MATHEMATICS OF COMPUTATION
Volume 74, Number 249, Pages 201–219
S 0025-5718(04)01664-3
Article electronically published on June 11, 2004

AN EXTENSION AND ANALYSIS OF THE SHU-OSHER REPRESENTATION OF RUNGE-KUTTA METHODS

L. FERRACINA AND M. N. SPIJKER

Abstract. In the context of solving nonlinear partial differential equations, Shu and Osher introduced representations of explicit Runge-Kutta methods, which lead to stepsize conditions under which the numerical process is total-variation-diminishing (TVD). Much attention has been paid to these representations in the literature.

In general, a Shu-Osher representation of a given Runge-Kutta method is not unique. Therefore, of special importance are representations of a given method which are best possible with regard to the stepsize condition that can be derived from them.

Several basic questions are still open, notably regarding the following issues: (1) the formulation of a simple and general strategy for finding a best possible Shu-Osher representation for any given Runge-Kutta method; (2) the question of whether the TVD property of a given Runge-Kutta method can still be guaranteed when the stepsize condition, corresponding to a best possible Shu-Osher representation of the method, is violated; (3) the generalization of the Shu-Osher approach to general (possibly implicit) Runge-Kutta methods.

In this paper we give an extension and analysis of the original Shu-Osher representation, by means of which the above questions can be settled. Moreover, we clarify analogous questions regarding properties which are referred to, in the literature, by the terms monotonicity and strong-stability-preserving (SSP).

1. Introduction

1.1. **The purpose of the paper.** In this paper we deal with the numerical solution of initial value problems, for systems of ordinary differential equations, 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 $n = 1, 2, 3, \ldots$; cf., e.g., Butcher [2], Dekker and Verwer [3], Hairer, Nørsett and Wanner [8], Hairer and Wanner [9]. The approximations u_n are defined in terms of

Received by the editor May 7, 2003 and, in revised form, August 3, 2003.

2000 *Mathematics Subject Classification.* Primary 65M20; Secondary 65L05, 65L06.

Key words and phrases. Initial value problem, conservation law, method of lines (MOL), Runge-Kutta formula, Shu-Osher representation, total-variation-diminishing (TVD), strong-stability-preserving (SSP), monotonicity.

u_{n-1} by the relations

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

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

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 $b_1 + b_2 + \cdots + b_m = 1$, and we call the Runge-Kutta method *explicit* if $a_{ij} = 0$ (for $j \ge i$). We define the $m \times m$ matrix A by $A = (a_{ij})$ and the column vector $b \in \mathbb{R}^m$ by $b = (b_1, b_2, b_3, \ldots, b_m)^T$, 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 scalar conservation law of the form

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

In this situation, the function F occurring in (1.1) can be regarded as a function from

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

into itself; see, e.g., Laney [16], LeVeque [17], Toro [25]. The actual function values $F(y)$ depend on the given f as well as on the MOL semidiscretization being used. In the literature (see, e.g., Gottlieb, Shu and Tadmor [7], Shu [21], Shu and Osher [22], Spiteri and Ruuth [24]) much attention has been paid to solving the semidiscrete problem (1.1) by Runge-Kutta processes (1.2) which are *total-variation-diminishing (TVD)* in the sense that

$$\|u_n\|_{TV} \ \le \ \|u_{n-1}\|_{TV}; \tag{1.4}$$

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 relevance of the TVD property in the numerical solution of (1.3); see, e.g., Harten [10], Kröner [15], Laney [16], LeVeque [17], Toro [25].

By Shu and Osher [22] (see also Shu [20]) a clever representation of explicit Runge-Kutta methods was introduced which facilitates the proof of property (1.4) in the situation where, for some $\tau_0 > 0$,

$$\|v + \tau F(v)\|_{TV} \ \le \ \|v\|_{TV} \quad (\text{whenever } 0 < \tau \le \tau_0 \text{ and } v \in \mathbb{R}^\infty). \tag{1.5}$$

Clearly, (1.5) amounts to assuming that the semidiscretization of equation (1.3) has been performed in such a manner that the simple forward Euler method, applied to problem (1.1), is TVD when the stepsize τ is suitably restricted.

In order to describe the representation, given by Shu and Osher [22], we consider an arbitrary explicit coefficient scheme (A, b). We assume that λ_{ij} (for $2 \le i \le m+1$ and $1 \le j \le i-1$) are any real parameters with

$$\lambda_{i1} + \lambda_{i2} + \cdots + \lambda_{i,i-1} \ = \ 1 \quad (2 \le i \le m+1), \tag{1.6}$$

and we define corresponding values μ_{ij} (for $2 \leq i \leq m+1$ and $1 \leq j \leq i-1$) by

$$\mu_{ij} = a_{ij} - \sum_{k=j+1}^{i-1} \lambda_{ik}\, a_{kj} \quad (2 \leq i \leq m,\ 1 \leq j \leq i-1), \tag{1.7.a}$$

$$\mu_{m+1,j} = b_j - \sum_{k=j+1}^{m} \lambda_{m+1,k}\, a_{kj} \quad (1 \leq j \leq m) \tag{1.7.b}$$

(where the sums occurring in the above expressions defining μ_{ij} and $\mu_{m+1,j}$ should be interpreted as 0, when $j = i-1$ and $j = m$, respectively).

Theorem 1.1, to be given below, tells us that the relations (1.2) can be rewritten in the form

$$\begin{aligned} y_1 &= u_{n-1}, \\ y_i &= \sum_{j=1}^{i-1} [\lambda_{ij}\, y_j + \Delta t \cdot \mu_{ij} F(y_j)] \quad (2 \leq i \leq m+1), \\ u_n &= y_{m+1}. \end{aligned} \tag{1.8}$$

We shall refer to (1.8) as a *Shu-Osher representation* of the explicit Runge-Kutta method (1.2).

The following Theorem 1.1 also specifies a stepsize restriction, of the form

$$0 < \Delta t \leq c \cdot \tau_0, \tag{1.9}$$

under which the TVD property (1.4) is valid, when u_n is computed from u_{n-1} according to (1.8). In the theorem, we shall consider the situation where

$$\lambda_{ij} \ \geq\ 0 \quad (1 \leq j < i \leq m+1). \tag{1.10}$$

Furthermore, we shall deal with a coefficient c defined by

$$c = \min\{c_{ij} : 1 \leq j < i \leq m+1\}, \qquad \text{where } c_{ij} = \begin{cases} \lambda_{ij}/\mu_{ij} & \text{if } \mu_{ij} > 0, \\ \infty & \text{if } \mu_{ij} = 0, \\ 0 & \text{if } \mu_{ij} < 0. \end{cases} \tag{1.11}$$

Theorem 1.1 (Shu and Osher)**.** *Let* (A, b) *specify an explicit Runge-Kutta method and assume* λ_{ij}, μ_{ij} *are as in* (1.6), (1.7). *Then the following conclusions* (i) *and* (ii) *are valid.*

(i) *The Runge-Kutta relations* (1.2) *are equivalent to* (1.8).

(ii) *Assume additionally that* (1.10) *holds and that the coefficient* c *is defined by* (1.11). *Let* F *be a function from* $\mathbb{R}^\infty$ *to* $\mathbb{R}^\infty$, *satisfying* (1.5). *Then, under the stepsize restriction* (1.9), *process* (1.8) *is TVD; i.e.* (1.4) *holds whenever* u_n *is computed from* u_{n-1} *according to* (1.8).

The above theorem is essentially due to Shu and Osher [22]. The proof of the above statement (i) is straightforward. Furthermore, the proof of (ii) relies on noting that, for $2 \leq i \leq m+1$, the vector y_i in (1.8) can be rewritten as a convex combination of the vectors $[y_j + \Delta t \cdot (\mu_{ij}/\lambda_{ij}) F(y_j)]$ with $1 \leq j \leq i-1$ and on applying (1.5) (with $v = y_j$).

It is evident that a combination of the above statements (i) and (ii) immediately leads to *a conclusion which is highly relevant to the original Runge-Kutta method* (A, b)*: if* (1.6), (1.7), (1.10) (1.11) *are fulfilled, then the conditions* (1.5), (1.9) *guarantee the TVD property* (1.4) *for* u_n *computed from* u_{n-1} *by* (1.2).

