Monografías de la Real Academia de Ciencias de Zaragoza **33**, 125–139, (2010).

Positivity properties for the classical fourth order Runge-Kutta method

I. Higueras

Departamento de Ingeniería Matemática e Informática

Universidad Pública de Navarra, 31006 Pamplona, Spain

Dedicated to Prof. Manuel Calvo, on the occasion of his 65th birthday

Abstract

Over the last few years a great effort has been done to develop Runge-Kutta (RK) methods that preserve properties such as monotonicity or contractivity for convex funtionals, or positivity. Provided that these properties hold for the explicit Euler scheme under certain stepsize restriction, it has been proved that these properties can also be maintained by some higher order RK methods under a modified stepsize. As this restriction includes the radius of absolute monotonicity of the RK scheme, strictly positive radius are required in order to obtain the desired properties with non trivial stepsizes. However, at least from the numerical positivity point of view, some authors have reported fairly good numerical results for some RK methods with zero radius, e.g. the classical fourth order four stages RK scheme. In this paper, we analyze this method and prove that, for some class of problems, it also preserves positivity. The study done strongly relies on the concept of region of absolute monotonicity for additive RK methods.

Keywords: Runge-Kutta, positivity, SSP, monotonicity, contractivity.

AMS subject classification: 65L06, 65L05, 65M20.

1 Introduction

Initial value problems for ordinary differential systems (ODEs)

$$\begin{aligned} \frac{d}{dt}u(t) &= f(t,u(t)) \qquad t \geq t_0 \\ u(t_0) &= u_0\,, \end{aligned} \tag{1}$$

arise directly in the modeling process of different phenomena, or after a method of lines approximation of evolutive partial differential equations. Quite often, the exact solution to (1) has certain property $\mathcal{P}$ (e.g., contractivity, monotonicity, positivity), usually with a physical meaning, which is relevant in the context where it appears. For example, we may have:

- Contractivity property: the solutions $u(t)$ and $\tilde{u}(t)$ satisfy

$$\|\tilde{u}(t) - u(t)\| \leq \|\tilde{u}(t_0) - u(t_0)\|\,, \qquad \text{for all } \; t \geq t_0\,, \tag{2}$$

  where $\|\cdot\|$ is a given convex function (norm, seminorm, entropy function, ...).

- Monotonicity property: the solution $u(t)$ satisfies

$$\|\tilde{u}(t)\| \leq \|u(t_0)\|\,, \qquad \text{for all } \; t \geq t_0\,, \tag{3}$$

  where again, $\|\cdot\|$ is a given convex function (norm, seminorm, entropy function, ...).

- Positivity property: if $u_0 \geq 0$, the solution $u(t)$ satisfies

$$u(t) \geq 0 \quad \text{for all} \quad t \geq t_0\,, \tag{4}$$

  where the inequalities should be understood component-wise.

In this situation, when the ODE (1) is solved numerically, it is natural to require the same qualitative property $\mathcal{P}$ to the numerical solution, $u_n \approx u(t_n)$. For this reason, when the exact solution to (1) satisfies whichever property (2)-(4), we will try to obtain, respectively,

$$\begin{aligned} &\|\tilde{u}_{n+1} - u_{n+1}\| \leq \|\tilde{u}_n - u_n\|\,, \\ &\|\tilde{u}_{n+1}\| \leq \|\tilde{u}_n\|\,, \\ &u_n \geq 0\,. \end{aligned} \tag{5}$$

Moreover, when a property $\mathcal{P}$ holds numerically, as the numerical solution depends on the stepsize h, a natural question is whether it holds for all step sizes $h > 0$, or it only holds under a step size restriction of the form $h \leq H$.

As a rule, when these issues are studied, there are four crucial aspects to consider:

i) How property $\mathcal{P}$ is obtained for the continuous problem (1).

ii) The class of problems $\mathcal{C}$ considered (e.g., linear, non linear).

iii) The type of function (*"norms"*) involved in property $\mathcal{P}$ (general convex functions, arbitrary norms, inner product norms, ...).

iv) The class of numerical methods used (Runge-Kutta, multistep, implicit or explicit schemes,...).

Depending on these aspects, different results can be obtained. Obviously, the most general conditions on problems, methods, norms,... will lead to more restricted results, whereas with more stringent conditions, sharper results will be obtained. Because of this, in order to get optimal results, it is important to analyze and determine the class of problems, the used *"norms"* and the methods we are dealing with.

Once the theoretical stepsize restrictions have been attained, it is mandatory to check their sharpness with numerical experiments on concrete problems. Although sometimes it is possible to construct a problem where the predicted and observed stepsize bounds fit, very often, for a wide class of problems, there is a great discrepancy between the effective stepsize restrictions and theoretical ones. This situation arises for example in the context of numerical positivity, studied for Runge-Kutta methods e.g. in [10, 11]. In this setting, some authors [12, 11] have reported good numerical results for schemes such that, according to the theory developed, they are not good. This the case for a widely used Runge-Kutta scheme: the fourth order four stages method (RK4). For this scheme, the radius of absolute monotonicity is trivial and therefore numerical positivity cannot be ensured. However, for many problems, RK4 method give fair good results.

