A Unifying Theory for Runge–Kutta-like Time Integrators: Convergence and Stability

By

Thomas Izgin

A thesis submitted in partial fulfillment for the degree of
Doktor der Naturwissenschaften (Dr. rer. nat.)

in the

Faculty of Mathematics and Natural Sciences

University of Kassel

Date of Submission: November 10, 2023
Disputation Date: February 2, 2024

First Reviewer: **Prof. Dr. Andreas Meister**

Second Reviewer: **Prof. Dr. Chi-Wang Shu**

Kassel, February 14, 2024

Contents

Acknowledgements

This work was done during my doctoral studies at the University of Kassel and was largely funded by the German Research Foundation (DFG, project number 466355003), for which I would like to express my gratitude at this point.

First and foremost, I would like to thank my doctoral supervisor Prof. Dr. Andreas Meister for his excellent supervision and extraordinary commitment to my advancement. He has not only enabled me to further my education but also to travel to many international conferences and scientists, which have supported me to an incomparable extent. With individual support and a multitude of scientific discussions, he has influenced the present work in many ways.

I would also like to thank Dr. Stefan Kopecz very much for his constructive criticism, which significantly helped me to present research results in a more structured and transparent way. In this context, I would also like to thank him very much for the many scientific discussions, each of which had great value for me.

At this point, I would also like to express my gratitude to all my colleagues from Department 10, who made me feel very welcome. In particular, I would like to single out Veronika Straub, Stephanie Thomas, Stefan Dingel, and Andreas Linß, who became close friends during my doctoral studies and with whom I spent many hours in fruitful discussions.

However, my deepest gratitude also goes to the international scientists who have been invaluable to my research questions. At this point, I would especially like to thank Prof. Dr. Chi-Wang Shu from Brown University, Prof. Dr. Juntao Huang from Texas Tech University, and Prof. Dr. David I. Ketcheson from King Abdullah University of Science and Technology (KAUST). I would also like to thank Dr. Philipp Öffner for many good advices and great cooperation.

Finally, I would like to thank my wife Daniela from the bottom of my heart, who lovingly accompanied me through every phase of my PhD time.

Chapter 1

Introduction

Many realistic phenomena in the natural sciences, epidemiology and ecology are modeled by systems of differential equations that are constrained by restrictions linked to the nature of the problem [LD21, CD16, Koo00]. Solving these equations analytically is not possible in general, necessitating the use of numerical methods to approximate the solution. However, given the model assumptions and the presence of measurement errors, an exact representation of reality cannot be expected anyway. Rather, the goal of the numerical approximation is to retain all properties of the underlying process while achieving approximations within the limits of measurement accuracy. Two important examples of physical properties are the conservation of quantities and the positivity of certain solution components. For instance in the context of chemical reactions such as the stratospheric reaction problem [San01] or the Robertson problem [HW10], the total mass is conserved and the modeled densities are non-negative.

Often, the underlying process to be modeled consists of converting one quantity into the other, which can be represented in a more abstract framework using a special system of ordinary differential equations (ODEs), a so-called *conservative production-destruction system* (PDS). Conservativity in this context means that the production of one quantity is equivalent to the destruction of another, and vice versa. As a result of conservativity, the sum of constituents remains constant in time. A numerical method that mimics this behavior on a discrete level for every chosen time step size $\Delta t > 0$ is called *unconditionally conservative*. Similarly, if the method produces positive approximations for any $\Delta t > 0$ whenever the initial value is positive, the scheme is called *unconditionally positive*. In many cases additional terms exist that have no counterpart. In such a situation, the corresponding non-conservative PDS may be understood as the sum of a conservative PDS and rest terms. Hence, a non-conservative PDS can always be interpreted as a so-called *production-destruction-rest system* (PDRS) with a conservative PDS part.

Besides the scientifically induced requirement of preserving specific solution properties such as conservativity and positivity, the preservation of these two particular properties also hold significant importance from a purely numerical perspective. First, a numerical method that does not preserve all linear invariants such as conservativity may produce a qualitatively wrong behavior [Sha86, BDM03, LD21]. Second, the preservation of positivity is a desirable property because negative approximations can lead to the failure of the method, see for instance [STKB05] and the literature mentioned therein. Preserving the positivity of certain solution components is also crucial in the context of partial differential equations (PDEs). For instance, the calculation of the speed of sound when solving the compressible Euler equations requires the positivity of pressure and density. Another system of PDEs that emphasizes the importance of generating positive approximations is given in [KM19a], where the right-hand sides of the so-called NPZD model ($n$utrients, $p$hytoplankton, $z$ooplankton, and $d$etritus) [BDM05] were used as stiff source terms. In the numerical solution of the resulting PDE, the occurrence of negative approximations can lead to the divergence of the method and therefore necessitates a severe time step constraint for methods that are not unconditionally positive, see [KM19a].

While high order general linear methods [HW10, Jac09] such as Runge–Kutta and linear multistep schemes [But16, HNW93, HW10] preserve all linear invariants of the system, unconditional positivity is much harder to obtain. Among the class of linear integrators, unconditional positivity is restricted to first order [San02, BC78]. The implicit Euler method indeed grants the positivity, although methods for solving nonlinear systems coming from implicit schemes do not guarantee positive approximations. Higher order linear methods can only guarantee positivity by restricting the time step size, leading to a significant increase in computational time [San02, Ber96].

Positive and linear invariants preserving schemes based on projection techniques were proposed in [San01, NRK21a], where at each time step, the negative approximations or the weights of the Runge–Kutta method are changed to guarantee positivity while maintaining the order of the method. More recently, the issue of positivity preservation was addressed in [BIM22], where splitting and exponential methods were combined to construct positive and conservative integrators up to 3rd order for solving nonlinear mass conservative systems of the type $\mathbf{y}'(t) = \mathbf{A}(\mathbf{y}(t), t)\mathbf{y}(t)$, where $\mathbf{A}(\mathbf{y}(\cdot), \cdot)$ is an $N \times N$ matrix-valued function.

Another approach for preserving positivity is to apply the Patankar-trick [Pat80] to an RK method resulting in a Patankar–Runge–Kutta (PRK) scheme, which guarantees the unconditional positivity of the numerical approximation. However, the PRK method in general does not preserve linear invariants such as conservativity anymore. Still, PRK methods are of interest due to their unconditional positivity. Furthermore, in the context of conservative production-destruction systems, it is possible to improve the PRK method obtaining modified PRK (MPRK) schemes, originally introduced in [BDM03], which additionally are unconditionally conservative. Second and third order MPRK schemes have been developed and numerically investigated in [KM18a, KM18b, KM19b]. The idea was then carried out in the context of strong-stability preserving (SSP) Runge–Kutta methods in [HS19, HZS19], where the resulting SSPMPRK schemes have been applied to solve reactive Euler equations. In [ÖT20], the authors used the Patankar-trick to develop MPDeC methods, which are modified Patankar (MP) schemes of arbitrary order based on deferred correction methods (DeC). It is worth mentioning that the 5th order MPDeC method was used to preserve a positive water height when solving the shallow water equations [CMÖT22]. Furthermore, an implicit first order MP scheme based on a 3rd order SDIRK method was presented in [MO14] and applied to the shallow water equations to guarantee a positive water height. Thereby, it was also proven that the method is of third order away from the wet-dry transition zone. All these schemes are mass conservative and unconditionally positive. Moreover their efficiency and robustness was proven numerically while integrating stiff PDS.

Among the positive and linear invariants preserving integrators for biochemical systems, 1st and 2nd order generalized BBKS (gBBKS), which were developed in

[BBKS07, BRBM08, AKM20] and named after the authors Bruggeman, Burchard, Kooi, and Sommeijer, and Geometric Conservative (GeCo) schemes [MCD20] have been introduced in recent literature. These methods fall in the class of non-standard integrators [Mic21], as they result as non-standard versions of explicit first and second order Runge–Kutta schemes, where the advancement in time is modulated by a nonlinear functional dependency on the temporal step size and on the approximation itself. The step size modification thereby guarantees the numerical solution to be unconditionally positive while keeping the accuracy of the underlying method. While GeCo schemes are explicit integrators, the gBBKS step size modification function leads to an implicit scheme. Nevertheless, nonlinear implicit equations that arise from gBBKS schemes may be reduced to a scalar nonlinear equation in one single unknown [AKM20].

We want to emphasize that the application of the modified Patankar approach on an RK scheme has a great impact on its structure. Indeed, the resulting MPRK scheme is not an RK method anymore. Even more, MPRK schemes do not belong to the class of general linear methods. Therefore, the excessive theory for RK schemes cannot be applied directly to deduce the properties of MPRK methods. As a result, the first constructions of 2nd and 3rd order MPRK schemes in [KM18a, KM18b] were interlinked with technical proofs using Taylor series expansions. Moreover, due to the nonlinear nature of Patankar-type methods, also a stability analysis for these schemes is not straightforward, yet of high importance.

The first part of my thesis is concerned with developing a comprehensive theory for deriving order conditions of Patankar-type methods. To that end, we generalized the theory of NB-series [AMSS97] by interpreting Patankar-type methods as Runge–Kutta-like schemes with solution-dependent Butcher tableau, which we referred to as non-standard additive Runge–Kutta (NSARK) methods in [IKM23b]. Thereby, the main idea was to revisit Butchers approach from [But16] concerning order conditions for RK schemes and apply his techniques to the results for additive Runge–Kutta methods [AMSS97]. Furthermore, we adapted Butcher's proofs in such a way that they remain valid even for the case of solution-dependent Butcher tableaux. In particular, we provided a theorem for arbitrary high order NSARK methods. However, these order conditions may be implicit or not fully reduced. Nevertheless, we were able to trace the reduction of order conditions back to the investigation of polynomial systems of equations, which we were able to solve using the Gröbner basis theory from commutative algebra. We applied this approach deriving the known order conditions for GeCo and MPRK methods from [MCD20, KM18a, KM18b] in a compact manner. Moreover, within the same work [IKM23b], we derived for the first time explicit conditions for 3rd and 4th order GeCo methods as well as 4th order MPRK schemes.

Even though the first MPRK schemes were introduced about two decades ago in [BDM03] and followed by many further works on positivity-preserving methods, the corresponding theories for a stability analysis and deriving order conditions were first developed in my PhD project. In particular, I present in this work a unifying theory for the analysis of Patankar-type schemes concerning their stability and convergence. To that end, we review and extend the corresponding results that were already published during my PhD time.

A first step in my approach of investigating the stability of MPRK schemes was the observation that the scalar Dahlquist equation $y'(t) = \lambda y(t)$ with $\lambda \in \mathbb{C}^-$ could not be used for the analysis. The reason for that is the fact that MP schemes are applied to real valued systems of equations. One is thus tempted to consider the decoupled PDS

$$\begin{pmatrix} y_1'(t) \\ y_2'(t) \end{pmatrix} = \begin{pmatrix} \lambda y_1(t) \\ -\lambda y_1(t) \end{pmatrix}, \quad \lambda \in \mathbb{R}^-,$$

whose first component represents the Dahlquist equation with $\lambda \in \mathbb{R}^-$. However, it turned out that the analysis of this equation is not even sufficient to understand the stability behavior in a more general system with two equations [IKM22a, IKM22b], let alone larger systems. Instead, the main idea was to use the theory of center manifolds for maps from dynamical systems [Car81, SH98, MM76] to analyze the behavior of the numerical method near steady states when applied to general linear autonomous problems. This approach was first carried out for systems of two equations [IKM22a] and later generalized to arbitrary large linear PDS [IKM22b], already analyzing a second order family of MPRK schemes. The very first stability analysis of further Patankar-type methods followed shortly, which resulted in several publications [IÖ23, IKMM23, HIK$^+$23] during my PhD time. We also want to note that the theory is not limited to linear problems, but can also be applied in the context of certain nonlinear PDS [IKM23c]. Furthermore, we derived a necessary condition for avoiding unrealistic oscillations in [IÖT22], underlining the numerical results from [TÖR22], where different modified Patankar methods from [ÖT20, KM18a, KM18b] were analyzed with respect to oscillatory behavior. Also, recently we investigated the hypothesis that the stability properties may be of global nature when the MPRK scheme is based on a non-negative Butcher tableau [IKMS23], which is mostly based on the master thesis [Sch23].

Altogether, this thesis represents a collection of my work as first author with several collaborators on the stability and convergence of nonlinear time stepping methods. Additionally, I unify in this framework the stability analysis for the above mentioned MPRK schemes by deriving a stability function for NSARK methods. Moreover, we also investigate RK schemes generalizing the notion of A-stability.

The remainder of the thesis is divided into six chapters and an appendix.

We first review the theoretical fundamentals in Chapter 2. In particular, RK and additive RK (ARK) methods are introduced. Additionally, we recall the main theorems concerning their stability and order of convergence. Furthermore, we introduce the notation for the production-destruction-rest systems together with the main properties of interest.

In Chapter 3 we present the previously mentioned Patankar-type schemes and write them as Runge–Kutta-like methods with solution-dependent Butcher tableau.

In the following Chapter 4 we then turn to order conditions for Patankar-type methods giving a unifying and comprehensive theory based on the order conditions for ARK methods. In particular, we investigate GeCo and MPRK reproducing the known order conditions in a compact manner. Furthermore, we give explicit formulations for the conditions of 3rd and 4th order GeCo and 4th order MPRK methods. We also construct a 4th order MPRK method and confirm its order of convergence numerically.

In Chapter 5 we present the stability theory based on the center manifold theorem for maps and investigate Patankar-type methods as well as Runge–Kutta schemes. We also provide necessary conditions for non-oscillatory schemes and validate the theoretical results with numerical experiments.

Finally, we come to a conclusion in Chapter 6, where we also discuss open questions for future work.

Chapter 2

Theoretical Fundamentals

2.1 Runge–Kutta Methods

Runge–Kutta (RK) methods are numerical schemes to approximate the solution $\mathbf{y}\colon [t_0, t_{\text{end}}] \to \mathbb{R}^N$ of the initial value problem (IVP)

$$\mathbf{y}'(t) = \mathbf{f}(\mathbf{y}(t), t), \quad \mathbf{y}(t_0) = \mathbf{y}^0. \tag{2.1}$$

Hereafter, we use superscript indices for vectors to better distinguish between iterates of a numerical method and their respective components. For the sake of simplicity, let us consider a fixed time step size Δt and set $t_n = t_0 + n\Delta t$ for $n = 1, \dots, k$ so that $t_n \in [t_0, t_{\text{end}}]$. A time integrator such as a Runge–Kutta method aims to generate approximations $\mathbf{y}^n$ to $\mathbf{y}(t_n)$. In the case of RK schemes, intermediate times

$$\xi_j = t_n + c_j\Delta t, \quad c_j \in [0, 1], \quad j = 1, \dots, s$$

are introduced and a quadrature formula is used to obtain

$$\mathbf{y}(t_{n+1}) - \mathbf{y}(t_n) = \int_{t_n}^{t_{n+1}} \mathbf{y}'(t)\mathrm{d}t = \int_{t_n}^{t_{n+1}} \mathbf{f}(\mathbf{y}(t), t)\mathrm{d}t \approx \Delta t \sum_{j=1}^{s} b_j \mathbf{f}(\mathbf{y}(\xi_j), \xi_j),$$

where b_j depend on the particular quadrature formula, and $\sum_{j=1}^{s} b_j = 1$ holds true if an interpolatory quadrature formula is used. Since the value of $\mathbf{y}$ at the intermediate times ξ_i, $i = 1, \dots, s$, is not known in general, we approximate them in a similar manner, i. e.

$$\mathbf{y}(\xi_i) - \mathbf{y}(t_n) = \int_{t_n}^{\xi_i} \mathbf{y}'(t)\mathrm{d}t = \int_{t_n}^{t_n + c_i\Delta t} \mathbf{f}(\mathbf{y}(t), t)\mathrm{d}t \approx \Delta t \sum_{j=1}^{s} a_{ij} \mathbf{f}(\mathbf{y}(\xi_j), \xi_j),$$

where a_{ij} again depend on the chosen quadrature rule and $\sum_{\nu=1}^{s} a_{ij} = c_i$ holds true for interpolatory quadrature formulae.

Now, denoting the approximation to $\mathbf{y}(\xi_i)$ by $\mathbf{y}^{(i)}$, the corresponding s-stage

Runge–Kutta method for the solution of the IVP (2.1) is given by

$$\mathbf{y}^{(i)} = \mathbf{y}^n + \Delta t \sum_{j=1}^{s} a_{ij} \mathbf{f}(\mathbf{y}^{(j)}, t_n + c_j \Delta t), \quad i = 1, \dots, s, \tag{2.2a}$$

$$\mathbf{y}^{n+1} = \mathbf{y}^n + \Delta t \sum_{j=1}^{s} b_j \mathbf{f}(\mathbf{y}^{(j)}, t_n + c_j \Delta t). \tag{2.2b}$$

It is worth mentioning that a Runge–Kutta method is characterized by its coefficients a_{ij}, b_j, c_j for $i, j = 1, \dots, s$ and can be represented by the *Butcher tableau*

$$\begin{array}{c|c} \mathbf{c} & \mathbf{A} \\ \hline & \mathbf{b} \end{array}$$

with $\mathbf{A} = (a_{ij})_{i,j=1,\dots,s}$, $\mathbf{b} = (b_1, \dots, b_s)$ and $\mathbf{c} = (c_1, \dots, c_s)^T$. If $\mathbf{A}$ is a strict lower left triangular matrix, the *stage vectors* $\mathbf{y}^{(i)}$ can be computed explicitly using (2.2a), which is why the corresponding RK method is called *explicit*. Otherwise, the scheme is called *implicit*. If $\mathbf{f}$ is nonlinear and the RK scheme is implicit, the stage vectors $\mathbf{y}^{(i)}$ are the solution to a nonlinear system of equations. Nevertheless, the existence of a unique solution can be guaranteed under some time step constrains for Lipschitz continuous (with respect to $\mathbf{y}$) right-hand sides $\mathbf{f}$ [HNW93, Theorem 7.2].

Remark 2.1 ([HNW93, Section II.2],[DB02])**.** Given the non-autonomous IVP (2.1), one may rather consider solving the corresponding autonomous problem

$$\mathbf{Y}'(t) = \mathbf{F}(\mathbf{Y}(t)) \tag{2.3}$$

with $\mathbf{Y}(t) = \binom{\mathbf{y}(t)}{t}$ and $\mathbf{F}(\mathbf{Y}(t)) = \begin{pmatrix} \mathbf{f}(\mathbf{Y}(t)) \\ 1 \end{pmatrix}$. If the stage vectors are uniquely determined and

$$\sum_{j=1}^{s} a_{ij} = c_i \tag{2.4}$$

holds, then the approximations of an RK method to the solution $\mathbf{y}$ of (2.1) are identical regardless of whether the method was applied to (2.1) or (2.3).

2.1.1 Additive Runge–Kutta Methods

A generalization of Runge–Kutta methods are *additive Runge–Kutta* (ARK) schemes, which approximate the solution of the initial value problem, where the right-hand side is split into a sum, that is

$$\mathbf{y}'(t) = \mathbf{f}(\mathbf{y}(t), t) = \sum_{\nu=1}^{N} \mathbf{f}^{[\nu]}(\mathbf{y}(t), t), \quad \mathbf{y}(t_0) = \mathbf{y}^0 \in \mathbb{R}^d. \tag{2.5}$$

The main idea of an ARK method is to apply very different RK schemes determined by $\mathbf{A}^{[\nu]}, \mathbf{b}^{[\nu]}, \mathbf{c}^{[\nu]}$ to the different addends $\mathbf{f}^{[\nu]}$. A popular class of ARK schemes are Implicit-Explicit (IMEX) RK methods [Cro80, ARS97]. For internal consistency, we require that the different RK schemes actually do not differ in $\mathbf{c}$, i. e.

$$c_i = c_i^{[\nu]} = \sum_{j=1}^{s} a_{ij}^{[\nu]} \tag{2.6}$$

for $i = 1, \dots, s$ and $\nu = 1, \dots, N$, see [SG15]. For standard RK methods this reduces to (2.4). The resulting ARK method reads

$$
\begin{aligned}
\mathbf{y}^{(i)} &= \mathbf{y}^n + \Delta t \sum_{j=1}^{s} \sum_{\nu=1}^{N} a_{ij}^{[\nu]} \mathbf{f}^{[\nu]}(\mathbf{y}^{(j)}, t_n + c_j \Delta t), \quad i = 1, \dots, s, \\
\mathbf{y}^{n+1} &= \mathbf{y}^n + \Delta t \sum_{j=1}^{s} \sum_{\nu=1}^{N} b_j^{[\nu]} \mathbf{f}^{[\nu]}(\mathbf{y}^{(j)}, t_n + c_j \Delta t),
\end{aligned} \tag{2.7}
$$

and the corresponding extended Butcher tableau is given by

$$
\begin{array}{c|c|c|c|c}
\mathbf{c} & \mathbf{A}^{[1]} & \mathbf{A}^{[2]} & \cdots & \mathbf{A}^{[N]} \\
\hline
 & \mathbf{b}^{[1]} & \mathbf{b}^{[2]} & \cdots & \mathbf{b}^{[N]}
\end{array},
$$

where $\mathbf{A}^{[\nu]} = (a_{ij}^{[\nu]})_{i,j=1,\dots,s}$ and $\mathbf{b}^{[\nu]} = (b_1^{[\nu]}, \dots, b_s^{[\nu]})$. The statement of Remark 2.1 follows also in the case of ARK methods from the internal consistency condition (2.6), see [SG15]. As a consequence, it suffices to investigate autonomous systems to understand the order of the method, if (2.6) is satisfied.

2.2 NB-Series and Order Conditions for ARK Methods

Runge–Kutta (RK) and additive RK schemes belong to *one-step* methods since there exists an *incremental map* $\mathbf{\Phi}$ generating the iterates according to

$$
\mathbf{y}^{n+1} = \mathbf{y}^n + \Delta t \mathbf{\Phi}(t_n, \mathbf{y}^n, \Delta t), \quad \mathbf{y}^0 = \mathbf{y}(t_0), \tag{2.8}
$$

where implicit schemes are formally represented in their explicit form. For one-step methods, we consider the following notions and results.

Definition 2.2 ([HNW93, HW10])**.** Let $\mathbf{y} \colon [t_0, t_{\text{end}}] \to \mathbb{R}^N$ be the solution to the IVP (2.5). A one-step method for solving (2.5)

a) has and *order of consistency* p, if the *local truncation error*

$$
\boldsymbol{\eta}(t, \Delta t) = \mathbf{y}(t) + \Delta t \mathbf{\Phi}(t, \mathbf{y}(t), \Delta t) - \mathbf{y}(t + \Delta t)
$$

with $t \in [t_0, t_{\text{end}}]$ and $0 \le \Delta t \le t_{\text{end}} - t$ satisfies

$$
\boldsymbol{\eta}(t, \Delta t) = \mathcal{O}(\Delta t^{p+1}), \quad \Delta t \to 0
$$

for all $t \in [t_0, t_{\text{end}}]$.

b) has an *order of convergence* p, if the *global error*

$$
\mathbf{e}(t_n, \Delta t) = \mathbf{y}^n - \mathbf{y}(t_n)
$$

satisfies

$$
\mathbf{e}(t_n, \Delta t) = \mathcal{O}(\Delta t^p), \quad \Delta t \to 0
$$

for any $t_n = t_0 + n\Delta t \in [t_0, t_{\text{end}}]$.

While the local error represents the error of the method generated by a single step starting with exact data, the global error is determined by the difference of the numerical and analytical solutions after n steps. These two notions are deeply interlinked by the following result.

