Stepsize conditions for boundedness in numerical initial value problems

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

2009, January 7, Report MI 2009-01, Math. Inst., Leiden Univ.

Abstract. For Runge-Kutta methods (RKMs), linear multistep methods (LMMs) and classes of general linear methods (GLMs) much attention has been paid, in the literature, to special nonlinear stability requirements indicated by the terms total-variation-diminishing (TVD), strong stability preserving (SSP) and monotonicity. Stepsize conditions, guaranteeing these properties, were derived by Shu & Osher [J. Comput. Phys., 77 (1988) pp. 439-471] and in numerous subsequent papers. These special stability requirements imply essential boundedness properties for the numerical methods, among which the property of being total-variation-bounded. Unfortunately, for many well-known methods, the above special requirements are violated, so that one cannot conclude in this way that the methods are (total-variation-)bounded.

In this paper, we focus on stepsize conditions for boundedness directly - rather than via the detour of the above special stability properties. We present a generic framework for deriving best possible stepsize conditions which guarantee boundedness of actual RKMs, LMMs and GLMs - whether or not the methods under consideration have the special stability properties mentioned above.

Key words. 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.

AMS subject classifications. 65L05, 65L06, 65L20, 65M20.

1 Introduction

1.1 Monotonicity and boundedness

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

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

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

Monotonicity of Runge-Kutta methods

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

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

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

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

*CWI, P.O Box 94079, NL-1090-GB Amsterdam, Nederland. Email: Willem.Hundsdorfer@cwi.nl

†CWI, P.O Box 94079, NL-1090-GB Amsterdam, Nederland. Email: A.Mozartova@cwi.nl

‡Math. Inst., Leiden Univ., P.O. Box 9512, NL-2300-RA Leiden, Nederland. Email: spijker@math.leidenuniv.nl

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

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

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

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

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

Satisfying (1.4) is of crucial importance for suitable convergence properties when $\Delta t \to 0$, and constitutes one of the underlying reasons why attention has been paid in the literature to (1.3), cf. e.g. LeVeque (2002), Hundsdorfer & Verwer (2003).

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

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

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

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

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

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

Monotonicity of linear multistep methods

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

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

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

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

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

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

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

Boundedness

Unfortunately, there are well known RKMs and LMMs, with a record of practical success, for which there exist *no positive stepsize-coefficients* γ such that (1.5), (1.6) always imply (1.3) or (1.8), respectively - among which the Adams methods and BDFs with $k \geq 2$ as well as the Dormand-Prince formula, cf. e.g. Hairer, Nørsett & Wanner (1987). Moreover, no second order (implicit) RKMs or LMMs exist with $\gamma = \infty$, see e.g. Spijker (1983, Sections 2.2, 3.2). These circumstances suggest that there are situations where monotonicity may be too strong a theoretical demand, and that it is worthwhile to study, along with monotonicity, also directly the following weaker *boundedness properties* for methods (1.2) and (1.7), respectively:

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

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

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

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

1.2 Scope of the paper

Boundedness of general linear methods

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

In this paper, we shall consider, for GLMs, boundedness properties, similar to (1.10), (1.11). A generic framework will be presented which facilitates the computation of stepsize-coefficients γ related to such properties. Besides being helpful in finding stepsize conditions that are *sufficient* for boundedness, the framework leads to *necessary* conditions as well.

The theory in the present paper can be viewed as a (nontrivial) extension of an approach to monotonicity of GLMs given earlier in the literature, cf. Spijker (2007). Its usefulness will be illustrated briefly in the present paper, whereas in future work the theory will be applied in a more general analysis for classes of GLMs, cf. Hundsdorfer, Mozartova & Spijker (2009a), (2009b).

Organization of the paper

Section 2 deals with stepsize-coefficients γ related to explicit bounds for the output vectors of a generic numerical process. Our main theorems, Theorems 2.2 and 2.4, provide an algebraic criterion in terms of γ, viz. (2.12), for these bounds to be valid in situations of practical relevance.

In Section 3, we give results related to Theorems 2.2 and 2.4. In Section 3.1, we apply the theorems so as to obtain simplified conditions for bounding the generic process. We also recover easily a concise criterion for monotonicity obtained earlier in the literature (but derived differently), cf. Spijker (2007). In Section 3.2, a lemma is presented which is helpful when applying the main theorems in the boundedness analysis of actual GLMs. In Section 3.3, we illustrate the significance of the general theory shortly, by applying it in resolving the question of boundedness for some concrete numerical methods.

In Section 4 we give the proofs of Theorems 2.2, 2.4.

2 Bounds for a generic numerical process

In this section, we shall study bounds for the output vectors of a generic numerical process. We are interested in these bounds, primarily because they facilitate significantly the derivation of actual boundedness results for given GLMs. In Section 2.1 we first describe GLMs, whereas in Section 2.2 we introduce the generic numerical process and relate it to GLMs. In the Sections 2.3 and 2.4 we present criteria for the existence of the above mentioned bounds for the generic process.

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

2.1 General linear methods

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

$$v_i^{[n]} = \sum_{j=1}^{l} \alpha_{ij}\, u_j^{[n-1]} + \Delta t \cdot \sum_{j=1}^{q} \beta_{ij}\, F((n-1+c_j)\,\Delta t,\, v_j^{[n]}) \quad (1 \le i \le q), \tag{2.1.a}$$

$$u_i^{[n]} = v_{q-l+i}^{[n]} \quad (1 \le i \le l). \tag{2.1.b}$$

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

Obviously, the Runge-Kutta method (1.2) is an example of (2.1), with $l = 1,\ q = s+1,\ u_1^{[n]} = u_n \simeq u(n\cdot\Delta t)$ and $\alpha_{i1} = 1,\ \beta_{ij} = a_{ij}$ (for $1 \le j \le s$), $\beta_{ij} = 0$ (for $j = s+1$).

The linear multistep method (1.7) is another example of (2.1), with $l = k,\ q = k+1$ and $u_i^{[n]} = u_{n-1+i}$ $(1 \le i \le k,\ n \ge 0)$, $v_i^{[n]} = u_{n-2+i}$ $(1 \le i \le k+1,\ n \ge 1)$. Method (1.7) can be written in the form (2.1) with $c_j = j-1,\ A = \binom{I}{a},\ B = \binom{O}{b}$, where I denotes the $k\times k$ identity matrix, O the $k\times(k+1)$ zero matrix and $a = (a_k, \dots, a_1),\ b = (b_k, \dots, b_0)$.

For completeness, we note that GLMs are often represented differently from (2.1), viz. in a partitioned form with parameters $u_{ij}, v_{ij}, a_{ij}, b_{ij}, c_j$, as follows:

$$Y_i = \sum_{j=1}^{l} u_{ij}\, y_j^{[n-1]} + \Delta t \cdot \sum_{j=1}^{s} a_{ij}\, F((n-1+c_j)\,\Delta t,\, Y_j) \quad (1 \le i \le s), \tag{2.2.a}$$

$$y_i^{[n]} = \sum_{j=1}^{l} v_{ij}\, y_j^{[n-1]} + \Delta t \cdot \sum_{j=1}^{s} b_{ij}\, F((n-1+c_j)\,\Delta t,\, Y_j) \quad (1 \le i \le l), \tag{2.2.b}$$

see e.g. Hairer & Wanner (1991, p. 313), Butcher (2003, p. 358). Here s is the number of internal approximations Y_i, and l is again the number of vectors $y_i^{[n]}$ which propagate from step to step. Clearly, (2.2) is formally of type (2.1) with $q = l+s$ and $u_i^{[n]},\ v_i^{[n]}$ defined, with obvious vector notations, by $u^{[n]} = y^{[n]}, \quad v^{[n]} = \binom{Y}{y^{[n]}}$. In this paper, we aim at bounding simultaneously Y and $y^{[n]}$, in terms of $y^{[0]}$, so that we find it convenient to use a representation of the GLM in which Y and $y^{[n]}$ are lumped together. In the following, we shall thus deal with representation (2.1) rather than (2.2).

Definition 2.1. (Boundedness of general linear methods)
We define method (2.1) to be bounded, with constant μ (for given stepsize Δt, vector space $\mathbb{V}$, seminorm $\|.\|$ and function F), if

$$\|v_i^{[n]}\| \le \mu \cdot \max_{1\le j\le l} \|u_j^{[0]}\| \quad (\textit{for all } n \ge 1 \textit{ and } 1 \le i \le q), \tag{2.3}$$

whenever $u_i^{[n-1]},\ u_i^{[n]},\ v_i^{[n]} \in \mathbb{V}$ *satisfy (2.1) (for all* $n \geq 1$*).*

Note that (2.3) implies (1.10) or (1.11), respectively, if method (2.1) stands for a RKM or LMM in the way indicated above.

