Journal of Scientific Computing manuscript No.
(will be inserted by the editor)

Optimal monotonicity–preserving perturbations of a given Runge-Kutta method

Inmaculada Higueras · David I. Ketcheson · Tihamér A. Kocsis

January 15, 2018

Abstract Perturbed Runge–Kutta methods (also referred to as downwind Runge–Kutta methods) can guarantee monotonicity preservation under larger step sizes relative to their traditional Runge–Kutta counterparts. In this paper we study the question of how to optimally perturb a given method in order to increase the radius of absolute monotonicity (a.m.). We prove that for methods with zero radius of a.m., it is always possible to give a perturbation with positive radius. We first study methods for linear problems and then methods for nonlinear problems. In each case, we prove upper bounds on the radius of a.m., and provide algorithms to compute optimal perturbations. We also provide optimal perturbations for many known methods.

Keywords Strong Stability Preserving, Monotonicity, Runge-Kutta methods, time discretization

Mathematics Subject Classification (2000) 65L06, 65L20, 65M20

1 Introduction

In this work we are concerned with the numerical solution of initial value ordinary differential equations:

$$u'(t) = f(u(t)), \qquad u(0) = u_0. \tag{1}$$

In many physical problems f is dissipative, i.e. the exact solution satisfies

$$\frac{d}{dt}\|u(t)\| \leq 0, \tag{2}$$

The first author was supported by Ministerio de Economía y Competividad, Spain, Projects MTM2014-53178-P and MTM2016-77735-C3-2-P. The second and third authors were supported by KAUST Award No. FIC/2010/05-2000000231. The third author was also supported by TÁMOP-4.2.2.A-11/1/KONV-2012-0012: Basic research for the development of hybrid and electric vehicles, supported by the Hungarian Government and co-financed by the European Social Fund.

I. Higueras
Public University of Navarre, Pamplona 31006, Spain. E-mail: higueras@unavarra.es

D.I. Ketcheson
King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia. E-mail: david.ketcheson@kaust.edu.sa

T.A. Kocsis
Széchenyi István University, Győr, H-9026, Hungary. E-mail: katihi@sze.hu

where $\|\cdot\|$ denotes a convex functional (e.g., a norm or semi–norm). A sufficient condition for (2) is that f be monotone under an explicit Euler step:

$$\|v + hf(v)\| \leq \|v\|, \qquad \text{for all } v, \text{ and for } h \text{ satisfying } 0 \leq h \leq h_0, \tag{3}$$

where $h_0 > 0$ (in general h_0 may depend on v). We refer to [12, p. 1-2] and [23, p. 501] for details.[1]

Let u_n, u_{n+1} denote approximations, computed by some numerical integrator, to the solution at successive time steps t_n and $t_{n+1} = t_n + h$. Under the forward Euler monotonicity condition (3), it is possible to prove that many Runge–Kutta and linear multistep methods also give monotone solutions; i.e., solutions that satisfy

$$\|u_{n+1}\| \leq \|u_n\|, \qquad \text{for } h \text{ satisfying } 0 \leq h \leq R\, h_0. \tag{4}$$

Such methods are known as strong stability preserving (SSP) methods, and the factor R is known as the radius of absolute monotonicity or SSP coefficient of the method. SSP methods necessarily have non-negative coefficients, since the monotonicity property is proved using (3) and convexity. Results on numerical preservation of some other properties, like non-negativity [16] or discrete maximum-principle [30], can also be obtained in the SSP framework.

Monotonicity cannot be ensured using only assumption (3) for methods with negative coefficients [23, Thm. 4.2], or even for some methods (such as the classical fourth-order Runge–Kutta method) with non-negative coefficients [23, Thm. 9.6]. However, in some problems (such as those of Section 1.2 below) it happens that f satisfies property (3) also for negative step sizes. In other cases, f is a dissipative approximation of a conservative operator, in which case one may devise a second approximation $\tilde{f}$ that is dissipative for negative step sizes; i.e.

$$\|v - h\tilde{f}(v)\| \leq \|v\|, \qquad \text{for all } v, \text{ and for } h \text{ satisfying } 0 \leq h \leq \tilde{h}_0, \tag{5}$$

where $\tilde{h}_0 > 0$. This situation arises naturally in the context of hyperbolic PDE semi-discretizations, where f is upwind-biased and $\tilde{f}$ is downwind-biased; typically $\tilde{h}_0 = h_0$.

The function $\tilde{f}$ is to be used in place of f wherever a negative coefficient appears in the time integration method, in order to ensure monotonicity of the overall method. Introduction of $\tilde{f}$ makes it possible to ensure monotonicity for a broader class of methods, including the classical Runge–Kutta method of order four. It also makes it possible to ensure monotonicity for many methods under larger step sizes.

During the last quarter century, a number of additional authors have studied monotonicity for methods that use $\tilde{f}$ (see e.g., [29,9,28,12,13,27,8,3,14,21]). The main motivation for this work has been to break the "order barrier" that restricts explicit Runge–Kutta methods to order four and to find new methods with larger SSP coefficient [29,9,28,12,13,27,8,21]), or to explain why some non-SSP methods preserve strong stability properties like non-negativity and a discrete maximum principle [3,14]. In this context, numerical optimization of the SSP coefficient for Runge–Kutta methods with negative coefficients was conducted for explicit methods in [28,27,8] and for implicit methods in [21]. In each case, optimization was carried out over methods with a specified order and number of stages.

Methods that use both f and $\tilde{f}$ can naturally be viewed as *perturbed Runge–Kutta methods*. Although they are also connected to additive Runge–Kutta methods (see [12,13]), in the present work we will employ the perturbation viewpoint, and refer to methods that use downwind discretization as perturbed Runge–Kutta methods.

[1] Although the results in [23] are given in the context of contractivity, they are also relevant to the preservation of monotonicity. In [23, Thm. 5.1], quotients $m_\tau[x, y]$ and one-sided Gateaux variations ($m_+[x, y]$, $m_-[x, y]$), are used for $x = u - \tilde{u}$ and $y = f(u) - f(\tilde{u})$ to obtain the contractivity property $\|u(t) - \tilde{u}(t)\| \leq \|u(t_0) - \tilde{u}(t_0)\|$ for $t \geq t_0$. In the context of monotonicity, we simply take $x = u$ and $y = f(u)$ to obtain the monotonicity property $\|u(t)\| \leq \|u(t_0)\|$ for $t \geq t_0$.

1.1 Perturbed Runge–Kutta methods

A Runge-Kutta method applied to the initial value problem (1) computes approximations $u_n \approx u(t_n)$ by

$$Y = u_n e + hKF\,, \tag{6a}$$

$$u_{n+1} = Y_{s+1}. \tag{6b}$$

Here s is the number of stages, e is a vector whose entries are equal to one, Y is the vector containing the stage values and the numerical solution, $Y = (Y_1, \ldots, Y_s, Y_{s+1})^t$, $[F]_i = f(Y_i)$, and K is the $(s+1) \times (s+1)$ matrix of Butcher coefficients:

$$K = \begin{pmatrix} A & 0 \\ b^t & 0 \end{pmatrix}.$$

In this work we study perturbations of a Runge–Kutta method K to solve problem (1). To define a perturbed method, we introduce a second coefficient matrix

$$\widetilde{K} = \begin{pmatrix} \tilde{A} & 0 \\ \tilde{b}^t & 0 \end{pmatrix},$$

where the matrix $\tilde{A}$ has the same structure (strictly lower-triangular, lower-triangular, or full) as the matrix A. We also introduce a function $\tilde{f}$ such that $\tilde{f} \approx f$. We assume that f and $\tilde{f}$ satisfy the explicit Euler assumptions (3) and (5), respectively, with $\tilde{h}_0 = h_0$.

Definition 1 A *perturbed Runge–Kutta method* $(K, \widetilde{K})$ takes the form

$$Y = u_n e + hKF + h\widetilde{K}(F - \widetilde{F})\,, \tag{7a}$$

$$u_{n+1} = Y_{s+1}, \tag{7b}$$

where $[\widetilde{F}]_i = \tilde{f}(Y_i)$.

As far as we know, the first attempt to perturb a given Runge–Kutta method to obtain non-trivial SSP coefficient was made in [29], where the classical fourth–order Runge–Kutta scheme is perturbed to obtain a non-trivial SSP coefficient. The goals of this paper are to perform a rigorous study of perturbed Runge–Kutta methods and propose algorithms to obtain perturbations of a given Runge-Kutta method with optimal SSP coefficient.

Remark 1 (Perturbed methods as additive schemes) Observe that method (7) may be viewed as approximating the solution of the perturbed problem

$$u'(t) = f(u) + (f(u) - \tilde{f}(u)),$$

where $\tilde{f} \approx f$, with the additive method $(K, \widetilde{K})$. □

1.1.1 Relation between the unperturbed method and the perturbed method

The present work is based on the premise that the behavior of a perturbed method $(K, \widetilde{K})$ is related to the properties of the unperturbed method with coefficients K. To see why this is the case, observe first that the perturbed method (7) reduces to the Runge–Kutta method (6) when $\tilde{f} = f$. Furthermore, as $\tilde{f} \to f$, the perturbed method solution (given by (7)) obviously tends to the unperturbed method solution (given by (6)). In practice, for hyperbolic problems, $\tilde{f}$ and f are discretizations of a spatial differential operator [29, p.144], and their difference can be made arbitrarily small by increasing the accuracy of these discretizations. As far as convergence is concerned, the reasoning in [12, p. 933-934] shows that, for stable Runge–Kutta methods, the perturbed method

retains the order of the unperturbed one, provided that $f - \tilde{f}$ is small enough. Herein we are particularly interested in high-order time discretizations, intended to be paired with high-order spatial discretizations, for which the difference $f - \tilde{f}$ is very small.

Given the close relationship between the perturbed method and its unperturbed counterpart, it makes sense to consider developing perturbed versions of existing methods, in order to take advantage of the substantial amount of work that has gone into designing those methods.

1.2 Two motivating examples

To demonstrate the usefulness of the present work, we consider two numerical experiments. In both, the Runge–Kutta methods are used in the standard way, and an equivalent reformulation allows us to analyze their behavior using the formalism of perturbed schemes with $\tilde{f} = f$; in other words, the Runge–Kutta methods are perturbed fictitiously.

The first example shows how this work can better be used to *understand the behavior of standard (unperturbed) Runge–Kutta methods*. The second one shows that "perturbing" a robust Runge–Kutta method with many important features can be advantageous versus using an optimized SSP perturbed method.

1.2.1 Example 1

We integrate the initial value problem

$$u'(t) = \text{sign}(\sin(t))u(t)(1 - u(t)) \tag{8}$$

on the interval $t \in [0, 100]$ with initial condition $u(0) \in (0, 1)$. The true solution remains in the interval $(0, 1)$, and the explicit Euler method keeps the solution in this interval if the step size satisfies $-1 \leq h < 1$ (note that negative step sizes are included here). We will apply some well-known Runge–Kutta methods to this problem and consider two initial values: $u(0) = 10^{-8}$ and $u(0) = 1 - 10^{-8}$ (these values are chosen because initial values very close to zero or unity are the most challenging; testing other initial values in $[0, 1]$ does not seem to change the results found below).

We first consider the explicit midpoint method:

$$\begin{aligned} y_1 &= u_n\,, \\ y_2 &= u_n + \frac{h}{2} f(y_1)\,, \\ u_{n+1} &= u_n + h f(y_2). \end{aligned}$$

For this method, the formula for y_2 is an Euler step, but the formula for u_{n+1} cannot be written as a convex combination of forward Euler steps, so the standard theory of strong stability preservation does not guarantee invariance of the interval $(0, 1)$ under any step size. Nevertheless, experimentally we observe that the interval is preserved for step sizes up to $h \approx 0.73$ (see Table 1).

The theory in the present paper explains this result rather precisely. Since f satisfies both (3) and (5) for $h_0 = \tilde{h}_0 = 1$, we can formally introduce a function $\tilde{f} = f$ to facilitate the analysis. We thus "perturb" the midpoint method, replacing the formula for u_{n+1} with the equivalent expressions

$$u_{n+1} = u_n + h f(y_2) + h\,\frac{r}{2}\,\big(f(y_1) - \tilde{f}(y_1)\big) \tag{9}$$

$$= r\left(y_2 + \frac{h}{r} f(y_2)\right) + (1 - r)\left(y_1 - \frac{h}{r}\tilde{f}(y_1)\right) \tag{10}$$

Method	$R(K)$	h_{obs}	$R^{\text{opt}}(K)$
Forward Euler	1	1.00	1
Midpoint RK2	0	0.73	0.732
Heun33 [11]	0	0.91	0.776
RK4 (Kutta)	0	1.24	0.685
Merson [25]	0	0.29	0.242

Table 1 Theoretical ($R(K)$) and observed (h_{obs}) step sizes for preserving the invariant interval $(0, 1)$ for problem (8). Values in the last column ($R^{\text{opt}}(K)$) are obtained using the tools presented herein.

where $r = \sqrt{3} - 1$. The last formula above shows that u_{n+1} can be written as a convex combination of forward Euler steps with step size h/r, one using f and one using $\tilde{f}$. Of course, since $\tilde{f} = f$ this is in fact the same midpoint method, but writing it this way allows us to prove that it preserves the interval $(0, 1)$ for step sizes up to $\sqrt{3} - 1 \approx 0.73$. The perturbed forms (9) and (10) correspond to expressions (7) and (36), respectively, for the explicit midpoint method (see (74)).

Results for some additional methods are given in Table 1. The value $R(K)$ is the SSP coefficient, which is also the theoretical maximum step size for preservation of the interval $(0, 1)$ that can be guaranteed based on considering only condition (3). The value h_{obs} is the largest step size (truncated to 2 decimal places) observed to preserve the invariant interval in practice. Finally, the value $R^{\text{opt}}(K)$ gives the step size that can be guaranteed to preserve the interval using the tools developed in the present work, by finding an optimal perturbation. The values $R^{\text{opt}}(K)$ do a much better job of predicting (or explaining) the behavior of the methods for this problem.

