Monotonicity and Boundedness in general Runge-Kutta methods

Ferracina, L.

Chapter IV

Stepsize restrictions for total-variation-boundedness in general Runge-Kutta procedures

The contents of this chapter are equal to: Ferracina L., Spijker M.N. (2005): Stepsize restrictions for total-variation-boundedness in general Runge-Kutta procedures, *Appl. Numer. Math.* **53**, 265–279.

Abstract

In the literature, on the numerical solution of nonlinear time dependent partial differential equations, much attention has been paid to numerical processes which have the favourable property of being total variation bounded (TVB). A popular approach to guaranteeing the TVB property consists in demanding that the process has the stronger property of being total variation diminishing (TVD).

For Runge-Kutta methods - applied to semi-discrete approximations of partial differential equations - conditions on the time step were established which guarantee the TVD property; see e.g. Shu & Osher (1988), Gottlieb & Shu (1998), Gottlieb, Shu & Tadmor (2001), Ferracina & Spijker (2004), Higueras (2004), Spiteri & Ruuth (2002). These conditions were derived under the assumption that the simple explicit Euler time stepping process is TVD.

However, for various important semi-discrete approximations, the Euler process is TVB but *not* TVD - see e.g. Shu (1987), Cockburn & Shu (1989). Accordingly, the above stepsize conditions for Runge-Kutta methods are not directly relevant

to such approximations, and there is a need for stepsize restrictions with a wider range of applications.

In this paper, we propose a general theory yielding stepsize restrictions which cover a larger class of semi-discrete approximations than covered thus far in the literature. In particular, our theory gives stepsize restrictions, for general Runge-Kutta methods, which guarantee total-variation-boundedness in situations where the Euler process is TVB but not TVD.

1 Introduction

1.1 The purpose of the paper

In this paper we deal with the numerical solution of initial value problems (IVPs), for systems of ordinary differential equations (ODEs), which can be written in the form

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

The general Runge-Kutta method, applied to problem (1.1), provides us with numerical approximations u_n to $U(n\Delta t)$, where Δt denotes a positive time step and $n = 1, 2, 3, ...$; see e.g. Hairer, Nørsett & Wanner (1993), Hairer & Wanner (1996), Butcher (2003), Hundsdorfer & Verwer (2003). The approximations u_n are defined in terms of u_{n-1} by the relations

$$y_i = u_{n-1} + \Delta t \sum_{j=1}^{m} a_{ij} F(y_j) \quad (1 \leq i \leq m), \tag{1.2.a}$$

$$u_n = u_{n-1} + \Delta t \sum_{j=1}^{m} b_j F(y_j). \tag{1.2.b}$$

Here a_{ij} and b_j are real parameters, specifying the Runge-Kutta method, and y_i are intermediate approximations needed for computing u_n from u_{n-1}. As usual, we assume that $b_1 + b_2 + ... + b_m = 1$, and we call the Runge-Kutta method *explicit* if $a_{ij} = 0$ (for $j \geq i$). We define the $m \times m$ matrix A by $A = (a_{ij})$ and the column vector $b \in \mathbb{R}^m$ by $b = (b_1, b_2, b_3, ..., b_m)^T$, so that we can identify the Runge-Kutta method with its *coefficient scheme* (A, b).

In order to introduce the questions to be studied in this paper, we assume that (1.1) results from applying the method of lines (MOL) to a Cauchy problem for a partial differential equation (PDE) of the form

$$\frac{\partial}{\partial t}u(x,t) + \frac{\partial}{\partial x}f(u(x,t)) = 0 \quad (t \geq 0, \quad -\infty < x < \infty). \tag{1.3}$$

Here f stands for a given (possibly nonlinear) scalar function, so that the PDE is a simple instance of a conservation law. In this situation, the function F occurring

in (1.1) can be regarded as a function from

$$\mathbb{R}^\infty = \{y : y = (..., \eta_{-1}, \eta_0, \eta_1, ...) \text{ with } \eta_j \in \mathbb{R} \text{ for } j = 0, \pm 1, \pm 2, ...\}$$

into itself; it depends on the given function f as well as on the process of semi-discretization being used. Further, $u_0 \in \mathbb{R}^\infty$ depends on the initial data of the original Cauchy problem. The solution $U(t)$ to (1.1) now stands for a (time dependent) vector in $\mathbb{R}^\infty$ with components $U_j(t)$ which are to approximate the desired true solution values $u(x_j, t)$ (or cell averages thereof) corresponding to grid points x_j $(j = 0, \pm 1, \pm 2, ...)$. For detailed explanations of the MOL, see e.g. Laney (1998), Toro (1999), LeVeque (2002), Hundsdorfer & Verwer (2003).

In the situation just specified, where (1.1) stands for a semi-discrete version of a conservation law, it is desirable that the corresponding (fully discrete) process (1.2) has a property which is referred to in the literature as *total variation boundedness (TVB)*. In discussing this property, we shall use below the total variation seminorm $\|.\|_{TV}$ and the vector space $\mathbb{R}^\infty_{TV}$, which are defined as follows:

$$\begin{aligned} \|y\|_{TV} &= \sum_{j=-\infty}^{+\infty} |\eta_j - \eta_{j-1}| \quad (\text{for } y \in \mathbb{R}^\infty \text{ with components } \eta_j), \\ \mathbb{R}^\infty_{TV} &= \{y : \ y \in \mathbb{R}^\infty \text{ and } \|y\|_{TV} < \infty\}. \end{aligned}$$

Total variation boundedness of process (1.2) means that, for initial vector $u_0 \in \mathbb{R}^\infty_{TV}$ and $T > 0$, there is a positive constant B and value $\Delta t_0 > 0$ such that

$$\|u_n\|_{TV} \leq B \quad (0 < \Delta t \leq \Delta t_0, \quad 0 < n\Delta t \leq T). \tag{1.4}$$

For more details and an explanation of the importance of the TVB property in the numerical solution of nonlinear conservation laws, in particular in the context of convergence proofs, see e.g. Harten (1984), Shu (1987), Cockburn & Shu (1989), Kröner (1997), Laney (1998), LeVeque (2002).

A popular approach to guaranteeing the TVB property, consists in demanding that the total variation be non-increasing as time evolves, so that, at any positive time level, the total variation of the approximate solution u_n is bounded by the total variation of the initial vector u_0. Following the terminology in the literature, we will say that process (1.2) is *total variation diminishing (TVD)* if

$$\|u_n\|_{TV} \leq \|u_{n-1}\|_{TV}, \quad \text{for } u_n \text{ and } u_{n-1} \text{ satisfying (1.2)}. \tag{1.5}$$

In the literature, crucial stepsize restrictions of the form

$$0 < \Delta t \leq \Delta t_0 \tag{1.6}$$

were given ensuring the TVD property (1.5); see e.g. Shu (1988), Shu & Osher (1988), Gottlieb & Shu (1998), Gottlieb, Shu & Tadmor (2001), Ferracina & Spijker

(2004), Higueras (2004), Spiteri & Ruuth (2002) and Section 2.2 below. These stepsize restrictions were derived under the assumption that, for some positive τ_0,

$$(1.7)\quad F:\ \mathbb{R}^{\infty}_{TV} \longrightarrow \mathbb{R}^{\infty}_{TV} \quad \text{satisfies} \quad \|v+\tau_0\,F(v)\|_{TV}\ \le\ \|v\|_{TV} \quad (v\in\mathbb{R}^{\infty}_{TV}).$$

Clearly, (1.7) amounts to assuming that the semi-discretization of equation (1.3) has been performed in such a manner that the simple forward Euler method, applied to problem (1.1), is TVD for some suitably chosen stepsize τ_0.