2.2 A generic numerical process, with a simple form

For studying boundedness of (2.1), it is convenient to represent in a concise form all relations, involved in specifying $v_i^{[N]}$ (for any given $N \geq 1$). We describe now a standard representation of N consecutive steps of the GLM, to which we will refer in the following as the *canonical representation*. We combine all vectors $v_i^{[n]}$ (with $1 \leq i \leq q$ and $1 \leq n \leq N$) into one single vector $y = [y_i] \in \mathbb{V}^m$, where $m = N \cdot q$, and $y_i \in \mathbb{V}$ $(1 \leq i \leq m)$. Furthermore, we introduce shorthand notations for $u_i^{[0]}$ and $F((n-1+c_i)\Delta t, v)$. Defining, for $1 \leq i \leq l$ and $1 \leq j \leq q$,

$$x_i = u_i^{[0]},\ \ y_{(n-1)q+j} = v_j^{[n]},\ \ F_{(n-1)q+j}(v) = F((n-1+c_j)\Delta t, v), \tag{2.4}$$

we can rewrite the relations (2.1) (for $n = 1, \ldots N$) in the following form:

$$y_i = \sum_{j=1}^{l} s_{ij}\ x_j\ +\ \Delta t \cdot \sum_{j=1}^{m} t_{ij}\ F_j(y_j) \quad (1 \leq i \leq m). \tag{2.5}$$

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

$$S_n \ = \ A\,(A_0)^{n-1}, \tag{2.6.a}$$

$$T_{n,j} \ = \ O \ \text{(for } j > n), \quad T_{n,n} = B, \quad T_{n,j} = A\,(A_0)^{n-j-1}\,B_0 \ \text{(for } n > j). \tag{2.6.b}$$

Furthermore, when $F : \mathbb{R} \times \mathbb{V} \to \mathbb{V}$ satisfies (1.5), then definition (2.4) implies

$$\|v + \tau_0\,F_i(v)\| \ \leq\ \|v\| \quad (\text{for } 1 \leq\ i \leq\ m, \text{ and } v \in \mathbb{V}). \tag{2.7}$$

For analysing boundedness of (2.1), it is sometimes also handy to use *non*-canonical representations, of N steps of the method - see e.g. Section 3.3.2. Such representations share with the canonical representation the form (2.5), with property (2.7), but violate (2.6). Therefore, unless specified otherwise, in the following discussion of (2.5) we shall *not* assume S, T to satisfy (2.6), so that the conclusions, to be obtained about (2.5), can be applied both to canonical and non-canonical representations of method (2.1).

We shall interpret $x_i \in \mathbb{V}$ and $y_i \in \mathbb{V}$ as *input* and *output vectors*, respectively, of the *generic process* (2.5). In the situation (2.5), (2.7), we shall focus on the bound

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

We shall say that *process (2.5) satisfies the bound (2.8)* (for given stepsize Δt, vector space $\mathbb{V}$, seminorm $\|.\|$ and functions $F_i : \mathbb{V} \to \mathbb{V}$), if (2.8) holds whenever x_i and $y_i \in \mathbb{V}$ satisfy (2.5).

Clearly when (2.5) stands, as above, for N versions of (2.1) via the relations (2.4), (2.6), then boundedness of the GLM, defined in Section 2.1, corresponds to the situation where process (2.5) satisfies the bound (2.8) - with constant μ independent of $N = 1, 2, 3, \ldots$.

In Sections 2.3, 2.4, we shall present, without proof, the basic results of the paper, Theorems 2.2, 2.4. The theorems give conditions, on the ratio $\Delta t/\tau_0$, in order that process (2.5), with arbitrary parameter matrices $S = (s_{ij})$, $T = (t_{ij})$, satisfies the bound (2.8).

2.3 Satisfying the bound (2.8) for arbitrary functions F_i

In this subsection, we shall give our first main result, Theorem 2.2. The theorem deals with γ and μ such that the following general and fundamental property is present:

(2.9) Condition $0 < \Delta t \leq \gamma \cdot \tau_0$ implies that process (2.5) satisfies the bound (2.8), whenever $\mathbb{V}$ is a vector space with seminorm $\| \cdot \|$, and arbitrary functions $F_i : \mathbb{V} \to \mathbb{V}$ satisfy (2.7).

Theorem 2.2 concerns not only the above property (2.9), but also the following weaker property (2.10), in which the focus is on the *maximum norm*, defined by $\|x\|_\infty = \max_i |\xi_i|$ (for vectors $x \in \mathbb{R}^m$ with components ξ_i).

(2.10) Condition $\Delta t = \gamma \cdot \tau_0$ implies that process (2.5) satisfies the bound (2.8), when $\mathbb{V} = \mathbb{R}^m$, $\| \cdot \| = \| \cdot \|_\infty$, and arbitrary $F_i : \mathbb{R}^m \to \mathbb{R}^m$ satisfy (2.7).

The theorem below will show that the general property (2.9) is already present as soon as (2.10) is in force. Moreover, the theorem will give an algebraic criterion, in terms of γ, μ, for (2.9), (2.10) to be valid.

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

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

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

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

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

Our criterion - for properties (2.9), (2.10) - involves the following requirements:

(2.12.a) $I + \gamma T$ is invertible,

(2.12.b) $\mathrm{spr}(|P|) < 1$,

(2.12.c) $\| (I - |P|)^{-1} \, | R | \, \|_\infty \quad \leq \quad \mu.$

Theorem 2.2. (Criterion for the bound (2.8), when arbitrary F_i satisfy (2.7))
Consider process (2.5), with arbitrary coefficient matrices $S = (s_{ij})$ and $T = (t_{ij})$, and let positive τ_0, γ, μ be given. Then condition (2.12) is necessary and sufficient for property (2.9), as well as for (2.10).

Since property (2.9) is a-priori stronger than (2.10), the essence of the above theorem is that the algebraic condition (2.12) implies the (strong) statement (2.9), whereas already the (weaker) statement (2.10) implies (2.12).

Clearly, when γ satisfies (2.12.a), (2.12.b), the theorem shows that the smallest μ, for which statements (2.9), (2.10) hold, is equal to

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

In many practical situations, condition (2.12.c) is the essential requirement rather than conditions (2.12.a) or (2.12.b). One easily sees that the last two conditions will be satisfied, with any $\gamma > 0$, if T is lower triangular with nonnegative diagonal entries. This applies notably to the situation where T is strictly lower triangular, which corresponds to a numerical process that is explicit.

2.4 Satisfying the bound (2.8) for restricted functions F_i

Our second main result, Theorem 2.4 below, deals with important situations not adequately covered by Theorem 2.2. It is often *not* natural to allow - as in Theorem 2.2 - that all functions F_i are different from each other.

For instance, if in (2.1) we have $c_i = c_j$ for some $i \neq j$, or if the differential equation is autonomous, then N successive applications of (2.1) are represented canonically - via (2.4), (2.6) - by a process (2.5) with $F_i = F_j$ for some, or all, indices $i \neq j$.

Also when $c_i \neq c_j$ (for all $i \neq j$), and the differential equation is *non*-autonomous, it can happen that the canonical representation, obtained via (2.4), (2.6), amounts to a process (2.5) with $F_i = F_j$ for some indices $i \neq j$. According to (2.4), this situation occurs as soon as $n_1 + c_i = n_2 + c_j$ for some n_1, n_2, i, j with $n_1\, q + i \neq n_2\, q + j$. When a general LMM, cf. (1.7), is represented as a GLM as indicated in Section 2.1, then $N \geq 2$ applications of the GLM provide an example of this situation.

Below we shall see that, in cases where some of the functions F_i are equal to each other, condition (2.12) can be an unnecessarily strong requirement on γ in order that the stepsize restriction $0 < \Delta t \leq \gamma \cdot \tau_0$ implies the bound (2.8).

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

(2.14) $\quad \mathcal{I}_1, \ldots, \mathcal{I}_r$ are nonempty and mutually disjoint, with $\mathcal{I}_1 \cup \cdots \cup \mathcal{I}_r = \{1, \ldots, m\}$,

(2.15) $\quad F_i = F_j$ whenever i and j belong to the same index set $\mathcal{I}_\rho$.

Below, we shall deal with the following variant of property (2.9), in which the functions F_i are restricted according to (2.15):

(2.16) Condition $0 < \Delta t \leq \gamma \cdot \tau_0$ implies that process (2.5) satisfies the bound (2.8), whenever $\mathbb{V}$ is a vector space with seminorm $\| \cdot \|$, and functions $F_i : \mathbb{V} \to \mathbb{V}$ satisfy (2.7), (2.15).

We will see that finding a criterion for (2.16) is more subtle an issue than for (2.9). It will turn out to be convenient to consider, in addition to the above property (2.16), the following weaker version:

(2.17) Condition $\Delta t = \gamma \cdot \tau_0$ implies that process (2.5) satisfies the bound (2.8), whenever $\mathbb{V} = \mathbb{R}^m$ with seminorm $\| \cdot \|$, and $F_i : \mathbb{R}^m \to \mathbb{R}^m$ satisfy (2.7), (2.15).

Note that, because arbitrary seminorms occur in (2.17), this weaker version is *not* related to the original property (2.16), in the same way as the weaker version (2.10) is related to (2.9). An adaptation of (2.10), for the situation at hand, reads as follows:

(2.18) Condition $\Delta t = \gamma \cdot \tau_0$ implies that process (2.5) satisfies the bound (2.8), when $\mathbb{V} = \mathbb{R}^m$, $\| \cdot \| = \| \cdot \|_\infty$, and $F_i : \mathbb{R}^m \to \mathbb{R}^m$ satisfy (2.7), (2.15).

By Theorem 2.2, condition (2.12) is still *sufficient* in order that (2.16), (2.17), (2.18) hold. But, the following simple Example 2.3 shows that the condition is *no longer necessary* - cf. also Section 3.3.2 for a more natural, but less simple, counterexample.

Example 2.3. *Consider process* (2.5) *with* $l = 1$, $m = 2$ *and* $s_{i,1} = 1$, $t_{i,1} = 3$, $t_{i,2} = -2$. *Suppose (2.14), (2.15) with* $r = 1$, $\mathcal{I}_1 = \{1, 2\}$ *- i.e.* $F_1 = F_2$ *- and consider* $\gamma \geq 1/4$.

One easily sees that requirement (2.12.a) is fulfilled, and $spr(|P|) \geq 1$. *Therefore, condition (2.12.b) is violated.*

On the other hand, the process at hand is nothing but the (backward Euler) method $y_2 = y_1 = x_1 + \Delta t\, F_1(y_1)$, *which is of the form (2.5) - with* $\tilde{l} = \widetilde{m} = 1$ *and* $\tilde{S} = 1$, $\widetilde{T} = 1$. *Condition (2.12) is fulfilled by* $\tilde{S}$, $\widetilde{T}$, *with* $\mu = 1$, *for any* $\gamma > 0$.

In line with Theorem 2.2 (applied with $\tilde{S}$, $\widetilde{T}$*), we can conclude that the original process (with* $m = 2$*) must have property (2.16), with* $\mu = 1$, *for any* $\gamma > 0$, *although (2.12) is violated for* $\gamma \geq 1/4$.

