2), and let $\gamma > 0$. Then the following propositions are valid*:

(I) *Condition (3.4) is necessary and sufficient for property (3.10) as well as for property (3.11).*
(II) *If $S \geq 0$, then condition (3.5) is necessary and sufficient for property (3.10) as well as for property (3.11).*

(III) *If $S \geq 0$ has no row equal to zero, then the classical condition* (2.2) *is necessary and sufficient for property* (3.10) *as well as for property* (3.11).

Proof Part (I) follows from a combination of Theorem 2.5 and Lemma 3.2.

Part (II) follows from part (I) and Lemma 3.2.

In order to prove statement (III), assume $S \geq 0$ has no row equal to zero. Combining part (II) of Theorem 3.5 and part (III) of Lemma 3.2, it follows that property (3.10) as well as (3.11) is equivalent to condition (3.7). Because $S \geq 0$, the last condition is equivalent to the classical condition (2.2). □

Property (3.11) is a-priori stronger than (3.10). Therefore the essence of the above theorem is that conditions (3.4), (3.5) and (2.2), under the appropriate assumptions on S, imply the strong statement (3.11), whereas already the weaker statement (3.10), under the same assumptions on S, implies conditions (3.4), (3.5) and (2.2).

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

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

Example 3.6 Consider the functionals $\|v\| = \|v\|_+$ and $\|v\| = \|v\|_-$ defined by

$$\|v\|_+ = \max_i v_i, \qquad \|v\|_- = -\min_i v_i \tag{3.12}$$

for $v = (v_1, v_2, \ldots, v_M)^T \in \mathbb{V} = \mathbb{R}^M$. These two functionals are no seminorms. But, they are highly relevant to *discrete maximum principles* for actual numerical processes, cf. [16, p. 118], [22, p. 1235].

Example 3.7 Another useful functional which fails to be a seminorm, is given by

$$\|v\|_0 = -\min\{0, v_1, \ldots, v_M\} \tag{3.13}$$

for $v = (v_1, v_2, \ldots, v_M)^T \in \mathbb{V} = \mathbb{R}^M$. For this non-negative functional we have $\|v\|_0 = 0$ if and only if $v \geq 0$, where this inequality is to be interpreted componentwise. One sees that any boundedness property $\max_i \|y_i\|_0 \leq \mu \cdot \max_j \|x_j\|_0$ implies the *preservation-of-nonnegativity* property: $y_i \geq 0$ (for $1 \leq i \leq m$) whenever all $x_j \geq 0$. For the practical relevance of this property, e.g. in the numerical solution of reaction-diffusion-convection equations, one may consult e.g. [16].

Since the above functionals $\|v\|_+$, $\|v\|_-$ and $\|v\|_0$ violate the seminorm condition (2.8), the material of Sects. 2.3 and 3.3 does *not* apply. It is therefore natural to look for versions of Theorems 2.5, 3.4 and 3.5 which are relevant to classes of functionals that are larger than the one specified by (2.8). Below we shall focus on functionals $\|\cdot\|$ which are only required to be *sublinear*, i.e.

$$\|\alpha v + \beta w\| \leq \alpha\|v\| + \beta\|w\| \quad \text{(for all } \alpha, \beta \geq 0 \text{ and } v, w \in \mathbb{V}\text{)}. \tag{3.14}$$

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

In line with the above, we shall study the question for which values $\gamma > 0$ the process (1.2) has either of the following two general boundedness properties:

> Condition $0 < \Delta t \leq \gamma \cdot \tau_0$ implies the bound (2.6), whenever $\mathbb{V}$ is a vector space, $\|\cdot\|$ a sublinear functional on $\mathbb{V}$, and the functions $F_i : \mathbb{V} \to \mathbb{V}$ satisfy the basic assumption (1.4). (3.15)

> Condition $0 < \Delta t \leq \gamma \cdot \tau_0$ implies the bound (2.7), whenever $\mathbb{V}$ is a vector space, $\|\cdot\|$ a sublinear functional on $\mathbb{V}$, and the functions $F_i : \mathbb{V} \to \mathbb{V}$ satisfy the basic assumption (1.4). (3.16)

The following theorem may be viewed as a variant of Theorem 3.4 tuned to sublinear functionals. It shows, somewhat surprisingly, that we loose nothing by focusing on bounds with the coefficients (3.1).

Theorem 3.8 (Expressions for μ_i and μ_{ij})

(I) *If $\gamma > 0$ and μ_i are such that property (3.15) is present, then $\mu_i = \sum_j |s_{ij}|$ $(1 \leq i \leq m)$ and $S \geq 0$.*

(II) *If $\gamma > 0$ and μ_{ij} are such that property (3.16) is present, then $\mu_{ij} = |s_{ij}|$ $(1 \leq i \leq m, 1 \leq j \leq l)$ and $S \geq 0$.*

Proof (I) It follows from Theorem 3.4 that

$$\sum_j |s_{ij}| \leq \mu_i \quad (\text{for } 1 \leq i \leq m). \tag{3.17}$$

Applying property (3.15) to the situation where $\mathbb{V} = \mathbb{R}$, $\|v\| = v$, $F_i(v) \equiv 0$, and choosing successively all $x_j = 1$ and all $x_j = -1$, we find $\sum_j s_{ij} \leq \mu_i$ and $(-\sum_j s_{ij}) \leq (-\mu_i)$, respectively. Hence

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

Combining this equality and (3.17), we arrive at proposition (I).

(II) It follows from Theorem 3.4 that

$$\sum_j |s_{ij}| \leq \sum_j \mu_{ij} \quad (\text{for } 1 \leq i \leq m). \tag{3.18}$$

Applying property (3.16) to the situation where $\mathbb{V} = \mathbb{R}$, $\|v\| = v$, $F_i(v) \equiv 0$, we conclude that $\sum_j s_{ij} x_j = y_i \leq \sum_j \mu_{ij} x_j (1 \leq i \leq m)$, for all real values x_j. This implies

$$\mu_{ij} = s_{ij} \quad (\text{for } 1 \leq i \leq m, 1 \leq j \leq l).$$

Combining this equality and (3.18), we arrive at proposition (II). □

Theorem 3.8 shows that the special bounds (3.2), (3.3), respectively, are the only bounds of type (2.6), (2.7) which make sense in the context of general sublinear

functionals $\|\cdot\|$. Accordingly, we shall focus on the following special versions of the general properties (3.15) and (3.16), respectively:

> Condition $0 < \Delta t \le \gamma \cdot \tau_0$ implies that process (1.2) satisfies the special bound (3.2), whenever $\mathbb{V}$ is a vector space, $\|\cdot\|$ a sublinear functional on $\mathbb{V}$, and the functions $F_i : \mathbb{V} \to \mathbb{V}$ satisfy the basic assumption (1.4), (3.19)

> Condition $0 < \Delta t \le \gamma \cdot \tau_0$ implies that process (1.2) satisfies the special bound (3.3), whenever $\mathbb{V}$ is a vector space, $\|\cdot\|$ a sublinear functional on $\mathbb{V}$, and the functions $F_i : \mathbb{V} \to \mathbb{V}$ satisfy the basic assumption (1.4). (3.20)

Our main result about these two properties has been formulated in Theorem 3.9. The theorem can be regarded as a neat version of Theorem 3.5, parts (I) and (III), adapted to sublinear functionals.