But this conclusion regarding the Runge-Kutta method (1.2) would be of no, or little, value if the coefficient c given by (1.11) were zero, or positive and so small that the stepsize restriction (1.9) is too severe for any practical purposes—in fact, the less restrictions on Δt the better. Therefore, it is important to note that the coefficient c, given by (1.11), not only depends on the underlying Runge-Kutta method (A, b), but also on the parameters λ_{ij} actually chosen. Suppose $\tilde{\lambda}_{ij}$ are parameters which are best possible, in the sense that the corresponding coefficient $\tilde{c}$, obtained via (1.11), satisfies $\tilde{c} \geq c$, for any other coefficient c obtainable by applying Theorem 1.1 to the method (A, b) in question. Then $\tilde{c}$ depends only on the coefficient scheme (A, b) so that we can write $\tilde{c} = c(A, b)$, and the following natural question arises: how can we determine (in a transparent and simple way) parameters $\tilde{\lambda}_{ij}$ leading to the maximal coefficient $c(A, b)$?

Another—and second—natural question is related to the circumstance that one may be tempted to take the magnitude of the coefficient $c(A, b)$ into account, when assessing the qualities of a given explicit Runge-Kutta method (A, b). It is evident that such a use of $c(A, b)$ could be quite misleading if, for the Runge-Kutta method (A, b) in question, there exists a coefficient c (*not* obtainable from Theorem 1.1) which is (much) larger than $c(A, b)$ and for which the conditions (1.5), (1.9) still guarantee the TVD property (1.4) for process (1.2). Accordingly, we arrive at the fundamental question of whether such coefficients c do exist.

The above two questions are strongly related to the problem of determining a method (A, b), belonging to a given class of explicit Runge-Kutta methods, which is optimal in the sense of its coefficient $c(A, b)$. Much attention has been paid to this problem in the literature—usually with terminology and notation somewhat different from the above (see, e.g., Gerisch and Weiner [5], Gottlieb and Shu [6], Ruuth and Spiteri [19], Shu [21], Shu and Osher [22], Spiteri and Ruuth [24]). In fact, for various values of m and p, optimal methods (A, b) were determined within the class of explicit m-stage Runge-Kutta methods with order of accuracy p—either by clever ad hoc arguments or by numerical computations based on optimization with respect to the parameters λ_{ij}, μ_{ij}—but, neither of the above two questions were resolved (in general).

A third natural question is of whether the Shu-Osher Theorem 1.1 can be generalized so as to also become relevant to Runge-Kutta methods which are *not explicit*. Partial results related to this question, but no complete answers, were obtained by Gottlieb, Shu and Tadmor [7, Section 6.2] and Hundsdorfer and Verwer [13].

The purpose of this paper is to give a generalization and analysis of the Shu-Osher representation (1.8) by means of which the above three natural questions, as well as related ones, can be settled.

1.2. **Outline of the rest of the paper.**

In Section 2 we shall give generalizations of the Shu-Osher representation (1.8) and of the above Shu-Osher Theorem 1.1; our generalizations are relevant to arbitrary Runge-Kutta methods (A, b)—either explicit or not.

It was noted (see, e.g., Gottlieb, Shu and Tadmor [7], Shu and Osher [22]) that the convexity arguments, used in proving conclusion (ii) of Theorem 1.1, also show that $\|y_i\|_{TV} \leq \|u_{n-1}\|_{TV}$ $(2 \leq i \leq m)$ and also apply in the more general setting of arbitrary Banach spaces $\mathbb{V}$ and nonnegative convex functions $\|.\|$ (rather than $\mathbb{R}^\infty$ and $\|.\|_{TV}$). Therefore, a useful version of Theorem 1.1 is valid in that context as well. Accordingly, we shall present our material in Section 2 using a similar general framework.

In Section 2.1 we shall introduce concepts and notation which are basic for the rest of our paper. A generalization will be given of the Shu-Osher process (1.8) and of the properties (1.4) and (1.5). In Section 2.2 we shall present Theorem 2.2, which constitutes the first of the two main theorems of our paper. This theorem settles completely the question, about the generalization of Theorem 1.1, raised above at the end of Section 1.1. Conclusion (I) of Theorem 2.2 generalizes conclusion (i) of Theorem 1.1. For any given Runge-Kutta method (A, b), it gives a generalized Shu-Osher representation which is specified by an $(m+1) \times m$ parameter matrix $L = (\lambda_{ij})$; the corresponding numerical process can thus be identified with a coefficient scheme (A, b, L). Conclusion (II) of Theorem 2.2 generalizes conclusion (ii) of Theorem 1.1; it provides us with a coefficient $c = c(A, b, L)$ having properties generalizing those of c (see (1.11)) mentioned in conclusion (ii) of Theorem 1.1. In Section 2.3 we shall give the proof of Theorem 2.2.

In Section 3 we shall study, for given Runge-Kutta schemes (A, b), the maximum of $c(A, b, L)$ over all relevant parameter matrices $L = (\lambda_{ij})$. In preparation for the actual study of this maximum, we shall recall in Section 3.1 the concept of irreducibility for general Runge-Kutta methods, and we shall review the important quantity $R(A, b)$, introduced by Kraaijevanger [14]. In Section 3.2 we shall present (without proof) the second of our two main theorems, Theorem 3.4. This theorem is relevant to arbitrary irreducible Runge-Kutta schemes (A, b); it gives a special parameter matrix $L^* = (\lambda^*_{ij})$ such that $c(A, b, L^*) = \max_L c(A, b, L)$. Moreover, the theorem brings to light that there exists no coefficient c that is larger than $c(A, b, L^*)$ and which shares with $c(A, b, L^*)$ properties analogous to those of c mentioned in part (ii) of Theorem 1.1. Finally, the theorem relates the optimal coefficient $c(A, b, L^*)$ to Kraaijevanger's quantity $R(A, b)$. The proof of Theorem 3.4 will be given in Section 3.3, making use of Lemma 3.5.

For completeness we mention that also in Ferracina and Spijker [4] and Higueras [11] the quantity $R(A, b)$ was related to the TVD properties of method (1.2). In fact, Lemma 3.5 is an immediate consequence of a theorem in the first of these papers. But, apart from this lemma, the material in Section 3 is essentially different from and no consequence of those papers.

In Section 4 we shall present some applications and illustrations to the theorems derived in Sections 2 and 3.

In Section 4.1 we shall apply Theorems 2.2 and 3.4 to general Runge-Kutta methods so as to arrive at Corollaries 4.1 and 4.2. The former of these corollaries says that $c(A, b, L)$ is finite, for every scheme (A, b) which is more than first order, whereas the latter corollary amounts to an extension of a monotonicity result in Ferracina and Spijker [4].

In Section 4.2, the two questions will be answered which were raised above in Section 1.1, in connection to the coefficient $c(A, b)$. For any given explicit method (A, b), Theorem 4.3 gives special parametes $\lambda_{ij} = \tilde{\lambda}_{ij}$ and $\mu_{ij} = \tilde{\mu}_{ij}$, satisfying (1.6), (1.7), (1.10) such that the corresponding coefficient $c = \tilde{c}$, obtained from (1.11), is the largest one obtainable with *any* parameters λ_{ij}, μ_{ij} satisfying (1.6), (1.7), (1.10) (i.e., $\tilde{c} = c(A, b)$). Moreover, Theorem 4.3 says that $\tilde{c} = c(A, b)$ is equal to the largest coefficient c for which the conditions (1.5), (1.9) guarantee (1.4). This result is relevant to justifying the practice of considering $c(A, b)$ when assessing the qualities of a given Runge-Kutta method (A, b). At the end of Section 4.2, we apply Theorem 4.3 so as to relate results, obtained in the literature on optimization of $c(A, b)$, to material of Kraaijevanger [14].

In Section 4.3 we shall shortly illustrate our theory by applying it in the analysis of (generalized) Shu-Osher representations for two given Runge-Kutta schemes.

