SIAM J. NUMER. ANAL.
Vol. 42, No. 3, pp. 1073–1093

STEPSIZE RESTRICTIONS FOR THE TOTAL-VARIATION-DIMINISHING PROPERTY IN GENERAL RUNGE–KUTTA METHODS*

L. FERRACINA† AND M. N. SPIJKER†

Abstract. Much attention has been paid in the literature to total-variation-diminishing (TVD) numerical processes in the solution of nonlinear hyperbolic differential equations. For special Runge–Kutta methods, conditions on the stepsize were derived that are sufficient for the TVD property; see, e.g., Shu and Osher [*J. Comput. Phys.*, 77 (1988), pp. 439–471] and Gottlieb and Shu [*Math. Comp.*, 67 (1998), pp. 73–85]. Various basic questions are still open regarding the following issues: 1. the extension of the above conditions to more general Runge–Kutta methods; 2. simple restrictions on the stepsize which are not only sufficient but at the same time necessary for the TVD property; and 3. the determination of optimal Runge–Kutta methods with the TVD property.

In this paper we propose a theory by means of which we are able to clarify the above questions. Moreover, by applying our theory, we settle analogous questions regarding the related strong-stability-preserving (SSP) property (see, e.g., Gottlieb, Shu, and Tadmor [*SIAM Rev.*, 43 (2001), pp. 89–112] and Shu [*Collected Lectures on the Preservation of Stability under Discretization*, D. Estep and S. Tavener, eds., SIAM, Philadelphia, 2002]). Our theory can be viewed as a variant to a theory of Kraaijevanger [*BIT*, 31 (1991), pp. 482–528] on the contractivity of Runge–Kutta methods.

Key words. initial value problem, conservation law, method of lines, Runge–Kutta formula, total-variation-diminishing (TVD), strong-stability-preserving (SSP), monotonicity

AMS subject classifications. Primary, 65M20; Secondary, 65L05, 65L06

DOI. 10.1137/S0036142902415584

1. Introduction.

1.1. The purpose of the paper. In this paper we shall address some natural questions arising in the numerical solution of certain partial differential equations (PDEs). In order to formulate these questions, we consider an initial value problem for a system of ordinary differential equations (ODEs) of the form

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

We assume that (1.1) results from an application of the method of lines to a Cauchy problem for a PDE of the form

$$\frac{\partial}{\partial t}u(x,t) + \frac{\partial}{\partial x}f(u(x,t)) = 0 \quad (t \geq 0, \quad -\infty < x < \infty).$$

Here f stands for a given (possibly nonlinear) scalar function, so that the PDE is a simple instance of a conservation law; cf., e.g., Kröner (1997) and LeVeque (2002).

The solution $U(t)$ to (1.1) stands for a (time dependent) vector in $\mathbb{R}^\infty = \{y : y = (\ldots, \eta_{-1}, \eta_0, \eta_1, \ldots)$ with $\eta_j \in \mathbb{R}$ for $j = 0, \pm 1, \pm 2, \ldots\}$. The components $U_j(t)$ of $U(t)$ are to approximate the desired true solution values $u(j\Delta x, t)$ (or cell averages

*Received by the editors October 4, 2002; accepted for publication (in revised form) September 26, 2003; published electronically July 29, 2004.
http://www.siam.org/journals/sinum/42-3/41558.html

†Mathematical Institute, Leiden University, Niels Bohrweg 1, 2333 CA Leiden, The Netherlands (ferra@math.leidenuniv.nl, spijker@math.leidenuniv.nl). The research of the first author was partially supported by a Padova University grant, Dipartimento di Matematica Pura e Applicata, Università di Padova.

thereof); here Δx denotes a (positive) mesh-width. Furthermore, F stands for a function from $\mathbb{R}^\infty$ into $\mathbb{R}^\infty$; it depends on the given function f as well as on the process of semidiscretization being used. Finally, $u_0 \in \mathbb{R}^\infty$ depends on the initial data of the original Cauchy problem.

Any Runge–Kutta method, when applied to problem (1.1), yields approximations u_n to $U(n\Delta t)$, where $\Delta t > 0$ denotes the time step and $n = 1, 2, 3, \ldots$. Since $\frac{d}{dt}U(t) = F(U(t))$ stands for a semidiscrete version of a conservation law, it is desirable that the (fully discrete) process be *total-variation-diminishing (TVD)* in the sense that

$$\|u_n\|_{TV} \le \|u_{n-1}\|_{TV}; \tag{1.2}$$

here the function $\|.\|_{TV}$ is defined by

$$\|y\|_{TV} = \sum_{j=-\infty}^{+\infty} |\eta_j - \eta_{j-1}| \quad (\text{for } y \in \mathbb{R}^\infty \text{ with components } \eta_j).$$

For an explanation of the importance of the TVD property, particularly in the numerical solution of nonlinear conservation laws, see, e.g., Harten (1983), Laney (1998), Toro (1999), LeVeque (2002), and Hundsdorfer and Verwer (2003).

By Shu and Osher (1988) (see also, e.g., Gottlieb, Shu, and Tadmor (2001) and Shu (2002)) a simple but very useful approach was described for obtaining (high order) Runge–Kutta methods leading to TVD numerical processes. They considered explicit m-stage Runge–Kutta methods, written in the special form

$$\begin{aligned} y_1 &= u_{n-1}, \\ y_i &= \sum_{j=1}^{i-1} [\lambda_{ij}\, y_j + \Delta t \cdot \mu_{ij} F(y_j)] \quad (2 \le i \le m+1), \\ u_n &= y_{m+1}. \end{aligned} \tag{1.3}$$

Here λ_{ij}, μ_{ij} are real coefficients specifying the Runge–Kutta method, and y_i are intermediate vectors in $\mathbb{R}^\infty$, depending on u_{n-1}, used for computing u_n (for $n = 1, 2, 3, \ldots$). Theorem 1.1 will state one of the conclusions formulated in the three papers just mentioned. It applies to the situation where the semidiscretization of the conservation law has been carried out in such a manner that the forward Euler method, applied to $\frac{d}{dt}U(t) = F(U(t))$, yields a fully discrete process which is TVD, when the stepsize Δt is suitably restricted, i.e.,

$$\|v + \Delta t\, F(v)\|_{TV} \le \|v\|_{TV} \quad (\text{whenever } 0 < \Delta t \le \tau_0 \text{ and } v \in \mathbb{R}^\infty). \tag{1.4}$$

Furthermore, in the theorem it is assumed that

$$\lambda_{i1} + \lambda_{i2} + \cdots + \lambda_{i,i-1} = 1 \quad (2 \le i \le m+1), \tag{1.5a}$$

$$\lambda_{ij} \ge 0, \quad \mu_{ij} \ge 0 \quad (1 \le j < i \le m+1), \tag{1.5b}$$

and the following notation is used:

$$c_{ij} = \lambda_{ij}/\mu_{ij} \ \ (\text{for } \mu_{ij} \ne 0), \quad c_{ij} = \infty \ \ (\text{for } \mu_{ij} = 0), \tag{1.6a}$$

$$c = \min_{i,j} c_{ij}. \tag{1.6b}$$

THEOREM 1.1 (Shu and Osher). *Assume* (1.5)*, and let* c *be defined by* (1.6)*. Suppose* (1.4) *holds, and*

$$0 < \Delta t \le c \cdot \tau_0. \tag{1.7}$$

Then process (1.3) *is TVD; i.e.,* (1.2) *holds whenever* u_n *is computed from* u_{n-1} *according to* (1.3).

It was remarked, notably in Shu and Osher (1988) and Gottlieb, Shu, and Tadmor (2001), that, under the assumptions (1.5), (1.6), the above theorem can be generalized. Let $\mathbb{V}$ be an arbitrary linear subspace of $\mathbb{R}^\infty$ and let $\|.\|$ denote any corresponding seminorm (i.e., $\|u+v\| \le \|u\| + \|v\|$ and $\|\lambda v\| = |\lambda| \cdot \|v\|$ for all $\lambda \in \mathbb{R}$ and $u, v \in \mathbb{V}$). A straightforward generalized version of Theorem 1.1 says that if $F : \mathbb{V} \to \mathbb{V}$ and

$$\|v + \Delta t F(v)\| \le \|v\| \quad (\text{whenever } 0 < \Delta t \le \tau_0 \text{ and } v \in \mathbb{V}), \tag{1.8}$$

then (1.7) still implies that

$$\|u_n\| \le \|u_{n-1}\|, \tag{1.9}$$

when u_n is computed from $u_{n-1} \in \mathbb{V}$ according to (1.3). In the last mentioned paper, time discretization methods for which a positive constant c exists such that (1.7), (1.8) always imply (1.9) were called *strong-stability-preserving (SSP)*. Property (1.9) is important, also with seminorms different from $\|.\|_{TV}$, and also when solving certain differential equations different from conservation laws—see, e.g., Dekker and Verwer (1984), LeVeque (2002), and Hundsdorfer and Verwer (2003).

Clearly, it would be awkward if the factor c, defined in (1.6), would be so small that (1.7) would reduce to a stepsize restriction which is too severe for any practical purposes—in fact, the less restrictions on Δt, the better. One might thus be tempted to take the magnitude of c into account when comparing the effectiveness of different Runge–Kutta processes (1.3), (1.5) to each other. However, it is evident that such a use of c, defined by (1.6), could be quite misleading if, for a given process (1.3), (1.5), the conclusion in Theorem 1.1 would also be valid with some factor c which is (much) larger than the one given by (1.6).

For any given method (1.3) satisfying (1.5), the question thus arises what is the largest factor c, *not necessarily defined via* (1.6), such that the conclusion in Theorem 1.1 is still valid. Moreover, a second question is of whether there exists a positive constant c such that (1.4), (1.7) imply (1.2), also for methods (1.3) satisfying (1.5a) but violating (1.5b). Two analogous questions arise in connection with the generalized version of Theorem 1.1, related to the SSP property, mentioned above.

The purpose of this paper is to propose a general theory which allows us to answer the above questions, as well as related ones.

1.2. Outline of the rest of the paper. In section 2 we present our general theory, just mentioned at the end of section 1.1. Section 2.1 contains notations and definitions which are basic for the rest of our paper. We review here the concept of *monotonicity*, which generalizes the TVD-property (1.2) in the context of arbitrary vector spaces $\mathbb{V}$, with seminorms $\|.\|$, and of general Runge–Kutta schemes (A, b). Furthermore, we introduce the notion of a *stepsize-coefficient* for monotonicity, which formalizes and generalizes the property of the coefficient c as stated in Theorem 1.1. In section 2.2 we recall the concept of irreducibility for general Runge–Kutta schemes (A, b), and we review the crucial quantity $R(A, b)$, introduced by Kraaijevanger (1991). In section 2.3 we present (without proof) our main result, Theorem 2.5. This theorem can be regarded as a variant to a theorem, on contractivity of Runge–Kutta methods, of Kraaijevanger (1991). Theorem 2.5 is relevant to arbitrary irreducible Runge–Kutta schemes (A, b); it tell us that, in the important situations specified by (2.9), (2.10), (2.11), respectively, the largest stepsize-coefficient for monotonicity is equal to $R(A, b)$.

