Strong Stability of Singly-Diagonally-Implicit Runge-Kutta Methods

L. Ferracina* and M. N. Spijker†

2007, June 4

Abstract. This paper deals with the numerical solution of initial value problems, for systems of ordinary differential equations, by Runge-Kutta methods (RKMs) with special nonlinear stability properties indicated by the terms total-variation-diminishing, strongly stable and monotonic. Stepsize conditions, guaranteeing these properties, were studied earlier, see e.g. Shu & Osher (1988), Gottlieb, Shu & Tadmor (2001), Shu (2002), Hundsdorfer & Ruuth (2003), Higueras (2004, 2005), Gottlieb (2005), Ferracina & Spijker (2004, 2005).

Special attention was paid to RKMs which are optimal, in that the corresponding stepsize conditions are as little restrictive as possible within a given class of methods. Extensive searches for such optimal methods were made in the class of explicit RKMs with a prescribed number of stages s and order of accuracy p, see e.g. Gottlieb & Shu (1998), Spiteri & Ruuth (2002, 2003), Ruuth (2006).

In the present paper we focus on the interesting class of singly-diagonally-implicit Runge-Kutta methods, with s stages and order p. We determine methods that are optimal in the above sense, within this class.

Key words. initial value problem, method of lines (MOL), ordinary differential equation (ODE), singly-diagonally-implicit Runge-Kutta method (SDIRK), total-variation-diminishing (TVD), strong-stability-preserving (SSP), monotonicity.

AMS subject classifications. 65L05, 65L06, 65L20, 65M20.

1 Introduction

1.1 Purpose of the paper

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

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

The general Runge-Kutta method (RKM), applied to problem (1.1), can provide 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, 2003), Hairer, Nørsett & Wanner (1987), 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}.

*CWI, P.O. Box 94079, NL-1090-GB Amsterdam, Nederland. Email: Luca.Ferracina@cwi.nl

†Math. Inst., Leiden Univ., P.O. Box 9512, NL-2300-RA Leiden, Nederland. Email: spijker@math.leidenuniv.nl

In the following, $\mathbb{V}$ stands for the vector space on which the differential equation is defined, and $\|\cdot\|$ denotes a convex function on $\mathbb{V}$ (i.e.: $\|\lambda v + (1-\lambda)w\| \leq \lambda\|v\| + (1-\lambda)\|w\|$ for $0 \leq \lambda \leq 1$ and $v, w \in \mathbb{V}$). Much attention has been paid in the literature to the property

$$\|y_i\| \leq \|u_{n-1}\| \quad (\text{for } 1 \leq i \leq s+1). \tag{1.3}$$

Clearly, (1.3) implies $\|u_n\| \leq \|u_{n-1}\|$. The latter property, as well as property (1.3), is often referred to by the term *strong stability* or *monotonicity*; it is of particular importance in situations where (1.1) results from (method of lines) semidiscretizations of time-dependent partial differential equations. Choices for $\|\cdot\|$ which occur in that context, include e.g. the *supremum norm* $\|x\| = \|x\|_\infty = \sup_i |\xi_i|$ and the *total variation seminorm* $\|x\| = \|x\|_{TV} = \sum_i |\xi_{i+1} - \xi_i|$ (for vectors x with components ξ_i).

Numerical processes, satisfying $\|u_n\|_{TV} \leq \|u_{n-1}\|_{TV}$, play a special role in the solution of nonlinear hyperbolic differential equations and are called *total-variation-diminishing* (TVD), cf. e.g. Harten (1983), Shu (1988), Shu & Osher (1988), LeVeque (2002), Hundsdorfer & Verwer (2003).

We note that, for practical calculations, special importance has been attached, by various authors, to the inequality $\|y_i\| \leq \|u_{n-1}\|$ being fulfilled for *all* i with $1 \leq i \leq s+1$ (rather than just for $i = s+1$) - see e.g. Shu (2002), Gottlieb (2005).

In the literature, conditions on Δt which guarantee (1.3) were given in the situation where, for given $\tau_0 > 0$,

$$\|v + \tau_0 F(v)\| \leq \|v\| \quad (\text{for all } v \in \mathbb{V}). \tag{1.4}$$

Assumption (1.4) means that the explicit Euler method, with stepsize τ_0, is strongly stable. It can be interpreted as a condition on the manner in which the semidiscretization is performed, in case $\frac{d}{dt}U(t) = F(U(t))$ stands for a semidiscrete version of a partial differential equation.

In the literature, *stepsize-coefficients* c were determined such that strong stability, in the sense of (1.3), is present for all Δt with

$$0 < \Delta t \leq c \cdot \tau_0. \tag{1.5}$$

For explicit RKMs, this was done by rewriting the right-hand members of (1.2.a) as convex combinations of explicit Euler steps - see e.g. Shu & Osher (1988), Gottlieb & Shu (1998), Shu (2002). For more general RKMs, stepsize-coefficients were obtained e.g. in Gottlieb, Shu & Tadmor (2001), Ketcheson (2004), Higueras (2004, 2005), Ferracina & Spijker (2004, 2005). We note that, in the context of discretizations for hyperbolic differential equations, the above coefficients c are sometimes called *CFL coefficients*, see e.g. Gottlieb & Shu (1998), Shu (2002).

