Stepsize conditions for general monotonicity in numerical initial value problems

M. N. Spijker*

2006, June 3, Report MI 2006-08, Math. Inst., Leiden Univ.

Abstract. For Runge-Kutta Methods and Linear Multistep Methods, much attention has been paid, in the literature, to special nonlinear stability properties indicated by the terms total-variation-diminishing (TVD), strong stability preserving (SSP) and monotonicity. Stepsize conditions, guaranteeing these properties, were studied e.g. by Shu & Osher (1988), Gottlieb, Shu & Tadmor (2001), Hundsdorfer & Ruuth (2003), Higueras (2004a, 2005), Spiteri & Ruuth (2002), Gottlieb (2005), Ferracina & Spijker (2004, 2005).

In this paper, we obtain a special stepsize condition guaranteeing the above properties, for a generic numerical process. This condition is best possible in a well defined and natural sense. It is applicable to the important class of General Linear Methods, and it can also be used to answer some open questions, for methods of which the above stability properties were studied earlier.

Key words. initial value problem, method of lines (MOL), ordinary differential equation (ODE), General Linear Method, total-variation-diminishing (TVD), strong-stability-preserving (SSP), monotonicity.

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

1 Introduction

1.1 Maximal stepsize-coefficients for monotonicity

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

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

We shall deal with step-by-step-methods for finding numerical approximations u_n to the true solution values $U(n\Delta t)$, where Δt denotes a positive stepsize and $n = 1, 2, 3, \ldots$.

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

$$\text{(1.2.a)} \qquad y_i = u_{n-1} + \Delta t \cdot \sum_{j=1}^{s} a_{ij}\, f((n-1+c_j)\Delta t, y_j) \quad (1 \leq i \leq s+1),$$

$$\text{(1.2.b)} \qquad u_n = y_{s+1}.$$

Here $a_{i,j}$ and c_j are parameters defining the method, whereas y_i $(1 \leq i \leq s)$ are intermediate approximations used for computing $u_n = y_{s+1}$ from u_{n-1}. If $a_{ij} = 0$ (for $j \geq i$), the method is called *explicit*.

In the following, $\mathbb{V}$ stands for the vector space on which the differential equation is defined, and $\|\cdot\|$ denotes a convex function on $\mathbb{V}$ (i.e.: $\|\lambda v + (1-\lambda)w\| \leq \lambda\|v\| + (1-\lambda)\|w\|$ for $0 \leq \lambda \leq 1$ and $v, w \in \mathbb{V}$). Much attention has been paid in the literature to the property

$$\|y_i\| \leq \|u_{n-1}\| \quad (\text{for } 1 \leq i \leq s+1). \tag{1.3}$$

Clearly, (1.3) implies $\|u_n\| \leq \|u_{n-1}\|$. The latter property, as well as property (1.3), is often referred to by the term *monotonicity* or *strong stability*; it is of particular importance in

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

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 $\|u_n\|_{TV} \leq \|u_{n-1}\|_{TV}$, play a special role in the solution of hyperbolic conservation laws and are called *total variation diminishing*, cf. e.g. Harten (1983), Shu (1988), Shu & Osher (1988), LeVeque (2002), Hundsdorfer & Verwer (2003). We note that, for practical calculations, special importance has been attached to the inequality $\|y_i\| \leq \|u_{n-1}\|$ being fulfilled for *all* i with $1 \leq i \leq s+1$ (rather than just for $i = s+1$) - see e.g. Shu (2002), Gottlieb (2005).

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

$$\|v + \tau_0\, f(t,v)\| \;\leq\; \|v\| \quad (\text{for } t \in \mathbb{R},\, v \in \mathbb{V}). \tag{1.4}$$

Assumption (1.4) 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.

In the literature, *stepsize-coefficients* c were determined such that monotonicity, in the sense of (1.3), is present for all Δt with

$$0 < \Delta t \leq c \cdot \tau_0. \tag{1.5}$$

For explicit RKMs, this was done by rewriting the right-hand members of (1.2.a) as convex combinations of forward Euler steps - see e.g. Shu & Osher (1988), Spiteri & Ruuth (2002), Ruuth (2006). For more general RKMs, stepsize-coefficients were obtained e.g. in Gottlieb, Shu & Tadmor (2001), Higueras (2004a, 2005), Ferracina & Spijker (2004, 2005). We note that, in the context of discretizations for hyperbolic problems, the above coefficients c are sometimes called *CFL coefficients*, see e.g. Gottlieb & Shu (1998), Shu (2002).

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

$$u_n = \sum_{j=1}^{k} \alpha_j\, u_{n-j} \;+\; \Delta t \cdot \sum_{j=0}^{k} \beta_j\, f((n-j)\Delta t, u_{n-j}). \tag{1.6}$$

Here α_j, β_j are parameters defining the method, and u_n is computed from $u_{n-k}, \ldots, u_{n-1}$. If $\beta_0 = 0$, the method is called *explicit*.

Monotonicity has been studied, for (1.6), in the sense of the inequality

$$\|u_n\| \;\leq\; \max_{1 \leq j \leq k} \|u_{n-j}\|. \tag{1.7}$$

For explicit LMMs, stepsize-coefficients c, with the property that (1.4), (1.5) guarantee (1.7), were determined, by rewriting the right-hand member of (1.6) as a convex combination of forward Euler steps - see e.g. Shu (1988). Stepsize-coefficients, relevant to more general LMMs, were given e.g. in Gottlieb, Shu & Tadmor (2001), Hundsdorfer & Ruuth (2003), Hundsdorfer, Ruuth & Spiteri (2003).

Clearly, the larger c, the less restrictive is condition (1.5). For any given method, the *maximal stepsize-coefficient* c, with the property that (1.4), (1.5) imply monotonicity, is thus an important and characteristic quantity. When comparing the computational efficiency of different methods, it is natural to take these characteristic quantities into account.

Special attention was paid to the problem of determining, for any given RKM, the corresponding maximal stepsize-coefficient; in Higueras (2004a), Ferracina & Spijker (2004, 2005) conditions were given under which this coefficient equals the famous coefficient $R(A, b)$, which was introduced by Kraaijevanger (1991). For completeness, we note that also much attention was paid to the related - but different - problem of optimizing, over given *classes* of RKMs or LMMs, the *special* stepsize-coefficients obtainable via convex combinations of Euler steps, see e.g. Shu (1988), Shu & Osher (1988), Gottlieb & Shu (1998), Gottlieb (2005), Spiteri & Ruuth (2002), Ruuth (2006).

Both RKMs and LMMs 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. No theory seems to be available in the literature for determining maximal stepsize-coefficients for arbitrary GLMs.

In the present paper, we determine the maximal stepsize-coefficient for a generic numerical process. This result enables us to obtain maximal stepsize-coefficients for arbitrary GLMs, and to gain new insights for numerical methods of which the monotonicity properties were studied earlier.

For completeness we note that, already in Burrage & Butcher (1980), monotonicity of GLMs was studied - but, for seminorms $\|\cdot\|$ generated by (pseudo) inner products, excluding e.g. the seminorm $\|\cdot\|_{TV}$. The present paper deals with arbitrary convex functions $\|\cdot\|$; as a result, our analysis is largely different from the one in the paper just mentioned.

1.2 Scope of the paper

Section 2 contains our theory for the generic numerical process mentioned above. In Section 2.1, we specify GLMs as well as the generic numerical process, and characterize them by a pair of matrices S, T. We also give a formal definition of monotonicity.

In Section 2.2, we introduce in an algebraic way a coefficient $c(S, T)$, which can be viewed as a generalization of Kraaijevanger's coefficient $R(A, b)$. We state typical properties of $c(S, T)$ in Theorem 2.2. This theorem extends earlier results about $R(A, b)$.

In Section 2.3 we state, without proof, the basic results of the paper: Theorems 2.4 and 2.7. These theorems specify situations in which the maximal stepsize-coefficient, for monotonicity of the generic numerical process, is equal to $c(S, T)$. Theorem 2.7 has a wider scope than Theorem 2.4; but the latter theorem has a more simple structure and is of independent interest.

Section 3 contains examples and applications of the theory given in Section 2. In Section 3.1, we focus on arbitrary GLMs. Theorem 3.1 tells us that the maximal stepsize-coefficient for these methods equals $c(S, T)$. Corollaries 3.2, 3.3, respectively, show that $c(S, T)$ is not only relevant to monotonicity, but also to a *discrete maximum principle* and *numerical contractivity* of GLMs.

In Section 3.2, we apply the preceding theory to RKMs, LMMs and a class of multistep-multistage methods (MMMs). We arrive at conclusions supplementing earlier results about these methods. In particular, we find (optimal) second-order and third-order MMMs which we have not seen elsewhere.

In Section 3.3, we apply material of Section 2 to the interesting class of *additive Runge-Kutta methods*. In this way we obtain Theorem 3.6, which answers an open and fundamental question about these methods.

Section 4 contains the proof of Theorems 2.4 and 2.7. In Section 4.1, we prove $c(S, T)$ to be a stepsize-coefficient for the generic numerical process, and in Section 4.2 we prove it to be maximal.

2 A theory for monotonicity

2.1 Monotonicity in a general setting

2.1.1 General linear methods

The general linear method, for solving (1.1), depends on parameters c_j $(1 \le j \le m)$ and parameter matrices $S = (s_{i,j}) \in \mathbb{R}^{m\times l}$, $\quad T = (t_{i,j}) \in \mathbb{R}^{m\times m}$, where $1 \le l \le m$. The method can be written in the following form:

$$\text{(2.1.a)} \qquad y_i \;=\; \sum_{j=1}^{l} s_{ij}\, u_j^{(n-1)} \;+\; \Delta t \cdot \sum_{j=1}^{m} t_{ij}\, f((n-1+c_j)\Delta t,\, y_j) \quad (1 \le i \le m),$$

$$\text{(2.1.b)} \qquad u_i^{(n)} = \; y_{m-l+i} \quad (1 \le i \le l).$$

Here $u_i^{(n-1)}$ are input vectors available at the n-th step of the method, whereas y_i are (intermediate) approximations used for computing the input vectors $u_i^{(n)}$ for the next step; c.f. e.g. Butcher (1987, pp. 336-338), Butcher (2003, p. 358), Hairer, Nørsett & Wanner (1987, p. 390) for related representations of GLMs.