In section 3 we apply Theorem 2.5 to a generalized version of process (1.3). After the introductory section 3.1, we clarify in the sections 3.2 and 3.3, respectively, the questions raised at the end of section 1.1 regarding the TVD and SSP properties of process (1.3). Section 3.4 gives two examples illustrating the superiority of the quantity $R(A, b)$ (to the factor c, given by (1.6)) as a guide to stepsize restrictions for the TVD and SSP properties.

Section 4 is mainly devoted to explicit Runge–Kutta schemes which are optimal, in the sense of their stepsize-coefficients for monotonicity. After the introductory section 4.1, we review, in section 4.2, conclusions of Kraaijevanger (1991) regarding the optimization of $R(A, b)$, in various classes of explicit Runge–Kutta schemes (A, b). Combining these conclusions and our Theorem 2.5, we are able to extend and shed new light on (recent) results in the literature about the optimization of c defined by (1.6). In section 4.3 we describe an algorithm for computing $R(A, b)$, which may be useful in determining further optimal Runge–Kutta methods. Section 4.4 contains a brief discussion of a few important related issues.

In order to look at our main result in the right theoretical perspective, we give in the final section, section 5, not only the formal proof of Theorem 2.5, but we present a short account of related material from Kraaijevanger (1991) as well. In section 5.1 we review Kraaijevanger's theorem mentioned above, and we compare it with our Theorem 2.5. In section 5.2 we give the proof of our main result.

We have framed our paper purposefully in the way just described: the reader who is primarily interested in our Theorem 2.5 and its applications (rather than in the underlying theory) will not be hampered by unnecessary digressions when reading sections 2, 3, and 4.

2. A general theory for monotonic Runge–Kutta processes.

2.1. Stepsize-coefficients for monotonicity in a general context. We want to study properties like (1.2) and (1.9) in a general setting. For that reason, we assume that $\mathbb{V}$ is an arbitrary real vector space, and that $F(v)$ is a given function, defined for all $v \in \mathbb{V}$, with values in $\mathbb{V}$. We consider a formal generalization of (1.1),

$$\frac{d}{dt}U(t) = F(U(t)) \quad (t \geq 0), \quad U(0) = u_0, \tag{2.1}$$

where u_0 and $U(t)$ stand for vectors in $\mathbb{V}$.

The general Runge–Kutta method with m stages, (formally) applied to the abstract problem (2.1), provides us with vectors $u_1, u_2, u_3, \ldots$ in $\mathbb{V}$ (see, e.g., Dekker and Verwer (1984), Butcher (1987), and Hairer and Wanner (1996)). Here u_n is related to u_{n-1} by the formula

$$u_n = u_{n-1} + \Delta t \sum_{j=1}^{m} b_j F(y_j), \tag{2.2a}$$

where the vectors y_j in $\mathbb{V}$ satisfy

$$y_i = u_{n-1} + \Delta t \sum_{j=1}^{m} a_{ij} F(y_j) \quad (1 \leq i \leq m). \tag{2.2b}$$

In these formulas, $\Delta t > 0$ denotes the stepsize and b_j, a_{ij} are real parameters, specifying the Runge–Kutta method. We always assume that $b_1 + b_2 + \cdots + b_m = 1$. If

$a_{ij} = 0$ (for $j \geq i$), the Runge–Kutta method is called *explicit*. Defining the $m \times m$ matrix A by $A = (a_{ij})$ and the column vector $b \in \mathbb{R}^m$ by $b = (b_1, b_2, b_3, \ldots, b_m)^T$, we can identify the Runge–Kutta method with the *coefficient scheme* (A, b).

Let $\|.\|$ denote an arbitrary seminorm on $\mathbb{V}$ (i.e., $\|u+v\| \leq \|u\| + \|v\|$ and $\|\lambda v\| = |\lambda| \cdot \|v\|$ for all real λ and $u, v \in \mathbb{V}$). The following inequality generalizes (1.2) and (1.9):

$$\|u_n\| \leq \|u_{n-1}\|. \tag{2.3}$$

We shall say that the Runge–Kutta method is *monotonic* (for the stepsize Δt, function F, and seminorm $\|.\|$) if (2.3) holds whenever the vectors u_{n-1} and u_n in $\mathbb{V}$ are related to each other as in (2.2). Our use of the term "monotonic" is nicely in agreement with earlier use of this term, e.g., by Burrage and Butcher (1980), Dekker and Verwer (1984, p. 263), Spijker (1986), Butcher (1987, p. 392), and Hundsdorfer, Ruuth, and Spiteri (2003). Property (2.3) is related to what sometimes is called *practical stability* or *strong stability*; see, e.g., Morton (1980) and Gottlieb, Shu, and Tadmor (2001).

In order to study stepsize restrictions for monotonicity, we start from a given stepsize $\tau_0 \in (0, \infty)$. We shall deal with the situation where F is a function from $\mathbb{V}$ into $\mathbb{V}$, satisfying

$$\|v + \tau_0 F(v)\| \leq \|v\| \quad (\text{for all } v \in \mathbb{V}). \tag{2.4}$$

The last inequality implies, for $0 < \Delta t \leq \tau_0$, that $\|v + \Delta t F(v)\| = \|(1 - \Delta t/\tau_0)v + (\Delta t/\tau_0)(v + \tau_0 F(v))\| \leq \|v\|$. Consequently, (2.4) is equivalent to the following generalized version of (1.4) and (1.8):

$$\|v + \Delta t F(v)\| \leq \|v\| \quad (\text{whenever } 0 < \Delta t \leq \tau_0 \text{ and } v \in \mathbb{V}).$$

Let a Runge–Kutta method (A, b) be given. We shall study monotonicity of the method under arbitrary stepsize restrictions of the form

$$0 < \Delta t \leq c \cdot \tau_0. \tag{2.5}$$

Definition 2.1 (stepsize-coefficient for monotonicity). *A value $c \in (0, \infty]$ is called a stepsize-coefficient for monotonicity (with respect to $\mathbb{V}$ and $\|.\|$) if the Runge–Kutta method is monotonic, as in* (2.3)*, whenever F is a function from $\mathbb{V}$ to $\mathbb{V}$ satisfying* (2.4)*, and Δt is a (finite) stepsize satisfying* (2.5)*.*

It is easily verified that this definition is independent of the above value τ_0: if c is a stepsize-coefficient for monotonicity, with respect to $\mathbb{V}$ and $\|.\|$, using one particular value $\tau_0 > 0$, then c will have the same property when using any other value, say $\tau_0' > 0$.

The concept of a stepsize-coefficient as introduced in the above definition, corresponds to what is sometimes called a *CFL coefficient* in the context of discretizations for hyperbolic problems; see, e.g., Gottlieb and Shu (1998) and Shu (2002).

In subsection 2.3 we shall give maximal stepsize-coefficients for monotonicity with respect to various spaces $\mathbb{V}$ and seminorms $\|.\|$.

2.2. Irreducible Runge–Kutta schemes and the quantity $R(A, b)$. In this subsection we give some definitions which will be needed when we formulate our results, in subsection 2.3, about maximal stepsize-coefficients c. We start with the fundamental concepts of reducibility and irreducibility.

DEFINITION 2.2 (reducibility and irreducibility). *An m-stage Runge–Kutta scheme (A, b) is called reducible if (at least) one of the following two statements* (i), (ii) *is true; it is called irreducible if neither* (i) *nor* (ii) *is true.*

(i) *There exist nonempty, disjoint index sets M, N with $M \cup N = \{1, 2, \ldots, m\}$ such that $b_j = 0$ (for $j \in N$) and $a_{ij} = 0$ (for $i \in M$, $j \in N$);*

(ii) *there exist nonempty, pairwise disjoint index sets $M_1, M_2, \ldots, M_r$, with $1 \leq r < m$ and $M_1 \cup M_2 \cup \cdots \cup M_r = \{1, 2, \ldots, m\}$, such that $\sum_{k \in M_q} a_{ik} = \sum_{k \in M_q} a_{jk}$ whenever $1 \leq p \leq r$, $1 \leq q \leq r$, and $i, j \in M_p$.*

In case the above statement (i) is true, the vectors y_j in (2.2) with $j \in N$ have no influence on u_n, and the Runge–Kutta method is equivalent to a method with less than m stages. Also in case of (ii), the Runge–Kutta method essentially reduces to a method with less then m stages; see, e.g., Dekker and Verwer (1984) or Hairer and Wanner (1996). Clearly, for all practical purposes, it is enough to consider only Runge–Kutta schemes which are irreducible.

Next, we turn to a very useful characteristic quantity for Runge–Kutta schemes introduced by Kraaijevanger (1991). Following this author, we shall denote his quantity by $R(A, b)$, and in defining it, we shall use, for real ξ, the notations

$$A(\xi) = A(I - \xi A)^{-1}, \qquad b(\xi) = (I - \xi A)^{-T} b,$$
$$e(\xi) = (I - \xi A)^{-1} e, \qquad \varphi(\xi) = 1 + \xi b^T (I - \xi A)^{-1} e.$$

Here $^{-T}$ stands for transposition after inversion, I denotes the identity matrix of order m, and e stands for the column vector in $\mathbb{R}^m$, all of whose components are equal to 1. We shall focus on values $\xi \leq 0$ for which

$$I - \xi A \text{ is invertible}, \quad A(\xi) \geq 0, \quad b(\xi) \geq 0, \quad e(\xi) \geq 0, \quad \text{and} \quad \varphi(\xi) \geq 0. \tag{2.6}$$

The first inequality in (2.6) should be interpreted entrywise, the second and the third ones componentwise. Similarly, all inequalities for matrices and vectors occurring below are to be interpreted entrywise and componentwise, respectively.

DEFINITION 2.3 (the quantity $R(A, b)$). *Let (A, b) be a given coefficient scheme. In case $A \geq 0$ and $b \geq 0$, we define*

$$R(A, b) = \sup\{r : \ r \geq 0 \ \text{ and } (2.6) \text{ holds for all } \xi \text{ with } -r \leq \xi \leq 0\}.$$

In case (at least) one of the inequalities $A \geq 0$, $b \geq 0$ is violated, we define $R(A, b) = 0$.

