Monotonicity and Boundedness in general Runge-Kutta methods

Proefschrift

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van de Rector Magnificus Dr. D.D. Breimer,
hoogleraar in de faculteit der Wiskunde en
Natuurwetenschappen en die der Geneeskunde,
volgens besluit van het College voor Promoties
te verdedigen op dinsdag 6 september 2005
klokke 15.15 uur

door

Luca Ferracina

geboren te Vicenza, Italië
in 1973

Samenstelling van de promotiecommissie:

promotors:	Prof.dr. M.N. Spijker Prof.dr. J.G. Verwer (UvA/CWI)
referent:	Dr. W. Hundsdorfer (CWI)
overige leden:	Prof.dr. G. van Dijk Prof.dr.ir. L.A. Peletier Prof.dr. S.M. Verduyn Lunel

Alla mia famiglia:
papà, mamma, Marco, Anna.

Monotonicity and Boundedness in general Runge-Kutta methods

Thomas Stieltjes Institute
for Mathematics

Preface

This thesis consists of an introduction and four papers which appeared (or were submitted for publication) in scientific journals. The introduction has been written with the intention to be understandable also for the reader who is not specialized in the field. The papers, which are listed below, are essentially self-contained, and each of them may be read independently of the others.

FERRACINA L., SPIJKER M.N. (2004): Stepsize restrictions for the total-variation-diminishing property in general Runge-Kutta methods, *SIAM J. Numer. Anal.* **42**, 1073–1093.

FERRACINA L., SPIJKER M.N. (2005): An extension and analysis of the Shu-Osher representation of Runge-Kutta methods, *Math. Comp.* **249**, 201–219.

FERRACINA L., SPIJKER M.N. (2005): Computing optimal monotonicity-preserving Runge-Kutta methods, submitted for publication, report Mathematical Institute, Leiden University, MI 2005-07.

FERRACINA L., SPIJKER M.N. (2005): Stepsize restrictions for total-variation-boundedness in general Runge-Kutta procedures, *Appl. Numer. Math.* **53**, 265–279.

Contents

Introduction

1 Monotonicity for Runge-Kutta methods

The growth in power and availability of digital computers during the last half century has led to an increasing use of sophisticated mathematical models in science, engineering and economics. *Systems of ordinary differential equations (ODEs)* frequently occur in such models as they naturally arise when modelling processes that evolve in time. A system of ODEs, for example, often models the time evolution of chemical or biological species. Many other interesting examples can be found in, e.g., Arrowsmith & Place (1982) and Strogatz (1994).

Usually, the state of the process is known at a particular (initial) moment whereas its evolution has to be determined. One then arrives at an *initial value problem (IVP)* for a system of ODEs.

In this thesis we consider *IVPs for systems of ODEs* that can be written in the form

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

Here u_0 is a given vector in a real vector space $\mathbb{V}$ and F stands for a given function from $\mathbb{V}$ into itself. The problem is then to find $U(t) \in \mathbb{V}$ for $t > 0$.

In most problems of this form that arise in practise, an analytical expression for the solution cannot be obtained whereas often precise data are desired. Therefore, it is common to seek approximate solutions of (1.1) by means of numerical methods.

There exists an extensive literature on numerical methods to approximate the solution of IVP (1.1), see, e.g., Butcher (2003), Hairer, Nørsett & Wanner (1993), Hairer & Wanner (1996). In this thesis we consider the important class of *Runge-Kutta methods*.

Runge-Kutta methods constitute a canonical class of so-called step-by-step methods. In these methods, each step starts from a given approximation u_{n-1} of $U(t)$ at a point $t = t_{n-1} \geq 0$. A stepsize $\Delta t > 0$ is selected and t_n is set equal to $t_{n-1} + \Delta t$. An approximation u_n of $U(t_n)$ is then computed from u_{n-1}. The result of this step, u_n, is then the starting value for the next step.

In particular, when a general Runge-Kutta method is applied to problem (1.1),

the approximations u_n of $U(t_n)$, can be defined in terms of u_{n-1} by the relations

$$\begin{cases} y_i = u_{n-1} + \Delta t \sum_{j=1}^{s} \kappa_{ij} F(y_j) \quad (1 \le i \le s+1), \\ u_n = y_{s+1}, \end{cases} \tag{1.2}$$

cf. e.g. Butcher (2003), Dekker & Verwer (1984), Hairer, Nørsett & Wanner (1993), Hundsdorfer & Verwer (2003).

Here κ_{ij} are real parameters, specifying the Runge-Kutta method, and y_i ($1 \le i \le s$) are intermediate approximations needed for computing $u_n = y_{s+1}$ from u_{n-1}. For the sake of simplicity, and to avoid unnecessary heavy notation later on, we define the $(s+1) \times s$ matrix K by $K = (\kappa_{ij})$. Then we can identify the Runge-Kutta method with the coefficient matrix K. If $\kappa_{ij} = 0$ (for $1 \le i \le j \le s$) then the intermediate approximations y_i can be computed directly from u_{n-1} and the already known y_j ($j < i$); otherwise a system of (nonlinear) equations has to be solved to obtain y_i. Accordingly, we call the Runge-Kutta method K *explicit* in the first case, *implicit* otherwise.

In the literature, much attention has been paid to solving (1.1) by processes (1.2) having a property which is called *monotonicity* (or *strong stability*). There are a number of closely related monotonicity concepts; see e.g. Hundsdorfer & Ruuth (2003), Hundsdorfer & Verwer (2003), Gottlieb, Shu & Tadmor (2001), Shu (2002), Shu & Osher (1988), Spiteri & Ruuth (2002). In this thesis we shall deal with a quite general monotonicity concept, and we shall study the problem of finding Runge-Kutta methods which have optimal properties regarding this kind of monotonicity.

We will deal with processes (1.2) which are monotonic in the sense that the vectors $u_n \in \mathbb{V}$ computed from $u_{n-1} \in \mathbb{V}$, via (1.2), satisfy

$$\|u_n\| \le \|u_{n-1}\| \tag{1.3}$$

– here we assume $\|.\|$ to be a seminorm on the real vector space $\mathbb{V}$ (i.e. $\|u+v\| \le \|u\| + \|v\|$ and $\|\lambda v\| \le |\lambda| \|v\|$ for all $\lambda \in \mathbb{R}$ and $u, v \in \mathbb{V}$).

Although there are other situations where (1.3) is a desirable property or a natural demand – see Harten (1983), Laney (1998), LeVeque (2002), Hundsdorfer & Ruuth (2003), Hundsdorfer & Verwer (2003) – Runge-Kutta methods with the property (1.3) have been designed specifically for solving IVPs, of form (1.1), coming from a (method of lines) semi-discretization of time dependent partial differential equations (PDEs), especially of conservation laws of the type

$$\frac{\partial}{\partial t} u(x,t) + \frac{\partial}{\partial x} \Phi(u(x,t)) = 0. \tag{1.4}$$

In order to illustrate why property (1.3) plays a key role when solving IVPs, of form (1.1), arising from the application of the method of lines (MOL) to the solution of time dependent PDEs, we shall elaborate, in the next section, a simple

example based on a test PDE of the form (1.4). We will start with briefly explaining the MOL. Then we will apply it when solving a Cauchy problem for the Burgers equation. With such an example, we hope to clarify the importance of property (1.3) in the context described above.

2 A numerical illustration

The application of the method of lines to a Cauchy problem for equation (1.4) consists of two steps.

> First a space-discretization (based, e.g., on finite-difference, finite-element or finite-volume methods) is applied to (1.4). This will yield an IVP of the form (1.1) with t as continuous variable - the so-called *semi-discrete* system. In this situation, the function F occurring in (1.1) depends on the given Φ as well as on the process of semi-discretization being used, and u_0 depends on the initial data of the original Cauchy problem.
>
> Secondly, a time-integration (e.g. a Runge-Kutta method or a multistep method) is applied to the so-obtained IVP (1.1) to derive a *fully-discrete* numerical process.

In order to clarify the approach described above, consider the Cauchy (Riemann) problem for the test scalar Burgers equation (of the form (1.4))

$$\frac{\partial}{\partial t}u(x,t) + \frac{\partial}{\partial x}\left(\frac{1}{2}u^2(x,t)\right) = 0 \qquad t \geq 0,\ -\infty < x < \infty, \tag{2.1.a}$$

$$u(x,0) = \begin{cases} 1 & \text{for } x < 0, \\ 0 & \text{for } x > 0. \end{cases} \tag{2.1.b}$$

The function

$$u(x,t) = \begin{cases} 1 & \text{for } x < t/2 \\ 0 & \text{for } x > t/2 \end{cases} \tag{2.2}$$

is the exact (weak) solution of problem (2.1).

Clearly, there is no need to seek an approximate solution to problem (2.1), but for illustration purpose only, we will apply the MOL. The solution of (2.1) will be approximated by combining a space-discretization, based on the finite-difference method, and a Runge-Kutta method as time integrator. Since the exact (weak) solution is known, it can be compared to the numerical approximation and it becomes easy to see whether the numerical solution is a reliable approximation or not.

Given the mesh-width $\Delta x = 1$, consider the point-grid in space $\mathbb{G} = \{x_j \,|\, x_j = j\Delta x,\ j = 0, \pm 1, \pm 2, ...\}$. The solution to (2.1) will be approximated at points (x_j, t) and we will denote by $U_j(t)$ these approximations. To that end, the quantity

$\frac{\partial}{\partial x}\left(\frac{1}{2}u^2(x_j,t)\right)$ is replaced by a (conservative) difference quotient $\frac{1}{\Delta x}\left[\frac{1}{2}\big(U_j(t)\big)^2 - \frac{1}{2}\big(U_{j-1}(t)\big)^2\right]$. Then we obtain the following *semi-discrete* system

$$\frac{d}{dt}U_j(t) = -\frac{1}{\Delta x}\left[\frac{1}{2}\big(U_j(t)\big)^2 - \frac{1}{2}\big(U_{j-1}(t)\big)^2\right].$$

Using the vector notation $U(t) = (...,U_{-1}(t),U_0(t),U_1(t),...) \in \mathbb{R}^\infty$, we arrive at the IVP (1.1), where $\mathbb{V} = \mathbb{R}^\infty$.

Since (1.1) now stands for a semi-discrete version of the conservation law (1.4), it is important that the *fully discrete* process (consisting of an application of (1.2) to (1.1)) is monotonic in the sense of (1.3) where $\|.\|$ denotes the *total-variation* seminorm

$$\|y\|_{TV} = \sum_{j=-\infty}^{+\infty} |\eta_j - \eta_{j-1}| \quad \text{(for } y \in \mathbb{R}^\infty \text{ with components } \eta_j\text{)}. \tag{2.3}$$

With this seminorm, the monotonicity property (1.3) reduces to the so-called *total-variation-diminishing* (TVD) property – see, e.g., Harten (1983), Laney (1998), Toro (1999), LeVeque (2002), and Hundsdorfer & Verwer (2003).

We will now see why guaranteeing monotonicity (TVD property) in the numerical approximation is important. To that end we solve (1.1) by applying two different explicit Runge-Kutta methods. The first method is defined by the relations

$$\begin{cases} y_1 = u_{n-1}, \\ y_2 = u_{n-1} + \Delta t F(y_1), \\ y_3 = u_{n-1} + \Delta t\Big(\frac{1}{2}F(y_1) + \frac{1}{2}F(y_2)\Big), \\ u_n = y_3, \end{cases} \tag{2.4}$$

and the second by

$$\begin{cases} y_1 = u_{n-1}, \\ y_2 = u_{n-1} - 20\Delta t F(y_1), \\ y_3 = u_{n-1} + \Delta t\Big(\frac{41}{40}F(y_1) - \frac{1}{40}F(y_2)\Big), \\ u_n = y_3 \end{cases} \tag{2.5}$$

– these two methods are taken from Gottlieb & Shu (1998).

It is easy to verify that the two methods coincide if F is linear. However, since the function F under consideration is not linear, we may and do observe different results when the two methods are applied. We use the same fixed time step $\Delta t = 0.75$ in both methods. In Figure 1, top and bottom, we show the results of the first and the second method, respectively, after 53 time steps – so that the

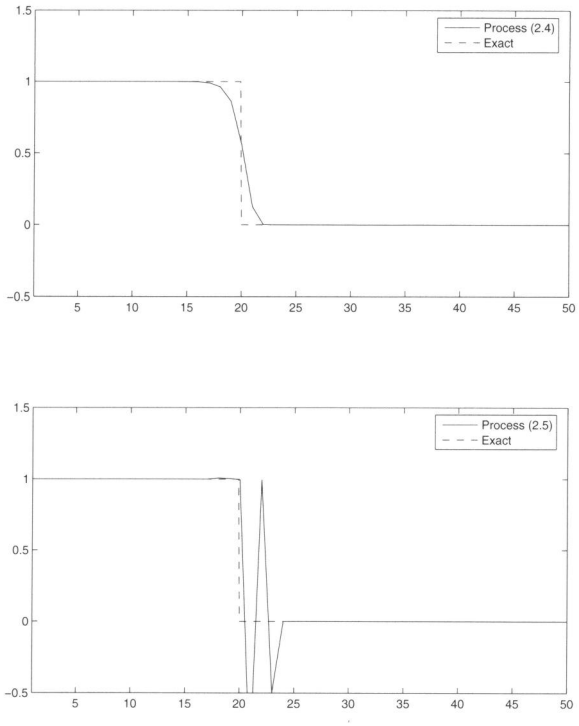

Figure 1: Top: solution with process (2.4). Bottom: solution with process (2.5).

profile of the true solution has moved over about 20 grid points. We clearly see that the second result is oscillatory while the first one is not. Clearly the solution on top approximates the true solution (2.2) well, while the solution on the bottom does not. This is strongly connected to the fact that the Runge-Kutta method (2.4) has property (1.3) (with $\|.\| = \|.\|_{TV}$) while method (2.5) does not.

We finally note that demanding the TVD property (monotonicity) from the numerical solution is a natural request. In fact, if we denote the restriction of the solution (2.2) on the point-grid $\mathbb{G}$ by $u(t) = (..., u(x_{-1}, t), u(x_0, t), u(x_1, t), ...)$, we clearly have

$$\|u(t_1)\|_{TV} \leq \|u(t_2)\|_{TV}$$

for every $t_1 \geq t_2$.

By the above, one is left with two questions: "How can we guarantee the Runge-Kutta method (1.2) to have property (1.3)?" Whether a given Runge-Kutta method (1.2) has property (1.3) or not depends (among other things) on the stepsize. Accordingly, the second question is "How should one select a good stepsize?" The answers to these questions are (essentially) the subject of this thesis.

3 Guaranteeing the monotonicity property: reviewing some literature

By Shu & Osher (1988) (see also Shu (1988)) a clever representation of explicit Runge-Kutta methods was introduced which facilitates the proof of property (1.3) in the situation where, for some $\tau_0 > 0$,

$$\|v + \tau_0 F(v)\| \leq \|v\| \quad \text{(for all } v \in \mathbb{V}\text{)}. \tag{3.1}$$

Clearly, in case (1.1) stands for a semi-discrete version of (1.4), then (3.1) can be interpreted as requiring that the semi-discretization has been performed in such a manner that the simple forward Euler method, applied to problem (1.1), is monotonic with stepsize τ_0.

In order to describe the representation introduced in Shu & Osher (1988), suppose an arbitrary explicit Runge-Kutta methods (1.2) is given with coefficient matrix $K = (\kappa_{ij})$.

We assume that λ_{ij} $(1 \leq j < i \leq s+1)$ are any real parameters with

$$\lambda_{i1} + \lambda_{i2} + ... + \lambda_{i,i-1} = 1 \quad (2 \leq i \leq s+1), \tag{3.2}$$

and we define corresponding coefficients μ_{ij} by

$$\mu_{ij} = \kappa_{ij} - \sum_{l=j+1}^{i-1} \lambda_{il}\kappa_{lj} \quad (1 \leq j < i \leq s+1) \tag{3.3}$$

(where the last sum should be interpreted as 0, when $j = i-1$).

Statement (i) of Theorem 3.1, to be given below, tells us that the relations (1.2) can be rewritten in the form

$$\begin{cases} y_1 = u_{n-1}, \\ y_i = \sum_{j=1}^{i-1} [\lambda_{ij}\, y_j + \Delta t \cdot \mu_{ij} F(y_j)] \quad (2 \leq i \leq s+1), \\ u_n = y_{s+1}. \end{cases} \tag{3.4}$$

We shall refer to (3.4) as a *Shu-Osher representation* of the explicit Runge-Kutta method (1.2).