Theorem 2.3 ([SM03, Theorem 12.2, 12.3],[DB02, Theorem 4.10])**.**
Let $\mathbf{y}\colon [t_0, t_{\text{end}}] \to \mathbb{R}^N$ be the sufficiently smooth solution to the IVP (2.5). Furthermore, let the incremental map $\mathbf{\Phi}$ of the one-step method (2.8) for solving (2.5) be continuous. In addition let $\mathbf{\Phi}$ be locally Lipschitz with respect to its second input argument in the sense that

$$\|\mathbf{\Phi}(t,\mathbf{x},\Delta t) - \mathbf{\Phi}(t,\mathbf{z},\Delta t)\| \leq L_{\mathbf{\Phi}}\|\mathbf{x}-\mathbf{z}\| \quad \text{on} \quad D \times [0, \Delta t_0]$$

for some $0 < \Delta t_0 \leq t_{\text{end}}$ and

$$D = \{(t,\mathbf{z}) \mid t_0 \leq t \leq t_M, \|\mathbf{z} - \mathbf{y}^0\| \leq C\}$$

with some $t_M \leq t_{\text{end}}$ and $C > 0$. If the one-step method is consistent of order p, then it is also convergent of order p.

If the incremental map satisfies a certain Lipschitz condition specified in Theorem 2.3, it thus suffices to study the local truncation error of a method to understand its accuracy, i. e. to deduce the order of convergence.

The accuracy of standard RK methods can be understood through the use of trees and B-series, which are formal power series used to represent exact and approximate solutions of an autonomous initial value problem [But16, HW74]. Similarly, ARK methods can be studied using colored trees and NB-Series [AMSS97], which we briefly review in the upcoming subsection.

2.2.1 Colored Rooted Trees

A *rooted tree* is a cycle-free, connected graph with one node designated as the root [But16]. More precisely, a rooted tree can be understood as the underlying undirected graph of an arborescence, for which the root is the uniquely determined node with no incoming arc [KV12]. We consider colored rooted trees, in which each node possesses one of N possible colors from the set $\{1,\ldots,N\}$. We denote the set of all colored rooted trees, the so-called *N-trees*, by NT. We indicate the color $\nu \in \{1,\ldots,N\}$ of the tree represented by $\bullet$ by writing $\bullet^{[\nu]}$. In general, a colored rooted tree τ with a root color ν can be written in terms of its colored *children* $\tau_1,\ldots,\tau_k$ by writing

$$\tau = [\tau_1,\ldots,\tau_k]^{[\nu]} = [\tau_1^{m_1},\ldots,\tau_r^{m_r}]^{[\nu]}, \tag{2.9}$$

where the children $\tau_1,\ldots,\tau_k$ are the connected components of τ when the root together with its edges are removed. Moreover, the neighbors of the root of τ are the roots of the corresponding children. In the latter representation of τ in (2.9), m_i is the number of copies of τ_i within $\tau_1,\ldots,\tau_k$, which already includes the fact that we do not distinguish between trees whose children are permuted.

Example 2.4. For simplicity, we consider only one color in this example, that is $N = 1$ and $\bullet^{[1]} = \bullet$. The children of the tree $\tau =$ are given by $\tau_1 = \bullet$ and $\tau_2 =$ and the respective roots are the lowest nodes. In terms of the representation (2.9) we can write $= [\bullet,$ $] = [\bullet, [\bullet, \bullet]] = [\bullet, [\bullet^2]] = [[\bullet^2], \bullet]$.

The *order* of a colored tree τ is denoted by $|\tau|$ and equals the number of its nodes. We introduce the set NT_q of all N-trees up to order q. We set $NT_0 = \emptyset$

and note that the sets NT_q for $q = 1, 2, 3$ read

$$\begin{aligned}
NT_1 &= \{\bullet^{[\nu]} \mid \nu = 1, \ldots, N\},\\
NT_2 &= NT_1 \cup \left\{ \begin{smallmatrix}\bullet & [\nu]\\ | & \\ \bullet & [\mu]\end{smallmatrix} \,\middle|\, \nu, \mu = 1, \ldots, N \right\},\\
NT_3 &= NT_2 \cup \left\{ \begin{smallmatrix}\bullet & [\xi]\\ | & \\ \bullet & [\nu]\\ | & \\ \bullet & [\mu]\end{smallmatrix} \,\middle|\, \nu, \mu, \eta = 1, \ldots, N \right\} \cup \left\{ \begin{smallmatrix}\bullet\,[\nu] & & \bullet\,[\xi]\\ & \diagdown\!\diagup & \\ & \bullet\,[\mu] & \end{smallmatrix} \,\middle|\, \nu, \mu, \eta = 1, \ldots, N \right\},
\end{aligned} \tag{2.10}$$

where we used the representation $\left[\bullet^{[\nu]}\right]^{[\mu]} = \begin{smallmatrix}\bullet & [\nu]\\ | & \\ \bullet & [\mu]\end{smallmatrix}$, $\left[\left[\bullet^{[\xi]}\right]^{[\nu]}\right]^{[\mu]} = \begin{smallmatrix}\bullet & [\xi]\\ | & \\ \bullet & [\nu]\\ | & \\ \bullet & [\mu]\end{smallmatrix}$ as well as $\left[\bullet^{[\nu]}, \bullet^{[\xi]}\right]^{[\mu]} = \begin{smallmatrix}\bullet\,[\nu] & & \bullet\,[\xi]\\ & \diagdown\!\diagup & \\ & \bullet\,[\mu] & \end{smallmatrix}$. Lastly, the *symmetry* σ and *densitity* γ of τ from (2.9) are defined by

$$\begin{aligned}
\sigma(\tau) &= \prod_{j=1}^{r} m_j!\sigma(\tau_j), \quad \sigma(\bullet^{[\nu]}) = 1, \quad \nu = 1, \ldots, N,\\
\gamma(\tau) &= |\tau| \prod_{i=1}^{k} \gamma(\tau_i), \quad \gamma(\bullet^{[\nu]}) = 1, \quad \nu = 1, \ldots, N.
\end{aligned} \tag{2.11}$$

Observe that σ depends on the coloring of τ, while γ does not since already $|\tau|$ is independent of the coloring. For instance we find $\sigma(\left[\bullet^{[1]}, \bullet^{[2]}\right]^{[3]}) = 1$ since the children are not identical, while $\sigma(\left[\bullet^{[1]}, \bullet^{[1]}\right]^{[3]}) = 2$. Meanwhile, we observe $\gamma(\left[\bullet^{[1]}, \bullet^{[2]}\right]^{[3]}) = \gamma(\left[\bullet^{[1]}, \bullet^{[1]}\right]^{[3]}) = 3$. The symmetry and density are crucial quantities to describe the expansions of the analytical solution as we will see in the next subsection.

2.2.2 Elementary Differentials

For the following analysis, we assume for simplicity that the system (2.5) is autonomous, i. e. $\mathbf{f}^{[\nu]}(\mathbf{y}, t) = \mathbf{f}^{[\nu]}(\mathbf{y})$. We first introduce elementary differentials $\mathcal{F}\colon NT \to \mathcal{C}(\mathbb{R}^d, \mathbb{R}^d)$ for colored trees, see [AMSS97], which are recursively defined by

$$\begin{aligned}
\mathcal{F}(\bullet^{[\nu]})(\mathbf{y}) &= \mathbf{f}^{[\nu]}(\mathbf{y}),\\
\mathcal{F}([\tau_1, \ldots, \tau_k]^{[\nu]})(\mathbf{y}) &= \sum_{i_1, \ldots, i_k = 1}^{d} \partial_{i_1 \ldots i_k} \mathbf{f}^{[\nu]}(\mathbf{y}) \mathcal{F}_{i_1}(\tau_1)(\mathbf{y}) \cdots \mathcal{F}_{i_k}(\tau_k)(\mathbf{y})
\end{aligned} \tag{2.12}$$

for $\nu = 1, \ldots, N$. An important result in [AMSS97, But16] is the representation of the analytical solution of (2.5) in terms of an NB-series

$$\mathrm{NB}(u, \mathbf{y}) = \mathbf{y} + \sum_{\tau \in NT} \frac{\Delta t^{|\tau|}}{\sigma(\tau)} u(\tau) \mathcal{F}(\tau)(\mathbf{y}),$$

where $u\colon NT \to \mathbb{R}$, $\mathbf{y} \in \mathbb{R}^d$ and σ is the previously introduced symmetry. Note that $\mathrm{NB}(u, \mathbf{y})$ is defined only if $\mathbf{f}^{[\mu]} \in \mathcal{C}^\infty$ for $\mu = 1, \ldots, N$. For $\mathbf{f}^{[\mu]} \in \mathcal{C}^{p+1}$, we truncate the NB-series and introduce

$$\mathrm{NB}_p(u, \mathbf{y}) = \mathbf{y} + \sum_{\tau \in NT_p} \frac{\Delta t^{|\tau|}}{\sigma(\tau)} u(\tau) \mathcal{F}(\tau)(\mathbf{y}),$$

and point out that $\mathrm{NB}_0(u,\mathbf{y}) = \mathbf{y}$. With that, we can formulate a theorem concerning the NB-series expansion of the solution to the differential equation (2.5) at some time $t+\Delta t$.

Theorem 2.5 ([AMSS97, Theorem 1])**.** Let the functions $\mathbf{f}^{[\mu]}$ from (2.5) satisfy $\mathbf{f}^{[\mu]} \in \mathcal{C}^{p+1}$ for $\mu = 1,\dots,N$. If $\mathbf{y}$ solves (2.5), then

$$\mathbf{y}(t+\Delta t) = \mathrm{NB}_p(\tfrac{1}{\gamma}, \mathbf{y}(t)) + \mathcal{O}(\Delta t^{p+1}),$$

where γ is the density defined in (2.11).

The numerical solution given by one step of an ARK method can also be written as an NB-series $\mathrm{NB}(u,\mathbf{y}^n)$, with coefficients u recursively determined by

$$\begin{aligned}
u(\tau) &= \sum_{\nu=1}^{N}\sum_{i=1}^{s} b_i^{[\nu]} g_i^{[\nu]}(\tau), & & \\
g_i^{[\nu]}(\bullet^{[\mu]}) &= \delta_{\nu\mu}, & & \nu,\mu = 1,\dots,N, \\
g_i^{[\nu]}([\tau_1,\dots,\tau_l]^{[\mu]}) &= \delta_{\nu\mu}\prod_{j=1}^{l} d_i(\tau_j), & & \nu,\mu = 1,\dots,N \text{ and} \\
d_i(\tau) &= \sum_{\nu=1}^{N}\sum_{j=1}^{s} a_{ij}^{[\nu]} g_j^{[\nu]}(\tau) & &
\end{aligned} \tag{2.13}$$

with the Kronecker delta $\delta_{\nu\mu}$, see [AMSS97]. From Theorem 2.5 and the fact that elementary differentials are linearly independent [But16, AMSS97], we obtain the following result.

Theorem 2.6 ([AMSS97])**.** An ARK method (2.7) applied to (2.5) with $\mathbf{f}^{[\nu]} \in \mathcal{C}^{p+1}$ is of order p if and only if u determined by (2.13) satisfies

$$u(\tau) = \frac{1}{\gamma(\tau)} \qquad \text{for all } \tau \in NT_p. \tag{2.14}$$

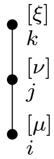
Remark 2.7. Based on [But16, Lemma 312B], the value of u can be read off from a colored and labeled rooted tree τ. Thereby, a node labeled by i and colored in μ is represented by $\bullet_i^{[\mu]}$. It is convenient to also associate with each edge a color; we denote the edge connecting parent node i to child node j by $e_{ij}^{[v]}$, where ν is the color of node j. We denote the set of labels by $L(\tau)$ and the set of colored edges by $E(\tau)$.

For computing $u(\tau)$, let the root of τ be labeled by i and colored in μ. Then form the product

$$b_i^{[\mu]} \prod_{e_{jk}^{[\nu]} \in E(\tau)} a_{jk}^{[\nu]}$$

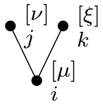
and sum over all elements of $L(\tau)$ ranging over the index set $\{1,\dots,s\}$. The result of the sum equals $u(\tau)$.

Example 2.8. We label the colored rooted tree $\tau = \left[\left[\bullet^{[\xi]}\right]^{[\nu]}\right]^{[\mu]}$ and represent the result by



so that $u(\tau) = \sum_{i,j,k=1}^{s} b_i^{[\mu]} a_{ij}^{[\nu]} a_{jk}^{[\xi]}$ since $E(\tau) = \left\{e_{ij}^{[\nu]}, e_{jk}^{[\xi]}\right\}$ and $L(\tau) = \{i, j, k\}$.

For the tree $\tau = \left[\bullet^{[\nu]}, \bullet^{[\xi]}\right]^{[\mu]}$, which we label and represent by



the value of $u(\tau)$ is $\sum_{i,j,k=1}^{s} b_i^{[\mu]} a_{ij}^{[\nu]} a_{ik}^{[\xi]}$ as $E(\tau) = \left\{e_{ij}^{[\nu]}, e_{ik}^{[\xi]}\right\}$ and $L(\tau) = \{i, j, k\}$.

2.3 Linear Stability of Runge–Kutta Methods

The linear stability of a time integration method is usually tackled by the application of the scheme to the linear test equation

$$y'(t) = \lambda y(t), \quad \lambda \in \mathbb{C}^- = \{z \in \mathbb{C} \mid \mathrm{Re}(z) < 0\}, \tag{2.15}$$

which was introduced in 1963 by Dahlquist [Dah63]. The basic idea behind stability is that the numerical method should replicate the qualitative behavior of the analytic solution in some sense. The central notion linked to the Dahlquist equation is A-stability.

Definition 2.9 ([Dah63])**.** A time integration method is called *A-stable*, if the sequence of iterates y^n of the method tends to zero, as $n \to \infty$, when applied with fixed $\Delta t > 0$ to any differential equation of the form (2.15).

The reason why A-stability is of interest may be based on the following heuristic. Consider the difference of two solutions $\mathbf{w}, \mathbf{y}$, denoted by $\mathbf{u}$, of a nonlinear system $\mathbf{y}' = \mathbf{f}(\mathbf{y})$. Note that $\mathbf{u} = \mathbf{w} - \mathbf{y}$ can be seen as a perturbation. We then linearize the disturbed system $(\mathbf{y} + \mathbf{u})' = \mathbf{w}' = \mathbf{f}(\mathbf{w}) = \mathbf{f}(\mathbf{y} + \mathbf{u})$, which results in

$$\mathbf{u}' = (\mathbf{y} + \mathbf{u})' - \mathbf{y}' = \mathbf{f}(\mathbf{y} + \mathbf{u}) - \mathbf{f}(\mathbf{y}) \approx \mathbf{f}(\mathbf{y}) + \mathbf{Df}(\mathbf{y})\mathbf{u} - \mathbf{f}(\mathbf{y}) = \mathbf{Df}(\mathbf{y})\mathbf{u}.$$

Freezing the Jacobian $\mathbf{Df}(\mathbf{y})$ at a given time T yields a linear system $\mathbf{u}' = \boldsymbol{\Lambda}\mathbf{u}$ for the perturbation, where $\boldsymbol{\Lambda}$ possibly has complex eigenvalues λ. Moreover, the perturbation should disappear as $t \to \infty$, and hence, we consider $\lambda \in \mathbb{C}^-$ in (2.15) rather than $\lambda \in \mathbb{C}$. Since this heuristic is not rigorous, I would rather prefer to point out the following motivation. A numerical method that is not capable of mimicking the behavior of the analytical solution to a (scalar) linear test problem is not worth considering for more complex problems.

Later, the notion of L-stability was introduced [HNW93]. Moreover, for the case of $\lambda \in \mathbb{R}^-$, the notions A_0-stable [Cry73] and L_0-stable arise [TGA96]. We also note that more theories have been developed, some of which are suitable for the analysis of RK schemes applied to stiff nonlinear ODEs [DK06, SVV18].

For multistep methods zero-stability is a fundamental notion [SM03]. Some stability properties even introduce a class of schemes, e.g. so-called positive and elementary stable non-standard (PESN) schemes [DK06].

In this work we focus on A-stability. In the case of an RK method, there exists a rational function $R=\frac{P}{Q}$ such that the method applied to the Dahlquist equation (2.15) reads $y^{n+1}=R(\Delta t\lambda)y^n$. Hence, the RK method is A-stable if and only if $|R(z)|<1$ for all $z\in\mathbb{C}^-$, which is why R is also called the *stability function* of the Runge–Kutta method. Indeed, if we apply the RK method to a linear system

$$\mathbf{y}'=\mathbf{\Lambda}\mathbf{y},\quad \mathbf{y}(0)=\mathbf{y}^0,\quad \sigma(\mathbf{\Lambda})\subseteq\mathbb{C}^-, \tag{2.16}$$

where $\sigma(\mathbf{\Lambda})$ denotes the spectrum, then the RK method has the same stability properties as applied to the Dahlquist equation with λ passing through the eigenvalues of $\mathbf{\Lambda}$, see for instance [DB02, Chapter 6]. Hence, if an RK method is A-stable, then $\lim_{n\to\infty}\mathbf{y}^n=\mathbf{0}$ holds also for general linear problems (2.16).

Even though Definition 2.9 does not require the method to be linear, some nonlinear schemes are constructed only for systems of equations as is the case for modified Patankar (MP) methods, see Chapter 3. Even more, as mentioned in the introduction, the investigation of

$$\begin{pmatrix} y_1'(t)\\ y_2'(t)\end{pmatrix}=\begin{pmatrix}\lambda y_1(t)\\ -\lambda y_1(t)\end{pmatrix},\quad \lambda\in\mathbb{R}^-,$$

whose first component represents the Dahlquist equation with $\lambda\in\mathbb{R}^-$ is not enough for understanding the stability properties of an MP method applied to more complex linear systems [IKM22a, IKM22b]. Hence, for nonlinear methods it is necessary to investigate general linear systems $\mathbf{y}'=\mathbf{\Lambda}\mathbf{y}$ rather than a scalar equation. To generalize the notion of A-stability in a meaningful way also for nonlinear methods, we consider stability in the sense of Lyapunov, which we recall in the upcoming section.

2.4 Stability in the Sense of Lyapunov

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 $\mathcal{C}^1$-map $\mathbf{f}$.

Dahlquist already considered to generalize the notion of A-stability in [Dah63] by considering stability in the sense of Lyapunov, which is defined for arbitrary systems of ODEs. Here, the stability near steady states is investigated.

Definition 2.10. Let $\mathbf{y}^*\in\mathbb{R}^N$ be a steady state solution of a differential equation $\mathbf{y}'=\mathbf{f}(\mathbf{y})$, that is $\mathbf{f}(\mathbf{y}^*)=\mathbf{0}$.

a) Then $\mathbf{y}^*$ 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 $\|\mathbf{y}(t)-\mathbf{y}^*\|\to 0$ for $t\to\infty$, we call $\mathbf{y}^*$ *asymptotically stable.*

c) A steady state solution that is not Lyapunov stable is said to be *unstable.*

In the following, we will also briefly speak of stability instead of Lyapunov stability. Note that in contrast to A-stability, these notions are only *global* if δ

and c can be chosen arbitrarily large. Considering the linear system (2.16), the stability of $\mathbf{y}^*$ is fully determined by the spectrum $\sigma(\mathbf{\Lambda})$.

Theorem 2.11. ([DB02, Theorem 3.23]) A steady state $\mathbf{y}^*$ of $\mathbf{y}' = \mathbf{\Lambda}\mathbf{y}$ with a matrix $\mathbf{\Lambda} \in \mathbb{R}^{N\times N}$

a) is stable if and only if $\max_{\lambda\in\sigma(\mathbf{\Lambda})} \operatorname{Re}(\lambda) \leq 0$ and all λ with $\operatorname{Re}(\lambda) = 0$ are associated with a Jordan block of size 1.

b) is asymptotically stable if and only if $\max_{\lambda\in\sigma(\mathbf{\Lambda})} \operatorname{Re}(\lambda) < 0$.

As we are interested in numerical schemes mimicking the stability behavior of the exact solution, we shall consider the following definition, noting that steady states should correspond to fixed points of the method.

Definition 2.12. 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) Then $\mathbf{y}^*$ 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 $\|\mathbf{y}^n - \mathbf{y}^*\| \to 0$ for $n \to \infty$, we call $\mathbf{y}^*$ *asymptotically stable.*

c) A fixed point that is not Lyapunov stable is said to be *unstable.*

As before, we may only speak of stability in the following. For linear methods, such as RK schemes, we have the following result.

Theorem 2.13 ([DB02, Theorem 3.33])**.** A fixed point $\mathbf{y}^*$ of $\mathbf{y}^{n+1} = \mathbf{R}\mathbf{y}^n$ with $\mathbf{R} \in \mathbb{R}^{N\times N}$

a) is stable if and only if the spectral radius ρ satisfies $\rho(\mathbf{R}) \leq 1$ and all $\lambda \in \sigma(\mathbf{R})$ with $|\lambda| = 1$ are associated with a Jordan block of size 1.

b) is asymptotically stable if and only if $\rho(\boldsymbol{R}) < 1$.

Remark 2.14. According to Theorem 2.11, $y^* = 0$ is the unique globally asymptotically stable solution of the Dahlquist equation. Also, Theorem 2.13 tells us that an RK method is A-stable if and only 0 is an asymptotically stable fixed point of the method when applied to the Dahlquist equation. This also demonstrates that 0 is a globally asymptotically stable fixed point of the A-stable RK method. We also note that in some literature, such as [But16, HW10], A-stability of an RK scheme is defined by requiring $|R(z)| \leq 1$ for all $z \in \overline{\mathbb{C}^-} = \{z \in \mathbb{C} \mid \operatorname{Re}(z) \leq 0\}$. The idea behind this adaptation is that we may only require that the numerical solution is bounded for bounded solutions of the Dahlquist equation. However, with this notion of A-stability, the generalization to linear systems is more involved as Theorem 2.13 suggests.

If the method is not linear, the stability properties are a priori only of local nature and their investigation is more complex. As stated by the next theorem, it is in some cases sufficient to investigate the linearized method in order to understand the stability properties of a fixed point.

Theorem 2.15 ([SH98, Theorem 1.3.7])**.** Let $\mathbf{y}^{n+1} = \mathbf{g}(\mathbf{y}^n)$ be an iteration scheme with fixed point $\mathbf{y}^*$. Suppose the Jacobian $\mathbf{Dg}(\mathbf{y}^*)$ exists. Then

a) $\mathbf{y}^*$ is asymptotically stable if $\rho(\mathbf{Dg}(\mathbf{y}^*)) < 1$.

b) $\mathbf{y}^*$ is unstable if $\rho(\mathbf{Dg}(\mathbf{y}^*)) > 1$.

The above theorem gives sufficient conditions for the stability of fixed points that are *hyperbolic* in the following sense.

Definition 2.16 ([SH98, Definition 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 λ of $\mathbf{Dg}(\mathbf{y}^*)$. If a fixed point is not hyperbolic, it is called *non-hyperbolic*.

A generalization of Theorem 2.15 is the Hartman-Grobman Theorem, which states that a nonlinear iteration scheme and its linearization share the same behavior near hyperbolic fixed points, see [SH98, Theorem 1.6.2] for the precise statement.

In this work, we will also analyze schemes that require $\dim(\ker(\mathbf{\Lambda})) = k > 0$. In such a case, the linear system $\mathbf{y}' = \mathbf{\Lambda}\mathbf{y}$ possesses a subspace of steady state solutions, each of which can be stable according to Theorem 2.11 but none of them is asymptotically stable. If the numerical method is steady state preserving, it thus possesses a subspace of fixed points, each of them being non-hyperbolic as we will find out in Chapter 5. Hence, it is also of high importance to understand the stability of non-hyperbolic fixed points. However, for schemes outside the class of general linear methods the stability behavior of a single non-hyperbolic fixed point is in general not captured by the eigenvalues of the corresponding Jacobian, i. e. is not guaranteed by $\rho(\mathbf{D}(\mathbf{g}(\mathbf{y}^*))) = 1$ as the following example illustrates.