2. An extension of the Shu-Osher approach to arbitrary Runge-Kutta methods

2.1. **A generalization of the Shu-Osher process (1.8).** We want to consider generalized versions of the Shu-Osher process (1.8) in a versatile framework. For that reason we assume in all of the following (unless specified otherwise) 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 $\mathbb{V}$. Our generalization of the Shu-Osher process (1.8) is as follows:

$$\text{(2.1.a)} \quad y_i = \left(1 - \sum_{j=1}^{m} \lambda_{ij}\right) u_{n-1} + \sum_{j=1}^{m} [\lambda_{ij}\, y_j + \Delta t \cdot \mu_{ij} F(y_j)] \quad (1 \le i \le m),$$

$$\text{(2.1.b)} \quad u_n = \left(1 - \sum_{j=1}^{m} \lambda_{m+1,j}\right) u_{n-1} + \sum_{j=1}^{m} [\lambda_{m+1,j}\, y_j + \Delta t \cdot \mu_{m+1,j} F(y_j)].$$

Here λ_{ij} and μ_{ij} are real coefficients specifying the numerical process (2.1), and Δt denotes again a positive stepsize. Furthermore, y_i are intermediate vectors in $\mathbb{V}$ needed for computing u_n in $\mathbb{V}$ from a given vector $u_{n-1} \in \mathbb{V}$. We shall write

$$\text{(2.2.a)} \quad L = \begin{pmatrix} L_0 \\ L_1 \end{pmatrix}, \quad L_0 = \begin{pmatrix} \lambda_{11} & \dots & \lambda_{1m} \\ \vdots & & \vdots \\ \lambda_{m1} & \dots & \lambda_{mm} \end{pmatrix}, \quad L_1 = (\lambda_{m+1,1}, \dots, \lambda_{m+1,m})$$

and

$$\text{(2.2.b)} \quad M = \begin{pmatrix} M_0 \\ M_1 \end{pmatrix}, \quad M_0 = \begin{pmatrix} \mu_{11} & \dots & \mu_{1m} \\ \vdots & & \vdots \\ \mu_{m1} & \dots & \mu_{mm} \end{pmatrix}, \quad M_1 = (\mu_{m+1,1}, \dots, \mu_{m+1,m}).$$

Clearly, if the above parameters $\lambda_{ij},\ \mu_{ij}$ satisfy $\lambda_{ij} = \mu_{ij} = 0$ (for $1 \le i \le j \le m$) and $\sum_{j=1}^{m} \lambda_{ij} = 1$ (for $2 \le i \le m+1$), then process (2.1) neatly reduces to an algorithm of the form (1.8). Therefore, the above process (2.1), with arbitrary matrices L and M, amounts to a generalization of the original Shu-Osher process (1.8).

In all of the following (unless specified otherwise) we shall denote by $\|.\|$ an *arbitrary real convex function* on $\mathbb{V}$, i.e., $\|v\| \in \mathbb{R}$ and $\|\lambda v + (1-\lambda)w\| \le \lambda\|v\| + (1-\lambda)\|w\|$ for all $v, w \in \mathbb{V}$ and $0 \le \lambda \le 1$.

We shall be interested in situations where, for given $F,\ \Delta t$, and convex function $\|.\|$,

$$\text{(2.3.a)} \quad \|y_i\| \le \|u_{n-1}\| \quad (1 \le i \le m),$$

$$\text{(2.3.b)} \quad \|u_n\| \le \|u_{n-1}\|,$$

when u_{n-1}, u_n and $y_i \in \mathbb{V}$ are related to each other as in (2.1). Clearly, property (2.3) extends and generalizes the TVD property (1.4); it is important, also with $\|.\|$ different from $\|.\|_{TV}$, and also when solving differential equations different from conservation laws (see, e.g., Dekker and Verwer [3], Hundsdorfer and Verwer [13], LeVeque [17]). Property (2.3.b), with $\|.\|$ not necessarily equal to $\|.\|_{TV}$, has been studied extensively in the literature and corresponds to what is often called *monotonicity*, *practical stability* or *strong stability* (see, e.g., Butcher [2, p. 392], Dekker and Verwer [3, p. 263], Gottlieb, Shu and Tadmor [7], Hundsdorfer, Ruuth and Spiteri [12], Morton [18]).

In the next subsection we shall study property (2.3) in the situation where, for some $\tau_0 > 0$, the function $F : \mathbb{V} \to \mathbb{V}$ satisfies

$$\|v + \tau_0 F(v)\| \le \|v\| \quad \text{(whenever } v \in \mathbb{V}). \tag{2.4}$$

Clearly, this condition is more general than (1.5)—in case $\mathbb{V} = \mathbb{R}^\infty$ and $\|.\| = \|.\|_{TV}$, assumption (1.5) implies (2.4).

In Theorem 2.2, to be presented below, we shall give conditions under which (2.1) is equivalent to (1.2). Moreover, we shall give restrictions on the stepsize Δt guaranteeing (2.3) for functions $F : \mathbb{V} \to \mathbb{V}$ satisfying (2.4).

2.2. **A generalization of the Shu-Osher Theorem 1.1.** Let an arbitrary Runge-Kutta method (A, b) be given. In order to represent it in the form (2.1), we assume that $L = (\lambda_{ij})$ is a given matrix of type (2.2.a). We define a corresponding matrix $M = (\mu_{ij})$ of type (2.2.b) by

$$M_0 = A - L_0 A, \quad M_1 = b^T - L_1 A. \tag{2.5}$$

The way of defining M_0 and M_1 in (2.5) can be viewed as a generalization of the definition of μ_{ij} in (1.7).

The coefficients μ_{ij}, corresponding to M_0, M_1 as in (2.5), depend only on the given Runge-Kutta scheme (A, b) and on the choice of the $(m+1) \times m$ parameter matrix $L = (\lambda_{ij})$. This justifies the following definition.

Definition 2.1. Process (2.1) is said to be generated by the coefficient scheme (A, b, L) if the coefficients μ_{ij} occurring in (2.1) are chosen according to (2.2), (2.5).

Theorem 2.2 below gives a condition on L under which the original Runge-Kutta process (1.2) is equivalent to the process (2.1) generated by (A, b, L). The theorem also specifies a stepsize restriction, of the form

$$0 < \Delta t \le c \cdot \tau_0, \tag{2.6}$$

under which (2.3) is valid for u_{n-1}, u_n, y_i satisfying (2.1).

Below we shall deal with matrices $L = (\lambda_{ij})$ of the form (2.2.a) which are such that

$$I - L_0 \text{ is invertible.} \tag{2.7}$$

Here, as well as in the following, we denote by I the $m \times m$ identity matrix. In Theorem 2.2 we shall pay special attention to the situation where the matrix $L = (\lambda_{ij})$ has been chosen in such a way that, in addition to (2.7),

$$\lambda_{ij} \ge 0 \quad \text{and} \quad \sum_{k=1}^{m} \lambda_{ik} \le 1 \quad (\text{ for } 1 \le i \le m+1,\ 1 \le j \le m). \tag{2.8}$$

This condition, on the parameters λ_{ij}, can be viewed as a generalization of the requirement that (1.6), (1.10) hold.

Furthermore, for given coefficient schemes (A, b, L), we shall use the notation

$$c(A,b,L) \;=\; \min\{c_{ij} :\; 1 \le i \le m+1,\;\; 1 \le j \le m\} \quad \text{where} \tag{2.9}$$

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

and the values $\lambda_{ij},\ \mu_{ij}$ are defined by (2.2), (2.5).

This notation can be regarded as a generalization of (1.11), (1.7). We note that there are two distinct situations in which the above values c_{ij} vanish: we have $c_{ij} = 0$ if either $\mu_{ij} < 0$ or $\lambda_{ij} = 0$, $\mu_{ij} > 0$, $i \neq j$.

The following theorem amounts to a generalization of Theorem 1.1, relevant to arbitrary Runge-Kutta methods (1.2). It constitutes the first of the two main theorems of our paper.

Theorem 2.2 (Generalization of the Shu-Osher theorem). *Let* (A, b) *specify an arbitrary Runge-Kutta method* (1.2). *Let* $L = (\lambda_{ij})$ *be any parameter matrix satisfying* (2.2.a), (2.7) *and consider the corresponding process* (2.1) *generated by* (A, b, L) *(cf. Definition 2.1). Then the following conclusions* (I) *and* (II) *are valid.*

(I) *The Runge-Kutta relations* (1.2) *are equivalent to* (2.1).

(II) *Assume additionally that* (2.8) *holds and the coefficient* c *is equal to* $c(A, b, L)$ *(see* (2.9)*). Let* F *be a function from* $\mathbb{V}$ *to* $\mathbb{V}$ *satisfying* (2.4). *Then, under the stepsize restriction* (2.6)*, process* (2.1) *has property* (2.3)*; i.e., the inequalities* (2.3) *are fulfilled whenever* u_{n-1}, u_n, *and* y_i *are related to each other as in* (2.1).

