Eötvös Loránd University
Faculty of Science

Embedded pairs for optimal strong stability preserving Runge–Kutta methods

Master Thesis

Ákos Horváth
Applied Mathematician MSc
Applied Analysis Specialization

Supervisor: Imre Fekete
Department of Applied Analysis and
Computational Mathematics

Budapest
2019

Contents

1 Introduction

Let us consider the initial value problem (IVP)

$$u'(t) = F(t, u(t)), \quad u(t_0) = u_0. \tag{1}$$

An IVP (1) could arise in the sense of method of line approach, i.e. when we spatially discretize partial differential equations (PDEs) and we obtain a large system of ordinary differential equations (ODEs) in time. In this thesis, we will be focusing on a special time discretization family which was developed for the time evolution of spatially discretized hyperbolic PDEs and mainly used for hyperbolic conservation laws.

Since in this class the exact solutions may develop discontinuities (e.g. shockwaves) even for smooth initial data, therefore solving these problems numerically are very challenging. Because of this reason, a huge effort has been put into the development of high-order spatial discretizations that can handle discontinuities like essentially non-oscillatory (ENO), weighted ENO finite difference and finite volume schemes [1], [2], [3], [4].

As an important example let us consider the one-dimensional hyperbolic law in the form

$$u_t(t, x) + f_x(u(t, x)) = 0, \tag{2}$$

where the subscripts refer to the partial derivatives. Applying an appropriate spatial discretization scheme (ENO, WENO, finite volume schemes) to (2), we obtain a system in the form

$$u_t = F(u). \tag{3}$$

Typically

$$F(u) = -f_x(u) + \mathcal{O}(\Delta x^k),$$

where Δx and k refer to mesh size and order of accuracy in space, respectively. Now we would like to discretize (3) in time such that the preserved stability properties of the original PDE (2) will be maintained during the time discretization process.

Using the first order forward Euler method in time the fully discretized scheme has the form

$$u^{n+1} = u^n + \Delta t F(u^n), \tag{4}$$

where Δt denotes the mesh size in time. The problem with the full scheme (4) is that even if we are using high-oder spatial discretization methods the global order is one due to forward Euler method.

Therefore we would like to design high-order time discretization methods such that we are maintaining the required stability properties. In the sequel we assume that the semi-discretization (3) and a convex functional $||\cdot||$ (or norm, semi-norm) are given, and that there exists a Δt_{FE} forward Euler time step such that the so-called *forward Euler condition*

$$||u + \Delta t F(u)|| \leq ||u|| \text{ for } 0 \leq \Delta t \leq \Delta t_{\mathrm{FE}} \tag{5}$$

holds for all u. The method is called *strong stability preserving* (SSP) if the estimate

$$||u_{n+1}|| \leq ||u_n||$$

holds for the numerical solution of (3), whenever (5) holds and $\Delta t \leq \mathcal{C}\Delta t_{\mathrm{FE}}$. The constant $\mathcal{C}$ is called the SSP coefficient.

In Chapter 2. we briefly summarize the connection between the Butcher and the so-called *Shu-Osher* representations following the lines of [5]. In Chapter 3. we list the optimal SSP Runge–Kutta (RK) methods. In Chapter 4. we introduce certain criteria designing embedded pairs for optimal SSP RK methods. For the implicit case we give our new results in details based on paper [6].

2 Representing Runge–Kutta methods

In 1988 Shu and Osher introduced a special class of RK methods to maintain total variation diminishing (TVD) and total variation bounded (TVB) properties of semi-discrete problem (3) coupled with high-order RK methods [1]. Their key observation was that certain RK methods can be rewritten as convex combinations of the forward Euler method and these methods can maintain the desired stability properties.

The classical representation of the s-stage explicit RK (ERK) methods is

$$\begin{aligned} u^{(i)} &= u^n + \Delta t \sum_{j=1}^{i-1} a_{ij} F(u^{(j)}) \qquad (1 \le i \le s) \\ u^{n+1} &= u^n + \Delta t \sum_{j=1}^{s} b_j F(u^{(j)}), \end{aligned}$$

where a_{ij}, b_j are real parameters. They are represented in the Bucher tableau [7]. As we mentioned Shu and Osher rewrote the stages and the step as convex combinations of forward Euler method. ERK methods can be written in the form

$$\begin{aligned} u^{(0)} &= u^n \\ u^{(i)} &= \sum_{j=0}^{i-1} \Big(\alpha_{ij} u^{(j)} + \Delta t \beta_{ij} F(u^{(j)}) \Big) \\ u^{n+1} &= u^{(s)}. \end{aligned} \tag{6}$$

Consistency requires the condition $\sum_{j=0}^{i-1} \alpha_{ij} = 1$. Furthermore, if all the coefficients α_{ij} and β_{ij} are non-negative, then it can be manipulated into convex

combinations of forward Euler steps with a modified time step. This representation is called *Shu-Osher representation.*

The following theorem gives us a time step restriction maintaining the strong stability property for RK methods.

Theorem 1 ([5], Thm. 2.1.)**.** Let us apply the forward Euler method to (3). If (5) holds and $\alpha_{ij}, \beta_{ij} \geq 0$, then the solution obtained by (6) satisfies the strong stability bound

$$||u^{n+1} \leq ||u^n||$$

under the time step restriction

$$\Delta t \leq \mathcal{C}(\alpha, \beta)\Delta t_{FE},$$

where $\mathcal{C}(\alpha, \beta) = \min_{i,j} \frac{\alpha_{ij}}{\beta_{ij}}$ and the ratio is understood as infinite if $\beta_{ij} = 0$.

Although Theorem 1 is really useful, it does not help us calculating the largest SSP coefficient $\mathcal{C}(\alpha, \beta)$, since the representation is not unique. To demonstrate this we will consider a two-stage second order ERK method in three ways.

Example ([5], Example 2.2.)**.** *Let us consider the method*

$$\begin{aligned} u^{(0)} &= u^n \\ u^{(1)} &= u^{(0)} + \Delta t F(u^{(0)}) \\ u^{n+1} &= u^{(0)} + \frac{1}{2}\Delta t F(u^{(0)}) + \frac{1}{2}\Delta t F(u^{(1)}). \end{aligned} \tag{7}$$

Based on Theorem 1. the corresponding SSP coefficient is 0. However, one can rewrite (7) *as*

$$u^{n+1} = \frac{3}{4}u^{(0)} + \frac{1}{4}\Delta t F(u^{(0)}) + \frac{1}{4}u^{(1)} + \frac{1}{2}\Delta t F(u^{(1)}),$$

which yields $\mathcal{C}(\alpha,\beta) = 1/2$. *As a third option we have*

$$u^{n+1} = \frac{1}{2}u^{(0)} + \frac{1}{2}u^{(1)} + \frac{1}{2}\Delta t F(u^{(1)}),$$

which yields $\mathcal{C}(\alpha,\beta) = 1$.

As we have seen in Example 2., the three representation of the same method gives us different SSP coefficients. Obviously we would like to get the largest SSP coefficient in order to allow the largest step size. Until so far the problem is that we have infinitely many Shu-Osher representations. We will make a step further towards the direction of unique representation.