Definition 2.3 suggests that it may be difficult to determine $R(A, b)$ for given coefficient schemes (A, b). However, in section 4 we shall see that (for explicit Runge–Kutta methods) *a simple algorithm exists for computing $R(A, b)$*. Moreover, Kraaijevanger (1991, p. 497) gave the following simple criterion (2.7) for determining whether $R(A, b) = 0$ or $R(A, b) > 0$. For any given $k \times l$ matrix $B = (b_{ij})$, we define the corresponding $k \times l$ *incidence matrix* by

$$\mathrm{Inc}(B) = (c_{ij}), \quad \text{with } c_{ij} = 1 \text{ (if } b_{ij} \neq 0) \text{ and } c_{ij} = 0 \text{ (if } b_{ij} = 0).$$

THEOREM 2.4 (about positivity of $R(A, b)$). *Let (A, b) be a given irreducible coefficient scheme. Then $R(A, b) > 0$ if and only if*

$$A \geq 0, \quad b > 0, \quad \text{and} \quad \mathrm{Inc}(A^2) \leq \mathrm{Inc}(A). \tag{2.7}$$

Proof. For ξ sufficiently close to zero, the matrix $I - \xi A$ is invertible and $e(\xi) \geq 0$, $\varphi(\xi) \geq 0$. Therefore, it is sufficient to analyze the inequalities $A(\xi) \geq 0$ and $b(\xi) \geq 0$. With no loss of generality, we assume $A \geq 0$, $b \geq 0$.

For ξ close to zero, we have

$$A(\xi) = (A + \xi A^2) \sum_{k=0}^{\infty} (\xi A)^{2k} \quad \text{and} \quad b(\xi)^T = (b^T + \xi b^T A) \sum_{k=0}^{\infty} (\xi A)^{2k}.$$

From these two expressions, one easily sees that there exists a positive r, with

$$A(\xi) \geq 0 \quad \text{and} \quad b(\xi)^T \geq 0 \quad (\text{for } -r \leq \xi \leq 0)$$

if and only if $\text{Inc}(A^2) \leq \text{Inc}(A)$ and $\text{Inc}(b^T A) \leq \text{Inc}(b^T)$. Since statement (i) in Definition 2.2 is *not* true, we conclude that the last inequality is equivalent to $b > 0$. □

We note that, in Kraaijevanger (1991), one can find various other interesting properties related to $R(A, b)$, among them characterizations different from Definition 2.3.

2.3. Formulation of our main theorem. In this subsection we shall determine maximal stepsize-coefficients (Definition 2.1) with respect to general spaces $\mathbb{V}$ and seminorms $\|.\|$. Moreover, we shall pay special attention to the particular (semi)norms

$$\|y\|_\infty = \sup_{-\infty<j<\infty} |\eta_j|, \quad \|y\|_1 = \sum_{-\infty}^{\infty} |\eta_j|, \quad \|y\|_{TV} = \sum_{-\infty}^{\infty} |\eta_j - \eta_{j-1}|$$

for $y = (\ldots, \eta_{-1}, \eta_0, \eta_1, \ldots) \in \mathbb{R}^\infty$. Furthermore, for integers $s \geq 1$ and vectors $y \in \mathbb{R}^s$ with components η_j $(1 \leq j \leq s)$, we shall focus on the (semi)norms

$$\|y\|_\infty = \max_{1\leq j\leq s} |\eta_j|, \quad \|y\|_1 = \sum_{j=1}^{s} |\eta_j|, \quad \|y\|_{TV} = \sum_{j=2}^{s} |\eta_j - \eta_{j-1}|$$

(where $\sum_{j=2}^{s} |\eta_j - \eta_{j-1}| = 0$ for $s = 1$). In our Theorem 2.5, the following inequality will play a prominent part:

$$c \leq R(A, b). \tag{2.8}$$

Here is our main theorem, about stepsize-coefficients of irreducible Runge–Kutta schemes (Definitions 2.1 and 2.2).

THEOREM 2.5 (relating monotonicity to $R(A, b)$). *Consider an arbitrary irreducible Runge–Kutta scheme (A, b). Let c be a given value with $0 < c \leq \infty$. Choose one of the three (semi)norms $\|.\|_\infty$, $\|.\|_1$, or $\|.\|_{TV}$, and denote it by $|\!|.|\!|$. Then each of the following three statements is equivalent to* (2.8).

(2.9) *c is a stepsize-coefficient for monotonicity, with respect to all vector spaces $\mathbb{V}$ and seminorms $\|.\|$ on $\mathbb{V}$;*

(2.10) *c is a stepsize-coefficient for monotonicity, with respect to the special space $\mathbb{V} = \{y : \ y \in \mathbb{R}^\infty \text{ and } |\!|y|\!| < \infty\}$ and seminorm $\|.\| = |\!|.|\!|$;*

(2.11) *c is a stepsize-coefficient for monotonicity, with respect to the finite dimensional space $\mathbb{V} = \mathbb{R}^s$ and seminorm $\|.\| = |\!|.|\!|$ for $s = 1, 2, 3, \ldots$.*

Clearly, (2.9) is a priori a stronger statement than (2.10) or (2.11). Accordingly, the essence of Theorem 2.5 is that the (algebraic) property (2.8) implies the (strong) statement (2.9), whereas already either of the (weaker) statements (2.10) or (2.11) implies (2.8).

The above theorem highlights the importance of Kraaijevanger's quantity $R(A, b)$. Theorem 2.5 shows that, with respect to each of the three situations specified in (2.9), (2.10), and (2.11), *the maximal stepsize-coefficient for monotonicity is equal to* $R(A, b)$.

The above theorem will be compared with a theorem on nonlinear contractivity of Kraaijevanger (1991) in section 5.1, and it will be proved in section 5.2.

3. The application of our main theorem to the questions raised in subsection 1.1.

3.1. The equivalence of (a generalized version of) process (1.3) to method (2.2). In section 3 we study time stepping processes producing numerical approximations $u_n \in \mathbb{R}^\infty$ to $U(n\Delta t)$ (for $n \geq 1$), where $U(t) \in \mathbb{R}^\infty$ satisfies (1.1). We focus on processes of the form

$$y_1 = u_{n-1}, \tag{3.1a}$$

$$y_i = \sum_{j=1}^{m} [\lambda_{ij} y_j + \Delta t \cdot \mu_{ij} F(y_j)] \quad (2 \leq i \leq m), \tag{3.1b}$$

$$u_n = \sum_{j=1}^{m} [\lambda_{m+1,j} y_j + \Delta t \cdot \mu_{m+1,j} F(y_j)]. \tag{3.1c}$$

Here λ_{ij}, μ_{ij} are arbitrary real coefficients with

$$\lambda_{i1} + \lambda_{i2} + \cdots + \lambda_{im} = 1 \quad (2 \leq i \leq m+1). \tag{3.2a}$$

Clearly, if $\lambda_{ij} = \mu_{ij} = 0$ (for $j \geq i$), the above process reduces to algorithm (1.3). Moreover, process (3.1) is sufficiently general to also cover other algorithms, such as the one in Gottlieb, Shu, and Tadmor (2001, p. 109), which was considered recently for solving (1.1).

In order to relate (3.1) to a Runge–Kutta method in the standard form (2.2), we define $\lambda_{ij} = \mu_{ij} = 0$ (for $i = 1$ and $1 \leq j \leq m$), and we introduce the $(m+1) \times m$ matrices $L = (\lambda_{ij})$, $M = (\mu_{ij})$. The $m \times m$ submatrices composed of the first m rows of L and M, respectively, will be denoted by L_0 and M_0. Furthermore, the last rows of L and M—that is, $(\lambda_{m+1,1}, \ldots, \lambda_{m+1,m})$ and $(\mu_{m+1,1}, \ldots, \mu_{m+1,m})$, respectively—will be denoted by L_1 and M_1, so that

$$L = \begin{pmatrix} L_0 \\ L_1 \end{pmatrix} \quad \text{and} \quad M = \begin{pmatrix} M_0 \\ M_1 \end{pmatrix}. \tag{3.2b}$$

We assume that

$$\text{the } m \times m \text{ matrix } I - L_0 \text{ is invertible.} \tag{3.2c}$$

We shall now show that the relations (3.1) imply (2.2), with matrix $A = (a_{ij})$ and column vector $b = (b_i)$ specified by

$$A = (I - L_0)^{-1} M_0 \quad \text{and} \quad b^T = M_1 + L_1 A. \tag{3.3}$$

We denote the entries of the matrix $(I - L_0)^{-1}$ by γ_{ij}, and note that the relations (3.1a), (3.1b) can be rewritten as

$$\sum_{k=1}^{m} (\delta_{jk} - \lambda_{jk}) y_k = \delta_{j,1} u_{n-1} + \sum_{k=1}^{m} \mu_{jk} F_k \quad (\text{for } 1 \leq j \leq m), \tag{3.4}$$

where δ_{jk} is the Kronecker index and $F_k = \Delta t \cdot F(y_k)$. Multiplying (3.4) by γ_{ij} and summing over $j = 1, 2, \ldots, m$, we obtain, for $1 \le i \le m$, the equality $y_i = (\sum_{j=1}^{m} \gamma_{ij}\delta_{j,1})u_{n-1} + \sum_{k=1}^{m}(\sum_{j=1}^{m} \gamma_{ij}\mu_{jk})F_k$. In view of (3.2a), the first sum in the right-hand member of the last equality is equal to 1; hence (2.2b) holds with $(a_{ij}) = (I - L_0)^{-1}M_0$. Furthermore, in view of (3.1c), we easily arrive at (2.2a) with $(b_1, b_2, \ldots, b_m) = M_1 + L_1 A$.

Similarly to the above, the relations (2.2), (3.3) can be proved to imply (3.1), so that the following conclusion is valid.

LEMMA 3.1. *Let λ_{ij} and μ_{ij} be given coefficients satisfying* (3.2a), (3.2b), (3.2c). *Define the Runge–Kutta scheme (A, b) by* (3.3). *Then the relations* (3.1) *are equivalent to* (2.2).

In the following subsections, we shall use this lemma for relating the monotonicity properties of process (3.1) to those of the corresponding Runge–Kutta scheme (A, b) given by (3.3).