In order to explain the favorable results observed, one should consider the possibility that the set of problems used to test the desired qualitative behavior belongs to a subclass $\tilde{\mathcal{C}}$ of the problems considered, $\tilde{\mathcal{C}} \subset \mathcal{C}$, and that the method used performs well on this class $\tilde{\mathcal{C}}$. This idea is not new and in the context of positivity, one could say that it is contained in the approach followed in [11] for linear and quasilinear problems; in fact, the reduction of initial values to a set of positive vectors done in [11] can be considered as a restriction of the class of problems.

In this paper we consider RK4 scheme and we will try to explain why it gives good results for certain class of problems. The approach followed differers from the one done in [11] in the sense that we do not impose any restriction on initial values but on the class of problems itself. On the other hand, we deal with non linear problems. The study done here strongly relies on the concept of region of absolute monotonicity for additive RK methods.

The rest of the paper is organized as follows. In sections 2 and 3 we introduce the methods used and we review the most relevant definitions and results concerning numerical

monotonicity. A simple example showing how RK4 scheme performs is given in section 4. A theoretical framework to explain this behavior is given in section 5. Next, these results are used for the example in section 4. The paper ends with some conclusions and forthcoming work.

2 Runge-Kutta and additive Runge-Kutta methods

A common class of one step methods to solve numerically (1) are the Runge-Kutta (RK) methods. An s-stages RK method is defined by an $s \times s$ real matrix $\mathcal{A}$ and a real vector $b \in \mathbb{R}^s$. From u_n, the numerical approximation of the solution $u(t)$ at $t = t_n$, we obtain u_{n+1}, the numerical approximation of the solution at $t_{n+1} = t_n + h$ from

$$u_{n+1} = u_n + \sum_{i=1}^{s} b_i \, f(t_n + c_i h, U_{ni}) \,, \tag{6}$$

where

$$U_{ni} = u_n + h \sum_{j=1}^{s} a_{ij} \, f(t_n + c_j h, U_{nj}) \,. \tag{7}$$

If the matrix $\mathcal{A}$ is strictly lower triangular, the method is explicit, otherwise the method is implicit. For nonlinear problems, implicit methods require the resolution of nonlinear systems of dimension $s \cdot m$, with m the dimension of the ODE system (1). Denoting the coefficients of the RK method by

$$\mathbb{A} = \begin{pmatrix} \mathcal{A} & 0 \\ b^t & 0 \end{pmatrix},$$

we can write (6)-(7) in compact form as

$$U = e \otimes u_n + h \, (\mathbb{A} \otimes I) \, F(U) \,, \tag{8}$$

where we have denoted by $e = (1, \ldots, 1)^t \in \mathbb{R}^{s+1}$, $U = (U_1^t, \ldots, U_s^t, u_{n+1}^t)^t \in \mathbb{R}^{(s+1)m}$, $F(U) = (f(U_1)^t, \ldots, f(U_s)^t, 0)^t \in \mathbb{R}^{(s+1)m}$, and similarly $\tilde{F}(U)$. The symbol $\otimes$ denotes the Kronecker product (see e.g. [2, Section 12.1])

$$A \otimes B = \begin{pmatrix} a_{11}B & \cdots & a_{1m}B \\ \vdots & & \vdots \\ a_{m1}B & \cdots & a_{mm}B \end{pmatrix}.$$

Explicit schemes are easy to implement but they are not adequate to solve stiff ODEs because they require small stepsizes; on the other hand, many implicit schemes do not suffer from these stepsize restrictions, but with them, one has to deal with the numerical

resolution (difficult sometimes) of nonlinear systems. However, many times, stiffness is only associated with a part of the problem; that is to say, the ODE can be written as

$$\frac{d}{dt}u(t) \;=\; f(t,u(t)) + \tilde{f}(t,u(t)) \qquad t \geq t_0 \tag{9}$$

where f contains the non stiff terms and $\tilde{f}$ contains the stiff ones. In this case, we can use IMplicit-EXplicit (IMEX) RK methods, where an explicit method is used for the non stiff terms, and an implicit one for the stiff part. In compact form, IMEX RK methods with coefficients $(\mathbb{A}, \tilde{\mathbb{A}})$, where $\mathbb{A}$ denotes the explicit method and $\tilde{\mathbb{A}}$ the implicit one, are given by

$$U = e \otimes u_n + h\,(\mathbb{A} \otimes I)\,F(U) + h\,(\tilde{\mathbb{A}} \otimes I)\,\tilde{F}(U)\,. \tag{10}$$