2.1 The modified Shu-Osher form

The generalization of Shu-Osher form was independently introduced by Higueras [8], and Ferracina and Spijker [9]. It has the form

$$\begin{aligned} u^{(i)} &= v_i u^n + \sum_{j=1}^{s} \Big(\alpha_{ij} u^{(j)} + \Delta t \beta_{ij} F(u^{(j)})\Big) \quad (1 \le i \le s+1) \\ u^{n+1} &= u^{(s+1)} \end{aligned}$$

which can represent both explicit and implicit RK methods. An immediate advantage of this *modified Shu-Osher form* is the indexing since it agrees with the stage numbering in the Butcher form. Rearranging it we have

$$u^{(i)} = v_i u^n + \sum_{j=1}^{s} \alpha_{ij} \Big(u^{(j)} + \Delta t \frac{\beta_{ij}}{\alpha_{ij}} F(u^{(j)})\Big) \quad (1 \le i \le s+1).$$

Since consistency requires that

$$v_i + \sum_{j=1}^{s} \alpha_{ij} = 1 \quad (1 \le i \le s+1),$$

then if $\alpha_{ij}, \beta_{ij}, v_j \geq 0$, then each stage $u^{(i)}$ is a convex combination of forward Euler steps.

To generalize Theorem 1. first we have to exclude the so-called *non-zero-well-defined methods*. An RK method is zero-well-defined if the stage equations have a unique solution when the method is applied to the scalar problem (3).

Theorem 2 ([5], Thm 3.1.)**.** If the forward Euler method applied to (3) is strongly stable under the time step restriction $\Delta t \leq \Delta t_{FE}$, then the solution obtained by a zero-well-defined Runge–Kutta method satisfies the strong stability bound $\|u^{n+1}\| \leq \|u^n\|$ under the time step restriction

$$0 \leq \Delta t \leq \mathcal{C}(\alpha, \beta)\Delta t_{FE},$$

where

$$\mathcal{C}(\alpha, \beta) = \begin{cases} \min\limits_{i,j} \dfrac{\alpha_{ij}}{\beta_{ij}}, & \text{if } \alpha_{ij}, \beta_{ij}, v_j \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

The ratio is understood as infinite if $\beta_{ij} = 0$.

We will introduce a compact notation which helps to represent better RK methods. Let us define the matrices $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ as

$$(\boldsymbol{\alpha})_{ij} = \begin{cases} \alpha_{ij} & 1 \leq i \leq s+1, 1 \leq j \leq s \\ 0 & j = s+1 \end{cases}$$

$$(\boldsymbol{\beta})_{ij} = \begin{cases} \beta_{ij} & 1 \leq i \leq s+1, 1 \leq j \leq s \\ 0 & j = s+1. \end{cases}$$

We will use the notation $\boldsymbol{u}^n$ and $\boldsymbol{u}^{(i)}$ since the solution u^n and the stages $u^{(i)}$

are usually vectors. We define the vectors

$$\boldsymbol{y}_i = \boldsymbol{u}^{(i)}$$
$$\boldsymbol{f}_i = F(\boldsymbol{u}^{(i)})$$

and the matrices by using Kronecker-products

$$\bar{\boldsymbol{v}} = \boldsymbol{I} \otimes \boldsymbol{v}$$
$$\bar{\boldsymbol{\alpha}} = \boldsymbol{I} \otimes \boldsymbol{\alpha}$$
$$\bar{\boldsymbol{\beta}} = \boldsymbol{I} \otimes \boldsymbol{\beta}.$$

Now the RK method can be written in a compact form

$$\begin{aligned} \boldsymbol{y} &= \bar{\boldsymbol{v}}\boldsymbol{u}^n + \bar{\boldsymbol{\alpha}}\boldsymbol{y} + \Delta t\bar{\boldsymbol{\beta}}\boldsymbol{f} \\ \boldsymbol{u}^{n+1} &= \boldsymbol{y}_{s+1}. \end{aligned} \tag{8}$$

Here, we would like to emphasize that neither the Shu-Osher nor its modified version is unique. In order to get the maximum SSP coefficient over all representation we need a way of unique representation of a particular method.

2.2 Unique respresentation

In this subsection we make the final step having unique representations. First, solving (8) for $\boldsymbol{y}$ yields

$$\begin{aligned} \boldsymbol{y} &= \bar{\boldsymbol{v}}\boldsymbol{u}^n + \bar{\boldsymbol{\alpha}}\boldsymbol{y} + \Delta t\bar{\boldsymbol{\beta}}\boldsymbol{f} \\ (\boldsymbol{I} - \bar{\boldsymbol{\alpha}})\boldsymbol{y} &= \bar{\boldsymbol{v}}\boldsymbol{u}^n + \Delta t\bar{\boldsymbol{\beta}}\boldsymbol{f} \end{aligned}$$

and assuming that the inverse $(\boldsymbol{I} - \boldsymbol{\alpha})^{-1}$ exist (it exists when a method is zero-well-defined), then

$$\boldsymbol{y} = (\boldsymbol{I} - \bar{\boldsymbol{\alpha}})^{-1}\bar{\boldsymbol{v}}\boldsymbol{u}^n + \Delta t(\boldsymbol{I} - \bar{\boldsymbol{\alpha}})^{-1}\bar{\boldsymbol{\beta}}\boldsymbol{f}.$$

Rearranging the consistency condition we have

$$\boldsymbol{y} = \boldsymbol{e}\boldsymbol{u}^n + \Delta t(\boldsymbol{I} - \bar{\boldsymbol{\alpha}})^{-1}\bar{\boldsymbol{\beta}}\boldsymbol{f}.$$

From the relation $\boldsymbol{y} = \bar{\boldsymbol{v}}\boldsymbol{u}^n + \bar{\boldsymbol{\alpha}}\boldsymbol{y} + \Delta t\bar{\boldsymbol{\beta}}\boldsymbol{f}$ and taking $\boldsymbol{\alpha} = 0$, consistency yields $\boldsymbol{v} = \boldsymbol{e}$. For the moment denoting the unknown coefficients β_{ij} by $\bar{\boldsymbol{\beta}}_{\mathbf{0}}$, the method reads as

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{e}\boldsymbol{u}^n + \Delta t\bar{\boldsymbol{\beta}}_0\boldsymbol{f} \\ \boldsymbol{u}^{n+1} &= \boldsymbol{y}_{s+1}. \end{aligned}$$

So we can see the relation with the Butcher form since

$$\boldsymbol{\beta}_0 = (\boldsymbol{I} - \boldsymbol{\alpha})^{-1}\boldsymbol{\beta} = \begin{pmatrix} \mathbf{A} & 0 \\ \mathbf{b}^T & 0 \end{pmatrix}.$$

The matrix $(\boldsymbol{I} - \boldsymbol{\alpha})$ singular for trivial class of methods ([5], Lemma 3.1.). The remaining problem is that even using the Butcher form it is possible to represent a method in multiple ways. This problem can be handled by using DJ-reducible RK methods (see for further details [5], Section 3.2.2.).

