

## ON LYAPUNOV STABILITY OF POSITIVE AND CONSERVATIVE TIME INTEGRATORS AND APPLICATION TO SECOND ORDER MODIFIED PATANKAR–RUNGE–KUTTA SCHEMES

THOMAS IZGIN\*, STEFAN KOPECZ AND ANDREAS MEISTER

**Abstract.** Since almost twenty years, modified Patankar–Runge–Kutta (MPRK) methods have proven to be efficient and robust numerical schemes that preserve positivity and conservativity of the production-destruction system irrespectively of the time step size chosen. Due to these advantageous properties they are used for a wide variety of applications. Nevertheless, until now, an analytic investigation of the stability of MPRK schemes is still missing, since the usual approach by means of Dahlquist’s equation is not feasible. Therefore, we consider a positive and conservative 2D test problem and provide statements usable for a stability analysis of general positive and conservative time integrator schemes based on the center manifold theory. We use this approach to investigate the Lyapunov stability of the second order MPRK22( $\alpha$ ) and MPRK22ncs( $\alpha$ ) schemes. We prove that MPRK22( $\alpha$ ) schemes are unconditionally stable and derive the stability regions of MPRK22ncs( $\alpha$ ) schemes. Finally, numerical experiments are presented, which confirm the theoretical results.

**Mathematics Subject Classification.** 65L05, 65L06, 65L20.

Received November 22, 2021. Accepted March 22, 2022.

### 1. INTRODUCTION

In recent years, there has been a strong interest in the development of numerical schemes that preserve properties of the solutions of differential equations. Modified Patankar–Runge–Kutta (MPRK) methods, see [7, 17, 18, 22, 23, 30], guarantee positivity and conservativity of the numerical solution of positive and conservative production-destruction systems (PDS). For other recent approaches which facilitate positive and conservative numerical approximations, we refer to [1, 2, 5, 26, 29].

A PDS

$$\mathbf{y}' = \mathbf{P}(\mathbf{y}) - \mathbf{D}(\mathbf{y}), \quad \mathbf{y}(0) = \mathbf{y}^0,$$

with  $\mathbf{y} = (y_1, \dots, y_N)^T$  and  $\mathbf{P}, \mathbf{D} \geq \mathbf{0}$ , is called positive if  $\mathbf{y}^0 > \mathbf{0}$  implies  $\mathbf{y}(t) > \mathbf{0}$  for all times  $t > 0$  and is called conservative if  $\sum_{i=1}^N y_i(t) = \sum_{i=1}^N y_i^0$  for all times  $t > 0$ . Conditions which ensure the positivity of a PDS are given in [13]. A conservative PDS can always be written in the form

---

*Keywords and phrases.* Modified Patankar–Runge–Kutta schemes, production-destruction systems, unconditionally positive and conservative schemes, Lyapunov stability analysis, center manifold theorem for maps.

Department of Mathematics, University of Kassel, Kassel, Germany.

\*Corresponding author: [izgin@mathematik.uni-kassel.de](mailto:izgin@mathematik.uni-kassel.de)

$$y'_i = \sum_{j=1}^N (p_{ij}(\mathbf{y}) - d_{ij}(\mathbf{y})) \quad \text{with} \quad p_{ij}(\mathbf{y}) = d_{ji}(\mathbf{y}) \quad \text{and} \quad p_{ij}(\mathbf{y}), d_{ij}(\mathbf{y}) \geq 0$$

for all  $\mathbf{y} \geq \mathbf{0}$  and  $i, j = 1, \dots, N$  in which  $p_{ij}(\mathbf{y})$  refers to a production term of the  $i$ th equation with corresponding destruction term  $d_{ji}(\mathbf{y})$  in the  $j$ th equation. Analogously,  $d_{ij}(\mathbf{y})$  denotes a destruction term of equation  $i$  with associated production term  $p_{ji}(\mathbf{y})$  in equation  $j$ . In summary, the solution of a positive and conservative PDS remains positive for all times  $t > 0$  and the sum of the solutions components remains constant for all times  $t > 0$ .

Originally introduced in [7], there has been a considerable interest in the development of MPRK schemes in recent years. In [22–24], MPRK schemes of second and third order were introduced. These were generalized in the context of SSP Runge–Kutta methods in [17, 18] and applied to solve reactive Euler equations. In [30], the idea of [7] was used to develop mPDeC schemes, which are MPRK schemes of arbitrary order based on deferred correction schemes. All these schemes are unconditionally positive and conservative and have proven their efficiency and robustness while integrating stiff PDS.

MPRK schemes have been used in a wide range of applications. The first order modified Patankar–Euler scheme, introduced in [7], is used in a global ocean mercury model with a methylation cycle [34]. The second order MPRK scheme of [7] is applied to an ecosystem model for the simulation of the cyanobacteria life cycle [15, 16] or that of dinoflagellates [36]. In [3, 4, 27], this scheme is also used to model the phosphorus cycle in rivers and lakes. Moreover this scheme was found to be beneficial when applied to NPZD-models in [8] and is also implemented in the General Ocean Turbulence Model (GOTM) [9]. Further applications can be found in the context of magneto-thermal winds [14] or warm-hot intergalactic mediums [21].

Often MPRK schemes are used within a splitting ansatz as a time integrator for the reactive part of the considered system of partial differential equations in order to avoid additional time step restrictions arising from stiff reaction terms. In [11], mPDeC schemes are used as time integrators for the shallow water equations to ensure unconditionally positivity of the water height. In [33], it was demonstrated that the second order MPRK scheme of [7] surpasses standard Runge–Kutta and Rosenbrock methods for the solution of conservative biochemical models in performance. This was also confirmed in [6] where the Brusselator PDS was solved with different time integration methods. In [37], a third order MPRK scheme from [23] was successfully used in a high-order operator-splitting method for the numerical solution of the SIR epidemic model.

To the authors knowledge, no general stability analysis of MPRK schemes has been carried out so far. There are several reasons for the lack of such an analysis. First of all, unlike Runge–Kutta methods, MPRK schemes cannot be applied to the scalar Dahlquist equation

$$y' = \lambda y, \quad \lambda \in \mathbb{C}^-, \quad (1.1)$$

since it is unclear how to treat the complex term  $\lambda y$  in the production-destruction setting. This issue can be handled by choosing  $\lambda \in \mathbb{R}^-$  and considering the system  $y'_1 = \lambda y_1$ ,  $y'_2 = -\lambda y_1$ . For this system the first order MPRK scheme of [7] is equivalent to the L-stable backward Euler method. Also second order MPRK schemes show an excellent stability behavior when applied to this system, as we show in Section 3.1. Unfortunately, this stability behavior can only be observed for specific MPRK schemes in more general cases, which requires a more detailed analysis.

The scalar Dahlquist equation is so valuable, since it makes a direct stability analysis of the linear system

$$\mathbf{y}' = \mathbf{A}\mathbf{y} \quad (1.2)$$

with  $\mathbf{A} \in \mathbb{R}^{N \times N}$  unnecessary. A Runge–Kutta method applied to (1.2) has the same stability properties as applied to the  $N$  scalar equations (1.1) with  $\lambda$  passing through the  $N$  eigenvalues of  $\mathbf{A}$ , see for instance Chapter 6 of [12].

Since the direct application of MPRK schemes to (1.1) is not possible, stability should be investigated for a linear system (1.2), where we also require the system to be positive and conservative. In [31], it is pointed

out that the inherent nonlinear nature of MPRK schemes makes even a linear stability analysis difficult. To see this, we follow [31] and consider the second order MPRK scheme of [7] applied to a conservative and positive linear PDS of the form (1.2). The resulting scheme is given by

$$\mathbf{y}^{(2)} = \mathbf{y}^n + \Delta t \mathbf{A} \mathbf{y}^{(2)}, \quad \mathbf{y}^{n+1} = \mathbf{y}^n + \frac{\Delta t}{2} \mathbf{A} (\mathbf{W}^n + \mathbf{I}) \mathbf{y}^{n+1}$$

with a positive time step size  $\Delta t$ ,  $\mathbf{W}^n = \text{diag}(y_i^n / y_i^{(2)})$  and  $\mathbf{I}$  denoting the identity matrix in  $\mathbb{R}^{N \times N}$ . Hence, we find  $\mathbf{y}^{n+1} = \mathbf{R}^n \mathbf{y}^n$  with

$$\mathbf{R}^n = \left( \mathbf{I} - \frac{\Delta t}{2} \mathbf{A} (\mathbf{W}^n + \mathbf{I}) \right)^{-1}.$$

This shows that  $\mathbf{y}^{n+1}$  depends nonlinearly on  $\mathbf{y}^n$  even when a linear PDS is considered, which complicates the analysis significantly.

The system matrix  $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{N \times N}$  of a positive and conservative linear PDS written in the form (1.2) must satisfy  $a_{ii} \leq 0$ ,  $a_{ij} \geq 0$  for  $i \neq j$  and  $\sum_{i=1}^N a_{ij} = 0$  for  $j = 1, \dots, N$ . Hence, the system

$$\mathbf{y}' = \mathbf{A} \mathbf{y}, \quad \mathbf{A} = \begin{pmatrix} -a & b \\ a & -b \end{pmatrix}, \quad a, b \geq 0, \quad a + b > 0 \quad (1.3)$$

represents all positive and conservative linear PDS of size  $2 \times 2$ , except the case with  $a = b = 0$ , which needs no stability analysis. In order to see that (1.3) is indeed a PDS, we set

$$p_{12}(\mathbf{y}) = d_{21}(\mathbf{y}) = by_2, \quad p_{21}(\mathbf{y}) = d_{12}(\mathbf{y}) = ay_1, \quad p_{ii}(\mathbf{y}) = d_{ii}(\mathbf{y}) = 0, \quad i \in \{1, 2\} \quad (1.4)$$

and obtain  $y'_1 = p_{12}(\mathbf{y}) - d_{12}(\mathbf{y})$  and  $y'_2 = p_{21}(\mathbf{y}) - d_{21}(\mathbf{y})$ . The eigenvalues of  $\mathbf{A}$  are  $\lambda = -(a + b) < 0$  and 0. Given an initial value  $\mathbf{y}^0 = (y_1^0, y_2^0)^T$ , the solution of the initial value problem associated with (1.3) is

$$\mathbf{y}(t) = \frac{y_1^0 + y_2^0}{a + b} \begin{pmatrix} b \\ a \end{pmatrix} + \frac{ay_1^0 - by_2^0}{a + b} \begin{pmatrix} 1 \\ -1 \end{pmatrix} e^{\lambda t} = \frac{1}{a + b} \begin{pmatrix} b + ae^{\lambda t} & b - be^{\lambda t} \\ a - ae^{\lambda t} & a + be^{\lambda t} \end{pmatrix} \mathbf{y}^0. \quad (1.5)$$

Since  $\lambda < 0$  we have  $\mathbf{y}(t) > 0$  for  $t \geq 0$  if  $\mathbf{y}^0 > 0$ , which shows that the PDS (1.3) is positive. Moreover, summation in (1.5) shows  $y_1(t) + y_2(t) = y_1^0 + y_2^0$  for all  $t \geq 0$ , which confirms that the PDS is also conservative.

The system (1.3) is also considered in [20], where it is used to study a linearization of second order MPRK schemes. Section 3 extends the results of [20] to the nonlinear case.

In the following, we introduce a framework to study the Lyapunov stability of positive and conservative time integrators, when applied to (1.3). Within this framework we analyze the stability of the second order MPRK22( $\alpha$ ) and MPRK22ncs( $\alpha$ ) schemes introduced in [22]. If we want to refer to both schemes we use MPRK22 schemes as an abbreviation. The MPRK22 schemes are given by

$$y_i^{(1)} = y_i^n, \quad (1.6a)$$

$$y_i^{(2)} = y_i^n + \alpha \Delta t \sum_{j=1}^N \left( (1 - \gamma) p_{ij}(\mathbf{y}^{(1)}) + p_{ij}(\mathbf{y}^{(1)}) \frac{y_j^{(2)}}{y_j^{(1)}} \gamma - d_{ij}(\mathbf{y}^{(1)}) \frac{y_i^{(2)}}{y_i^{(1)}} \right), \quad (1.6b)$$

$$y_i^{n+1} = y_i^n + \Delta t \sum_{j=1}^N \left( \left( \left( 1 - \frac{1}{2\alpha} \right) p_{ij}(\mathbf{y}^{(1)}) + \frac{1}{2\alpha} p_{ij}(\mathbf{y}^{(2)}) \right) \frac{y_j^{n+1}}{(y_j^{(2)})^{\frac{1}{\alpha}} (y_j^{(1)})^{1-\frac{1}{\alpha}}} \right. \\ \left. - \left( \left( 1 - \frac{1}{2\alpha} \right) d_{ij}(\mathbf{y}^{(1)}) + \frac{1}{2\alpha} d_{ij}(\mathbf{y}^{(2)}) \right) \frac{y_i^{n+1}}{(y_i^{(2)})^{\frac{1}{\alpha}} (y_i^{(1)})^{1-\frac{1}{\alpha}}} \right), \quad (1.6c)$$

for  $i = 1, \dots, N$  with  $\alpha \geq \frac{1}{2}$ . The methods with  $\gamma = 1$  in (1.6b) are called MPRK22( $\alpha$ ) schemes. With  $\gamma = 0$  they are named MPRK22ncs( $\alpha$ ) schemes. For MPRK22( $\alpha$ ) schemes  $\mathbf{y}^{(2)}$  is also conservative, in the sense that  $\sum_{i=1}^N y_i^{(2)} = \sum_{i=1}^N y_i^0$ . This is not the case for MPRK22ncs( $\alpha$ ) methods, where “ncs” is an abbreviation for “non-conservative stages”, and will result in inferior stability properties as shown in Section 3.

The outline of the paper is as follows. In Section 2, we summarize the center manifold theory and prove the main theorem concerning the Lyapunov stability of general positive and conservative time integration schemes. This theorem is used to analyze the local stability of the schemes (1.6) when applied to the positive and conservative linear PDS (1.3) in Section 3. We prove that the MPRK22( $\alpha$ ) methods are unconditionally stable. Whereas, MPRK22ncs( $\alpha$ ) is also unconditionally stable for  $\alpha \geq 1$  and requires time step restrictions in the case  $\alpha < 1$ . Finally, we provide numerical experiments confirming the theoretical results in Section 4.

## 2. CENTER MANIFOLD THEORY AND STABILITY OF POSITIVE AND CONSERVATIVE TIME INTEGRATION SCHEMES

In this section we recall the definitions of stable and asymptotically stable steady state solutions of differential equations and the corresponding definitions for fixed points of iteration schemes. We also recap theorems that are helpful to identify the stability properties of a given fixed point. These show that for hyperbolic fixed points stability is solely determined by the eigenvalues of the Jacobian of the underlying map, which is not true for non-hyperbolic fixed points. The center manifold theory is an important tool to investigate the stability of non-hyperbolic fixed points. Using this theory we present Theorem 2.9, which provides for the first time a criteria to assess the stability of general positive and conservative schemes applied to (1.3).

In the following, we use  $\|\cdot\|$  to represent an arbitrary norm in  $\mathbb{R}^l$  for  $l \in \mathbb{N}$  and  $\mathbf{Df}$  denotes the Jacobian of a map  $\mathbf{f}$ .

**Definition 2.1.** Let  $\mathbf{y}^*$  be a steady state solution of a differential equation  $\mathbf{y}' = \mathbf{f}(\mathbf{y})$ , that is  $\mathbf{f}(\mathbf{y}^*) = \mathbf{0}$ .

- (a) The steady state solution  $\mathbf{y}^*$  is called *Lyapunov stable* if, for any  $\epsilon > 0$ , there exists a  $\delta = \delta(\epsilon) > 0$  such that  $\|\mathbf{y}(0) - \mathbf{y}^*\| < \delta$  implies  $\|\mathbf{y}(t) - \mathbf{y}^*\| < \epsilon$  for all  $t \geq 0$ .
- (b) If in addition to (a), there exists a constant  $c > 0$  such that  $\|\mathbf{y}(0) - \mathbf{y}^*\| < c$  implies  $\|\mathbf{y}(t) - \mathbf{y}^*\| \rightarrow 0$  for  $t \rightarrow \infty$ , we call the steady state solution  $\mathbf{y}^*$  *asymptotically stable*.
- (c) A steady state solution that is not stable is said to be *unstable*.

The PDS (1.3) has infinitely many steady state solutions, since every  $\mathbf{y}^*$  in the nullspace of  $\mathbf{A}$ , that is  $\mathbf{y}^* = \theta(b, a)^T$  with  $\theta \in \mathbb{R}$ , is a steady state of (1.3). In geometrical terms, all steady states lie on the line  $ay_1 - by_2 = 0$  in the  $y_1$ - $y_2$ -coordinate system. With respect to the asymptotic behavior of (1.5), we see

$$\lim_{t \rightarrow \infty} \mathbf{y}(t) = \frac{y_1^0 + y_2^0}{a + b} \begin{pmatrix} b \\ a \end{pmatrix} + \lim_{t \rightarrow \infty} \frac{ay_1^0 - by_2^0}{a + b} \begin{pmatrix} 1 \\ -1 \end{pmatrix} e^{\lambda t} = \frac{y_1^0 + y_2^0}{a + b} \begin{pmatrix} b \\ a \end{pmatrix},$$

since  $\lambda < 0$ . Thereby, given an initial value  $\mathbf{y}^0 = (y_1^0, y_2^0)^T$ , the solution monotonically approaches the steady state

$$\mathbf{y}^* = \frac{y_1^0 + y_2^0}{a + b} \begin{pmatrix} b \\ a \end{pmatrix}$$

along the line  $y_1 + y_2 = y_1^0 + y_2^0$  in the  $y_1$ - $y_2$ -coordinate system. Hence, the steady state solutions of (1.3) cannot be asymptotically stable, as there are infinitely many other steady state solutions in every neighborhood of a steady state. But it can be shown that they are stable in the sense of Definition 2.1 (a), see Theorem 3.23 of [12].

When applied to a differential equation, a time integration scheme should preserve as many properties of the differential equation as possible. In particular, the fixed points of the iteration scheme, should be the steady state solutions of the differential equation with equal stability properties.

**Definition 2.2.** Let  $\mathbf{y}^*$  be a fixed point of an iteration scheme  $\mathbf{y}^{n+1} = \mathbf{g}(\mathbf{y}^n)$ , that is  $\mathbf{y}^* = \mathbf{g}(\mathbf{y}^*)$ .

- (a) The fixed point  $\mathbf{y}^*$  is called *Lyapunov stable* if, for any  $\epsilon > 0$ , there exists a  $\delta = \delta(\epsilon) > 0$  such that  $\|\mathbf{y}^0 - \mathbf{y}^*\| < \delta$  implies  $\|\mathbf{y}^n - \mathbf{y}^*\| < \epsilon$  for all  $n \geq 0$ .
- (b) If in addition to (a), there exists a constant  $c > 0$  such that  $\|\mathbf{y}^0 - \mathbf{y}^*\| < c$  implies  $\|\mathbf{y}^n - \mathbf{y}^*\| \rightarrow 0$  for  $n \rightarrow \infty$ , the fixed point  $\mathbf{y}^*$  is called *asymptotically stable*.
- (c) A fixed point that is not stable is said to be *unstable*.

In the following, we will also briefly speak of stability instead of Lyapunov stability.

Next, we summarize theorems which are helpful to investigate the stability of fixed points of iteration schemes.

**Theorem 2.3** ([35], Thm. 1.3.7). Let  $\mathbf{y}^{n+1} = \mathbf{g}(\mathbf{y}^n)$  be an iteration scheme with fixed point  $\mathbf{y}^*$ . Then

- (a)  $\mathbf{y}^*$  is asymptotically stable if  $|\lambda| < 1$  for all eigenvalues  $\lambda$  of  $\mathbf{Dg}(\mathbf{y}^*)$ .
- (b)  $\mathbf{y}^*$  is unstable if  $|\lambda| > 1$  for one eigenvalue  $\lambda$  of  $\mathbf{Dg}(\mathbf{y}^*)$ .

