A necessary condition for non oscillatory and positivity preserving time-integration schemes

Th. Izgin and P. Öffner and D. Torlo

Abstract Modified Patankar (MP) schemes are conservative, linear implicit and unconditionally positivity preserving time-integration schemes constructed for production-destruction systems. For such schemes, a classical stability analysis does not yield any information about the performance. Recently, two different techniques have been proposed to investigate the properties of MP schemes. In Izgin *et al.* [ESAIM: M2AN, 56 (2022)], inspired from dynamical systems, the Lyapunov stability properties of such schemes have been investigated, while in Torlo *et al.*[Appl. Numer. Math., 182 (2022)] their oscillatory behaviour has been studied. In this work, we investigate the connection between the oscillatory behaviour and the Lyapunov stability and we prove that a condition on the Lyapunov stability function is necessary to avoid oscillations. We verify our theoretical result on several numerical tests.

1 Introduction

Consider a production–destruction system (PDS) of ODEs

$$\frac{\mathrm{d}\, y_i(t)}{\mathrm{d}\, t} = \sum_{j=1}^{I} \left(p_{ij}(y(t)) - d_{ij}(y(t)) \right), \qquad i = 1, \dots, I, \quad t \in \mathbb{R}_0^+, \tag{1}$$

Thomas Izgin
Department of Mathematics, University of Kassel, Germany, e-mail: izgin@mathematik.uni-kassel.de

Philipp Öffner
Institute of Mathematics, Johannes Gutenberg University, Mainz, Germany e-mail: poeffner@uni-mainz.de

Davide Torlo
mathLab SISSA, SISSA International School of Advanced Studies, Trieste, Italy e-mail: davide.torlo@sissa.it

where $p_{ij}, d_{ij} : \mathbb{R}^I \to \mathbb{R}_0^+$ are Lipschitz continuous production and destruction functions, respectively, such that $p_{ij}(y) = d_{ji}(y)$ and $\lim_{y_i \to 0} d_{ij}(y) = 0$. Then the system (1) is conservative, i.e., $\sum_i y_i(t) = \sum_i y_i(0)$, and positive, that is, if $y_i(0) \geq 0$ for all i, then $y_i(t) \geq 0$ for all i. These systems arise in various fields, e.g. chemical reactions and biological processes, but can be also obtained from spatial discretisations of hyperbolic conservation/balance laws, e.g. shallow water equations or Euler equations.

Modified Patankar (MP) schemes are conservative, linear implicit and unconditionally positivity preserving time-integration schemes constructed for PDS, inspired by Patankar's original work [17]. In recent years, many different MP schemes have been developed [2, 13, 7, 16], they have been applied to different applications [3, 6, 15] and their properties have been studied [12, 5, 9, 11, 14, 19].

In the following, we compare the oscillations observed in 2 dimensional systems in [19] and the Lyapunov stability function studied in [8]. Indeed, it is possible to show that a condition on the Lyapunov stability function is necessary to have oscillations–free schemes. In Section 2, we present the proof of this result; in Section 3, we list some stability function of some MP schemes and in Section 4 we show how the numerical results validate the theoretical findings.

2 Connection between oscillations and Lyapunov stability

We restrict to a linear 2–dimensional problem, in order to have a clear definition of oscillations [19]. All 2–dimensional linear systems of ODEs that are positive and conservative can be rewritten, with a change of variables, as the following IVP

$$\begin{cases} \mathbf{y}'(t) = \mathbf{A}_\theta \mathbf{y}(t), \\ \mathbf{y}(0) = \mathbf{y}^0 > \mathbf{0}, \end{cases} \qquad \mathbf{A}_\theta = \begin{pmatrix} -\theta & 1-\theta \\ \theta & -(1-\theta) \end{pmatrix}, \quad \theta \in (0,1), \tag{2}$$

where this can be seen as PDS, with $p_{12} = d_{21} = (1-\theta)y_2$, $d_{12} = p_{21} = \theta y_1$ and all other entries zero. Let us also consider a one step numerical method whose iterates are generated by a map $\mathbf{g}$, i. e. $\mathbf{y}^{n+1} := \mathbf{g}(\mathbf{y}^n)$. Note that $\mathbf{g}$ might be given implicitly.