3.2. The total-variation-diminishing property of process (3.1). Our following Theorem 3.2 gives a stepsize restriction guaranteeing the TVD-property for the general process (3.1). Since (3.1) is more general than process (1.3), our theorem is highly relevant to (1.3). In the theorem, we shall use the notation

$$\mathbb{R}^{\infty}_{TV} = \{y : \ y \in \mathbb{R}^{\infty} \text{ with } \|y\|_{TV} < \infty\},$$

where $\|.\|_{TV}$ has the same meaning as in subsection 1.1. We shall deal with functions F from $\mathbb{R}^{\infty}_{TV}$ into $\mathbb{R}^{\infty}_{TV}$, satisfying

$$\|v + \tau_0 F(v)\|_{TV} \le \|v\|_{TV} \quad (\text{whenever } v \in \mathbb{R}^{\infty}_{TV}), \tag{3.5}$$

and with stepsize restrictions of the form

$$0 < \Delta t \le R(A, b) \cdot \tau_0 \tag{3.6}$$

(see Definition 2.3).

THEOREM 3.2 (optimal stepsize restriction for the TVD-property in process (3.1)). *Let λ_{ij} and μ_{ij} be given coefficients satisfying* (3.2a), (3.2b), (3.2c). *Define the matrix A and the vector b by* (3.3), *and suppose that the coefficient scheme (A, b) is irreducible (Definition* 2.2). *Let F be a function from $\mathbb{R}^{\infty}_{TV}$ into $\mathbb{R}^{\infty}_{TV}$ satisfying* (3.5), *and let Δt be a (finite) stepsize satisfying* (3.6).

Then, process (3.1) *is TVD; i.e., the inequality* (1.2) *holds whenever $u_{n-1}, u_n \in \mathbb{R}^{\infty}_{TV}$ are related to each other as in* (3.1).

Proof. We apply Lemma 3.1, and consider the Runge–Kutta scheme (A, b) specified by the lemma. Next, we apply Theorem 2.5: choosing $c = R(A, b)$, we have (2.8) so that (2.10) must be fulfilled with $|\!|.|\!| = \|.\|_{TV}$. An application of Definition 2.1 completes the proof of the theorem. □

Remark 3.3. The above theorem has a *wider scope than Theorem* 1.1. The class of numerical methods (3.1) satisfying (3.2a), (3.2b), (3.2c) encompasses all processes (1.3) satisfying (1.5a), as well as other (implicit) procedures. Specifically, unlike Theorem 1.1, the above Theorem 3.2 is relevant to processes (1.3) satisfying (1.5a) but violating (1.5b)—see Example 3.7 in subsection 3.4 for an illustration.

Remark 3.4. The above theorem, when applied to any process (1.3) satisfying (1.5a), (1.5b), gives a *stronger conclusion than Theorem* 1.1. By Theorem 2.5, property (2.10) with $|\!|.|\!| = \|.\|_{TV}$ implies inequality (2.8). Therefore the coefficient c, given by Theorem 1.1, satisfies $c \le R(A, b)$; this means that the stepsize restriction (3.6) of

Theorem 3.2 is, in general, less severe than the restriction (1.7) of Theorem 1.1—see Example 3.8 in subsection 3.4 for an illustration.

Remark 3.5. Theorem 3.2 gives a stepsize restriction which is *optimal* in that the conclusion of the theorem would no longer be valid if the factor $R(A,b)$ in (3.6) would be replaced by any factor $c > R(A,b)$. This follows again from Theorem 2.5.

3.3. The strong-stability-preserving property of process (3.1). Let $\mathbb{V}$ be an arbitrary linear subspace of $\mathbb{R}^\infty$, and let $\|.\|$ denote any seminorm on $\mathbb{V}$. For functions $F : \mathbb{V} \longrightarrow \mathbb{V}$ satisfying

$$\|v + \tau_0 F(v)\| \leq \|v\| \quad (\text{whenever } v \in \mathbb{V}), \tag{3.7}$$

we shall consider process (3.1) under a stepsize restriction of the form

$$0 < \Delta t \leq c \cdot \tau_0. \tag{3.8}$$

Following the terminology of Gottlieb, Shu, and Tadmor (2001), already reviewed in subsection 1.1, we shall say that process (3.1) is *strong-stability-preserving* (SSP) if a positive constant c exists (only depending on λ_{ij} and μ_{ij}) such that (1.9) holds whenever (3.1), (3.7), (3.8) are fulfilled.

THEOREM 3.6 (criterion for the SSP property of process (3.1)). *Let λ_{ij} and μ_{ij} be given coefficients satisfying* (3.2a), (3.2b), (3.2c). *Define the matrix A and vector b by* (3.3), *and suppose that the coefficient scheme (A,b) is irreducible (Definition* 2.2). *Then process* (3.1) *is SSP if and only if* (2.7) *holds.*

Proof. By Lemma 3.1 and Theorem 2.5, process (3.1) is SSP if and only if $R(A,b) > 0$. According to Theorem 2.4, the last inequality is equivalent to (2.7). ∎

It is clear that the above Theorem 3.6, similarly as Theorem 3.2, is highly relevant to all numerical processes (1.3) satisfying (1.5a); see Examples 3.7 and 3.8 below for illustrations.

3.4. Illustrations to Theorems 3.2 and 3.6. We give two examples illustrating Theorems 3.2 and 3.6.

Example 3.7. Consider process (1.3), with $m = 3$ and coefficients λ_{ij}, μ_{ij} given by the relations

$$\begin{pmatrix} \lambda_{21} & & \\ \lambda_{31} & \lambda_{32} & \\ \lambda_{41} & \lambda_{42} & \lambda_{43} \end{pmatrix} = \begin{pmatrix} 1 & & \\ \frac{1}{4} & \frac{3}{4} & \\ 1 & 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} \mu_{21} & & \\ \mu_{31} & \mu_{32} & \\ \mu_{41} & \mu_{42} & \mu_{43} \end{pmatrix} = \begin{pmatrix} 1 & & \\ -\frac{1}{2} & \frac{1}{4} & \\ \frac{1}{6} & \frac{1}{6} & \frac{2}{3} \end{pmatrix}.$$

Since $\mu_{31} < 0$, condition (1.5b) is violated; therefore Theorem 1.1 does not apply.

For the corresponding matrix $A = (a_{ij})$ and vector $b = (b_i)$ (see (3.3)), we have $a_{ij} = 0$ $(j \geq i)$, $a_{21} = 1$, $a_{31} = a_{32} = 1/4$ and $b_1 = b_2 = 1/6$, $b_3 = 2/3$, respectively. It is very easy to see that (2.7) holds; by virtue of Theorem 3.6, the numerical process is thus SSP. Moreover, according to Kraaijevanger (1991, Theorem 9.4), for this process we have $R(A,b) = 1$. By Theorem 3.2 we conclude that the process is TVD, under the assumption (3.5) if $0 < \Delta t \leq \tau_0$. We note that essentially the same numerical process was presented earlier by Shu and Osher (1988); we shall come back to it in section 4.2 (Remark 4.4; $m = p = 3$).

Example 3.8. Consider process (1.3), with $m = 2$ and

$$\begin{pmatrix} \lambda_{21} & \\ \lambda_{31} & \lambda_{32} \end{pmatrix} = \begin{pmatrix} 1 & \\ 1 & 0 \end{pmatrix}, \quad \begin{pmatrix} \mu_{21} & \\ \mu_{31} & \mu_{32} \end{pmatrix} = \begin{pmatrix} 1/2 & \\ 1/2 & 1/2 \end{pmatrix}.$$

The conditions (1.5a), (1.5b) are neatly fulfilled, but the coefficient c, defined by (1.6), is equal to 0.

For the corresponding Runge–Kutta scheme (A, b), defined by (3.3), we have $a_{ij} = 0$ $(j \geq i)$, $a_{21} = 1/2$ and $b_1 = b_2 = 1/2$. Clearly, (2.7) is fulfilled, guaranteeing the SSP property (see Theorem 3.6). Moreover, according to Kraaijevanger (1991, Theorem 9.2), we have $R(A, b) = 2$. Therefore, by Theorem 3.2, the numerical process is TVD, under assumption (3.5), if $0 < \Delta t \leq 2 \cdot \tau_0$. We note that the same method was presented by Spiteri and Ruuth (2002); we shall come back to it in section 4.2 (Remark 4.4; $m = 2$, $p = 1$).

4. Optimal Runge–Kutta methods.

4.1. Preliminaries. For integer values $m \geq 1$ and $p \geq 1$, we shall denote by $E_{m,p}$ the class of all explicit m-stage Runge–Kutta methods (A, b) with (classical) order of accuracy at least p. Considerable attention has been paid, in the literature, to identifying methods of class $E_{m,p}$ of the special form (1.3), (1.5) which are optimal in the sense of the coefficient c given by (1.6); see notably Shu and Osher (1988), Gottlieb and Shu (1998), Ruuth and Spiteri (2002), Shu (2002), and Spiteri and Ruuth (2002). Independently of this work, Kraaijevanger (1991) dealt with the optimization, in the full class $E_{m,p}$, of his quantity $R(A, b)$. Our theory (section 2) can be used to relate his conclusions to the work just mentioned about optimization of c defined in (1.6).

In section 4.2 we shall briefly review some of Kraaijevanger's conclusions so as to arrive at extensions and completions of the material, referred to above, on optimality in the sense of c (1.6). Furthermore, we shall consider scaled stepsize-coefficients which reflect the efficiency of the methods better than the unscaled coefficients; in Table 4.1 we shall display optimal scaled stepsize-coefficients. Next, in section 4.3, we shall focus on an algorithm for computing $R(A, b)$; the authors feel that it can be useful in (future) calculations for determining, numerically, optimal Runge–Kutta methods. Finally, in section 4.4 we touch upon a few important related issues.

4.2. Optimal methods in the class $E_{m,p}$. We start with the following fundamental lemma, which gives a simple upper bound for $R(A, b)$ in the class $E_{m,p}$.

LEMMA 4.1 (Kraaijevanger (1991, p. 517)). *Let $1 \leq p \leq m$, and consider an arbitrary Runge–Kutta method (A, b) of class $E_{m,p}$. Then $R(A, b) \leq m - p + 1$.*

Remark 4.2. Ruuth and Spiteri (2002, Theorem 3.1) showed that, for Runge–Kutta methods in class $E_{m,p}$ of the special form (1.3), (1.5), the coefficient c defined by (1.6) satisfies $c \leq m-p+1$. Clearly, a combination of the above lemma and our theory (section 2) yields an extension and improvement over the last bound on c: for *any* Runge–Kutta method of class $E_{m,p}$, *any* stepsize-coefficient for monotonicity, say c', and *any* of the situations covered by (2.9), (2.10), or (2.11), we have $c' \leq m - p + 1$.

The following theorem specifies methods (A, b) for which the upper bound $R(A, b) \leq m - p + 1$ of Lemma 4.1 becomes an equality.