Obviously, the Runge-Kutta method (1.2) is an example of (2.1), with $l = 1,\ m = s+1$, $u_i^{(n)} = u_n \simeq U(n \cdot \Delta t)$ and $s_{i1} = 1,\ \ t_{ij} = a_{ij}$ (for $1 \le j \le s$), $\ t_{ij} = 0$ (for $j = s+1$).

The linear multistep method (1.6) is another example of (2.1), with $l = k,\ m = k+1$ and $u_i^{(n)} = u_{n-l+i}$ $(1 \le i \le l)$, $\ y_i = u_{n-m+i}$ $(1 \le i \le m)$. Method (1.6) can be written in the form (2.1) with $c_j = j - k,\ S = \binom{I}{A},\ T = \binom{O}{B}$, where I denotes the $k \times k$ identity matrix, O the $k \times (k+1)$ zero matrix and $A = (\alpha_k, \ldots, \alpha_1),\ B = (\beta_k, \ldots, \beta_0)$.

We denote the vector space on which the differential equation is defined again by $\mathbb{V}$, and assume $\|\cdot\|$ to be a convex function on $\mathbb{V}$. We will say that method (2.1) is *monotonic* (for the stepsize Δt, function f, and convex function $\|.\|$) if

$$\text{(2.2)} \qquad \|y_i\| \;\le\; \max_{1\le j\le l} \|u_j^{(n-1)}\| \quad (\text{for } 1 \le i \le m),$$

whenever $u_i^{(n-1)}$ and y_i satisfy (2.1.a). Note that the inequalities (2.2) reduce to (1.3) or (1.7), respectively, if method (2.1) stands for (1.2) or (1.6) in the way just indicated.

In the following, we shall assume that the parameters $s_{i,j}$ satisfy

$$\text{(2.3)} \qquad s_{i\,1} + s_{i\,2} + \cdots + s_{i\,l} = 1 \quad (1 \le i \le m).$$

This condition is fulfilled if (2.1) stands, in the above way, for (1.2) or (1.6) (provided $\sum_j \alpha_j = 1$). Moreover, the condition can be seen to be no essential restriction for the general process (2.1): any (preconsistent) GLM can be transformed into an equivalent method satisfying (2.3) - see Butcher (2003, pp. 358-360) for transformations of GLMs.

2.1.2 A generic numerical process, with a simple form

The relations (2.1.a) can be rewritten a bit more compactly. Defining

$$\text{(2.4)} \qquad x_i = u_i^{(n-1)} \ (\text{for } 1 \le i \le l),\quad f_i(v) = f((n-1+c_i)\Delta t, v) \ (\text{for } 1 \le i \le m,\ v \in \mathbb{V}),$$

the relations (2.1.a) reduce to

$$\text{(2.5)} \qquad y_i = \sum_{j=1}^{l} s_{ij}\, x_j \;+\; \Delta t \cdot \sum_{j=1}^{m} t_{ij}\, f_j(y_j) \quad (1 \le i \le m).$$

Furthermore, when $f : \mathbb{R} \times \mathbb{V} \to \mathbb{V}$ satisfies (1.4), then definition (2.4) implies

$$\|v + \tau_0 \, f_i(v)\| \; \le \; \|v\| \quad (\text{for } 1 \le \; i \le \; m, \; \text{ and } v \in \mathbb{V}). \tag{2.6}$$

In the rest of Section 2 we shall deal with (2.5), rather than (2.1.a) - not only because (2.5) has a more simple form, but also because this widens, in a natural way, the range of applications: in Section 3.3 we shall apply our results, to be obtained for the generic process (2.5), to numerical methods which, strictly speaking, are *not* of the form (2.1).

We shall interpret $x_i \in V$ and $y_i \in V$ as *input* and *output vectors*, respectively, of process (2.5). In the general situation (2.3), (2.5), (2.6), we shall focus on the bound

$$\|y_i\| \; \le \; \max_{1 \le j \le l} \|x_j\| \quad (\text{for } 1 \le i \le m), \tag{2.7}$$

and we will say that process (2.5) is *monotonic* (for the stepsize Δt, functions f_i, and convex function $\|.\|$) if (2.7) holds whenever x_i and y_i satisfy (2.5). Clearly, when (2.5) stands for (2.1.a) via the relations (2.4), then monotonicity of (2.5) corresponds to monotonicity as defined above for the GLM.

In Section 2.3 we shall present the basic results of the paper - best possible stepsize conditions which guarantee monotonicity of process (2.5). In formulating these results we need a coefficient which we first introduce in the next Section 2.2.

2.2 The coefficient $c(S, T)$

Throughout this section we denote by $S \in \mathbb{R}^{m \times l}$ and $T \in \mathbb{R}^{m \times m}$ arbitrary matrices, with property (2.3). In Section 2.3 we shall use a coefficient $c(S, T)$ which can be adjoined to S and T. The definition of this coefficient involves the following condition, in which γ denotes a real variable:

$$I + \gamma \, T \text{ is invertible, and } (I + \gamma T)^{-1} \big[\, S \quad \gamma T \, \big] \ge 0. \tag{2.8}$$

Here I denotes the $m \times m$ identity matrix, and $\big[\, S \;\; \gamma T \, \big]$ stands for the $m \times (l + m)$ matrix whose first l columns equal to those of S and whose last m columns equal those of γT. The inequality in (2.8) should be interpreted entry-wise; all inequalities for matrices occurring below are to be interpreted in the same way.

Definition 2.1 (The coefficient $c(S, T)$).
We define $c(S, T) = 0$ *if there is no* $\gamma > 0$ *satisfying* (2.8)*; otherwise*

$$c(S, T) = \sup\{\gamma \;\; : \;\; \gamma \text{ satisfies (2.8)}\}.$$

The above definition may seem to appear out of the blue. The author was led to it, however, by important earlier work of Kraaijevanger (1991) and Higueras (2005). In case (2.1) stands for the Runge-Kutta method (1.2) in the way indicated in Section 2.1.1, then $c(S, T)$ can be seen to reduce to the coefficient introduced, and denoted by $R(A, b)$, in Kraaijevanger(1991); see also Section 3.2.1 of the present paper. In Higueras (2005) the original conditions used by Kraaijevanger for defining his coefficient were simplified to an elegant form which has a resemblance to the above condition (2.8).

By Definition 2.1 we have $c(S, T) \ge 0$. Part (i) of the following Theorem 2.2 makes it relatively easy to see whether $c(S, T)$ is zero or not. If $c(S, T) > 0$, part (ii) of Theorem 2.2 is useful for simplifying the (numerical) computation of $c(S, T)$ - e.g. by using a bisection-type algorithm as in Ferracina & Spijker (2004; Section 4.3), Kraaijevanger (1991, p. 498).

In part (i) of Theorem 2.2 we use, for any given matrix $M = (m_{ij})$, the notation $\mathrm{Inc}(M)$ to denote the *incidence matrix* of M (which has the same dimensions as M), given by

$$\mathrm{Inc}(M) = (\widetilde{m}_{ij}), \quad \text{with } \widetilde{m}_{ij} = 1 \text{ (if } m_{ij} \ne 0) \quad \text{and } \widetilde{m}_{ij} = 0 \text{ (if } m_{ij} = 0).$$

Theorem 2.2 (Properties of $c(S,T)$).
(i) $c(S,T)>0$ *if and only if :* $S\geq 0,\ T\geq 0,\ Inc(T\,S)\leq Inc(S)$ *and* $Inc(T^2)\leq Inc(T)$.
(ii) Suppose $0<\gamma<\infty$ *with* $\gamma\leq c(S,T)$. *Let* $D=diag(\delta_1,\ldots,\delta_m)$, *where* $0\leq\delta_i\leq 1$. *Then* (2.8) *holds, with* T *replaced by* TD.

We note that part (ii) of the theorem is already nontrivial and useful in the simple case where D equals the identity matrix I. The theorem can be viewed as an extension (and improvement) of earlier results in the literature; for related results - about $R(A,b)$ - see Kraaijevanger (1991, Theorem 4.2, Lemma 4.4), Higueras (2004b, Proposition 2.11), Horváth (1998, Theorem 4).

Below we shall prove Theorem 2.2 using Lemma 2.3. We think the lemma is of independent interest: it gives an interesting interpretation of (2.8). We shall use for $x\in\mathbb{R}^n$ (with components ξ_i) and arbitrary $A=(a_{i,j})\in\mathbb{R}^{m\times n}$ the notations $\|x\|_\infty=\max_i|\xi_i|$, $\|A\|_\infty=\max_{x\neq 0}\frac{\|Ax\|_\infty}{\|x\|_\infty}$ and the well known formula $\|A\|_\infty=\max_i\sum_j|a_{i,j}|$.

Lemma 2.3 (Interpretation of (2.8)). *Let* $0<\gamma<\infty$. *Then* (2.8) *holds if and only if* $(I+\gamma T)$ *is invertible, and* $\|(I+\gamma T)^{-1}\big[\ S\ \ \gamma T\ \big]\|_\infty\leq 1$.

Proof of Lemma 2.3.
For any integer $q\geq 1$ we denote by e_q the vector in $\mathbb{R}^q$ with all components equal to 1. We assume that $0<\gamma<\infty$ and $I+\gamma T$ is invertible.