Statement (ii) of Theorem 3.1 also specifies a stepsize restriction, of the form

$$0 < \Delta t \leq c \cdot \tau_0, \tag{3.5}$$

under which the monotonicity property (1.3) is valid, when u_n is computed from u_{n-1} according to (3.4). In the theorem, we shall consider the situation where

$$\lambda_{ij} \geq 0, \quad \mu_{ij} \geq 0 \quad (1 \leq j < i \leq s+1). \tag{3.6}$$

Furthermore under condition (3.6), we shall deal with a coefficient c defined by

$$c = \min\{\lambda_{ij}/\mu_{ij} : 1 \leq j < i \leq s+1\}, \tag{3.7}$$

where we use the convention $\lambda/\mu = \infty$ for $\lambda \geq 0$, $\mu = 0$.

Theorem 3.1 (Shu and Osher).
Let $K = (\kappa_{ij})$ specify an explicit Runge-Kutta method and assume λ_{ij}, μ_{ij} are as in (3.2), (3.3). *Then the following conclusions* (i) *and* (ii) *are valid.*

(i) *The Runge-Kutta relations* (1.2) *are equivalent to* (3.4).

(ii) *Assume additionally* (3.6) *holds, and that the coefficient c is defined by* (3.7). *Let F be a function from $\mathbb{V}$ to $\mathbb{V}$, satisfying* (3.1). *Then, under the stepsize restriction* (3.5), *process* (3.4) *is monotonic; i.e.* (1.3) *holds whenever u_n is computed from u_{n-1} according to* (3.4).

The above theorem is essentially due to Shu & Osher (1988). The proof of the above statement (i) is straightforward. Furthermore, the proof of (ii) relies on noting that, for $2 \leq i \leq s+1$, the vector y_i in (3.4) 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 (essentially) (3.1) (with $v = y_j$).

4 The limitation of the approach in the literature

4.1 Stepsize restrictions guaranteeing monotonicity for general Runge-Kutta methods

It is evident that a combination of Statements (i) and (ii), of Theorem 3.1, immediately leads to *a conclusion which is highly relevant to the original Runge-Kutta method K*. We emphasize such a result in the following corollary.

Corollary 4.1.
Let $K = (\kappa_{ij})$ specify an explicit Runge-Kutta method and assume λ_{ij}, μ_{ij} are as in (3.2), (3.3) (3.6). *Let c be defined by* (3.7). *Then the conditions* (3.1), (3.5) *guarantee the monotonicity property* (1.3) *for u_n computed from u_{n-1} by* (1.2).

Clearly, it would be awkward if the factor c, defined in (3.7), were zero, or positive and so small that (3.5) reduces to a stepsize restriction which is too severe for any practical purposes – in fact, the less restrictions on Δt, the better. One might thus be tempted to take the magnitude of c into account when comparing the effectiveness of different Runge-Kutta methods K. However, it is evident that such a use of the coefficient c defined by (3.7), could be quite misleading if, for a given Runge-Kutta (1.2), the conclusion in Corollary 4.1 were also valid with some factor c which is (much) larger than the c defined by (3.7).

For any given Runge-Kutta method K, the question thus arises what is the largest factor $\mathscr{C}(K)$, *not necessarily defined via* (3.7), such that the conclusion in Corollary 4.1 is still valid.

4.2 Optimal Shu-Osher representations

Consider once more Corollary 4.1. It is important to note that the coefficient c, given by (3.7), not only depends on the underlying Runge-Kutta method $K = (\kappa_{ij})$, but also on the parameters λ_{ij} actually chosen – the coefficients μ_{ij} are fixed by (3.3). Denoting by L the $(s+1)\times s$ matrix defined by $L = (\lambda_{ij})$, $1 \leq j < i \leq s+1$ and 0 otherwise, we then indicate with $c(K, L)$ the coefficient c define by (3.7).

Suppose $\tilde{L} = (\tilde{\lambda}_{ij})$ are parameters which are best possible, in the sense that the corresponding coefficient $c(K, \tilde{L})$, obtained via (3.7), satisfies $c(K, \tilde{L}) \geq c(K, L)$, for any other Shu-Osher representation of the given method K in question. Then $c(K, \tilde{L})$ depends only on the coefficient scheme K so that we can write $c(K, \tilde{L}) = c(K)$. Then, a second question is: how can we determine (in a transparent and simple way) parameters $\tilde{L} = (\tilde{\lambda}_{ij})$ leading to the coefficient $c(K)$?

A third natural question, related to Section 4.1, then arises: can $\mathscr{C}(K)$ be larger than $c(K)$?

A fourth question is of whether the Shu-Osher Theorem 3.1 can be generalized so as to become also relevant to Runge-Kutta methods which are *not necessarily explicit.*

4.3 Computing optimal monotonic Runge-Kutta methods

In the following we denote by $E_{s,p}$ the class of all explicit s-stage Runge-Kutta methods with (classical) order of accuracy at least p.

The questions formulated in the previous two sections are strongly related to the problem of determining a method K, belonging to $E_{s,p}$ which is optimal with regard to the size of its coefficient $\mathscr{C}(K)$. In spite of the (possible) limitations of the coefficient $c(K)$ for guaranteeing monotonicity of Runge-Kutta methods K, much attention has been paid in the literature to optimizing $c(K)$ – usually with a terminology and notation somewhat different from the above - see e.g. Gerisch & Weiner (2003), Gottlieb & Shu (1998), Ruuth & Spiteri (2002), Shu (2002), Shu & Osher (1988), Spiteri & Ruuth (2002).

In fact, for various values of s and p, optimal methods K, w.r.t. $c(K)$, were determined within the class of $E_{s,p}$ – see, e.g., Shu & Osher (1988), Gottlieb & Shu (1998), Ruuth & Spiteri (2004), Spiteri & Ruuth (2003), Ruuth (2004).

For given s and p, the numerical searches carried out in the last three papers, are essentially based on the following optimization problem (4.1), in which λ_{ij}, μ_{ij}, γ are the independent variables and $f(\lambda_{ij}, \mu_{ij}, \gamma) = \gamma$ is the objective function.

(4.1) Maximize γ, subject to the following constraints:

$$\lambda_{ij} - \gamma\, \mu_{ij} \geq 0 \quad (1 \leq j < i \leq s+1);$$

λ_{ij}, μ_{ij} satisfy (3.2), (3.3) (3.6)

the coefficients κ_{ij}, satisfying (3.3), specify a Runge-Kutta method (1.2) belonging to class $E_{s,p}$.

Clearly, the variable γ in (4.1) corresponds to c in (3.7), and parameters λ_{ij}, μ_{ij}, γ solving the optimization problem (4.1) yield a Shu-Osher process in $E_{s,p}$ which is optimal in the sense of c, (3.7).

It should be evident how the answers to the questions mentioned in the previous two sections could strongly influence the relevance of Ruuth & Spiteri's approach (4.1). In particular, it would be of great interest to know whether their approach can be improved and/or generalized so as to guarantee optimality w.r.t. $\mathscr{C}(K)$. Moreover, it would be of much interest if optimizations, with regard to $\mathscr{C}(K)$, could also be carried out within classes of methods K which are not necessarily explicit.

4.4 Boundedness for general Runge-Kutta methods

In the Shu-Osher Theorem 3.1 (and Corollary 4.1), conditions on the stepsize were established which guarantee monotonicity property (1.3). These conditions were derived under the assumption that the simple Euler method, applied to problem (1.1), is monotonic, for the stepsize τ_0 – i.e., (3.1) holds.

However, important semi-discrete versions (1.1) of (1.4), cannot be modelled suitably via condition (3.1), see, e.g., Shu (1987), Cockburn & Shu (1989). Clearly, in such cases the stepsize restrictions which are relevant to the situation (3.1), do not allow us to conclude any longer that a Runge-Kutta procedure is monotonic.

Although for these semi-discretizations condition (3.1) does not apply, the following weaker condition provides an appropriate model:

$$\|v + \tau_0 F(v)\| \leq (1 + \alpha_0 \tau_0)\|v\| + \beta_0 \tau_0 \quad (\text{for all } v \in \mathbb{V}). \tag{4.2}$$

Here τ_0 is again positive, and α_0, β_0 are nonnegative constants. Condition (4.2) can be interpreted, analogously to (3.1), as a bound on the increase of the seminorm, when the explicit Euler time stepping is applied to (1.1) with time step τ_0.

In the situation where property (4.2) is present, it is natural to look for an analogous property in the general Runge-Kutta process (1.2), namely

$$\|u_n\| \leq (1 + \alpha \Delta t)\|u_{n-1}\| + \beta \Delta t. \tag{4.3}$$

Here α, β denote nonnegative constants.

Suppose (4.3) would hold under a stepsize restriction of the form $0 < \Delta t \leq \Delta t_0$. By applying (4.3) recursively and noting that $(1 + \alpha \Delta t)^n \leq \exp(\alpha\, n \Delta t)$, we then obtain

$$\|u_n\| \leq e^{\alpha T}\|u_0\| + \frac{\beta}{\alpha}(e^{\alpha T} - 1) \quad (\text{for } 0 < \Delta t \leq \Delta t_0 \text{ and } 0 < n\Delta t \leq T)$$

– here $\frac{\beta}{\alpha}(e^{\alpha T} - 1)$ stands for βT, in the special case where $\alpha = 0$. Hence, property (4.3) (for $0 < \Delta t \leq \Delta t_0$) amounts to *boundedness*, in that

$$\|u_n\| \leq B \quad (\text{for } 0 < \Delta t \leq \Delta t_0 \text{ and } 0 < n\Delta t \leq T)$$

with $B = e^{\alpha T}\|u_0\| + \frac{\beta}{\alpha}(e^{\alpha T} - 1)$.

Since (4.2) and (4.3) reduce to (3.1) and (1.3), respectively, when $\alpha_0 = \beta_0 = \alpha = \beta = 0$, it is natural to look for extensions, to the boundedness context, of the results in the literature pertinent to the monotonicity property. More specifically, the natural question arises of whether stepsize restrictions of the form (3.5) can be established which guarantee property (4.3) when condition (4.2) is fulfilled.

5 Scope of this thesis

In this thesis we propose a theory by means of which, among other things, the open questions posed in Section 4 can be settled.

Chapter I is essentially addressed to the question raised in Section 4.1. First we review the crucial quantity $R(K)$ introduced by Kraaijevanger (1991). Then we solve the question by proving that the factor $\mathscr{C}(K)$ equals $R(K)$ – such a conclusion is given for arbitrary Runge-Kutta methods, either explicit or not. The contents of this chapter are equal to Ferracina L., Spijker M.N. (2004): Stepsize restrictions for the total-variation-diminishing property in general Runge-Kutta methods, *SIAM J. Numer. Anal.* **42**, 1073–1093.

In Chapter II we answer the questions of Sections 4.2. We give generalizations of the Shu-Osher representation (3.4) and of the Shu-Osher Theorem 3.1; our generalizations are relevant to arbitrary Runge-Kutta methods K – either explicit or not. With the help of such generalizations we are able to give, in a simple way, special parameters $\tilde{L} = (\tilde{\lambda}_{ij})$ leading to the coefficient $c(K)$. Moreover, we prove that $\mathscr{C}(K)$ is never larger than $c(K, \tilde{L}) = c(K)$. The contents of this chapter are equal to Ferracina L., Spijker M.N. (2005): An extension and analysis of the Shu-Osher representation of Runge-Kutta methods, *Math. Comp.* **249**, 201–219.

In Chapter III we solve the questions of Section 4.3. We continue the analysis of Shu-Osher representations so as to arrive naturally at a generalization and improved version of Ruuth & Spiteri's approach (4.1). Our procedure guarantees optimality with respect to $\mathscr{C}(K)$. Moreover it is, unlike (4.1), also relevant to Runge-Kutta methods which are implicit. The contents of this chapter are equal to Ferracina L., Spijker M.N. (2005): Computing optimal monotonicity-preserving Runge-Kutta methods, submitted for publication, report Mathematical Institute MI 2005-07.

In Chapter IV we settle the question raised at the end of Section 4.4. We propose a general theory yielding stepsize restrictions which cover a larger class of semidiscrete approximations than covered so far in the literature. In particular our theory gives stepsize restrictions, of the form (3.5), which guarantee, for general Runge-Kutta methods (1.2), property (4.3) when condition (4.2) is fulfilled.

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.

For a more detailed introduction to the topics of this thesis, and for related literature, we refer to the beginning of each chapter.

Bibliography

[1] ARROWSMITH D. K., PLACE C. M. (1982): *Ordinary differential equations. A qualitative approach with applications.* Chapman & Hall (London).

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

[3] 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.

[4] 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).

[5] GERISCH A., WEINER R. (2003): The positivity of low-order explicit Runge-Kutta schemes applied in splitting methods. *Comput. Math. Appl.*, 45 No. 1-3, 53–67. Numerical methods in physics, chemistry, and engineering.

[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. (1983): High resolution schemes for hyperbolic conservation laws. *J. Comput. Phys.*, 49 No. 3, 357–393.

[11] HUNDSDORFER W., RUUTH S. J. (2003): Monotonicity for time discretizations. Procs. Dundee Conference 2003. Eds. D.F. Griffiths, G.A. Watson, Report NA/217, Univ. of Dundee.

[12] 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).

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

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

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

[16] Ruuth S. J. (2004): Global optimization of explicit strong-stability-preserving Runge-Kutta methods. Tech. rep., Department of Mathematics Simon Fraser University.

[17] Ruuth S. J., Spiteri R. J. (2002): Two barriers on strong-stability-preserving time discretization methods. *J. Sci. Comput.*, 17 No. 1-4, 211–220.

[18] Ruuth S. J., Spiteri R. J. (2004): High-order strong-stability-preserving runge–kutta methods with downwind-biased spatial discretizations. *SIAM J. Numer. Anal.*, 42 No. 3, 974–996.

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

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

[21] 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).

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

[23] 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).

[24] Spiteri R. J., Ruuth S. J. (2003): Non-linear evolution using optimal fourth-order strong-stability-preserving Runge-Kutta methods. *Math. Comput. Simulation*, 62 No. 1-2, 125–135.

[25] Strogatz S. H. (1994): *Nonlinear dynamics and chaos : with applications in physics, biology, chemistry, and engineering*. Perseus Books (Reading, MA).

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

Chapter I

Stepsize restrictions for the total-variation-dimishing property in general Runge-Kutta methods

The contents of this chapter are equal to: Ferracina L., Spijker M.N. (2004): Stepsize restrictions for the total-variation-diminishing property in general Runge-Kutta methods, *SIAM J. Numer. Anal.* **42**, 1073–1093.

Abstract

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

In this paper we propose a theory by means of which we are able to clarify the above questions. Moreover, by applying our theory, we settle analogous questions regarding the related strong-stability-preserving (SSP) property (see e.g. Gottlieb, Shu & Tadmor (2001), Shu (2002)). Our theory can be viewed as a variant to a theory of Kraaijevanger (1991) on contractivity of Runge-Kutta methods.

1 Introduction

1.1 The purpose of the paper

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

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

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

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

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

The solution $U(t)$ to (1.1) stands for a (time dependent) vector in $\mathbb{R}^\infty = \{y : y = (..., \eta_{-1}, \eta_0, \eta_1, ...)$ with $\eta_j \in \mathbb{R}$ for $j = 0, \pm 1, \pm 2, ...\}$. The components $U_j(t)$ of $U(t)$ are to approximate the desired true solution values $u(j\Delta x, t)$ (or cell averages thereof); here Δx denotes a (positive) mesh-width. Furthermore, F stands for a function from $\mathbb{R}^\infty$ into $\mathbb{R}^\infty$; it depends on the given function f as well as on the process of semidiscretization being used. Finally, $u_0 \in \mathbb{R}^\infty$ depends on the initial data of the original Cauchy problem.

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

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

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

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

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

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

explicit m-stage Runge-Kutta methods, written in the special form

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

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

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

Furthermore, in the theorem it is assumed that

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

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

and the following notation is used:

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

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

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

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

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

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

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

then (1.7) still implies that

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

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

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

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

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

1.2 Outline of the rest of the paper

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

In Section 3 we apply Theorem 2.5 to a generalized version of process (1.3). After the introductory Section 3.1, we clarify in the Sections 3.2 and 3.3, respectively, the questions raised at the end of Section 1.1 regarding the TVD and SSP

properties of process (1.3). Section 3.4 gives two examples illustrating the superiority of the quantity $R(A,b)$ (to the factor c, given by (1.6)) as a guide to stepsize restrictions for the TVD and SSP properties.