The above theorem will be proved in Section 2.3. Obviously, a combination of the above statements (I) and (II) immediately leads to *a conclusion which is highly relevant to the original Runge-Kutta method* (A, b)*: if* $L = (\lambda_{ij})$ *is any matrix satisfying* (2.2.a)*,* (2.7)*,* (2.8) *and* $c = c(A, b, L)$ *(see* (2.9)*), then the conditions* (2.4)*,* (2.6) *guarantee the monotonicity properties* (2.3) *whenever* u_{n-1}*,* u_n*,* y_i *satisfy* (1.2).

Let the Runge-Kutta method (A, b) be explicit. Choose any $(m+1)\times m$ matrix $L = (\lambda_{ij})$ such that its $m \times m$ submatrix L_0 (cf. (2.2.a)) is strictly lower triangular and $\sum_{j=1}^{m} \lambda_{ij} = 1$ (for $2 \le i \le m+1$). One easily sees that the corresponding process (2.1), generated by the coefficient scheme (A, b, L), coincides with the original Shu-Osher representation (1.8). Since L_0 is strictly lower triangular, condition (2.7) is fulfilled, and Theorem 2.2 can thus be applied so as to arrive easily at statements (i) and (ii) of Theorem 1.1. This shows that *Theorem 2.2 can be viewed as a neat generalization of Theorem 1.1.*

We note that the special implicit Runge-Kutta processes, analysed by Gottlieb, Shu and Tadmor [7, Section 6.2], are covered by our general formulation (2.1). In the analysis in the paper just mentioned, it was assumed that the first order implicit Euler discretization is unconditionally monotonic, i.e., $\|v\| \le \|v - \tau F(v)\|$ (for all $v \in \mathbb{V}$ and all positive stepsizes τ). This assumption is not required (explicitly) in our Theorem 2.2; we require instead condition (2.4) to be fulfilled. (Note that (2.4) implies $\|v\| = (1+\tau/\tau_0)\|v\| - (\tau/\tau_0)\|v\| \le (1+\tau/\tau_0)\|v\| - (\tau/\tau_0)\|v + \tau_0 F(v)\| \le \|(1+\tau/\tau_0)v - (\tau/\tau_0)(v + \tau_0 F(v))\| = \|v - \tau F(v)\|$; consequently, (2.4) implies that the above assumption about the implicit Euler discretization is automatically fulfilled.)

2.3. **Proving Theorem 2.2.**

Before giving the actual proof of Theorem 2.2, we introduce some notation which will be used below.

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

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

Let $B = (b_{ij})$ denote any (real) $l \times m$ matrix. 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 $w_i = \sum_{j=1}^m b_{ij} v_j$ (for $1 \le i \le l$). Clearly, if B and C are $l \times m$ matrices and D is an $m \times k$ matrix, then $(B+C)_{\mathbb{V}} = B_{\mathbb{V}} + C_{\mathbb{V}}$, $(\lambda B)_{\mathbb{V}} = \lambda \cdot B_{\mathbb{V}}$, $(BD)_{\mathbb{V}} = B_{\mathbb{V}} \cdot D_{\mathbb{V}}$. Here the addition and multiplication, occurring in the last three left-hand members, stand for the usual algebraic operations for matrices, whereas the addition and multiplication in the right-hand members apply to linear operators.

For clarity, we will also use the following simplified notation: $\mathbf{b}^T = (b^T)_{\mathbb{V}}$, $\mathbf{A} = A_{\mathbb{V}}$, $\mathbf{M}_0 = (M_0)_{\mathbb{V}}$, $\mathbf{M}_1 = (M_1)_{\mathbb{V}}$, $\mathbf{L}_0 = (L_0)_{\mathbb{V}}$ and $\mathbf{L}_1 = (L_1)_{\mathbb{V}}$. Furthermore, we define $\mathbf{I} = (I)_{\mathbb{V}}$ and $\mathbf{e} = (e)_{\mathbb{V}}$, where I is the $m \times m$ identity matrix and e is the column vector in $\mathbb{R}^m$ all of whose components are equal to 1.

The actual proof of Theorem 2.2.

(1) For proving conclusion (I), we have to show that the relations (2.1) are equivalent to (1.2). Using (2.5), (2.7), one easily sees that

$$\begin{aligned} \text{(2.1.a)} \quad &\iff \quad (\mathbf{I} - \mathbf{L}_0)[y_i] = (\mathbf{I} - \mathbf{L}_0)\mathbf{e}u_{n-1} + \Delta t \mathbf{M}_0[F(y_i)] \\ &\iff \quad [y_i] = \mathbf{e}u_{n-1} + \Delta t(\mathbf{I} - \mathbf{L}_0)^{-1}\mathbf{M}_0[F(y_i)] \iff \text{(1.2.a)}, \end{aligned}$$

so that (2.1.a) and (1.2.a) are equivalent. Therefore, assuming (2.1.a) or (1.2.a), we also have

$$\begin{aligned} \text{(2.1.b)} \quad &\iff \quad u_n = (1 - L_1 e)u_{n-1} + \mathbf{L}_1[y_i] + \Delta t \mathbf{M}_1[F(y_i)] \\ &\iff \quad u_n = (1 - L_1 e)u_{n-1} + \mathbf{L}_1\{\mathbf{e}u_{n-1} + \Delta t \mathbf{A}[F(y_i)]\} + \Delta t \mathbf{M}_1[F(y_i)] \\ &\iff \quad u_n = u_{n-1} + \Delta t(\mathbf{L}_1\mathbf{A} + \mathbf{M}_1)[F(y_i)] \iff \text{(1.2.b)}. \end{aligned}$$

This completes the proof of the equivalence of (2.1) and (1.2).

(2) If $c(A, b, L) = 0$, then conclusion (II) is trivially fulfilled. Therefore, in the following proof of (II), we assume $c(A, b, L) > 0$. This implies that, for all i, j,

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

We have to show (2.3) under the assumptions stated in Theorem 2.2. To this end, we put

$$x_i = \tau_0 F(y_i), \quad \alpha_i = \mu_{ii}\Delta t/\tau_0 \quad \text{and} \quad \beta_{ij} = \Delta t(\tau_0 c_{ij})^{-1},$$

where β_{ij} stands for zero in case $c_{ij} = \infty$. With this notation we obtain from (2.1.a), by using the convexity of the function $\|.\|$,

$$\text{(2.10)} \qquad \|y_i - \alpha_i x_i\| \le (1 - \sum_{j=1}^m \lambda_{ij})\|u_{n-1}\| + \lambda_{ii}\|y_i\| + \sum_{j \ne i} \lambda_{ij}\|y_j + \beta_{ij} x_j\|,$$

for $1 \le i \le m$. From (2.4) we have $\|y_i + x_i\| \le \|y_i\|$. Therefore, by using the relation $(1+\alpha_i)y_i = (y_i - \alpha_i x_i) + \alpha_i(y_i + x_i)$, we obtain $\|y_i\| \le \theta\|y_i - \alpha_i x_i\| + (1-\theta)\|y_i\|$, with $\theta = (1+\alpha_i)^{-1}$. Hence

$$\|y_i - \alpha_i x_i\| \;\ge\; \|y_i\|. \tag{2.11}$$

Similarly, by using the relation $y_j + \beta_{ij}x_j = (1-\beta_{ij})y_j + \beta_{ij}(y_j + x_j)$, we see that

$$\|y_j + \beta_{ij}x_j\| \;\le\; \|y_j\|. \tag{2.12}$$

Combining inequalities (2.10), (2.11) and (2.12), we obtain a bound for $\|y_i\|$ $(1 \le i \le m)$ which can be written compactly in the form

$$(I - L_0)\;[\|y_i\|] \;\le\; \|u_{n-1}\|(I - L_0)e. \tag{2.13}$$

This inequality, between two vectors in $\mathbb{R}^m$, should be interpreted componentwise.

From (2.13) we easily obtain (2.3.a), provided the entries r_{ij} of the matrix $R = (r_{ij}) = (I - L_0)^{-1}$ are nonnegative. In view of (2.7) and (2.8), we see that the matrix $K(t) = (I - tL_0)^{-1}$ (for $0 \le t \le 1$) exists and depends continuously on t. For $0 \le t < 1$ we have $K(t) = I + tL_0 + (tL_0)^2 + \cdots$ so that the entries of $K(t)$ are nonnegative. Therefore, the entries r_{ij} of $R = K(1)$ must be nonnegative as well, which thus proves (2.3.a).

