Issues with Positivity-Preserving Patankar-type Schemes

Davide Torlo*, Philipp Öffner†, Hendrik Ranocha‡

November 19, 2021

Patankar-type schemes are linearly implicit time integration methods designed to be unconditionally positivity-preserving by going outside of the class of general linear methods. Thus, classical stability concepts cannot be applied and there is no satisfying stability or robustness theory for these schemes. We develop a new approach to study a few related issues that impact some Patankar-type methods. In particular, we demonstrate problematic behaviors of these methods that can lead to undesired oscillations or order reduction on very simple linear problems. Extreme cases of the latter manifest as spurious steady states. We investigate various classes of Patankar-type schemes based on classical Runge-Kutta methods, strong stability preserving Runge-Kutta methods, and deferred correction schemes using our approach. Finally, we strengthen our analysis with challenging applications including stiff nonlinear problems.

Key words. Patankar-type methods, Runge–Kutta methods, deferred correction methods, implicit-explicit methods, semi-implicit methods
AMS subject classification. 65L06, 65L20, 65L04

1. Introduction

Many differential equations in biology, chemistry, physics, and engineering are naturally equipped with constraints such as the positivity of certain solution components (e.g., density, energy, pressure) and conservation (e.g., total mass, momentum, energy). In particular, reaction equations are often of this form. Typically, such reaction systems can also be stiff. We consider such ordinary differential equations (ODEs)

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

that can be written as a production destruction system (PDS) [5]

$$f_i(u) = \sum_{j \in I} (p_{ij}(u) - d_{ij}(u)), \quad \forall i \in I, \tag{2}$$

where $p_{ij}, d_{ij} \geq 0$ are the production and destruction terms, respectively. Sometimes, these terms are conveniently written as matrices $p(u) = (p_{ij}(u))_{i,j}$ and $d(u) = (d_{ij}(u))_{i,j}$.

[*] davide.torlo@inria.fr, Inria Bordeaux - Sud-Ouest, 200 avenue de la Vieille Tour 33405 Talence cedex, France.
[†] poeffner@uni-mainz.de, Institut für Mathematik, Johannes Gutenberg Universität, Staudingerweg 9, 55099 Mainz, Germany
[‡] mail@ranocha.de, Applied Mathematics, University of Münster, Orléans-Ring 10, 48149 Münster, Germany.

Definition 1.1. An ODE (1) is called *positive*, if positive initial data $u_0 > 0$ result in positive solutions $u(t) > 0, \forall t$. Here, inequalities for vectors are interpreted componentwise, i.e., $u(t) > 0$ means $\forall i \in I \colon u_i(t) > 0$. A production destruction system (2) is called *conservative*, if $\forall i, j \in I, \forall u \colon p_{ij}(u) = d_{ji}(u)$.

A slight generalization of the PDS (2) is given by the production destruction rest system (PDRS)

$$f_i(u) = r_i(u) + \sum_{j \in I} (p_{ij}(u) - d_{ij}(u)), \quad \forall i \in I, \tag{3}$$

where p_{ij}, d_{ij} are as before and additional rest terms r_i are introduced. These can of course violate the conservative nature of a PDS but can still result in a positive solution if $r_i \geq 0$. The rest term can be interpreted as additional force/source term.

To ensure physically meaningful and robust numerical approximations, we would like to preserve positivity and conservation discretely.

Definition 1.2. A numerical method computing $u^{n+1} \approx u(t_{n+1})$ given $u^n \approx u(t_n)$ is called *conservative*, if $\sum_i u_i^{n+1} = \sum_i u_i^n$. It is called *unconditionally positive*, if $u^n > 0$ implies $u^{n+1} > 0$.

There are several ways to study positivity of numerical methods [8], e.g., based on the concept of strong stability preserving (SSP) [10] or adaptive Runge–Kutta (RK) methods [26]. However, general linear methods methods are restricted to conditional positivity if they are at least second order accurate [4]. One way to circumvent such order restrictions is given by diagonally split RK methods, which can be unconditionally positive [2, 12, 15]. However, they are less accurate than the unconditionally positive implicit Euler method for large step sizes in practice [21].

Another approach to unconditionally positivity-preserving methods is based on the so-called Patankar trick [28, Section 7.2-2]. First- and second-order accurate conservative methods based thereon were introduced in [5]. Later, these were extended to families of second- and third-order accurate modified Patankar–Runge–Kutta (MPRK) methods based on the Butcher coefficients [17, 19] and the Shu–Osher form [13, 14]. Related deferred correction (DeC) methods were proposed recently [27]. Positive but not conservative methods using the Patankar trick have been proposed and studied in [6], although the connection to Patankar methods seems to be unknown up to now. Other related numerical schemes are inflow-implicit/outflow-explicit methods [9, 24, 25]. Ideas from Patankar-type methods have also been used in numerical methods based on limiters [20].

The methods mentioned above are based on explicit RK methods. To guarantee positivity, the schemes are modified to be linearly implicit, which seems to introduce some stabilization mechanism. In fact, Patankar-type methods have been applied successfully to some stiff systems [6, 16, 17, 19]. However, up to the authors' knowledge, there have been no stability or robustness investigations of Patankar-type methods which are applicable to systems of positive equations.

1.1. Motivating example

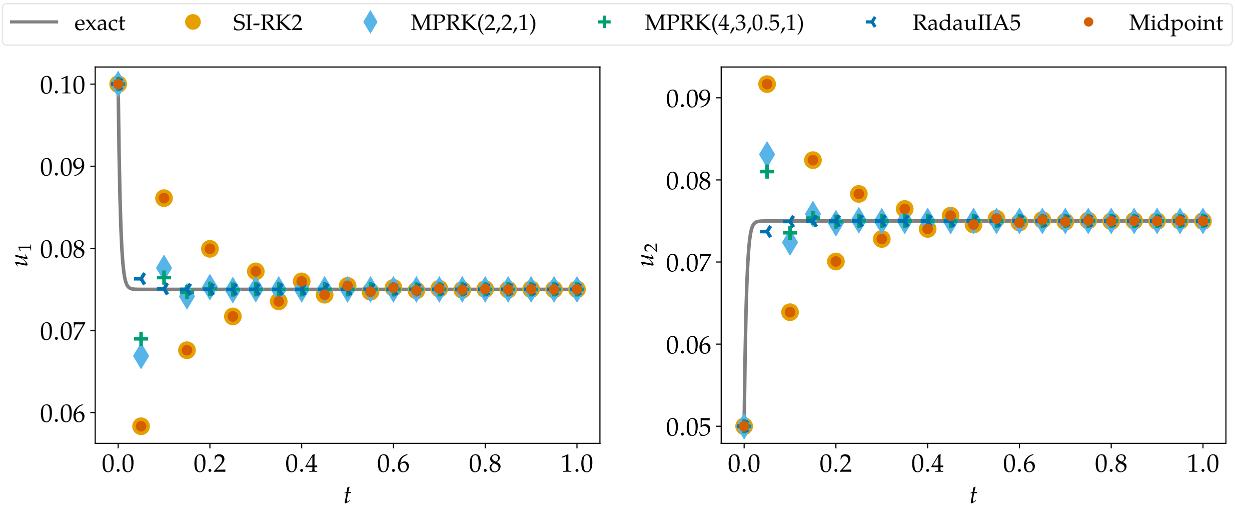
Consider the normal linear system

$$u'(t) = 10^2 \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} u(t), \quad u(0) = u_0 = \begin{pmatrix} 0.1 \\ 0.05 \end{pmatrix}, \tag{4}$$

which can be written as a production destruction system with

$$p(u) = \begin{pmatrix} 0 & 10^2 u_2 \\ 10^2 u_1 & 0 \end{pmatrix}, \quad d(u) = \begin{pmatrix} 0 & 10^2 u_1 \\ 10^2 u_2 & 0 \end{pmatrix}. \tag{5}$$

The eigenvalues of the normal matrix in (4) are 0 and -200. The second order method SI-RK2 of [6] has been shown to be $A(\alpha)$-stable with $\alpha = \pi/4$ and stiff decay [6, Theorem 2.3]. Hence, one could expect that this scheme results in non-oscillating numerical solutions for a normal linear



(a) Component u_1.

(b) Component u_2.

Figure 1: Numerical solutions (time step $\Delta t = 0.05$) of the normal linear system (4) with real and non-positive eigenvalues obtained using different Patankar-type schemes as well as two implicit Runge–Kutta methods.

system such as (4). Similar investigations for scalar equations (3) can be done for the second- and third-order accurate modified Patankar–Runge–Kutta schemes MPRK(2,2,α) and MPRK(4,3,α,β) [17, 19].

However, numerical solutions of (4) are oscillating for three Patankar-type schemes SI-RK2, MPRK(2,2,1) and MPRK(4,3,0.5,1), as visualized in Figure 1 (the precise description of the methods is given in the next section). We are interested in non-oscillating numerical solutions such as the one obtained by the fifth-order, three stage RadauIIA5 scheme [11] implemented in DifferentialEquations.jl [29] in Julia [3].

Another problematic behavior of modified Patankar schemes can be observed on the same problem (4) when one initial condition is very close to zero. Let us consider the IC $u_0 = (1, 10^{-300})^T$, we would expect a fast convergence to the steady state $u_1^* = u_2^* = 1/2$. In Figure 2, we observe that for several modified Patankar methods, *inter alia* MPRK(2,2,2), mPDeC9eq and MPRK(4,3,2,0.5), need long times before leaving the zero state of one of the components. Classical implicit Runge–Kutta method as well as other modified Patankar schemes do not show this behavior and their first time step approaches quickly the steady state value. This issue is linked with a loss of consistency in the limit for an initial condition approaching zero.

1.2. Scope of the article

Positive and conservative schemes naturally satisfy some bounds on the maximum norm of the numerical solution, which is loosely related to the classical concept of A-stability. Motivated by the numerical example above, we are interested in a stability concept excluding the dominant appearance of spurious oscillations, which is loosely related to L-stability. Since Patankar-type methods are not compatible with an eigenvalue analysis, we propose to use a simple and generic linear system as test problem (instead of the classical scalar linear test equation).

We have focused on different types of systems (stiff, dissipative ones, etc.) and considered several quantities like the dissipation of some norms or Lyapunov functionals, cf. [30–34]. However, theses results have not been sufficient to describe the properties of the schemes in an adequate way. Thus, we will measure the amount of spurious oscillations directly.

The rest of the article is structured as follows. The numerical schemes studied in this article are introduced in Section 2. Thereafter, we introduce the oscillation measure and the generic linear system in Section 3. We continue with an analytical investigation of specific schemes on the loss

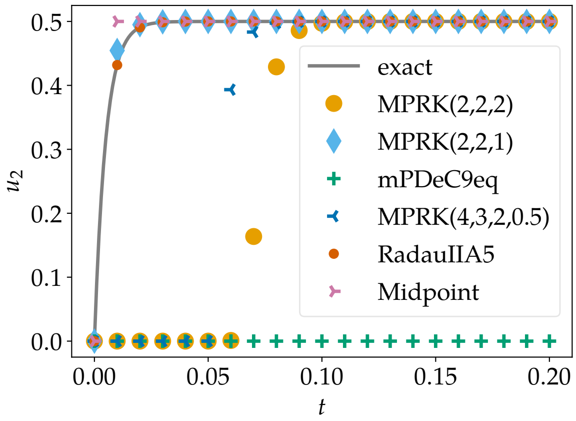


Figure 2: Numerical solutions (time step $\Delta t = 0.01$) of linear system (4) with initial condition $u_0 = (1, 10^{-300})^T$ using several modified Patankar schemes and two implicit Runge–Kutta

of consistency in the limit of vanishing initial condition in Section 4. In Section 5, a numerical study on linear systems derives the results on all schemes that theoretical studies of the previous sections could not found: bounds on time step for oscillation–free schemes and consistency in the vanishing initial condition regime. In Section 6, we extend the numerical study to nonlinear and stiff problems. Finally, we summarize and discuss our results in Section 7.

2. Numerical schemes

Here, we introduce Patankar-type methods proposed in the literature that we will investigate later. In addition, we propose a new MPRK method and give a heuristic how to construct such schemes in general.

2.1. Modified Patankar–Euler method

The explicit Euler method $u^{n+1} = u^n + \Delta t f(u^n)$ can be modified by the Patankar trick [28, Section 7.2-2] for a PDR system (3) to get the positive Patankar–Euler method

$$u_i^{n+1} = u_i^n + \Delta t r_i(u^n) + \Delta t \sum_j \left(p_{ij}(u^n) - d_{ij}(u^n) \frac{u_i^{n+1}}{u_i^n} \right). \tag{6}$$

Indeed, given $r, p, d \geq 0$, the new numerical solution u^{n+1} is obtained by solving a linear system with positive diagonal entries, vanishing off-diagonal entries, and a positive right-hand side.

Since the Patankar–Euler method (6) is not conservative, the modified Patankar–Euler method

$$u_i^{n+1} = u_i^n + \Delta t r_i(u^n) + \Delta t \sum_j \left(p_{ij}(u^n) \frac{u_j^{n+1}}{u_j^n} - d_{ij}(u^n) \frac{u_i^{n+1}}{u_i^n} \right), \tag{MPE}$$

has been introduced in [5] (with additional rest terms r here). The modification of the production terms makes the method conservative if the rest terms r vanish. Nevertheless, the method is still positive, because the arising linear systems has positive diagonal entries, negative off-diagonal entries, and is strictly diagonally dominant. Hence, the system matrix is an M matrix and the solution u^{n+1} given a positive right-hand side is positive [1, Section 6.1]. We observe that, when dealing with the scalar linear test problem $u' = \lambda u$ with $\lambda < 0$, the Patankar–Euler method coincides with the implicit Euler method. Similarly, MPE coincides with the implicit Euler method if we deal with positive and conservative linear PDS. Indeed, the destruction terms $d_i(u) = \sum_j d_{ij}(u)$ must go to 0 if $u_i \longrightarrow 0$ [5]. Since the system is linear, $d_{ij}(u^n) = \tilde{d}_{ij} u_i^n$ with $\tilde{d}_{ij} \in \mathbb{R}_0^+$. Exploiting the conservation properties, we have $p_{ji}(u^n) = \tilde{d}_{ij} u_i^n$. Substituting these formulae in MPE leads to the implicit Euler method.

2.2. MPRK methods using Butcher coefficients

A one-parameter family of MPRK schemes based on the Butcher coefficients of a two stage, second-order RK method was introduced in [17]. Given a parameter $\alpha \in [1/2, \infty)$, the method is

$$\begin{aligned}
y^1 &= u^n, \\
y_i^2 &= u_i^n + \alpha \Delta t r_i(y^1) + \alpha \Delta t \sum_j \left(p_{ij}(y^1) \frac{y_j^2}{y_j^1} - d_{ij}(y^1) \frac{y_i^2}{y_i^1} \right), \\
u_i^{n+1} &= u_i^n + \Delta t \left(\frac{2\alpha - 1}{2\alpha} r_i(y^1) + \frac{1}{2\alpha} r_i(y^2) \right) \\
&\quad + \Delta t \sum_j \left(\left(\frac{2\alpha - 1}{2\alpha} p_{ij}(y^1) + \frac{1}{2\alpha} p_{ij}(y^2) \right) \frac{u_j^{n+1}}{(y_j^2)^{1/\alpha} (y_j^1)^{1-1/\alpha}} \right. \\
&\quad \left. - \left(\frac{2\alpha - 1}{2\alpha} d_{ij}(y^1) + \frac{1}{2\alpha} d_{ij}(y^2) \right) \frac{u_i^{n+1}}{(y_i^2)^{1/\alpha} (y_i^1)^{1-1/\alpha}} \right).
\end{aligned} \tag{MPRK(2,2,α)}$$

The scheme for the choice $\alpha = 1$ is based on Heun's method and has been proposed already in [5]. Heun's method can be also written as a strong stability preserving Runge–Kutta method (SSPRK) and we will denote it by SSPRK(2,2) [10].