In view of (2.3) we have $Se_l=e_m$. Introducing the matrices

$$P=(p_{i,j})=(I+\gamma T)^{-1}(\gamma T),\ \ Q=(q_{i,j})=(I+\gamma T)^{-1},\ \ R=(r_{i,j})=Q\,S, \tag{2.9}$$

we thus have $R=(I-P)S$ and $\big[\ R\ \ P\ \big]e_{l+m}=R\,e_l+P\,e_m=e_m$. Consequently,

$$\sum_{j=1}^{l}r_{i,j}+\sum_{j=1}^{m}p_{i,j}=1\quad(\text{for } 1\leq i\leq m). \tag{2.10}$$

If (2.8) holds, then all $r_{i,j},\ p_{i,j}$ are nonnegative, so that (2.10) implies $\|\big[\ R\ \ P\ \big]\|_\infty\leq 1$. Conversely, if $\|\big[\ R\ \ P\ \big]\|_\infty\leq 1$, then $\sum_{j=1}^{l}|r_{i,j}|+\sum_{j=1}^{m}|p_{i,j}|\leq\sum_{j=1}^{l}r_{i,j}+\sum_{j=1}^{m}p_{i,j}$. The last inequality proves that all $r_{i,j},\ p_{i,j}$ are nonnegative, so that (2.8) holds. □

We now turn to the proof of Theorem 2.2.

Proof of Theorem 2.2, part (i).
In view of part (ii) of Theorem 2.2 (to be proved below), we have $c(S,T)>0$ if and only if there is a $\gamma_0>0$ such that the matrix $M(\gamma)=(I+\gamma T)^{-1}\big[\ S\ \ \gamma T\ \big]$ is nonnegative for all $\gamma\in[0,\gamma_0]$. Therefore, we can assume with no loss of generality that $S\geq 0,\ T\geq 0$.

We have, for $\gamma>0$ sufficiently small, $M(\gamma)=\{\sum_{k=0}^{\infty}(\gamma T)^{2\,k}\}\big[\ (I-\gamma T)S\ \ (I-\gamma T)\gamma T\ \big]$. It follows that $M(\gamma)\geq 0$ for $\gamma\downarrow 0$ if and only if $\mathrm{Inc}(TS)\leq\mathrm{Inc}(S)$ and $\mathrm{Inc}(TT)\leq\mathrm{Inc}(T)$, which proves (i).

Proof of Theorem 2.2, part (ii).
Let γ_i be any finite values with $0\leq\gamma_i\leq c(S,T)$ and put $\Gamma=\mathrm{diag}(\gamma_1\ldots\gamma_m)$. In order to prove statement (ii), it is enough to assume $c(S,T)>0$ and to show that (2.8) holds with matrix γT replaced throughout by the product $T\,\Gamma$.

Choose any finite γ satisfying (2.8) with $0<\gamma\leq c(S,T)$ and put $E=\mathrm{diag}(\varepsilon_1\ldots\varepsilon_m)$, where $\varepsilon_i=(\gamma-\gamma_i)\,\gamma^{-1}$. In order to prove the invertibility of $I+T\,\Gamma$, we write

$$I+T\,\Gamma=(I+\gamma T)(I-X),\ \text{ with } X=P\,E \text{ and } P \text{ as in (2.9)}.$$

In view of Lemma 2.3, we have $\|P\|_\infty \le 1$, so that $\|X\|_\infty \le \|E\|_\infty$.

First, consider the special case where $c(S,T) < \infty$ and $\gamma_i = c(S,T)$ (for $1 \le i \le m$). Choosing the above γ sufficiently close to $c(S,T)$, we can arrange that $\|E\|_\infty < 1$, which implies that in this special case $I + T\,\Gamma = I + c(S,T)\,T$ is invertible. Using a continuity argument it follows that (2.8) holds with $\gamma = c(S,T)$.

Next, consider again the general case of arbitrary finite γ_i with $0 \le \gamma_i \le c(S,T) \le \infty$. In view of the above, we can choose a positive γ satisfying (2.8), with $\gamma \ge \max_i \gamma_i$. With this γ we have $0 \le X = P\,E \le P$, which implies that the spectral radii of X and P satisfy $\mathrm{spr}(X) \le \mathrm{spr}(P) \le \|P\|_\infty \le 1$. In view of (2.9), the matrix $Q = I - P$ is invertible, so that P has no eigenvalue equal to 1. From the Perron-Frobenius theory (see e.g. Horn & Johnson (1988, p.503)) it follows that $\mathrm{spr}(P) < 1$. Hence $\mathrm{spr}(X) < 1$, so that $I + T\,\Gamma = (I+\gamma T)(I-X)$ has an inverse equal to $(I-X)^{-1}(I+\gamma\,T)^{-1}$. Using $(I+T\,\Gamma)^{-1} = \big(\sum_0^\infty X^k\big)(I+\gamma\,T)^{-1}$, it follows that (2.8) is valid with $\gamma\,T$ replaced by $T\,\Gamma$. $\square$

2.3 Stepsize-coefficients for monotonicity

In this section we give, without proof, the basic results of the paper, Theorems 2.4 and 2.7. Throughout the section, $S \in \mathbb{R}^{m\times l}$ and $T \in \mathbb{R}^{m\times m}$ are again arbitrary matrices satisfying (2.3). We study stepsize conditions $0 < \Delta t \le c \cdot \tau_0$, guaranteeing monotonicity of process (2.5) when f_i satisfy (2.6). The following inequality will turn out to be of crucial importance:

(2.11) $$c \ \le \ c(S,T).$$

Our first result is as follows:

Theorem 2.4 (Monotonicity for arbitrary f_i satisfying (2.6)). *Consider numerical process (2.5). Let τ_0, c be given with $0 < \tau_0 < \infty$, $0 < c \le \infty$. Then each of the following statements (2.12), (2.13) is equivalent to (2.11).*

(2.12) *Condition $0 < \Delta t \le c \cdot \tau_0$ implies monotonicity, whenever $\mathbb{V}$ is a vector space, $\|\cdot\|$ a convex function on $\mathbb{V}$, and arbitrary $f_i : \mathbb{V} \to \mathbb{V}$ satisfy (2.6);*

(2.13) *Condition $0 < \Delta t \le c \cdot \tau_0$ implies monotonicity, when $\mathbb{V} = \mathbb{R}^m$, $\|\cdot\| = \|\cdot\|_\infty$, and arbitrary $f_i : \mathbb{V} \to \mathbb{V}$ satisfy (2.6).*

Clearly, (2.12) is a-priori a stronger statement than (2.13). Accordingly, the essence of Theorem 2.4 is that the (algebraic) property (2.11) implies the (strong) statement (2.12), whereas already the (weaker) statement (2.13) implies (2.11). The theorem highlights the importance of the quantity $c(S,T)$: Theorem 2.4 shows that - with respect to the situations specified in (2.12), (2.13) - *the maximal stepsize-coefficient* c, with the property that condition $0 < \Delta t \le c \cdot \tau_0$ guarantees monotonicity, is equal to $c(S,T)$.

Our second result, Theorem 2.7, deals with important situations not adequately covered by Theorem 2.4: it is often *not* natural to allow - as in Theorem 2.4 - that all functions f_i are different from each other. For instance, if in method (2.1) we have $c_i = c_j$ for some $i \ne j$, or if the differential equation is autonomous, then (2.1) is represented by a process (2.5) with $f_i = f_j$ for some, or all, indices $i \ne j$. In Section 3.3 we will come across another situation where the functions f_i in (2.5) are not independent of each other. In order to cover all of such cases, we consider index sets I_q with $I_q \subset \{1,\ldots,m\}$ (for $1 \le q \le r$), and functions $f_i : \mathbb{V} \to \mathbb{V}$ (for $1 \le i \le m$), such that

(2.14) $I_1, \ldots, I_r$ are nonempty and mutually disjoint, with $I_1 \cup \cdots \cup I_r = \{1,\ldots,m\}$,

(2.15) $f_i = f_j$ whenever i and j belong to the same index set I_q.

According to Theorem 2.4, also when (2.14), (2.15) hold, inequality (2.11) is *sufficient* in order that condition $0 < \Delta t \leq c \cdot \tau_0$ implies monotonicity of numerical process (2.5); but the following counterexample shows that - under the assumptions (2.14), (2.15) - the *maximal stepsize-coefficient* $c = c_{max}$ can be larger than $c(S,T)$.

Example 2.5. *Consider process* (2.5) *with* $l = 1$, $m = 2$ *and* $s_{i,1} = 1$, $t_{i,1} = 2$, $t_{i,2} = -1$ *(for* $i = 1, 2$*). Suppose (2.14), (2.15) with* $r = 1$, $I_1 = \{1, 2\}$*. Since condition* $T \geq 0$ *in part (i) of Theorem 2.2 is violated, we have* $c(S,T) = 0$*. But, with* $f_1 = f_2 = f$*, the process reduces to the (backward Euler) method* $y_1 = x_1 + \Delta t\, f(y_1)$*, which is again of the form* (2.5) *- with* $\tilde{l} = \widetilde{m} = 1$ *and* $c(\tilde{S}, \widetilde{T}) = \infty$*. In line with Theorem 2.4, the maximal stepsize-coefficient* $c = c_{max}$ *such that condition* $0 < \Delta t \leq c \cdot \tau_0$ *implies monotonicity (for the original process with* $m = 2$ *and (2.6), (2.14), (2.15) in force), is equal to* $c_{max} = \infty > c(S,T) = 0$.

Theorem 2.7 will make clear that the inequality $c_{max} > c(S,T)$, in Example 2.5, is an anomaly related to reducibility of the method. We shall use the following formal definition of reducibility and irreducibility, with regard to index sets $I_1, \ldots, I_r$ satisfying (2.14):

Definition 2.6 (Reducibility and irreducibility).
Process (2.5) *is called reducible with respect to* $I_1, \ldots, I_r$*, if indices* i*,* j*,* q *exist with the following properties:* $i \in I_q$*,* $j \in I_q$ *and* $i \neq j$*, whereas the* i*-th and the* j*-th row of the matrix* $\begin{bmatrix} S & T \end{bmatrix}$ *are equal to each other. Process (2.5) is called irreducible with respect to* $I_1, \ldots, I_r$*, if such indices* i*,* j*,* q *do not exist.*

Clearly, if $r < m$ and there is reducibility with respect to $I_1, \ldots, I_r$, then process (2.5) - with f_i satisfying (2.15) - is equivalent to a process (2.5) with a smaller value of m.