1.2.2 Example 2

We integrate the problem

$$u'(t) = 5\,u\,(1 - u)\left(u - \frac{1}{2}\right) \tag{11}$$

with initial condition $u(0) = 0.49$ up to time $T = 10$. This problem was proposed in the context of hyperbolic PDEs in [24] and has been studied extensively. The values $u = 0$ and $u = 1$ are stable equilibria while the value $u = 1/2$ is unstable; the exact solution remains in the interval $[0, 1]$. It can be shown that the forward Euler method preserves the interval $u \in [0, 1]$ for step sizes in the approximate range $-16/5 \le \Delta t \le 2/5$ (note that negative step sizes are included here) [3, Lemma 8.1].

We consider two methods: the 5th order Bogacki-Shampine method [1] (BS75) and the optimized 5th order 7-stage SSP perturbed method (SSP75) [28]. Method BS75 has negative coefficients, so traditional SSP theory does not guarantee invariance of $[0, 1]$ under any non-zero step size; using the approach in this paper it can be proven to preserve this interval for step sizes up to approximately 0.125. The SSP75 method can be proven to preserve this interval for step sizes up to approximately 0.558. The values 0.125 and 0.558 are obtained from (40) for $h_0 = 2/5$ and the data in Table 3.

In Figure 1 we plot the error versus the time step for each method. Open circles indicate solutions that contain values outside the interval $[0, 1]$. Notice that the BS75 method performs better both in terms of accuracy and strong stability preservation.

Recall that the scheme SSP75 has been optimized to achieve the largest SSP coefficient but many other relevant properties are not taken into account. However, the Bogacki-Shampine method has been carefully constructed to optimize several properties, including accuracy. The difference in the error constants for Bogacki-Shampine and SSP75 methods, approximately 2.2×10^{-5} and 2.7×10^{-3}, respectively, explains the observed accuracies. Clearly, a method optimized for properties other than the SSP coefficient may be useful even for problems where strong stability properties are paramount.

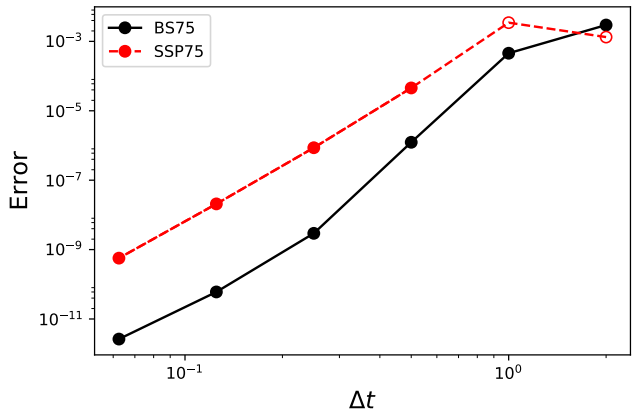

Fig. 1 Accuracy of BS75 and SSP75 methods for (11). Open circles indicate the presence of negative solution values.

Production implementations of modern IVP solvers include many important features, such as continuous output, error estimation, and automatic step size control [10]. The BS75 method, for instance, includes all of these features. None of these have been developed for existing high-order optimal downwind SSP methods, so such methods may not be a reasonable option when an efficient and robust solution is required. By instead using a perturbation of an existing method, all of these features can be used in the usual way.

1.3 Scope and outline

In the present work we seek to answer the following questions:

1. Can every Runge–Kutta method be perturbed in a way that yields a positive radius of absolute monotonicity?
2. What *a priori* limits are there on the radius of absolute monotonicity obtained by perturbing a given method?
3. Can optimal SSP methods from the literature be perturbed in order achieve an even larger radius of absolute monotonicity?
4. Given a fixed Runge–Kutta method, what perturbation results in the largest radius of absolute monotonicity for the perturbed method?
5. How can that perturbed method be found?
6. To begin with, what are the answers to the questions above if only linear problems are considered?

In Theorem 3, we prove that the answer to question 1 is affirmative. Theorems 4 and 5 answer question 2 by providing simple upper bounds; Theorem 5 implies that the answer to question 3 is negative for most known optimal SSP methods. In Section 3.5 we answer questions 4 and 5 by giving two algorithms for computing optimal perturbations. The first is provably correct but approximate, while the second is heuristic but exact and agrees with the first in all cases we have tested. Both are applicable only to explicit methods. These algorithms have been implemented in the free open-source software package Nodepy [22] and can easily be applied to any desired method. We conclude Section 3 with an application of the theorems and algorithms to optimal perturbations

of some Runge–Kutta methods from the literature and a numerical test. Among the results is the first truly optimal perturbation for the classical 4th-order method of Kutta.

We deal with application to linear problems first, in Section 2. For explicit methods applied to linear problems, the questions above can be cast in terms of absolute monotonicity of the (bivariate) stability polynomial. In Section 2.1.1 we prove a general upper bound on the radius of absolute monotonicity of the stability polynomial of an explicit perturbed Runge–Kutta method with s stages and linear order p. In Section 2.1.2, we provide an algorithm for computing tighter bounds, and tabulate some of the resulting numerical values. Examples of optimal methods are given in Section 2.2.

Section 4 contains some conclusions as well as some open questions to be studied in the future.

In Section 5 we give the proofs of the results in the paper together with an auxiliary lemma. We have collected them in a separate section in order to not interrupt the reading of the paper.

Finally, in the Appendix we give some details on perturbations for the family of second order 2-stage methods and the classical fourth order Runge–Kutta method.

Computer code to reproduce some of the examples in this paper, including the two examples above, can be found at [15].

2 Explicit perturbed Runge–Kutta methods for linear problems

To study the behavior of the perturbed Runge–Kutta method $(K, \widetilde{K})$ for linear problems, we apply it to a linear scalar test problem, setting $f(u) = \lambda u$ and $\tilde{f}(u) = \tilde{\lambda} u$ in (7). This results in the iteration

$$u_{n+1} = \phi_{(K,\widetilde{K})}(z, -\tilde{z})\, u_n,$$

where $z = h\lambda$, $\tilde{z} = h\tilde{\lambda}$ and

$$\phi_{(K,\widetilde{K})}(z,\tilde{z}) = 1 + \left(zb^t + (z+\tilde{z})\tilde{b}^t\right)\left(I - zA - (z+\tilde{z})\tilde{A}\right)^{-1} e\,. \tag{12}$$

Definition 2 We refer to (12) as the *stability function of the perturbed Runge–Kutta method* $(K, \widetilde{K})$.

Following the steps in [10, Prop. 3.2], function (12) can also be written as

$$\phi_{(K,\widetilde{K})}(z,\tilde{z}) = \frac{\det\left(I - z(A - eb^t) - (z+\tilde{z})(\tilde{A} - e\tilde{b}^t)\right)}{\det\left(I - zA - (z+\tilde{z})\tilde{A}\right)}\,. \tag{13}$$

The stability function ϕ in (12) is a rational function $\psi = P/Q$, where P and Q are polynomials in the complex variables z and $\tilde{z}$, both with real coefficients. A function ψ of this type is said to be absolutely monotonic (a.m.) at a given point $(\xi, \tilde{\xi}) \in \mathbb{R}^2$ if $Q(\xi, \tilde{\xi}) \neq 0$ and $(d^{j+k}\psi/dz^j d\tilde{z}^k)(\xi, \tilde{\xi}) \geq 0$, $j = 0, 1, \ldots$, $k = 0, 1, \ldots$ (see, e.g., [13, Def. 2.7]).

This definition is an extension of the one given in [23, Def. 2.1] for the unidimensional case: given a rational function $\psi = P/Q$, where P and Q are polynomials in the complex variable z, both with real coefficients, we say that ψ is absolutely monotonic (a.m.) at a given point $\xi \in \mathbb{R}$ if $Q(\xi) \neq 0$ and all the derivatives $(d^k\psi/dz^k)(\xi) \geq 0$, $k = 0, 1, 2, \ldots$

Definition 3 Given a function $\psi(z, \tilde{z})$, we define the *radius of absolute monotonicity* as

$$R(\psi) = \sup\left\{r \in \mathbb{R} \,|\, r = 0, \text{ or } r > 0\,, \text{ and } \psi(z,\tilde{z}) \text{ is a.m. at } (-r,-r)\right\}\,. \tag{14}$$

Observe that, if $\psi(z,\tilde{z})$ is a bivariate polynomial of combined degree s, for $r \leq R(\psi)$ we can write

$$\psi(z,\tilde{z}) = \sum_{j=0}^{s} \sum_{\ell=0}^{j} \gamma_{j\ell} \left(1+\frac{z}{r}\right)^{j-\ell} \left(1+\frac{\tilde{z}}{r}\right)^{\ell}, \qquad \text{with } \gamma_{j\ell} = \frac{r^j}{j!} \frac{\partial^j \psi}{\partial z^{j-\ell} \partial \tilde{z}^{\ell}}(-r,r), \tag{15}$$

where the coefficients $\gamma_{j\ell}$ are non-negative.

Definition 4 Given a perturbed Runge–Kutta method (7) with coefficients $(K,\widetilde{K})$, we define the *threshold factor* $R_{\text{Lin}}(K,\widetilde{K})$ as the radius of absolute monotonicity of its stability function:

$$R_{\text{Lin}}(K,\widetilde{K}) = R(\phi_{(K,\widetilde{K})}) \,. \tag{16}$$

The quantity $R_{\text{Lin}}(K,\widetilde{K})$ is referred to as the *threshold factor* due to its role in the step size for monotonicity. The following theorem appeared previously as [20, Thm. 4.6.2]. Its proof, given in Section 5, is based on the fact that, for explicit schemes, the stability function $\phi_{(K,\widetilde{K})}(z,\tilde{z})$ is a bivariate polynomial of combined degree s and thus, for $r \leq R_{\text{Lin}}(K,\widetilde{K})$, it can be written in the form (15) with non-negative coefficients $\gamma_{j\ell}$.

Theorem 1 *Let a consistent perturbed s-stage explicit Runge–Kutta method $(K,\widetilde{K})$ be given with stability function $\phi_{(K,\widetilde{K})}$, and let $\|\cdot\|$ be a convex functional. Consider the numerical solution*

$$u_{n+1} = \phi_{(K,\widetilde{K})}(hL, -h\tilde{L})\, u_n \,, \tag{17}$$

where L and $\tilde{L}$ are linear operators such that $L\tilde{L} = \tilde{L}L$ and

$$\|I + hL\| \leq 1 \,, \quad \|I - h\tilde{L}\| \leq 1 \,, \qquad 0 \leq h \leq h_0 \,.$$

Then the numerical solution (17) *satisfies the monotonicity condition* (4) *for step sizes*

$$0 \leq h \leq R_{Lin}(K,\widetilde{K})\, h_0.$$

Consequently, the larger $R_{\text{Lin}}(K,\widetilde{K})$ is, the larger is the step size restriction for montonicity. For a given Runge–Kutta method (6) with coefficients K, we are interested in determining perturbations $\widetilde{K}$ that give the largest threshold factor.

Definition 5 The *threshold factor of the optimal perturbation* is given by

$$R^{\text{opt}}_{\text{Lin}}(K) = \sup_{\widetilde{K}} R_{\text{Lin}}(K,\widetilde{K}) \,, \tag{18}$$

where, in order to preserve the explicit nature of the method, the supremum in (18) is taken over all strictly lower triangular matrices $\widetilde{K}$. A perturbation $\widetilde{K}$ such that

$$R_{\text{Lin}}(K,\widetilde{K}) = R^{\text{opt}}_{\text{Lin}}(K) \,,$$

will be called an *optimal perturbation of the method K for the linear problem.*

Taking $\widetilde{K} = 0$ gives a (not perturbed) Runge–Kutta method (6) and a (not perturbed) stability function ϕ_K. In this case we denote the threshold factor $R_{\text{Lin}}(K,0)$ simply by $R(\phi_K)$. Clearly

$$R(\phi_K) \leq R^{\text{opt}}_{\text{Lin}}(K) \,. \tag{19}$$

In the next section, we give upper bounds on $R^{\text{opt}}_{\text{Lin}}(K)$.

2.1 Upper bounds on the threshold factor for optimal perturbations

In this section we consider the set $\widetilde{\Pi}_{s,p}$, with $p \leq s$, defined as follows.

Definition 6 We define $\widetilde{\Pi}_{s,p}$, with $p \leq s$, as the set of bivariate polynomials with the following properties:

1. $\psi(z,\tilde{z}) = \displaystyle\sum_{j=0}^{p} \frac{z^j}{j!} + \sum_{j=p+1}^{s} \sigma_j z^j + (z+\tilde{z})\,\Psi(z,\tilde{z})$;
2. Ψ is a polynomial of combined degree at most $s-1$.

Observe that if $\psi(z,\tilde{z}) \in \widetilde{\Pi}_{s,p}$, then

$$\psi(z,-z) = \exp(z) + \mathcal{O}(z^{p+1})\,. \tag{20}$$

The following result explains the interest in studying the set $\widetilde{\Pi}_{s,p}$.

Proposition 1 *Let K be an explicit s-stage Runge–Kutta method with linear order p. If $\phi_{(K,\widetilde{K})}$ is the stability function of the perturbed Runge–Kutta method $(K,\widetilde{K})$, then $\phi_{(K,\widetilde{K})} \in \widetilde{\Pi}_{s,p}$.*

The aim of this section is to investigate

$$\widetilde{R}_{s,p} = \sup\left\{ R(\psi) \,|\, \psi(z,\tilde{z}) \in \widetilde{\Pi}_{s,p} \right\}. \tag{21}$$

Clearly, by Proposition 1, $\widetilde{R}_{s,p}$ is an upper bound of the threshold factor of the optimal perturbation defined by (18) (see too (16)),

$$R^{\mathrm{opt}}_{\mathrm{Lin}}(K) \leq \widetilde{R}_{s,p}\,. \tag{22}$$

Remark 2 (Realizable polynomials) We remark that not all polynomials in $\widetilde{\Pi}_{s,p}$ can be realized as the stability function of an s-stage perturbed Runge-Kutta method (7). Thus, inequality (22) is often strict (see Example 1 below). In case the optimal polynomial is realizable, the corresponding method may be of interest for the integration of linear systems. □

The rest of the section is organized as follows. In Subsection 2.1.1 we give an upper bound for $\widetilde{R}_{s,p}$. In Subsection 2.1.2, we give an algorithm to compute, $\widetilde{R}_{s,p}$ for given s and p, along with numerical values.

2.1.1 Upper bound on $\widetilde{R}_{s,p}$

The following upper bound on $\widetilde{R}_{s,p}$ is proved in Section 5.