3 Monotonicity preserving methods (or Strong Stability Preserving methods)

Over the last years, a big effort has been done to develop methods such that monotonicity (contractivity, positivity) is preserved numerically. For RK methods, it is proven that these properties can be ensured under a stepsize restriction of the form

$$\Delta t \leq \tau_0 \cdot \mathcal{R}(\mathbb{A})\,. \tag{11}$$

In (11), τ_0 is a problem dependent parameter and $\mathcal{R}(\mathbb{A})$ is a method dependent parameter. To be more precise, τ_0 is a constant that ensures property $\mathcal{P}$ for the explicit Euler method, $u_{n+1} = u_n + \tau\, f(u_n)$, whenever $0 \leq \tau \leq \tau_0$, that is to say,

$$\begin{aligned}
&\|u_n + \tau\, f(u_n)\| \leq \|u_n\|\,, && \text{(monotonicity)}\\
&\|u_n - v_n + \tau\,(f(u_n) - f(v_n))\| \leq \|u_n - v_n\|\,, && \text{(contractivity)}\\
&u_n \geq 0 \quad \Longrightarrow \quad u_{n+1} = u_n + \tau\, f(u_n) \geq 0\,, && \text{(positivity)}
\end{aligned} \tag{12}$$

and $\mathcal{R}(\mathbb{A})$ is the radius of absolute monotonicity defined as follows.

Definition 3.1 *[13, Definition 2.4] An s-stage RK method with coefficients $\mathbb{A}$ is said to be absolutely monotonic at a given point $\xi \leq 0$ if $I - \xi\mathbb{A}$ is non singular, and*

$$(I - \xi\mathbb{A})^{-1}\mathbb{A} \geq 0\,, \qquad\qquad (I - \xi\mathbb{A})^{-1}e \geq 0\,, \tag{13}$$

where $e = (1, 1, \ldots, 1)^t \in \mathbb{R}^{s+1}$, and the vector inequalities are understood componentwise. Further, the method is said to be absolutely monotonic on a given set $\Omega \subset \mathbb{R}$ if it is absolutely monotonic at each $\xi \in \Omega$. The radius of absolute monotonicity $\mathcal{R}(\mathbb{A}))$ is defined by

$$\mathcal{R}(\mathbb{A}) = \sup\{\, r \,|\, r \geq 0 \text{ and } \mathbb{A} \text{ is absolutely monotonic on } [-r, 0]\,\}\,.$$

If there is no $r > 0$ such that $\mathbb{A}$ is absolutely monotonic on $[-r, 0]$, we set $\mathcal{R}(\mathbb{A}) = 0$.

As a result, if $\mathcal{R}(\mathbb{A}) = 0$, from (11) we obtain a trivial stepsize restriction. At this point we should remark that, as proven in [3], the stepsize restriction for monotonicity (11) for the class of problems (12) is optimal.

Observe that conditions (13) are

$$\begin{pmatrix} (I-\xi\mathcal{A})^{-1} & 0 \\ \xi\, b^t(I-\xi\mathcal{A})^{-1} & 1 \end{pmatrix} \begin{pmatrix} e \\ 1 \end{pmatrix} \geq 0\,, \qquad \begin{pmatrix} (I-\xi\mathcal{A})^{-1} & 0 \\ \xi\, b^t(I-\xi\mathcal{A})^{-1} & 1 \end{pmatrix} \begin{pmatrix} \mathcal{A} & 0 \\ b^t & 0 \end{pmatrix} \geq 0\,.$$

and hence absolute monotonicity at a given point ξ is equivalent to the following sign conditions:

$$\begin{aligned} \phi(\xi) &= 1 + \xi\, b^t(I-\xi\,\mathcal{A})^{-1}e \geq 0\,, \\ \mathcal{A}(\xi) &= \mathcal{A}(I-\xi\mathcal{A})^{-1} \geq 0\,, \\ b(\xi)^t &= b^t(I-\xi\mathcal{A})^{-1} \geq 0\,, \\ e(\xi) &= (I-\xi\mathcal{A})^{-1}e \geq 0\,, \end{aligned}$$

where now $e = (1, 1, \ldots, 1) \in \mathbb{R}^s$. Observe that $\phi(\xi)$ is the the stability function of the RK method.

For additive RK methods (10), the concept of radius of absolute monotonicity is extended to the region of absolute monotonicity.

Definition 3.2 *[9, Definition 2.3] An s-stage additive RK method $(\mathbb{A}, \tilde{\mathbb{A}})$ is said to be absolutely monotonic (a.m.) at a given point $(\xi, \tilde{\xi})$ with $\xi, \tilde{\xi} \leq 0$ if the matrix $I - \xi\mathbb{A} - \tilde{\xi}\tilde{\mathbb{A}}$ is invertible and*

$$\mathsf{A}(\xi, \tilde{\xi}) = (I - \xi\mathbb{A} - \tilde{\xi}\tilde{\mathbb{A}})^{-1}\mathbb{A} \geq 0\,, \tag{14}$$

$$\tilde{\mathsf{A}}(\xi, \tilde{\xi}) = (I - \xi\mathbb{A} - \tilde{\xi}\tilde{\mathbb{A}})^{-1}\tilde{\mathbb{A}} \geq 0\,, \tag{15}$$

$$\mathsf{e}(\xi, \tilde{\xi}) = (I - \xi\mathbb{A} - \tilde{\xi}\tilde{\mathbb{A}})^{-1}e \geq 0\,. \tag{16}$$

Further, the additive method is said to be absolutely monotonic on a given set $\Omega \in \mathbb{R}^2$ if it is absolutely monotonic at each $(\xi, \tilde{\xi}) \in \Omega$.