2.3 Canonical Shu-Osher form

Now we have a unique Butcher form for irreducible methods but it does not reveal the SSP coefficient. In this section we will have a form that reveals the SSP coefficient of the method. Finding the SSP coefficient is easy when we have a particular modified Shu-Osher form. Namely, where $r = \alpha_{ij}/\beta_{ij}$ is the same for all i, j such that $\beta_{ij} \neq 0$. The method coefficients are $\boldsymbol{\alpha}_r$ and $\boldsymbol{\beta}_r$ such that $\boldsymbol{\alpha}_r = r\boldsymbol{\beta}_r$. Since

$$\boldsymbol{\beta}_0 = (\boldsymbol{I} - \boldsymbol{\alpha})^{-1}\boldsymbol{\beta},$$

therefore for $\boldsymbol{\beta} = \boldsymbol{\beta}_r$ we have

$$\boldsymbol{\beta}_0 = (\boldsymbol{I} - r\boldsymbol{\beta}_r)^{-1}\boldsymbol{\beta}_r.$$

If the method is zero-well-defined then $\boldsymbol{I} - r\boldsymbol{\beta}_r = \boldsymbol{I} - \boldsymbol{\alpha}_r$ is invertible. Then

$$\begin{aligned}
\boldsymbol{\beta}_r = (\boldsymbol{I} - r\boldsymbol{\beta}_r)\boldsymbol{\beta}_0 &= \boldsymbol{\beta}_0 - r\boldsymbol{\beta}_r\boldsymbol{\beta}_0 \\
\boldsymbol{\beta}_r + r\boldsymbol{\beta}_r\boldsymbol{\beta}_0 &= \boldsymbol{\beta}_0 \\
\boldsymbol{\beta}_r(\boldsymbol{I} + r\boldsymbol{\beta}_0) &= \boldsymbol{\beta}_0.
\end{aligned}$$

Then if the inverse of the matrix $\boldsymbol{I} + r\boldsymbol{\beta}_0$ exists, then we have

$$\begin{aligned}
\boldsymbol{\beta}_r &= \boldsymbol{\beta}_0(\boldsymbol{I} + r\boldsymbol{\beta}_0)^{-1} \\
\boldsymbol{\alpha}_r &= r\boldsymbol{\beta}_r = r\boldsymbol{\beta}_0(\boldsymbol{I} + r\boldsymbol{\beta}_0)^{-1} \\
\boldsymbol{v}_r &= (\boldsymbol{I} - \boldsymbol{\alpha}_r)\boldsymbol{e} = \big(\boldsymbol{I} - r\boldsymbol{\beta}_0(\boldsymbol{I} + r\boldsymbol{\beta}_0)^{-1}\big)\boldsymbol{e} \\
&= \big((\boldsymbol{I} + r\boldsymbol{\beta}_0)(\boldsymbol{I} + r\boldsymbol{\beta}_0)^{-1} - r\boldsymbol{\beta}_0(\boldsymbol{I} + r\boldsymbol{\beta}_0)^{-1}\big)\boldsymbol{e} \\
&= \big((\boldsymbol{I} + r\boldsymbol{\beta}_0 - r\boldsymbol{\beta}_0)(\boldsymbol{I} + r\boldsymbol{\beta}_0)^{-1}\big)\boldsymbol{e} = (\boldsymbol{I} + r\boldsymbol{\beta}_0)^{-1}\boldsymbol{e}.
\end{aligned} \tag{9}$$

We will refer the form given by (9) as *the canonical Shu-Osher form*

$$\boldsymbol{y} = \bar{\boldsymbol{v}}_r\boldsymbol{u}^{n-1} + \bar{\boldsymbol{\alpha}}_r\Big(\boldsymbol{y} + \frac{\Delta t}{r}\boldsymbol{f}\Big). \tag{10}$$

3 Optimal SSP Runge–Kutta methods

The following theorem states that RK methods with SSP coefficient $\mathcal{C}$ can always be rewritten in the form (10).

Theorem 3 ([5], Thm. 3.2.)**.** Consider a Runge–Kutta method with Butcher coefficient array $\boldsymbol{\beta}_0$. The SSP coefficient of the method is

$$\mathcal{C} = \max\left\{r \geq 0 \quad | \quad (\boldsymbol{I} + r\boldsymbol{\beta}_0)^{-1}\text{exists},\ \boldsymbol{\alpha}_r \geq 0,\ \boldsymbol{v}_r \geq 0\right\}.$$

During the proof of Theorem 3. it turns out that for a zero-well-defined method

- the matrix $\boldsymbol{I} + r\boldsymbol{\beta}_0$ is invertible for all $0 \leq r \leq \mathcal{C}(\alpha, \beta)$,
- for all $0 \leq r \leq \mathcal{C}(\alpha, \beta)$ the condition

$$(\boldsymbol{I} + r\boldsymbol{\beta}_0)^{-1}\text{exists},\ \boldsymbol{\alpha}_r \geq 0,\ \boldsymbol{v}_r \geq 0$$

  holds.

Taking into account the above statements we are able to compute the SSP coefficients $\mathcal{C}$ by using the bisection method. Another option calculating the SSP coefficients $\mathcal{C}$ is to form an optimization problem in terms of Butcher array. For a given order and number of stages the corresponding optimization problem ([5], Section 3.4.) is

$$\begin{aligned} &\text{maximize } r \text{ subject to} \\ &\boldsymbol{\beta}_0(\boldsymbol{I} + r\boldsymbol{\beta}_0)^{-1} \geq 0 \\ &||r\boldsymbol{\beta}_0(\boldsymbol{I} + r\boldsymbol{\beta}_0)^{-1}||_\infty \leq 1 \\ &\tau_k(\boldsymbol{\beta}_0) = 0, \end{aligned}$$

where the last row represents the set of order conditions for order $k \leq p$. Now we have algorithmic tools calculating the optimal SSP coefficients.

3.1 The optimal explicit and implicit methods

An extensive effort have been made to find optimal explicit and implicit SSP RK methods. Without giving the details in this section we would like to list those optimal methods which will be used during the thesis. Other methods and theoretical observations will be mentioned and the corresponding references will be given.

In the sequel we will denote the optimal explicit SSP RK and implicit SSP RK methods with stage number s and order p by SSPERK(s, p) and SSPIRK(s, p), respectively. We introduce the *effective SSP coefficient*

$$\mathcal{C}_{\text{eff}} = \frac{\mathcal{C}}{s}$$

in order to compare the listed methods.

Optimal first order explicit SSP Runge–Kutta methods consist simply of repeated forward Euler steps. These methods all have $\mathcal{C} = 1$ and so $\mathcal{C}_{\text{eff}} = 1$. These methods are equivalent to simply using the forward Euler method.