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

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

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

2 A general theory for monotonic Runge-Kutta processes

2.1 Stepsize-coefficients for monotonicity in a general context

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

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

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

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

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

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

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

In these formulas, $\Delta t > 0$ denotes the stepsize and b_j, a_{ij} are real parameters, specifying the Runge-Kutta method. We always assume that $b_1 + b_2 + ... + b_m = 1$. If $a_{ij} = 0$ (for $j \geq i$), the Runge-Kutta method is called *explicit*. Defining the $m \times m$ matrix A by $A = (a_{ij})$ and the column vector $b \in \mathbb{R}^m$ by $b = (b_1, b_2, b_3, ..., b_m)^T$, we can identify the Runge-Kutta method with the *coefficient scheme* (A, b).

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

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

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

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

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

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

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

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

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

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

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

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

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

2.2 Irreducible Runge-Kutta schemes and the quantity $R(A, b)$

In this subsection we give some definitions which will be needed when we formulate our results, in Subsection 2.3, about maximal stepsize-coefficients c. We start with the fundamental concepts of reducibility and irreducibility.

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

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

(ii) *there exist nonempty, pairewise 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 (2.2) with $j \in N$ have no influence on u_n, and the Runge-Kutta method is equivalent to a method with less than m stages. Also in case of (ii), the Runge-Kutta method essentially reduces to a method with less then m stages; see, e.g., Dekker & 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 characteristic quantity for Runge-Kutta schemes introduced by Kraaijevanger (1991). Following this author, we shall denote his quantity by $R(A, b)$, and in defining it, we shall use, for real ξ, the notations:

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

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

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

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

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

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

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

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

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

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

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

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

For ξ close to zero, we have

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

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

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

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

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

2.3 Formulation of our main theorem

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

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

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

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

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

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

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

Theorem 2.5 (Relating monotonicity to $R(A, b)$).
Consider an arbitrary irreducible Runge-Kutta scheme (A, b). Let c be a given value with $0 < c \leq \infty$. Choose one of the three (semi)norms $\|.\|_\infty$, $\|.\|_1$, or $\|.\|_{TV}$, and denote it by $|\!|.|\!|$. Then each of the following statements (2.9), (2.10) *and* (2.11) *is equivalent to* (2.8).

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

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

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

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

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

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

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

3.1 The equivalence of (a generalized version of) process (1.3) to method (2.2)

In this Section 3 we study time stepping processes producing numerical approximations $u_n \in \mathbb{R}^\infty$ to $U(n\Delta t)$ (for $n \geq 1$), where $U(t) \in \mathbb{R}^\infty$ satisfies (1.1). We focus on processes of the form

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

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

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

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

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

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

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

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

We assume that

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

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

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

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

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

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

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

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

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

3.2 The total-variation-diminishing property of process (3.1)

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

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

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

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

and with stepsize restrictions of the form

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

(see Definition 2.3).

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

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

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

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

Remark 3.4. The above theorem, when applied to any process (1.3) satisfying (1.5.a), (1.5.b), gives a *stronger conclusion than Theorem 1.1.* By Theorem 2.5, property (2.10) with $|\!|.|\!| = \|.\|_{TV}$ implies inequality (2.8). Therefore the coefficient c, given by Theorem 1.1, satisfies $c \leq R(A, b)$; this means that the stepsize restriction (3.6) of Theorem 3.2 is, in general, less severe than the restriction (1.7) of Theorem 1.1 – see Example 3.8 in Subsection 3.4 for an illustration. ◇

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

3.3 The strong-stability-preserving property of process (3.1)

Let $\mathbb{V}$ be an arbitrary linear subspace of $\mathbb{R}^\infty$, and let $\|.\|$ denote any seminorm on $\mathbb{V}$. For functions $F : \mathbb{V} \longrightarrow \mathbb{V}$ satisfying

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

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

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

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

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

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

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

3.4 Illustrations to the Theorems 3.2 and 3.6

We give two examples illustrating the Theorems 3.2 and 3.6.

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

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

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

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

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

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

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

For the corresponding Runge-Kutta scheme (A, b), defined by (3.3), we have $a_{ij} = 0 \quad (j \geq i)$, $a_{21} = 1/2$ and $b_1 = b_2 = 1/2$. Clearly, (2.7) is fulfilled, guaranteeing the SSP property (see Theorem 3.6). Moreover, according to Kraaijevanger (1991, Theorem 9.2), we have $R(A, b) = 2$. Therefore, by Theorem 3.2, the numerical process is TVD, under assumption (3.5), if $0 < \Delta t \leq 2 \cdot \tau_0$. We note that

the same method was presented by Spiteri & Ruuth (2002); we shall come back to it in Section 4.2 (Remark 4.4; $m = 2,\ p = 1$). ◇

4 Optimal Runge-Kutta methods

4.1 Preliminaries

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

In Section 4.2 we shall briefly review some of Kraaijevanger's conclusions so as to arrive at extensions and completions of the material, referred to above, on optimality in the sense of c, (1.6). Furthermore, we shall consider scaled stepsize-coefficients which reflect the efficiency of the methods better than the unscaled coefficients; in Table I.1 we shall display optimal scaled stepsize-coefficients. Next, in Section 4.3, we shall focus on an algorithm for computing $R(A, b)$; the authors feel that it can be useful in (future) calculations for determining, numerically, optimal Runge-Kutta methods. Finally, in Section 4.4 we touch upon a few important related issues.

4.2 Optimal methods in the class $E_{m,p}$

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

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

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

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

Theorem 4.3 (Kraaijevanger (1991; pp. 518-520)).

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

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

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

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

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

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

Kraaijevanger (1991, pp. 522-523) constructed numerically an explicit 5-stage method (A, b) of order 4, with $R(A, b) \approx 1.508$. It is interesting to note that the same method was found by Spiteri & Ruuth (2002) in a numerical search within

the class of methods (1.3) satisfying (1.5). By a similar search, the last authors also found a 5-stage method of order 3 with $c \approx 2.651$ (given by (1.6)). In view of Kraaijevanger (1986, Theorem 5.3), Spijker (1983) and our Theorem 2.5, we can conclude that this method has a value $R(A, b) \approx 2.651$, and is optimal in a *stronger sense* and over a *larger class* of methods than follows from Spiteri & Ruuth (2002).

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

In Table I.1 we display scaled stepsize-coefficients of Runge-Kutta methods (A, b), which were reviewed above and are optimal in $E_{m,p}$ with respect to $R(A, b)$.

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

Table I.1: Scaled stepsize-coefficients $R(A, b)/m$ for optimal Runge-Kutta methods in $E_{m,p}$.

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

4.3 An algorithm for computing $R(A, b)$, for methods of class $E_{m,p}$

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

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

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

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

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

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

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

4.4 Final remarks

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

Above, in Section 4, we dealt exclusively with explicit Runge-Kutta schemes. However, in Kraaijevanger (1991) also (a few) results were obtained, regarding the size of $R(A, b)$, relevant to implicit schemes – see below. A combination of these results with our Theorem 2.5 immediately leads to interesting conclusions about stepsize-coefficients for monotonicity.

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

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

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

5 Kraaijevanger's theory and our proof of Theorem 2.5

5.1 A theorem of Kraaijevanger on contractivity

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

5.2 The proof of Theorem 2.5

5.2.1 Preliminaries

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

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

The second implication is trivial, whereas the third and fourth implication will be proved in Section 5.2.3. The proofs, in this section, are *not* related to arguments used in Kraaijevanger (1991), but are based on Lemma 5.6. This lemma gives a general framework in which the property of c being a stepsize-coefficient for monotonicity can be carried over from a space $\mathbb{Y}$ with seminorm $\|.\|_{\mathbb{Y}}$ to another space $\mathbb{X}$ with seminorm $\|.\|_{\mathbb{X}}$.

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

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

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

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

5.2.2 Statement (2.8) $\Longrightarrow$ statement (2.9)

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

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

Furthermore, for any (real) $l \times m$ matrix $B = (b_{ij})$, we define a corresponding linear operator $B_{\mathbb{V}}$, from $\mathbb{V}^m$ to $\mathbb{V}^l$, by $B_{\mathbb{V}}(v) = w$, for $v = [v_j] \in \mathbb{V}^m$, where $w = [w_i] \in \mathbb{V}^l$ with $w_i = \sum_{j=1}^m b_{ij} v_j$ $(1 \leq i \leq l)$. 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 multiplications occurring in the last three left-hand members stand for the usual algebraic operations for matrices, whereas the addition and multiplications in the right-hand members apply to linear operators. The last three equalities will underlie part of our subsequent calculations.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

This completes the proof of (2.9).

5.2.3 Statement (2.10) $\Longrightarrow$ statement (2.11); and statement (2.11) with $|.| = \|.\|_{TV}$ $\Longrightarrow$ statement (2.11) with $|.| = \|.\|_1$

We start this subsection by giving Lemma 5.6. The lemma deals with a general situation where

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

5.2.4 (2.11) with $|.| = \|.\|_1$ or $\|.\|_\infty \Longrightarrow$ (2.8)

Throughout this subsection we shall use, for $p = 1, \infty$ and $s \times s$ matrices G, the notation $\|G\|_p = \max \|Gv\|_p / \|v\|_p$, where the maximum is over all nonzero vectors v in $\mathbb{R}^s$. Furthermore, we shall denote the $s \times s$ identity matrix by I.

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

Choose $\Delta t = c\, \tau_0$ and write $Z_i = \Delta t G_i$. Then the above linear variants of (2.2)

and (2.4), respectively can be written in the form

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

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

and

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

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

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

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

The inequality

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

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

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

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

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

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

Proof. The proof will be given in three steps.
Step 1. Let

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

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

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

We shall prove that

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Acknowledgement

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

Bibliography

[1] Burrage K., Butcher J. C. (1980): Nonlinear stability of a general class of differential equation methods. *BIT*, 20 No. 2, 185–203.

[2] Butcher J. C. (1987): *The numerical analysis of ordinary differential equations. Runge Kutta and general linear methods.* A Wiley-Interscience Publication. John Wiley & Sons Ltd. (Chichester).

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

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

[6] 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.

[7] Harten A. (1983): High resolution schemes for hyperbolic conservation laws. *J. Comput. Phys.*, 49 No. 3, 357–393.

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

[9] 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).

[10] Jeltsch R., Nevanlinna O. (1981): Stability of explicit time discretizations for solving initial value problems. *Numer. Math.*, 37 No. 1, 61–91.

[11] Kennedy C. A., Carpenter M. H., Lewis R. M. (2000): Low-storage, explicit Runge-Kutta schemes for the compressible Navier-Stokes equations. *Appl. Numer. Math.*, 35 No. 3, 177–219.

[12] Kraaijevanger J. F. B. M. (1986): Absolute monotonicity of polynomials occurring in the numerical solution of initial value problems. *Numer. Math.*, 48 No. 3, 303–322.

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

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

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

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

[17] Morton K. W. (1980): Stability of finite difference approximations to a diffusion-convection equation. *Internat. J. Numer. Methods Engrg.*, 15 No. 5, 677–683.

[18] Ruuth S. J., Spiteri R. J. (2002): Two barriers on strong-stability-preserving time discretization methods. *J. Sci. Comput.*, 17 No. 1-4, 211–220.

[19] Schönbeck S. O. (1967): On the extension of lipschitz maps. *Ark. Mat.*, 7 201–209.

[20] 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).

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

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

[23] 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).

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

Chapter II

An extension and analysis of the Shu-Osher representation of Runge-Kutta methods

The contents of this chapter are equal to: Ferracina L., Spijker M.N. (2005): An extension and analysis of the Shu-Osher representation of Runge-Kutta methods, *Math. Comp.* **249**, 201–219.

Abstract

In the context of solving nonlinear partial differential equations, Shu & Osher (1988) 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, see e.g. Gerisch & Weiner (2003), Gottlieb & Shu (1998), Gottlieb, Shu & Tadmor (2001), Ruuth & Spiteri (2002), Shu (2002), Spiteri & Ruuth (2002).

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 $n = 1, 2, 3, ...$; cf. e.g. Butcher (1987), Dekker & Verwer (1984), Hairer, Nørsett & Wanner (1993), Hairer & Wanner (1996). 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 scalar conservation law of the form

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

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, see e.g. Laney (1998), LeVeque (2002), Toro (1999). The actual function values $F(y)$ depend on the given f as well as on the MOL semi-discretization being used. In the literature – see e.g. Gottlieb, Shu & Tadmor (2001), Shu (2002), Shu

& Osher (1988), Spiteri & Ruuth (2002) – much attention has been paid to solving the semi-discrete problem (1.1) by Runge-Kutta processes (1.2) which are *total-variation-diminishing (TVD)* in the sense that

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

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 (1983), Kröner (1997), Laney (1998), LeVeque (2002), Toro (1999).

By Shu & Osher (1988) (see also Shu (1988)) a clever representation of explicit Runge-Kutta methods was introduced which facilitates the proof of property (1.4) in the situation where, for some $\tau_0 > 0$,

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

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 & Osher (1988), we consider an arbitrary explicit coefficient scheme (A, b). We assume that λ_{ij} (for $2 \leq i \leq m+1$ and $1 \leq j \leq i-1$) are any real parameters with

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

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

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

Further, we shall deal with a coefficient c defined by

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

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 & Osher (1988). The proof of the above statement (i) is straightforward. Further, 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) would be 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 would exist 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 a terminology and notation somewhat different from the above – see e.g. Gerisch & Weiner (2003), Gottlieb & Shu (1998), Ruuth & Spiteri (2002), Shu (2002), Shu & Osher (1988), Spiteri & Ruuth (2002). 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 become also relevant to Runge-Kutta methods which are *not explicit*. Partial results related to this question, but no complete answers, were obtained by Gottlieb, Shu & Tadmor (2001, Section 6.2) and Hundsdorfer & Verwer (2003).

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 & Tadmor (2001), Shu & Osher (1988) – 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 notations 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 to 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 (1991). 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 & Spijker (2004) and Higueras (2004) 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 the Sections 2 and 3.

In Section 4.1 we shall apply the Theorems 2.2 and 3.4 to general Runge-Kutta methods so as to arrive at the 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 & Spijker (2004).

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 (1991).

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:

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

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

Here λ_{ij} and μ_{ij} are real coefficients specifying the numerical process (2.1), and Δt denotes again a positive stepsize. Further, 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

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

and

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

Clearly, if the above parameters λ_{ij}, μ_{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 $\|\lambda v + (1-\lambda)w\| \leq \lambda\|v\| + (1-\lambda)\|w\|$ for all $v, w \in \mathbb{V}$ and $0 \leq \lambda \leq 1$.

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

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

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

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 & Verwer (1984), Hundsdorfer & Verwer (2003), LeVeque (2002). 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 (1987, p.392), Dekker & Verwer (1984, p.263), Gottlieb, Shu & Tadmor (2001), Hundsdorfer, Ruuth & Spiteri (2003), Morton (1980).

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

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

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

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

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 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),

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

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

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

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

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

and the values λ_{ij}, μ_{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 the 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 & Tadmor (2001, 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 $\|v\|=(1+\tau/\tau_0)\|v\|-(\tau/\tau_0)\|v\|\le(1+\tau/\tau_0)\|v\|-(\tau/\tau_0)\|v+\tau_0F(v)\|\le\|(1+\tau/\tau_0)v-(\tau/\tau_0)(v+\tau_0F(v))\|=\|v-\tau F(v)\|$; 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 notations which will be used below.

For any vectors $v_1,v_2,...,v_m$ 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 $1 \leq i \leq l$). 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 multiplications, occurring in the last three left-hand members, stand for the usual algebraic operations for matrices, whereas the addition and multiplications in the right-hand members apply to linear operators.