Clearly, the larger c, the less restrictive is condition (1.5). For any given method, the maximal stepsize-coefficient c, with the property that (1.4), (1.5) still imply (1.3), is thus an important *characteristic coefficient* of the method. When comparing the computational efficiency of different methods, it is natural to take these characteristic coefficients into account.

In order to single out efficient RKMs, with a given number of stages s and order of accuracy p, much attention has been paid to the interesting problem of optimizing, over given classes of RKMs, the *special* stepsize-coefficients obtainable via convex combinations of Euler steps, see e.g. Shu (1988), Shu & Osher (1988), Gottlieb & Shu (1998), Gottlieb, Shu & Tadmor (2001), Spiteri & Ruuth (2002, 2003), Gottlieb (2005), Ruuth (2006). The focus in these papers is on RKMs which are *explicit*.

Not many results seem to be available about the problem of optimizing the above mentioned *characteristic* coefficients, over given classes of *implicit* RKMs.

Since the beginning of the seventies, much attention has been paid to solving (stiff) ODEs with a special type of implicit Runge-Kutta methods, viz. the so-called *singly-diagonally-implicit Runge-Kutta* (SDIRK) methods. These methods are characterized by matrices $K = (\kappa_{ij})$ with

$\kappa_{ij} = 0$ $(j > i)$ and $\kappa_{ii} = \kappa_{11} \neq 0$ $(2 \leq i \leq s)$. Thanks to this property of the coefficients κ_{ij}, the s equations (1.2.a) for y_i can be solved in s successive stages, rather than in one (very costly) stage involving all y_i. Moreover, if Newton-type iterations are used in the s stages, one may hope to use repeatedly the stored LU-factorization of a *single* coefficient matrix $I - \kappa_{11}\,\Delta t\, F'(v)$.

There exists a vast literature dealing with SDIRK methods, see e.g. Norsett (1974), Crouzeix (1975), Alexander (1977), Kværnø, Nørsett & Owren (1996), Butcher (1987), Hairer, Nørsett & Wanner (1987), Hairer & Wanner (1996), Dekker & Verwer (1984) and the references therein. We note that SDIRK methods have not only been considered in their own right, but also as implicit part of implicit-explicit (IMEX) Runge-Kutta methods; see e.g. Asher, Ruuth & Spiteri (1997), Calvo, Frutos & Novo (2001), Pareschi & Russo (2005).

The purpose of the present paper is to analyse the strong stability properties of SDIRK methods. We shall focus on the problem of determining methods which are optimal, with respect to the characteristic coefficients discussed above, within the class of s-stage SDIRK methods with given order of accuracy p.

1.2 Scope of the paper

Various theorems have been given, in the literature, which specify the above mentioned characteristic coefficient, for a given RKM, in terms of the matrix K, see Higueras (2004), Ketchenson (2004), Ferracina & Spijker (2004, 2005), Spijker (2007). In Section 2 we give a concise review of some of these results. Corollary 2.5 states that, for all SDIRK methods, the characteristic coefficient is equal to the famous coefficient $R(A, b)$, which was introduced by Kraaijevanger (1991).

Section 3 contains the main results of the paper. We use the material of Section 2 in a systematic search for optimal methods within the class of SDIRK methods with s stages and order of accuracy p. In Sections 3.1 - 3.4 we give optimal methods of orders 1 - 4, whereas Section 3.5 deals with the case $p > 4$. Section 3.6 gives a summary and discussion of the optimal SDIRK methods found in the preceding sections.

In Section 4, we report shortly on a numerical experiment involving the TVD properties of various optimal SDIRK methods given in Section 3. We apply the methods to a nonlinear hyperbolic test equation, the 1-dimensional Buckley-Leverett equation.

2 Strong stability and Kraaijevanger's coefficient

2.1 Kraaijevanger's coefficient $r(K)$

Consider a given s-stage Runge-Kutta method (1.2), with $(s+1) \times s$ coefficient matrix $K = (\kappa_{ij})$. We denote the $s \times s$ matrix obtained from K by omitting its last row by K_s.

Below, in Section 2.2, we shall relate the characteristic coefficient of the method, discussed in Section 1.1, to an important coefficient introduced by Kraaijevanger (1991). In the present Section 2.1, we define the latter coefficient and list some of its properties.

The definition of Kraaijevanger's coefficient involves the following condition, in which γ denotes a real variable:

$$(2.1) \qquad (I + \gamma K_s) \text{ is invertible}, \quad \gamma K (I + \gamma K_s)^{-1} \geq 0, \quad \gamma K (I + \gamma K_s)^{-1} E_s \leq E_{s+1}.$$

Here I denotes the $s \times s$ identity matrix, and E_s, E_{s+1}, respectively, stand for the $s \times 1$ and the $(s+1) \times 1$ matrix, with all entries equal to 1. The inequalities in (2.1) should be interpreted entry-wise; all inequalities for matrices occurring below are to be interpreted in the same way.