Example 2.17 ([Osi12])**.** Consider the generating map defined by

$$\mathbf{g}(x, y) = \left(x + xy, \frac{1}{2}(y + x^2) + 2x^2y + y^3\right)^T. \tag{2.17}$$

We observe $\mathbf{g}(0,0) = (0,0)^T$ and $\mathbf{Dg}(0,0) = \left(\begin{smallmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{smallmatrix}\right)$ Now, defining $h(x) = x^2$, we see that the graph of h is invariant under $\mathbf{g}$ since

$$\mathbf{g}(x, h(x)) = (x + x^3, x^2 + 2x^4 + x^6)^T = (x + x^3, h(x + x^3))^T.$$

Focusing on the x-component, i. e. $x + x^3 = x(1 + x^2)$, we find that the iterates distance from the origin along the graph of h, see Figure 2.1 for an illustration. From this, we can conclude that the origin is unstable even though the eigenvalues

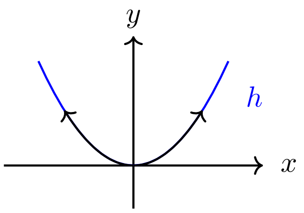


Figure 2.1: Graph of h defined by $h(x) = x^2$. The arcs indicate the action of $\mathbf{g}$ from (2.17) on the graph of h.

of the Jacobian are 1 and $\frac{1}{2}$.

This example demonstrates that, in general, higher-order terms have to be included within the stability analysis of nonlinear methods. One possibility to decrease the complexity of such a stability analysis is to use the center manifold theory, which allows to assess the stability based on a corresponding iteration on a lower dimensional manifold. Indeed, in Example 2.17 the map h represents the center manifold.

2.5 Center Manifold Theory

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

$$\mathbf{G}(\mathbf{w}_1, \mathbf{w}_2) = \begin{pmatrix} \mathbf{U}\mathbf{w}_1 + \mathbf{u}(\mathbf{w}_1, \mathbf{w}_2) \\ \mathbf{V}\mathbf{w}_2 + \mathbf{v}(\mathbf{w}_1, \mathbf{w}_2) \end{pmatrix}, \tag{2.18}$$

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 $|\lambda| = 1$ holds for all eigenvalues λ of $\mathbf{U}$ and each eigenvalue μ of $\mathbf{V}$ satisfies $|\mu| < 1$. The functions $\mathbf{u}$ and $\mathbf{v}$ are in $\mathcal{C}^1$ and $\mathbf{u}, \mathbf{v}$ as well as their first order derivatives vanish at the origin, that is

$$\mathbf{u}(\mathbf{0},\mathbf{0}) = \mathbf{0}, \qquad \mathbf{Du}(\mathbf{0},\mathbf{0}) = \mathbf{0}, \qquad \mathbf{v}(\mathbf{0},\mathbf{0}) = \mathbf{0}, \qquad \mathbf{Dv}(\mathbf{0},\mathbf{0}) = \mathbf{0},$$

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 $\mathbf{y}^*$ as we point out in the proof of Theorem 5.4.

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.

Theorem 2.18. (Center Manifold Theorem, [MM76, Theorem 2.1, Remark 2.6]) Let $\mathbf{G}$ be defined as in (2.18) with Lipschitz continuous derivatives on $\mathcal{M}$.

a) (Existence): There exists a center manifold for $\mathbf{G}$, which is locally representable as the graph of a function $\mathbf{h}\colon \mathbb{R}^m \to \mathbb{R}^l$. This means, for some $\epsilon > 0$ there exists a $\mathcal{C}^1$-function $\mathbf{h}\colon \mathbb{R}^m \to \mathbb{R}^l$ with $\mathbf{h}(\mathbf{0}) = \mathbf{0}$ and $\mathbf{Dh}(\mathbf{0}) = \mathbf{0}$ such that $\|\mathbf{w}_1^0\|, \|\mathbf{w}_1^1\| < \epsilon$ and $(\mathbf{w}_1^1, \mathbf{w}_2^1)^T = \mathbf{G}(\mathbf{w}_1^0, \mathbf{h}(\mathbf{w}_1^0))$ imply $\mathbf{w}_2^1 = \mathbf{h}(\mathbf{w}_1^1)$.

b) (Local Attractivity): If in addition to a) the iterates $(\mathbf{w}_1^n, \mathbf{w}_2^n)^T$ generated by

$$\begin{pmatrix} \mathbf{w}_1^{n+1} \\ \mathbf{w}_2^{n+1} \end{pmatrix} = \mathbf{G}(\mathbf{w}_1^n, \mathbf{w}_2^n) = \begin{pmatrix} \mathbf{U}\mathbf{w}_1^n + \mathbf{u}(\mathbf{w}_1^n, \mathbf{w}_2^n) \\ \mathbf{V}\mathbf{w}_2^n + \mathbf{v}(\mathbf{w}_1^n, \mathbf{w}_2^n) \end{pmatrix}, \quad \begin{pmatrix} \mathbf{w}_1^0 \\ \mathbf{w}_2^0 \end{pmatrix} \in \mathcal{M} \tag{2.19}$$

satisfy $\|\mathbf{w}_1^n\|, \|\mathbf{w}_2^n\| < \epsilon$ 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 $n \to \infty$, i. e. $\|\mathbf{w}_2^n - \mathbf{h}(\mathbf{w}_1^n)\| \to 0$ for $n \to \infty$.

As will be seen in Theorem 2.19, the existence of a center manifold enables the investigation of the stability properties of the origin based on a system with reduced dimension. This reduced system is obtained by restricting (2.18) to the

[1] See the proof of Theorem 5.4 for the details of this transformation.

center manifold, i. e. using $\mathbf{w}_2 = \mathbf{h}(\mathbf{w}_1)$ which leads to the map

$$\mathcal{G}(\mathbf{w}_1) = \mathbf{U}\mathbf{w}_1 + \mathbf{u}(\mathbf{w}_1, \mathbf{h}(\mathbf{w}_1)). \tag{2.20}$$

Theorem 2.19. ([Car81, Theorem 8]) (Stability): Suppose the fixed point $\mathbf{0} \in \mathbb{R}^m$ of $\mathcal{G}$ from (2.20) is stable, asymptotically stable or unstable. Then the fixed point $\mathbf{0} \in \mathbb{R}^N$ of $\mathbf{G}$ from (2.18) is stable, asymptotically stable or unstable, respectively.

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

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

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

Theorem 2.20. ([Car81, Theorem 7]) Let $\mathbf{h}$ be a center manifold for $\mathbf{G}$ and $\mathbf{\Phi}$ be a $\mathcal{C}^1(\mathbb{R}^m, \mathbb{R}^l)$-map with $\mathbf{\Phi}(\mathbf{0}) = \mathbf{0}$ and $\mathbf{D\Phi}(\mathbf{0}) = \mathbf{0}$. If

$$\mathbf{\Phi}(\mathbf{U}\mathbf{w}_1 + \mathbf{u}(\mathbf{w}_1, \mathbf{\Phi}(\mathbf{w}_1))) - (\mathbf{V}\mathbf{\Phi}(\mathbf{w}_1) + \mathbf{v}(\mathbf{w}_1, \mathbf{\Phi}(\mathbf{w}_1))) = \mathcal{O}(\lVert \mathbf{w}_1 \rVert^q)$$

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

Before we go to theoretical fundamentals on production-destruction-rest systems, let us summarize the sections on stability. We started with A-stability which is the central notion for capturing the linear stability properties of general linear methods such as Runge–Kutta schemes. However, we discussed that analyzing a scalar equation is not sufficient to capture the stability behavior of nonlinear methods. Hence, we generalized A-stability by considering stability in the sense of Lyapunov. Moreover, we presented tools for analyzing general numerical methods with hyperbolic and non-hyperbolic fixed points, where the analysis of the latter is more challenging as more techniques such as the approximation of the center manifold is required. However, for our purposes this is the interesting case when analyzing Patankar-type methods.

2.6 Production-Destruction-Rest Systems

In this work we are interested in methods that are capable of producing positive approximations for any chosen time step size. First focusing on autonomous problems, it is convenient to rewrite the system of ODEs into the form of a *production-destruction system* (PDS)

$$y_k'(t) = f_k(\mathbf{y}(t)) = \sum_{\nu=1}^{N} (p_{k\nu}(\mathbf{y}(t)) - d_{k\nu}(\mathbf{y}(t))), \quad \mathbf{y}(0) = \mathbf{y}^0 > \mathbf{0}, \tag{2.21}$$

where $p_{k\nu}(\mathbf{y}(t)), d_{k\nu}(\mathbf{y}(t)) \geq 0$ for all $\mathbf{y}(t) \geq 0$. Note that every real valued right-hand side f_k can be split into production and destruction terms setting

$$p_{k1}(\mathbf{y}) = \max\{0, f_k(\mathbf{y})\}, \quad d_{k1}(\mathbf{y}) = -\min\{0, f_k(\mathbf{y})\}, \quad p_{k\nu} = d_{k\nu} = 0, \quad \nu \neq 1.$$

However, using this splitting the production and destruction terms are generally not differentiable. Nevertheless, in view of Theorem 2.3 we note that if f_k is locally Lipschitz continuous, then so are p_{k1} and d_{k1} as they are the composition of two locally Lipschitz mappings.

Definition 2.21. The PDS (2.21) is called *positive*, if $\mathbf{y}(0) > \mathbf{0}$ implies $\mathbf{y}(t) > \mathbf{0}$ for all $t > 0$. Similarly, a *non-negative* PDS are defined.

Proposition 2.22 ([BDM03])**.** For non-negative initial data, the PDS (2.21) is non-negative if $d_{k\nu}(\mathbf{y}) \to 0$ as $y_k \to 0$ for $k, \nu = 1, \dots, N$.

Definition 2.23. We call the PDS (2.21) *conservative*, if $p_{k\nu} = d_{\nu k}$ for all $k, \nu = 1, \dots, N$. If in addition we have $p_{kk} = d_{kk} = 0$, the PDS is called *fully conservative*.

Remark 2.24. Since $p_{kk} = d_{kk}$ cancel out in (2.21) for a conservative PDS we can assume without loss of generality that $p_{kk} = d_{kk} = 0$, i. e. that the PDS is always fully conservative.

For a conservative PDS, we know that $\sum_{k=1}^N y_k'(t) = 0$, and hence, the sum of the constituents remains constant in time. In general, if a linear combination $\mathbf{n}^T\mathbf{y}$ remains constant in time, we call it a *linear invariant*.

It is also worth mentioning that the additive splitting into production and destruction terms is not uniquely determined. For instance, considering

$$\mathbf{y}' = \begin{pmatrix} y_2 + y_4 - y_1 \\ y_1 - y_2 \\ y_1 - y_3 \\ y_3 - y_1 - y_4 \end{pmatrix},$$

the terms $p_{14}(\mathbf{y}) = y_4$, $p_{12}(\mathbf{y}) = y_2$ and $p_{43}(\mathbf{y}) = y_3$ are a straightforward choice, however, both,

$$p_{21}(\mathbf{y}) = y_1, \quad p_{34}(\mathbf{y}) = y_1$$

and

$$p_{31}(\mathbf{y}) = y_1, \quad p_{24}(\mathbf{y}) = y_1$$

complete the splitting into a PDS, where we set $p_{mn} = 0$ for the remaining production terms and $p_{k\nu} = d_{\nu k}$.

In this work, we are also interested in positive PDS which are non-autonomous and not conservative. For a transparent notation we split the PDS into a conservative part and rest terms, leading to a *production-destruction-rest system* (PDRS)

$$y_k'(t) = f_k(\mathbf{y}(t), t) = r_k(\mathbf{y}(t), t) + \sum_{\nu=1}^N (p_{k\nu}(\mathbf{y}(t), t) - d_{k\nu}(\mathbf{y}(t), t)), \quad \mathbf{y}(0) = \mathbf{y}^0 > \mathbf{0} \tag{2.22}$$

with $k = 1, \dots, N$ and $p_{k\nu} = d_{\nu k}$. Additionally, the rest term is also split according to

$$r_k(\mathbf{y}(t), t) = r_k^p(\mathbf{y}(t), t) - r_k^d(\mathbf{y}(t), t) \tag{2.23}$$

with $r_k^p, r_k^d \geq 0$ for $t \geq 0$, $\mathbf{y}(t) \geq \mathbf{0}$. Note that r_k^p and r_k^d can always be constructed, for example by using the functions max and min as above. The autonomous version of the PDRS (2.22) was already considered in [TÖR22] and the existence, uniqueness and positivity of the solution of (2.22) was discussed in [FS11a]. In what follows, we are assuming that such a positive solution exists. For later use it is also beneficial to rewrite the PDS as an additive splitting of the form (2.5).

Remark 2.25. Any PDRS (2.22), (2.23) may be rewritten as an additive splitting of the form

$$\mathbf{f}(\mathbf{y}(t), t) = \sum_{\nu=1}^{N+1} \mathbf{f}^{[\nu]}(\mathbf{y}(t), t) \in \mathbb{R}^N$$

using $\mathbf{f}^{[N+1]}(\mathbf{y}(t), t) = (r_1^p(\mathbf{y}(t), t), \dots, r_N^p(\mathbf{y}(t), t))^T$ and

$$f_k^{[\nu]}(\mathbf{y}(t), t) = \begin{cases} p_{k\nu}(\mathbf{y}(t), t), & k \neq \nu, \\ -\left(r_k^d(\mathbf{y}(t), t) + \sum_{\mu=1}^N d_{k\mu}(\mathbf{y}(t), t)\right), & k = \nu \end{cases}$$

for $k, \nu = 1, \dots, N$. To see this, we first point out that $p_{kk} = d_{kk} = 0$ can be assumed, see Remark 2.24. Hence,

$$\begin{aligned} f_k = \sum_{\nu=1}^{N+1} f_k^{[\nu]} &= \sum_{\substack{\nu=1 \\ \nu \neq k}}^{N} f_k^{[\nu]} + f_k^{[k]} + f_k^{[N+1]} \\ &= \sum_{\nu=1}^{N} p_{k\nu} - \left(r_k^d + \sum_{\mu=1}^{N} d_{k\mu} \right) + r_k^p = r_k + \sum_{\nu=1}^{N} (p_{k\nu} - d_{k\nu}). \end{aligned}$$

Chapter 3

Numerical Schemes

In this chapter we review positivity-preserving schemes that additionally preserve at least one linear invariant. For other recent approaches which facilitate positive and conservative numerical approximations, we refer to [AGKM21, NRK21b, BIM21], some of which even conserve all linear invariants. The following schemes are one-step methods, for which we briefly recall the definition of unconditional conservativity and positivity.

Definition 3.1. Let $\mathbf{y}^n$ denote an approximation of $\mathbf{y}(t_n)$ at time level t_n. The corresponding one-step method is called

- *unconditionally conservative*, if

  $$\sum_{k=1}^{N} y_k^{n+1} = \sum_{k=1}^{N} y_k^n$$

  is satisfied for all $n \in \mathbb{N}_0$ and $\Delta t > 0$.

- *unconditionally positive*, if $\mathbf{y}^n > 0$ implies $\mathbf{y}^{n+1} > 0$ for all $n \in \mathbb{N}_0$ and $\Delta t > 0$.

3.1 Non-standard Additive Runge–Kutta Methods

Non-standard additive Runge–Kutta (NSARK) methods are based on ARK schemes (2.7), where the Butcher tableau is allowed to also depend on the step size and the solution. In particular, NSARK methods are of the form

$$\begin{aligned}\mathbf{y}^{(i)} &= \mathbf{y}^n + \Delta t \sum_{j=1}^{s}\sum_{\nu=1}^{N} a_{ij}^{[\nu]}(\mathbf{y}^n, t_n, \Delta t)\mathbf{f}^{[\nu]}(\mathbf{y}^{(j)}, t_n + c_j\Delta t), \quad i = 1, \dots, s,\\ \mathbf{y}^{n+1} &= \mathbf{y}^n + \Delta t \sum_{j=1}^{s}\sum_{\nu=1}^{N} b_{j}^{[\nu]}(\mathbf{y}^n, t_n, \Delta t)\mathbf{f}^{[\nu]}(\mathbf{y}^{(j)}, t_n + c_j\Delta t).\end{aligned} \tag{NSARK}$$

Note that the stages $\mathbf{y}^{(i)} = \mathbf{y}^{(i)}(\mathbf{y}^n)$ may be interpreted as functions of $\mathbf{y}^n$, so that the dependence of $a_{ij}^{[\nu]}, b_j^{[\nu]}$ on $\mathbf{y}^n$ might be given implicitly. As a result of this notation, an NSARK method is called *explicit*, if the matrices $\mathbf{A}^{[\nu]}$ are strict lower left triangular matrices and the dependence of $\mathbf{A}^{[\nu]}$ as well as $\mathbf{b}^{[\nu]}$ on $\mathbf{y}^n$ is only explicit. Otherwise, the NSARK method is called *implicit*.

As we will discover in this chapter, all MP methods based on RK schemes can be written as an NSARK method. Moreover, given a Butcher tableau defined by $\mathbf{A}, \mathbf{b}, \mathbf{c}$, the corresponding MP methods are of the form

$$\begin{aligned} a_{ij}^{[\nu]}(\mathbf{y}^n, t_n, \Delta t) &= a_{ij}\gamma_\nu^{[i]}(\mathbf{y}^n, t_n, \Delta t),\\ b_j^{[\nu]}(\mathbf{y}^n, t_n, \Delta t) &= b_j\delta_\nu(\mathbf{y}^n, t_n, \Delta t) \end{aligned} \tag{3.1}$$

for some scheme-dependent functions $\gamma_\nu^{[i]}$ and δ_ν, which we refer to as *non-standard weights* (NS weights). Investigating NSARK methods allows the comprehensive derivation of a general stability function as well as order conditions for different families of methods. In particular, it turns out that NSARK methods are a valuable formulation for the analysis of so-called modified Patankar–Runge–Kutta (MPRK) methods. Nevertheless, we will be able to deduce also some results for Geometric Conservative (GeCo) schemes in this work and discuss how to generalize or adapt NSARK schemes to investigate even more nonlinear methods.

The following proposition formulates sufficient conditions under which an NSARK scheme produces the same approximations for the transformed autonomous system mentioned in Remark 2.1.

Proposition 3.2. Let $\mathbf{A}, \mathbf{b}, \mathbf{c}$ describe an RK method satisfying $\sum_{j=1}^s a_{ij} = c_i$ and $\sum_{j=1}^s b_j = 1$. Let the stages of the corresponding NSARK method (NSARK) be uniquely determined for some $\Delta t > 0$ and transform the IVP (2.5) into the autonomous system $\mathbf{Y}'(t) = \sum_{\nu=1}^{N+1} \mathbf{F}^{[\nu]}(\mathbf{Y}(t))$ using

$$\mathbf{Y}(t) = \begin{pmatrix} \mathbf{y}(t)\\ t\end{pmatrix}, \quad \mathbf{F}^{[\mu]}(\mathbf{Y}(t)) = \begin{pmatrix} \mathbf{f}^{[\mu]}(\mathbf{Y}(t))\\ 0\end{pmatrix},\ 1 \leq \mu \leq N, \quad \mathbf{F}^{[N+1]}(\mathbf{Y}(t)) = \begin{pmatrix} \mathbf{0}\\ 1\end{pmatrix}.$$

If $\gamma_{N+1}^{[i]}, \delta_{N+1}^{[j]} = 1$, then the approximations for the solution of the IVP (2.5) using the NSARK method coincide irrespective of whether the autonomous or non-autonomous system is solved.

Proof. Since $\gamma_{N+1}^{[i]}, \delta_{N+1}^{[j]} = 1$, the NSARK method applied to the autonomous system reads

$$\begin{aligned} Y_k^{(i)} &= Y_k^n + \Delta t \sum_{j=1}^s \left(\sum_{\nu=1}^N a_{ij}^{[\nu]}(\mathbf{Y}^n, \Delta t) F_k^{[\nu]}(\mathbf{Y}^{(j)}) + a_{ij} F_k^{[N+1]}(\mathbf{Y}^{(j)}) \right),\\ Y_k^{n+1} &= Y_k^n + \Delta t \sum_{j=1}^s \left(\sum_{\nu=1}^N b_j^{[\nu]}(\mathbf{Y}^n, \Delta t) F_k^{[\nu]}(\mathbf{Y}^{(j)}) + b_j F_k^{[N+1]}(\mathbf{Y}^{(j)}) \right). \end{aligned} \tag{3.2}$$

Thus, for $k = N+1$ we find

$$\begin{aligned} t_{(i)} &= t_n + \Delta t \sum_{j=1}^s a_{ij} = t_n + c_i \Delta t,\\ t_{n+1} &= t_n + \Delta t \sum_{j=1}^s b_j = t_n + \Delta t. \end{aligned} \tag{3.3}$$

Furthermore, for $k \leq N$, we end up with

$$y_k^{(i)} = y_k^n + \Delta t \sum_{j=1}^s \sum_{\nu=1}^N a_{ij}^{[\nu]}(\mathbf{y}^n, t_n, \Delta t) f_k^{[\nu]}(\mathbf{y}^{(j)}, t_{(j)}),$$

$$y_k^{n+1} = y_k^n + \Delta t \sum_{j=1}^{s} \sum_{\nu=1}^{N} b_j^{[\nu]}(\mathbf{y}^n, t_n, \Delta t) f_k^{[\nu]}(\mathbf{y}^{(j)}, t_{(j)}).$$

Substituting (3.3) into these equations, the proof is finished by noting that the resulting systems always possess a unique solution due to our preconditions. □

As a consequence of Proposition 3.2 we may consider only autonomous problems for deriving order conditions, if the method satisfies the assumptions of the proposition.

3.2 Modified Patankar–Runge–Kutta

The main idea of modified Patankar–Runge–Kutta (MPRK) methods [BDM03, KM18a, KM18b, KM19b] is to apply an explicit Runge–Kutta (RK) method to a production-destruction systems (PDS) (2.21) and use the modified Patankar-trick. We extend this approach also to production-destruction-rest systems (PDRS) (2.22), (2.23) where we only apply the Patankar-trick to the rest term. This means, that r_k^p will not be weighted and r_k^d will be treated like a destruction term.

Definition 3.3. Given an explicit s-stage RK method described by a non-negative Butcher array, i. e. $\mathbf{A}, \mathbf{b}, \mathbf{c} \geq \mathbf{0}$ we define the corresponding MPRK schemes applied to (2.22), (2.23) by

$$\begin{aligned} y_k^{(i)} =& y_k^n + \Delta t \sum_{j=1}^{i-1} a_{ij} \Bigg(r_k^p(\mathbf{y}^{(j)}, t_n + c_j\Delta t) + \sum_{\nu=1}^{N} p_{k\nu}(\mathbf{y}^{(j)}, t_n + c_j\Delta t) \frac{y_\nu^{(i)}}{\pi_\nu^{(i)}} \\ &- \left(r_k^d(\mathbf{y}^{(j)}, t_n + c_j\Delta t) + \sum_{\nu=1}^{N} d_{k\nu}(\mathbf{y}^{(j)}, t_n + c_j\Delta t) \right) \frac{y_k^{(i)}}{\pi_k^{(i)}} \Bigg), \quad k = 1, \ldots, s, \\ y_k^{n+1} =& y_k^n + \Delta t \sum_{j=1}^{s} b_j \Bigg(r_k^p(\mathbf{y}^{(j)}, t_n + c_j\Delta t) + \sum_{\nu=1}^{N} p_{k\nu}(\mathbf{y}^{(j)}, t_n + c_j\Delta t) \frac{y_\nu^{(i)}}{\sigma_\nu} \\ &- \left(r_k^d(\mathbf{y}^{(j)}, t_n + c_j\Delta t) + \sum_{\nu=1}^{N} d_{k\nu}(\mathbf{y}^{(j)}, t_n + c_j\Delta t) \right) \frac{y_k^{(i)}}{\sigma_k} \Bigg), \end{aligned} \tag{MPRK}$$

where $\pi_\nu^{(i)}, \sigma_\nu$ are the so-called *Patankar-weight denominators* (PWDs) and positive for any $\Delta t \geq 0$ as well as independent of the corresponding numerators $y_k^{(i)}$ and y_k^{n+1}, respectively.