Unfortunately, for important semi-discrete versions (1.1) of (1.3), condition (1.7) is *not* fulfilled see e.g. Shu (1987), Cockburn & Shu (1989). Clearly, in such cases the above stepsize restrictions (1.6), which are relevant to the situation (1.7), do not allow us to conclude that a Runge-Kutta procedure is TVD (and therefore TVB).

We note that a notorious weakness, of most TVD schemes, is that their accuracy degenerates to first order at smooth extrema of the solution - see e.g. Osher & Chakravarthy (1984). The semi-discretizations just mentioned, proposed by Shu (1987), Cockburn & Shu (1989) and others, were introduced to overcome this weakness. Although, for these semi-discretizations, condition (1.7) is violated, the following weaker condition is fulfilled:

$$(1.8)\quad F:\mathbb{R}^{\infty}_{TV} \to \mathbb{R}^{\infty}_{TV} \text{ satisfies } \|v+\tau_0F(v)\|_{TV}\le(1+\alpha_0\tau_0)\|v\|_{TV}+\beta_0\tau_0 \quad (v\in\mathbb{R}^{\infty}_{TV}).$$

Here τ_0 is again positive, and α_0, β_0 are nonnegative constants. Condition (1.8) can be interpreted, analogously to (1.7), as a bound on the increase of the total variation, when the explicit Euler time stepping is applied to (1.1) with time step τ_0.

In the situation where property (1.8) is present, it is natural to look for an analogous property in the general Runge-Kutta process (1.2), namely

$$(1.9)\quad \|u_n\|_{TV} \le (1+\alpha\Delta t)\|u_{n-1}\|_{TV} + \beta\Delta t, \quad \text{for } u_n \text{ and } u_{n-1} \text{ satisfying (1.2).}$$

Here α, β denote nonnegative constants.

Suppose (1.9) would hold under a stepsize restriction of the form (1.6). By applying (1.9) recursively and noting that $(1+\alpha\Delta t)^n \le \exp(\alpha\, n\Delta t)$, we then would obtain

$$(1.10)\quad \|u_n\|_{TV} \le e^{\alpha T}\|u_0\|_{TV} + \frac{\beta}{\alpha}(e^{\alpha T}-1) \quad (0<\Delta t\le\Delta t_0,\ \ 0<n\Delta t\le T).$$

Hence, *property* (1.9) (for $0<\Delta t\le\Delta t_0$) *amounts to total variation boundedness*, in that (1.4), is fulfilled with $B = e^{\alpha T}\|u_0\|_{TV} + \frac{\beta}{\alpha}(e^{\alpha T}-1)$. The last expression stands for $\|u_0\|_{TV} + \beta T$, in the special case where $\alpha = 0$.

Since (1.8) and (1.9) reduce to (1.7) and (1.5), respectively, when $\alpha_0 = \beta_0 = \alpha = \beta = 0$, it is natural to look for extensions, to the TVB context, of the results in the literature pertinent to the TVD property. More specifically, the natural

question arises of whether stepsize restrictions of the form (1.6) can be established which guarantee property (1.9) when condition (1.8) is fulfilled.

Partial results related to the last question, but no complete answers, were indicated, for special explicit Runge-Kutta methods, by Gottlieb, Shu & Tadmor (2001, Section 2.1), Shu (2002, Section 2).

The purpose of this paper is to propose a general theory by means of which the above question, as well as related ones, can completely be clarified.

1.2 Outline of the rest of the paper

In Section 2, we recall some concepts which are basic for the rest of the paper, and we give a short review of relevant results from the literature.

Section 2.1 deals with the concept of irreducibility of Runge-Kutta methods (A, b) and with Kraaijevanger's coefficient $R(A, b)$. Theorem 2.3 gives a condition which is necessary and sufficient in order that $R(A, b)$ is positive. This theorem will be used later in the Sections 3, 4 and 5.

Theorem 2.4, in Section 2.2, gives a stepsize condition of the form (1.6) which is known to be necessary and sufficient for the TVD property (1.5) under assumption (1.7). This condition is also known to be relevant to versions of properties (1.5), (1.7) which are more general, than the original properties, in that they involve an arbitrary vector space $\mathbb{V}$ with seminorm $\|.\|$, rather than $\mathbb{R}^{\infty}_{TV}$ and $\|.\|_{TV}$. Theorem 2.4 serves as a preparation and motivation for the material in Section 3.

In Section 3, we propose an extension of the theory reviewed in Section 2.2. Our extension is applicable in the situation where (a generalized version of) condition (1.8) is fulfilled.

In Section 3.1, we consider versions of (1.8), (1.9) in the context of arbitrary vector spaces $\mathbb{V}$ with seminorm $\|.\|$. Further, we introduce, for arbitrary Runge-Kutta methods (A, b), an important characteristic quantity, which we denote by $S(A, b)$. This quantity will play, together with $R(A, b)$, a prominent part in Section 3.2.

The latter section contains our main result, Theorem 3.2. This theorem is relevant to arbitrary Runge-Kutta methods (*not* necessarily explicit). It can be viewed as a convenient variant of Theorem 2.4 adapted to the situation where (1.5) and (1.7) are replaced by (1.9) and (1.8), respectively. Theorem 3.2 amply answers the question mentioned above at the end of Section 1.1. The proof of the theorem requires arguments different from those underlying Theorem 2.4. In fact, our proof of Theorem 3.2 relies substantially on the use of Lemma 3.6. This lemma, which is of independent interest, gives general upper bounds for the seminorms of vectors u_n, y_i satisfying (1.2). In order not to interrupt the presentation of our results, we have postponed the proof of the lemma to the last section of the paper.

In Section 4 we shortly present some applications and illustrations of Theorem 3.2 and Lemma 3.6.

In Section 5 we prove Lemma 3.6. Our proof is based on a convenient representation of general Runge-Kutta methods, which is of a similar type as considered

recently in Ferracina & Spijker (2005), Higueras (2003).

2 Kraaijevanger's coefficient and the TVD property

2.1 Irreducible Runge-Kutta methods and the coefficient $R(A, b)$

The following definition is of fundamental importance in the rest of our paper.

