STEPSIZE CONDITIONS FOR GENERAL MONOTONICITY IN NUMERICAL INITIAL VALUE PROBLEMS

M. N. SPIJKER†


Key words. initial value problem, method of lines, ordinary differential equation, general linear method, total-variation-diminishing, strong-stability-preserving, monotonicity

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

DOI. 10.1137/060661739

1. Introduction.

1.1. Maximal stepsize-coefficients for monotonicity. Consider an initial value problem for a system of ordinary differential equations of type

\[
\frac{d}{dt} U(t) = f(t, U(t)) \quad (t \geq 0), \quad U(0) = u_0.
\]

We shall deal with step-by-step-methods for finding numerical approximations \( u_n \) to the true solution values \( U(n\Delta t) \), where \( \Delta t \) 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

\[
y_i = u_{n-1} + \Delta t \cdot \sum_{j=1}^{s} a_{ij} f((n-1+c_j)\Delta t, y_j) \quad (1 \leq i \leq s+1),
\]

\[
u_n = y_{s+1}.
\]

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, \( V \) stands for the vector space on which the differential equation is defined, and \( \| \cdot \| \) denotes a convex function on \( V \) (i.e., \( \| \lambda v + (1-\lambda)w \| \leq \lambda \|v\| + (1- \ldots

*Received by the editors June 1, 2006; accepted for publication (in revised form) November 17, 2006; published electronically May 22, 2007.

†Mathematical Institute, Leiden University, P. O. Box 9512, NL-2300-RA Leiden, The Netherlands (spieler@math.leidenuniv.nl).
\( \lambda \|w\| \) for \( 0 \leq \lambda \leq 1 \) and \( v, w \in V \). Much attention has been paid in the literature to the property
\[
(1.3) \quad \|y_i\| \leq \|u_{n-1}\| \quad (\text{for } 1 \leq i \leq s + 1).
\]
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 \textit{monotonicity} or \textit{strong stability}; it is of particular importance in situations where (1.1) results from \textit{(method of lines) semidiscretizations} of time-dependent partial differential equations. Choices for \( \|\cdot\| \), which occur in that context, include, e.g., the \textit{supremum norm} \( \|x\| = \|x\|_{\infty} = \sup_i |\xi_i| \) and the \textit{total variation seminorm} \( \|x\| = \|x\|_{TV} = \sum_i |\xi_{i+1} - \xi_i| \) (for vectors \( x \) with components \( \xi_i \)). 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 \textit{total variation diminishing}; cf., e.g., Harten [13], Shu [28], Shu and Osher [30], LeVeque [26], and Hundsdorfer and Verwer [21]. We note that, for practical calculations, special importance has been attached to the inequality \( \|y_i\| \leq \|u_{n-1}\| \) being fulfilled for \textit{all} \( i \) with \( 1 \leq i \leq s + 1 \) (rather than just for \( i = s + 1 \)); see, e.g., Shu [29] and Gottlieb [8].

Conditions on \( \Delta t \) which guarantee (1.3) were given in the literature, mainly for \textit{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
\[
(1.4) \quad \|v + \tau_0 f(t, v)\| \leq \|v\| \quad (\text{for } t \in \mathbb{R}, v \in V).
\]
Assumption (1.4) means that the forward Euler method is monotonic with stepsize \( \tau_0 \). 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, \textit{stepsize-coefficients} \( c \) were determined such that monotonicity, in the sense of (1.3), is present for all \( \Delta t \) with
\[
(1.5) \quad 0 < \Delta t \leq c \cdot \tau_0.
\]
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 and Osher [30], Spiteri and Ruuth [31], and Ruuth [27]. For more general RKMs, stepsize-coefficients were obtained, e.g., in Gottlieb, Shu, and Tadmor [10], Higueras [14, 16], and Ferracina and Spijker [6, 7]. We note that, in the context of discretizations for hyperbolic problems, the above coefficients \( c \) are sometimes called \textit{CFL coefficients}; see, e.g., Gottlieb and Shu [9] and Shu [29].

The linear multistep method (LMM) for computing \( u_n \) can be written in the form
\[
(1.6) \quad u_n = \sum_{j=1}^{k} \alpha_j u_{n-j} + \Delta t \sum_{j=0}^{k} \beta_j f((n - j)\Delta t, u_{n-j}).
\]
Here \( \alpha_j, \beta_j \) 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 \textit{explicit}.

Monotonicity has been studied for (1.6) in the sense of the inequality
\[
(1.7) \quad \|u_n\| \leq \max_{1 \leq j \leq k} \|u_{n-j}\|.
\]
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 [28]. Stepsize-coefficients, relevant to more general LMMs, were given, e.g., in Gottlieb, Shu, and Tadmor [10], Hundsdorfer and Ruuth [19], and Hundsdorfer, Ruuth, and Spiteri [20].