The above theorem does not give any information if the spectral radius of  $\mathbf{Dg}(\mathbf{y}^*)$  is equal to 1. Hence, it is reasonable to introduce the following definition.

**Definition 2.4** ([35], Def. 1.3.6). A fixed point  $\mathbf{y}^*$  of an iteration scheme  $\mathbf{y}^{n+1} = \mathbf{g}(\mathbf{y}^n)$  is called *hyperbolic* if  $|\lambda| \neq 1$  for all eigenvalues  $\lambda$  of  $\mathbf{Dg}(\mathbf{y}^*)$ . If a fixed point is not hyperbolic, it is called *non-hyperbolic*.

A generalization of Theorem 2.3 is the Hartman–Grobman Theorem, which states that a nonlinear iteration scheme and its linearization share the same behavior near hyperbolic fixed points, see Theorem 1.6.2 of [35] for the precise statement. One can show, that for non-hyperbolic fixed points nonlinear terms have to be taken into account in order to investigate the stability. Thereby, the theory of center manifolds is an important tool and will be explained in the following section.

## 2.1. Center manifold theory

To study the stability of a non-hyperbolic fixed point  $\mathbf{y}^*$  of an iteration scheme with  $\mathcal{C}^2$ -map  $\mathbf{g}$ , we make use of an affine linear transformation<sup>1</sup> to obtain a  $\mathcal{C}^2$ -map  $\mathbf{G}: D \rightarrow \mathbb{R}^N$ , with  $D \subset \mathbb{R}^N$  being a neighborhood of the origin, which has the form

$$\mathbf{G}(\mathbf{w}_1, \mathbf{w}_2) = \begin{pmatrix} \mathbf{U}\mathbf{w}_1 + \mathbf{u}(\mathbf{w}_1, \mathbf{w}_2) \\ \mathbf{V}\mathbf{w}_2 + \mathbf{v}(\mathbf{w}_1, \mathbf{w}_2) \end{pmatrix}, \quad (2.1)$$

with  $\mathbf{w}_1 \in \mathbb{R}^m$ ,  $\mathbf{w}_2 \in \mathbb{R}^l$  and  $m + l = N$ . The square matrices  $\mathbf{U} \in \mathbb{R}^{m \times m}$  and  $\mathbf{V} \in \mathbb{R}^{l \times l}$  are such that  $|\lambda| = 1$  for all eigenvalues  $\lambda$  of  $\mathbf{U}$  and  $|\mu| < 1$  for all eigenvalues  $\mu$  of  $\mathbf{V}$ . The functions  $\mathbf{u}$  and  $\mathbf{v}$  are in  $\mathcal{C}^2$  and  $\mathbf{u}, \mathbf{v}$  as well as their first order derivatives vanish at the origin, that is

$$\mathbf{u}(\mathbf{0}, \mathbf{0}) = \mathbf{0}, \quad \mathbf{D}\mathbf{u}(\mathbf{0}, \mathbf{0}) = \mathbf{0}, \quad \mathbf{v}(\mathbf{0}, \mathbf{0}) = \mathbf{0}, \quad \mathbf{D}\mathbf{v}(\mathbf{0}, \mathbf{0}) = \mathbf{0},$$

where  $\mathbf{0}$  stands for the zero vector or matrix of appropriate size, respectively. In particular, the fixed point  $\mathbf{y}^*$  of  $\mathbf{g}$  is mapped to  $\mathbf{0}$ , which is a fixed point of  $\mathbf{G}$  with equal stability properties as  $\mathbf{y}^*$ .

Hence, it is sufficient to study the stability of the origin with respect to  $\mathbf{G}$ , which is a simplification due to the existence of a center manifold.

<sup>1</sup>See the proof of Theorem 2.9 for the details of this transformation.

**Theorem 2.5** (Center Manifold Theorem). *Let  $\mathbf{G}$  be defined as in (2.1) and*

$$\begin{pmatrix} \mathbf{w}_1^{n+1} \\ \mathbf{w}_2^{n+1} \end{pmatrix} = \mathbf{G}(\mathbf{w}_1^n, \mathbf{w}_2^n) = \begin{pmatrix} \mathbf{U}\mathbf{w}_1^n + \mathbf{u}(\mathbf{w}_1^n, \mathbf{w}_2^n) \\ \mathbf{V}\mathbf{w}_2^n + \mathbf{v}(\mathbf{w}_1^n, \mathbf{w}_2^n) \end{pmatrix}, \quad \begin{pmatrix} \mathbf{w}_1^0 \\ \mathbf{w}_2^0 \end{pmatrix} \in \mathbb{R}^{m+l}. \quad (2.2)$$

- (a) (Existence): *there exists a center manifold for  $\mathbf{G}$ , which is locally representable as the graph of a function  $\mathbf{h}: \mathbb{R}^m \rightarrow \mathbb{R}^l$ . This means, for some  $\epsilon > 0$  there exists a  $\mathcal{C}^1$ -function  $\mathbf{h}: \mathbb{R}^m \rightarrow \mathbb{R}^l$  with  $\mathbf{h}(\mathbf{0}) = \mathbf{0}$  and  $\mathbf{D}\mathbf{h}(\mathbf{0}) = \mathbf{0}$  such that  $\|\mathbf{w}_1^0\| < \epsilon$  and  $(\mathbf{w}_1^1, \mathbf{w}_2^1)^T = \mathbf{G}(\mathbf{w}_1^0, \mathbf{h}(\mathbf{w}_1^0))$  implies  $\mathbf{w}_2^1 = \mathbf{h}(\mathbf{w}_1^1)$ .*
- (b) (Local Attractivity): *let  $(\mathbf{w}_1^n, \mathbf{w}_2^n)^T$ ,  $n \in \mathbb{N}_0$  represent the sequence generated by (2.2). If  $\|\mathbf{w}_1^n\|, \|\mathbf{w}_2^n\| < \epsilon$  for all  $n \in \mathbb{N}_0$ , then the distance of  $(\mathbf{w}_1^n, \mathbf{w}_2^n)$  to the center manifold tends to zero for  $n \rightarrow \infty$ , i.e.  $\|\mathbf{w}_2^n - \mathbf{h}(\mathbf{w}_1^n)\| \rightarrow 0$  for  $n \rightarrow \infty$ .*

*Proof.* See Theorem 2.1 of [25], Theorem 6 of [10], Theorem 4 of [32] for existence and Theorem 2.1 of [25], Chapter V, Theorem 2 of [19] for local attractivity.  $\square$

The existence of a center manifold allows the study of a system with reduced dimensionality to determine the stability properties of the origin. Restricting the iteration (2.2) to the center manifold, i.e.  $\mathbf{w}_2^0 = \mathbf{h}(\mathbf{w}_1^0)$ , gives

$$\mathbf{w}_1^{n+1} = \mathcal{G}(\mathbf{w}_1^n), \quad \mathcal{G}(\mathbf{w}_1) = \mathbf{U}\mathbf{w}_1 + \mathbf{u}(\mathbf{w}_1, \mathbf{h}(\mathbf{w}_1)) \quad (2.3)$$

for  $\|\mathbf{w}_1^n\| < \epsilon$ . The next theorem states that stability of the origin with respect to  $\mathcal{G}$  already implies stability of the origin with respect to  $\mathbf{G}$ .

**Theorem 2.6** ([10], Thm. 8 (Stability)). *Suppose the fixed point  $\mathbf{0} \in \mathbb{R}^m$  of  $\mathcal{G}$  from (2.3) is stable, asymptotically stable or unstable. Then the fixed point  $\mathbf{0} \in \mathbb{R}^N$  of  $\mathbf{G}$  from (2.2) is stable, asymptotically stable or unstable.*

In summary, the stability of a non-hyperbolic fixed point  $\mathbf{y}^* \in \mathbb{R}^N$  of a map  $\mathbf{g}$  can be determined by investigating the fixed point  $\mathbf{0} \in \mathbb{R}^m$  of  $\mathcal{G}$ , which has a lower complexity due to the reduced dimension  $m < N$ .

To actually calculate the center manifold we need to solve

$$(\mathbf{w}_1^1, \mathbf{h}(\mathbf{w}_1^1))^T = \mathbf{G}(\mathbf{w}_1^0, \mathbf{h}(\mathbf{w}_1^0)) = \begin{pmatrix} \mathbf{U}\mathbf{w}_1^0 + \mathbf{u}(\mathbf{w}_1^0, \mathbf{h}(\mathbf{w}_1^0)) \\ \mathbf{V}\mathbf{h}(\mathbf{w}_1^0) + \mathbf{v}(\mathbf{w}_1^0, \mathbf{h}(\mathbf{w}_1^0)) \end{pmatrix},$$

which can be rewritten as

$$\mathbf{h}(\mathbf{U}\mathbf{w}_1^0 + \mathbf{u}(\mathbf{w}_1^0, \mathbf{h}(\mathbf{w}_1^0))) = \mathbf{V}\mathbf{h}(\mathbf{w}_1^0) + \mathbf{v}(\mathbf{w}_1^0, \mathbf{h}(\mathbf{w}_1^0))$$

or

$$\mathbf{h}(\mathbf{G}(\mathbf{w}_1^0, \mathbf{h}(\mathbf{w}_1^0)))_1 = \mathbf{G}(\mathbf{w}_1^0, \mathbf{h}(\mathbf{w}_1^0))_2.$$

The above invariance property offers a way to approximate the center manifold up to an arbitrary order.

**Theorem 2.7** ([10], Thm. 7). *Let  $\mathbf{h}$  be a center manifold for  $\mathbf{G}$  and  $\Phi \in \mathcal{C}^1(\mathbb{R}^m, \mathbb{R}^l)$  with  $\Phi(\mathbf{0}) = \mathbf{0}$  and  $\mathbf{D}\Phi(\mathbf{0}) = \mathbf{0}$ . If*

$$\Phi(\mathbf{G}(\mathbf{w}_1, \Phi(\mathbf{w}_1)))_1 - \mathbf{G}(\mathbf{w}_1, \Phi(\mathbf{w}_1))_2 = \mathcal{O}(\|\mathbf{w}_1\|^q)$$

*as  $\mathbf{w}_1 \rightarrow \mathbf{0}$  for some  $q > 1$ , then  $\mathbf{h}(\mathbf{w}_1) = \Phi(\mathbf{w}_1) + \mathcal{O}(\|\mathbf{w}_1\|^q)$  as  $\mathbf{w}_1 \rightarrow \mathbf{0}$ .*

Now, we consider a general positive and conservative iteration scheme  $\mathbf{y}^{n+1} = \mathbf{g}(\mathbf{y}^n)$ , i.e.  $\mathbf{y}^n > \mathbf{0}$  for all  $n \in \mathbb{N}$  if  $\mathbf{y}^0 > \mathbf{0}$  and  $\|\mathbf{y}^{n+1}\|_1 = \|\mathbf{y}^n\|_1$ , in two dimensions. We further assume that all fixed points of the iteration scheme are located on a line through the origin, which is the case for the steady states of (1.3). Under these circumstances, Theorem 2.9 gives a sufficient condition for stability of the iteration scheme based on the eigenvalues of the corresponding Jacobian  $\mathbf{D}\mathbf{g}(\mathbf{y}^*)$ . Furthermore, the theorem states that stability implies convergence towards a fixed point with equal 1-norm.

The following lemma is used in the proof of Theorem 2.9.

**Lemma 2.8.** Let  $\mathbf{y}^0 = (y_1^0, y_2^0)^T > \mathbf{0}$  and  $L = \{\mathbf{y} \in \mathbb{R}_{\geq 0}^2 \mid \|\mathbf{y}\|_1 = \|\mathbf{y}^0\|_1\}$ , then

$$L = \{\mathbf{y} \in \mathbb{R}^2 \mid \mathbf{y} = \mathbf{y}^0 + s\bar{\mathbf{y}}, -y_1^0 < s < y_2^0\},$$

where  $\bar{\mathbf{y}} = (1, -1)^T$ . If in addition,  $\mathbf{g}: \mathbb{R}_{\geq 0}^2 \rightarrow \mathbb{R}_{\geq 0}^2$  is a conservative map, i.e.  $\|\mathbf{g}(\mathbf{y})\|_1 = \|\mathbf{y}\|_1$ , then for every  $\mathbf{y}^0 + t\bar{\mathbf{y}} > \mathbf{0}$  with  $t \in \mathbb{R}$ , there exists a  $s(t)$  with  $-y_1^0 < s(t) < y_2^0$ , such that

$$\mathbf{g}(\mathbf{y}^0 + t\bar{\mathbf{y}}) = \mathbf{y}^0 + s(t)\bar{\mathbf{y}}.$$

*Proof.* Let  $\mathbf{y}^0, \mathbf{y} > \mathbf{0}$ , then the condition  $\|\mathbf{y}\|_1 = \|\mathbf{y}^0\|_1$  is equivalent to  $y_1 + y_2 = y_1^0 + y_2^0$ . Geometrically, this describes a line in the  $y_1$ - $y_2$ -coordinate system with normal vector  $(1, 1)^T$ . A parameter form of this line is given by  $\mathbf{y} = \mathbf{y}^0 + s(t)\bar{\mathbf{y}}$ ,  $s(t) \in \mathbb{R}$ . To ensure  $\mathbf{y} > \mathbf{0}$ , we must have  $-y_1^0 < s(t) < y_2^0$ .

Now, define  $\mathbf{w} = \mathbf{g}(\mathbf{y}^0 + t\bar{\mathbf{y}})$  with  $\mathbf{y}^0 + t\bar{\mathbf{y}} > \mathbf{0}$ . Since  $\|\mathbf{w}\|_1 = \|\mathbf{y}^0 + t\bar{\mathbf{y}}\|_1 = \|\mathbf{y}^0\|_1$ , we have  $\mathbf{g}(\mathbf{y}^0 + t\bar{\mathbf{y}}) = \mathbf{w} = \mathbf{y}^0 + s(t)\bar{\mathbf{y}}$  with  $-y_1^0 < s(t) < y_2^0$ .  $\square$

Next, we present the main theorem of this section, which provides criteria to assess the stability of general positive and conservative schemes. Application of a general positive and conservative scheme to (1.3) results in a nonlinear iteration  $\mathbf{y}^{n+1} = \mathbf{g}(\mathbf{y}^n)$ , for which the steady states  $\mathbf{y}^*$  of (1.3) should be non-hyperbolic fixed points of  $\mathbf{g}$ . The theorem shows that even in this nonlinear case the investigation of the eigenvalues of the Jacobian  $\mathbf{Dg}(\mathbf{y}^*)$  is sufficient to analyze stability. To the authors knowledge there are no similar results focusing on general positive and conservative schemes in the literature even though the statements are of fundamental importance.

**Theorem 2.9.** Let  $\mathbf{g} \in \mathcal{C}^2(\mathbb{R}_{\geq 0}^2)$  with fixed point  $\mathbf{y}^* > \mathbf{0}$ , such that all  $r\mathbf{y}^*$  are fixed points of  $\mathbf{g}$  for all  $r > 0$ . In addition, let the iterates of the iteration scheme  $\mathbf{y}^{n+1} = \mathbf{g}(\mathbf{y}^n)$  satisfy  $\|\mathbf{y}^{n+1}\|_1 = \|\mathbf{y}^n\|_1$  for all  $n \in \mathbb{N}_0$ . Then, the spectrum of the Jacobian  $\mathbf{Dg}(\mathbf{y}^*)$  is  $\sigma(\mathbf{Dg}(\mathbf{y}^*)) = \{1, R\}$  with  $R \in \mathbb{R}$ , and the following statements apply. If  $|R| < 1$ , then

- (a)  $\mathbf{y}^*$  is stable.
- (b) there exists a  $\delta > 0$ , such that  $\|\mathbf{y}^0\|_1 = \|\mathbf{y}^*\|_1$  and  $\|\mathbf{y}^0 - \mathbf{y}^*\| < \delta$  imply  $\mathbf{y}^n \rightarrow \mathbf{y}^*$  as  $n \rightarrow \infty$ .

*Proof.* Throughout this proof, we use the notation  $\bar{\mathbf{y}} = (1, -1)^T$  and  $\mathbf{e}_1 = (1, 0)^T$ ,  $\mathbf{e}_2 = (0, 1)^T$  to denote the standard unit vectors.

First, we compute the eigenvalues and eigenvectors of  $\mathbf{Dg}(\mathbf{y}^*)$ . Since  $\mathbf{g}$  is differentiable in  $\mathbf{y}^*$  the directional derivatives  $\partial_{\mathbf{v}}\mathbf{g}(\mathbf{y}^*) = \mathbf{Dg}(\mathbf{y}^*)\mathbf{v}$  exist for all directions  $\mathbf{v} \in \mathbb{R}^2$ . Hence,

$$\mathbf{Dg}(\mathbf{y}^*)\mathbf{y}^* = \partial_{\mathbf{y}^*}\mathbf{g}(\mathbf{y}^*) = \lim_{h \rightarrow 0} \frac{1}{h}(\mathbf{g}(\mathbf{y}^* + h\mathbf{y}^*) - \mathbf{g}(\mathbf{y}^*)) = \lim_{h \rightarrow 0} \frac{1}{h}(\mathbf{g}((1+h)\mathbf{y}^*) - \mathbf{y}^*).$$

As  $r\mathbf{y}^* > \mathbf{0}$  is a fixed point of  $\mathbf{g}$  for all  $r > 0$ , we see

$$\mathbf{Dg}(\mathbf{y}^*)\mathbf{y}^* = \lim_{h \rightarrow 0} \frac{1}{h}((1+h)\mathbf{y}^* - \mathbf{y}^*) = \mathbf{y}^*.$$

Thus,  $\mathbf{y}^*$  is an eigenvector of  $\mathbf{Dg}(\mathbf{y}^*)$  with associated eigenvalue 1. To compute the other eigenvalue and eigenvector, we consider the directional derivative

$$\mathbf{Dg}(\mathbf{y}^*)\bar{\mathbf{y}} = \partial_{\bar{\mathbf{y}}}\mathbf{g}(\mathbf{y}^*) = \lim_{h \rightarrow 0} \frac{1}{h}(\mathbf{g}(\mathbf{y}^* + h\bar{\mathbf{y}}) - \mathbf{g}(\mathbf{y}^*)). \quad (2.4)$$

Since  $\|\mathbf{g}(\mathbf{y})\|_1 = \|\mathbf{y}\|_1$ , we can use Lemma 2.8 to see

$$\mathbf{g}(\mathbf{y}^* + h\bar{\mathbf{y}}) = \mathbf{y}^* + s(h)\bar{\mathbf{y}},$$

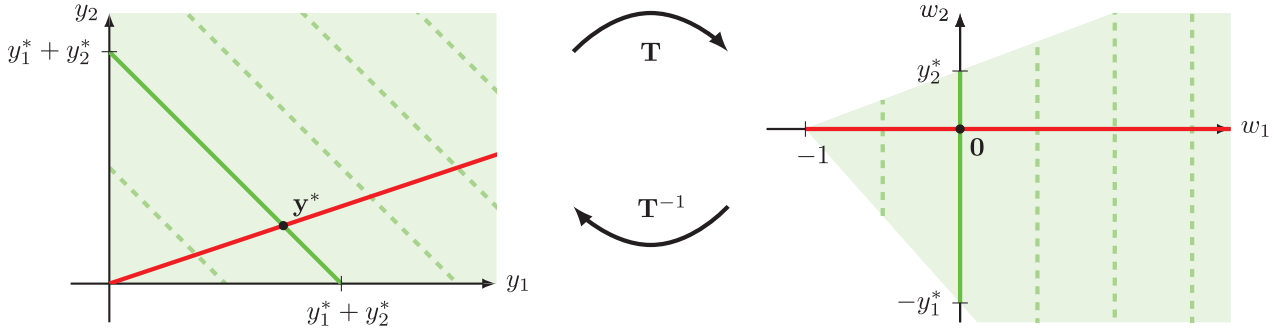


FIGURE 1. Visualisation of the affine transformation  $\mathbf{T}$ . The red and green line segments are mapped onto each other.

for sufficiently small  $h$  and some function  $s: \mathbb{R} \rightarrow \mathbb{R}$ ,  $h \mapsto s(h)$ . Inserting this into (2.4) yields

$$\mathbf{Dg}(\mathbf{y}^*)\bar{\mathbf{y}} = \lim_{h \rightarrow 0} \frac{1}{h} (\mathbf{y}^* + s(h)\bar{\mathbf{y}} - \mathbf{y}^*) = \left( \lim_{h \rightarrow 0} \frac{s(h)}{h} \right) \bar{\mathbf{y}}.$$

The above limit exists, since  $\mathbf{g}$  is differentiable in  $\mathbf{y}^*$ . Setting  $R = \lim_{h \rightarrow 0} \frac{s(h)}{h} \in \mathbb{R}$ , we see  $\mathbf{Dg}(\mathbf{y}^*)\bar{\mathbf{y}} = R\bar{\mathbf{y}}$ , i.e.  $\bar{\mathbf{y}}$  is an eigenvector of  $\mathbf{Dg}(\mathbf{y}^*)$  with corresponding eigenvalue  $R$ . Hence, the spectrum of  $\mathbf{Dg}(\mathbf{y}^*)$  is given by  $\sigma(\mathbf{Dg}(\mathbf{y}^*)) = \{1, R\}$ . Introducing the matrix of eigenvectors

$$\mathbf{S} = (\mathbf{y}^* \bar{\mathbf{y}}), \quad (2.5)$$

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

$$\mathbf{S}^{-1} \mathbf{Dg}(\mathbf{y}^*) \mathbf{S} = \text{diag}(1, R), \quad (2.6)$$

where  $\text{diag}(\mathbf{y}) \in \mathbb{R}^{2 \times 2}$  denotes the diagonal matrix with  $(\text{diag}(\mathbf{y}))_{ii} = y_i$  for  $i = 1, 2$ .