A similar two-parameter family MPRK(4,3,α,β) of four stage, third-order accurate schemes was introduced and studied in [18, 19]. The family under consideration can be found in the Appendix A for completeness.

2.3. MPRK methods using Shu–Osher coefficients

A two-parameter family of MPRK schemes based on the Shu–Osher coefficients of a two stage, second-order RK method was introduced in [13]. Given parameters α, β, the method is

$$\begin{aligned}
y^1 &= u^n, \\
y_i^2 &= y_i^1 + \beta \Delta t r_i(y^1) + \beta \Delta t \sum_j \left(p_{ij}(y^1) \frac{y_j^2}{y_j^1} - d_{ij}(y^1) \frac{y_i^2}{y_i^1} \right), \\
u_i^{n+1} &= (1-\alpha) y_i^1 + \alpha y_i^2 + \Delta t \left((1 - \frac{1}{2\beta} - \alpha\beta) r_i(y^1) + \frac{1}{2\beta} r_i(y^2) \right) \\
&\quad + \Delta t \sum_j \left(\left((1 - \frac{1}{2\beta} - \alpha\beta) p_{ij}(y^1) + \frac{1}{2\beta} p_{ij}(y^2) \right) \frac{u_j^{n+1}}{(y_j^2)^{\gamma} (y_j^1)^{1-\gamma}} \right. \\
&\quad \left. - \left((1 - \frac{1}{2\beta} - \alpha\beta) d_{ij}(y^1) + \frac{1}{2\beta} d_{ij}(y^2) \right) \frac{u_i^{n+1}}{(y_i^2)^{\gamma} (y_i^1)^{1-\gamma}} \right),
\end{aligned} \tag{MPRKSO(2,2,α,β)}$$

where the parameters are restricted to $\alpha \in [0, 1]$, $\beta \in (0, \infty)$, $\alpha\beta + \frac{1}{2\beta} \leq 1$, and

$$\gamma = \frac{1 - \alpha\beta + \alpha\beta^2}{\beta(1 - \alpha\beta)}, \tag{7}$$

in order to be positive. In our simulations, we will exchange the weights of production and destruction when the coefficients are negative. In the next section we will give an example of such inversion. An extension to four stage, third-order accurate methods MPRKSO(4,3) was developed in [14] and can be found in the Appendix A.

2.4. Modified Patankar deferred correction schemes

Arbitrarily high-order conservative and positive modified Patankar deferred correction schemes (mPDeC) were introduced in [27]. A time step $[t^n, t^{n+1}]$ is divided into M subintervals, where $t^{n,0} = t^n$ and $t^{n,M} = t^{n+1}$. For every subinterval, the Picard-Lindelöf theorem is mimicked as follows. At each subtimestep $t^{n,m}$, an approximation y^m is calculated. An iterative procedure of K correction steps improves the approximation by one order of accuracy at each iteration. The modified Patankar trick is introduced inside the basic scheme to guarantee positivity and conservation of the intermediate approximations. Using the fact that initial states $y_i^{0,(k)} = u_i^n$ are identical for any correction k, the mPDeC correction steps can be rewritten for $k = 1, \dots, K$, $m = 1, \dots, M$ and $\forall i \in I$ as

$$\begin{aligned} & y_i^{m,(k)} - y_i^0 - \sum_{r=0}^{M} \theta_r^m \Delta t r_i \left(y^{r,(k-1)}\right) - \\ & \sum_{l=0}^{M} \theta_l^m \Delta t \sum_{j=1} \left(p_{ij}(y^{l,(k-1)}) \frac{y^{m,(k)}_{\gamma(j,i,\theta_r^m)}}{y^{m,(k-1)}_{\gamma(j,i,\theta_l^m)}} - d_{ij}(y^{l,(k-1)}) \frac{y^{m,(k)}_{\gamma(i,j,\theta_l^m)}}{c^{m,(k-1)}_{\gamma(i,j,\theta_l^m)}} \right) = 0, \end{aligned} \tag{mPDeC}$$

where θ_r^m are the correction weights and the $\gamma(j, i, \theta_r^m)$ are the indicator functions depending on θ if the values are positive or negative, see [27] for details. This allow to obtain always positive terms in the diagonal terms and nonpositive in the offdiagonal terms of the system matrix. Finally, the new numerical solution is $u_i^{n+1} = y^{M,(K)}$.

The choice of the distribution and the number of subtimesteps M and the number of iterations K determines the order of accuracy of the scheme. In the following, we will use equispaced and Gauss–Lobatto points. To reach order d, we use $M = d - 1$ subintervals and $K = d$ corrections. We will denote the pth-order mPDeC method as mPDeCp. Note that mPDeC1 is equivalent to MPE and mPDeC2 is equivalent to MPRK(2,2,1).

2.5. A new MPRK method

We proposed the following new three stage, second-order MPRK method based on SSPRK(3,3):

$$
\begin{aligned}
y_i^1 &= u_i^n,\\
y_i^2 &= u_i^n + \Delta t r_i(y^1) + \Delta t \sum_j \left(p_{ij}(y^1) \frac{y_j^2}{y_j^1} - d_{ij}(y^1) \frac{y_i^2}{y_i^1} \right),\\
y_i^3 &= u_i^n\\
&\quad + \Delta t \frac{r_i(y^1) + r_i(y^2)}{4} + \Delta t \sum_j \left(\frac{p_{ij}(y^1) + p_{ij}(y^2)}{4} \frac{y_j^3}{y_j^2} - \frac{d_{ij}(y^1) + d_{ij}(y^2)}{4} \frac{y_i^3}{y_i^2} \right),\\
u_i^{n+1} &= u_i^n + \Delta t \frac{r_i(y^1) + r_i(y^2) + 4 r_i(y^3)}{6}\\
&\quad + \Delta t \sum_j \left(\frac{p_{ij}(y^1) + p_{ij}(y^2) + 4 p_{ij}(y^3)}{6} \frac{u_j^{n+1}}{y_j^2} \right.\\
&\qquad \left. - \frac{d_{ij}(y^1) + d_{ij}(y^2) + 4 d_{ij}(y^3)}{6} \frac{u_i^{n+1}}{y_i^2} \right).
\end{aligned}
\tag{MPRK(3,2)}
$$

For explicitly time-dependent problems, the abscissae are the ones of SSPRK(3,3) [10], i.e., $c = (0, 1, 0.5)$. As will be seen later, this scheme has some desirable robustness. MPRK(3,2) is second-order accurate. This can be seen through the following observation. The second stage y_i^2 is an approximation of order one. Next, the midpoint rule is applied as the quadrature which is second-order accurate. In the final stage, the Simpson rule is applied, where we get only second order accuracy since we use the first-order approximation which is multiplied by Δt. At the end, the scheme is second-order accurate.

Remark 2.1. The construction of higher-order MPRK schemes can be done in a similar way. The basic idea is to create a method with increasing stage order, similar to the construction of mPDeC. Starting from a high-order RK scheme, by applying the modified Patankar trick in the substeps in combination with quadrature rules should lead to high-order modified Patankar RK schemes. Essential in the construction is the fact that more stages have to be applied compared to classical RK schemes. This is in accordance with the result of [18] on the existence of third-order, three stages MPRK schemes. There is work in progress to describe a general recipe to construct MPRK schemes of arbitrary order and to study the properties of these schemes.

2.6. Semi-implicit methods

The semi-implicit methods of [6] are also based on the Shu–Osher representation of SSP RK methods, which can be decomposed into convex combinations of the previous step value and explicit Euler steps. Instead of introducing Patankar weights multiplying all destruction terms for a step/stage update, a Patankar weight is introduced for the destruction terms of each Euler stage which is used to compute the new value. Since this procedure limits the order of accuracy of the resulting scheme to first order, an additional function evaluation is used to correct the final solution and get second order of accuracy.

The two methods proposed in [6] are

$$\begin{aligned}
y^1 &= u^n, \\
y_i^2 &= \frac{u_i^n + \Delta t r_i(y^1) + \Delta t \sum_j p_{ij}(y^1)}{1 + \Delta t \sum_j d_{ij}(y^1)/y_i^1}, \\
y_i^3 &= \frac{1}{2} u_i^n + \frac{1}{2} \frac{y_i^2 + \Delta t r_i(y^2) + \Delta t \sum_j p_{ij}(y^2)}{1 + \Delta t \sum_j d_{ij}(y^2)/y_i^2}, \\
u_i^{n+1} &= \frac{y_i^3 + \Delta t^2 \bigl(r_i(y^3) + \sum_j p_{ij}(y^3)\bigr) \sum_j d_{ij}(y^3)/y_i^3}{1 + \bigl(\Delta t \sum_j d_{ij}(y^3)/y_i^3\bigr)^2},
\end{aligned} \tag{SI-RK2}$$

which uses three stages and is based on SSPRK(2,2), and

$$\begin{aligned}
y^1 &= u^n, \\
y_i^2 &= \frac{u_i^n + \Delta t r_i(y^1) + \Delta t \sum_j p_{ij}(y^1)}{1 + \Delta t \sum_j d_{ij}(y^1)/y_i^1}, \\
y_i^3 &= \frac{3}{4} u_i^n + \frac{1}{4} \frac{y_i^2 + \Delta t r_i(y^2) + \Delta t \sum_j p_{ij}(y^2)}{1 + \Delta t \sum_j d_{ij}(y^2)/y_i^2}, \\
y_i^4 &= \frac{1}{3} u_i^n + \frac{2}{3} \frac{y_i^3 + \Delta t r_i(y^3) + \Delta t \sum_j p_{ij}(y^3)}{1 + \Delta t \sum_j d_{ij}(y^3)/y_i^3}, \\
u_i^{n+1} &= \frac{y_i^4 + \Delta t^2 \bigl(r_i(y^4) + \sum_j p_{ij}(y^4)\bigr) \sum_j d_{ij}(y^4)/y_i^4}{1 + \bigl(\Delta t \sum_j d_{ij}(y^4)/y_i^4\bigr)^2},
\end{aligned} \tag{SI-RK3}$$

which uses four stages and is based on SSPRK(3,3).

The relation to Patankar schemes becomes obvious by rewriting the computation of the stage y^2 of (SI-RK2) as

$$y_i^2 = u_i^n + \Delta t r_i(y^1) + \Delta t \sum_j \left(p_{ij}(y^1) - d_{ij}(y^1) \frac{y_i^2}{y_i^1} \right), \tag{8}$$

which is the Patankar–Euler method (6). As for the Patankar–Euler method, the semi-implicit methods of [6] are not conservative, i.e., it is not guaranteed that $\sum_i u_i^n = \sum_i u_i^{n+1}$ when the system is conservative.

2.7. Steady state preservation

Motivated by the investigations of [6], steady state preservation for (modified) Patankar–Runge–Kutta methods will be studied here. Except for the SI-RK2 and SI-RK3 methods [6], such investigations cannot be found in the literature.

Definition 2.2. A method is steady state preserving if, given a time step Δt and $u^n = u^*$ with $r_i(u^*) + \sum_j p_{ij}(u^*) - d_{ij}(u^*) = 0$, then $u^{n+1} = u^n = u^*$.

Proposition 2.3. *All (modified) Patankar methods described above are steady state preserving.*

Proof. The solution to each stage and the new step value are unique. If the initial condition is a steady state, this steady state is also a valid solution to all stage and step equations. Hence, the steady state is preserved. □

This theorem is important, since some related modifications of explicit Runge–Kutta methods such as IMEX methods are not necessarily steady state preserving [6]. For (stiff) systems with an initial condition near a steady state, the ability to preserve this steady state exactly is desirable and usually results in a better approximation of solutions nearby or decaying to steady state. This result is also interesting since we will observe spurious steady states of Patankar-type methods in the following section.

3. Oscillation–free Patankar-type schemes for linear problems

Here, we introduce a new approach to study the behavior of Patankar-type schemes. Recall that a classical stability analysis using scalar problems does not generalize to systems for Patankar-type methods, since these do not commute with diagonalization. Thus, instead of Dahlquist's equation, which is not a PDS, we propose to use a 2 × 2 linear system similar to (4) as test problem. More precisely, we consider the general 2 × 2 production-destruction linear system

$$\begin{pmatrix} u_1' \\ u_2' \end{pmatrix} = \begin{pmatrix} -a & b \\ a & -b \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}. \tag{9}$$

Rescaling the time, we can simplify this system to a one parameter system setting $a + b = 1$ and $0 \leq \theta = a \leq 1$, i.e.,

$$\begin{pmatrix} u_1' \\ u_2' \end{pmatrix} = \begin{pmatrix} -\theta & (1-\theta) \\ \theta & -(1-\theta) \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}. \tag{10}$$

We can also rescale any initial condition $u^0 = (u_1^0, u_2^0)^T$ to sum up to one (scaling by a factor $\frac{1}{u_1^0 + u_2^0}$). Thus, we consider the initial condition

$$\begin{pmatrix} u_1^0 \\ u_2^0 \end{pmatrix} = \begin{pmatrix} 1-\varepsilon \\ \varepsilon \end{pmatrix}, \tag{11}$$

with $0 < \varepsilon < 1$. The exact solution of the problem is

$$\begin{pmatrix} u_1(t) \\ u_2(t) \end{pmatrix} = \begin{pmatrix} (1-\theta) + (\theta - \varepsilon)e^{-t} \\ \theta + (\varepsilon - \theta)e^{-t} \end{pmatrix}, \tag{12}$$

and the steady state of the system is $u^* = (1-\theta, \theta)^T$.

In this section we try to find schemes that do not show oscillatory behavior around the steady state as the one presented in Figure 1. This reduces to finding schemes that do not overshoot/undershoot the steady state solution. In particular, if $u_1^0 > u_1^*$ and $u_2^0 < u_2^*$, the first step should be above the steady state solution $u^* = (u_1^*, u_2^*)^T$ for the u_1 component and below for the u_2 component, as RadauIIA5 is doing in Figure 1. Thus, we consider the *oscillation measure*

$$\max \left\{ |u_1^{n+1} - u_1^n| - |u_1^n - u_1^*|, 0 \right\} \tag{13}$$

to detect whenever a scheme is not behaving in this way. This oscillation measure vanishes for non-oscillatory schemes and increases with the amplitude of oscillations. When the initial conditions and the system taken in consideration are arbitrary, i.e., checking for all $0 < \varepsilon < 1$, we can use this measure to find oscillation–free schemes. This measure cannot detect all the possible oscillations, e.g. oscillations entirely contained above the steady state solution. In all the run experiments, such situation never appeared for all the proposed schemes and for simple schemes we can prove that it cannot happen (more on this in Remark 3.1 and in Appendix B). Hence, the measure (13) helps us in obtaining a very simple criterion on oscillation-free solutions studying just one time step.

Remark 3.1 (Oscillations around other states and initial direction)**.** We have never observed modified Patankar schemes oscillating around other states rather than the steady state for the PDS (10). For some schemes one can prove that w.l.o.g.

$$\text{if } u_1^0 > u_1^*, \text{ then } u_1^1 < u_1^0. \tag{14}$$

In Appendix B we show it for MPE and for MPRK(2,2,α) for $\alpha \leq 1$ and for MPRKSO(2,2,α,β) for $\gamma \geq 1$. For MPRK(3,2) we show it in the Mathematica [37] notebook `MPRK_3_2.nb`. For all other methods we conduct a numerical study. This condition is met for almost all schemes. The few exceptions are very high order methods. Nevertheless, also in these exceptions, we do not observe oscillations, rather the approximation (monotonically) converges to a spurious steady state $\tilde{u}_1 \neq u_1^*$ for very low initial conditions and θ. A summary of theoretical and numerical results on the property (14) is done in Appendix B.