Theorem 3.9 (Criteria for the properties (3.19) and (3.20)) *Consider an arbitrary generic process* (1.2), *and let* $\gamma > 0$. *Then the following propositions are valid*:

(I) *Condition* (3.6) *is necessary and sufficient for property* (3.19) *as well as for property* (3.20).
(II) *If* $S \ge 0$ *has no row equal to zero, then the classical condition* (2.2) *is necessary and sufficient for property* (3.19) *as well as for property* (3.20).

Proof

(I) We prove necessity and sufficiency of (3.6) separately.
1 (Sufficiency). It is easy to see that property (3.20) implies (3.19). Therefore, it is enough to prove that condition (3.6) implies (3.20). The last implication can be proved by almost the same arguments as used in part 1 of the proof of Theorem 2.4 in Sect. 2. Note that again the inequalities $I + \gamma T \ge 0$ and $S \ge 0$ are needed, which follow now from: $I + \gamma T = (I - P)^{-1} = I + P + P^2 + \cdots \ge 0$ and $S = (I - P)^{-1} R \ge 0$.
2 (Necessity). For proving the necessity it is enough to show that (3.19) implies (3.6). To prove this implication, we (only) assume (3.19) to hold in the situation where

$$\mathbb{V} = \mathbb{R}^m, \qquad \|v\| = \max_k v^{[k]} \quad (\text{for } v \in \mathbb{V} \text{with components } v^{[k]} (1 \le k \le m)).$$

We define functions $F_j : \mathbb{V} \to \mathbb{V}$ by

$$F_j(v) = \tau_0^{-1}(-y_j + z_j) \quad (\text{for } v = y_j), \qquad F_j(v) = 0 \quad (\text{otherwise}),$$

where y_j, z_j are vectors in $\mathbb{V}$—to be specified below—satisfying

$$\|z_j\| \le \|y_j\| \quad (1 \le j \le m). \tag{3.21}$$

Clearly the functions F_j defined in this fashion satisfy the basic assumption (1.4).

We consider the matrices $P = (p_{ij})$, $R = (r_{ij})$ (cf. (2.1)) and define the components of $x_j, z_j \in \mathbb{V}$ by $x_j^{[k]} = -1$ (if $r_{kj} < 0$), $x_j^{[k]} = 0$ (otherwise), and $z_j^{[k]} = -1$ (if $p_{kj} < 0$), $z_j^{[k]} = 0$ (otherwise). We define the vectors $y_i \in \mathbb{V}$ by $y_i = \sum_{j=1}^{l} r_{ij} x_j + \sum_{j=1}^{m} p_{ij} z_j$ $(1 \le i \le m)$. A short calculation shows that x_i, y_i satisfy the relations (1.2) with the functions F_j as defined above and $\Delta t = \gamma \tau_0$.

We denote by ρ_i the sum of the absolute values of the negative entries in the i-th row of R, and by π_i the sum of the absolute values of the negative entries in the i-th row of P. By the definition of y_i, we have $\|y_i\| \ge y_i^{[i]} = \rho_i + \pi_i$ $(1 \le i \le m)$. Because $\|z_i\| \le 0$, the inequalities (3.21) are in force, so that the basic assumption (1.4) is valid.

Applying property (3.19) to the situation at hand, there follows

$$\rho_i + \pi_i \le \|y_i\| \le \left(\sum_j |s_{ij}|\right) \cdot \max_j \|x_j\| \le 0 \quad (1 \le i \le m),$$

which proves $P \ge 0$, $R \ge 0$. The remaining inequality, $\mathrm{spr}(P) < 1$, follows e.g. by applying Theorem 3.5, part (I).

(II) Let the classical condition (2.2) be fulfilled. Then (3.7) holds as well. So, by Lemma 3.2, part (III), condition (3.6) is fulfilled. From part (I) of Theorem 3.9 we conclude that (3.19) and (3.20) hold.

Conversely, assume property (3.19) or (3.20). By Theorem 3.5, part (III), we arrive at (2.2).

□

Since property (3.20) is a-priori stronger than (3.19), the essence of the above theorem is that conditions (3.6), (2.2) (under the appropriate assumptions on S) imply the strong statement (3.20), whereas already the weaker statement (3.19) implies (3.6) and (2.2) (under the same assumptions on S).

3.5 Various natural questions

In this section we ask and answer five natural questions about possible simplifications or extensions of Lemma 3.2 and Theorems 3.5, 3.9. For each of these questions we will provide counterexamples.

Question 3.10 Because all of the conditions (3.5), (3.6), (3.7) and (2.2) are more simple in appearance than condition (3.4), the question arises of whether the last condition can be replaced by any of the first four conditions in Lemma 3.2 (part (I)) or in Theorem 3.5 (part (I)).

To answer this question, consider the generic process (1.2) with $l = 2$, $m = 1$ and $s_{11} = -2$, $s_{12} = 1$, $t_{11} = 1$. Let $\gamma > 0$. It is easy to see that condition (3.4) is fulfilled. Hence, the properties (3.10) and (3.11) are present. But, we do *not* have $R \ge 0$, so that the conditions (3.5), (3.6), (3.7) and (2.2) are violated. Therefore, none of the last four conditions can replace condition (3.4) in Lemma 3.2 (part (I)) or in Theorem 3.5 (part (I)).

Question 3.11 Because the conditions (3.6), (3.7) and (2.2) are more simple than (3.5), the question arises of whether condition (3.5) can be replaced by one of the first three conditions, in Lemma 3.2 (part (II)) or in Theorem 3.5 (part (II)).

The following counterexample proves that such a replacement is *not* possible. Consider process (1.2) with $l = m = 1$ and $s_{11} = 0$, $t_{11} = -1$. Let $\gamma = 1/4$. One easily sees that condition (3.5) is fulfilled, so that the properties (3.10) and (3.11) are present. But, we do not have $P \geq 0$, so that (3.6), (3.7) and (2.2) are violated. Therefore, none of the last three conditions can replace condition (3.5) in Lemma 3.2 (part (II)) or in Theorem 3.5 (part (II)).