Theorem 2.7 (Monotonicity when f_i satisfy (2.6), (2.15)). *Assume (2.14) and irreducibility of process (2.5) with respect to* $I_1, \ldots, I_r$*. Let* τ_0*,* c *be given with* $0 < \tau_0 < \infty$*,* $0 < c \leq \infty$*. Then each of the following statements (2.16), (2.17) is equivalent to (2.11).*

(2.16) *Condition* $0 < \Delta t \leq c \cdot \tau_0$ *implies monotonicity, whenever* $\mathbb{V}$ *is a vector space,* $\|\cdot\|$ *a convex function on* $\mathbb{V}$*, and functions* $f_i : \mathbb{V} \to \mathbb{V}$ *satisfy* (2.6)*,* (2.15)*;*

(2.17) *Condition* $0 < \Delta t \leq c \cdot \tau_0$ *implies monotonicity, when* $\mathbb{V} = \mathbb{R}^m$*,* $\|\cdot\| = \|\cdot\|_\infty$*, and functions* $f_i : \mathbb{V} \to \mathbb{V}$ *satisfy* (2.6)*,* (2.15)*.*

Theorem 2.7 implies that, in the situations specified by (2.16) and (2.17), the *maximal stepsize-coefficient* c, such that condition $0 < \Delta t \leq c \cdot \tau_0$ implies monotonicity, is still equal to $c(S,T)$ - provided there is irreducibility with respect to the relevant index sets.

The following counterexample shows that the dimension of the space $\mathbb{V}$ in statements (2.13) and (2.17) cannot be replaced, in general, by an integer smaller than m.

Example 2.8. *Consider numerical process* (2.5) *with* $l = 1$, $m = 2$ *and* $s_{1,1} = s_{2,1} = 1$, $t_{1,1} = 1/3$, $t_{1,2} = 8/3$, $t_{2,1} = 0$, $t_{2,2} = 1$*. A straightforward calculation yields* $c(S,T) = 3/5$*. On the other hand, it can be proved that propositions* (2.13) *and* (2.17) *would be valid with* $c = 3/2 > c(S,T)$*, if the space* $\mathbb{V} = \mathbb{R}^m = \mathbb{R}^2$ *would be replaced by* $\mathbb{V} = \mathbb{R}^1$.

Theorem 2.4 can formally be viewed as a special case of Theorem 2.7 - the latter theorem with $r = m$ reduces to the former. We have formulated Theorem 2.4 separately in view of its importance and simplicity: it does not need (2.14), (2.15) nor Definition 2.6. Furthermore, in Section 4, where the theorems are proved, we will see that it is convenient to focus first on Theorem 2.4, and to use (arguments used in the proof of) that theorem for proving Theorem 2.7.

3 Examples and applications

3.1 Applications to arbitrary GLMs

In this section we consider method (2.1). We assume (2.3), and give some results which follows readily from the above theory. We focus on stepsize-coefficients c such that

(3.1) Condition $0 < \Delta t \leq c \cdot \tau_0$ implies monotonicity, whenever $\mathbb{V}$ is a vector space, $\|\cdot\|$ a convex function on $\mathbb{V}$, and $f : \mathbb{R} \times \mathbb{V} \to \mathbb{V}$ satisfies (1.4).

In the following theorem, the columns of the matrix $T = (t_{ij})$ are denoted by T_i $(1 \leq i \leq m)$, and the rows of the $m \times (l+m)$ matrix $[\,S \;\; T\,]$ by R_i $(1 \leq i \leq m)$.

Theorem 3.1 (Monotonicity of GLMs). *Consider method (2.1), and given* $\tau_o > 0$.
(i) *Let* $c \leq c(S,T)$. *Then statement (3.1) is valid.*
(ii) *Assume the method is irreducible in the sense that* $R_i \neq R_j$ *as soon as* $i \neq j$, $T_i \neq 0$, $T_j \neq 0$, $c_i = c_j$. *Then, conversely, statement (3.1) implies that* $c \leq c(S,T)$.

Note that the irreducibility assumption in (ii) is trivially fulfilled if the method is *non-confluent* - i. e. if $c_i \neq c_j$ (for all $i \neq j$). Moreover, in case $c_i = c_j$ (for some $i \neq j$), the assumption of irreducibility is *no* strong restriction, because any given method, violating the assumption, is equivalent to a method (with a smaller number of stages) which is irreducible. The theorem highlights the importance of $c(S,T)$ for (irreducible) GLMs: it implies that the *maximal stepsize-coefficient* c, with property (3.1), is equal to $c(S,T)$.

Proof of Theorem 3.1.
In order to prove (i), it is enough to apply Theorem 2.4 to process (2.5), where x_i and f_i are defined via (2.4).

For proving (ii), note that, when $T_k = 0$, the value of the parameter c_k is *irrelevant* to monotonicity of method (2.1). We can thus arrange, without loss of generality, that $c_k \neq c_i$ (for $T_k = 0$ and $i \neq k$). Under the irreducibility assumption in (ii), we thus have

$$R_i \neq R_j \quad \text{whenever} \quad i \neq j,\; c_i = c_j.$$

In order that we can apply Theorem 2.7, we specify index sets I_q by the following requirement: indices i, j belong to the same index set, if and only if $c_i = c_j$. Since process (2.5) is now irreducible in the sense of Definition 2.6, we can apply Theorem 2.7. For proving (ii) it is thus enough to show that (3.1) (for the GLM) implies (2.16) (for process (2.5)).

In order to prove the last implication, we assume (3.1) and suppose x_i, y_i satisfy (2.5) with $0 < \Delta t \leq c \cdot \tau_0$ and functions $f_i : \mathbb{V} \to \mathbb{V}$ satisfying (2.6), (2.15).

We shall show that (2.1) holds with $u_i^{(n-1)} = x_i$ and some function f satisfying (1.4). In defining f we use the notations $t_i = (n-1+c_i)\Delta t$, $\alpha = \min t_i$, $\beta = \max t_i$. We put $f(t_i, v) = f_i(v)$, and extend this function to a function $f : \mathbb{R} \times \mathbb{V} \to \mathbb{V}$ by linear interpolation for $\alpha \leq t \leq \beta$, and by setting $f(t,v) = f(\alpha, v)$ (for $t < \alpha$) and $f(t,v) = f(\beta, v)$ (for $t > \beta$). This function f satisfies (1.4); and (2.1) holds with $u_i^{(n-1)} = x_i$.

Applying (3.1), it follows that (2.2) - and therefore also (2.7) - is fulfilled. This implies (2.16) and concludes the proof. □

Remark 3.2.
The Theorems 2.4, 2.7, used in the above proof, can also be applied to prove a variant of Theorem 3.1 tuned to *autonomous* differential equations. In such a variant, property (3.1) is modified by including that f is independent of t, and the irreducibility condition in (ii) becomes: $R_i \neq R_j$ whenever $i \neq j$, $T_i \neq 0$, $T_j \neq 0$. ■

The subsequent corollaries to Theorem 3.1 involve two properties different in appearance from (2.2).

Property 1 (Discrete maximum principle)
Let $\mathbb{V} = \mathbb{R}^N$, $N \geq 1$. Suppose vectors y_i, $u_i^{(n-1)} \in \mathbb{R}^N$ satisfy (2.1.a). We denote the components of these vectors by y_{pi} and $u_{pi}^{(n-1)}$, respectively $(1 \leq p \leq N)$. The property

$$(3.2) \qquad \min_{1\leq j\leq l} \min_{1\leq q\leq N} u_{qj}^{(n-1)} \leq y_{pi} \leq \max_{1\leq j\leq l} \max_{1\leq q\leq N} u_{qj}^{(n-1)} \quad (\text{for } 1 \leq i \leq m,\ 1 \leq p \leq N)$$

can be interpreted as a *discrete maximum principle*. It is of importance in the solution of partial differential equations (via the method of lines), and can be associated with the absence of undesirable overshoots and undershoots - cf. e.g. Hundsdorfer & Verwer (2003, p. 9, 118). Below we denote the components of $f(t, x) \in \mathbb{R}^N$ by $f_p(t, x)$ $(1 \leq p \leq N)$.

Corollary 3.3 (Discrete maximum principle for GLMs). *Let $f : \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}^N$ and $\tau_0 > 0$ be such that, for $x \in \mathbb{R}^N$ with components ξ_p,*

$$\min_{1\leq q\leq N} \xi_q \leq \xi_p + \tau_0 \cdot f_p(t, x) \leq \max_{1\leq q\leq N} \xi_q \quad (\textit{for } 1 \leq p \leq N).$$

Then (3.2) holds, whenever $u_i^{(n-1)}$ and y_i satisfy (2.1.a) with $0 < \Delta t \leq c(S,T) \cdot \tau_0$.

Proof of Corollary 3.3. Define the convex functions $\|x\|_+ = \max_p \xi_p$ and $\|x\|_- = -\min_p \xi_p$ for $x \in \mathbb{V}$. The assumption in the corollary, about f and τ_0, can be rewritten as

$$\|x + \tau_0\, f(t, x)\|_- \leq \|x\|_- , \quad \|x + \tau_0\, f(t, x)\|_+ \leq \|x\|_+ ,$$

so that (1.4) holds with $\|\cdot\|$ equal to $\|\cdot\|_-$ and $\|\cdot\|_+$, respectively.

Assume (2.1.a) with $0 < \Delta t \leq c(S,T) \cdot \tau_0$. Choosing $c = c(S,T)$ and applying Theorem 3.1, part (i), we get from (3.1) the inequalities $\|y_i\|_+ \leq \max_{1\leq j\leq l} \|u_i^{(n-1)}\|_+$ and $\|y_i\|_- \leq \max_{1\leq j\leq l} \|u_i^{(n-1)}\|_-$, which imply (3.2). □

Property 2 (Contractivity)
Let $\|.\|$ be a convex function on the vector space $\mathbb{V}$. We consider the *contractivity* property

$$(3.3) \qquad \|\tilde{y}_i - y_i\| \leq \max_{1\leq j\leq l} \|\tilde{u}_j^{(n-1)} - u_j^{(n-1)}\| \quad (\text{for } 1 \leq i \leq m),$$

where $u_i^{(n-1)}$, y_i and $\tilde{u}_i^{(n-1)}$, $\tilde{y}_i$ satisfy (2.1.a) with the same stepsize $\Delta t > 0$. Contractivity of numerical processes was studied earlier in various frameworks, cf. e.g. Kraaijevanger (1991), Hairer & Wanner (1996).