In the following, we will see that violation of condition (2.12) while (2.16) is valid - as in the above example - is a phenomenon related to reducibility of the generic process (2.5). We will deal below with two irreducibility assumptions under which (2.12) cannot be violated.

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

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

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

We will use the *irreducibility assumption*

$$[\,S\ T\,](i,:) \neq [\,S\ T\,](j,:) \ \ (\text{if } i \neq j \text{ are in the same } \mathcal{I}_\rho \text{ and } T(:,i) \neq 0,\ T(:,j) \neq 0), \tag{2.19}$$

as well as the slightly stronger assumption

$$[\,S\ \widehat{T}\,](i,:) \neq [\,S\ \widehat{T}\,](j,:) \ \ (\text{if } i \neq j \text{ are in the same } \mathcal{I}_\rho \text{ and } T(:,i) \neq 0,\ T(:,j) \neq 0). \tag{2.20}$$

Clearly, if $r < m$ and there is *no* irreducibility in the sense of (2.19), then process (2.5) - with F_i satisfying (2.15) - is equivalent to a process (2.5) with a smaller value of m.

Theorem 2.4. (Criterion for the bound (2.8), when F_i satisfy (2.7), (2.15))
Consider process (2.5), with arbitrary coefficient matrices $S = (s_{ij})$ and $T = (t_{ij})$. Let positive τ_0, γ, μ be given, and assume (2.14).

(i) Assume irreducibility in the sense of (2.19). Then condition (2.12) is necessary and sufficient for property (2.16), as well as for (2.17).

(ii) Assume irreducibility in the sense of (2.20). Then condition (2.12) is necessary and sufficient for property (2.16), as well as for (2.18).

The above statement (i) shows that, under the irreducibility assumption (2.19), property (2.17) implies the algebraic property (2.12). On the other hand, statement (ii) reveals that under the stronger irreducibility assumption (2.20), already the weaker property (2.18) implies (2.12). The natural question thus arises of whether statements (i), (ii) can be combined and strengthened into the following proposition:

(iii) Assume irreducibility in the sense of (2.19). Then condition (2.12) is necessary and sufficient for property (2.16), as well as for (2.18).

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

Example 2.5. *Consider process (2.5) with $l = 1$, $m = 3$ and $s_{1,1} = 0$, $s_{2,1} = s_{3,1} = 1$, $t_{i,1} = i$, $t_{1,2} = t_{1,3} = 0$, $t_{2,2} = t_{3,2} = 3$, $t_{2,3} = t_{3,3} = -2$. Suppose (2.14), (2.15) with $r = 1$, $\mathcal{I}_1 = \{1,2,3\}$ - i.e. $F_1 = F_2 = F_3$ - and consider $\gamma = 1/4$.*

The irreducibility assumption (2.19) is fulfilled. Furthermore, one easily sees that requirement (2.12.a) is fulfilled, but $spr(|P|) = 1$. Therefore, condition (2.12) is violated.

On the other hand, for $\Delta t = \tau_0/4$ and $\mathbb{V}$, $\|\cdot\|$, F_i as in (2.18), it can be seen that $\|y_1\| = \|\Delta t\, F(y_1)\| = 0$, $\|y_2\| = \|y_3\| \leq \|x_1\|$. With $\mu = 1$, we thus have property (2.18).

Theorem 2.2 can formally be viewed as a special case of Theorem 2.4 - the latter theorem, with $r = m$ and the trivial index sets $\mathcal{I}_\rho = \{\rho\}$, implies the former. We have formulated Theorem 2.2 separately in view of its importance and simplicity: it does not need (2.14), (2.15) nor (2.19), (2.20). Moreover, by formulating first Theorem 2.2 explicitly, we could show in a natural way, via Example 2.3, that some additional (irreducibility) assumption is needed in order that condition (2.12) is the appropriate criterion when some F_i are equal.

3 Results related to the main theorems

3.1 Alternative conditions for properties (2.9), (2.16)

In this Section 3.1, we study process (2.5) with arbitrary coefficient matrices $S = (s_{ij})$ and $T = (t_{ij})$. We shall give conditions, for properties (2.9) and (2.16), which are in general simpler and easier to check than (2.12). In deriving these conditions, we shall use a lemma about condition (3.2.b) which will be presented first in Section 3.1.1.

The same notations will be used as in Section 2, notably (2.11), and any inequalities between matrices or vectors should be understood entry-wise or component-wise, respectively.

3.1.1 Background regarding condition (2.12.b)

The following lemma, about condition (2.12.b), will be used in Sections 3, 4.

Lemma 3.1. (Interpretations of (2.12.b)
Assume (2.12.a). Then each of the following three requirements is equivalent to (2.12.b).
(i) $I - |P|$ *is invertible, with* $(I - |P|)^{-1} \geq 0$*;*
(ii) $I - |P|$ *is invertible, and* $spr(|P|) \leq 1$*;*
(iii) There exist no real scalar λ *and vector* $\varphi \in \mathbb{R}^m$ *with:*

$$(\lambda\, I - |P|)\, \varphi = 0, \quad \varphi \neq 0, \quad \varphi \geq 0, \quad \lambda \geq 1. \tag{3.1}$$

Proof of Lemma 3.1.

One easily sees that (2.12.b) implies each of the properties (i), (ii), (iii). Conversely, applying the Perron-Frobenius theory as presented e.g. in Horn & Johnson (1988, p. 503), it follows that (2.12.b) is implied by (ii) as well as by (iii).

We shall complete the proof of the lemma by assuming (i) and proving (iii). Suppose, (iii) does *not* hold, i.e. there are λ, φ satisfying (3.1). Then $0 \geq -\varphi = (I - |P|)^{-1}\{(\lambda - 1)\, \varphi\} \geq 0$, so that $\varphi = 0$, which contradicts (3.1). □

3.1.2 Simplified conditions for properties (2.9), (2.16), with arbitrary μ

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

$$I + \gamma\, T \ \text{ is invertible,} \tag{3.2.a}$$

$$P \ \geq \ 0, \tag{3.2.b}$$

$$R \ \geq \ 0. \tag{3.2.c}$$

Assume (3.2.a) and (3.2.b) are fulfilled. We then see from Lemma 3.1 and the formula

$$I - P = Q = (I + \gamma\, T)^{-1} \tag{3.3}$$

(which follows from (2.11)), that condition (2.12.b) is equivalent to: $\mathrm{spr}(P) \ \leq \ 1$.

For matrices S, T satisfying (3.2), (2.12.b), we have $\|(I - |P|)^{-1}|R|\|_\infty = \|(I - P)^{-1}R\|_\infty = \|Q^{-1}Q\, S\|_\infty = \|S\|_\infty$. For such matrices we have also $S = (I - P)^{-1}\, R$, with $(I - P)^{-1} \geq 0$, so that $S \geq 0$ and $\|(I - |P|)^{-1}|R|\|_\infty = \max_i \sum_j s_{ij}$.

Consequently, under assumption (3.2), the conditions (2.12.b), (2.12.c) are equivalent to

$$\mathrm{spr}(\, P\,) \leq 1 \qquad \text{and} \qquad \sum_j s_{i\,j} \ \leq \ \mu \quad (1 \leq i \leq m). \tag{3.4}$$

In view of this equivalency, we have the following useful corollary to Theorems 2.2, 2.4:

Corollary 3.2. (Criterion for properties (2.9), (2.16), when $P \geq 0$, $R \geq 0$)
Let arbitrary matrices $S = (s_{ij})$, $T = (t_{ij})$ and positive values τ_0, γ, μ be given, such that (3.2) is fulfilled. Then the following two statements are valid.
(i) Condition (3.4) is necessary and sufficient for property (2.9).
(ii) Assume (2.14), (2.19). Then (3.4) is necessary and sufficient for property (2.16).

The following corollary to Theorem 2.2 is useful in cases where (3.2.a), (3.2.b) hold, but (3.2.c) is violated. It can be applied when constants ϱ_j, σ, τ are available such that the matrices $R = (r_{ij})$, $T = (t_{ij})$, $P = (p_{ij})$ satisfy

$$\text{spr}(P) \leq 1 \quad \text{and} \quad \sum_k |r_{jk}| \leq \varrho_j, \quad \sum_j \varrho_j \leq \sigma, \quad \max_{i,j} |t_{ij}| \leq \tau. \tag{3.5}$$

Corollary 3.3. (Condition for property (2.9), when $P \geq 0$)
Let arbitrary matrices $S = (s_{ij})$, $T = (t_{ij})$ and positive values τ_0, γ be given, such that (3.2.a), (3.2.b) are fulfilled. Then condition (3.5) guarantees property (2.9) with

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

Proof of Corollary 3.3.
In view of Theorem 2.2, it is sufficient to prove (2.12.b), (2.12.c) for the above μ. Condition (2.12.b) follows, from Lemma 3.1 and (3.3), as above. Furthermore, $\|(I-|P|)^{-1}\,|R|\|_\infty = \|(I-P)^{-1}\,|R|\|_\infty = \|(I+\gamma\,T)\,|R|\|_\infty \leq \|R\|_\infty + \gamma\,\|T\,|R|\|_\infty \leq \max_j \varrho_j + \gamma \cdot \max_i \sum_j |t_{ij}|\,\varrho_j$. □

3.1.3 Simplified criterion for properties (2.9), (2.16), with $\mu = 1$

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

$$s_{i\,1} + s_{i\,2} + \cdots + s_{i\,l} = 1 \quad (1 \leq i \leq m). \tag{3.6}$$

Assumption (3.6) is e.g. fulfilled when (2.5) stands for the canonical representation of N steps of a method (2.1) with coefficients α_{ij} satisfying

$$\alpha_{i\,1} + \alpha_{i\,2} + \cdots + \alpha_{i\,l} = 1 \quad (1 \leq i \leq q) \tag{3.7}$$

- this follows easily from (2.6.a). GLMs are often represented with coefficients α_{ij} such that (3.7) is in force, cf. e.g. the examples in Section 3.3.