Question 3.12 Because the classical condition (2.2) is more simple than (3.6), the question arises of whether condition (3.6) can be replaced by (2.2) in Theorem 3.9 (part (I)).

The following counterexample proves that such replacement is *not* possible. Consider process (1.2) with $l = m = 1$ and $s_{11} = 0$, $t_{11} = -1$. Let $\gamma = 2$. One easily sees that condition (3.6) is violated, so that the properties (3.19) and (3.20) are not present. But (2.2) is fulfilled. Therefore, condition (2.2) cannot replace (3.6) in Theorem 3.9 (part (I)).

Question 3.13 One may ask whether the condition $S \geq 0$ can be omitted in Theorem 3.5 (part (III)) or in Theorem 3.9 (part (II)).

To answer this question, consider process (1.2) with $l = m = 1$ and $s_{11} = -1$, $t_{11} = -1$. Let $\gamma = 2$. It is easy to see that condition (2.2) is fulfilled, but *not* (3.4) or (3.6). Hence, all of the special boundedness properties (3.10), (3.11), (3.19) and (3.20) are not present. Therefore, the condition $S \geq 0$ cannot be omitted in Theorem 3.5 (part (III)) or in Theorem 3.9 (part (II)).

Question 3.14 Finally, we consider the question of whether the condition of S having no row equal to zero, can be omitted in Theorem 3.9 (part (II)). A negative answer to this question easily follows from the counterexample used above in resolving Question 3.12.

4 Applications of the theory

4.1 Preliminaries

Below we shall illustrate the preceding theory by applying it to some well-known numerical methods. In these applications, we will restrict ourselves, for ease of presentation, to autonomous problems, i.e. F in the initial value problem (1.1) is independent of t. Accordingly, in the generic process (1.2), we assume $F_j = F$, and the basic assumption (1.4) takes the form

$$\|v + \tau_0 F(v)\| \leq \|v\| \quad \text{(for all } v \in \mathbb{V}\text{)}. \tag{4.1}$$

In Sect. 4.2 we shall deal with the two-step ($k = 2$) Adams-Bashforth LMM and in Sect. 4.3 with a class of k-step 2-stage methods. All of these methods generate vectors

$u_n \in \mathbb{V}$ (for $n \geq k$) from starting vectors $u_0, \ldots, u_{k-1} \in \mathbb{V}$, where $u_n \approx u(n \cdot \Delta t)$ and k is fixed. We call a k-step method *bounded with factor* μ (for given stepsize Δt, vector space $\mathbb{V}$, functional $\|\cdot\|$ and function F) if

$$\|u_n\| \leq \mu \cdot \max_{0 \leq j \leq k-1} \|u_j\| \quad (k \leq n \leq k-1+N), \tag{4.2}$$

whenever $N \geq 1$ and $u_n \in \mathbb{V}$ $(k \leq n \leq k-1+N)$ are generated from any $u_0, \ldots, u_{k-1} \in \mathbb{V}$ by N successive applications of the method. Boundedness with factor $\mu = 1$ will be referred to as *monotonicity* of the method.

We recall that boundedness and monotonicity with the so-called total-variation-seminorm (defined by $\|x\| = \|x\|_{TV} = \sum_i |\xi_{i+1} - \xi_i|$ for vectors x with components ξ_i) correspond to the important concepts *total-variation-bounded* and *total-variation-diminishing*, respectively, cf. e.g. [16, 18].

In the following we shall focus on the situation where the functional $\|\cdot\|$ is a seminorm. We shall consider stepsize-coefficients $\gamma > 0$ and factors μ such that

> Condition $0 < \Delta t \leq \gamma \cdot \tau_0$ implies boundedness with factor μ, whenever $\mathbb{V}$ is a vector space with seminorm $\|\cdot\|$, and $F : \mathbb{V} \to \mathbb{V}$ satisfies the basic assumption (4.1). (4.3)

In case γ, μ satisfy (4.3), we will say that γ is a *stepsize-coefficient for boundedness of the method with factor* μ; in case γ satisfies (4.3) with $\mu = 1$, we will call it a *stepsize-coefficient for monotonicity*. Below we shall look for stepsize-coefficients γ with property (4.3) by considering representations (1.2) of N consecutive steps of the method under consideration.

4.2 The two-step Adams-Bashforth method

The well-known 2-step Adams-Bashforth method reads

$$u_n = u_{n-1} + \Delta t \left[\frac{3}{2} F(u_{n-1}) - \frac{1}{2} F(u_{n-2}) \right]; \tag{4.4}$$

it yields numerical approximations $u_n \approx u(n\Delta t)$ $(n = 2, 3, \ldots)$, starting from u_0 and $u_1 \approx u(\Delta t)$. In this section we shall look at the relevance of Theorems 2.2, 2.4, 3.5, 3.9 to this method, thereby representing N consecutive numerical steps in two different ways as a process of type (1.2).

In order to describe our first (and most natural) representation, we put $l = 2$, $m = N + 2$ and $x_1 = u_0$, $x_2 = u_1$, $y_i = u_{i-1}$ $(1 \leq i \leq m)$. Clearly, the equalities (4.4) hold for $2 \leq n \leq N + 1$ if and only if

$$\begin{aligned} y_1 &= x_1, \\ y_2 &= x_2, \\ y_i &= x_2 - \frac{1}{2}\Delta t F(y_1) + \Delta t \sum_{j=2}^{i-2} F(y_j) + \frac{3}{2}\Delta t F(y_{i-1}) \quad (3 \leq i \leq m). \end{aligned} \tag{4.5}$$

These relations are equivalent to the relations in (1.2), with coefficients s_{ij}, t_{ij} defined by:

$$(s_{ij}) = S = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix}, \qquad (t_{ij}) = T = \begin{pmatrix} 0 & & & & & \\ 0 & 0 & & & & \\ -\frac{1}{2} & \frac{3}{2} & 0 & & & \\ -\frac{1}{2} & 1 & \frac{3}{2} & 0 & & \\ \vdots & \vdots & \ddots & \ddots & \ddots & \\ -\frac{1}{2} & 1 & \cdots & 1 & \frac{3}{2} & 0 \end{pmatrix}.$$

With these definitions, the equalities (4.4) (for $2 \le n \le N+1$) thus hold if and only if (1.2) is fulfilled.