Definition 2.1 (Kraaijevanger's coefficient $r(K)$).

$$r(K) = \sup\{\gamma \ : \ \gamma \geq 0 \text{ and } (2.1) \text{ holds}\}.$$

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

$$r(K) = \sup\{r \ : \ r \in \mathbb{R} \text{ and (2.1) holds for all } \gamma \in [0, r]\}.$$

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 2.2 (Fulfillment of condition (2.1)). *Let K be given and let γ be any finite value with $0 \leq \gamma \leq r(K)$ (Definition (2.1)). Then γ satisfies (2.1).*

Theorem 2.2 can be viewed as a (somewhat stronger) version of earlier results about $r(K)$ in the literature – for related material, see Kraaijevanger (1991, Lemma 4.4), Higueras (2006, Proposition 2.11), Horváth (1998, Theorem 4). Theorem 2.2 follows easily from a (more general) theorem given in Spijker (2007, Theorem 2.2 (ii) and Section 3.2.1).

Clearly, we have always $r(K) \geq 0$. The subsequent theorem makes it quite easy to determine whether $r(K) > 0$ or $r(K) = 0$.

Theorem 2.3 (Criterion for positivity of $r(K)$).
We have $r(K) > 0$ if and only if: $K \geq 0$ and, for all i, j, the following implication is valid

$$\kappa_{i\,j} = 0 \implies \kappa_{i\,m}\,\kappa_{m\,j} = 0 \text{ (for } 1 \leq m \leq s). \tag{2.2}$$

Proof. By Theorem 2.2, we have $r(K) > 0$ if and only if there is a $\gamma_0 > 0$ such that (2.1) holds for all $\gamma \in [0, \gamma_0]$. Therefore, in view of the first inequality in (2.1), we can assume with no loss of generality, that $K \geq 0$.

For $\gamma > 0$ sufficiently small, the matrix $(I + \gamma K_s)$ is invertible and the second inequality in (2.1) is automatically fulfilled. Therefore, we have $r(K) > 0$ if and only if there is a $\gamma_0 > 0$ such that

$$(K - \gamma K K_s)\Big\{\sum_{j=0}^{\infty}(\gamma K_s)^{2j}\Big\} = K\,(I + \gamma K_s)^{-1} \geq 0 \quad \text{(for all } \gamma \in [0, \gamma_0]),$$

which is equivalent to the requirement that $\kappa_{ij} > 0$ as soon as the entry in the i-th row and j-th column of the matrix KK_s is positive. This last requirement is equivalent to (2.2) (because $K \geq 0$). □

For theorems closely related to Theorem 2.3, see e.g. Kraaijevanger (1991, Theorem 4.2), Higueras (2005, Proposition 2.4), Spijker (2007, Theorem 2.2 (i)).

2.2 Relating the characteristic stepsize-coefficient to $r(K)$

We consider stepsize-coefficients c such that, for method (1.2), the following property is present:

(2.3) Condition $0 < \Delta t \leq c \cdot \tau_0$ implies strong stability, in the sense of (1.3), whenever $\mathbb{V}$ is a vector space, $\|\cdot\|$ a convex function on $\mathbb{V}$, and $F : \mathbb{V} \to \mathbb{V}$ satisfies (1.4).

It is easily verified that this property is independent of the valueτ_0: if c has property (2.3) using one particular value $\tau_0 > 0$, then c will have the same property when using any other value, say $\tau_0' > 0$.

Clearly, if $\widehat{c}$ is the *maximal stepsize-coefficient* c with property (2.3), then $\widehat{c}$ equals the *characteristic coefficient* of the RKM, discussed in Section 1.1.

Theorem 2.4 (Strong Stability of RKMs). *Consider method (1.2), and given $\tau_o > 0$.*
(i) Let $c \leq r(K)$. Then statement (2.3) is valid.
(ii) Assume method (1.2) is irreducible in the sense that the rows of the $s \times s$ matrix K_s are different from each other. Then, conversely, statement (2.3) implies that $c \leq r(K)$.

Note that any given RKM, violating the irreducibility assumption in (*ii*), produces approximations u_n which can also be obtained from an irreducible method, with a smaller number of stages.

The theorem highlights the importance of $r(K)$ for (irreducible) RKMs: it implies that for RKMs which are irreducible in the sense of (*ii*), *the characteristic coefficient equals* $r(K)$.

Statements (*i*), (*ii*) supplement related material in Higueras (2004, 2005), Ketcheson (2004), Spijker & Ferracina (2004, 2005). The irreducibility condition in (*ii*) is essentially weaker than in these papers, whereas property (2.3) is stronger than in (some of) the papers. Theorem 2.4 follows easily from (more general) results given in Spijker (2007, Sections 3.1 and 3.2.1).

Since SDIRK methods automatically satisfy the irreducibility requirement occurring in the above statement (*ii*), the above theorem yields immediately the following

Corollary 2.5 (Strong Stability of SDIRK methods). *Consider an arbitrary SDIRK method with coefficient matrix $K = (\kappa_{ij})$. Then the largest value c, for which (2.3) holds, is equal to $r(K)$; i.e. the characteristic coefficient of the method is equal to Kraaijevanger's coefficient.*

3 Optimal SDIRK methods

In the following, we denote by $S_{s,p}$ the class of all s-stage SDIRK methods with order of accuracy at least p. We consider the problem of determining a method in the class which is *optimal*, in that it has property (2.3) with a value c which is maximal in $S_{s,p}$. We identify RKMs with their coefficient matrices $K = (\kappa_{i,j})$, and denote the corresponding characteristic coefficients, discussed in Sections 1.1, 2.2, by $c(K)$. The following three remarks are basic for the rest of Section 3.