- (a) In this part we assume  $|R| < 1$  and use the center manifold Theorem 2.5 (a) in combination with Theorem 2.6 to conclude that this implies that  $\mathbf{y}^*$  is a stable fixed point. The theorem requires a map  $\mathbf{G}$  of form (2.1), which shall be obtained from  $\mathbf{g}$  by means of an affine linear transformation.

We consider the affine transformation  $\mathbf{T}: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ ,  $\mathbf{y} \mapsto \mathbf{w} = \mathbf{T}(\mathbf{y}) = \mathbf{S}^{-1}(\mathbf{y} - \mathbf{y}^*)$ , where  $\mathbf{S}$  is given in (2.5) and the inverse transformation  $\mathbf{T}^{-1}$  is given by  $\mathbf{T}^{-1}(\mathbf{w}) = \mathbf{S}\mathbf{w} + \mathbf{y}^*$ . By construction, the line segment of fixed points  $\{\mathbf{y} \in \mathbb{R}_{>0}^2 \mid \mathbf{y} = r\mathbf{y}^*, r > 0\}$  is mapped onto the  $w_1$ -axis, as  $\mathbf{T}(r\mathbf{y}^*) = \mathbf{S}^{-1}(r\mathbf{y}^* - \mathbf{y}^*) = (r - 1)\mathbf{S}^{-1}\mathbf{y}^* = (r - 1)\mathbf{e}_1$ . The line segment  $\{\mathbf{y} \in \mathbb{R}_{>0}^2 \mid \|\mathbf{y}\|_1 = \|\mathbf{y}^*\|_1\} = \{\mathbf{y} \in \mathbb{R}^2 \mid \mathbf{y} = \mathbf{y}^* + s\bar{\mathbf{y}}, -y_1^* < s < y_2^*\}$ , see Lemma 2.8, is mapped onto the  $w_2$ -axis, since  $\mathbf{T}(\mathbf{y}^* + s\bar{\mathbf{y}}) = \mathbf{S}^{-1}(\mathbf{y}^* + s\bar{\mathbf{y}} - \mathbf{y}^*) = s\mathbf{S}^{-1}\bar{\mathbf{y}} = s\mathbf{e}_2$ . In particular,  $\mathbf{y}^*$  is mapped to the origin and parallel lines are mapped onto parallel lines, since  $\mathbf{T}$  is affine. See Figure 1 for a sketch of this situation.

Now we define

$$\mathbf{G}: \mathbf{T}(\mathbb{R}_{>0}^2) \rightarrow \mathbf{T}(\mathbb{R}_{>0}^2), \quad \mathbf{G}(\mathbf{w}) = \mathbf{T}(\mathbf{g}(\mathbf{T}^{-1}(\mathbf{w}))) \quad (2.7)$$

and observe that the origin is a fixed point of  $\mathbf{G}$ . To represent  $\mathbf{G}$  in the form (2.1), we write  $\mathbf{g}$  as

$$\mathbf{g}(\mathbf{y}) = \mathbf{g}(\mathbf{y}^*) + \mathbf{Dg}(\mathbf{y}^*)(\mathbf{y} - \mathbf{y}^*) + \mathbf{Q}_{\mathbf{y}^*}(\mathbf{y}) = \mathbf{y}^* + \mathbf{Dg}(\mathbf{y}^*)(\mathbf{y} - \mathbf{y}^*) + \mathbf{Q}_{\mathbf{y}^*}(\mathbf{y}), \quad (2.8)$$

where the remainder  $\mathbf{Q}_{\mathbf{y}^*}(\mathbf{y})$  can be written in Lagrange form

$$(\mathbf{Q}_{\mathbf{y}^*}(\mathbf{y}))_i = \frac{1}{2}(\mathbf{y} - \mathbf{y}^*)^T \mathbf{H}g_i(\mathbf{y}^* + c(\mathbf{y} - \mathbf{y}^*))(\mathbf{y} - \mathbf{y}^*), \quad i = 1, 2, \quad (2.9)$$

with  $\mathbf{H}g_i$  denoting the Hessian of  $g_i$  and  $c \in (0, 1)$  depends on  $\mathbf{y}$  and  $\mathbf{y}^*$ . This is possible because  $\mathbf{g}$  is assumed to be in  $\mathcal{C}^2$  on the convex set  $\mathbb{R}_{>0}^2$ . In particular, we have

$$\mathbf{Q}_{\mathbf{y}^*}(\mathbf{y}^*) = \mathbf{0}, \quad \mathbf{D}\mathbf{Q}_{\mathbf{y}^*}(\mathbf{y}^*) = \mathbf{0}. \quad (2.10)$$

By inserting (2.8) in (2.7) we obtain

$$\mathbf{G}(\mathbf{w}) = \mathbf{S}^{-1}(\mathbf{D}\mathbf{g}(\mathbf{y}^*)(\mathbf{T}^{-1}(\mathbf{w}) - \mathbf{y}^*) + \mathbf{Q}_{\mathbf{y}^*}(\mathbf{T}^{-1}(\mathbf{w}))) = \mathbf{S}^{-1}\mathbf{D}\mathbf{g}(\mathbf{y}^*)\mathbf{S}\mathbf{w} + \mathbf{S}^{-1}\mathbf{Q}_{\mathbf{y}^*}(\mathbf{T}^{-1}(\mathbf{w}))$$

and using (2.6), we see

$$\mathbf{G}(\mathbf{w}) = \text{diag}(1, R)\mathbf{w} + \mathbf{S}^{-1}\mathbf{Q}_{\mathbf{y}^*}(\mathbf{T}^{-1}(\mathbf{w})). \quad (2.11)$$

With  $\mathbf{w} = (w_1, w_2)^T$  this can be rewritten as

$$\mathbf{G}(w_1, w_2) = \begin{pmatrix} Uw_1 + u(w_1, w_2) \\ Vw_2 + v(w_1, w_2) \end{pmatrix} \quad (2.12)$$

where

$$U = 1, \quad V = R, \quad u(w_1, w_2) = (\mathbf{S}^{-1}\mathbf{Q}_{\mathbf{y}^*}(\mathbf{T}^{-1}(w_1, w_2)))_1, \quad v(w_1, w_2) = (\mathbf{S}^{-1}\mathbf{Q}_{\mathbf{y}^*}(\mathbf{T}^{-1}(w_1, w_2)))_2. \quad (2.13)$$

The eigenvalue of  $U$  has absolute value 1 and the eigenvalue of  $V$  satisfies  $|R| < 1$ . Furthermore, we conclude from (2.10) that  $u(0, 0) = v(0, 0) = 0$ , since  $\mathbf{Q}_{\mathbf{y}^*}(\mathbf{y}^*) = \mathbf{0}$ , and  $\mathbf{D}u(0, 0) = \mathbf{D}v(0, 0) = 0$ , since  $\mathbf{D}\mathbf{Q}_{\mathbf{y}^*}(\mathbf{y}^*) = \mathbf{0}$ . Thus, (2.11) is of form (2.1).

Now, the center manifold Theorem 2.5 (a) states that for some  $\epsilon > 0$  there exists a  $\mathcal{C}^1$  function  $h: \mathbb{R} \rightarrow \mathbb{R}$  with  $h(0) = 0$  and  $h'(0) = 0$ , such that  $(w_1^1, w_2^1)^T = \mathbf{G}(w_1^0, h(w_1^0))$  and  $|w_1^0| < \epsilon$  implies  $w_2^1 = h(w_1^1)$ . Furthermore,  $\epsilon$  can be further reduced, if necessary, to guarantee  $(w_1, h(w_1)) \in \mathbf{T}(\mathbb{R}_{>0}^2)$  for  $|w_1| < \epsilon$  as  $h$  is continuous with  $h(0) = 0$ . We know from Lemma 2.8, that for all  $\mathbf{y}^0 > \mathbf{0}$  the map  $\mathbf{g}$  is invariant on the line segment  $\{\mathbf{y} \in \mathbb{R}_{>0}^2 \mid \|\mathbf{y}\|_1 = \|\mathbf{y}^0\|_1\} = \{\mathbf{y} \in \mathbb{R}^2 \mid \mathbf{y} = \mathbf{y}^0 + s\bar{\mathbf{y}}, -y_1^0 < s < y_2^0\}$ . The transformation  $\mathbf{T}$  maps such a line segment onto a segment of a line parallel to the  $w_2$ -axes, on which  $\mathbf{G}$  must then be invariant. To verify this, we use Lemma 2.8 and compute

$$\begin{aligned} \mathbf{G}(w_1, w_2) &= \mathbf{G}(w_1\mathbf{e}_1 + w_2\mathbf{e}_2) = \mathbf{T}(\mathbf{g}(\mathbf{T}^{-1}(w_1\mathbf{e}_1 + w_2\mathbf{e}_2))) = \mathbf{T}(\mathbf{g}(w_1\mathbf{S}\mathbf{e}_1 + w_2\mathbf{S}\mathbf{e}_2 + \mathbf{y}^*)) \\ &= \mathbf{T}(\mathbf{g}((1 + w_1)\mathbf{y}^* + w_2\bar{\mathbf{y}})) = \mathbf{T}((1 + w_1)\mathbf{y}^* + s(w_2)\bar{\mathbf{y}}) \\ &= \mathbf{S}^{-1}(w_1\mathbf{y}^* + s(w_2)\bar{\mathbf{y}}) = w_1\mathbf{e}_1 + s(w_2)\mathbf{e}_2 = (w_1, s(w_2))^T \end{aligned}$$

for some suitable  $s = s(w_2) \in \mathbb{R}$ . In particular, we have shown

$$\mathbf{G}(w_1, w_2)_1 = w_1. \quad (2.14)$$

We can now consider the iteration scheme

$$w_1^{n+1} = \mathcal{G}(w_1^n), \quad \mathcal{G}(w_1) = Uw_1 + u(w_1, h(w_1))$$

for  $|w_1^n| < \epsilon$ , where  $U$  and  $u$  are given in (2.13). According to Theorem 2.6, the fixed point  $\mathbf{0} \in \mathbb{R}^2$  of  $\mathbf{G}$  is stable, if the fixed point  $0 \in \mathbb{R}$  is a stable fixed point of  $\mathcal{G}$ . From (2.12) and (2.14) we see

$$\mathcal{G}(w_1) = \mathbf{G}(w_1, h(w_1))_1 = w_1,$$

which implies  $w_1^n = \mathcal{G}(w_1^{n-1}) = w_1^0$  for all  $n \in \mathbb{N}$  and every  $w_1^0$  with  $|w_1^0| < \epsilon$ . Consequently, for every  $\tilde{\epsilon} > 0$  we define  $\tilde{\delta} = \min(\tilde{\epsilon}, \epsilon)$  to obtain that  $|w_1^0| < \tilde{\delta}$  implies  $|w_1^n| = |w_1^0| < \tilde{\delta} \leq \tilde{\epsilon}$ . Thus, 0 is a stable fixed point

of  $\mathcal{G}$  in the sense of Definition 2.2 (a). Furthermore, by Theorem 2.6 the fixed point  $\mathbf{0} \in \mathbb{R}^2$  of  $\mathbf{G}$  is stable as well.

As a last step, we show that the above implies that  $\mathbf{y}^*$  is a stable fixed point of  $\mathbf{g}$ . We know that  $\mathbf{0}$  is a stable fixed point of the iteration scheme  $\mathbf{w}^{n+1} = \mathbf{G}(\mathbf{w}^n)$ , that is for every  $\epsilon_w > 0$  exists  $\delta_w > 0$  such that  $\|\mathbf{w}^0\| < \delta_w$  implies  $\|\mathbf{w}^n\| < \epsilon_w$ . Now, let  $\epsilon > 0$ ,  $\epsilon_w = \epsilon/\|\mathbf{S}\|$  and  $\delta = \delta_w/\|\mathbf{S}^{-1}\|$ . If  $\|\mathbf{y}^0 - \mathbf{y}^*\| < \delta$ , then  $\|\mathbf{w}^0\| = \|\mathbf{T}(\mathbf{y}^0)\| = \|\mathbf{S}^{-1}(\mathbf{y}^0 - \mathbf{y}^*)\| \leq \|\mathbf{S}^{-1}\| \|\mathbf{y}^0 - \mathbf{y}^*\| < \|\mathbf{S}^{-1}\| \delta = \delta_w$  and consequently  $\|\mathbf{w}^n\| < \epsilon_w$ . Furthermore,  $\mathbf{w}^n = \mathbf{T}(\mathbf{y}^n) = \mathbf{S}^{-1}(\mathbf{y}^n - \mathbf{y}^*)$  is equivalent to  $\mathbf{S}\mathbf{w}^n = \mathbf{y}^n - \mathbf{y}^*$  and hence,  $\|\mathbf{y}^n - \mathbf{y}^*\| \leq \|\mathbf{S}\| \|\mathbf{w}^n\| < \|\mathbf{S}\| \epsilon_w = \epsilon$ . Thus, we have shown that  $\mathbf{y}^*$  is a stable fixed point of the iteration  $\mathbf{y}^{n+1} = \mathbf{g}(\mathbf{y}^n)$ .

- (b) In the following we use that the center manifold is given by  $\{(w_1, w_2) \in \mathbb{R}^2 \mid w_2 = 0, |w_1| < \epsilon\}$ , *i.e.*  $h(w_1) = 0$ , for a sufficiently small  $\epsilon > 0$ . This can be shown with Theorem 2.7 as follows. The function  $\Phi: \mathbb{R} \rightarrow \mathbb{R}$ ,  $\Phi(w_1) = 0$  satisfies  $\Phi(0) = \Phi'(0) = 0$ . Furthermore, all points on the  $w_1$ -axes are fixed points of  $\mathbf{G}$ , since

$$\begin{aligned} \mathbf{G}(w_1, 0) &= \mathbf{G}(w_1 \mathbf{e}_1) = \mathbf{T}(\mathbf{g}(\mathbf{T}^{-1}(w_1 \mathbf{e}_1))) = \mathbf{T}(\mathbf{g}(w_1 \mathbf{S} \mathbf{e}_1 + \mathbf{y}^*)) \\ &= \mathbf{T}(\mathbf{g}((1 + w_1) \mathbf{y}^*)) = \mathbf{T}((1 + w_1) \mathbf{y}^*) \\ &= \mathbf{S}^{-1}(w_1 \mathbf{y}^*) = w_1 \mathbf{e}_1 = (w_1, 0)^T. \end{aligned}$$

Hence, it follows that

$$\Phi(\mathbf{G}(w_1, \Phi(w_1)))_1 - \mathbf{G}(w_1, \Phi(w_1))_2 = -\mathbf{G}(w_1, 0)_2 = 0.$$

By Theorem 2.7  $\Phi$  is an approximation of  $h$  for any order  $q > 1$ . Thus,  $h(w_1) = \Phi(w_1) = 0$  for  $|w_1| < \epsilon$ . Now, we prove that the iteration scheme  $\mathbf{y}^{n+1} = \mathbf{g}(\mathbf{y}^n)$  is locally convergent to  $\mathbf{y}^*$ , if the starting value  $\mathbf{y}^0$  is sufficiently close to  $\mathbf{y}^*$  and satisfies  $\|\mathbf{y}^0\|_1 = \|\mathbf{y}^*\|_1$ . Taking advantage of the transformation  $\mathbf{T}$ , this is equivalent to proving that  $\mathbf{w}^{n+1} = \mathbf{G}(\mathbf{w}^n)$  is locally convergent to  $\mathbf{0}$ , if the starting value  $\mathbf{w}^0$  is sufficiently close to  $\mathbf{0}$  and satisfies  $w_1^0 = 0$ . Moreover, since  $\mathbf{G}$  leaves the first component of its argument fixed, see (2.14), we only need to show  $w_2^n \rightarrow 0$  for  $n \rightarrow \infty$ . According to Theorem 2.5 (b) the distance of  $(w_1^n, w_2^n) \in \mathbb{R}^2$  to the center manifold tends to zero for  $n \rightarrow \infty$ , *i.e.*  $|w_2^n| \rightarrow 0$  for  $n \rightarrow \infty$ , if  $|w_1^n|, |w_2^n| < \epsilon$  for all  $n \in \mathbb{N}_0$  and some sufficiently small  $\epsilon > 0$ . Finally, since the origin is a stable fixed point of  $\mathbf{G}$ , as shown in (a), there exists  $\delta > 0$  such that  $\|\mathbf{w}^0\|_\infty < \delta$  implies  $\|\mathbf{w}^n\|_\infty = \max\{|w_1^n|, |w_2^n|\} < \epsilon$  for all  $n \in \mathbb{N}$ . If necessary,  $\|\mathbf{w}^0\|_\infty < \epsilon$  can be assured by choosing  $\delta < \epsilon$ . Altogether, this proves  $\|\mathbf{w}^n\|_\infty = \max\{|w_1^n|, |w_2^n|\} < \epsilon$  for all  $n \in \mathbb{N}_0$  and thus the assertion.  $\square$

We like to highlight that Theorem 2.9 is valid for general positive and conservative schemes and not restricted to MPRK schemes.

### 3. STABILITY OF MPRK22 SCHEMES

In this section we use Theorem 2.9 to examine the stability properties of MPRK22( $\alpha$ ) and MPRK22ncs( $\alpha$ ) schemes. The main task is to express the schemes in the form  $\mathbf{y}^{n+1} = \mathbf{g}(\mathbf{y}^n)$  and to compute the eigenvalues of the Jacobian  $\mathbf{Dg}(\mathbf{y}^*)$  for a given fixed point  $\mathbf{y}^*$ . We show that MPRK22( $\alpha$ ) schemes for all permissible  $\alpha$  and MPRK22ncs( $\alpha$ ) schemes with  $\alpha \geq 1$  are unconditionally stable. Hence, the iterates of these methods locally converge towards a fixed point with equal 1-norm. Furthermore, for MPRK22ncs( $\alpha$ ) schemes with  $\alpha < 1$  time step restrictions are necessary to achieve the same behavior.