Clearly, the larger $c$ is, 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 [14] and Ferracina and Spijker [6, 7] conditions were given under which this coefficient equals the famous coefficient $R(A, b)$, which was introduced by Kraaijevanger [24]. For completeness, we note also that 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 [28], Shu and Osher [30], Gottlieb and Shu [9], Gottlieb [8], Spiteri and Ruuth [31], and Ruuth [27].

Both RKMs and LMMs are examples of methods belonging to the important and very large class of general linear methods (GLMs), introduced by Butcher [3], and studied extensively in the literature; see, e.g., Butcher [4, 5], Hairer, Norsett, and Wanner [12], Hairer and Wanner [11], and the references therein. No theory seems to be available in the literature for determining maximal stepsize-coefficients for arbitrary GLMs.

In this 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 and Butcher [2], monotonicity of GLMs was studied, but, for seminorms $\| \cdot \|$ generated by (pseudo) inner products, excluding, e.g., the seminorm $\| \cdot \|_{TV}$. This 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 3.3 and 3.4, respec-
tively, 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 from section 2 to the interesting class of additive \textit{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 GLM for solving (1.1) depends on parameters \( c_j \) \((1 \leq j \leq m)\) and parameter matrices \( S = (s_{ij}) \in \mathbb{R}^{m \times l} \), \( T = (t_{ij}) \in \mathbb{R}^{m \times m} \), where \( 1 \leq l \leq m \). The method can be written in the following form:

\[
\begin{align*}
(2.1.a) \quad y_i &= \sum_{j=1}^{l} s_{ij} u_j^{(n-1)} + \Delta t \sum_{j=1}^{m} t_{ij} f((n - 1 + c_j)\Delta t, y_j) \quad (1 \leq i \leq m), \\
(2.1.b) \quad u_i^{(n)} &= y_{m-i+i} \quad (1 \leq i \leq l).
\end{align*}
\]

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; cf., e.g., Butcher \cite[pp. 336–338]{Butcher} and \cite[p. 358]{Hairer} and Wanner \cite[p. 390]{Wanner} for related representations of GLMs.

Obviously, the RKM (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 \( s_{1i} = 1, t_{ij} = a_{ij} \) (for \( 1 \leq j \leq s \)), \( t_{ij} = 0 \) (for \( j = s + 1 \)).

The LMM (1.6) is another example of (2.1), with \( l = k, m = k + 1, \) and \( u_i^{(n)} = u_{n-l+i} \) \((1 \leq i \leq l), \) \( y_i = u_{n-m+i} \) \((1 \leq i \leq m)\). Method (1.6) can be written in the form (2.1) with \( c_j = j - k, S = \begin{pmatrix} A \end{pmatrix}, T = \begin{pmatrix} B \end{pmatrix} \), 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 \( V \), and assume \( \| \cdot \| \) to be a convex function on \( V \). We will say that method (2.1) is \textit{monotonic} (for the stepsize \( \Delta t \), function \( f \), and convex function \( \| \cdot \| \)) if

\[
\| y_i \| \leq \max_{1 \leq j \leq l} \| u_j^{(n-1)} \| \quad \text{(for } 1 \leq i \leq m),
\]

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 \( s_{ij} \) satisfy

\[
(2.3) \quad s_{i1} + s_{i2} + \cdots + s_{il} = 1 \quad (1 \leq i \leq m).
\]

This condition is fulfilled if (2.1) stands, in the above way, for (1.2) or (1.6) (provided
\[\sum \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 \[5, pp. 358–360\] 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

\[x_i = u_i^{(n-1)} \quad (for \ 1 \leq i \leq l), \quad f_i(v) = f((n - 1 + c_i)\Delta t, v) \quad (for \ 1 \leq i \leq m, \ v \in V),\]

the relations \(2.1.a\) reduce to

\[y_i = \sum_{j=1}^{l} s_{ij} x_j + \Delta t \cdot \sum_{j=1}^{m} t_{ij} f_j(y_j) \quad (1 \leq i \leq m).\]

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

\[\|v + \tau_0 f_i(v)\| \leq \|v\| \quad (for \ 1 \leq i \leq m \ and \ v \in V).\]