We first describe oscillations for 2–dimensional linear ODEs through the solution and the steady state. It is known that the exact solution does not overshoot the steady state.

Definition 1 a) A method is *not overshooting* the steady state of (2) if $y_2^1 < \theta$ and $y_1^1 > 1-\theta$ for any given initial state $\mathbf{y}^0 = (1-\varepsilon, \varepsilon)^\intercal$ with $\varepsilon < \theta$, while when $\varepsilon > \theta$ the method is *not overshooting* the steady state if $y_2^1 > \theta$ and $y_1^1 < 1-\theta$.
b) Otherwise the method is said to be *overshooting* the steady state of (2).

Theorem 1 *Let any positive steady state of* (2) *be a fixed point of a map* $\mathbf{g} \in C^2(\mathbb{R}^2_{>0})$. *In addition, let the iterates generated by* $\mathbf{y}^{n+1} = \mathbf{g}(\mathbf{y}^n)$ *satisfy* $\|\mathbf{y}^{n+1}\|_1 = \|\mathbf{y}^n\|_1$ *for all* $n \in \mathbb{N}_0$. *Finally, let* $\mathbf{y}^*$ *be the unique positive steady state of* (2).

Then, the spectrum of the Jacobian $\mathbf{Dg}(\mathbf{y}^*)$ *is* $\sigma(\mathbf{Dg}(\mathbf{y}^*)) = \{1, R\}$ *with* $R \in \mathbb{R}$. *Furthermore, if* $R < 0$, *then the method generated by* $\mathbf{g}$ *is overshooting the steady state of* (2).

Proof Throughout this proof, we use $\mathbf{e}_1 = (1,0)^\intercal, \mathbf{e}_2 = (0,1)^\intercal$ to denote the standard unit vectors as well as the notation $\bar{\mathbf{y}} = (1,-1)^\intercal$. In the proof of [8, Theorem 2.9], it is shown that $\mathbf{Dg}(\mathbf{y}^*)\mathbf{y}^* = \mathbf{y}^*$ and $\mathbf{Dg}(\mathbf{y}^*)\bar{\mathbf{y}} = R\bar{\mathbf{y}}$ with $R \in \mathbb{R}$, which means that the matrix of eigenvectors

$$\mathbf{S} = (\mathbf{y}^* \;\; \bar{\mathbf{y}}) \tag{3}$$

is invertible since $\bar{\mathbf{y}}$ cannot be a multiple of the positive vector $\mathbf{y}^*$. In particular, we obtain

$$\mathbf{S}^{-1}\mathbf{Dg}(\mathbf{y}^*)\mathbf{S} = \operatorname{diag}(1, R),$$

where $\operatorname{diag}(\mathbf{y}) \in \mathbb{R}^{2\times 2}$ denotes the diagonal matrix with $(\operatorname{diag}(\mathbf{y}))_{ii} = y_i$ for $i = 1, 2$. Following the lines of the proof of [8, Theorem 2.9], we introduce the affine linear transformation $\mathbf{T}\colon \mathbb{R}^2 \to \mathbb{R}^2$,

$$\mathbf{y} \mapsto \mathbf{w} = \mathbf{T}(\mathbf{y}) = \mathbf{S}^{-1}(\mathbf{y} - \mathbf{y}^*),$$

where $\mathbf{S}$ is given in (3) and the inverse transformation $\mathbf{T}^{-1}$ is given by

$$\mathbf{T}^{-1}(\mathbf{w}) = \mathbf{S}\mathbf{w} + \mathbf{y}^*.$$

To see that the method defined by $\mathbf{g}$ is overshooting $\mathbf{y}^*$, we show that the transformed method given by the map

$$\mathbf{G}\colon \mathbf{T}(\mathbb{R}^2_{>0}) \to \mathbf{T}(\mathbb{R}^2_{>0}), \quad \mathbf{G}(\mathbf{w}) = \mathbf{T}(\mathbf{g}(\mathbf{T}^{-1}(\mathbf{w})))$$