To give a clear representation of the MPRK iterations in the form  $\mathbf{y}^{n+1} = \mathbf{g}(\mathbf{y}^n)$ , we define some auxiliary matrices and functions. First, we split the matrix  $\mathbf{A}$  into a production part  $\mathbf{A}_P$  and a destruction part  $\mathbf{A}_D$ , *i.e.*

$$\mathbf{A} = \mathbf{A}_P - \mathbf{A}_D, \quad \mathbf{A}_P = \mathbf{A} - \text{diag}(\mathbf{A}) = \begin{pmatrix} 0 & b \\ a & 0 \end{pmatrix}, \quad \mathbf{A}_D = -\text{diag}(\mathbf{A}) = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}. \quad (3.1)$$

Based on this splitting we define the matrices

$$\mathbf{B}_\gamma = \gamma(\mathbf{I} - \alpha\Delta t\mathbf{A})^{-1} + (1 - \gamma)(\mathbf{I} + \alpha\Delta t\mathbf{A}_D)^{-1}(\mathbf{I} + \alpha\Delta t\mathbf{A}_P) \quad (3.2)$$

and

$$\mathbf{C}_\gamma = \left(1 - \frac{1}{2\alpha}\right)\mathbf{I} + \left(\frac{1}{2\alpha}\right)\mathbf{B}_\gamma, \quad (3.3)$$

where  $\mathbf{I} \in \mathbb{R}^{2 \times 2}$  denotes the identity matrix. Furthermore, we need the nonlinear functions  $\boldsymbol{\sigma}(\mathbf{y}) = (\sigma_1(\mathbf{y}), \sigma_2(\mathbf{y}))^T$  and  $\boldsymbol{\tau}(\mathbf{y}) = (\tau_1(\mathbf{y}), \tau_2(\mathbf{y}))^T$  with

$$\sigma_i(\mathbf{y}) = (\mathbf{B}_\gamma \mathbf{y})_i^{\frac{1}{\alpha}} (y_i)^{1 - \frac{1}{\alpha}} \quad (3.4)$$

and

$$\tau_i(\mathbf{y}) = \frac{(\mathbf{C}_\gamma \mathbf{y})_i}{\sigma_i(\mathbf{y})} \quad (3.5)$$

for  $i = 1, 2$ .

**Proposition 3.1.** *The MPRK22 schemes (1.6) applied to the test equation (1.3), can be written in the form  $\mathbf{y}^{n+1} = \mathbf{g}(\mathbf{y}^n)$ , where the map  $\mathbf{g}: \mathbb{R}_{>0}^2 \rightarrow \mathbb{R}_{>0}^2$  is given by*

$$\mathbf{g}(\mathbf{y}) = \left( \mathbf{I} + \frac{\Delta t}{1 + \Delta t(a\tau_1(\mathbf{y}) + b\tau_2(\mathbf{y}))} \mathbf{A} \operatorname{diag}(\boldsymbol{\tau}(\mathbf{y})) \right) \mathbf{y}, \quad (3.6)$$

with  $\boldsymbol{\tau}$  defined by (3.5).

*Proof.* We start with (1.6b) to express  $\mathbf{y}^{(2)}$  in terms of  $\mathbf{y}^n$ . Using (1.4) and (1.6a), the stage value (1.6b), when applied to (1.3), is given by

$$\begin{aligned} y_1^{(2)} &= y_1^n + \alpha\Delta t \left( (1 - \gamma)by_2^n + \gamma by_2^{(2)} - ay_1^{(2)} \right), \\ y_2^{(2)} &= y_2^n + \alpha\Delta t \left( (1 - \gamma)ay_1^n + \gamma ay_1^{(2)} - by_2^{(2)} \right). \end{aligned} \quad (3.7)$$

Incorporating the splitting (3.1), the linear system (3.7) becomes

$$\mathbf{y}^{(2)} = \mathbf{y}^n + \alpha\Delta t \left( (1 - \gamma)\mathbf{A}_P \mathbf{y}^n + \gamma\mathbf{A}_P \mathbf{y}^{(2)} - \mathbf{A}_D \mathbf{y}^{(2)} \right)$$

and solving for  $\mathbf{y}^{(2)}$  shows

$$\mathbf{y}^{(2)} = (\mathbf{I} - \alpha\Delta t(\gamma\mathbf{A}_P - \mathbf{A}_D))^{-1}(\mathbf{I} + \alpha\Delta t(1 - \gamma)\mathbf{A}_P)\mathbf{y}^n.$$

In the case  $\gamma = 1$ , we have

$$\mathbf{y}^{(2)} = (\mathbf{I} - \alpha\Delta t\mathbf{A})^{-1}\mathbf{y}^n, \quad (3.8)$$

and for  $\gamma = 0$ , we get

$$\mathbf{y}^{(2)} = (\mathbf{I} + \alpha\Delta t\mathbf{A}_D)^{-1}(\mathbf{I} + \alpha\Delta t\mathbf{A}_P)\mathbf{y}^n. \quad (3.9)$$

Using (3.2), the cases (3.8) and (3.9) can be combined to obtain

$$\mathbf{y}^{(2)} = \mathbf{B}_\gamma \mathbf{y}^n. \quad (3.10)$$

Next, we consider (1.6c) to express  $\mathbf{y}^{n+1}$  in terms of  $\mathbf{y}^n$ . With (1.4) and (3.4) the approximation step (1.6c) reads

$$\begin{aligned} y_1^{n+1} &= y_1^n + \Delta t \left( \left( \left( 1 - \frac{1}{2\alpha} \right) b y_2^n + \left( \frac{1}{2\alpha} \right) b y_2^{(2)} \right) \frac{y_2^{n+1}}{\sigma_2(\mathbf{y}^n)} - \left( \left( 1 - \frac{1}{2\alpha} \right) a y_1^n + \left( \frac{1}{2\alpha} \right) a y_1^{(2)} \right) \frac{y_1^{n+1}}{\sigma_1(\mathbf{y}^n)} \right), \\ y_2^{n+1} &= y_2^n + \Delta t \left( \left( \left( 1 - \frac{1}{2\alpha} \right) a y_1^n + \left( \frac{1}{2\alpha} \right) a y_1^{(2)} \right) \frac{y_1^{n+1}}{\sigma_1(\mathbf{y}^n)} - \left( \left( 1 - \frac{1}{2\alpha} \right) b y_2^n + \left( \frac{1}{2\alpha} \right) b y_2^{(2)} \right) \frac{y_2^{n+1}}{\sigma_2(\mathbf{y}^n)} \right). \end{aligned} \quad (3.11)$$

Here we note that  $\mathbf{y}^n > \mathbf{0}$  implies  $\mathbf{y}^{(2)} > \mathbf{0}$ , see [22], and in addition

$$\sigma_i(\mathbf{y}^n) = (\mathbf{B}_\gamma \mathbf{y}^n)_i^{\frac{1}{\alpha}} (y_i^n)^{1-\frac{1}{\alpha}} = \left( y_i^{(2)} \right)^{\frac{1}{\alpha}} (y_i^n)^{1-\frac{1}{\alpha}} > 0. \quad (3.12)$$

The system (3.11) can be written as

$$\begin{aligned} \mathbf{y}^{n+1} &= \mathbf{y}^n + \Delta t \left( \left( 1 - \frac{1}{2\alpha} \right) \mathbf{A}_P \operatorname{diag}(\mathbf{y}^{n+1}) \operatorname{diag}(\boldsymbol{\sigma}(\mathbf{y}^n))^{-1} \mathbf{y}^n + \left( \frac{1}{2\alpha} \right) \mathbf{A}_P \operatorname{diag}(\mathbf{y}^{n+1}) \operatorname{diag}(\boldsymbol{\sigma}(\mathbf{y}^n))^{-1} \mathbf{y}^{(2)} \right. \\ &\quad \left. - \left( 1 - \frac{1}{2\alpha} \right) \mathbf{A}_D \operatorname{diag}(\mathbf{y}^{n+1}) \operatorname{diag}(\boldsymbol{\sigma}(\mathbf{y}^n))^{-1} \mathbf{y}^n - \left( \frac{1}{2\alpha} \right) \mathbf{A}_D \operatorname{diag}(\mathbf{y}^{n+1}) \operatorname{diag}(\boldsymbol{\sigma}(\mathbf{y}^n))^{-1} \mathbf{y}^{(2)} \right), \end{aligned} \quad (3.13)$$

where  $\operatorname{diag}(\mathbf{y}) \in \mathbb{R}^{2 \times 2}$  denotes the diagonal matrix with  $(\operatorname{diag}(\mathbf{y}))_{ii} = y_i$  for  $i = 1, 2$ . As  $\operatorname{diag}(\mathbf{v})\mathbf{w} = \operatorname{diag}(\mathbf{w})\mathbf{v}$  and  $\operatorname{diag}(\mathbf{v})\operatorname{diag}(\mathbf{w}) = \operatorname{diag}(\mathbf{w})\operatorname{diag}(\mathbf{v})$  for all  $\mathbf{v}, \mathbf{w} \in \mathbb{R}^2$  as well as  $\mathbf{A} = \mathbf{A}_P - \mathbf{A}_D$ , equation (3.13) can be rewritten as

$$\mathbf{y}^{n+1} = \mathbf{y}^n + \Delta t \left( \left( 1 - \frac{1}{2\alpha} \right) \mathbf{A} \operatorname{diag}(\mathbf{y}^n) \operatorname{diag}(\boldsymbol{\sigma}(\mathbf{y}^n))^{-1} + \left( \frac{1}{2\alpha} \right) \mathbf{A} \operatorname{diag}(\mathbf{y}^{(2)}) \operatorname{diag}(\boldsymbol{\sigma}(\mathbf{y}^n))^{-1} \right) \mathbf{y}^{n+1}.$$

Utilizing (3.10), this can be further simplified to

$$\mathbf{y}^{n+1} = \mathbf{y}^n + \Delta t \mathbf{A} \left( \left( 1 - \frac{1}{2\alpha} \right) \operatorname{diag}(\mathbf{y}^n) + \left( \frac{1}{2\alpha} \right) \operatorname{diag}(\mathbf{B}_\gamma \mathbf{y}^n) \right) \operatorname{diag}(\boldsymbol{\sigma}(\mathbf{y}^n))^{-1} \mathbf{y}^{n+1}.$$

Using  $\alpha \operatorname{diag}(\mathbf{v}) + \beta \operatorname{diag}(\mathbf{w}) = \operatorname{diag}(\alpha \mathbf{v} + \beta \mathbf{w})$  for all  $\alpha, \beta \in \mathbb{R}$  and  $\mathbf{v}, \mathbf{w} \in \mathbb{R}^2$  together with (3.3) we see

$$\begin{aligned} \mathbf{y}^{n+1} &= \mathbf{y}^n + \Delta t \mathbf{A} \operatorname{diag} \left( \left( 1 - \frac{1}{2\alpha} \right) \mathbf{y}^n + \left( \frac{1}{2\alpha} \right) \mathbf{B}_\gamma \mathbf{y}^n \right) \operatorname{diag}(\boldsymbol{\sigma}(\mathbf{y}^n))^{-1} \mathbf{y}^{n+1} \\ &= \mathbf{y}^n + \Delta t \mathbf{A} \operatorname{diag}(\mathbf{C}_\gamma \mathbf{y}^n) \operatorname{diag}(\boldsymbol{\sigma}(\mathbf{y}^n))^{-1} \mathbf{y}^{n+1} \end{aligned}$$

or equivalently

$$\mathbf{y}^{n+1} = \mathbf{M}(\mathbf{y}^n)^{-1} \mathbf{y}^n,$$

where

$$\mathbf{M}(\mathbf{y}^n) = \mathbf{I} - \Delta t \mathbf{A} \operatorname{diag}(\mathbf{C}_\gamma \mathbf{y}^n) \operatorname{diag}(\boldsymbol{\sigma}(\mathbf{y}^n))^{-1}. \quad (3.14)$$

Hence, we have

$$\mathbf{y}^{n+1} = \mathbf{g}(\mathbf{y}^n)$$

with

$$\mathbf{g}(\mathbf{y}) = \mathbf{M}(\mathbf{y})^{-1} \mathbf{y} = (\mathbf{I} - \Delta t \mathbf{A} \operatorname{diag}(\mathbf{C}_\gamma \mathbf{y}) \operatorname{diag}(\boldsymbol{\sigma}(\mathbf{y}))^{-1})^{-1} \mathbf{y}.$$

Next, we want to find an explicit representation of  $\mathbf{M}(\mathbf{y}^n)^{-1}$ . The diagonal matrix  $\text{diag}(\mathbf{C}_\gamma \mathbf{y}^n)$  is nonsingular for all  $\mathbf{y}^n > \mathbf{0}$ , since

$$\mathbf{C}_\gamma \mathbf{y}^n = \left(1 - \frac{1}{2\alpha}\right) \mathbf{y}^n + \left(\frac{1}{2\alpha}\right) \mathbf{B}_\gamma \mathbf{y}^n = \left(1 - \frac{1}{2\alpha}\right) \mathbf{y}^n + \left(\frac{1}{2\alpha}\right) \mathbf{y}^{(2)} > \mathbf{0} \quad (3.15)$$

holds due to  $\alpha \geq \frac{1}{2}$ . This implies that the matrix  $\mathbf{A} \text{diag}(\mathbf{C}_\gamma \mathbf{y}^n) \text{diag}(\boldsymbol{\sigma}(\mathbf{y}^n))^{-1}$  has rank one, just like  $\mathbf{A}$ . Hence, according to [28], the inverse of  $\mathbf{M}(\mathbf{y}^n)$  is given by

$$\mathbf{M}(\mathbf{y}^n)^{-1} = \mathbf{I} + \left(1 - \text{trace}(\Delta t \mathbf{A} \text{diag}(\mathbf{C}_\gamma \mathbf{y}^n) \text{diag}(\boldsymbol{\sigma}(\mathbf{y}^n))^{-1})\right)^{-1} \Delta t \mathbf{A} \text{diag}(\mathbf{C}_\gamma \mathbf{y}^n) \text{diag}(\boldsymbol{\sigma}(\mathbf{y}^n))^{-1}. \quad (3.16)$$

Using (3.5), we obtain

$$\text{diag}(\mathbf{C}_\gamma \mathbf{y}^n) \text{diag}(\boldsymbol{\sigma}(\mathbf{y}^n))^{-1} = \text{diag}(\boldsymbol{\tau}(\mathbf{y}^n)).$$

and (3.16) becomes

$$\mathbf{M}(\mathbf{y}^n)^{-1} = \mathbf{I} + \frac{\Delta t}{1 + \Delta t(a\tau_1(\mathbf{y}^n) + b\tau_2(\mathbf{y}^n))} \mathbf{A} \text{diag}(\boldsymbol{\tau}(\mathbf{y}^n)).$$

□

**Remark 3.2.** The representation of  $\mathbf{M}(\mathbf{y}^n)$  in (3.14) is also valid for arbitrary linear positive and conservative PDS  $\mathbf{y}' = \mathbf{A}\mathbf{y}$  with  $\mathbf{A} \in \mathbb{R}^{N \times N}$ , where the splitting is given by  $\mathbf{A} = \mathbf{A}_P - \mathbf{A}_D$  with  $\mathbf{A}_D = \text{diag}(\mathbf{A})$  and  $\mathbf{A}_P = \mathbf{A} - \mathbf{A}_D$ . The representation of the inverse  $\mathbf{M}(\mathbf{y}^n)^{-1}$  in (3.16) is based on the assumption that  $\mathbf{A}$  has rank one, which must not be satisfied for a general linear positive and conservative PDS.

Next, we verify that steady state solutions of (1.3) are indeed fixed points of the MPRK22 schemes. As a consequence every  $\mathbf{y}^* = r(b, a)^T$  with  $r > 0$  is a fixed point of the MPRK22 schemes. This satisfies the requirement of Theorem 2.9 to have fixed points on a line.

**Lemma 3.3.** *Any steady state solution  $\mathbf{y}^* > \mathbf{0}$  of (1.3) is a fixed point of the MPRK22 schemes, i.e. a fixed point of the map  $\mathbf{g}$  given in (3.6). Moreover, we have*

$$\mathbf{B}_\gamma \mathbf{y}^* = \mathbf{y}^*, \quad \mathbf{C}_\gamma \mathbf{y}^* = \mathbf{y}^*, \quad \boldsymbol{\sigma}(\mathbf{y}^*) = \mathbf{y}^*, \quad \tau_1(\mathbf{y}^*) = \tau_2(\mathbf{y}^*) = 1.$$

*Proof.* Let  $\mathbf{y}^*$  be a steady state solution of (1.3), i.e.  $\mathbf{A}\mathbf{y}^* = \mathbf{0}$ . At first we show, that  $\mathbf{y}^*$  is a fixed point of  $\mathbf{B}_\gamma$  defined in (3.2). Considering the first term in  $\mathbf{B}_\gamma$  we find  $(\mathbf{I} - \alpha \Delta t \mathbf{A})\mathbf{y}^* = \mathbf{y}^*$  and equivalently  $\mathbf{y}^* = (\mathbf{I} - \alpha \Delta t \mathbf{A})^{-1} \mathbf{y}^*$ . Furthermore,  $\mathbf{0} = \mathbf{A}\mathbf{y}^* = \mathbf{A}_P \mathbf{y}^* - \mathbf{A}_D \mathbf{y}^*$  is equivalent to  $\mathbf{A}_P \mathbf{y}^* = \mathbf{A}_D \mathbf{y}^*$ . Hence,  $(\mathbf{I} + \alpha \Delta t \mathbf{A}_D)\mathbf{y}^* = (\mathbf{I} + \alpha \Delta t \mathbf{A}_P)\mathbf{y}^*$  or  $\mathbf{y}^* = (\mathbf{I} + \alpha \Delta t \mathbf{A}_D)^{-1} (\mathbf{I} + \alpha \Delta t \mathbf{A}_P)\mathbf{y}^*$ . Altogether, we find

$$\mathbf{y}^{(2)} = \mathbf{B}_\gamma \mathbf{y}^* = \mathbf{y}^*. \quad (3.17)$$

As a direct consequence, it follows from (3.3) that

$$\mathbf{C}_\gamma \mathbf{y}^* = \left(1 - \frac{1}{2\alpha}\right) \mathbf{y}^* + \left(\frac{1}{2\alpha}\right) \mathbf{B}_\gamma \mathbf{y}^* = \mathbf{y}^* \quad (3.18)$$

and in addition, we conclude from (3.4) that

$$\boldsymbol{\sigma}(\mathbf{y}^*) = \mathbf{y}^*, \quad (3.19)$$

where we used  $\mathbf{y}^* > \mathbf{0}$ . Finally, this implies

$$\tau_i(\mathbf{y}^*) = \frac{(\mathbf{C}_\gamma \mathbf{y}^*)_i}{\sigma_i(\mathbf{y}^*)} = 1, \quad (3.20)$$

for  $i = 1, 2$ , and thus, according to (3.6),

$$\mathbf{g}(\mathbf{y}^*) = \left( \mathbf{I} + \frac{\Delta t}{1 + \Delta t(a\tau_1(\mathbf{y}^*) + b\tau_2(\mathbf{y}^*))} \mathbf{A} \operatorname{diag}(\boldsymbol{\tau}(\mathbf{y}^*)) \right) \mathbf{y}^* = \mathbf{y}^* + \frac{\Delta t}{1 + \Delta t(a + b)} \mathbf{A} \mathbf{y}^* = \mathbf{y}^*.$$

□

Based on the knowledge that steady state solutions  $\mathbf{y}^* > \mathbf{0}$  of (1.3) are fixed points of the MPRK schemes, it is now necessary to compute the eigenvalues of the Jacobian  $\mathbf{Dg}(\mathbf{y}^*)$ . Therefore we have to show that all partial derivatives of  $\mathbf{g}$  exist. The next lemma even states that  $\mathbf{g} \in \mathcal{C}^\infty(\mathbb{R}_{>0}^2)$ , so that all requirements of Theorem 2.9 are fulfilled.

