Issues with Positivity-Preserving Patankar-type Schemes

Davide Torlo*, Philipp Öffner†, Hendrik Ranocha‡

March 20, 2022

Patankar-type schemes are linearly implicit time integration methods designed to be unconditionally positivity-preserving. However, there is no generally accepted stability or robustness theory for these schemes. We suggest two approaches to analyze the performance and robustness of these methods. In particular, we demonstrate problematic behaviors of these methods that can lead to undesired oscillations and order reduction even on very simple linear problems. Finally, we demonstrate in numerical simulations that our theoretical results for linear problems apply analogously to nonlinear stiff problems.

Keywords. 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) [6]

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

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 $d(u) = (d_{ij}(u))_{i,j}$.

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 $\forall i, j \in I, \forall u \colon p_{ij}(u) = d_{ji}(u)$.

*davide.torlo@sissa.it, SISSA mathLab, Mathematics Area, SISSA, via Bonomea 265, Trieste, Italy.

†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.

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.

The existence, uniqueness and positivity of the solution of a PDS can be proven under the following assumptions [10].

Theorem 1.2. *The PDS with initial conditions $u^0 \geq 0$ has a unique solution $u \in [C^1(\mathbb{R}^+)]^{|I|}$ and $u_i(t) > 0$ if $u_i^0 > 0$, if*

1. *for all $i, j \in I$ d_{ij} is locally Lipschitz continuous in $\mathbb{R}^{|I|}$,*
2. *$d_{ij}(u) = 0$ for all $i, j \in I$ if $u = 0$,*
3. *$d_{ij}(u) = \tilde{d}_{ij}(u)u_i$ with $\tilde{d}_{ij} \in C((\overline{\mathbb{R}^+})^{|I|})$ and $\tilde{d}_{ij}(u) > 0$ if $u > 0$ and $\tilde{d}_{ij}(u) = 0$ if $u = 0$.*

In [6] the previous assumptions 2 and 3 are replaced by the condition $d_{ij}(u) \to 0$ if $u_i \to 0$. It can be easily shown that this condition plus the Lipschitz continuity of the destruction terms lead to similar structures. Let C be the maximum of the Lipschitz continuity constants of the destruction terms and consider $u = v$ except for the i-th component for which $v_i = 0$. We have that

$$d_{ij}(u) = |d_{ij}(u) - d_{ij}(v)| \leq C||u - v||_2 = Cu_i. \tag{4}$$

Hence, we can define

$$\tilde{d}_{ij}(u) := \frac{d_{ij}(u)}{u_i} \leq C. \tag{5}$$

This condition is less restrictive and it does not guarantee the continuity of $\tilde{d}_{ij}$ in $u_i = 0$. For the rest of the paper, we will consider assumptions of theorem 1.2. Also, all the physically/chemically/biologically relevant cases, of which we are aware, fall in this definition.

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

Definition 1.3. 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 [9], e.g., based on the concept of strong stability preserving (SSP) [12] or adaptive Runge–Kutta (RK) methods [32]. However, general linear methods are restricted to conditional positivity if they are at least second order accurate [5]. One way to circumvent such order restrictions is given by diagonally split RK methods, which can be unconditionally positive [3, 14, 17]. However, they are less accurate than the unconditionally positive implicit Euler method for large step sizes in practice [26].

Another approach to unconditionally positivity-preserving methods is based on the so-called Patankar trick [34, Section 7.2-2]. First- and second-order accurate conservative methods based thereon were introduced in [6]. Later, these were extended to families of second- and third-order accurate modified Patankar–Runge–Kutta (MPRK) methods based on the Butcher coefficients [22, 24] and the Shu–Osher form [15, 16]. Related deferred correction (DeC) methods were proposed recently [33]. Positive but not conservative methods using the Patankar trick have been proposed and studied in [7], although the connection to Patankar methods seems to be unknown up to now. Other related numerical schemes are inflow-implicit/outflow-explicit methods [11, 30, 31]. Ideas from Patankar-type methods have also been used in numerical methods based on limiters [25].

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 [7, 21, 22, 24]. Recently, Patankar methods have been investigated using Lyapunov stability theory [18–20]. We will point out the relation between their approach and our investigations. Lately, BBKS and GeCo, two geometric integrators, have been introduce to simulate biochemistry models preserving not only positivity and conservation, but also all linear invariants of a system [27].

1.1. Motivating example

Consider the normal linear system

$$u'(t) = 10^2 \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} u(t), \quad u(0) = u_0 = \begin{pmatrix} 0.1 \\ 0 \end{pmatrix}, \tag{6}$$

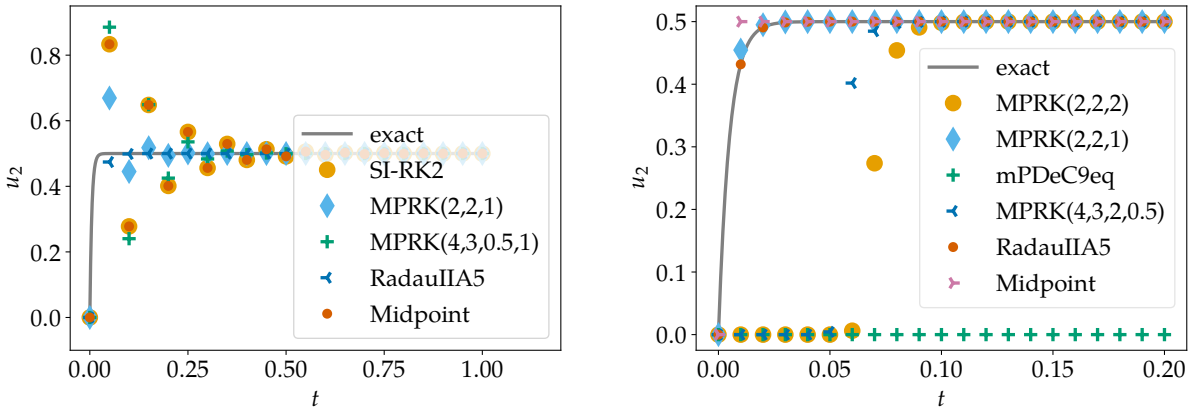
which can be written as a production destruction system with

$$p(u) = \begin{pmatrix} 0 & 10^2 u_2 \\ 10^2 u_1 & 0 \end{pmatrix}, \quad d(u) = \begin{pmatrix} 0 & 10^2 u_1 \\ 10^2 u_2 & 0 \end{pmatrix}. \tag{7}$$

On (6), we can demonstrate all the problematic behaviors mentioned above. We solve (6) with several different methods. In details, we apply the second order method SI-RK2 of [7], the second- and third-order accurate modified Patankar–Runge–Kutta schemes MPRK(2,2,α) and MPRK(4,3,α,β) from [22, 24] with different parameter selections, the implicit Midpoint rule and fifth-order, three stage RadauIIA5 scheme [13] implemented in DifferentialEquations.jl [36] in Julia [4]. The solutions are shown in Figure 1a. It can be recognized that even for this simple test case, most of the methods are oscillating for the selected time step but with different amplitudes while RadauIIA5 results in an oscillating–free approximation.
Another problem rises if we use other Patankar schemes. These methods are constructed for strictly positive PDS, therefore we have to substitute the zero initial condition with something very small, i.e., $u_2(0) = 10^{-250}$. We observe in Figure 1b that some of the methods replicate the initial condition for some time steps while others do not leave it at all in the considered time interval. On the other side 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 accuracy in the limit for an initial condition approaching zero.
In our investigation, we want to find the methods that have those undesirable behaviors and avoid them.

1.2. Scope of the article

There is no accepted stability theory for positive and conservative schemes up to this point. Motivated by our numerical examples above we are interested in concepts that would detect the dominant appearance of spurious oscillations and the loss of accuracy in the limit process. 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. [37–41]. However, the obtained results have not been sufficient for us to describe the properties of the schemes in an adequate way. Thus, we will directly measure the amount of spurious oscillations using a generic linear system as a test problem, and focus as well as on the loss of accuracy in the limit process. Our investigation leads to a deeper understanding of the basic properties of Patankar-type methods.
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 on the loss of the order of accuracy in the limit of vanishing initial condition in Section 4. At the end, we give also a remark to the recently started investigations from [18–20]. In Section 5, a numerical study on linear systems



(a) $\Delta t = 0.05$ (b) $\Delta t = 0.01$

Figure 1: Numerical solutions of the normal linear system (6) with real and non-positive eigenvalues obtained using different Patankar-type schemes as well as two implicit Runge–Kutta methods (only second component depicted) with initial condition $u_0 = (1, 10^{-250})^T$.

derives the results on all schemes that theoretical studies of the previous sections could not found: bounds on time step for oscillation–free schemes. 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 on 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 [34, Section 7.2-2] for a PDR system (3) to get the positive Patankar–Euler method

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

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 (8) 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 [6] (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, since the right-hand side is positive, the solution u^{n+1} is positive [2, 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 $u_i \to 0$ [6]. 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 [22]. 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 [6]. 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) [12].

A similar two-parameter family MPRK(4,3,α,β) of four stage, third-order accurate schemes was introduced and studied in [23, 24]. The family under consideration can be found in the 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 [15]. 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

$$
\gamma = \frac{1 - \alpha\beta + \alpha\beta^2}{\beta(1 - \alpha\beta)}, \tag{9}
$$

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 [16] and can be found in the A.