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 the process \(2.5\). In the general situation \(2.3\), \(2.5\), \(2.6\), we shall focus on the bound

\[\|y_i\| \leq \max_{1 \leq j \leq l} \|x_j\| \quad (for \ 1 \leq i \leq m),\]

and we will say that process \(2.5\) is monotonic (for the stepsize \(\Delta t\), functions \(f_i\), and convex function \(\|\cdot\|\)) 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, the best possible stepsize conditions which guarantee monotonicity of process \(2.5\). In formulating these results we need a coefficient which we first introduce in 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 \(\gamma\) denotes a real variable:

\[I + \gamma T \text{ is invertible and } (I + \gamma T)^{-1} \begin{bmatrix} S & \gamma T \end{bmatrix} \geq 0.\]

Here \(I\) denotes the \(m \times m\) identity matrix, and \(\begin{bmatrix} S & \gamma T \end{bmatrix}\) stands for the \(m \times (l + m)\) matrix whose first \(l\) columns equal those of \(S\) and whose last \(m\) columns equal those of \(\gamma T\). The inequality in \(2.8\) should be interpreted entrywise; 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 previous definition may seem to appear out of the blue. The author was led to it, however, by important earlier work of Kraaijevanger [24] and Higueras [16]. In case (2.1) stands for the RKM (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 [24]; see also section 3.2.1 of this paper. In Higueras [16] the original conditions used by Kraaijevanger for defining his coefficient were simplified to an elegant form which has a resemblance to condition (2.8).

By Definition 2.1 we have $c(S,T) \geq 0$. Part (i) of 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 and Spijker [6, section 4.3] and Kraaijevanger [24, p. 498].

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

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

**Theorem 2.2** (properties of $c(S,T)$).

(i) $c(S,T) > 0$ if and only if $S \geq 0$, $T \geq 0$, $\text{Inc}(T S) \leq \text{Inc}(S)$, and $\text{Inc}(T^2) \leq \text{Inc}(T)$.

(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 concerning $R(A,b)$, see Kraaijevanger [24, Theorem 4.2, Lemma 4.4], Higueras [15, Proposition 2.11], and Horváth [18, 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 $x_i$) and arbitrary $A = (a_{ij}) \in \mathbb{R}^{m \times n}$ the notations $\|x\|_\infty = \max_i |x_i|$, $\|A\|_\infty = \max_{x \neq 0} \frac{\|Ax\|_\infty}{\|x\|_\infty}$, and the well known formula $\|A\|_\infty = \max_i \sum_j |a_{ij}|$.

**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 $\|((I + \gamma T)^{-1} [S \gamma T])\|_\infty \leq 1$.

**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 $S e_i = e_m$. Introducing the matrices

\begin{equation}
P = (p_{i,j}) = (I + \gamma T)^{-1}(\gamma T), \quad Q = (q_{i,j}) = (I + \gamma T)^{-1}, \quad R = (r_{i,j}) = QS,
\end{equation}

we thus have $R = (I - P)S$ and $[R \, \, P] e_l + m = Re_l + P e_m = e_m$. Consequently,

\begin{equation}
\sum_{j=1}^l r_{i,j} + \sum_{j=1}^m p_{i,j} = 1 \quad (\text{for } 1 \leq i \leq m).
\end{equation}

If (2.8) holds, then all $r_{i,j}$, $p_{i,j}$ are nonnegative, so that (2.10) implies $\|R \, \, P\|_\infty \leq 1$. Conversely, if $\|R \, \, P\|_\infty \leq 1$, 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. \( \square \)
We now turn to the proof of Theorem 2.2.

Proof of Theorem 2.2(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 $M(\gamma) = (I + \gamma T)^{-1} [S \gamma T]$ is nonnegative for all $\gamma \in [0, \gamma_0]$. Therefore, we can assume without loss of generality that $S \geq 0$, $T \geq 0$.

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

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

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

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

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

First, consider the special case where $c(S, T) < \infty$ and $\gamma_i = c(S, T)$ (for $1 \leq i \leq m$). Choosing the above $\gamma$ 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 $\gamma_i$ with $0 \leq \gamma_i \leq c(S, T) \leq \infty$. In view of the above, we can choose a positive $\gamma$ satisfying (2.8), with $\gamma \geq \max_i \gamma_i$. With this $\gamma$ we have $0 \leq X = PE \leq P$, which implies that the spectral radii of $X$ and $P$ satisfy $\text{spr}(X) \leq \text{spr}(P) \leq \|P\|_{\infty} \leq 1$. In view of (2.9), the matrix $Q = I - P$ is invertible, so that $P$ has no eigenvalue equal to 1. Applying the Perron–Frobenius theory (see, e.g., Horn and Johnson [17, p. 503]), it follows that $\text{spr}(P) < 1$. Hence $\text{spr}(X) < 1$, so that $I + T \Gamma = (I + \gamma T)(I - X)$ has an inverse equal to $(I - X)^{-1}(I + \gamma T)^{-1}$. Using $(I + T \Gamma)^{-1} = (\sum_{k=0}^{\infty} X^k)(I + \gamma T)^{-1}$, it follows that (2.8) is valid with $\gamma T$ replaced by $T \Gamma$.