THEOREM 4.3 (Kraaijevanger (1991, pp. 518–520)).

(a) *Let $p = 1 \leq m$. Then there is a unique method (A, b) of class $E_{m,p}$ with $R(A, b) = m$; it is given by $a_{ij} = 1/m$ $(1 \leq j < i \leq m)$ and $b_i = 1/m$ $(1 \leq i \leq m)$.*

(b) *Let $p = 2 \leq m$. Then there is a unique method (A, b) of class $E_{m,p}$ with $R(A, b) = m - 1$; it is given by $a_{ij} = 1/(m - 1)$ $(1 \leq j < i \leq m)$ and $b_i = 1/m$ $(1 \leq i \leq m)$.*

(c) *Let $p = 3$, $m = 3$. Then there is a unique method (A, b) of class $E_{m,p}$ with $R(A, b) = 1$; it is given by $a_{21} = 1$, $a_{31} = a_{32} = 1/4$, $b_1 = b_2 = 1/6$, and $b_3 = 2/3$.*

(d) *Let $p = 3$, $m = 4$. Then there is a unique method (A, b) of class $E_{m,p}$ with $R(A, b) = 2$; it is given by $a_{21} = a_{31} = a_{32} = b_4 = 1/2$ and $a_{4,i} = b_i = 1/6$ $(1 \le i \le 3)$.*

Remark 4.4. Essentially the same methods as specified in the above theorem, for $m = p = 2$ and $m = p = 3$, were already found by Shu and Osher (1988) in a search for methods in $E_{m,p}$, of the special type (1.3), (1.5), with maximal c (defined in (1.6)); Gottlieb and Shu (1998) proved optimality for these two methods with respect to c, (1.6). In an analogous search, Spiteri and Ruuth (2002) arrived at all other methods specified by the theorem, and proved optimality in the sense of c, (1.6). Similarly as in Remark 4.2, our theory (section 2) can be used here to conclude that all methods given in Theorem 4.3 are optimal (with respect to their stepsize-coefficients for monotonicity) in a *stronger sense*, and over a *larger class* of Runge–Kutta methods, than can be concluded from the three papers just mentioned.

Kraaijevanger (1991) did not specify analytically any methods (A, b) in $E_{m,p}$ with maximal $R(A, b)$, for pairs p, m different from those in Theorem 4.3. However, he arrived at interesting (negative) conclusions: if method (A, b) is of class $E_{m,p}$ and $p = 3$, $m \ge 5$, then $R(A, b) < m-p+1$; and if (A, b) belongs to $E_{m,p}$ with $p = m = 4$ or $p \ge 5$, then $R(A, b) = 0$. Moreover, by combining Kraaijevanger (1986, Theorem 5.1), Spijker (1983), and our Theorem 2.5, one can conclude that $R(A, b) < m - p + 1$ also for all (A, b) in $E_{m,p}$ with $p = 4$, $m \ge 6$. A combination of these conclusions and our theory (section 2) amounts to a far-reaching extension of related results obtained in Ruuth and Spiteri (2002).

Kraaijevanger (1991, pp. 522–523) constructed numerically an explicit 5-stage method (A, b) of order 4, with $R(A, b) \approx 1.508$. It is interesting to note that the same method was found by Spiteri and Ruuth (2002) in a numerical search within the class of methods (1.3) satisfying (1.5). By a similar search, the last authors also found a 5-stage method of order 3 with $c \approx 2.651$ (given by (1.6)). In view of Kraaijevanger (1986, Theorem 5.3), Spijker (1983), and our Theorem 2.5, we can conclude that this method has a value $R(A, b) \approx 2.651$, and is optimal in a *stronger sense* and over a *larger class* of methods than follows from Spiteri and Ruuth (2002).

Clearly, when comparing two explicit Runge–Kutta methods to each other, one cannot simply say that the one with the largest value $R(A, b)$ is the most efficient one. However, assuming that the stepsize Δt, used for solving (1.1) over some interval $[0, T]$, is governed by monotonicity (TVD) demands, it seems reasonable to use the quantity $m \cdot T/R(A, b)$ as a measure of the amount of computational labor of a Runge–Kutta method (A, b) with m stages—cf. Jeltsch and Nevanlinna (1981), Kraaijevanger (1986), and Spiteri and Ruuth (2002) for related considerations. In line with the terminology in the first two of these papers, we shall refer to the ratio $R(A, b)/m$ as a *scaled stepsize-coefficient*. The above mentioned measure, for the amount of computational labor, is inversely proportional to $R(A, b)/m$, so the scaled stepsize-coefficient is a more realistic guide than $R(A, b)$ for comparing the efficiency of different methods to each other.

In Table 4.1 we display scaled stepsize-coefficients of Runge–Kutta methods (A, b), which were reviewed above and are optimal in $E_{m,p}$ with respect to $R(A, b)$.

From the table, one may conclude that, for given p, it is advantageous to use optimal methods with relatively large m. Clearly, this conclusion is (only) justifiable under the above assumption about Δt being determined by monotonicity demands. For related numerical experiments, see, e.g., Gottlieb and Shu (1998) and Spiteri and Ruuth (2002).

TABLE 4.1
Scaled stepsize-coefficients $R(A,b)/m$ for optimal Runge–Kutta methods in $E_{m,p}$.

	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$
$p=1$	1	1	1	1	1
$p=2$		0.500	0.667	0.750	0.800
$p=3$			0.333	0.500	0.530
$p=4$					0.302

4.3. An algorithm for computing $R(A,b)$, for methods of class $E_{m,p}$. Below we will describe a simple algorithm for computing $R(A,b)$ whenever (A,b) is an irreducible Runge–Kutta scheme of class $E_{m,p}$. The following lemma plays a fundamental role in the algorithm.

LEMMA 4.5 (Kraaijevanger (1991, pp. 497–498)). *Let (A,b) be an irreducible coefficient scheme and r a positive real number. Then $R(A,b) \geq r$ if and only if $A \geq 0$ and the conditions* (2.6) *are fulfilled at $\xi = -r$.*

It was noted by Kraaijevanger (1991) that the above lemma simplifies calculating $R(A,b)$ if $A \geq 0$: for checking the conditions (2.6) on the whole of an interval $[-r,0]$, it is sufficient to consider only the left endpoint $\xi = -r$.

Let Test1 and Test2(x) be boolean functions defined by

$$\mathsf{Test1} = \begin{cases} \mathsf{true} & \text{if (2.7) holds,} \\ \mathsf{false} & \text{otherwise;} \end{cases} \qquad \mathsf{Test2(x)} = \begin{cases} \mathsf{true} & \text{if (2.6) holds at } \xi = \mathsf{x}, \\ \mathsf{false} & \text{otherwise.} \end{cases}$$

From Lemma 4.1 we know that if (A,b) is a coefficient scheme of class $E_{m,p}$, then $R(A,b) \leq m - p + 1$. In view of the last inequality, Theorem 2.4, and Lemma 4.5, we can calculate $R(A,b)$ with the wanted precision Tol, by using the above boolean functions as well as two pointers LeftExtr and RightExtr. The following algorithm finds $R(A,b)$ with error $\leq$ Tol.

```
x=0
if Test1
   LeftExtr=-(m-p+1),   RightExtr=0,   x=LeftExtr
   while (RightExtr-LeftExtr ≥ 2·Tol)
      if Test2(x)
         RightExtr=x,   x=(LeftExtr+RightExtr)/2
      else
         LeftExt=x,   x=(LeftExtr+RightExtr)/2
      end
   end
end
R(A,b)=−x.
```

4.4. Final remarks. For completeness, we note that Gottlieb and Shu (1998), Shu (2002), and Spiteri and Ruuth (2002) gave useful results regarding the optimization of c, (1.6), over classes of low-storage schemes of the (special) form (1.3), (1.5). Furthermore, Kennedy, Carpenter, and Lewis (2000) obtained interesting related results regarding the optimization of $R(A,b)$ over general classes of low-storage schemes (A,b). Clearly, our theory (section 2) is fit to put also this work in a wider perspective.

Above, in section 4, we dealt exclusively with explicit Runge–Kutta schemes. However, in Kraaijevanger (1991) also (a few) results were obtained, regarding the size of $R(A,b)$, relevant to implicit schemes—see below. A combination of these results

with our Theorem 2.5 immediately leads to interesting conclusions about stepsize-coefficients for monotonicity.

For arbitrary (possibly implicit) schemes (A, b) of order p, the following general results were obtained in Kraaijevanger (1991, pp. 514, 516): if $p \geq 2$, then $R(A, b) < \infty$; and if $p \geq 7$, then $R(A, b) = 0$. Moreover (on p. 516 of that article), a notable implicit method (A, b) was given, with a value $R(A, b)$ exceeding the upper bound of Lemma 4.1: the method with $m = 2$, $a_{1,1} = a_{1,2} = 0$, $a_{2,1} = a_{2,2} = 3/8$, $b_1 = 1/3$, $b_2 = 2/3$ is of order $p = 2$ and has a value $R(A, b) = 8/3$. The last value is considerably larger than the optimal value $m - p + 1 = 1$, which can be achieved in $E_{2,2}$ (cf. section 4.2); but this advantage should of course be balanced against the additional amount of work per step due to the implicitness of the method.

We think that it would be very useful to perform a systematic search for implicit methods which are optimal, for given m and p, in the sense of $R(A, b)$. Because such a search is beyond the scope of our present work, we do not go further into this matter here.

Finally, we note that our algorithm in section 4.3 can easily be adapted so as to compute $R(A, b)$ also for methods (A, b), of order at least 2, which are implicit: we still base the algorithm on Lemma 4.5, and (instead of using Lemma 4.1) we start with $\mathsf{LeftExtr} = \xi$, where ξ is a negative value at which (2.6) is violated; in view of the bound $R(A, b) < \infty$, such a ξ can be found, e.g., by a simple doubling process.

5. Kraaijevanger's theory and our proof of Theorem 2.5.

5.1. A theorem of Kraaijevanger on contractivity. Kraaijevanger (1991) presented an interesting theory, relevant to method (2.2) in the situation where F is a function from $\mathbb{R}^s$ into $\mathbb{R}^s$, and $\|.\|$ is a norm on $\mathbb{R}^s$. The focus in his paper is on numerical processes which, for given F, $\|.\|$, and Δt, are *contractive* in the sense that

$$\|\tilde{u}_n - u_n\| \leq \|\tilde{u}_{n-1} - u_{n-1}\| \tag{5.1}$$

whenever both the vectors u_{n-1}, u_n and the vectors $\tilde{u}_{n-1}, \tilde{u}_n$ are related to each other as in (2.2). Kraaijevanger studied property (5.1) for functions F satisfying

$$\|F(\tilde{v}) - F(v) + \rho(\tilde{v} - v)\| \leq \rho\|\tilde{v} - v\| \quad (\text{for all } v, \tilde{v} \in \mathbb{R}^s). \tag{5.2}$$

Here ρ is a positive constant; in the literature on numerical ODEs one often refers to (5.2) as a *circle condition* (with radius ρ) on the function F—cf. Kraaijevanger (1991).