Corollary 3.4 (Contractivity for GLMs). *Let $f : \mathbb{R} \times \mathbb{V} \to \mathbb{V}$ and $\tau_0 > 0$ be such that, $\|\tilde{v} - v + \tau_0 \cdot (f(t, \tilde{v}) - f(t, v))\| \leq \|\tilde{v} - v\|$ (for $t \in \mathbb{R}$ and $v, \tilde{v} \in \mathbb{V}$). Then (3.3) holds, whenever $u_i^{(n-1)}$, y_i and $\tilde{u}_i^{(n-1)}$, $\tilde{y}_i$ satisfy (2.1.a) with $0 < \Delta t \leq c(S,T) \cdot \tau_0$.*

Proof of Corollary 3.4. The corollary follows from Theorem 3.1, using arguments similar to those in Burrage & Butcher (1980, p.190): we introduce the auxiliary space $\mathbb{W} = \mathbb{V} \times \mathbb{V}$, and put $\|w\| = \|\tilde{v} - v\|$, $g(t, w) = (f(t, \tilde{v}), f(t, v))$ (for $w = (\tilde{v}, v)$ with $\tilde{v}, v \in \mathbb{V}$). The above assumption, about f and τ_0, implies that $\|w + \tau_0 \cdot g(t, w)\| \leq \|w\|$ (for $w \in \mathbb{W}$).

Let $u_i^{(n-1)}$, y_i and $\tilde{u}_i^{(n-1)}$, $\tilde{y}_i$ satisfy (2.1.a). Defining $U_i = (\tilde{u}_i^{(n-1)}, u_i^{(n-1)})$, $Y_i = (\tilde{y}_i, y_i)$, we have $Y_i = \sum_j s_{ij}\, U_j + \Delta t \cdot \sum_j t_{ij}\, g((n-1+c_j)\Delta t,\ Y_j)$ and $\|Y_i\| = \|\tilde{y}_i - y_i\|$, $\|U_i\| = \|\tilde{u}_i^{(n-1)} - u_i^{(n-1)}\|$. An application of Theorem 3.1, part (i) (to the space $\mathbb{W}$ and the function g) proves the proposition. □

3.2 Applications to RKMs, LMMs and a multistep-multistage method

We illustrate the preceding theory by applying it to some concrete numerical methods.

3.2.1 Runge-Kutta methods

Consider method (1.2). We denote by A_{s+1} the $(s+1)\times s$ matrix with entries a_{ij}, and by A_s the matrix of order s obtained from A_{s+1} by omitting its last row. By E_{s+1} and E_s, respectively, we denote the $(s+1)\times 1$ and the $s\times 1$ matrix with all entries equal to 1. In Section 2.1.1, method (1.2) was already represented as a GLM of form (2.1), with $l=1,\ m=s+1$ and

$$S = E_{s+1},\ \ T = [\,A_{s+1}\ O\,].$$

Monotonicity of this GLM amounts to (1.3). Hence, according to Theorem 3.1, the largest stepsize-coefficient c, such that (3.1) holds for the RKM, is essentially equal to $c(S,T)$. Below we reformulate this result in a more explicit form.

For S, T just defined, it follows easily, similarly as in Higueras (2005), that (2.8) is equivalent to the following condition:

(3.4) $\quad I+\gamma\,A_s$ is invertible, and $A_{s+1}(I+\gamma\,A_s)^{-1}\geq O,\ \ E_{s+1}\geq\gamma\,A_{s+1}(I+\gamma\,A_s)^{-1}\,E_s.$

In view of Definition 2.1, it thus follows - after a simple application of Theorem 2.2, part (i) - that $c(S,T)=\Gamma$, where

(3.5) $\qquad \Gamma=\sup\{\gamma:\ \gamma \text{ satisfies (3.4)}\}\ \ (\text{if } A_{s+1}\geq O), \qquad \text{and}\ \ \Gamma=0\ \ (\text{otherwise}).$

We denote the rows of the $s\times s$ matrix A_s by $r_1,\ldots,r_s$. Applying Theorem 3.1, we immediately arrive at the following two conclusions:

(i) *For method (1.2), statement (3.1) is valid with* $c=\Gamma$, where Γ is given by (3.5).
(ii) *Assume the RKM is irreducible in the sense that* $r_i\neq r_j$, *as soon as* $i\neq j,\ c_i=c_j$. *Then the value* $c=\Gamma$ *in conclusion (i) is best possible, in that (3.1) is not valid with* $c>\Gamma$.

These results imply that, for (irreducible) RKMs, the *maximal stepsize-coefficient c* with property (3.1) equals Γ. Statements (i), (ii) supplement related material in Higueras (2004a, 2005), Spijker & Ferracina (2004, 2005). The irreducibility condition in (ii) is essentially weaker than in these papers, whereas the monotonicity property (3.1) in (i), (ii) is stronger than in (most of) the papers.

Definition (3.5) can be viewed as a smooth variant of similar definitions in the papers just mentioned. For many RKMs, the corresponding value of Γ is explicitly known, because it equals the coefficient introduced - and denoted by $R(A,b)$ - in Kraaijevanger (1991); the equality $\Gamma=R(A,b)$ is an easy consequence of Theorem 2.2, part (ii). For various interesting RKMs, the actual value of $R(A,b)$ was studied and computed in the last mentioned paper - cf. e.g. also Higueras (2004a), Ferracina & Spijker (2004).

Versions of (i), (ii), tuned to *autonomous* differential equations can easily be obtained by applying the variant of Theorem 3.1 mentioned in Remark 3.2. In these versions, the assumption on f in (3.1) includes that f is independent of t, and the irreducibility condition on the RKM becomes: $r_i\neq r_j$ (whenever $i\neq j$).

3.2.2 Linear multistep methods

Consider method (1.6), with $\sum_1^k\alpha_i=1$. In Section 2.1.1, the method was represented as a GLM of form (2.1), with $l=k,\ m=k+1,\ c_j=j-k,\ S=\binom{I}{A},\ T=\binom{O}{B}$ where $A=(\alpha_k,\ldots,\alpha_1),\ B=(\beta_k,\ldots,\beta_0)$. This GLM is irreducible in the sense of Theorem 3.1, because $c_i\neq c_j$ (for $i\neq j$); and its monotonicity amounts to (1.7). Theorem 3.1 thus implies that the largest c, for which the LMM has property (3.1), is equal to $c=c(S,T)$.

In order to find a convenient expression for $c(S,T)$, we consider, for $\gamma > 0$, condition (2.8) with S, T as defined above. One easily sees (using $\sum_1^k \alpha_i = 1$) that (2.8) is equivalent to the requirement that $\beta_0 \geq 0$ and $\alpha_i \geq 0,\ \beta_i \geq 0,\ \alpha_i - \gamma\,\beta_i \geq 0\ \ (1 \leq i \leq k)$. By Definition 2.1, we obtain $c(S,T) = \Gamma$, where

$$(3.6)\qquad \Gamma = \min_{1\leq i\leq k} \alpha_i/\beta_i \ \text{ (if all } \alpha_i,\ \beta_i \text{ are nonegative)}, \qquad \text{and } \ \Gamma = 0 \ \text{ (otherwise)}.$$

Here we use the convention that $a/0 = \infty$, for all $a \geq 0$. In view of the above, we have:

(i) *For method (1.6), statement (3.1) is valid with* $c = \Gamma$*, where* Γ *is given by (3.6).*
(ii) *The value* $c = \Gamma$ *in conclusion (i) is best possible, in that (3.1) is not valid with* $c > \Gamma$*.*

Statements (i), (ii) imply that the *maximal stepsize-coefficient* c with property (3.1) equals Γ. Results similar to (i) were given earlier, see e. g. Shu (1988), Gottlieb, Shu & Tadmor (2001), Hundsdorfer & Ruuth (2003), Hundsdorfer, Ruuth & Spiteri (2003).

As an illustration we consider the following LMM, taken from Shu (1988):

$$(3.7)\qquad u_n = \tfrac{3}{4}\, u_{n-1} + \tfrac{1}{4}\, u_{n-3} + \tfrac{3}{2}\,\Delta t\, f((n-1)\Delta t, u_{n-1}).$$

For this second-order method, we have $\Gamma = 1/2$, so that (3.1) holds with $c = 1/2$. In Gottlieb, Shu & Tadmor (2001), the method was proved to be optimal, in that there exists no explicit second-order method (1.6), with $k = 3$ and $\Gamma > 1/2$. In view of statement (ii), it follows that (3.7) is even *optimal in a wider and more fundamental sense* than stated in the last paper: there exists no explicit second-order method (1.6), with $k = 3$, satisfying (3.1) with $c > 1/2$.

Versions of (i), (ii), for *autonomous* differential equations, follow again from the variant of Theorem 3.1 mentioned in Remark 3.2. No explicit irreducibility assumption, about the LMM, is needed in these versions.