We shall find that condition (3.2) is the appropriate criterion for properties (2.9), (2.16), by proving the equivalence of (3.2) and (2.12) (with $\mu = 1$). In our proof we shall use the notation E_k to denote the $k \times 1$ matrix with all entries equal to 1.

First, assume (3.2). In order to prove (2.12.b), (2.12.c) (with $\mu = 1$), we note that $P\,E_m = P\,S\,E_l = (I-Q)\,S\,E_l = E_m - R\,E_l \leq E_m$. It follows that $\|P\|_\infty \leq 1$, so that $\text{spr}(P) \leq 1$. Hence, (3.4) is in force, which in Section 3.1.2 was proved to be equivalent to (2.12.b), (2.12.c).

Conversely, assume (2.12) (with $\mu = 1$). We have $E_m = S\,E_l = (I-P)^{-1}\,R\,E_l \leq (I - |P|)^{-1}\,|R|\,E_l \leq E_m$. Hence $(I-|P|)^{-1}\,|R|\,E_l = E_m$, which implies $|P|\,E_m + |R|\,E_l = E_m = P\,E_m + R\,E_l$. Therefore, $(|P| - P)\,E_m + (|R| - R)\,E_l = 0$, so that $P = |P| \geq 0$, $R = |R| \geq 0$, i.e. (3.2).

In view of the equivalency of (3.2) and (2.12), the Theorems 2.2, 2.4 yield the following corollary, which is closely related to a monotonicity result formulated earlier in the literature (but derived differently), cf. Spijker (2007).

Corollary 3.4. (Criterion for properties (2.9), (2.16), with $\mu = 1$)
Let arbitrary matrices $S = (s_{ij})$, $T = (t_{ij})$ and positive τ_0, γ be given. Assume (3.6). Then the following two statements are valid.
(i) Condition (3.2) is necessary and sufficient for property (2.9) with $\mu = 1$.
(ii) Assume (2.14), (2.19). Then (3.2) is necessary and sufficient for (2.16) with $\mu = 1$.

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

By representing N steps of method (2.1) in the form (2.5) canonically - cf. (2.4), (2.6) - and a subsequent application of one of the Theorems 2.2, 2.4 or Corollaries 3.2, 3.3, 3.4, one can obtain conditions for boundedness of the GLM. Because such conditions involve the corresponding T, P and R - cf. (2.11) - we shall study these matrices, in the subsequent Lemma 3.5. The lemma will be applied in Section 3.3.

From (2.6), (2.11), we see that the matrices S, T, P, Q, R, respectively, corresponding to the canonical representation of N steps of (2.1) reduce, for $N = 1$, simply to:

$$A = (\alpha_{i\,j}), \quad B = (\beta_{i\,j}), \quad K = (I + \gamma\, B)^{-1}(\gamma B), \quad L = (I + \gamma\, B)^{-1}, \quad M = L\, A. \tag{3.8}$$

The following lemma relates (conditions on) T, P, R for *any* $N \geq 1$, directly to the simple matrices (3.8). We denote by K_0, M_0 the matrices consisting of the last l rows of K and M, respectively. Note that M_0 equals the $l \times l$ stability matrix $M(z)$ of the GLM at the point $z = -\gamma$, cf. e.g. Butcher (2003, p. 381).

Lemma 3.5. (On the matrices T, P, R of the canonical representation)
For given $\gamma > 0$, $\mu > 0$ and integer $N \geq 1$, the following statements are valid.
(i) Matrix T satisfies (2.12.a) if and only if $I + \gamma\, B$ is invertible.
(ii) If (2.12.a) holds, then matrix P satisfies (2.12.b) if and only if $spr(|K|) < 1$.
(iii) If (2.12.a) holds, then R is made up of $q \times l$ blocks R_n, and P of $q \times q$ blocks $P_{n,j}$, where $1 \leq n \leq N$, $1 \leq j \leq N$ and

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

Proof.
Part (i) follows from (2.6.b), and (ii) follows from the expressions for $P_{n,j}$ given in (iii).

To analyse the blocks R_n, we rewrite $(I + \gamma\, T)\, R = S$ in terms of these blocks, using (2.6): $\gamma \sum_{j=1}^{n-1} A\,(A_0)^{n-j-1}\, B_0\, R_j + (I + \gamma\, B)\, R_n = A\,(A_0)^{n-1}$ $(n \geq 1)$. To give this relation a more convenient form, we introduce the $l \times q$ matrix $H = [\, O\ I\,]$, composed of the $l \times (q - l))$ zero matrix O and the $l \times l$ identity matrix I. Clearly, $A_0 = H\, A$, $B_0 = H\, B$, $K_0 = H\, K$, $M_0 = H\, M$. We put $\bar{A} = A\, H$, $\bar{M} = M\, H$, so that

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

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

To complete the proof, we conclude from $(I + \gamma\, T)\, P = \gamma\, T$ and (2.6.b), that P has a block Toeplitz structure, with $q \times q$ blocks $P_{n,j} = P_{n-j+1}$ where $P_k = 0$ $(k \leq 0)$, $P_1 = K$. Similarly as above we find $\gamma \sum_{j=1}^{k-1} \bar{A}^{\,k-j} B\, P_j + (I + \gamma\, B)\, P_k = \gamma\, \bar{A}^{\,k-1} B$ $(k \geq 1)$ and $(I + \gamma\, B)\, P_k = \bar{A}\, P_{k-1}$, so that $P_k = (\bar{M})^{k-1} K = M\, M_0^{k-2} K_0$ $(k \geq 2)$. $\square$

3.3 Examples of actual boundedness results obtainable from the theory

This section only serves to make evident the practical relevance of the generic process (2.5) and the applicability of the above theory to the boundedness analysis of given GLMs, see Definition 2.1. Accordingly, below we will illustrate the theory by applying it just to a few actual numerical methods. In future work, cf. Hundsdorfer, Mozartova & Spijker (2009a), (2009b), we intend to use the theory for a more general analysis of classes of GLMs.

For ease of presentation, and also to illustrate (2.14), (2.15) and Theorem 2.4 with $r < m$, we deal throughout this section with autonomous problems - i.e. F in (1.1) is independent of t, and

(1.5) reduces to

$$\|v + \tau_0 F(v)\| \leq \|v\| \quad (\text{for } v \in \mathbb{V}). \tag{3.9}$$

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

(3.10) Condition $0 < \Delta t \leq \gamma \cdot \tau_0$ implies boundedness with constant μ, cf. Definition 2.1, whenever $\mathbb{V}$ is a vectorspace with seminorm $\|\cdot\|$ and $F : \mathbb{V} \to \mathbb{V}$ satisfies (3.9).

Clearly, when (3.10) holds with $\mu = 1$, then γ is a stepsize coefficient for *monotonicity*.

3.3.1 Two explicit RKMs

Following Gottlieb & Shu (1998), we consider two explicit RKMs (1.2), with $s = 2$, the nonzero coefficients of which are given by (3.11) and (3.12), respectively:

$$a_{2\,1} = 1, \;\; a_{3\,1} = a_{3\,2} = 1/2, \tag{3.11}$$

$$a_{2\,1} = -20, \;\; a_{3\,1} = 41/40, \;\; a_{3\,2} = -1/40. \tag{3.12}$$

Both methods are of second order and yield identical numerical approximations when applied to linear autonomous problems. The first method is monotonic ((3.10) with $\mu = 1$) with stepsize-coefficient $\gamma = 1$, whereas for method (3.12) there exists *no* positive stepsize-coefficient γ for monotonicity, cf. e.g. the paper just mentioned and Ferracina & Spijker (2004) or Higueras(2004).

To analyse for both methods the boundedness property (3.10) (with arbitrary $\mu \geq 1$), we represent the methods as GLMs (2.1) with coefficient matrices A, B - as indicated in Section 2.1 - and consider the corresponding canonical representation of $N \geq 1$ steps, cf. (2.4), (2.5), (2.6). Because F is independent of t, we have properties (2.14), (2.15) with $r = 1$, $\mathfrak{I}_1 = \{1, \cdots, m\}$. From (2.6) one sees that (2.19) and (2.20) are fulfilled, so that Theorem 2.4 can be applied. It follows that property (3.10) is present if and only if condition (2.12) is fulfilled (for all $N \geq 1$). From Lemma 3.5 we see that conditions (2.12.a), (2.12.b) are fulfilled, with any $\gamma > 0$, for both methods. In order to express the dependence of (2.12.c) on N, we put $\mu_N = ||(I - |P|)^{-1}\,|R|||_\infty$.

For method (3.11), it is possible to find by a computation based on Lemma 3.5 that, when $N \geq 1$,

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

Hence, for *any* given $\mu \geq 1$, the largest stepsize-coefficient γ, for which method (3.11) has the boundedness property (3.10), is equal to $\gamma = 1$.

For method (3.12), a similar computation yields $\mu_N = (1 + \frac{\gamma}{20} + \gamma^2)^{N-1}\,(1 + 40\,\gamma)$ (for $N \geq 1$ and $0 < \gamma \leq 2$). From this expression we can conclude that there exists *no* positive γ for which method (3.12) has the boundedness property (3.10) with *any* $\mu \geq 1$.

We think these conclusions, about methods (3.11), (3.12), nicely supplement and confirm the discussion of the methods, as presented in Gottlieb & Shu (1998): method (3.11) is superior to (3.12) not only regarding monotonicity, but also with respect to boundedness.

We have not displayed the details of the computations leading to the above expressions for μ_N, because we want to keep the size of the paper within reasonable limits, and intend to report on this kind of computations, in detail and greater generality, in Hundsdorfer, Mozartova & Spijker (2009a).