Definition 2.1 (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 $M \cup N = \{1, 2, ..., m\}$ 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 $M_1, M_2, ..., M_r$, with $1 \leq r < m$ and $M_1 \cup M_2 \cup ... \cup M_r = \{1, 2, ..., m\}$, 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 (1.2) 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 & Verwer (1984) or Hairer & 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 coefficient for arbitrary Runge-Kutta schemes (A, b) introduced by Kraaijevanger (1991). Following this author, we shall denote his coefficient by $R(A, b)$, and in defining it, we shall use, for real ξ, the following notations:

$$\begin{array}{ll} A(\xi) = A(I - \xi A)^{-1}, & b(\xi) = (I - \xi A)^{-T} b, \\ e(\xi) = (I - \xi A)^{-1} e, & \varphi(\xi) = 1 + \xi b^T (I - \xi A)^{-1} e. \end{array} \tag{2.1}$$

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

$$I - \xi A \text{ is invertible, } A(\xi) \geq 0,\ b(\xi) \geq 0,\ e(\xi) \geq 0, \text{ and } \varphi(\xi) \geq 0. \tag{2.2}$$

The first inequality in (2.2) should be interpreted entry-wise; the second and the third ones component-wise. Similarly, all inequalities for matrices and vectors occurring below are to be interpreted entry-wise and component-wise, respectively.

Definition 2.2 (The coefficient $R(A, b)$).
Let (A, b) be a given Runge-Kutta scheme. In case $A \geq 0$ and $b \geq 0$, we define

$$R(A, b) = \sup\{r : r \geq 0 \text{ and (2.2) holds for all } \xi \in [-r, 0]\}.$$

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

Definition 2.2 may suggest that it is difficult to determine $R(A, b)$ for given Runge-Kutta schemes (A, b). But, Kraaijevanger (1991) showed that it is relatively simple to decide whether $R(A, b) = 0$ or $R(A, b) = \infty$ and to compute numerically the value of $R(A, b)$ in the intermediate cases - see also Ferracina & Spijker (2004, 2005).

We give below a criterion for positivity of $R(A, b)$ due to Kraaijevanger (1991; Theorem 4.2). The criterion will be used later in proving Theorem 3.2, Lemma 3.6 and Theorem 4.1. In order to formulate the criterion concisely, we define for any $m \times m$ matrix $B = (b_{ij})$, the corresponding $m \times m$ *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 2.3 (Kraaijevanger's criterion for positivity of $R(A, b)$).
Let (A, b) be a given irreducible coefficient scheme. Then $R(A, b) > 0$ if and only if

$$A \geq 0, \quad b > 0 \quad \text{and} \quad \mathrm{Inc}(A^2) \leq \mathrm{Inc}(A). \tag{2.3}$$

2.2 Stepsize restrictions from the literature for the TVD property

In this subsection, we will review a known stepsize restriction, for property (1.5) and for a generalized version thereof.

In order to formulate this generalized version, we consider an arbitrary real vector space $\mathbb{V}$ with seminorm $\|.\|$ (i.e. $\|u + v\| \leq \|u\| + \|v\|$ and $\|\lambda v\| = |\lambda| \cdot \|v\|$ for all real λ and $u, v \in \mathbb{V}$). In this general setting, the following property (2.4) replaces (1.5):

$$\|u_n\| \leq \|u_{n-1}\|, \quad \text{for } u_n \text{ and } u_{n-1} \text{ satisfying (1.2).} \tag{2.4}$$

The above property (2.4) is important, also with seminorms $\|.\|$ different from $\|.\|_{TV}$, and also when solving certain differential equations different from conservation laws. In the recent literature, property (2.4) was studied extensively and referred to as *strong stability* or *monotonicity*, see e.g. Gottlieb, Shu & Tadmor (2001), Spiteri & Ruuth (2002), Ferracina & Spijker (2004), Hundsdorfer, Ruuth & Spiteri (2003), Hundsdorfer & Verwer (2003).

The following theorem gives a stepsize condition guaranteeing (1.5) under the assumption (1.7), as well as a stepsize condition for property (2.4) under the assumption that, for $\tau_0 > 0$,

$$F : \mathbb{V} \longrightarrow \mathbb{V} \quad \text{satisfies} \quad \|v + \tau_0 F(v)\| \leq \|v\| \quad (v \in \mathbb{V}). \tag{2.5}$$

The theorem deals with stepsize restrictions of the form

$$0 < \Delta t \leq \rho \cdot \tau_0, \tag{2.6}$$

where ρ denotes a positive factor. The following condition will play a prominent part:

$$\rho \leq R(A, b). \tag{2.7}$$

Theorem 2.4.
Consider an arbitrary irreducible Runge-Kutta method (A, b)*, and let* ρ *be any given positive factor. Then each of the following statements* (i) *and* (ii) *is equivalent to* (2.7)*.*

(i) *The stepsize restriction* (2.6) *implies property* (2.4)*, whenever* $\mathbb{V}$ *is real vector space, with seminorm* $\|.\|$*, and* F *satisfies* (2.5)*.*

(ii) *The stepsize restriction* (2.6) *implies the TVD property* (1.5) *whenever* F *satisfies* (1.7)*.*

The above theorem is an immediate consequence of Ferracina & Spijker (2004, Theorem 2.5).

Clearly, (i) is a-priori a stronger statement that (ii). Accordingly, the essence of Theorem 2.4 is that the (algebraic) property (2.7) implies the (strong) statement (i), whereas already the (weaker) statement (ii) implies (2.7).

3 TVB Runge-Kutta processes

3.1 Preliminaries

In the present Section 3 we shall focus on stepsize conditions for property (1.9) and for a generalized version thereof.

In formulating this generalized version, we deal, similarly as in Section 2.2, with an arbitrary real vector space $\mathbb{V}$ with seminorm $\|.\|$. In this setting, the following property (3.1) corresponds to the TVB property (1.9):

$$\|u_n\| \leq (1 + \alpha\Delta t)\|u_{n-1}\| + \beta\Delta t \quad \text{for } u_n \text{ and } u_{n-1} \text{ satisfying (1.2).} \tag{3.1}$$

Here α and β denote again nonnegative constants.

The following condition (3.2) amounts to a natural generalization of (1.8) to the situation at hand:

$$F : \mathbb{V} \longrightarrow \mathbb{V} \quad \text{satisfies} \quad \|v + \tau_0 F(v)\| \leq (1 + \alpha_0\tau_0)\|v\| + \beta_0\tau_0 \quad (v \in \mathbb{V}). \tag{3.2}$$

Here τ_0 is again positive, and α_0, β_0 are nonnegative constants. This condition was also considered recently in Hundsdorfer & Ruuth (2004), in connection to

boundedness properties of linear multistep methods. Clearly, (3.1) and (3.2) reduce to (2.4) and (2.5), respectively, in case $\alpha = \beta = \alpha_0 = \beta_0 = 0$.

The above Theorem 2.4 shows that, in the situations (i) and (ii) of the theorem, the crucial stepsize restriction is of the form (2.6), with ρ satisfying (2.7). In the situation, where (3.2) or (1.8) is in force, the crucial stepsize restriction for property (3.1) or (1.9), respectively, will turn out to be less simple. In fact, not only the coefficient $R(A,b)$ will play a role, but also the quantity $S(A,b)$ defined below.