For clarity, we will also use the following simplified notations: $\mathbf{b}^T = (b^T)_{\mathbb{V}}$, $\mathbf{A} = A_{\mathbb{V}}$, $\mathbf{M}_0 = (M_0)_{\mathbb{V}}$, $\mathbf{M}_1 = (M_1)_{\mathbb{V}}$, $\mathbf{L}_0 = (L_0)_{\mathbb{V}}$ and $\mathbf{L}_1 = (L_1)_{\mathbb{V}}$. Further, 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

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

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

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

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,

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

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 these notations we obtain from (2.1.a), by using the convexity of the function $\|.\|$,

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

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

$$(2.11) \qquad \|y_i - \alpha_i x_i\| \geq \|y_i\|.$$

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

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

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

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

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

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 $0 \leq t \leq 1$) exists and depends continuously on t. For $0 \leq t < 1$ we have $K(t) = I + tL_0 + (tL_0)^2 + \ldots$ 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

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

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

$$\|u_n\| \leq \theta\|u_{n-1}\| + \sum_{j=1}^{m} \lambda_{m+1,j}\,\|y_j\| \leq (\theta + \sum_{j=1}^{m} \lambda_{m+1,j})\|u_{n-1}\| = \|u_{n-1}\|. \quad \blacksquare$$

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

3.1 Irreducible Runge-Kutta schemes and the quantity $R(A, b)$

In this subsection we give some definitions which will be needed when we formulate our results, in Subsection 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 $1 \leq r < m$ and $M_1 \cup M_2 \cup \ldots \cup M_r = \{1, 2, \ldots, m\}$, such that $\sum_{k \in M_q} a_{ik} = \sum_{k \in M_q} a_{jk}$ whenever $1 \leq p \leq r$, $1 \leq q \leq r$ and $i, j \in M_p$.*

In case the above statement (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 then m stages, see e.g. Dekker & Verwer (1984) or Hairer & Wanner (1996). 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 (1991). Following this author, we shall denote his quantity by $R(A, b)$, and in defining it, we shall use, for real ξ, the notations:

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

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 $\xi \leq 0$ for which

(3.1) $I - \xi A$ is invertible, $A(\xi) \geq 0$, $b(\xi) \geq 0$, $e(\xi) \geq 0$, and $\varphi(\xi) \geq 0$.

The first inequality in (3.1) 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 3.2 (The quantity $R(A, b)$).
Let (A, b) be a given coefficient scheme. In case $A \geq 0$ and $b \geq 0$, we define

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

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

Definition 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 & Spijker (2004; Section 4.3), Kraaijevanger (1991, 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

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

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

(i) *$R(A, b) > 0$ if and only if : $A \geq 0$, $b > 0$ and $Inc(A^2) \leq Inc(A)$.*

(*ii*) *Let* $0 < r < \infty$. *Then* $R(A,b) \geq r$ *if and only if :* $A \geq 0$ *and the 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,*
- $A \geq 0$ *and* $A^{-1}e \geq 0$,
- $b^T A^{-1} \geq 0$ *and* $b^T A^{-1} e \leq 1$.

The Parts $(i), (ii), (iii)$ of the above theorem have been taken almost literally from Kraaijevanger (1991; 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 gives also interesting properties of this maximal value.

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

$$\text{(3.2.a)} \qquad L^* = \begin{pmatrix} L_0^* \\ L_1^* \end{pmatrix}, \; L_0^* = \begin{pmatrix} \lambda_{11}^* & \dots & \lambda_{1m}^* \\ \vdots & & \vdots \\ \lambda_{m1}^* & \dots & \lambda_{mm}^* \end{pmatrix}, \; L_1^* = (\lambda_{m+1,1}^*, ..., \lambda_{m+1,m}^*),$$

with

$$\text{(3.2.b)} \quad L_0^* = \gamma A(I+\gamma A)^{-1}, \;\; L_1^* = \gamma b^T (I+\gamma A)^{-1}, \;\; \gamma = R(A,b)$$
$$(\text{ if } \;\; 0 \leq R(A,b) < \infty),$$

$$\text{(3.2.c)} \quad L_0^* = I - \gamma P, \;\; L_1^* = b^T P, \;\; \gamma = (\max_i p_{ii})^{-1}, \text{ where } P = (p_{ij}) = A^{-1}$$
$$(\text{ if } \;\; R(A,b) = \infty).$$

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 (1991; 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). *Further, 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 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) $c(A,b,L^*) = R(A,b)$ *(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 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 $0 \le R(A,b) < \infty$.
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 $L_0^* \ge 0$ and $(I - L_0^*)e = (I + \gamma A)^{-1}e \ge 0$. Consequently, $\lambda_{ij} = \lambda^*_{ij}$ satisfy the requirements occurring in (2.8) for $1 \le i \le m$. Similarly, using Theorem 3.3 once more, we see that $L_1^* \ge 0$ and $1 - L_1^* e = 1 - \gamma b^T (I + \gamma A)^{-1} e \ge 0$ 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 $(I + \gamma A) \ge 0$. Since $\sum b_j = 1$, it follows that there is an index k with:

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

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 & Spijker (2004; 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 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). *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 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). Hence, by Lemma 3.5,

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

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 also property (II) follows from (III). ■

4 Applications and illustrations of the Theorems 2.2 and 3.4

4.1 Applications to general Runge-Kutta methods

In Kraaijevanger (1991), 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 Ferracina & Spijker (2004, 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 Kraaijevanger (1991; 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* $\mathbb{V}$, *the 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 the 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 would be true for some $c > R(A, b)$. Then, with this c, the 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 $c \leq R(A, b)$, which yields a contradiction. ■

The above corollary can be viewed as a variant to one of the results given in Ferracina & Spijker (2004; 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 the 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 & Verwer (1984, pp. 38,39), Gottlieb, Shu & Tadmor (2001, 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 following 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 $0 \leq R(A, b) < \infty$, 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

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

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

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 $j \geq i$) 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 & Shu (1998), Ruuth & Spiteri (2002), Shu (2002), Shu & Osher (1988), Spiteri & Ruuth (2002). Independently of this work, in Kraaijevanger (1991), methods (A, b) were identified that are optimal in $E_{m,p}$ with respect to $R(A, b)$. In Ferracina & Spijker (2004; Section 4), the remarkable fact was noted (but not explained!) that the methods identified in Kraaijevanger (1991) 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 non-existence of methods (A,b) with $c(A,b) > 0$ in $E_{4,4}$ and in $E_{m,5}$ – as proved in Gottlieb & Shu (1998), Ruuth & Spiteri (2002), respectively. According to Kraaijevanger (1991, 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 the Theorems 3.4 and 4.3

We give two examples illustrating the 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 (1991; 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:

$$\begin{pmatrix} \tilde{\lambda}_{21} & & & \\ \tilde{\lambda}_{31} & \tilde{\lambda}_{32} & & \\ \tilde{\lambda}_{41} & \tilde{\lambda}_{42} & \tilde{\lambda}_{43} & \\ \tilde{\lambda}_{51} & \tilde{\lambda}_{52} & \tilde{\lambda}_{53} & \tilde{\lambda}_{54} \end{pmatrix} = \begin{pmatrix} 1 & & & \\ 0 & 1 & & \\ \frac{2}{3} & 0 & \frac{1}{3} & \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} \tilde{\mu}_{21} & & & \\ \tilde{\mu}_{31} & \tilde{\mu}_{32} & & \\ \tilde{\mu}_{41} & \tilde{\mu}_{42} & \tilde{\mu}_{43} & \\ \tilde{\mu}_{51} & \tilde{\mu}_{52} & \tilde{\mu}_{53} & \tilde{\mu}_{54} \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & & & \\ 0 & \frac{1}{2} & & \\ 0 & 0 & \frac{1}{6} & \\ 0 & 0 & 0 & \frac{1}{2} \end{pmatrix}.$$

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 λ_{ij}, μ_{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 (2002). 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 (1982). 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 the conditions (1.5), (1.9) guarantee (1.4) (whenever u_{n-1}, u_n, y_i satisfy (1.2)).

Bibliography

[1] Burrage K. (1982): Efficiently implementable algebraically stable Runge-Kutta methods. *SIAM J. Numer. Anal.*, 19 No. 2, 245–258.

[2] Butcher J. C. (1987): *The numerical analysis of ordinary differential equations. Runge Kutta and general linear methods.* A Wiley-Interscience Publication. John Wiley & Sons Ltd. (Chichester).

[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] Gerisch A., Weiner R. (2003): The positivity of low-order explicit Runge-Kutta schemes applied in splitting methods. *Comput. Math. Appl.*, 45 No. 1-3, 53–67. Numerical methods in physics, chemistry, and engineering.

[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. (1983): High resolution schemes for hyperbolic conservation laws. *J. Comput. Phys.*, 49 No. 3, 357–393.

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

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

[13] 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).

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

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

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

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

[18] Morton K. W. (1980): Stability of finite difference approximations to a diffusion-convection equation. *Internat. J. Numer. Methods Engrg.*, 15 No. 5, 677–683.

[19] Ruuth S. J., Spiteri R. J. (2002): Two barriers on strong-stability-preserving time discretization methods. *J. Sci. Comput.*, 17 No. 1-4, 211–220.

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

[21] 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).

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

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

[24] 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).

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

Chapter III

Computing optimal monotonicity-preserving Runge-Kutta methods

The contents of this chapter are equal to: Ferracina L., Spijker M.N. (2005): Computing optimal monotonicity-preserving Runge-Kutta methods, submitted for publication, report Mathematical Institute, Leiden University, MI 2005-07.

Abstract

This paper deals with the numerical solution of initial value problems, for systems of ordinary differential equations, by Runge-Kutta methods which are monotonicity preserving - also called strong stability preserving (SSP). In the context of solving partial differential equations by the method of lines, Shu & Osher (1988) introduced representations of explicit Runge-Kutta methods which lead to stepsize conditions under which monotonicity is preserved. Recently, a numerical procedure, based on such representations, was employed for finding explicit Runge-Kutta methods which are optimal with respect to the above stepsize conditions; see Spiteri & Ruuth (2002, 2003), Ruuth & Spiteri (2004), Ruuth (2004).

In the present paper we continue the analysis, of Shu-Osher representations, given earlier in Higueras (2003, 2004), Ferracina & Spijker (2005). In this way we arrive naturally at a generalized and improved version of the numerical procedure mentioned above. Our procedure is, unlike the earlier one, also relevant to Runge-Kutta methods which are implicit. We illustrate our procedure in a numerical search for some optimal methods within the class of singly-diagonally-implicit Runge-Kutta methods, and we exemplify the monotonicity properties of these optimal methods in the solution of the Buckley-Leverett equation. Finally,

we formulate some open questions and conjectures.

1 Introduction

1.1 Monotonic Runge-Kutta processes

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

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

The general Runge-Kutta method, applied to problem (1.1), provides us with numerical approximations u_n of $U(n\Delta t)$, where Δt denotes a positive time step and $n = 1, 2, 3, ...$; cf. e.g. Butcher (1987), Hairer, Nørsett & Wanner (1993), Hundsdorfer & Verwer (2003). The approximations u_n can be defined in terms of u_{n-1} by the relations

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

$$u_n = y_{s+1}. \tag{1.2.b}$$

Here κ_{ij} are real parameters, specifying the Runge-Kutta method, and y_i $(1 \leq i \leq s)$ are intermediate approximations needed for computing $u_n = y_{s+1}$ from u_{n-1}. As usual, we call the Runge-Kutta method *explicit* if $\kappa_{ij} = 0$ (for $1 \leq i \leq j \leq s$), and *implicit* otherwise.

In the literature, much attention has been paid to solving (1.1) by processes (1.2) having a property which is called *monotonicity*, or *strong stability*. There are a number of closely related monotonicity concepts; see e.g. Hundsdorfer & Ruuth (2003), Hundsdorfer & Verwer (2003), Gottlieb, Shu & Tadmor (2001), Shu (2002), Shu & Osher (1988), Spiteri & Ruuth (2002).

In this paper we shall deal with a quite general monotonicity concept, and we shall study the problem of finding Runge-Kutta methods which have optimal properties regarding this kind of monotonicity. As we want to address this problem in a general setting, we assume F to be a mapping from an arbitrary real vector space $\mathbb{V}$ into itself and $\|.\|$ to be a real convex function on $\mathbb{V}$ (i.e. $\|v\| \in \mathbb{R}$ and $\|\lambda v + (1-\lambda)w\| \leq \lambda\|v\| + (1-\lambda)\|w\|$ for all $v, w \in \mathbb{V}$ and $0 \leq \lambda \leq 1$). We will deal with processes (1.2) which are monotonic in the sense that the vectors $u_n \in \mathbb{V}$ computed from $u_{n-1} \in \mathbb{V}$, via (1.2), satisfy

$$\|u_n\| \leq \|u_{n-1}\|. \tag{1.3}$$

In order to illustrate the general property (1.3), we consider the numerical solution of a Cauchy problem for the hyperbolic partial differential equation,

$$\frac{\partial}{\partial t}u(x,t) + \frac{\partial}{\partial x}\Phi(u(x,t)) = 0, \tag{1.4}$$

where $t \geq 0$, $-\infty < x < \infty$. Here Φ stands for a given (possibly nonlinear) scalar function, so that (1.4) is a simple instance of a conservation law, cf., e.g., Laney (1998), LeVeque (2002). Suppose (1.1) originates from a (method of lines) semi-discretization of (1.4). 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; the actual function values $F(y)$ depend on the given Φ as well as on the process of semi-discretization being used - see loc. cit.. Since $\frac{d}{dt}U(t) = F(U(t))$ now stands for a semi-discrete version of the conservation law (1.4), it is desirable that the fully discrete process (consisting of an application of (1.2) to (1.1)) be monotonic in the sense of (1.3), where $\|.\|$ denotes the *total-variation* seminorm

$$\|y\|_{TV} = \sum_{j=-\infty}^{+\infty} |\eta_j - \eta_{j-1}| \quad (\text{for } y \in \mathbb{R}^\infty \text{ with components } \eta_j). \tag{1.5}$$

With this seminorm, the monotonicity property (1.3) reduces to the so-called *total-variation-diminishing* (TVD) property. For an explanation of the importance of the last property, as well as for further examples, where (1.3) is a desirable property or a natural demand, we refer to Harten (1983), Laney (1998), LeVeque (2002), Hundsdorfer & Ruuth (2003), Hundsdorfer & Verwer (2003).

In order to place the study, to be carried out in the present paper, in the right context, we shall first review, in Section 1.2, an approach of Shu & Osher (1988) to proving the general property (1.3) for certain explicit Runge-Kutta methods. Next, in Section 1.3, we shall briefly review a numerical procedure used in Spiteri & Ruuth (2002, 2003), Ruuth & Spiteri (2004), Ruuth (2004) for finding explicit Runge-Kutta methods which are optimal with respect to stepsize conditions guaranteeing (1.3). Finally, in Section 1.4, we shall outline the study to be presented in the rest of our paper.

1.2 The Shu-Osher representation

By Shu & Osher (1988) (see also Shu (1988)) a representation of explicit Runge-Kutta methods (1.2) was introduced which is very useful for proving property (1.3). In order to describe this representation, we consider an arbitrary explicit Runge-Kutta method (1.2) specified by coefficients κ_{ij}. We assume that λ_{ij} (for $1 \leq j < i \leq s+1$) are any real parameters with

$$\lambda_{ij} \geq 0, \qquad \lambda_{i1} + \lambda_{i2} + ... + \lambda_{i,i-1} = 1 \quad (1 \leq j < i \leq s+1), \tag{1.6}$$

and we define corresponding coefficients μ_{ij} by

$$\mu_{ij} = \kappa_{ij} - \sum_{l=j+1}^{i-1} \lambda_{il}\kappa_{lj} \quad (1 \leq j < i \leq s+1) \tag{1.7}$$

(where the last sum should be interpreted as 0, when $j = i-1$).

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

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

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

The representation (1.8) is very relevant in the situation where, for some $\tau_0 > 0$,

$$\|v + \tau_0\, F(v)\| \le \|v\| \quad (\text{for all } v \in \mathbb{V}). \tag{1.9}$$

Clearly, in case (1.1) results from applying the method of lines to a given partial differential equation, (1.9) amounts to a condition on the actual manner in which the semi-discretization has been performed. In general, (1.9) can be interpreted as monotonicity of the forward Euler process with stepsize τ_0, cf. e.g. Hundsdorfer & Verwer (2003). We also note that, for $0 \le \tau < \tau_0$, condition (1.9) implies $\|v + \tau F(v)\| \le \|(\tau/\tau_0)(v + \tau_0 F(v)) + (1 - \tau/\tau_0)v\| \le \|v\|$ – i.e. the Euler process is still monotonic with any stepsize $\tau \in [0, \tau_0)$.