is overshooting the transformed steady state which is $\mathbf{w}^* = \mathbf{0}$. As demonstrated in [8, Theorem 2.9], $\mathbf{y}^0$ is transformed onto the w_2-axis and due to the conservation of the map $\mathbf{g}$, it is proven that $\mathbf{G}(\mathbf{w}^0) \in \operatorname{span}(\mathbf{w}^0)$ for $\mathbf{w}^0 = (0, w_2^0)^\intercal$. Moreover,

$$\mathbf{G}(\mathbf{w}) = \operatorname{diag}(1, R)\mathbf{w} + \mathbf{S}^{-1}\bar{\boldsymbol{R}}(\mathbf{T}^{-1}(\mathbf{w}))$$

holds, where $\bar{\boldsymbol{R}}$ denotes the Lagrangian remainder

$$(\bar{\boldsymbol{R}}(\mathbf{y}))_i = \frac{1}{2}(\mathbf{y} - \mathbf{y}^*)^\intercal \mathbf{H}g_i(\mathbf{y}^* + c_i(\mathbf{y} - \mathbf{y}^*))(\mathbf{y} - \mathbf{y}^*), \quad i = 1, 2 \tag{4}$$

for some $c_i \in (0,1)$ depending on $\mathbf{y}$ and $\mathbf{y}^*$ and where $\mathbf{H}g_i$ are the Hessian matrices of g_i for $i = 1, 2$. We consider from now on the iterates given by

$$\mathbf{w}^{n+1} = \begin{pmatrix} 1 & 0 \\ 0 & R \end{pmatrix}\mathbf{w}^n + \mathbf{S}^{-1}\bar{\boldsymbol{R}}(\mathbf{T}^{-1}(\mathbf{w}^n)), \quad \mathbf{w}^0 = (0, w_2^0)^\intercal.$$

Here, using $\mathbf{S}^{-1} = (\widetilde{s}_{ij})_{i,j=1,2}$ and $w_1^n = 0$ it follows from (4) that

$$(\mathbf{S}^{-1}\bar{\boldsymbol{R}}(\mathbf{T}^{-1}(\mathbf{w}^0)))_1 = 0 \tag{5}$$

since $(\mathbf{G}(\mathbf{w}))_1 = w_1$. Furthermore,

$$\begin{aligned}(\mathbf{S}^{-1}\bar{\boldsymbol{R}}(\mathbf{T}^{-1}(\mathbf{w}^0)))_2 =& \frac{1}{2}\sum_{i=1}^{2}\widetilde{s}_{2i}(\mathbf{T}^{-1}(\mathbf{w}^0)-\mathbf{y}^*)^\intercal \mathbf{H}g_i(\xi_i^0)(\mathbf{T}^{-1}(\mathbf{w}^0)-\mathbf{y}^*)\\ =& \frac{1}{2}\sum_{i=1}^{2}\widetilde{s}_{2i}(w_2^0\mathbf{S}\mathbf{e}_2)^\intercal \mathbf{H}g_i(\xi_i^0)(w_2^0\mathbf{S}\mathbf{e}_2)\\ =& \frac{1}{2}\sum_{i=1}^{2}\widetilde{s}_{2i}(w_2^0\bar{\mathbf{y}})^\intercal \mathbf{H}g_i(\xi_i^0)(w_2^0\bar{\mathbf{y}})\\ =& C(\xi_1^0,\xi_2^0)\cdot(w_2^0)^2,\end{aligned} \tag{6}$$

where $\xi_i^0 = \mathbf{y}^* + c_i^0(\mathbf{y}^0 - \mathbf{y}^*)$ and $c_i^0 \in (0,1)$. Also note that the mapping $C\colon \mathbb{R}^2\times\mathbb{R}^2 \to \mathbb{R}$ depends on the entries of the Hessians as well as $\mathbf{S}^{-1}$.