SSPERK$(s,2)$**, [10]**

The coefficients are

$$\beta_{i,i-1} = \begin{cases} \frac{1}{s-1} & 1 \le i \le s-1 \\ \frac{1}{s} & i = s \end{cases}$$

$$\alpha_{i,i-1} = \begin{cases} 1 & 1 \le i \le s-1 \\ \frac{s-1}{s} & i = s \end{cases}$$

$$\alpha_{s,0} = \frac{1}{s}.$$

The SSP coefficient in this family is $s-1$, so $\mathcal{C}_{\text{eff}} = \frac{s-1}{s}$.

The optimal three-stage third order explicit SSP Runge–Kutta method was reported in [1]. It has $\mathcal{C} = 1$ and $\mathcal{C}_{\text{eff}} = \frac{1}{3}$. The optimal four-stage third order explicit SSP Runge–Kutta method was reported in [11]. It has $\mathcal{C} = 2$ and $\mathcal{C}_{\text{eff}} = \frac{1}{2}$.

SSPERK$(n^2,3)$**, [12]**

In this family $n > 2$ and $s = n^2$. The corresponding coefficients are

$$\alpha_{i,i-1} = \begin{cases} \frac{n-1}{2n-1} & i = \frac{n(n+1)}{2} \\ 1 & \text{otherwise} \end{cases}$$

$$\alpha_{\frac{n(n+1)}{2},\frac{(n-1)(n-2)}{2}} = \frac{n}{2n-1}$$

$$\beta_{i,i-1} = \frac{\alpha_{i,i-1}}{n^2-n}.$$

In this family the SSP coefficient $\mathcal{C}$ is $n^2 - n = s - \sqrt{s}$, therefore we have $\mathcal{C}_{\text{eff}} = 1 - \frac{1}{n} = 1 - \frac{1}{\sqrt{s}}$.

In [12] a ten-stage fourth order method is given. It has been proven analytically that it is optimal for linear problems with $\mathcal{C}_{\text{eff}} = \frac{6}{10}$. It is also

known that any irreducible explicit RK method with $\mathcal{C} > 0$ has order $p \leq 4$ ([5], Observation 5.4).

Now we list the optimal implicit methods. The backward Euler method is unconditionally SSP.

SSPIRK$(s, 2)$**, [13]**

The coefficients are

$$\begin{aligned} \alpha_{i+1,i} =& 1 \\ \beta_{i,i} = \beta_{i+1,i} =& \frac{1}{2s}. \end{aligned}$$

In this family $\mathcal{C} = 2s$, so $\mathcal{C}_{\text{eff}} = 2$.

SSPIRK$(s, 3)$**, [13]**

The coefficients are

$$\alpha_{i+1,i} = \begin{cases} \frac{(s+1)(s-1+\sqrt{s^2-1})}{s(s+1+\sqrt{s^2-1})} & i = s \\ 1 & \text{otherwise} \end{cases}$$

$$\beta_{i,i} = \frac{1}{2}\left(1 - \sqrt{\frac{s-1}{s+1}}\right)$$

$$\beta_{i+1,i} = \begin{cases} \frac{s+1}{s(s+1+\sqrt{s^2-1})} & i = s \\ \frac{1}{2}\left(\sqrt{\frac{s+1}{s-1}} - 1\right) & \text{otherwise.} \end{cases}$$

These methods have $\mathcal{C} = s - 1 + \sqrt{s^2 - 1}$ and $\mathcal{C}_{\text{eff}} = 1 - \frac{1}{s} + \frac{\sqrt{s^2-1}}{s}$.

The order barrier for the implicit case is six, so there is no implicit SSP RK method with $\mathcal{C} > 0$ and order $p \geq 7$. Optimal fourth through sixth order

methods have been derived for certain stage number. The corresponding coefficients are listed in Section 7.4. in [5].

Its worth emphasizing that for explicit methods we have $\mathcal{C}_{\text{eff}} \leq 1$. So we cannot guarantee a larger step size than the step size of forward Euler in the effective sense but at least we increase the order. In the implicit case we can achieve a larger effective step size.

3.2 Converting the Shu-Osher and Butcher forms

SSP analysis relies mostly on Shu-Osher forms, but computing the SSP coefficient is more convenient using the Butcher form. Here we give two algorithms to convert a Runge-Kutta method from one form to the other.

Let us consider the Shu-Osher coefficient matrices $\alpha, \beta \in \mathbb{R}^{(s+1)\times s}$. First we create matrices $\hat{\alpha}, \hat{\beta} \in \mathbb{R}^{s\times s}$ by simply removing the last row from α and β. The last rows are denoted by $\alpha_{\text{last}}, \beta_{\text{last}}$. Then

$$\begin{aligned} X =& I - \hat{\alpha} \\ A =& X^{-1}\hat{\beta} \\ b^T =& \beta_{\text{last}} + \alpha_{\text{last}} A \end{aligned}$$

gives us the Butcher form.

In order to convert the Butcher form to the optimal Shu-Osher form, first we

need to compute the SSP coefficient $\mathcal{C}$ by using the bisection method. Then

$$
\begin{aligned}
K =& \begin{pmatrix} \mathrm{A} & 0 \\ \mathrm{b}^T & 0 \end{pmatrix} \\
\beta =& K(I + \mathcal{C}K)^{-1} \\
\alpha =& \mathcal{C}\beta \\
v =& (I - \alpha)e.
\end{aligned}
$$

Since in the sequel we will use the more convenient Butcher form, therefore we rephrase Theorem 3.

Theorem 4. Let us consider the matrix

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

and the SSP conditions

$$K(I + rK)^{-1} \geq 0 \tag{11a}$$

$$rK(I + rK)^{-1}e \leq e. \tag{11b}$$

Then, the SSP coefficient is

$$\mathcal{C} = \sup \left\{ r : (I + rK)^{-1} \text{exists and conditions (11a)-(11b) hold} \right\}.$$

4 Embedded pairs

Embedded pairs provide an estimate of local truncation error [14] and they are used for automatic error control. For further details in case of Runge–Kutta methods see [15] and [16].

In this section we would like to design embedded pairs for the previously listed optimal explicit and implicit SSP Runge–Kutta methods. In the explicit case we briefly summarize the results of paper [17]. In the implicit case we give our own results.

The extended Butcher tableau is

$$\begin{array}{c|c} c & A \\ \hline & b^T \\ & \hat{b}^T \end{array}$$

where the pair b^T corresponds to the higher order RK method. The embedded pair and the full method are denoted by $\hat{b}^T$ and $\mathrm{RK}(A, b^T, \hat{b}^T)$, respectively. The order conditions up to order three are

$$\begin{aligned} b^T e &= 1 && (p=1) \\ b^T c &= \frac{1}{2} && (p=2) \\ b^T c^2 &= \frac{1}{2} && (p=3) \\ b^T (c^2/2 - Ac) &= 0. && (p=3) \end{aligned}$$

Following [17] the embedded pair should fulfill the below listed properties:

a) The embedded method is order of $p-1$, i.e., it has one order less than the SSPRK method.

b) The embedded method is non-defective, i.e., it violates all of the p-th order conditions.

c) Whenever possible, the embedded method has rational coefficients and a simple structure.

d) The embedded method has maximum SSP coefficient $\hat{\mathcal{C}}$, where $\hat{\mathcal{C}}$ is the SSP coefficient of the optimal SSPRK method of order $p-1$. If this is not the case, then we are looking for embedded SSPRK methods with smaller SSP coefficient or simply embedded RK methods.