For the matrix T at hand, we see that $I + \gamma T$ is invertible for all $\gamma > 0$. Furthermore, because the preconsistency condition (1.5) is fulfilled, one might hope to be able to prove the monotonicity property (1.7) and its variant (2.4), for some $\gamma > 0$, by applying Theorems 2.2 and 2.4. If this were possible, such a γ would be a stepsize-coefficient for monotonicity in the sense specified in Sect. 4.1.

However, a short calculation shows that the matrix $P = (I + \gamma T)^{-1}(\gamma T)$ has a negative entry (for any $\gamma > 0$ and all $N \ge 1$), so that we cannot conclude, by applying the Theorems 2.2, 2.4, that there is $\gamma > 0$ for which the properties (1.7), (2.4) hold. Similarly, Theorems 3.5, 3.9 cannot be applied here so as to arrive at the boundedness property (4.3) with positive γ. The following negative statement can be proved, e.g. by applying the material in [21, Theorem 3.3]:

Proposition 4.1 *For the two-step Adams-Bashforth method* (4.4) *there exists no positive stepsize-coefficient γ for monotonicity.*

In spite of this statement, we will see below that a positive stepsize-coefficient for *boundedness* can be determined by applying Theorem 3.5 and representing the equalities (4.4) (for $2 \le n \le N+1$) in the generic form (1.2) with less obvious matrices S, T than used above.

We consider the representation in the generic form (1.2), with $l = 2, m = N$, $y_i = u_{i+1}$ $(1 \le i \le m)$ and input vectors

$$x_1 = u_1 + \frac{3}{2}\Delta t F(u_1) - \frac{1}{2}\Delta t F(u_0), \qquad x_2 = -\frac{1}{2}\Delta t F(u_1).$$

Clearly, the equalities (4.4) (for $2 \le n \le N+1$) amount to

$$\begin{aligned} y_1 &= x_1, \\ y_i &= x_1 + x_2 + \Delta t \sum_{j=1}^{i-2} F(y_j) + \frac{3}{2}\Delta t F(y_{i-1}) \quad (2 \le i \le m). \end{aligned} \tag{4.6}$$

The first N steps of the Adams-Bashforth method can thus be represented by the generic process (1.2), with $l = 2, m = N$ and

$$(s_{ij}) = S = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{pmatrix}, \qquad (t_{ij}) = T = \begin{pmatrix} 0 & & & & \\ \frac{3}{2} & 0 & & & \\ 1 & \frac{3}{2} & 0 & & \\ \vdots & \ddots & \ddots & \ddots & \\ 1 & \cdots & 1 & \frac{3}{2} & 0 \end{pmatrix}. \tag{4.7}$$

Note that this matrix S violates the preconsistency condition (1.5), so that the monotonicity theory of Sect. 2.2 is not relevant here. But, the special boundedness theory of Sect. 3 still applies.

To be able to apply Theorem 3.5, we shall determine expressions for P and R corresponding to S, T just defined. A short calculation shows that

$$(I + \gamma T)^{-1} = \begin{pmatrix} q_0 & & & \\ q_1 & q_0 & & \\ \vdots & \ddots & \ddots & \\ q_{m-1} & \cdots & q_1 & q_0 \end{pmatrix},$$

where $q_0 = 1$, $q_1 = -\frac{3}{2}\gamma$ and $q_i = (1 - \frac{3}{2}\gamma)q_{i-1} + \frac{1}{2}\gamma q_{i-2}$ for $i \geq 2$. It follows that

$$R = \begin{pmatrix} r_0 & 0 \\ r_1 & r_0 \\ \vdots & \vdots \\ r_{m-1} & r_{m-2} \end{pmatrix}, \qquad P = -\begin{pmatrix} 0 & & & \\ q_1 & 0 & & \\ \vdots & \ddots & \ddots & \\ q_{m-1} & \cdots & q_1 & 0 \end{pmatrix},$$

where $r_i = q_0 + q_1 + \cdots + q_i$. Using the recurrence relation satisfied by q_i, one finds for $0 < \gamma \leq \frac{4}{9}$ and $i \geq 1$ that $q_i \leq 0$ and $\gamma \cdot r_i = -[(1 - \gamma)q_i + \frac{\gamma}{2}q_{i-1}] \geq 0$. Hence, the classical condition (2.2) is fulfilled for any $\gamma \in (0, \frac{4}{9}]$. In the rest of this section we assume $\gamma = \frac{4}{9}$.

From proposition (III) of Theorem 3.5, we conclude that the generic process (1.2) (with coefficients given by (4.7)) has the special boundedness property (3.11). Using this property and the definition of x_1, x_2 in force, it follows that condition $0 < \Delta t \leq \gamma \cdot \tau_0$ implies:

$$\|u_n\| \leq \left\| u_1 + \frac{3}{2}\Delta t F(u_1) - \frac{1}{2}\Delta t F(u_0) \right\| + \left\| -\frac{1}{2}\Delta t F(u_1) \right\| \tag{4.8}$$

for $2 \leq n \leq N + 1$, whenever u_n is generated by applying the Adams-Bashforth method under the basic assumption (4.1). Here $\|\cdot\|$ stands for an arbitrary seminorm on the vector space $\mathbb{V}$.

For $0 < \Delta t \leq \gamma \cdot \tau_0$ and any seminorm $\|\cdot\|$, we have

$$\|\Delta t F(v)\| = (\Delta t/\tau_0)\| - v + (v + \tau_0 F(v))\| \leq 2\gamma\|v\|,$$

which can be seen to imply

$$\left\| u_1 + \frac{3}{2}\Delta t F(u_1) - \frac{1}{2}\Delta t F(u_0) \right\| \le \|u_1\| + \gamma \|u_0\|;$$

hence,

$$\left\| u_1 + \frac{3}{2}\Delta t F(u_1) - \frac{1}{2}\Delta t F(u_0) \right\| + \left\| -\frac{1}{2}\Delta t F(u_1) \right\| \le (1+\gamma)\|u_1\| + \gamma \|u_0\|. \quad (4.9)$$

Combining this inequality with the above bound for $\|u_n\|$, we arrive at the following:

Proposition 4.2 *For the two-step Adams-Bashforth method* (4.4), *the stepsize condition* $0 < \Delta t \le \frac{4}{9}\tau_0$ *implies boundedness with factor* $\mu = 17/9$, *whenever* $\mathbb{V}$ *is a vector space with seminorm* $\|\cdot\|$, *and* $F : \mathbb{V} \to \mathbb{V}$ *satisfies the basic assumption* (4.1).