3.2.3 A multistep-multistage method

We illustrate Theorems 2.2 and 3.1 with the method

$$(3.8.a)\qquad v_n = \gamma_1\, u_{n-1} + \gamma_2\, u_{n-2} + \Delta t \cdot [\delta_0\, f_n + \delta_1\, f_{n-1} + \delta_2\, f_{n-2} + \delta_3\, g_n],$$

$$(3.8.b)\qquad u_n = \alpha_1\, u_{n-1} + \alpha_2\, u_{n-2} + \Delta t \cdot [\beta_0\, f_n + \beta_1\, f_{n-1} + \beta_2\, f_{n-2} + \beta_3\, g_n].$$

Here v_n is an intermediate approximation used for computing u_n from u_{n-1}, u_{n-2}, and $f_n = f(n\Delta t,\, u_n)$, $g_n = f((n-\sigma)\Delta t,\, v_n)$. We assume $\alpha_1+\alpha_2 = \gamma_1+\gamma_2 = 1$ and, in order to prevent reducibility, that the coefficient vectors $(\alpha_1, \alpha_2, \beta_0, \beta_1, \beta_2, \beta_3)$ and $(\gamma_1, \gamma_2, \delta_0, \delta_1, \delta_2, \delta_3)$ are different. Method (3.8) can be viewed as a modified LMM or a two-step RKM, cf. e.g. Butcher(2003), Jackiewicz & Tracogna (1995). In Gottlieb, Shu & Tadmor (2001, pp. 102, 103), methods of type (3.8) were explored which are *explicit*, i.e. $\beta_0 = \delta_0 = \delta_3 = 0$.

The general method (3.8) can be written as a GLM (2.1), with $l = 2$, $m = 4$ and with matrices S, T, determined by α_i, β_i, γ_i, δ_i, such that $y_1 = v_n$, $y_2 = u_{n-2}$, $y_3 = u_{n-1}$, $y_4 = u_n$. The monotonicity relation (2.2) reduces to: $\max\{\|v_n\|,\, \|u_n\|\} \leq \max\{\|u_{n-1}\|,\, \|u_{n-2}\|\}$. From Theorem 3.1, we conclude that the largest c, for which method (3.8) satisfies (3.1), is equal to $c(S,T)$. By combining this conclusion with Theorem 2.2 part (i), we arrive after a short calculation at the following proposition:

For method (3.8), a positive c exists with property (3.1), if and only if all coefficients α_i, β_i, γ_i, δ_i *are nonnegative, with*

$$\frac{\alpha_i}{\beta_i + \beta_3\,\gamma_i} > 0,\ \frac{\gamma_i}{\delta_i + \delta_0\,\alpha_i} > 0\ (i = 1, 2)\ \textit{and}\ \frac{\beta_{i-1}}{\beta_3\,\delta_{i-1}} > 0,\ \frac{\delta_i}{\delta_0\,\beta_i} > 0\ (i = 1, 2, 3).$$

Here we use again the convention that $a/0 = \infty > 0$ for $a \geq 0$.

With E_2 we denote the class of all *explicit* methods (3.8) with *order of accuracy 2*. We consider the problem of determining a method in the class which is *optimal*, in that it has property (3.1) with a value c which is maximal in E_2. In view of Theorem 3.1, this problem amounts to finding the maximum of $c(S,T)$ over the class E_2. According to Definition 2.1, this maximum can be computed by performing an optimization, with objective function γ and search variables α_i, β_i, γ_i, δ_i, σ, γ, under the constraints (2.8), supplemented by the order conditions. We performed a numerical search along these lines (using MATLAB), and obtained an (optimal) method of form (3.8), for which we found that the nonzero parameters can be represented (up to 13 decimal digits) as follows:

$$\alpha_1 = 2(\sqrt{2}-1),\ \ \alpha_2 = 3-2\sqrt{2},\ \ \beta_1=\beta_3 = 2-\sqrt{2},\ \ \gamma_1 = 1,\ \ \delta_1 = \sqrt{2}/2,\ \ \sigma = 1-\sqrt{2}/2.$$

This method is of order 2, and it satisfies (3.1) with stepsize-coefficient $c = \sqrt{2}$.

Second-order methods, with a larger stepsize-coefficient, can be found in the class G_2 of *general second-order* methods (3.8). By a numerical search in G_2, similar to the above, we arrived at an (optimal) method with the following nonzero parameters:

$$\alpha_1 = \gamma_1 = 1,\ \ \beta_0 = \beta_1 = \delta_1 = \delta_3 = 1/4,\ \ \beta_3 = 1/2,\ \ \sigma = 1/2.$$

This method is of order 2, and it satisfies (3.1) with stepsize-coefficient $c = 4$. Note that the method is equivalent to two applications of the Trapezoidal Rule (TR), starting from u_{n-1}, and using stepsize $\Delta t/2$; we refer to Lenferink (1991, p. 180) for a related interesting optimality property of the TR.

We also performed a similar numerical search in the class E_3 of all *explicit third-order* methods (3.8). Our search resulted in an (optimal) method for which the nonzero parameters can be represented (up to 13 decimal digits) as follows:

$$\alpha_1 = 6\sqrt{3}-10,\ \ \alpha_2 = 11-6\sqrt{3},\ \ \beta_1 = 4-2\sqrt{3},\ \ \beta_2 = 2-\sqrt{3},\ \ \beta_3 = 6-3\sqrt{3},$$
$$\gamma_1 = 2/3,\ \ \gamma_2 = 1/3,\ \ \delta_1 = (1+\sqrt{3})/3,\ \ \sigma = 1-\sqrt{3}/3.$$

The method is of order 3 and satisfies (3.1) with stepsize-coefficient $c = \sqrt{3}-1 \approx 0.732$. This result extends an earlier numerical search in Gottlieb, Shu & Tadmor (2001, pp.102, 103), where a special method of class E_3 was found which can be implemented - at the cost of an additional function evaluation $\tilde{f}_{n-1}$ - such that (3.1) holds with $c \approx 0.473$.

We did also a numerical search in the class G_3 of *general third-order* methods (3.8). The best method we could find has a coefficient $\delta_0 = 0$, and it satisfies (3.1) with stepsize-coefficient $c \approx 3.233$. But, we did not succeed in finding simple closed-form expressions, similarly as above, for the parameters specifying the method.

We do not go here into details of higher order methods. We just refer to Gottlieb, Shu & Tadmor (2001, p. 103) for an interesting proposition about explicit fourth-order methods, and note that fifth-order methods with $\delta_0 = 0$ exist, satisfying (3.1) with positive c.

3.3 Applications to additive RKMs

Numerical methods of the form

$$\text{(3.9.a)}\quad y_i \;=\; u_{n-1} \;+\; \Delta t\cdot\sum_{j=1}^{s} a_{ij}\, f(y_j) \;+\; \Delta t\cdot\sum_{j=1}^{s}\widehat{a}_{ij}\,\widehat{f}(y_j)\quad (1\le i\le s+1),$$

$$\text{(3.9.b)}\quad u_n \;=\; y_{s+1}$$

have been considered for the efficient solution of equations $\frac{d}{dt}U(t) = f(U(t)) + \widehat{f}(U(t))$, where f and $\widehat{f}$ have different stiffness properties - cf. e.g. Ascher, Ruuth & Spiteri (1997), Kennedy

& Carpenter (2003). The methods are known as *additive Runge-Kutta methods*; and also as *implicit-explicit (IMEX) methods* in case the RKM with coefficients a_{ij} is implicit, and the one with $\widehat{a}_{ij}$ explicit. Furthermore, methods of the form (3.9) have been studied under the name of *perturbed Runge-Kutta methods*, in the context of solving semidiscrete versions of hyperbolic problems. In that situation, (3.9) is equivalent to a *Shu-Osher implementation* of a standard RKM where some a_{ij} are negative, cf. Higueras (2004b, 2005).

In the last mentioned papers, monotonicity - in the sense of (1.3) - was studied for (3.9), under the assumption

$$\|v + \tau_0 f(v)\| \le \|v\|, \quad \|v + \widehat{\tau}_0 \widehat{f}(v)\| \le \|v\| \qquad \text{(for all } v \in \mathbb{V}); \tag{3.10}$$

and a stepsize $(\Delta t)^*$ was presented with the following crucial property:

(3.11) Condition $0 < \Delta t \le (\Delta t)^*$ implies monotonicity of (3.9), whenever $\mathbb{V}$ is a vector space, $\|.\|$ a convex function on $\mathbb{V}$, and $f,\, \widehat{f} : \mathbb{V} \to \mathbb{V}$ satisfy (3.10).

In order to specify $(\Delta t)^*$, we introduce the $(s+1)\times s$ matrices $A = (a_{ij})$, $\widehat{A} = (\widehat{a}_{ij})$ and the $(s+1)\times(s+1)$ matrices $K = [A\; O]$, $\widehat{K} = [\widehat{A}\; O]$. We define $\mathcal{R}$ to be the set of all pairs $(\gamma,\, \widehat{\gamma}) \in \mathbb{R}^2$ such that:

$$I + \gamma\, K + \widehat{\gamma}\, \widehat{K} \text{ is invertible, and } (I + \gamma\, K + \widehat{\gamma}\, \widehat{K})^{-1} \big[\, E \quad \gamma\, K \quad \widehat{\gamma}\, \widehat{K}\, \big] \ge 0. \tag{3.12}$$

Here E stands for the $(s+1)\times 1$ matrix with all entries equal to 1, and the inequality in (3.12) is to be interpreted entry-wise. For given $\tau_0 > 0$, $\widehat{\tau}_0 > 0$, we put

$$\begin{aligned} (\Delta t)^* &= 0 \quad \text{if there is no pair } (\gamma,\, \widehat{\gamma}) \text{ in } \mathcal{R} \text{ with } \gamma\, \tau_0 = \widehat{\gamma}\, \widehat{\tau}_0 > 0; \text{ otherwise} \\ (\Delta t)^* &= \sup\{\tau :\; \tau = \gamma\, \tau_0 = \widehat{\gamma}\, \widehat{\tau}_0 > 0 \text{ with } (\gamma,\, \widehat{\gamma}) \text{ in } \mathcal{R}\}. \end{aligned} \tag{3.13}$$

The following theorem follows immediately from the material in Higueras (2004b):

Theorem 3.5 (Higueras, 2004). *Consider method (3.9), and let $\tau_0 > 0$, $\widehat{\tau}_0 > 0$ be given. Then statement (3.11) is valid, with $(\Delta t)^*$ defined by (3.13).*

In Higueras (2004b), sets $\mathcal{R}$ were computed for a series of important additive RKMs. For any given τ_0, $\widehat{\tau}_0$, these sets allow the immediate calculation of $(\Delta t)^*$ defined by (3.13). One may be tempted to view these sets as important characteristics of the underlying methods, and to compare the efficiency of different methods by taking (the magnitude of) the corresponding sets $\mathcal{R}$ into account. However, if (3.11) would also be valid for some $(\Delta t)^*$ which is *greater* than the one given by (3.13), such a use of these sets might be misleading. The natural question arises of whether the value $(\Delta t)^*$, given in the above theorem, is best possible. We think this fundamental question has not yet been answered in the literature.