(i) For any SDIRK method K, the computation of the characteristic coefficient $c(K)$ can be based on the formula $c(K) = r(K)$ – see the above Corollary 2.5.

(ii) For any SDIRK method K of order $p > 1$, we have $c(K) < \infty$ – see Spijker (1983), and e.g. Kraaijevanger (1991, Theorem 8.3.).

(iii) The order of an s-stage SDIRK method cannot exceed $s + 1$ – see Nørsett (1974), and e.g. Nørsett & Wolfbrandt (1977), Dekker & Verwer (1984, Theorem 3.5.11).

3.1 Optimal SDIRK methods of order $p = 1$

It is well known that the implicit Euler method $K = (\kappa_{ij})$, specified by $s = 1$ and $\kappa_{1,1} = \kappa_{2,1} = 1$, has order $p = 1$ and characteristic coefficient $c(K) = r(K) = \infty$; see e.g. Kraaijevanger (1991, Lemma 4.5). Consequently, a search for optimal methods in $S_{s,p}$ with $p = 1$ is simple: the s-stage SDIRK method consisting of s consecutive applications of the implicit Euler method, with timestep $\Delta t/s$, has order 1 and $c(K) = \infty$. The coefficients of this SDIRK method satisfy $\kappa_{ij} = 1/s$ (for $1 \leq j \leq i \leq s$ and for $i = s + 1,\ 1 \leq j \leq s$).

3.2 Optimal SDIRK methods of order $p = 2$

Optimal method with $s = 1$ stage

It is well known that there is a unique SDIRK method $K = (\kappa_{ij})$ in $S_{1,2}$, viz. the implicit midpoint rule. For this method we have $\kappa_{1,1} = 1/2$, $\kappa_{2,1} = 1$, and one easily sees that $c(K) = 2$.

Optimal method with $s = 2$ stages

It can be proved that, for the case $(s, p) = (2, 2)$, the method K with $\kappa_{1,1} = \kappa_{2,2} = 1/4$; $\kappa_{2,1} = \kappa_{3,1} = \kappa_{3,2} = 1/2$ is optimal in $S_{s,p}$ with regard to its value $c(K)$, see Ferracina & Spijker (2005, Section 4.3). The corresponding characteristic coefficient equals $c(K) = 4$.

Optimal methods with $s \geq 3$ stages

In view of the equality $c(K) = r(K)$, we consider the maximization of $r(K)$ over the classes $S_{s,p}$, with $p = 2$ and given $s \geq 3$. According to Definition 2.1, this maximum can be computed by performing an optimization, with objective function γ and search variables κ_{ij}, γ, under the constraints (2.1), supplemented by the order conditions (see e.g. Hairer & Wanner (1996, Section IV.6, Table 6.1)).

We performed a numerical search along these lines (using MATLAB), and obtained methods $K = (\kappa_{ij})$ with maximal $c(K)$ in $S_{s,p}$. We found that κ_{ij} and $c(K)$ can be represented (up to all computed digits) by the following formulas:

$$c(K) = 2s, \quad \text{and} \quad \kappa_{ij} = \begin{cases} \frac{1}{2s} \text{ if } i = j,\ 1 \leq i \leq s, \\ \frac{1}{s} \text{ if } 1 \leq j < i \leq s+1, \\ 0 \text{ otherwise.} \end{cases} \tag{3.1}$$

The formulas in (3.1) were obtained via a numerical search, which provides no formal proof of optimality. We are thus led, in a natural way, to Conjecture 3.1, stated below. The conjecture is supported by the fact that the optimal second order SDIRK methods with 1 and 2 stages, given above, nicely fulfill (3.1) with $s = 1, 2$, respectively.

Conjecture 3.1. *Let $p = 2$ and $s \geq 3$. Then the SDIRK method $K = (\kappa_{ij})$, defined by* (3.1)*, is optimal, with respect to $c(K)$, in $S_{s,p}$.*

3.3 Optimal SDIRK methods of order $p = 3$

Optimal method with $s = 2$ stages

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), Butcher (1987, Section 347), Dekker and Verwer (1984, Section 3.5). From these expressions, one easily sees that just one of the two methods has a nonnegative coefficient matrix K (which is necessary in order that $c(K) > 0$, see Theorem 2.3). For this method we have $\kappa_{1,1} = \kappa_{2,2} = \frac{3-\sqrt{3}}{6}$, $\kappa_{2,1} = \frac{1}{\sqrt{3}}$, $\kappa_{3,1} = \kappa_{3,2} = \frac{1}{2}$, and the corresponding characteristic coefficient equals $c(K) = 1+\sqrt{3} \simeq 2.732\ 050\ 807\ 568$.

Optimal methods with $s \geq 3$ stages

By a numerical search in $S_{s,3}$ (using MATLAB), similar to the search in Section 3.2, we obtained methods K with maximal $c(K)$. By trial and error, we found for these methods that (up to all computed digits) the following explicit formulas are valid:

$$c(K) = s - 1 + \sqrt{s^2 - 1}, \text{ and } \kappa_{ij} = \begin{cases} \frac{1}{2}\big(1 - \sqrt{\frac{s-1}{s+1}}\big) & \text{if } i = j,\ 1 \leq i \leq s, \\ \frac{1}{\sqrt{s^2-1}} & \text{if } 1 \leq j < i \leq s, \\ \frac{1}{s} & \text{if } i = s+1,\ 1 \leq j \leq s, \\ 0 & \text{otherwise.} \end{cases} \tag{3.2}$$

Since the formulas (3.2) were obtained via numerical computations, we have again no formal proof of optimality. We are led to Conjecture 3.2, stated below. The corollary is supported by the fact that the optimal method in $S_{2,3}$, given above, fulfills (3.2) with $s = 2$.