Theorem 2 *The coefficient $\widetilde{R}_{s,p}$ defined by* (21) *has the following upper bound*

$$\widetilde{R}_{s,p} \leq \sqrt[p]{s(s-1)\cdots(s-p+1)}\,. \tag{23}$$

Consequently, from (22), we obtain the following bound for the threshold factor of the optimal perturbation

$$R^{\mathrm{opt}}_{\mathrm{Lin}}(K) \leq \sqrt[p]{s(s-1)\cdots(s-p+1)}\,.$$

s p	1	2	3	4	5	6	7	8	9	10
1	1.00									
2	2.00	1.41								
3	3.00	2.45	1.60							
4	4.00	3.46	2.49	2.00						
5	5.00	4.47	3.20	2.94	2.18					
6	6.00	5.48	4.00	3.65	3.11	2.58				
7	7.00	6.48	4.86	4.45	3.88	3.55	2.76			
8	8.00	7.48	5.77	5.31	4.57	4.32	3.72	3.15		
9	9.00	8.49	6.62	6.22	5.24	5.02	4.52	4.14	3.33	
10	10.00	9.49	7.42	7.09	5.95	5.70	5.25	4.96	4.32	3.73

Table 2 $\widetilde{R}_{s,p}$: upper bounds on $R^{\text{opt}}_{\text{Lin}}(K)$, the threshold factors for optimal perturbations

2.1.2 Numerical computation of $\widetilde{R}_{s,p}$

In this section we provide a means to compute tighter values of $\widetilde{R}_{s,p}$ using linear programming. The material in this section closely follows [20, Sect. 4.6.2].

In order to obtain these bounds, for each (s,p) we are going to construct functions $\psi(z,\tilde{z}) \in \widetilde{\Pi}_{s,p}$ that can be written in the form (15) for some $r > 0$ with non-negative coefficients $\gamma_{j\ell}$. Observe that these polynomials can be constructed if $\gamma_{j\ell}$ and r are given. From (15), after considerable manipulation we find that $\psi(z,-z) = \sum_{i=0}^{s} C_i z^i$ where

$$C_i(r,\gamma) = \sum_{j=i}^{s} \sum_{\ell=0}^{j} \gamma_{j\ell} \sum_{m=\max(0,i-\ell)}^{\min(i,j-\ell)} \binom{j-\ell}{m} \binom{\ell}{i-m} \frac{(-1)^{i-m}}{r^i},$$

where γ is a vector whose components are the coefficients $\gamma_{j,\ell}$.

Hence we have the following problem for existence of a polynomial (15) with perturbed threshold factor at least r and order at least p:

Given $r > 0$, find γ such that

$$\gamma_{j\ell} \geq 0 \qquad 0 \leq \ell \leq j \leq s \tag{24a}$$

$$C_i(r,\gamma) = \frac{1}{i!} \qquad 0 \leq i \leq p. \tag{24b}$$

Since (24b) is a system of linear equations (in γ) then for any given value of r (24) represents a linear programming feasibility problem. Hence we can use bisection and an LP solver to find the largest value of r satisfying (24), as was done for similar problems in [18,19]. Table 2 gives the computed values of $\widetilde{R}_{s,p}$ for s and p up to ten.

2.2 Examples

In Section 2.2.1 we give some examples of polynomials achieving $\widetilde{R}_{s,p}$; in Section 2.2.2 we study optimal threshold factors for perturbations $R^{\text{opt}}_{\text{Lin}}(K)$ of specified Runge–Kutta methods K.

2.2.1 Polynomials achieving $\widetilde{R}_{s,p}$

The algorithm just described also provides coefficients for an optimal polynomial $\psi_{s,p}(z,\tilde{z})$, which may or may not be realizable as the stability function of a perturbed Runge–Kutta method. Observe that all of them belong to $\widetilde{\Pi}_{s,p}$ and thus $\psi_{s,p}(z,-z)$ is an order p approximation of $\exp(z)$ (see (20)).

By computing optimal polynomials with $p = 1$ and $p = 2$ we arrived at the following results.

Proposition 2 *For $p = 1$ we have $\widetilde{R}_{s,1} = s$. This value is attained by the following polynomial in $\widetilde{\Pi}_{s,1}$*

$$\psi_{s,1}(z, \tilde{z}) = \left(1 + \frac{z}{s}\right)^s,$$

which corresponds to performing s iterated forward Euler steps of size h/s.

The proposition can be proved by checking the radius of absolute monotonicity and noticing that it achieves the bound (23). Thus the optimal first-order perturbed methods for linear problems are the same as the optimal unperturbed methods for linear problems.

Proposition 3 *For $p = 2$ we have $\widetilde{R}_{s,2} = \sqrt{s(s-1)}$. This value is attained by the following polynomial in $\widetilde{\Pi}_{s,2}$*

$$\psi_{s,2}(z, \tilde{z}) = \frac{2(s+r)-1}{2(s+r)} \left(1 + \frac{z}{r}\right)^s + \frac{1}{2(s+r)} \left(1 + \frac{\tilde{z}}{r}\right)^s, \tag{25}$$

where $r = \widetilde{R}_{s,2}$.

Again, the proposition can be proved by checking the radius of absolute monotonicity and noticing that it achieves the bound (23).

Some of the other optimal polynomials also have rational coefficients. Two optimal degree-four fourth order polynomials we found are

$$\psi^1_{4,4}(z, \tilde{z}) = \frac{1}{3}\left(1 + \frac{z}{r}\right)^2 + \frac{17}{48}\left(1 + \frac{z}{r}\right)^4 + \frac{14}{48}\left(1 + \frac{z}{r}\right)^2 \left(1 + \frac{\tilde{z}}{r}\right)^2 + \frac{1}{48}\left(1 + \frac{\tilde{z}}{r}\right)^4,$$

and

$$\psi^2_{4,4}(z, \tilde{z}) = \frac{7}{16}\left(1 + \frac{z}{r}\right)^4 + \frac{3}{8}\left(1 + \frac{z}{r}\right)^2 \left(1 + \frac{\tilde{z}}{r}\right)^2 + \frac{1}{6}\left(1 + \frac{z}{r}\right)^3 \left(1 + \frac{\tilde{z}}{r}\right) + \frac{1}{48}\left(1 + \frac{\tilde{z}}{r}\right)^4,$$

where $r = \widetilde{R}_{4,4} = 2$. Thus the optimal polynomial in $\widetilde{\Pi}_{s,p}$ is in general not unique.

Remark 3 As noted already, not all polynomials of the form (15) can be realized as the stability function of a perturbed Runge–Kutta method (7) with s stages. For example, the polyomial (25) with $s = 2$ is not the stability function of any two-stage method (i.e., using only evaluations of $f(u_n), \tilde{f}(u_n), f(y_1), \tilde{f}(y_1)$). It can be realized as the stability function of a method that has three stages, using evaluations of $f(u_n), \tilde{f}(u_n), f(y_1), \tilde{f}(y_2)$. The difference in cost between such methods depends on the nature of $f, \tilde{f}$; see [8]. For this reason, we stress that the values in Table 2 are only *upper bounds* on what can be achieved. We do not pursue the topic further here. □

2.2.2 Optimal threshold factors for perturbations of specified Runge–Kutta methods

We have no general method for finding $R^{\mathrm{opt}}_{\mathrm{Lin}}(K)$ nor a corresponding method. In this section we report results of some symbolic searches. In the case of the second-order methods, due to the small number of free parameters, it is not difficult to prove that the results below are truly optimal.

Example 1 We consider explicit perturbed second-order 2-stage Runge–Kutta methods

$$\begin{array}{c|cc} 0 & 0 & 0 \\ \alpha & \alpha & 0 \\ \hline K & 1 - \frac{1}{2\alpha} & \frac{1}{2\alpha} \end{array} \qquad \begin{array}{c|cc} & 0 & 0 \\ & \tilde{a}_{21} & 0 \\ \hline \widetilde{K} & \tilde{b}_1 & \tilde{b}_2 \end{array}. \tag{26}$$

For these methods, function (12) can be expanded as

$$\phi_{(K,\widetilde{K})}(z,\tilde{z}) = 1 + z + \frac{1}{2}z^2 + \beta_{11}z(z+\tilde{z}) + \beta_1(z+\tilde{z}) + \beta_2(z+\tilde{z})^2\,, \tag{27}$$

where

$$\beta_{11} = b^t\tilde{A}e + \tilde{b}^tAe = \tilde{b}_2a_{21} + b_2\tilde{a}_{21}\,, \qquad \beta_1 = \tilde{b}^te = \tilde{b}_1 + \tilde{b}_2\,, \qquad \beta_2 = \tilde{b}^t\tilde{A}e = \tilde{b}_2\tilde{a}_{21}\,. \tag{28}$$

The polynomial (27) is realizable (in the sense that it corresponds to a 2-stage Runge–Kutta method (26)) if the first and last equations in (28) can be solved for $\tilde{a}_{21}$ and $\tilde{b}_2$ in $\mathbb{R}$. A simple computation gives that a necessary condition is $\beta_{11}^2 - 2\beta_2 \geq 0$.

With the help of the symbolic computation program Mathematica, we have computed the largest r such that (27) is a.m. at $(-r,-r)$ and the polynomial is realizable (see [15]). We have obtained that the optimal perturbation, denoted by $\widetilde{K}_L$, satisfies $\tilde{b}_2 = \tilde{a}_{21} = 0$ and $\tilde{b}_1 = \frac{1}{3}\left(\sqrt{7}-2\right)$. Note that for these values the stability function (27) is independent of α. Furthermore,

$$R_{\text{Lin}}^{\text{opt}}(K) = \frac{1}{3}\left(1+\sqrt{7}\right) \approx 1.21525\,. \tag{29}$$

Observe that $R_{\text{Lin}}^{\text{opt}}(K) < \widetilde{R}_{2,2} = \sqrt{2}$. The stability function (27) for the optimal perturbed method is

$$\tilde{\phi}_{(K,\widetilde{K}_L)}(z,\tilde{z}) = \frac{1}{9}\left(4+\sqrt{7}\right)\left(1+\frac{z}{r}\right)^2 + \frac{1}{9}\left(5-\sqrt{7}\right)\left(1+\frac{\tilde{z}}{r}\right)\,,$$

where $r = R_{\text{Lin}}^{\text{opt}}(K)$, the value given in (29). □

Example 2 We consider now perturbations of the classical fourth–order Runge-Kutta method, of the form

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \hline K & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array} \qquad \begin{array}{c|cccc} & 0 & 0 & 0 & 0 \\ & 0 & 0 & 0 & 0 \\ & \tilde{a}_{31} & 0 & 0 & 0 \\ & \tilde{a}_{41} & \tilde{a}_{42} & 0 & 0 \\ \hline \widetilde{K} & \tilde{b}_1 & \tilde{b}_2 & 0 & 0 \end{array} \tag{30}$$

We consider these perturbations because, in order to obtain a nonzero SSP coefficient for nonlinear problems, the analysis done in [12] shows that only the entries $\tilde{a}_{31}$, $\tilde{a}_{41}$, $\tilde{a}_{42}$, $\tilde{b}_1$ and $\tilde{b}_2$ in $\widetilde{K}$ need be nonzero. To study SSP coefficients for the linear case, we have to analyze the perturbed stability function, that in this case is of the form

$$\phi_{(K,\widetilde{K})}(z,\tilde{z}) = 1 + z + \frac{1}{2}z^2 + \frac{1}{6}z^3 + \frac{1}{24}z^4 + \beta_1(z+\tilde{z}) + \beta_{11}z(z+\tilde{z}) + \beta_{21}z^2(z+\tilde{z}) \tag{31}$$

where

$$\beta_1 = \tilde{b}_1 + \tilde{b}_2\,, \qquad \beta_{11} = \frac{1}{6}\left(3\,\tilde{b}_2 + 2\,\tilde{a}_{31} + \tilde{a}_{41} + \tilde{a}_{42}\right)\,, \qquad \beta_{21} = \frac{1}{12}\left(2\,\tilde{a}_{31} + \tilde{a}_{42}\right)\,.$$

Next, we construct the Taylor expansion of (31) in terms of a general value r, and we compute the largest r such that all the coefficients in the Taylor expansion are nonnegative; in this case, polynomial (31) is always realizable (in the sense that it corresponds to a perturbation of the form (30)). After some computations, we obtain a coefficient $R_{\text{Lin}}(K,\widetilde{K}) \approx 1.66728$, that is the positive root of the polynomial $15\,x^4 - 4\,x^3 - 12\,x^2 - 24\,x - 24 = 0$, and the coefficients

$$\beta_1 = \frac{7\,r_0^3 - 2\,r_0^2 - 6\,r_0 - 12}{12}\,, \qquad \beta_{11} = \frac{5\,r_0^2 - 2\,r_0 - 6}{12}\,, \qquad \beta_{21} = \frac{r_0 - 1}{6}\,,$$

where $r_0 = R_{\text{Lin}}(K, \widetilde{K})$. With these values, the perturbed stability function can be written as

$$\phi_{(K,\tilde{K})}(z,\tilde{z}) = \gamma_{01}\left(1+\frac{\tilde{z}}{r_0}\right)+\gamma_{11}\left(1+\frac{z}{r_0}\right)\left(1+\frac{\tilde{z}}{r_0}\right)+\gamma_{21}\left(1+\frac{z}{r_0}\right)^2\left(1+\frac{\tilde{z}}{r_0}\right)+\gamma_{40}\left(1+\frac{z}{r_0}\right)^4,$$

where

$$\gamma_{01} = \frac{r_0\left(2\,r_0^3 - r_0^2 - 6\right)}{6}, \quad \gamma_{11} = \frac{r_0^2\left(r_0^2 + 2\,r_0 - 6\right)}{12}, \quad \gamma_{21} = \frac{r_0^3\left(r_0 - 1\right)}{6}, \quad \gamma_{40} = \frac{r_0^4}{24}.$$

This perturbed stability function can be realized with the family of perturbations

$$\tilde{a}_{31} = \frac{1}{2}\left(2\,r_0 - 2 - \tilde{a}_{42}\right), \tilde{a}_{41} = \frac{1}{2}\left(5\,r_0^2 - 6\,r_0 - 2 - 6\,\tilde{b}_2\right), \tilde{b}_1 = \frac{1}{12}\left(7\,r_0^3 - 2\,r_0^2 - 6\,r_0 - 12 - 12\,\tilde{b}_2\right).$$

Observe that $R_{\text{Lin}}(K, \widetilde{K})$ is independent of the choice of $\tilde{a}_{42}$ and $\tilde{b}_2$. Thus, for $\tilde{a}_{42} = \tilde{b}_2 = 0$, we obtain the same value of $R_{\text{Lin}}(K, \widetilde{K})$ with a perturbation (30) whose nontrivial elements are only in the first column of $\widetilde{K}$.