Since we are interested in non-oscillatory behavior, we need to check whether

$$|u_1^1 - u_1^0| \leq |u_1^0 - u_1^*| \tag{15}$$

for every initial condition (IC) $0 < \varepsilon < 1$ and for every system $0 \leq \theta \leq 1$. We can simplify the search exploiting the symmetry of the system, for example considering only $0 < \varepsilon \leq 0.5$. It suffices to check the initial step, since every other step will fall back in another IVP (10) with a different IC.

Remark 3.2 (Equivalent non–oscillatory condition)**.** We can rewrite the previous condition as a positivity condition for the diagonalized system. Rewriting it into a matrix formulation

$$u' = Au = \begin{pmatrix} -\theta & (1-\theta) \\ \theta & -(1-\theta) \end{pmatrix} u,$$

we can obtain the diagonal form $A = R\Lambda R^{-1}$ of the system, i.e.,

$$\Lambda = \begin{pmatrix} -1 & 0 \\ 0 & 0 \end{pmatrix}; \quad R = \begin{pmatrix} 1 & 1-\theta \\ -1 & \theta \end{pmatrix}; \quad R^{-1} = \begin{pmatrix} \theta & -(1-\theta) \\ 1 & 1 \end{pmatrix}.$$

So for $v = R^{-1}u$ we can write an always positive (or always negative) exact solution for the first component

$$v_1 = \theta u_1 - (1-\theta)u_2; \qquad v_1' = -v_1; \quad v_1 = e^{-t}v_1^0.$$

The second equation corresponds to the total mass conservation. This tells us also what we want to preserve: the sign of v_1. If it starts positive, it should stay positive; if it starts negative, it should stay negative.

For a general linear method such as RK methods, this sign condition and the *oscillation-free* condition are equivalent, as we can always pass to the diagonal form. For example, the implicit–Euler is unconditionally positive and thus also unconditionally oscillations-free. Since Patankar-type methods are not general linear methods, they are not necessarily oscillations-free, even if they are unconditionally positive.

3.1. Oscillatory-free restrictions of MPRK(2,2,1)

The method MPRK(2,2,α) with $\alpha = 1$ is equivalent to mPDeC2. Since it is simple enough, a detailed analysis for the simplified linear systems (10) is feasible.

Theorem 3.3 (Time restriction for mPDeC2 for 2×2 linear systems)**.** *Consider the system* (10) *with the initial conditions* (11)*. mPDeC2 is oscillation-free in the sense of* (15) *for any initial condition* $0 < \varepsilon < 1$ *and any system* $0 \leq \theta \leq 1$ *under the time step restriction* $\Delta t \leq 2$*. For the general linear system* (9) *the time restriction is* $\Delta t \leq \frac{2}{a+b}$.

Proof. First of all, the cases $\theta = 0$ and $\theta = 1$ are trivially verified as the steady state solutions are $(1,0)^T$ and $(0,1)^T$, respectively. Since the scheme is positive, $0 < u_1^n, u_2^n < 1$ holds for any possible initial condition and timestep, verifying the *oscillation-free* condition.

Secondly, the case $\varepsilon = \theta$ implies that the initial condition is the steady state. Since all modified Patankar schemes are able to unconditionally preserve the steady state, the solution will be steady.

In the general case, we can write the solution at the first time step as ratio of polynomials that are of degree one in Δt and θ and of degree two in ε. We refer to the computations inside `MPRK_2_2_1_generalSystem.nb` in the reproducibility repository [36]. The condition (15) simplifies to $u_2^1 \geq \theta$ in the case $\varepsilon > \theta$ and to $u_2^1 \leq \theta$ if $\varepsilon < \theta$. In `MPRK_2_2_1_generalSystem.nb` [36], we show how both of these conditions lead to the condition $\Delta t \leq z$, where z is the only positive zero of the polynomial $p_{\varepsilon,\theta}(x)$

$$p_{\varepsilon,\theta}(x) = x^3 - x^2 - 2\left(\frac{\varepsilon}{\theta} + \frac{1-\varepsilon}{1-\theta}\right)x - 2\frac{\varepsilon(1-\varepsilon)}{\theta(1-\theta)}. \tag{16}$$

In `MPRK_2_2_1_generalSystem.nb` we also check that the discriminant of this polynomial is positive and, hence, its three roots are real and distinct. Denoting with $y \leq w \leq z$ the three zeros of $p_{\varepsilon,\theta}(x)$, we see that they have to satisfy

$$\begin{cases} y + w + z = 1, \\ yz + wz + yw = -2\left(\frac{\varepsilon}{\theta} + \frac{1-\varepsilon}{1-\theta}\right) < -2, \\ ywz = 2\frac{\varepsilon(1-\varepsilon)}{\theta(1-\theta)} > 0. \end{cases} \tag{17}$$

Since ywz is positive and $yz + wz + yw$ is negative, it is clear that only one root is positive, while the other two are negative, w.l.o.g. $y \leq w < 0 < z$. From the second equation of (17), we see that

$$z(w + y) < z(w + y) + wy = yz + wz + yw < -2, \tag{18}$$

$$w + y < -\frac{2}{z}. \tag{19}$$

Using then the first equation of (17), we have that

$$0 = z + y + w - 1 < z - \frac{2}{z} - 1, \quad 0 < z^2 - z - 2, \tag{20}$$

which has positive solutions only for $z > 2$. Hence, $\Delta t \leq 2$ in order to avoid oscillations for all systems (10). The bound is sharp in the sense that it can be reached for the limit polynomial $\lim_{\theta\to 0}\lim_{\varepsilon\to 0} p_{\varepsilon,\theta}(x)$. We can observe that when $\varepsilon \longrightarrow 0$, the third equation in (17) tells us that $w \longrightarrow 0^-$. Hence, from the second equation we can see that $y \longrightarrow -2\frac{1}{(1-\theta)z}$. Finally, the third zero will converge to

$$z \longrightarrow \frac{1 + \sqrt{1 + \frac{8}{1-\theta}}}{2}.$$

For $\theta \longrightarrow 0$, z goes to 2. □

Unfortunately, the computational complexity increases significantly for all other schemes considered in this article. Thus, we will perform numerical studies for all methods, using different initial conditions (ε), systems (θ), and step sizes (Δt) to find the largest possible timestep without oscillations in Section 5.

4. Spurious steady states

Another particular behavior we observe for some modified Patankar schemes is the loss of consistency when one component of the initial condition tends to zero. In that case, available analytical consistency and convergence results do not hold as these suppose $u_i^0 \geq \varepsilon > 0$. Nevertheless, this condition is of general interest in many applications, where physical/chemical/biological constituents might be zero.

In this section, we show for few schemes and a particular test problem when this situation occurs and demonstrate consequences. A more detailed study is performed in Sections 5 and 6.

Here, we consider the linear initial value problem (10) with $\theta = 0.5$, i.e.,

$$u'(t) = f(u(t)) = \frac{1}{2}\begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} u(t), \quad u(0) = u^0 = \begin{pmatrix} 1-\varepsilon \\ \varepsilon \end{pmatrix}. \tag{21}$$

To use the modified Patankar schemes with a generic implementation, ε must be strictly positive to avoid division by zero. As recommended in the literature [17], we set ε to the smallest positive number that can be represented as floating point number with given precision (usually 64 bit) whenever we are interested in the limit $\varepsilon \to 0$. In the following we first study the behavior of some of the previously presented schemes for $\varepsilon \to 0$, then we show where this study is meaningful and where it is less.

4.1. Inconsistency for MPRK(2,2,α)

In order to explain the kind of computations used in the following, we start with a simple example using again MPRK(2,2,α) with $\alpha = 1$ to show how we study the consistency of a method in the limit of $\varepsilon \to 0$. Recall that the system (21) conserves the total mass $u_1(t) + u_2(t) = 1$ and can be formulated as

$$\begin{aligned} u_1'(t) &= \frac{-u_1(t) + u_2(t)}{2} = \frac{1}{2} - u_1(t), \\ u_2'(t) &= \frac{u_1(t) - u_2(t)}{2} = \frac{1}{2} - u_2(t), \end{aligned}$$

with steady state solution $u_1^* = u_2^* = \frac{1}{2}$ and exact solutions

$$u_1(t) = \frac{1}{2}(1 + e^{-t}(1-2\varepsilon)) \text{ and } u_2(t) = 1 - u_1(t).$$

Example 4.1 (Accuracy of MPRK(2,2,1)). We investigate the behavior of MPRK(2,2,1) applied to (21). Due to the conservation property and the symmetry of the problem, it suffices to focus on the first component u_1. For the first non-trivial stage, we obtain

$$y_1^2 = u_1^0 + \frac{\Delta t}{2}\left(u_2^0 \frac{y_2^2}{u_2^0} - u_1^0 \frac{y_1^2}{u_1^0}\right) = u_1^0 + \frac{\Delta t}{2}\left(1 - 2y_1^2\right) \iff y_1^2 = \frac{u_1^0 + \frac{\Delta t}{2}}{1 + \Delta t}. \tag{22}$$

Using this expression for y_1^2, the new numerical solution satisfies

$$\begin{aligned} u_1^1 &= u_1^0 + \frac{\Delta t}{4}\left(\left((1-u_1^0) + (1-y_1^2)\right)\frac{1-u_1^1}{1-y_1^2} - \left(u_1^0 + y_1^2\right)\frac{u_1^1}{y_1^2}\right) \\ &= u_1^0 + \frac{\Delta t}{4}\left(\frac{(1-u_1^0)(1-u_1^1)}{1-y_1^2} + 1 - u_1^1 - \frac{u_1^0 u_1^1}{y_1^2} - u_1^1\right). \end{aligned} \tag{23}$$

This can be reformulated as

$$\left(1 + \frac{\Delta t}{4}\frac{1-u_1^0}{1-y_1^2} + \frac{\Delta t}{2} + \frac{\Delta t}{4}\frac{u_1^0}{y_1^2}\right) u_1^1 = u_1^0 + \frac{\Delta t}{4}\frac{1-u_1^0}{1-y_1^2} + \frac{\Delta t}{4}. \tag{24}$$

Passing to the limit $\varepsilon \to 0$ with $u_1^0 = 1$ and $\lim_{\varepsilon\to 0} y_1^2 = (1 + \Delta t/2)/(1 + \Delta t)$ yields

$$\begin{aligned} &\left(1 + \frac{\Delta t}{2} + \frac{\Delta t}{4}\frac{(1+\Delta t)}{1 + \Delta t/2}\right)\lim_{\varepsilon\to 0} u_1^1 = 1 + \frac{\Delta t}{4} \\ \iff \quad &\lim_{\varepsilon\to 0} u_1^1 = \frac{8 + 6\Delta t + \Delta t^2}{8 + 10\Delta t + 4\Delta t^2} = 1 - \frac{\Delta t}{2} + \frac{\Delta t^2}{4} - \frac{\Delta t^3}{16} + \mathcal{O}(\Delta t^4). \end{aligned} \tag{25}$$

This tells us that the solution is consistent with the exact one and that, as expected, the second order error is an $\mathcal{O}(\Delta t^3)$ for the first time step, indeed

$$u_1(\Delta t) = 1 - \frac{\Delta t}{2} + \frac{\Delta t^2}{4} - \frac{\Delta t^3}{12} + \mathcal{O}(\Delta t^4). \tag{26}$$

In general this is not true. We apply now for different α the same procedure on MPRK(2,2,α) for the symmetric problem (21). We observe the following behaviors.

- For $\alpha > 1$, taking the limit $\varepsilon \longrightarrow 0$ results in $u^1 = u^0$ and, by induction, $u^n \equiv u^0$. This can also be observed numerically if sufficiently high accuracy is used, e.g. `BigFloat` in Julia. For `Float64`, the first few steps do almost nothing and later steps result in the desired behavior. The number of steps necessary to actually do something increases for $\alpha \gg 1$.
- For $\alpha \in [1/2, 1]$, the schemes are consistent also for $\varepsilon \longrightarrow 0$, but we lose one order of accuracy.

These are analyzed in detail in the following.

Theorem 4.2. *For the test problem* (21) *in the limit* $\varepsilon \longrightarrow 0$ *the initial state becomes a spurious steady state for MPRK(2,2,α) with* $\alpha > 1$.

Proof. This proof makes use of explicit calculations using Mathematica [37]. All calculations can be found in the notebook `MPRK_2_2_alpha.nb` in the accompanying reproducibility repository [36].

For $\alpha > 1$ and $\varepsilon > 0$, the first step of (21) can be computed explicitly. The second component after the first step is of the form $u_2^1 = h_1(\varepsilon)/h_2(\varepsilon)$, where $\lim_{\varepsilon\to 0} h_1(\varepsilon) = \lim_{\varepsilon\to 0} h_2(\varepsilon) = 0$. Defining $\tilde{h}_i(\varepsilon) := \frac{h_i(\varepsilon)}{\varepsilon}$ for $i = 1, 2$, we can rewrite $u_2^1 = \tilde{h}_1(\varepsilon)/\tilde{h}_2(\varepsilon)$, where

$$\lim_{\varepsilon \longrightarrow 0} \tilde{h}_1(\varepsilon) = 2^{\frac{1}{\alpha}} \Delta t(\Delta t - 4 - 4\Delta t\alpha)\left(\frac{\Delta t\alpha}{1 + \Delta t\alpha}\right)^{\frac{1}{\alpha}} \tag{27}$$

and this hold for all $\alpha > \frac{1}{2}$, while the denominator is

$$\lim_{\varepsilon \longrightarrow 0} \tilde{h}_2(\varepsilon) = \infty, \quad \alpha > 1, \tag{28}$$

results in $\lim_{\varepsilon\to 0} u_2^1 = 0 = \lim_{\varepsilon\to 0} u_2^0$ for $\alpha > 1$. Since the sum of all components of u is conserved, $\lim_{\varepsilon\to 0} u^0 = (1, 0)$ is a spurious steady state. □

We remark that numerically this is appreciable also because we lose the consistency when $\varepsilon \longrightarrow 0$. Indeed, we will have that $\lim_{\varepsilon\to 0} u_1^1 = 1$, which is not consistent with the exact solution.

Theorem 4.3. *Consider the application of MPRK(2,2,α) with* $\alpha \in [0.5, 1]$ *to the test problem* (21) *in the limit* $\varepsilon \longrightarrow 0$*. The order of accuracy for* $\alpha < 1$ *reduces to 1.*

Proof. This proof makes use of explicit calculations using Mathematica [37]. All calculations can be found in the notebook `MPRK_2_2_alpha.nb` in the accompanying reproducibility repository [36].