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 stepsizes conditions $0 < \Delta t \leq c \cdot \tau_0$, guaranteeing monotonicity of process (2.5) when $f_i$ satisfies (2.6). The following inequality will be of crucial importance:

$$c \leq c(S, T). \tag{2.11}$$

Our first result is as follows.

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

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

(2.13) Condition $0 < \Delta t \leq c \cdot \tau_0$ implies monotonicity when $V = \mathbb{R}^{m}$, $\| \cdot \| = \| \cdot \|_{\infty}$, and arbitrary $f_i : V \to V$ 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 $0 < \Delta t \leq c \cdot \tau_0$ 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 $i \neq j$, 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 $i \neq j$. 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 $1 \leq q \leq r$) and functions $f_i: \mathbb{V} \rightarrow \mathbb{V}$ (for $1 \leq i \leq m$), 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 $c = c_{\text{max}}$ 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 Theorem 2.2(i) 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 $\tilde{l} = \tilde{m} = 1$ and $c(\tilde{S}, \tilde{T}) = \infty$. In line with Theorem 2.4, the maximal stepsize-coefficient $c = c_{\text{max}}$ 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 $c_{\text{max}} = \infty > c(S, T) = 0$.

Theorem 2.7 will make clear that the inequality $c_{\text{max}} > c(S, T)$, 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 $[S \ T]$ 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 $\tau_0, c$ 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 $\mathcal{V}$ is a vector space,
$\| \cdot \|$ a convex function on $\mathcal{V}$, and functions $f_i : \mathcal{V} \to \mathcal{V}$ satisfy (2.6), (2.15);

(2.17) Condition $0 < \Delta t \leq c \cdot \tau_0$ implies monotonicity whenever $\mathcal{V} = \mathbb{R}^m$, $\| \cdot \| = \| \cdot \|_\infty$, and functions $f_i : \mathcal{V} \to \mathcal{V}$ 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 $\mathcal{V}$ 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 $\mathcal{V} = \mathbb{R}^m = \mathbb{R}^2$ would be replaced by $\mathcal{V} = \mathbb{R}^1$.

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) or 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 follow 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 $\mathcal{V}$ is a vector space,
$\| \cdot \|$ a convex function on $\mathcal{V}$, and functions $f : \mathbb{R} \times \mathcal{V} \to \mathcal{V}$ satisfy (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 $\tau_0 > 0$.

(i) Let $c \leq c(S, T)$. Then statement (3.1) is valid.

(ii) Assume the method is irreducible in the sense that $R_i \neq R_j$ for all $i, j$ with $i \neq j$, $T_i \neq 0$, $T_j \neq 0$, $c_i = c_j$. 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 nonconfluent, 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 to 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 \( f_i : \mathbb{V} \rightarrow \mathbb{V} \) 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 \( f(t, v) = f_i(v) \), and extend this function to a function \( f : \mathbb{R} \times \mathbb{V} \rightarrow \mathbb{V} \) 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. Theorems 2.4 and 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 \( u_{qi}^{(n-1)} \), respectively (\( 1 \leq p \leq N \)). The property

\[ \min_{1 \leq j \leq l} \min_{1 \leq q \leq N} u_{qi}^{(n-1)} \leq y_{pi} \leq \max_{1 \leq j \leq l} \max_{1 \leq q \leq N} u_{qi}^{(n-1)} \quad (\text{for } 1 \leq i \leq m, \ 1 \leq p \leq N) \]

(3.2)

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; see, e.g., Hundsdorfer and Verwer [21, pp. 9 and 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 \( f : \mathbb{R} \times \mathbb{R}^N \rightarrow \mathbb{R}^N \) and \( \tau_0 > 0 \) be such that, for \( x \in \mathbb{R}^N \) with components \( \xi_p \),

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

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 \( \tau_0 \), 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(i), we get from (3.1) the inequalities \(\|y_i\|_+ \leq \max_{1 \leq j \leq 1} \|u_j^{(n-1)}\|_+\) and
\(\|y_i\|_- \leq \max_{1 \leq j \leq 1} \|u_j^{(n-1)}\|_-\), which imply (3.2).