In order to be able to reformulate one of Kraaijevanger's main results in such a way that it can easily be compared to our Theorem 2.5, we consider stepsize-restrictions of the form

$$0 < \Delta t \leq c/\rho. \tag{5.3}$$

Furthermore, adapting our Definition 2.1 to the situation at hand, we arrive at the following definition.

DEFINITION 5.1 (stepsize-coefficient for contractivity). *A value $c \in (0, \infty]$ is a stepsize-coefficient for contractivity (with respect to $\mathbb{R}^s$ and $\|.\|$) if the Runge–Kutta method is contractive, as in* (5.1)*, whenever $F : \mathbb{R}^s \to \mathbb{R}^s$ satisfies* (5.2) *and Δt is a (finite) stepsize satisfying* (5.3).

The subsequent theorem is an easy consequence of Kraaijevanger (1991, Theorem 5.4); it relates stepsize-coefficients for contractivity to the inequality

$$c \leq R(A, b). \tag{5.4}$$

THEOREM 5.2 (relating contractivity to $R(A,b)$). *Consider an arbitrary irreducible Runge–Kutta scheme (A,b). Let c be a given value with $0 < c \leq \infty$. Then both of the following statements are equivalent to* (5.4)*:*

(5.5) *c is a stepsize-coefficient for contractivity, with respect to $\mathbb{R}^s$ and $\|.\|$ for each $s \geq 1$ and each norm $\|.\|$ on $\mathbb{R}^s$;*

(5.6) *c is a stepsize-coefficient for contractivity, with respect to $\mathbb{R}^s$ and the special norm $\|.\|_\infty$ for each $s \geq 1$.*

Since condition (5.2) is equivalent to requiring that the forward Euler method with stepsize $\tau_0 = 1/\rho$ is contractive, there is a close resemblance between (5.2) and (2.4) (with $\mathbb{V} = \mathbb{R}^s$). Accordingly, one might think that (part of) our Theorem 2.5 is a simple consequence of Theorem 5.2. However, the following three remarks indicate that the relation between the two theorems is far from being that simple.

Remark 5.3. Let c be as in statement (2.11), with seminorm $\|.\| = \|.\|_1$ or $\|.\| = \|.\|_{TV}$. Theorem 2.5 claims that this coefficient c must satisfy $c \leq R(A,b)$. This claim cannot be expected to follow from the above Theorem 5.2; at best, it might follow from a version of that theorem in which the norm $\|.\|_\infty$ (in (5.6)) would simply be replaced by $\|.\|_1$ or $\|.\|_{TV}$. However, it is not known whether such a version is actually valid—Kraaijevanger's proof, underlying Theorem 5.2 as formulated above, makes an essential use of a specific (geometric) property of the norm $\|.\|_\infty$ which is *not* valid for $\|.\|_1$ or $\|.\|_{TV}$; cf. Kraaijevanger (1991, p. 505) and Schönbeck (1967, Theorem 2.4) for more details.

Remark 5.4. Let c be as in (2.11), with $\|.\| = \|.\|_\infty$. Even in this more convenient situation, it is not evident how the inequality $c \leq R(A,b)$, claimed by Theorem 2.5, could follow from Theorem 5.2. The fact is that (2.11) (with $\|.\| = \|.\|_\infty$) does not imply (5.6), because, in general, monotonicity does *not* imply contractivity.

Remark 5.5. Suppose $c \leq R(A,b)$. Then Theorem 2.5 claims that (2.9) is valid so that c would certainly be a stepsize-coefficient for monotonicity, with respect to $\mathbb{R}^s$ and any norm on $\mathbb{R}^s$. Even this last property of c does not follow from a simple application of Theorem 5.2, because it is no obvious consequence of (5.5)—note that (2.4) (with $\mathbb{V} = \mathbb{R}^s$) does *not* imply (5.2) (with $\rho = 1/\tau_0$).

The above three remarks make clear that our Theorem 2.5 can be viewed as a variant of Theorem 5.2 covering essentially new situations.

5.2. The proof of Theorem 2.5.

5.2.1. Preliminaries. Throughout this section 5.2 we assume, unless specified otherwise, that (A,b), c, and $|\!|.|\!|$ are as explained at the beginning of Theorem 2.5. With no loss of generality, we assume that c is finite. Below we shall prove the theorem by showing that the following five implications are valid: (2.8) $\Longrightarrow$ (2.9), (2.9) $\Longrightarrow$ (2.10), (2.10) $\Longrightarrow$ (2.11), [(2.11) with $|\!|.|\!| = \|.\|_{TV}$] $\Longrightarrow$ [(2.11) with $|\!|.|\!| = \|.\|_1$], and finally [(2.11) with $|\!|.|\!| = \|.\|_1$ or $\|.\|_\infty$] $\Longrightarrow$ (2.8).

The first implication will be proved in section 5.2.2, using arguments which are analogous to arguments for proving that (5.4) implies (5.5) (see Kraaijevanger (1991, pp. 502–504)).

The second implication is trivial, whereas the third and fourth implication will be proved in section 5.2.3. The proofs, in this section, are *not* related to arguments used in Kraaijevanger (1991), but are based on Lemma 5.6. This lemma gives a general framework in which the property of c being a stepsize-coefficient for monotonicity

can be carried over from a space $\mathbb{Y}$ with seminorm $\|.\|_{\mathbb{Y}}$ to another space $\mathbb{X}$ with seminorm $\|.\|_{\mathbb{X}}$.

The proof of the fifth implication will be given in section 5.2.4.

In that section we shall first deal with a linear variant of process (2.2). Lemma 5.7 tells us that a monotonicity property of that variant implies (2.8); the lemma is relevant to the norms $\|.\|_p$, with $p = 1$ and $p = \infty$. This lemma, with value $p = \infty$, was used implicitly by Kraaijevanger (1991, pp. 507–508) in a proof related to the implication (5.6) $\Longrightarrow$ (5.4) (cf. Theorem 5.2).

Next, we shall give Lemma 5.8, which states that property (2.11), with $|\!|.|\!| = \|.\|_p$ and $p = 1$ or $p = \infty$, implies the monotonicity property of the linear variant considered in Lemma 5.7. A combination of Lemmas 5.7 and 5.8 proves the fifth implication. Our proof of Lemma 5.8 has no relation to arguments in Kraaijevanger (1991); it makes use, among other things, of arguments employed earlier in Spijker (1986).

For completeness we mention that no counterpart of Lemma 5.8 is known to the authors which is relevant to contractivity with respect to $\mathbb{R}^s$ and $\|.\|_1$—cf. Remark 5.3 and Kraaijevanger (1991, p. 505).

5.2.2. Statement (2.8) $\Longrightarrow$ statement (2.9). We start this subsection by introducing some notation relevant to the vector space $\mathbb{V}$. For any vectors $v_1, v_2, \ldots, v_m$ in $\mathbb{V}$, we shall denote the vector in $\mathbb{V}^m$ with components v_j by

$$v = [v_j] = \begin{pmatrix} v_1 \\ \vdots \\ v_m \end{pmatrix} \in \mathbb{V}^m.$$

Furthermore, for any (real) $l \times m$ matrix $B = (b_{ij})$, we define a corresponding linear operator $B_{\mathbb{V}}$, from $\mathbb{V}^m$ to $\mathbb{V}^l$, by $B_{\mathbb{V}}(v) = w$, for $v = [v_j] \in \mathbb{V}^m$, where $w = [w_i] \in \mathbb{V}^l$ with $w_i = \sum_{j=1}^m b_{ij} v_j$ $(1 \le i \le l)$. Clearly, if B and C are $l \times m$ matrices and D is an $m \times k$ matrix, then $(B + C)_{\mathbb{V}} = B_{\mathbb{V}} + C_{\mathbb{V}}$, $(\lambda B)_{\mathbb{V}} = \lambda \cdot B_{\mathbb{V}}$, and $(BD)_{\mathbb{V}} = B_{\mathbb{V}} \cdot D_{\mathbb{V}}$. Here, the addition and multiplications occurring in the last three left-hand members stand for the usual algebraic operations for matrices, whereas the addition and multiplications in the right-hand members apply to linear operators. The last three equalities will underlie part of our subsequent calculations.

Assume (2.8), and let F be a function from $\mathbb{V}$ to $\mathbb{V}$ satisfying (2.4). We have to prove that c is a stepsize-coefficient for monotonicity; i.e., $0 < \Delta t \le c \cdot \tau_0$ implies $\|u_n\| \le \|u_{n-1}\|$ whenever u_n and u_{n-1} are related to each other by (2.2).

Assuming (2.2), with $0 < \Delta t \le c \cdot \tau_0$, we obtain

$$u_n = u_{n-1} + \sum_{j=1}^m b_j w_j, \tag{5.7a}$$

$$y_i = u_{n-1} + \sum_{j=1}^m a_{ij} w_j \quad (1 \le i \le m), \tag{5.7b}$$

where $w_j = \Delta t F(y_j)$. Putting $\gamma = \Delta t/\tau_0$, we have $\|w_i + c y_i\| = \gamma \|(c/\gamma) y_i + \tau_0 F(y_i)\| \le \gamma\{(c/\gamma - 1)\|y\| + \|y_i + \tau_0 F(y_i)\|\}$. Therefore, in view of (2.4),

$$\|w_i + c y_i\| \le c \|y_i\|. \tag{5.8}$$

Defining $y = [y_i] \in \mathbb{V}^m$, $w = [w_i] \in \mathbb{V}^m$, and $e = (1, \ldots, 1)^T \in \mathbb{R}^m$, we can rewrite (5.7) as

$$u_n = u_{n-1} + \mathbf{b}^T w, \tag{5.9a}$$

$$y = \mathbf{e}u_{n-1} + \mathbf{A}w, \tag{5.9b}$$

where $\mathbf{b}^T = (b^T)_{\mathbb{V}}$, $\mathbf{e} = (e)_{\mathbb{V}}$, and $\mathbf{A} = A_{\mathbb{V}}$. Denoting the identity in $\mathbb{V}^m$ by $\mathbf{I}$, we see from (5.9b) that $(\mathbf{I} + c\mathbf{A})y = \mathbf{e}u_{n-1} + \mathbf{A}w + c\mathbf{A}y = \mathbf{e}u_{n-1} + \mathbf{A}(w + cy)$. From Lemma 4.5, we conclude that (2.6) holds with $\xi = -c$ and that $A \geq 0$. Therefore, $\mathbf{I} + c\mathbf{A}$ is invertible and

$$y = (\mathbf{I} + c\mathbf{A})^{-1}\mathbf{e}u_{n-1} + \mathbf{A}(\mathbf{I} + c\mathbf{A})^{-1}(w + cy). \tag{5.10}$$