Observe too that the perturbation in (30) does not contain all the possible nonnegative elements in a strictly lower triangular matrix (see Definition 1), and therefore we cannot claim that the value $R_{\text{Lin}}(K, \widetilde{K}) \approx 1.66728$ is the threshold factor of the optimal perturbation $R^{\text{opt}}_{\text{Lin}}(K)$. With the study done, we have that $1.66728 \le R^{\text{opt}}_{\text{Lin}}(K) \le \widetilde{R}_{4,4} = 2$. □

3 Perturbed Runge–Kutta methods for nonlinear problems

In this section we seek to answer the questions posed in Section 1.3 for nonlinear problems. We begin with an introduction section where we collect some known results from the literature.

3.1 Introduction

In this section the introduce some notation used in the rest of the paper and we collect some known results from the literature.

First, it is convenient to write scheme (6) in canonical Shu-Osher form [7]

$$Y = v_r u_n + \alpha_r\left(Y + \frac{h}{r}F\right) \tag{33}$$

where

$$v_r = (I + rK)^{-1}e, \qquad \alpha_r = r(I + rK)^{-1}K. \tag{34}$$

Observe that matrices K and α_r have the same structure (strictly lower triangular, lower triangular or full).

Definition 7 The *radius of absolute monotonicity* of a Runge–Kutta method (6) is the largest r such that v_r and α_r in (34) exist and are non-negative:

$$R(K) = \sup\left\{r \,|\, r = 0 \text{ or } r > 0, (I + rK)^{-1} \text{ exists, and } \alpha_r, v_r \ge 0\right\}. \tag{35}$$

Recall that Definition 7 is the one in [23, Def. 2.4] using the notation given in [12, Eq. (1.21)]. The quantity $R(K)$ is also known as the SSP coefficient or Kraaijevanger coefficient. As usual, the inequalities above should be understood component–wise.

In [6, Thm. 2.5] step size restrictions to obtain monotonicity are given in terms of the radius of absolute monotonicity of the method. Thus, the larger $R(K)$ is, the larger is the step size restriction for monotonicity; in particular, if $R(K) = 0$, numerical monotonicity cannot be ensured.

Next, we consider perturbed Runge–Kutta methods (7). To study absolute monotonicity of perturbed Runge–Kutta methods, we write method (7) also in a canonical Shu-Osher-like form

$$Y = \gamma_r u_n + \alpha_r^{\rm up} \left(Y + \frac{h}{r} F \right) + \alpha_r^{\rm down} \left(Y - \frac{h}{r} \widetilde{F} \right) , \tag{36}$$

where

$$\gamma_r = (I + rK + 2r\widetilde{K})^{-1} e \, , \tag{37a}$$

$$\alpha_r^{\rm up} = r(I + rK + 2r\widetilde{K})^{-1} (K + \widetilde{K}) \, , \tag{37b}$$

$$\alpha_r^{\rm down} = r(I + rK + 2r\widetilde{K})^{-1} \widetilde{K} \, . \tag{37c}$$

Observe that method (36), with $\gamma_r = (I - \alpha_r^{\rm up} - \alpha_r^{\rm down})e$, is a perturbed Runge–Kutta scheme with Butcher coefficients

$$K = \frac{1}{r}(I - \alpha_r^{\rm up} - \alpha_r^{\rm down})^{-1}(\alpha_r^{\rm up} - \alpha_r^{\rm down}) \, , \qquad \widetilde{K} = \frac{1}{r}(I - \alpha_r^{\rm up} - \alpha_r^{\rm down})^{-1} \alpha_r^{\rm down} \, , \tag{38}$$

provided that $(I - \alpha_r^{\rm up} - \alpha_r^{\rm down})^{-1}$ exists.

Definition 8 [12, Def. 3.1] The *radius of absolute monotonicity* of a perturbed Runge–Kutta method $(K, \widetilde{K})$ is the largest r such that γ_r, $\alpha_r^{\rm up}$ and $\alpha_r^{\rm down}$ in (37) exist and are non-negative:

$$R(K, \widetilde{K}) = \sup \left\{ r \, | \, r = 0 \text{ or } \ r > 0, \, (I + rK + 2r\widetilde{K})^{-1} \text{ exists, and } \gamma_r, \, \alpha_r^{\rm up}, \, \alpha_r^{\rm down} \geq 0 \right\} . \tag{39}$$

For perturbation $(K, \widetilde{K})$, step size restrictions to obtain monotonicity are given in terms of $R(K, \widetilde{K})$ [12, Thm. 3.5],

$$h \leq R(K, \widetilde{K}) \, h_0 \, . \tag{40}$$

Thus, perturbations with large values of $R(K, \widetilde{K})$ ensure larger step size restrictions for monotonicity.

Remark 4 (Fictitious perturbations) As it has been pointed out in Section 1.2, if function f in (1) satisfies both (3) and (5), we can formally introduce a function $\tilde{f} = f$ to perturb fictitiously the Runge–Kutta method (see (7)). In this way, the standard (unperturbed) Runge–Kutta method can be written as (36). Thus, the results in this paper can also be used to ensure monotonicity for step size restrictions larger than the ones given in terms of the (unperturbed) SSP coefficient. □

Remark 5 (Property C) Most previous works, including [28, 27], have focused on methods with the following property: for each value of j

$$\widetilde{K}_{ij} \neq 0 \text{ (for some } i) \quad \Longrightarrow \quad K_{ij} = 0 \text{ (for all } i). \tag{41}$$

In this case, we will say that a perturbation $\widetilde{K}$ to a Runge–Kutta method K possesses *property C*. In words, property C means that in the jth column, only one of $K, \widetilde{K}$ has any nonzero entries. Thus, only one of $f(y_j), \tilde{f}(y_j)$ need ever be evaluated, so only s total function evaluations are required per step. In [8] it was shown that for WENO discretizations, the cost of computing both $f(y_j)$ and $\tilde{f}(y_j)$ is much less than twice the cost of computing $f(y_j)$ alone. Therefore methods without property C may also be of practical interest. In the present work, we do not assume property C. □

3.1.1 Zero-well-defined perturbations

Regularity of $(I-\alpha_r^{\rm up}-\alpha_r^{\rm down})$ is evidently important in our study. Observe that from (37) we have

$$(I-\alpha_r^{\rm up}-\alpha_r^{\rm down})(I+rK)=(I-2\alpha_r^{\rm down})\,. \tag{42}$$

Consequently, if $I+rK$ is regular for some r, then $(I-\alpha_r^{\rm up}-\alpha_r^{\rm down})$ is regular if and only if $(I-2\alpha_r^{\rm down})$ is regular.

If $I-2\alpha_r^{\rm down}$ is singular, then the stage equations do not have a unique solution even for the trivial ODE given by $f=0$. This motivates the following deinition.

Definition 9 Let a perturbed Runge–Kutta method (36) be given. If $I-2\alpha_r^{\rm down}$ (defined by (37c)) is non-singular, we say that the perturbation is *zero-well-defined*.

See [7, Chap. 3] for the analogous definition in the context of traditional Runge–Kutta methods.

3.2 Optimal perturbations

In this section we answer question 1 of Section 1.3 by showing that every method can be perturbed so as to give a method with strictly positive SSP coefficient.

Theorem 3 *Let K be a Runge-Kutta method that belongs to a specified class of methods (explicit, diagonally implicit, or fully implicit). Then it is always possible to find a perturbation $\widetilde{K}$ within the same class such that $R(K,\widetilde{K})>0$.*

Thus it makes sense to deal with perturbations that give the largest SSP coefficient. We formalize this idea in the following definition.

Definition 10 The *optimal perturbed SSP coefficient* of a Runge–Kutta method K is denoted by

$$R^{\rm opt}(K)=\sup_{\widetilde{K}} R(K,\widetilde{K})\,.$$

For a given method K that is (explicit/diagonally implicit/fully implicit), we consider the supremum over perturbations $\widetilde{K}$ that are zero-well-defined and correspond to the same class of methods. A matrix $\widetilde{K}$ such that $R(K,\widetilde{K})=R^{\rm opt}(K)$ is called an *optimal perturbation*.

Observe that a perturbed Runge–Kutta method $(K,\widetilde{K})$ can be interpreted as an additive Runge–Kutta method $(K+\widetilde{K},\widetilde{K})$ for functions $(f,\tilde{f})$ (see Remark 1), and conditions (37) are the ones required for the absolute monotonicity of this additive scheme at $(z,\tilde{z})=(-r,-r)$ (see [13]). From Lemma 2.8 in [13], we obtain that the stability function $\phi_{(K,\widetilde{K})}$ defined by (12), is absolutely monotonic at $(\xi,\tilde{\xi})=(-r,-r)$. Consequently,

$$R(K,\widetilde{K})\le R_{\rm Lin}(K,\widetilde{K})\le R^{\rm opt}_{\rm Lin}(K)\,. \tag{43}$$

Furthermore, from SSP theory and inequality (19), we have

$$R(K)\le R(\phi_K)\le R^{\rm opt}_{\rm Lin}(K)\,,\qquad R(K)\le R^{\rm opt}(K)\le R^{\rm opt}_{\rm Lin}(K)\,. \tag{44}$$

The following example illustrates that $R(\phi_K)$ can be either larger or smaller than $R^{\rm opt}(K)$.

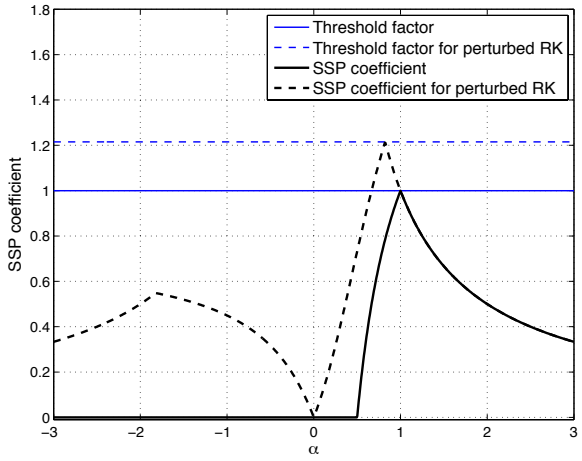

Fig. 2 Family of second order 2-stage methods: SSP coefficients for unperturbed methods and optimal SSP coefficients for perturbed methods.

Example 3 We consider the family of second order 2-stage Runge-Kutta methods (26) for $\alpha \in \mathbb{R}$. For this family we have

$$v_r \geq 0 \iff \alpha > 0 \quad \text{and} \quad 0 \leq r \leq \frac{1}{\alpha},$$
$$\alpha_r \geq 0 \iff \alpha \geq \frac{1}{2} \quad \text{and} \quad 0 \leq r \leq \frac{2\alpha - 1}{\alpha}.$$

Thus

$$R(K) = \begin{cases} 0, & \text{if} \quad \alpha \leq \frac{1}{2}, \\ \frac{2\,\alpha - 1}{\alpha}, & \text{if} \quad \frac{1}{2} < \alpha \leq 1, \\ \frac{1}{\alpha}, & \text{if} \quad 1 < \alpha. \end{cases} \tag{45}$$

In Figure 2 we show the threshold factor $R(\phi_K)$ (thin solid blue line) and the SSP coefficient $R(K)$ (thick solid black line). We also show the corresponding optimal coefficients for perturbed methods, namely, the optimal threshold factor $R^{\text{opt}}_{\text{Lin}}(K)$ (thin dashed blue line) given by (29) in Example 1, and the optimal SSP coefficient $R^{\text{opt}}(K)$ for the perturbed method (thick dashed black line) given by (72).

We see that for optimal SSP method ($\alpha = 1$) it is not possible to increase the SSP coefficient by means of perturbations. However, for $\alpha = (\sqrt{7}-1)/2$ it is possible to obtain a perturbation that raises the SSP coefficient to $R^{\text{opt}}(K) = R^{\text{opt}}_{\text{Lin}}(K) = \left(1+\sqrt{7}\right) \approx 1.21525$ (see (29)).

We have that $R(\phi_K) = R^{\text{opt}}(K) = 1$ for $\alpha = 2/3, 1$. For $2/3 < \alpha < 1$ we obtain $R(\phi_K) < R^{\text{opt}}(K)$, whereas for $0 < \alpha < 2/3$ and for $1 < \alpha$ we have $R^{\text{opt}}(K) < R(\phi_K)$.

Coefficients of the perturbations that give rise to these values are given in Appendix 6.1. □

3.3 Upper bounds on the SSP coefficient for perturbed Runge–Kutta methods

In this section we answer questions 2 and 3 of Section 1.3 We begin by exploring some upper bounds on the SSP coefficient $R^{\text{opt}}(K)$ where K is an s-stage order p Runge–Kutta method. A

straightforward upper bound is obtained from inequality (44) and Theorem 23:

$$R^{\text{opt}}(K) \leq \sqrt[p]{s(s-1)\dots(s-p+1)}\,. \tag{46}$$

Another bound is given by the next Theorem.

Theorem 4 *Consider an explicit Runge-Kutta method K and let r_e be the largest positive value such that vector v_r in* (34) *is non-negative. Then*

$$R^{\text{opt}}(K) \leq r_e\,. \tag{47}$$

From Theorem 4 we obtain that

$$R(K) \leq R^{\text{opt}}(K) \leq r_e\,. \tag{48}$$

Consequently, for those methods such that $R(K) = r_e$, the SSP coefficient cannot be increased by perturbation. This is the case for the family of second-order two-stage methods. For $\alpha \geq 1$, $R(K) = r_e = 1/\alpha$ (see Example 3).

On the other hand, if $R(K) < r_e$ one can try to find a perturbation to increase the SSP coefficient. This is the case for the classical 4-stage order 4 method for which $R(K) = 0$ and $r_e \approx 1.2956$, the real root of $x^3 - 2x^2 + 4x - 4 = 0$.

Finally, another interesting bound, for explicit methods only, can be obtained in terms of the Butcher coefficients of the Runge-Kutta method K.