In order to prove (2.3.b), we note that (2.1.b) implies

$$\|u_n\| \;\le\; \theta\|u_{n-1}\| + \sum_{j=1}^{m} \lambda_{m+1,j}\|y_j + \beta_{m+1,j}x_j\|,$$

where $\theta = 1 - \sum_{j=1}^{m} \lambda_{m+1,j}$. Hence,

$$\begin{aligned}\|u_n\| &\le \theta\|u_{n-1}\| + \sum_{j=1}^{m} \lambda_{m+1,j}\|y_j\| \\ &\le (\theta + \sum_{j=1}^{m} \lambda_{m+1,j})\|u_{n-1}\| = \|u_{n-1}\|.\end{aligned}$$ □

3. Maximizing the coefficient $c(A, b, L)$

3.1. **Irreducible Runge-Kutta schemes and the quantity** $R(A, b)$**.** In this section we give some definitions which will be needed when we formulate our results, in Section 3.2, about the maximum value of the important coefficient $c(A, b, L)$ (see (2.9)). We start with the fundamental concepts of reducibility and irreducibility.

Definition 3.1 (Reducibility and irreducibility)**.** An m-stage Runge-Kutta scheme (A, b) is called reducible if (at least) one of the following two statements (a), (b) is true; it is called irreducible if neither (a) nor (b) is true.

(a) 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 $i \in M$, $j \in N$).

(b) There exist nonempty, pairwise disjoint index sets $M_1, M_2, \ldots M_r$, with $1 \le r < m$ 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 $1 \le p \le r$, $1 \le q \le r$ and $i, j \in M_p$.

In case the above statement (a) is true, the vectors y_j in (1.2) with $j \in N$ have no influence on u_n, so that the Runge-Kutta method is equivalent to a method with less than m stages. Also in case of (b), the Runge-Kutta method essentially reduces to a method with less than m stages; see, e.g., Dekker and Verwer [3] or Hairer and Wanner [9]. Clearly, from a practical point of view, it is enough to consider only Runge-Kutta schemes which are irreducible.

Next, we turn to an important characteristic quantity for Runge-Kutta schemes introduced by Kraaijevanger [14]. Following this author, we shall denote his quantity by $R(A,b)$, and in defining it, we shall use, for real ξ, the notation

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

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 $\xi \le 0$ for which

$$\text{(3.1)} \qquad I-\xi A \text{ is invertible}, \quad A(\xi)\ge 0, \quad b(\xi)\ge 0, \quad e(\xi)\ge 0, \quad \text{and} \quad \varphi(\xi)\ge 0.$$

The first inequality in (3.1) should be interpreted entrywise, the second and third ones componentwise. Similarly, all inequalities for matrices and vectors occurring below are to be interpreted entrywise and componentwise, respectively.

Definition 3.2 (The quantity $R(A,b)$)**.** Let (A,b) be a given coefficient scheme. In case $A\ge 0$ and $b \ge 0$, we define

$$R(A,b) = \sup\{r: \ r\ge 0 \ \text{ and (3.1) holds for all } \xi \text{ with } -r \le \xi \le 0\}.$$

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

Definition 3.2 may suggest that it is difficult to determine the quantity $R(A,b)$ for a given coefficient scheme (A,b). But, parts (i) and (iii) of the following Theorem 3.3 show that it is relatively easy to decide whether $R(A,b)=0$ or $R(A,b)=\infty$. Moreover, part (ii) of the theorem can be exploited for simplifying the (numerical) computation of $R(A,b)$, if $0 < R(A,b) < \infty$; cf. Ferracina and Spijker [4, Section 4.3], Kraaijevanger [14, p. 498].

In order to formulate part (i) of Theorem 3.3 concisely, we define, for any given $m\times m$ matrix $B=(b_{ij})$, the corresponding $m\times m$ *incidence matrix* by

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

Theorem 3.3 (Kraaijevanger)**.** *Let (A,b) be an irreducible coefficient scheme. Then*

(i) $R(A,b)>0$ *if and only if* $A\ge 0$, $b>0$ *and* $\mathrm{Inc}(A^2)\le \mathrm{Inc}(A)$.

(ii) *Let* $0<r<\infty$*. Then* $R(A,b)\ge r$ *if and only if* $A\ge 0$ *and conditions* (3.1) *hold at* $\xi=-r$.

(iii) $R(A,b)=\infty$ *if and only if*

- *A is invertible and all off-diagonal entries of A^{-1} are nonpositive,*
- *$A\ge 0$ and $A^{-1}e\ge 0$,*
- *$b^TA^{-1}\ge 0$ and $b^TA^{-1}e\le 1$.*

Parts (i), (ii), (iii) of the above theorem have been taken almost literally from Kraaijevanger [14, Theorem 4.2, Lemma 4.4 and Theorem 4.7, respectively].

We shall make use of the quantity $R(A,b)$ in formulating our results below in Section 3.2, whereas Theorem 3.3 will be essential for proving our results, in Section 3.3.

3.2. **The special parameter matrix L^*.** The following Theorem 3.4 constitutes the second of the two main theorems of our paper. It resolves the problem of finding a parameter matrix $L = (\lambda_{ij})$ such that the crucial coefficient $c(A, b, L)$ (see (2.9)) attains its maximal value and it also gives interesting properties of this maximal value.

In the theorem, the focus will be on the following matrix L^*:

(3.2.a)

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

with

(3.2.b)

$$L_0^* = \gamma A(I + \gamma A)^{-1}, \qquad L_1^* = \gamma b^T (I + \gamma A)^{-1}, \quad \gamma = R(A, b)$$
$$(\text{if} \quad 0 \le R(A, b) < \infty),$$

(3.2.c)

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

The above matrix L^* seems to appear out of the blue. But, the authors were led to introduce this matrix by analysing calculations of Kraaijevanger [14, Sections 5.3 and 6]. For more details, we refer the interested reader to that important paper.

Theorem 3.4 (The largest coefficient $c(A, b, L)$)**.** *Let the Runge-Kutta method* (1.2) *be specified by an arbitrary irreducible coefficient scheme* (A, b)*. Then the inverses occurring in* (3.2.b), (3.2.c) *do exist, so that we can define the matrix* $L^* = (\lambda_{ij}^*)$ *by* (3.2)*. Furthermore, the matrix* $L = L^*$ *satisfies* (2.2.a), (2.7), (2.8), *and the corresponding coefficient* $c(A, b, L^*)$ *(see* (2.9)*) has the following properties:*

(I) $c(A, b, L^*) = \max_L c(A, b, L)$, *where the maximum is over all matrices* $L = (\lambda_{ij})$ *satisfying* (2.2.a), (2.7), (2.8).

(II) $c(A, b, L^*)$ *is equal to the maximal coefficient* c *for which conditions* (1.5), (1.9) *imply the TVD property* (1.4) *whenever* u_{n-1}, u_n, $y_i \in \mathbb{R}^\infty$ *satisfy* (1.2).

(III) $c(A, b, L^*) = R(A, b)$ *(see Definition 3.2).*

The above theorem will be proved in Section 3.3. Clearly, the above property (I) shows how to maximize the coefficient $c(A, b, L)$ over all relevant matrices L, whereas property (II) brings to light the fact that the coefficient $c(A, b, L^*)$ is optimal—not only in the context of maximizing $c(A, b, L)$ but also—in the important context of optimizing arbitrary stepsize restrictions (of type (1.9)) which guarantee the TVD property (1.4) for process (1.2). Finally, property (III) gives a neat expression for the maximal coefficient $c(A, b, L^*)$. We shall come back to the relevance of Theorem 3.4 in Section 4.

3.3. **Proving Theorem 3.4.**

3.3.1. *The proof that L^* satisfies* (2.7), (2.8) *and* (III).

(1) *Assume* $0 \leq R(A, b) < \infty$.

One easily sees, from Theorem 3.3, that the inverse occurring in (3.2.b) exists. We consider the $(m+1) \times m$ matrix $L^* = (\lambda^*_{ij})$ defined by (3.2.a), (3.2.b). From (3.2.b) we see that $I - L_0^* = (I + \gamma A)^{-1}$ so that $L_0 = L_0^*$ satisfies (2.7).