Property 2 (contractivity). Let \(\|\cdot\|\) be a convex function on the vector space \(\mathbb{V}\).
We consider the contractivity property
\[
(3.3) \quad \|\tilde{y}_i - y_i\| \leq \max_{1 \leq j \leq l} \|\tilde{u}_j^{(n-1)} - u_j^{(n-1)}\| \quad (\text{for } 1 \leq i \leq m),
\]
where \(u_j^{(n-1)}, y_i\) and \(\tilde{u}_j^{(n-1)}, \tilde{y}_i\) satisfy (2.1.a) with the same stepsize \(\Delta t > 0\). Contraction
of numerical processes were studied earlier in various frameworks; cf., e.g.,

**Corollary 3.4** (contractivity for GLMs). Let \(f : \mathbb{R} \times \mathbb{V} \to \mathbb{V}\) and \(\tau_0 > 0\) be such
that \(\|\tilde{v} - v + \tau_0 \cdot (f(t, \tilde{v}) - f(t, v))\| \leq \|\tilde{v} - v\|\) \((\text{for } t \in \mathbb{R} \text{ and } v, \tilde{v} \in \mathbb{V})\). Then (3.3)
holds, whenever \(u_j^{(n-1)}, y_i\) and \(\tilde{u}_j^{(n-1)}, \tilde{y}_i\) 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 and Butcher [2, p. 190]: We introduce the auxiliary space \(\mathbb{W} = \mathbb{V} \times \mathbb{V}\) and put \(\|w\| = \|\tilde{v} - v\|, \; g(t, w) = (f(t, \tilde{v}), f(t, v))\) \((\text{for } w = (\tilde{v}, v) \text{ with } \tilde{v}, v \in \mathbb{V})\). The above assumption, about \(f\) and \(\tau_0\), implies that
\(\|w + \tau_0 \cdot g(t, w)\| \leq \|w\|\) \((\text{for } w \in \mathbb{W})\).

Let \(u_i^{(n-1)}, \gamma_i, \tilde{u}_i^{(n-1)}\), \(\tilde{y}_i\) satisfy (2.1.a). Defining \(U_i = (u_i^{(n-1)}, \tilde{u}_i^{(n-1)}), \; Y_i = (\tilde{y}_i, y_i)\), we have \(Y_i = \sum s_j U_j + \Delta t \cdot \sum t_{ij} g((n - 1 + c_j)\Delta t, Y_j)\) 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(i) \((\text{to the space } \mathbb{W} \text{ and the function } g)\) proves the proposition.

**3.2. Applications to RKMs, LMMs, and a MMM.** 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
\(S = E_{s+1}, \quad T = [A_{s+1} \ O]\).

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 Higuera [16], that (2.8) is
equivalent to the following condition:
\[
(3.4) \quad I + \gamma A_s \text{ is invertible and } A_{s+1}(I + \gamma A_s)^{-1} \geq O, \quad E_{s+1} \geq \gamma A_{s+1}(I + \gamma A_s)^{-1} E_s.
\]
In view of Definition 2.1, it thus follows, after a simple application of Theorem 2.2(i),
that \(c(S, T) = \Gamma, \) where
\[
(3.5) \quad \Gamma = \sup \{ \gamma : \gamma \text{ satisfies (3.4)} \} \quad (\text{if } A_{s+1} \geq O) \quad \text{and} \quad \Gamma = 0 \quad (\text{otherwise}).
\]
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 $\Gamma$ is given by (3.5).

(ii) Assume the RKM is irreducible in the sense that $r_i \neq r_j$ for all $i, j$ with $i \neq j$, $c_i = c_j$. Then the value $c = \Gamma$ in conclusion (i) is optimal, 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 $\Gamma$. Statements (i) and (ii) supplement related material in Higueras [14, 16] and Ferracina and Spijker [6, 7]. The irreducibility condition in (ii) is essentially weaker than in these papers, whereas the monotonicity property (3.1) in (i) and (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 $\Gamma$ is explicitly known, because it equals the coefficient introduced, and denoted by $R(A, b)$, in Kraaijevanger [24]; the equality $\Gamma = R(A, b)$ is an easy consequence of Theorem 2.2(ii). For various interesting RKMs, the actual value of $R(A, b)$ was studied and computed in the last mentioned paper; see also Higueras [14] and Ferracina and Spijker [6].

Versions of (i) and (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 $\sum_{i=1}^{k} \alpha_i = 1$. In section 2.1.1, the method was represented as a GLM of form (2.1), with $l = k$, $m = k + 1$, $c_j = j - k$, $S = (\hat{A})$, $T = (\hat{O})$, where $A = (\alpha_0, \ldots, \alpha_1)$, $B = (\beta_0, \ldots, \beta_0)$. This GLM is irreducible in the sense of Theorem 3.1, because $c_{ij} \neq c_{ij}$ (for $i \neq j$). 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 $\sum_{i=1}^{k} \alpha_i = 1$) 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

$$\Gamma = \min_{1 \leq i \leq k} \frac{\alpha_i}{\beta_i} \quad \text{if all} \quad \alpha_i, \beta_i \quad \text{are nonnegative} \quad \text{and} \quad \Gamma = 0 \quad \text{(otherwise)}.$$  

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

(i) For method (1.6), statement (3.1) is valid with $c = \Gamma$, where $\Gamma$ is given by (3.6).