Conjecture 3.2. *Let $p = 3$ and $s \geq 3$. Then the method $K = (\kappa_{ij})$ defined in (3.2) is optimal, with respect to $c(K)$, in $S_{s,p}$.*

3.4 Optimal SDIRK methods of order $p = 4$

Optimal method with $s = 3$ stages

There exist three SDIRK methods with 3 stages and order 4. Explicit expressions for the coefficients κ_{ij} are available – see e.g. Kværnø, Nørsett & Owren (1996, Table1), Butcher (1987, Section 347), Dekker and Verwer (1984, Section 3.5). From these expressions, one easily sees that just one of the three methods satisfies $K \geq 0$. For this method we have:

$$c(K) = \frac{4\xi}{4\xi^2 - 6\xi + 1}, \qquad K = \begin{pmatrix} \xi & 0 & 0 \\ \frac{1}{2} - \xi & \xi & 0 \\ 2\xi & 1 - 4\xi & \xi \\ \frac{1}{6(2\xi-1)^2} & \frac{2(6\xi^2-6\xi+1)}{3(2\xi-1)^2} & \frac{1}{6(2\xi-1)^2} \end{pmatrix}, \tag{3.3}$$

where ξ is the smallest solution of the equation $\xi^3 - \frac{3}{2}\xi^2 + \frac{1}{2}\xi - \frac{1}{24} = 0$. We have (up to the number of given digits) $\xi = 0.128\ 886\ 400\ 515$ and $c(K) = 1.758\ 770\ 483\ 143$.

Optimal methods with $s \geq 4$ stages

By a numerical search, along similar lines as in Sections 3.2, 3.3, we obtained methods K with maximal $c(K)$ in $S_{s,4}$. We did not succeed in finding for these methods simple closed-form expressions such as (3.1), (3.2). Coefficients specifying the optimal methods $K = (\kappa_{ij})$, and corresponding characteristic coefficients $c(K)$, are given below. We display only the nonzero entries of the matrices K.

Method with 4 *stages*
The optimal method, in $S_{4,4}$, has a characteristic coefficient $c(K) = 4.208\ 135\ 414\ 418$, and its coefficient matrix $K = (\kappa_{ij})$ is as follows:

```
0.097961082941
0.262318069183  0.097961082941
0.230169419019  0.294466719347  0.097961082941
0.210562684389  0.269382888280  0.307008634881  0.097961082941
0.222119403264  0.282060762166  0.236881213175  0.258938621395
```

Method with 5 *stages*
The optimal method, in $S_{5,4}$, has a characteristic coefficient $c(K) = 5.747\ 429\ 371\ 524$, and its coefficient matrix $K = (\kappa_{ij})$ is as follows:

```
0.078752939968
0.222465723027  0.078752939968
0.203192361700  0.230847263068  0.078752939968
0.188022704389  0.191735630027  0.209922288451  0.078752939968
0.188025114093  0.191739898281  0.209907601860  0.252726086329  0.078752939968
0.192143833571  0.200935182974  0.205799262036  0.200553844640  0.200567876778
```

Method with 6 *stages*
The optimal method, in $S_{6,4}$, has a characteristic coefficient $c(K) = 7.549\ 977\ 007\ 094$, and its coefficient matrix $K = (\kappa_{ij})$ is as follows:

0.067410767219					
0.194216850802	0.067410767219				
0.194216850802	0.199861501713	0.067410767219			
0.162188551749	0.166902343330	0.145120313717	0.067410767219		
0.165176818500	0.169977460026	0.150227711763	0.181214258555	0.067410767219	
0.165176818500	0.169977460026	0.150227711763	0.181214258555	0.199861501713	0.067410767219
0.168954170460	0.173864595628	0.156683775305	0.157643002581	0.173864725004	0.168989731022

Method with 7 *stages*
The optimal method, in $S_{7,4}$, has a characteristic coefficient $c(K) = 8.671\ 030\ 957\ 620$, and its coefficient matrix $K = (\kappa_{ij})$ is as follows:

0.056879041592						
0.172205581756	0.056879041592					
0.135485903539	0.135485903539	0.056879041592				
0.133962606568	0.133962606568	0.170269437596	0.056879041592			
0.133962606568	0.133962606568	0.170269437596	0.172205581756	0.056879041592		
0.138004377067	0.133084723451	0.152274237527	0.154005757170	0.154005757170	0.056879041592	
0.139433665640	0.134719607258	0.145910607076	0.147569765489	0.147569765489	0.165009008641	0.056879041592
0.138370770799	0.134572540279	0.150642940425	0.152355910489	0.152355910489	0.132951737506	0.138750190012

Method with 8 *stages*
The optimal method, in $S_{8,4}$, has a characteristic coefficient $c(K) = 10.269\ 965\ 214\ 352$, and its coefficient matrix $K = (\kappa_{ij})$ is as follows:

0.050353353407							
0.147724666662	0.050353353407						
0.114455029802	0.114455029802	0.050353353407					
0.114147680771	0.114147680771	0.147327977820	0.050353353407				
0.114163314686	0.114163314686	0.147259379853	0.147655883990	0.050353353407			
0.114163314686	0.114163314686	0.147259379853	0.147655883990	0.147724666662	0.050353353407		
0.118472990244	0.118472990244	0.128349529304	0.128695117609	0.128755067770	0.128755067770	0.050353353407	
0.118472990244	0.118472990244	0.128349529304	0.128695117609	0.128755067770	0.128755067770	0.147724666662	0.050353353407
0.117592883046	0.117592883046	0.132211234288	0.132567220450	0.132628974356	0.132293123539	0.117556840638	0.117556840638

3.5 SDIRK methods of order $p > 4$

We have the following negative result.