MPRK schemes are of considerable interest and widely used such as in the context of ecosystems [HB10a, HB10b, WHK13, BMZ07, BMZ09, MB10] or ocean models [SD17, BBK+06]. Further applications can be found in the context of magneto-thermal winds [Gre17] or warm-hot intergalactic mediums [KM10] as well as in that of the SIR epidemic model [WS22].

Remark 3.4. In matrix notation, (MPRK) can be rewritten as

$$\begin{aligned} \mathbf{M}^{(i)}\mathbf{y}^{(i)} &= \mathbf{y}^n + \Delta t \sum_{j=1}^{i-1} a_{ij} \mathbf{r}^p(\mathbf{y}^{(j)}, t_n + c_j\Delta t), \quad i = 1, \ldots, s, \\ \mathbf{M}\mathbf{y}^{n+1} &= \mathbf{y}^n + \Delta t \sum_{j=1}^{s} b_j \mathbf{r}^p(\mathbf{y}^{(j)}, t_n + c_j\Delta t), \end{aligned} \tag{3.4}$$

where $\mathbf{r}^p = (r_1^p, \ldots, r_N^p)^T$ and $\mathbf{M}^{(i)} = (m_{k\nu}^{(i)})_{1\le k,\nu\le N}$ with

$$m_{kk}^{(i)} = 1 + \Delta t \sum_{j=1}^{i-1} a_{ij} \left(r_k^d(\mathbf{y}^{(j)}, t_n + c_j\Delta t) + \sum_{\nu=1}^{N} d_{k\nu}(\mathbf{y}^{(j)}, t_n + c_j\Delta t) \right) \frac{1}{\pi_\nu^{(i)}},$$

$$m_{k\nu}^{(i)} = -\Delta t \sum_{j=1}^{i-1} a_{ij} p_{k\nu}(\mathbf{y}^{(j)}, t_n + c_j\Delta t) \frac{1}{\pi_\nu^{(i)}}, \quad k \neq \nu$$

as well as, using $\mathbf{M} = (m_{k\nu})_{1\le k,\nu\le N}$,

$$m_{kk} = 1 + \Delta t \sum_{j=1}^{s} b_j \left(r_k^d(\mathbf{y}^{(j)}, t_n + c_j\Delta t) + \sum_{\nu=1}^{N} d_{k\nu}(\mathbf{y}^{(j)}, t_n + c_j\Delta t) \right) \frac{1}{\sigma_\nu},$$

$$m_{k\nu} = -\Delta t \sum_{j=1}^{s} b_j p_{k\nu}(\mathbf{y}^{(j)}, t_n + c_j\Delta t) \frac{1}{\sigma_\nu}, \quad k \neq \nu.$$

Remark 3.5. We require σ_ν to be independent of y_ν^{n+1} to ensure that the scheme is positive and linear implicit. To see this, recall that the choice $\sigma_\nu = y_\nu^{n+1}$ and $\pi_\nu^{(i)} = y_\nu^{(i)}$ would lead to the original Runge–Kutta scheme, which is not unconditionally positive. Moreover, if σ_ν would allowed to be a nonlinear function of y_ν^{n+1} we would have to solve a nonlinear system instead of a linear one to compute y_ν^{n+1}. For the same reason we require $\pi_\nu^{(i)}$ to be independent of $y_\nu^{(i)}$.

The following two lemmas state that MPRK schemes as defined in Definition 3.3 are indeed unconditionally positive and conservative. Both lemmas are slight generalizations of lemmas from [BDM03, KM18a].

Lemma 3.6. An MPRK scheme (MPRK) applied to a conservative PDS, i. e. $\mathbf{r} = \mathbf{0}$, is unconditionally conservative. The same holds for all stage values, that is $\sum_{k=1}^N y_k^{(i)} = \sum_{k=1}^N y_k^n$ for $i = 1, \ldots, s$.

Lemma 3.7. An MPRK scheme (MPRK) is unconditionally positive. The same holds for all the stages of the scheme, this is for all $\Delta t > 0$ and $\mathbf{y}^n > 0$ we have $\mathbf{y}^{(i)} > 0$ for $i = 1, \ldots, s$. In particular, the inverses $(\mathbf{M}^{(i)})^{-1}, (\mathbf{M})^{-1}$ exist and their entries lie in the interval $[0, 1]$

We also note that this scheme always produces positive approximations, if $\mathbf{y}^0 > \mathbf{0}$. However, if it is known that the analytic solution is not positive due to the existence of the rest term $\mathbf{r}$, then one may consider choosing $\mathbf{r}^d = \mathbf{0}$ and $\mathbf{r}^p = \mathbf{r}$ in the MPRK scheme (MPRK). This essentially means that we drop the non-negativity constrain on $\mathbf{r}^p$, so that the right-hand sides in (3.4) are allowed to be negative, and thus, the stage vectors and iterates of the MPRK scheme are not forced to stay positive anymore.

Remark 3.8. Definition 3.3 is formulated for non-negative Runge–Kutta parameters. But MPRK schemes with negative Runge–Kutta parameters can be devised as well. In this case, the weighting of the production and destruction terms which get multiplied by the negative weight must be interchanged. To be precise, the index ν of the PWDs $\pi_\nu^{(i)}$ and σ_ν in the formula (MPRK) is replaced by the value of the *index function*

$$\gamma(\nu, k, x) = \begin{cases} \nu, & x \geq 0 \\ k, & x < 0 \end{cases} \tag{3.5}$$

at $x = a_{ij}$ and $x = b_j$, respectively. Similarly, the index k is replaced by $\gamma(k, \nu, a_{ij})$ and $\gamma(k, \nu, b_j)$ for $\pi_k^{(i)}$ and σ_k, respectively.

This procedure will ensure the unconditional positivity of the scheme, but one may argue that this has an impact on the necessary requirements to obtain a certain order of accuracy. Fortunately, we will discover that this is not the case in Chapter 4. To avoid multiple case distinctions we demand for positive Runge–Kutta parameters in the remainder of this thesis.

Next, we want to explain in what sense the given definition of MPRK schemes generalizes the existing ones from [KM18a, TÖR22]. First, MPRK schemes can be understood as NSARK methods using the splitting of the right-hand side mentioned in Remark 2.25. Substituting this into (MPRK) and setting $t_j = t_n + c_j\Delta t$, we see

$$\begin{aligned}
y_k^{(i)} &= y_k^n + \Delta t \sum_{j=1}^{i-1} a_{ij} \left(\sum_{\substack{\nu=1\\ \nu\neq k}}^{N} f_k^{[\nu]}(\mathbf{y}^{(j)}, t_j) \frac{y_\nu^{(i)}}{\pi_\nu^{(i)}} + f_k^{[k]}(\mathbf{y}^{(j)}, t_j) \frac{y_k^{(i)}}{\pi_k^{(i)}} + f_k^{[N+1]}(\mathbf{y}^{(j)}, t_j) \right) \\
&= y_k^n + \Delta t \sum_{j=1}^{i-1} a_{ij} \left(\sum_{\nu=1}^{N} \frac{y_\nu^{(i)}}{\pi_\nu^{(i)}} f_k^{[\nu]}(\mathbf{y}^{(j)}) + f_k^{[N+1]}(\mathbf{y}^{(j)}, t_j) \right) \\
&= y_k^n + \Delta t \sum_{j=1}^{i-1} \sum_{\nu=1}^{N+1} a_{ij}^{[\nu]}(\mathbf{y}^n, \Delta t) f_k^{[\nu]}(\mathbf{y}^{(j)}), \\
y_k^{n+1} &= y_k^n + \Delta t \sum_{j=1}^{s} b_j \left(\sum_{\substack{\nu=1\\ \nu\neq k}}^{N} f_k^{[\nu]}(\mathbf{y}^{(j)}) \frac{y_\nu^{n+1}}{\sigma_\nu} + f_k^{[k]}(\mathbf{y}^{(j)}) \frac{y_k^{n+1}}{\sigma_k} + f_k^{[N+1]}(\mathbf{y}^{(j)}, t_j) \right) \\
&= y_k^n + \Delta t \sum_{j=1}^{s} \sum_{\nu=1}^{N+1} b_j^{[\nu]}(\mathbf{y}^n, \Delta t) f_k^{[\nu]}(\mathbf{y}^{(j)}),
\end{aligned}$$

where the solution-dependent coefficients are given by

$$a_{ij}^{[\nu]}(\mathbf{y}^n, \Delta t) = \begin{cases} a_{ij} \frac{y_\nu^{(i)}}{\pi_\nu^{(i)}}, & \nu \leq N, \\ a_{ij}, & \nu = N+1 \end{cases} \quad \text{and} \quad b_j^{[\nu]}(\mathbf{y}^n, \Delta t) = \begin{cases} b_j \frac{y_\nu^{n+1}}{\sigma_\nu}, & \nu \leq N, \\ b_j, & \nu = N+1. \end{cases} \tag{3.6}$$

This means that the NS weights are

$$\gamma_\nu^{[i]} = \begin{cases} \frac{y_\nu^{(i)}}{\pi_\nu^{(i)}}, & \nu \leq N, \\ 1, & \nu = N+1 \end{cases} \quad \text{and} \quad \delta_\nu = \begin{cases} \frac{y_\nu^{n+1}}{\sigma_\nu}, & \nu \leq N, \\ 1, & \nu = N+1, \end{cases} \tag{3.7}$$

see (3.1).

Remark 3.9. In view of the index function (3.5), the NS weights for MPRK schemes based on RK methods with negative entries in the Butcher tableau not only depend on the step size, solution, and splitting of the right-hand side but also vary with its components. Hence, our formulation (NSARK) actually does not capture this case as we used vector notation. However, for the sake of simplicity and the reading flow, we rather discuss this special case in the particular sections than complicating the notation at this point.

If in the context of an MPRK method, constant addends in the right-hand side splitting are treated as rest terms, then $\mathbf{F}^{[N+1]}$ in Proposition 3.2 will be integrated explicitly, which means that the condition $\gamma_{N+1}^{[i]}, \delta_{N+1}^{[j]} = 1$ is satisfied as this term is not multiplied with a PWD. Hence, with this convention it suffices to study autonomous problems for deriving order conditions. Moreover we are also in the position to apply Theorem 2.3, if the production, destruction and rest terms as well as the PWDs are in $\mathcal{C}^1$ because of the following. The linear systems always possess a unique solution and the implicit function theorem tells us that the resulting incremental map is in $\mathcal{C}^1$, and hence, locally Lipschitz with respect to its second argument. We will later see that the PWDs fulfill these requirements for the particular MPRK schemes.

Hereafter, we present schemes for the conservative and autonomous PDS (2.21). The formulation for general PDRS is straightforward. In particular, (MPRK) reduces in this case to

$$y_k^{(i)} = y_k^n + \Delta t \sum_{j=1}^{i-1} a_{ij} \sum_{\nu=1}^{N} \left(p_{k\nu}(\mathbf{y}^{(j)}) \frac{y_\nu^{(i)}}{\pi_\nu^{(i)}} - d_{k\nu}(\mathbf{y}^{(j)}) \frac{y_k^{(i)}}{\pi_k^{(i)}} \right), \quad i = 1, \ldots, s, \tag{3.8a}$$

$$y_k^{n+1} = y_k^n + \Delta t \sum_{j=1}^{s} b_j \sum_{\nu=1}^{N} \left(p_{k\nu}(\mathbf{y}^{(j)}) \frac{y_\nu^{n+1}}{\sigma_\nu} - d_{k\nu}(\mathbf{y}^{(j)}) \frac{y_k^{n+1}}{\sigma_k} \right), \quad k = 1, \ldots, N. \tag{3.8b}$$

First Order MPRK Scheme

Based on the explicit Euler method, the first MPRK method, the so-called modified Patankar Euler (MPE) scheme was developed in [BDM03]. It is proven to be first order accurate when applied to (2.21) within the same work and reads

$$\begin{aligned} y_k^{(1)} &= y_k^n, \\ y_k^{n+1} &= y_k^n + \Delta t \sum_{\nu=1}^{N} \left(p_{k\nu}(\mathbf{y}^{(1)}) \frac{y_\nu^{n+1}}{y_\nu^n} - d_{k\nu}(\mathbf{y}^{(1)}) \frac{y_k^{n+1}}{y_k^n} \right), \end{aligned} \tag{MPE}$$

for $k = 1, \ldots, N$, that is $\sigma_\nu = y_\nu^n$. Here, σ_ν is obviously a $\mathcal{C}^1$-mapping.

Second Order MPRK Schemes

The explicit 2-stage RK method based on the Butcher array

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

is second order accurate. Moreover, the entries of the array are non-negative for $\alpha \geq \frac{1}{2}$. With that as a starting point, the authors from [KM18a] derived a 1-parameter family of second order accurate MPRK schemes using $\pi_\nu^{(2)} = y_\nu^n$ and $\sigma_\nu = (y_\nu^{(2)})^{\frac{1}{\alpha}} (y_\nu^n)^{1-\frac{1}{\alpha}}$ for $i = 1, \ldots, N$. For simplicity, we again present the resulting MPRK22(α) scheme for solving (2.21), i. e.

$$\begin{aligned}
y_k^{(1)} &= y_k^n,\\
y_k^{(2)} &= y_k^n + \alpha\Delta t\sum_{\nu=1}^N\left(p_{k\nu}(\mathbf{y}^{(1)})\frac{y_\nu^{(2)}}{y_\nu^n} - d_{k\nu}(\mathbf{y}^{(1)})\frac{y_k^{(2)}}{y_k^n}\right),\\
y_k^{n+1} &= y_k^n + \Delta t\sum_{\nu=1}^N\Bigg(\left(\left(1-\frac{1}{2\alpha}\right)p_{k\nu}(\mathbf{y}^{(1)}) + \frac{1}{2\alpha}p_{k\nu}(\mathbf{y}^{(2)})\right)\frac{y_\nu^{n+1}}{(y_\nu^{(2)})^{\frac{1}{\alpha}}(y_\nu^n)^{1-\frac{1}{\alpha}}}\\
&\qquad -\left(\left(1-\frac{1}{2\alpha}\right)d_{k\nu}(\mathbf{y}^{(1)}) + \frac{1}{2\alpha}d_{k\nu}(\mathbf{y}^{(2)})\right)\frac{y_k^{n+1}}{(y_k^{(2)})^{\frac{1}{\alpha}}(y_k^n)^{1-\frac{1}{\alpha}}}\Bigg) \qquad \text{(MPRK22)}
\end{aligned}$$

for $k=1,\dots,N$ with $\alpha\geq\frac{1}{2}$. Since $y_\nu^{(2)}=y_\nu^{(2)}(\mathbf{y}^n)$ is in $\mathcal{C}^1$ due to the implicit function theorem, the same holds for the PWDs.

Third Order MPRK Schemes

Assuming a non-negative Butcher tableau from an explicit 3-stage RK method, third order MPRK schemes have been constructed in [KM18b] for solving (2.21) using the denominator weights

$$\begin{aligned}
\pi_\nu^{(2)} &= y_\nu^n,\\
\pi_\nu^{(3)} &= (y_\nu^{(2)})^{\frac{1}{p}}(y_\nu^n)^{1-\frac{1}{p}}, \quad p = 3a_{21}(a_{31}+a_{32})b_3,\\
\sigma_k &= y_k^n + \Delta t\sum_{\nu=1}^N\Bigg(\left(\beta_1 p_{k\nu}(\mathbf{y}^n) + \beta_2 p_{k\nu}(\mathbf{y}^{(2)})\right)\frac{\sigma_\nu}{(y_\nu^{(2)})^{\frac{1}{a_{21}}}(y_\nu^n)^{1-\frac{1}{a_{21}}}}\\
&\qquad - \left(\beta_1 d_{k\nu}(\mathbf{y}^n) + \beta_2 d_{k\nu}(\mathbf{y}^{(2)})\right)\frac{\sigma_k}{(y_k^{(2)})^{\frac{1}{a_{21}}}(y_k^n)^{1-\frac{1}{a_{21}}}}\Bigg)
\end{aligned} \tag{3.9}$$

for $\nu,k=1,\dots,N$, $\beta_1=1-\beta_2$ and $\beta_2=\frac{1}{2a_{21}}$. Note, that solving another system of linear equations is necessary to calculate $\boldsymbol{\sigma}=(\sigma_1,\dots,\sigma_N)$. Hence, the resulting MPRK scheme may be based on 3-stage RK methods but can be viewed as 4-stage schemes, where we note that $\boldsymbol{\sigma}$ can be computed simultaneously with $\mathbf{y}^{(3)}$. We also point out that there are no additional right-hand side evaluations required for computing $\boldsymbol{\sigma}$. The final scheme for conservative and autonomous PDS takes the form

$$\begin{aligned}
y_k^{(1)} &= y_k^n,\\
y_k^{(2)} &= y_k^n + a_{21}\Delta t\sum_{\nu=1}^N\left(p_{k\nu}(\mathbf{y}^n)\frac{y_\nu^{(2)}}{y_\nu^n} - d_{k\nu}(\mathbf{y}^n)\frac{y_k^{(2)}}{y_k^n}\right),\\
y_k^{(3)} &= y_k^n + \Delta t\sum_{\nu=1}^N\Bigg(\left(a_{31}p_{k\nu}(\mathbf{y}^n) + a_{32}p_{k\nu}(\mathbf{y}^{(2)})\right)\frac{y_\nu^{(3)}}{(y_\nu^{(2)})^{\frac{1}{p}}(y_\nu^n)^{1-\frac{1}{p}}}\\
&\qquad - \left(a_{31}d_{k\nu}(\mathbf{y}^n) + a_{32}d_{k\nu}(\mathbf{y}^{(2)})\right)\frac{y_k^{(3)}}{(y_k^{(2)})^{\frac{1}{p}}(y_k^n)^{1-\frac{1}{p}}}\Bigg),
\end{aligned}$$

$$
\begin{aligned}
\sigma_k = y_k^n + \Delta t \sum_{\nu=1}^{N} \Bigg(&\Big(\beta_1 p_{k\nu}\big(\mathbf{y}^n\big) + \beta_2 p_{k\nu}\big(\mathbf{y}^{(2)}\big)\Big) \frac{\sigma_\nu}{\big(y_\nu^{(2)}\big)^{\frac{1}{q}} \big(y_\nu^n\big)^{1-\frac{1}{q}}} \\
&- \Big(\beta_1 d_{k\nu}\big(\mathbf{y}^n\big) + \beta_2 d_{k\nu}\big(\mathbf{y}^{(2)}\big)\Big) \frac{\sigma_k}{\big(y_k^{(2)}\big)^{\frac{1}{q}} \big(y_k^n\big)^{1-\frac{1}{q}}} \Bigg), \\
y_k^{n+1} = y_k^n + \Delta t \sum_{\nu=1}^{N} \Bigg(&\Big(b_1 p_{k\nu}\big(\mathbf{y}^n\big) + b_2 p_{k\nu}\big(\mathbf{y}^{(2)}\big) + b_3 p_{k\nu}\big(\mathbf{y}^{(3)}\big)\Big) \frac{y_\nu^{n+1}}{\sigma_\nu} \\
&- \Big(b_1 d_{k\nu}\big(\mathbf{y}^n\big) + b_2 d_{k\nu}\big(\mathbf{y}^{(2)}\big) + b_3 d_{k\nu}\big(\mathbf{y}^{(3)}\big)\Big) \frac{y_k^{n+1}}{\sigma_k} \Bigg),
\end{aligned}
\tag{MPRK43}
$$

where $p = 3a_{21}\,(a_{31}+a_{32})\,b_3,\ q = a_{21},\ \beta_2 = \frac{1}{2a_{21}}$ and $\beta_1 = 1-\beta_2$. As before, the PWDs are in $\mathcal{C}^1$, if the production and destruction terms are.

MPRK43(α,β)

All entries of the Butcher array

$$
\begin{array}{c|ccc}
0 & & & \\
\alpha & \alpha & & \\
\beta & \frac{3\alpha\beta(1-\alpha)-\beta^2}{\alpha(2-3\alpha)} & \frac{\beta(\beta-\alpha)}{\alpha(2-3\alpha)} & \\
\hline
 & 1+\frac{2-3(\alpha+\beta)}{6\alpha\beta} & \frac{3\beta-2}{6\alpha(\beta-\alpha)} & \frac{2-3\alpha}{6\beta(\beta-\alpha)}
\end{array}
\tag{3.10}
$$

with

$$
\begin{cases}
2/3 \le \beta \le 3\alpha(1-\alpha) \\
3\alpha(1-\alpha) \le \beta \le 2/3 \\
\frac{3\alpha-2}{6\alpha-3} \le \beta \le 2/3
\end{cases}
\quad \text{for} \quad
\begin{cases}
1/3 \le \alpha < \frac{2}{3}, \\
2/3 < \alpha < \alpha_0, \\
\alpha > \alpha_0,
\end{cases}
\tag{3.11}
$$

and $\alpha_0 \approx 0.89255$ are non-negative [KM18b, Lemma 6], see Figure 3.1 for an illustration of the feasible domain.

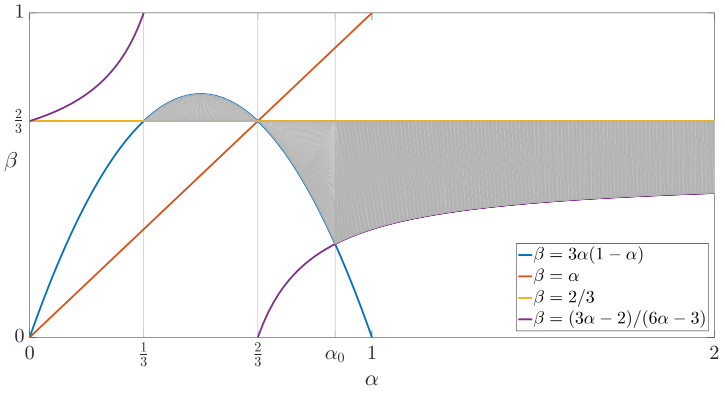


Figure 3.1: The gray area represents all (α,β) pairs which fulfill the conditions (3.11), i. e. for which the Butcher tableau (3.10) is non-negative [KM18b].

The resulting MPRK43(α,β) method is determined by (MPRK43) using (3.10) and

$$\begin{aligned} p &= 3a_{21}(a_{31}+a_{32})b_3 = \alpha\frac{2-3\alpha}{2(\beta-\alpha)}, & q &= a_{21} = \alpha, \\ \beta_2 &= \frac{1}{2a_{21}} = \frac{1}{2\alpha}, & \beta_1 &= 1-\beta_2 = 1-\frac{1}{2\alpha}. \end{aligned} \tag{3.12}$$

MPRK43(γ)

It was also proven in [KM18b, Lemma 6] that all entries of the tableau

$$\begin{array}{c|ccc} 0 & & & \\ \frac{2}{3} & \frac{2}{3} & & \\ \frac{2}{3} & \frac{2}{3}-\frac{1}{4\gamma} & \frac{1}{4\gamma} & \\ \hline & \frac{1}{4} & \frac{3}{4}-\gamma & \gamma \end{array} \tag{3.13}$$

are non-negative for $\frac{3}{8} \leq \gamma \leq \frac{3}{4}$. The corresponding MPRK scheme is denoted by MPRK43(γ) and can be obtained from (MPRK43) by substituting (3.13) and

$$\begin{aligned} p &= 3a_{21}(a_{31}+a_{32})b_3 = \frac{4}{3}\gamma, & q &= a_{21} = \frac{2}{3}, \\ \beta_2 &= \frac{1}{2a_{21}} = \frac{3}{4}, & \beta_1 &= 1-\beta_2 = \frac{1}{4}. \end{aligned} \tag{3.14}$$

3.3 Strong-Stability Preserving MPRK

Strong-stability preserving Runge–Kutta (SSPRK) methods were introduced in [SO88] and developed for the time integration of the semi-discretization of hyperbolic conservation laws. The main idea was to rewrite an explicit RK method into *Shu–Osher form*. With that, the authors in [SO88] present higher order methods that preserve any convex functional bound such as positivity or total variation diminishing (TVD) property whenever the forward Euler method possesses the respective property. To obtain unconditional positivity, strong-stability preserving modified Patankar–Runge–Kutta (SSPMPRK) methods were constructed in [HS19] and proven to be of second order. Later, also third order methods were constructed [HZS19]. Moreover, the schemes are also conservative and there exist analogues of Lemma 3.6 and Lemma 3.7 for these methods.