Using Theorem 3.3, we easily arrive at the inequalities $L_0^* \geq 0$ and $(I - L_0^*)e = (I + \gamma A)^{-1}e \geq 0$. Consequently, $\lambda_{ij} = \lambda^*_{ij}$ satisfy the requirements occurring in (2.8) for $1 \leq i \leq m$. Similarly, using Theorem 3.3 once more, we see that $L_1^* \geq 0$ and $1 - L_1^* e = 1 - \gamma b^T (I + \gamma A)^{-1} e \geq 0$ so that $\lambda_{ij} = \lambda^*_{ij}$ satisfy the requirements in condition (2.8) also for $i = m + 1$.

In order to prove (III), we consider the $(m+1) \times m$ matrix $M^* = (\mu^*_{ij})$ defined by $M^* = \begin{pmatrix} M_0^* \\ M_1^* \end{pmatrix}$, where M_0^*, M_1^* are given by (2.5) (with L_0, L_1, M_0, M_1 replaced by $L_0^*, L_1^*, M_0^*, M_1^*$, respectively). Clearly,

$$L_0^* = \gamma M_0^*, \quad L_1^* = \gamma M_1^*. \tag{3.3}$$

In view of (2.5), (3.3) and Theorem 3.3, we have $b^T = M_1^* + L_1^* A = M_1^*(I + \gamma A)$ with $(I + \gamma A) \geq 0$. Since $\sum b_j = 1$, it follows that there is an index k with:

$$1 \leq k \leq m \quad \text{and} \quad \mu^*_{m+1,k} > 0. \tag{3.4}$$

If all $\mu^*_{ij} \geq 0$, then we see from (2.9), (3.3), (3.4) that $c(A, b, L^*) = \gamma$, i.e., (III). On the other hand, if there is a $\mu^*_{ij} < 0$, then we conclude from (2.9), (3.3) that $c(A, b, L^*) = 0$ and $\gamma = 0$, i.e., again (III).

(2) *Assume* $R(A, b) = \infty$.

One easily sees, from Theorem 3.3, that the inverse A^{-1} occurring in (3.2.c) exists. Since $p_{ii} a_{ii} = 1 - \sum_{k \neq i} p_{ik} a_{ki}$, we can also conclude from Theorem 3.3 that $p_{ii} > 0$, so that γ in (3.2.c) is well defined, with $0 < \gamma < \infty$.

Defining L^* by (3.2.a), (3.2.c), and $M^* = \begin{pmatrix} M_0^* \\ M_1^* \end{pmatrix}$ again by (2.5) (with L_0, L_1, M_0, M_1 replaced by $L_0^*, L_1^*, M_0^*, M_1^*$, respectively), one has

$$M_0^* = \gamma I, \quad M_1^* = 0.$$

Consequently, $c(A, b, L^*)$ (see (2.9)) satisfies (III).

From (3.2.c) it follows that $I - L_0^* = \gamma A^{-1}$ so that $L_0 = L_0^*$ satisfies (2.7).

Using Theorem 3.3 and the definition of γ, it is easy to prove $L_0^* \geq 0$, $L_1^* \geq 0$, $(I - L_0^*)e = \gamma A^{-1} e \geq 0$ and $1 - L_1^* e = 1 - b^T A^{-1} e \geq 0$. The last four inequalities imply that the matrix $L = L^*$ satisfies (2.8).

3.3.2. *The proof of* (I) *and* (II). In proving the remaining properties (I), (II), we shall make use of the following lemma, which immediately follows from Ferracina and Spijker [4, Theorem 2.5].

Lemma 3.5. *Consider an arbitrary irreducible Runge-Kutta scheme* (A, b). *Let* c *be any value, with* $0 \leq c \leq \infty$, *such that conditions* (1.5), (1.9) *imply the TVD property* (1.4) *whenever* $u_{n-1},\ u_n,\ y_i \in \mathbb{R}^\infty$ *satisfy* (1.2). *Then* $c \leq R(A, b)$.

From Theorem 2.2 we see that, given any matrix L satisfying (2.2.a), (2.7), (2.8), the coefficient $c = c(A, b, L)$, defined via (2.9), is such that conditions (1.5), (1.9) imply the TVD property (1.4) whenever $u_{n-1},\ u_n,\ y_i \in \mathbb{R}^\infty$ satisfy (1.2). Hence, by Lemma 3.5,

$$c(A,b,L) \le R(A,b) \quad \text{(whenever } L \text{ satisfies (2.2.a), (2.7), (2.8))}.$$

This shows that property (I) follows from property (III). Moreover, by using Lemma 3.5 once more and applying Theorem 2.2 with matrix L^*, we see that property (II) also follows from (III). □

4. Applications and illustrations of Theorems 2.2 and 3.4

4.1. **Applications to general Runge-Kutta methods.** In [14], interesting relations were revealed between the order of accuracy p, of m-stage Runge-Kutta schemes (A,b), and the size of $R(A,b)$ (Definition 3.2); in [4, Section 4] a review of these results was presented. Combining Kraaijevanger's findings with our Theorem 3.4, one easily obtains interesting relations between the order p and the size of $c(A,b,L)$. As an important illustration, we give the following corollary to Theorem 3.4; for the concept of irreducibility, occurring in the corollary, see Definition 3.1.

Corollary 4.1. *Let the Runge-Kutta method* (1.2) *be specified by an arbitrary irreducible coefficient scheme* (A,b)*. Assume the method has an order of accuracy greater than one. Then, for any matrix* $L = (\lambda_{ij})$*, satisfying* (2.2.a)*,* (2.7)*,* (2.8)*, the corresponding coefficient* $c(A,b,L)$ *(see* (2.9)*) is finite.*

Proof. In [14, p. 514], it was shown that $R(A,b) < \infty$ if the order of the method is greater than one. An application of Theorem 3.4 (parts (I), (III)) completes the proof. □

Next, we turn to a corollary obtainable by combining Theorems 2.2 and 3.4.

Corollary 4.2. *For any given irreducible Runge-Kutta scheme* (A,b) *the following two statements are valid.*

(I) *Let* $c = R(A,b)$*. Then, for all vector spaces* $\mathbb{V}$ *and convex functions* $\|.\|$ *on* $\mathbb{V}$*, conditions* (2.4)*,* (2.6) *guarantee the monotonicity properties* (2.3)*, whenever* $u_{n-1},\ u_n,\ y_i$ *satisfy* (1.2)*.*

(II) *The value* $c = R(A,b)$ *in the above statement* (I) *is optimal in that, for any value* $c > R(A,b)$*, the general conclusion as given in statement* (I) *is no longer true.*

Proof. In order to prove (I), we note that by Theorem 3.4 the coefficient $c = R(A,b)$ is equal to $c(A,b,L^*)$, where $L = L^*$ satisfies (2.2.a), (2.7), (2.8). An application of parts (I), (II) of Theorem 2.2, with $L = L^*$, thus shows that conditions (2.4), (2.6) imply (2.3) for $u_{n-1},\ u_n,\ y_i$ satisfying (1.2).

In order to prove statement (II) of the corollary, suppose that the general conclusion as given in statement (I) of the corollary is true for some $c > R(A,b)$. Then, with this c, conditions (1.5), (1.9) would imply (1.4) for $u_{n-1},\ u_n,\ y_i$ satisfying (1.2). Lemma 3.5 shows that $c \le R(A,b)$, which yields a contradiction. □

The above corollary can be viewed as a variant to one of the results given in [4, Theorem 2.5]. The conclusion, given above in statement (I), is stronger than an analogous monotonicity result in the paper just mentioned—because (I) deals with arbitrary convex functions (rather than seminorms) and property (2.3) gives not only a bound for $\|u_n\|$ but also for $\|y_i\|$.

We finally note that the relevance of Theorem 3.4 is not restricted to properties (1.4) and (2.3). The theorem may be applied as well in the analysis of other interesting (stability and boundedness) properties studied in the literature; cf., e.g., Dekker and Verwer [3, pp. 38, 39], Gottlieb, Shu and Tadmor [7, p. 92].

4.2. **Applications to explicit Runge-Kutta methods.** In this section, we shall make use of Theorem 3.4 in resolving, for explicit Runge-Kutta methods (A, b), the two questions related to the coefficient $c(A, b)$ as raised at the end of Section 1.1. Due to the restriction $\sum_j \lambda_{ij} = 1$ (cf. (1.6)), which occurs in the original Shu-Osher representation but not in our generalized representation (cf. Sections 2, 3), Theorem 3.4 will have to be applied with some care.