Definition 3.1 (The coefficient $S(A,b)$).
Let (A,b) be a given Runge-Kutta scheme. Then

$$S(A,b) = \sup\{r \ : \ r > 0 \text{ and } I - \xi A \text{ is invertible for all } \xi \in [0,r]\}.$$

We note that the quantity $S(A,b)$ allows of a simple interpretation by looking at the special function $F(v) = \alpha_0 v$, with $\alpha_0 > 0$: for this function, the system (1.2.a) has a proper solution, when $0 < \Delta t \leq \Delta t_0$, if and only if the product $\alpha_0 \Delta t_0$ is smaller than the above value $S(A,b)$.

3.2 Formulation and proof of the main result

The following Theorem 3.2 constitutes the main result of this paper. It can be viewed as a convenient variant of Theorem 2.4 which is applicable in the situations (1.8), (3.2), which were not yet covered by the latter theorem. Theorem 3.2 gives stepsize restrictions guaranteeing (1.9) and (3.1), respectively, under the assumptions (1.8) and (3.2). These restrictions are of the form

$$0 < \Delta t \leq \min\{\rho \cdot \tau_0, \, \sigma/\alpha_0\}, \tag{3.3}$$

where ρ and σ are positive factors and τ_0, α_0 are as in (1.8), (3.2). Note that, in case $\alpha_0 = 0$, condition (3.3) neatly reduces to (2.6). The following conditions on ρ and σ will play a crucial role:

$$\rho \leq R(A,b) \quad \text{and} \quad \sigma < S(A,b). \tag{3.4}$$

Theorem 3.2 (Main Theorem).
Consider an arbitrary irreducible Runge-Kutta method (A,b), and let ρ, σ be any given positive values. Then each of the following statements (I) *and* (II) *is equivalent to* (3.4).

(I) *There exists a finite γ such that the stepsize restriction* (3.3) *implies property* (3.1) *with $\alpha = \gamma\alpha_0$, $\beta = \gamma\beta_0$, whenever $\mathbb{V}$ is a real vector space with seminorm $\|.\|$ and F satisfies* (3.2).

(II) *There exists a finite γ such that the stepsize restriction* (3.3) *implies the TVB property* (1.9) *with $\alpha = \gamma\alpha_0$, $\beta = \gamma\beta_0$, whenever F satisfies* (1.8).

The proof of Theorem 3.2 will be given at the end of this section, by using the important Lemma 3.6 to be formulated below.

Remark 3.3. Clearly, (I) is a-priori a stronger statement than (II). The essence of Theorem 3.2 thus lies in the fact that the (algebraic) property (3.4) implies the (strong) statement (I), whereas already the (weaker) statement (II) implies (3.4). The fact that (3.4) implies (II) answers the natural question that was considered at the end of Section 1.1: we see that condition (1.6) with $\Delta t_0 = \min\{R(A,b) \cdot \tau_0,\ \sigma/\alpha_0\}$, $0 < \sigma < S(A,b)$, guarantees property (1.9) whenever condition (1.8) is fulfilled. ◇

Remark 3.4. The coefficient γ in (I) and (II), whose existence under condition (3.4) is insured by Theorem 3.2, can be chosen independently of ρ. In fact, an explicit value for γ is given in the proof of the theorem; see (3.7). This value depends only on the Runge-Kutta method (A,b) and on σ. ◇

Remark 3.5. Consider an arbitrary irreducible Runge-Kutta method (A,b) that is *explicit*. We then have $S(A,b) = \infty$, so that (3.4) is equivalent to (2.7). Condition (3.3), with $\rho = R(A,b)$ and $\sigma/\alpha_0 \geq \rho \cdot \tau_0$, reduces to

$$0 < \Delta t \leq R(A,b) \cdot \tau_0. \tag{3.5}$$

According to Theorem 3.2, condition (3.5) guarantees the TVB property (1.9), with $\alpha = \gamma\alpha_0$, $\beta = \gamma\beta_0$, for F satisfying (1.8). Moreover, it can be seen (from Theorem 2.4) that (3.5) is an *optimal stepsize restriction* in that property (1.9) can no longer be guaranteed, in the same fashion, if the factor $R(A,b)$ in (3.5) would be replaced by any factor $\rho > R(A,b)$. ◇

The following lemma gives upper bounds for $\|y_i\|$ and $\|u_n\|$, in the situation where the basic assumptions (3.2), (3.3), (3.4), occurring in Theorem 3.2, are fulfilled. In order not to interrupt our presentation, we postpone the proof of the lemma to Section 5.

Lemma 3.6.
Consider an arbitrary irreducible Runge-Kutta method (A,b) *and let* $\rho,\ \sigma \in (0,+\infty)$ *satisfy* (3.4). *Then, for any vector space* $\mathbb{V}$ *with seminorm* $\|.\|$, *the conditions* (3.2), (3.3) *imply*

$$[\|y_i\|] \leq e(\alpha_0\Delta t)\|u_{n-1}\| + \beta_0\Delta t(I - \alpha_0\Delta t A)^{-1}Ae, \tag{3.6.a}$$

$$\|u_n\| \leq \varphi(\alpha_0\Delta t)\|u_{n-1}\| + \beta_0\,\frac{\varphi(\alpha_0\Delta t) - 1}{\alpha_0}, \tag{3.6.b}$$

whenever u_{n-1}, u_n *and* y_i *are related to each other as in* (1.2). *Here* $[\|y_i\|] = (\|y_1\|, \|y_2\|, ..., \|y_m\|)^T$ *belongs to* $\mathbb{R}^m$, *and* $e(\xi)$, $\varphi(\xi)$ *are defined in* (2.1). *Further, the right-hand member of* (3.6.b) *stands for* $\|u_{n-1}\| + \beta_0\Delta t$ *in case* $\alpha_0 = 0$.

Remark 3.7. Consider the *linear scalar* function $F(v) = \alpha_0 v + \beta_0$ (for $v \in \mathbb{R}$), with $\alpha_0 \geq 0$, $\beta_0 \geq 0$. Clearly, this function satisfies (3.2) with $\mathbb{V} = \mathbb{R}$ and $\|.\| = |.|$. Further, it is easy to verify that, for this simple F, the *upper bounds* (3.6) *of Lemma* 3.6 *are sharp*, in that the vectors $e(\alpha_0\Delta t)$, $\beta_0\Delta t(I - \alpha_0\Delta t A)^{-1}Ae$ and the scalars $\varphi(\alpha_0\Delta t)$, $\beta_0 \frac{\varphi(\alpha_0\Delta t)-1}{\alpha_0}$ in (3.6) cannot be replaced by any smaller quantities. Lemma 3.6 tells us that - in the situation (3.3), (3.4) - the upper bounds which are best possible for the above simple F, are also literally valid for any *nonlinear vector-valued* F satisfying (3.2).