In order to adapt SSPMPRK methods into our framework of NSARK schemes, we would have to introduce the ARK methods in Shu–Osher form and then consider solution-dependent coefficients. This together with the corresponding generalization of the results from [HS19, HZS19, HIK$^+$23] along the theory developed in [IKM23b] is object to future work. We also want to note here that in [HS19, HZS19], the SSPMPRK methods were also used as time integrators in the context of reactive Euler equations.

Second Order SSPMPRK Schemes The second order SSPMPRK scheme for solving (2.21), introduced in [HS19], is given by

$$
\begin{aligned}
y_i^{(1)} &= y_i^n + \beta\Delta t \left(\sum_{j=1}^N p_{ij}(\mathbf{y}^n) \frac{y_j^{(1)}}{y_j^n} - \sum_{j=1}^N d_{ij}(\mathbf{y}^n) \frac{y_i^{(1)}}{y_i^n} \right), \\
y_i^{n+1} &= (1-\alpha)y_i^n + \alpha y_i^{(1)} + \Delta t \Bigg(\sum_{j=1}^N \Big(\beta_{20} p_{ij}(\mathbf{y}^n) + \beta_{21} p_{ij}(\mathbf{y}^{(1)}) \Big) \frac{y_j^{n+1}}{(y_j^n)^{1-s}(y_j^{(1)})^s} \\
&\qquad - \sum_{j=1}^N \Big(\beta_{20} d_{ij}(\mathbf{y}^n) + \beta_{21} d_{ij}(\mathbf{y}^{(1)}) \Big) \frac{y_i^{n+1}}{(y_i^n)^{1-s}(y_i^{(1)})^s} \Bigg),
\end{aligned}
\tag{SSPMPRK2}
$$

where $\beta_{20} = 1 - \frac{1}{2\beta} - \alpha\beta$, $\beta_{21} = \frac{1}{2\beta}$ and $s = \frac{1-\alpha\beta+\alpha\beta^2}{\beta(1-\alpha\beta)}$. Thereby, the free parameters α and β are subject to

$$
0 \leq \alpha \leq 1, \quad \beta > 0, \quad \alpha\beta + \frac{1}{2\beta} \leq 1. \tag{3.15}
$$

We refer to the above scheme as SSPMPRK2(α, β).

Third Order SSPMPRK Schemes The third order method applied to (2.21) can be written as

$$
\begin{aligned}
y_i^{(1)} &= \alpha_{10} y_i^n + \beta_{10}\Delta t \left(\sum_{j=1}^N p_{ij}(\mathbf{y}^n) \frac{y_j^{(1)}}{y_j^n} - \sum_{j=1}^N d_{ij}(\mathbf{y}^n) \frac{y_i^{(1)}}{y_i^n} \right), \\
\rho_i &= n_1 y_1^{(1)} + n_2 y_i^n \left(\frac{y_i^{(1)}}{y_i^n} \right)^2, \\
y_i^{(2)} &= \alpha_{20} y_i^n + \alpha_{21} y_i^{(1)} + \Delta t \Bigg(\sum_{j=1}^N \Big(\beta_{20} p_{ij}(\mathbf{y}^n) + \beta_{21} p_{ij}(\mathbf{y}^{(1)}) \Big) \frac{y_j^{(2)}}{\rho_j} \\
&\qquad - \sum_{j=1}^N \Big(\beta_{20} d_{ij}(\mathbf{y}^n) + \beta_{21} d_{ij}(\mathbf{y}^{(1)}) \Big) \frac{y_i^{(2)}}{\rho_i} \Bigg), \\
\gamma_i &= \eta_1 y_i^n + \eta_2 y_i^{(1)} + \Delta t \Bigg(\sum_{j=1}^N \Big(\eta_3 p_{ij}(\mathbf{y}^n) + \eta_4 p_{ij}(\mathbf{y}^{(1)}) \Big) \frac{\gamma_j}{(y_j^n)^{1-s}(y_j^{(1)})^s} \\
&\qquad - \sum_{j=1}^N \Big(\eta_3 d_{ij}(\mathbf{y}^n) + \eta_4 d_{ij}(\mathbf{y}^{(1)}) \Big) \frac{\gamma_i}{(y_i^n)^{1-s}(y_i^{(1)})^s} \Bigg),
\end{aligned}
\tag{3.16}
$$

$$
\begin{aligned}
\sigma_i &= \gamma_i + \zeta y_i^n \frac{y_i^{(2)}}{\rho_i}, \\
y_i^{n+1} &= \alpha_{30} y_i^n + \alpha_{31} y_i^{(1)} + \alpha_{32} y_i^{(2)} \\
&\qquad + \Delta t \Bigg(\sum_{j=1}^N \Big(\beta_{30} p_{ij}(\mathbf{y}^n) + \beta_{31} p_{ij}(\mathbf{y}^{(1)}) + \beta_{31} p_{ij}(\mathbf{y}^{(2)}) \Big) \frac{y_j^{n+1}}{\sigma_j} \qquad (3.17) \\
&\qquad - \sum_{j=1}^N \Big(\beta_{30} d_{ij}(\mathbf{y}^n) + \beta_{31} d_{ij}(\mathbf{y}^{(1)}) + \beta_{32} d_{ij}(\mathbf{y}^{(2)}) \Big) \frac{y_i^{n+1}}{\sigma_i} \Bigg),
\end{aligned}
\tag{SSPMPRK3}
$$

where we use the parameters

$$\begin{aligned}
\alpha_{10} &= 1, & \alpha_{20} &= 9.2600312554031827 \cdot 10^{-1},\\
\alpha_{21} &= 7.3996874459681783 \cdot 10^{-2}, & \alpha_{30} &= 7.0439040373427619 \cdot 10^{-1},\\
\alpha_{31} &= 2.0662904223744017 \cdot 10^{-10}, & \alpha_{32} &= 2.9560959605909481 \cdot 10^{-1},\\
\beta_{10} &= 4.7620819268131703 \cdot 10^{-1}, & \beta_{20} &= 7.7545442722396801 \cdot 10^{-2},\\
\beta_{21} &= 5.9197500149679749 \cdot 10^{-1}, & \beta_{30} &= 2.0044747790361456 \cdot 10^{-1},\\
\beta_{31} &= 6.8214380786704851 \cdot 10^{-10}, & \beta_{32} &= 5.9121918658514827 \cdot 10^{-1},\\
\zeta &= 0.62889380778287493358, & \eta_1 &= 0.37110619221712506642 - \eta_2,\\
\eta_3 &= -1.2832127371313151768\eta_2 & \eta_4 &= 2.2248760403511226405,\\
&\quad + 0.6146025595987523739, & &\\
n_1 &= 0.25690460257320105191, & n_2 &= 1 - n_1
\end{aligned} \tag{3.18}$$

in accordance with [HZS19]. Here, η_2 is a free parameter satisfying $\eta_2 \in [0, r_1]$ with $r_1 = 0.37110619221712506642$, so that we refer to this scheme as SSPMPRK3(η_2). For more details on the parameters we refer to the Maple code in the reproducibility repository [HIK$^+$22].

3.4 Modified Patankar Deferred Correction

Arbitrarily high-order conservative and positive modified Patankar Deferred Correction schemes (MPDeC) were introduced in [ÖT20] which are based on the Deferred Correction (DeC) approach developed in [DGR00]. To that end, a time step $[t_n, t_{n+1}]$ is transformed to $[0, 1]$ and then divided into M subintervals determined by $0 = t_{n,0} < \cdots < t_{n,M} = 1$. The idea of the scheme is to mimic the Picard iterations on a discrete level as follows. At each subtime step $t_{n,m}$ an approximation y^m is calculated. An iterative procedure of K correction steps improves the approximation by one order of accuracy at each iteration. The modified Patankar-trick is introduced inside the basic scheme to guarantee positivity and conservation of the intermediate approximations.

The MPDeC correction steps can be rewritten for $k = 1, \dots, K$, $m = 1, \dots, M$ and $i = 1, \dots, N$ as

$$y_i^{m,(k)} = y_i^0 + \sum_{r=0}^{M} \theta_r^m \Delta t \sum_{j=1}^{N} \left(p_{ij}(y^{r,(k-1)}) \frac{y^{m,(k)}_{\gamma(j,i,\theta_r^m)}}{y^{m,(k-1)}_{\gamma(j,i,\theta_r^m)}} - d_{ij}(y^{r,(k-1)}) \frac{y^{m,(k)}_{\gamma(i,j,\theta_r^m)}}{y^{m,(k-1)}_{\gamma(i,j,\theta_r^m)}} \right), \tag{MPDeC}$$

where $\theta_r^m = \int_0^{t_{n,m}} \varphi_r(t) \mathrm{d}t$ are the *correction weights*, and

$$\gamma(j, i, \theta_r^m) = \begin{cases} j, & \theta_r^m \geq 0 \\ i, & \theta_r^m < 0 \end{cases}$$

is the index function (3.5). Here, φ_r is the rth Lagrangian polynomial defined by the subtime nodes $\{t_{n,m}\}_{m=0}^M$. As a result of $\theta_r^0 = 0$, the initial states $y_i^{0,(k)} = y_i^n$ are identical for any correction k. The new numerical solution at time $t_n + \Delta t$ is $\mathbf{y}^{n+1} = \mathbf{y}^{M,(K)}$.

Remark 3.10. Formally, MPDeC methods can be interpreted as RK schemes by viewing the correction steps as additional stages. Consequently, MPDeC methods can be written as NSARK schemes. However, similarly to the case discussed in Remark 3.9, the NS weights of MPDeC depend on the components of the vector $\mathbf{y}^{m,(k)}$ whenever the correction weights are negative, which is already the case for $K > 2$. Nevertheless, since the weights are similar to those of MPRK methods we can conclude that the order of accuracy of MPDeC methods is also maintained for non-autonomous PDS and that Theorem 2.3 can be applied.

The order of accuracy of the MPDeC scheme is the minimum between K and the accuracy of the quadrature formula given by the M subtime steps. In view of the existing literature, we will focus on equispaced (EQ) and Gauss–Lobatto (GL) subtime steps [ÖT20]. To obtain order p, a number of $K = p$ iterations is required, while we need $M = \max\{p-1, 1\}$ EQ subtime steps or $M = \lceil \frac{p}{2} \rceil$ GL subtime steps. To indicate the quadrature formula we introduce the notation MPDeCGL(p) and MPDeCEQ(p) for MPDeC methods of order p using GL or EQ subtime steps, respectively.

Obviously (MPDeC) is due to this iterative process more complicated than the previous schemes, especially since the index function changes productive and destructive part inside the underlying PDS. However, these methods are arbitrary high order, unconditionally positive and conservative. Additionally, they have been applied successfully in the context of the shallow water equations guaranteeing a positive water height [CMÖT22].

3.5 Geometric Conservative

A class of numerical methods that preserve all linear invariants and still guarantee positivity is given by GeCo schemes introduced in [MCD20]. These methods fall in the class of non-standard integrators [Mic21] as they result as non-standard versions of explicit first and second order Runge–Kutta schemes, where the advancement in time is modulated by a nonlinear functional dependency on the temporal step size and on the approximation itself. The step size modification thereby guarantees the numerical solution to be unconditionally positive while keeping the accuracy of the underlying method. GeCo schemes are applied to general biochemical systems [FS11b, BBKS07]

$$\mathbf{y}' = \mathbf{f}(\mathbf{y}, t) = \mathbf{S}\mathbf{r}(\mathbf{y}, t), \qquad \mathbf{y}(0) = \mathbf{y}^0, \tag{3.19}$$

where $\mathbf{S} \in \mathbb{R}^{N\times M}$ is the *stoichiometric* matrix with entries s_{ij} for $i = 1, \dots, N$ and $j = 1, \dots, M$, and $\mathbf{r}(\mathbf{y}) = (r_1(\mathbf{y}), \dots, r_M(\mathbf{y}))^T$ is the vector of the *reaction* functions. The following assumptions, stated in [FS11b], assure the well-posedness of the system (3.19) and the positivity of the solutions.

a) For $j = 1, \dots, M$ we have $r_j \in \mathcal{C}^0\left(\mathbb{R}^N_{\geq 0}, \mathbb{R}_{\geq 0}\right)$ and $r_j(\cdot, t)$ is locally Lipschitz in $\mathbb{R}^N$, uniformly in t.

b) There holds $\mathbf{r}(\mathbf{y}, t) > \mathbf{0}$ if $\mathbf{y} > \mathbf{0}$, and $\mathbf{r}(\mathbf{y}, t) = \mathbf{0}$ if $\mathbf{y} = \mathbf{0}$ for all $t > 0$.

c) If $s_{ij} < 0$, there exists a $q_j \in \mathcal{C}^0\left(\mathbb{R}^N_{\geq 0}, \mathbb{R}_{\geq 0}\right)$ such that $r_j(\mathbf{y}, t) = q_j(\mathbf{y}, t) y_i$.

The GeCo methods are of the form

$$\begin{aligned}\mathbf{y}^{(i)} &= \mathbf{y}^n + \phi_i(\mathbf{y}^n, t_n, \Delta t)\Delta t \sum_{j=1}^{i-1} a_{ij}\mathbf{f}(\mathbf{y}^{(j)}), \quad i = 1, \ldots, s, \\ \mathbf{y}^{n+1} &= \mathbf{y}^n + \phi_{n+1}(\mathbf{y}^n, t_n, \Delta t)\Delta t \sum_{j=1}^{s} b_j\mathbf{f}(\mathbf{y}^{(j)}),\end{aligned} \tag{GeCo}$$

see [MCD20], where we point out that our formulation includes non-autonomous biochemical problems. Note that ϕ here corresponds to the function Φ of [MCD20] divided by Δt, and that the value of ϕ_1 has no effect since $a_{1j} = 0$. The idea is to choose the functions ϕ_i and ϕ_{n+1} in a way that guarantees the positivity of the stages and the updated solution. At the same time, these functions must be chosen in a way that does not compromise the order of accuracy. Up to now, only conditions for first and second order GeCo schemes are available.

To interpret (GeCo) as a non-standard RK (NSRK) method, we absorb the factors ϕ_i, ϕ_{n+1} into the RK coefficients, which we can write formally in the notation of Section 3.1 via the coefficients:

$$a^{[1]}_{ij}(\mathbf{y}^n, t_n, \Delta t) = a_{ij}\phi_i(\mathbf{y}^n, t_n, \Delta t), \quad b^{[1]}_j(\mathbf{y}^n, t_n, \Delta t) = b_j\phi_{n+1}(\mathbf{y}^n, t_n, \Delta t) \tag{3.20}$$

for $i, j = 1, \ldots, s$. This means that the NS weights are $\gamma^{[1]}_i = \phi_i$ and $\delta_i = \phi_{n+1}$.

First Order GeCo Scheme For the construction of the NS weights of GeCo methods, the vector field $\mathbf{f}(\mathbf{y}, t) = \mathbf{S}\mathbf{r}(\mathbf{y}, t)$ is split into production and destruction parts as

$$\mathbf{f}(\mathbf{y}, t) = \mathbf{f}^{[P]}(\mathbf{y}, t) - \mathbf{f}^{[D]}(\mathbf{y}, t), \quad \mathbf{f}^{[P]}(\mathbf{y}, t) = \mathbf{S}^{+}\mathbf{r}(\mathbf{y}, t), \quad \mathbf{f}^{[D]}(\mathbf{y}, t) = \mathbf{S}^{-}\mathbf{r}(\mathbf{y}, t) \tag{3.21}$$

with $\mathbf{S}^+, \mathbf{S}^- \geq \mathbf{0}$. The first order GeCo scheme (GeCo1) applied to a general biochemical system (3.19), (3.21) is defined as

$$\mathbf{y}^{n+1} = \mathbf{y}^n + \Delta t \varphi\left(\Delta t \sum_{i=1}^{N} \frac{f^{[D]}_i(\mathbf{y}^n, t_n)}{y^n_i}\right)\mathbf{f}(\mathbf{y}^n, t_n), \tag{GeCo1}$$

where the function $\varphi \in \mathcal{C}^2$ is defined as

$$\varphi(x) = \begin{cases} \dfrac{1 - e^{-x}}{x}, & x > 0, \\ 1, & x = 0. \end{cases} \tag{3.22}$$

In the notation of a general GeCo method, we have

$$\phi_{n+1}(\mathbf{y}^n, t_n, \Delta t) = \varphi\left(\Delta t \sum_{i=1}^{N} \frac{f^{[D]}_i(\mathbf{y}^n, t_n)}{y^n_i}\right).$$

Remark 3.11. Even though (GeCo1) can be interpreted as an NSARK method with φ being the NS weight, the scheme is not an additive method since the whole right-hand side $\mathbf{f}$ is weighted by the same factor. Hence, we are not in the position to apply Proposition 3.2 directly. However, considering the autonomous problem with $\mathbf{F}(\mathbf{Y}) = (\mathbf{f}(\mathbf{Y}), 1)^T$ and $\mathbf{Y} = (\mathbf{y}, t)^T$, one can see from (GeCo1) that the last component of the method reads

$$t_{n+1} = t_n + \Delta t \varphi \left(\Delta t \sum_{i=1}^{N} \frac{f_i^{[D]}(\mathbf{y}^n, t_n)}{y_i^n} \right),$$

which is why it is not clear whether or not the condition for first order from [MCD20] is sufficient also for non-autonomous problems.

We also note that the NS weight is in $\mathcal{C}^1$ whenever $\mathbf{f}^{[D]} \in \mathcal{C}^1$, so that we can also apply Theorem 2.3 to prove the order of convergence.

Second Order GeCo Scheme The second order GeCo (GeCo2) scheme for a general biochemical system (3.19), (3.21) is based on Heun's methods and takes the form

$$\begin{aligned}
\mathbf{y}^{(1)} &= \mathbf{y}^n, \\
\mathbf{y}^{(2)} &= \mathbf{y}^n + \Delta t \varphi \left(\Delta t \sum_{i=1}^{N} \frac{f_i^{[D]}(\mathbf{y}^n, t_n)}{y_i^n} \right) \mathbf{f}(\mathbf{y}^n, t_n), \\
\mathbf{y}^{n+1} &= \mathbf{y}^n + \frac{\Delta t}{2} \varphi \left(\Delta t \sum_{i=1}^{N} \frac{w_i^+(\mathbf{y}^n, t_n)}{y_i^n} \right) \left(\mathbf{f}(\mathbf{y}^n, t_n) + \mathbf{f}(\mathbf{y}^{(2)}, t_n + \Delta t) \right),
\end{aligned} \tag{GeCo2}$$

where

$$w_i^+(\mathbf{y}^n, t_n) = \max(0, w_i(\mathbf{y}^n, t_n)), \quad i = 1, \ldots, N$$

with

$$\mathbf{w}(\mathbf{y}^n, t_n) = 2\varphi \left(\Delta t \sum_{j=1}^{N} \frac{f_i^{[D]}(\mathbf{y}^n, t_n)}{y_i^n} \right) \mathbf{f}(\mathbf{y}^n, t_n) - \mathbf{f}(\mathbf{y}^n, t_n) - \mathbf{f}(\mathbf{y}^{(2)}, t_n + \Delta t).$$

Since φ is in $\mathcal{C}^1$ we see that w_i^+ is the composition of locally Lipschitz continuous mappings if $\mathbf{f}^{[D]} \in \mathcal{C}^1$, and hence, itself locally Lipschitz continuous. Thus, we can apply Theorem 2.3 to deduce the order of convergence.

Moreover, as discussed in Remark 3.11 for GeCo1, the order conditions derived in [MCD20] for GeCo2 hold for autonomous problems and it is not clear if the method is of second order for non-autonomous problems. Investigating this question for GeCo methods will be part of future work.

3.6 Generalized BBKS

The generalized BBKS (gBBKS) schemes, named after the authors Bruggeman, Burchard, Kooi and Sommeijer, were developed in [BBKS07, BRBM08, AKM20] and represent a class of schemes that are unconditionally positive while preserving all linear invariants of the underlying ordinary differential equation $\mathbf{y}' = \mathbf{f}(\mathbf{y}, t)$. Similarly to GeCo methods, the idea is to weight the function $\mathbf{f} \colon \mathbb{R}^N \times \mathbb{R} \to \mathbb{R}^N$ with a positivity-preserving factor. As a result, gBBKS schemes can also be interpreted as NSRK methods with the positivity-preserving factor being the NS weight. So far, first and second order accurate methods have been constructed which we briefly review in the following.

First Order gBBKS Schemes The first order gBBKS schemes (gBBKS1) can be written as

$$\mathbf{y}^{n+1} = \mathbf{y}^n + \Delta t \mathbf{f}(\mathbf{y}^n, t_n) \left(\prod_{m \in M^n} \frac{y_m^{n+1}}{\sigma_m^n} \right)^{r^n}, \tag{gBBKS1}$$

where $r^n, \sigma_m^n > 0$ are free parameters, but need to be chosen independently of $\mathbf{y}^{n+1}$, and

$$M^n = \{m \in \{1, \dots, N\} \mid f_m(\mathbf{y}^n, t_n) < 0\}.$$

For instance, the BBKS1 scheme from [BBKS07, AKM20] is given by setting $\sigma_m^n = y_m^n$ and $r^n = 1$. As discussed for GeCo methods in Remark 3.11, it is not straightforward to see whether or not the proven first order of (gBBKS1) is maintained for non-autonomous problems.

Moreover, as the number of factors in the NS weight $\left(\prod_{m \in M^n} \frac{y_m^{n+1}}{\sigma_m^n} \right)^{r^n}$ depends on $\mathbf{y}^n$, further investigation is needed to conclude the order of convergence of the method by means of Theorem 2.3.

Second Order gBBKS Schemes The second order gBBKS schemes, denoted by gBBKS2(α), have a free parameter $\alpha \geq \frac{1}{2}$ and can be written as

$$\begin{aligned}
\mathbf{y}^{(1)} &= \mathbf{y}^n, \\
\mathbf{y}^{(2)} &= \mathbf{y}^n + \alpha \Delta t \mathbf{f}(\mathbf{y}^n, t_n) \left(\prod_{j \in J^n} \frac{y_j^{(2)}}{\pi_j^n} \right)^{q^n}, \\
\mathbf{y}^{n+1} &= \mathbf{y}^n + \Delta t \left(\left(1 - \frac{1}{2\alpha}\right) \mathbf{f}(\mathbf{y}^n, t_n) + \frac{1}{2\alpha} \mathbf{f}(\mathbf{y}^{(2)}, t_n + \alpha \Delta t) \right) \left(\prod_{m \in M^n} \frac{y_m^{n+1}}{\sigma_m^n} \right)^{r^n}
\end{aligned} \tag{gBBKS2}$$

with $\pi_j^n, q^n > 0$ being free parameters chosen independently of $\mathbf{y}^{(2)}$, while we require $\sigma_m^n, r^n > 0$ to be independent of $\mathbf{y}^{n+1}$. To give an example, the BBKS2(1) scheme from [BRBM08, AKM20] uses $\pi_m^n = \sigma_m^n = y_m^n$ and $q^n = r^n = 1$. Moreover, the sets J^n and M^n are given by

$$\begin{aligned}
J^n &= \{j \in \{1, \dots, N\} \mid f_j(\mathbf{y}^n, t_n) < 0\}, \\
M^n &= \left\{ m \in \{1, \dots, N\} \;\middle|\; \left(1 - \frac{1}{2\alpha}\right) f_m(\mathbf{y}^n, t_n) + \frac{1}{2\alpha} f_m(\mathbf{y}^{(2)}, t_n + \alpha \Delta t) < 0 \right\}.
\end{aligned}$$

We want to note that M^n always refers to the last step of the corresponding method. As before, the same concerns arise for (gBBKS2) when it comes to the order of convergence in general and in the case of non-autonomous problems.