Assume (1.9). Then, for $2 \le i \le s+1$, the vectors y_i in (1.8) can be rewritten as convex combinations of Euler steps with stepsizes $\tau = \Delta t(\mu_{ij}/\lambda_{ij})$. From this observation, it follows easily that (1.3) is now valid, under a stepsize restriction of the form

$$0 < \Delta t \le c \cdot \tau_0, \tag{1.10}$$

where $c = \min_{ij} \gamma_{ij}$, with $\gamma_{ij} = \lambda_{ij}/\mu_{ij}$ (if $\mu_{ij} \ge 0$), $\gamma_{ij} = 0$ (if $\mu_{ij} < 0$) – here, as well as below, we use the convention $\lambda/\mu = \infty$ for $\lambda \ge 0$, $\mu = 0$.

Clearly, in order that $c > 0$, it is necessary that all μ_{ij} are nonnegative. Using an idea of Shu (1988), Shu & Osher (1988), one can avoid this condition on μ_{ij} in certain cases. Suppose, for instance, that $\frac{d}{dt}U(t) = F(U(t))$ approximates (1.4); then, for $\mu_{ij} < 0$, the quantity $\mu_{ij}F(y_j)$ in (1.8) should be replaced by $\mu_{ij}\tilde{F}(y_j)$, where $\tilde{F}$ approximates $-\frac{\partial}{\partial x}\Phi$ to the same order of accuracy as F, but satisfies (instead of (1.9))

$$\|v - \tau_0\, \tilde{F}(v)\| \le \|v\| \quad (\text{for all } v \in \mathbb{V}). \tag{1.11}$$

E.g., if $\frac{\partial}{\partial x}\Phi(u(x,t)) = \frac{\partial}{\partial x}u(x,t)$, $F_i(y) = (\eta_{i-1} - \eta_i)/\Delta x$, $\|.\| = \|.\|_{TV}$ and $\tau_0 = 1/\Delta x$, then $\tilde{F}_i(y) = (\eta_i - \eta_{i+1})/\Delta x$ would do. Clearly, after such a (partial) replacement of F by $\tilde{F}$, property (1.3) is still valid under a stepsize condition of the form (1.10), with

$$c = \min_{ij} \frac{\lambda_{ij}}{|\mu_{ij}|}. \tag{1.12}$$

If every coefficient μ_{ij} is nonnegative, then the number of function evaluations, in process (1.8), is equal to the number of stages, s. However, if both $F(y_j)$ and $\tilde{F}(y_j)$ were required for some j, then the number of function evaluations, needed for computing u_n from u_{n-1}, would be greater than s. Therefore, in order to avoid this unfavourable situation, it is natural to demand that, for each given j, all non-zero coefficients μ_{ij} (with $j < i \leq s+1$) have the same sign; cf. e.g. Ruuth & Spiteri (2004). Accordingly, we assume that, for $1 \leq j \leq s$, *sign indicators* $\sigma_j = \pm 1$ can be associated to the coefficients μ_{ij} such that

$$\mu_{ij} \geq 0 \text{ (whenever } \sigma_j = 1), \text{ and } \mu_{ij} \leq 0 \text{ (whenever } \sigma_j = -1). \tag{1.13}$$

For completeness we note that one can rewrite any process (1.8), for which *no* σ_j exist satisfying (1.13), in the form of a different Shu-Osher process, with more stages, satisfying (1.13).

The following theorem summarizes our above discussion of the Shu-Osher process (1.8).

Theorem 1.1 (Shu and Osher).

(i) *Consider an explicit Runge-Kutta method* (1.2) *specified by coefficients* κ_{ij}, *and assume* (1.6) *and* (1.7). *Then processes* (1.2) *and* (1.8) *are equivalent.*

(ii) *Assume* (1.6), (1.13) *and let* c *be defined by* (1.12). *Consider any vector space* $\mathbb{V}$ *and convex function* $\|.\|$ *on* $\mathbb{V}$; *assume* (1.9), (1.11). *Then stepsize condition* (1.10) *guarantees property* (1.3), *for process* (1.8) *where* $F(y_j)$ *is replaced throughout by* $\tilde{F}(y_j)$ *when* $\sigma_j = -1$.

The above propositions (i) and (ii) are essentially due to Shu & Osher (1988) - in that paper the starting-point was just a slightly stronger assumption, than above, regarding $\|.\|$, F and $\tilde{F}$; see loc. cit.

Clearly, if for a given Runge-Kutta method a representation (1.8) exists such that the assumptions of Theorem 1.1 are fulfilled with $c > 0$, then the Runge-Kutta process maintains monotonicity of the Euler processes in (1.9), (1.11), under the stepsize restriction (1.10). For that reason, Runge-Kutta methods for which such a positive c exists, may be called *monotonicity-preserving* or *strong-stability-preserving* – cf. Gottlieb, Shu & Tadmor (2001), Ferracina & Spijker (2004).

For future reference, we note that the implementation of process (1.8) involving F and $\tilde{F}$, as discussed above, can be written in the form

$$\begin{aligned} y_1 &= u_{n-1}, \\ y_i &= \sum_{j=1}^{i-1} [\lambda_{ij}\, y_j + \Delta t \cdot \mu_{ij} f_j(y_j)] \quad (2 \leq i \leq s+1), \\ u_n &= y_{s+1}, \end{aligned} \tag{1.14}$$

where $f_j(y_j) = F(y_j)$ for $\sigma_j = 1$, and $f_j(y_j) = \tilde{F}(y_j)$ for $\sigma_j = -1$. In view of (1.9), (1.11), these functions f_j satisfy

$$\|v + \tau_0\, \sigma_j f_j(v)\| \leq \|v\| \quad (1 \leq j \leq s, \quad v \in \mathbb{V}). \tag{1.15}$$

1.3 A numerical procedure used by Ruuth & Spiteri

Below we denote by $E_{s,p}$ the class of all explicit s-stage Runge-Kutta methods with (classical) order of accuracy at least p.

Clearly, it would be awkward if the coefficient c, occurring in Theorem 1.1 (ii), were zero or so small that (1.10) reduces to a stepsize restriction which is to severe for any practical purposes – in fact, the less restrictions on Δt the better. Accordingly, for given s and p, much attention has been paid in the literature to determining Shu-Osher processes (1.8), (1.13) in $E_{s,p}$ which are optimal with regard to the size of c. Extensive numerical searches in $E_{s,p}$ for optimal Shu-Osher processes (1.8), (1.13), were recently carried out in Ruuth & Spiteri (2004), Spiteri & Ruuth (2003), Ruuth (2004).

For given s and p, the numerical searches carried out in the last three papers, are essentially based on the following optimization problem (1.16), in which λ_{ij}, μ_{ij}, γ are the independent variables and $f(\lambda_{ij}, \mu_{ij}, \gamma) = \gamma$ is the objective function.

(1.16.a) maximize γ, subject to the following constraints:

(1.16.b) $\lambda_{ij} - \gamma\,|\mu_{ij}| \geq 0 \quad (1 \leq j < i \leq s+1)$;

(1.16.c) λ_{ij} satisfy (1.6), and there are $\sigma_j = \pm 1$ such that (1.13) holds;

(1.16.d) the coefficients κ_{ij}, satisfying (1.7), specify a Runge-Kutta method (1.2) belonging to class $E_{s,p}$.

Clearly, the variable γ in (1.16) corresponds to c in (1.12), and parameters λ_{ij}, μ_{ij}, γ solving the optimization problem (1.16) yield a Shu-Osher process in $E_{s,p}$ which is optimal in the sense of c, (1.12).

For completeness we note that, also for the special case where all σ_j in (1.13) are required to satisfy $\sigma_j = 1$, optimal Shu-Osher processes (1.8) were determined in $E_{s,p}$ – either by clever ad hoc arguments, or by numerical computations based on an earlier version of (1.16); see Shu & Osher (1988), Spiteri & Ruuth (2002).

Problem (1.16), as well as the earlier version just mentioned, were solved numerically by Ruuth and Spiteri – initially using Matlab's Optimization Toolbox, subsequently with the optimization software package BARON; see Ruuth & Spiteri (2004), Spiteri & Ruuth (2002, 2003), Ruuth (2004) and references therein. In this way optimal methods were found in $E_{s,p}$, for $1 \leq s \leq 10, \quad 1 \leq p \leq 5$.

1.4 Outline of the rest of the paper

Various generalizations and refinements of Theorem 1.1 were given recently, notably in Higueras (2003, 2004), Ferracina & Spijker (2004, 2005). In Section 2 we shall give a concise review, and an extension, of some of these results.

In Section 3, we shall use the material of Section 2 so as to arrive at a generalized and improved version of Ruuth & Spiteri's approach (1.16) to finding optimal methods.

Our approach is, unlike (1.16), not restricted to explicit methods. Accordingly, in Section 4, we shall illustrate our new version of (1.16) in a numerical search for some optimal methods within the important class of singly-diagonally-implicit Runge-Kutta (SDIRK) methods. In this way we shall arrive at optimal s-stage methods of orders 2, and 3.

In Section 5, we shall exemplify the preceding material with a simple numerical experiment in which various optimal SDIRK methods are applied to a scalar conservation law, the 1-dimensional Buckley-Leverett equation.

The material of Sections 4 and 5 leads to some conjectures and open questions which will be formulated in our last section, Section 6.

2 An extension and analysis of the Shu-Osher representation

2.1 A generalization of Theorem 1.1

As in the previous section, $\mathbb{V}$ denotes an arbitrary real vector space. Furthermore, $f_j(v)$ denote given functions, defined for all $v \in \mathbb{V}$, with values in $\mathbb{V}$. We shall deal with the following general process:

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

$$\text{(2.1.b)} \quad u_n = y_{s+1}.$$

Here λ_{ij}, μ_{ij} denote arbitrary real coefficients. Clearly, this general process reduces to (1.14) in case $\mu_{ij} = \lambda_{ij} = 0$ (for $1 \le i \le j \le s$), $\sum_{j=1}^{s} \lambda_{ij} = 1$ (for $2 \le i \le s+1$).

Along with (2.1), we consider the following generalization of (1.2):

$$\text{(2.2.a)} \qquad y_i = u_{n-1} + \Delta t \sum_{j=1}^{s} \kappa_{ij} f_j(y_j) \quad (1 \le i \le s+1),$$

$$\text{(2.2.b)} \qquad u_n = y_{s+1}.$$

We define the $(s+1) \times s$ coefficient matrices K, L, M as

$$\text{(2.3)} \qquad K = (\kappa_{ij}), \quad L = (\lambda_{ij}), \quad M = (\mu_{ij}),$$

so that the numerical methods (2.1) and (2.2), respectively, can be identified with the pair (L, M) and the matrix K.

Below we shall relate (2.1) to (2.2). We shall denote the $s \times s$ identity matrix

by I, and we shall use the following definitions and assumptions:

$$(2.4)\quad K_0 = \begin{pmatrix} \kappa_{11} & \dots & \kappa_{1s} \\ \vdots & & \vdots \\ \kappa_{s1} & \dots & \lambda_{ss} \end{pmatrix}, \quad L_0 = \begin{pmatrix} \lambda_{11} & \dots & \lambda_{1s} \\ \vdots & & \vdots \\ \lambda_{s1} & \dots & \lambda_{ss} \end{pmatrix}, \quad M_0 = \begin{pmatrix} \mu_{11} & \dots & \mu_{1s} \\ \vdots & & \vdots \\ \mu_{s1} & \dots & \mu_{ss} \end{pmatrix},$$

$$(2.5)\quad M = K - LK_0,$$

$$(2.6)\quad I - L_0 \text{ is invertible.}$$

Clearly, (2.5) is a straightforward generalization of (1.7); and (2.6) is automatically fulfilled if (2.1) stands for (1.14).

We shall deal with monotonicity of process (2.1), under the following generalized version of condition (1.6):

$$(2.7)\qquad L \geq 0, \quad Le_s \leq e_{s+1}.$$

Here, and in the following, e_m stands for the column vector in $\mathbb{R}^m$ with all components equal to 1 (for $m = s,\ s+1$). Furthermore, the first inequality in (2.7) should be interpreted entry-wise, whereas the second inequality is to be interpreted component-wise. All inequalities between matrices or vectors, to be stated below, should be interpreted in the same way.

In addition to (2.7), we shall assume that sign indicators $\sigma_j = \pm 1$ can be adjoined to the columns of M, such that

$$(2.8)\quad \mu_{ij} \geq 0 \ \ (1 \leq i \leq s+1 \text{ and } \sigma_j = 1), \quad \mu_{ij} \leq 0 \ \ (1 \leq i \leq s+1 \text{ and } \sigma_j = -1).$$

For arbitrary $(s+1) \times s$ matrices $L = (\lambda_{ij})$, $M = (\mu_{ij})$, we define

$$(2.9)\quad c(L, M) = \min\{\gamma_{ij} : 1 \leq i \leq s+1,\ 1 \leq j \leq s\}, \ \ \gamma_{ij} = \begin{cases} \lambda_{ij}/\mu_{ij} & \text{if } \mu_{ij} > 0, \\ \infty & \text{if } \mu_{ij} = 0, \\ 0 & \text{if } \mu_{ij} < 0, \end{cases}$$

and we put

$$(2.10)\qquad |M| = (|\mu_{ij}|).$$

The following theorem can be viewed as an extension, of the original Shu-Osher Theorem 1.1, to the general processes (2.1), (2.2).

Theorem 2.1.
With the notations (2.3), (2.4), *the following statements are valid.*

(I) *Assume* (2.5), (2.6). *Then the general processes* (2.1) *and* (2.2) *are equivalent.*

(II) *Assume* (2.6), (2.7), (2.8). *Let* $c = c(L, |M|)$ – *see* (2.9), (2.10). *Then, for any vector space* $\mathbb{V}$ *and convex function* $\|.\|$ *on* $\mathbb{V}$, *conditions* (1.10), (1.15) *guarantee the monotonicity property* (1.3), *whenever* u_{n-1}, u_n, y_i *satisfy* (2.1).

In view of Theorems 1.1, 2.1, we shall call any process (2.1), satisfying (2.5), (2.6), (2.7), a *generalized Shu-Osher representation* of the Runge-Kutta process (2.2). From Theorem 2.1, we immediately obtain the following corollary relevant to the Runge-Kutta process (2.2):

Corollary 2.2. *Assume* (2.5), (2.6), (2.7), (2.8), *and let* $c = c(L, |M|)$. *Then for any vector space* $\mathbb{V}$ *and convex function* $\|.\|$ *on* $\mathbb{V}$, *conditions* (1.10), (1.15) *guarantee the monotonicity property* (1.3), *whenever* u_{n-1}, u_n, y_i *satisfy the Runge-Kutta relations* (2.2).

Remark 2.3.

(a) Assume (2.5), (2.6), (2.7), (2.8). Let F, $\tilde{F}$ be as in (1.9), (1.11) and consider the Runge-Kutta process (2.2) with $f_j = F$ (if $\sigma_j = 1$), $f_j = \tilde{F}$ (if $\sigma_j = -1$). From Corollary 2.2 we easily conclude that the stepsize condition $0 \leq \Delta t \leq c(L, |M|) \cdot \tau_0$ guarantees property (1.3), whenever u_{n-1}, u_n, y_i satisfy (2.2).

(b) Runge-Kutta procedures of the form (2.2) occur also very naturally in the solution of *nonautonomous* equations $U'(t) = F(t, U(t))$; notably with $f_j(v) = F(\tau_j, v)$, $\tau_j = [(n-1+\gamma_j)]\Delta t$, $\gamma_j = \sum_{k=1}^{s} \kappa_{jk}$ – see e.g. Butcher (1987), Hairer, Nørset & Wanner (1993), Hundsdorfer & Verwer (2003). Accordingly, the above corollary (with all $\sigma_j = 1$) is highly relevant to establishing monotonicity for such Runge-Kutta procedures: assuming that $\|v + \tau_0 F(\tau_j, v)\| \leq \|v\|$ (for $1 \leq j \leq s$ and $v \in \mathbb{V}$), one arrives at monotonicity of the Runge-Kutta process, under the stepsize condition $0 \leq \Delta t \leq c(L, M) \cdot \tau_0$.

(c) Consider a Runge-Kutta method of the form (1.2), and assume that matrices L, M, satisfying (2.5) – (2.8) exist, with $c(L, |M|) > 0$. Then, in view of Remark 2.3 (a), and in line with the terminology in Section 1.2, we will say that the Runge-Kutta method under consideration is *monotonicity-preserving*.