2.4. Modified Patankar deferred correction schemes

Arbitrarily high-order conservative and positive modified Patankar deferred correction schemes (mPDeC) were introduced in [33]. 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. At each subtimestep $t^{n,m}$, an approximation y^m is calculated. In the formulation of [1] 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

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

where θ_r^m are the correction weights and the $\gamma(j, i, \theta_r^m)$ takes value j if $\theta_r^m > 0$ and i otherwise, see [33] for details. This allows 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 $u_i^{n+1} = y_i^{M,(K)}$.

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

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

For explicitly time-dependent problems, the abscissae are the ones of SSPRK(3,3) [12], 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 [23] 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 [7] are also based on the Shu–Osher representation of SSPRK 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 [7] are

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

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

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

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

$$y_i^2 = u_i^n + \Delta t r_i(y^1) + \Delta t \sum_j \left(p_{ij}(y^1) - d_{ij}(y^1) \frac{y_i^2}{y_i^1} \right), \tag{10}$$

which is the Patankar–Euler method (8). As for the Patankar–Euler method, the semi-implicit methods of [7] 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 [7], steady state preservation for (modified) Patankar–Runge–Kutta methods will be studied here. Except for the SI-RK2 and SI-RK3 methods [7], 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. Indeed, the Patankar weights reduce to 1 and the simple rest-production-destruction forms remains and their sum is 0 in the steady state. 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 [7]. 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.

In our discussion, it will be useful to check not only the preservation of the steady state, but also how this state is approached, for example, if in a monotone manner or not.

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

Here, we introduce a new approach to study the behavior of Patankar-type schemes. Thus, instead of Dahlquist's equation, which is not a PDS, we propose to use a 2×2 linear system similar to (6) as test problem. This is the simplest PDS that can be considered. More precisely, we consider the general 2×2 production-destruction linear system as also done later in similar form in [18]

$$\begin{pmatrix} u_1' \\ u_2' \end{pmatrix} = \begin{pmatrix} -a & b \\ a & -b \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}. \tag{11}$$

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.,

$$\begin{pmatrix} u_1' \\ u_2' \end{pmatrix} = \begin{pmatrix} -\theta & (1-\theta) \\ \theta & -(1-\theta) \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}. \tag{12}$$

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

$$\begin{pmatrix} u_1^0 \\ u_2^0 \end{pmatrix} = \begin{pmatrix} 1-\varepsilon \\ \varepsilon \end{pmatrix}, \tag{13}$$

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

$$\begin{pmatrix} u_1(t) \\ u_2(t) \end{pmatrix} = \begin{pmatrix} (1-\theta) + (\theta - \varepsilon)e^{-t} \\ \theta + (\varepsilon - \theta)e^{-t} \end{pmatrix}, \tag{14}$$

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 as the one presented in Figure 1a. This reduces to finding schemes that for every u^n and every system θ has a monotone behavior and 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 between the previous state and the steady state $u_1^0 > u_1^1 > u_1^*$ and $u_2^0 < u_2^1 < u_2^*$, as RadauIIA5 is doing in Figure 1a. Thus, we consider the *oscillation measure*

$$\mathsf{osc}(u_1^0, u_1^1, u_1^*) := \begin{cases} \max\left\{(u_1^1 - u_1^0)^+, (u_1^* - u_1^1)^+\right\} & \text{if } u_1^0 > u_1^*, \\ \max\left\{(u_1^0 - u_1^1)^+, (u_1^1 - u_1^*)^+\right\} & \text{if } u_1^0 < u_1^*, \end{cases} \tag{15}$$

to detect whenever a scheme is not behaving in this way. Here, $(\cdot)^+$ denotes the positive part of a real number. This oscillation measure vanishes for monotone 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 $0 < \varepsilon, \theta < 1$, we can use this measure to find oscillation–free schemes. Hence, the measure (15) helps us in obtaining a very simple criterion on oscillation-free solutions studying just one time step.

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

$$\mathsf{osc}(u_1^0, u_1^1, u_1^*) = 0 \tag{16}$$

for every initial condition (IC) $0 < \varepsilon < 1$ and for every system $0 \le \theta \le 1$. We can simplify the search exploiting the symmetry of the system, for example considering only $0 < \varepsilon \le 0.5$. It suffices to check the initial step, since every other step will fall back in another IVP (12) with a different IC.

Remark 3.1 (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 conservation property. 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 one side of the *oscillation-free* condition (u_1^1 not over/undershooting the steady state) 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 (12) is feasible.

Theorem 3.2 (Time restriction for mPDeC2 for 2×2 linear systems)**.** *Consider the system* (12) *with the initial conditions* (13)*. mPDeC2 is oscillation-free in the sense of* (16) *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* (11) *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 three in Δt, and degree two in θ and ε. Here, for brevity we write one of the two component $u_2^1 = \frac{N}{D}$, where

$$\begin{aligned} N =& 2(1-\varepsilon)\varepsilon^2 + 2\Delta t\,\varepsilon(\varepsilon(1-\theta) + 2(1-\varepsilon)\theta) \\ &+ \Delta t^2\left((1-\varepsilon)\varepsilon\theta + 3\varepsilon(1-\theta)\theta + 2(1-\varepsilon)\theta^2\right) + \Delta t^3\left((1-\varepsilon)\theta^2 + (1-\theta)\theta^2\right) > 0, \\ D =& 2(1-\varepsilon)\varepsilon + \Delta t(2(1-\varepsilon)\varepsilon + 2\varepsilon(1-\theta) + 2(1-\varepsilon)\theta) \\ &+ \Delta t^2((1-\theta)(2\varepsilon+\theta) + (1-\varepsilon)(\varepsilon+2\theta)) + \Delta t^3(\varepsilon(1-\theta) + (1-\varepsilon)\theta) > 0. \end{aligned}$$

The condition (16) simplifies to $\varepsilon \geq u_2^1 \geq \theta$ in the case $\varepsilon > \theta$ and to $\varepsilon \leq u_2^1 \leq \theta$ if $\varepsilon < \theta$. The inequality regarding u_2^1 and ε is proven in B in Theorem B.3 for all MPRK(2,2,α) schemes with $\alpha \leq 1$. To prove the inequality regarding θ and u_2^1 we analyze the sign of $D\theta - N$, where N and D are the numerator and the denominator of u_2^1 respectively, clearly both positive. For $\varepsilon > \theta$ we want to have $D\theta - N < 0$ not to overshoot the steady state, while for $\varepsilon < \theta$ we should have $D\theta - N > 0$ or, in other words, $\frac{D\theta - N}{\varepsilon - \theta} < 0$. We have that

$$\frac{D\theta - N}{\varepsilon - \theta} = -2\varepsilon(1-\varepsilon) - \Delta t 2(\theta(1-\varepsilon) + \varepsilon(1-\theta)) - \Delta t^2\theta(1-\theta) + \Delta t^3\theta(1-\theta) < 0, \tag{17}$$

which is a third degree polynomial inequality for Δt and can be rewritten as

$$p_{\varepsilon,\theta}(\Delta t) = \Delta t^3 - \Delta t^2 - 2\left(\frac{\varepsilon}{\theta} + \frac{1-\varepsilon}{1-\theta}\right)\Delta t - 2\frac{\varepsilon(1-\varepsilon)}{\theta(1-\theta)} < 0. \tag{18}$$

There are two options for real coefficients cubic polynomials. If the discriminant $\Delta \geq 0$ then the roots are all real, while if $\Delta < 0$ there are two complex conjugated roots and a real one [35]. Only if $\Delta = 0$ the roots are multiple. Let us consider first the case $\Delta \geq 0$. Denoting with $y \leq w \leq z$ the three real roots of $p_{\varepsilon,\theta}(x)$, we see that they have to satisfy

$$\begin{cases} y + w + z = 1, \\ yz + wz + yw = -2\left(\frac{\varepsilon}{\theta} + \frac{1-\varepsilon}{1-\theta}\right) < -2, \\ ywz = 2\frac{\varepsilon(1-\varepsilon)}{\theta(1-\theta)} > 0. \end{cases} \tag{19}$$

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 (19), we see that

$$z(w+y) < z(w+y) + wy = yz + wz + yw < -2, \tag{20}$$

$$w + y < -\frac{2}{z}. \tag{21}$$

Using then the first equation of (19), we have that

$$0 = z + y + w - 1 < z - \frac{2}{z} - 1, \quad 0 < z^2 - z - 2, \tag{22}$$

which has positive solutions only for $z > 2$. Hence, $\Delta t \leq 2$ in order to avoid oscillations for all systems (12). 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 $\varepsilon \to 0$, the first and third equations in (19) tell us that $w \to 0^-$. Hence, from the second equation we can see that $y \to -2\frac{1}{(1-\theta)z}$. Finally, the third zero will converge to

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

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

If $\Delta < 0$ then there are one real root z and two complex conjugated roots $y = a + ib$, $\bar{y} = a - ib$ [35]. These roots must verify

$$\begin{cases} 2a + z = 1, \\ 2az + a^2 + b^2 = -2\left(\frac{\varepsilon}{\theta} + \frac{1-\varepsilon}{1-\theta}\right) < -2, \\ (a^2+b^2)z = 2\frac{\varepsilon(1-\varepsilon)}{\theta(1-\theta)} > 0. \end{cases} \tag{23}$$

Since $(a^2 + b^2)z$ is positive, z is positive. From the second equation of (23), we see that

$$2az < 2az + a^2 + b^2 < -2, \tag{24}$$

$$a < -\frac{1}{z}. \tag{25}$$

Using then the first equation of (23), we have that

$$0 = z + 2a - 1 < z - \frac{2}{z} - 1, \quad 0 < z^2 - z - 2, \tag{26}$$

which has positive solutions only for $z > 2$. □

Remark 3.3. The discriminant of $p_{\varepsilon,\theta}$

$$\begin{aligned} \Delta =& 4\theta(1-\theta)\big[\varepsilon^2(1-\theta)^3\theta + (1-\varepsilon)^2(1-\theta)\theta^3 + 8\varepsilon^3(1-\theta)^3 + 8(1-\varepsilon)^3\theta^3 \\ &+ 6(1-\varepsilon)\varepsilon^2(1-\theta)^2\theta + 6(1-\varepsilon)^2\varepsilon(1-\theta)\theta^2 - 27(1-\varepsilon)^2\varepsilon^2(1-\theta)\theta\big] \end{aligned} \tag{27}$$

is positive in the square $0 < \varepsilon, \theta < 1$. This has been verified in `MPRK_2_2_1_generalSystem.nb` in [42]. Hence, the case $\Delta < 0$ never happens for $0 < \varepsilon, \theta < 1$.