By applying Theorem 3.9, instead of Theorem 3.5, we find similarly as above that the bound (4.8) is valid under the basic assumption (4.1), when $\|\cdot\|$ is an arbitrary *sublinear functional* on the vector space $\mathbb{V}$. But, in the general situation of sublinear functionals, we *cannot* derive similarly as above that the inequality (4.9) is valid.

To give a simple illustration of the estimate (4.8), with a sublinear functional $\|\cdot\|$ which is no seminorm, we consider $\mathbb{V} = \mathbb{R}^M$ with functional $\|\cdot\| = \|\cdot\|_0$ (given by (3.13)). Applying Theorem 3.9 to the situation at hand, and defining $v \ge 0$ by non-negativity of all components of $v \in \mathbb{V}$, yields:

Proposition 4.3 *Consider the two-step Adams-Bashforth method* (4.4) *in the situation where* $\mathbb{V} = \mathbb{R}^M$ *and* $\|\cdot\| = \|\cdot\|_0$ (*see* (3.13)). *Assume* $F : \mathbb{V} \to \mathbb{V}$ *satisfies the basic assumption* (4.1). *Then the stepsize condition* $0 < \Delta t \le \frac{4}{9}\tau_0$ *implies that*

$$u_n \ge 0 \quad (2 \le n \le N+1),$$

whenever u_n *is obtained from* u_0, u_1 *with* $u_1 + \frac{3}{2}\Delta t F(u_1) \ge \frac{1}{2}\Delta t F(u_0)$ *and* $F(u_1) \le 0$.

We note that there exists *no* positive stepsize-coefficient γ, such that the inequalities $u_n \ge 0$ are valid for $0 < \Delta t \le \gamma \cdot \tau_0$, under the more natural assumption that

$$u_0 \ge 0, \qquad u_1 \ge 0 \quad \text{and} \quad v + \tau_0 F(v) \ge 0 \quad (\text{for all } v \in \mathbb{R}^M \text{ with } v \ge 0).$$

This can be seen, for example, by considering $\mathbb{V} = \mathbb{R}$, $F(v) \equiv v$ and $u_0 = 1$, $u_1 = 0$.

4.3 Predictor-corrector methods and hybrid multistep methods

4.3.1 Notations

Using an explicit linear multistep method (LMM), with coefficients $\hat{a}_j$, $\hat{b}_j$, as a predictor for an implicit LMM, with coefficients a_j, b_j, results in a numerical process

of type

$$v_n = \sum_{j=1}^{k} \hat{a}_j u_{n-j} + \Delta t \sum_{j=1}^{k} \hat{b}_j F(u_{n-j}), \tag{4.10a}$$

$$u_n = \sum_{j=1}^{k} a_j u_{n-j} + \Delta t \sum_{j=1}^{k} b_j F(u_{n-j}) + \Delta t b_0 F(v_n), \tag{4.10b}$$

where $k \geq 1$ is fixed and $n = k, k+1, \ldots$, cf. e.g. [3, 8, 12]. The starting values for this method are $u_0, u_1, \ldots, u_{k-1} \in \mathbb{V}$.

Throughout this section we assume $b_0 > 0$, $\sum_{j=1}^{k} \hat{a}_j = 1$, $\sum_{j=1}^{k} a_j = 1$, as well as zero-stability, i.e. all roots of the equation $\xi^k = \sum_{j=1}^{k} a_j \xi^{k-j}$ have a modulus $|\xi| \leq 1$, and the roots with $|\xi| = 1$ are simple.

Methods of type (4.10) are called predictor-corrector methods if u_n and v_n, respectively, are final and tentative approximations to the solution at $t_n = n\Delta t$. If a predictor (4.10a) corresponds to a method with order of accuracy k, and a corrector (4.10b) to a method with order $k+1$, then the predictor-corrector method (4.10) has order $k+1$. The most popular schemes of this type are obtained by combining the explicit Adams-Bashforth and implicit Adams-Moulton methods, cf. the literature mentioned above.

The formulas (4.10) can also stand for so-called hybrid multistep methods, also known as modified linear multistep methods, where v_n approximates the solution at a point $\bar{t}_n = (n - \kappa)\Delta t$, with an extra parameter $\kappa \neq 0$; cf. the above literature.

We shall represent $N \geq 1$ steps of the general method (4.10) as a process of type (1.2), were $y = [y_i] \in \mathbb{V}^m$, $m = 2N$, with

$$y_i = u_{k-1+i}, \qquad y_{N+i} = v_{k-1+i} \quad \text{for } 1 \leq i \leq N. \tag{4.11}$$

For the input vector we take $x = [x_j] \in \mathbb{V}^l$, $l = 2k$, defined by

$$x_i = \sum_{j=i}^{k} a_j u_{k-1+i-j} + \Delta t \sum_{j=i}^{k} b_j F(u_{k-1+i-j}) \quad (1 \leq i \leq k), \tag{4.12a}$$

$$x_{i+k} = \sum_{j=i}^{k} \hat{a}_j u_{k-1+i-j} + \Delta t \sum_{j=i}^{k} \hat{b}_j F(u_{k-1+i-j}) \quad (1 \leq i \leq k). \tag{4.12b}$$

To write the relations (4.10), (4.11) specifying $y_1, y_2, \ldots, y_m$ in a compact way, we give the following definitions. For any $m \times r$ matrix $S = (s_{ij})$ we denote by the boldface symbol $\boldsymbol{S}$ the corresponding linear map from $\mathbb{V}^r$ to $\mathbb{V}^m$, that is, $y = \boldsymbol{S}x$ if $y_i = \sum_{j=1}^{r} s_{ij} x_j \in \mathbb{V}$ $(1 \leq i \leq m)$. Let I be the $N \times N$ identity matrix. Let $J_0 \in \mathbb{R}^{N \times k}$ be the matrix that consists of either the first N rows of the $k \times k$ identity matrix (when $1 \leq N < k$), or the first k columns of I (when $N \geq k$). Furthermore, let $A_0 \in \mathbb{R}^{N \times N}$ be the lower triangular Toeplitz matrix with diagonal entries 0, entries a_j on the j-th lower diagonal $(1 \leq j \leq \min\{k, N-1\})$ and with the remaining entries 0 again. The matrices B_0, $\hat{A}_0$, $\hat{B}_0 \in \mathbb{R}^{N \times N}$ are defined likewise with coefficients $b_j, \hat{a}_j, \hat{b}_j$ $(1 \leq$

$j \le \min\{k, N-1\}$), respectively (the coefficient b_0 does not enter into the matrix B_0).