Theorem 3.1. *There exists no SDIRK method with positive charateristic coefficient and order of accuracy greater than four.*

Proof. Suppose $K = (\kappa_{ij})$ is a coefficient matrix specifying an SDIRK method with characteristic coefficient $c(K) > 0$. In view of Corollary 2.5 and Theorem 2.3, we have for all i, j the inequality $\kappa_{ij} \geq 0$.

We put $b_j = \kappa_{(s+1),\,j}$ and define index sets M, N with $M \cup N = \{1, 2, \ldots, s\}$ by

$$b_j > 0 \ \text{ (for } j \in M) \quad \text{ and } \quad b_j = 0 \ \text{ (for } j \in N).$$

Using Corollary 2.5 and Theorem 2.3 once more, it follows that implication (2.2) is valid. An application of (2.2), with $i = s + 1$, shows that $b_m \cdot \kappa_{m,\,n} = 0$ for $n \in N$ and $1 \leq m \leq s$. Hence $\kappa_{m,\,n} = 0$ for $m \in M,\ n \in N$.

We delete the n-th row and n-th column of K (for all $n \in N$), and denote the resulting matrix by $\widehat{K}$. The latter matrix specifies an SDIRK method with the property that all entries in the last row of its coefficient matrix are positive. Any SDIRK method, with this property, has order of accuracy at most 4, see e.g. Dekker & Verwer (1984, Theorem 3.6.16) or Hairer & Wanner (1996, Section IV.13).

The proof of the theorem is completed by noting that the order of the original SDIRK method K is equal to the order of the method given by $\widehat{K}$. $\square$

Results, related to Theorem 3.1, can be found e.g. in Dekker & Verwer (1984, Corollary 6.2.8), Kraaijevanger (1991, Corollary 8.7).

3.6 Summary and discussion of the optimal SDIRK methods

Table 1 summarizes some results of the above search for optimal methods in $S_{s,p}$. It displays the characteristic coefficients $c(K)$ of optimal SDIRK methods $K = (\kappa_{ij})$ with s-stages and order p.

	$p=1$	$p=2$	$p=3$	$p=4$
$s=1$	∞	2	-	-
$s=2$	∞	4	2.7321	-
$s=3$	∞	6	4.8284	1.7588
$s=4$	∞	8	6.8730	4.2081
$s=5$	∞	10	8.8990	5.7474
$s=6$	∞	12	10.9161	7.5500
$s=7$	∞	14	12.9282	8.6710
$s=8$	∞	16	14.9373	10.2700

Table 1: Characteristic coefficients $c(K)$ of optimal methods K in $S_{s,p}$.

The table clearly shows that, for given p, the characteristic coefficients $c(K)$ become larger when s increases. A larger $c(K)$ means that strong stability in the situation (1.4) can be guaranteed under a milder stepsize restriction (1.5) (with $c = c(K)$).

Of course, a larger $c(K)$ does not automatically imply a better overall efficiency in practice – because also the computational labor per step and accuracy should be taken into account – cf. e.g. Spiteri & Ruuth (2002, Section 3), Ferracina & Spijker (2004, Section 4.2) for related considerations.

It may be natural to compare the coefficients in Table 1 with the maximal characteristic coefficients obtainable in the class of s-stage p-th order explicit RKMs. We denote the latter coefficients by $e_{s,p}$. It is known that $e_{s,p} \leq s - p + 1$ (for $s \geq p \geq 1$), whereas $e_{s,p} = 0$ (for $p \geq 5$), see Kraaijevanger (1991) and e.g. Ferracina & Spijker (2004). For $p = 1$, 2, and for $p = 3$, $s = 3$, 4, the above upperbound for $e_{s,p}$ becomes an equality. Moreover, for $p = 3$, 4 and various $s \geq 5$, values for $e_{s,p}$ were found that are not much smaller than $s - p + 1$, see the above papers and Spiteri & Ruuth (2002, 2003).

It follows that, for given s, p, the characteristic coefficient in the above table is larger than $e_{s,p}$, certainly, but it is not evident that the difference is so big that the optimal SDIRK methods, with $p \geq 2$, can beat in practice the explicit optimal methods. In particular, the question poses itself of whether the size of the coefficients in the table offsets the amount of work that is necessary, for SDIRK methods, to solve the s (nonlinear) equations (1.2.a).

The size of the coefficients $c(K)$, on which we have been focussing, is a *theoretical* aid to assess a-priori the behavior of RKMs, for nonlinear problems which can be modelled via assumption (1.4). Therefore, extensive numerical *experiments* may be essential for supplementing this assessment.