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. Loss of the order of accuracy for vanishing initial conditions

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

In this section, we show for which Patankar and modified Patankar schemes there is an order reduction for a very simple linear problem. Some numerical experiments validate this study in Sections 5 and 6.

4.1. Strong loss of order accuracy for vanishing initial conditions

Different modified Patankar schemes behave differently for vanishing initial condition, some are not affected, some become second order accurate, some first order accurate. We tested different modified Patankar schemes on (12) with $\theta = 0.5$ comparing $\varepsilon = 0.01$ and $\varepsilon = 10^{-250}$. We can see in figure 2 that some MPRK(2,2,α) and MPRK(4,3,α, β) fall into first order accuracy, while other third and fourth order schemes become second order ones in this situation.
The fall back to first and second order is due to an error in the first timesteps when one initial condition is close to 0. As soon as this component becomes large enough the error disappears. This leaves either a shift of some Δt on the solution or a first time step with a second order error. To grasp why we lose order of accuracy, we need to understand what happens in the limit of our schemes for $\varepsilon \to 0$ for the first time step. We remark that, in the linear system case, $\tilde{p}_{ij}$ and $\tilde{d}_{ij}$ defined in theorem 1.2 are positive and constant. As an example, we can see the role of these production/destruction rates in the MPE

$$u_i^1 = u_i^n + \Delta t \sum_j \left(\tilde{p}_{ij}(u^0) \cancel{u_j^0} \frac{u_j^1}{\cancel{u_j^0}} - \tilde{d}_{ij}(u^0) \cancel{u_i^0} \frac{u_i^1}{\cancel{u_i^0}} \right), \tag{28}$$

$$u_i^1 = \frac{u_i^0 + \Delta t \sum_j \tilde{p}_{ij}(u^0) u_j^1}{1 + \Delta t \sum_j \tilde{d}_{ij}(u^0)} = u_i^0 + \Delta t \sum_{j \in I} \left(\tilde{p}_{ij}(u^0) u_j^1 - \tilde{d}_{ij}(u^0) u_i^0 \right) + O(\Delta t^2). \tag{29}$$

Hence, we see that the method that we obtain for vanishing initial condition $\varepsilon \to 0$, is well defined and, in this case, it leads to a consistent and first order scheme.
This is not true for MPRK(2,2,α) for all α. The first stage of the scheme is a MPE step and it does not introduce issues. The second stage depends on the coefficient α. Let us define $\omega = \frac{1}{2\alpha}$, the second stage reads

$$u_i^1 = u_i^0 + \Delta t \sum_j \left[\left(\frac{(1-\omega) p_{ij}(y^1) + \omega p_{ij}(y^2)}{(y_j^2)^{1/\alpha} (y_j^1)^{1-1/\alpha}} \right) u_j^1 - \left(\frac{(1-\omega) d_{ij}(y^1) + \omega d_{ij}(y^2)}{(y_i^2)^{1/\alpha} (y_i^1)^{1-1/\alpha}} \right) u_i^1 \right]. \tag{30}$$

Here, we cannot simplify as before the linear terms of destructions and productions. If we focus on the destruction term for the vanishing constituent, i.e., $u_i^n = y_i^1 \to 0$, and if we suppose that the first step is such that $y_i^2 \geq C_2 \Delta t$, this is true for example for linear problems with nonzero production, we have that

$$\lim_{y_i^1 \to 0} \frac{(1-\omega) d_{ij}(y^1) + \omega d_{ij}(y^2)}{(y_i^2)^{1/\alpha} (y_i^1)^{1-1/\alpha}} = \begin{cases} 0, & \text{if } 1 - 1/\alpha < 0 \Leftrightarrow \alpha < 1, \\ \omega \tilde{d}_{ij}(y^2), & \text{if } 1 - 1/\alpha = 0 \Leftrightarrow \alpha = 1, \\ \infty, & \text{if } 1 - 1/\alpha > 0 \Leftrightarrow \alpha > 1. \end{cases} \tag{31}$$

Let us make it more formal with Landau symbols in the simplified linear system (12) case.

Theorem 4.1 (Accuracy of the first time step for MPRK(2,2,α))**.** *The scheme MPRK(2,2,α) on problem* (12) *for vanishing initial condition, i.e., $\varepsilon \ll \Delta t$, performs a first time step with an error of $O(\Delta t^3)$ if $\alpha = 1$, with an $O(\Delta t^2)$ if $\frac{1}{2} \leq \alpha < 1$ and with an $O(\Delta t)$ for $\alpha > 1$. More precisely, for $\alpha > 1$,*

$$u_2^1 = O\left(\left(\frac{\varepsilon}{\Delta t} \right)^{1-1/\alpha} \right).$$

Proof. Let us apply the scheme for one time step to the initial condition $u^0 = y^1 = (1-\varepsilon, \varepsilon)$ for the system (12) with $0 < \theta < 1$ and with $\varepsilon \ll \Delta t$. We obtain

$$y_2^2 = \frac{\varepsilon + \alpha \Delta t \theta}{1 + \alpha \Delta t}, \qquad y_1^2 = \frac{1 - \varepsilon + \alpha \Delta t (1-\theta)}{1 + \alpha \Delta t}, \tag{32}$$

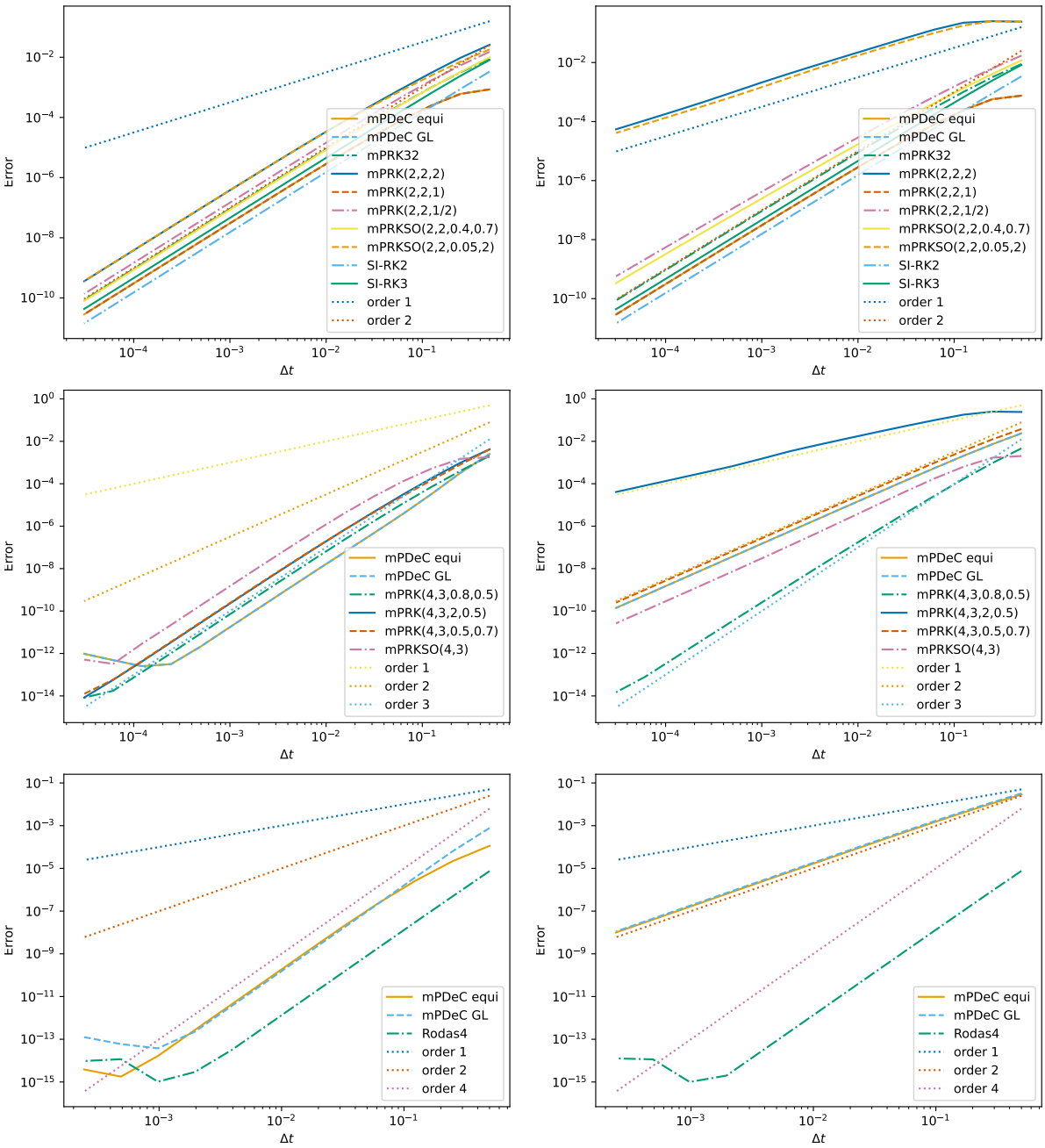


Figure 2: Error decay for the system (12) with $\theta = 0.5$, $T = 1$ and $\varepsilon = 10^{-2}$ on the left and $\varepsilon = 10^{-250}$ on the right. Various scheme and different order of accuracy from top to bottom

$$\left[1 + \Delta t\theta A + \Delta t(1-\theta)B\right] u_2^1 = \varepsilon + \Delta t\theta A, \tag{33}$$

$$A := \frac{2\alpha - 1}{2\alpha}\left(\frac{y_1^1}{y_1^2}\right)^{\frac{1}{\alpha}} + \frac{1}{2\alpha}\left(\frac{y_1^2}{y_1^1}\right)^{1-\frac{1}{\alpha}}, \tag{34}$$

$$B := \frac{2\alpha - 1}{2\alpha}\left(\frac{y_2^1}{y_2^2}\right)^{\frac{1}{\alpha}} + \frac{1}{2\alpha}\left(\frac{y_2^2}{y_2^1}\right)^{1-\frac{1}{\alpha}}. \tag{35}$$