Observe that for RK we work in $\mathbb{R}$ but additive RK methods we have to work in $\mathbb{R}^2$. For this reason we define the region and the curve of absolute monotonicity as follows.

Definition 3.3 *[9, Definition 2.4] The region of absolute monotonicity, denoted by $\mathcal{R}(\mathbb{A}, \tilde{\mathbb{A}})$, is defined by*

$$\mathcal{R}(\mathbb{A}, \tilde{\mathbb{A}}) = \{\, (r, \tilde{r}) \,|\, r \geq 0\,, \tilde{r} \geq 0 \text{ and } (\mathbb{A}, \tilde{\mathbb{A}}) \text{ is a.m. on } [-r, 0] \times [-\tilde{r}, 0] \,\}\,.$$

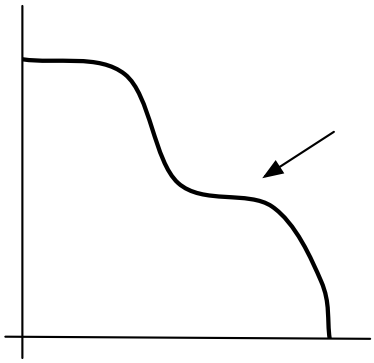

Figure 1.— Curve of absolute monotonicity

The curve of absolute monotonicity, denoted by $\partial\mathcal{R}(\mathbb{A},\tilde{\mathbb{A}})$*, is the frontier of the set* $\mathcal{R}(\mathbb{A},\tilde{\mathbb{A}})$ *excluding the coordinate axis (see figure 1). If there is no* $r>0$*,* $\tilde{r}>0$ *such that* $(\mathbb{A},\tilde{\mathbb{A}})$ *is absolutely monotonic on* $[-r,0]\times[-\tilde{r},0]$*, we set* $\partial\mathcal{R}(\mathbb{A},\tilde{\mathbb{A}})=(0,0)$*.*

For additive RK methods, it is assumed that numerical monotonicity holds when explicit Euler method is used for both functions f, $\tilde{f}$, i.e. there exists some fixed $\tau_0,\tilde{\tau}_0>0$ such that

$$\|u_n+\tau\,f(u_n)\|\le\|u_n\|\,,\qquad\qquad\|u_n+\tilde{\tau}\,\tilde{f}(u_n)\|\le\|u_n\|\,. \tag{17}$$

Under these assumptions, numerical monotonicity can be ensured for the additive RK method $(\mathbb{A},\tilde{\mathbb{A}})$ under the stepsize restriction

$$h\le\ \min\{r\,\tau_0,\,\tilde{r}\,\tilde{\tau}_0\}\ , \tag{18}$$

where r and $\tilde{r}$ are such that the point $(r,\tilde{r})\in\mathcal{R}(\mathbb{A},\tilde{\mathbb{A}})$ (see [9] for details). As it is proven in [19], stepsize restriction (18) is optimal for the class of problems (17).

Monotonicity properties of numerical schemes have also been deeply studied in the context of hyperbolic systems of conservation laws. In this setting, monotone schemes for the Total Variation (TV) seminorm are known as Total Variation Diminishing (TVD) or Strong Stability preserving methods (SSP). The class of ODEs considered in this context arise from a method of lines approximation of this class of partial differential equations, and a simple numerical example given in [5], shows that the use of non-SSP methods for the time discretization of these ODEs has the potential to produce an undesirable overshoot.

In the seminal paper [16], Shu & Osher consider SSP (or TVD) spatial discretizations such that

$$\|u_n+h\,f(u_n)\|_{TV}\le\|u_n\|_{TV}\,,\qquad h\le\Delta t_{FE}\,.$$

However, as the forward Euler method has the drawback of its low order of accuracy, higher order SSP methods are of great interest, and over the last few years a great effort

has been done to develop high order SSP methods ([16, 17, 6, 14, 15, 20], see [5, 18, 7] for reviews on this topic). It is important to point out that explicit RK methods in [16] are not written in the standard form (6)-(7), but as

$$\begin{aligned} u^{(1)} &= u_n \\ u^{(i)} &= \sum_{k=1}^{i-1} \left(\alpha_{ik} u^{(k)} + h\, \beta_{ik} f\left(u^{(k)}\right) \right) , \qquad i = 2, \ldots, s+1 \\ u_{n+1} &= u^{(s+1)} \end{aligned} \tag{19}$$

where $\alpha_{ik} \geq 0$ for all i, j, and $\sum_{k=1}^{i-1} \alpha_{ik} = 1$, $i = 2, \ldots, s+1$. It is also imposed that