Chapter 4

Order Conditions for NSARK Methods

In this chapter we are interested in deriving order conditions for general NSARK methods. As an application of the upcoming theory, we will reproduce known order conditions for MPRK and GeCo methods from [KM18a, KM18b, MCD20]. Additionally, we present reduced conditions for MPRK and GeCo schemes up to order four.

The main idea is to follow [But16] and to adapt Theorem 2.6 for schemes with solution-dependent coefficients.

4.1 Main Result on Order Conditions

In the appendix, we prove modified versions of theorems from [But16] to demonstrate that for an NSARK scheme we can take the formula for u from (2.13) and replace the constant coefficients with the solution-dependent ones from (NSARK), i. e. that the solution-dependent $u = u(\tau, \mathbf{y}^n, \Delta t)$ in the case of an NSARK method is given by

$$\begin{aligned} u(\tau, \mathbf{y}^n, \Delta t) &= \sum_{\nu=1}^{N} \sum_{i=1}^{s} b_i^{[\nu]}(\mathbf{y}^n, \Delta t) g_i^{[\nu]}(\tau, \mathbf{y}^n, \Delta t), \\ g_i^{[\nu]}(\bullet^{[\mu]}, \mathbf{y}^n, \Delta t) &= \delta_{\nu\mu}, \quad \nu, \mu = 1, \dots, N, \\ g_i^{[\nu]}([\tau_1, \dots, \tau_l]^{[\mu]}, \mathbf{y}^n, \Delta t) &= \delta_{\nu\mu} \prod_{j=1}^{l} d_i(\tau_j, \mathbf{y}^n, \Delta t), \quad \nu, \mu = 1, \dots, N \text{ and} \\ d_i(\tau, \mathbf{y}^n, \Delta t) &= \sum_{\nu=1}^{N} \sum_{j=1}^{s} a_{ij}^{[\nu]}(\mathbf{y}^n, \Delta t) g_j^{[\nu]}(\tau, \mathbf{y}^n, \Delta t). \end{aligned} \tag{4.1}$$

As a result of this claim, we would be in the position to formulate an analogous condition to (2.14) for an NSARK method to have an order of at least p.

To prove our main result, we introduce in Theorem 4.1 a generalization of NB-series, in which the coefficients of the series are allowed to depend on $\mathbf{y}^n$ and Δt. We note that such a series is *not* a Taylor expansion in Δt, but instead can be understood as an asymptotic expansion in expressions depending on powers of Δt and the solution-dependent coefficients of the Butcher tableau. As a result of this approach, we do not require at this point any regularity of $a_{ij}^{[\nu]}(\mathbf{y}^n, \Delta t)$

or $b_j^{[\nu]}(\mathbf{y}^n, \Delta t)$. But for our present purposes the current representation is more convenient. The results in this section are analogous to results in [But16], and we follow many of the ideas employed therein. The proofs of the intermediate results can be found in the appendix, so that we directly present and prove the main theorem analogously to Theorem 313B in [But16].

Moreover, since we have already discussed the circumstances under which the analysis of the convergence order can be reduced to the study of autonomous problems, we will limit ourselves to this case for the sake of simplicity.

Theorem 4.1. Let d_i, $g_i^{[\nu]}$ and u be defined as in (4.1) for $i = 1, \dots, s$ and $\nu = 1, \dots, N$. Suppose that for small enough Δt there exists a solution to the stage equations (NSARK) of the NSARK method, that $\mathbf{f}^{[\nu]} \in \mathcal{C}^{p+1}$ for $p \in \mathbb{N}$ is Lipschitz continuous, and that $a_{ij}^{[\nu]} = \mathcal{O}(1)$ (with respect to Δt, as $\Delta t \to 0$) for all $\nu = 1, \dots, N$. Then the stages, stage derivatives and output of the NSARK method can be expressed as

$$\mathbf{y}^{(i)} = \mathbf{y}^n + \sum_{\tau \in NT_p} \frac{\Delta t^{|\tau|}}{\sigma(\tau)} d_i(\tau, \mathbf{y}^n, \Delta t)\mathcal{F}(\tau)(\mathbf{y}^n) + \mathcal{O}(\Delta t^{p+1}), \tag{4.2a}$$

$$\Delta t \mathbf{f}^{[\nu]}(\mathbf{y}^{(i)}) = \sum_{\tau \in NT_p} \frac{\Delta t^{|\tau|}}{\sigma(\tau)} g_i^{[\nu]}(\tau, \mathbf{y}^n, \Delta t)\mathcal{F}(\tau)(\mathbf{y}^n) + \mathcal{O}(\Delta t^{p+1}), \tag{4.2b}$$

$$\mathbf{y}^{n+1} = \mathbf{y}^n + \sum_{\tau \in NT_p} \frac{\Delta t^{|\tau|}}{\sigma(\tau)} u(\tau, \mathbf{y}^n, \Delta t)\mathcal{F}(\tau)(\mathbf{y}^n) + \mathcal{O}(\Delta t^{p+1}). \tag{4.2c}$$

for $i = 1, \dots, s$ and $\nu = 1, \dots, N$.

Proof. We follow the idea from [But16, Theorem 313B]. For approximating the stage $\mathbf{y}^{(i)}$, define the sequence

$$\begin{aligned}
\mathbf{y}_{[0]}^{(i)} &= \mathbf{y}^n, \\
\mathbf{y}_{[m]}^{(i)} &= \mathbf{y}^n + \Delta t \sum_{j=1}^{s} \sum_{\nu=1}^{N} a_{ij}^{[\nu]}(\mathbf{y}^n, \Delta t)\mathbf{f}^{[\nu]}(\mathbf{y}_{[m-1]}^{(j)}),
\end{aligned} \tag{4.3}$$

where we want to point out that $a_{ij}^{[\nu]}(\mathbf{y}^n, \Delta t)$ here only depends on the solution $\mathbf{y}^n$, the step size Δt and, potentially, the assumed solution to the stage equations, but *not* on the iterates $\mathbf{y}_{[m]}^{(i)}$.

Next, we demonstrate that for $m \le p$, this expression for $\mathbf{y}_{[m]}^{(i)}$ agrees with the expression for $\mathbf{y}^{(i)}$ from (4.2a) within an error of $\mathcal{O}(\Delta t^{m+1})$. For $m = 0$, this is obvious. By induction we suppose that

$$\mathbf{y}_{[m-1]}^{(i)} = \mathbf{y}^n + \sum_{\tau \in NT_{m-1}} \frac{\Delta t^{|\tau|}}{\sigma(\tau)} d_i(\tau, \mathbf{y}^n, \Delta t)\mathcal{F}(\tau)(\mathbf{y}^n) + \mathcal{O}(\Delta t^m).$$

By Lemma B.3, we see that

$$\Delta t \mathbf{f}^{[\nu]}(\mathbf{y}_{[m-1]}^{(i)}) = \sum_{\tau \in NT_m} \frac{\Delta t^{|\tau|}}{\sigma(\tau)} g_i^{[\nu]}(\tau, \mathbf{y}^n, \Delta t)\mathcal{F}(\tau)(\mathbf{y}^n) + \mathcal{O}(\Delta t^{m+1}).$$

Substituting this into (4.3), we see from (4.1) that

$$\begin{aligned}\mathbf{y}^{(i)}_{[m]} &= \mathbf{y}^n + \sum_{\tau\in NT_m} \frac{\Delta t^{|\tau|}}{\sigma(\tau)} \sum_{j=1}^{s}\sum_{\nu=1}^{N} a^{[\nu]}_{ij}(\mathbf{y}^n,\Delta t) g^{[\nu]}_j(\tau,\mathbf{y}^n,\Delta t)\mathcal{F}(\tau)(\mathbf{y}^n) + \mathcal{O}(\Delta t^{m+1}) \\ &= \mathbf{y}^n + \sum_{\tau\in NT_m} \frac{\Delta t^{|\tau|}}{\sigma(\tau)} d_i(\tau,\mathbf{y}^n,\Delta t)\mathcal{F}(\tau)(\mathbf{y}^n) + \mathcal{O}(\Delta t^{m+1}).\end{aligned} \tag{4.4}$$

We have shown now that (4.4) is true for all $m \le p$. Indeed, by the same reasoning we have even proven that

$$\mathbf{y}^{(i)}_{[m]} = \mathbf{y}^n + \sum_{\tau\in NT_p} \frac{\Delta t^{|\tau|}}{\sigma(\tau)} d_i(\tau,\mathbf{y}^n,\Delta t)\mathcal{F}(\tau)(\mathbf{y}^n) + \mathcal{O}(\Delta t^{p+1}) \quad \text{for all } m \ge p.$$

Moreover, for Δt small enough we know that $a^{[\nu]}_{ij}$ is bounded since we assumed $a^{[\nu]}_{ij}(\mathbf{y}^n,\Delta t) = \mathcal{O}(1)$ as $\Delta t \to 0$. Together with the Lipschitz continuity of $\mathbf{f}^{[\nu]}$, we thus conclude that for small enough Δt the iteration (4.3) is a contraction with $\mathbf{y}^{(i)} = \lim_{m\to\infty}\mathbf{y}^{(i)}_{[m]}$ being the unique limit. Thus, for Δt small enough and $\epsilon = \Delta t^{p+1} > 0$, there exist $N_\epsilon \in \mathbb{N}$ such that $\|\mathbf{y}^{(i)}_{[m]} - \mathbf{y}^{(i)}\| < \Delta t^{p+1}$ for all $m \ge N_\epsilon$. Without loss of generality we can choose $N_\epsilon \ge p$, so that we find $m \ge p$. This implies that

$$\mathbf{y}^{(i)} = \mathbf{y}^{(i)}_{[m]} + \mathcal{O}(\Delta t^{p+1}) = \mathbf{y}^n + \sum_{\tau\in NT_p} \frac{\Delta t^{|\tau|}}{\sigma(\tau)} d_i(\tau,\mathbf{y}^n,\Delta t)\mathcal{F}(\tau)(\mathbf{y}^n) + \mathcal{O}(\Delta t^{p+1}),$$

from which equation (4.2a) follows. Furthermore, (4.2b) then follows from Lemma B.3. Finally, computing $\mathbf{y}^{n+1}$ according to (NSARK), also taking into account equation (4.1), we obtain

$$\begin{aligned}\mathbf{y}^{n+1} &= \mathbf{y}^n + \sum_{j=1}^{s}\sum_{\nu=1}^{N} b^{[\nu]}_j(\mathbf{y}^n,\Delta t)\Delta t\mathbf{f}^{[\nu]}(\mathbf{y}^{(j)}) \\ &= \mathbf{y}^n + \sum_{\tau\in NT_p} \frac{\Delta t^{|\tau|}}{\sigma(\tau)} \sum_{j=1}^{s}\sum_{\nu=1}^{N} b^{[\nu]}_j(\mathbf{y}^n,\Delta t) g^{[\nu]}_j(\tau,\mathbf{y}^n,\Delta t)\mathcal{F}(\tau)(\mathbf{y}^n) + \mathcal{O}(\Delta t^{p+1}) \\ &= \mathbf{y}^n + \sum_{\tau\in NT_p} \frac{\Delta t^{|\tau|}}{\sigma(\tau)} u(\tau,\mathbf{y}^n,\Delta t)\mathcal{F}(\tau)(\mathbf{y}^n) + \mathcal{O}(\Delta t^{p+1}),\end{aligned}$$

finishing the proof. □

Note that under the assumptions of this theorem, any solution of the stage equations has the same expansion up to the order p. Moreover, we obtain the following order conditions as a result of this theorem, where the expression $\mathbf{A}^{[\nu]}(\mathbf{y}^n,\Delta t) = \mathcal{O}(1)$ should be understood component-wise and in the limit $\Delta t \to 0$.

Corollary 4.2. Let u be defined as in (4.1) and $\mathbf{f}^{[\nu]} \in \mathcal{C}^{p+1}$ for $\nu = 1, \dots, N$ be Lipschitz continuous. Furthermore, let $\mathbf{A}^{[\nu]}(\mathbf{y}^n, \Delta t) = \mathcal{O}(1)$. If the stage equations of the NSARK method possess a solution for small enough Δt, then the NSARK scheme (NSARK) is of order at least p if and only if

$$u(\tau, \mathbf{y}^n, \Delta t) = \frac{1}{\gamma(\tau)} + \mathcal{O}(\Delta t^{p+1-|\tau|}), \quad \forall \tau \in NT_p. \tag{4.5}$$

Corollary 4.3. Under the assumptions of Theorem 4.1, if

$$a_{ij}^{[\nu]}(\mathbf{y}^n, \Delta t) = a_{ij}^{[\nu]} + \mathcal{O}(\Delta t^{p-1}) \quad \text{and} \quad b_j^{[\nu]}(\mathbf{y}^n, \Delta t) = b_j^{[\nu]} + \mathcal{O}(\Delta t^p),$$

for $i, j = 1, \dots, s$ and $\nu = 1, \dots, N$, the NSARK method (NSARK) applied to autonomous problems is of order p, if $\mathbf{A}^{[\nu]} = (a_{ij}^{[\nu]})_{i,j=1,\dots,s}$, $\mathbf{b}^{[\nu]} = (b_1^{[\nu]}, \dots, b_N^{[\nu]})$ define an ARK method of order p.

Proof. Inserting the assumptions into (NSARK) yields

$$\begin{aligned}
\mathbf{y}^{(i)} &= \mathbf{y}^n + \Delta t \sum_{j=1}^{s} \sum_{\nu=1}^{N} a_{ij}^{[\nu]} \mathbf{f}^{[\nu]}(\mathbf{y}^{(j)}) + \mathcal{O}(\Delta t^p), \\
\mathbf{y}^{n+1} &= \mathbf{y}^n + \Delta t \sum_{j=1}^{s} \sum_{\nu=1}^{N} b_j^{[\nu]} \mathbf{f}^{[\nu]}(\mathbf{y}^{(j)}) + \mathcal{O}(\Delta t^{p+1}).
\end{aligned}$$

According to Theorem 4.1 and Lemma B.3 we see

$$\Delta t \mathbf{f}^{[\nu]}(\mathbf{y}^{(i)}) = \sum_{\tau \in NT_p} \frac{\Delta t^{|\tau|}}{\sigma(\tau)} g_i^{[\nu]}(\tau) \mathcal{F}(\tau)(\mathbf{y}^n) + \mathcal{O}(\Delta t^{p+1}).$$

Consequently, (2.13) implies that

$$\mathbf{y}^{n+1} = \mathbf{y}^n + \sum_{\tau \in NT_p} \frac{\Delta t^{|\tau|}}{\sigma(\tau)} u(\tau) \mathcal{F}(\tau)(\mathbf{y}^n) + \mathcal{O}(\Delta t^{p+1}).$$

Finally, since the corresponding underlying ARK scheme is of order p the claim follows. □

In order to grasp the condition (4.5) from Corollary 4.2, we collect the value of u for all $\tau \in NT_4$ in Table 4.1.

Remark 4.4. Using Corollary 4.2 and Table 4.1, the condition for $p = 1$ reads

$$\sum_{i=1}^{s} b_i^{[\mu]}(\mathbf{y}^n, \Delta t) = 1 + \mathcal{O}(\Delta t), \quad \mu = 1, \dots, N. \tag{4.6}$$

For $p = 2$ we find the conditions

$$\begin{aligned}
\sum_{i=1}^{s} b_i^{[\mu]}(\mathbf{y}^n, \Delta t) &= 1 + \mathcal{O}(\Delta t^2), \qquad \mu = 1, \dots, N, \\
\sum_{i,j=1}^{s} b_i^{[\mu]}(\mathbf{y}^n, \Delta t) a_{ij}^{[\nu]}(\mathbf{y}^n, \Delta t) &= \frac{1}{2} + \mathcal{O}(\Delta t), \quad \mu, \nu = 1, \dots, N,
\end{aligned} \tag{4.7}$$

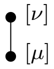
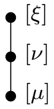
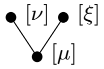
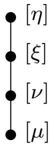
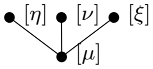
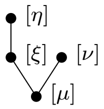
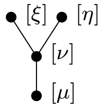
τ	$\gamma(\tau)$	$u(\tau,\mathbf{y}^n,\Delta t)$
	1	$\sum_{i=1}^s b_i^{[\mu]}(\mathbf{y}^n,\Delta t)$
	2	$\sum_{i,j=1}^s b_i^{[\mu]}(\mathbf{y}^n,\Delta t)a_{ij}^{[\nu]}(\mathbf{y}^n,\Delta t)$
	6	$\sum_{i,j,k=1}^s b_i^{[\mu]}(\mathbf{y}^n,\Delta t)a_{ij}^{[\nu]}(\mathbf{y}^n,\Delta t)a_{jk}^{[\xi]}(\mathbf{y}^n,\Delta t)$
	3	$\sum_{i,j,k=1}^s b_i^{[\mu]}(\mathbf{y}^n,\Delta t)a_{ij}^{[\nu]}(\mathbf{y}^n,\Delta t)a_{ik}^{[\xi]}(\mathbf{y}^n,\Delta t)$
	24	$\sum\limits_{i,j,k,l=1}^s b_i^{[\mu]}(\mathbf{y}^n,\Delta t)a_{ij}^{[\nu]}(\mathbf{y}^n,\Delta t)a_{jk}^{[\xi]}(\mathbf{y}^n,\Delta t)a_{kl}^{[\eta]}(\mathbf{y}^n,\Delta t)$
	4	$\sum\limits_{i,j,k,l=1}^s b_i^{[\mu]}(\mathbf{y}^n,\Delta t)a_{il}^{[\eta]}(\mathbf{y}^n,\Delta t)a_{ij}^{[\nu]}(\mathbf{y}^n,\Delta t)a_{ik}^{[\xi]}(\mathbf{y}^n,\Delta t)$
	8	$\sum\limits_{i,j,k,l=1}^s b_i^{[\mu]}(\mathbf{y}^n,\Delta t)a_{il}^{[\nu]}(\mathbf{y}^n,\Delta t)a_{ij}^{[\xi]}(\mathbf{y}^n,\Delta t)a_{jk}^{[\eta]}(\mathbf{y}^n,\Delta t)$
	12	$\sum\limits_{i,j,k,l=1}^s b_i^{[\mu]}(\mathbf{y}^n,\Delta t)a_{ij}^{[\nu]}(\mathbf{y}^n,\Delta t)a_{jk}^{[\xi]}(\mathbf{y}^n,\Delta t)a_{jl}^{[\eta]}(\mathbf{y}^n,\Delta t)$

Table 4.1: Density γ from (2.11) and value of u from (4.1) for $\tau \in NT_4$.

and for $p=3$ we obtain

$$\begin{aligned} \sum_{i=1}^s b_i^{[\mu]}(\mathbf{y}^n,\Delta t) &= 1+\mathcal{O}(\Delta t^3), && \mu=1,\dots,N,\\ \sum_{i,j=1}^s b_i^{[\mu]}(\mathbf{y}^n,\Delta t)a_{ij}^{[\nu]}(\mathbf{y}^n,\Delta t) &= \frac{1}{2}+\mathcal{O}(\Delta t^2), && \mu,\nu=1,\dots,N,\\ \sum_{i,j,k=1}^s b_i^{[\mu]}(\mathbf{y}^n,\Delta t)a_{ij}^{[\nu]}(\mathbf{y}^n,\Delta t)a_{ik}^{[\xi]}(\mathbf{y}^n,\Delta t) &= \frac{1}{3}+\mathcal{O}(\Delta t), && \mu,\nu,\xi=1,\dots,N,\\ \sum_{i,j,k=1}^s b_i^{[\mu]}(\mathbf{y}^n,\Delta t)a_{ij}^{[\nu]}(\mathbf{y}^n,\Delta t)a_{jk}^{[\xi]}(\mathbf{y}^n,\Delta t) &= \frac{1}{6}+\mathcal{O}(\Delta t), && \mu,\nu,\xi=1,\dots,N. \end{aligned} \tag{4.8}$$

As we derive also 4th order conditions for GeCo and MPRK schemes in the next sections, we also present the general conditions for $p = 4$ reading

$$\begin{aligned}
\sum_{i=1}^{s} b_i^{[\mu]}(\mathbf{y}^n, \Delta t) &= 1 + \mathcal{O}(\Delta t^4),\\
\sum_{i,j=1}^{s} b_i^{[\mu]}(\mathbf{y}^n, \Delta t) a_{ij}^{[\nu]}(\mathbf{y}^n, \Delta t) &= \frac{1}{2} + \mathcal{O}(\Delta t^3),\\
\sum_{i,j,k=1}^{s} b_i^{[\mu]}(\mathbf{y}^n, \Delta t) a_{ij}^{[\nu]}(\mathbf{y}^n, \Delta t) a_{ik}^{[\xi]}(\mathbf{y}^n, \Delta t) &= \frac{1}{3} + \mathcal{O}(\Delta t^2),\\
\sum_{i,j,k=1}^{s} b_i^{[\mu]}(\mathbf{y}^n, \Delta t) a_{ij}^{[\nu]}(\mathbf{y}^n, \Delta t) a_{jk}^{[\xi]}(\mathbf{y}^n, \Delta t) &= \frac{1}{6} + \mathcal{O}(\Delta t^2),\\
\sum_{i,j,k,l=1}^{s} b_i^{[\mu]}(\mathbf{y}^n, \Delta t) a_{il}^{[\nu]}(\mathbf{y}^n, \Delta t) a_{ij}^{[\xi]}(\mathbf{y}^n, \Delta t) a_{jk}^{[\eta]}(\mathbf{y}^n, \Delta t) &= \frac{1}{8} + \mathcal{O}(\Delta t),\\
\sum_{i,j,k,l=1}^{s} b_i^{[\mu]}(\mathbf{y}^n, \Delta t) a_{il}^{[\eta]}(\mathbf{y}^n, \Delta t) a_{ij}^{[\nu]}(\mathbf{y}^n, \Delta t) a_{ik}^{[\xi]}(\mathbf{y}^n, \Delta t) &= \frac{1}{4} + \mathcal{O}(\Delta t),\\
\sum_{i,j,k,l=1}^{s} b_i^{[\mu]}(\mathbf{y}^n, \Delta t) a_{ij}^{[\nu]}(\mathbf{y}^n, \Delta t) a_{jk}^{[\xi]}(\mathbf{y}^n, \Delta t) a_{kl}^{[\eta]}(\mathbf{y}^n, \Delta t) &= \frac{1}{4!} + \mathcal{O}(\Delta t),\\
\sum_{i,j,k,l=1}^{s} b_i^{[\mu]}(\mathbf{y}^n, \Delta t) a_{ij}^{[\nu]}(\mathbf{y}^n, \Delta t) a_{jk}^{[\xi]}(\mathbf{y}^n, \Delta t) a_{jl}^{[\eta]}(\mathbf{y}^n, \Delta t) &= \frac{1}{12} + \mathcal{O}(\Delta t)
\end{aligned} \tag{4.9}$$

for $\mu, \nu, \xi, \eta = 1, \ldots, N$.