**Lemma 3.4.** *The elements  $g_i$  of  $\mathbf{g}$  from (3.6) satisfy  $g_i \in \mathcal{C}^\infty(\mathbb{R}_{>0}^2)$  for  $i \in \{1, 2\}$ .*

*Proof.* Recall from (3.5) and (3.6) that

$$\mathbf{g}(\mathbf{y}) = \left( \mathbf{I} + \frac{\Delta t}{1 + \Delta t(a\tau_1(\mathbf{y}) + b\tau_2(\mathbf{y}))} \mathbf{A} \operatorname{diag}(\boldsymbol{\tau}(\mathbf{y})) \right) \mathbf{y},$$

with  $\tau_i(\mathbf{y}) = (\mathbf{C}_\gamma \mathbf{y})_i \sigma_i(\mathbf{y})^{-1}$  for  $i = 1, 2$ . The functions  $\sigma_i(\mathbf{y}) = (\mathbf{B}_\gamma \mathbf{y})_i^{\frac{1}{\alpha}} (y_i)^{1-\frac{1}{\alpha}}$  and  $(\mathbf{C}_\gamma \mathbf{y})_i$  are in  $\mathcal{C}^\infty(\mathbb{R}_{>0}^2)$ . Furthermore, we know  $\sigma_i(\mathbf{y}) > 0$  for  $\mathbf{y} > \mathbf{0}$  according to (3.12), which yields  $\tau_i \in \mathcal{C}^\infty(\mathbb{R}_{>0}^2)$  for  $i = 1, 2$  due to the quotient rule. Also  $\boldsymbol{\tau}(\mathbf{y}) > \mathbf{0}$  for  $\mathbf{y} > \mathbf{0}$  holds because of (3.15). Thus,  $1 + \Delta t(a\tau_1(\mathbf{y}) + b\tau_2(\mathbf{y}))$  is always positive. Consequently even  $g_i \in \mathcal{C}^\infty(\mathbb{R}_{>0}^2)$  for  $i = 1, 2$ . □

Next, we give an explicit representation of the Jacobian  $\mathbf{Dg}(\mathbf{y}^*)$ . As  $\mathbf{g}$  is defined on  $\mathbb{R}_{>0}^2$ , we have to ensure  $\mathbf{y}^* > \mathbf{0}$ , which requires  $a, b > 0$ . The cases  $a = 0$  or  $b = 0$  are special, in that no steady states of (1.3) are contained in  $\mathbb{R}_{>0}^2$ . These cases will be discussed separately in Section 3.1.

**Lemma 3.5.** *Let  $\mathbf{g}$  be defined by (3.6) and  $\mathbf{y}^* > \mathbf{0}$  a steady state of (1.3), i.e.  $\mathbf{Ay}^* = \mathbf{0}$ . Then, the Jacobian  $\mathbf{Dg}(\mathbf{y}^*)$  is given by*

$$\mathbf{Dg}(\mathbf{y}^*) = \mathbf{I} + \frac{\Delta t}{1 + \Delta t(a + b)} \mathbf{A} \left( \mathbf{I} + \frac{1}{2\alpha} (\mathbf{I} - \mathbf{B}_\gamma) \right). \quad (3.21)$$

*Proof.* Setting

$$\tilde{\boldsymbol{\tau}}(\mathbf{y}) = \frac{1}{d(\mathbf{y})} \begin{pmatrix} \tau_1(\mathbf{y})y_1 \\ \tau_2(\mathbf{y})y_2 \end{pmatrix}, \quad d(\mathbf{y}) = 1 + \Delta t(a\tau_1(\mathbf{y}) + b\tau_2(\mathbf{y})),$$

the map  $\mathbf{g}$  from (3.6) can be written as

$$\mathbf{g}(\mathbf{y}) = \mathbf{y} + \Delta t \mathbf{A} \tilde{\boldsymbol{\tau}}(\mathbf{y})$$

and consequently we have

$$\mathbf{Dg}(\mathbf{y}) = \mathbf{I} + \Delta t \mathbf{A} \mathbf{D} \tilde{\boldsymbol{\tau}}(\mathbf{y}). \quad (3.22)$$

Making use of the notation  $\partial_i = \frac{\partial}{\partial y_i}$ , we obtain

$$\begin{aligned} \partial_1 \tilde{\tau}_1(\mathbf{y}) &= \frac{(\partial_1 \tau_1(\mathbf{y})y_1 + \tau_1(\mathbf{y}))d(\mathbf{y}) - \tau_1(\mathbf{y})y_1 \partial_1 d(\mathbf{y})}{d(\mathbf{y})^2}, & \partial_2 \tilde{\tau}_1(\mathbf{y}) &= \frac{\partial_2 \tau_1(\mathbf{y})y_1 d(\mathbf{y}) - \tau_1(\mathbf{y})y_1 \partial_2 d(\mathbf{y})}{d(\mathbf{y})^2}, \\ \partial_1 \tilde{\tau}_2(\mathbf{y}) &= \frac{\partial_1 \tau_2(\mathbf{y})y_2 d(\mathbf{y}) - \tau_2(\mathbf{y})y_2 \partial_1 d(\mathbf{y})}{d(\mathbf{y})^2}, & \partial_2 \tilde{\tau}_2(\mathbf{y}) &= \frac{(\partial_2 \tau_2(\mathbf{y})y_2 + \tau_2(\mathbf{y}))d(\mathbf{y}) - \tau_2(\mathbf{y})y_2 \partial_2 d(\mathbf{y})}{d(\mathbf{y})^2} \end{aligned}$$

and thus

$$\mathbf{D} \tilde{\boldsymbol{\tau}}(\mathbf{y}) = \frac{1}{d(\mathbf{y})} (\operatorname{diag}(\mathbf{y}) \mathbf{D} \boldsymbol{\tau}(\mathbf{y}) + \operatorname{diag}(\boldsymbol{\tau}(\mathbf{y}))) - \frac{1}{d(\mathbf{y})^2} \operatorname{diag}(\boldsymbol{\tau}(\mathbf{y})) \mathbf{y} (\operatorname{grad} d(\mathbf{y}))^T.$$

Inserting this into (3.22), the Jacobian of  $\mathbf{g}$  in  $\mathbf{y}^*$  is given by

$$\mathbf{Dg}(\mathbf{y}^*) = \mathbf{I} + \frac{\Delta t}{d(\mathbf{y}^*)} \mathbf{A}(\text{diag}(\mathbf{y}^*) \mathbf{D}\boldsymbol{\tau}(\mathbf{y}^*) + \text{diag}(\boldsymbol{\tau}(\mathbf{y}^*))) - \frac{\Delta t}{d(\mathbf{y}^*)^2} \mathbf{A} \text{diag}(\boldsymbol{\tau}(\mathbf{y}^*)) \mathbf{y}^* (\text{grad } d(\mathbf{y}^*))^T.$$

From (3.20) we know that  $\text{diag}(\boldsymbol{\tau}(\mathbf{y}^*)) = \mathbf{I}$  and together with  $\mathbf{Ay}^* = \mathbf{0}$  we find

$$\mathbf{Dg}(\mathbf{y}^*) = \mathbf{I} + \frac{\Delta t}{d(\mathbf{y}^*)} \mathbf{A}(\text{diag}(\mathbf{y}^*) \mathbf{D}\boldsymbol{\tau}(\mathbf{y}^*) + \mathbf{I}).$$

To finish the proof, we have to show  $\text{diag}(\mathbf{y}^*) \mathbf{D}\boldsymbol{\tau}(\mathbf{y}^*) = \frac{1}{2\alpha}(\mathbf{I} - \mathbf{B}_\gamma)$ . Therefore, we need to express the partial derivatives of  $\tau_i(\mathbf{y}) = (\mathbf{C}_\gamma \mathbf{y})_i \sigma_i(\mathbf{y})^{-1}$  for  $i, j = 1, 2$  in terms of  $\mathbf{B}_\gamma = (b_{ij}^\gamma)_{i,j=1,2}$ . The partial derivatives of  $\boldsymbol{\sigma}$  are

$$\begin{aligned} \partial_1 \sigma_1(\mathbf{y}) &= \frac{1}{\alpha} (\mathbf{B}_\gamma \mathbf{y})_1^{\frac{1}{\alpha}-1} b_{11}^\gamma y_1^{1-\frac{1}{\alpha}} + (\mathbf{B}_\gamma \mathbf{y})_1^{\frac{1}{\alpha}} \left(1 - \frac{1}{\alpha}\right) y_1^{-\frac{1}{\alpha}}, \\ \partial_2 \sigma_1(\mathbf{y}) &= \frac{1}{\alpha} (\mathbf{B}_\gamma \mathbf{y})_1^{\frac{1}{\alpha}-1} b_{12}^\gamma y_1^{1-\frac{1}{\alpha}}, \\ \partial_1 \sigma_2(\mathbf{y}) &= \frac{1}{\alpha} (\mathbf{B}_\gamma \mathbf{y})_2^{\frac{1}{\alpha}-1} b_{21}^\gamma y_2^{1-\frac{1}{\alpha}}, \\ \partial_2 \sigma_2(\mathbf{y}) &= \frac{1}{\alpha} (\mathbf{B}_\gamma \mathbf{y})_2^{\frac{1}{\alpha}-1} b_{22}^\gamma y_2^{1-\frac{1}{\alpha}} + (\mathbf{B}_\gamma \mathbf{y})_2^{\frac{1}{\alpha}} \left(1 - \frac{1}{\alpha}\right) y_2^{-\frac{1}{\alpha}} \end{aligned}$$

and as  $\mathbf{B}_\gamma \mathbf{y}^* = \mathbf{y}^*$  we have

$$\begin{aligned} \partial_1 \sigma_1(\mathbf{y}^*) &= \frac{1}{\alpha} b_{11}^\gamma + 1 - \frac{1}{\alpha}, & \partial_2 \sigma_1(\mathbf{y}^*) &= \frac{1}{\alpha} b_{12}^\gamma, \\ \partial_1 \sigma_2(\mathbf{y}^*) &= \frac{1}{\alpha} b_{21}^\gamma, & \partial_2 \sigma_2(\mathbf{y}^*) &= \frac{1}{\alpha} b_{22}^\gamma + 1 - \frac{1}{\alpha}. \end{aligned}$$

Furthermore, due to (3.3) together with  $\mathbf{C}_\gamma = (c_{ij}^\gamma)_{i,j=1,2}$  we see

$$\begin{aligned} \partial_1(\mathbf{C}_\gamma \mathbf{y})_1 &= c_{11}^\gamma = \left(1 - \frac{1}{2\alpha}\right) + \frac{1}{2\alpha} b_{11}^\gamma, & \partial_2(\mathbf{C}_\gamma \mathbf{y})_1 &= c_{12}^\gamma = \frac{1}{2\alpha} b_{12}^\gamma, \\ \partial_1(\mathbf{C}_\gamma \mathbf{y})_2 &= c_{21}^\gamma = \frac{1}{2\alpha} b_{21}^\gamma, & \partial_2(\mathbf{C}_\gamma \mathbf{y})_2 &= c_{22}^\gamma = \left(1 - \frac{1}{2\alpha}\right) + \frac{1}{2\alpha} b_{22}^\gamma. \end{aligned}$$

Now we are ready to compute the partial derivatives of  $\boldsymbol{\tau}$ , whereby we repeatedly use (3.18) and (3.19), *i.e.*  $\mathbf{C}_\gamma \mathbf{y}^* = \mathbf{y}^*$  and  $\sigma(\mathbf{y}^*) = \mathbf{y}^*$ , which yield

$$\begin{aligned} \partial_1 \tau_1(\mathbf{y}^*) &= \frac{\partial_1(\mathbf{C}_\gamma \mathbf{y})_1 \sigma_1(\mathbf{y}) - (\mathbf{C}_\gamma \mathbf{y})_1 \partial_1 \sigma_1(\mathbf{y})}{(\sigma_1(\mathbf{y}))^2} \Big|_{\mathbf{y}=\mathbf{y}^*} = \frac{\left(1 - \frac{1}{2\alpha} + \frac{1}{2\alpha} b_{11}^\gamma\right) y_1^* - y_1^* \left(\frac{1}{\alpha} b_{11}^\gamma + 1 - \frac{1}{\alpha}\right)}{(y_1^*)^2} = \frac{1 - b_{11}^\gamma}{2\alpha y_1^*}, \\ \partial_2 \tau_1(\mathbf{y}^*) &= \frac{\partial_2(\mathbf{C}_\gamma \mathbf{y})_1 \sigma_1(\mathbf{y}) - (\mathbf{C}_\gamma \mathbf{y})_1 \partial_2 \sigma_1(\mathbf{y})}{(\sigma_1(\mathbf{y}))^2} \Big|_{\mathbf{y}=\mathbf{y}^*} = \frac{\frac{1}{2\alpha} b_{12}^\gamma y_1^* - y_1^* \frac{1}{\alpha} b_{12}^\gamma}{(y_1^*)^2} = -\frac{b_{12}^\gamma}{2\alpha y_1^*}, \\ \partial_1 \tau_2(\mathbf{y}^*) &= \frac{\partial_1(\mathbf{C}_\gamma \mathbf{y})_2 \sigma_2(\mathbf{y}) - (\mathbf{C}_\gamma \mathbf{y})_2 \partial_1 \sigma_2(\mathbf{y})}{(\sigma_2(\mathbf{y}))^2} \Big|_{\mathbf{y}=\mathbf{y}^*} = \frac{\frac{1}{2\alpha} b_{21}^\gamma y_2^* - y_2^* \frac{1}{\alpha} b_{21}^\gamma}{(y_2^*)^2} = -\frac{b_{21}^\gamma}{2\alpha y_2^*}, \\ \partial_2 \tau_2(\mathbf{y}^*) &= \frac{\partial_2(\mathbf{C}_\gamma \mathbf{y})_2 \sigma_2(\mathbf{y}) - (\mathbf{C}_\gamma \mathbf{y})_2 \partial_2 \sigma_2(\mathbf{y})}{(\sigma_2(\mathbf{y}))^2} \Big|_{\mathbf{y}=\mathbf{y}^*} = \frac{\left(1 - \frac{1}{2\alpha} + \frac{1}{2\alpha} b_{22}^\gamma\right) y_2^* - y_2^* \left(\frac{1}{\alpha} b_{22}^\gamma + 1 - \frac{1}{\alpha}\right)}{(y_2^*)^2} = \frac{1 - b_{22}^\gamma}{2\alpha y_2^*}. \end{aligned}$$

Altogether we see

$$\mathbf{D}\boldsymbol{\tau}(\mathbf{y}) = \frac{1}{2\alpha} \text{diag}(\mathbf{y}^*)^{-1} (\mathbf{I} - \mathbf{B}_\gamma),$$

which completes the proof.  $\square$

We are now in a position to compute the eigenvalues of  $\mathbf{Dg}(\mathbf{y}^*)$ , which are needed to evaluate the stability of the fixed point  $\mathbf{y}^*$ .

**Lemma 3.6.** *Every steady state  $\mathbf{y}^* > \mathbf{0}$  of (1.3) is a non-hyperbolic fixed point of the MPRK22 schemes. In particular, we have*

$$\mathbf{Dg}(\mathbf{y}^*)\mathbf{y}^* = \mathbf{y}^*$$

and

$$\mathbf{Dg}(\mathbf{y}^*)\bar{\mathbf{y}} = R_\gamma(-\Delta ta, -\Delta tb)\bar{\mathbf{y}},$$

where  $\bar{\mathbf{y}} = (1, -1)^T$  and

$$R_1(z_a, z_b) = \frac{2 - 2\alpha(z_a + z_b) - (z_a + z_b)^2}{2(1 - (z_a + z_b))(1 - \alpha(z_a + z_b))} \quad (3.23a)$$

as well as

$$R_0(z_a, z_b) = \frac{2 - (z_a + z_b)\left(\frac{z_a}{1 - \alpha z_a} + \frac{z_b}{1 - \alpha z_b}\right)}{2(1 - (z_a + z_b))}. \quad (3.23b)$$

*Proof.* From the proof of Theorem 2.9 we know that  $\mathbf{y}^*$  is an eigenvector of  $\mathbf{Dg}(\mathbf{y}^*)$  with associated eigenvalue 1, which can be checked with the straightforward calculation

$$\mathbf{Dg}(\mathbf{y}^*)\mathbf{y}^* = \mathbf{y}^* + \frac{\Delta t}{1 + \Delta t(a + b)}\mathbf{A}\left(\mathbf{I} + \frac{1}{2\alpha}(\mathbf{I} - \mathbf{B}_\gamma)\right)\mathbf{y}^* = \mathbf{y}^* + \frac{\Delta t}{1 + \Delta t(a + b)}\mathbf{A}\mathbf{y}^* = \mathbf{y}^*,$$

where we used (3.21) and (3.17). Hence,  $\mathbf{y}^*$  is a non-hyperbolic fixed point of  $\mathbf{g}$ .

Furthermore, we know from the proof of Theorem 2.9 that  $\bar{\mathbf{y}} = (1, -1)^T$  is another eigenvector of  $\mathbf{Dg}(\mathbf{y}^*)$  and we need to compute the associated eigenvalue.

First, we consider the case  $\gamma = 1$ . The vector  $\bar{\mathbf{y}}$  is an eigenvector of the matrix  $\mathbf{A}$  from (1.3) with associated eigenvalue  $\lambda = -(a + b)$ , i.e.  $\mathbf{A}\bar{\mathbf{y}} = \lambda\bar{\mathbf{y}}$ . If  $\gamma = 1$ ,  $\mathbf{B}_\gamma$  from (3.2) becomes  $\mathbf{B}_1 = (\mathbf{I} - \alpha\Delta t\mathbf{A})^{-1}$  and  $(\mathbf{I} - \alpha\Delta t\mathbf{A})\bar{\mathbf{y}} = \bar{\mathbf{y}} - \alpha\Delta t\lambda\bar{\mathbf{y}} = (1 - \alpha\Delta t\lambda)\bar{\mathbf{y}}$  implies  $\mathbf{B}_1\bar{\mathbf{y}} = (1 - \alpha\Delta t\lambda)^{-1}\bar{\mathbf{y}}$ . Hence, using (3.21), we see that

$$\mathbf{Dg}(\mathbf{y}^*)\bar{\mathbf{y}} = \bar{\mathbf{y}} + \frac{\Delta t}{1 - \Delta t\lambda}\mathbf{A}\left(\bar{\mathbf{y}} + \frac{1}{2\alpha}(\mathbf{I} - \mathbf{B}_1)\bar{\mathbf{y}}\right) = \left(1 + \frac{\Delta t\lambda}{1 - \Delta t\lambda}\left(1 + \frac{1}{2\alpha}\left(1 - \frac{1}{1 - \alpha\Delta t\lambda}\right)\right)\right)\bar{\mathbf{y}}.$$

Setting  $z = \Delta t\lambda$ , we find that the eigenvalue associated with  $\bar{\mathbf{y}}$  for  $\gamma = 1$  is given by

$$1 + \frac{z}{1 - z}\left(1 + \frac{1}{2\alpha}\left(1 - \frac{1}{1 - \alpha z}\right)\right) = \frac{2 - 2\alpha z - z^2}{2(1 - z)(1 - \alpha z)}.$$

With  $z_a = -\Delta ta$  and  $z_b = -\Delta tb$  we get  $z = -\Delta t(a + b) = z_a + z_b$  and thus

$$\frac{2 - 2\alpha z - z^2}{2(1 - z)(1 - \alpha z)} = \frac{2 - 2\alpha(z_a + z_b) - (z_a + z_b)^2}{2(1 - (z_a + z_b))(1 - \alpha(z_a + z_b))} = R_1(z_a, z_b).$$

Next, we consider the case  $\gamma = 0$  and compute  $\mathbf{AB}_0\bar{\mathbf{y}}$ . From (3.1), (3.2) and  $\mathbf{A}\bar{\mathbf{y}} = \lambda\bar{\mathbf{y}} = \lambda\mathbf{I}\bar{\mathbf{y}}$  we see

$$\begin{aligned} \mathbf{AB}_0\bar{\mathbf{y}} &= \mathbf{A}(\mathbf{I} + \Delta t\alpha\mathbf{A}_D)^{-1}(\mathbf{I} + \Delta t\alpha\mathbf{A}_P)\bar{\mathbf{y}} = \mathbf{A}(\mathbf{I} + \Delta t\alpha\mathbf{A}_D)^{-1}(\mathbf{I} + \Delta t\alpha(\mathbf{A} + \mathbf{A}_D))\bar{\mathbf{y}} \\ &= \mathbf{A}(\mathbf{I} + \Delta t\alpha\mathbf{A}_D)^{-1}(\mathbf{I} + \Delta t\alpha\mathbf{A}_D + \Delta t\alpha\lambda\mathbf{I})\bar{\mathbf{y}} = \lambda\bar{\mathbf{y}} + \alpha\Delta t\lambda\mathbf{A}(\mathbf{I} + \Delta t\alpha\mathbf{A}_D)^{-1}\bar{\mathbf{y}}. \end{aligned}$$