We remark that

$$y_2^2 = \theta\alpha\Delta t + O(\Delta t^2) + O(\varepsilon).$$

If $\alpha > 1$, then $1 - 1/\alpha > 0$, and we have that

$$A = 1 + O(\Delta t) + O(\varepsilon), \tag{36}$$

$$B = \frac{2\alpha - 1}{2\alpha}\left(\frac{\varepsilon}{\theta\alpha\Delta t}\right)^{\frac{1}{\alpha}} + \frac{1}{2\alpha}\left(\frac{\theta\alpha\Delta t}{\varepsilon}\right)^{1-\frac{1}{\alpha}} + O\left(\left(\frac{\Delta t^2}{\varepsilon}\right)^{1-\frac{1}{\alpha}}\right) = O\left(\left(\frac{\Delta t}{\varepsilon}\right)^{1-\frac{1}{\alpha}}\right). \tag{37}$$

Hence, the solution at the first time step is

$$u_2^1 = O\left(\left(\frac{\varepsilon}{\Delta t}\right)^{1-\frac{1}{\alpha}}\right) = u_2(\Delta t) + O(\Delta t). \tag{38}$$

In the regime $\varepsilon \ll \Delta t$ this gives an error of $O(\Delta t)$ at the first time step. This means that several timesteps are needed before the regime $\varepsilon \ll \Delta t$ is left. Experimentally, the number of these timesteps was observed to be independent of Δt, hence, after a time interval of length proportional to Δt, classical accuracy is restored, with a global error of $O(\Delta t)$.
The case $\frac{1}{2} \le \alpha < 1$ leads to negative $-1 \le 1 - \frac{1}{\alpha} < 0$. Hence, $0 < \frac{1}{\alpha} - 1 \le 1$ is positive and

$$A = 1 - \left(1 - \frac{\theta}{2}\right)\Delta t + O(\Delta t^2) + O(\varepsilon), \tag{39}$$

$$B = O\left(\left(\frac{\varepsilon}{\Delta t}\right)^{\frac{1}{\alpha}-1}\right). \tag{40}$$

This means that the non-zero-th order terms of A dominates on B in the left hand side of (33), hence,

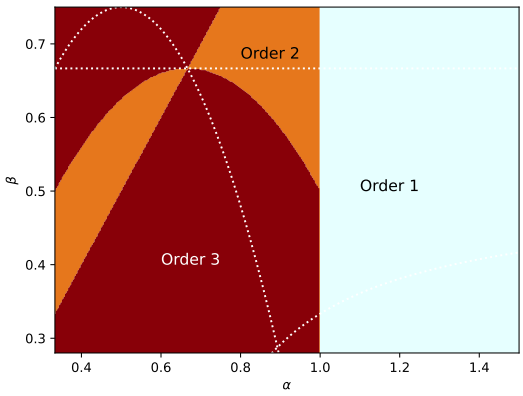
$$u_2^1 = \frac{\theta\Delta t A}{1 + \theta\Delta t A} + O(\Delta t^{\frac{1}{\alpha}}\varepsilon^{\frac{1}{\alpha}-1}) = \theta\Delta t + \left(\theta - \frac{3\theta^2}{2}\right)\Delta t^2 + O(\Delta t^3) + O(\Delta t^{\frac{1}{\alpha}}\varepsilon^{\frac{1}{\alpha}-1}) \tag{41}$$

$$= u_2(\Delta t) + O(\Delta t^2) + O(\Delta t^{\frac{1}{\alpha}}\varepsilon^{\frac{1}{\alpha}-1}). \tag{42}$$

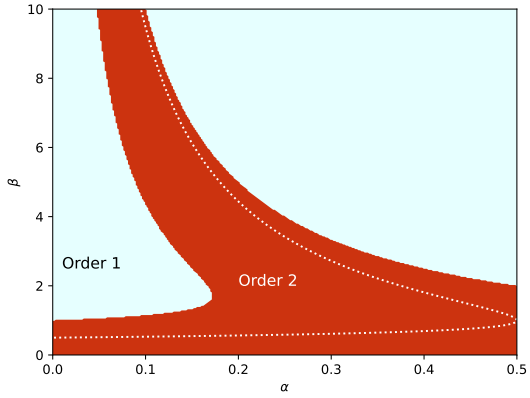
This means that the first step has an error of $O(\Delta t^2)$ in the limit of $\varepsilon \to 0$, while starting from the second time step we leave the regime $\varepsilon \ll \Delta t$ and we restore the original accuracy. Since the method is anyway second order, the error of the first step does not affect the global order of accuracy.
Finally, in case $\alpha = 1$, the exponent $1 - 1/\alpha = 0$ simplifies some terms, and we obtain

$$A = 1 - \frac{\theta\Delta t}{2} + O(\Delta t^3) + O(\varepsilon), \tag{43}$$

$$B = \frac{1}{2} + O(\varepsilon). \tag{44}$$



(a) MPRK(4,3,α, β) orders: light blue first order, orange second order, brown third order



(b) MPRKSO(2,2,α,β) orders: light blue first order, red second order

Figure 3: Order of accuracy of some schemes for vanishing initial conditions. The white dashed lines bound the positive RK coefficients area [15, 24].

Expanding in Taylor series the solution u_2^1 we obtain

$$u_2^1 = \theta\left(\Delta t - \frac{1}{2}\Delta t^2\right) + O(\Delta t^3) + O(\varepsilon) = u_2(\Delta t) + O(\Delta t^3) + O(\varepsilon). \tag{45}$$

This means that the error of the first time step is of order 3, and contributes to an overall second order method. □

Also for the other modified Patankar schemes we can derive similar estimations. We summarize the results and we sketch the proofs for all these schemes in the following theorem.

Theorem 4.2 (Accuracy of MP schemes for vanishing data)**.** *Consider the system of ODEs* (12) *with* $u_0 = (1-\varepsilon, \varepsilon)$ *with vanishing initial condition, i.e.,* $\varepsilon \ll \Delta t$*. Then,*

- *MPRK(4,3,*α*,* β*) is first order accurate if* $\alpha > 1$*, otherwise it is second order when* $p > 1$ *and it is third order for all other parameters, see figure 3a;*
- *MPRKSO(2,2,*α*,*β*) is first order accurate when* $q > 1$ *or* $\gamma < 1$*, otherwise it is second order accurate, see figure 3b;*
- *MPRKSO(4,3) is second order accurate;*
- *mPDeC are second order accurate, except some very high order methods with equispaced subtimesteps with some negative* θ_r^M *that are first order accurate;*
- *MPRK(3,2), SI-RK2 and SI-RK3 are second order accurate.*

Sketch of the proof. In the following, we explain which are the terms that introduce a loss of order of accuracy.

- MPRK(4,3,α, β)
  As we saw in figure 2, there is no loss of order of accuracy for MPRK(4,3,α, β) for $q = \alpha < 1$ and $p = \frac{\alpha(2-3\alpha)}{2(\beta-\alpha)} < 1$. $q = \alpha < 1$ affects the equation of σ_i, as for MPRK(2,2,α) with $\alpha < 1$ by an error of $O(\Delta t^2)$. In the weighting procedure in the final stage, this introduced an error of $O(\Delta t^3)$ for only the first time step, not changing the order of accuracy of the scheme. The same happens when $p < 1$ and the equation of y_i^3 is affected by an error of $O(\Delta t^2)$, which introduces an error of $O(\Delta t^3)$ in the final stage, without affecting the order of accuracy. When

$q > 1$ some σ_i are an $O(\varepsilon)$, so also the final stage has divisions by $O(\varepsilon)$ terms. This leads to an error of $O(\Delta t)$ for few steps and a global first order of accuracy. Finally, when $p > 1$ the term y_i^3 carries an error of $O(\Delta t)$, which results in an $O(\Delta t^2)$ in the final stage, hence, reducing to second order of accuracy. The orders are summarized in figure 3a.

- MPRKSO(2,2,α,β)
  As for (MPRK(2,2,α)) for $\alpha > 1$, also for MPRKSO(2,2,α,β) when $\gamma = \frac{1-\alpha\beta+\alpha\beta^2}{\beta(1-\alpha\beta)} < 1$, there exist some destruction terms that are an $O\left(\frac{\Delta t}{\varepsilon}\right)$. Hence, for $\gamma < 1$ the scheme is only first order accurate.