By applying the theorems of Section 2, one can recover the above theorem and essentially answer the question just raised (in the positive); we have

Theorem 3.6 (Upper bound for $(\Delta t)^*$ in (3.11)). *Let (3.9) be irreducible, in the sense that the first s rows of the $(s+1)\times 2\,s$ matrix $\big[\, A \quad \widehat{A}\, \big]$ are different from each other. Let $\tau_0 > 0$, $\widehat{\tau}_0 > 0$ be given. If $(\Delta t)^*$ is such that statement (3.11) holds, then $(\Delta t)^*$ cannot exceed the value given in (3.13).*

Proof of Theorems 3.5, 3.6 using the theory of Section 2.
1. Let $\tau_0 > 0$, $\widehat{\tau}_0 > 0$ be given. We shall relate (3.9) to a numerical process of the form (2.5): we put $l = 1,\ m = 2(s+1)$ and $S = (s_{ij}) = \binom{E}{E}$, $T = (t_{ij}) = \begin{pmatrix} K & \delta\widehat{K} \\ K & \delta\widehat{K}\end{pmatrix}$, where $\delta = \tau_0/\widehat{\tau}_0$. We

define index sets $I_1 = \{1, \ldots, s\}$, $I_2 = \{s+1\}$, $I_3 = \{s+2, \ldots, 2s+1\}$, $I_4 = \{2(s+1)\}$.

Let y_i, u_{n-1} satisfy (3.9.a) with f, $\widehat{f}$ as in (3.10). Then $x_1 = u_{n-1}$ and y_i, with

$$y_{s+1+i} = y_i \ \text{ (for } 1 \le i \le s+1), \tag{3.14}$$

can be seen to fulfill (2.5), with some functions f_i satisfying (2.6).

Conversely, let x_i, y_i fulfill (2.5) with f_i satisfying (2.6), (2.15). Then (3.14) holds, so that y_i and $u_{n-1} = x_1$ satisfy (3.9.a) with some f, $\widehat{f}$ as in (3.10).

2. From the above, it follows that (3.11) holds, with $(\Delta t)^* = c \cdot \tau_0$, as soon as (2.12) is in force for process (2.5). According to Theorem 2.4, we can choose $c = c(S,T)$. Hence (3.11) is valid with $(\Delta t)^* = c(S,T) \cdot \tau_0$. A straightforward calculation shows that $c(S,T) \cdot \tau_0$ is equal to the value $(\Delta t)^*$ defined by (3.13). This proves Theorem 3.5.

3. Assume (3.11) holds for some $(\Delta t)^*$. We see, from the above, that property (2.16) must be valid for process (2.5), with $c = (\Delta t)^*/\tau_0$. The irreducibility assumption in Theorem 3.6 implies that process (2.5) is irreducible with respect to $I_1, \ldots, I_4$ - so that Theorem 2.7 can be applied. It follows that the largest value c in (2.16) equals $c(S,T)$, which implies $(\Delta t)^*/\tau_0 \le c(S,T)$. Using again that $c(S,T) \cdot \tau_0$ equals the value defined by (3.13), we arrive at Theorem 3.6. □

Remark 3.7.

The set $\mathcal{R}$, defined above, has the following interesting property:

$$\text{If } (\gamma, \widehat{\gamma}) \in \mathcal{R}, \text{ then } (\beta, \widehat{\beta}) \in \mathcal{R} \text{ whenever } 0 \le \beta \le \gamma,\ 0 \le \widehat{\beta} \le \widehat{\gamma}. \tag{3.15}$$

This can be proved by defining S, T similarly as in the above proof and applying Theorem 2.2, part (ii). For related material, see Higueras (2004b). ■

4 Proof of Theorems 2.4 and 2.7

4.1 Sufficiency of the inequality (2.11)

In order to write (2.5) and similar relations more concisely, we introduce some notations relevant to the vector space $\mathbb{V}$. For any integer $n \ge 1$ and vectors $x_1, \ldots, x_n \in \mathbb{V}$, we denote the vector in $\mathbb{V}^n$ with components x_i by

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

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

We combine the vectors x_i and y_i, occurring in (2.5), into the 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 F, from $\mathbb{V}^m$ to $\mathbb{V}^m$, by $F(y) = [f_i(y_i)] \in \mathbb{V}^m$ for $y = [y_i] \in \mathbb{V}^m$. With these notations, the relations (2.5) can be written as an equality in $\mathbb{V}^m$:

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

The following simple Lemma 4.1 will be quite useful: in the present section for proving that (2.11) implies (2.12) and (2.16); in the next section for proving that (2.13) and (2.17) imply (2.11). In the lemma we use the notations (2.9), and we relate (4.1) - with f_i satisfying (2.6), (2.15) - to the conditions

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

$$z_i = z_j, \ \text{ whenever: } \ y_i = y_j \text{ and } i, j \text{ belong to the same index set } I_q. \tag{4.2.b}$$

Lemma 4.1 (Reformulation of (4.1) with f_i satisfying (2.6), (2.15)).
Let $\tau_0 > 0,\ \Delta t > 0,\ \gamma = \Delta t/\tau_0$, *and* $I + \gamma T$ *be invertible. Assume* (2.14), *and let* $x = [x_i] \in \mathbb{V}^l$ *and* $y = [y_i] \in \mathbb{V}^m$ *be given. Then (4.1) holds for some* $f_i : \mathbb{V} \to \mathbb{V}$ *satisfying (2.6), (2.15), if and only if there exists a vector* $z = [z_i] \in \mathbb{V}^m$ *such that (4.2) holds.*

Proof of Lemma 4.1.
Assume (4.1), (2.6), (2.15), and define $z_i = y_i + \tau_0\, f_i(y_i),\ z = [z_i] \in \mathbb{V}^m$. Applying (2.6), (2.9) and the equality $y = \mathbf{S}\, x + \gamma\, \mathbf{T}\,(-y + z)$, we arrive at (4.2.a); and by applying (2.15) we obtain (4.2.b).

Conversely, suppose (4.2) holds. For $i \in I_q$ we define $f_i : \mathbb{V} \to \mathbb{V}$ by:

$$f_i(v) = (1/\tau_0)(-y_j + z_j) \ \ (\text{if } v = y_j,\ j \in I_q), \ \ \text{and} \ \ f_i(v) = 0 \ \ (\text{otherwise}) .$$

Using again (2.9), it follows easily that (4.1) holds with f_i satisfying (2.6), (2.15). $\square$

Proof that inequality (2.11) *implies* (2.12) *and* (2.16).
Assume $0 < \tau_0 < \infty,\ 0 < c \leq c(S,T)$. We shall prove (2.12) - which is enough, because (2.16) follows from (2.12).

Let $\mathbb{V}$, $\|.\|$, f_i be as assumed in (2.12), and suppose x_i, y_i satisfy (2.5) with $0 < \Delta t \leq c{\cdot}\tau_0$. We put $\gamma = \Delta t/\tau_0$, so that $0 < \gamma \leq c(S,T)$. From Theorem 2.2, part (ii), it thus follows that γ satisfies (2.8), so that $I + \gamma\, T$ is invertible.

Since (4.1) holds with f_i satisfying (2.6), we can apply Lemma 4.1, with the trivial index sets $I_q = \{q\}$ for $1 \leq q \leq m$. It follows that (4.2.a) holds; so that, with the notations (2.9),

$$y_i \ = \ \sum_{j=1}^{l} r_{ij}\, x_j \ + \ \sum_{j=1}^{m} p_{ij}\, z_j, \quad \|z_i\| \leq \|y_i\| \quad (1 \leq i \leq m).$$

In view of (2.8), (2.9), we have $r_{ij} \geq 0$, $p_{ij} \geq 0$; and similarly as in the proof of Lemma 2.3 we have (2.10), i.e. $\sum_j r_{ij} + \sum_j p_{ij} = 1$.

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 z_i. Using the convexity of the function $\|\cdot\|$, it thus follows that $[\|y_i\|] \leq R\,[\|x_i\|] + P\,[\|z_i\|] \leq Q\,S\,[\|x_i\|] + (I - Q)\,[\|y_i\|]$, i.e. $Q\,[\|y_i\|] \leq Q\,S\,[\|x_i\|]$. Multiplying the last inequality by the matrix $Q^{-1} = I + \gamma\, T$ (which is nonnegative, in view of Theorem 2.2, part (i)), we get

$$\|y_i\| \ \leq \ \sum_{j=1}^{l} s_{ij}\, \|x_j\| \quad (1 \leq i \leq m). \tag{4.3}$$

Using (2.3) and the nonnegativity of s_{ij} (cf. Theorem 2.2, part (i)), we obtain (2.7). $\square$

4.2 Necessity of the inequality (2.11)

In proving that (2.13) and (2.17) imply (2.11), we shall use the following lemma:

Lemma 4.2 (Invertibility of $I + \gamma T$). *Let* $\tau_0 > 0$, $\Delta t > 0$ *be given, and* $\gamma = \Delta t/\tau_0$. *Assume* $\mathbb{V} = \mathbb{R}^m$, $\|\cdot\| = \|\cdot\|_\infty$, *and let* $I_1, \ldots, I_r$ *be index sets as in* (2.14). *Suppose process* (2.5) *is monotonic for all functions* f_i *satisfying* (2.6), (2.15). *Then* $I + \gamma\, T$ *is invertible.*

Proof of Lemma 4.2.
Suppose $(I + \gamma\, T)\, \eta = 0$ for some vector $\eta = [\eta_i] \in \mathbb{R}^m$. Define $f_i(v) = -(1/\tau_0)\, v$ (for all $v \in \mathbb{V}$). Then (2.5) is satisfied by the vectors $x_i = 0$ $(1 \leq i \leq l)$ and $y_i = \eta_i\, e_m$ $(1 \leq i \leq m)$, where e_m is the vector in $\mathbb{R}^m$ with all components equal to 1. Since the functions f_i satisfy

(2.6), (2.15), it follows that $|\eta_i| = \|y_i\| \leq \max_j \|x_j\| = 0$, so that $\eta = 0$. $\square$

Proof that (2.13) implies (2.11).
Let τ_0, c be given with $0 < \tau_0 < \infty$, $0 < c \leq \infty$, and assume (2.3), (2.13). We choose $\Delta t = \gamma\,\tau_0$, where γ is an arbitrary finite value with $0 < \gamma \leq c$; and we define $\mathbb{V} = \mathbb{R}^m$.