3.3.2 A two-stage RKM depending on a parameter θ

We shall give an example showing that the canonical representation of N steps of an (irreducible) RKM can fail to satisfy the irreducibility condition (2.19), with the result that Theorem 2.4 does *not* yield a necessary condition for boundedness. The example will also provide an instance of a *non-canonical* representation yielding a boundedness result that is *not* obtainable via the canonical

representation. Finally, it will show, unlike the examples in Section 3.3.1, that the restrictions on γ for *boundedness* of RKMs can be less severe than for *monotonicity*.

We consider the two-stage RKM, given by (1.2) with $s = 2,\ a_{1,1} = a_{1,2} = 0,\ a_{2,1} = a_{3,1} = 1-\theta,\ a_{2,2} = a_{3,2} = \theta$, with real parameter θ. We write the method concisely as (2.1) with $l = 1,\ q = 2,\ A = \begin{pmatrix} 1 \\ 1 \end{pmatrix},\ B = \begin{pmatrix} 0 & 0 \\ 1-\theta & \theta \end{pmatrix}$, and consider the corresponding canonical representation (2.5) of N consecutive steps of the method. We see from Lemma 3.5 that (2.12.a), (2.12.b) hold, if and only if $1 + 2\,\gamma\,\theta > 0$. Assuming this inequality to be fulfilled, it is possible to find by a computation using Lemma 3.5, that $\mu_N = ||(I-|P|)^{-1}\,|R|||_\infty$ equals $\mu_N = \lambda^N \ \ (N \geq 1)$, where $\lambda = \frac{|1+\gamma\,(\theta-1)|+\gamma\,|\theta-1|}{1+\gamma\,\theta-\gamma\,|\theta|} \geq 1$. We see that $\lambda = 1$, if and only if

$$0 \ \leq\ \theta \ \leq\ 1, \quad \gamma\,(1-\theta) \ \leq\ 1. \tag{3.13}$$

This does *not* allow us to conclude via Theorem 2.4 - with $r = 1,\ \mathfrak{I}_1 = \{1, \cdots, m\}$ as in Section 3.3.1 - that condition (3.13) is *necessary* for *boundedness* (property (3.10) with any fixed $\mu \geq 1$), because the irreducibility condition (2.19) on $[\,S\,T\,]$ is violated for $N \geq 2$.

On the other hand, Theorem 2.4 can be applied - with $r = 1,\ \mathfrak{I}_1 = \{1, \cdots, m\}$ - to the canonical representation for $N = 1$, because $[\,S\,T\,] = [\,A\,B\,]$ satisfies (2.19). Since $\mu_1 = \lambda$, condition (3.13) is *necessary and sufficient* for *monotonicity* ((3.10) with $\mu = 1$) - this follows also e.g. from Corollary 3.4, from Ferracina & Spijker (2004) or Higueras (2004).

To prove that boundedness is possible under a weaker condition than (3.13), we represent N steps of the method - *not* canonically - by (2.5) with $l = 1,\ m = N,\ s_{n,1} = 1,\ t_{nj} = 0\ (j > n),\ t_{nj} = \theta\ (j = n),\ t_{nj} = 1\ (j < n)$ and $y_n = u_n,\ x_1 = u_0 + \Delta t\,(1-\theta)\,F(u_0)$. Since $[\,S\,T\,]$ now satisfies (2.19) (with $r = 1,\ I = \{1, \ldots, m\}$), we can apply e.g. Corollary 3.4 to the situation at hand. A computation shows that (3.2) holds if and only if $0 \leq \theta,\ \gamma\,(1-\theta) \leq 1$. Hence, for any $\theta > 1\ \gamma > 0$, the conditions (1.6), (3.9) imply that $\|u_n\| \leq \ \|x_1\| = \|(1 + \frac{(\theta-1)\,\Delta\,t}{\tau_0})\,u_0 \ - \ \frac{(\theta-1)\,\Delta t}{\tau_0}\,(u_0 + \tau_0\,F(u_0))\| \leq \mu\,\|u_0\|$, with $\mu = 1 + 2\,(\theta-1)\,\gamma$.

In conclusion, for $\theta > 1$, there exists no positive stepsize coefficient for monotonicity, whereas any $\gamma > 0$ is a stepsize coefficient corresponding to the boundedness property (3.10), with $\mu = 1 + 2\,(\theta-1)\,\gamma$.

3.3.3 One-leg Adams-Bashforth method

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

$$u_n = u_{n-1} + \Delta t\,F\Big(\tfrac{3}{2}\,u_{n-1} - \tfrac{1}{2}\,u_{n-2}\Big), \tag{3.14}$$

cf. e.g. Butcher (1987), Hairer, Nørsett & Wanner (1987), Hairer & Wanner (1996). This method is *not monotonic*, in that there exists *no* positive γ with the property that (3.9), (3.14), (1.6) always imply $\|u_n\| \leq \max\{\|u_{n-1}\|, \|u_{n-2}\|\}$. This follows e.g. directly from Spijker (1983, Theorem 3.3).

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

$$\|u_n\| \ \leq\ \mu \cdot \max\{\|u_0\|, \|u_1\|\} \ \text{(for } 0 < \Delta t < \gamma \cdot \tau_0, \text{ and all } n \geq 2), \tag{3.15}$$

as soon as (3.9) and (3.14) (for $n \geq 2$) are in force.

Below we shall prove this boundedness result by rewriting method (3.14) as a GLM, and applying Corollary 3.3 in combination with Lemma 3.5 to the canonical representation, cf. (2.4), (2.5), (2.6).

We introduce, for $n \geq 1$, the vectors $v_1^{[n]} = -\frac{1}{2}\,u_{n-1} + \frac{3}{2}\,u_n,\ v_2^{[n]} = u_n,\ v_3^{[n]} = u_{n+1}$ and $u_1^{[n-1]} = u_{n-1},\ u_2^{[n-1]} = u_n$, so that (3.14) is equivalent to the GLM (2.1), with

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

Clearly, if this GLM satisfies (3.10) with positive γ, μ, then method (3.14) has the boundedness property mentioned above, cf. (3.15).

In order to apply Corollary 3.3 to the canonical representation of the GLM, we have to check conditions (3.2.a), (3.2.b) and (3.5). Because B is strictly lower triangular, we see directly from Lemma 3.5 (i) that (3.2.a) is fulfilled for any $\gamma > 0$.

To analyse (3.2.b) we consider, for any $\gamma > 0$, the expressions for the blocks $P_{n,j}$ given by Lemma 3.5 (iii). One easily sees that $P_{n,j} \geq 0$ ($j \geq n$). Furthermore, it can be seen that $P_{n,j} \geq 0$ (for $j = n-1$ and $j = n-2$) if and only if $\gamma \leq 4/9$. From now on we assume $\gamma = 4/9$. In the analysis of $P_{n,j}$ with $j \leq n-3$, via Lemma 3.5 (iii), it is convenient to use the following representation for the powers of M_0:

$$(M_0)^k = \begin{pmatrix} x_{k-1} & y_{k-1} \\ x_k & y_k \end{pmatrix}, \text{ where } \left\{ \begin{array}{ll} x_{k+1} = \frac{1}{3}\, x_k + \frac{2}{9}\, x_{k-1}, & x_0 = 0,\ x_1 = \frac{2}{9} \\ y_{k+1} = \frac{1}{3}\, y_k + \frac{2}{9}\, y_{k-1}, & y_0 = 1,\ y_1 = \frac{1}{3} \end{array} \right\} \text{ (for } k \geq 1).$$

Substituting this representation (with $k = n-j-1$) in the expression for $P_{n,j}$ of Lemma 3.5 (iii), it can be seen that $P_{n,j} \geq 0$ (for $j \leq n-3$), which proves (3.2.b).

The first inequality in (3.5) is fulfilled - with $\mathrm{spr}(P) = 0$ - because the blocks $P_{n,n}$ are strictly lower triangular. A computation, using the above representation for $(M_0)^k$, shows that the remaining inequalities in (3.5) are fulfilled as well, with $\varrho_j = 2$ (for $j = 1$), $\varrho_j = 3^{-n}[2^n - (-1)^n]$ (for $j = 3n-2$, $n \geq 2$), $\varrho_j = 3^{-n-1}[2^{n+2} - (-1)^n]$ (for $j = 3n-1$, $n \geq 1$), $\varrho_j = 3^{-n-2}[2^{n+3} + (-1)^n]$ (for $j = 3n$, $n \geq 1$) and $\sigma = 31/4$, $\tau = 3/2$. The upperbound $(1 + \gamma\,\tau)\,\sigma$ of Corollary 3.3 thus amounts to $155/12$, from which we conclude that method (3.14) has the boundedness property (3.15), with $\gamma = 4/9$ and $\mu = 155/12 \simeq 12.9$.

A smaller value for μ can be obtained by a straightforward - but slightly longer - computation of the expression $\mu = \max_j \varrho_j + \gamma \cdot \max_i \sum_j |t_{ij}|\, \varrho_j$, see Corollary 3.3. In this way one can arrive at a similar conclusion as above, but with $\gamma = 4/9$ and the better value $\mu = 31/9 \simeq 3.4$.

We note, for completeness, that the above results could not have been obtained by a similar application of Corollary 3.2, instead of Corollary 3.3, because condition (3.2.c) is violated, in the situation at hand, for all $N \geq 1$ and $\gamma > 0$.

3.3.4 A two-stage GLM

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

$$u_1^{[n]} = -u_1^{[n-1]} + 2\, u_2^{[n-1]}, \tag{3.16.a}$$

$$u_2^{[n]} = u_2^{[n-1]} + \Delta t \cdot F(u_1^{[n]}), \tag{3.16.b}$$

where $u_1^{[n-1]} \simeq u((n-1/2)\,\Delta t)$ and $u_2^{[n-1]} \simeq u(n\,\Delta t)$ ($n = 1, 2, 3, \ldots$). We write the method as (2.1), with $l = q = 2$, $A = \begin{pmatrix} -1 & 2 \\ 0 & 1 \end{pmatrix}$, $B = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$, and consider the corresponding canonical representation (2.5), cf. (2.4), (2.6). Because the matrix $[\,S\ T\,]$ satisfies the irreducibility condition (2.19) with $r = 1$, $\mathfrak{I}_1 = \{1, \ldots, m\}$, we can apply Theorem 2.4 in the situation at hand.