We note that upper bounds, closely related to (3.6.b), were given earlier in Spijker (1983; Theorem 3.3) for the special case where F is a linear operator from $\mathbb{V}$ to $\mathbb{V}$ (satisfying (3.2) with $\beta_0 = 0$). ◇

Proof of Theorem 3.2.
The proof will be given by showing that the following three implications are valid: (3.4) ⇒ (I); (I) ⇒ (II) and (II) ⇒ (3.4). The first implication will be proved in step 1; the second implication is trivial; the third one will be proved in step 2.

Step 1. Assume (3.4). For proving statement (I), it is (in view of Lemma 3.6) sufficient to specify a suitable factor γ such that

$$\varphi(\alpha_0\Delta t) \leq 1 + \gamma\alpha_0\Delta t \quad \text{(for all } \Delta t \text{ satisfying (3.3))}.$$

We define

$$\gamma = \sup_{0<x\leq\sigma} \frac{\varphi(x) - 1}{x}. \tag{3.7}$$

Since $\varphi(x)$ is a differentiable for $0 \leq x \leq \sigma$ with $\varphi'(0) = \varphi(0) = 1$, we see that $\gamma \in [1, \infty)$ is as required. This proves (I).

Step 2. Assume (II); we shall prove (3.4).

In order to obtain the inequality $\rho \leq R(A, b)$, we consider an arbitrary function F satisfying (1.7), i.e. (1.8) with $\alpha_0 = \beta_0 = 0$. From (II) it follows that, for $0 < \Delta t \leq \rho \cdot \tau_0$, property (1.9) is present with $\alpha = \beta = 0$, which is the same as (1.5). An application of Theorem 2.4 (statement (ii) implies (2.7)) shows that $\rho \leq R(A, b)$.

The second inequality in (3.4) will be proved by reductio ad absurdum. With no loss of generality, we assume $S(A, b) < \infty$, $0 < \rho \leq R(A, b)$ and we suppose $\sigma \geq S(A, b)$.

In proving that this supposition leads to a contradiction, we will make use of a vector $x = (\xi_1, \xi_2, ..., \xi_m)^T \in \mathbb{R}^m$ satisfying

$$(I - \sigma_0 A)x = 0, \quad \text{with } \sigma_0 = S(A, b) > 0, \tag{3.8.a}$$

$$b_1\xi_1 + b_2\xi_2 + ... + b_m\xi_m > 0. \tag{3.8.b}$$

In order to prove the existence of such an x, we note that $\lambda_0 = 1/\sigma_0$ is a eigenvalue of A and, by definition of $S(A, b)$, there is no real eigenvalue $\lambda > \lambda_0$. Theorem

2.3 shows that $A \geq 0$ and $b > 0$. From the Perron-Frobenius theory (see e.g. Lancaster & Tismenetsky (1985), p.543), it thus follows that there exists a vector $x \in \mathbb{R}^m$, with $(\lambda_0 I - A)x = 0$, $x \geq 0$, $x \neq 0$. Consequently, (3.8.a) holds, and because all $b_i > 0$, we also have (3.8.b)

Let $\alpha_0 > 0$ be given, and let the linear function F, from $\mathbb{R}^\infty_{TV}$ into itself, be defined by $F(v) = \alpha_0 v$. It satisfies condition (1.8) with $\beta_0 = 0$ and any positive τ_0. We choose $\tau_0 = \sigma_0/(\alpha_0\rho)$, so that the stepsize $\Delta t = \sigma_0/\alpha_0$ satisfies condition (3.3). Let $w \in \mathbb{R}^\infty_{TV}$, with $\|w\|_{TV} > 0$. From (3.8), it follows immediately that, for the above F and Δt, the Runge-Kutta relations (1.2) are fulfilled, with $u_{n-1} = 0$, $y_i = \xi_i w$ and $u_n = \sigma_0(b^T x)w$, so that

$$\|u_{n-1}\|_{TV} = 0, \quad \|u_n\|_{TV} = \sigma_0 b^T x \|w\|_{TV} > 0.$$

Statement (II) implies that there exists a finite γ such that $\|u_n\|_{TV} \leq (1 + \gamma\sigma_0)\|u_{n-1}\|_{TV} + \gamma\sigma_0\beta_0/\alpha_0$. Since $\|u_{n-1}\|_{TV} = \beta_0 = 0$, it follows that $\|u_n\|_{TV} = 0$, which is impossible. ■

4 Applications and illustrations of Theorem 3.2 and Lemma 3.6

4.1 TVB preserving Runge-Kutta methods

Consider an arbitrary Runge-Kutta method (A, b). If there exist positive factors ρ, σ for which Statement (II) (of Theorem 3.2) is valid, the Runge-Kutta method will be said to be *TVB preserving*. Clearly, in this situation the TVB property of the explicit Euler method, (1.8), is carried over to the Runge-Kutta method (see (1.9)) for $\Delta t > 0$ sufficiently small. The following theorem gives a characterization of TVB preserving Runge-Kutta methods.

Theorem 4.1 (Criterion for TVB preserving Runge-Kutta methods).
Let (A, b) specify an arbitrary irreducible Runge-Kutta method. Then the method is TVB preserving if and only if (2.3) *holds.*

Proof of Theorem 4.1.
From Theorem 3.2 we see that the method (A, b) is TVB preserving if and only if $R(A, b) > 0$ and $S(A, b) > 0$. In view of Definition 3.1, we have $S(A, b) > 0$. Moreover, by Theorem 2.3 the inequality $R(A, b) > 0$ is equivalent to (2.3). ■

We note that a characterization related to the one in Theorem 4.1 was given in Ferracina & Spijker (2004, Theorem 3.6). In that paper the same class of Runge-Kutta methods satisfying (2.3) was found in a search for so-called *strong stability preserving* Runge-Kutta methods.

4.2 Two examples

In the following we will give two simple examples, illustrating the theory of Section 3.2 with an implicit and an explicit Runge-Kutta method, respectively.

Example 4.2 (An implicit Runge-Kutta method).

Consider the 1-stage second order Runge-Kutta method given by $A = (1/2)$ and $b = (1)$ (implicit midpoint rule). A simple calculation shows that $R(A, b) = S(A, b) = 2$.

Let $0 < \sigma < 2$. Then, according to Theorem 3.2 and Remark 3.4, there is a factor γ such that (1.9) holds with $\alpha = \gamma\alpha_0$, $\beta = \gamma\beta_0$, whenever F satisfies (1.8) and $0 < \Delta t \leq \min\{2\tau_0, \sigma/\alpha_0\}$. Using formula (3.7), we arrive at the following actual value for γ:

$$\gamma = \frac{2}{(2-\sigma)}.$$

Example 4.3 (An explicit Runge-Kutta method).

Consider the explicit Runge-Kutta method, with 3 stages, specified by

$$A = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1/4 & 1/4 & 0 \end{pmatrix} \quad \text{and} \quad b^T = (1/6, 1/6, 2/3).$$

This method was studied earlier, notably in Shu & Osher (1988), Kraaijevanger (1991), Gottlieb & Shu (1998), Gottlieb, Shu & Tadmor (2001), Spiteri & Ruuth (2002), Ferracina & Spijker (2004). In Kraaijevanger (1991, Theorem 9.4) it was proved that this method is of third order, with $R(A, b) = 1$, whereas there exists no other explicit third order method with $m = 3$ and $R(A, b) \geq 1$. Obviously, for the above method, $S(A, b) = \infty$.