As in the proof of Theorem 4.2, we evaluate the limit $\varepsilon \longrightarrow 0$ of $u_2^1 = \tilde{h}_1(\varepsilon)/\tilde{h}_2(\varepsilon)$. The expression of $\lim_{\varepsilon\to 0} \tilde{h}_1(\varepsilon)$ is given in (27), while for $\tilde{h}_2$ we have a different expression. In case $\alpha < 1$, we have

$$\lim_{\varepsilon \longrightarrow 0} \tilde{h}_2(\varepsilon) = \left(\frac{\Delta t\alpha}{1 + \Delta t\alpha}\right)^{\frac{1}{\alpha}} \left(-8(1 + \Delta t\alpha)(2 + \Delta t\alpha)^{\frac{1}{\alpha}} - 2^{\frac{1}{\alpha}} \Delta t(4 - \Delta t + 4\alpha\Delta t)\right), \tag{29}$$

while for $\alpha = 1$ we have the extra term $-2\frac{2+\Delta t}{1+\Delta t}\Delta t^2$. Using (27) from before and Taylor expansion in Δt,

$$\lim_{\varepsilon \longrightarrow 0} u_2^1(\varepsilon) = \frac{\lim_{\varepsilon\to 0} \tilde{h}_1(\varepsilon)}{\lim_{\varepsilon\to 0} \tilde{h}_2(\varepsilon)} = \begin{cases} 1 - \dfrac{\Delta t}{2} + \dfrac{\Delta t^2}{4} - \dfrac{\Delta t^3}{16} + \mathcal{O}(\Delta t^4), & \alpha = 1, \\ 1 - \dfrac{\Delta t}{2} + \dfrac{\Delta t^2}{8} + \dfrac{\alpha \Delta t^3}{16} + \mathcal{O}(\Delta t^4), & \alpha \in [0.5, 1). \end{cases} \tag{30}$$

Hence, the term in Δt^2 for $\alpha < 1$ does not coincide with the Taylor expansion of the exact solution (26), resulting in a first order accurate scheme for $\varepsilon \longrightarrow 0$. Note the discontinuity at $\alpha = 1$ of $\lim_{\varepsilon\to 0} u_2^1(\varepsilon)$. □

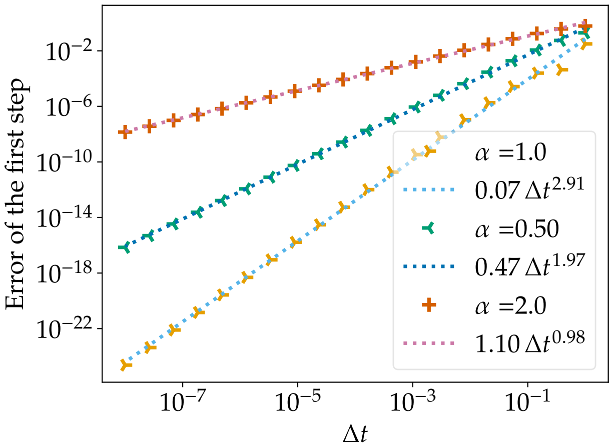


Figure 3: Convergence study of the error of the first step for different members of the family MPRK(2,2,α) for the test problem (21) with ε = `eps(BigFloat)`.

Remark 4.4. This result does not demonstrate that there is an error in the proofs of the order of accuracy of MPRK(2,2,α) [17]. Indeed, studies of the order of accuracy focus on fixed $\varepsilon > 0$ and the limit $\Delta t \to 0$. Numerical experiments with $\alpha > 1$ suggest that the numerical solutions stays approximately constant for a certain number of steps determined by ε (and with less sensitivity also by Δt) until small changes have accumulated and the exponential decay of u_1 becomes visible. In particular, the limits $\Delta t \to 0$ and $\varepsilon \to 0$ are not interchangeable.

Expanding Remark 4.4, a careful error analysis can be conduced by constructing Taylor expansions of the error after the first step for $\Delta t \to 0$ and $\varepsilon \to 0$ in both possible orders. Expanding at first around $\Delta t = 0$ shows that the leading order errors contain terms proportional to ε^{-1} for $\alpha \neq 1$, both for $\alpha < 1$ and for $\alpha > 1$. However, these leading order terms are in agreement with the analysis of [17], i.e., they are proportional to Δt^3.

More insights can be gained by studying the expansions first for $\varepsilon \to 0$, expanding around $\Delta t = 0$ afterwards. Then, the leading order terms in ε are $O(\Delta t^3)$ for $\alpha = 1$, $O(\Delta t^2)$ for $\alpha = 0.5$, and $O(\Delta t)$ for $\alpha = 2$, see `MPRK_2_2_alpha.nb` in [36]. This can also be observed in numerical experiments using `BigFloat` in Julia [3] as shown in Figure 3.

4.2. Consistency study for other MP methods

A similar analysis can be conducted for the mPDeC algorithm. However, the approach we used with Mathematica was only able to give results up to third order schemes. We study the multivariate function $\eta(\varepsilon, \Delta t) := u_1^1 - u_1(\Delta t)$, where the initial conditions depend on ε and the time step is Δt, for different limits procedure. Letting Δt go to zero faster than ε and *vice versa*. In practice, we compute a Taylor expansion first in Δt and than in ε and then the opposite in the Mathematica notebooks `mPDeC.nb` and `MPRK_3_2.nb` [36].

Since mPDeC2 is equivalent to MPRK(2,2,1), we get the same results as before. In particular, expanding Δt around 0 first and then ε we obtain

$$\eta(\varepsilon, \Delta t) = \left(\frac{1}{12} - \frac{\varepsilon}{6} + O(\varepsilon^3) \right) \Delta t^3 + O(\Delta t^4),$$

while, doing the opposite, we obtain

$$\eta(\varepsilon, \Delta t) = \left(\frac{\Delta t^3}{48} + O(\Delta t^4) \right) + \left(\frac{\Delta t^2}{8} - \frac{11\Delta t^3}{48} + O(\Delta t^4) \right) \varepsilon + \\ \left(-\frac{\Delta t}{4} + \frac{\Delta t^2}{16} + \frac{3\Delta t^3}{32} + O(\Delta t^4) \right) \varepsilon^2 + O(\varepsilon^3).$$

Note that $O(\varepsilon)$ terms can be ignored when evaluating the order of accuracy. Hence, we see that in both cases we have an error of $O(\Delta t^3)$ for the first step, i.e., a second-order accurate method.

For the third-order algorithm mPDeC3, we have a different behavior and an order reduction for small ε: expanding first Δt and then ε in 0 we obtain

$$\eta(\varepsilon,\Delta t) = \left(-\frac{1}{13824\varepsilon^2} - \frac{5}{1152\varepsilon} + \frac{1789}{13824} - \frac{1697\varepsilon}{6912} + \frac{7\varepsilon^2}{1536} + O(\varepsilon^3)\right)\Delta t^4 \\ + O(\Delta t^5),$$

while, doing the opposite, we obtain

$$\eta(\varepsilon,\Delta t) = \left(-\frac{\Delta t^2}{6} + O(\Delta t^3)\right) + \left(112\Delta t + O(\Delta t^2)\right)\varepsilon - 74880\varepsilon^2 \\ + O(\Delta t\varepsilon^2) + O(\varepsilon^3).$$

If we let $\varepsilon \to 0$ before $\Delta t \to 0$, we have a reduction to first order accuracy. The computations for these tests can be found in `MPDEC.nb` in the accompanying reproducibility repository [36].

For the second-order MPRK(3,2) proposed in this article, we observe a consistent second order accuracy in the limit case $\varepsilon \to 0$, i.e., expanding first Δt and then ε in 0 we obtain

$$\eta(\varepsilon,\Delta t) = \left(\frac{1}{4} - \frac{\varepsilon}{2} + O(\varepsilon^3)\right)\Delta t^3 + O(\Delta t^4),$$

while, doing the opposite, we obtain

$$\eta(\varepsilon,\Delta t) = \left(\frac{\Delta t^3}{6} + O(\Delta t^4)\right) + \left(\frac{\Delta t^2}{6} - \frac{71\Delta t^3}{96} + O(\Delta t^4)\right)\varepsilon + \\ \left(-\frac{\Delta t}{3} + \frac{35\Delta t^2}{96} + \frac{139\Delta t^3}{1152} + O(\Delta t^4)\right)\varepsilon^2 + O(\varepsilon^3).$$

This results are computed Mathematica and the related notebook `MPRK_3_2.nb` is in the accompanying reproducibility repository [36].

Remark 4.5. We have also analyzed MPRKSO$(2,2,\alpha,\beta)$ with selected parameters α, β. We do not present these analyses here; in general, they all agree with the numerical studies presented in the following.

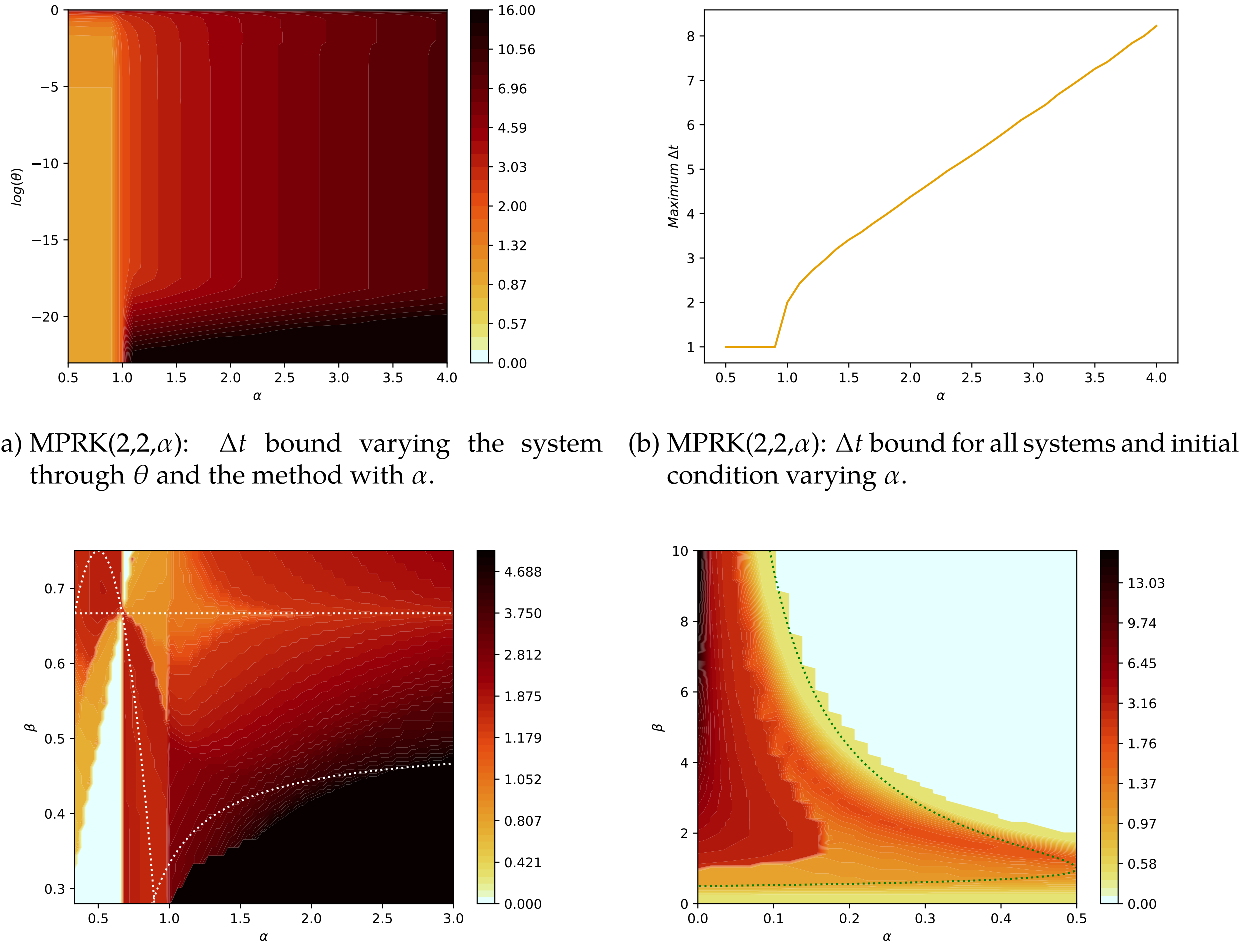
4.3. Need for a numerical study

In the previous studies, we observed two issues with the modified Patankar methods. First, there can be oscillations around the steady state, which can be avoided by a CFL-like restriction on Δt. Moreover, there can be spurious steady states when $u_1^0 \to 0$, leading in practice to a loss of order of accuracy or even an inconsistency.

We want to understand if oscillations happen for all possible initial conditions and all test systems. The ultimate goal is to find conditions on the time step and schemes which do not produce oscillations but are also consistent. Since analytical approaches become infeasible with increasing number of parameters, we will resort to a numerical study in the following.

5. Numerical experiments for simplified linear systems

As described in Section 3, we consider the simplified 2×2 system (10) with initial condition $u^0 = (1-\varepsilon, \varepsilon)^T$. The goal of this study is to find the largest timestep Δt for all possible systems parameterized by $0 \le \theta \le 1$ and initial conditions $0 < \varepsilon < 1$, such that the oscillation-free condition



(a) MPRK(2,2,α): Δt bound varying the system through θ and the method with α.

(b) MPRK(2,2,α): Δt bound for all systems and initial condition varying α.

(c) Δt bound for MPRK(4,3,α, β) varying α and β. The white dashed lines bound the positive RK coefficients area [19].

(d) Δt bound for MPRKSO(2,2,α,β) varying α and β. The green dashed lines bound the positive RK coefficients area [13].

Figure 4: Numerical search of the Δt bound for having an oscillation-free first time step, in the sense of (15), for problem (10) varying IC and system parameter θ: MPRK(2,2,α), MPRK(4,3,α, β) and MPRKSO(2,2,α,β).

(15) is satisfied. We exploit the symmetry of the system studying only the $\varepsilon < 0.5$ case, as the other can be obtain substituting $\tilde{\varepsilon} = 1 - \varepsilon$ and $\tilde{\theta} = 1 - \theta$.

In the following tests, we compare different methods and families presented above: MPRK(2,2,α), MPRK(4,3,α, β), MPRKSO(2,2,α,β), MPRKSO(4,3), mPDeC both for equispaced and Gauss–Lobatto subtimesteps, MPRK(3,2), SI-RK2, and SI-RK3.

We apply all methods to a variety of $\varepsilon \in [0, 0.5]$ and $\theta \in [0, 0.5]$, which are uniformly distributed in a logarithmic scale. For θ, we also consider the symmetrized values for $[0.5, 1]$. We run the simulations for all these schemes and initial conditions for one time step Δt of varying size, uniformly distributed in a logarithmic scale between 2^{-6} and 2^{6}. The maximum Δt that gives no oscillations in the sense of (15) will be denoted as our bound.

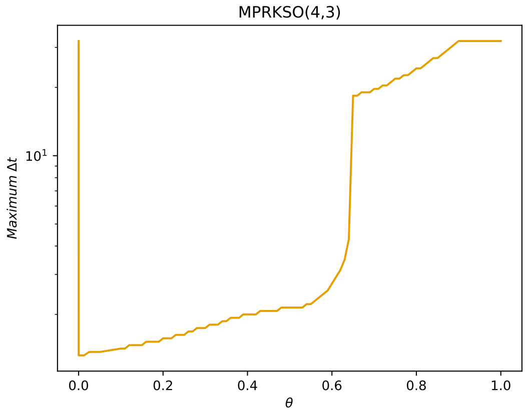
In Figure 4, 5 and 6, we present the results for the all the modified Patankar methods and for the semi-implicit Runge-Kutta methods. We highlight that the evaluation of condition (15) is done with a tolerance of 5 × machine epsilon. Some tests can be sensitive to this tolerance, in particular for (mPDeC) equispaced schemes with high odd order of accuracy, when the Δt bound is large. There the number of stages is large and the machine error can sum up to non-negligible errors.

The second investigation of this section aims at understanding whether the schemes lose conmPDeC

Equispaced	
p	Δt bound
1	∞
2	2.0
3	1.19
4	1.11
5	1.07
6	1.04
7	1.04
8	1.37
9	6.96
10	1.0
11	16.0
12	1.0
13	40.79
14	1.07
15	27.85
16	1.80

Gauss-Lobatto	
p	Δt bound
1	∞
2	2.0
3	1.19
4	1.07
5	1.04
6	1.0
7	1.0
8	1.0
9	1.0
10	1.0
11	1.0
12	1.0
13	1.0
14	1.0
15	1.0
16	1.0

(a) Δt bound for mPDeC of order p with equispaced and Gauss–Lobatto subtimesteps. In red the schemes with spurious steady state.