- MPRKSO(4,3)
  The MPRKSO(4,3) follows a pattern similar to the MPRK(2,2,α) for $\alpha < 1$. Indeed, in the case $\varepsilon \to 0$, we see that $\mu_2 \to +\infty$, hence, it destroy the destruction terms in $\tilde{a}_i$ leading to an $O(\Delta t)$ error in it and in σ_i. This accumulate also in u_i^1 through the weights $\frac{u_i^1}{\sigma_i}$, which leads to an $O(\Delta t^2)$ in the first stage. Hence, the accuracy is reduced to second order.

- mPDeC
  Also mPDeC for high orders loses orders of accuracy. The reason is the switch between production and destruction weighting due to the negative θ_l^m coefficients. In those situations, we have some terms of the type

$$\frac{p_{21}(u^{m,(0)})}{u_2^{m,(0)}} = \frac{\theta u_1^{m,(0)}}{u_2^{m,(0)}} = \frac{\theta(1-\varepsilon)}{\varepsilon} \tag{46}$$

  and, when the limit for $u_2^{m,(0)} = \varepsilon \to 0$ is applied, that term goes to infinity, as the production is not proportional to $u_2^{m,(0)}$. Hence, it pushes to 0 the $y_i^{m,(1)}$ for which there are $\theta_l^m < 0$. In most of the cases (Gauss–Lobatto subtimesteps and low order equispaced subtimesteps) there are no negative θ_l^M at the last stage, hence, the last stage is not directly affected by this infinity term. On the other side, some intermediate stages are affected by an $O(\Delta t)$ error which becomes an $O(\Delta t^2)$ at the final stage of the first timestep. This explains the reduction to second order of accuracy of all the high order mPDeC in figure 2. For some mPDeC with equispaced subtimesteps and order larger than 8 also the final stage has negative weights. Those methods reduce to first order of accuracy, similarly to MPRK(2,2,α) for $\alpha > 1$.

- MPRK(3,2)
  The term y_2^1 appear at the denominator only in the first stage, which is a MPE Euler stage. Hence, it does not effect the accuracy of the scheme.

- Semi implicit schemes
  In the semi implicit schemes SI-RK2 and SI-RK3, the destruction is never explicitly used, but it is always presented in the form of $\tilde{d}$. Hence, all the cancellations happen smoothly in the stages and no order reduction is observed.

□

For an automatic detection of such order reduction in the first step of the scheme, one can use symbolic tools and write, for specific problems and methods the Taylor expansion of the solution at the first timestep first in ε and then in Δt. As an example, we show here the Taylor expansion for the error $\eta(\varepsilon, \Delta t) := u_1^1(\varepsilon, \Delta t) - u_1^{ex}(\varepsilon, \Delta t)$ for mPDeC3. Expanding first Δt and then ε in 0 we obtain

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

which means third order of accuracy for non vanishing ε, while, letting $\varepsilon \to 0$ first, we obtain

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

and, hence, we have an error of $O(\Delta t^2)$ for the first step and a global second order of accuracy. More Taylor expansions can be found in the supplementary material [43] and the computations for these tests can be found in Mathematica notebooks in the accompanying reproducibility repository [42].

In order to validate this theoretical results, in the next section we will also search through all the schemes for first order ones, but before we give the following remark on the stability theory from [18–20].

Remark 4.3. There is no classical accepted stability theory for Patankar type methods. Recently, in [18–20] a promising ansatz to investigate the behavior of conservative and positivity preserving methods has been proposed. In their work, the main idea is to use the center manifold theory corresponding to fixed-point investigations. First applied on 2×2 systems in [18], the theory has been extended to general $n \times n$ systems in [19]. The main idea is the following: a generic linear system $y' = Ay$ with $A \in R^{n\times n}$ with initial condition $y_0 > 0$ possessing $k > 0$ linear invariants is considered. In such a case, zero is always an eigenvalue of A which implies the existence of nontrivial steady state solutions, cf. [19]. The steady state solutions are fixed-points for any reasonable time integration method. Due to the nonlinear character of Patankar-type schemes (actually for all higher-order positivity preserving schemes), a nonlinear iteration process is obtained. Here, additional techniques have to be used to investigate the stability properties. The authors of [18–20] proved a theorem based on the central manifold theorem which gives sufficient conditions for the stability of all such methods. It is further demonstrated that MPRK22(α) is stable in such context meaning for all $\Delta t > 0$ it will converge to such fixed-points at any rate.
As shown in [18], we suspect that most of modified Patankar schemes are stable in the fixed-point sense. In our investigation, we do not deal with this type of stability, but, rather, we look for some more restricted schemes that show monotone character for monotone problems and that do not completely lose the high order accuracy. A stability analysis of all the considered methods in respect to the method proposed in [19] is work in progress. Furthermore, the connection between our observations and the obtained eigenvalues of the iterative process will be considered and compared in the future.

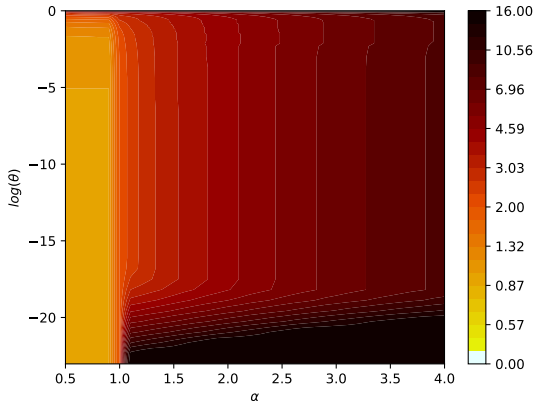
5. Numerical experiments for simplified linear systems

As described in Section 3, we consider the simplified 2×2 system (12) 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 (16) 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 $\tilde{\theta} = 1 - \theta$.

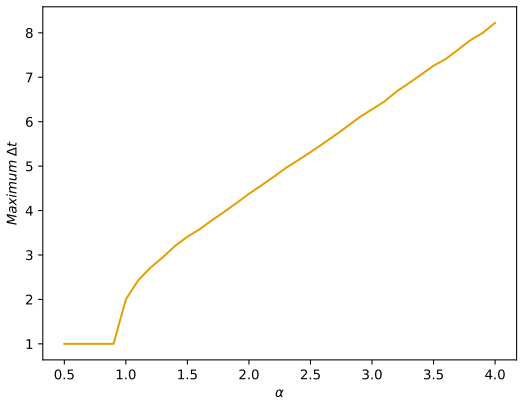
In the following tests, we compare different methods and families presented above: MPRK(2,2,α), MPRK(4,3,α, β), MPRKSO(2,2,α,β), MPRKSO(4,3), mPDeC both for equispaced and Gauss–Lobatto subtimesteps, MPRK(3,2), SI-RK2, and SI-RK3.

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 (16) will be denoted as our bound.

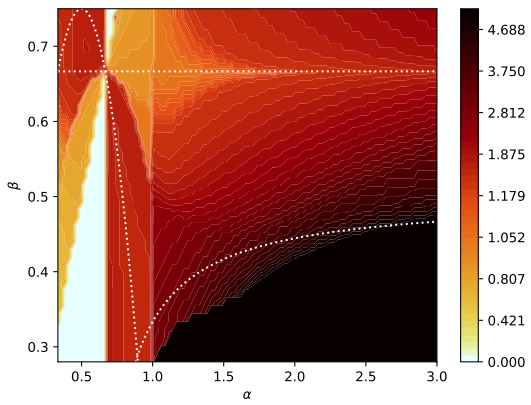
In Figure 4 and 5, we present the results for the all the modified Patankar methods and for the semi-implicit Runge-Kutta methods. We highlight that the evaluation of condition (16) is done with a tolerance of $5\times$ machine epsilon. Some tests can be sensitive to this tolerance, in particular



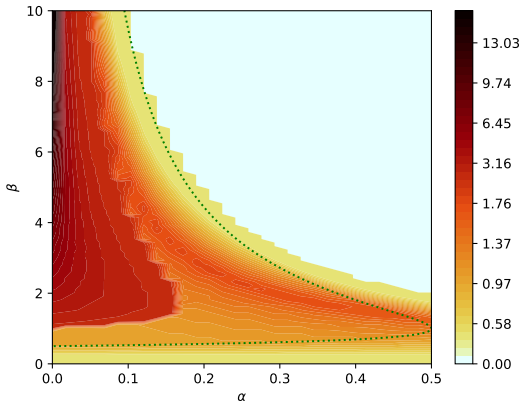
(a) MPRK(2,2,α): Δt bound varying the system through θ and the method with α.



(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 [24].



(d) Δt bound for MPRKSO(2,2,α,β) varying α and β. The green dashed lines bound the positive RK coefficients area [15].

Figure 4: Numerical search of the Δt bound for having an oscillation-free first time step, in the sense of (16), for problem (12) varying IC and system parameter θ: MPRK(2,2,α), MPRK(4,3,α,β) and MPRKSO(2,2,α,β).

mPDeC

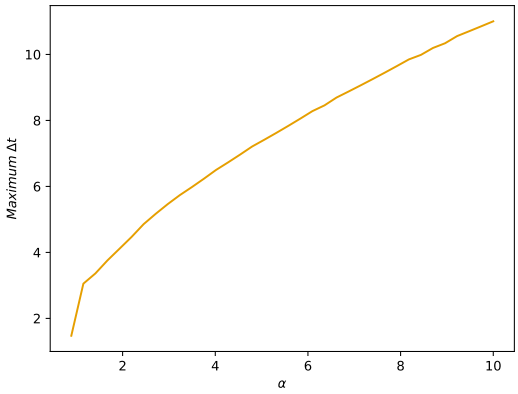
Equispaced	
p	Δt bound
1	∞
2	2.0
3	1.19
4	1.11
5	1.07
6	1.04
7	1.04
8	1.37
9	6.96
10	1.0
11	15.5
12	1.0
13	35.51
14	1.07
15	12.13
16	1.80

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 first order accuracy for vanishing initial conditions.