It is easy to see that the relations (4.10) (for $k \le n \le k-1+N$) are equivalent to

$$y = \boldsymbol{J}x + \boldsymbol{A}y + \Delta t \boldsymbol{B}\boldsymbol{F}(y), \tag{4.13}$$

where $\boldsymbol{F}(y) = [F(y_j)] \in \mathbb{V}^m$, and $J \in \mathbb{R}^{m\times l}$, $A, B \in \mathbb{R}^{m\times m}$ are given by

$$J = \begin{pmatrix} J_0 & 0 \\ 0 & J_0 \end{pmatrix}, \quad A = \begin{pmatrix} A_0 & 0 \\ \hat{A}_0 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} B_0 & b_0 I \\ \hat{B}_0 & 0 \end{pmatrix}. \tag{4.14}$$

The generic form (1.2) is thus obtained with coefficient matrices $(s_{ij}) = S = (I - A)^{-1}J$ and $(t_{ij}) = T = (I - A)^{-1}B$.

4.3.2 Monotonicity for predictor-corrector methods and hybrid multistep methods

Let us first take a brief look at standard monotonicity with respect to the starting vectors $u_0, \ldots, u_{k-1}$. For this, it is convenient to introduce $\check{a}_j = a_j - \gamma b_0 \hat{a}_j$ and $\check{b}_j = b_j - \gamma b_0 \hat{b}_j$ (for $j = 1, \ldots, k$). The relations (4.10) imply that

$$u_n = \sum_{j=1}^{k} \check{a}_j u_{n-j} + \Delta t \sum_{j=1}^{k} \check{b}_j F(u_{n-j}) + \gamma b_0 \left(v_n + \frac{\Delta t}{\gamma} F(v_n) \right).$$

By combining this equality with (4.10a), we arrive at the following theorem; see also e.g. [6, 12, 22].

Theorem 4.4 *Consider method (4.10) with $n = k, k+1, \ldots, k-1+N$. Let $\|\cdot\|$ be a convex functional on the vector space $\mathbb{V}$, and assume $F : \mathbb{V} \to \mathbb{V}$ satisfies the basic assumption (4.1). Let $\gamma > 0$ be such that*

$$\hat{a}_j \ge \gamma \hat{b}_j \ge 0, \qquad \check{a}_j \ge \gamma \check{b}_j \ge 0 \quad (j = 1, \ldots, k). \tag{4.15}$$

Then the stepsize restriction $0 < \Delta t \le \gamma \cdot \tau_0$ implies that

$$\|u_n\| \le \max_{0\le j\le k-1} \|u_j\| \quad (k \le n \le k-1+N). \tag{4.16}$$

Note that, under a weak irreducibility assumption, condition (4.15) is not only sufficient but also necessary for the above bound (4.16), see [22].

However, the methods (4.10) with coefficients satisfying condition (4.15) (with $\gamma > 0$) form a small class, excluding popular schemes, for instance obtained by combining explicit and implicit Adams-type methods as indicated above. Furthermore, in view of results for LMMs of [19], one can expect that the stepsize requirement $\Delta t \le \gamma \cdot \tau_0$ (with γ such that (4.15) holds) may be unnecessarily restrictive if γ is only required to be a stepsize-coefficient for boundedness (in the sense of Sect. 4.1).

Below we apply the theory of Sect. 3 in an analysis of the methods (4.10) which is also relevant in cases where condition (4.15) is violated.

4.3.3 Special bounds for predictor-corrector methods and hybrid multistep methods

Below we shall look for stepsize-coefficients for boundedness using the representation of (4.10) in the generic form (1.2) with the matrices S, T specified in Sect. 4.3.1.

For this T, the matrix $I+\gamma T$ is invertible for all $\gamma > 0$. To prove this, we consider the alternative ordering

$$y_{2i-1} = v_{k-1+i}, \qquad y_{2i} = u_{k-1+i} \quad (1 \le i \le N), \tag{4.17}$$

which yields a representation of type (4.13) with strictly lower triangular matrices, say, $\underline{A}, \underline{B}$. The corresponding matrix $\underline{T} = (I-\underline{A})^{-1}\underline{B}$ is also strictly lower triangular. With our original ordering, viz. (4.11), we thus have a matrix $T = V\underline{T}V^{-1}$, where V is a permutation matrix, and therefore $I+\gamma T$ is invertible. To derive boundedness results it will be convenient to use the original ordering (4.11).

Substituting the expressions for S and T (given at the end of Sect. 4.3.1) into the definition (2.1) of P and R, we arrive at

$$R = KJ, \qquad P = \gamma KB, \qquad K = (I - A + \gamma B)^{-1}. \tag{4.18}$$

Because $P = V\underline{P}V^{-1}$, with $\underline{P} = \gamma\underline{T}(I+\gamma\underline{T})^{-1}$ and $\mathrm{spr}(\underline{P}) = 0$, we have also $\mathrm{spr}(P) = 0$.

Let $\check{K}_0 = (I - \check{A}_0 + \gamma\check{B}_0)^{-1}$, $\check{A}_0 = A_0 - \gamma b_0\hat{A}_0$, $\check{B}_0 = B_0 - \gamma b_0\hat{B}_0$. It can be seen that

$$K = \begin{pmatrix} I - A_0 + \gamma B_0 & \gamma b_0 I \\ -\hat{A}_0 + \gamma\hat{B}_0 & I \end{pmatrix}^{-1} = \begin{pmatrix} \check{K}_0 & -\gamma b_0\check{K}_0 \\ (\hat{A}_0 - \gamma\hat{B}_0)\check{K}_0 & (I - A_0 + \gamma B_0)\check{K}_0 \end{pmatrix}.$$

This gives

$$R = \begin{pmatrix} \check{K}_0 J_0 & -\gamma b_0\check{K}_0 J_0 \\ (\hat{A}_0 - \gamma\hat{B}_0)\check{K}_0 J_0 & (I - A_0 + \gamma B_0)\check{K}_0 J_0 \end{pmatrix}. \tag{4.19}$$

Using the fact that lower triangular Toeplitz matrices commute, it is found that

$$P = \gamma \begin{pmatrix} (B_0 - \gamma b_0\hat{B}_0)\check{K}_0 & b_0\check{K}_0 \\ \big((I - A_0)\hat{B}_0 + \hat{A}_0 B_0\big)\check{K}_0 & b_0(\hat{A}_0 - \gamma\hat{B}_0)\check{K}_0 \end{pmatrix}. \tag{4.20}$$

We have

$$S = \begin{pmatrix} (I - A_0)^{-1}J_0 & O \\ \hat{A}_0(I - A_0)^{-1}J_0 & J_0 \end{pmatrix}.$$

By considering the upper-right blocks of R, P, S and $PS, |P|S$ it can be seen that none of conditions (3.4)–(3.7) is fulfilled (for any $\gamma > 0$ and all $N \ge 1$). Hence, Theorem 3.5 cannot be applied here directly so as to arrive at property (4.3) with positive γ. However, we shall see below that a positive *stepsize-coefficient for boundedness* can be found by modifying the matrix S and applying Theorem 3.9.