(b) Δt bound for MPRKSO(4,3) varying the system through θ. Minimum Δt is 1.31.

Figure 5: Numerical search of the Δt bound for having an oscillation-free first time step, in the sense of (15), for problem (10) varying IC and system parameter θ: mPDeC and MPRKSO(4,3)

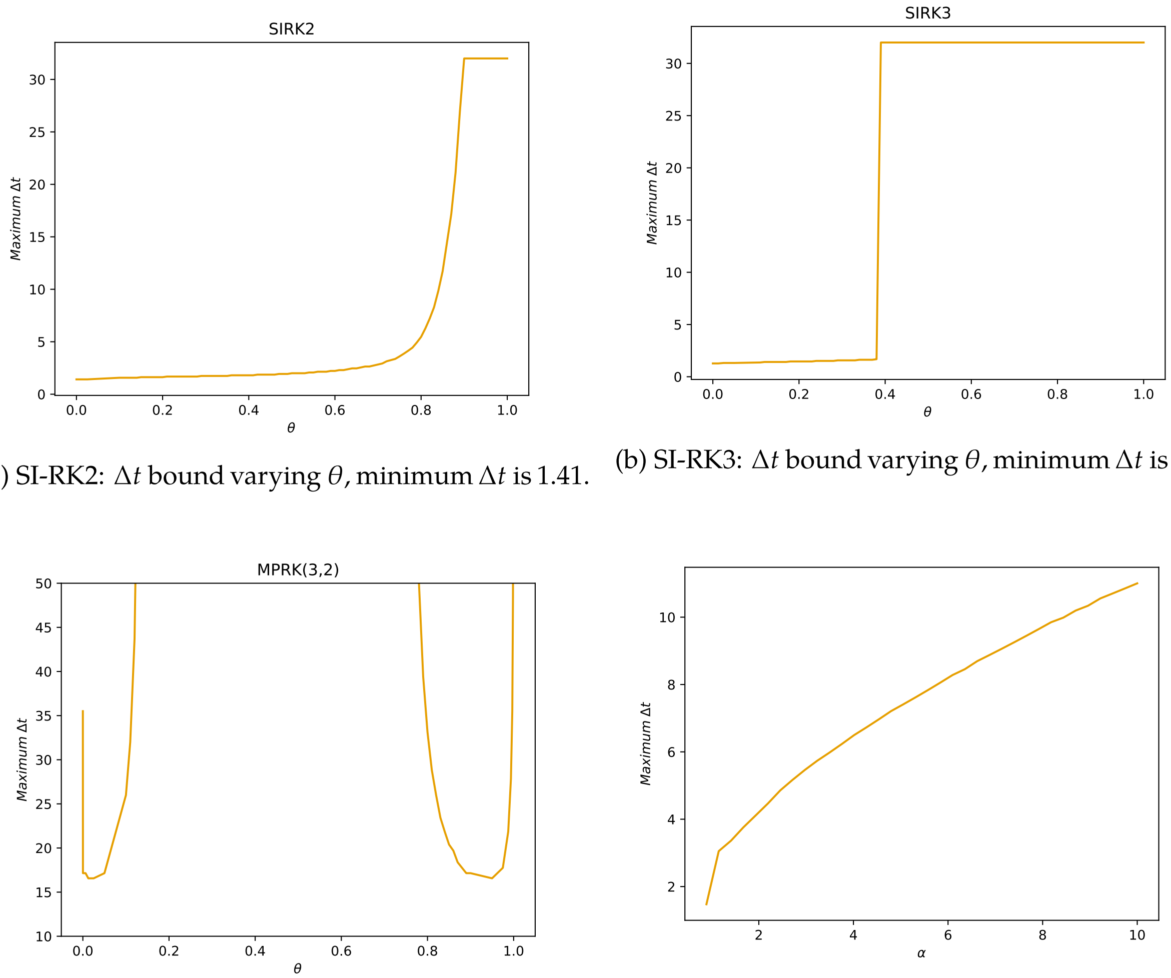
sistency when $\varepsilon \to 0$. For this, we consider the symmetric system (21), and $\varepsilon = 10^{-300}$ and we run the schemes for one large time step $\Delta t = 1$. The exact solution at time 1 is $u_1(1) \approx 0.56$. If the approximation is such that $u_1^1 > 0.999$ we denote the scheme as inconsistent. By numerical experiments, we can say that this definition is robust with respect the system chosen and the tolerance on u_1^1. The interested reader can try different parameters in the repository code [36].

For MPRK(2,2,α), we see in Figures 4a and 4b that the bound on Δt is 1 for $\alpha < 1$, 2 for $\alpha = 1$, and is increasing with $\alpha > 1$. Recall that the methods with $\alpha > 1$ become inconsistent in the limit $\varepsilon \to 0$, preserving the initial condition as spurious steady state. This must be kept in mind when choosing the scheme one wants to use. Varying the system parameter θ influences the bound on the time step, as shown in Figure 4a.

For MPRK(4,3,α,β), we observe areas where the Δt bound reaches very low values ($\ll 1$) and other areas where it is larger than one, independently on the positivity of the RK coefficients. It must be noted that in the areas where the Δt bound is large, we observe inconsistency problems for $\varepsilon \to 0$ as the one shown in Section 4.2. The precise area where this happens is denoted in brown in Figure 7a. It is noticeable that around the curve $\beta(6\alpha - 3) = 3\alpha - 2$, which is a boundary for nonnegative coefficients [19], the Δt bound is particularly large. Hence, in Figure 6d we plot the values for that specific curve, and indeed they are larger than other methods. On the other side, all the schemes given by these parameters show inconsistency when starting close to 0, i.e., $u_1^0 \to 0$.

For MPRKSO(2,2,α,β), we observe that a large area of the α, β plane has Δt bound around unity. The bounds increase close to the line $\alpha = 0$. For this family of methods, we also observe inconsistencies as $\varepsilon \to 0$ for large Δt. The precise area where this happens is denoted in brown in Figure 7b. In the area of negative RK coefficients we observe very low Δt bounds for the oscillation-free condition.

For mPDeC, we observe very different behaviors between equispaced and Gauss–Lobatto points. The two formulations coincide up to third order. The second order mPDeC shows the $\Delta t = 2$ bound that was derived analytically in Section 3. The methods based on Gauss–Lobatto nodes have a time step restriction of unity for orders four and higher. Moreover, we have no evidence of order



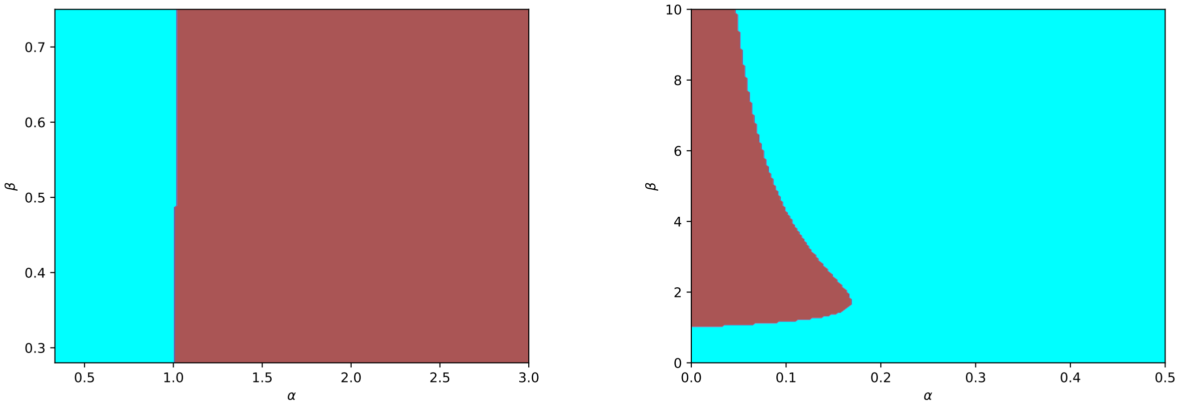
(a) SI-RK2: Δt bound varying θ, minimum Δt is 1.41.

(b) SI-RK3: Δt bound varying θ, minimum Δt is 1.27.

(c) MPRK(3,2): Δt bound varying θ. Minimum Δt is 16.56.

(d) Δt bound varying α for the family MPRK(4,3,α, β) on the curve $\beta(6\alpha - 3) = 3\alpha - 2$ for all the systems through θ of the method.

Figure 6: Numerical search of the Δt bound for having an oscillation-free first time step, in the sense of (15), for problem (10) varying IC and system parameter θ: SI-RK2, SI-RK3, MPRK(3,2) and MPRK(4,3,α, β). For MPRK(4,3,α, β) we consider the minimum Δt over all θ values and we plot the value as a function of α on the curve $\beta(6\alpha - 3) = 3\alpha - 2$.



(a) MPRK(4,3,α, β)

(b) MPRKSO(2,2,α,β)

Figure 7: Inconsistency area for MPRK(4,3,α, β) and MPRKSO(2,2,α,β). Brown for inconsistent, light blue for consistent schemes.

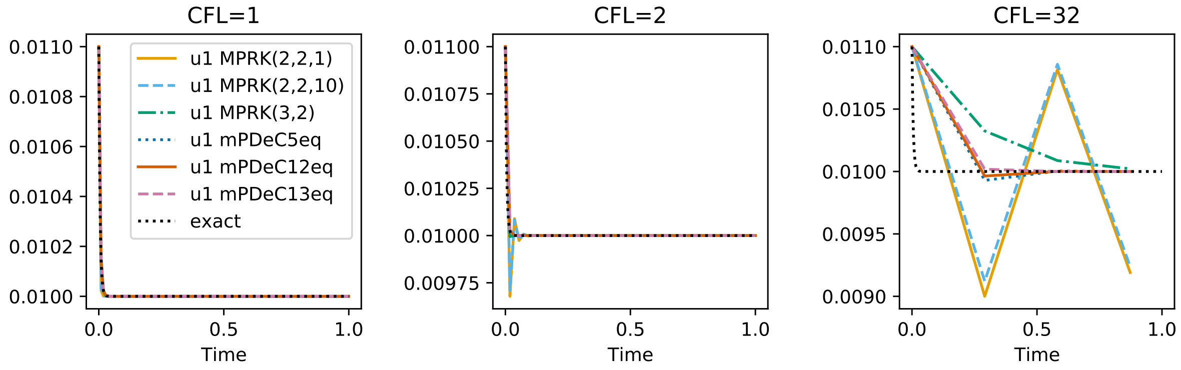


Figure 8: Simulations of (31) at different CFLs for some schemes.

reduction or inconsistency when $\varepsilon \to 0$. For equispaced nodes, we obtain larger Δt bounds, in particular for schemes with odd order of accuracy. In contrast to Gauss–Lobatto nodes, we often observe order reduction/inconsistency problems shown in Section 4.2 for high order schemes, more precisely for order 9 and order greater or equal to 11.

The MPRKSO(4,3) scheme has a Δt bound of 1.31, as shown in Figure 5b. Moreover, it does not show inconsistencies in the numerical tests. Indeed, MPRK(3,2) has maybe the best conditions of all the schemes, see Figure 6c. Its Δt bound is around 16 and it never shows inconsistencies.

Finally, in Figures 6a and 6b, the semi-implicit schemes are presented. Both show similar behaviors with Δt slightly larger than unity. For these methods, we do not observe the same inconsistency shown in previous methods.

6. Validation on nonlinear problems

6.1. Scalar nonlinear problem

The second problem on which we are testing our methods on is a scalar ODE with a source term [6]. Find $u : [0, 0.15] \to \mathbb{R}$, with $u(0) = 1.1\sqrt{1/k}$, where $k > 0$ is a coefficient of the problem, and

$$u' = -k|u|u + 1. \tag{31}$$

The solution for this problem is monotonically decreasing and converging to $u^* = \sqrt{1/k}$. The schemes can be applied to this problem following simple prescriptions.

- The source shall be integrated in time without considering the Patankar trick, simply using the coefficients of the original schemes.
- The productions and destruction terms must be rewritten as $d_{11} = k|u|u$ and $p_{11} = 0$.

We want to extend the linear analysis of the previous two sections, trying to understand if the linear Δt bound can be useful in the nonlinear case as well. Aiming at that, we check the first time step, which often shows overshoots with respect to the steady state, for different time steps.

In particular, we can observe that the (local) Lipschitz constant of the right-hand side of (31) is $C(k) := \max_u k|u| = k|u_0| = 1.1\sqrt{k}$. Hence, inspired by the theory for numerical PDEs, we use a CFL number in $\mathbb{R}^+$ through which we set the Δt step as $\Delta t := \text{CFL}/C(k)$. In this way, we study the bound on Δt setting a condition on the CFL number instead. Doing so, we essentially get rid of the dependence on k, through a rescaling factor both for time and amplitude on the solution. Hence, the CFL number should be comparable with the Δt bound found in the previous sections for arbitrary k. We fix $k = 10^4$ for the following simulations; analogous results can be obtained for other k.

Table 1: Oscillation measure for problem (31) with some selected methods.

CFL	0.5	1.0	2.0	4.0	8.0	16.0	32.0
MPDeC1eq	0	0	2.7e-04	5.3e-04	7.0e-04	8.0e-04	8.5e-04
MPDeC2eq	0	0	3.2e-04	6.1e-04	8.1e-04	9.3e-04	1.0e-03
MPDeC5GL	0	0	0	2.8e-06	6.2e-05	2.0e-04	3.4e-04
MPDeC5eq	0	0	0	0	0	2.0e-05	7.1e-05
MPDeC9eq	0	0	0	0	0	0	0
MPDeC9GL	0	0	0	0	0	0	0
MPDeC11eq	0	0	0	0	0	0	0
MPDeC11GL	0	0	0	0	1.4e-06	1.9e-05	9.0e-05
MPRK(2,2,10.0)	0	0	2.9e-04	5.5e-04	7.2e-04	8.2e-04	8.8e-04
MPRK(4,3,1.25,0.39)	0	0	0	0	0	0	0
MPRKSO(2,2,0.1,1.5)	0	0	3.6e-04	6.7e-04	8.8e-04	1.0e-03	1.1e-03
MPRKSO(4,3)	0	0	5.1e-05	7.7e-05	0	0	0
MPRK(3,2)	0	0	5.2e-06	0	0	0	0
SIRK2	0	0	2.9e-04	5.4e-04	7.0e-04	8.0e-04	8.5e-04
SIRK3	0	0	1.0e-04	3.3e-05	0	0	0

Figure 8 shows the simulations for different CFLs. For low CFLs, we observe no oscillations for essentially all methods. Increasing the CFL number, we observe that most of the schemes go below u^* for the first timestep.

We analyze now all the methods at the first step. We list in Table 1 some representative methods and their oscillations (13) with different CFLs. In the supplementary material [35], we include more tables with many more schemes and parameters, which we summarize in the following.

For mPDeC methods with equispaced and Gauss–Lobatto subtimesteps, we notice that many schemes overshoot the steady values when increasing the CFL. In particular, whenever we are below the Δt bound of Figure 5a, we do not observe oscillations. In some cases, we also do not have oscillations above this bounds, but this might depend on the problem itself. Surprisingly, MPE is not so well performing as in the linear tests, where it was unconditionally not oscillating. For this nonlinear problem, it shows oscillations for CFL > 1.