We now prove that the method defined by $\mathbf{G}$ is overshooting $\mathbf{w}^* = \mathbf{0}$ by proving the existence of $w_2^0 \in \mathbb{R}$ such that $\operatorname{sgn}(w_2^1) \neq \operatorname{sgn}(w_2^0)$. We set

$$L = \left\{\mathbf{y}\in\mathbb{R}^2 \middle| \exists s \in \left[-\tfrac{y_1^*}{2}, \tfrac{y_2^*}{2}\right] : \mathbf{y} = \mathbf{y}^* + s\bar{\mathbf{y}}\right\} \subseteq \mathbb{R}^2_{>0}$$

and observe that there exists a $K > 0$ such that $\sup_{\xi\in L\times L}\{|C(\xi_1,\xi_2)|\} \leq K < \infty$ since $\mathbf{g}\in C^2$ has bounded second derivatives on the compact set L.

Next, we restrict to $\mathbf{w}^0$ satisfying $|w_2^0| < \min\left\{\frac{y_1^*}{2}, \frac{y_2^*}{2}, \frac{|R|}{K}\right\}$. As a result, $\mathbf{w}^0 = w_2^0\mathbf{e}_2$ yields $\mathbf{y}^0 = \mathbf{T}^{-1}(\mathbf{w}^0) = \mathbf{S}\mathbf{w}^0 + \mathbf{y}^* = w_2^0\bar{\mathbf{y}} + \mathbf{y}^* \in L$, which means that

$$\xi_i^0 = \mathbf{y}^* + c_i^0(\mathbf{y}^0 - \mathbf{y}^*) = \mathbf{y}^* + c_i^0 w_2^0\bar{\mathbf{y}} \in L$$

for $i = 1, 2$. Now, according to (6), we have

$$w_2^1 = Rw_2^0 + C(\xi_1^0,\xi_2^0)\cdot(w_2^0)^2 = (R + C(\xi_1^0,\xi_2^0)w_2^0)w_2^0. \tag{7}$$

as well as

$$C(\xi_1^0,\xi_2^0)w_2^0 \leq |C(\xi_1^0,\xi_2^0)||w_2^0| < |C(\xi_1^0,\xi_2^0)|\frac{|R|}{K} \leq |R|. \tag{8}$$

Because of $R < 0$, the inequality (8) turns into the statement

$$R + C(\xi_1^0,\xi_2^0)w_2^n < 0,$$

and thus, $\operatorname{sgn}(w_2^1) \neq \operatorname{sgn}(w_2^0)$ due to (7). This proves that the method defined by $\mathbf{G}$ is overshooting $\mathbf{w}^*$ and consequently, the method with iterates given by the map $\mathbf{g}$ is overshooting $\mathbf{y}^*$. □

Remark 1 It was proven in [8] that if $|R| < 1$ holds true, then $\mathbf{y}^*$ is a Lyapunov stable fixed point of the method, whereas it is already well-known that if $|R| > 1$ the corresponding fixed point $\mathbf{y}^*$ is unstable, see [18] for more details. Furthermore, we

want to note that for a numerical time-integration method, the eigenvalue R depends on the time step size Δt, so that R can be interpreted as a *stability function* giving rise to the investigation of stability regions. The result from [8] was generalized, see [9, Theorem 2.9], and applied to many positivity-preserving schemes in [5, 9, 11]. To that end, the corresponding stability functions have been computed, so that we only need to investigate the location of their zeros for investigating the methods with respect to the property of overshooting the steady state of (2).

3 Analysis of Modified Patankar Schemes

In the following, we list the stability functions of some MP schemes. For brevity, we refer to other references for the explicit computations, when available. As derived in [8], the stability function of the second order family of MPRK22(α) schemes, first introduced in [2], is given by

$$R(z) = \frac{-z^2 - 2\alpha z + 2}{2(1-\alpha z)(1-z)}. \tag{9}$$

This function has negative values for negative real part of z if $\mathrm{Re}(z) < -\alpha - \sqrt{\alpha^2+2}$. Hence, for the problem (2) we obtain the necessary condition

$$\Delta t < \Delta t_0(\alpha) := \alpha + \sqrt{\alpha^2+2} \tag{10}$$

for the method not to overshoot the steady state.

The stability functions of the families of MPRK(4,3,α, β) and MPRK(4,3,γ) [13] and the simple MPRK32 [19] are computed in [11] and not reported here for brevity.

