Computing optimal monotonicity-preserving Runge-Kutta methods

L. Ferracina* and M. N. Spijker†

April 7, 2005

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.

Key words: initial value problem, Shu-Osher representation, total-variation-diminishing (TVD), monotonicity, strong-stability-preserving (SSP), singly-diagonally-implicit Runge-Kutta formula (SDIRK).

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.

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

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

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

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

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

Here κ_{ij} are real parameters, specifying the Runge-Kutta method, and y_i $(1 \le i \le s)$ 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 $1 \le i \le j \le s$), 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 $\|\lambda v + (1-\lambda)w\| \le \lambda\|v\| + (1-\lambda)\|w\|$ for all $v, w \in \mathbb{V}$ and $0 \le \lambda \le 1$). 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

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

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 $t \ge 0$, $-\infty < x < \infty$. 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

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

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 $\tau_0 > 0$,

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

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 $0 \leq \tau < \tau_0$, condition (1.9) implies $\|v + \tau F(v)\| \leq \|(\tau/\tau_0)(v + \tau_0 F(v)) + (1 - \tau/\tau_0)v\| \leq \|v\|$ – i.e. the Euler process is still monotonic with any stepsize $\tau \in [0, \tau_0)$.

Assume (1.9). Then, for $2 \leq i \leq s+1$, 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

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

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

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))

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

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

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

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

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

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 as 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) $\lambda_{ij} - \gamma\, |\mu_{ij}| \ge 0 \quad (1 \le j < i \le s+1)$;

(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:

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

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

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 $1 \leq i \leq j \leq s$), $\sum_{j=1}^{s} \lambda_{ij} = 1$ (for $2 \leq i \leq s+1$).

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

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

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

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

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

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:

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

$$M=K-LK_0, \tag{2.5}$$

$$I-L_0 \text{ is invertible.} \tag{2.6}$$

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):

$$L\ge 0,\quad Le_s\le e_{s+1}. \tag{2.7}$$

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

$$\mu_{ij}\ge 0 \ \text{ (for } 1\le i\le s+1 \text{ and } \sigma_j=1),\quad \mu_{ij}\le 0 \ \text{ (for } 1\le i\le s+1 \text{ and } \sigma_j=-1). \tag{2.8}$$

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

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

and we put

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

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:

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

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) $\kappa_{ij} \geq 0$ (for $1 \leq i \leq s{+}1$ and $\sigma_j = 1$), $\kappa_{ij} \leq 0$ (for $1 \leq i \leq s{+}1$ 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 (3.4), 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 (3.5) 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 (3.5), 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 $|M_0|=M_0D=(K_0-L_0K_0)D=(I-L_0)K_0D$, i.e. $K_0D=(I-L_0)^{-1}|M_0|$. 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 $K_0D\geq|M_0|\geq 0$. Consequently, $K_0D=|K_0|$ 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 $\sigma_j = \pm 1\,(1 \leq j \leq s)$ 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

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

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

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

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 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 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:

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

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

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

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 $0 \le x \le 1,\ 0 \le t \le 1/8$, with (periodic) boundary condition $u(0,t) = u(1,t)$ and initial condition

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

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 $1 \le j \le 50$), $u_{0,j} = 1/2$ (for $51 \le j \le 100$). 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.

In Figure 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

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

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

We found that the Euler method is monotonic (TVD) for $0 < \Delta t \le \tau \approx 0.0025$, and the SDIRK$(s,p)$ methods for $0 < \Delta t \le \Delta t_{s,p}$, 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 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 $\Delta t_{s,2} \ge s\,\Delta t_{1,2}$ 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 $\Delta t/s$.

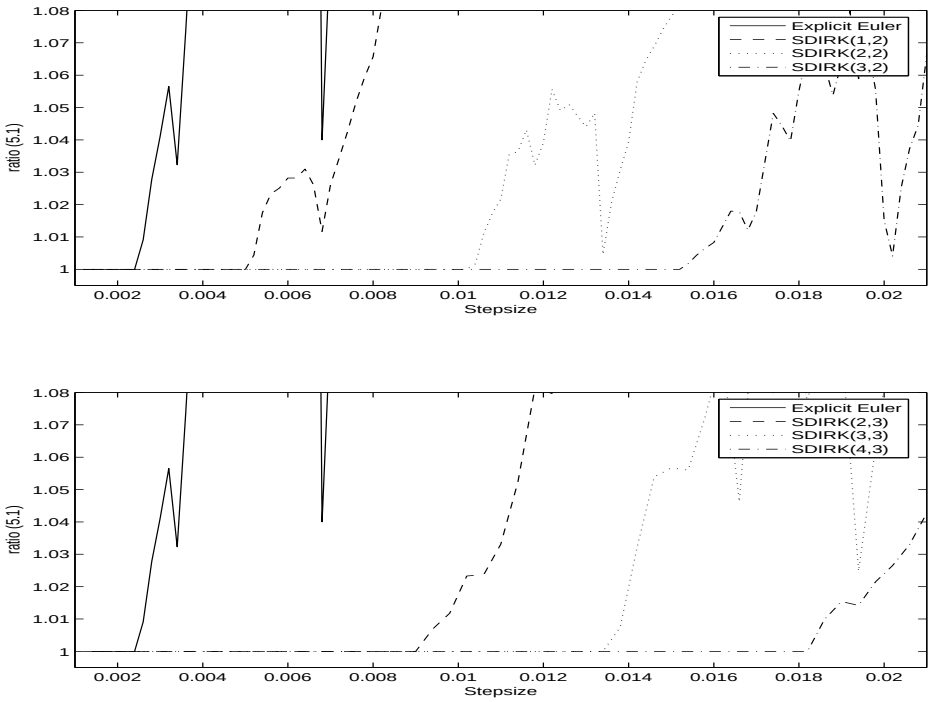

Figure 1: The ratio (5.1) vs. the stepsize Δt.

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.

References

[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. H., Ruuth S. J. (2003): Monotonicity for time discretizations. Procs. Dundee Conference 2003, pp. 85-94. Eds. D.F. Griffiths, G.A. Watson, Report NA/217, Univ. of Dundee.

[13] Hundsdorfer W. H., 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.