(ii) The value $c = \Gamma$ in conclusion (i) is optimal, in that (3.1) is not valid with $c > \Gamma$.

Statements (i) and (ii) imply that the maximal stepsize-coefficient $c$ with property (3.1) equals $\Gamma$. Results similar to (i) were given earlier; see, e.g., Shu [28], Gottlieb, Shu, and Tadmor [10], Hundsdorfer and Ruuth [19], and Hundsdorfer, Ruuth, and Spiteri [20].

As an illustration we consider the following LMM, taken from Shu [28]:

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

$$u_n = \frac{3}{4} u_{n-1} + \frac{1}{4} u_{n-3} + \frac{3}{2} \Delta t \hat{f}((n-1)\Delta t, u_{n-1}).$$
For this second order method, we have $\Gamma = 1/2$, so that (3.1) holds with $c = 1/2$. In Gottlieb, Shu, and Tadmor [10], 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) and (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

\begin{align}
(3.8.a) & \quad u_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) & \quad 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].
\end{align}

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 [5] and Jackiewicz and Tracogna [22]. In Gottlieb, Shu, and Tadmor [10, pp. 102 and 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 $\alpha_i, \beta_i, \gamma_i, \delta_i$, 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 $\max\{\|v_n\|, \|u_n\|\} \leq \max\{\|u_{n-1}\|, \|u_{n-2}\|\}$. Applying 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(i), we arrive after a short calculation at the following proposition.

**Proposition 3.1.** For method (3.8), a positive $c$ exists with property (3.1), if and only if all coefficients $\alpha_i, \beta_i, \gamma_i, \delta_i$ are nonnegative, with $\frac{\alpha_0}{\alpha_1+\beta_0} > 0$, $\frac{\gamma_1}{\delta_1+\gamma_0} > 0$ ($i = 1, 2, 3$).

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 $\gamma$ and search variables $\alpha_i, \beta_i, \gamma_i, \delta_i, \sigma, \gamma_i$, 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:

\begin{align}
&\alpha_1 = 2(\sqrt{2}-1), \quad \alpha_2 = 3-2\sqrt{2}, \quad \beta_1 = \beta_3 = 2-\sqrt{2}, \quad \gamma_1 = 1, \quad \delta_1 = \sqrt{2}/2, \quad \sigma = 1-\sqrt{2}/2.
\end{align}

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:

\begin{align}
&\alpha_1 = \gamma_1 = 1, \quad \beta_0 = \beta_1 = \delta_1 = \delta_3 = 1/4, \quad \beta_3 = 1/2, \quad \sigma = 1/2.
\end{align}
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 [25, 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:

$$\alpha_1 = 6\sqrt{3} - 10, \quad \alpha_2 = 11 - 6\sqrt{3}, \quad \beta_1 = 4 - 2\sqrt{3}, \quad \beta_2 = 2 - \sqrt{3}, \quad \beta_3 = 6 - 3\sqrt{3},$$

$$\gamma_1 = 2/3, \quad \gamma_2 = 1/3, \quad \delta_1 = (1 + \sqrt{3})/3, \quad \sigma = 1 - \sqrt{3}/3.$$

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, and Tadmor [10, pp. 102 and 103], where a special method of class $E_3$ was found which can be implemented, at the cost of an additional function evaluation $\hat{f}_{n-1}$, such that (3.1) holds with $c \approx 0.473$.

We also did 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 into details here of higher order methods. We just refer to Gottlieb, Shu, and Tadmor [10, 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

\begin{align}
(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} \hat{a}_{ij} \hat{f}(y_j) \quad (1 \leq i \leq s + 1), \\
(3.9.b) \quad u_n &= y_{s+1}
\end{align}

have been considered for the efficient solution of equations $\frac{d}{dt} U(t) = f(U(t)) + \hat{f}(U(t))$, where $f$ and $\hat{f}$ have different stiffness properties; cf., e.g., Ascher, Ruuth, and Spiteri [1] and Kennedy and Carpenter [23]. 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 $\hat{a}_{ij}$ 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. Higuera [15, 16].