Rephrasing the SSP conditions (11a)-(11b) and taking into account the listed required properties a)-d) our task to find $\hat{b}^T$ such that

$$K(I+\mathcal{C}K)^{-1} \geq 0,$$

$$||\mathcal{C}K(I+\mathcal{C}K)^{-1}||_\infty \leq 1,$$

appropriate order conditions and a), b) properties are fulfilled,

c), d) properties are desired.

4.1 Embedded pairs for optimal SSPERK methods

In this section we briefly summarize the results of paper [17].

SSPERK$(s, 2)$

First, we consider optimal SSPERK$(s, 2)$ methods. The corresponding Butcher tableau is

$$
\begin{array}{c|ccccc}
0 & & & & & \\
\frac{1}{s-1} & \frac{1}{s-1} & & & & \\
\frac{2}{s-1} & \frac{1}{s-1} & \frac{1}{s-1} & & & \\
\vdots & \vdots & \vdots & \ddots & & \\
1 & \frac{1}{s-1} & \frac{1}{s-1} & \cdots & \frac{1}{s-1} & \\
\hline
 & \frac{1}{s} & \frac{1}{s} & \cdots & \frac{1}{s} & \frac{1}{s}
\end{array}
$$

In [17] the authors recommended the embedded pair

$$
\hat{b}^T = \left[\frac{s+1}{s^2}, \frac{1}{s}, \ldots, \frac{1}{s}, \frac{s-1}{s^2} \right].
$$

The corresponding absolute stability region is given in Figure 1.

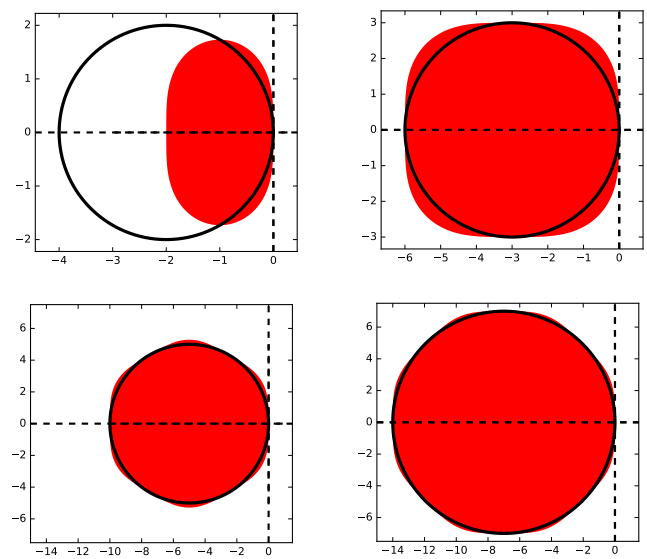


Figure 1: The absolute stability regions of SSPERK$(s, 2)$ methods (red) and the black contours of the embedded SSPERK$(s, 1)$ methods from left to right and top to bottom for $s = 2, 4, 6, 8$.

SSPERK$(n^2, 3)$

The Butcher tableau of optimal SSPERK$(n^2, 3)$ methods is

$$A = \begin{pmatrix} 0 & & & & & & & & & & \\ \frac{1}{n(n-1)} & \ddots & & & & & & & & & \\ \frac{1}{n(n-1)} & & \ddots & & & & & & & & \\ \frac{1}{n(n-1)} & & & \ddots & & & & & & & \\ \vdots & & & & \ddots & & & & & & \\ \vdots & & & & & \ddots & & & & & \\ \frac{1}{n(n-1)} & \cdots & \frac{1}{n(n-1)} & \cdots & \cdots & \frac{1}{n(n-1)} & & & & & \\ \frac{1}{n(n-1)} & & \frac{1}{n(n-1)} & \frac{1}{n(2n-1)} & \cdots & \frac{1}{n(2n-1)} & \frac{1}{n(n-1)} & & & & \\ \vdots & & \vdots & \vdots & & \vdots & \vdots & \ddots & & & \\ \underbrace{\frac{1}{n(n-1)} \cdots \frac{1}{n(n-1)}}_{\frac{(n-2)(n-1)}{2}} & & & \frac{1}{n(2n-1)} & \cdots & \frac{1}{n(2n-1)} & \underbrace{\frac{1}{n(n-1)} \cdots \frac{1}{n(n-1)}}_{\frac{n(n-1)}{2}-1} & & & 0 \end{pmatrix} \in \mathbb{R}^{n^2 \times n^2}, \tag{12}$$

where the submatrix in the rectangle is a $\left(\frac{n(n-1)}{2}\right) \times (2n-1)$ dimensional matrix and

$$b^T = \left[\underbrace{\frac{1}{n(n-1)}, \ldots, \frac{1}{n(n-1)}}_{\frac{(n-1)(n-2)}{2}}, \underbrace{\frac{1}{n(2n-1)}, \ldots, \frac{1}{n(2n-1)}}_{2n-1}, \underbrace{\frac{1}{n(n-1)}, \ldots, \frac{1}{n(n-1)}}_{\frac{n(n-1)}{2}} \right] \in \mathbb{R}^{n^2}. \tag{13}$$

In [17] they recommended the embedded pair

$$\hat{b}^T = \left[\frac{1}{s}, \ldots, \frac{1}{s} \right].$$

Their numerical searches failed to find any embedded pair with $\mathcal{C} = n^2 - n$, but computations showed that the method has $\mathcal{C} > 0$ and nice absolute stability regions. in Figure 2. shows the corresponding absolute stability regions.

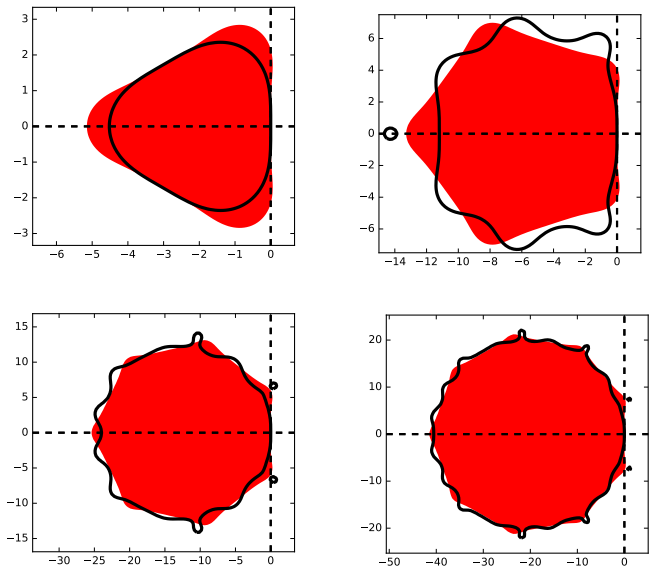


Figure 2: The absolute stability regions of SSPERK$(n^2, 3)$ methods (red) and the black contours of the embedded methods from left to right and top to bottom for $n^2 = 4, 9, 16, 25$.