Method	Δt bound
MPRKSO(4,3)	1.31
SI-RK2	1.41
SI-RK3	1.27
MPRK(3,2)	16.56

(b) Nonparamteric Patankar schemes and their Δt bounds.



(c) 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.

Method	Δt bound
ImplicitMidpoint	2.0
Trapezoid	2.0
TRBDF2	2.38
SDIRK2	2.64
RadauIIA5	∞

(d) Other methods and their Δt bounds.

Figure 5: Numerical search of the Δt bound for having an oscillation-free first time step, in the sense of (16), for problem (12) varying IC and system parameter θ.

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 validating the loss of accuracy of the schemes when they fall back to first order methods for $\varepsilon \to 0$. For this, we consider the system (12) with $\theta = 0.5$, 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 say that the scheme is at most first order accurate. 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 [42].

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 $\alpha > 1$. We recall that the methods with $\alpha > 1$ lose the order of accuracy in the limit $\varepsilon \to 0$, preserving the initial condition as spurious steady state for few time steps. 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 only first order accuracy for problems with $\varepsilon \to 0$ as one can compare with figure 3a. It is noticeable that around the curve $\beta(6\alpha - 3) = 3\alpha - 2$, which is a boundary for nonnegative coefficients [24], the Δt bound is particularly large. Hence, in Figure 5c 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 are only first order accurate for vanishing initial conditions.

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 recall that as $\varepsilon \to 0$ we lose the order of accuracy for small α and large β. The precise area where this happens is denoted in brown in figure 3b. 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, all the schemes reduce to order 2 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 observe also order reduction to first order for high order schemes, more precisely for order 9 and order greater or equal to 11, when there are some negative θ_l^M.

The MPRKSO(4,3) scheme has a Δt bound of 1.31, as shown in Figure 5b. Moreover, it does show a reduction only to order 2 for the numerical tests with vanishing initial conditions. MPRK(3,2) has maybe the best conditions of all the schemes, see Figure 5b. Its Δt bound is around 16 and it keeps its second order accuracy.

In Figures 5b, the semi-implicit schemes are presented. Both show similar behaviors with Δt slightly larger than unity. For these methods, there is no loss of accuracy.

In Figure 5d, we report the Δt bound for some other standard time discretizations. Their implementation is available in the DifferentialEquations.jl [36] package in Julia [4]. We observe that many classical implicit schemes have a bound of around 2, while RadauIIA5 is unconditionally monotone. Clearly all these methods do not suffer of order reduction for vanishing initial conditions.

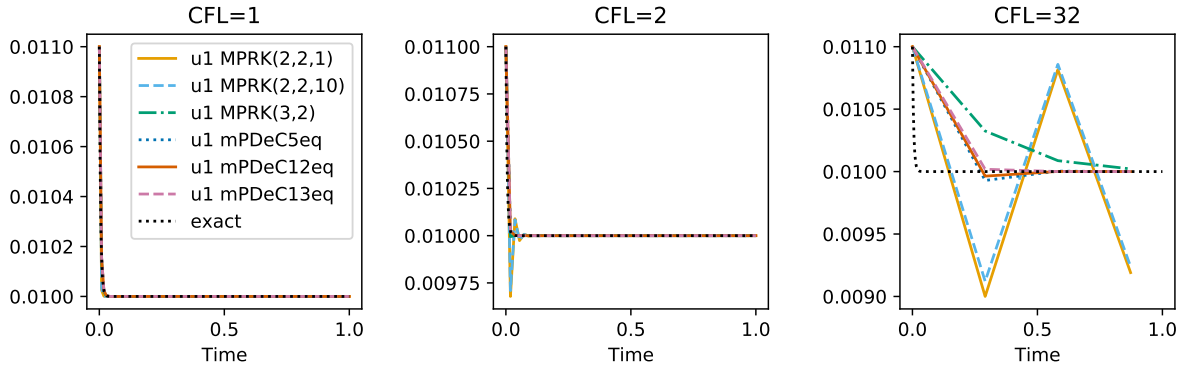


Figure 6: Simulations of (47) at different CFLs for some schemes.

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 [7]. 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

$$u' = -k|u|u + 1. \tag{47}$$

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 (47) 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.

Figure 6 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 (15) with different CFLs. In the supplementary material [43], 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.

Table 1: Oscillation measure for problem (47) 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

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 5c.

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 $\approx$ 16, which shows anyway a very small oscillation only for CFL=2, see Figure 6 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 5b.

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.

6.2. Robertson problem

The Robertson problem [28, 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 [22] with non-zero components

$$p_{12}(u) = d_{21}(u) = k_3 u_2 u_3, \quad p_{21}(u) = d_{12}(u) = k_1 u_1, \quad p_{32}(u) = d_{23}(u) = k_2 u_2, \tag{48}$$

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 [22]. 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 loss of accuracy problem as some constituents are not present

as initial conditions. In Figure 7, 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 order reduction to 1 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) [13]. The original problem HIRES [28, Section II.1] can be rewritten as a nine-dimensional production–destruction system with

$$\begin{aligned}
&r_1(u) = \sigma, && d_{12}(u) = k_1 u_1, && d_{21}(u) = k_2 u_2,\\
&d_{24}(u) = k_3 u_2, && d_{34}(u) = k_1 u_3, && d_{31}(u) = k_6 u_3,\\
&d_{43}(u) = k_2 u_4, && d_{46}(u) = k_4 u_4, && d_{56}(u) = k_1 u_5,\\
&d_{53}(u) = k_5 u_5, && d_{65}(u) = k_2 u_6, && d_{75}(u) = \frac{k_2}{2} u_7,\\
&d_{76}(u) = \frac{k_-}{2} u_7, && d_{79}(u) = \frac{k_*}{2} u_7, && d_{67}(u) = k_+ u_6 u_8,\\
&d_{87}(u) = k_+ u_6 u_8, && d_{78}(u) = \frac{k_- + k_* + k_2}{2} u_7,
\end{aligned} \tag{49}$$

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

$$\begin{aligned}
&k_1 = 1.71, \quad k_2 = 0.43, \quad k_3 = 8.32, \quad k_4 = 0.69, \quad k_5 = 0.035,\\
&k_6 = 8.32, \quad k_+ = 280, \quad k_- = 0.69, \quad k_* = 0.69, \quad \sigma = 0.0007.
\end{aligned} \tag{50}$$

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 inaccuracy of some methods also for this problem as some constituents are close to 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 [28] 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 [43]. For brevity, we plot in Figure 8 just a sample.

For mPDeC, we observe the loss of accuracy only for equispaced time steps for high odd orders (9, 11, 13 and so on). In Figure 8, 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 great loss of accuracy only for $\alpha > 1$. This is demonstrated in Figure 8 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 better 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 spurious steady state. An oscillatory behavior can be observed, though, also in the first simulation, which is shown in Figure 8. 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 loss of accuracy only for the second one, in accordance with the linear tests. For MPRKSO(4,3), MPRK(3,2), SI-RK2 and