Numerical experiments on any large scale are beyond the scope of the present paper. In the next section a small experiment is performed, just to obtain a first idea about the relation between Table 1 and the actual numerical behavior of optimal SDIRK methods.

4 A numerical experiment.

In the following we denote by $K_{s,\,p}$ the (coefficient matrices of the) optimal methods in $S_{s,\,p}$ which are specified in Section 3. We consider the numerical behavior of the methods, with $s = p-1,\ p,\ p+1$, for a nonlinear test equation. Our focus is on the TVD property, mentioned in Section 1.1, rather than on a detailed accuracy study.

We deal with the numerical solution of a nonlinear hyperbolic equation, the 1-dimensional Buckley-Leverett equation, defined by

$$\frac{\partial}{\partial t}u(x,t) + \frac{\partial}{\partial x}\Phi(u(x,t)) = 0, \tag{4.1}$$

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 the problem, using a uniform grid with mesh-points $x_j = j\Delta x$, where $j = 1,\ 2, ..., N$ and $\Delta x = 1/N,\ N = 100$. Equation (4.1) is approximated 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 $p \equiv q \pmod N$. 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 $51 \le j \le 100$). We solved the resulting initial value problem, of the form (1.1), by the explicit Euler method and by the optimal methods $K_{s,\,p}$ 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 $[0, 1/8]$, is plotted as a function of the stepsize; i.e, the quantity

$$\mu(\Delta t) = \max\left\{\frac{\|u_n\|_{TV}}{\|u_{n-1}\|_{TV}} :\ n \ge 1 \text{ with } n\Delta t \le 1/8\right\} \tag{4.2}$$

is plotted as a function of Δt. Clearly, the value $\mu(\Delta t) = 1$ corresponds to the situation where $\|u_n\|_{TV} \le \|u_{n-1}\|_{TV}$ for all $n \ge 1$, $n\Delta t \le 1/8$, i.e. the method is total-variation-diminishing on whole of the interval $[0, 1/8]$. In Figure 1, the notation SDIRK(s, p) (for $s = p-1,\ p,\ p+1$) refers to the optimal method $K_{s,\,p}$.

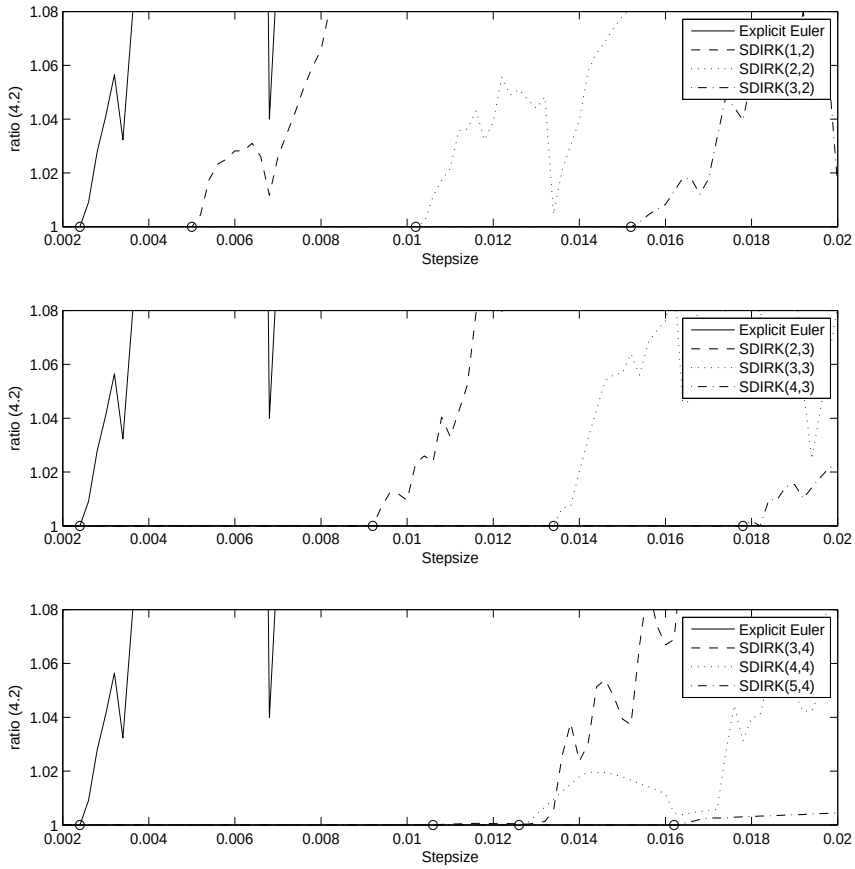

Figure 1: The ratio $\mu(\Delta t)$ vs. the stepsize Δt.

We found that the explicit Euler method is TVD for $0 < \Delta t \leq \tau \simeq 0.0025$. Furthermore, the methods $K_{s,\,p}$ are TVD for $0 < \Delta t \leq c_{s,p} \cdot \tau$, where

$$c_{1,2} \simeq 2.00, \quad c_{2,3} \simeq 3.68, \quad c_{3,4} \simeq 4.24,$$

$$c_{2,2} \simeq 4.08, \quad c_{3,3} \simeq 5.36, \quad c_{4,4} \simeq 5.04,$$

$$c_{3,2} \simeq 6.08, \quad c_{4,3} \simeq 7.12, \quad c_{5,4} \simeq 6.48.$$

One may compare these numerically observed coefficients $c_{s,\,p}$ to the maximal coefficients c for which the methods $K_{s,\,p}$ satisfy (2.3). The latter coefficients equal $c(K_{s,\,p})$ and are displayed in Table 1; from the table we have $c(K_{s+1,\,p}) > c(K_{s,\,p}) > 1$.