Choosing $\rho = R(A, b) = 1$ and $0 < \sigma < S(A, b) = \infty$, condition (3.4) is fulfilled, and the stepsize restriction (3.3) reduces to

$$0 < \Delta t \leq \min\{\tau_0, \sigma/\alpha_0\}. \tag{4.1}$$

According to Theorem 3.2, there is a factor γ such that (1.8), (4.1) imply (1.9) with $\alpha = \gamma\alpha_0$, $\beta = \gamma\beta_0$. In view of Remark 3.4, we can apply (3.7) so as to arrive at the value

$$\gamma = 1 + \frac{\sigma}{2} + \frac{\sigma^2}{6}. \tag{4.2}$$

Moreover, using Lemma 3.6 directly, we can get a bound on $\|u_n\|_{TV}$ which is more complicated than (1.9) but more refined. For the Runge-Kutta method under consideration, relation (3.6.b), with $\|.\| = \|.\|_{TV}$, reduces to

$$\begin{aligned} \|u_n\|_{TV} \quad \leq \quad & [1 + \alpha_0\Delta t + \frac{1}{2}(\alpha_0\Delta t)^2 + \frac{1}{6}(\alpha_0\Delta t)^3]\|u_n\|_{TV} + \\ & + [1 + \frac{1}{2}\alpha_0\Delta t + \frac{1}{6}(\alpha_0\Delta t)^2]\beta_0\Delta t. \end{aligned} \tag{4.3}$$

From Lemma 3.6 it can be seen that (4.3) is valid, whenever F satisfies (1.8) and $0 < \Delta t \leq \tau_0$.

4.3 A special semi-discretization given by Shu (1987)

Applying the special semi-discretization devised by Shu (1987) to equation (1.3), we obtain a semi-discrete system of equations which can be modeled as $\frac{d}{dt}U(t) = F(U(t))$ where

$$(4.4)\quad F:\ \mathbb{R}^{\infty}_{TV} \longrightarrow \mathbb{R}^{\infty}_{TV} \ \text{ satisfies } \ \|v + \tau_0 F(v)\|_{TV} \leq \|v\|_{TV} + \beta_0\tau_0 \quad (v \in \mathbb{R}^{\infty}_{TV}).$$

Here $\tau_0 > 0$ and $\beta_0 > 0$. The basic assumption (1.7) of the TVD theory, reviewed in Section 2.2, is *not* fulfilled here. On the other hand, the above situation (4.4) is nicely covered by Theorem 3.2 and Lemma 3.6 (with $\alpha_0 = 0$).

We consider the application of an arbitrary irreducible Runge-Kutta method (A, b), in the situation (4.4), with a stepsize Δt satisfying

$$(4.5)\qquad 0 < \Delta t \leq R(A, b) \cdot \tau_0$$

Using Theorem 3.2 or Lemma 3.6 (with $\alpha_0 = 0$), one sees that (4.4), (4.5) imply

$$(4.6)\qquad \|u_n\|_{TV} \leq \|u_{n-1}\|_{TV} + \beta_0\Delta t, \quad \text{for } u_n \text{ and } u_{n-1} \text{ satisfying (1.2).}$$

Hence, in the situation (4.4), the Runge-Kutta approximations u_n satisfy (1.4), with $B = \|u_0\|_{TV} + \beta_0 T$ and $\Delta t_0 = R(A, b) \cdot \tau_0$.

It is worthwhile to note that the last value Δt_0 is positive if and only if the Runge-Kutta method (A, b) satisfies (2.3) - this is evident from Theorem 2.3.

5 The proof of Lemma 3.6

In our following proof of Lemma 3.6, we shall make use of the subsequent Lemmas 5.1 and 5.2.

Lemma 5.1 deals with the situation where

$$(5.1.a)\qquad B \geq 0,$$
$$(5.1.b)\qquad I - tB \quad \text{is invertible for } t_0 \leq t \leq t_1,$$
$$(5.1.c)\qquad (I - t_0 B)^{-1} \geq 0.$$

Here B stands for an $m \times m$ matrix and I denotes the $m \times m$ identity matrix.

Lemma 5.1.
The assumptions (5.1) *imply that*

$$(5.2)\qquad (I - tB)^{-1} \geq 0 \quad \textit{for} \quad t_0 \leq t \leq t_1.$$

Proof of Lemma 5.1.
Assume (5.1) and suppose (5.2) is not true. Let T be the greatest lower bound of the values $t \in [t_0, t_1]$ where the inequality $(I - tB)^{-1} \geq 0$ is violated. One easily sees (by continuity arguments) that $(I - TB)^{-1} \geq 0$ and $t_0 \leq T < t_1$. For all sufficient small $\varepsilon > 0$, we have

$$I - (T + \varepsilon)B = I - TB - \varepsilon B = (I - TB)(I - (I - TB)^{-1}\varepsilon B),$$

so that

$$[I - (T + \varepsilon)B]^{-1} = \left\{ \sum_{k=0}^{\infty} [\varepsilon (I - TB)^{-1} B]^k \right\} (I - TB)^{-1} \geq 0.$$

This contradicts the definition of T. Hence (5.2) must be true. ■

In the actual proof of Lemma 3.6, the Runge-Kutta process (1.2) will be represented in the following form:

$$\text{(5.3.a)} \quad y_i = \left(1 - \sum_{j=1}^{m} \lambda_{ij}\right) u_{n-1} + \sum_{j=1}^{m} [\lambda_{ij}\, y_j + \Delta t \cdot \mu_{ij} F(y_j)] \quad (1 \leq i \leq m),$$

$$\text{(5.3.b)} \quad u_n = \left(1 - \sum_{j=1}^{m} \lambda_{m+1,j}\right) u_{n-1} + \sum_{j=1}^{m} [\lambda_{m+1,j}\, y_j + \Delta t \cdot \mu_{m+1,j} F(y_j)].$$

Here λ_{ij} and μ_{ij} denote real parameters. We define corresponding matrices L, M by:

$$\text{(5.4.a)} \quad L = \begin{pmatrix} L_0 \\ L_1 \end{pmatrix}, \quad L_0 = \begin{pmatrix} \lambda_{11} & \dots & \lambda_{1m} \\ \vdots & & \vdots \\ \lambda_{m1} & \dots & \lambda_{mm} \end{pmatrix}, \quad L_1 = (\lambda_{m+1,1}, ..., \lambda_{m+1,m}),$$

$$\text{(5.4.b)} \quad M = \begin{pmatrix} M_0 \\ M_1 \end{pmatrix}, \; M_0 = \begin{pmatrix} \mu_{11} & \dots & \mu_{1m} \\ \vdots & & \vdots \\ \mu_{m1} & \dots & \mu_{mm} \end{pmatrix}, \; M_1 = (\mu_{m+1,1}, ..., \mu_{m+1,m}).$$

Lemma 5.2, to be given below, gives a condition under which the processes (1.2) and (5.3) are equivalent.