4.2 Embedded pairs for optimal SSPIRK methods

In this section we presents our own results regarding the embedded pairs for optimal SSPIRK$(s, 2)$ and SSPIRK$(s, 3)$ methods.

SSPIRK(s,2)

The Butcher tableau of optimal SSPIRK$(s, 2)$ methods is

$$\begin{array}{c|ccccc}
\frac{1}{2s} & \frac{1}{2s} & & & & \\
\frac{3}{s} & \frac{1}{s} & \frac{1}{2s} & & & \\
\frac{5}{s} & \frac{1}{s} & \frac{1}{s} & \frac{1}{2s} & & \\
\vdots & \vdots & \vdots & \ddots & \ddots & \\
\frac{2s-1}{2s} & \frac{1}{s} & \frac{1}{s} & \cdots & \frac{1}{s} & \frac{1}{2s} \\
\hline
 & \frac{1}{s} & \frac{1}{s} & \cdots & \frac{1}{s} & \frac{1}{s}
\end{array}$$

Numerical searches failed to find any embedded pair with $\mathcal{C} = 2s$.

Theorem 5. There is no embedded pair for SSPIRK(s,2) with $\mathcal{C} = 2s$.

Proof. For the sake of simplicity the proof is given for the optimal SSIRK$(2, 2)$ method but it can be generalized. The corresponding Butcher tableau is

$$\begin{array}{c|cc}
\frac{1}{4} & \frac{1}{4} & 0 \\
\frac{3}{4} & \frac{1}{2} & \frac{1}{4} \\
\hline
 & \frac{1}{2} & \frac{1}{2}
\end{array}$$

Let us introduce the matrix $\hat{K}$ as

$$\hat{K} = \begin{pmatrix} \frac{1}{4} & 0 & 0 \\ \frac{1}{2} & \frac{1}{4} & 0 \\ \hat{b_1} & \hat{b_2} & 0 \end{pmatrix}.$$

In this case the SSP conditions are

$$M = \hat{K}(I + r\hat{K})^{-1} = \begin{pmatrix} \frac{1}{4+r} & 0 & 0 \\ \frac{8}{(4+r)^2} & \frac{1}{4+r} & 0 \\ \frac{4(-2r\hat{b}_2+(4+r)\hat{b}_1)}{(4+r)^2} & \frac{4\hat{b}_2}{4+r} & 0 \end{pmatrix}$$

and

$$d = rMe = \begin{bmatrix} \frac{r}{4+r} \\ \frac{r(12+r)}{(4+r)^2} \\ \frac{4r((4-r)\hat{b}_2+(4+r)\hat{b}_1)}{(4+r)^2} \end{bmatrix}.$$

The SSP conditions require that $M \geq 0$ and $d \leq 1$. Since $r = \mathcal{C} = 4$ we have the relation

$$\hat{b}_1 \geq \hat{b}_2,$$
$$\hat{b}_1 \leq \frac{1}{2}.$$

Taking into account the first order condition

$$\hat{b}_1 + \hat{b}_2 = 1$$

we can conclude that

$$\hat{b}_1 = \hat{b}_2 = \frac{1}{2}.$$

However, this implies that this method is second order since $\hat{b}^T c = \frac{1}{2}$. For the general stage number s we have

$$\hat{b}_i \geq \hat{b}_i + 1, \quad (1 \leq i \leq s-1),$$
$$\hat{b}_1 \leq \frac{1}{s},$$
$$\sum_{i=1}^{s} \hat{b}_i = 1,$$

hence

$$\hat{b}_i = \frac{1}{s}, \qquad (i = 1, \ldots, s).$$

This means that method is second order. $\square$

Taking into account Theorem 5. now we are looking for embedded pairs with smaller $\mathcal{C}$. Namely we choose the case $\mathcal{C} = s$. Our searches suggested three pairs which satisfy the properties a), b), and d), the appropriate order conditions and the SSP conditions. These pairs are

$$\hat{b}_1^T = \left[\frac{2}{s+1}, \ldots, \frac{2}{s+1}, \frac{3}{s+1}\right]^T,$$
$$\hat{b}_2^T = \left[\frac{1}{s}, \ldots, \frac{1}{s}, \frac{5}{4s}, \frac{3}{4s}\right]^T \text{ and}$$
$$\hat{b}_3^T = \left[\frac{1}{s}, \ldots, \frac{1}{s}, \frac{13}{12s}, \frac{10}{12s}, \frac{10}{12s}, \frac{15}{12s}\right]^T.$$

Similarly to the explicit cases we demonstrate the effectiveness of our embedded pairs by plotting their absolute stability regions in Figure 3. by using NodePy Python package [18].