Theorem 5 *Consider an explicit Runge-Kutta method K with perturbed SSP coefficient $R^{\text{opt}}(K) > 0$. Let $K = (a_{ij})$. Then*

$$R^{\text{opt}}(K) \leq \frac{1}{\max_{ij} |a_{ij}|}\,. \tag{49}$$

Consequently,

$$R(K) \leq R^{\text{opt}}(K) \leq \frac{1}{\max_{ij} |a_{ij}|}\,.$$

For those methods such that $R(K) = 1/\max_{ij} |a_{ij}|$ it is not possible to increase the SSP coefficient by perturbing the method. This is the case for all known optimal explicit SSP Runge–Kutta methods of orders one through three, with any number of stages [18].

For the restricted class of perturbations considered in [28], similar results were obtained in [28, Thms. 3.1, 3.4, 3.5 and 3.6]. Theorem 5 extends those results, showing that no improvement in the radius of absolute monotonicity is possible for many optimal SSP methods, even when more general perturbations are considered. Those methods include all optimal methods of order one or two, the optimal methods of order three with n^2 stages (for any integer n), and the optimal methods with $(s, p) \in \{(5, 3), (6, 3), (10, 4)\}$. Interestingly, the widely-used optimal five-stage fourth-order method is an exception; it has $R(K) \approx 1.508$ while the Theorem 5 gives an upper bound of 1.8349.... Numerical computations suggest that it can be perturbed to achieve $R^{\text{opt}}(K) \approx 1.63979$.

3.4 Relations among the Butcher and canonical Shu-Osher representations

To answer questions 3 and 4 in Section 1.3, two algorithms are proposed. In this section we prove some technical results that justify some steps in Algorithms 1 and 2 below.

For explicit Runge–Kutta methods (or other methods with one stage equal to u_n), some components of vector v_r in (34) can be moved to the first column of matrix α_r [12, Remark 1]. This simple transformation may yield a larger value of $R(K)$ – and never yields a smaller value–. A similar transformation, that may give a larger value of $R(K, \widetilde{K})$ but never smaller values, exists for perturbations of this class of schemes.

Proposition 4 *Let an s-stage explicit perturbed Runge–Kutta method with coefficients $\gamma_r, \alpha_r^{\rm up}, \alpha_r^{\rm down} \geq 0$, be given where r is the radius of absolute monotonicity of the method. Consider the perturbed method with coefficients*

$$\widehat{\gamma} = (1, 0, \ldots, 0)^t, \tag{50a}$$
$$\widehat{\alpha}_{i,1}^{\rm up} = \alpha_{i,1}^{\rm up} + (\gamma_r)_i/2 \qquad 2 \leq i \leq s \tag{50b}$$
$$\widehat{\alpha}_{i,1}^{\rm down} = \alpha_{i,1}^{\rm down} + (\gamma_r)_i/2 \qquad 2 \leq i \leq s. \tag{50c}$$

Then the perturbed method with coefficients $(\gamma_r, \alpha_r^{\rm up}, \alpha_r^{\rm down})$ and the modified perturbed method with coefficients $(\widehat{\gamma}, \widehat{\alpha}^{\rm up}, \widehat{\alpha}^{\rm down})$ correspond to the same Runge–Kutta method K. The modified perturbed method has radius of absolute monotonicity at least equal to r.

Remark 6 Proposition 4 is also valid for Runge-Kutta methods whose first row is equal to zero. □

In the Butcher form (7) it is obvious which perturbed methods $(K, \widetilde{K})$ correspond to a given method K. In the canonical Shu-Osher form it is less obvious. The following lemma characterizes which methods of the form (36) are perturbations of a given method (33).

Lemma 1 *If method* (36) *is a perturbation of method* (33)*, then their coefficients are related as follows:*

$$(I - 2\alpha_r^{\rm down})\alpha_r = (\alpha_r^{\rm up} - \alpha_r^{\rm down}) \tag{51a}$$
$$(I - 2\alpha_r^{\rm down})v_r = \gamma_r. \tag{51b}$$

Furthermore, if (51) *holds and the perturbation is zero-well-defined, then* (36) *is a perturbation of* (33)*.*

Lemma 1 does *not* imply that the perturbation (36) is unique for a given r; see Proposition 4.

Remark 7 The necessity of the zero-well-defined condition in the second part of Lemma 1 can be seen from the following example. We take the implicit trapezoidal Runge-Kutta method

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1/2 & 1/2 & 1/2 \\ \hline & 1/2 & 1/2 \end{array}.$$

The canonical form (33) is then

$$\alpha_r = \begin{pmatrix} 0 & 0 & 0 \\ \frac{r}{r+2} & \frac{r}{r+2} & 0 \\ \frac{r}{r+2} & \frac{r}{r+2} & 0 \end{pmatrix}, \qquad v_r = \begin{pmatrix} 1 \\ \frac{2-r}{r+2} \\ \frac{2-r}{r+2} \end{pmatrix}.$$

Then (51) is satisfied – for any r – by

$$\alpha^{\rm up} = \alpha^{\rm down} = \begin{pmatrix} 1/3 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 1/2 & 0 \end{pmatrix}, \qquad \gamma = \begin{pmatrix} 1/3 \\ 0 \\ 0 \end{pmatrix}.$$

However, this method – which involves a perturbation that is not zero-well-defined – is not a perturbation of the original method. □

3.5 Computing optimal perturbations

In this section we present two algorithms to symbolically or numerically find the optimal perturbed SSP coefficient and a corresponding perturbation of a given Runge–Kutta method. The first algorithm is proven to approximate the optimal value to any accuracy, contingent on the computational solution of linear program subproblems. It is only valid for explicit perturbations. The second algorithm uses no floating-point approximations, and can be applied to both explicit and implicit methods, but it is not proven to give the optimal value. The results of the two algorithms coincide (to high precision) for all explicit methods on which we have tested them.

3.5.1 Provably correct algorithm for finding optimal explicit perturbations

In the foregoing, we have shown that finding an optimal perturbation consists of determining the largest r such that there exists a splitting satisfying (51) with positive coefficients. Note that the range of values for which a method $(K, \widetilde{K})$ is absolutely monotonic is always the interval $[0, R(K, \widetilde{K})]$. Therefore, one way to find the largest r is to devise a method for testing for a given r whether there exists a perturbation $\widetilde{K}$ such that $R(K, \widetilde{K}) \geq r$. For given method (6) and value of r, the system of equations (51) together with the inequalities $\alpha_r^{\text{up}}, \alpha_r^{\text{down}} \geq 0$ constitutes a linear programming (LP) feasibility problem. The following theorem is an immediate consequence of Lemma 1.

Theorem 6 *Let an s-stage Runge–Kutta method K and a positive number r be given. There exists a perturbation $\widetilde{K}$ with $R(K, \widetilde{K}) \geq r$ if and only if there exists an $(s+1) \times (s+1)$ matrix α_r^{down} such that $(I - 2\alpha_r^{\text{down}})$ is regular and the following componentwise inequalities hold:*

$$(I - 2\alpha_r^{down})\alpha_r + \alpha_r^{down} \geq 0 \tag{52a}$$

$$(I - 2\alpha_r^{down})v_r \geq 0 \tag{52b}$$

$$\alpha_r^{\text{down}} \geq 0. \tag{52c}$$

The linear program (52) can be solved by standard LP solvers. By embedding this solution in a one-dimensional root-finding algorithm, optimal perturbations can be found. An algorithm based on bisection is given as Algorithm 1. For a prescribed tolerance ϵ, it returns a value that is less than or equal to optimal perturbed radius of absolute monotonicity, and is within ϵ of that value.

Algorithm 1 Optimal explicit perturbation

Input: K, ϵ
  $r_{\max} := 1/\max|a_{ij}|, r_{\min} := 0.$
  while $r_{\max} - r_{\min} > \epsilon$ **do**
    $r = (r_{\max} + r_{\min})/2.$
    Compute the coefficient matrices α_r, v_r using (34).
    Solve the LP given by (52).
    if it is feasible **then**
      $r_{\min} := r$
    else
      $r_{\max} := r$
    end if
  end while
  return $r_{\min}$

Assuming the solution of the LP is correct, the algorithm provably finds an optimal explicit perturbation. However, for implicit perturbations the LP solver may converge to a solution (like the method in Remark 7 above) for which $I - 2\alpha_r^{\text{down}}$ is singular.

3.5.2 Iterated splitting algorithm

We next investigate how to choose $\alpha_r^{\rm up}, \alpha_r^{\rm down}$ directly so as to find a perturbation with radius of a.m. at least r. The following result suggests an approach.

Lemma 2 *Given an explicit Runge–Kutta method* (33)*, let* $\alpha_r^{\rm up} \geq 0, \alpha_r^{\rm down} \geq 0$ *denote coefficients of a zero-well-defined perturbation of* (33)*. Then there exist matrices* $\alpha^+ \geq 0$*,* $\alpha^- \geq 0$*, such that*

$$\alpha_r^{\rm up} = (I + 2\alpha^-)^{-1}\alpha^+ , \qquad \alpha_r^{\rm down} = (I + 2\alpha^-)^{-1}\alpha^-, \tag{53}$$

and $\alpha_r = \alpha^+ - \alpha^-$.

Thus, zero-well defined perturbations with $\alpha_r^{\rm up} \geq 0, \alpha_r^{\rm down} \geq 0$ come from splittings of the matrix α_r and expressions (53).

In the next algorithm we use the following notation:

$$((x)^+)_{ij} = \begin{cases} x_{ij} & \text{if } x_{ij} \geq 0 \\ 0 & \text{if } x_{ij} < 0. \end{cases} \qquad ((x)^-)_{ij} = \begin{cases} 0 & \text{if } x_{ij} \geq 0 \\ -x_{ij} & \text{if } x_{ij} < 0\,, \end{cases}$$

and thus $x = (x)^+ - (x)^-$ is a sign splitting of matrix x, with $(x)^+ \geq 0$, $(x)^- \geq 0$.

Given a perturbed Runge-Kutta method (36) with $\gamma_r = e_1$, where $e_1 = (1, 0, \ldots, 0)^t$, and $\alpha^{\rm up}$ or $\alpha^{\rm down}$ containing negative values, we construct

$$\tilde{\gamma}_r = \left(I + 2\,(\alpha_r^{\rm up})^- + 2\,(\alpha_r^{\rm down})^-\right)^{-1} e_1\,, \tag{54a}$$

$$\tilde{\alpha}_r^{\rm up} = \left(I + 2\,(\alpha_r^{\rm up})^- + 2\,(\alpha_r^{\rm down})^-\right)^{-1} \left((\alpha_r^{\rm up})^+ + (\alpha_r^{\rm down})^-\right), \tag{54b}$$

$$\tilde{\alpha}_r^{\rm down} = \left(I + 2\,(\alpha_r^{\rm up})^- + 2\,(\alpha_r^{\rm down})^-\right)^{-1} \left((\alpha_r^{\rm up})^- + (\alpha_r^{\rm down})^+\right), \tag{54c}$$

where $\alpha^{\rm up} = (\alpha_r^{\rm up})^+ - (\alpha_r^{\rm up})^-$, $\alpha_r^{\rm down} = (\alpha_r^{\rm down})^+ - (\alpha_r^{\rm down})^-$, provided that $I + 2(\alpha_r^{\rm up})^- + 2(\alpha_r^{\rm down})^-$ is invertible. Using Lemma 1, it is straightforward to prove that, if method $\alpha_r^{\rm up}$, $\alpha_r^{\rm down}$ is a perturbation of (33), then (54) is also perturbation of (33).

Next, for explicit methods, we perform transformation (50). In this way, (54) followed by transformation (50) gives a perturbation of the form (36) with $\gamma_r = e_1$, that we denote by $\hat{\alpha}_r^{\rm up}$, $\hat{\alpha}_r^{\rm down}$. If $\hat{\alpha}_r^{\rm up} \geq 0$, $\hat{\alpha}_r^{\rm down} \geq 0$, then r is an SSP coefficient; otherwise, we can repeat the above process.

The following lemma studies the sign of $(\hat{\alpha}_r^{\rm up})_{ij}$, $(\hat{\alpha}_r^{\rm down})_{ij}$ when $(\alpha_r^{\rm up})_{ij} < 0$ or $(\alpha_r^{\rm down})_{ij} < 0$. For the sake of clarity, we drop the index r.

Lemma 3 *We consider a perturbed explicit Runge-Kutta method with coefficients* $\gamma = e_1$*,* $\alpha^{\rm up}$*,* $\alpha^{\rm down}$*, and the perturbation* $\hat{\alpha}^{\rm up}$*,* $\hat{\alpha}^{\rm down}$ *obtained by computing* (54) *followed by transformation* (50)*. Assume that* $j_0 \geq 2$ *is the first row with negative terms in* $\alpha^{\rm up}$ *or* $\alpha^{\rm down}$*. Let* m_0 *be the largest index* $m_0 \geq 1$ *such that* $\alpha_{j_0,m_0}^{\rm up} < 0$ *or* $\alpha_{j_0,m_0}^{\rm down} < 0$*. Then*

1. *For first to* $(j_0 - 1)$*-th row, we have:* $\hat{\alpha}_{i,j}^{\rm up} = \alpha_{i,j}^{\rm up}$ *and* $\hat{\alpha}_{i,j}^{\rm down} = 0$ *for* $1 \leq i \leq j_0 - 1$*,* $1 \leq j \leq j_0 - 2$*.*
2. *For the* j_0*-th row, we have:*
   (a) *If* $m_0 = 1$*, then* $\hat{\alpha}_{j_0,1}^{\rm up} < 0$ *or* $\hat{\alpha}_{j_0,1}^{\rm down} < 0$*.*
   (b) *If* $m_0 \geq 2$*, then,* $\hat{\alpha}_{j_0,m_0}^{\rm up} \geq 0$ *and* $\hat{\alpha}_{j_0,m_0}^{\rm down} \geq 0$*.*
   (c) *For* $1 \leq m_0 \leq j_0 - 2$*, we have* $\hat{\alpha}_{j_0,\ell}^{\rm up} \geq 0$ *and* $\hat{\alpha}_{j_0,\ell}^{\rm down} \geq 0$ *for* $\ell = m_0 + 1, \ldots, j_0 - 1$*.*

Consequently, if matrices $\alpha_r^{\rm up}$ and $\alpha_r^{\rm down}$ contain negative elements in the second or later columns, an iterated construction of perturbations $\hat\alpha_r^{\rm up}$, $\hat\alpha_r^{\rm down}$ removes these negative values obtaining a perturbation with non-negative elements from second column on. However, if in a row j_0 we have:

$$\alpha^{\rm up}_{j_0,1} < 0 \qquad \text{and} \qquad \alpha^{\rm up}_{j_0,\ell} \geq 0 \qquad \ell = 2, \ldots, j_0 - 1\,, \tag{55}$$

or

$$\alpha^{\rm down}_{j_0,1} < 0 \qquad \text{and} \qquad \alpha^{\rm down}_{j_0,\ell} \geq 0 \qquad \ell = 2, \ldots, j_0 - 1\,, \tag{56}$$

the new perturbation $\hat\alpha_r^{\rm up}$, $\hat\alpha_r^{\rm down}$ will also contain negative elements in the first column.