We also observe oscillations for all methods of the family MPRK(2,2,α) for CFL > 1. This happen even if in the linear tests we had a larger Δt bound for schemes with $\alpha > 1$.

Testing MPRK(4,3,α,β), with some interesting parameters, we found oscillations according to the Δt bound found in Figure 4c almost everywhere, while, on the bottom curve $\beta(6\alpha-3) = 3\alpha-2$, we observe no oscillations for $\alpha \geq 1$, which is slightly better then expected, considering the (large but not so large) Δt bounds in Figure 6d.

Another disappointing result comes from the schemes MPRKSO(2,2,α,β) for which, even on the line $\alpha = 0$, we do not have oscillation-free simulations with large Δt as predicted by Figure 4d on linear problems. Conversely, for the other parameters we have, as expected, oscillations for almost all CFLs larger than 1.

For MPRK(3,2), MPRKSO(4,3), and SI-RK3, the oscillations appear for CFL neither too small nor too large. This is surprising, first of all for MPRK(3,2) of which we expected no oscillations up to CFL ≈ 16, which shows anyway a very small oscillation only for CFL=2, see Figure 8 and Table 1. The amplitude of this oscillation is comparable only with ones produced by very high order schemes. For MPRKSO(4,3) and SI-RK3, we have slightly better results than expected for large CFLs. For SI-RK2, the results are exactly following the Δt bounds found in Figure 6a.

Conclusion 6.1. For this test, most of the schemes behaves as predicted based on the linear example, with few exceptions for second-order methods. The bounds of the linear case can mostly be transferred to the considered nonlinear problem. The linear analysis gives some meaningful results also for more challenging problems.

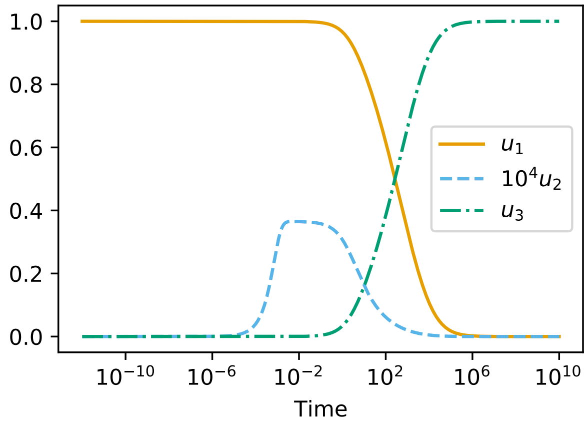


Figure 9: Robertson problem: reference solution obtained with 3000 time steps with mPDeC2.

6.2. Robertson problem

The Robertson problem [22, Section II.10] with parameters $k_1 = 0.04$, $k_2 = 3 \cdot 10^7$, and $k_3 = 10^4$ is a stiff system of three nonlinear ODEs. It can be written as a PDS [17] with non-zero components

$$p_{12}(u)=d_{21}(u)=k_3u_2u_3, \quad p_{21}(u)=d_{12}(u)=k_1u_1, \quad p_{32}(u)=d_{23}(u)=k_2u_2, \tag{32}$$

with initial conditions $u(0) = (1, 0, 0)^T$. Reactions in this problem scale with different orders of magnitudes. To reasonably capture the behavior of the solution, it is necessary to use exponentially increasing time steps [17]. A reference solution can be found in Figure 9. To apply generic modified Patankar schemes, we have to modify the initial condition u^0 slightly, replacing 0 by $\varepsilon > 0$; here, we use $\varepsilon = 10^{-180}$.

For this problem, oscillations are not so clearly defined, because the steady state $u^* = (0, 0, 1)^T$ cannot be exceeded since all the schemes are positive (and the modified Patankar also conservative). Nevertheless, we might encounter the inconsistency problem as one of more constituents approach 0. In Figure 10, we observe that many methods do not catch the behavior of u_2 and remain close to zero. In some cases, even u_3 stays close to zero. All these phenomena are in accordance with the results found for the linear problem. Indeed, among the computed tests we see that MPRK(2,2,α) for $\alpha > 1$, MPRK(4,3,10,0.5), MPRKSO(2,2,0.001,10) and mPDeC11 with equispaced subtimesteps had spurious steady states for $\varepsilon \to 0$ and in this problem, they cannot properly describe the behavior of u_2 (and u_3). Both semi-implicit methods SI-RK2 and SI-RK3 go to infinity as they do not conserve the total sum of the constituents. Hence, we are not showing their simulations.

6.3. HIRES

We consider the "High Irradiance RESponse" problem (HIRES) [11]. The original problem HIRES [22, Section II.1] can be rewritten as a nine-dimensional production–destruction system with

$$\begin{aligned}
&r_1(u) = \sigma, && d_{12}(u) = k_1u_1, && d_{21}(u) = k_2u_2,\\
&d_{24}(u) = k_3u_2, && d_{34}(u) = k_1u_3, && d_{31}(u) = k_6u_3,\\
&d_{43}(u) = k_2u_4, && d_{46}(u) = k_4u_4, && d_{56}(u) = k_1u_5,\\
&d_{53}(u) = k_5u_5, && d_{65}(u) = k_2u_6, && d_{75}(u) = \frac{k_2}{2}u_7,\\
&d_{76}(u) = \frac{k_-}{2}u_7, && d_{79}(u) = \frac{k_*}{2}u_7, && d_{67}(u) = k_+u_6u_8,\\
&d_{87}(u) = k_+u_6u_8, && d_{78}(u) = \frac{k_- + k_* + k_2}{2}u_7,
\end{aligned} \tag{33}$$

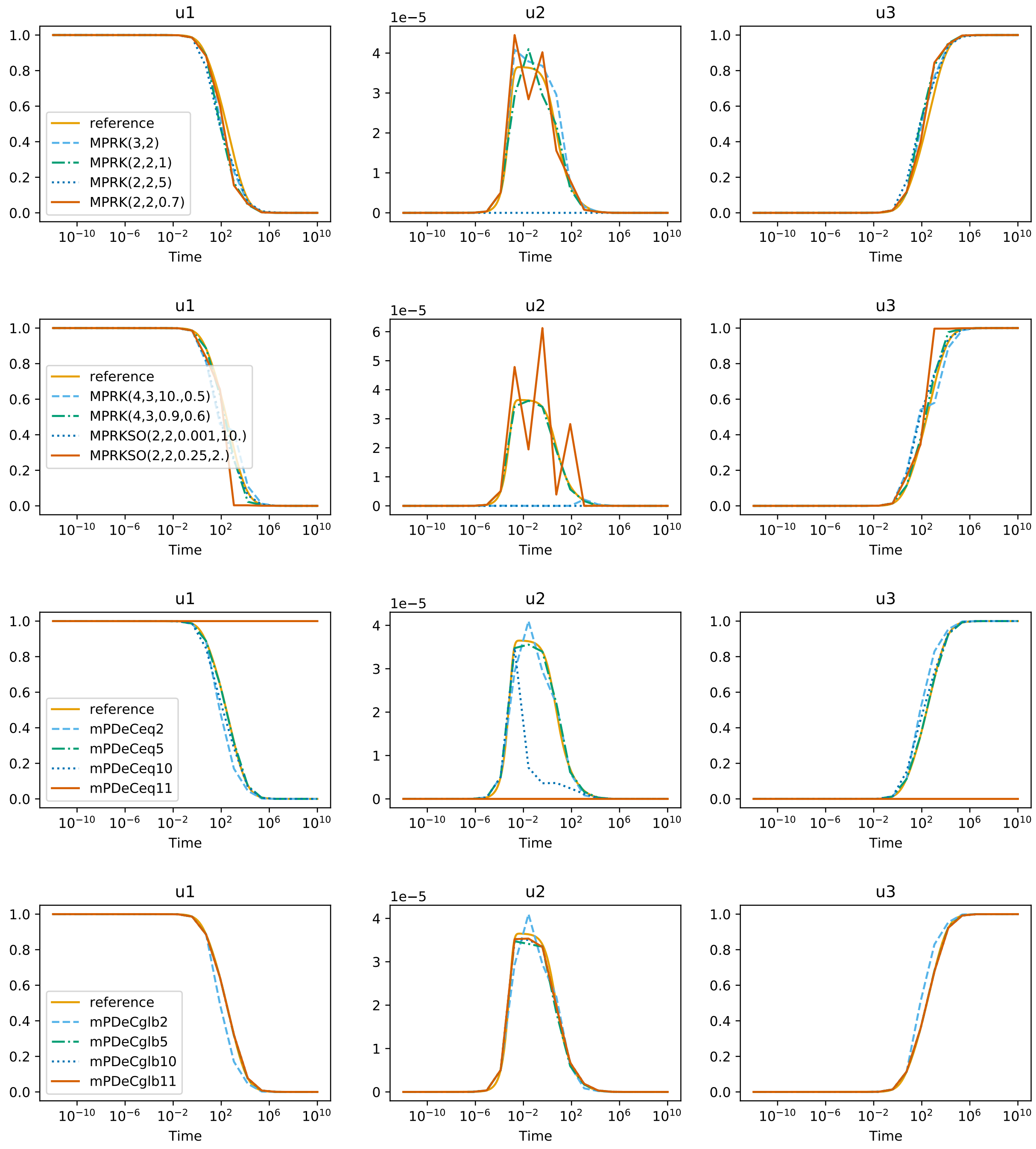


Figure 10: Robertson problem with different methods and 20 time steps.

$p_{ij}(u) = d_{ji} \, \forall \, i,j$ and parameters

$$\begin{aligned} &k_1 = 1.71, \quad k_2 = 0.43, \quad k_3 = 8.32, \quad k_4 = 0.69, \quad k_5 = 0.035, \\ &k_6 = 8.32, \quad k_+ = 280, \quad k_- = 0.69, \quad k_* = 0.69, \quad \sigma = 0.0007. \end{aligned} \tag{34}$$

The initial condition is $u(0) = (1, 0, 0, 0, 0, 0, 0, 0.0057, 0)^T$, where numerically we used 10^{-35} instead of zero for vanishing initial constituents. The time interval is $t \in [0, 321.8122]$.

For this test, the concept of oscillation is not clear as well. Nevertheless, we can observe inconsistencies of some methods also for this problem as some constituents are 0 or almost 0. We compute the reference solution with 10^5 uniform time steps using mPDeC5 with equispaced subtimesteps, which is in accordance with the reference solution [22] up to the fourth significant digit for all constituents.

Testing with $N = 10^3$ uniform time steps, we spot troubles with the *inconsistent* methods found in Section 5. We test the problem with many schemes presented above and we include the relative plots in the supplementary material [35]. For brevity, we plot in Figure 11 just a sample.

For mPDeC, we observe the inconsistency problem only for equispaced time steps for high odd orders ($9, 11, 13$ and so on). In Figure 11, we see the simulation for mPDeC6 with Gauss–Lobatto points. We observe that the high accuracy helps in obtaining a good result at the end of the simulation, when u_7 and u_8 react. The moment at which this change happens is hard to catch and only high order methods are able to obtain it within this number of time steps.

We run the MPRK(2,2,α) with $\alpha \in \{1, 5\}$. As for the linear case, we observe inconsistencies only for $\alpha > 1$. This is demonstrated in Figure 11 for $\alpha = 5$, where the evolution of some constituents is completely missed, e.g., u_2, u_3, u_5, u_9, while for $\alpha = 1$ we obtain consistent results.

We test MPRKSO(2,2,α,β) with $\alpha = 0.3$, $\beta = 2$ and $\alpha = 0$, $\beta = 8$. As expected, the second one shows the inconsistent spurious steady state. An oscillatory behavior can be observed, though, also in the first simulation, which is shown in Figure 11. This is probably due to the CFL condition; refining the time discretization, the oscillations disappear.

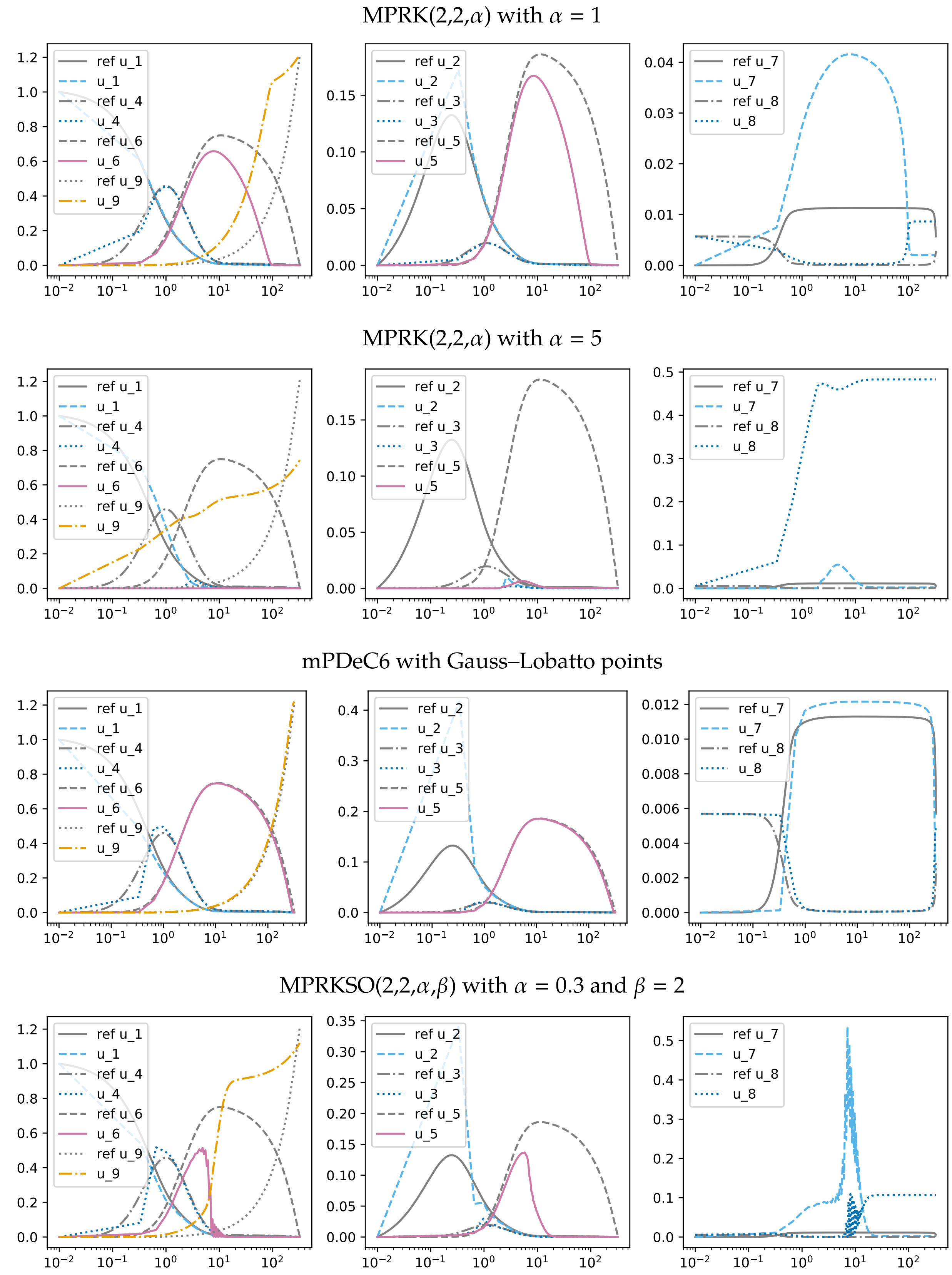
For MPRK(4,3,α, β), we test $\alpha = 0.9$, $\beta = 0.6$ and $\alpha = 5$, $\beta = 0.5$, observing inconsistencies only for the second one, in accordance with the linear tests. For MPRKSO(4,3), MPRK(3,2), SI-RK2 and SI-RK3, we do not observe inconsistencies, as in the linear test, nor other particular behaviors.