$$\beta_{i,j} = 0 \qquad \text{whenever} \qquad \alpha_{ij} = 0\,. \tag{20}$$

It is straightforward to check that, if $\beta_{ij} \geq 0$, convex combinations of the forward Euler method are obtained in (19). In this case, the new method will also be strongly stable, with a modified step size restriction

$$h \leq c\, \Delta t_{FE}\,,$$

where the CFL coefficient c is given by

$$c = \min_{ik} \frac{\alpha_{ik}}{\beta_{ik}}\,. \tag{21}$$

Given a RK method in the Shu & Osher representation (19), if we denote by $\Lambda = (\alpha_{ij})$, $\Gamma = (\beta_{ij})$, it is not difficult to see that the Butcher matrix of the RK scheme is given by $\mathbb{A} = (I - \Lambda)^{-1}\Gamma$; with this notation, the sign conditions on α_{ij}, β_{ij} imply that $\Lambda \geq 0$, $\Gamma \geq 0$, the CFL coefficient in (21) is given by

$$\Lambda - c\,\Gamma \geq 0\,, \tag{22}$$

and condition (20) trivially follows from (22).

However, as many authors have pointed out, given a RK method $\mathbb{A}$, its representation Λ, Γ is not unique. For this reason, as the CFL coefficient (21) depends on the representation available, a problem of great interest and deeply studied in the SSP community has been how to obtain optimal representations. This problem was solved in [4, 8] where the connection between optimal Shu & Osher representations (19) and the radius of absolute monotonicity $\mathcal{R}(\mathbb{A})$ is given.

In the Shu & Osher representation, RK methods with $\mathcal{R}(\mathbb{A}) = 0$ require negative coefficients β_{ij}. This is the case the classical fourth order four stage RK method, whose Butcher coefficients are given by

$$
\begin{array}{c|cccc}
0 & 0 & & & \\
\frac{1}{2} & \frac{1}{2} & 0 & & \\
\frac{1}{2} & 0 & \frac{1}{2} & 0 & \\
1 & 0 & 0 & 1 & 0 \\
\hline
 & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6}
\end{array}
$$

In [16] when a negative value β_{ij} is required, function f is replaced by an associated operator $\tilde{f}$ corresponding to stepping backward in time. It is assumed that $\tilde{f}$ approximates the same spatial derivatives as f and

$$\|u_n - h\,\tilde{f}(u_n)\| \le \|u_n\|\,, \qquad \text{for } h \le \Delta t_{FE}\,, \tag{23}$$

where the stepsize restriction Δt_{FE} is the stepsize restriction needed to obtain monotonicity when the explicit Euler scheme is used for f in forward time,

$$\|u_n + h\,f(u_n)\| \le \|u_n\|\,, \qquad \text{for } h \le \Delta t_{FE}\,. \tag{24}$$

When negative values are required, the CFL coefficient with the Shu & Osher representations is computed from

$$c = \min_{ik} \frac{\alpha_{ik}}{|\,\beta_{ik}\,|}\,. \tag{25}$$

See [16] for details.

For example, the four stages RK scheme is written in [16] as

$$
\begin{aligned}
u^{(1)} &= u^{(0)} \\
u^{(2)} &= u^{(1)} + \tfrac{1}{2}hf(u^{(1)}) \\
u^{(3)} &= \tfrac{1}{2}u^{(1)} - \tfrac{1}{4}h\tilde{f}(u^{(1)}) + \tfrac{1}{2}u^{(2)} + \tfrac{1}{2}hf(u^{(2)}) \\
u^{(4)} &= \tfrac{6431}{80000}u^{(1)} - \tfrac{18769}{160000}h\tilde{f}(u^{(1)}) + \tfrac{18769}{80000}u^{(2)} - \tfrac{137}{400}h\tilde{f}(u^{(2)}) + \tfrac{137}{200}u^{(3)} + hf(u^{(3)}) \\
u^{(5)} &= \tfrac{1}{3}u^{(2)} + \tfrac{1}{6}hf(u^{(2)}) + \tfrac{1}{3}u^{(3)} + \tfrac{1}{3}u^{(4)} + \tfrac{1}{6}hf(u^{(4)})
\end{aligned}
\tag{26}
$$

In [8], Shu & Osher representations with negative coefficients were interpreted as perturbations of the original RK method $\mathbb{A}$ with a perturbation matrix $\tilde{\mathbb{A}}$. More precisely,

$$U = e \otimes u_n + h(\mathbb{A} \otimes I)F(U) + h(\tilde{\mathbb{A}} \otimes I)\left(F(U) - \tilde{F}(U)\right)\,. \tag{27}$$

We can separate the terms in f and $\tilde{f}$ and consider scheme (27) in additive form,

$$U = e \otimes u_n + h((\mathbb{A} + \tilde{\mathbb{A}}) \otimes I)F(U) - h(\tilde{\mathbb{A}} \otimes I)\tilde{F}(U)\,.$$

Observe that in this case we are assuming (23)-(24), and hence we have that $\tau_0 = \tilde{\tau}_0 = \Delta t_{FE}$. Applying the results for additive RK methods [9], we obtain monotonicity under the stepsize restriction (see (18))

$$h \leq r\, \Delta t_{FE}\,,$$

where r is such that $(r, r) \in \mathcal{R}(\mathbb{A} + \tilde{\mathbb{A}}, \tilde{\mathbb{A}})$.