We now give Algorithm 2 to determine whether there exists a perturbation with a.m. radius r for a given method.

Algorithm 2 Existence of a perturbation with radius r

Input: r, K

Compute the coefficient matrices α_r, v_r using (34).
Set $\alpha^{\rm up} = \alpha_r$ and $\alpha^{\rm down} = 0$.
while $\alpha^{\rm up}$ or $\alpha^{\rm down}$ has any negative entries **do**
If K has a zero row, perform the transformation (50).
If $\alpha^{\rm up}, \alpha^{\rm down} \geq 0$, stop. This is a feasible perturbation.
If condition (55) or (56) hold, stop. A feasible perturbation cannot be found.
Set $\alpha^- = (\alpha^{\rm up})^- + (\alpha^{\rm down})^+$ and $\alpha^+ = (\alpha^{\rm up})^+ + (\alpha^{\rm down})^-$
Compute a new splitting:

$$\alpha^{\rm up} = \left(I + 2((\alpha^{\rm up})^- + (\alpha^{\rm down})^-)\right)^{-1} \alpha^+$$
$$\alpha^{\rm down} = \left(I + 2((\alpha^{\rm up})^- + (\alpha^{\rm down})^-)\right)^{-1} \alpha^-$$

end while

Remark 8 The difficulty in proving the correctness of Algorithm 2 for explicit methods is that one could use $(\alpha_r)^+ + \delta$, $(\alpha_r)^- + \delta$, in place of $(\alpha_r)^+, (\alpha_r)^-$, where δ is any non-negative matrix. □

3.6 Examples and numerical tests

3.6.1 Examples

In this section we compute optimal perturbations of some existing methods, using the algorithms described in the last section.

We have computed optimal perturbations for several known explicit methods using the two algorithms described above. In all cases, the two algorithms gave the same values. It thus seems possible that Algorithm 2 also gives truly optimal results in general, but we do not have a proof. Properties of the methods studied are given in Table 3. The values found have been truncated to three decimal places but are known to greater precision.

For the 4-stage, order-four method of Kutta, the three-digit value of $R^{\rm opt}(K)$ given in the table matches the value found by Shu and Osher. However, the exact (irrational) value is slightly larger and is given in the Appendix. The methods SSP75, SSP85, and SSP95 are optimal methods found in [28], with property C (see Remark 5). By considering perturbed methods without property C, we obtain slightly larger coefficients for perturbations of SSP75 and SSP95. On the other hand, relaxing the column assumption gives no benefit in the case of the SSP85 method.

Order	Stages	Method	$R(K)$	$R^{\text{opt}}(K)$	Bound (49)	Bound (46)	Property C (41)
1	1	Forward Euler	1	1	1	1	True
2	2	Midpoint	0	0.732	1	1.414	True
	2	Min. trunc. error	0.5	1	1.333	1.414	True
	2	SSP22 [29]	1	1	1	1.414	True
	2	SSP22* [8]	0.784	1.215	1.215	1.414	True
3	3	Heun33 [11]	0	0.776	1.333	1.817	False
	3	SSP33 [29]	1	1	1	1.817	True
4	4	RK44 (Kutta)	0	0.685	1	2.213	False
	5	Merson [25]	0	0.242	0.5	3.309	False
	10	SSP104 [18]	6	6	6	8.425	False
5	6	Fehlberg [5]	0	0.057	0.125	3.727	False
	7	Dormand-Prince [4]	0	0.040	0.086	4.789	False
	7*	Bogacki-Shampine [1]	0	0.313	0.859	5.827	False
	7	SSP75 [28]	0	1.396	1.792	4.789	False
	8	SSP85 [28]	0	1.875	1.919	5.827	True
	9	SSP95 [28]	0	2.738	3.198	6.853	False
6	9	Calvo [2]	0	0.021	0.059	6.265	False
8	13	Prince-Dormand [26]	0	0.013	0.059	9.212	False

Table 3 Properties of some Runge–Kutta methods and their optimal perturbations. The optimal perturbed radius of absolute monotonicity was computed by both the linear programming algorithm and the iterated splitting algorithm; in every case they gave identical results (up to roundoff errors). Decimal values have been truncated to the number of digits shown. The Bogacki-Shampine method uses 8 stages but is first-same-as-last, so it is as efficient as if it had 7 stages for non-rejected steps.

Order	Stages	Method	$R(K)$	$h_{[0,1]}$	$R^{\text{opt}}(K)$	$\tilde{h}_{[0,1]}$
1	1	Forward Euler	1	1.00	1	1.00
2	2	Midpoint	0	0.02	0.732	0.73
	2	Min. trunc. error	0.5	0.54	1	1.00
	2	SSP22 [29]	1	1.09	1	1.00
	2	SSP22* [8]	0.784	0.81	1.215	1.21
3	3	Heun33 [11]	0	0.00	0.776	0.90
	3	SSP33 [29]	1	1.00	1	1.00
4	4	RK44 (Kutta)	0	0.17	0.685	0.68
	5	Merson [25]	0	0.01	0.242	0.29
	10	SSP104 [18]	6	6.04	6	6.00
5	6	Fehlberg [5]	0	0.01	0.057	0.05
	7	Dormand-Prince [4]	0	0.02	0.040	0.04
	7*	Bogacki [1]	0	0.06	0.313	0.31
	7	SSP75 [28]	0	0.06	1.396	1.56
	8	SSP85 [28]	0	0.08	1.875	1.87
	9	SSP95 [28]	0	0.10	2.738	2.82
6	9	Calvo [2]	0	0.04	0.021	0.02
8	13	Prince-Dormand [26]	0	0.06	0.013	0.01

Table 4 Problem (57): actual and predicted step size restrictions for preservation of the numerical solution in the interval $[0, 1]$; $R(K)$ and $R^{\text{opt}}(K)$ are the SSP coefficient and the optimal perturbed SSP coefficient, respectively, and $h_{[0,1]}$ and $\tilde{h}_{[0,1]}$ are the largest step sizes that preserve the interval $[0, 1]$ in practice for the unperturbed method and the optimal perturbed method, respectively.

3.6.2 Numerical test

We also apply the methods – both perturbed and unperturbed – to the variable-coefficient advection problem

$$\begin{aligned} u_t + (a(x,t)u)_x &= 0\,, \\ u(0,t) &= 0\,, \\ u(x,0) &= g(x)\,, \\ a(x,t) &= \cos^4(200x + 400t)\,, \end{aligned} \tag{57}$$

representing a highly oscillatory flow field. If $g(x) \in [0,1]$, then the exact solution remains in $[0,1]$ for all time. We consider the domain $0 < x < 1$ and we semi-discretize using first-order upwind differencing in space on an equispaced grid with 20 points. The forward invariance of the interval $[0,1]$ is then preserved by the explicit Euler method as long as $0 \le h \le 1$. Application of any Runge–Kutta method to this semi-discretization yields an iteration of the form

$$u_{n+1} = M(t_n, h)u_n$$

where $M(t_n, h)$ is a square matrix. The initial vector u_0 is obtained from $g(x)$ at the spatial grid points; hence, if $g(x) \in [0,1]$, vector u_0 is also in $[0,1]$. Consequently, the numerical solution of (57) will remain in $[0,1]$ if $g(x) \in [0,1]$ and M satisfies

$$m_{ij} \ge 0 \qquad \text{for all } i,j \tag{58a}$$

$$\sum_j m_{ij} \le 1 \qquad \text{for all } i. \tag{58b}$$

In Table 4, in the column labeled $h_{[0,1]}$, we give the largest step size for which the corresponding unperturbed method preserves the interval $[0,1]$; i.e. the largest step size for which M satisfies (58). Similarly, in the column labeled $\tilde{h}_{[0,1]}$, we give the largest step size for which the optimal perturbation of the method (with first-order downwind differencing for the downwind operator) preserves the interval $[0,1]$. The values given are truncated (not rounded) to two decimal places. Most of the actual values agree very well with the theoretical bounds.

Some additional interesting patterns are evident in the table and are discussed in the conclusions below.

4 Conclusions

In this work we have studied SSP coefficients for perturbations of a given explicit Runge-Kutta method. We have considered both the linear and the nonlinear case, and have obtained useful bounds on the threshold factor and on the radius of absolute monotonicity for perturbed Runge–Kutta methods. We have also provided an algorithm for computing optimal perturbations of explicit Runge-Kutta methods, and given optimal perturbations for many methods from the literature. From Table 4 we see that

- For most optimal SSP methods (up to order three), perturbation cannot yield a larger coefficient. This is evident already from the bound (49). For all other methods, some improvement is achieved.
- Consistent with Theorem 5, for every method considered, it is possible to achieve $R^{\text{opt}} > 0$ by some perturbation.
- The simple bound (49) predicts the optimal coefficient to within a factor of three in every case.

This work seems to provide a complete picture for the case of most interest: explicit methods applied to nonlinear problems. Nevertheless, some other interesting issues remain unsolved. These include:

- A method to compute optimal perturbations for linear problems.
- An algorithm for obtaining optimal splittings of implicit methods.

Besides, in this paper we have only considered perturbations $\tilde{f}$ such that $\tilde{h}_0 = h_0$ (see (3)-(5)), but the study done can be extended to the case $\tilde{h}_0 \neq h_0$. In this way, a wider class of perturbations $\tilde{f}$ can be considered and larger SSP coefficients may be obtained. In a similar way, for fictitious perturbations (see Remark 4), monotonicity can be ensured with step size restrictions larger than the ones obtained with the results in this paper.

These may be a starting point for future work.

5 Proofs of the results in the paper

This section contains the proofs of the different results in the paper (Theorems 1-5, Propositions 1 and 4, and Lemmas 1-3), and an auxiliary lemma; the proofs of Propositions 2 and 3 and Theorem 6 are straightforward and they are omitted.

*Proof of **Theorem 1.***
The stability function $\phi_{(K,\widetilde{K})}(z,\tilde{z})$ is a bivariate polynomial and thus, for $r \leq R(\phi_{(K,\widetilde{K})})$, it can be written in the form (15), where the coefficients $\gamma_{j\ell}$ are non-negative and (by consistency of the method) sum to unity. Letting $z = hL$ and $\tilde{z} = h\tilde{L}$, applying $\|\cdot\|$, and using convexity shows that $\|\phi_{(K,\widetilde{K})}(hL, -h\tilde{L})u\| \leq \|u\|$. □

*Proof of **Proposition 1.***
From (13), the stability function (12) of any s-stage explicit perturbed Runge–Kutta method is a bivariate polynomial of combined degree at most s. Furthermore, as $\phi_{(K,\widetilde{K})}(z,-z)$ is the stability function of an s-stage Runge–Kutta scheme of linear order p, we have that $\phi_{(K,\widetilde{K})}(z,-z) = \sum_{j=0}^{p} z^j/j! + \sum_{j=p+1}^{s} \sigma_j z^j$. Thus, there is a bivariate polynomial Ψ such that $\phi_{(K,\widetilde{K})}(z,\tilde{z}) = \phi_{(K,\widetilde{K})}(z,-z) + (z+\tilde{z})\Psi(z,\tilde{z})$. As $\phi_{(K,\widetilde{K})}$ has combined degree at most s, trivially Ψ is a polynomial of combined degree at most $s-1$. □

Lemma 4 *Let $\varphi(z)$ be a polynomial satisfying*

$$\begin{aligned} \varphi(z) &= 1 + \gamma_1 z + \cdots + \gamma_p z^p + \gamma_{p+1} z^{p+1} + \cdots + \gamma_s z^s \\ \gamma_j &\geq \frac{1}{j!}, \quad j = 1, \ldots, p\,. \end{aligned} \tag{59}$$

Then the radius of absolute monotonicity of φ satisfies

$$R(\varphi) \leq \sqrt[p]{s(s-1)\ldots(s-p+1)}\,. \tag{60}$$

*Proof of **Lemma 4.***
If $R(\varphi) = 0$, inequality (60) is trivial. Let $\varphi(z)$ satisfy (59) and be absolutely monotonic at $-r$ with $r > 0$. Then it can be written as

$$\varphi(z) = \sum_{j=0}^{s} \alpha_j \left(1 + \frac{z}{r}\right)^j = \sum_{j=0}^{s} \alpha_j \left(\sum_{\ell=0}^{j} \frac{z^\ell}{r^\ell} \binom{j}{\ell}\right) = \sum_{\ell=0}^{s} \left(\sum_{j=\ell}^{s} \alpha_j \binom{j}{\ell}\right) \frac{z^\ell}{r^\ell}\,, \tag{61}$$

where $\alpha_j \geq 0$. Observe that from (59) we get $\varphi(0) = 1$, and thus in (61) we have $\sum_j \alpha_j = 1$. As φ is of the form (59), the coefficient of z^p is larger than $1/p!$. Some computations give

$$\frac{1}{p!} \leq \left(\sum_{j=p}^{s} \alpha_j \binom{j}{p} \right) \frac{1}{r^p} \leq \left(\sum_{j=p}^{s} \alpha_j \right) \binom{s}{p} \frac{1}{r^p} \leq \binom{s}{p} \frac{1}{r^p} = \frac{s\,(s-1)\cdots(s-p+1)}{p!\; r^p}.$$

Consequently, $r \leq \sqrt[p]{s(s-1)\cdots(s-p+1)}$. □

We remark that equality in (60) is obtained for the polynomial

$$\varphi(z) = \left(1 + \frac{z}{r}\right)^s, \tag{62}$$

where $r = \sqrt[p]{s(s-1)\cdots(s-p+1)}$.