7. Summary and discussion

We proposed an analysis for Patankar-type schemes focused on two issues that some of these schemes present: oscillations around the steady state and inconsistency when a constituent is not present at the initial state. Focusing on a generic 2×2 linear test problem, we introduced an oscillation measure. Based thereon, we derived a CFL-like time step restriction avoiding oscillations for all methods under consideration, either analytically (whenever feasible) or numerically. Moreover, we investigated these methods near vanishing components, discovering spurious behavior including order reduction, inconsistency and artificial steady states in the limit of vanishing initial condition. Finally, we applied the methods to more challenging problems including stiff nonlinear ones. We observed that our proposed oscillation-free and consistency analysis generalizes reasonably well to these other problems.

From our point of view, this is a first step toward stability investigations of Patankar-type schemes. Extensions and further studies could be based on various Lyapunov functionals instead of our oscillation measure. Moreover, different test systems could be considered. Nevertheless, we would like to stress that our current approach seems promising and generalizes well to other demanding problems.

In the future, our investigation should be extended to hyperbolic conservation laws. Using the structure of corresponding spatial semidiscretizations, the resulting ODE can be written as a production destruction rest system [7, 14, 23]. Here, the relation between the time step restrictions derived in this work and classical CFL conditions will be the major focus of research.



MPRK(2,2,α) with $\alpha = 1$

MPRK(2,2,α) with $\alpha = 5$

mPDeC6 with Gauss–Lobatto points

MPRKSO(2,2,α,β) with $\alpha = 0.3$ and $\beta = 2$

Figure 11: Simulations run with different schemes with $N = 10^3$ time steps, plot in logarithmic scale in time.

Acknowledgments

D. T. is funded by Team CARDAMOM in Inria–Bordeaux Sud–Ouest, France. P.Ö. gratefully acknowledge support of the Gutenberg Research College, JGU Mainz and the UZH Postdoc Scholarship (Number FK-19-104). H. R. was supported by the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft) under Germany's Excellence Strategy EXC 2044-390685587, Mathematics Münster: Dynamics-Geometry-Structure. We would like to thank Stefan Kopecz and David Ketcheson for fruitful discussion at the beginning of this project. This project has started with the visit by H.R. in Zurich in 2019 which was supported by the SNF project (Number 175784) and the King Abdullah University of Science and Technology (KAUST).

A. Third order modified Patankar Runge–Kutta methods

In the following part, the third order accurate MPRK(4,3,α,β) from [18, 19] is repeated for completeness. Please note that the investigated version is called $MPRK43I(\alpha,\beta)$ in their papers. It is given by

$$\begin{aligned}
y^1 &= u^n,\\
y_i^2 &= u_i^n + a_{21}\Delta t r_i(y^1) + a_{21}\Delta t \sum_j \left(p_{ij}(y^1)\frac{y_j^2}{y_j^1} - d_{ij}(y^1)\frac{y_i^2}{y_i^1}\right),\\
y_i^3 &= u_i^n + \Delta t \left(a_{31} r_i(y^1) + a_{32} r_i(y^2)\right)\\
&\quad + \Delta t \sum_j \Bigg(\left(a_{31}p_{ij}(y^1) + a_{32}p_{ij}(y^2)\right) \frac{y_j^3}{\left(y_j^2\right)^{1/p}\left(y_j^1\right)^{1/p-1}}\\
&\qquad - \left(a_{31}d_{ij}(y^1) + a_{32}d_{ij}(y^2)\right) \frac{y_i^3}{\left(y_i^2\right)^{1/p}\left(y_i^1\right)^{1/p-1}}\Bigg),\\
\sigma_i &= u_i^n + \Delta t \sum_j \Bigg(\left(\beta_1 p_{ij}(y^1) + \beta_2 p_{ij}(y^2)\right) \frac{\sigma_j}{\left(y_j^2\right)^{1/q}\left(y_j^1\right)^{1/q-1}}\\
&\quad - \left(\beta_1 d_{ij}(y^1) + \beta_2 d_{ij}(y^2)\right) \frac{\sigma_i}{\left(y_i^2\right)^{1/q}\left(y_i^1\right)^{1/q-1}}\Bigg)\\
u_i^{n+1} &= u_i^n + \Delta t \left(b_1 r_i(y^1) + b_2 r_i(y^2) + b_3 r_i(y^3)\right)\\
&\quad + \Delta t \sum_j \Bigg(\left(b_1 p_{ij}(y^1) + b_2 p_{ij}(y^2) + b_3 p_{ij}(y^3)\right) \frac{u_j^{n+1}}{\sigma_j}\\
&\qquad - \left(b_1 d_{ij}(y^1) + b_2 d_{ij}(y^2) + b_3 d_{ij}(y^3)\right) \frac{u_i^{n+1}}{\sigma_i}\Bigg),
\end{aligned} \tag{MPRK(4,3,α,β)}$$

where $p = 3a_{21}\left(a_{31} + a_{32}\right) b_3,\ q = a_{21},\ \beta_2 = \frac{1}{2a_{21}}$ and $\beta_1 = 1 - \beta_2$. The Butcher tableaus in respect to the two parameters

$$\begin{array}{c|ccc}
0 & & & \\
\alpha & \alpha & & \\
\beta & \frac{3\alpha\beta(1-\alpha)-\beta^2}{\alpha(2-3\alpha)} & \frac{\beta(\beta-\alpha)}{\alpha(2-3\alpha)} & \\
\hline
 & 1 + \frac{2-3(\alpha+\beta)}{6\alpha\beta} & \frac{3\beta-2}{6\alpha(\beta-\alpha)} & \frac{2-3\alpha}{6\beta(\beta-\alpha)}
\end{array} \tag{35}$$

with positive coefficients for

$$\left.\begin{array}{l}
2/3 \le \beta \le 3\alpha(1-\alpha)\\
3\alpha(1-\alpha) \le \beta \le 2/3\\
(3\alpha-2)/(6\alpha-3) \le \beta \le 2/3
\end{array}\right\} \text{ for } \begin{cases}
1/2 \le \alpha < \frac{2}{3},\\
2/3 \le \alpha < \alpha_0,\\
\alpha > \alpha_0,
\end{cases}$$

and $\alpha_0 \approx 0.89255$. When the coefficients are negative we swap the weights of production and destruction terms as for (mPDeC).

Next, also the MPRKSO(4,3) from [14] is repeated. It is given by

$$
\begin{aligned}
y^1 &= u^n,\\
y_i^2 &= y_i^1 + a_{10}\Delta t r_i(y^1) + \Delta t \sum_j b_{10}\left(p_{ij}(y^1)\frac{y_j^2}{y_j^1} - d_{ij}(y^1)\frac{y_i^2}{y_i^1}\right),\\
\varrho_i &= n_1 y_i^2 + n_2 y_i^1 \left(\frac{y_i^2}{y_i^1}\right)^2\\
y_i^3 &= (a_{20}y_i^1 + a_{21}y_i^2) + \Delta t\left(b_{20}r_i(y^1) + b_{21}r_i(y^2)\right)\\
&\quad + \Delta t \sum_j \left(\left(b_{20}p_{ij}(y^1) + b_{21}p_{ij}(y^2)\right)\frac{y_j^2}{\varrho_j} - \left(b_{20}d_{ij}(y^1) + b_{21}d_{ij}(y^2)\right)\frac{y_i^2}{\varrho_i}\right),\\
\mu_i &= y_i^1 \left(\frac{y_i^2}{y_i^1}\right)^s \qquad\qquad\qquad\qquad \text{(MPRKSO(4,3))}\\
\tilde{a}_i &= \eta_1 y_i^1 + \eta_2 y_i^2 + \Delta t \sum_j \left(\left(\eta_3 p_{ij}(y^1) + \eta_4 p_{ij}(y^2)\right)\frac{\tilde{a}_j}{\mu_j} - \left(\eta_3 d_{ij}(y^1) + \eta_4 d_{ij}(y^2)\right)\frac{\tilde{a}_i}{\mu_i}\right)\\
\sigma_i &= \tilde{a}_i + z y_i^1 \frac{y_i^2}{\varrho_i}\\
u_i^{n+1} &= \left(a_{30}y_i^1 + a_{31}y_i^2 + a_{32}y_i^3\right) + \Delta t\left(b_{30}r_i(y^1) + b_{31}r_i(y^2) + b_{32}r_i(y^3)\right)\\
&\quad + \Delta t \sum_j \Bigg(\left(b_{30}p_{ij}(y^1) + b_{31}p_{ij}(y^2) + b_{32}p_{ij}(y^2)\right)\frac{u_j^{n+1}}{\sigma_j}\\
&\qquad\qquad - \left(b_{30}d_{ij}(y^1) + b_{31}d_{ij}(y^2) + b_{32}d_{ij}(y^2)\right)\frac{u_i^{n+1}}{\sigma_i}\Bigg).
\end{aligned}
$$

Here, the optimal SSP coefficients determined in [14] will be used. They are given by

$$
\begin{aligned}
n_1 &= 2.5690460257320 11E-01, & n_2 &= 7.430953974267989E-01,\\
a_{10} &= 1, & a_{20} &= 9.2600312554031827E-01,\\
a_{21} &= 7.3996874459681783E-02, & a_{31} &= 2.0662904223744017E-10,\\
b_{10} &= 4.7620819268131703E-01, & a_{30} &= 7.0439040373427619E-01,\\
a_{32} &= 2.9560959605909481E-01, & b_{20} &= 7.7545442722396801E-02,\\
b_{21} &= 5.9197500149679749E-01, & b_{31} &= 6.8214380786704851E-10,\\
b_{30} &= 2.0044747790361456E-01, & b_{32} &= 5.9121918658514827E-01,\\
\eta_1 &= 3.777285888379173E-02, & \eta_2 &= 1/3,\\
\eta_3 &= 1.868649805549811E-01, & \eta_3 &= 2.224876040351123,\\
z &= 6.288938077828750E-01, & s &= 5.721964308755304.
\end{aligned}
$$

B. Initial direction of Patankar schemes

In this section we investigate the direction of the first step of a method. In particular, if we know that the direction of the first step is always towards the steady state, for any initial condition, we know that oscillations are possible only around the steady state. We will first present some theoretical results for very few schemes, then we summarize some numerical results we obtained varying ε and θ.

Let us define the property that we will check along this section.

Definition B.1. A numerical method *has the correct direction* for the linear PDS (10) if $u_1^0 > (1-\theta)$ implies that $u_1^1 < u_1^0$ and $u_1^0 < (1-\theta)$ implies that $u_1^1 > u_1^0$.

For symmetry we will check only the first condition on the whole range of $0 < \varepsilon \leq \theta < 1$.

Theorem B.2 (Direction of MPE)**.** *MPE has the correct direction of the first time step, i.e., if the initial condition is above the steady state, then the first step will be below the initial condition, or, in other words,*

$$u_1^0 > (1-\theta) \Longrightarrow u_1^0 > u_1^1. \tag{36}$$

Proof. We write the MPE for the system (10) in the first equation, making use of the conservation property and we collect all the implicit terms.

$$\begin{aligned}
u_1^1 &= u_1^0 + \alpha\Delta t \left((1-\theta)(1-u_1^0)\frac{1-u_1^1}{1-u_1^0} - \theta u_1^0 \frac{u_1^1}{u_1^0} \right), && \text{(37a)}\\
u_1^1 &= u_1^0 + \Delta t \left((1-\theta)(1-u_1^1) - \theta u_1^1 \right), && \text{(37b)}\\
u_1^1(1+\Delta t) &= y_1^1 + \Delta t(1-\theta), && \text{(37c)}\\
u_1^1 &= \frac{u_1^0 + \alpha\Delta t(1-\theta)}{(1+\Delta t)} < \frac{u_i^0(1+\Delta t)}{(1+\Delta t)} = u_1^0. && \text{(37d)}
\end{aligned}$$

Here, we have simply used the hypothesis on $u_1^0 > (1-\theta)$ and we obtain the thesis of the theorem. □

Theorem B.3 (Direction of MPRK(2,2,α) with $\alpha \leq 1$)**.** *MPRK(2,2,α) for $\alpha \leq 1$ applied on the simplified system* (10) *has the correct direction of the first time step.*

Proof. The first stage consists in a first MPE step with time step $\alpha\Delta t$. So we obtain that $y_1^2 < y_1^1 = u_1^0$. For the second stage we can proceed analogously, exploiting the conservation property, the system (10), collecting all the implicit terms and using the hypothesis $u_1^0 > (1-\theta)$.

$$\begin{aligned}
u_1^{n+1} = u_1^n + \Delta t \Bigg(&\left(\frac{2\alpha-1}{2\alpha}(1-\theta)(1-y_1^1) + \frac{1}{2\alpha}(1-\theta)(1-y_1^2) \right) \frac{1-u_1^{n+1}}{(1-y_1^2)^{1/\alpha}(1-y_1^1)^{1-1/\alpha}} \\
&- \left(\frac{2\alpha-1}{2\alpha}\theta y_1^1 + \frac{1}{2\alpha}\theta y_1^2 \right) \frac{u_1^{n+1}}{(y_1^2)^{1/\alpha}(y_1^1)^{1-1/\alpha}} \Bigg),
\end{aligned} \tag{38a}$$

$$\begin{aligned}
u_1^{n+1} = u_1^n + \Delta t \Bigg(&\left(\frac{2\alpha-1}{2\alpha}(1-\theta)\left(\frac{1-y_1^1}{1-y_1^2}\right)^{1/\alpha} + \frac{1}{2\alpha}(1-\theta)\left(\frac{1-y_1^2}{1-y_1^1}\right)^{1-1/\alpha} \right)(1-u_1^{n+1}) \\
&- \left(\frac{2\alpha-1}{2\alpha}\theta \left(\frac{y_1^1}{y_1^2}\right)^{1/\alpha} + \frac{1}{2\alpha}\theta\left(\frac{y_1^2}{y_1^1}\right)^{1-1/\alpha} \right) u_1^{n+1} \Bigg),
\end{aligned} \tag{38b}$$

$$\begin{aligned}
&\left(1+\Delta t \left(\frac{2\alpha-1}{2\alpha}(1-\theta)\left(\frac{1-y_1^1}{1-y_1^2}\right)^{1/\alpha} + \frac{1}{2\alpha}(1-\theta)\left(\frac{1-y_1^2}{1-y_1^1}\right)^{1-1/\alpha} \right) + \right.\\
&\left. \Delta t \left(\frac{2\alpha-1}{2\alpha}\theta\left(\frac{y_1^1}{y_1^2}\right)^{1/\alpha} + \frac{1}{2\alpha}\theta\left(\frac{y_1^2}{y_1^1}\right)^{1-1/\alpha} \right) \right) u_1^{n+1} = \\
&u_1^n + \Delta t \left(\left(\frac{2\alpha-1}{2\alpha}(1-\theta)\left(\frac{1-y_1^1}{1-y_1^2}\right)^{1/\alpha} + \frac{1}{2\alpha}(1-\theta)\left(\frac{1-y_1^2}{1-y_1^1}\right)^{1-1/\alpha} \right) \right) < \\
&u_1^n \left(1 + \Delta t \left(\left(\frac{2\alpha-1}{2\alpha}\left(\frac{1-y_1^1}{1-y_1^2}\right)^{1/\alpha} + \frac{1}{2\alpha}\left(\frac{1-y_1^2}{1-y_1^1}\right)^{1-1/\alpha} \right) \right) \right).
\end{aligned} \tag{38c}$$