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

\begin{align}
(3.10) \quad &\|v + \tau_0 f(v)\| \leq \|v\|, \quad \|v + \hat{\tau}_0 \hat{f}(v)\| \leq \|v\| \quad (\text{for all } v \in \mathbb{V}); \\
\end{align}

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

\begin{align}
(3.11) \quad \text{Condition } 0 < \Delta t \leq (\Delta t)^* \quad \text{implies monotonicity of (3.9) whenever } \mathbb{V} \text{ is a vector space, } \|\cdot\| \text{ a convex function on } \mathbb{V}, \text{ and functions } f, \hat{f} : \mathbb{V} \rightarrow \mathbb{V} \text{ satisfy (3.10).}
\end{align}
In order to specify \((\Delta t)^*\), we introduce the \((s + 1) \times s\) matrices \(A = (a_{ij}), \hat{A} = (\hat{a}_{ij})\), and the \((s + 1) \times (s + 1)\) matrices \(K = [A\ O], \hat{K} = [\hat{A}\ O]\). We define \(R\) to be the set of all pairs \((\gamma, \hat{\gamma})\in\mathbb{R}^2\) such that

\[
(I + \gamma K + \hat{\gamma} \hat{K})\text{ is invertible and } (I + \gamma K + \hat{\gamma} \hat{K})^{-1}[E\ \gamma K + \hat{\gamma} \hat{K}] \geq 0.
\]

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 entrywise. For a given \(\tau_0 > 0, \hat{\tau}_0 > 0\), we put

\[
(3.13) \quad (\Delta t)^* = 0 \text{ if there is no pair } (\gamma, \hat{\gamma}) \in R \text{ with } \gamma \tau_0 = \hat{\gamma} \hat{\tau}_0 > 0; \text{ otherwise (3.13) holds, then (\Delta t)^* cannot exceed the value given in (3.13).}
\]

The following theorem follows immediately from the material in Higuera [15].

**Theorem 3.5** (Higuera, 2006). Consider method (3.9) and let \(\tau_0 > 0, \hat{\tau}_0 > 0\) be given. Then statement (3.11) is valid, with \((\Delta t)^*\) defined by (3.13).

In Higuera [15], sets \(R\) were computed for a series of important additive RKMs. For any given \(\tau_0, \hat{\tau}_0\), 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 \(R\) 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 the following theorem.

**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 \((s + 1) \times 2s\) matrix \([A\ \hat{A}]\) are different from each other. Let \(\tau_0 > 0, \hat{\tau}_0 > 0\) 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 and 3.6 using the theory of section 2.** (i) Let \(\tau_0 > 0, \hat{\tau}_0 > 0\) be given. We shall relate (3.9) to a numerical process of the form (2.5): we put \(l = 1, m = 2(s + 1), S = (s_{ij}) = (\hat{E})^T, T = (t_{ij}) = (\hat{K} + \hat{K})\), where \(\delta = \tau_0 / \hat{\tau}_0\). 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 \(f, \hat{f}\) as in (3.10). Then \(x_i = u_{n-1}\) and \(y_i\), with

\[
y_{s+1+i} = y_i \quad (\text{for } 1 \leq i \leq s + 1),
\]

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 \(u_{n-1}\) fulfill (3.9.a) with some \(f_i\), \(\hat{f}\) as in (3.10).

(ii) In view of 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.

(iii) Assume (3.11) holds for some \((\Delta t)^*\). We see now 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 $(\Delta t)^*/\tau_0 \leq c(S, T)$. 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 $R$ has the following interesting property:

$$\text{(3.15) If } (\gamma, \hat{\gamma}) \in R, \text{ then } (\beta, \hat{\beta}) \in R \text{ whenever } 0 \leq \beta \leq \gamma, \; 0 \leq \hat{\beta} \leq \hat{\gamma}. \text{ This can be proved by defining } S, T \text{ similarly as in the above proof and applying Theorem 2.2(ii).} \text{ For related material, see Higueras [15].}$$

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 $V$. For any integer $n \geq 1$ and vectors $x_1, \ldots, x_n \in V$, we denote the vector in $V^n$ with components $x_i$ by

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

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

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

$$y = Sx + \Delta t \cdot TF(y).$$

The simple Lemma 4.1 will be quite useful, in the present section for proving that (2.11) implies (2.12) and (2.16), and 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

(4.2a) $y = Rx + Pz$, with $\|z_i\| \leq \|y_i\|$ ($1 \leq i \leq m$),

(4.2b) $z_i = z_j$, whenever $y_i = y_j$ and $i, j$ belong to the same index set $I_q$.

**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 $x = [x_i] \in V^l$ and $y = [y_i] \in V^m$ be given. Then (4.1) holds for some $f_i : V \to V$ satisfying (2.6), (2.15), if and only if there exists a vector $z = [z_i] \in V^m$ such that (4.2) holds.