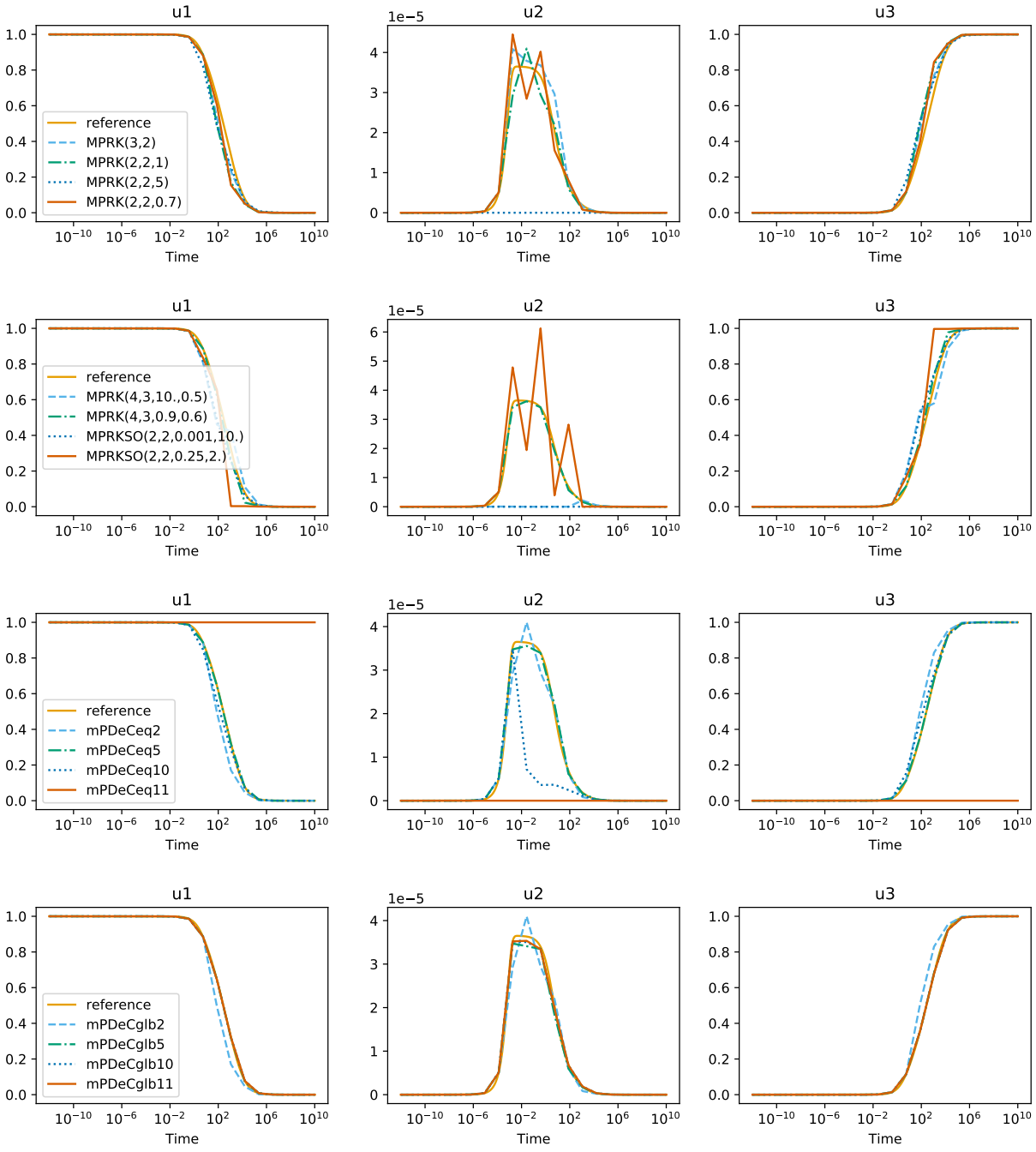


Figure 7: Robertson problem with different methods and 20 time steps.

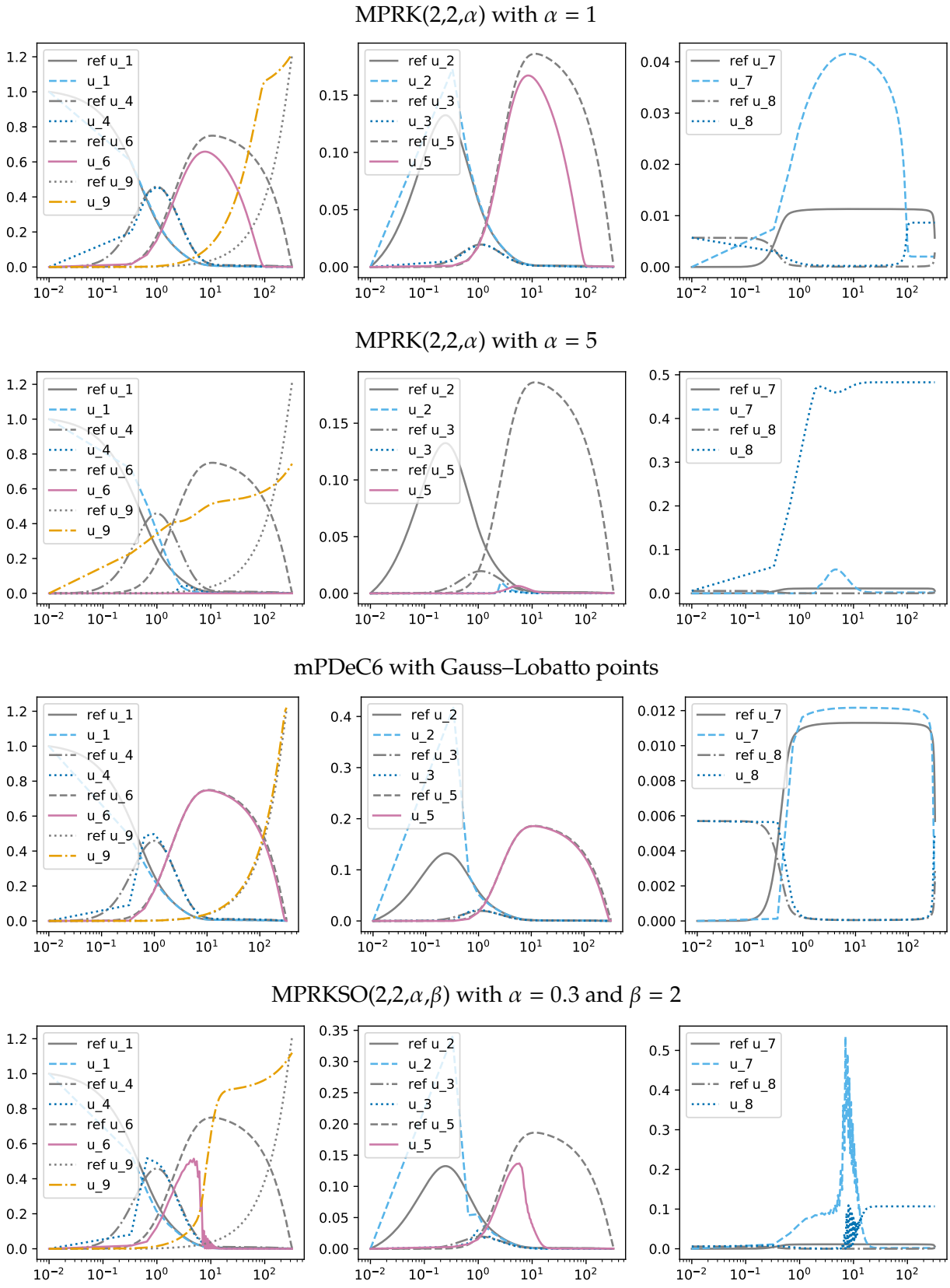


Figure 8: Simulations run with different schemes with $N = 10^3$ time steps, plot in logarithmic scale in time.

SI-RK3, we do not observe significant loss of accuracy, 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 loss of accuracy 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 up to first order of accuracy. Finally, we applied the methods to more challenging problems including stiff nonlinear ones. We observed that our proposed oscillation-free and accuracy analysis generalizes reasonably well to these other problems.

From our point of view, this is a first step toward further investigations on Patankar-type schemes. Extensions 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.

As mentioned in Remark 4.3, a stability analysis of all the considered methods in respect to [19] is work in progress. Furthermore, the connection between our observations and the obtained eigenvalues of the iterative process will be considered and compared in the future.

We plan also to extend our investigation to hyperbolic conservation laws. After a spatial semidiscretizations, we obtain ODEs that can be written as a production–destruction–rest system [8, 16, 29]. Here, the relation between the time step restrictions derived in this work and classical CFL conditions will be the major focus of research.

Acknowledgments

D. T. was funded by Team CARDAMOM in Inria–Bordeaux Sud–Ouest, France and by a SISSA Mathematical Fellowship, Italy. 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 [23, 24] is repeated for completeness. Please note that the investigated version is called $MPRK43I(\alpha,\beta)$ in their papers. It is given by

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

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

$$
\begin{array}{c|ccc}
0 & & & \\
\alpha & \alpha & & \\
\beta & \frac{3\alpha\beta(1-\alpha)-\beta^2}{\alpha(2-3\alpha)} & \frac{\beta(\beta-\alpha)}{\alpha(2-3\alpha)} & \\
\hline
 & 1+\frac{2-3(\alpha+\beta)}{6\alpha\beta} & \frac{3\beta-2}{6\alpha(\beta-\alpha)} & \frac{2-3\alpha}{6\beta(\beta-\alpha)}
\end{array}
\tag{51}
$$

with positive coefficients for

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

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 [16] is repeated. It is given by

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

Here, the optimal SSP coefficients determined in [16] will be used. They are given by

$$
\begin{aligned}
n_1 &= 2.569046025732011E-01, & n_2 &= 7.430953974267989E-01, \\
a_{10} &= 1, & a_{20} &= 9.2600312554031827E-01, \\
a_{21} &= 7.3996874459681783E-02, & a_{31} &= 2.0662904223744017E-10, \\
b_{10} &= 4.7620819268131703E-01, & a_{30} &= 7.0439040373427619E-01, \\
a_{32} &= 2.9560959605909481E-01, & b_{20} &= 7.7545442722396801E-02, \\
b_{21} &= 5.9197500149679749E-01, & b_{31} &= 6.8214380786704851E-10, \\
b_{30} &= 2.0044747790361456E-01, & b_{32} &= 5.9121918658514827E-01, \\
\eta_1 &= 3.777285888379173E-02, & \eta_2 &= 1/3, \\
\eta_3 &= 1.868649805549811E-01, & \eta_3 &= 2.224876040351123, \\
z &= 6.288938077828750E-01, & s &= 5.721964308755304.
\end{aligned}
$$

B. Initial direction of Patankar schemes

As seen in Section 3, we are looking for schemes that do not oscillate. To check this, there are two inequalities that each step must verify. Given an arbitrary initial condition, the first step should go towards the steady state and should not overshoot the steady state. 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 (12) if $u_1^0 = (1-\varepsilon) > (1-\theta)$ implies that $u_1^1 < u_1^0$ and $u_1^0 = (1-\varepsilon) < (1-\theta)$ 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,*

$$u_1^0 > (1-\theta) \Longrightarrow u_1^0 > u_1^1. \tag{52}$$

Proof. We write the MPE for the system (12) in the first equation, making use of the conservation property and we collect all the implicit terms.