Our Theorem 4.3 answers the two questions just mentioned. Property (I), in the theorem, makes clear how to choose parameters $\lambda_{ij} = \tilde{\lambda}_{ij}$ satisfying (1.6), (1.10) such that the corresponding coefficient $\tilde{c}$ (see (1.11), (1.7)) is maximal, i.e., $\tilde{c} = c(A, b)$. In addition, property (II), in the theorem, shows that no coefficient c greater than $\tilde{c} = c(A, b)$ exists for which the conditions (1.5), (1.9) still guarantee the TVD property (1.4) for process (1.2). Finally, property (III), in the theorem, relates the maximal coefficient $\tilde{c} = c(A, b)$ to Kraaijevanger's quantity $R(A, b)$. The proof of Theorem 4.3 will be based on Theorem 3.4.

The concept of irreducibility and the quantity $R(A, b)$, which occur in Theorem 4.3, are defined above in Section 3.1.

Theorem 4.3 (The largest coefficient c of the form (1.11))**.** *Consider an arbitrary irreducible explicit Runge-Kutta method* (A, b)*. Then* $0 \le R(A, b) < \infty$*, and the inverse occurring in* (3.2.b) *exists so that we can define the matrix* $L^* = (\lambda^*_{ij})$ *by* (3.2.a), (3.2.b)*. Let parameters* $\tilde{\lambda}_{ij}$ *be defined by*

$$\text{(4.1.a)} \qquad \tilde{\lambda}_{ij} = 1 - \sum_{k=2}^{m} \lambda^*_{ik} \quad (\text{for } 2 \le i \le m+1 \text{ and } j = 1),$$

$$\text{(4.1.b)} \qquad \tilde{\lambda}_{ij} = \lambda^*_{ij} \quad (\text{for } 2 \le i \le m+1 \text{ and } 2 \le j \le i-1),$$

and corresponding values $\tilde{\mu}_{ij}$ *via* (1.7)*. Then the parameters* $\lambda_{ij} = \tilde{\lambda}_{ij}$ *satisfy* (1.6), (1.10)*, and the corresponding coefficient* $c = \tilde{c}$ *(defined by* (1.11) *with* $\lambda_{ij} = \tilde{\lambda}_{ij}$ *and* $\mu_{ij} = \tilde{\mu}_{ij}$*) has the following properties:*

(I) $\tilde{c}$ *is the largest coefficient, obtainable from* (1.11) *with any parameters* λ_{ij}, μ_{ij} *satisfying* (1.6), (1.7), (1.10).

(II) $\tilde{c}$ *is equal to the largest coefficient* c *for which the conditions* (1.5), (1.9) *imply the TVD property* (1.4) *whenever* u_{n-1}, u_n, $y_i \in \mathbb{R}^\infty$ *satisfy* (1.2).

(III) $\tilde{c} = R(A, b)$.

Proof. Since A is strictly lower triangular, one easily sees from Theorem 3.3 that $R(A, b) < \infty$ and the inverse occurring in (3.2.b) exists.

Clearly, the parameters $\lambda_{ij} = \tilde{\lambda}_{ij}$ satisfy condition (1.6).

From Theorem 3.4 we know that $L = L^*$ satisfies (2.8), so that the parameters $\lambda_{ij} = \tilde{\lambda}_{ij}$ also satisfy (1.10).

Define $(m+1) \times m$ matrices, with a structure as in (2.2), by $\tilde{L} = (\tilde{\lambda}_{ij})$, $\tilde{M} = (\tilde{\mu}_{ij})$, where $\tilde{\lambda}_{ij}$, $\tilde{\mu}_{ij}$ (for $j < i$) satisfy (4.1) and (1.7), and $\tilde{\lambda}_{ij}$, $\tilde{\mu}_{ij}$ (for $j \ge i$) are defined to be zero. One easily sees that $L = \tilde{L}$ and $M = \tilde{M}$ satisfy (2.5), (2.7), (2.8), and that

$$\tilde{c} = c(A, b, \tilde{L}).$$

In order to be able to apply Theorem 3.4 to the situation at hand, we shall now relate $c(A, b, \tilde{L})$ to the coefficient $c(A, b, L^*)$.

From (3.2.b) we see that L_0^* is strictly lower triangular. This implies, in view of (4.1), that $\tilde{L}$ and L^* differ only in their first column and that $\tilde{L} \geq L^*$. Denoting by M^* the matrix which is related to L^* as in (2.5), it follows that $\tilde{M} - M^* = (L^* - \tilde{L})A = 0$. Consequently, $\tilde{M} = M^*$ so that $c(A, b, \tilde{L}) \geq c(A, b, L^*)$. In view of Theorem 3.4, we thus have

$$c(A, b, \tilde{L}) = c(A, b, L^*).$$

We conclude that $\tilde{c} = c(A, b, L^*)$, which in combination with Theorem 3.4 easily leads to the properties (I), (II), (III) of Theorem 4.3. □

Let $E_{m,p}$ denote the class of all explicit m-stage Runge-Kutta methods with (classical) order of accuracy at least p. As mentioned in Section 1.1, much attention has been paid in the literature to finding methods (A, b) of class $E_{m,p}$ which are optimal in $E_{m,p}$ with respect to the coefficient $c(A, b)$ (introduced in Section 1.1); see, e.g., Gottlieb and Shu [6], Ruuth and Spiteri [19], Shu [21], Shu and Osher [22], Spiteri and Ruuth [24]. Independently of this work, in [14], methods (A, b) were identified that are optimal in $E_{m,p}$ with respect to $R(A, b)$. In [4, Section 4], the remarkable fact was noted (but not explained!) that the methods identified in [14] coincide with methods (A, b) obtained in the above literature on optimization with respect to $c(A, b)$; cf. also Example 4.4 in Section 4.3 below. Theorem 4.3 allows us to fully understand this fact: by definition, $c(A, b)$ is equal to $\tilde{c}$ in property (I) of the theorem, so that, in view of property (III),

$$c(A, b) = R(A, b). \tag{4.2}$$

This equality makes clear that any method which is optimal in the sense of $c(A, b)$ is also optimal with respect to $R(A, b)$.

Relation (4.2) is also relevant, e.g., to the nonexistence of methods (A, b) with $c(A, b) > 0$ in $E_{4,4}$ and in $E_{m,5}$—as proved in [6] and [19], respectively. According to Kraaijevanger [14, pp. 516, 521], for any method (A, b) of class $E_{4,4}$ or $E_{m,5}$, we have $R(A, b) = 0$, which via (4.2) immediately leads to $c(A, b) = 0$.

4.3. **Illustrations to Theorems 3.4 and 4.3.**

We give two examples illustrating Theorems 3.4 and 4.3 in the construction of (generalized) Shu-Osher representations with maximal coefficients $c(A, b, L)$.

Example 4.4 (Illustration to Theorem 4.3)**.** Consider the explicit Runge-Kutta method (1.2), with $m = 4$ and

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \\ 1/6 & 1/6 & 1/6 & 0 \end{pmatrix}, \quad b^T = (1/6,\ 1/6,\ 1/6,\ 1/2). \tag{4.3}$$

Kraaijevanger [14, Theorem 9.5] proved that this method is of third order and $R(A, b) = 2$, whereas there exists no other explicit third order method with $m = 4$ and $R(A, b) \geq 2$.

Define parameters $\tilde{\lambda}_{ij}, \tilde{\mu}_{ij}$ as in Theorem 4.3. It is easy to see that the coefficients $\lambda_{ij} = \tilde{\lambda}_{ij}$, $\mu_{ij} = \tilde{\mu}_{ij}$ in the corresponding process (1.8) are as follows:

$$\begin{pmatrix} \tilde{\lambda}_{21} & & & \\ \tilde{\lambda}_{31} & \tilde{\lambda}_{32} & & \\ \tilde{\lambda}_{41} & \tilde{\lambda}_{42} & \tilde{\lambda}_{43} & \\ \tilde{\lambda}_{51} & \tilde{\lambda}_{52} & \tilde{\lambda}_{53} & \tilde{\lambda}_{54} \end{pmatrix} = \begin{pmatrix} 1 & & & \\ 0 & 1 & & \\ \frac{2}{3} & 0 & \frac{1}{3} & \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

$$\begin{pmatrix} \tilde{\mu}_{21} & & & \\ \tilde{\mu}_{31} & \tilde{\mu}_{32} & & \\ \tilde{\mu}_{41} & \tilde{\mu}_{42} & \tilde{\mu}_{43} & \\ \tilde{\mu}_{51} & \tilde{\mu}_{52} & \tilde{\mu}_{53} & \tilde{\mu}_{54} \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & & & \\ 0 & \frac{1}{2} & & \\ 0 & 0 & \frac{1}{6} & \\ 0 & 0 & 0 & \frac{1}{2} \end{pmatrix}.$$

We see that, as predicted by Theorem 4.3, the coefficient $\tilde{c}$, defined by (1.11) (with $\lambda_{ij} = \tilde{\lambda}_{ij},\ \mu_{ij} = \tilde{\mu}_{ij}$), satisfies

$$\tilde{c} = 2.$$

Moreover, applying Theorem 4.3 once more, we immediately arrive at the following two interesting conclusions.