Let any $\gamma > 0$ be given. From Lemma 3.5 we see easily that the corresponding matrices T, P satisfy conditions (2.12.a), (2.12.b). By Theorem 2.4, the boundedness property (3.10) thus holds, for any given μ, if and only if $\mu_N = \|(I - |P|)^{-1}\,|R|\|_\infty$ is such that $\sup\{\mu_N : N \geq 1\} \leq \mu$.

Because $(I - |P|)^{-1}\,|R| \geq |R|$, we see from Lemma 3.5 (iii) that $\mu_N \geq \|R\|_\infty \geq \|M^N\|_\infty$, with M as in (3.8). From the expression $M = \begin{pmatrix} -1 & 2 \\ \gamma & 1 - 2\,\gamma \end{pmatrix}$, it follows that $\mathrm{spr}(M) = \gamma + \sqrt{1 + \gamma^2} > 1$, so that $\mu_N \to \infty$ for $N \to \infty$.

We conclude that *there is no boundedness*, in the sense of (3.10), for any positive γ and μ.

4 Proof of Theorems 2.2, 2.4

Because Theorem 2.2 follows from Theorem 2.4 by choosing in the latter theorem the trivial index sets $\mathfrak{I}_\rho = \{\rho\}$ (for $1 \le \rho \le m = r$), it is enough to prove below Theorem 2.4.

The sufficiency of condition (2.12), in parts (i) and (ii) of Theorem 2.4, is a direct consequence of Lemma 4.1, to be given in Section 4.1, and the fact that (2.9) implies the three properties (2.16), (2.17) and (2.18) (for any index sets $\mathfrak{I}_\rho$ as in (2.14)).

The necessity of condition (2.12), in Theorem 2.4, follows directly from Lemma 4.3, to be given in Section 4.2, and the fact that property (2.16) implies both (2.17) and (2.18).

4.1 Sufficiency of condition (2.12)

Lemma 4.1. (Sufficiency of condition (2.12) for property (2.9))
Let $\tau_0 > 0$ be given, and assume γ, μ are positive constants such that (2.12) holds. Then process (2.5) has the boundedness property (2.9).

In the following proof of the lemma, we shall write (2.5) and similar relations more concisely, by using the following notations relevant to the vector space $\mathbb{V}$. For any integer $k \ge 1$ and vectors $x_1, \ldots, x_k \in \mathbb{V}$, we denote the vector in $\mathbb{V}^k$ with components x_i by

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

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

We combine the vectors x_i and y_i, occurring in (2.5), into vectors $x = [x_i] \in \mathbb{V}^l$ and $y = [y_i] \in \mathbb{V}^m$, respectively. Furthermore, for given functions $F_i : \mathbb{V} \to \mathbb{V}$ $(1 \le i \le m)$, we define a function $\boldsymbol{F}$, from $\mathbb{V}^m$ to $\mathbb{V}^m$, by $\boldsymbol{F}(y) = [F_i(y_i)] \in \mathbb{V}^m$ for $y = [y_i] \in \mathbb{V}^m$. With these notations, the relations (2.5) can be written as an equality in $\mathbb{V}^m$:

$$y \;=\; \boldsymbol{S}x \;+\; \Delta t \cdot \boldsymbol{T}\,\boldsymbol{F}(y). \tag{4.1}$$

The subsequent lemma is a variant to Spijker (2007, Lemma 4.1). It will be useful, in the present section for proving Lemma 4.1, and later on for proving Lemma 4.3. We shall use the notations (2.11), and relate (4.1) - with F_i satisfying (2.7), (2.15) - to the conditions

$$y \;=\; \boldsymbol{R}\,x \,+\, \boldsymbol{P}\,z, \quad \text{with} \quad \|z_i\| \le \|y_i\| \;\; (1 \le i \le m), \tag{4.2.a}$$

$$y_i \;\neq\; y_j \quad \text{whenever } z_i \neq z_j \text{ and } i,\, j \text{ belong to the same index set } \mathfrak{I}_\rho. \tag{4.2.b}$$

Lemma 4.2. (Reformulation of (4.1) with F_i satisfying (2.7), (2.15))
Let $\tau_0 > 0$, $\gamma > 0$, $I + \gamma T$ invertible, and assume (2.14). Let $x = [x_i] \in \mathbb{V}^l$ and $y = [y_i] \in \mathbb{V}^m$ be given. Then the following three statements are equivalent:

(4.3) *The vectors x, y satisfy (4.1) for some Δt with $0 < \Delta t \le \gamma \cdot \tau_0$ and some functions $F_i : \mathbb{V} \to \mathbb{V}$ satisfying (2.7), (2.15);*

(4.4) *The vectors x, y satisfy (4.1) with $\Delta t = \gamma \cdot \tau_0$ and some functions $F_i : \mathbb{V} \to \mathbb{V}$ satisfying (2.7), (2.15);*

(4.5) *There exists a vector $z = [z_i] \in \mathbb{V}^m$ such that (4.2) holds.*

Proof of Lemma 4.2.

Assume (4.3). In order to prove (4.4), we define $\theta = \Delta t/(\gamma\,\tau_0)$ and $\tilde{F}_i = \theta \cdot F_i$, so that x, y satisfy (4.1) also with $\Delta t = \gamma \cdot \tau_0$ and F_i replaced by $\tilde{F}_i$. Clearly, $\tilde{F}_i = \tilde{F}_j$ for i, j in the same index set, and $\|v + \tau_0\,\tilde{F}_i(v)\| = \|(1-\theta)\,v + \theta\,[v + \tau_0\,F_i(v)]\| \le \|v\|$. This implies (4.4).

Assume (4.4). In order to prove (4.5), we rewrite (4.1) as

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

from which we see that x, y satisfy (4.2.a) with $z = [z_i] = y + \tau_0\,\mathbf{F}(y)$. Furthermore, when $z_i \neq z_j$ and i, j belong to the same index set $\mathfrak{I}_\rho$, we have $y_i + \tau_0\,F_i(y_i) \neq y_j + \tau_0\,F_i(y_j)$, which implies (4.2.b). Hence, (4.5) holds.

Assume (4.5). We shall prove (4.3). For $i \in \mathfrak{I}_\rho$ we define $F_i(v) = (1/\tau_0)\,(z_k - y_k)$ (if $v = y_k,\ k \in \mathfrak{I}_\rho$) and $F_i(v) = 0$ (otherwise). In view of (4.2.b) this is a proper definition, and $F_i = F_j$ for i, j in the same index set, i.e. (2.15). Furthermore, we see that x, y satisfy (4.1) with $\Delta t = \gamma \cdot \tau_0$. Finally, for $i \in \mathfrak{I}_\rho$, we have $\|v + \tau_0\,F_i(v)\| = \|z_k\| \le \|v\|$ (if $v = y_k,\ k \in \mathfrak{I}_\rho$) and $\|v + \tau_0\,F_i(v)\| = \|v\|$ (otherwise), so that (2.7) is fulfilled. This completes the proof of (4.3). □

Proof of Lemma 4.1

Assume condition (2.12) is fulfilled, and consider x_i, y_i satisfying (2.5), in the situation where (2.7) holds and $0 < \Delta t \le \gamma \cdot \tau_0$. Applying Lemma 4.2 (with the trivial index sets $\mathfrak{I}_\rho = \{\rho\}$, $1 \le \rho \le r = m$), we have (4.2.a), from which we obtain

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

Consequently, $(I - |P|)\,[\,\|y_i\|\,] \ \le\ [\,\|r_i\|\,]$. By Lemma 3.1, the matrix $I - |P|$ is invertible with $(I - |P|)^{-1} \ge 0$. Therefore, $[\,\|y_i\|\,] \ \le\ (I - |P|)^{-1}\,[\,\|r_i\|\,]$, which implies

$$[\,\|y_i\|\,] \ \le\ (I - |P|)^{-1}\,|R|\,[\,\|x_i\|\,]. \tag{4.6}$$

An application of (2.12.c) shows that the components in the right-hand member of the last inequality do not exceed $\mu \cdot (\max_j \|x_j\|)$, which completes the proof of Lemma 4.1. □

4.2 Necessity of condition (2.12)

In this section we shall prove the necessity of condition (2.12) for properties (2.17) and (2.18), under the irreducibility assumptions (2.19) and (2.20), respectively.

We assume throughout the section that τ_0, γ, μ are given positive constants and, unless stated otherwise, that $\mathfrak{I}_\rho$ are arbitrary given index sets of type (2.14).

4.2.1 Formulation and proof of Lemma 4.3

Lemma 4.3. (Necessity of condition (2.12) for properties (2.17), (2.18))
(i) Assume property (2.17) and irreducibility in the sense of (2.19). Then (2.12) holds.
(ii) Assume property (2.18) and irreducibility in the sense of (2.20). Then (2.12) holds.

Our proof of Lemma 4.3 will need three lemmas, the first of which is

Lemma 4.4. (Invertibility of $I + \gamma\,T$)
Property (2.17), as well as property (2.18), implies that the matrix $I + \gamma\,T$ is invertible.

Proof of Lemma 4.4.

Assume (2.17) or (2.18). Let $\eta = [\eta_i] \in \mathbb{R}^m$ with $(I + \gamma\,T)\,\eta = 0$. We shall prove $\eta = 0$.

We define $F_i(v) = -(1/\tau_0)\,v$ (for all $v \in \mathbb{V} = \mathbb{R}^m$), so that (2.15), (2.7) are fulfilled with $\|\cdot\| = \|\cdot\|_\infty$. We see that (2.5) is satisfied, with $\Delta t = \gamma \cdot \tau_0$, by the vectors $x_i = 0$ $(1 \le i \le l)$ and $y_i = \eta_i\,e_1$ $(1 \le i \le m)$, where e_1 is the first unit vector in $\mathbb{V} = \mathbb{R}^m$.