Similarly, for SSPMPRK schemes we do not report the stability function of SSPMPRK22(α,β) [6], which can be found in [5], but we focus on the SSPMPRK43(η_2) for $\eta = \frac{1}{3}$ [7]. This scheme possesses the stability function $R(z) = \frac{\sum_{i=1}^{4} a_i z^i}{\sum_{j=1}^{4} b_j z^j}$, where, at double precision

$$\begin{aligned}
a_0 &= 1, & b_0 &= 1,\\
a_1 &= -3.349136322977521, & b_1 &= -4.349136322977523,\\
a_2 &= 2.049225690609540, & b_2 &= 5.898362013587063,\\
a_3 &= 0.6815805312568625, & b_3 &= -3.208879987508106,\\
a_4 &= -0.5093985705698671, & b_4 &= 0.6087426554481902.
\end{aligned}$$

For the Modified Patankar-Deferred-Correction (MPDeC) methods [16], we derive the stability functions as in [11] and we show some examples for different orders. The MPDeC schemes are a class of arbitrarily high order positivity preserving methods, based on the Deferred Correction (DeC) methods [4, 1]. At each stage of the DeC procedure the modified Patankar trick is adopted, carefully choosing the

production and destruction terms, according to the DeC coefficients. The MPDeC schemes are defined by M subtimesteps and K iterations. The order of accuracy of the MPDeC scheme is the minimum between K and the accuracy of the quadrature formula given by the M subtimesteps. We will focus on equispaced (EQ) and Gauss–Lobatto (GL) subtimesteps. To obtain order p, a number of $K = p$ iterations is required, while we need $M = \max\{p-1, 1\}$ EQ subtimesteps or $M = \left\lceil \frac{p}{2} \right\rceil$ GL subtimesteps. The definition of the subtimesteps $0 = t^0 < \cdots < t^M = 1$ leads to the definition of the coefficients $\theta_r^m := \int_0^{t^m} \varphi_r(t)dt$ that are the ground component of the MPDeC schemes. Here, φ_r is the r-th Lagrangian function defined by the subtimenodes $\{t^m\}_{m=0}^M$.

We denote the MPDeC scheme of order p by MPDeC(p) and the corresponding stability function R_p can be computed with the following steps:

$$
\begin{aligned}
R^{m,(1)}(z) &= \frac{1 + 2z\sum_{j=0}^{M} \theta_{j,-}^m}{1 - z\sum_{r=0}^{M} |\theta_r^m|}, \\
R^{m,(\hat{k})}(z) &= \frac{1 + \theta_0^m z + z\sum\limits_{\substack{j=1 \\ j\neq m}}^{M} \theta_j^m R^{j,(\hat{k}-1)}(z) - z\left(\sum\limits_{\substack{j=0 \\ j\neq m}}^{M} |\theta_j^m| - 2\theta_{m,-}^m\right) R^{m,(\hat{k}-1)}(z)}{1 - z\sum_{j=0}^{M} |\theta_j^m|.}, \\
R_p(z) &= R^{M,(K)}(z),
\end{aligned}
$$

for $\hat{k} = 2, \ldots, K$ and $m = 1, \ldots, M$, where $\theta_{r,\pm}^m = \frac{\theta_r^m \pm |\theta_r^m|}{2}$, see [11] for the details. We introduce the matrix $\Theta^{X,(p)} \in \mathbb{R}^{M\times(M+1)}$ satisfying $\Theta_{mr}^{X,(p)} = \theta_{r-1}^m$, where $X \in \{\mathrm{EQ}, \mathrm{GL}\}$ indicates either EQ or GL points. In the case of $p = 2$, i. e., $M = 1$ and $K = 2$ we have $\Theta^{\mathrm{EQ},(2)} = \Theta^{\mathrm{GL},(2)} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \end{pmatrix}$, that is $\theta_0^1 = \theta_1^1 = \frac{1}{2}$, and consequently

$$
R_2(z) = \frac{-z^2 - 2z + 2}{2(1-z)^2}, \tag{11}
$$

which equals the stability function of MPRK22(α) for $\alpha = 1$. This is no surprise since MPDeC(2) is the MPRK22(1) scheme. Next, for $p = 3$ we find

$$
\Theta^{\mathrm{EQ},(3)} = \Theta^{\mathrm{GL},(3)} = \begin{pmatrix} \frac{5}{24} & \frac{1}{3} & -\frac{1}{24} \\ \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \end{pmatrix}
$$

leading to

$$
R_3(z) = \frac{-331z^5 + 1830z^4 + 3096z^3 - 16452z^2 + 16416z - 5184}{36(-12+7z)^2(-1+z)^3}.
$$