In particular, for scheme (26), after some manipulations, we obtain that it is of the form (27) with coefficient matrices $(\mathbb{A}, \tilde{\mathbb{A}})$ given by

$$\mathbb{A} = \begin{pmatrix} 0 & & & & \\ \frac{1}{2} & 0 & & & \\ 0 & \frac{1}{2} & 0 & & \\ 0 & 0 & 1 & 0 & \\ \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} & 0 \end{pmatrix}, \qquad \tilde{\mathbb{A}} = \begin{pmatrix} 0 & & & & \\ 0 & 0 & & & \\ 1/4 & 0 & 0 & & \\ \frac{46169}{160000} & \frac{137}{400} & 0 & 0 & \\ \frac{28723}{160000} & \frac{137}{1200} & 0 & 0 & 0 \end{pmatrix}. \tag{28}$$

With the notation of additive RK method, the point $(0.685, 0.685) \in \mathcal{R}(\mathbb{A} + \tilde{\mathbb{A}}, \tilde{\mathbb{A}})$, and hence, for the perturbed RK method the CFL coefficient is 0.685.

The above examples show the potential of the study done in [9] for additive RK methods, and how we can transfer these results to different kind of schemes whenever they can be formally reinterpreted as additive RK methods. As we will see later on, these ideas can be used to explain why some non-SSP methods may perform well on certain classes of problems.

4 A simple example

As it has been pointed out above, the classical fourth order four stages RK scheme has $\mathcal{R}(\mathbb{A}) = 0$, and therefore monotonicity (or contractivity, positivity) cannot be ensured for the class of problems satisfying (12). However, in the context of positivity good results have been reported for this method [12]. In fact, it is not difficult to construct simple academic examples that give numerical positivity under nontrivial stepsizes.

Example 4.1 We consider the problem

$$y'(t) = y(t)\,(y(t) - 1)\,, \qquad y'(t_0) = y_0\,,$$

whose solution satisfies $y(t) \in [0, 1]$ whenever $y_0 \in [0, 1]$. It is easy to check that if $0 \leq y \leq 1$, we obtain that

$$0 \leq y + \tau\, y\,(y - 1) \leq 1 \qquad \text{for all} \;\; 0 \leq \tau \leq 1\,,$$

and hence this problem satisfies property (12) for $\tau_0 = 1$. Numerical positivity can be ensured for RK methods with coefficient matrix $\mathbb{A}$ under the stepsize restriction

$$h \leq \mathcal{R}(\mathbb{A})\,.$$

However, for RK4 scheme, after some computations, we obtain that $y_n \in [0,1]$ gives $y_{n+1} \in [0,1]$ under the non-trivial stepsize restriction $h \leq 1.2956$. □

In the next sections we give an explanation of this fact.

5 Main results

Given a RK method with coefficient matrix $\mathbb{A}$, we can formally reinterpret it as an additive RK method by splitting the coefficient matrix $\mathbb{A}$. For example, if we split $\mathbb{A}$ as $\mathbb{A} = \mathbb{A}_+ - \mathbb{A}_-$, with $\mathbb{A}$, $\tilde{\mathbb{A}} \geq 0$, the numerical scheme (8) can be written as

$$U = e \otimes u_n + h\,(\mathbb{A}_+ \otimes I)\,F(U) - h\,(\mathbb{A}_- \otimes I)\,F(U)\,, \tag{29}$$

that can be interpreted as an additive RK scheme with coefficients $\mathbb{A} = \mathbb{A}_+$, $\tilde{\mathbb{A}} = \mathbb{A}_-$ applied to the functions f and $-f$. In this case, following the ideas used in [9], we can rewrite the original RK method as

$$U = \mathsf{e}(-r,-\tilde{r}) \otimes u_n + (r\,\mathsf{A}(-r,-\tilde{r}) \otimes I)\left(U + \frac{h}{r}\,F(U)\right) + \left(\tilde{r}\,\tilde{\mathsf{A}}(-r,-\tilde{r}) \otimes I\right)\left(U - \frac{h}{\tilde{r}}\,F(U)\right)\,, \tag{30}$$

where $\mathsf{e}(-r,\tilde{r})$, $\mathsf{A}(-r,-\tilde{r})$ and $\tilde{\mathsf{A}}(-r,-\tilde{r})$ are given by (14)-(16), and $r, \tilde{r}$ are such that the matrix $I + r\,\mathbb{A} + \tilde{r}\,\tilde{\mathbb{A}}$ is invertible.