Realizing that  $\bar{\mathbf{y}} = (1, -1)^T$  is an eigenvector of

$$\mathbf{A}(\mathbf{I} + \Delta t\alpha\mathbf{A}_D)^{-1} = \begin{pmatrix} -\frac{a}{1 + \Delta t\alpha a} & \frac{b}{1 + \Delta t\alpha b} \\ \frac{a}{1 + \Delta t\alpha a} & -\frac{b}{1 + \Delta t\alpha b} \end{pmatrix}$$

with associated eigenvalue

$$\mu = -\left(\frac{a}{1 + \Delta t \alpha a} + \frac{b}{1 + \Delta t \alpha b}\right),$$

we obtain

$$\mathbf{A}\mathbf{B}_0\bar{\mathbf{y}} = (\lambda + \alpha\Delta t\lambda\mu)\bar{\mathbf{y}}.$$

This together with (3.21) shows

$$\begin{aligned} \mathbf{Dg}(\mathbf{y}^*)\bar{\mathbf{y}} &= \left(\mathbf{I} + \frac{\Delta t}{1 - \Delta t\lambda} \left(\mathbf{A} + \frac{1}{2\alpha}(\mathbf{A} - \mathbf{A}\mathbf{B}_0)\right)\right)\bar{\mathbf{y}} = \left(1 + \frac{\Delta t\lambda}{1 - \Delta t\lambda} + \frac{\Delta t\lambda}{2\alpha(1 - \Delta t\lambda)} - \frac{\Delta t\lambda + \alpha\Delta t\lambda\Delta t\mu}{2\alpha(1 - \Delta t\lambda)}\right)\bar{\mathbf{y}} \\ &= \left(1 + \frac{\Delta t\lambda}{1 - \Delta t\lambda} - \frac{\alpha\Delta t\lambda\Delta t\mu}{2\alpha(1 - \Delta t\lambda)}\right)\bar{\mathbf{y}} = \frac{2 - \Delta t\lambda\Delta t\mu}{2(1 - \Delta t\lambda)}\bar{\mathbf{y}}. \end{aligned}$$

Hence, using  $\Delta t\lambda = -\Delta t(a + b) = z_a + z_b$  as well as  $\Delta t\mu = \frac{z_a}{1 - \alpha z_a} + \frac{z_b}{1 - \alpha z_b}$  the eigenvalue corresponding to  $\bar{\mathbf{y}}$  for  $\gamma = 0$  is

$$\frac{2 - (z_a + z_b)\left(\frac{z_a}{1 - \alpha z_a} + \frac{z_b}{1 - \alpha z_b}\right)}{2(1 - (z_a + z_b))} = R_0(z_a, z_b).$$

□

**Remark 3.7.** The notation  $R_\gamma(z_a, z_b)$  for the eigenvalue in the above lemma was chosen on purpose. If the above analysis is carried out for a Runge–Kutta scheme, the corresponding Jacobian will have the eigenvalues 1 and  $R(z_a + z_b)$ , in which  $R$  denotes the stability function of the Runge–Kutta scheme. In this respect, the function  $R_\gamma$  of the MPRK22 schemes plays the same role as the stability function  $R$  of a Runge–Kutta scheme. In the following, we refer to  $R_\gamma$  as the stability function of the MPRK22 schemes.

To assess the stability of the non-hyperbolic fixed point  $\mathbf{y}^*$  we must investigate the absolute value of the stability function  $R_\gamma$  from (3.23). Theorem 2.3 states that if  $|R_\gamma(-\Delta ta, -\Delta tb)| > 1$ , the fixed point is unstable. If however  $|R_\gamma(-\Delta ta, -\Delta tb)| < 1$ , we can use Theorem 2.9 to conclude the stability of  $\mathbf{y}^*$ .

**Lemma 3.8.** *Let  $R_1$  be given by (3.23a), then the inequality  $|R_1(z_a, z_b)| < 1$  holds for all  $z_a, z_b < 0$  and  $\alpha \geq \frac{1}{2}$ .*

*Proof.* First, we realize that  $R_1(z_a, z_b)$  only depends on  $z_a + z_b$  and hence can be written as  $R_1(z) = \frac{2 - 2\alpha z - z^2}{2(1 - z)(1 - \alpha z)}$  with  $z = z_a + z_b$ . To prove the lemma, we need to show  $|R_1(z)| < 1$  for all  $z < 0$  and  $\alpha \geq \frac{1}{2}$ . Using  $-2z > 0$  and  $2\alpha \geq 1$ , this follows from

$$\begin{aligned} |2 - 2\alpha z - z^2| &\leq |2(1 - \alpha z)| + |z^2| = 2(1 - \alpha z) + z^2 \\ &< 2(1 - \alpha z) + 2\alpha z^2 - 2z = 2(1 - \alpha z) - 2z(1 - \alpha z) = 2(1 - z)(1 - \alpha z) = |2(1 - z)(1 - \alpha z)|. \end{aligned}$$

□

The above lemma states that the stability function of an MPRK22( $\alpha$ ) scheme has absolute value less than one for all time step sizes  $\Delta t > 0$  and matrix elements  $a, b > 0$ . This allows for the application of Theorem 2.9 to conclude that  $\mathbf{y}^*$  is indeed a stable fixed point of MPRK22( $\alpha$ ) schemes.

**Corollary 3.9.** *Let  $\mathbf{y}^* > \mathbf{0}$  be an arbitrary steady state of (1.3).*

- (a) *The MPRK22( $\alpha$ ) schemes are unconditionally stable, in the sense that any  $\mathbf{y}^*$  is a stable fixed point of the MPRK22( $\alpha$ ) schemes independent of the time step size  $\Delta t$ .*
- (b) *For every MPRK22( $\alpha$ ) scheme there exists a  $\delta > 0$ , such that  $\|\mathbf{y}^0\|_1 = \|\mathbf{y}^*\|_1$  and  $\|\mathbf{y}^0 - \mathbf{y}^*\| < \delta$  imply  $\mathbf{y}^n \rightarrow \mathbf{y}^*$  as  $n \rightarrow \infty$  independent of the time step size  $\Delta t$ .*

The situation is different for MPRK22ncs( $\alpha$ ) schemes, for which the absolute value of the stability function  $R_0$  may exceed 1, if  $\alpha$  and  $\Delta t$  are not chosen properly.

**Lemma 3.10.** *Let  $R_0$  be given by (3.23b). If  $\alpha \geq 1$  the inequality*

$$|R_0(z_a, z_b)| < 1$$

*is satisfied for all  $z_a, z_b < 0$ . If  $\frac{1}{2} \leq \alpha < 1$ , then*

$$|R_0(z_a, z_b)| < 1 \quad \text{for all } z_a < 0 \quad \text{and} \quad f(z_a) < z_b < 0$$

*and*

$$|R_0(z_a, z_b)| > 1 \quad \text{for all } z_a, z_b < 0 \quad \text{and} \quad z_b < f(z_a),$$

*where  $f$  is defined by*

$$f(\xi) = -\frac{p(\xi)}{2} - \sqrt{\frac{p(\xi)^2}{4} - q(\xi)} \quad (3.24)$$

*with*

$$p(\xi) = -\frac{2(1-\alpha\xi)(2\alpha+(1-\alpha)\xi+1)}{(2\alpha-1)(1-\alpha\xi)+\alpha\xi}, \quad q(\xi) = \frac{2(1-\alpha\xi)(2-\xi)-\xi^2}{(2\alpha-1)(1-\alpha\xi)+\alpha\xi}. \quad (3.25)$$

*Proof.* First, we show that  $R_0$  is strictly increasing with respect to  $z_a$  as well as  $z_b$ . To see this, we consider the partial derivative with respect to  $z_a$ , which is given by

$$\begin{aligned} \frac{\partial R_0}{\partial z_a}(z_a, z_b) &= \frac{-\left(\frac{z_a}{1-\alpha z_a} + \frac{z_b}{1-\alpha z_b}\right) - (z_a + z_b)\frac{1}{(1-\alpha z_a)^2}}{2(1-(z_a + z_b))} + \frac{2 - (z_a + z_b)\left(\frac{z_a}{1-\alpha z_a} + \frac{z_b}{1-\alpha z_b}\right)}{2(1-(z_a + z_b))^2} \\ &= \frac{-\left(\left(\frac{z_a}{1-\alpha z_a} + \frac{z_b}{1-\alpha z_b}\right) + \frac{z_a + z_b}{(1-\alpha z_a)^2}\right)(1-(z_a + z_b)) + 2 - (z_a + z_b)\left(\frac{z_a}{1-\alpha z_a} + \frac{z_b}{1-\alpha z_b}\right)}{2(1-(z_a + z_b))^2} \\ &= \frac{-\left(\frac{z_a}{1-\alpha z_a} + \frac{z_b}{1-\alpha z_b}\right) - \frac{z_a + z_b}{(1-\alpha z_a)^2} + \frac{(z_a + z_b)^2}{(1-\alpha z_a)^2} + 2}{2(1-(z_a + z_b))^2} = \frac{-\left(\frac{z_a}{1-\alpha z_a} + \frac{z_b}{1-\alpha z_b}\right) + \frac{(z_a + z_b)(z_a + z_b - 1)}{(1-\alpha z_a)^2} + 2}{2(1-(z_a + z_b))^2}. \end{aligned}$$

Since  $z_a, z_b < 0$  and  $\alpha > 0$ , we find  $\frac{\partial R_0}{\partial z_a}(z_a, z_b) > 0$ , i.e.  $R_0$  is strictly increasing with respect to  $z_a$ . In addition, due to the symmetry  $R_0(z_a, z_b) = R_0(z_b, z_a)$ , it is also strictly increasing with respect to  $z_b$ .

Furthermore, we have

$$\lim_{(z_a, z_b) \rightarrow (0,0)} R_0(z_a, z_b) = \lim_{(z_a, z_b) \rightarrow (0,0)} \frac{2 - (z_a + z_b)\left(\frac{z_a}{1-\alpha z_a} + \frac{z_b}{1-\alpha z_b}\right)}{2(1-(z_a + z_b))} = 1$$

and due to the monotonicity  $R_0(z_a, z_b) < 1$  for all  $z_a, z_b < 0$ . To compute a lower bound for  $R_0$ , we rewrite  $R_0$  in the form

$$R_0(z_a, z_b) = \frac{2(1-\alpha z_a)(1-\alpha z_b) - (z_a + z_b)(z_a(1-\alpha z_b) + z_b(1-\alpha z_a))}{2(1-(z_a + z_b))(1-\alpha z_a)(1-\alpha z_b)}, \quad (3.26)$$

which can also be written as

$$R_0(z_a, z_b) = \frac{2(1-\alpha z_a)\left(\frac{1}{z_b^2} - \frac{\alpha}{z_b}\right) - \left(\frac{z_a}{z_b} + 1\right)\left(z_a\left(\frac{1}{z_b} - \alpha\right) + (1-\alpha z_a)\right)}{2\left(\frac{1}{z_b} - \left(\frac{z_a}{z_b} + 1\right)\right)(1-\alpha z_a)\left(\frac{1}{z_b} - \alpha\right)}.$$

Now we see

$$\lim_{z_b \rightarrow -\infty} R_0(z_a, z_b) = \frac{-(-\alpha z_a + 1 - \alpha z_a)}{-2(1-\alpha z_a)(-\alpha)} = -\frac{1-2\alpha z_a}{2\alpha(1-\alpha z_a)}$$

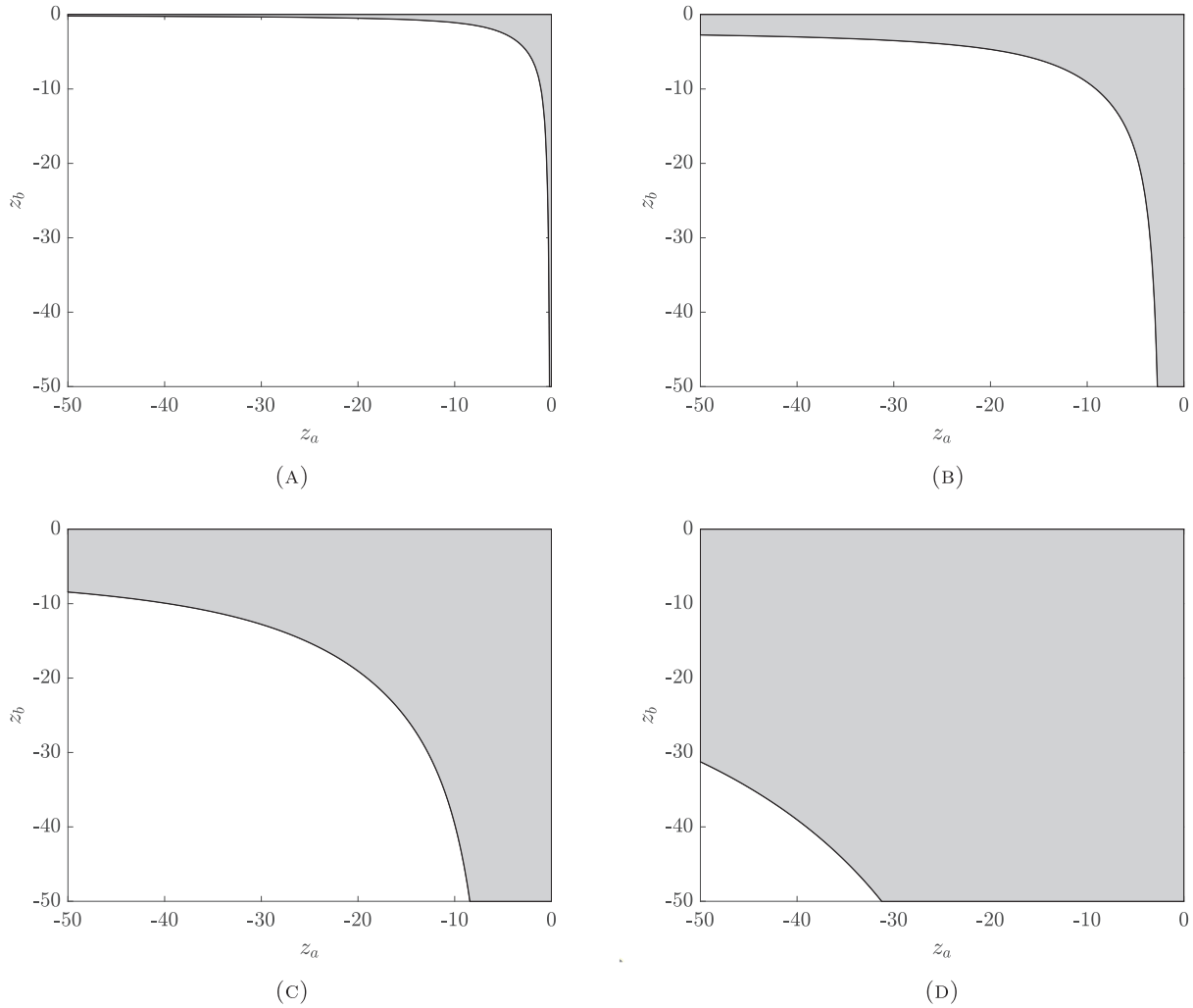


FIGURE 2. For each  $\alpha$  the grey area indicates the stability region  $S(\alpha)$  of the MPRK22ncs( $\alpha$ ) scheme for  $z_a, z_b \in [-50, 0)$ . (A)  $\alpha = 0.5$ . (B)  $\alpha = 0.8$ . (C)  $\alpha = 0.9$ . (D)  $\alpha = 0.95$ .

and

$$\lim_{z_a \rightarrow -\infty} \lim_{z_b \rightarrow -\infty} R_0(z_a, z_b) = \lim_{z_a \rightarrow -\infty} -\frac{1 - 2\alpha z_a}{2\alpha(1 - \alpha z_a)} = \lim_{z_a \rightarrow -\infty} -\frac{\frac{1}{z_a} - 2\alpha}{2\alpha(\frac{1}{z_a} - \alpha)} = -\frac{1}{\alpha}.$$

Since  $R_0$  is continuous, we even have  $\lim_{(z_a, z_b) \rightarrow (-\infty, -\infty)} R_0(z_a, z_b) = -\frac{1}{\alpha}$  and since  $R_0$  is strictly increasing in  $z_a$  and  $z_b$ , we obtain  $-\frac{1}{\alpha} < R_0(z_a, z_b)$ . Altogether, we know  $-\frac{1}{\alpha} < R_0(z_a, z_b) < 1$  for all  $z_a, z_b < 0$ , which implies

$$|R_0(z_a, z_b)| < 1 \quad \text{for all } z_a, z_b < 0 \text{ and } \alpha \geq 1.$$

However, if  $\alpha < 1$  then  $\lim_{(z_a, z_b) \rightarrow (-\infty, \infty)} R_0(z_a, z_b) = -\frac{1}{\alpha} < -1$  and hence there exist  $z_a$  and  $z_b$  for which  $|R_0(z_a, z_b)| > 1$ .

Due to the monotonicity and continuity of  $R_0$  there is exactly one  $z_b$  for a given  $z_a$ , such that  $R_0(z_a, z_b) = -1$ . To find this  $z_b$  we need to solve the equation  $R_0(z_a, z_b) = -1$  for  $z_b$ . According to (3.26), this is equivalent to

$$\frac{2(1 - \alpha z_a)(1 - \alpha z_b) - (z_a + z_b)(z_a(1 - \alpha z_b) + z_b(1 - \alpha z_a))}{2(1 - (z_a + z_b))(1 - \alpha z_a)(1 - \alpha z_b)} = -1$$

or

$$2(1 - \alpha z_a)(1 - \alpha z_b) - (z_a + z_b)(z_a(1 - \alpha z_b) + z_b(1 - \alpha z_a)) = -2(1 - (z_a + z_b))(1 - \alpha z_a)(1 - \alpha z_b).$$

A technical computation reveals that this equation can be rewritten in the form

$$((2\alpha - 1)(1 - \alpha z_a) + \alpha z_a)z_b^2 - (2(1 - \alpha z_a)(2\alpha + (1 - \alpha)z_a + 1))z_b + 2(1 - \alpha z_a)(2 - z_a) - z_a^2 = 0$$

and solving for  $z_b$  shows  $z_b = f(z_a)$ , with  $f$  defined in (3.24), (3.25). Due to the monotonicity of  $R_0$ , we have  $R_0(z_a, z_b) < -1$  for all  $z_a < 0$  and  $z_b < f(z_a)$  and  $-1 < R_0(z_a, z_b) < 1$  for all  $z_a < 0$  and  $f(z_a) < z_b < 0$ . Hence, we have also proven the statement of the lemma for the case  $\frac{1}{2} \leq \alpha < 1$ .  $\square$

An immediate consequence of the above lemma in combination with Theorems 2.3 and 2.9 is the following corollary.