4.1.1 Application to Geometric Conservative Methods

In this section we derive the known order conditions for Geometric Conservative (GeCo) schemes [MCD20] and present for the first time order conditions for 3rd and 4th order.

As GeCo schemes are NSRK methods we can interpret them formally as NSARK methods in order to use Corollary 4.2. The resulting order conditions can easily be simplified somewhat, using the fact that the original coefficients a_{ij}, b_j satisfy traditional RK order conditions. In view of (3.20), the first condition is

$$\sum_{i=1}^{s} b_i \phi_{n+1}(\mathbf{y}^n, \Delta t) = 1 + \mathcal{O}(\Delta t^p),$$

which implies simply $\phi_{n+1}(\mathbf{y}^n, \Delta t) = 1 + \mathcal{O}(\Delta t^p)$. This turns out to allow us to neglect the factor ϕ_{n+1} in all the remaining order conditions. For instance, the next condition is

$$\sum_{i=1}^{s} b_i c_i \phi_{n+1}(\mathbf{y}^n, \Delta t) \phi_i(\mathbf{y}^n, \Delta t) = \frac{1}{2} + \mathcal{O}(\Delta t^{p-1}),$$

which is equivalent to

$$\sum_{i=1}^{s} b_i c_i \phi_i(\mathbf{y}^n, \Delta t) = \frac{1}{2} + \mathcal{O}(\Delta t^{p-1}).$$

With more work, we can use these conditions to obtain direct conditions on the functions ϕ for specific cases of s and p, as demonstrated in the following theorem.

Theorem 4.5. Let $\mathbf{A}, \mathbf{b}$ be the coefficients of an explicit RK scheme of order p with s stages satisfying $\sum_{j=1}^{s} a_{ij} = c_i$. Assume $\phi_i(\mathbf{y}^n, \Delta t) = \mathcal{O}(1)$ as $\Delta t \to 0$ for $i = 2, \dots, s$ and that $\mathbf{f} \in \mathcal{C}^{p+1}$ is Lipschitz continuous. Then

a) if $p = 1$, (GeCo) is of order 1 if and only if $\phi_{n+1}(\mathbf{y}^n, \Delta t) = 1 + \mathcal{O}(\Delta t)$.

b) if $p = s = 2$, (GeCo) is of order 2 if and only if $\phi_2(\mathbf{y}^n, \Delta t) = 1 + \mathcal{O}(\Delta t)$ and $\phi_{n+1}(\mathbf{y}^n, \Delta t) = 1 + \mathcal{O}(\Delta t^2)$.

c) if $p = s = 3$, (GeCo) is of order 3 if and only if

$$\begin{aligned} \phi_{n+1}(\mathbf{y}^n, \Delta t) &= 1 + \mathcal{O}(\Delta t^3), \\ \sum_{i=2}^{3} b_i c_i \phi_i(\mathbf{y}^n, \Delta t) &= \frac{1}{2} + \mathcal{O}(\Delta t^2), \\ \phi_i(\mathbf{y}^n, \Delta t) &= 1 + \mathcal{O}(\Delta t), \quad i = 2, 3. \end{aligned}$$

d) if $p = s = 4$, (GeCo) is of order 4 if and only if

$$\begin{aligned} \phi_{n+1}(\mathbf{y}^n, \Delta t) &= 1 + \mathcal{O}(\Delta t^4), \\ \sum_{i=2}^{4} b_i c_i \phi_i(\mathbf{y}^n, \Delta t) &= \frac{1}{2} + \mathcal{O}(\Delta t^3), \\ \phi_i(\mathbf{y}^n, \Delta t) &= 1 + \mathcal{O}(\Delta t^2), \quad i = 2, 3, 4. \end{aligned}$$

Proof. First of all, the assumptions of Theorem 4.1 and Corollary 4.2 are met. Thus, we can use the order conditions (4.6) to (4.9) as a basis of this proof.

a) Substituting $\sum_{i=1}^{s} b_i = 1$ into (4.6) yields $\phi_{n+1}(\mathbf{y}^n, \Delta t) = 1 + \mathcal{O}(\Delta t)$.

b) Using $\sum_{i=1}^{s} b_i = 1$ now in (4.7) together with $\sum_{j=1}^{s} a_{ij} = c_i$, the order conditions reduce to

$$\phi_{n+1}(\mathbf{y}^n, \Delta t) = 1 + \mathcal{O}(\Delta t^2),$$

and

$$\sum_{i=1}^{s} b_i \phi_{n+1}(\mathbf{y}^n, \Delta t) c_i \phi_i(\mathbf{y}^n, \Delta t) = \frac{1}{2} + \mathcal{O}(\Delta t).$$

The latter condition can be further simplified to

$$b_2 c_2 \phi_2(\mathbf{y}^n, \Delta t) = \frac{1}{2} + \mathcal{O}(\Delta t),$$

since $s = 2$ and $\phi_{n+1}(\mathbf{y}^n, \Delta t) = 1 + \mathcal{O}(\Delta t^2)$. As $b_2 c_2 = \frac{1}{2}$, this means that

$$\phi_2(\mathbf{y}^n, \Delta t) = 1 + \mathcal{O}(\Delta t).$$

c) Similar as before we obtain from (4.8) the simplified conditions

$$\begin{aligned}\phi_{n+1}(\mathbf{y}^n,\Delta t)&=1+\mathcal{O}(\Delta t^3),\\ \sum_{i=2}^{3}b_ic_i\phi_i(\mathbf{y}^n,\Delta t)&=\frac{1}{2}+\mathcal{O}(\Delta t^2),\\ \sum_{i=2}^{3}b_ic_i^2(\phi_i(\mathbf{y}^n,\Delta t))^2&=\frac{1}{3}+\mathcal{O}(\Delta t),\\ \sum_{i,j=2}^{3}b_ia_{ij}c_j\phi_i(\mathbf{y}^n,\Delta t)\phi_j(\mathbf{y}^n,\Delta t)&=\frac{1}{6}+\mathcal{O}(\Delta t),\end{aligned}$$

which by Lemma B.4 with $N=1$, $\gamma_1^{(i)}=\phi_i(\mathbf{y}^n,\Delta t)$ and $\delta_1=\phi_{n+1}(\mathbf{y}^n,\Delta t)$ are equivalent to the conditions stated in the Theorem.

d) As in the previous part, the conditions (4.9) are simplified resulting in (B.6) with $N=1$, $\gamma_1^{(i)}$ and δ_1 as before. These conditions are then reduced by Lemma B.5 resulting in the conditions given in this theorem.

□

With this result, we have shown that the conditions from [MCD20, Theorem 1] are also necessary. Moreover, we provided the very first necessary and sufficient order conditions for the construction of 3rd and 4th order GeCo schemes.

4.1.2 Application to Modified Patankar–Runge–Kutta Methods

As we have discussed in Section 3.2, modified Patankar–Runge–Kutta methods (MPRK) originally were constructed for conservative and positive PDS of the form (2.21). Moreover, we concluded in that section that we may assume without loss of generality that the PDS is autonomous.

Until now, sufficient and necessary order conditions for MPRK schemes only up to order three were constructed and in the context of autonomous PDS. However, these order conditions are also valid in the context of a non-autonomous PDRS, as the NS weights are either the same as in the PDS case or equal to 1, see (3.2).

Thus, in order to obtain order conditions for even higher order MPRK schemes in the context of a PDRS, we actually can restrict to autonomous PDS where the solution-dependent Butcher tableau is determined by

$$a_{ij}^{[\nu]}(\mathbf{y}^n,\Delta t)=a_{ij}\frac{y_\nu^{(i)}}{\pi_\nu^{(i)}}\quad\text{and}\quad b_j^{[\nu]}(\mathbf{y}^n,\Delta t)=b_j\frac{y_\nu^{n+1}}{\sigma_\nu}.$$

Note that, since $a_{1j}=0$, the value of $\pi_\nu^{(1)}$ has no effect. In order to apply Theorem 4.1 and Corollary 4.2, we show in the next lemma that the stages are uniquely determined for any $\Delta t\geq 0$ and that $\mathbf{A}^{[\nu]}(\mathbf{y}^n,\Delta t)=\mathcal{O}(1)$. The key observation to prove this is that $\pi_\nu^{(i)},\sigma_\nu$ are positive even for $\Delta t=0$ by definition. Moreover, we assume that the PWDs are continuous functions of $\mathbf{y}^n$ and the stages, that is $\pi_\nu^{(i)}=\pi_\nu^{(i)}(\mathbf{y}^n,\mathbf{y}^{(1)},\dots,\mathbf{y}^{(i-1)})$ and $\sigma_\nu=\sigma_\nu(\mathbf{y}^n,\mathbf{y}^{(1)},\dots,\mathbf{y}^{(s)})$, which is fulfilled by the PWDs introduced so far. Also, as we will apply the lemma in the context of the local error analysis, we may start with some $\mathbf{y}^n$ representing the exact solution at a given time level t_n.

Lemma 4.6. An MPRK scheme (3.8) with a given positive vector $\mathbf{y}^n$ has uniquely determined stages and satisfies $\mathbf{y}^{(i)}, \mathbf{y}^{n+1} = \mathcal{O}(1)$ as $\Delta t \to 0$. Moreover, if $p_{k\nu}, d_{k\nu} \in \mathcal{C}$ and $\pi_\nu^{(i)}(\mathbf{y}^n, \mathbf{y}^{(1)}, \dots, \mathbf{y}^{(i-1)}), \sigma_\nu(\mathbf{y}^n, \mathbf{y}^{(1)}, \dots, \mathbf{y}^{(s)}) > 0$ are continuous functions of $\mathbf{y}^n$ and the stages, then $\pi_\nu^{(i)}, \sigma_\nu = \mathcal{O}(1)$ for $i, \nu = 1, \dots, N$. In addition, the modified Butcher coefficients from (3.6) satisfy $\mathbf{A}^{[\nu]}(\mathbf{y}^n, \Delta t) = \mathcal{O}(1)$ and $\mathbf{b}^{[\nu]}(\mathbf{y}^n, \Delta t) = \mathcal{O}(1)$.

Proof. According to Lemma 3.7, there exist unique matrices $\mathbf{M}^{(i)}$, $i = 1, \dots, s$ and $\mathbf{M}$, with inverses in $\mathcal{O}(1)$ as $\Delta t \to 0$, such that the stage vectors satisfy the equations $\mathbf{y}^{(i)} = \left(\mathbf{M}^{(i)}\right)^{-1} \mathbf{y}^n = \mathcal{O}(1)$ and $\mathbf{y}^{n+1} = \mathbf{M}^{-1}\mathbf{y}^n = \mathcal{O}(1)$. Now, since $p_{k\nu}, d_{k\nu}, \pi_\nu^{(i)}, \sigma_\nu \in \mathcal{C}$, we conclude by induction over i that the stage vectors are continuous functions of Δt themselves by pointing out that every entry in $\mathbf{M}^{(i)}, \mathbf{M}$ is a continuous function of Δt. Hence, even $\pi_\nu^{(i)}$ and σ_ν are continuous functions of Δt, so that we conclude $\pi_\nu^{(i)} = \mathcal{O}(1)$ and $\sigma_\nu = \mathcal{O}(1)$ as $\Delta t \to 0$. Since even $\pi_\nu^{(i)}, \sigma_\nu > 0$ for $\Delta t = 0$, we deduce from the continuity in Δt that there is a positive lower bound also for $\Delta t > 0$ small enough. This gives us $\frac{y_\nu^{(i)}}{\pi_\nu^{(i)}} = \mathcal{O}(1)$ and $\frac{y_\nu^{n+1}}{\sigma_\nu} = \mathcal{O}(1)$, from which the claim follows. □

Using this lemma and Corollary 4.2 we are able to provide necessary and sufficient conditions for arbitrary high order NSARK schemes, to which MPRK methods belong. However, those conditions are in general implicit, since the NS weights depend on the stages. In the next two subsections, for specific classes of MPRK methods we reformulate these conditions to be explicit.

Remark 4.7. At this point we should discuss the situation mentioned in Remark 3.9. To prove an analogue of Corollary 4.2 for MPRK schemes based on a Butcher tableau with partially negative entries, we first note that Remark 2.7 tells us that using the index function (3.5) to switch the PWDs corresponds to switching the colors in the corresponding N-tree. Now, since the condition (4.5) needs to be satisfied for all colored N-trees in NT_p, the order conditions for MPRK schemes do not depend on the sign of the Butcher tableau.

Order Conditions for MPRK Schemes from the Literature

In this subsection we focus on reformulating the order conditions from our theory deriving the sufficient and necessary conditions from the literature, that is the conditions up to order three. Note that, as discussed in Remark 4.7, the order conditions for an MPRK scheme do not depend on the sign of the entries of the Butcher tableau. In the following, we thus assume without loss of generality that $\mathbf{A}, \mathbf{b} \geq \mathbf{0}$, so that we can use the representation (3.8) of the MPRK scheme. Furthermore, we assume throughout this section that $\mathbf{f}^{[\nu]}$ is Lipschitz continuous for all $\nu = 1, \dots, N$ and $\sum_{j=1}^s a_{ij} = c_i$.

First we give a lemma that we will repeatedly use throughout this section without further notice.

Lemma 4.8. For given scalars x, y with $y \neq 0$ the identity

$$\frac{x + \mathcal{O}(\Delta t^p)}{y + \mathcal{O}(\Delta t^p)} = \frac{x}{y} + \mathcal{O}(\Delta t^p)$$

holds true.

Proof. Let $f, g = \mathcal{O}(\Delta t^p)$. Then, as $y \neq 0$, we find

$$\frac{x + f(\Delta t)}{y + g(\Delta t)} - \frac{x}{y} = \frac{yf(\Delta t) - xg(\Delta t)}{y(y + g(\Delta t))} = \frac{f(\Delta t)}{y + g(\Delta t)} - \frac{xg(\Delta t)}{y(y + g(\Delta t))}.$$

Now since $\lim_{\Delta t\to 0} g(\Delta t) = 0$, the denominators of both fractions on the right-hand side tend to constants as $\Delta t \to 0$. By definition, we know $\limsup_{\Delta t\to 0} \frac{f(\Delta t)}{\Delta t^p} < \infty$ and $\limsup_{\Delta t\to 0} \frac{g(\Delta t)}{\Delta t^p} < \infty$. Thus, the claim follows from

$$\limsup_{\Delta t\to 0} \frac{\frac{x+f(\Delta t)}{y+g(\Delta t)} - \frac{x}{y}}{\Delta t^p} \leq \limsup_{\Delta t\to 0} \frac{f(\Delta t)}{(y + g(\Delta t))\Delta t^p} + \left| \limsup_{\Delta t\to 0} \frac{xg(\Delta t)}{y(y + g(\Delta t))\Delta t^p} \right| < \infty.$$

□

To formulate the conditions up to the order $p = 3$, we observe from the general conditions (4.6), (4.7) and (4.8) that we should expand $a_{ij}^{[\nu]}(\mathbf{y}^n, \Delta t)$ up to an error of $\mathcal{O}(\Delta t^2)$. As we will see, it suffices for our current purposes to assume $\frac{y_\nu^{(i)}}{\pi_\nu^{(i)}} = 1 + \mathcal{O}(\Delta t)$ for deriving these expansions.

Lemma 4.9. Let $a_{ij}^{[\nu]}(\mathbf{y}^n, \Delta t) = a_{ij} \frac{y_\nu^{(i)}}{\pi_\nu^{(i)}}$, where $\mathbf{y}^{(i)}$ is the ith stage of an MPRK method (3.8). Moreover, let $\mathbf{f}^{[\nu]} \in \mathcal{C}^2$ for $\nu = 1, \dots, N$, and $\mathbf{f} = \sum_{\nu=1}^N \mathbf{f}^{[\nu]}$ be the right-hand side of (2.5). If

$$\frac{y_\nu^{(i)}}{\pi_\nu^{(i)}} = 1 + \mathcal{O}(\Delta t), \quad \nu = 1, \dots, N,$$

then

$$y_\nu^{(i)} = y_\nu^n + \Delta t c_i f_\nu(\mathbf{y}^n) + \mathcal{O}(\Delta t^2), \quad \nu = 1, \dots, N,$$

and in particular,

$$a_{ij}^{[\nu]}(\mathbf{y}^n, \Delta t) = a_{ij} \frac{y_\nu^n + \Delta t c_i f_\nu(\mathbf{y}^n)}{\pi_\nu^{(i)}} + \mathcal{O}(\Delta t^2), \quad \nu = 1, \dots, N.$$

Proof. The conditions for applying Theorem 4.1 with $k = 1$ are met due to Lemma 4.6, so that we can use the expansion of the stages to obtain

$$y_\nu^{(i)} = y_\nu^n + \Delta t \sum_{\mu=1}^N d_i(\bullet^{[\mu]}, \mathbf{y}^n, \Delta t)(\mathcal{F}(\bullet^{[\mu]})(\mathbf{y}^n))_\nu + \mathcal{O}(\Delta t^2), \quad \nu = 1, \dots, N. \tag{4.10}$$

Next, we substitute our assumption for the Patankar weights into $a_{ij}^{[\mu]}(\mathbf{y}^n, \Delta t)$ to receive

$$a_{ij}^{[\mu]}(\mathbf{y}^n, \Delta t) = a_{ij} \frac{y_\mu^{(i)}}{\pi_\mu^{(i)}} = a_{ij} + \mathcal{O}(\Delta t), \quad \mu = 1, \dots, N$$

and find

$$\begin{aligned} d_i(\bullet^{[\mu]}, \mathbf{y}^n, \Delta t) &= \sum_{\nu=1}^N \sum_{j=1}^s a_{ij}^{[\nu]}(\mathbf{y}^n, \Delta t) g_j^{[\nu]}(\bullet^{[\mu]}, \mathbf{y}^n, \Delta t) = \sum_{j=1}^s a_{ij}^{[\mu]}(\mathbf{y}^n, \Delta t) \\ &= \sum_{j=1}^s a_{ij} + \mathcal{O}(\Delta t) = c_i + \mathcal{O}(\Delta t). \end{aligned}$$

Finally, the claim follows from $\sum_{\mu=1}^N \mathcal{F}(\bullet^{[\mu]})(\mathbf{y}^n) = \sum_{\mu=1}^N \mathbf{f}^{[\mu]}(\mathbf{y}^n) = \mathbf{f}(\mathbf{y}^n)$. □

Another helpful result for deriving the known order conditions from the literature is the following.

Lemma 4.10. Let $\mathbf{A}, \mathbf{b}, \mathbf{c}$ describe an explicit s-stage Runge–Kutta method of at least order p for some $p \in \mathbb{N}$. Consider the corresponding MPRK scheme (3.8) and assume $\mathbf{F} \in \mathcal{C}^{p+1}$. If the MPRK method is of order p, then

$$\sigma_\mu = (\mathrm{NB}_{p-1}(\tfrac{1}{\gamma}, \mathbf{y}^n))_\mu + \mathcal{O}(\Delta t^p), \quad \mu = 1, \dots, N.$$

This means that σ_μ defines an embedded method of order $p-1$.

Proof. The MPRK scheme is of order p, i. e. $\mathbf{y}^{n+1} = \mathrm{NB}_p(\frac{1}{\gamma}, \mathbf{y}^n) + \mathcal{O}(\Delta t^{p+1})$. Next, according to Lemma 4.6 we can apply Corollary 4.2 to see

$$u(\bullet^{[\mu]}, \mathbf{y}^n, \Delta t) = \sum_{j=1}^s b_j \frac{y_\mu^{n+1}}{\sigma_\mu} = 1 + \mathcal{O}(\Delta t^p)$$

for $\mu = 1, \dots, N$, which is equivalent to $\frac{y_\mu^{n+1}}{\sigma_\mu} = 1 + \mathcal{O}(\Delta t^p)$ for $\mu = 1, \dots, N$ as $\sum_{j=1}^s b_j = 1$. Using Lemma 4.6 once again we find $\sigma_\mu = \mathcal{O}(1)$ yielding

$$\sigma_\mu = y_\mu^{n+1} + \mathcal{O}(\Delta t^p) = (\mathrm{NB}_{p-1}(\tfrac{1}{\gamma}, \mathbf{y}^n))_\mu + \mathcal{O}(\Delta t^p), \quad \mu = 1, \dots, N.$$

□

We are now in the position to derive the known order conditions from [KM18a, KM18b] for MPRK schemes up to order 3.

Theorem 4.11. Let $\mathbf{f}^{[\nu]} \in \mathcal{C}^2$ for $\nu = 1, \dots, N$ and $\mathbf{A}, \mathbf{b}, \mathbf{c}$ describe an explicit s-stage Runge–Kutta method of order at least 1. The corresponding MPRK scheme (3.8) is of order at least 1 if and only if

$$\sigma_\mu = y_\mu^n + \mathcal{O}(\Delta t), \quad \mu = 1, \dots, N. \tag{4.11}$$

Proof. The condition (4.6) for an MPRK scheme to be at least of order $p = 1$ reads

$$u(\bullet^{[\mu]}, \mathbf{y}^n, \Delta t) = \sum_{j=1}^s b_j \frac{y_\mu^{n+1}}{\sigma_\mu} = 1 + \mathcal{O}(\Delta t), \quad \mu = 1, \dots, N,$$

which can be simplified to $u(\bullet^{[\mu]}, \mathbf{y}^n, \Delta t) = \frac{y_\mu^{n+1}}{\sigma_\mu} = 1 + \mathcal{O}(\Delta t)$ for $\mu = 1, \dots, N$ as $\sum_{j=1}^s b_j = 1$. From Lemma 4.10 the condition $\sigma_\mu = y_\mu^n + \mathcal{O}(\Delta t)$ for $\mu = 1, \dots, N$ can be deduced.

Now let (4.11) be satisfied. It follows immediately from Lemma 4.6 and (4.2c) that $y_\nu^{n+1} = y_\nu^n + \mathcal{O}(\Delta t)$. Comparing with (4.11), this gives us

$$u(\bullet^{[\mu]}, \mathbf{y}^n, \Delta t) = \frac{y_\mu^{n+1}}{\sigma_\mu} = 1 + \mathcal{O}(\Delta t)$$

for $\mu = 1, \dots, N$ proving that (4.11) is sufficient and necessary. □

Theorem 4.12. Let $\mathbf{f}^{[\nu]} \in \mathcal{C}^3$ for $\nu = 1, \dots, N$ and $\mathbf{A}, \mathbf{b}, \mathbf{c}$ describe an explicit 2-stage Runge–Kutta method of order two. Then the corresponding MPRK scheme (3.8) is of order two if and only if

$$\pi_\nu^{(2)} = y_\nu^n + \mathcal{O}(\Delta t), \qquad \nu = 1, \dots, N, \tag{4.12a}$$

$$\sigma_\mu = (\mathrm{NB}_1(\tfrac{1}{\gamma}, \mathbf{y}^n))_\mu + \mathcal{O}(\Delta t^2), \qquad \mu = 1, \dots, N. \tag{4.12b}$$

Proof. First we reduce the necessary and sufficient conditions for $p = 2$ from (4.7), which state

$$u(\bullet^{[\mu]}, \mathbf{y}^n, \Delta t) = \sum_{i=1}^{s} b_i \frac{y_\mu^{n+1}}{\sigma_\mu} = 1 + \mathcal{O}(\Delta t^2), \qquad \mu = 1, \dots, N,$$

$$u([\bullet^{[\nu]}]^{[\mu]}, \mathbf{y}^n, \Delta t) = \sum_{i,j=1}^{s} b_i \frac{y_\mu^{n+1}}{\sigma_\mu} a_{ij} \frac{y_\nu^{(i)}}{\pi_\nu^{(i)}} = \frac{1}{2} + \mathcal{O}(\Delta t), \quad \mu, \nu = 1, \dots, N.$$

Since $\sum_{i=1}^s b_i = 1$, the first equation can be reduced to $\frac{y_\mu^{n+1}}{\sigma_\mu} = 1 + \mathcal{O}(\Delta t^2)$ for $\mu = 1, \dots, N$. Plugging this information into the second condition and using $\sum_{j=1}^s a_{ij} = c_i$, we end up with the condition $\sum_{i=1}^s b_i c_i \frac{y_\nu^{(i)}}{\pi_\nu^{(i)}} = \frac{1}{2} + \mathcal{O}(\Delta t)$ for $\nu = 1, \dots, N$. Since we assumed $s = 2$, we can use $c_1 = 0$ to obtain the equivalent conditions

$$\frac{y_\mu^{n+1}}{\sigma_\mu} = 1 + \mathcal{O}(\Delta t^2),$$
$$\frac{y_\nu^{(2)}}{\pi_\nu^{(2)}} = 1 + \mathcal{O}(\Delta t)$$

for $\mu, \nu = 1, \dots, N$.