If the sign conditions (14)-(16) hold for $\mathsf{e}(-r,\tilde{r})$, $\mathsf{A}(-r,\tilde{r})$ and $\tilde{\mathsf{A}}(-r,\tilde{r})$, expression (30) is simply a convex combination of forward and backward Euler steps. Hence, imposing property $\mathcal{P}$ for explicit Euler steps for f and $-f$ with coefficients τ_+, τ_- respectively, we can obtain preservation of property $\mathcal{P}$ under a stepsize restriction of the form (18),

$$h \leq \min\left\{r\,\tau_+,\, \tilde{r}\,\tau_-\right\}\,, \tag{31}$$

where r and $\tilde{r}$ are such that the point $(r,\tilde{r}) \in \mathcal{R}(\mathbb{A}_+, \mathbb{A}_-)$.

From (31), the minimum value is obtained when $r\,\tau_+ = \tilde{r}\,\tau_-$. Hence, we can take $\tilde{r} = r\,\tau_+/\tau_-$ and compute the largest value r such that

$$\left(r,\, r\,\frac{\tau_+}{\tau_-}\right) \in \mathcal{R}(\mathbb{A}_+, \mathbb{A}_-)\,.$$

Proceeding in this way, (31) is $h \leq r\,\tau_+$. Observe that r depends on τ_+/τ_-, and hence, if we denote by $y = \tau_+/\tau_-$, we obtain the stepsize restriction

$$h \leq r(y)\,\tau_+\,. \tag{32}$$

We have finished if for each $y = \tau_+/\tau_-$ we are able to compute a value $r(y)$ and a splitting $\mathbb{A}_+$, $\mathbb{A}_-$, such that the point $(r(y),\, r(y)\, y)$ belongs to the region of absolute monotonicity of $(\mathbb{A}_+, \mathbb{A}_-)$, i.e.

$$(r(y),\, r(y)\, y) \in \mathcal{R}(\mathbb{A}_+, \mathbb{A}_-)\,.$$

Furthermore, by (32), we are interested in the largest value $r(y)$.

Observe that we have not modified the original scheme and hence, if the problem function f has the desired property for explicit Euler method not only in forward time but also in backward time, and the numerical method has splittings that lead to conditions (14)-(16), we can observe good results for larger stepsize restrictions.

With the notation of section 1, in the above analysis we are not considering the class

$$\mathcal{C} = \{ \text{ problems with property } \mathcal{P} \text{ for Euler steps in forward time } \}$$

but the subclass

$$\tilde{\mathcal{C}} = \{ \text{ problems with property } \mathcal{P} \text{ for Euler steps in forward and backward time } \}\,.$$

This idea can also been used for implicit-explicit Runge-Kutta methods [1].

We finish this section pointing out that this way of proceeding is closely related to the one followed by Shu & Osher in [16] when negative coefficients β_{ij} are required. The difference is that with our approach (29), due to the good properties of f, we do not need to use a different operator $\tilde{f}$.

6 A simple example (revisited)

We can now explain the good results obtained in section 4. The first step is to check if we have property $\mathcal{P}$ for $-f$. In this case, it is easy to check that

$$0 \le y - \tau\, y\,(y-1) \le 1 \qquad \text{for all } \; 0 \le \tau \le 1\,,$$

and hence we obtain $\tau_- = 1$.

The next step is to study the used scheme, RK4 in this case. Using a numerical optimization method, we have obtained for each y the largest value $r(y)$ such that there is a splitting $\mathbb{A}_+$, $\mathbb{A}_-$, with

$$(r(y),\, r(y)\, y) \in \mathcal{R}(\mathbb{A}_+, \mathbb{A}_-)\,.$$

In this process we have used the results in (see [8, p. 939]) that establish the compulsory nonzero elements in matrix $\mathbb{A}_-$ to obtain non trivial regions $\mathcal{R}(\mathbb{A}_+, \mathbb{A}_-)$, namely, the elements a_{31}, a_{41}, a_{51}, a_{42}, a_{52}. The values obtained are shown in figure 2.

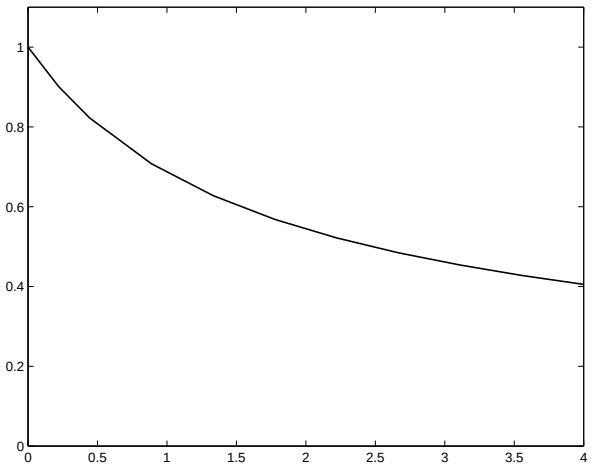

Figure 2.— Function $r(y)$ in (23), with $y = \tau_+/\tau_-$

In particular, we get $r(1) = 0.685$, and hence we obtain property $\mathcal{P}$ under the non-trivial stepsize restriction $h \leq 0.685$ (see (32)). Although this bound is not sharp for this problem, it is better than the trivial one obtained from $\mathcal{R}(\mathbb{A})$ (see (11)).