So we have that

$$u_1^{n+1} < u_1^n \frac{N}{D} \tag{38d}$$

with $N > 0$ and $D > 0$ deducible from (38c). If $N < D$ we have our result, or, in other words, if $N - D < 0$. So, let us compute

$$\begin{aligned}\frac{N-D}{\Delta t} =& \frac{2\alpha-1}{2\alpha}\left(\frac{1-y_1^1}{1-y_1^2}\right)^{1/\alpha} + \frac{1}{2\alpha}\left(\frac{1-y_1^2}{1-y_1^1}\right)^{1-1/\alpha} - \\ & \frac{2\alpha-1}{2\alpha}(1-\theta)\left(\frac{1-y_1^1}{1-y_1^2}\right)^{1/\alpha} - \frac{1}{2\alpha}(1-\theta)\left(\frac{1-y_1^2}{1-y_1^1}\right)^{1-1/\alpha} - \\ & \frac{2\alpha-1}{2\alpha}\theta\left(\frac{y_1^1}{y_1^2}\right)^{1/\alpha} - \frac{1}{2\alpha}\theta\left(\frac{y_1^2}{y_1^1}\right)^{1-1/\alpha},\end{aligned} \tag{38e}$$

$$\begin{aligned}\frac{N-D}{\Delta t} =& \frac{2\alpha-1}{2\alpha}\theta\left(\frac{1-y_1^1}{1-y_1^2}\right)^{1/\alpha} + \frac{1}{2\alpha}\theta\left(\frac{1-y_1^2}{1-y_1^1}\right)^{1-1/\alpha} - \\ & \frac{2\alpha-1}{2\alpha}\theta\left(\frac{y_1^1}{y_1^2}\right)^{1/\alpha} - \frac{1}{2\alpha}\theta\left(\frac{y_1^2}{y_1^1}\right)^{1-1/\alpha} = \\ & \frac{2\alpha-1}{2\alpha}\theta\left(\left(\frac{1-y_1^1}{1-y_1^2}\right)^{1/\alpha} - \left(\frac{y_1^1}{y_1^2}\right)^{1/\alpha}\right) + \frac{1}{2\alpha}\theta\left(\left(\frac{1-y_1^2}{1-y_1^1}\right)^{1-1/\alpha} - \left(\frac{y_1^2}{y_1^1}\right)^{1-1/\alpha}\right).\end{aligned} \tag{38f}$$

Now, we know that $y_1^1 > y_1^2$, hence

$$\frac{y_1^1}{y_1^2} > 1 > \frac{1-y_1^1}{1-y_1^2},$$

so, considering $0 < \alpha \leq 1$, we have that $1/\alpha > 0$ and $1 - 1/\alpha \leq 0$, we have

$$\left(\left(\frac{1-y_1^1}{1-y_1^2}\right)^{1/\alpha} - \left(\frac{y_1^1}{y_1^2}\right)^{1/\alpha}\right) < 0 \text{ and } \left(\left(\frac{1-y_1^2}{1-y_1^1}\right)^{1-1/\alpha} - \left(\frac{y_1^2}{y_1^1}\right)^{1-1/\alpha}\right) < 0.$$

Hence, $\frac{N-D}{\Delta t} < 0$ and the proof is complete. □

For the case with $\alpha > 1$ it is not so easy to derive an estimation as the two terms have opposite signs.

Theorem B.4 (Direction of MPRKSO(2,2,α,β) with $\gamma \geq 1$)**.** *MPRKSO(2,2,α,β) applied on the simplified system* (10) *for positive RK coefficients and for*

$$\gamma = \frac{1-\alpha\beta+\alpha\beta^2}{\beta(1-\alpha\beta)} \geq 1$$

has the correct direction of the first time step.

Proof. The proof follows the same step of proof of Theorem B.3. The condition on the exponent of the weights here is precisely $\gamma \geq 1$. □

Remark B.5 (Consistency area)**.** We want to remark that the area in the (α, β) plane where $\gamma \geq 1$ and the RK coefficients are positive is defined by

$$\alpha \leq \frac{\beta-1}{2\beta^2-\beta} \text{ with } \beta \geq 1,$$

and this area coincide with the consistency area of MPRKSO(2,2,α,β) found in Figure 7b.

B.1. Other schemes

For all other schemes it is not so easy to prove directly that the direction of the first step is the correct one. Nevertheless, we checked symbolically (when feasible) and numerically (otherwise) this property. The numerical computations are included in `CheckingDirection.ipynb` in the repository [36], while the only theoretical result is in `MPRK_3_2.nb`. We summarize in the following the results we obtained.

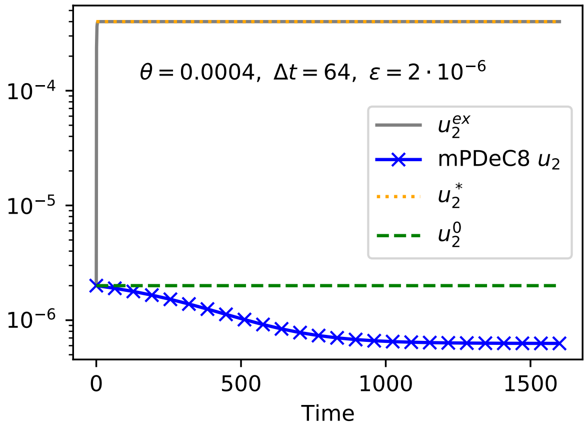


Figure 12: Simulation of (10) with $\theta = 4 \cdot 10^{-4}$ and $u_2^0 = \varepsilon = 2 \cdot 10^{-6}$ with mPDeC8 with equispaced points for $\Delta t = 64$

- MPRK(3,2) has the correct direction and we proved it in the Mathematica notebook `MPRK_3_2.nb`;
- MPRK(2,2,α) have the correct direction for all $1/2 \leq \alpha \leq 4$;
- MPRKSO(2,2,α,β) have the correct direction in an area slightly larger than the positive RK weights area displayed in Figure 4d, which coincide with the strictly positive Δt bound area there;
- MPRK(4,3,α, β) have the correct direction except in a small area around $\alpha = 2/3$ where the RK coefficients are negative;
- MPRKSO(4,3) has the correct direction;
- mPDeC
  - with Gauss–Lobatto points have the correct direction up to order 16;
  - with equispaced points have the correct direction up to order 7, for order 8, 9 and 15 we found wrong directions for large $\Delta t(\geq 30)$ and very small initial conditions and θ, all other mPDeC with orders up to 16 have the correct direction;
- SI-RK2 and SI-RK3 have the correct direction.

In Figure 12, we show an example for mPDeC8 where the correct direction is not followed. We see that even if we go away from the steady state, the scheme does not oscillate.

References

[1] O. Axelsson. *Iterative Solution Methods*. Cambridge: Cambridge University Press, 1996. DOI: 10.1017/CBO9780511624100.

[2] A. Bellen and L. Torelli. “Unconditional Contractivity in the Maximum Norm of Diagonally Split Runge–Kutta Methods.” In: *SIAM Journal on Numerical Analysis* 34.2 (1997), pp. 528–543. DOI: 10.1137/S0036142994267576.

[3] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. “Julia: A Fresh Approach to Numerical Computing.” In: *SIAM Review* 59.1 (2017), pp. 65–98. DOI: 10.1137/141000671. arXiv: 1411.1607 [cs.MS].

[4] C. Bolley and M. Crouzeix. “Conservation de la positivité lors de la discrétisation des problèmes d’évolution paraboliques.” In: *RAIRO. Analyse numérique* 12.3 (1978), pp. 237–245.

[5] H. Burchard, E. Deleersnijder, and A. Meister. “A high-order conservative Patankar-type discretisation for stiff systems of production–destruction equations.” In: *Applied Numerical Mathematics* 47.1 (2003), pp. 1–30. DOI: 10.1016/S0168-9274(03)00101-6.

[6] A. Chertock, S. Cui, A. Kurganov, and T. Wu. "Steady state and sign preserving semi-implicit Runge–Kutta methods for ODEs with stiff damping term." In: *SIAM Journal on Numerical Analysis* 53.4 (2015), pp. 2008–2029. DOI: 10.1137/151005798.

[7] M. Ciallella, L. Micalizzi, P. Öffner, and D. Torlo. *An Arbitrary High Order and Positivity Preserving Method for the Shallow Water Equations*. arXiv preprint: https://arxiv.org/abs/2108.07347. 2021. arXiv: 2110.13509 [math.NA].

[8] I. Fekete, D. I. Ketcheson, and L. Lóczi. "Positivity for convective semi-discretizations." In: *Journal of Scientific Computing* 74.1 (2018), pp. 244–266. DOI: 10.1007/s10915-017-0432-9.

[9] P. Frolkovic. "Semi-implicit methods based on inflow implicit and outflow explicit time discretization of advection." In: *Proceedings of ALGORITMY*. 2016, pp. 165–174.

[10] S. Gottlieb, D. I. Ketcheson, and C.-W. Shu. *Strong stability preserving Runge–Kutta and multistep time discretizations*. Singapore: World Scientific, 2011.

[11] E. Hairer and G. Wanner. "Stiff differential equations solved by Radau methods." In: *Journal of Computational and Applied Mathematics* 111.1-2 (1999), pp. 93–111. DOI: 10.1016/S0377-0427(99)00134-X.

[12] Z. Horváth. "Positivity of Runge–Kutta and diagonally split Runge–Kutta methods." In: *Applied Numerical Mathematics* 28.2-4 (1998), pp. 309–326. DOI: 10.1016/S0168-9274(98)00050-6.

[13] J. Huang and C.-W. Shu. "Positivity-Preserving Time Discretizations for Production–Destruction Equations with Applications to Non-equilibrium Flows." In: *Journal of Scientific Computing* 78.3 (2019), pp. 1811–1839. DOI: 10.1007/s10915-018-0852-1.

[14] J. Huang, W. Zhao, and C.-W. Shu. "A Third-Order Unconditionally Positivity-Preserving Scheme for Production–Destruction Equations with Applications to Non-equilibrium Flows." In: *Journal of Scientific Computing* 79.2 (2019), pp. 1015–1056. DOI: 10.1007/s10915-018-0881-9.

[15] K. J. in' t Hout. "A note on unconditional maximum norm contractivity of diagonally split Runge–Kutta methods." In: *SIAM Journal on Numerical Analysis* 33.3 (1996), pp. 1125–1134. DOI: 10.1137/0733055.

[16] S. Kopecz and A. Meister. "A comparison of numerical methods for conservative and positive advection–diffusion–production–destruction systems." In: *PAMM* 19.1 (2019). DOI: 10.1002/pamm.201900209.

[17] S. Kopecz and A. Meister. "On order conditions for modified Patankar–Runge–Kutta schemes." In: *Applied Numerical Mathematics* 123 (2018), pp. 159–179. DOI: 10.1016/j.apnum.2017.09.004.

[18] S. Kopecz and A. Meister. "On the existence of three-stage third-order modified Patankar–Runge–Kutta schemes." In: *Numerical Algorithms* (2019), pp. 1–12. DOI: 10.1007/s11075-019-00680-3.

[19] S. Kopecz and A. Meister. "Unconditionally positive and conservative third order modified Patankar–Runge–Kutta discretizations of production–destruction systems." In: *BIT Numerical Mathematics* 58.3 (2018), pp. 691–728. DOI: 10.1007/s10543-018-0705-1.

[20] D. Kuzmin. "Entropy stabilization and property-preserving limiters for $\mathbb{P}^1$ discontinuous Galerkin discretizations of scalar hyperbolic problems." In: *Journal of Numerical Mathematics* (2020).

[21] C. B. Macdonald, S. Gottlieb, and S. J. Ruuth. "A numerical study of diagonally split Runge–Kutta methods for PDEs with discontinuities." In: *Journal of Scientific Computing* 36.1 (2008), pp. 89–112. DOI: 10.1007/s10915-007-9180-6.

[22] F. Mazzia and C. Magherini. *Test Set for Initial Value Problem Solvers*. Technical Report Release 2.4. Italy: Department of Mathematics, University of Bari, Feb. 2008.

[23] A. Meister and S. Ortleb. "A positivity preserving and well-balanced DG scheme using finite volume subcells in almost dry regions." In: *Applied Mathematics and Computation* 272 (2016), pp. 259–273.

[24] K. Mikula and M. Ohlberger. "Inflow-implicit/outflow-explicit scheme for solving advection equations." In: *Finite Volumes for Complex Applications VI Problems & Perspectives*. Vol. 4. Springer Proceedings in Mathematics. Berlin, Heidelberg: Springer, 2011, pp. 683–691. DOI: 10.1007/978-3-642-20671-9_72.

[25] K. Mikula, M. Ohlberger, and J. Urbán. "Inflow-implicit/outflow-explicit finite volume methods for solving advection equations." In: *Applied Numerical Mathematics* 85 (2014), pp. 16–37. DOI: 10.1016/j.apnum.2014.06.002.

[26] S. Nüßlein, H. Ranocha, and D. I. Ketcheson. "Positivity-Preserving Adaptive Runge-Kutta Methods." In: *Communications in Applied Mathematics and Computational Science* 16.2 (Nov. 2021), pp. 155–179. DOI: 10.2140/camcos.2021.16.155. arXiv: 2005.06268 [math.NA].

[27] P. Öffner and D. Torlo. "Arbitrary high-order, conservative and positivity preserving Patankar-type deferred correction schemes." In: *Applied Numerical Mathematics* 153 (2020), pp. 15–34.

[28] S. V. Patankar. *Numerical Heat Transfer and Fluid Flow*. Washington: Hemisphere Publishing Corporation, 1980.

[29] C. Rackauckas and Q. Nie. "DifferentialEquations.jl – A Performant and Feature-Rich Ecosystem for Solving Differential Equations in Julia." In: *Journal of Open Research Software* 5.1 (2017), p. 15. DOI: 10.5334/jors.151.

[30] H. Ranocha. "On strong stability of explicit Runge–Kutta methods for nonlinear semibounded operators." In: *IMA Journal of Numerical Analysis* 41.1 (2021), pp. 654–682.

[31] H. Ranocha and D. I. Ketcheson. "Energy Stability of Explicit Runge–Kutta Methods for Nonautonomous or Nonlinear Problems." In: *SIAM Journal on Numerical Analysis* 58.6 (2020), pp. 3382–3405.

[32] H. Ranocha and P. Öffner. "L_2 Stability of Explicit Runge–Kutta Schemes." In: *Journal of Scientific Computing* 75.2 (May 2018), pp. 1040–1056. DOI: 10.1007/s10915-017-0595-4.

[33] Z. Sun and C.-W. Shu. "Stability of the fourth order Runge–Kutta method for time-dependent partial differential equations." In: *Annals of Mathematical Sciences and Applications* 2.2 (2017), pp. 255–284. DOI: 10.4310/AMSA.2017.v2.n2.a3.

[34] Z. Sun and C.-W. Shu. "Strong Stability of Explicit Runge–Kutta Time Discretizations." In: *SIAM Journal on Numerical Analysis* 57.3 (2019), pp. 1158–1182. DOI: 10.1137/18M122892X. arXiv: 1811.10680 [math.NA].

[35] D. Torlo, P. Öffner, and H. Ranocha. *A New Stability Approach for Positivity-Preserving Patankar-type Schemes*. arXiv preprint: https://arxiv.org/abs/2108.07347. 2021. arXiv: 2108.07347 [math.NA].

[36] D. Torlo, P. Öffner, and H. Ranocha. *Stability of Positivity Preserving Patankar-Type Schemes*. Git repository: https://git.math.uzh.ch/abgrall_group/patankar-stability. Aug. 2021.

[37] Wolfram Research, Inc. *Mathematica*. Version 12.0. 2019. URL: https://www.wolfram.com.