Moreover, for $p = 4$ and EQ subtimesteps we have

$$\Theta^{\mathrm{EQ},(4)} = \begin{pmatrix} \frac{1}{8} & \frac{19}{72} & -\frac{5}{72} & \frac{1}{72} \\ \frac{1}{9} & \frac{4}{9} & \frac{1}{9} & 0 \\ \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{1}{8} \end{pmatrix}$$

resulting in

$$R_4^{\mathrm{EQ}}(z) = \frac{\sum_{j=0}^{10} d_j z^j}{1536(-36+17z)^3(-3+2z)^3(-1+z)^4},$$

where

$$\begin{aligned} &d_0 = 1934917632, && d_1 = -12415721472, && d_2 = 3402678067, \\ &d_3 = -51295431168, && d_4 = 45088151040, && d_5 = -22031034912, \\ &d_6 = 4329437784, && d_7 = 82352116, && d_8 = -534268140, \\ &d_9 = 64784148, && d_{10} = 1805344. \end{aligned}$$

On the other hand, for GL and $p = 4$ we use $\Theta^{\mathrm{GL},(4)} = \Theta^{\mathrm{GL},(3)}$ with $K = 4$ and $M = 2$, obtaining a rational function with a polynomial of degree 7 in the numerator and denominator, which can be represented by

$$R_4^{\mathrm{GL}}(z) = \frac{\sum_{j=0}^{7} c_j z^j}{(7z\sqrt{5}+5z-60)^3(7z\sqrt{5}+31z-60)^4},$$

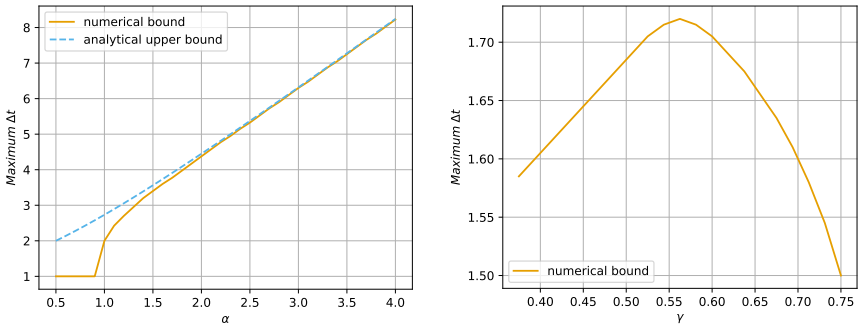
where

$$\begin{aligned} &c_0 = -279936 \cdot 10^7, && c_1 = (1982880\sqrt{5}+5062176) \cdot 10^6, \\ &c_2 = (-28409616\sqrt{5}-58953744) \cdot 10^5, && c_3 = (157481496\sqrt{5}+347034456) \cdot 10^4, \\ &c_4 = -262068264000\sqrt{5} - 617156712000, && c_5 = -55771610400\sqrt{5} - 129811572000, \\ &c_6 = 13763385600\sqrt{5} + 34116840000, && c_7 = 1038579760\sqrt{5} + 2083625200. \end{aligned}$$

For higher order and other schemes, we refer to the Maple code in the reproducibility repository [10].

4 Numerical Comparison

In this section, we compare the numerical bound Δt_0 for Δt not to be oscillating [19] with the necessary condition given by the Lyapunov stability function derived following [8]. The Julia Jupyter notebook used to compute the numerical bound and the Maple notebook where the Lyapunov stability functions are computed are available in the reproducibility repository [10]. Those notebooks can be used also to compute the bounds for different parameters of the presented schemes that could not fit in this work.