In the lemma the following relation will play a crucial role:

$$\text{(5.5)} \qquad M_0 = A - L_0 A, \quad M_1 = b^T - L_1 A.$$

Further, the following hypothesis will be used:

$$\text{(5.6)} \qquad I - L_0 \text{ is invertible.}$$

Lemma 5.2.
Let (A, b) specify an arbitrary Runge-Kutta method (1.2). *Let $L = (\lambda_{ij})$ be any parameter matrix satisfying* (5.4.a) *and* (5.6). *Consider the corresponding matrix M defined by* (5.4.b), (5.5). *Then the Runge-Kutta relations* (1.2) *are equivalent to* (5.3).

This lemma was proved in Ferracina & Spijker (2005, Theorem 2.2), Higueras (2003, Section 2). The proof is easy and involves only simple algebraic manipulations. Therefore, we do not repeat it here but refer to the papers just mentioned for details.

For matrices L and M of the form (5.4), we define the coefficient $c(L, M)$ by:

$$c(L,M) \;=\; \min\{c_{ij} :\; 1 \le i \le m+1,\;\; 1 \le j \le m\}, \tag{5.7}$$

$$c_{ij} = \begin{cases} \lambda_{ij}/\mu_{ij} & \text{if } \mu_{ij} > 0 \text{ and } i \ne j, \\ \infty & \text{if } \mu_{ij} > 0 \text{ and } i = j, \\ \infty & \text{if } \mu_{ij} = 0, \\ 0 & \text{if } \mu_{ij} < 0. \end{cases}$$

The actual proof of Lemma 3.6, to be given below, consists of two parts. In the first part we shall consider the situation where

$$\lambda_{ij} \ge 0 \quad \text{and} \quad \sum_{k=1}^{m} \lambda_{ik} \le 1 \quad (\text{for } 1 \le i \le m+1,\; 1 \le j \le m), \tag{5.8}$$

and

$$0 < \Delta t \le c(L,M) \cdot \tau_0. \tag{5.9}$$

It will be shown that (3.2), (5.3), (5.8), (5.9) imply

$$(I - L_0 - \alpha_0 \Delta t M_0)\,[\|y_i\|] \;\le\; \|u_{n-1}\|(I - L_0)e + \beta_0 \Delta t M_0\, e, \tag{5.10.a}$$

$$\|u_n\| \;\le\; (1 - L_1 e)\|u_{n-1}\| + (L_1 + \alpha_0 \Delta t M_1)[\|y_i\|] + \beta_0 \Delta t M_1 e. \tag{5.10.b}$$

The above relation (5.10.a) stands for an inequality between two vectors in $\mathbb{R}^m$, which should be interpreted component-wise. Further, we denote again by e the vector in $\mathbb{R}^m$ all of whose components are equal to 1.

In the second part of the actual proof, we shall choose a special parameter matrix L and define M by (5.4.b), (5.5). It will be seen that $I - L_0$ is invertible so that, by Lemma 5.2, the process (5.3) under consideration is equivalent to (1.2). Moreover, the conditions (5.8) are fulfilled and $c(L, M) = R(A, b)$. The proof of Lemma 3.6 will be completed by showing that, in the situation (5.5), (3.3), (3.4), the inequalities (5.10) imply (3.6).

The actual proof of Lemma 3.6.
Part 1. Assume (3.2), (5.3), (5.8), (5.9). We sall prove (5.10).

Condition (5.9) implies that, for all i, j,

$$0 < c_{ij} \leq \infty \quad \text{and} \quad 0 \leq \mu_{ij} < \infty.$$

From (5.3.a), we obtain for $1 \leq i \leq m$

$$\begin{aligned} \|y_i - \Delta t \mu_{ii} F(y_i)\| \quad &\leq \quad (1 - \sum_{j=1}^{m} \lambda_{ij}) \|u_{n-1}\| + \lambda_{ii} \|y_i\| + \\ &\quad + \sum_{j \neq i} \lambda_{ij} \|y_j + \Delta t c_{ij}^{-1} F(y_j)\|, \end{aligned} \tag{5.11}$$

where c_{ij}^{-1} stands for 0 in case $c_{ij} = \infty$.

Using the relation $(1+\mu_{ii}\Delta t/\tau_0)y_i = (y_i - \Delta t \mu_{ii} F(y_i)) + (\mu_{ii}\Delta t/\tau_0)(y_i + \tau_0 F(y_i))$ we obtain $(1+\mu_{ii}\Delta t/\tau_0)\|y_i\| \leq \|y_i - \Delta t \mu_{ii} F(y_i)\| + \{(1+\alpha_0\tau_0)\|y_i\| + \beta_0\tau_0\}\mu_{ii}\Delta t/\tau_0$. Hence

$$(1 - \mu_{ii}\alpha_0 \Delta t)\|y_i\| - \beta_0 \mu_{ii} \Delta t \leq \|y_i - \Delta t \mu_{ii} F(y_i)\|. \tag{5.12}$$

Similarly, by using the relation

$$y_j + \Delta t c_{ij}^{-1} F(y_j) = (1 - \Delta t (\tau_0 c_{ij})^{-1}) y_j + \Delta t (\tau_0 c_{ij})^{-1} (y_j + \tau_0 F(y_j)),$$

we see that

$$\|y_j + \Delta t c_{ij}^{-1} F(y_j)\| \leq \{1 + \alpha_0 \Delta t c_{ij}^{-1}\} \|y_j\| + \beta_0 \Delta t c_{ij}^{-1}. \tag{5.13}$$

Combining the inequalities (5.11), (5.12) and (5.13), we obtain a bound for $\|y_i\|$ $(1 \leq i \leq m)$ which can be written compactly in the form (5.10.a).

In order to prove (5.10.b), we note that (5.3.b) implies

$$\|u_n\| \leq \left(1 - \sum_{j=1}^{m} \lambda_{m+1,j}\right) \|u_{n-1}\| + \sum_{j=1}^{m} \lambda_{m+1,j} \|y_j + \Delta t \cdot \, c_{m+1,j}^{-1} F(y_j)\|.$$

Applying (5.13) with $i = m + 1$, we obtain (5.10.b).

Part 2. Assume (3.2), (1.2), (3.3), (3.4). We shall prove (3.6).