We note that Theorem 2.1 can be viewed as an extension of conclusions, regarding process (2.1), formulated in the recent literature. The equivalence of (2.1) and (2.2), in the special situation where $f_j = F$ $(1 \leq j \leq s)$, as well as the monotonicity of (2.1) when $f_j = F$ (for $\sigma_j = 1$), $f_j = \tilde{F}$ (for $\sigma_j = -1$), were treated earlier – cf. Higueras (2003, 2004), Ferracina & Spijker (2005). Although Theorem 2.1 covers situations which were not considered in the above papers, its proof can easily be given by arguments which are almost literally the same as in these papers. Therefore, we refer the reader for the proof of Theorem 2.1 to loc. cit.

2.2 The maximal size of $c(L, |M|)$

Let a Runge-Kutta method, with coefficient matrix K, be given. For any matrices L, M as in Corollary 2.2, the coefficient $c = c(L, |M|)$ yields a stepsize condition (1.10) which can guarantee monotonicity for the Runge-Kutta process – cf. Corollary 2.2, Remark 2.3 (a). Consequently, the larger $c(L, |M|)$ the better. The natural question thus arises, for the given matrix K, what is the maximal size

of $c(L, |M|)$. Theorem 2.6, below, will specify this maximal size in terms of the Runge-Kutta matrix K.

In Theorem 2.6, a coefficient introduced by Kraaijevanger (1991) will play a prominent part. In defining this coefficient, we deal with K, K_0 as in (2.3), (2.4) and we consider, for real γ, the following conditions:

(2.11) $(I + \gamma K_0)$ is invertible, $\gamma K(I + \gamma K_0)^{-1} \geq 0$, $\gamma K(I + \gamma K_0)^{-1} e_s \leq e_{s+1}$.

Definition 2.4 (Kraaijevanger's coefficient).
For arbitrary $(s+1) \times s$ matrices K, we define

$$R(K) = \sup\{\gamma \ : \ \gamma \geq 0 \text{ and } (2.11) \text{ holds}\}.$$

For completeness, we note that the original definition, given by Kraaijevanger (1991), is slightly more complicated and essentially amounts to

$$R(K) = \sup\{r \ : \ r \in \mathbb{R} \text{ and } (2.11) \text{ holds for all } \gamma \in [0, r]\}.$$

(Moreover, Kraaijevanger (1991) used the notation $R(A, b)$, instead of $R(K)$, but this difference is immaterial for our discussion.) The following theorem implies that the above two definitions of $R(K)$ are equivalent:

Theorem 2.5.
Let K be given and let γ be any finite value with $0 \leq \gamma \leq R(K)$ (Definition (2.4)). Then γ satisfies (2.11).

Theorem 2.5 can be viewed as a (somewhat stronger) version of earlier results in the literature – for related material, see Kraaijevanger (1991, Lemma 4.4), Higueras (2004, Proposition 2.11), Horváth (1998, Theorem 4).

In Section 2.3, we shall give an integrated proof of Theorem 2.5 and Theorem 2.6; the former theorem will be used in our proof of the latter.

In Theorem 2.6 we shall deal with coefficient matrices $K = (\kappa_{ij})$ satisfying

(2.12) $\kappa_{ij} \geq 0$ $(1 \leq i \leq s+1$ and $\sigma_j = 1)$, $\kappa_{ij} \leq 0$ $(1 \leq i \leq s+1$ and $\sigma_j = -1)$.

Theorem 2.6.
Let $K = (\kappa_{ij})$ and $\sigma_j = \pm 1$ $(1 \leq j \leq s)$ be given. Then there exist L, M satisfying (2.5) – (2.8) *if and only if K satisfies* (2.12). *Furthermore, if* (2.12) *is fulfilled, the following three statements are valid.*

(a) *We have* $\sup c(L, |M|) = R(|K|)$, *where the supremum is over all pairs (L, M) satisfying* (2.5) – (2.8).

(b) *We also have* $\sup c(L, |M|) = R(|K|)$, *where the supremum is only over all pairs (L, M) satisfying* (2.5) – (2.8), *with $L = \gamma\,|M|$, $\gamma \geq 0$.*

(c) *If $R(|K|) < \infty$, then the suprema in Statements* (a), (b) *are maxima.*

Theorem 2.6 combines and extends various results given earlier in the literature, see Higueras (2003, 2004), Ferracina & Spijker (2005).

2.3 Proof of Theorems 2.5, 2.6

Our proof below, of Theorems 2.5, 2.6, will be based on the following lemma, which can be viewed as an extension of related results in the literature; see Higueras (2003, 2004), Ferracina & Spijker (2005).

Lemma 2.7. *Let K be a given $(s+1) \times s$ matrix and $\gamma \geq 0$. Then Statements* (a), (b) *are valid.*

(a) *Suppose L, M are $(s+1) \times s$ matrices, with $L \geq \gamma M \geq 0$, satisfying* (2.5), (2.6), (2.7). *Then K and γ satisfy* (2.11).

(b) *Suppose, conversely, that* (2.11) *is fulfilled. Then there exist matrices L, M, with $L = \gamma M \geq 0$, satisfying* (2.5), (2.6), (2.7).

Proof. 1. Before going into the actual proof, we assume (2.6), (2.7) and consider an arbitrary $s \times s$ matrix E_0, with

$$0 \leq E_0 \leq L_0. \tag{2.13}$$

We shall prove that

$$I - E_0 \text{ is invertible, with } (I - E_0)^{-1} \geq I. \tag{2.14}$$

From (2.13) we conclude that the spectral radius of E_0 does not exceed the spectral radius, say r, of L_0; see, e.g., Horn & Johson (1985, Section 8.1). From $L_0 \geq 0$, $L_0 e_s \leq e_s$ we see that $r \leq 1$. Since $I - L_0$ is invertible, it follows – e.g. from a well known corollary to Perron's theorem, see Horn & Johson (1985, Section 8.3) – that $r < 1$. Consequently, the spectral radius of E_0 is less than 1. Hence, $I - E_0$ is invertible, with $(I - E_0)^{-1} = I + E_0 + (E_0)^2 + ... \geq I$, i.e. (2.14).

2. Assume (2.5), (2.6), (2.7) and $L \geq \gamma M \geq 0$. In order to prove (2.11), we define $E = L - \gamma M$, $E_0 = L_0 - \gamma M_0$. Note that, with this definition, (2.13) is fulfilled, so that (2.14) is valid as well.

From (2.5) we obtain $\gamma K_0 = (I - L_0)^{-1}(\gamma M_0) = (I - L_0)^{-1}(L_0 - E_0)$, and therefore $\gamma K_0 = -I + (I - L_0)^{-1}(I - E_0)$. Hence

$$I + \gamma K_0 \text{ is invertible and } (I + \gamma K_0)^{-1} = (I - E_0)^{-1}(I - L_0). \tag{2.15.a}$$

Since $\gamma K = \gamma M + L(\gamma K_0) = (L - E) + L(\gamma K_0)$, we find, by using our last expression for γK_0, that $\gamma K = -E + L(I - L_0)^{-1}(I - E_0)$. Combining this equality with (2.15.a), there follows

$$\gamma K(I + \gamma K_0)^{-1} = L - E(I - E_0)^{-1}(I - L_0). \tag{2.15.b}$$

The right-hand member of (2.15.b) is easily seen to be equal to $(L - E) + E(I - E_0)^{-1}(L_0 - E_0) \geq 0$. This implies the first inequality in (2.11). Furthermore, when we premultiply the vector e_s by the right-hand member of (2.15.b), we obtain the

vector $Le_s - E(I - E_0)^{-1}(I - L_0)e_s \leq Le_s \leq e_{s+1}$. Consequently, the second inequality in (2.11) is fulfilled as well – which completes the proof of Part (a) of the lemma.

3. In order to prove Part (b) of the lemma, we assume (2.11) and we define $M = K(I + \gamma K_0)^{-1}$, $L = \gamma M$. Clearly, (2.7) is fulfilled. Moreover $I - L_0 = (I + \gamma K_0)^{-1}$, which proves (2.6). Finally, a short calculation shows that (2.5) is fulfilled as well. ■

Proof of Theorem 2.5.
First suppose $0 \leq \gamma < R(K)$. Choose $\gamma' > \gamma$ such that γ' satisfies (2.11). Applying Lemma 2.7 (b) to γ', it follows that L, M exist satisfying (2.5), (2.6), (2.7) with $L = \gamma' M \geq \gamma M \geq 0$. An application of Lemma 2.7 (a) proves that γ satisfies (2.11).

Next, suppose $0 < \gamma = R(K) < \infty$, and (2.11) is violated. Using continuity arguments one sees that, in order to complete the proof of Theorem 2.5, it is enough to show that $(I + \gamma K_0)$ is invertible.

Let $\varepsilon \in (0, 1)$ be such that $\gamma' = \gamma/(1 + \varepsilon)$ satisfies (2.11). Then the matrix $P_0 = \gamma' K_0 (I + \gamma' K_0)^{-1}$ has a spectral radius not exceeding 1. We have $I + \gamma K_0 = (I + \gamma' K_0)(I + \varepsilon P_0)$, so that $I + \gamma K_0$ equals the product of two invertible matrices. Hence $I + \gamma K_0$ is invertible. ■

Proof of Theorem 2.6.
First, suppose K satisfies (2.12). Then the matrices $L = 0$, $M = K$ satisfy (2.5) – (2.8).

Next, suppose L, M satisfy (2.5) – (2.8). We shall denote by $|M_0|$ and $|K_0|$ the $s \times s$ matrices with entries $|\mu_{ij}|$ and $|\kappa_{ij}|$, respectively. Defining $D = \mathrm{diag}(\sigma_1, ..., \sigma_s)$, we have $|M_0| = M_0 D = (K_0 - L_0 K_0)D = (I - L_0)K_0 D$, i.e. $K_0 D = (I - L_0)^{-1}|M_0|$. In the first part of the proof of Lemma 2.7, we showed that (2.13) implies (2.14). Using this implication, with $E_0 = L_0$, we obtain $(I - L_0)^{-1} \geq I$, so that $K_0 D \geq |M_0| \geq 0$. Consequently, $K_0 D = |K_0|$ and therefore $KD = (M + LK_0)D = |M| + L|K_0|$. It follows that $KD \geq 0$, which proves (2.12).

Finally, assume again (2.12) and, without loss of generality, that $K \neq 0$. One easily sees that, in order to establish (a), (b), (c), it is enough to prove the following two implications:

(i) If L, M satisfy (2.5) – (2.8), then $c(L, |M|) \leq R(|K|)$.

(ii) If γ is a finite value with $0 < \gamma \leq R(|K|)$, then L, M exist satisfying (2.5) – (2.8) with $L = \gamma |M|$.

In order to prove (i), we assume (2.5) – (2.8). Using (2.9), (2.10) and our assumption $K \neq 0$, there follows

$$|M| = |K| - L|K_0|, \quad L \geq \gamma |M| \geq 0 \quad \text{with } \gamma = c(L, |M|) < \infty.$$

Applying Lemma 2.7 (a) to the pair $(L, |M|)$, we arrive at the inequality in (i).

In order to prove (ii), we consider a finite $\gamma \in (0, R(|K|)]$. Applying Theorem 2.5 and Lemma 2.7 (b) to the matrix $|K|$, we see that matrices $L, \tilde{M}$ exist with $L = \gamma\tilde{M} \geq 0$, $\tilde{M} = |K| - L|K_0|$, satisfying (2.6), (2.7). A multiplication of the last equality by $D = \mathrm{diag}(\sigma_1, ..., \sigma_s)$, yields $\tilde{M}D = K - LK_0$; so that (2.5) is fulfilled with $M = \tilde{M}D$. Since $\tilde{M} \geq 0$, we have $\tilde{M} = |M|$. Therefore L, M are as required in (ii). ■

3 Generalizing and improving Ruuth & Spiteri's procedure

In this section we shall give three General Procedures I, II and III, which can be viewed as variants to Ruuth & Spiteri's procedure (1.16). We think that our third procedure is the most attractive one; we present the other two mainly in order to put the third one in the right perspective and to compare it more easily with the approach (1.16).

Our procedures are relevant to arbitrary Runge-Kutta methods (not necessarily explicit). In line with Corollary 2.2 and Remark 2.3 (a), the procedures focus on optimizing $c(L, |M|)$ – which generalizes the optimization of (1.12), as in Ruuth & Spiteri's approach. We shall deal with maximization of $c(L, |M|)$, over all generalized Shu-Osher representations (L, M) of Runge-Kutta methods with coefficient matrices $K = (\kappa_{ij})$ belonging to a given class $\mathscr{C}$. We assume all $K \in \mathscr{C}$ to have the same number of columns, s, and for each individual $K \in \mathscr{C}$ we assume that sign indicators $\sigma_j = \pm 1\, (1 \leq j \leq s)$ exist, with property (2.12).

We denote by $\bar{\mathscr{C}}$ the set of all Shu-Osher pairs (L, M) satisfying (2.5) – (2.8), where K is any matrix of class $\mathscr{C}$ with sign indicators σ_j.

Below we give our three general procedures. We will use the notation (2.3), and with γ, κ_{ij}, λ_{ij}, μ_{ij} we denote independent variables.

GPI: General Procedure I

(3.1.a) maximize γ, subject to the constraints:
(3.1.b) $\lambda_{ij} - \gamma\,|\mu_{ij}| \geq 0 \quad (i = 1, 2, ..., s+1, \quad j = 1, 2, ..., s)$;
(3.1.c) $(L, M) \in \bar{\mathscr{C}}$.

GPII: General Procedure II

(3.2.a) maximize γ, subject to the constraints:
(3.2.b) $\lambda_{ij} - \gamma\,|\mu_{ij}| = 0 \quad (i = 1, 2, ..., s+1, \quad j = 1, 2, ..., s)$;
(3.2.c) $(L, M) \in \bar{\mathscr{C}}$.

GPIII: General Procedure III

(3.3.a) maximize γ, subject to the constraints:

(3.3.b) γ satisfies (2.11), with K_0, K replaced by $|K_0|$, $|K|$;

(3.3.c) $K = (\kappa_{ij}) \in \mathscr{C}$.

The variable γ, in the above three procedures, corresponds to $c(L, |M|)$. Furthermore, parameters λ_{ij}, μ_{ij}, γ, solving the optimization problems (3.1) or (3.2), yield a Shu-Osher pair (L, M) in $\bar{\mathscr{C}}$ which is optimal with respect to $c(L, |M|)$; similarly, parameters κ_{ij}, γ, solving (3.3), yield an optimal Runge-Kutta matrix K in $\mathscr{C}$. The following theorem relates the optimal value of $c(L, |M|)$ formally to the maximum of γ in the General Procedures I, II, III.

Theorem 3.1. *Let $\mathscr{C}$ be a given class of $(s+1) \times s$ coefficient matrices K such that, for each individual $K = (\kappa_{ij})$, sign indicators $\sigma_j = \pm 1 \, (1 \le j \le s)$ exist satisfying* (2.12). *Let $\bar{\mathscr{C}}$ be the set of all Shu-Osher pairs (L, M) satisfying* (2.5) – (2.8), *where K is any matrix of class $\mathscr{C}$ with sign indicators σ_j. Assume that $c^* = \max\{c(L, |M|) : (L, M) \in \bar{\mathscr{C}}\}$ exists and is finite. Then the maximum of γ, under the constraints as specified in any of the General Procedures* I, II *or* III, *exists and equals c^*.*

Proof.

1. Clearly, under the assumptions of the theorem, we have, for all $(L, M) \in \bar{\mathscr{C}}$, the equality

$$c(L, |M|) = \max\{\gamma : \lambda_{ij} - \gamma \, |\mu_{ij}| \ge 0 \ \ (\text{for all } i, j)\}. \tag{3.4}$$

This proves that the maximum of γ, specified in GPI, does exist and is equal to c^*.