The majority of the coefficients $c_{s,\,p}$ don't deviate substantially from the corresponding coefficients $c(K_{s,\,p})$; and the above inequalities for the latter coefficients are nicely reflected in the numerical experiments: we have also $c_{s+1,\,p} > c_{s,\,p} > 1$.

We think the above gives a nice impression of the relation between Table 1 and the TVD properties of the methods $K_{s,\,p}$ in connection to equation (4.1).

References

[1] Alexander R.K. (1977): *Diagonally implicit Runge-Kutta methods for stiff O.D.E.'s*, SIAM J. Numer. Anal. **14**, 1006-1021.

[2] Ascher U.M., Ruuth S.J., Spiteri R.J. (1997): *Implicit-explicit Runge-Kutta methods for time-dependent partial differential equations*, Appl. Numer. Math. **25**, 151-167.

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

[4] Butcher J.C. (2003): *Numerical methods for ordinary differential equations*, John Wiley, Chichester, UK.

[5] Calvo M.P., Frutos J. de, Novo J. (2001): *Linearly implicit Runge-Kutta methods for advection-diffusion-reaction problems*, Appl. Numer. Math. **37**, 535-549.

[6] Crouzeix M. (1975): *Sur l'approximation des équations différentielles opérationelles linéaires par des méthodes de Runge-Kutta*, Ph.D. Thesis, Université Paris (1975).

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

[8] 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.

[9] Ferracina L., Spijker M.N. (2005): *An extension and analysis of the Shu-Osher representation of Runge-Kutta methods*, Math. Comp. **74**, 201-219.

[10] Gottlieb S. (2005): *On high order strong stability preserving Runge-Kutta and multi step time discretizations*, Journ. Scientif. Computing **25**, 105-128.

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

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

[13] Hairer E., Wanner G. (1996): *Solving ordinary differential equations. II. Stiff and differential-algebraic problems*, Springer-Verlag, Berlin.

[14] Hairer E., Nørsett S.P., Wanner G. (1987): *Solving ordinary differential equations. I. nonstiff problems*, Springer-Verlag, Berlin.

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

[16] Higueras I. (2004): *On strong stability preserving time discretization methods*, Journ. Scientif. Computing **21**, 193-223.

[17] Higueras I. (2005): *Representations of Runge-Kutta methods and strong stability preserving methods*, SIAM J. Numer. Anal. **43**, 924-948.

[18] Higueras I. (2006): *Strong stability for additive Runge-Kutta methods*, SIAM J. Numer. Anal., **44**, 1735-1758.

[19] Horváth Z. (1998): *Positivity of Runge-Kutta and diagonally split Runge-Kutta methods*, Appl. Numer. Math. **28**, 309-326. Eighth Conference on the numerical treatment of differential equations (Alexisbad, 1997).

[20] 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. Dundee.

[21] Hundsdorfer W.H., Verwer J.G. (2003): *Numerical solution of time-dependent advection-diffusion-reaction equations*, Springer Ser. Comp. Math., Vol 33, Springer (Berlin)

[22] Ketcheson D. I. (2004): *An algebraic characterization of strong stability preserving Runge-Kutta schemes*, Undergraduate Thesis, Brigham Young University, Provo, Utah, USA.

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

[24] Kværnø A., Nørsett S.P., Owren B. (1996): *Runge-Kutta research in Trondheim*, Appl. Numer. Math. **22**, 263-277.

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

[26] Nørsett S.P. (1974): *Semi explicit Runge-Kutta methods*, Report Dept. Math. No. 6/74, Univ. Trondheim (1974).

[27] Nørsett S.P., Wolfbrandt A. (1977): *Attainable order of rational approximations to the exponential function with only real poles*, BIT **17**, 200-208 (1977).

[28] Pareschi L., Russo G. (2005): *Implicit-Explicit Runge-Kutta schemes and applications to hyperbolic systems with relaxation*, Journ. Scientif. Computing **25**, 129-155 (2005).

[29] Ruuth S.J. (2006): *Global optimization of explicit strong-stability-preserving Runge-Kutta methods*, Math. Comp. **75**, 183-207.

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

[31] Shu C.-W. (2002): *A survey of strong stability preserving high-order time discretizations*, Collected lectures on the preservation of stability under discretization, D. Estep and S. Tavener Eds., SIAM, Philadelphia, pp. 51-65.

[32] Shu C.-W., Osher S. (1988): *Efficient implementation of essentially nonoscillatory shock-capturing schemes*, J. Comput. Phys. **77**, 439-471.

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

[34] Spijker M.N. (2007): *Stepsize conditions for general monotonicity in numerical initial value problems*, To appear in SIAM J. Numer. Anal.

[35] 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**, 469-491.

[36] Spiteri R.J., Ruuth S.J. (2003): *Non-linear evolution using optimal fourth-order strong-stability-preserving Runge-Kutta methods*, Math. Comput. Simulation **62**, 125-135.