Let

$$\tilde{x}_i = x_i - \gamma b_0 x_{i+k}, \qquad \tilde{x}_{i+k} = x_{i+k} \quad \text{for } i = 1, \ldots, k. \tag{4.21}$$

Then $x = V\tilde{x}$ with $V = \left(\begin{smallmatrix} I & \gamma b_0 I \\ 0 & I \end{smallmatrix}\right)$. Below we shall deal with process (4.13) written in the equivalent form

$$y = \tilde{\boldsymbol{S}}\tilde{x} + \Delta t \boldsymbol{T}\boldsymbol{F}(y), \tag{4.22}$$

where $\tilde{S} = (\tilde{s}_{ij}) = (I - A)^{-1} JV = SV$. Defining $\tilde{R} = (I + \gamma T)^{-1}\tilde{S}$ (cf. (2.1)) we get in view of (4.18)

$$\tilde{R} = KJV = \begin{pmatrix} \check{K}_0 J_0 & 0 \\ (\hat{A}_0 - \gamma \hat{B}_0)\check{K}_0 J_0 & J_0 \end{pmatrix}. \tag{4.23}$$

We now have $\tilde{R} \geq 0$ (for all $N \geq 1$) whenever

$$P \geq 0 \quad \text{(for all } N \geq 1\text{)}. \tag{4.24}$$

This leads directly to the following result.

Lemma 4.5 *Consider N consecutive steps of method (4.10) written in the form (4.22). Let $\|\cdot\|$ be a sublinear functional on the vector space $\mathbb{V}$. Assume $F : \mathbb{V} \to \mathbb{V}$ satisfies the basic assumption (4.1) and $\gamma > 0$ is such that (4.24) holds. Then the step-size restriction $0 < \Delta t \leq \gamma \cdot \tau_0$ implies that the output vectors y_i defined by (4.11) satisfy*

$$\|y_i\| \leq \tilde{\mu}_i \cdot \max_{1 \leq j \leq l} \|\tilde{x}_j\| \quad (1 \leq i \leq 2N),$$

with $\tilde{\mu}_i = \sum_j |\tilde{s}_{ij}|$.

Proof To prove this lemma, we apply part (I) of Theorem 3.9 with S replaced by $\tilde{S}$. □

Consider $\tilde{\mu} = \max_i \tilde{\mu}_i = \|\tilde{S}\|_\infty$. Using the definition (4.14) and the expression (4.23), there follows after a little calculation that

$$\tilde{S} = \begin{pmatrix} I & \gamma b_0 I \\ \hat{A}_0 & I - \check{A}_0 \end{pmatrix} \begin{pmatrix} S_0 & 0 \\ 0 & S_0 \end{pmatrix},$$

with $S_0 = (I - A_0)^{-1} J_0$. We find that $\tilde{\mu} \leq \|(I - A_0)^{-1} J_0\|_\infty \cdot \max\{1 + \gamma b_0, 1 + \sum_{j=1}^k (|\hat{a}_j| + |\check{a}_j|)\}$. Due to the assumption of zero-stability we have $\sup_{N \geq 1} \|S_0\|_\infty < \infty$, so that $\tilde{\mu}$ can be bounded, uniformly with respect to N.

Consider $\gamma > 0$ such that (4.24) holds and let $0 < \Delta t \leq \gamma \cdot \tau_0$. Then from Lemma 4.5 and (4.12), (4.21), it follows that

$$\|u_n\| \leq \tilde{\mu} \cdot \max\left\{ \sum_{j=1}^k \left(|\check{a}_j - \gamma \check{b}_j| + |\gamma \check{b}_j|\right), \sum_{j=1}^k \left(|\hat{a}_j - \gamma \hat{b}_j| + |\gamma \hat{b}_j|\right) \right\} \cdot \max_{0 \leq j \leq k-1} \|u_j\|$$

for $k \leq n \leq k - 1 + N$, whenever u_n is generated from $u_0, \ldots, u_{k-1} \in \mathbb{V}$ by applying method (4.10) under the basic assumption (4.1), where $\|\cdot\|$ is a seminorm on the vector space $\mathbb{V}$. Thus we arrive at the following theorem.

Theorem 4.6 *Assume $\gamma > 0$ is such that $P \geq 0$ (for all $N \geq 1$). Then γ is a stepsize-coefficient for boundedness of the method (4.10) (in the sense of Sect. 4.1).*

4.3.4 Results for third order explicit two-step methods of the form (4.10)

In this section we study method (4.10) with $k = 2$, $u_n \approx u(n\Delta t)$, $v_n \approx u((n-\kappa)\Delta t)$. Requiring order $p = 3$ leaves 3 free parameters $a_1, \hat{a}_1, \kappa$ and the remaining coefficients can be computed by the formulas: $a_2 = 1 - a_1$, $b_0 = (4 + a_1)/(6(1-\kappa)(2-\kappa))$, $b_1 = (8 - 12\kappa - (4 - 3\kappa)a_1)/(6(1-\kappa))$, $b_2 = (4 - (5 - 3\kappa)a_1)/(6(2-\kappa))$, $\hat{a}_2 = 1 - \hat{a}_1$, $\hat{b}_1 = 2 - \frac{\hat{a}_1}{2} - 2\kappa + \frac{\kappa^2}{2}$, $\hat{b}_2 = -\frac{\hat{a}_1}{2} + \kappa - \frac{\kappa^2}{2}$. The method is zero-stable if and only if $a_1 \in [0, 2)$.

For these methods we will compute the maximal values of γ such that $P \geq 0$ for $N = 1, \ldots, 1000$; it was verified that with larger N the results did not differ anymore noticeably.

First we study the methods with $\kappa = 0$, corresponding to the classical two-step predictor-corrector methods. The result is shown in the left panel of Fig. 1. We note that there are no methods in this class for which the monotonicity condition (4.15) holds with $\gamma > 0$. The displayed values of γ for boundedness with these predictor-corrector methods are rather low; the maximal value is approximately 0.36, corresponding to $a_1 \approx 0.765$, $\hat{a}_1 \approx 1.673$.