Observe that we have the same CFL coefficient obtained for the Shu & Osher representation (26). The splitting of the matrix $\mathbb{A}$ obtained for $y = 1$ is

$$\mathbb{A}_+ = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 0 & 0 \\ 0.2142 & 0.5 & 0 & 0 & 0 \\ 0.2640 & 0.3425 & 1. & 0 & 0 \\ 0.2295 & 0.3727 & 0.3333 & 0.1667 & 0 \end{pmatrix}, \qquad \mathbb{A}_- = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0.2142 & 0 & 0 & 0 & 0 \\ 0.2640 & 0.3425 & 0 & 0 & 0 \\ 0.0628 & 0.0394 & 0 & 0 & 0 \end{pmatrix}.$$

If we compare these matrices with the ones obtained with the perturbed method (28),

$$\mathbb{A}+\tilde{\mathbb{A}} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 0 & 0 \\ 0.25 & 0.5 & 0 & 0 & 0 \\ 0.2886 & 0.3425 & 1. & 0 & 0 \\ 0.3462 & 0.4475 & 0.3333 & 0.1667 & 0. \end{pmatrix}, \qquad \tilde{\mathbb{A}} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0.25 & 0 & 0 & 0 & 0 \\ 0.2886 & 0.3425 & 0 & 0 & 0 \\ 0.1795 & 0.1142 & 0 & 0 & 0 \end{pmatrix}.$$

we observe that they are different but some values are exactly the same. This fact shows that there is not uniqueness of splittings/perturbations for a given method to achieve the maximum stepsize restriction.

We should point out that for $y = 0$ we obtain $r(0) = 1$. However, as the value $y = 0$ corresponds to $\tau_+ = 0$, the stepsize restriction in (23) is trivial.

Finally, we would like to remark that for our problem we have $\tau_+ = \tau_- = 1$, but the analysis done is valid for other values. We simply have to compute the ratio $y = \tau_+/\tau_-$, and the corresponding value $r(y)$ to obtain the stepsize restricion (23).

7 Conclussions and forthcoming work

In this paper we have studied why, in the context of positivity, RK4 scheme gives good results for some problems. The results obtained strongly rely on the concept of region of absolute monotonicity for additive RK methods.

Although we have focused on positivity, the analysis done is valid for other properties $\mathcal{P}$. The basic requirement for the function problem f is the fulfillment of property $\mathcal{P}$ for explicit Euler method in forward and backward time.

We have centered on RK4 scheme, but the study can be also done for some other well known methods, both implicit and explicit.

Acknowledgements

The author acknowledge support from project MTM2008-00785

References

[1] R. Donat, I. Higueras, A. Martinez-Gavara, On stability issues for IMEX schemes applied to hyperbolic equations with stiff reaction terms. *Submitted*

[2] P. Lancaster, M. Tismenetsky, The theory of matrices, Academic Press, San Diego, CA, 1985.

[3] L. Ferracina and M.N. Spijker, Stepsize restrictions for the total-variation-diminishing property in general Runge-Kutta methods, SIAM J. Numer. Anal., 42 (2004), pp. 1073–1093.

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

[5] S. Gottlieb, C.W. Shu, and E. Tadmor, Strong stability-preserving high order time discretization methods, SIAM Rev. 43 (2001), 89–112.

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

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

[8] I. Higueras, Representations of Runge–Kutta methods and strong stability preserving methods, SIAM J. Numer. Anal. 43 (2005), no. 3, 924–948.

[9] I. Higueras, Strong stability for additive Runge-Kutta methods, SIAM J. Numer. Anal. 44 (2006), no. 4, 1735–1758.

[10] Z. Horváth, Positivity of Runge-Kutta and diagonally split Runge-Kutta methods , Appl. Numer. Math. 28 (1998), 309–326.

[11] Z. Horváth, On the positivity stepsize threshold of Runge-Kutta methods, Appl. Numer. Math. 53 (2005), 341–356.

[12] W. Hundsdorfer, B. Koren, M. van Loon, J.G. Verwer, A positive finite-difference advection scheme, J. Comput. Phys. 117 (1995) 3–46.

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

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

[15] S.J. Ruuth and R.J. Spiteri, High-order strong-stability-preserving Runge-Kutta methods with downwind-biased spatial discretizations, SIAM J. Numer. Anal. 42(2004), 974–996.

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

[17] C.W. Shu, Total variation diminishing time discretizations, SIAM J. Sci. Comput. 9(1988), 1073–1084.

[18] C.W. Shu, A survey of strong stability preserving high order time discretizations, Collected lectures on the preservation of stability under discretization. D. Estep and T. Tavener Editors. Proceedings in Applied Mathematics 109, SIAM 2002, 51–65.

[19] M.N. Spijker, Stepsize conditions for general monotonicity in numerical initial value problems, SIAM Journal on Numerical Analysis, 45 (2007), 1226–1245.

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