**Corollary 3.11.** *Let  $\mathbf{y}^* > \mathbf{0}$  be an arbitrary steady state of (1.3) and  $\Delta t_\alpha^* > 0$  be the unique solution of  $f(-\Delta t_\alpha^* a) = -\Delta t_\alpha^* b$ , where  $f$  is defined in (3.24).*

- (a) *If  $\alpha \geq 1$  or  $\Delta t < \Delta t_\alpha^*$ , then the  $\text{MPRK22ncs}(\alpha)$  schemes are stable, in the sense that any  $\mathbf{y}^*$  is a stable fixed point of the  $\text{MPRK22ncs}(\alpha)$  schemes.*
- (b) *For every  $\text{MPRK22ncs}(\alpha)$  scheme with  $\alpha \geq 1$  or  $\Delta t < \Delta t_\alpha^*$  there exists a  $\delta > 0$ , such that  $\|\mathbf{y}^0\|_1 = \|\mathbf{y}^*\|_1$  and  $\|\mathbf{y}^0 - \mathbf{y}^*\| < \delta$  imply  $\mathbf{y}^n \rightarrow \mathbf{y}^*$  as  $n \rightarrow \infty$ .*
- (c) *If  $\Delta t > \Delta t_\alpha^*$  and  $\frac{1}{2} \leq \alpha < 1$ , then the  $\text{MPRK22ncs}(\alpha)$  schemes are unstable, in the sense that every steady state  $\mathbf{y}^*$  of (1.3) is an unstable fixed point of the  $\text{MPRK22ncs}(\alpha)$  schemes.*

According to part (a) of Corollary 3.11, we can define the stability regions for  $\text{MPRK22ncs}(\alpha)$  schemes with  $\frac{1}{2} \leq \alpha < 1$  as the set of points  $S(\alpha) \subset \mathbb{R}^- \times \mathbb{R}^-$  lying above the graph of the function  $f$  from (3.24), see Figure 2. We also want to pay attention to the fact that the  $\text{MPRK22ncs}(\alpha)$  schemes are stable for  $\alpha \geq 1$ , which coincides with the expansion of the stability region  $S(\alpha)$  for  $\alpha \rightarrow 1$  that can be observed within Figure 2.

### 3.1. The cases $a = 0$ or $b = 0$

As mentioned before, the cases  $a = 0$  and  $b = 0$  are special, in that no steady states of (1.3) are contained in  $\mathbb{R}_{>0}^2$ . In the following, we only discuss the case  $b = 0$ , i.e.

$$y_1' = -ay_1, \quad y_2 = ay_1, \quad a > 0, \quad (3.27)$$

since interchanging the roles of  $y_1$  and  $y_2$  leads to the case  $a = 0$ . This system (1.3) has the steady states  $\mathbf{y}^* = (0, y_2^*) \in \mathbb{R}^2$  with  $y_2^* \geq 0$  and we show that the MPRK22 iterations with an arbitrary initial condition  $\mathbf{y}^0 > \mathbf{0}$  converge to the steady state  $\mathbf{y}^* = (0, y_1^0 + y_2^0)$ , which is the steady state of the continuous problem, see (1.5). Since the MPRK22 schemes are conservative, it is sufficient to prove  $y_1^n \rightarrow 0$  as  $n \rightarrow \infty$ , since this directly implies  $y_2^n \rightarrow y_1^0 + y_2^0$  as  $n \rightarrow \infty$ .

Application of the MPRK22 schemes (1.6)–(3.27) results in

$$y_1^{(1)} = y_1^n, \quad (3.28a)$$

$$y_1^{(2)} = y_1^n - \alpha \Delta t a y_1^{(1)} \frac{y_1^{(2)}}{y_1^{(1)}} = y_1^n - \alpha \Delta t a y_1^{(2)}, \quad (3.28b)$$

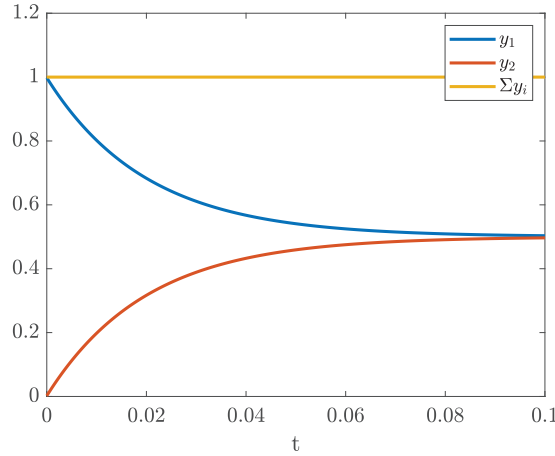


FIGURE 3. Exact solution (4.2) of the test problem (4.1).

$$y_1^{n+1} = y_1^n - \Delta t \left( \left( 1 - \frac{1}{2\alpha} \right) a y_1^{(1)} + \frac{1}{2\alpha} a y_1^{(2)} \right) \frac{y_1^{n+1}}{\left( y_1^{(2)} \right)^{\frac{1}{\alpha}} \left( y_1^{(1)} \right)^{1 - \frac{1}{\alpha}}}. \quad (3.28c)$$

We note that there is no dependence on  $\gamma$ , hence, the following applies to both  $\text{MPRK22}(\alpha)$  and  $\text{MPRK22ncs}(\alpha)$  schemes. Solving (3.28b) for  $y_1^{(2)}$  yields

$$y_1^{(2)} = \frac{y_1^n}{1 + \alpha \Delta t a}.$$

and inserting this into (3.28c) shows

$$\begin{aligned} y_1^{n+1} &= y_1^n - \Delta t a \left( \left( 1 - \frac{1}{2\alpha} \right) y_1^n + \frac{1}{2\alpha} \cdot \frac{y_1^n}{1 + \alpha \Delta t a} \right) \frac{(1 + \alpha \Delta t a)^{\frac{1}{\alpha}}}{(y_1^n)^{\frac{1}{\alpha}}} \cdot \frac{y_1^{n+1}}{(y_1^n)^{1 - \frac{1}{\alpha}}} \\ &= y_1^n - \Delta t a \left( \left( 1 - \frac{1}{2\alpha} \right) + \frac{1}{2\alpha(1 + \alpha \Delta t a)} \right) (1 + \alpha \Delta t a)^{\frac{1}{\alpha}} y_1^{n+1} \\ &= y_1^n - \frac{\Delta t a (1 + (\alpha - \frac{1}{2}) \Delta t a)}{(1 + \alpha \Delta t a)^{1 - \frac{1}{\alpha}}} y_1^{n+1}. \end{aligned}$$

Solving for  $y_1^{n+1}$  leads to

$$y_1^{n+1} = R(-\Delta t a) y_1^n,$$

where  $R$  is defined by

$$R(z) = \left( 1 - \frac{z(1 - (\alpha - \frac{1}{2})z)}{(1 - \alpha z)^{1 - \frac{1}{\alpha}}} \right)^{-1} = \frac{(1 - \alpha z)^{1 - \frac{1}{\alpha}}}{(1 - \alpha z)^{1 - \frac{1}{\alpha}} - z(1 - (\alpha - \frac{1}{2})z)}.$$

For  $z < 0$  and  $\alpha \geq \frac{1}{2}$  we have  $1 - \alpha z > 0$  and  $-z(1 - (\alpha - \frac{1}{2})z) > 0$ , which implies  $0 < R(z) < 1$ . Hence, the  $\text{MPRK22}(\alpha)$  and  $\text{MPRK22ncs}(\alpha)$  schemes converge monotonically towards the correct steady state along the line  $y_1 + y_2 = y_1^0 + y_2^0$ , just like the solution of the continuous problem.

As we have seen, there is no difference between  $\text{MPRK22}(\alpha)$  and  $\text{MPRK22ncs}(\alpha)$  schemes in the cases  $a = 0$  or  $b = 0$ , which might lead to the conclusion that both schemes have equal stability properties in general. In particular, since (3.27) is a straightforward extension of Dahlquist's equation to positive and conservative

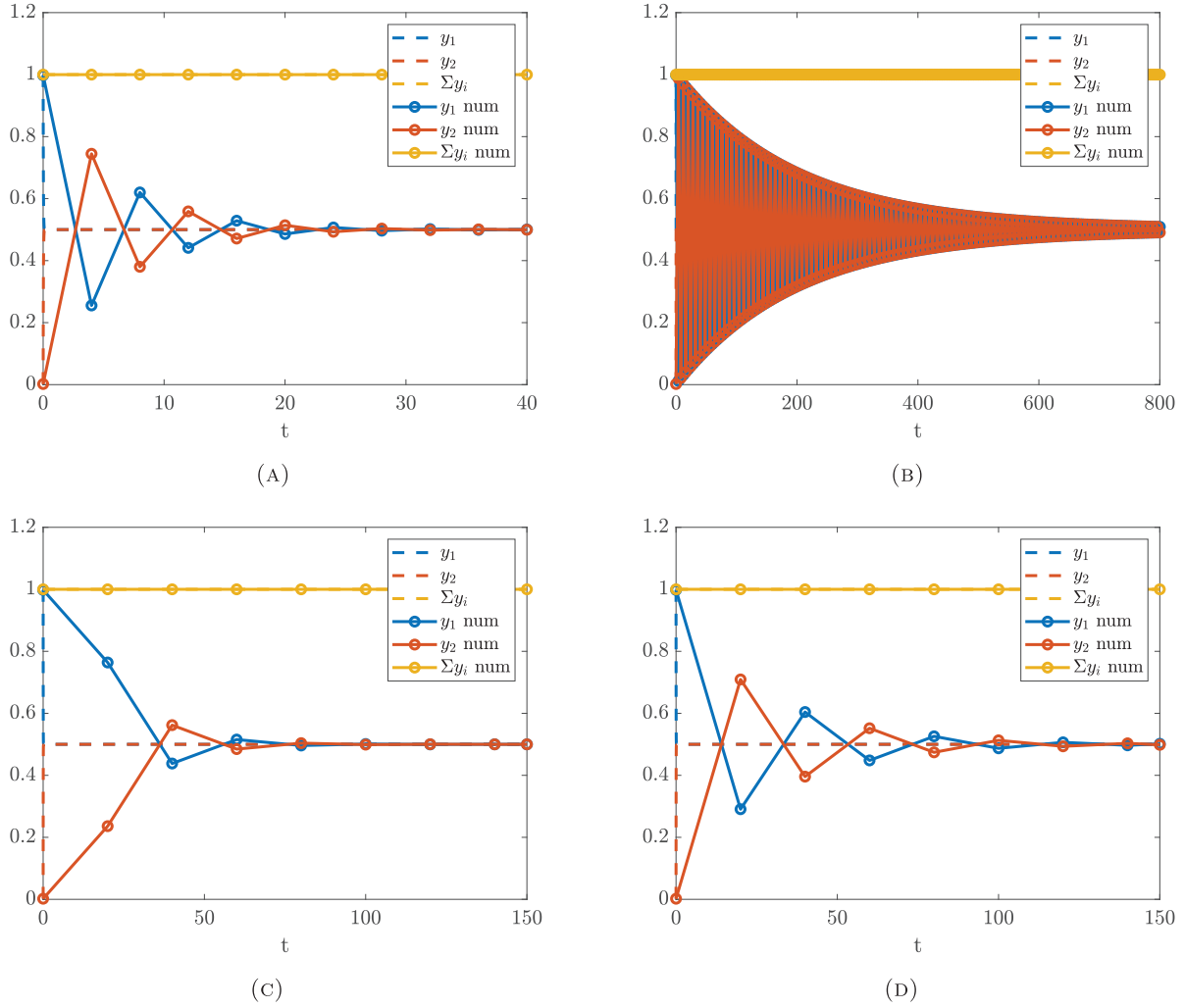


FIGURE 4. Numerical approximations of (4.1). The dashed lines indicate the exact solution (4.2). (A) MPRK22(1) with  $\Delta t = 4$ . (B) MPRK22ncs(1) with  $\Delta t = 4$ . (C) MPRK22(2) with  $\Delta t = 20$ . (D) MPRK22ncs(2) with  $\Delta t = 20$ .

PDS. But the analysis on the general system (1.3) in Section 3 shows significant differences with respect to the stability of MPRK22( $\alpha$ ) and MPRK22ncs( $\alpha$ ) schemes. Hence, it is insufficient to use (3.27) to evaluate the stability of schemes which do not belong to the class of general linear methods.

#### 4. NUMERICAL EXPERIMENTS

In this section we perform numerical experiments to confirm the stability properties of both MPRK22( $\alpha$ ) and MPRK22ncs( $\alpha$ ) schemes. For this purpose, we consider the test equation (1.3) for the case  $a = b = 25$ . In particular, we consider the initial value problem

$$\mathbf{y}' = \begin{pmatrix} -25 & 25 \\ 25 & -25 \end{pmatrix} \mathbf{y} \quad \text{with} \quad \mathbf{y}^0 = \begin{pmatrix} 0.998 \\ 0.002 \end{pmatrix}. \quad (4.1)$$

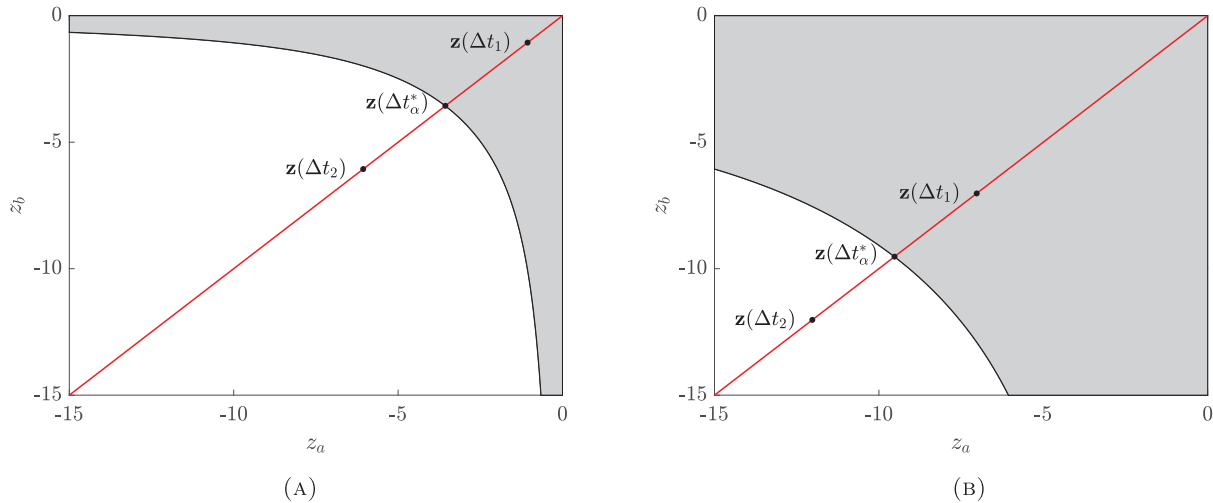


FIGURE 5. The marked points are  $\mathbf{z}(\Delta t_2) = -25(\Delta t_\alpha^* + 10^{-1})(1, 1)^T$ ,  $\mathbf{z}(\Delta t_\alpha^*) = -25\Delta t_\alpha^*(1, 1)^T$  and  $\mathbf{z}(\Delta t_1) = -25(\Delta t_\alpha^* - 10^{-1})(1, 1)^T$ , which lie on the red line  $z_a = z_b$ . The grey areas represent the stability regions for the MPRK22ncs( $\alpha$ ) schemes with  $\alpha = 0.5$  and  $\alpha = 0.8$  respectively, see also Figure 2. (A)  $\alpha = 0.5$  and  $\Delta t_\alpha^* = (\sqrt{17} + 3)/50 \approx 0.14$ . (B)  $\alpha = 0.8$  and  $\Delta t_\alpha^* = (\sqrt{101} + 9)/50 \approx 0.38$ .

The nonzero eigenvalue is  $\lambda = -(a + b) = -50$  and the analytic solution is given by

$$\mathbf{y}(t) = \frac{y_1^0 + y_2^0}{a + b} \begin{pmatrix} b \\ a \end{pmatrix} + \frac{ay_1^0 - by_2^0}{a + b} \begin{pmatrix} 1 \\ -1 \end{pmatrix} e^{\lambda t} = \frac{1}{2} \begin{pmatrix} 1 \\ 1 \end{pmatrix} + 0.498 \begin{pmatrix} 1 \\ -1 \end{pmatrix} e^{-50t}. \quad (4.2)$$

As depicted in Figure 3, it can be observed that the equilibrium state

$$\mathbf{y}^* = \frac{y_1^0 + y_2^0}{a + b} \begin{pmatrix} b \\ a \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

is approximately already reached at time  $t = 0.1$ .

To verify the theoretical statements of Corollaries 3.9 and 3.11, we investigate at first the case of  $\alpha \geq 1$ , for which we have proven that MPRK22( $\alpha$ ) as well as MPRK22ncs( $\alpha$ ) schemes are stable and locally convergent to the correct steady state of (1.3), irrespectively of the chosen time step size. To illustrate these positive properties of the methods, we choose  $\Delta t = 4$  and  $\Delta t = 20$ , in order to use very large time step sizes compared to the time scale of the exact solution of 0.1 as mentioned above. Numerical approximations obtain with these time step sizes can be seen in Figure 4. Thereby, the predicted stability of the schemes is clearly demonstrated. In the case of  $\Delta t = 4$ , the iterates of both MPRK22(1) and MPRK22ncs(1) converge to the steady state of the exact solution shown as a dashed lines. It is worth mentioning that the result of MPRK22(1) shows significantly less oscillations compared to the result of MPRK22ncs(1) and is also very close to the steady state already for  $t = 30$ , while MPRK22ncs(1) requires much more iterations to approach the steady state. The same behavior can be observed for  $\Delta t = 20$ , for which both schemes are stable and convergent to the steady state. Nevertheless, even in this case the MPRK22(2) shows smaller amplitudes and less oscillation compared to MPRK22ncs(2). In summary, the numerical experiments verify the theoretical results for both MPRK22( $\alpha$ ) and MPRK22ncs( $\alpha$ ) schemes in the case of  $\alpha \geq 1$ .