Since $(I + cA)^{-1}e = e(-c) \geq 0$ and $A(I + cA)^{-1} = A(-c) \geq 0$ we arrive at the inequality $[\|y_i\|] \leq \|u_{n-1}\|(I + cA)^{-1}e + A(I + cA)^{-1}[\|w_i + cy_i\|]$. In view of (5.8), there follows $[\|y_i\|] \leq \|u_{n-1}\|(I + cA)^{-1}e + cA(I + cA)^{-1}[\|y_i\|]$, which is the same as $(I + cA)^{-1}[\|y_i\|] \leq \|u_{n-1}\|(I + cA)^{-1}e$. Multiplying the last inequality by the matrix $I + cA \geq 0$, we can conclude that

$$\|y_i\| \leq \|u_{n-1}\| \quad (1 \leq i \leq m). \tag{5.11}$$

Using (5.9a), (5.10), we obtain

$$\begin{aligned} u_n &= u_{n-1} + \mathbf{b}^T w = u_{n-1} - c\mathbf{b}^T y + \mathbf{b}^T(w + cy) \\ &= u_{n-1} - c\mathbf{b}^T\{(\mathbf{I} + c\mathbf{A})^{-1}\mathbf{e}u_{n-1} + \mathbf{A}(\mathbf{I} + c\mathbf{A})^{-1}(w + cy)\} + \mathbf{b}^T(w + cy) \\ &= \{1 - cb^T(I + cA)^{-1}e\}u_{n-1} + \mathbf{b}^T(\mathbf{I} + c\mathbf{A})^{-1}(w + cy). \end{aligned}$$

Since $\varphi(-c) \geq 0$, $b(-c) \geq 0$, and (5.8), (5.11) are valid, we see from the last expression for u_n that

$$\begin{aligned} \|u_n\| &\leq \{1 - cb^T(I + cA)^{-1}e\}\|u_{n-1}\| + b^T(I + cA)^{-1}[\|w_i + cy_i\|] \\ &\leq \{1 - cb^T(I + cA)^{-1}e\}\|u_{n-1}\| + (cb^T(I + cA)^{-1}e)\|u_{n-1}\| = \|u_{n-1}\|. \end{aligned}$$

This completes the proof of (2.9).

5.2.3. Statement (2.10) $\Longrightarrow$ statement(2.11); and statement (2.11) with $|.| = \|.\|_{TV} \Longrightarrow$ statement (2.11) with $|.| = \|.\|_1$. We start this subsection by giving Lemma 5.6. The lemma deals with a general situation where

(5.12a) $\mathbb{X}$ and $\mathbb{Y}$ are vector spaces, with seminorms $\|.\|_{\mathbb{X}}$ and $\|.\|_{\mathbb{Y}}$, respectively,

(5.12b) $S : \mathbb{X} \to \mathbb{Y}$ is a linear operator,

(5.12c) $Sx = 0$ only for $x = 0$, and

(5.12d) $\|x\|_{\mathbb{X}} = \|Sx\|_{\mathbb{Y}}$ (for all $x \in \mathbb{X}$).

LEMMA 5.6. *Assume* (5.12) *and let c be a stepsize-coefficient for monotonicity, with respect to $\mathbb{Y}$ and $\|.\|_{\mathbb{Y}}$. Then c is also a stepsize-coefficient for monotonicity, with respect to $\mathbb{X}$ and $\|.\|_{\mathbb{X}}$.*

Proof. Let Δt be a stepsize with $0 < \Delta t \leq c \cdot \tau_0$, and let $F : \mathbb{X} \to \mathbb{X}$ with

$$\|x + \tau_0 F(x)\|_{\mathbb{X}} \leq \|x\|_{\mathbb{X}} \quad (\text{on } \mathbb{X}). \tag{5.13a}$$

Suppose the relations (2.2) are fulfilled. We have to prove that

$$\|u_n\|_{\mathbb{X}} \leq \|u_{n-1}\|_{\mathbb{X}}. \tag{5.13b}$$

We define the subspace $\mathbb{Y}_0 = \{y : \ y = Sx \text{ for some } x \in \mathbb{X}\}$ and we introduce a linear transformation T, from $\mathbb{Y}_0$ onto $\mathbb{X}$, by $Ty = x$ (for $y = Sx \in \mathbb{Y}_0$).

In view of (2.2), the vector $v_n = Su_n$ is generated from $v_{n-1} = Su_{n-1}$ by applying the Runge–Kutta method to the function $G_0 : \mathbb{Y}_0 \to \mathbb{Y}_0$, defined by $G_0(y) = SFT(y)$ (for $y \in \mathbb{Y}_0$). Using (5.12d) and (5.13a), one easily sees that $\|y + \tau_0 G_0(y)\|_{\mathbb{Y}} \le \|y\|_{\mathbb{Y}}$ (for all $y \in \mathbb{Y}_0$).

We define $G : \mathbb{Y} \to \mathbb{Y}$ by $G(y) = G_0(y)$ (for $y \in \mathbb{Y}_0$) and $G(y) = 0$ (for $y \in \mathbb{Y} \setminus \mathbb{Y}_0$). Clearly $\|y + \tau_0 G(y)\|_{\mathbb{Y}} \le \|y\|_{\mathbb{Y}}$ (for all $y \in \mathbb{Y}$). Moreover, the vector v_n can be viewed as being generated from v_{n-1} by applying the Runge–Kutta method, with stepsize Δt, to the function G. Consequently, $\|v_n\|_{\mathbb{Y}} \le \|v_{n-1}\|_{\mathbb{Y}}$. Combining this inequality and (5.12d), we arrive at (5.13b). □

Now assume (2.10). We shall prove (2.11) by applying Lemma 5.6.

We define $\mathbb{X} = \mathbb{R}^s$, $\mathbb{Y} = \{y : \ y \in \mathbb{R}^\infty$, and $\left|y\right| < \infty\}$, and $\|x\|_{\mathbb{X}} = \left|x\right|$, $\|y\|_{\mathbb{Y}} = \left|y\right|$ (for $x \in \mathbb{X}$ and $y \in \mathbb{Y}$, respectively). Furthermore, we introduce the operator S by

$$Sx = \begin{cases} (\dots, 0, 0, x_1, x_2, .., x_s, 0, 0 \dots) & \text{if } \ \left|.\right| = \|.\|_\infty \text{ or } \|.\|_1, \\ (\dots, x_1, x_1, x_1, x_2, \dots, x_s, x_s, x_s \dots) & \text{if } \ \left|.\right| = \|.\|_{TV} \end{cases}$$

for $x = (x_1, x_2, \dots, x_s) \in \mathbb{X}$.

With these definitions, the conditions (5.12) are fulfilled. In view of (2.10), we can apply Lemma 5.6 so as to conclude that (2.11) holds.

Finally assume (2.11) with $\left|.\right| = \|.\|_{TV}$. Let $s \ge 1$ and $\mathbb{X} = \mathbb{R}^s$, $\|x\|_{\mathbb{X}} = \|x\|_1$ (for $x \in \mathbb{X}$). We want to prove that c is a stepsize-coefficient for monotonicity with respect to $\mathbb{X}$ and $\|.\|_{\mathbb{X}}$.

In order to be able to apply Lemma 5.6 to the situation at hand, we define $\mathbb{Y} = \mathbb{R}^{s+1}$, $\|y\|_{\mathbb{Y}} = \|y\|_{TV}$ (for $y \in \mathbb{Y}$). Furthermore, for $x = (x_1, x_2, \dots, x_s) \in \mathbb{X}$ we define $Sx = (y_1, \dots, y_{s+1})$ with $y_1 = 0$ and $y_i = x_1 + x_2 + \cdots + x_{i-1}$ (for $2 \le i \le s+1$).

One easily sees that, with the above definitions, all assumptions of Lemma 5.6 are fulfilled. Hence, c has the required property.

5.2.4. (2.11) with $\left|.\right| = \|.\|_1$ or $\|.\|_\infty \implies$ (2.8). Throughout this subsection we shall use, for $p = 1, \infty$ and $s \times s$ matrices G, the notation $\|G\|_p = \max \|Gv\|_p / \|v\|_p$, where the maximum is over all nonzero vectors v in $\mathbb{R}^s$. Furthermore, we shall denote the $s \times s$ identity matrix by I.

Let $G_1, G_2, \dots, G_m$ be given $s \times s$ matrices. We consider a linear variant of (2.2) (with $n = 1$, $u_0 \in \mathbb{V} = \mathbb{R}^s$) in which all vectors $F(y_j)$ are replaced by $G_j\, y_j$. Furthermore, we consider the following linear variant of condition (2.4): $\|I + \tau_0 G_i\|_p \le 1$ $(1 \le i \le m)$.

Choose $\Delta t = c\,\tau_0$ and write $Z_i = \Delta t G_i$. Then the above linear variants of (2.2) and (2.4), respectively, can be written in the form

$$u_1 = u_0 + \sum_{j=1}^{m} b_j Z_j y_j, \tag{5.14a}$$

$$y_i = u_0 + \sum_{j=1}^{m} a_{ij} Z_j y_j \quad (1 \le i \le m), \tag{5.14b}$$

and

$$\|cI + Z_i\|_p \le c \quad (1 \le i \le m). \tag{5.15}$$

In the following we shall focus on ordered m-tuples $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_m)$, where the Z_i are $s \times s$ matrices, such that (5.15) holds and the system of equations (5.14b) has a unique solution $y_1, y_2, \ldots, y_m$. The set consisting of all of these $\mathbf{Z}$ will be denoted by $\mathscr{D}_p(c, s)$.

For any $\mathbf{Z}$ in $\mathscr{D}_p(c, s)$, the vector u_1 in (5.14) depends uniquely and linearly on u_0; we denote the $s \times s$ matrix transforming u_0 into u_1 by $\mathbf{K}(\mathbf{Z})$. We thus have

$$u_1 = \mathbf{K}(\mathbf{Z})u_0 \text{ whenever } \mathbf{Z} \in \mathscr{D}_p(c, s) \text{ and } u_0, u_1 \in \mathbb{R}^s \text{ satisfy (5.14).} \tag{5.16}$$

The inequality

$$\|\mathbf{K}(\mathbf{Z})\|_p \leq 1 \text{ (for all } \mathbf{Z} \in \mathscr{D}_p(c, s) \text{ and } s \geq 1) \tag{5.17}$$

amounts to a monotonicity condition on process (5.14). It will be related to (2.8) and to (2.11) in Lemmas 5.7 and 5.8, respectively.