Proof of ***Theorem 2.***
If $R(\psi) = 0$ for all $\psi \in \widetilde{\Pi}_{s,p}$, then $\widetilde{R}_{s,p} = 0$ and inequality (23) is true. Otherwise, there exists a function $\psi \in \widetilde{\Pi}_{s,p}$ a.m. at $(-r, -r)$ with $r > 0$. By [13, Lemmas 2.9 and 2.10], ψ is a.m. at the points (ξ, ξ), with $\xi \in [-r, 0]$. Writing $\psi(z, \tilde{z}) = \sum\sum \mu_{jk} z^j \tilde{z}^k$ and differentiating shows that all coefficients μ_{jk} are non-negative since ψ is a.m. at $(0, 0)$. Thus $\psi(z, z)$ (viewed as a function of one variable) is of the form (59) and is a.m. at $-r$. Application of Lemma 4 gives the desired result. □

Proof of ***Theorem 3.***
From [12, Prop. 3.7], we have $R(K, \widetilde{K}) > 0$ if and only if the Butcher coefficients satisfy

$$K + \widetilde{K} \geq 0\,, \qquad \widetilde{K} \geq 0\,, \tag{63}$$

and the following inequalities hold,

$$\text{Inc}\,((K + 2\widetilde{K})(K + \widetilde{K})) \leq \text{Inc}\,(K + \widetilde{K})\,, \tag{64a}$$

$$\text{Inc}\,((K + 2\widetilde{K})\widetilde{K}) \leq \text{Inc}\,(\widetilde{K})\,, \tag{64b}$$

where Inc (F) denotes the incidence matrix of matrix F defined as Inc $(F) = (g_{ij})$ where $g_{ij} = 1$ if $f_{ij} \neq 0$, and $g_{ij} = 0$ if $f_{ij} = 0$.

Consider first the implicit case. By making all entries of $\widetilde{K}$ positive we can satisfy (64), and by making them large enough we can satisfy (63). For the explicit and diagonally implicit cases, note that if $K, \widetilde{K}$ are (strictly) lower-triangular, then the left-hand sides of (64) are also. Thus by making all the (strictly) lower-triangular entries of $\widetilde{K}$ positive, and by taking them large enough, we can satisfy the above inequalities. □

Proof of ***Theorem 4.***
Let $r = R^{\text{opt}}(K)$. Then $\gamma_r = (I - \alpha^{\text{up}} - \alpha^{\text{down}})e \geq 0$, and thus from (34) and (42) we get

$$(I - 2\alpha_r^{\text{down}})v_r \geq 0\,. \tag{65}$$

As $\alpha_r^{\text{down}} \geq 0$, and since we consider only explicit, zero-well-defined perturbations, $I - 2\alpha_r^{\text{down}}$ is an M matrix. Thus $(I - 2\alpha_r^{\text{down}})^{-1} \geq 0$. If we multiply (65) by $(I - 2\alpha_r^{\text{down}})^{-1}$ we obtain that $v_r \geq 0$. □

Proof of ***Theorem 5.***
The proof is similar to that of [28, Lemma 3.2]. Consider an optimal perturbation $\widetilde{K}$ and set $r = R^{\text{opt}}(K, \widetilde{K}) > 0$; consider too the canonical representation (36). Let $\Lambda = \alpha_r^{\text{up}} + \alpha_r^{\text{down}} = (\alpha_{ij})$,

$\Gamma = \alpha_r^{\text{up}}/r = (\beta_{ij})$, $\tilde{\Gamma} = \alpha_r^{\text{down}}/r = (\tilde{\beta}_{ij})$; observe that $\Lambda, \Gamma, \tilde{\Gamma} \geq 0$, and that $\Lambda = r(\Gamma + \tilde{\Gamma})$. As $(I - \Lambda)e = \gamma_r \geq 0$ and $\alpha_{ik} \geq 0$, we have $\alpha_{ik} \leq 1$; as $(I - \Lambda)K = \Gamma - \tilde{\Gamma}$, we have

$$a_{ik} = \beta_{ik} - \tilde{\beta}_{ik} + \sum_{j=k+1}^{i-1} \alpha_{ij} a_{jk} \,. \tag{66}$$

As $\alpha_{ik} = r(\beta_{ik} + \tilde{\beta}_{ik})$, then $\beta_{ik} + \tilde{\beta}_{ik} = \alpha_{ik}/r \leq 1/r$. In particular, from (66),

$$|a_{21}| = \left|\beta_{21} - \tilde{\beta}_{21}\right| \leq \beta_{21} + \tilde{\beta}_{21} \leq \frac{1}{r} \,.$$

We proceed by induction on row ℓ of K. Assume that $|a_{ij}| \leq 1/r$, for $i = 2, \ldots, \ell$, $j = 1, \ldots, \ell - 1$, and consider row $\ell + 1$. Then, from (66),

$$\begin{aligned} |a_{\ell+1,1}| &= \left| \beta_{\ell+1,1} - \tilde{\beta}_{\ell+1,1} + \sum_{j=2}^{\ell} \alpha_{\ell+1,j}\, a_{j,1} \right| \leq \beta_{\ell+1,1} + \tilde{\beta}_{\ell+1,1} + \sum_{j=2}^{\ell} \alpha_{\ell+1,j} |\, a_{j,1}| \\ &\leq \frac{1}{r}\alpha_{\ell+1,1} + \frac{1}{r}\sum_{j=2}^{\ell} \alpha_{\ell+1,j} \leq \frac{1}{r}\sum_{j=1}^{\ell} \alpha_{\ell+1,j} \leq \frac{1}{r} \,. \end{aligned}$$

A similar argument can be used to show that $|a_{\ell+1,j}| \leq 1/r$, $j = 2, \ldots, \ell$. The Theorem follows by induction. □

Proof of ***Proposition 4.***
It is easily seen that the modified method is equivalent to the original one when $f = \tilde{f}$, so they correspond to the same unperturbed method. Meanwhile, the transformation never leads to negative coefficients, so the modified method is a.m. at r. □

Proof of ***Lemma 1.***
To prove the first part, take $\tilde{f} = f$ in (36) to obtain:

$$Y = \gamma_r u_n + (\alpha_r^{\text{up}} + \alpha_r^{\text{down}})Y + (\alpha_r^{\text{up}} - \alpha_r^{\text{down}})\frac{h}{r}F.$$

Subtract $2\alpha_r^{\text{down}}Y$ from both sides to get

$$(I - 2\alpha_r^{\text{down}})Y = \gamma_r u_n + (\alpha_r^{\text{up}} - \alpha_r^{\text{down}}) \left(Y + \frac{h}{r}F \right). \tag{67}$$

Substituting (33) in the above gives

$$(I - 2\alpha_r^{\text{down}})v_r u_n + (I - 2\alpha_r^{\text{down}})\alpha_r \left(Y + \frac{h}{r}F \right) = \gamma_r u_n + (\alpha_r^{\text{up}} - \alpha_r^{\text{down}}) \left(Y + \frac{h}{r}F \right),$$

Equating coefficients yields (51).

To prove the second part, assume $I - 2\alpha_r^{\text{down}}$ is invertible and write (51) as

$$\alpha_r = (I - 2\alpha_r^{\text{down}})^{-1}(\alpha_r^{\text{up}} - \alpha_r^{\text{down}}) \tag{68a}$$

$$v_r = (I - 2\alpha_r^{\text{down}})^{-1}\gamma_r. \tag{68b}$$

Substitute (68) in (33), multiply on the left by $(I - 2\alpha^{\text{down}})^{-1}$, and follow the steps above in reverse. □

Proof of ***Lemma 2.***
Since the perturbation is zero-well-defined, we can define

$$\alpha^+ = (I - 2\alpha^{\text{down}})^{-1}\alpha_r^{\text{up}}\,, \qquad \alpha^- = (I - 2\alpha^{\text{down}})^{-1}\alpha_r^{\text{down}}\,. \tag{69}$$

Then, by (51a), $\alpha_r = \alpha^+ - \alpha^-$. Furthermore, since $I - 2\alpha_r^{\text{down}}$ is an M-matrix, we have $\alpha^+ \geq 0$ and $\alpha^- \geq 0$. Solving (69) for $\alpha_r^{\text{up}}, \alpha_r^{\text{down}}$ gives (53). □

Proof of ***Lemma 3.***
If j_0 is the first row with negative terms in α^{up} or α^{down}, straightforward computations give that $\hat{\alpha}_{i,j}^{\text{up}} = \alpha_{i,j}^{\text{up}}$ and $\hat{\alpha}_{i,j}^{\text{down}} = 0$ for $1 \leq i \leq j_0 - 1$, $1 \leq j \leq j_0 - 2$, and

$$\hat{\alpha}_{j_0,1}^{\text{up}} = \alpha_{j_0,1}^{\text{up}} - 2\sum_{i=2}^{j_0-1}\left((\alpha_{j_0,i}^{\text{up}})^- + (\alpha_{j_0,i}^{\text{down}})^-\right)\alpha_{i,1}^{\text{up}}\,, \tag{70a}$$

$$\hat{\alpha}_{j_0,\ell}^{\text{up}} = (\alpha_{j_0,\ell}^{\text{up}})^+ + (\alpha_{j_0,\ell}^{\text{down}})^- - 2\sum_{i=\ell+1}^{j_0-1}\left((\alpha_{j_0,i}^{\text{up}})^- + (\alpha_{j_0,i}^{\text{down}})^-\right)\alpha_{i,\ell}^{\text{up}}\,, \quad \ell = 2,\dots,j_0-1\,. \tag{70b}$$

and

$$\hat{\alpha}_{j_0,1}^{\text{down}} = \alpha_{j_0,1}^{\text{down}} - 2\sum_{i=2}^{j_0-1}\left((\alpha_{j_0,i}^{\text{up}})^- + (\alpha_{j_0,i}^{\text{down}})^-\right)\alpha_{i,1}^{\text{down}}\,, \tag{71a}$$

$$\hat{\alpha}_{j_0,\ell}^{\text{down}} = (\alpha_{j_0,\ell}^{\text{up}})^- + (\alpha_{j_0,\ell}^{\text{down}})^+ - 2\sum_{i=\ell+1}^{j_0-1}\left((\alpha_{j_0,i}^{\text{up}})^- + (\alpha_{j_0,i}^{\text{down}})^-\right)\alpha_{i,\ell}^{\text{down}}\,, \quad \ell = 2,\dots,j_0-1\,. \tag{71b}$$

Let m_0 be the largest index $m_0 \geq 1$ such that $\alpha_{j_0,m_0}^{\text{up}} < 0$ or $\alpha_{j_0,m_0}^{\text{down}} < 0$. In this case, $\alpha_{j_0,i}^{\text{up}} \geq 0$, $\alpha_{j_0,i}^{\text{down}} \geq 0$ for $i = m_0 + 1,\dots,j_0 - 1$, and thus

$$(\alpha_{j_0,i}^{\text{up}})^+ = \alpha_{j_0,i}^{\text{up}}\,, \quad (\alpha_{j_0,i}^{\text{down}})^+ = \alpha_{j_0,i}^{\text{down}}\,, \quad (\alpha_{j_0,i}^{\text{up}})^- = (\alpha_{j_0,i}^{\text{down}})^- = 0\,, \qquad i = m_0+1,\dots,j_0-1\,.$$

If $m_0 = 1$, from (70a) and (71a) we get $\hat{\alpha}_{j_0,1}^{\text{up}} = \alpha_{j_0,1}^{\text{up}}$ and $\hat{\alpha}_{j_0,1}^{\text{down}} = \alpha_{j_0,1}^{\text{down}}$, and thus $\hat{\alpha}_{j_0,1}^{\text{up}} < 0$ or $\hat{\alpha}_{j_0,1}^{\text{down}} < 0$. If $m_0 \geq 2$, from (70b) and (71b) we get

$$\hat{\alpha}_{j_0,m_0}^{\text{up}} = (\alpha_{j_0,m_0}^{\text{up}})^+ + (\alpha_{j_0,m_0}^{\text{down}})^- \geq 0\,, \qquad \hat{\alpha}_{j_0,m_0}^{\text{down}} = (\alpha_{j_0,m_0}^{\text{up}})^- + (\alpha_{j_0,m_0}^{\text{down}})^+ \geq 0\,.$$

Finally, for $1 \leq m_0 \leq j_0 - 2$, from (70b) and (71b) we get that, for $\ell = m_0 + 1,\dots,j_0 - 1$, we have

$$\hat{\alpha}_{j_0,\ell}^{\text{up}} = (\alpha_{j_0,\ell}^{\text{up}})^+ \geq 0\,, \qquad \hat{\alpha}_{j_0,\ell}^{\text{down}} = (\alpha_{j_0,\ell}^{\text{down}})^+ \geq 0\,.$$

□

6 Appendix

In this section we give additional details on SSP coefficients and optimal perturbations of second order 2-stage Runge–Kutta methods and the classical 4-stage fourth order Runge–Kutta method.

6.1 Second order 2-stage methods

We consider the family of 2-stage second order methods (26). In example 1 we studied perturbations that increase the SSP coefficient for the linear case. For nonlinear problems, in example 3, figure 2 shows the values of $R^{\text{opt}}(K)$ for $\alpha \in [-3, 3]$.

In this section, for each α, we give the expressions for $R^{\text{opt}}(K)$ and we show optimal perturbations $\widetilde{K}_{NL}$ such that $R(K, \widetilde{K}_{NL}) = R^{\text{opt}}(K)$. It is important to point out the convenience of choosing $\widetilde{K}_{NL} = \widetilde{K}_L$, where $\widetilde{K}_L$ denotes the optimal perturbation for the linear case. In this case, we have not only $R(K, \widetilde{K}_L) = R^{\text{opt}}(K)$ but also $R_{\text{Lin}}(K, \widetilde{K}_{NL}) = R^{\text{opt}}_{\text{Lin}}(K)$. The computations required to obtain the results in this section have been done with the symbolic computation program *Mathematica*.

If we denote by $r = R^{\text{opt}}(K)$, we have that

$$r = \begin{cases} \dfrac{1}{|\alpha|}\,, & \text{if} \quad \alpha \in \left(-\infty, -\dfrac{1}{2}\left(1+\sqrt{7}\right)\right] \bigcup \left[\dfrac{1}{2}\left(-1+\sqrt{7}\right), \infty\right)\,, \\ \dfrac{-1+\alpha+\sqrt{3\alpha^2-2\alpha+1}}{|\alpha|}\,, & \text{if} \quad \alpha \in \left(-\dfrac{1}{2}\left(1+\sqrt{7}\right), 0\right) \bigcup \left(0, \dfrac{1}{2}\left(-1+\sqrt{7}\right)\right)\,. \end{cases} \tag{72}$$

Next we give optimal perturbations $\widetilde{K}_{NL}$.