2. Let $(L^*, M^*) \in \bar{\mathscr{C}}$ be an optimal pair, i.e., $c(L^*, |M^*|) = c^* < \infty$; and let $K^* \in \mathscr{C}$ be such that (L^*, M^*) satisfies (2.5) – (2.8) for $K = K^*$. By applying Theorem 2.6, Part (a), one can conclude that

$$c^* = c(L^*, |M^*|) = \max_{\bar{\mathscr{C}}} c(L, |M|) = R(|K^*|) = \max_{\mathrm{C}} R(|K|) < \infty. \tag{3.5}$$

From Theorem 2.5, we see that, for each $K \in \mathscr{C}$, the value $R(|K|)$ equals the maximum over all γ satisfying (2.11) with K_0, K replaced by $|K_0|$, $|K|$. In view of (3.5), we thus see that GPIII yields the value c^*.

3. By virtue of Theorem 2.6, we have $c^* = \max c(L, |M|)$ where the maximum is over all $(L, M) \in \bar{\mathscr{C}}$, with $L = \gamma |M|$, $\gamma \in \mathbb{R}$. For any pair (L, M) of this type, we see from (3.4) that $c(L, |M|) = \gamma$. Consequently, also GPII yields the value c^*. ■

Clearly, General Procedure I can be viewed as a direct generalization of Ruuth & Spiteri's procedure (1.16) for $E_{s,p}$, to arbitrary classes $\mathscr{C}$ of general Runge-Kutta methods.

General Procedure II can be regarded as an improvement over GPI, because the number of independent variables has essentially been reduced by (almost) 50%. Clearly, GPII can be expected to be considerably more efficient than GPI.

Finally, although (3.3.b) is usually more complicated than (3.2.b), we still think that General Procedure III constitutes a (further) improvement over GPII (and a-fortiori over GPI). The fact is that condition (3.3.c) is simpler to handle than (3.2.c). To see this, suppose we want to search for optimal methods in $\mathscr{C} = E_{s,p}$, using GPII. Then the pairs (L, M) of class $\bar{\mathscr{C}}$ must be specified by using the algebraic conditions for the order p. Similarly as in the original procedure (1.16), the order conditions, known in terms of K, would have to be rewritten in terms of L and M via complicated (and time consuming) routines; see, e.g., Spiteri & Ruuth (2002), Ruuth (2004) and references therein. Similar reformulations would have to be performed in case we were interested in methods with special structures of the matrix K, e.g., low-storage schemes or singly-diagonally-implicit schemes. When seen in this light, GPIII has an advantage over GPII because, in the former procedure, the order conditions (and special structures) can easily and directly be implemented in terms of K.

For completeness, we note that the above General Procedures I, I, III are also highly relevant to the important search for methods $K \in \mathscr{C}$ which are optimal with respect to $c(L, M)$ and $R(K)$ (rather than $c(L, |M|)$ and $R(|K|)$. When looking for such methods, one can simply apply the general procedures, with $\mathscr{C}$ replaced by $\mathscr{C}_+ = \{K \ : \ K \in \mathscr{C} \text{ and } K \geq 0\}$; because for any $K = (\kappa_{ij})$, with a negative entry κ_{ij}, we have $R(K) = c(L, M) = 0$ (see Theorem 2.5 and (2.11), (2.9)).

4 Illustrating our General Procedure III in a search for some optimal singly-diagonally-implicit Runge-Kutta methods

In the literature, much attention has been paid to a special class of implicit Runge-Kutta methods, the so-called *singly-diagonally-implicit Runge-Kutta* (SDIRK) methods, i.e. methods $K = (\kappa_{ij})$ with $\kappa_{ij} = 0$ $(j > i)$ and $\kappa_{11} \neq 0$, $\kappa_{ii} = \kappa_{11}$ $(2 \leq i \leq s)$. For a discussion of SDIRK methods, and their computational advantages over other (fully) implicit Runge-Kutta methods, see, e.g., Butcher (1987), Hairer, Nørsett & Wanner (1993), Hairer & Wanner (1996), Kværnø, Nørsett & Owren (1996) and the references therein.

In the present section, we shall illustrate our General Procedure III in a search for some optimal SDIRK methods. We shall denote by $S_{s,p}$ the class of all singly-diagonally-implicit s-stage Runge-Kutta methods $K = (\kappa_{ij})$ with order of accuracy at least p, such that $\kappa_{ii} > 0$ and sign indicators $\sigma_j = \pm 1$ exist satisfying (2.12). Clearly, for any $K \in S_{s,p}$, all σ_j must be equal to 1. Consequently, in line with Remark 2.3 (a) and Theorem 2.6, only the function F itself (and no additional $\tilde{F}$ as in (1.11)) would be needed when a method of class $S_{s,p}$ is applied in the situation

(1.1), (1.9). Clearly, for all $K \in S_{s,p}$ and $(L,M) \in \bar{S}_{s,p}$, we have $K \geq 0$, $M \geq 0$, so that $R(|K|) = R(K)$, $c(L,|M|) = c(L,M)$.

It is well known that the implicit Euler method $K = (\kappa_{ij})$, with $s = 1$, $\kappa_{1,1} = \kappa_{2,1} = 1$, has an order $p = 1$ and the (optimal) value $R(K) = \infty$; see, e.g., Kraaijevanger (1991, Lemma 4.5). Consequently, any search for optimal methods in $S_{s,p}$ with $p = 1$ is superfluous. Below we shall focus on computing optimal methods K in $S_{s,p}$ with $p = 2$, 3.

We applied GPIII to $\mathscr{C} = S_{s,p}$ for $s = 1, ..., 10$ and $p = 2, 3$, and we implemented it by using Matlab's Optimization Toolbox. In Table III.1 we have collected the maximal coefficients $c_{s,p} = \max\{c(L,M) : (L,M) \in \bar{S}_{s,p}\} = \max\{R(K) : K \in S_{s,p}\}$, which we obtained with this implementation of PGIII.

	$s=1$	$s=2$	$s=3$	$s=4$	$s=5$	$s=6$	$s=7$	$s=8$	$s=9$	$s=10$
$p=2$	2	4	6	8	10	12	14	16	18	20
$p=3$	-	2.7321	4.8284	6.8730	8.8990	10.9161	12.9282	14.9373	16.9443	18.9499

Table III.1: The maximal coefficients $c_{s,p} = c(L,M) = R(K)$ for generalized Shu-Osher representations (L,M) (in $\bar{S}_{s,p}$) and SDIRK methods K (in $S_{s,p}$).

The table clearly shows that, for given p, the stepsize coefficients $c_{s,p}$, corresponding to the optimal methods in $S_{s,p}$, become larger when s increases. A larger value of $c_{s,p}$ means that monotonicity preservation can be guaranteed under a milder stepsize restriction (1.10) (with $c = c_{s,p}$), but this does not automatically imply a better overall efficiency – because, e.g., also the computational labor per step should be taken into account – cf. Spiteri and Ruuth (2002, Section 3), Ferracina & Spijker (2004, Section 4.2) for related considerations.

By trial and error, we found explicit formulae for the optimal methods K, and corresponding values $R(K)$, which coincide, up to all computed decimal digits, to the values which we obtained numerically using GPIII. For the optimal methods $K = (\kappa_{ij})$, in $S_{s,2}$, we found the following explicit formulae:

$$R(K) = c_{s,2} = 2s, \quad \text{and} \quad \kappa_{ij} = \begin{cases} \frac{1}{2s} & \text{if } i = j,\ 1 \leq i \leq s, \\ \frac{1}{s} & \text{if } 1 \leq j < i \leq s+1, \\ 0 & \text{otherwise.} \end{cases} \tag{4.1}$$

For the optimal methods $K = (\kappa_{ij})$, in $S_{s,3}$, we found

$$R(K) = c_{s,3} = s-1+\sqrt{s^2-1}, \quad \kappa_{ij} = \begin{cases} \frac{1}{2}\left(1-\sqrt{\frac{s-1}{s+1}}\right) & \text{if } i = j,\ 1 \leq i \leq s, \\ \frac{1}{\sqrt{s^2-1}} & \text{if } 1 \leq j < i \leq s, \\ \frac{1}{s} & \text{if } i = s+1,\ 1 \leq j \leq s, \\ 0 & \text{otherwise.} \end{cases} \tag{4.2}$$

In the following, we shall refer to the SDIRK methods (4.1) and (4.2) as SDIRK$(s,2)$ and SDIRK$(s,3)$, respectively.

5 A numerical illustration

In this section, we shall give a simple numerical illustration to the material presented above. We shall focus on the TVD properties of the methods SDIRK(s,p) for $s = p-1,\ p,\ p+1$.

We will apply the methods in the numerical solution of the 1-dimensional Buckley-Leverett equation, defined by (1.4) with $\Phi(v) = \frac{3v^2}{3v^2+(1-v)^2}$; see, e.g., LeVeque (2002). We consider this equation for $0 \leq x \leq 1,\ 0 \leq t \leq 1/8$, with (periodic) boundary condition $u(0,t) = u(1,t)$ and initial condition

$$u(x,0) = \begin{cases} 0 & \text{for } 0 < x \leq \frac{1}{2}, \\ \frac{1}{2} & \text{for } \frac{1}{2} < x \leq 1. \end{cases}$$

We semi-discretize this Buckley-Leverett problem using a uniform grid with mesh-points $x_j = j\Delta x$, where $j = 1, ..., N$, $\Delta x = 1/N$ and $N = 100$. The partial differential equation is replaced by the system of ordinary differential equations

$$U_j'(t) = \frac{1}{\Delta x}\left(\Phi(U_{j-\frac{1}{2}}(t)) - \Phi(U_{j+\frac{1}{2}}(t))\right) \quad (j = 1, 2, ..., N),$$

where $U_j(t)$ is to approximate $u(x_j,t)$. Following Hundsdorfer & Verwer (2003, III, Section 1), we define

$$U_{j+\frac{1}{2}} = U_j + \frac{1}{2}\varphi(\theta_j)(U_{j+1} - U_j),$$

where $\varphi(\theta)$ is a (limiter) function due to Koren – see, loc. cit. – defined by

$$\varphi(\theta) = \max(0, \min(2, \frac{2}{3} + \frac{1}{3}\theta, 2\theta)),$$

and

$$\theta_j = \frac{U_j - U_{j-1}}{U_{j+1} - U_j}.$$

In line with the periodicity of the boundary condition, we use the convention $U_p = U_q$ if $p \equiv q \mod N$. We thus arrive at a system of $N = 100$ ordinary differential equations that can be written in the form $\frac{d}{dt}U(t) = F(U(t))$.

We define u_0 to be the vector in $\mathbb{R}^N$, $N = 100$, with components $u_{0,j} = 0$ (for $1 \leq j \leq 50$), $u_{0,j} = 1/2$ (for $51 \leq j \leq 100$). The resulting initial value problem, of the form (1.1), was integrated by the forward Euler method and by the SDIRK(s,p) methods mentioned above.

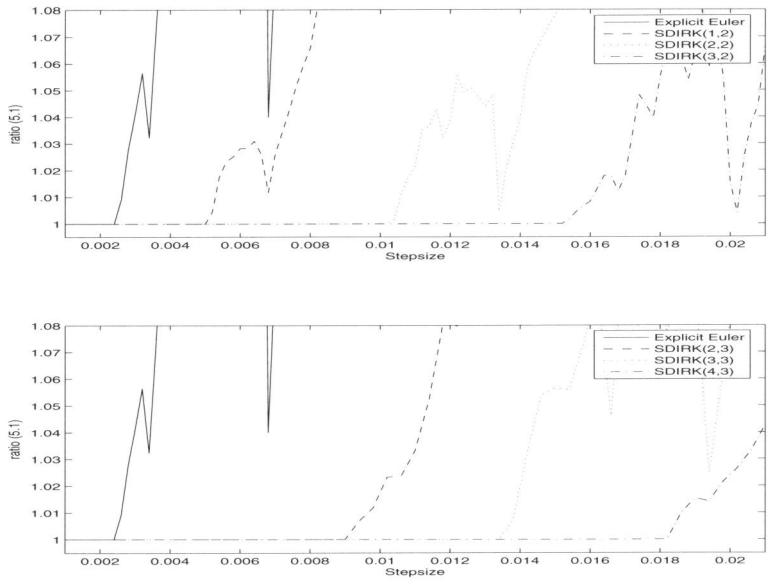

Figure III.1: The ratio (5.1) vs. the stepsize Δt.

In Figure III.1, the maximal ratio of the TV-seminorm $\|y\|_{TV} = \sum_{j=1}^{N} |\eta_j - \eta_{j-1}|$ (where $y = (\eta_1, ..., \eta_N)$, $\eta_0 = \eta_N$) of two consecutive numerical approximations, in the time interval $[0, \frac{1}{8}]$, is plotted as a function of the stepsize; i.e, the quantity

$$r(\Delta t) = \max \left\{ \frac{\|u_n\|_{TV}}{\|u_{n-1}\|_{TV}} : n \geq 1 \text{ with } n\Delta t \leq \frac{1}{8} \right\} \tag{5.1}$$

is plotted as a function of Δt. We note that in Figure III.1, the value $r(\Delta t) = 1$ corresponds to the monotonicity-preserving situation where $\|u_n\|_{TV} \leq \|u_{n-1}\|_{TV}$ for all $n \geq 1$, $n\Delta t \leq 1/8$.

We found that the Euler method is monotonic (TVD) for $0 < \Delta t \leq \tau \approx 0.0025$, and the SDIRK$(s, p)$ methods for $0 < \Delta t \leq \Delta t_{s,p}$, where $\Delta t_{1,2} \approx 0.0050$, $\Delta t_{2,2} \approx 0.0102$, $\Delta t_{3,2} \approx 0.0152$, $\Delta t_{2,3} \approx 0.0092$, $\Delta t_{3,3} \approx 0.0136$, $\Delta t_{4,3} \approx 0.0184$. Clearly, these numerically observed thresholds $\Delta t_{s,p}$ are amply larger than the threshold τ for the Euler method and, for given p, they increase when s increases. This can be viewed as a numerical reflection (and confirmation) of Remark 2.3 (a) (with all $\sigma_j = 1$) and of the fact that, in Table III.1, the coefficients $c_{s,p}$ satisfy: $1 < c_{s,p} < c_{s+1,p}$.

For $p = 2$, we see from the above that $\Delta t_{s,p}/\tau \approx c_{s,p} = 2s$. In this connection, it is interesting to note that the relation $\Delta t_{s,2} \geq s\, \Delta t_{1,2}$ follows directly from our formula (4.1) for SDIRK$(s, 2)$. In fact, from (4.1) we see that SDIRK$(s, 2)$ amounts

to applying SDIRK$(1,2)$ s times in succession, with Δt replaced by $\Delta t/s$.

6 Conjectures, open questions and final remarks

The optimal methods (4.1), (4.2) were obtained via a numerical search based on our General Procedure III. Clearly, this does not provide us with a formal proof of the optimality of these methods. Since the matrices K which we found numerically, correspond to (4.1), (4.2) up to all computed digits, we are naturally led to the following

Conjecture 6.1.

(a) Let $p = 2$ and $s \geq 1$. Then there is a unique method $K = (\kappa_{ij})$ in $S_{s,p}$ which is optimal with respect to $R(K)$, and this optimal method satisfies (4.1).

(b) Let $p = 3$ and $s \geq 2$. Then there is a unique method $K = (\kappa_{ij})$ in $S_{s,p}$ which is optimal with respect to $R(K)$, and this optimal method satisfies (4.2).

We can prove the conjecture in a straightforward way (only) for the special cases $(s,p) = (1,2)$, $(2,2)$ and $(s,p) = (2,3)$.

In fact, one easily sees that there is a unique SDIRK method $K = (\kappa_{ij})$ with $s = 1$ and $p = 2$, viz. the implicit midpoint rule, for which $\kappa_{1,1} = 1/2$, $\kappa_{2,1} = 1$, $R(K) = 2$. This proves Conjecture 6.1 (a) for the special case where $s = 1$. For the case $(s,p) = (2,2)$, a proof was given in Ferracina & Spijker (2005, Section 4.3).

Furthermore, there exist two different SDIRK methods $K = (\kappa_{ij})$ with $s = 2$ and $p = 3$, and explicit expressions for the coefficients κ_{ij} are available – see, e.g., Kværnø, Nørsett & Owren (1996, Table1). From these expressions, one easily sees that just one of the two methods belongs to $S_{2,3}$, and that it satisfies (4.2) with $s = 2$. This proves Conjecture 6.1 (b) for the special case where $s = 2$.