LEMMA 5.7. *Consider an arbitrary irreducible Runge–Kutta scheme* (A, b)*, and let* $p = 1$ *or* $p = \infty$*. Let* $0 < c < \infty$*, and assume condition* (5.17) *is fulfilled. Then* c *satisfies* (2.8).

Proof. In Kraaijevanger (1991) this lemma was proved (implicitly) for $p = \infty$. The proof in that paper is long and technical but is presented in a very clear way. Therefore, we do not repeat it here but note that the actual proof (given on pp. 507–508 of the paper) consists in a combination of conclusions regarding absolute monotonicity (on pp. 485–496) with Lemma 5.10 (on p. 505). The conclusions stated on pp. 485–496 are independent of the norm in $\mathbb{R}^s$, whereas Lemma 5.10 is tuned to the special norm $\|.\|_\infty$. It is not difficult to adapt the proof of the last mentioned lemma to the norm $\|.\|_1$ so as to conclude that Lemma 5.10 is verbatim valid for $\|.\|_1$ as well. As a result, the arguments in Kraaijevanger (1991, pp. 507–508) prove our Lemma 5.7 also for $p = 1$. □

A combination of the following lemma and Lemma 5.7 immediately leads to the desired implication ((2.11) with $|.| = \|.\|_1$ or $\|.\|_\infty \Longrightarrow$ (2.8)).

LEMMA 5.8. *Consider an arbitrary irreducible Runge–Kutta schema* (A, b)*, and let* $p = 1$ *or* $p = \infty$*. Let* $0 < c < \infty$*, and assume* (2.11) *with* $|.| = \|.\|_p$*. Then condition* (5.17) *is fulfilled.*

Proof. The proof will be given in three steps.

Step 1. Let

$$s \geq 1,\ u_0 \in \mathbb{R}^s,\ \mathbf{Z} = (Z_1, \ldots, Z_m) \in \mathscr{D}_p(c, s), \tag{5.18}$$

and assume that the corresponding vectors y_i, defined by (5.14b), satisfy

$$y_i \neq y_j \quad (\text{for } i \neq j). \tag{5.19}$$

We shall prove that

$$\|\mathbf{K}(\mathbf{Z})u_0\|_p \leq \|u_0\|_p. \tag{5.20}$$

Choose any $\tau_0 > 0$, and define $F : \mathbb{R}^s \to \mathbb{R}^s$ by $F(v) = (c\tau_0)^{-1} Z_i y_i$ (for $v = y_i$) and $F(v) = 0$ (for all other $v \in \mathbb{R}^s$). In view of (5.15), the function F satisfies (2.4) with $\mathbb{V} = \mathbb{R}^s$, $\|.\| = \|.\|_p$. Furthermore, we see from (5.14), (5.16) that the vector $\mathbf{K}(\mathbf{Z})u_0$ is generated from u_0 by applying the Runge–Kutta method with stepsize $\Delta t = c\tau_0$ to the function F. By virtue of (2.11) (with $|.| = \|.\|_p$), we conclude that (5.20) holds.

Step 2. Due to the restriction (5.19) in Step 1, the proof of (5.17) is not yet complete. Below, in Step 3, we shall get rid of this restriction by using (real) values γ_i, η_i (for $1 \le i \le m$) with the following properties:

$$0 < \gamma_i < c \quad (1 \le i \le m); \tag{5.21a}$$

$$\text{the } m \times m \text{ matrix } I + A \cdot \mathrm{diag}(\gamma_i) \text{ is invertible;} \tag{5.21b}$$

$$\eta_i = 1 - \sum_{j=1}^{m} a_{ij}\gamma_j\eta_j \quad (1 \le i \le m); \tag{5.21c}$$

$$\eta_i \ne \eta_j \quad (\text{whenever } i \ne j). \tag{5.21d}$$

In this (second) step we shall prove the existence of γ_i, η_i satisfying (5.21).

Since (A, b) is irreducible, statement (ii) (of Definition 2.2) is not true. It follows that the polynomials $p_i(t) = \sum_{j=1}^{m} a_{ij}t^j$ are different from each other. Therefore, there is a positive t_0 with $p_i(t_0) \ne p_j(t_0)$ (for all $i \ne j$). Writing $t_i = (t_0)^i$, we thus have

$$\sum_{k=1}^{m} a_{ik}t_k \ne \sum_{k=1}^{m} a_{jk}t_k \quad (\text{whenever } i \ne j).$$

Let $\gamma_i = \lambda t_i$, with $\lambda > 0$. We choose λ sufficiently small to guarantee (5.21a) and (5.21b). The corresponding values $\eta_i = \eta_i(\lambda)$, solving (5.21c), satisfy

$$\eta_i(\lambda) = 1 - \lambda \sum_{k=1}^{m} a_{ik}t_k + O(\lambda^2) \quad (\text{for } \lambda \downarrow 0).$$

Choosing λ sufficiently small, we conclude that γ_i, η_i exist satisfying (5.21).

Step 3. Assume (5.18). We shall prove (5.20).

Let y_i satisfy (5.14b), and choose any γ_i, η_i as in (5.21). We choose $\varepsilon > 0$ and define

$$u_0^* = \begin{pmatrix} u_0 \\ \varepsilon \end{pmatrix}, \qquad Z_i^* = \begin{pmatrix} Z_i & 0 \\ 0 & -\gamma_i \end{pmatrix}, \qquad y_i^* = \begin{pmatrix} y_i \\ \varepsilon\eta_i \end{pmatrix}.$$

Since $\mathbf{Z} \in \mathscr{D}_p(c, s)$ and (5.21a), (5.21b) hold, the m-tuple $\mathbf{Z}^* = (Z_1^*, Z_2^*, \dots, Z_m^*)$ belongs to $\mathscr{D}_p(c, s+1)$. Furthermore, $y_i^* = u_0^* + \sum_{j=1}^{m} a_{ij}Z_j^*y_j^*$ $(1 \le i \le m)$ and $y_i^* \ne y_j^*$ (for $i \ne j$). Consequently, the conclusion of the above Step 1 can be applied (to $u_0^* \in \mathbb{R}^{s+1}$ and $\mathbf{Z}^* \in \mathscr{D}_p(c, s+1)$) so as to obtain $\|\mathbf{K}(\mathbf{Z}^*)u_0^*\|_p \le \|u_0^*\|_p$.

Since $\|\mathbf{K}(\mathbf{Z})u_0\|_p \le \|\mathbf{K}(\mathbf{Z}^*)u_0^*\|_p$ and $\|u_0^*\|_p \le \|u_0\|_p + \varepsilon$, we arrive at (5.20) by letting $\varepsilon \to 0$. □

Acknowledgments. The authors are most thankful to Dr. W. H. Hundsdorfer for useful discussions and information regarding the topic of this paper. Moreover, they are indebted to three anonymous referees for constructive criticism regarding an earlier version of the paper.

REFERENCES

K. Burrage and J. C. Butcher (1980), *Non-linear stability of a general class of differential equation methods*, BIT, 20, pp. 185–203.

J. C. Butcher (1987), *The Numerical Analysis of Ordinary Differential Equations*, John Wiley, Chichester, UK.

K. Dekker and J. G. Verwer (1984), *Stability of Runge-Kutta Methods for Stiff Nonlinear Differential Equations*, North–Holland, Amsterdam.

S. Gottlieb and C.-W. Shu (1998), *Total-variation-diminishing Runge-Kutta schemes*, Math. Comp., 67, pp. 73–85.

S. Gottlieb, C.-W. Shu, and E. Tadmor (2001), *Strong-stability-preserving high-order time discretization methods*, SIAM Rev., 43, pp. 89–112.

E. Hairer and G. Wanner (1996), *Solving Ordinary Differential Equations* II. *Stiff and Differential-Algebraic Problems*, 2nd revised ed., Springer-Verlag, Berlin.

A. Harten (1983), *High resolution schemes for hyperbolic conservation laws*, J. Comput. Phys., 49, pp. 357–393.

W. Hundsdorfer, S. J. Ruuth, and R. J. Spiteri (2003), *Monotonicity-preserving linear multistep methods*, SIAM J. Numer. Anal., 41, pp. 605–623.

W. Hundsdorfer and J. G. Verwer (2003), *Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations*, Springer Ser. Comput. Math. 33, Springer-Verlag, Berlin.

R. Jeltsch and O. Nevanlinna (1981), *Stability of explicit time discretizations for solving initial value problems*, Numer. Math., 37, pp. 61–91.

C. K. Kennedy, M. H. Carpenter, and R. M. Lewis (2000), *Low-storage, explicit Runge-Kutta schemes for the compressible Navier-Stokes equations*, Appl. Numer. Math, 35, pp. 177–219.

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

J. F. B. M. Kraaijevanger (1986), *Absolute monotonicity of polynomials occurring in the numerical solution of initial value problems*, Numer. Math., 48, pp. 303–322.

D. Kröner (1997), *Numerical Schemes for Conservation Laws*, Wiley, Chichester, UK, and Teubner, Stuttgart, Germany.

C. B. Laney (1998), *Computational Gasdynamics*, Cambridge University Press, Cambridge, UK.

R. J. LeVeque (2002), *Finite Volume Methods for Hyperbolic Problems*, Cambridge University Press, Cambridge, UK.

K. W. Morton (1980), *Stability of difference approximations to a diffusion-convection equation*, Internat. J. Numer. Methods Engrg., 15, pp. 677–683.

S. Ruuth and R. Spiteri (2002), *Two barriers on strong-stability-preserving time discretization methods*, J. Sci. Comput., 17, pp. 211–220.

S. O. Schönbeck (1967), *On the extension of Lipschitz maps*, Ark. Mat., 7, pp. 201–209.

C.-W. Shu (2002), *A survey of strong stability preserving high-order time discretizations*, in Collected Lectures on the Preservation of Stability Under Discretization, D. Estep and S. Tavener, eds., SIAM, Philadelphia, pp. 51–65.

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

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

M. N. Spijker (1986), *Monotonicity and boundedness in implicit Runge-Kutta methods*, Numer. Math., 50, pp. 97–109.

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

E. F. Toro (1999), *Riemann Solvers and Numerical Methods for Fluid Dynamics*, 2nd ed., Springer-Verlag, Berlin.