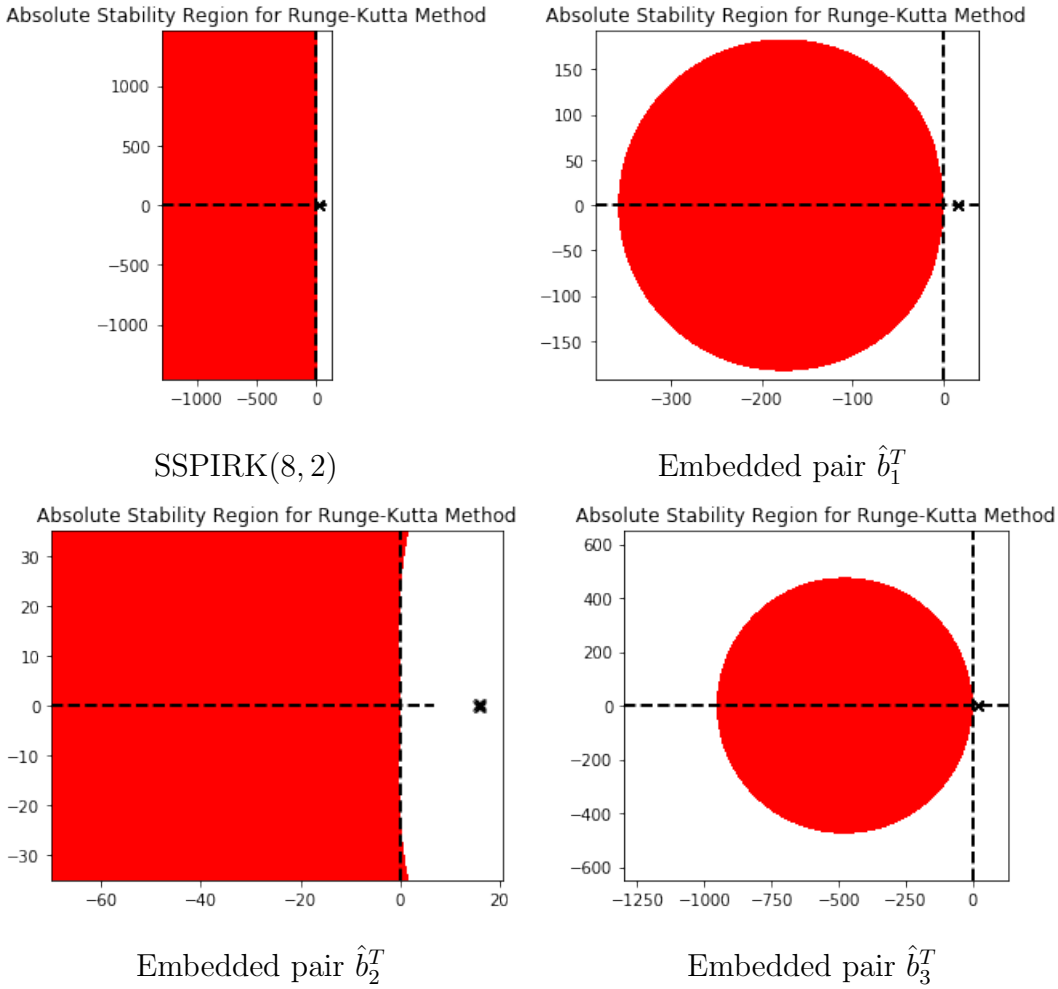


Figure 3: The absolute stability regions of SSPIRK$(8, 2)$ method and its recommended embedded pairs from left to right and top to bottom.

Based on Figure 3. it is obvious that we choose embedded pair $\hat{b}_2^T$. As we increase the number of stages we can see similar results.

SSPIRK(s,3)

The Butcher tableau of optimal SSPIRK$(s,3)$ methods is

$$\begin{array}{c|ccccc}
\beta_1 & \beta_1 & & & & \\
2\beta_1+\beta_2 & \beta_1+\beta_2 & \beta_1 & & & \\
3\beta_1+2\beta_2 & \beta_1+\beta_2 & \beta_1+\beta_2 & \beta_1 & & \\
\vdots & \vdots & \vdots & \ddots & \ddots & \\
s\beta_1+(s-1)\beta_2 & \beta_1+\beta_2 & \beta_1+\beta_2 & \dots & \beta_1+\beta_2 & \beta_1 \\
\hline
 & \frac{1}{s} & \frac{1}{s} & \dots & \frac{1}{s} & \frac{1}{s}
\end{array}$$

where

$$\beta_1 = \frac{1}{2}\left(1-\sqrt{\frac{s-1}{s+1}}\right) \text{ and } \beta_2 = \frac{1}{2}\left(\sqrt{\frac{s+1}{s-1}}-1\right).$$

Our numerical searches failed to find any embedded pair with SSP coefficient $\mathcal{C} = s-1+\sqrt{s^2-1}$.

Theorem 6. There is no embedded pair for SSPIRK$(s,3)$ with SSP coefficient $\mathcal{C} = s-1+\sqrt{s^2-1}$.

Proof. For the sake of simplicity we give the proof in details for the two-stage case. The corresponding Butcher tableau is

$$\begin{array}{c|cc}
\beta_1 & \beta_1 & 0 \\
2\beta_1+\beta_2 & \beta_1+\beta_2 & \beta_1 \\
\hline
 & \frac{1}{s} & \frac{1}{s},
\end{array}$$

where

$$\beta_1 = \frac{1}{2}\left(1-\sqrt{\frac{s-1}{s+1}}\right) \text{ and } \beta_2 = \frac{1}{2}\left(\sqrt{\frac{s+1}{s-1}}-1\right).$$

Let us introduce the matrix $\hat{K}$ as

$$\hat{K} = \begin{pmatrix} \beta_1 & 0 & 0 \\ \beta_1 + \beta_2 & \beta_1 & 0 \\ \hat{b_1} & \hat{b_2} & 0 \end{pmatrix}.$$

The SSP conditions are

$$M = \hat{K}(I + r\hat{K})^{-1} = \begin{pmatrix} \frac{1}{3+\sqrt{3}+r} & 0 & 0 \\ \frac{12\sqrt{3}}{(-6+(\sqrt{3}-3)r)^2} & \frac{1}{3+\sqrt{3}+r} & 0 \\ \frac{6(\hat{b}_2 2\sqrt{3}r+\hat{b}_1(-6+(-3+\sqrt{3})r))}{(-6+(\sqrt{3}-3)r)^2} & \frac{6\hat{b}_2}{6+3r-\sqrt{3}r} & 0 \end{pmatrix}$$

and

$$d = rMe = \begin{bmatrix} 1 + \frac{6}{-6+(-3+\sqrt{3})r} \\ \frac{6r(3+\sqrt{3}-(-2+\sqrt{3})r)}{(-6+(-3+\sqrt{3})r)^2} \\ \frac{6r(3\hat{b}_2(2+r-\sqrt{3})+\hat{b}_1(6-(-3+\sqrt{3})r)}{(-6+(-3+\sqrt{3})r)^2} \end{bmatrix}.$$

The SSP conditions require that $M \geq 0$ and $d \leq 1$. Since $r = \mathcal{C} = 1 + \sqrt{3}$ we have the relations

$$\hat{b}_1 \geq \hat{b}_2 \text{ and } \hat{b}_1 \leq \frac{1}{\sqrt{3}}.$$

The order conditions up to order two are

$$\hat{b}_1 + \hat{b}_2 = 1 \text{ and } \hat{b}^T c = \frac{1}{2}.$$

The abscissa of the method is

$$c = \left[\tfrac{1}{2}\left(1 - \tfrac{1}{\sqrt{3}}\right), \quad 1 - \tfrac{1}{\sqrt{3}} + \tfrac{1}{2}(-1 + \sqrt{3})\right].$$

Using the substitution

$$\hat{b}_1 = 1 - \hat{b}_2$$

we have

$$\begin{aligned}
\hat{b}^T c = \hat{b}^T c &= \Big(\frac{1}{2} - \frac{1}{2\sqrt{3}}\Big)(1-\hat{b}_2) + \Big(1 - \frac{1}{\sqrt{3}} - \frac{1}{2} + \frac{\sqrt{3}}{2}\Big)\hat{b}_2 = \\
&= \frac{1}{2} - \frac{1}{2\sqrt{3}} + \hat{b}_2\Big(\frac{1}{2} - \frac{1}{\sqrt{3}} + \frac{\sqrt{3}}{2} - \frac{1}{2} + \frac{1}{2\sqrt{3}}\Big) = \\
&= \frac{1}{2} - \frac{1}{2\sqrt{3}} + \hat{b}_2\Big(\frac{\sqrt{3}}{2} - \frac{1}{2\sqrt{3}}\Big) = \frac{1}{2} - \frac{1}{2\sqrt{3}} + \hat{b}_2\frac{2}{2\sqrt{3}} = \frac{1}{2}, \\
\hat{b}_2 &= \frac{1}{2\sqrt{3}}\frac{2\sqrt{3}}{2} = \frac{1}{2}, \\
\hat{b}_1 &= 1 - \hat{b}_2 = \frac{1}{2}.
\end{aligned}$$

This provides us a third order method, since

$$\begin{aligned}
\hat{b}^T c^2 =& \frac{1}{3}, \\
\hat{b}^T(c^2/2 - Ac) =& 0.
\end{aligned}$$