1. For any explicit third order method with four stages, different from (4.3), there exist no parameters $\lambda_{ij},\ \mu_{ij}$, satisfying (1.6), (1.7), (1.10), such that the corresponding coefficient c (see (1.11)) satisfies $c \geq 2$.
2. For any explicit third order method with four stages, different from (4.3), there exists no coefficient $c \geq 2$ such that the conditions (1.5), (1.9) guarantee (1.4) (whenever $u_{n-1},\ u_n,\ y_i$ satisfy (1.2)).

It is interesting to note that the numerical process (1.8) with the above parameter values $\lambda_{ij} = \tilde{\lambda}_{ij},\ \mu_{ij} = \tilde{\mu}_{ij}$ was also recently found by numerical computations based on optimization of c, (1.11), with respect to the parameters λ_{ij}, μ_{ij}; see Spiteri and Ruuth [24]. However, the last-mentioned paper gives no proof of our two conclusions stated above.

Example 4.5 (Illustration to Theorem 3.4)**.** Consider the singly diagonally implicit Runge-Kutta (SDIRK) method (1.2), with $m = 2$ and

$$A = \begin{pmatrix} 1/4 & 0 \\ 1/2 & 1/4 \end{pmatrix}, \quad b^T = (1/2,\ 1/2). \tag{4.4}$$

This method is algebraically stable and of second order; see Burrage [1]. A simple calculation shows that $R(A, b) = 4$. Moreover, it can be seen, by straightforward calculations using Theorem 3.3, that method (4.4) is optimal in that there exists no other second order SDIRK method with $m = 2$ and $R(A, b) \geq 4$.

We define matrices $L = L^* = (\lambda^*_{ij})$ and $M = M^* = (\mu^*_{ij})$, corresponding to (4.4), by (2.2), (2.5), (3.2). These matrices are as follows:

$$\begin{pmatrix} \lambda^*_{11} & \lambda^*_{12} \\ \lambda^*_{21} & \lambda^*_{22} \\ \lambda^*_{31} & \lambda^*_{32} \end{pmatrix} = \begin{pmatrix} 1/2 & 0 \\ 1/2 & 1/2 \\ 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} \mu^*_{11} & \mu^*_{12} \\ \mu^*_{21} & \mu^*_{22} \\ \mu^*_{31} & \mu^*_{32} \end{pmatrix} = \begin{pmatrix} 1/8 & 0 \\ 1/8 & 1/8 \\ 0 & 1/4 \end{pmatrix}.$$

We see that, as predicted by Theorem 3.4, the coefficient $c(A, b, L^*)$, computed from (2.9) (with $L = L^*$), satisfies

$$c(A, b, L^*) = 4.$$

Moreover, applying Theorem 3.4 once more, we obtain the following two interesting conclusions.

1. For any second order SDIRK method with two stages, different form (4.4), there exists no matrix $L = (\lambda_{ij})$ satisfying (2.2.a), (2.7), (2.8), such that the corresponding coefficient $c(A, b, L)$ (see (2.9)) satisfies $c(A, b, L) \geq 4$.
2. For any second order SDIRK method with two stages, different form (4.4), there exists no coefficient $c \geq 4$ such that conditions (1.5), (1.9) guarantee (1.4) (whenever $u_{n-1},\ u_n,\ y_i$ satisfy (1.2)).

References

[1] K. Burrage (1982): Efficiently implementable algebraically stable Runge-Kutta methods, *SIAM J. Numer. Anal.* **19**, 245-258. MR83d:65235

[2] J.C. Butcher (1987): *The numerical analysis of ordinary differential equations*, John Wiley (Chichester). MR88d:65002

[3] K. Dekker and J.G. Verwer (1984): *Stability of Runge-Kutta methods for stiff nonlinear differential equations*, North-Holland Publ. Comp. (Amsterdam). MR86g:65003

[4] L. Ferracina and M.N. Spijker (2002): Stepsize restrictions for the total-variation-diminishing property in general Runge-Kutta methods, Report MI 2002-21 (2002), Mathematical Institute, University of Leiden. To appear in *SIAM J. Numer. Anal.*

[5] A. Gerish and R. Weiner (2003): On the positivity of low order explicit Runge-Kutta schemes applied in splitting methods, *Computers and Mathematics with Applications* **45**, 53–67.

[6] S. Gottlieb and C.-W. Shu (1998): Total-variation-diminishing Runge-Kutta schemes, *Math. Comp.* **67**, 73–85. MR98c:65122

[7] S. Gottlieb, C.-W. Shu, and E. Tadmor (2001): Strong-stability-preserving high-order time discretization methods, *SIAM Review* **43**, 89–112. MR2002f:65132

[8] E. Hairer, S.P. Nørsett, and G. Wanner (1987): *Solving ordinary differential equations I*, Springer Verlag (Berlin). MR87m:65005

[9] E. Hairer and G. Wanner (1996): *Solving ordinary differential equations II. Stiff and differential-algebraic problems*, Second Revised Edition, Springer (Berlin). MR97m:65007

[10] A. Harten (1983): High resolution schemes for hyperbolic conservation laws, *J. Comput. Phys.* **49**, 357–393. MR84g:65115

[11] I. Higueras (2002): On strong stability preserving time discretization methods, Report n.2 (2002), Departamento de Matemática e Informática, Universidad Pública de Navarra.

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

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

[14] J.F.B.M. Kraaijevanger (1991): Contractivity of Runge-Kutta methods, *BIT* **31**, 482–528. MR92i:65120

[15] D. Kröner (1997): *Numerical schemes for conservation laws*, Wiley, Teubner (Chichester, Stuttgart). MR98b:65003

[16] C.B. Laney (1998): *Computational gas dynamics*, Cambridge University Press (Cambridge). MR2000e:76086

[17] R.J. LeVeque (2002): *Finite volume methods for hyperbolic problems*, Cambridge University Press (Cambridge). MR2003h:65001

[18] K.W. Morton (1980): Stability of difference approximations to a diffusion-convection equation, *Int. J. Num. Meth. Eng.* **15**, 677–683. MR82i:76080

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

[20] C.-W. Shu (1988): Total-variation-diminishing time discretizations, *SIAM J. Sci. Stat. Comput.* **9**, 1073-1084. MR90a:65196

[21] C.-W. Shu (2002): A survey of strong stability preserving high-order time discretizations, *Collected Lectures on the Preservation of Stability under Discretization*, D. Estep, S. Tavener Editors, SIAM (Philadelphia, PA), 51–65.

[22] C.-W. Shu and S. Osher (1988): Efficient implementation of essentially non-oscillatory shock-capturing schemes, *J. Comput. Phys.* **77**, 439–471. MR89g:65113

[23] M.N. Spijker (1983): Contractivity in the numerical solution of initial value problems, *Numer. Math.* **42**, 271–290. MR85b:65067

[24] R.J. Spiteri and S.J. Ruuth (2002): A new class of optimal high-order strong-stability-preserving time discretization methods, *SIAM J. Numer. Anal.* **40**, 469–491. MR2003g:65083

[25] E.F. Toro (1999): *Riemann Solvers and Numerical Methods for Fluid Dynamics*, Second Edition, Springer-Verlag (Berlin). MR2000f:76091

Mathematical Institute, Leiden University, Niels Bohrweg 1, 2333 CA Leiden, The Netherlands
E-mail address: ferra@math.leidenuniv.nl

Mathematical Institute, Leiden University, Niels Bohrweg 1, 2333 CA Leiden, The Netherlands
E-mail address: spijker@math.leidenuniv.nl