$$u_1^1 = u_1^0 + \Delta t \left((1-\theta)(1-u_1^0)\frac{1-u_1^1}{1-u_1^0} - \theta u_1^0 \frac{u_1^1}{u_1^0} \right), \tag{53a}$$

$$u_1^1 = u_1^0 + \Delta t \left((1-\theta)(1-u_1^1) - \theta u_1^1 \right), \tag{53b}$$

$$u_1^1(1+\Delta t) = y_1^1 + \Delta t(1-\theta), \tag{53c}$$

$$u_1^1 = \frac{u_1^0 + \Delta t(1-\theta)}{(1+\Delta t)} < \frac{u_i^0(1+\Delta t)}{(1+\Delta t)} = u_1^0. \tag{53d}$$

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* (12) *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 (12), collecting all the implicit terms and using the hypothesis $u_1^0 > (1-\theta)$.

$$\begin{aligned} u_1^1 = u_1^0 + \Delta t \Bigg(& \left(\frac{2\alpha-1}{2\alpha}(1-\theta)(1-y_1^1) + \frac{1}{2\alpha}(1-\theta)(1-y_1^2) \right) \frac{1-u_1^1}{(1-y_1^2)^{1/\alpha}(1-y_1^1)^{1-1/\alpha}} \\ & - \left(\frac{2\alpha-1}{2\alpha}\theta y_1^1 + \frac{1}{2\alpha}\theta y_1^2 \right) \frac{u_1^1}{(y_1^2)^{1/\alpha}(y_1^1)^{1-1/\alpha}} \Bigg), \end{aligned} \tag{54a}$$

$$\begin{aligned} u_1^1 = u_1^0 + \Delta t \Bigg(& \left(\frac{2\alpha-1}{2\alpha}(1-\theta)\left(\frac{1-y_1^1}{1-y_1^2}\right)^{1/\alpha} + \frac{1}{2\alpha}(1-\theta)\left(\frac{1-y_1^2}{1-y_1^1}\right)^{1-1/\alpha} \right)(1-u_1^1) \\ & - \left(\frac{2\alpha-1}{2\alpha}\theta\left(\frac{y_1^1}{y_1^2}\right)^{1/\alpha} + \frac{1}{2\alpha}\theta\left(\frac{y_1^2}{y_1^1}\right)^{1-1/\alpha} \right) u_1^1 \Bigg), \end{aligned} \tag{54b}$$

$$\begin{aligned} & \left(1 + \Delta t \left(\frac{2\alpha-1}{2\alpha}(1-\theta)\left(\frac{1-y_1^1}{1-y_1^2}\right)^{1/\alpha} + \frac{1}{2\alpha}(1-\theta)\left(\frac{1-y_1^2}{1-y_1^1}\right)^{1-1/\alpha} \right) + \right. \\ & \left. \Delta t \left(\frac{2\alpha-1}{2\alpha}\theta\left(\frac{y_1^1}{y_1^2}\right)^{1/\alpha} + \frac{1}{2\alpha}\theta\left(\frac{y_1^2}{y_1^1}\right)^{1-1/\alpha} \right) \right) u_1^1 = \\ & u_1^0 + \Delta t \left(\left(\frac{2\alpha-1}{2\alpha}(1-\theta)\left(\frac{1-y_1^1}{1-y_1^2}\right)^{1/\alpha} + \frac{1}{2\alpha}(1-\theta)\left(\frac{1-y_1^2}{1-y_1^1}\right)^{1-1/\alpha} \right) < \right. \\ & \left. u_1^0 \left(1 + \Delta t \left(\left(\frac{2\alpha-1}{2\alpha}\left(\frac{1-y_1^1}{1-y_1^2}\right)^{1/\alpha} + \frac{1}{2\alpha}\left(\frac{1-y_1^2}{1-y_1^1}\right)^{1-1/\alpha} \right) \right) \right). \end{aligned} \tag{54c}$$

So we have that

$$u_1^1 < u_1^0 \frac{N}{D} \tag{54d}$$

with $N > 0$ and $D > 0$ deducible from (54c). If $N < D$ we have our result, or, in other words, if $N - D < 0$. So, let us compute

$$\begin{aligned}
\frac{N-D}{\Delta t} =& \frac{2\alpha-1}{2\alpha}\left(\frac{1-y_1^1}{1-y_1^2}\right)^{1/\alpha} + \frac{1}{2\alpha}\left(\frac{1-y_1^2}{1-y_1^1}\right)^{1-1/\alpha} - \\
& \frac{2\alpha-1}{2\alpha}(1-\theta)\left(\frac{1-y_1^1}{1-y_1^2}\right)^{1/\alpha} - \frac{1}{2\alpha}(1-\theta)\left(\frac{1-y_1^2}{1-y_1^1}\right)^{1-1/\alpha} - \\
& \frac{2\alpha-1}{2\alpha}\theta\left(\frac{y_1^1}{y_1^2}\right)^{1/\alpha} - \frac{1}{2\alpha}\theta\left(\frac{y_1^2}{y_1^1}\right)^{1-1/\alpha},
\end{aligned} \tag{54e}$$

$$\begin{aligned}
\frac{N-D}{\Delta t} =& \frac{2\alpha-1}{2\alpha}\theta\left(\frac{1-y_1^1}{1-y_1^2}\right)^{1/\alpha} + \frac{1}{2\alpha}\theta\left(\frac{1-y_1^2}{1-y_1^1}\right)^{1-1/\alpha} - \\
& \frac{2\alpha-1}{2\alpha}\theta\left(\frac{y_1^1}{y_1^2}\right)^{1/\alpha} - \frac{1}{2\alpha}\theta\left(\frac{y_1^2}{y_1^1}\right)^{1-1/\alpha} = \\
& \frac{2\alpha-1}{2\alpha}\theta\left(\left(\frac{1-y_1^1}{1-y_1^2}\right)^{1/\alpha} - \left(\frac{y_1^1}{y_1^2}\right)^{1/\alpha}\right) + \frac{1}{2\alpha}\theta\left(\left(\frac{1-y_1^2}{1-y_1^1}\right)^{1-1/\alpha} - \left(\frac{y_1^2}{y_1^1}\right)^{1-1/\alpha}\right).
\end{aligned} \tag{54f}$$

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* (12) *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 (Accuracy 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 second order area for vanishing IC of MPRKSO(2,2,α,β) found in Figure 3b.

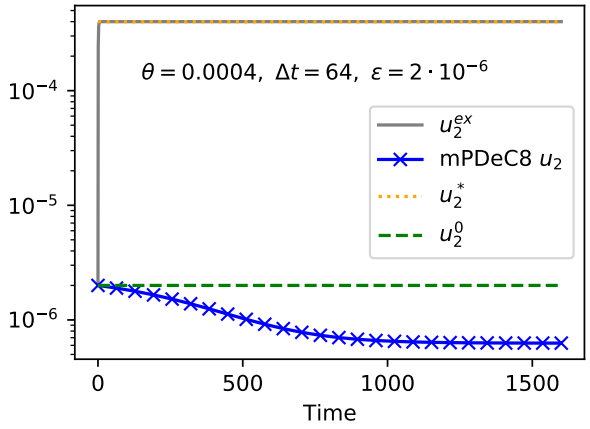


Figure 9: Simulation of (12) with $\theta = 4 \cdot 10^{-4}$ and $u_2^0 = \varepsilon = 2 \cdot 10^{-6}$ with mPDeC8 with equispaced points for $\Delta t = 64$

B.1. Initial direction of 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 [42], while the only theoretical result is in `MPRK_3_2.nb`. We summarize in the following the results we obtained.

- 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 (tested 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 9, 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] R. Abgrall. "High order schemes for hyperbolic problems using globally continuous approximation and avoiding mass matrices." In: *Journal of Scientific Computing* 73.2 (2017), pp. 461–494.

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

[3] 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.

[4] 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].

[5] 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.

[6] 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.

[7] 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.

[8] 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].

[9] 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.

[10] L. Formaggia and A. Scotti. "Positivity and Conservation Properties of Some Integration Schemes for Mass Action Kinetics." In: *SIAM Journal on Numerical Analysis* 49.3 (2011), pp. 1267–1288. DOI: 10.1137/100789592.

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

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] T. Izgin, S. Kopecz, and A. Meister. "On Lyapunov Stability of Positive and Conservative Time Integrators and Application to Second Order Modified Patankar–Runge–Kutta Schemes." In: *arXiv preprint arXiv:2202.01099* (2022).

[19] T. Izgin, S. Kopecz, and A. Meister. "On the Stability of Unconditionally Positive and Linear Invariants Preserving Time Integration Schemes." In: *arXiv preprint arXiv:2202.11649* (2022).

[20] T. Izgin, S. Kopecz, and A. Meister. "Recent Developments in the Field of Modified Patankar-Runge-Kutta-methods." In: *PAMM* 21.1 (2021), e202100027.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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).

[26] 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.

[27] A. Martiradonna, G. Colonna, and F. Diele. "GeCo: Geometric Conservative nonstandard schemes for biochemical systems." In: *Applied Numerical Mathematics* 155 (2020), pp. 38–57. DOI: 10.1016/j.apnum.2019.12.004.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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].

[33] 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.

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

[35] O. Pratt. *New and Easy Method of Solution of the Cubic Biquadratic Equations: Embracing Several New Formulas, Greatly Simplifying this Department of Mathematical Science*. Liverpool: Longmans, Green, Reader, and Dyer, 1866.

[36] 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.

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

[38] 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.

[39] 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.

[40] 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.

[41] 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].

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

[43] D. Torlo, P. Öffner, and H. Ranocha. *Issues with Positivity-Preserving Patankar-type Schemes*. arXiv preprint: https://arxiv.org/abs/2108.07347. 2021. arXiv: 2108.07347 [math.NA].