In case $0 \leq R(A, b) < \infty$, we know from Kraaijevanger (1991, Lemma 4.4) that the matrix $(I+\eta A)$, with $\eta = R(A, b)$, is invertible. Moreover, in case $R(A, b) = \infty$, it follows from Kraaijevanger (1991, Theorem 4.7) that the inverse A^{-1} exists, and that the diagonal elements of this inverse are positive. Therefore, we can define a matrix L of the form (5.4.a) in the following way:

$$L_0 = \eta A (I + \eta A)^{-1}, \quad L_1 = \eta b^T (I + \eta A)^{-1}, \quad \text{where } \eta = R(A, b) \tag{5.14.a}$$
$$(\text{if} \quad 0 \leq R(A, b) < \infty),$$

$$L_0 = I - \eta P, \quad L_1 = b^T P, \quad \eta = (\max_i p_{ii})^{-1}, \text{ where } P = (p_{ij}) = A^{-1} \tag{5.14.b}$$
$$(\text{if} \quad R(A, b) = \infty).$$

Similar matrices were introduced and analysed earlier in Ferracina & Spijker (2005), Higueras (2003). One easily sees that condition (5.6) is fulfilled. We define M by (5.4.b), (5.5), so that, according to Lemma 5.2, the relations (1.2) imply (5.3).

For the matrices L, M under consideration, it is known that (5.8) holds and that $c(L, M) = R(A, b)$ - see Ferracina & Spijker (2005, Theorem 3.4), Higueras (2003, Section 2). Therefore, our assumptions (3.3), (3.4) imply (5.9) and, according to the above Part 1, we can conclude that (5.10) holds. Below, we shall prove (3.6) by using (5.10), (5.5), (3.3), (3.4).

Using the equality $I - L_0 - \alpha_0 \Delta t M = (I - L_0)(I - \alpha_0 \Delta t A)$, one sees that (5.10.a) implies (3.6.a), provided the inverses $(I - L_0)^{-1}$, $(I - \alpha_0 \Delta t A)^{-1}$ exist and have only nonnegative entries. The existence of $(I - L_0)^{-1}$ was proved above, and its nonnegativity follows from an application of Lemma 5.1, with $B = L_0$, $t_0 = 0$, $t_1 = 1$ (note that, in view of (5.8), the eigenvalues of $I - tL_0$ are different from zero, for $0 \leq t < 1$). The existence of $(I - \alpha_0 \Delta t A)^{-1}$ is a consequence of (3.3), (3.4), and its nonnegativity follows by applying Theorem 2.3 and Lemma 5.1, with $B = A$, $t_0 = 0$, $t_1 = \alpha_0 \Delta t$. Finally, (3.6.b) follows by straightforward calculations using (3.6.a), (5.5). ■

Bibliography

[1] Butcher J. C. (2003): *Numerical methods for ordinary differential equations*. John Wiley & Sons Ltd. (Chichester).

[2] Cockburn B., Shu C.-W. (1989): TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws. II. General framework. *Math. Comp.*, 52 No. 186, 411–435.

[3] Dekker K., Verwer J. G. (1984): *Stability of Runge-Kutta methods for stiff nonlinear differential equations*, vol. 2 of *CWI Monographs*. North-Holland Publishing Co. (Amsterdam).

[4] Ferracina L., Spijker M. N. (2004): Stepsize restrictions for the total-variation-diminishing property in general Runge-Kutta methods. *SIAM J. Numer. Anal.*, 42 No. 3, 1073–1093.

[5] Ferracina L., Spijker M. N. (2005): An extension and analysis of the Shu-Osher representation of Runge-Kutta methods. *Math. Comp.*, 74 No. 249, 201–219.

[6] Gottlieb S., Shu C.-W. (1998): Total variation diminishing Runge-Kutta schemes. *Math. Comp.*, 67 No. 221, 73–85.

[7] Gottlieb S., Shu C.-W., Tadmor E. (2001): Strong stability-preserving high-order time discretization methods. *SIAM Rev.*, 43 No. 1, 89–112.

[8] Hairer E., Nørsett S. P., Wanner G. (1993): *Solving ordinary differential equations. I. Nonstiff problems*, vol. 8 of *Springer Series in Computational Mathematics*. Springer-Verlag (Berlin), second ed.

[9] Hairer E., Wanner G. (1996): *Solving ordinary differential equations. II. Stiff and differential-algebraic problems*, vol. 14 of *Springer Series in Computational Mathematics*. Springer-Verlag (Berlin), second ed.

[10] Harten A. (1984): On a class of high resolution total-variation-stable finite-difference schemes. *SIAM J. Numer. Anal.*, 21 No. 1, 1–23. With an appendix by Peter D. Lax.

[11] Higueras I. (2003): Representation of Runge-Kutta methods and strong stability preserving methods. Tech. rep., Departamento de Matemática e Informática, Universidad Pública de Navarra.

[12] Higueras I. (2004): On strong stability preserving time discretization methods. *J. Sci. Comput.*, 21 No. 2, 193–223.

[13] Hundsdorfer W., Ruuth S. J. (2004): On monotonicity and boundedness properties of linear multistep methods. Tech. rep., MAS-E0404, CWI-Centrum voor Wiskunde en Informatica (Amsterdam).

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

[15] Hundsdorfer W., Verwer J. G. (2003): *Numerical solution of time-dependent advection-diffusion-reaction equations*, vol. 33 of *Springer Series in Computational Mathematics*. Springer (Berlin).

[16] Kraaijevanger J. F. B. M. (1991): Contractivity of Runge-Kutta methods. *BIT*, 31 No. 3, 482–528.

[17] Kröner D. (1997): *Numerical schemes for conservation laws*. Wiley-Teubner Series Advances in Numerical Mathematics. John Wiley & Sons Ltd. (Chichester).

[18] Lancaster P., Tismenetsky M. (1985): *The theory of matrices*. Computer Science and Applied Mathematics. Academic Press Inc. (Orlando, FL), second ed.

[19] Laney C. B. (1998): *Computational gasdynamics*. Cambridge University Press (Cambridge).

[20] LeVeque R. J. (2002): *Finite volume methods for hyperbolic problems*. Cambridge Texts in Applied Mathematics. Cambridge University Press (Cambridge).

[21] Osher S., Chakravarthy S. (1984): High resolution schemes and the entropy condition. *SIAM J. Numer. Anal.*, 21 No. 5, 955–984.

[22] Shu C.-W. (1987): TVB uniformly high-order schemes for conservation laws. *Math. Comp.*, 49 No. 179, 105–121.

[23] Shu C.-W. (1988): Total-variation-diminishing time discretizations. *SIAM J. Sci. Statist. Comput.*, 9 No. 6, 1073–1084.

[24] Shu C.-W. (2002): A survey of strong stability preserving high-order time discretizations. In *Collected Lectures on the Preservation of Stability under Discretization*, S. T. E. D. Estep, Ed., pp. 51–65. SIAM (Philadelphia).

[25] Shu C.-W., Osher S. (1988): Efficient implementation of essentially nonoscillatory shock-capturing schemes. *J. Comput. Phys.*, 77 No. 2, 439–471.

[26] Spijker M. N. (1983): Contractivity in the numerical solution of initial value problems. *Numer. Math.*, 42 No. 3, 271–290.

[27] Spiteri R. J., Ruuth S. J. (2002): A new class of optimal high-order strong-stability-preserving time discretization methods. *SIAM J. Numer. Anal.*, 40 No. 2, 469–491 (electronic).

[28] Toro E. F. (1999): *Riemann solvers and numerical methods for fluid dynamics. A practical introduction.* Springer-Verlag (Berlin), second ed.