By (2.17) or (2.18), there follows $|\eta_i| = \|y_i\|_\infty \le \mu \cdot \max_j \|x_j\|_\infty = 0$, so that $\eta = 0$. □

In proving that property (2.17) implies (2.12), we shall make use of vectors $\xi = [\xi_j] \in \mathbb{R}^l$ and $\eta = [\eta_j]$, $\zeta = [\zeta_j] \in \mathbb{R}^m$ satisfying the following condition:

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

Furthermore, in proving that the (weaker) property (2.18) implies (2.12), we shall use vectors $\xi = [\xi_j] \in \mathbb{R}^l$ and $\eta = [\eta_j]$, $\zeta = [\zeta_j] \in \mathbb{R}^m$ satisfying the subsequent (stronger) condition:

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

We shall see that vectors ξ, η, ζ exist satisfying conditions (4.7) and (4.8), respectively, if the following (simplified) versions of assumptions (2.19) and (2.20) are fulfilled:

(4.9) $[\,S\;T\,](i,:) \neq [\,S\;T\,](j,:)$ (if $i \neq j$ belong to the same index set $\mathfrak{I}_\rho$),

(4.10) $[\,S\;\widehat{T}\,](i,:) \neq [\,S\;\widehat{T}\,](j,:)$ (if $i \neq j$ belong to the same index set $\mathfrak{I}_\rho$).

Our proof of Lemma 4.3 needs also the following two lemmas (to be proved in Section 4.2.2).

Lemma 4.5. (Relevance of (4.7), (4.8))
(i) Assume (2.17), and suppose ξ, η, ζ satisfy (4.7). Then condition (2.12) is fulfilled.
(ii) Assume (2.18), and suppose ξ, η, ζ satisfy (4.8). Then condition (2.12) is fulfilled.

Lemma 4.6. (Conditions for (4.7), (4.8))
Let $I + \gamma\,T$ be invertible. Then the following two implications are valid.
(i) Assumption (4.9) implies the existence of ξ, η, ζ satisfying (4.7).
(ii) Assumption (4.10) implies the existence of ξ, η, ζ satisfying (4.8).

Proof of Lemma 4.3.

Part 1. Because of Lemma 4.4, we can assume that $I + \gamma\,T$ is invertible.

Part 2. In order to prove Part (i) of Lemma 4.3, we assume (2.17), (2.19). We denote by I^0 the set of all indices i, with $1 \le i \le m$ and $T(:,i) = 0$.

First, assume there are *no* index sets $\mathfrak{I}_\rho$ containing a pair of indices $i \neq j$ with $j \in I^0$. Conditions (2.19) and (4.9) are then equivalent. Hence, combining Lemma 4.6 (i) and Lemma 4.5 (i), we obtain (2.12).

Next, assume there do exist sets $\mathfrak{I}_\rho$ containing indices $i \neq j$ where $j \in I^0$. We note that the functions F_j, with $j \in I^0$, *do not enter actually* in the basic relations (2.5). Accordingly, it is immaterial for these relations whether or not a given function F_j, with $j \in I^0$, is equal to any F_i with $i \neq j$. Therefore, we can refine the given partition $\mathfrak{I}_1 \cup \cdots \cup \mathfrak{I}_r = \{1, \cdots, m\}$ into one with regard to which properties (2.17) and (4.9) hold: the refined partition is obtained, from the original one, by creating new separate index sets for all indices $j \in I^0$ belonging to an (old) index set $\mathfrak{I}_\rho$ with at least two different indices.

From (the original) property (2.17) one sees that (2.17) is still present with regard to the new, refined partition. Moreover, the original property (2.19) implies that (4.9) is valid with regard to the new index sets. Therefore, we arrive at (2.12), again by combining Lemma 4.6 (i) and 4.5 (i) (in the situation of the new partition).

Part 3. To prove Part (ii) of Lemma 4.3, assume (2.18), (2.20), and define I^0 as above.

First assume there are *no* sets $\mathfrak{I}_\rho$ containing indices $i \neq j$ where $j \in I^0$. Conditions (2.20) and (4.10) are then equivalent. Hence, Lemmas 4.6 (ii) and 4.5 (ii) yield (2.12).

Next assume there do exist sets $\mathfrak{I}_\rho$ with indices $i \neq j$ where $j \in I^0$. Using the above refined partition, similarly as in Part 2 of the proof, we arrive again at (2.12) by combining Lemma 4.6 (ii) and 4.5 (ii). □

4.2.2 Proof of the Lemmas 4.5, 4.6

The sole purpose of the present section is to prove Lemmas 4.5, 4.6. Throughout the section we assume, with no loss of generality, that $I + \gamma T$ is invertible. We shall use the notation

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

Proof of Lemma 4.5

Part 1a. Assume (2.17), and let ξ, η, ζ satisfy (4.7). We shall prove (2.12.b) via Lemma 3.1, by assuming that λ and φ satisfy (3.1), and deducing a contradiction from that assumption.

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

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

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

$$\|z_j\|_\infty = |z_{ij}| = \varphi_j \leq y_{jj} = \|y_j\|_\infty \quad (1 \leq i \leq m,\ 1 \leq j \leq m). \tag{4.11}$$

First, suppose $y_j \neq y_k$ for all $j \neq k$ belonging to the same index set $\mathcal{I}_\rho$. Then x, y, z satisfy (4.2), with $\|\cdot\| = \|\cdot\|_\infty$, so that, by Lemma 4.2, the vectors x, y satisfy (4.4) with $\mathbb{V} = \mathbb{R}^m$, $\|\cdot\| = \|\cdot\|_\infty$. By property (2.17) and (4.11), there follows $\|\varphi\|_\infty \leq \max_j \|y_j\|_\infty \leq \mu \cdot \max_k \|x_k\|_\infty = 0$. Hence $\varphi = 0$, which contradicts (3.1) and thus proves (2.12.b).

Next, suppose $y_q = y_s$ for two indices $q < s$ belonging to the same set $\mathcal{I}_\rho$. In this situation, we modify (only) the q-th component of all x_j, y_j, z_j into $\tilde{x}_{qj} = \xi_j$, $\tilde{y}_{qj} = \eta_j$, $\tilde{z}_{qj} = \zeta_j$, and we denote the resulting vectors by $\tilde{x}_j$, $\tilde{y}_j$, $\tilde{z}_j$, respectively. The vectors $\tilde{x} = [\tilde{x}_j]$, $\tilde{y} = [\tilde{y}_j]$, $\tilde{z} = [\tilde{z}_j]$ satisfy the following variant of condition (4.2):

$$\tilde{y} = \boldsymbol{R}\,\tilde{x} + \boldsymbol{P}\,\tilde{z}, \quad \tilde{y}_j \neq \tilde{y}_k \ \text{ (for all } j \neq k \text{ in the same index set)}. \tag{4.12}$$

In order that $\tilde{x}$, $\tilde{y}$, $\tilde{z}$ actually fulfill (4.2), we define the special seminorm

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

Because y_j, z_j satisfy (4.11), we have

$$\|\tilde{z}_j\| = \|z_j\|_\infty \leq \|y_j\|_\infty = \|\tilde{y}_j\| \ \text{ (for } 1 \leq j \leq m) \tag{4.13}$$

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

Clearly, with the above special seminorm in $\mathbb{V}$, the vectors $\tilde{x}$, $\tilde{y}$, $\tilde{z}$ fulfill (4.2), so that $\tilde{x}$, $\tilde{y}$ satisfy (4.4). Using property (2.17) and the last equality in (4.13), we find $\max_j \|y_j\|_\infty = \max_j \|\tilde{y}_j\| \leq \mu \cdot \max_k \|\tilde{x}_k\| = 0$. In view of (4.11), it follows that $\varphi = 0$, which proves (2.12.b).

Part 1b. Assuming (2.17), (4.7), we shall prove (2.12.c). We have $\|\ (I - |P|)^{-1}\ |\,R\,|\ \|_\infty = \|\varphi\|_\infty$, with $\varphi = [\varphi_i] \in \mathbb{R}^m$, where the values $\varphi_i \geq 0$ satisfy the linear equations

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

Condition (2.12.c) is thus equivalent to

$$\|\varphi\|_\infty \leq \mu. \tag{4.14}$$

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

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

Clearly $y = \boldsymbol{R}\,x + \boldsymbol{P}\,z$, and because $y_{jj} = \sum_{k=1}^{l} |r_{jk}| + \sum_{k=1}^{m} |p_{jk}|\,\varphi_k = \varphi_j$, the relations (4.11) are again fulfilled.

First, suppose $y_j \neq y_k$ for all $j \neq k$ belonging to the same index set $\mathfrak{I}_\rho$. Then x, y, z satisfy (4.2), with $\|\cdot\| = \|\cdot\|_\infty$, so that, by Lemma 4.2, the vectors x, y satisfy (4.4) with $\mathbb{V} = \mathbb{R}^m$, $\|\cdot\| = \|\cdot\|_\infty$. By property (2.17) and (4.11), there follows $\|\varphi\|_\infty \leq \max_j \|y_j\|_\infty \leq \mu \cdot \max_k \|x_k\|_\infty = \mu$, which implies (4.14).

Next, suppose $y_q = y_s$, where $q < s$ belong to the same set $\mathfrak{I}_\rho$. We modify the q-th component of x_j, y_j, z_j as above in Part 1a of the proof. The resulting vectors $\tilde{x} = [\tilde{x}_j]$, $\tilde{y} = [\tilde{y}_j]$, $\tilde{z} = [\tilde{z}_j]$ satisfy again (4.12), and - in view of (4.11) - they satisfy also (4.13).