The choice of parameters  $\frac{1}{2} \leq \alpha < 1$  requires a more differentiated consideration since in this case MPRK22( $\alpha$ ) schemes are stable due to Corollary 3.9 while MPRK22ncs( $\alpha$ ) schemes postulate stability conditions with respect

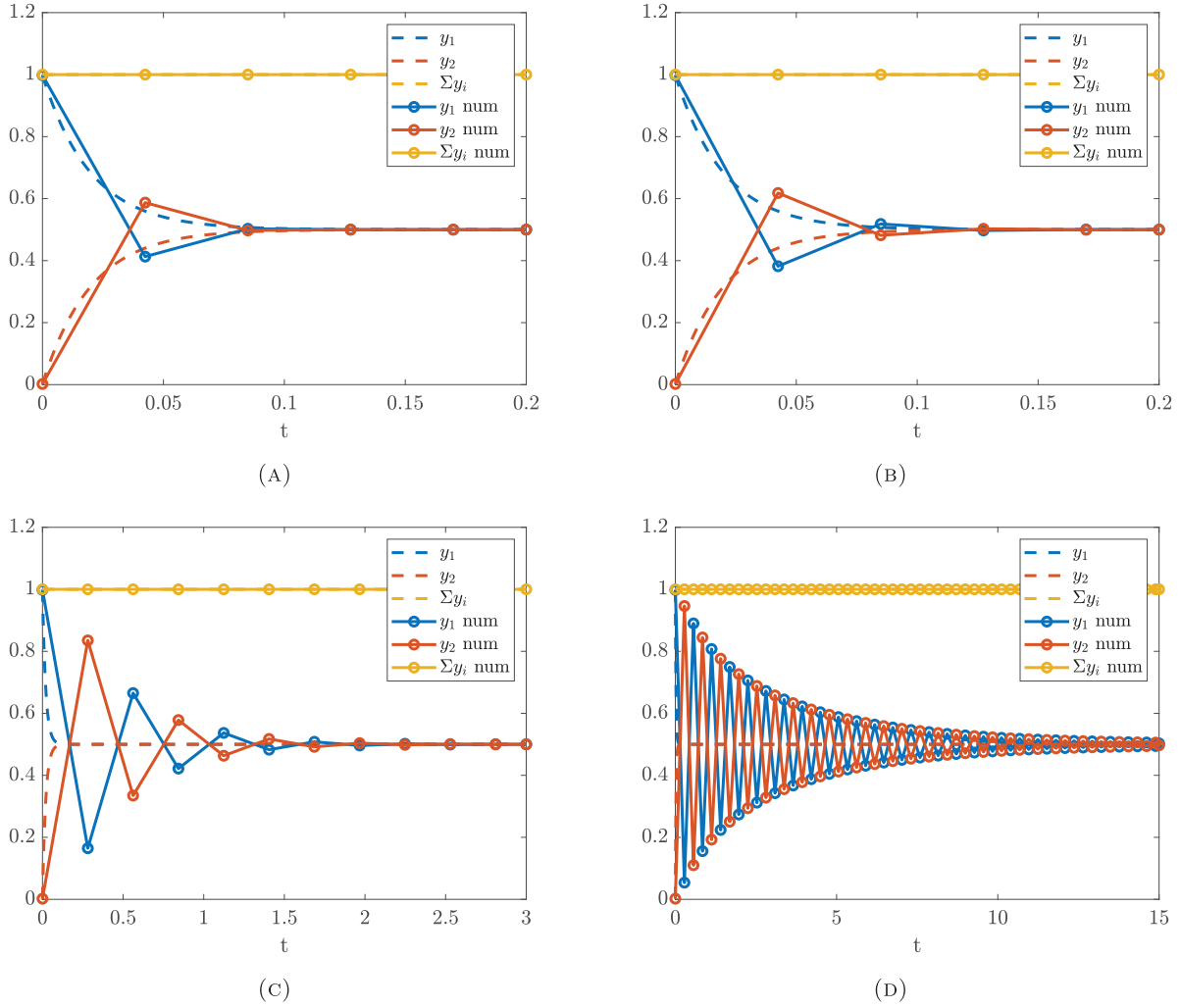


FIGURE 6. Numerical approximations of (4.1). The dashed lines indicate the exact solution (4.2). (A) MPRK22(0.5) with  $\Delta t_1 = \Delta t_\alpha^* - 10^{-1} \approx 0.04$ . (B) MPRK22ncs(0.5) with  $\Delta t_1 = \Delta t_\alpha^* - 10^{-1} \approx 0.04$ . (C) PRK22(0.8) with  $\Delta t_1 = \Delta t_\alpha^* - 10^{-1} \approx 0.28$ . (D) MPRK22ncs(0.8) with  $\Delta t_1 = \Delta t_\alpha^* - 10^{-1} \approx 0.28$ .

to the time step size according to Corollary 3.11. In the following we focus on  $\alpha = 0.5$  and  $\alpha = 0.8$ . To demonstrate the different stability behavior, we choose points of the form

$$\mathbf{z}(\Delta t) = \begin{pmatrix} z_a(\Delta t) \\ z_b(\Delta t) \end{pmatrix} = -\Delta t \begin{pmatrix} a \\ b \end{pmatrix} = -\Delta t \begin{pmatrix} 25 \\ 25 \end{pmatrix}$$

inside and outside the stability domain. The location of the points  $\mathbf{z}(\Delta t)$  is visualized for  $\alpha = 0.5$  as well as  $\alpha = 0.8$  by a red line in Figure 5. Following Corollary 3.11 the point  $\mathbf{z}(\Delta t_\alpha^*)$  lies on the boundary of the stability region and straightforward calculations yield  $\Delta t_\alpha^* = (\sqrt{17} + 3)/50 \approx 0.14$  for  $\alpha = 0.5$  and  $\Delta t_\alpha^* = (\sqrt{101} + 9)/50 \approx 0.38$  for  $\alpha = 0.8$ , respectively. The two remaining points within the figures result in each case from the choices  $\Delta t_1 = \Delta t_\alpha^* - 10^{-1}$  and  $\Delta t_2 = \Delta t_\alpha^* + 10^{-1}$ .

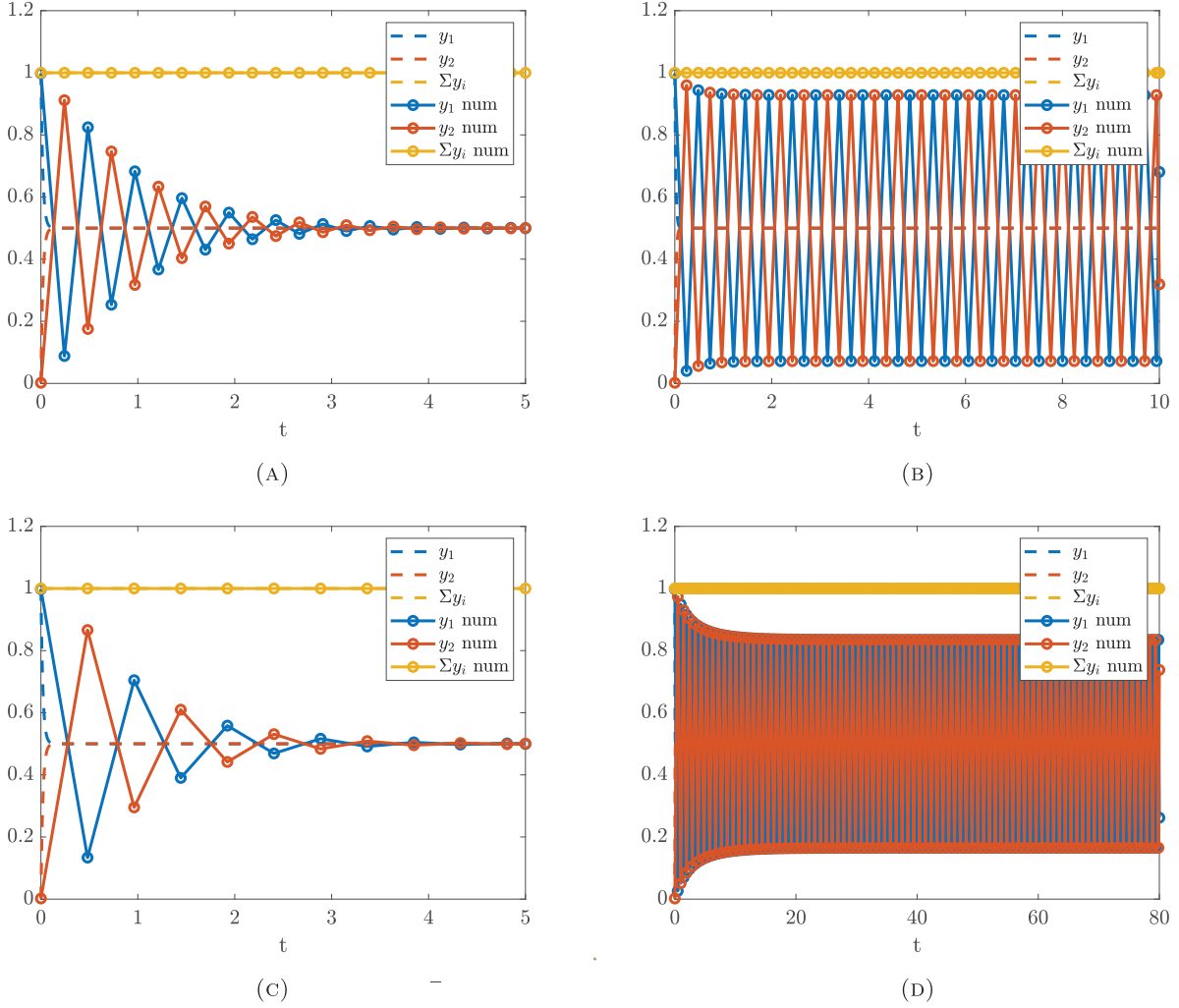


FIGURE 7. Numerical approximations of (4.1). The dashed lines indicate the exact solution (4.2). (A) MPRK22(0.5) with  $\Delta t = \Delta t_\alpha^* + 10^{-1} \approx 0.24$ . (B) MPRK22ncs(0.5) with  $\Delta t = \Delta t_\alpha^* + 10^{-1} \approx 0.24$ . (C) MPRK22(0.8) with  $\Delta t = \Delta t_\alpha^* + 10^{-1} \approx 0.48$ . (D) MPRK22ncs(0.8) with  $\Delta t = \Delta t_\alpha^* + 10^{-1} \approx 0.48$ .

According to Corollary 3.11, the MPRK22ncs method can be expected to be stable for  $\Delta t_1$  and unstable for  $\Delta t_2$ . The first expectation is confirmed by Figure 6 and additionally MPRK22( $\alpha$ ) is shown to be stable for both time step sizes, which coincides with the statement of Corollary 3.9.

In accordance with the presented theory, one can observe the superior stability behavior of MPRK22( $\alpha$ ) in Figure 7 even in the case that the time step size is chosen larger than the critical step size for MPRK22ncs( $\alpha$ ). However, the instability of MPRK22ncs( $\alpha$ ) for  $\Delta t_2 = \Delta t_\alpha^* + 10^{-1}$  can only be guessed by the illustration in Figure 7. To show the divergence of the method more clearly, we modify the initial condition within the initial value problem. Therefore, we consider

$$\mathbf{y}^0 = \mathbf{y}^* + 10^{-3} \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} 0.501 \\ 0.499 \end{pmatrix}$$

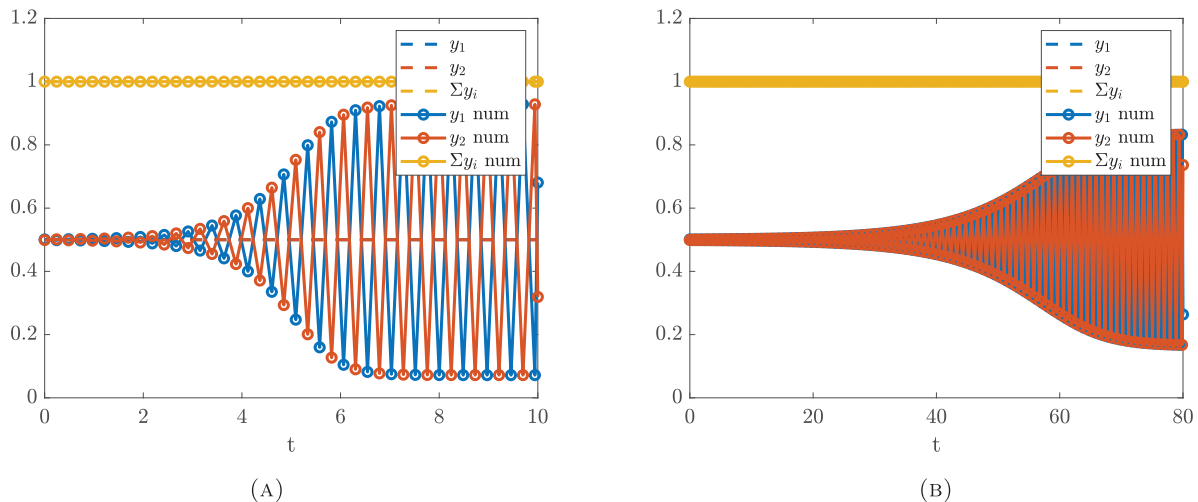


FIGURE 8. Numerical approximation of (1.3) with  $\mathbf{y}^0 = (y_1^* + 10^{-3}, y_2^* - 10^{-3})^T = (0.5 + 10^{-3}, 0.5 - 10^{-3})^T$ . The dashed lines indicate the exact solution (4.2). (A) MPRK22ncs(0.5) with  $\Delta t = \Delta t_\alpha^* + 10^{-1} \approx 0.24$ . (B) MPRK22ncs(0.8) with  $\Delta t = \Delta t_\alpha^* + 10^{-1} \approx 0.48$ .

which is much closer to the steady state than the previously used value. The results shown in Figure 8 clearly demonstrate the expected divergence from the steady state for  $\alpha = 0.5$  as well as  $\alpha = 0.8$ .

## 5. SUMMARY AND OUTLOOK

In this paper a stability analysis for general positive and conservative time integration schemes based on the center manifold theory for maps was presented for the first time. The theory shows that even for nonlinear positive and conservative time integrators the investigation of the eigenvalues of the Jacobian is sufficient to analyze stability. This novel theory was used to carry out a first stability analysis of MPRK schemes. Thereby, we discovered that for  $\alpha \geq 1$  both MPRK22( $\alpha$ ) and MPRK22ncs( $\alpha$ ) schemes possess stable fixed points irrespective of the chosen time step size  $\Delta t$ . If  $\alpha < 1$ , these stability properties are maintained by the MPRK22( $\alpha$ ) schemes, whereas the investigation of MPRK22ncs( $\alpha$ ) revealed time step restrictions to ensure stability. We also computed the corresponding stability regions for MPRK22ncs( $\alpha$ ) schemes.

Future research topics include the extension of the statement of Theorem 2.9 to higher dimensional linear and nonlinear systems, which requires the consideration of additional linear invariants. Furthermore, global stability properties are of interest. Also, due to the fact that Theorem 2.9 is applicable to general positive and conservative schemes a stability analysis of the schemes presented in [1, 2, 17, 18, 23, 26, 30] is now possible for the first time.

*Acknowledgements.* The author Th. Izgin gratefully acknowledges the financial support by the Deutsche Forschungsgemeinschaft (DFG) through grant ME 1889/10-1.

## REFERENCES

- [1] A.I. Ávila, S. Kopecz and A. Meister, A comprehensive theory on generalized BBKS schemes. *Appl. Numer. Math.* **157** (2020) 19–37.
- [2] A.I. Ávila, G.J. González, S. Kopecz and A. Meister, Extension of modified Patankar-Runge-Kutta schemes to nonautonomous production-destruction systems based on Oliver’s approach. *J. Comput. Appl. Math.* **389** (2021) 113350.

- [3] J. Benz, A. Meister and P.A. Zardo, A positive and conservative second order finite volume scheme applied to a phosphor cycle in canals with sediment. In: *PAMM: Proceedings in Applied Mathematics and Mechanics*. Vol. 7. Wiley Online Library (2007) 2040045–2040046.
- [4] J. Benz, A. Meister and P.A. Zardo, A conservative, positivity preserving scheme for advection-diffusion-reaction equations in biochemical applications. In: *Hyperbolic Problems: Theory, Numerics and Applications*. Vol. 67.2 of *Proceedings of Symposia in Applied Mathematics*, edited by E. Tadmor, J.-G. Liu and A. Tzavaras. American Mathematical Society, Providence, Rhode Island (2009) 399–408.
- [5] S. Blanes, A. Iserles and S. Macnamara, Positivity-preserving methods for population models. Preprint [arXiv:2102.08242](https://arxiv.org/abs/2102.08242) (2021).
- [6] L. Bonaventura and A. Della Rocca, Unconditionally strong stability preserving extensions of the TR-BDF2 method. *J. Sci. Comput.* **70** (2017) 859–895.
- [7] 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) 1–30.
- [8] H. Burchard, E. Deleersnijder and A. Meister, Application of modified Patankar schemes to stiff biogeochemical models for the water column. *Ocean Dyn.* **55** (2005) 326–337.
- [9] H. Burchard, K. Bolding, W. Kühn, A. Meister, T. Neumann and L. Umlauf, Description of a flexible and extendable physical–biogeochemical model system for the water column. *J. Marine Syst.* **61** (2006) 180–211.
- [10] J. Carr, Applications of Centre Manifold Theory. Vol. 35 of *Applied Mathematical Sciences*. Springer-Verlag, New York-Berlin (1981).
- [11] M. Ciallella, L. Micalizzi, P. Öffner and D. Torlo, An arbitrary high order and positivity preserving method for the shallow water equations. Preprint [arXiv:2110.13509](https://arxiv.org/abs/2110.13509) (2021).
- [12] P. Deuffhard and F. Bornemann, Scientific Computing with Ordinary Differential Equations. Translated from the 1994 German original by Werner C. Rheinboldt. Vol. 42 of *Texts in Applied Mathematics*. Springer-Verlag, New York (2002).
- [13] L. Formaggia and A. Scotti, Positivity and conservation properties of some integration schemes for mass action kinetics. *SIAM J. Numer. Anal.* **49** (2011) 1267–1288.
- [14] O. Gressel, Toward realistic simulations of magneto-thermal winds from weakly-ionized protoplanetary disks. *J. Phys. Conf. Ser.* **837** (2017) 012008.
- [15] I. Hense and A. Beckmann, The representation of cyanobacteria life cycle processes in aquatic ecosystem models. *Ecol. Modell.* **221** (2010) 2330–2338.
- [16] I. Hense and H. Burchard, Modelling cyanobacteria in shallow coastal seas. *Ecol. Modell.* **221** (2010) 238–244.
- [17] 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) 1811–1839.
- [18] 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) 1015–1056.
- [19] G. Iooss, Bifurcation of Maps and Applications. Elsevier (1979).
- [20] T. Izgin, S. Kopeck and A. Meister, Recent developments in the field of modified Patankar–Runge–Kutta–methods. In: *PAMM: Proceedings in Applied Mathematics and Mechanics*. Wiley Online Library (2021).
- [21] J.S. Klar and J.P. Mückel, A detailed view of filaments and sheets in the warm-hot intergalactic medium. *Astron. Astrophys.* **522** (2010) A114.
- [22] S. Kopeck and A. Meister, On order conditions for modified Patankar–Runge–Kutta schemes. *Appl. Numer. Math.* **123** (2018) 159–179.
- [23] S. Kopeck and A. Meister, Unconditionally positive and conservative third order modified Patankar–Runge–Kutta discretizations of production–destruction systems. *BIT Numer. Math.* **58** (2018) 691–728.
- [24] S. Kopeck and A. Meister, On the existence of three-stage third-order modified Patankar–Runge–Kutta schemes. *Numer. Algorithms* **81** (2019) 1473–1484.
- [25] J.E. Marsden and M. McCracken, The Hopf Bifurcation and its Applications. Vol. 19. Springer-Verlag (1976).
- [26] A. Martiradonna, G. Colonna and F. Diele, *GeCo*: geometric conservative nonstandard schemes for biochemical systems. *Appl. Numer. Math.* **155** (2020) 38–57.
- [27] A. Meister and J. Benz, Phosphorus Cycles in Lakes and Rivers: Modeling, Analysis, and Simulation. Springer Berlin Heidelberg, Berlin, Heidelberg (2010) 713–738.
- [28] K.S. Miller, On the inverse of the sum of matrices. *Math. Mag.* **54** (1981) 67–72.
- [29] S. Nüsslein, H. Ranocha and D.I. Ketcheson, Positivity-preserving adaptive Runge–Kutta methods. *Commun. Appl. Math. Comput. Sci.* **16** (2021) 155–179.
- [30] P. Öffner and D. Torlo, Arbitrary high-order, conservative and positivity preserving Patankar-type deferred correction schemes. *Appl. Numer. Math.* **153** (2020) 15–34.
- [31] S. Ortley and W. Hundsdorfer, Patankar-type Runge–Kutta schemes for linear PDEs. *AIP Conf. Proc.* **1863** (2017) 320008.
- [32] G. Osipenko, Center Manifolds, chapter 5. Springer New York, New York, NY (2009) 936–951.
- [33] B. Schippmann and H. Burchard, Rosenbrock methods in biogeochemical modelling – a comparison to Runge–Kutta methods and modified Patankar schemes. *Ocean Model.* **37** (2011) 112–121.
- [34] K. Semeniuk and A. Dastoor, Development of a global ocean mercury model with a methylation cycle: outstanding issues. *Global Biogeochem. Cycles* **31** (2017) 400–433.

- [35] A. Stuart and A.R. Humphries, *Dynamical Systems and Numerical Analysis*. Vol. 2. Cambridge University Press (1998).
- [36] A. Warns, I. Hense and A. Kremp, Modelling the life cycle of dinoflagellates: a case study with *Biecheleria baltica*. *J. Plankton. Res.* **35** (2013) 379–392.
- [37] S. Wei and R.J. Spiteri, Qualitative property preservation of high-order operator splitting for the SIR model. *Appl. Numer. Math.* **172** (2022) 332–350.

## Subscribe to Open (S2O)

### A fair and sustainable open access model



This journal is currently published in open access under a Subscribe-to-Open model (S2O). S2O is a transformative model that aims to move subscription journals to open access. Open access is the free, immediate, online availability of research articles combined with the rights to use these articles fully in the digital environment. We are thankful to our subscribers and sponsors for making it possible to publish this journal in open access, free of charge for authors.

#### **Please help to maintain this journal in open access!**

Check that your library subscribes to the journal, or make a personal donation to the S2O programme, by contacting [subscribers@edpsciences.org](mailto:subscribers@edpsciences.org)

More information, including a list of sponsors and a financial transparency report, available at: <https://www.edpsciences.org/en/math-s2o-programme>