(a) MPRK22(α) (b) MPRK43(γ), analytical upper bound 2.35

Fig. 1 Plot of numerical bound for Δt (orange) and Lyapunov stability Δt bound (10) (blue) for the MPRK22(α) and MPRK43(γ) families of schemes

In Figure 1(a), we show the two bounds on Δt for MPRK(2,2,α) [2] varying α. We observe that there is a very good agreement between the two conditions for $\alpha > 1.5$, while for smaller values the error is bounded by $\sqrt{3}$. For the MPRK(4,3,γ) [14] we observe that the numerical bound in Figure 1(b) is not as close as before to the Lyapunov stability bound 2.35 (independently on γ), but still it is giving an indication of the magnitude of the bound.

In Tables 2(a) and 2(b), we write the numerical Δt bound and the necessary condition given by the Lyapunov stability function in Theorem 1 for EQ and GL MPDeC, respectively. Here, we notice very different behaviors between EQ and GL MPDeC. In the EQ case, the bounds are widely varying across different orders of accuracy, in the numerical simulations, while for the theoretical bound, we get very large constraints that are not very useful. On the other side, for GL, the numerical bounds converge very quickly to 1 as the order increases. The Lyapunov stability function leads to a not so sharp bound, but much closer to the numerical one.

p	num. Δt_0	Lyap. Δt_0
1	∞	∞
2	2.0	2.73
3	1.19	3.31
4	1.11	3.83
5	1.07	4.19
6	1.04	∞
7	1.04	∞
8	1.37	∞
9	6.96	∞

(a) MPDeC EQ

p	num. Δt_0	Lyap. Δt_0
1	∞	∞
2	2.0	2.73
3	1.19	3.31
4	1.07	3.62
5	1.04	3.74
6	1.0	4.06
7	1.0	4.47
8	1.0	5.03
9	1.0	20.1

(b) MPDeC GL

Method	num. Δt_0	Lyap. Δt_0
SSPMPRK(4,3)	1.31	2.15
MPRK(3,2)	16.56	∞
MPRK(4,3,2,0.6)	1.89	3.07
MPRK(4,3,0.9,0.5)	1.59	2.82
MPRK(4,3,0.5,0.7)	1.74	2.00
MPRK(4,3,3,$\frac{7}{15}$)	5.37	5.62
SSPMPRK(2,2,0,1)	2	2.73
SSPMPRK(2,2,0,2)	4.36	4.45
SSPMPRK(2,2,0.4,1)	1.27	2.14
SSPMPRK(2,2,0.1,4)	2.10	2.37

(c) Other schemes

Fig. 2 Numerical bound for Δt and Lyapunov stability function Δt bound for various schemes

In Table 2(c), we summarize the results for a selection of other schemes for various parameters. In all cases, we observe, as predicted by Theorem 1, that the numerical bound is smaller than the Lyapunov stability function bound. The discrepancy between the two approaches vary a lot between different schemes and even between different parameters of the same method family, as already observed for MPRK(2,2,α). We observe, in general, lower discrepancy for second order schemes, e.g. SSPMPRK(2,2,0,2) and SSPMPRK(2,2,0.1,4), and higher discrepancy for higher order schemes, e.g. SSPMPRK(4,3) and MPRK(4,3,2,0.6). A special remark on MPRK(3,2) is necessary, as it is the second order scheme with the largest Δt_0. Its numerical bound is very large ≈ 16.5, while there is no Lyapunov stability function bound. This shows, again, that this scheme performs very robustly in these simulations.

5 Conclusion

We have shown that the oscillations that modified Patankar schemes show in two–dimensional systems are linked to the Lyapunov stability function. In particular, it is necessary that the Lyapunov stability function is nonnegative to have an oscillations–free method. In particular, these conditions are verified for $\Delta t \leq \Delta t_0$, where Δt_0 depends on the scheme. We validated the theoretical results with many numerical tests showing that the bound coming from the Lyapunov stability function is always larger than the numerical one.