For the general case we have

$$\begin{aligned}
\hat{b}_i &\geq \hat{b}_i + 1, \qquad (1 \leq i \leq s-1), \\
\hat{b}_1 &= \hat{b}_s \\
\sum_{i=1}^{s} \hat{b}_i &= 1,
\end{aligned}$$

hence

$$\hat{b}_i = \frac{1}{s}, \qquad (i = 1, \ldots, s).$$

It means that this is a third order method. □

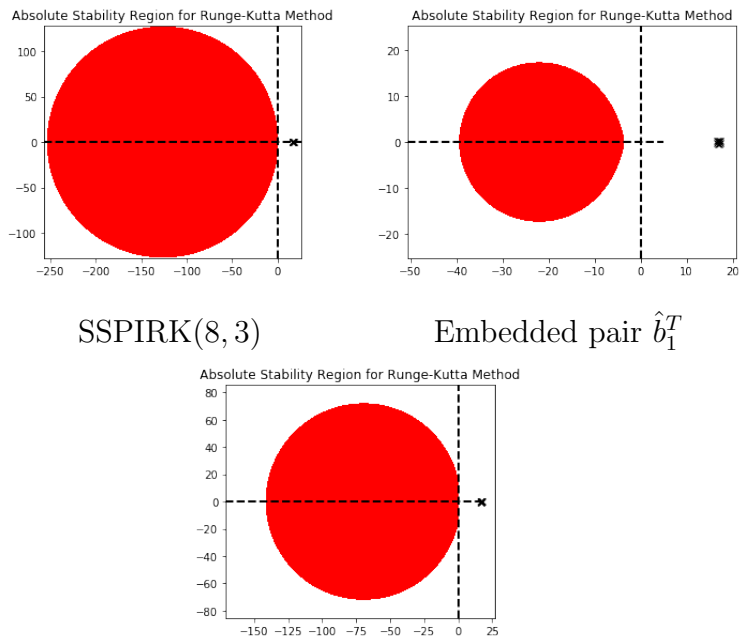
Similarly to the SSPIRK$(s, 2)$ case, we are looking for the halved SSP coefficient, i.e. $\mathcal{C} = \frac{s-1+\sqrt{s^2-1}}{2}$. Our searches suggested two pairs which

satisfy the properties a), b), and d), the appropriate order conditions and the SSP conditions. These pairs are

$$\hat{b}_1^T = \left[\frac{1}{\sqrt{s^2-1}}, \ldots, \frac{1}{\sqrt{s^2-1}}, \frac{s-1-\frac{s-2}{s-1}\sqrt{s^2-1}}{2}, \frac{3-s+\frac{s-2}{s+1}\sqrt{s^2-1}}{2}\right] \text{ and}$$

$$\hat{b}_2^T = \left[\frac{1}{s}, \ldots, \frac{1}{s}, \frac{21s+39-3\sqrt{s^2-1}}{16s^2+34s}, \frac{3s+12+3\sqrt{s^2-1}}{8s^2+17s}, \frac{21s+39-3\sqrt{s^2-1}}{16s^2+34s}\right].$$

Similarly to the previous implicit case we plot the absolute stability regions of the method and the embedded methods in Figure 4.



SSPIRK$(8,3)$ Embedded pair $\hat{b}_1^T$

Embedded pair $\hat{b}_2^T$

Figure 4: The absolute stability regions of SSPIRK$(8,3)$ method and its recommended embedded pairs from left to right and top to bottom.

Based on Figure 4. we recommend embedded pair $\hat{b}_2^T$. As we increase the number of stages we can see similar results.

Acknowledgements

The project has been supported by the European Union, co-financed by the European Social Fund (EFOP-3.6.3-VEKOP-16-2017-00002).

References

[1] Chi-Wang Shu and Stanley Osher. Efficient implementation of essentially non-oscillatory shock-capturing schemes. *Journal of computational physics*, 77(2):439–471, 1988.

[2] Chi-Wang Shu and Stanley Osher. Efficient implementation of essentially non-oscillatory shock-capturing schemes ii. *Journal of computational physics*, 83(1):32–78, 1989.

[3] Guang-Shan Jiang and Chi-Wang Shu. Efficient implementation of weighted eno schemes. *J. Comput. Phys.*, 126(1):202–228, June 1996.

[4] Randall J. LeVeque. *Finite Volume Methods for Hyperbolic Problems*. Cambridge Texts in Applied Mathematics. Cambridge University Press, 2002.

[5] Sigal Gottlieb, David I. Ketcheson, and Chi-Wang Shu. *Strong stability preserving Runge–Kutta and multistep time discretizations*. World Scientific, 2011.

[6] Imre Fekete and Ákos Horváth. Some embedded pairs for optimal implicit strong stability preserving Runge–Kutta methods. *Progress in Industrial Mathematics at ECMI 2018, accepted*, pages 1–6, 2019.

[7] John C. Butcher. On Runge–Kutta processes of high order. *Journal of the Australian Mathematical Society*, 4(2):179–194, 1964.

[8] Inmaculada Higueras. Representations of Runge–Kutta methods and strong stability preserving methods. *SIAM Journal on Numerical Analysis*, 43(3):924–948, 2005.

[9] Luca Ferracina and Marc Spijker. An extension and analysis of the shuosher representation of Runge–Kutta methods. *Mathematics of Computation*, 74(249):201–219, 2005.

[10] Raymond J. Spiteri and Steven J. Ruuth. A new class of optimal high-order strong-stability-preserving time discretization methods. *SIAM Journal on Numerical Analysis*, 40(2):469–491, 2002.

[11] Johannes Franciscus Bernardus Maria Kraaijevanger. Contractivity of Runge–Kutta methods. *BIT Numerical Mathematics*, 31(3):482–528, 1991.

[12] David I. Ketcheson. Highly efficient strong stability-preserving Runge–Kutta methods with low-storage implementations. *SIAM Journal on Scientific Computing*, 30(4):2113–2136, 2008.

[13] David I. Ketcheson, Colin B. Macdonald, and Sigal Gottlieb. Optimal implicit strong stability preserving Runge–Kutta methods. *Applied Numerical Mathematics*, 59(2):373 – 392, 2009.

[14] Ernst Hairer, Syvert P. Nørsett, and Gerhard Wanner. *Solving Ordinary Differential Equations I Nonstiff problems*. Springer, Berlin, second edition, 2000.

[15] Gustaf Söderlind. Automatic control and adaptive time-stepping. *Numerical Algorithms*, 31(1):281–310, Dec 2002.

[16] Gustaf Söderlind. Digital filters in adaptive time-stepping. *ACM Trans. Math. Softw.*, 29(1):1–26, March 2003.

[17] Sidafa Conde, Imre Fekete, and John N. Shadid. Embedded error estimation and adaptive step-size control for optimal explicit strong stability preserving Runge–Kutta methods. *arXiv preprint arXiv:1806.08693*, 2018.

[18] David I. Ketcheson. NodePy software version 0.6.1, http://github.com/ketch/nodepy/, 2015.