A numerical search revealed that larger values of γ can be found by allowing $\kappa \neq 0$. The right panel of Fig. 1 shows the values of γ with $\kappa = 1 - \frac{1}{3}\sqrt{3}$. The largest $\gamma \approx 0.73$ is found with $a_1 \approx 0.392$, $\hat{a}_1 \approx 0.667$ and this γ is optimal within the whole class (4.10) with $k = 2$, $p = 3$.

Rather surprisingly, this method coincides with the method found in [22, Sect. 3.2.3] which is optimal with respect to the monotonicity condition (4.15). The latter method corresponds to $a_1 = 6\sqrt{3} - 10$, $\hat{a}_1 = \frac{2}{3}$. These parameters coincide (up to four decimal digits) with the values for $a_1, \hat{a}_1$ obtained numerically by our search using condition (4.24), corresponding to the right panel in Fig. 1. In fact, if $\hat{a}_1 \leq \frac{2}{3}$ the monotonicity condition (4.15) seems to give the same γ as the boundedness condition (4.24). If $\hat{a}_1 > \frac{2}{3}$ then the method has some negative coefficient, so then there

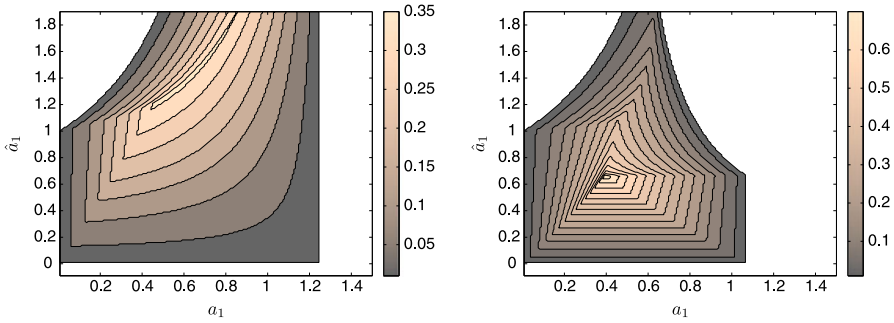

Fig. 1 Maximal values $\gamma > 0$ such that $P \geq 0$ for the methods (4.10) with $k = 2$ of order $p = 3$, with parameters $a_1 \in [0, 1.5]$ horizontally and $\hat{a}_1 \in [-0.1, 1.95]$ vertically. *Left panel*: standard predictor-corrector methods, $\kappa = 0$. *Right panel*: hybrid methods with $\kappa = 1 - \frac{1}{3}\sqrt{3}$. Contour levels at $j/20$, $j = 0, 1, \ldots$; for the 'white' areas, there is no positive γ

is no positive γ for monotonicity with arbitrary starting values. But, as shown by Fig. 1, for such $\hat{a}_1$ we can still have positive stepsize-coefficients γ for boundedness.

Acknowledgements We thank the referee for comments which have resulted in an improved presentation of our work. The work of A. Mozartova is supported by a grant from the Netherlands Organisation for Scientific Research NWO. The work of W. Hundsdorfer for this publication was partially supported by Award No. FIC/2010/05 from King Abdullah University of Science and Technology (KAUST).

References

1. Butcher, J.C.: On the convergence of numerical solutions to ordinary differential equations. Math. Comput. **20**, 1–10 (1966)
2. Butcher, J.C.: The Numerical Analysis of Ordinary Differential Equations. Wiley, Chichester (1987)
3. Butcher, J.C.: Numerical Methods for Ordinary Differential Equations. Wiley, Chichester (2003)
4. Ferracina, L., Spijker, M.N.: Stepsize restrictions for the total-variation-diminishing property in general Runge-Kutta methods. SIAM J. Numer. Anal. **42**, 1073–1093 (2004)
5. Gottlieb, S., Ketcheson, D.I, Shu, C.-W.: High order strong stability preserving time discretizations. J. Sci. Comput. **38**, 251–289 (2009)
6. Gottlieb, S., Shu, C.-W., Tadmor, E.: Strong stability-preserving high-order time discretization methods. SIAM Rev. **43**, 89–112 (2001)
7. Hairer, E., Wanner, G.: Solving Ordinary Differential Equations. II. Stiff and Differential-Algebraic Problems. Springer, Berlin (1996)
8. Hairer, E., Nørsett, S.P., Wanner, G.: Solving Ordinary Differential Equations. I. Nonstiff Problems. Springer, Berlin (1987)
9. Higueras, I.: On strong stability preserving time discretization methods. J. Sci. Comput. **21**, 193–223 (2004)
10. Higueras, I.: Representations of Runge-Kutta methods and strong stability preserving methods. SIAM J. Numer. Anal. **43**, 924–948 (2005)
11. Horn, R.A., Johnson, C.R.: Matrix Analysis. Cambridge University Press, Cambridge (1988)
12. Huang, C.: Strong stability preserving hybrid methods. Appl. Numer. Methods **59**, 891–904 (2009)
13. Hundsdorfer, W., Ruuth, S.J.: Monotonicity for time discretizations. In: Griffiths, D.F., Watson, G.A. (eds.) Procs. Dundee Conference 2003, pp. 85–94 (2003). Report NA/217, Univ. Dundee
14. Hundsdorfer, W., Ruuth, S.J.: On monotonicity and boundedness properties of linear multistep methods. Math. Comput. **75**, 655–672 (2006)
15. Hundsdorfer, W., Ruuth, S.J., Spiteri, R.J.: Monotonicity-preserving linear multistep methods. SIAM J. Numer. Anal. **41**, 605–623 (2003)
16. Hundsdorfer, W., Verwer, J.G.: Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations. Springer Ser. Comp. Math., vol. 33. Springer, Berlin (2003)
17. Hundsdorfer, W., Mozartova, A.S., Spijker, M.N.: Stepsize conditions for boundedness in numerical initial value problems. SIAM J. Numer. Anal. **47**, 3797–3819 (2009)
18. LeVeque, R.J.: Finite Volume Methods for Hyperbolic Problems. Cambridge University Press, Cambridge (2002)
19. Ruuth, S.J., Hundsdorfer, W.: High-order linear multistep methods with general monotonicity and boundedness properties. J. Comput. Phys. **209**, 226–248 (2005)
20. Shu, C.-W., Osher, S.: Efficient implementation of essentially nonoscillatory shock-capturing schemes. J. Comput. Phys. **77**, 439–471 (1988)
21. Spijker, M.N.: Contractivity in the numerical solution of initial value problems. Numer. Math. **42**, 271–290 (1983)
22. Spijker, M.N.: Stepsize conditions for general monotonicity in numerical initial value problems. SIAM J. Numer. Anal. **45**, 1226–1245 (2007)