The found results are useful to choose the time step to avoid oscillations. In many situations, the theoretical bound and the numerical one are actually very close and this gives an indication on how to adopt the time step. Furthermore, there are still open questions on the behavior of MP schemes, in particular for hyperbolic problems, where the positivity of various physical quantities is of paramount importance. We plan to extend this work to a stability analysis of fully discrete MP schemes hoping to find connections with the found oscillations bounds. Furthermore, it is of interest to investigate Lyapunov stability properties in the context of partial differential equations.

Acknowledgements

The author Th. Izgin gratefully acknowledges the financial support by the Deutsche Forschungsgemeinschaft (DFG) through grant ME 1889/10-1. P. Öffner was supported by the Gutenberg Research College, JGU Mainz. D. Torlo (Sissa, Italy) was supported by a SISSA Mathematical Fellowship.

References

1. R. Abgrall, *High order schemes for hyperbolic problems using globally continuous approximation and avoiding mass matrices*, J. Sci. Comput., 73 (2017), pp. 461–494.
2. H. Burchard, E. Deleersnijder, and A. Meister, *A high-order conservative Patankar-type discretisation for stiff systems of production–destruction equations*, Appl. Numer. Math., 47 (2003), pp. 1–30.
3. M. Ciallella, L. Micalizzi, P. Öffner, and D. Torlo, *An arbitrary high order and positivity preserving method for the shallow water equations*, Comput. Fluids, 247 (2022), p. 21. Id/No 105630.
4. A. Dutt, L. Greengard, and V. Rokhlin, *Spectral deferred correction methods for ordinary differential equations*, BIT, 40 (2000), pp. 241–266.
5. J. Huang, T. Izgin, S. Kopecz, A. Meister, and C.-W. Shu, *On the stability of strong-stability-preserving modified Patankar Runge-Kutta schemes*, https://arxiv.org/abs/2205.01488, (2022).
6. J. Huang and C.-W. Shu, *Positivity-preserving time discretizations for production-destruction equations with applications to non-equilibrium flows*, J. Sci. Comput., 78 (2019), pp. 1811–1839.
7. 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*, J. Sci. Comput., 79 (2019), pp. 1015–1056.
8. 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*, ESAIM: M2AN, 56 (2022), pp. 1053–1080.
9. ———, *On the stability of unconditionally positive and linear invariants preserving time integration schemes*, https://arxiv.org/abs/2202.11649, (2022).
10. T. Izgin, P. Öffner, and D. Torlo, *Modified Patankar: Oscillations and Lyapunov Stability (code)*. `https://github.com/accdavlo/Modified-Patankar-Oscillations-and-Lyapunov-Stability`, December 2022.
11. T. Izgin and P. Öffner, *On the stability of modified Patankar methods*, https://arxiv.org/abs/2206.07371, (2022).
12. S. Kopecz and A. Meister, *On order conditions for modified Patankar-Runge-Kutta schemes*, Appl. Numer. Math., 123 (2018), pp. 159–179.
13. S. Kopecz and A. Meister, *Unconditionally positive and conservative third order modified Patankar-Runge-Kutta discretizations of production-destruction systems*, BIT, 58 (2018), pp. 691–728.
14. ———, *On the existence of three-stage third-order modified Patankar-Runge-Kutta schemes*, Numer. Algorithms, 81 (2019), pp. 1473–1484.
15. A. Meister and S. Ortleb, *On unconditionally positive implicit time integration for the DG scheme applied to shallow water flows*, Int. J. Numer. Methods Fluids, 76 (2014), pp. 69–94.
16. P. Öffner and D. Torlo, *Arbitrary high-order, conservative and positivity preserving Patankar-type deferred correction schemes*, Applied Numerical Mathematics, (2020).
17. S. Patankar, *Numerical heat transfer and fluid flow*, CRC press, 1980.
18. A. Stuart and A. R. Humphries, *Dynamical systems and numerical analysis*, vol. 2, Cambridge University Press, 1998.
19. D. Torlo, P. Öffner, and H. Ranocha, *Issues with positivity-preserving Patankar-type schemes*, Appl. Numer. Math., 182 (2022), pp. 117–147.