Let $\mathscr{C}$ denote the class of *all SDIRK methods* K, with s stages and order at least p. Clearly, the class $\mathscr{C}_+ = \{K : K \in \mathscr{C} \text{ and } K \geq 0\}$ equals $S_{s,p}$. In line with the last paragraph of Section 3, and under the assumption that Conjecture 6.1 is true, we thus can conclude that the methods SDIRK(s,p) with $p = 2$, 3 – i.e (4.1), (4.2), respectively – are optimal (with respect to $R(K)$) not only in $S_{s,p}$, but even in the wider class $\mathscr{C}$.

The numerical experiments in Section 5 support the idea that the (optimal) methods (4.1), (4.2) allow a stepsize Δt which is large, compared to τ_0, while maintaining monotonicity, notably the TVD property. Because we want to keep the present work sufficiently concise, we have not entered into the (related) question when, and in how far, these methods are actually more efficient than other (explicit) Runge-Kutta methods. Likewise, we have not discussed the application of GPIII to other classes than $S_{s,2}$ and $S_{s,3}$ – e.g. (for given s, p) the class of *all* Runge-Kutta methods $K = (\kappa_{ij})$, with s stages and order at least p, satisfying (2.12). We hope to come back to these interesting questions in future work.

Bibliography

[1] Butcher J. C. (1987): *The numerical analysis of ordinary differential equations. Runge Kutta and general linear methods.* A Wiley-Interscience Publication. John Wiley & Sons Ltd. (Chichester).

[2] 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.

[3] 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.

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

[5] 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.

[6] 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.

[7] Harten A. (1983): High resolution schemes for hyperbolic conservation laws. *J. Comput. Phys.*, 49 No. 3, 357–393.

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

[9] Higueras I. (2004): Strong stability for additive Rung-Kutta methods. Tech. rep., Departamento de Matemática e Informática, Universidad Pública de Navarra.

[10] Horn R. A., Johnson C. R. (1985): *Matrix analysis.* Cambridge University Press (Cambridge).

[11] Horváth Z. (1998): Positivity of Runge-Kutta and diagonally split Runge-Kutta methods. *Appl. Numer. Math.*, 28 No. 2-4, 309–326. Eighth Conference on the Numerical Treatment of Differential Equations (Alexisbad, 1997).

[12] Hundsdorfer W., Ruuth S. J. (2003): Monotonicity for time discretizations. Procs. Dundee Conference 2003. Eds. D.F. Griffiths, G.A. Watson, Report NA/217, Univ. of Dundee.

[13] 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).

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

[15] Kværnø A., Nørsett S. P., Owren B. (1996): Runge-Kutta research in Trondheim. *Appl. Numer. Math.*, 22 No. 1-3, 263–277. Special issue celebrating the centenary of Runge-Kutta methods.

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

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

[18] Ruuth S. J. (2004): Global optimization of explicit strong-stability-preserving Runge-Kutta methods. Tech. rep., Department of Mathematics Simon Fraser University.

[19] Ruuth S. J., Spiteri R. J. (2004): High-order strong-stability-preserving runge–kutta methods with downwind-biased spatial discretizations. *SIAM J. Numer. Anal.*, 42 No. 3, 974–996.

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

[21] 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).

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

[23] 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).

[24] Spiteri R. J., Ruuth S. J. (2003): Non-linear evolution using optimal fourth-order strong-stability-preserving Runge-Kutta methods. *Math. Comput. Simulation*, 62 No. 1-2, 125–135.

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:

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

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

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

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. $\diamond$

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. $\diamond$

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 $\rho > R(A,b)$. $\diamond$

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 ρ, $\sigma \in (0,+\infty)$ 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) $F: \mathbb{R}^{\infty}_{TV} \longrightarrow \mathbb{R}^{\infty}_{TV}$ 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) $$0 < \Delta t \leq R(A, b) \cdot \tau_0$$

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

(4.6) $\|u_n\|_{TV} \leq \|u_{n-1}\|_{TV} + \beta_0 \Delta t$, for u_n and u_{n-1} 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) $B \geq 0$,

(5.1.b) $I - tB$ is invertible for $t_0 \leq t \leq t_1$,

(5.1.c) $(I - t_0 B)^{-1} \geq 0$.

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

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:

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

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

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:

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

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

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 $1 \leq i \leq m$

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

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:

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

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

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.

Samenvatting

Dit proefschrift handelt over de numerieke oplossing van beginwaardeproblemen voor gewone differentiaalvergelijkingen.

Er is voornamelijk gekeken naar Runga-Kutta methoden die monotoon zijn. Dit betekent dat de seminorm van de numerieke benaderingen niet toeneemt in de tijd. Deze eigenschap is zeer belangrijk wanneer men gewone differentiaalvergelijkingen oplost die afkomstig zijn van een toepassing van de zogenaamde lijnenmethode op tijdsafhankelijke partiële differentiaalvergelijkingen.

In de literatuur zijn representaties voor *speciale* Runge-Kutta methoden geïntroduceerd, waarmee het bewijzen van de monotoniciteitseigenschap, in de situatie waarin de Euler-methode zelf monotoon is, wordt vergemakkelijkt. Deze representaties leiden tot voorwaarden voor de stapgrootte die *voldoende* zijn voor de monotoniciteitseigenschap.

Dit proefschrift bevat een inleiding en vier in wetenschappelijke tijdschriften gepubliceerde of nog te publiceren artikelen.

De inleiding is geschreven met de bedoeling ook begrijpelijk te zijn voor lezers die niet gespecialiseerd zijn in het betreffende vakgebied.

In het eerste artikel wordt een theorie omtrent algemene Runge-Kutta methoden voorgesteld die leidt tot voorwaarden voor de stapgrootte die niet alleen voldoende maar ook *noodzakelijk* zijn voor de monotoniciteitseigenschap.

In het tweede artikel wordt een simpele en algemene manier van aanpak gegeven waarmee voor elke gegeven Runge-Kutta methode een *best mogelijke representatie* gevonden kan worden met betrekking tot de stapgroottevoorwaarden die daaruit afgeleid kunnen worden.

In het derde artikel wordt een *numerieke procedure* geïntroduceerd voor het vinden van *optimale Runge-Kutta methoden* (met betrekking tot de stapgroottevoorwaarden voor monotoniciteit).

Het vierde artikel bevat een algemene theorie, omtrent de voorwaarden voor de stapgrootte, die een grotere klasse van semidiscrete benaderingen omvat dan tot dusver beschouwd in de literatuur. Deze theorie geeft stapgroottevoorwaarden die in het bijzonder een *begrensdheidseigenschap* impliceren voor algemene Runge-Kutta methoden.

Curriculum Vitæ

Luca Ferracina was born in Vicenza, Italy, on January 11, 1973. After completing his primary studies, he attended the Scientific High School "Seminario Vescovile di Vicenza" from 1988 to 1992. In 1993 he started his studies in mathematics at the "Università degli Studi di Padova": he graduated in 1999. From 1999 to 2000 he was part-time member (during the period of civil service) of the numerical analysis research group at the "Dipartimento di Matematica, Università degli Studi di Padova". In the same institute, in 2000, he started a PhD in mathematics and in 2001–2002 he was a Visiting Fellow at the Mathematical Institute of Leiden University working with Prof.dr. M.N. Spijker. From 2002 to 2005 he was employed as Assistent in Opleiding (AiO) at the Mathematical Institute of Leiden University. During that period he did his PhD research under the supervision of Prof.dr. M.N. Spijker and Prof.dr. J.G. Verwer that is described in this thesis.

Stellingen

behorende bij het proefschrift

Monotonicity and Boundedness in general Runge-Kutta methods

van

Luca Ferracina

In the following propositions we deal with the numerical solution of initial value problems for sytems of ordinary differential equations that can be written in the form

$$\frac{d}{dt}U(t) = F(U(t)) \quad (t \geq 0), \quad U(0) = u_0. \tag{1}$$

We focus on the (irreducible) Runge-Kutta method (2) where, given an approximation u_{n-1} of $U(t_{n-1})$, a new approximation u_n of $U(t_{n-1} + \Delta t)$ is computed by the relations

$$\begin{cases} y_i = u_{n-1} + \Delta t \sum_{j=1}^{s} \kappa_{ij} F(y_j) \quad (1 \leq i \leq s+1), \\ u_n = y_{s+1}. \end{cases} \tag{2}$$

We identify the Runge-Kutta method with the $(s+1) \times s$ matrix $K = (\kappa_{ij})$ and we denote with K_0 the $s \times s$ submatrix $K_0 = (\kappa_{ij})$, $1 \leq i \leq j \leq s$. We are interested in coefficients c such that

$$\left.\begin{array}{c} \|v + \tau_0 F(v)\| \leq \|v\| \quad (\forall\, v \in \mathbb{V}) \\ 0 < \Delta t \leq c \cdot \tau_0 \end{array}\right\} \Rightarrow \ \|u_n\| \leq \|u_{n-1}\|. \tag{3}$$

Let $L = (\lambda_{ij})$ be any $(s+1) \times s$ matrix with submatrix $L_0 = (\lambda_{ij})$, $1 \leq i \leq j \leq s$ such that $L \geq 0$, $Le_s \leq e_{s+1}$ and $I - L_0$ is invertible – here, and in the following, I denotes the $s \times s$ identity matrix and $e_m \in \mathbb{R}^m$ stands for the column vector with all components equal to 1 (for $m = s,\ s+1$). Define the $(s+1) \times s$ matrix $M = (\mu_{ij})$ by $M = K - LK_0$ and consider the process

$$\begin{cases} y_i = \left(1 - \sum_{j=1}^{s} \lambda_{ij}\right) u_{n-1} + \sum_{j=1}^{s} [\lambda_{ij}\, y_j + \Delta t \cdot \mu_{ij} F(y_j)] \quad (1 \leq i \leq s+1), \\ u_n = y_{s+1}. \end{cases} \tag{4}$$

We identify the above process with the pair (L, M). In view of Statement 1, to be given below, we will refer to process (4) as an (L, M) representation of method K.

We define the following coefficient

$$c(L,M) = \min\{\gamma_{ij} : 1 \le i \le s+1,\ 1 \le j \le s\}, \quad \gamma_{ij} = \begin{cases} \lambda_{ij}/\mu_{ij} & \text{if } \mu_{ij} > 0, \\ \infty & \text{if } \mu_{ij} = 0, \\ 0 & \text{if } \mu_{ij} < 0. \end{cases}$$

1. Process (4) is a useful representation of process (2).
The following two statements are valid.

(i) Method (2) and process (4) are equivalent.

(ii) Let c be equal to $c(L,M)$ defined above. Then implication (3) holds whenever $\mathbb{V}$ is a real vector space with seminorm $\|.\|$, and u_n, u_{n-1} are related to each other as in (4).

See Chapters I and II of this thesis.

Given a Runge-Kutta method K, consider, for real γ, the following conditions:

(5) $\qquad (I+\gamma K_0)$ is invertible, $\quad \gamma K(I+\gamma K_0)^{-1} \ge 0, \quad \gamma K(I+\gamma K_0)^{-1} e_s \le e_{s+1}.$

We define the following coefficient

$$R(K) = \sup\{\gamma \ :\ \gamma \ge 0 \text{ and (5) holds}\}.$$

2. The largest c guaranteeing (3) for methods (2).
Let c be given with $0 < c \le \infty$. Then (I) and (II) are equivalent:

(I) $c \le R(K)$,

(II) implication (3) holds whenever $\mathbb{V}$ is a real vector space with seminorm $\|.\|$, and u_n, u_{n-1} are related to each other as in (2).

See Chapters I and II of this thesis.

3. Optimal (L,M) representations.
For any Runge-Kutta method K there exist an (L,M) representation with $c(L,M) = R(K)$.
See Chapter II of this thesis.

4. The optimal (L,M) representation is not unique.
The (L,M) representation mentioned in Statement 3 is in general not unique.

5. Optimal Runge-Kutta methods.
Let $\mathscr{C}$ be a given class of Runge-Kutta methods K such that $c^* = \max\{R(K) : K \in \mathscr{C}\}$ exists and is finite. We denote by $\bar{\mathscr{C}}$ the set of all (L, M) representaions of methods $K \in \mathscr{C}$. Then the following two statements are valid.

(i) The maximum of γ, specified in the following two procedures, exists and equals c^*.

(ii) The first procedure is, from a practical point of view, to be preferred over the second one.

Procedure 1	maximize γ,	subject to: γ satisfies (5) and $K \in \mathscr{C}$.
Procedure 2	maximize γ,	subject to: $L - \gamma M \geq 0$ and $(L, M) \in \bar{\mathscr{C}}$.

See Chapter III of this thesis.

6. Completing results in the literature.
In the literature, optimal (w.r.t. $R(K)$) explicit Runge-Kutta methods, with s stages and order of accuracy at least p, are available with $1 \leq p \leq 4$ and $p \leq s \leq 9$, except the case $(s, p) = (9, 4)$. It can be shown that the missing optimal method K has $R(K) = 4.9142$ (rounded to 5 decimal digits).

7. Boundedness.
Statement 2 can be generalized so as to become valid also when (3) is replaced by the following implication

$$\left.\begin{array}{l} \|v + \tau_0 F(v)\| \leq (1 + \alpha_0\tau_0)\|v\| + \beta_0\tau_0 \quad (\forall\, v \in \mathbb{V}) \\ 0 < \Delta t \leq c \cdot \tau_0 \end{array}\right\} \Rightarrow \ \|u_n\| \leq (1 + \alpha\Delta t)\|u_{n-1}\| + \beta\Delta t.$$

See Chapter IV of this thesis.

8. TVD does not avoid oscillations.
When dealing with numerical solutions of IVPs for ODEs (and PDEs), one should keep in mind the following remark.

"(...) some people believe that the TVD property (i.e. $\|u_n\|_{TV} \leq \|u_{n-1}\|_{TV}$ with $\|.\|_{TV}$ total variation seminorm) completely eliminates all spurious oscillations for all (Δx and) Δt. It does not. In fact, the TVD condition may allow large spurious oscillations (...)"

See C.B. Laney (1998), *Computational Gasdynamics*, Chapter 16.

In the following two statements we denote by $S_{s,p}$ the class of all singly-diagonally-implicit s-stage Runge-Kutta methods $K = (\kappa_{ij})$, with order of accuracy at least p and with all $\kappa_{ij} \geq 0$, $\kappa_{ii} > 0$.

9. Upper bound for the order of accuracy.
There are no methods, with $R(K) > 0$, in $S_{s,p}$ if $p > 4$.

10. Optimal method in $S_{3,4}$.
Consider the following Runge-Kutta method

$$K_{3,4} = \begin{pmatrix} \frac{1+\xi}{2} & 0 & 0 \\ -\frac{\xi}{2} & \frac{1+\xi}{2} & 0 \\ 1+\xi & -1-2\xi & \frac{1+\xi}{2} \\ \frac{1}{6\xi^2} & 1-\frac{1}{3\xi^2} & \frac{1}{6\xi^2} \end{pmatrix} \quad \text{with} \quad \xi = -\frac{2}{\sqrt{3}}\cos(\frac{5\pi}{18}).$$

Then $K_{3,4} \in S_{3,4}$, and for any other $K \in S_{3,4}$ we have $R(K) < R(K_{3,4}) = 2\,\dfrac{1+\xi}{\xi^2-\xi-1}$.

11. A model for studying the dispersion in the Venice Lagoon

This is a mesh with 1967 nodes and 3423 triangular elements modelling the Venice Lagoon. Discretizing in space the advection-diffusion equation

$$\frac{\partial u}{\partial t} + v \cdot \nabla u = \nabla \cdot (\mathbf{K} \cdot \nabla u) + s$$

with the finite element method (linear triangular elements), one obtains a semi-discrete system of ordinary differential equations that can be written as

$$\mathbf{M}\frac{d}{dt}U(t) + \mathbf{N}(t)U(t) + l(t) = 0.$$

Consider the simple time-discretization (fully-discrete system)

$$\mathbf{M}\frac{u_{n+1}-u_n}{\Delta t} + \mathbf{N}_n u_n + \theta[\mathbf{N}_{n+1}u_{n+1} - \mathbf{N}_n u_n] + l_n + \theta(l_{n+1} - l_n) = 0.$$

We then obtain a linear system (in $\mathbb{R}^m$, $m = 1967$), of the form

$$\mathbf{A}u_{n+1} = b,$$

that has to be solved at each time level. Because of the special structure of the above matrix $\mathbf{A}$ (and because of the limited memory space needed for storing it), the above system can be solved (repeatedly) 720 times, on a *IBM RISC 6000* workstation, in less than 200 seconds.