For $\alpha < 0$, we obtain that it is not possible to obtain a perturbation of the form (26) with $\tilde{b}_2 = 0$ and $\tilde{a}_{21} = 0$. Consequently, $\widetilde{K}_{NL} \neq \widetilde{K}_L$ and we always have that $R_{\text{Lin}}(K, \widetilde{K}_{NL}) < R^{\text{opt}}_{\text{Lin}}(K)$. Optimal perturbations of the form (26) for different values of $\alpha < 0$ must satisfy the following conditions.

- For $-\frac{1}{2}\left(1+\sqrt{7}\right) \leq \alpha < 0$, the coefficients $\tilde{a}_{21}$, $\tilde{b}_1$ and $\tilde{b}_2$ in $\widetilde{K}_{NL}$ must satisfy

$$-\alpha \leq \tilde{a}_{21} \leq \frac{1 - r\,\alpha}{2\,r}\,, \qquad \tilde{b}_1 = -\frac{r\,\tilde{a}_{21}}{2\,\alpha}\,, \qquad \tilde{b}_2 = -\frac{1}{2\,\alpha}\,,$$

  where $r = R^{\text{opt}}(K)$.
- For $\alpha \leq -\frac{1}{2}\left(1+\sqrt{7}\right)$, we should have

$$\tilde{a}_{21} = -\alpha, \qquad -\frac{1}{2\,\alpha} \leq \tilde{b}_1 \leq \frac{-2\,\alpha^2 - 2\,\alpha + 1}{4\,\alpha}\,, \qquad -\frac{1}{2\,\alpha} \leq \tilde{b}_2 \leq \frac{2\,\alpha\,\tilde{b}_1 - 1}{4\,\alpha}\,.$$

For $\alpha > 0$ we can find optimal perturbations with $\tilde{b}_2 = 0$ and $\tilde{a}_{21} = 0$. Coefficient $\tilde{b}_1$ must satisfy the following conditions.

- For $0 < \alpha \leq \left(-1+\sqrt{7}\right)/2$, we have that

$$\tilde{b}_1 = \frac{\sqrt{3\alpha^2 - 2\alpha + 1} - \alpha}{2\alpha}\,. \tag{73}$$

  Thus there is a unique $\widetilde{K}_{NL}$ of the form (26). In this case, we have $R(K) < R(K, \widetilde{K}_{NL}) = R^{\text{opt}}(K)$.
- For $\left(-1+\sqrt{7}\right)/2 < \alpha < 1$, we also get $R(K) < R^{\text{opt}}(K)$, but in this case the optimal perturbation $\widetilde{K}_{NL}$ is not unique. All the perturbations with $\tilde{b}_1$ satisfying

$$\frac{1-\alpha}{\alpha} \leq \tilde{b}_1 \leq \frac{2\alpha^2 - 2\alpha + 1}{4\alpha}\,,$$

  are optimal. In particular, we can take $\widetilde{K}_{NL} = \widetilde{K}_L$. With this choice, $R(K, \widetilde{K}_L) = R^{\text{opt}}(K) = 1/\alpha$ and $R_{\text{Lin}}(K, \widetilde{K}_L) = R^{\text{opt}}_{\text{Lin}}(K) \approx 1.22$. Furthermore, $\alpha = \left(-1+\sqrt{7}\right)/2$ provides the largest SSP coefficient within the family of 2-stage second order method (see figure 2).

– For $1 \le \alpha$, we have $R(K) = R^{\rm opt}(K) = 1/\alpha$ and the optimal perturbation $\widetilde{K}_{NL}$ is not unique. All the values
$$0 \le \tilde{b}_1 \le \frac{2\alpha^2 - 2\alpha + 1}{4\alpha}$$
give optimal perturbations. We can take $\widetilde{K}_{NL} = 0$, but in this case $R_{\rm Lin}(K, 0) < R^{\rm opt}_{\rm Lin}(K)$. A better choice is $\widetilde{K}_{NL} = \widetilde{K}_L$. Observe that, for $\alpha = 1$, we get the optimal SSP coefficient $R(K) = 1$ that cannot be increased by perturbations.

Next, we consider some concrete values of α to show the the expressions of the perturbations. For each value, we give the Butcher tableau of the perturbation and matrices $\alpha^{\rm up}$ and $\alpha^{\rm down}$ in (36).

– For $\alpha = 1/2$ we get method RK2a in [17] with $R(K) = 0$. With perturbation
$$\widetilde{K} = \begin{pmatrix} 0 & 0\,0 \\ 0 & 0\,0 \\ \tilde{b}_1 & 0\,0 \end{pmatrix}, \alpha^{\rm up} = \begin{pmatrix} 0 & 0 & 0 \\ \tilde{b}_1 & 0 & 0 \\ 0 & 2\tilde{b}_1 & 0 \end{pmatrix}, \alpha^{\rm down} = \begin{pmatrix} 0 & 0\,0 \\ 0 & 0\,0 \\ 1-2\tilde{b}_1 & 0\,0 \end{pmatrix}, \gamma = \begin{pmatrix} 1 \\ 1-\tilde{b}_1 \\ 0 \end{pmatrix}, \tag{74}$$
where $\tilde{b}_1 = \frac{1}{2}\left(\sqrt{3} - 1\right)$, we get $R(K, \widetilde{K}) = R_{\rm Lin}(K, \widetilde{K}) = \sqrt{3} - 1$.

– For $\alpha = 2/3$, we have a nontrivial SSP coefficient $R^{\rm opt}(K) = 1/2$, but we can increase this value to $R(K, \widetilde{K}_1) = R^{\rm opt}(K) = 1$ with perturbation
$$\widetilde{K}_1 = \begin{pmatrix} 0 & 0\,0 \\ 0 & 0\,0 \\ 1/4 & 0\,0 \end{pmatrix}, \alpha^{\rm up} = \begin{pmatrix} 0 & 0 & 0 \\ 2/3 & 0 & 0 \\ 0 & 3/4 & 0 \end{pmatrix}, \alpha^{\rm down} = \begin{pmatrix} 0 & 0\,0 \\ 0 & 0\,0 \\ 1/4 & 0\,0 \end{pmatrix}, \gamma = \begin{pmatrix} 1 \\ 1/3 \\ 0 \end{pmatrix}.$$
For this perturbation, $R(\phi_K) = R_{\rm Lin}(K, \widetilde{K}_1) = 1$. We can take $\gamma = (1, 0, 0)^t$ by modifying the first column of $\alpha^{\rm up}$ and $\alpha^{\rm down}$ according to (50),
$$\widetilde{K}_2 = \begin{pmatrix} 0 & 0\,0 \\ 1/6 & 0\,0 \\ 3/8 & 0\,0 \end{pmatrix}, \alpha^{\rm up} = \begin{pmatrix} 0 & 0 & 0 \\ 5/6 & 0 & 0 \\ 0 & 3/4 & 0 \end{pmatrix}, \alpha^{\rm down} = \begin{pmatrix} 0 & 0\,0 \\ 1/6 & 0\,0 \\ 1/4 & 0\,0 \end{pmatrix}, \gamma = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

– As it has been pointed out above, the largest value in the α-family of 2-stage second order schemes is $R^{\rm opt}(K) = (1 + \sqrt{7})/3$ and it is obtained for $\alpha = (\sqrt{7} - 1)/2$. The perturbation is of the form (26) with $\tilde{b}_1 = \left(\sqrt{7} - 2\right)/2$, and
$$\alpha^{\rm up} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & \frac{1}{9}\left(4 + \sqrt{7}\right) & 0 \end{pmatrix}, \alpha^{\rm down} = \begin{pmatrix} 0 & 0\,0 \\ 0 & 0\,0 \\ \frac{1}{9}\left(5 - \sqrt{7}\right) & 0\,0 \end{pmatrix}, \gamma_r = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$
This is the perturbation obtained in [8, Table V] by numerical search in the class of perturbations considered in [8].

6.2 Classical fourth order 4-stage method

For nonlinear problems, applying the analysis above, we find that the optimal perturbation of the classical method has SSP coefficient given by the real root of $x^3 + 2x^2 + 4x - 4 = 0$, which is approximately $R^{\rm opt}(K) \approx 0.685016$. The corresponding perturbation is *not unique.* For instance, we can take $\gamma_r = (1, 0, 0, 0, 0)$, and all entries of $\alpha_r^{\rm down}$ equal to zero except
$$(\alpha_r^{\rm down})_{31} = \frac{r^2}{4} \qquad\qquad (\alpha_r^{\rm down})_{42} = \frac{r^2}{2}, \tag{75}$$
where $r = R^{\rm opt}(K)$. However, there exist other optimal perturbations with additionally $(\alpha_r^{\rm down})_{42} = \epsilon$ where $0 \le \epsilon \le 0.782$.

We remark that nearly-optimal perturbations for this method are given in [29, p. 448] and [14]. Interestingly, these different perturbed methods have different values of $R_{\rm Lin}(K, \widetilde{K})$.

References

1. P. Bogacki and L. F. Shampine. An efficient Runge-Kutta (4, 5) pair. *Comput. Math. Appl.*, 32(6):15–28, 1996.
2. M. Calvo, J. I. Montijano, and L. Rández. A new embedded pair of Runge-Kutta formulas of orders 5 and 6. *Comput. Math. Appl.*, 20(1):15–24, 1990.
3. R. Donat, I. Higueras, A. Martínez-Gavara. On stability issues for IMEX schemes applied to hyperbolic equations with stiff reaction terms. *Math. Comp.*, 80:2097–2126, 2011.
4. J. R. Dormand and P. J. Prince. A family of embedded Runge-Kutta formulae. *J. Comput. Appl. Math.*, 6(1):19–26, 1980.
5. E. Fehlberg. Klassische Runge-Kutta-Formeln fünfter und siebenter Ordnung mit Schrittweiten-Kontrolle. *Computing*, 4(2):93–106, 1969.
6. L. Ferracina and M. N. Spijker. Stepsize restrictions for the total-variation-diminishing property in general Runge-Kutta methods. *SIAM J. Numer. Anal.*, 42:1073–1093, 2004.
7. S. Gottlieb, D. I. Ketcheson, and C. W. Shu. *Strong Stability Preserving Runge-Kutta and Multistep Time Discretizations*. World Scientific Publishing Company, 2011.
8. S. Gottlieb and S. J. Ruuth. Optimal strong-stability-preserving time-stepping schemes with fast downwind spatial discretizations. *J. Sci. Comput.*, 27:289–303, 2006.
9. S. Gottlieb and C. W. Shu. Total variation diminishing runge-kutta schemes. *Math. Comp.*, 67(221):73–85, 1998.
10. E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II*. Springer, Berlin, 1991.
11. K. Heun. Neue methoden zur approximativen integration der differentialgleichungen einer unabhängigen veränderlichen. *Z. Math. Phys*, 45:23–38, 1900.
12. I. Higueras. Representations of Runge-Kutta methods and strong stability preserving methods. *SIAM J. Numer. Anal.*, 43:924–948, 2005.
13. I. Higueras. Strong Stability for Additive Runge-Kutta Methods. *SIAM J. Numer. Anal.*, 44(4):1735–1758, 2006.
14. I. Higueras. Positivity properties for the classical fourth order Runge-Kutta methods. *Monografías de la Real Academia de Ciencias de Zaragoza*, 33:125–139, 2010.
15. I. Higueras, Ketcheson D. I., and Kocsis T. A. Repository for computations of optimal perturbations to Runge–Kutta methods. `http://dx.doi.org/10.5281/zenodo.1146916`.
16. Z. Horváth. On the positivity step size threshold of Runge-Kutta methods. *Appl. Numer. Math.*, 53:341–356, 2005.
17. W. Hundsdorfer, B. Koren, M. van Loon, and J. C. Verwer. A positive finite-difference advection scheme. *J. Comput. Phys.*, 117(1):35–46, 1995.
18. D. I. Ketcheson. Highly Efficient Strong Stability Preserving Runge-Kutta Methods with Low-Storage Implementations. *SIAM J. Sci. Comput.*, 30:2113–2136, 2008.
19. D. I. Ketcheson. Computation of optimal monotonicity preserving general linear methods. *Math. Comput.*, 78:1497–1513, 2009.
20. D. I. Ketcheson. *High Order Strong Stability Preserving Time Integrators and Numerical Wave Propagation Methods for Hyperbolic PDEs*. Doctoral thesis, University of Washington, 2009.
21. D. I. Ketcheson. Step Sizes for Strong Stability Preservation with Downwind-biased Operators. *SIAM J. Numer. Anal.*, 49(4):1649–1660, 2011.
22. D. I. Ketcheson. Nodepy software version 0.6.1, 2015. `http://github.com/ketch/nodepy`.
23. J. F. B. M. Kraaijevanger. Contractivity of Runge-Kutta Methods. *BIT*, 31:482–528, 1991.
24. R. J. LeVeque and H. C. Yee. A study of numerical methods for hyperbolic conservation laws with stiff source terms. *J. Comput. Phys.*, 210:187–210, 1990.
25. R. H. Merson. An operational method for the study of integration processes. In *Proc. Symp. Data Processing*, pages 1–25, 1957.
26. P. J. Prince and J. R. Dormand. High order embedded Runge-Kutta formulae. *J. Comput. Appl. Math.*, 7(1):67–75, 1981.
27. S. J. Ruuth. Global optimization of explicit strong-stability-preserving Runge-Kutta Methods. *Math. Comput.*, 75:183–207, 2006.
28. S. J. Ruuth and R. J. Spiteri. High-order strong-stability-preserving Runge-Kutta methods with downwind-biased spatial discretizations. *SIAM J. Numer. Anal.*, 42:974–996, 2004.
29. C. W. Shu and S. Osher. Efficient implementation of essentially non-oscillatory shock-capturing schemes. *J. Comput. Phys.*, 77(2):439–471, 1988.
30. X. Zhang and C. W. Shu. On maximum-principle-satisfying high order schemes for scalar conservation laws. *J. Comput. Phys.*, 229(9):3091–3120, 2010.