Consequently, $\tilde{x}$, $\tilde{y}$, $\tilde{z}$ fulfill condition (4.2), so that $\tilde{x}$, $\tilde{y}$ satisfy (4.4) with the special seminorm defined above. Using property (2.17) and the last equality in (4.13), we find $\max_j \|y_j\|_\infty = \max_j \|\tilde{y}_j\| \leq \mu \cdot \max_k \|\tilde{x}_k\| = \mu$, which proves again (4.14).

Part 2a. Assume (2.18) and (4.8). We shall again prove (2.12.b) via Lemma 3.1.

Denote by λ and $\varphi = [\varphi_i]$, $x = [x_j] = [[x_{ij}]]$, $y = [y_j] = [[y_{ij}]]$, $z = [z_j] = [[z_{ij}]]$ the same scalar and vectors as in Part 1a of the proof, so that (4.11) is again in force.

First, suppose $y_j \neq y_k$ for all $j \neq k$ belonging to the same index set $\mathfrak{I}_\rho$. Similarly as in Part 1a, we arrive at $\varphi = 0$, which proves (2.12.b).

Next, suppose $y_q = y_s$ where $q < s$ belong to the same set $\mathfrak{I}_\rho$. Define $\tilde{x}_j$, $\tilde{y}_j$, $\tilde{z}_j$ as in Part 1a, but now with ξ, η, ζ satisfying (4.8). We have again (4.12), (4.13), and therefore

$$\|\tilde{z}_j\|_\infty = \max\{\|\tilde{z}_j\|, |\zeta_j|\} \leq \max\{\|\tilde{y}_j\|, |\eta_j|\} = \|\tilde{y}_j\|_\infty. \tag{4.15}$$

Hence, $\tilde{x}_j$, $\tilde{y}_j$, $\tilde{z}_j$ satisfy (4.2) with $\|\cdot\| = \|\cdot\|_\infty$. Via Lemma 4.2 and property (2.18) we obtain $\|\tilde{y}_j\|_\infty \leq \mu \cdot \|\xi\|_\infty$, and in view of (4.11), (4.13) there follows $\|\varphi\|_\infty \leq \mu \cdot \|\xi\|_\infty$.

By suitable scaling of ξ, η, ζ, with property (4.8), we can achieve that $\|\xi\|_\infty$ is arbitrarily close to zero. Hence, $\varphi = 0$, which proves (2.12.b).

Part 2b. Assuming (2.18), (4.8), we shall prove (2.12.c).

The beginning of the proof runs as in Part 1b above, using (4.8) instead of (4.7). We arrive again at (2.12.c), via (4.14), if $y_j \neq y_k$ for all $j \neq k$ belonging to the same set $\mathfrak{I}_\rho$.

If $y_q = y_s$, for some $q < s$ belonging to the same $\mathfrak{I}_\rho$, we proceed as in Part 2a above, and introduce $\tilde{x}_j$, $\tilde{y}_j$, $\tilde{z}_j$ satisfying (4.2) with $\|\cdot\| = \|\cdot\|_\infty$. From Lemma 4.2 and property (2.18) it follows that $\|\tilde{y}_j\|_\infty \leq \mu \cdot \max_k\{1,\ \|\xi\|_\infty\}$, and in view of (4.11), (4.13) we obtain $\|\varphi\|_\infty \leq \mu \cdot \max_k\{1,\ \|\xi\|_\infty\}$.

By arranging that $\|\xi\|_\infty < 1$, we obtain (4.14) and therefore also (2.12.c). □

Proof of Lemma 4.6

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

$$\eta_i = \sum_k s_{ik}\,\xi_k + \sum_k t_{ik}\,\lambda_k, \quad \zeta_i = \eta_i + \lambda_i/\gamma \quad (1 \leq i \leq m). \tag{4.16}$$

The definition is easily seen to imply

$$\eta = R\,\xi + P\,\zeta. \tag{4.17}$$

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

Assuming (4.9), one can see that ξ_i, λ_i exist, such that η_i, defined by (4.16), satisfy

$$\eta_i \neq \eta_j \quad (\text{for any } i \neq j \text{ in the same index set } \mathfrak{I}_\rho). \tag{4.18}$$

Because (4.16) implies (4.17), it follows that ξ, η, ζ exist satisfying (4.7).

Part 2. Assuming (4.10), we shall determine scalars ε, μ_k, ξ_k, with

$$0 \leq \varepsilon\, \mu_k \leq 2\, \gamma, \tag{4.19}$$

such that the system of equations

$$\eta_i = \sum_k s_{ik}\, \xi_k - \varepsilon \sum_k t_{ik}\, \mu_k\, \eta_k \quad (1 \leq i \leq m) \tag{4.20}$$

has a solution $\eta = [\eta_i]$ satisfying (4.18). Using the implication (4.16) $\Rightarrow$ (4.17) (with $\lambda_i = -\varepsilon\, \mu_i\, \eta_i$), one sees that such scalars ε, μ_k, ξ_k lead to (4.8) (with $\zeta_i = (1 - \frac{\varepsilon\, \mu_i}{\gamma})\, \eta_i$).

To find ε, μ_k, ξ_k with the above properties, consider first any fixed μ_k, ξ_k, and note that the corresponding system (4.20) has a solution $\eta_i = \eta_i(\varepsilon)$, for $\varepsilon > 0$ small enough, with

$$\eta_i(\varepsilon) = \sigma_i - \varepsilon\, \tau_i + \mathcal{O}(\varepsilon^2) \ (\text{for } \varepsilon \downarrow 0), \quad \text{where} \quad \sigma_i = \sum_k s_{ik}\, \xi_k, \quad \tau_i = \sum_k t_{ik}\, \sigma_k\, \mu_k. \tag{4.21}$$

Aiming at (4.18) (with $\eta_i = \eta_i(\varepsilon)$), we are lead by (4.21) to fix ξ_k such that

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

Below we shall specify μ_k, in terms of values ϱ_k which are determined such that

$$\operatorname{sgn}(\varrho_k) = \operatorname{sgn}(\sigma_k) \ (\text{for } 1 \leq k \leq m) \quad \text{and} \quad \sum_k \hat{t}_{ik}\, \varrho_k \neq \sum_k \hat{t}_{jk}\, \varrho_k \ (\text{for } \hat{T}(i,:) \neq \hat{T}(j,:)).$$

We define $\mu_k = \varrho_k/\sigma_k$ (if $\sigma_k \neq 0$) and $\mu_k = 0$ (if $\sigma_k = 0$). It follows that

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

Because of (4.10), the values σ_i, τ_i corresponding to ξ_k, μ_k thus specified, satisfy

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

Combining these inequalities with (4.21), it follows that (4.18) (with $\eta_i = \eta_i(\varepsilon)$) and (4.19) hold for sufficiently small $\varepsilon > 0$. Hence ε, μ_k, ξ_k exist with the properties stated above. □

References

[1] Butcher J.C (1966): *On the convergence of numerical solutions to ordinary differential equations*, Math. Comp. **20**, 1-10.

[2] Butcher J.C (1987): *The numerical analysis of ordinary differential equations*, John Wiley, Chichester, UK.

[3] Butcher J.C (2003): *Numerical methods for ordinary differential equations*, John Wiley, Chichester, UK.

[4] Ferracina L, Spijker M.N (2004): *Stepsize restrictions for the total-variation-diminishing property in general Runge-Kutta methods*, SIAM J. Numer. Anal. **42**, 1073-1093.

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

[6] Gottlieb S, Shu C.-W (1998): *Total-variation-diminishing Runge-Kutta schemes*, Math. Comp. **67**, 73-85.

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

[8] Hairer E, Wanner G (1996): *Solving ordinary differential equations. II. Stiff and differential-algebraic problems*, Springer-Verlag, Berlin.

[9] Hairer E, Nørsett S.P, Wanner G (1987): *Solving ordinary differential equations. I. nonstiff problems*, Springer-Verlag, Berlin.

[10] Harten A (1983): *High resolution schemes for hyperbolic conservation laws*, J. Comput. Phys. **49**, 357-393.

[11] Higueras I (2004): *On strong stability preserving time discretization methods*, Journ. Scientif. Computing **21**, 193-223.

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

[13] Horn R.A, Johnson C.R (1998): *Matrix analysis*, Cambridge University Press, Cambridge.

[14] Hundsdorfer W, Mozartova A, Spijker M.N (2009a): *Boundedness and monotonicity of Runge-Kutta methods*, in preparation.

[15] Hundsdorfer W, Mozartova A, Spijker M.N (2009b): *Stepsize restrictions for monotonicity and boundedness of multistep methods*, in preparation.

[16] Hundsdorfer W, Ruuth S.J (2003): *Monotonicity for time discretizations*, Procs. Dundee Conference 2003, pp. 85-94. Eds. D.F. Griffiths, G.A. Watson, Report NA/217, Univ. Dundee.

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

[18] Hundsdorfer W, Verwer J.G (2003): *Numerical solution of time-dependent advection-diffusion-reaction equations*, Springer Ser. Comp. Math., Vol 33, Springer (Berlin)

[19] LeVeque R.J (2002): *Finite volume methods for hyperbolic problems*, Cambridge University Press, Cambridge.

[20] Ruuth S.J (2006): *Global optimization of explicit strong-stability-preserving Runge-Kutta methods*, Math. Comp. **75**, 183-207.

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

[22] Shu C.-W (1988): *Total-variation-diminishing time discretizations*, SIAM J. Sci. Statist. Comput., **9**, 1073-1084.

[23] Shu C.-W, Osher S (1988): *Efficient implementation of essentially nonoscillatory shock-capturing schemes*, J. Comput. Phys. **77**, 439-471.

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

[25] Spijker M.N (2007): *Stepsize conditions for general monotonicity in numerical initial value problems*, SIAM J. Numer. Anal. **45**, 1226-1245.

[26] Spiteri R.J, Ruuth S.J (2002): *A new class of optimal high-order strong-stability-preserving time discretization methods*, SIAM J. Numer. Anal. **40**, 469-491.