An application of Lemma 4.2, with the trivial index sets $I_q = \{q\}$ (for $1 \leq q \leq m$), shows that the matrix $I + \gamma T$ is invertible. We thus can use the notations (2.9), and apply Lemma 4.1 (again with the trivial index sets), so as to conclude that, for any $x \in \mathbb{V}^l$ and $y,\, z \in \mathbb{V}^m$, the relations

$$y \;=\; \boldsymbol{R}x \;+\; \boldsymbol{P}\,z, \quad \text{with} \;\; \|z_j\|_\infty \leq \|y_j\|_\infty \;(1 \leq j \leq m), \tag{4.4}$$

imply that

$$\|y_j\|_\infty \;\leq\; \max_{1\leq k\leq l} \|x_k\|_\infty \quad (\text{for } 1 \leq j \leq m). \tag{4.5}$$

Below we shall use this implication for proving that

$$\|\,[\,R \;\; P\,]\,\|_\infty \leq 1. \tag{4.6}$$

By Lemma 2.3, inequality (4.6) implies that $\gamma \leq c(S,T)$. Since γ was chosen arbitrarily in $(0, c]$, the last inequality implies (2.11).

In proving (4.6), we shall use the notation $\mathrm{sgn}(\alpha) = 1$ (for $\alpha \geq 0$), $\mathrm{sgn}(\alpha) = -1$ (for $\alpha < 0$). We put $x_{ij} = \mathrm{sgn}(r_{ij})$, $z_{ij} = \mathrm{sgn}(p_{ij})$, where r_{ij}, p_{ij} are the entries of R and P, and we consider the special vectors x_j, $z_j \in \mathbb{V} = \mathbb{R}^m$ with components x_{ij} and z_{ij}, respectively $(1 \leq i \leq m)$. We define $x \in \mathbb{V}^l$ and $y,\, z \in \mathbb{V}^m$ by $x = [x_j]$, $z = [z_j]$, $y = [y_j] = \boldsymbol{R}x + \boldsymbol{P}z$, and denote the components of the vectors y_j by y_{ij} $(1 \leq i \leq m)$.

The relations (4.4) hold, with these special vectors, because

$$\|y_j\|_\infty \geq y_{jj} = \sum_k r_{jk}\, x_{jk} \;+\; \sum_k p_{jk}\, z_{jk} \;=\; \sum_k |r_{jk}| \;+\; \sum_k |p_{jk}|,$$

and, in view of (2.10),

$$\|z_j\|_\infty = 1 \;=\; \sum_k r_{jk} \;+\; \sum_k p_{jk} \;\leq\; \sum_k |r_{jk}| \;+\; \sum_k |p_{jk}|.$$

Since (4.4) implies (4.5), we obtain $\sum_k |r_{jk}| \;+\; \sum_k |p_{jk}| \;\leq\; 1$, i.e. (4.6). $\square$

Proof that (2.17) implies (2.11).
1. Assume (2.3), (2.14) and irreducibility with respect to the index sets under consideration. Let τ_0, c be given with $0 < \tau_0 < \infty$, $0 < c \leq \infty$, and assume (2.17). We choose $\Delta t = \gamma\,\tau_0$, where γ is an arbitrary finite value with $0 < \gamma \leq c$; and we define $\mathbb{V} = \mathbb{R}^m$.

Similarly as in the last proof above, $I + \gamma T$ is invertible, and for any $x \in \mathbb{V}^l$ and $y,\, z \in \mathbb{V}^m$, the implication

$$(4.4) \;\; \& \;\; (4.2.b) \;\; \Rightarrow \;\; (4.5)$$

is valid. For completing the present proof, it is again enough to deduce (from the last implication) that (4.6) holds.

Below we shall denote by x_j, y_j, z_j the special vectors in $\mathbb{V} = \mathbb{R}^m$, with components x_{ij}, y_{ij}, z_{ij}, used in the above proof that (2.13) implies (2.11).

2. First, assume that $y_i \neq y_j$ whenever indices $i \neq j$ belong to the same index set. Clearly, under this assumption, (4.2.b) holds. Furthermore, just as in the proof above, we have (4.4), so that (4.5) is valid. This implies again (4.6).

3. Next, assume the last assumption is violated, i.e. there are indices s, q belonging to the same index set, with $s \neq q$ and $y_s = y_q$. In this situation, we modify (only) the q-th component of our special vectors x_j, y_j, z_j into $\tilde{x}_{qj} = \xi_j$, $\tilde{y}_{qj} = \eta_j$ and $\tilde{z}_{qj} = \zeta_j$, respectively. Here $\xi = [\xi_j] \in \mathbb{R}^l$, $\eta = [\eta_j] \in \mathbb{R}^m$, $\zeta = [\zeta_j] \in \mathbb{R}^m$ are vectors such that:

(4.7.a) $\qquad \eta = R\,\xi + P\,\zeta$, with $\|\xi\|_\infty \leq 1$, $\|\zeta\|_\infty \leq 1$,

(4.7.b) $\qquad \eta_i \neq \eta_j$ whenever $i \neq j$ belong to the same index set.

We will show that such vectors exist, below in part 4 of the proof. In order to distinguish the original vectors x_j, y_j, z_j from the modified ones, we denote the latter by $\tilde{x}_j$, $\tilde{y}_j$, $\tilde{z}_j$, respectively.

Clearly, for $\tilde{x} = [\tilde{x}_j]$, $\tilde{y} = [\tilde{y}_j]$, $\tilde{z} = [\tilde{z}_j]$, the equality $\tilde{y} = \boldsymbol{R}\tilde{x} + \boldsymbol{P}\tilde{z}$ holds. Furthermore, $\tilde{y}_i \neq \tilde{y}_j$, whenever $i \neq j$ belong to the same index set. Finally (using $|y_{ss}| = |y_{sq}| \leq \|\tilde{y}_q\|_\infty$) we see that $\|\tilde{z}_j\|_\infty \leq 1 \leq \|\tilde{y}_j\|_\infty$ $(1 \leq j \leq m)$. The modified vectors thus satisfy (4.4), (4.2.b). Consequently, they satisfy (4.5), which implies $\sum_k |r_{jk}| + \sum_k |p_{jk}| \leq 1$ (for all $j \neq q$). By interchanging the role of s and q, we see that the last inequality is also valid for all $j \neq s$. Hence, (4.6) holds.

4. In view of the irreducibility assumption, the polynomials $f_i(\lambda) = \sum_{k=1}^{l} s_{ik}\,\lambda^k + \gamma \cdot \sum_{k=1}^{m} t_{ik}\,\lambda^{l+k}$ satisfy $f_i \neq f_j$, if $i \neq j$ belong to the same index set. It follows that, for sufficiently small $\lambda > 0$, the vectors $\xi = [\xi_j]$, $\eta = [\eta_j]$, $\zeta = [\zeta_j]$, with $\xi_k = \lambda^k$ $(1 \leq k \leq l)$, $\eta_k = f_k(\lambda)$ $(1 \leq k \leq m)$, $\zeta_k = \lambda^{l+k} + f_k(\lambda)$ $(1 \leq k \leq m)$, satisfy (4.7). $\square$

Acknowledgement The author thanks Dr Karel in 't Hout and Dr Jaap van de Griend for helpful discussions related to the MATLAB calculations in Section 3.2.3.

References

[1] Ascher U.M., Ruuth S.J., Spiteri R.J. (1997): *Implicit-explicit Runge-Kutta methods for time-dependent partial differential equations*, Appl. Numer. Math. **25**, 151-167.

[2] Burrage K., Butcher J.C. (1980): *Non-linear stability of a general class of differential equation methods*, BIT **20**, 185-203.

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

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

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

[6] 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.

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

[8] Gottlieb S. (2005): *On high order strong stability preserving Runge-Kutta and multi step time discretizations*, Journ. Scientif. Computing **25**, 105-128.

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

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

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

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

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

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

[15] Higueras I. (2004b): *Strong stability for additive Runge-Kutta methods*, Report Dep. Mat. Inform, Univ. Publ. Navarra.

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

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

[18] Horvát Z. (1998): *Positivity of Runge-Kutta and diagonally split Runge-Kutta methods*, Appl. Numer. Math. **28**, 309-326. Eighth Conference on the numerical treatment of differential equations (Alexisbad, 1997).

[19] Hundsdorfer W.H., 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.

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

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

[22] Jackiewicz Z., Tracogna S. (1995): *A general class of two-step Runge-Kutta methods for ordinary differential equations*, SIAM J. Numer. Anal. **32**, 1390-1427. ,

[23] Kennedy C.A., Carpenter M.H. (2003): *Additive Runge-Kutta schemes for convection-diffusion-reaction equations*, Appl. Numer. Math. **44**, 139-181.

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

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

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

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

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

[29] Shu C.-W. (2002): *A survey of strong stability preserving high-order time discretizations*, Collected lectures on the preservation of stability under discretization, D. Estep and S. Tavener eds., SIAM, Philadelphia, pp. 51-65.

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

[31] 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.