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

Conversely, suppose (4.2a) and (4.2b) hold. For $i \in I_q$ we define $f_i : V \to V$ 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).

**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 \( V, \| \cdot \|, f_i \) 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 \( 0 < \gamma \leq c(S,T) \). Applying Theorem 2.2(ii), it thus follows that \( \gamma \) 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, so as to conclude that, for any 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 \( r_{ij} \geq 0, p_{ij} \geq 0 \); 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(i)), we get

\[
\| y_i \| \leq \sum_{j=1}^{l} s_{ij} \| x_j \| \quad (1 \leq i \leq m).
\]

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

**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 \( V = \mathbb{R}^m, \| \cdot \| = \| \cdot \|_{\infty} \), 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 \( f_j(v) = -(1/\tau_0) v \) (for all \( v \in V \)). 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 \( \eta = 0 \).

**Proof that (2.13) implies (2.11).** Let \( \tau_0, c \) 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 \( \gamma \) is an arbitrary finite value with \( 0 < \gamma \leq c \); and we define \( V = \mathbb{R}^m \).

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 \( x \in V \) and \( y, z \in V^m \), the relations

\[
y = Rx + Pz, \quad \| z_j \|_{\infty} \leq \| y_j \|_{\infty} \quad (1 \leq j \leq m),
\]

imply that

\[
\| y_j \|_{\infty} \leq \max_{1 \leq k \leq l} \| x_k \|_{\infty} \quad (1 \leq j \leq m).
\]
Below we shall use this implication for proving that

\[(4.6) \quad \| [R \ P] \|_\infty \leq 1.\]

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

In proving (4.6), we shall use the notation \( \text{sgn}(\alpha) = 1 \) (for \( \alpha \geq 0 \)), \( \text{sgn}(\alpha) = -1 \) (for \( \alpha < 0 \)). We put \( x_{ij} = \text{sgn}(r_{ij}) \), \( z_{ij} = \text{sgn}(p_{ij}) \), 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 \( x = [x_j], y, z \in \mathbb{V}^m \) by \( x = [x_j], z = [z_j], y = [y_j] = Rx + Pz \), 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

\[
\|y_{ij}\| \geq y_{ij} = \sum_k r_{jk} x_{jk} + \sum_k p_{jk} z_{jk} = \sum_k |r_{jk}| + \sum_k |p_{jk}|,
\]

and, in view of (2.10),

\[
\|z_{ij}\| = 1 = \sum_k r_{jk} + \sum_k p_{jk} \leq \sum_k |r_{jk}| + \sum_k |p_{jk}|.
\]

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

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

Similar to the proof above, \( I + \gamma T \) is invertible, and for any \( x \in \mathbb{V}_\ell \) and \( y, z \in \mathbb{V}^m \), the implication

\[(4.4) \quad \text{and} \quad (4.2.b) \Rightarrow (4.5)\]

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 previous proof that (2.13) implies (2.11).

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

(iii) 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 \( \xi = [\xi_j] \in \mathbb{R}^\ell, \eta = [\eta_j] \in \mathbb{R}^m \), and \( \zeta = [\zeta_j] \in \mathbb{R}^m \) are vectors such that

\[
(4.7.a) \quad \eta = R\xi + P\zeta, \quad \text{with} \quad \|\xi\|_\infty \leq 1, \|\zeta\|_\infty \leq 1,
(4.7.b) \quad \eta_i \neq \eta_j \text{ whenever } i \neq j \text{ belong to the same index set}.
\]

We will show that such vectors exist in part (iv) 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 $\tilde{y} = R\tilde{x} + P\tilde{z}$ holds. Furthermore, $\tilde{y}_i \neq \tilde{y}_j$, whenever $i \neq j$ belong to the same index set. Finally (using $|y_{ss}| = |y_{sq}| \leq ||y_q||_{\infty}$), 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.

(iv) In view of the irreducibility assumption, the polynomials $f_i(\lambda) = \sum_{k=1}^l s_{ik} \lambda^k + \gamma \cdot \sum_{k=1}^l t_{ik} \lambda^{k+1}$ 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 $\xi_k = \lambda^k$ ($1 \leq k \leq l$), $\eta_k = f_k(\lambda)$ ($1 \leq k \leq m$), $\zeta_k = \lambda^{k+1} + f_k(\lambda)$ ($1 \leq k \leq m$), satisfy (4.7). □

Acknowledgments. The author thanks Dr. Karel in’t Hout and Dr. Jaap van de Grieth for helpful discussions related to the MATLAB calculations in section 3.2.3.

REFERENCES