To prove the claim, first assume that the MPRK scheme has order $p = 2$. Then Lemma 4.9 and Lemma 4.10 yield the conditions from (4.12).

Now let (4.12) be satisfied. Using (4.12a) and Lemma 4.6 together with the expansion (4.2a) of the stages we see $y_\nu^{(2)} = y_\nu^n + \mathcal{O}(\Delta t)$, and hence,

$$\frac{y_\nu^{(2)}}{\pi_\nu^{(2)}} = 1 + \mathcal{O}(\Delta t).$$

Moreover, with (4.12b) we can apply Theorem 4.11 to find that the scheme is at least first order accurate, that is $y_\nu^{n+1} = (\mathrm{NB}_1(\frac{1}{\gamma}, \mathbf{y}^n))_\nu + \mathcal{O}(\Delta t^2)$. Comparing with (4.12b) we end up with $\frac{y_\nu^{n+1}}{\sigma_\nu} = 1 + \mathcal{O}(\Delta t^2)$. □

As one can see, the stage-dependent conditions for second order are

$$\frac{y_\mu^{n+1}}{\sigma_\mu} = 1 + \mathcal{O}(\Delta t^2),$$
$$\frac{y_\nu^{(2)}}{\pi_\nu^{(2)}} = 1 + \mathcal{O}(\Delta t)$$

and were reformulated in the above Theorem. Similarly, the simplified conditions from (B.4) with $\gamma_\nu^{(i)} = \frac{y_\nu^{(i)}}{\pi_\nu^{(i)}}$ and $\delta_\mu = \frac{y_\mu^{n+1}}{\sigma_\mu}$ are, by means of Lemma B.4, equivalent to

$$\frac{y_\mu^{n+1}}{\sigma_\mu} = 1 + \mathcal{O}(\Delta t^3), \qquad \mu = 1,\dots,N, \tag{4.13a}$$

$$\sum_{i=2}^{3} b_i c_i \frac{y_\nu^{(i)}}{\pi_\nu^{(i)}} = \frac{1}{2} + \mathcal{O}(\Delta t^2), \qquad \nu = 1,\dots,N, \tag{4.13b}$$

$$\frac{y_\nu^{(i)}}{\pi_\nu^{(i)}} = 1 + \mathcal{O}(\Delta t), \qquad \nu = 1,\dots,N, \quad i = 2,3. \tag{4.13c}$$

The next theorem decodes these conditions reformulating them in an explicit form.

Theorem 4.13. Let $\mathbf{A}, \mathbf{b}, \mathbf{c}$ describe an explicit 3-stage RK scheme and let $\mathbf{f}^{[\nu]} \in \mathcal{C}^4$ for $\nu = 1,\dots,N$. Then the corresponding MPRK scheme (3.8) is at least of order $p = s = 3$ if and only if

$$\sigma_\mu = (\mathrm{NB}_2(\tfrac{1}{\gamma}, \mathbf{y}^n))_\mu + \mathcal{O}(\Delta t^3), \quad \mu = 1,\dots,N, \tag{4.14a}$$

$$\sum_{i=2}^{3} b_i c_i \frac{y_\nu^n + \Delta t c_i f_\nu(\mathbf{y}^n)}{\pi_\nu^{(i)}} = \frac{1}{2} + \mathcal{O}(\Delta t^2), \qquad \nu = 1,\dots,N, \tag{4.14b}$$

$$\pi_\nu^{(i)} = y_\nu^n + \mathcal{O}(\Delta t), \qquad \nu = 1,\dots,N, \quad i = 2,3. \tag{4.14c}$$

Proof. We now show that the conditions (4.13) are equivalent to (4.14). First, assuming (4.13) is fulfilled, the MPRK scheme is of order 3. Thus, Lemma 4.10 implies (4.14a). Finally, with (4.13c) we are in the position to apply Lemma 4.9, which, together with (4.13b), yield the conditions (4.14c) and (4.14b).

Let's now suppose that (4.14) holds. The condition (4.13c) follows from (4.14c) and the expansion (4.2a) for the stages. Having derived (4.13c), we can apply Lemma 4.9 to obtain

$$y_\nu^{(i)} = y_\nu^n + \Delta t c_i f_\nu(\mathbf{y}^n) + \mathcal{O}(\Delta t^2), \quad \nu = 1,\dots,N.$$

Together with (4.14b) we can thus conclude (4.13b). Therefore, it remains to deduce condition (4.13a).

First of all, (4.14a) and Theorem 4.11 imply that the MPRK scheme is of order at least 1, which means that $\mathbf{y}^{n+1} = \mathrm{NB}_1(\frac{1}{\gamma}, \mathbf{y}^n) + \mathcal{O}(\Delta t^2)$. Comparing with (4.14a), we see

$$\frac{y_\mu^{n+1}}{\sigma_\mu} = 1 + \mathcal{O}(\Delta t^2), \quad \mu = 1,\dots,N.$$

Moreover, since we have already shown (4.13c), we can now verify that condition (4.7) is fulfilled which means that the MPRK scheme is even second order accurate. Therefore, we find $\mathbf{y}^{n+1} = \mathrm{NB}_2(\frac{1}{\gamma}, \mathbf{y}^n) + \mathcal{O}(\Delta t^3)$, so that a comparison with (4.14a) gives us (4.13a). □

We have now derived all known order conditions for MPRK schemes from the literature and even proved that they are valid for MPRK schemes based on $\mathbf{A}$ and $\mathbf{b}$ with negative entries, see Remark 4.7.

Reduced Order Conditions for 4th Order MPRK Methods

The main idea in deriving the known conditions for 3rd order MPRK schemes was to use Lemma B.4 for reducing the order conditions (4.8) and then substituting the expansions for the stages to obtain conditions depending only on $\mathbf{y}^n$ and Δt.

Similarly, we are in the position to derive conditions for 4th order by first using Lemma B.5 to reduce the order conditions (4.9). Now, in order to eliminate the dependency of the conditions on the stages, we need to expand $a_{ij}^{[\nu]}(\mathbf{y}^n,\Delta t)$ up to an error of $\mathcal{O}(\Delta t^3)$ giving an analogue to Lemma 4.9. However, since equation (B.7) in Lemma B.5 gives us $\gamma_\nu^{(i)}=1+\mathcal{O}(\Delta t^2)$, we will see that it suffices to prove the following lemma assuming $\frac{y_\nu^{(i)}}{\pi_\nu^{(i)}}=1+\mathcal{O}(\Delta t^2)$.

Lemma 4.14. Let $a_{ij}^{[\nu]}(\mathbf{y}^n,\Delta t)=a_{ij}\frac{y_\nu^{(i)}}{\pi_\nu^{(i)}}$, where $\mathbf{y}^{(i)}$ is the ith stage of an MPRK method (3.8). Moreover, let $\mathbf{f}^{[\nu]}\in\mathcal{C}^3$ for $\nu=1,\dots,N$, and $\mathbf{f}=\sum_{\nu=1}^N\mathbf{f}^{[\nu]}$ be the right-hand side of (2.5). If

$$\frac{y_\nu^{(i)}}{\pi_\nu^{(i)}}=1+\mathcal{O}(\Delta t^2),\quad \nu=1,\dots,N,$$

then

$$y_\nu^{(i)}=y_\nu^n+\Delta t c_i f_\nu(\mathbf{y}^n)+\frac12\Delta t^2\sum_{k=1}^s a_{ik}c_k(\mathbf{Df}(\mathbf{y}^n)\mathbf{f}(\mathbf{y}^n))_\nu+\mathcal{O}(\Delta t^3),\quad \nu=1,\dots,N,$$

and in particular,

$$a_{ij}^{[\nu]}(\mathbf{y}^n,\Delta t)=a_{ij}\frac{y_\nu^n+\Delta t c_i f_\nu(\mathbf{y}^n)+\frac12\Delta t^2\sum_{k=1}^s a_{ik}c_k(\mathbf{Df}(\mathbf{y}^n)\mathbf{f}(\mathbf{y}^n))_\nu}{\pi_\nu^{(i)}}+\mathcal{O}(\Delta t^3)$$

for $\nu=1,\dots,N$.

Proof. The conditions for applying Theorem 4.1 with $k=2$ are met due to Lemma 4.6, so that we can use the expansion of the stages to obtain

$$\begin{aligned}y_\nu^{(i)}=&y_\nu^n+\Delta t\sum_{\mu=1}^N d_i(\bullet^{[\mu]},\mathbf{y}^n,\Delta t)(\mathcal{F}(\bullet^{[\mu]})(\mathbf{y}^n))_\nu\\&+\frac12\Delta t^2\sum_{\mu,\eta=1}^N d_i\left(\left[\bullet^{[\eta]}\right]^{[\mu]},\mathbf{y}^n,\Delta t\right)\left(\mathcal{F}\left(\left[\bullet^{[\eta]}\right]^{[\mu]}\right)(\mathbf{y}^n)\right)_\nu+\mathcal{O}(\Delta t^3)\end{aligned}\tag{4.15}$$

for $\nu=1,\dots,N$. Moreover, we know that

$$a_{ij}^{[\mu]}(\mathbf{y}^n,\Delta t)=a_{ij}\frac{y_\mu^{(i)}}{\pi_\mu^{(i)}}=a_{ij}+\mathcal{O}(\Delta t^2),\quad \mu=1,\dots,N\tag{4.16}$$

as well as

$$d_i(\bullet^{[\mu]},\mathbf{y}^n,\Delta t)=\sum_{j=1}^s a_{ij}^{[\mu]}(\mathbf{y}^n,\Delta t)=\sum_{j=1}^s a_{ij}+\mathcal{O}(\Delta t^2)=c_i+\mathcal{O}(\Delta t^2)\tag{4.17}$$

by following the lines of the proof of Lemma 4.9. Moreover, in that proof we have already seen that $\sum_{\mu=1}^N\mathcal{F}(\bullet^{[\mu]})(\mathbf{y}^n)=\mathbf{f}(\mathbf{y}^n)$ so that we obtain the intermediate result

$$\begin{aligned}y_\nu^{(i)} =&y_\nu^n+\Delta t c_i f_\nu(\mathbf{y}^n)\\&+\frac{1}{2}\Delta t^2\sum_{\mu,\eta=1}^N d_i\left(\left[\bullet^{[\eta]}\right]^{[\mu]},\mathbf{y}^n,\Delta t\right)\left(\mathcal{F}\left(\left[\bullet^{[\eta]}\right]^{[\mu]}\right)(\mathbf{y}^n)\right)_\nu+\mathcal{O}(\Delta t^3).\end{aligned}$$

Turning to the coefficient of Δt^2, we first point out that, according to (4.1), we have

$$\begin{aligned}d_i\left(\left[\bullet^{[\eta]}\right]^{[\mu]},\mathbf{y}^n,\Delta t\right)&=\sum_{\nu=1}^N\sum_{j=1}^s a_{ij}^{[\nu]}(\mathbf{y}^n,\Delta t)g_j^{[\nu]}\left(\left[\bullet^{[\eta]}\right]^{[\mu]},\mathbf{y}^n,\Delta t\right)\\&=\sum_{j=1}^s a_{ij}^{[\mu]}(\mathbf{y}^n,\Delta t)d_j\left(\bullet^{[\eta]},\mathbf{y}^n,\Delta t\right)=\sum_{j=1}^s a_{ij}c_j+\mathcal{O}(\Delta t^2),\end{aligned}$$

where we used (4.16) and (4.17). Finally, using (2.12) we obtain

$$\begin{aligned}\sum_{\mu,\eta=1}^N\mathcal{F}\left(\left[\bullet^{[\eta]}\right]^{[\mu]}\right)(\mathbf{y}^n)&=\sum_{\mu,\eta=1}^N\sum_{i_1=1}^d\partial_{i_1}\mathbf{f}^{[\mu]}(\mathbf{y}^n)\mathcal{F}_{i_1}(\bullet^{[\eta]})(\mathbf{y})\\&=\sum_{\mu=1}^N\mathbf{Df}^{[\mu]}(\mathbf{y}^n)\sum_{\eta=1}^N\mathbf{f}^{[\eta]}(\mathbf{y}^n)=\mathbf{Df}(\mathbf{y}^n)\mathbf{f}(\mathbf{y}^n).\end{aligned}$$

The claim follows after substituting these equations into (4.15). □

With that lemma we now derive sufficient and necessary conditions for 4th order MPRK schemes.

The simplified conditions for an MPRK method of order 4 are given by (B.6) with $\gamma_\nu^{(i)}=\frac{y_\nu^{(i)}}{\pi_\nu^{(i)}}$ and $\delta_\mu=\frac{y_\mu^{n+1}}{\sigma_\mu}$. Using Lemma B.5 these conditions are equivalent to

$$\frac{y_\mu^{n+1}}{\sigma_\mu}=1+\mathcal{O}(\Delta t^4),\qquad \mu=1,\dots,N,\tag{4.18a}$$

$$\sum_{i=2}^4 b_ic_i\frac{y_\nu^{(i)}}{\pi_\nu^{(i)}}=\frac{1}{2}+\mathcal{O}(\Delta t^3),\qquad \nu=1,\dots,N,\tag{4.18b}$$

$$\frac{y_\nu^{(i)}}{\pi_\nu^{(i)}}=1+\mathcal{O}(\Delta t^2),\qquad \nu=1,\dots,N,\quad i=2,3,4.\tag{4.18c}$$

However, these conditions again depend on the stages. The next theorem gives us equivalent conditions depending only on $\mathbf{y}^n$ and Δt.

Theorem 4.15. Let $\mathbf{A},\mathbf{b},\mathbf{c}$ describe an explicit 4-stage RK scheme of order 4 with $\sum_{j=1}^s a_{ij}=c_i$, and let $\mathbf{f}^{[\nu]}\in\mathcal{C}^5$ for $\nu=1,\dots,N$. Then the corresponding MPRK scheme (3.8) is at least of order $p=s=4$ if and only if for $\mu,\nu=1,\dots,N$ and $i=2,3,4$ we have

$$\sigma_\mu=(\mathrm{NB}_3(\tfrac{1}{\gamma},\mathbf{y}^n))_\mu+\mathcal{O}(\Delta t^4), \tag{4.19a}$$

$$\sum_{i=2}^4 b_ic_i\frac{y_\nu^n+\Delta tc_if_\nu(\mathbf{y}^n)+\frac12\Delta t^2\sum_{k=1}^4 a_{ik}c_k(\mathbf{Df}(\mathbf{y}^n)\mathbf{f}(\mathbf{y}^n))_\nu}{\pi_\nu^{(i)}}=\frac12+\mathcal{O}(\Delta t^3), \tag{4.19b}$$

$$\pi_\nu^{(i)}=y_\nu^n+\Delta tc_if_\nu(\mathbf{y}^n)+\mathcal{O}(\Delta t^2). \tag{4.19c}$$

Proof. We start by assuming that (4.18) is fulfilled and note that this part works along the same lines as in Theorem 4.13. Nevertheless, we present it here for the sake of completeness.

Now, since (4.18) holds, the MPRK scheme is of order 4. Thus, Lemma 4.10 implies (4.19a). Finally, with (4.18c) we are in the position to apply Lemma 4.14, which, together with (4.18b), yield the conditions (4.19c) and (4.19b).

Let's now suppose that (4.19) holds. Using the expansion (4.2a) for the stages, we first observe with (4.19c) that $\frac{y_\nu^{(i)}}{\pi_\nu^{(i)}}=1+\mathcal{O}(\Delta t)$. Applying Lemma 4.9, we see $y_\nu^{(i)}=y_\nu^n+\Delta tc_if_\nu(\mathbf{y}^n)+\mathcal{O}(\Delta t^2)$, and thus, comparing with (4.19c), we derived (4.18c). As a result, we can now apply Lemma 4.14 to obtain

$$y_\nu^{(i)}=y_\nu^n+\Delta tc_if_\nu(\mathbf{y}^n)+\frac12\Delta t^2\sum_{k=1}^4 a_{ik}c_k(\mathbf{Df}(\mathbf{y}^n)\mathbf{f}(\mathbf{y}^n))_\nu+\mathcal{O}(\Delta t^3),\quad \nu=1,\dots,N.$$

As a direct consequence of this and (4.19b), we thus conclude (4.18b). Therefore, it remains to deduce condition (4.18a). First of all, (4.19a) and Theorem 4.11 imply that the MPRK scheme is of order at least 1, i. e. $\mathbf{y}^{n+1}=\mathrm{NB}_1(\frac{1}{\gamma},\mathbf{y}^n)+\mathcal{O}(\Delta t^2)$. Comparing with (4.19a), we see

$$\frac{y_\mu^{n+1}}{\sigma_\mu}=1+\mathcal{O}(\Delta t^2),\quad \mu=1,\dots,N.$$

Moreover, since we have already shown (4.18c), we can now verify that condition (4.7) is fulfilled which means that the MPRK scheme is even second order accurate. Therefore, we find $\mathbf{y}^{n+1}=\mathrm{NB}_2(\frac{1}{\gamma},\mathbf{y}^n)+\mathcal{O}(\Delta t^3)$, so that a comparison with (4.19a) gives us

$$\frac{y_\mu^{n+1}}{\sigma_\mu}=1+\mathcal{O}(\Delta t^3),\quad \mu=1,\dots,N.$$

Finally, using this and (4.18c) once again, we even fulfill the conditions (4.8) proving the 3rd order accuracy of the scheme, that is $\mathbf{y}^{n+1}=\mathrm{NB}_3(\frac{1}{\gamma},\mathbf{y}^n)+\mathcal{O}(\Delta t^4)$. Comparing a last time with (4.19a) gives us (4.18a). □

With this proof, we obtain for the first time necessary and sufficient order conditions for 4th order MPRK methods. A first intuitive, yet rather expensive way of achieving 4th order would be to use lower order MPRK methods for the computation of the PWDs. In particular, we propose the following method based on the classical Runge–Kutta method described by the Butcher tableau

$$\begin{array}{c|cccc}
0 & & & & \\
\frac{1}{2} & \frac{1}{2} & & & \\
\frac{1}{2} & 0 & \frac{1}{2} & & \\
1 & 0 & 0 & 1 & \\
\hline
 & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6}
\end{array}$$

as a proof of concept scheme. We know that the PWD $\boldsymbol{\sigma}$ needs to be a third order approximation to $\mathbf{y}^{n+1}$, for which we use the MPRK43$(0.5, 0.75)$ method derived in [KM18b]. Within this method, there is a second order scheme embedded, which we denote by $\hat{\boldsymbol{\sigma}}$ and use to compute $\boldsymbol{\pi}^{(i)}$ for $i = 2, 3, 4$ using $c_i\Delta t$ as a time step, resulting in $\frac{y_\nu^{(i)}}{\pi_\nu^{(i)}} = 1 + \mathcal{O}(\Delta t^3)$. Now, according to Corollary 4.3 the overall method is of order 4.

The third order scheme returns $\hat{\mathbf{y}}$ and consists of solving 4 linear systems, where $\hat{\boldsymbol{\sigma}}$ requires solving 2 systems. However, as $c_4 = 1$, we can actually use $\boldsymbol{\pi}^{(4)} = \hat{\boldsymbol{\sigma}}$, and since $c_2 = c_3$ we can use $\boldsymbol{\pi}^{(2)} = \boldsymbol{\pi}^{(3)}$. Finally, the MP trick applied to the classical RK method also adds 4 linear systems to our list. Altogether $\boldsymbol{\pi}^{(2)}$ and $\boldsymbol{\pi}^{(3)}$ yield a total of 2 linear systems, $\boldsymbol{\pi}^{(4)} = \hat{\boldsymbol{\sigma}}$ and $\boldsymbol{\sigma} = \hat{\mathbf{y}}$ need the solution of 4 linear systems and the MP approach applied to the classical RK scheme results in another 4 linear systems giving us a total of 10 stages and linear systems to solve. The optimal amount of linear systems of course would be 4 and to reduce the number of linear systems to be solved will be part of my future work. An indication that this is possible is given by MPDeC methods where fourth order is obtained by 7 stages for Gauss–Lobatto nodes. Nevertheless, our first attempt has as many stages as MPDeCEQ(4).

The experimental order of convergence of our first fourth order MPRK method, denoted by MPRKord4, is verified in Figure 4.1, where the linear system

$$\mathbf{y}'(t) = \left(\begin{smallmatrix} -5 & 1 \\ 5 & -1 \end{smallmatrix}\right)\mathbf{y}(t), \quad \mathbf{y}(0) = \left(\begin{smallmatrix} 0.9 \\ 0.1 \end{smallmatrix}\right) \tag{4.20}$$

is solved on $[0, 1.75]$ as suggested in [KM18b]. We plot the error of the numerical solution at the final time $t_{\text{end}} = 1.75$, where the reference solution was computed with the MATLAB ODE solver `ode45` using `RelTol = 1e-13` and `AbsTol = 1e-13`.

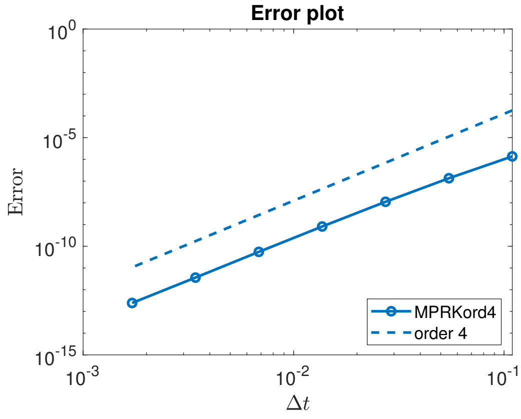


Figure 4.1: Error plot of MPRKord4 applied to (4.20). The error was computed at $t_{\text{end}} = 1.75$ using `ode45` as a reference solution.

Chapter 5

Stability Theory

5.1 Test Equations and Properties

As discussed in Section 2.3 and Section 2.4, we are interested in the stability properties of the positivity-preserving methods reviewed in Chapter 3 when applied to positive linear systems of ordinary differential equations $\mathbf{y}' = \mathbf{\Lambda}\mathbf{y}$. Before we formulate assumptions on the system matrix $\mathbf{\Lambda} = (\lambda_{ij})_{1\leq i,j\leq N}$, we introduce the algebraic multiplicity $\mu_{\mathbf{\Lambda}}(\lambda)$ of the eigenvalue $\lambda \in \sigma(\mathbf{\Lambda})$ as well as the corresponding geometric multiplicity $\gamma_{\mathbf{\Lambda}}(\lambda)$, where

$$\sigma(\mathbf{\Lambda}) \subseteq \overline{\mathbb{C}^-} = \{z \in \mathbb{C} \mid \operatorname{Re}(z) \leq 0\}$$

denotes the spectrum of $\mathbf{\Lambda}$.

In view of Theorem 2.15 for the hyperbolic case, we are particularly interested in problems possessing linear invariants such as conservativity. As mentioned in Section 2.6, the presence of $k > 0$ linear invariants means that there exist linearly independent vectors $\mathbf{n}_1, \dots, \mathbf{n}_k \in \mathbb{R}^N \setminus \{\mathbf{0}\}$ such that $\mathbf{n}_i^T\mathbf{y}(t) = \mathbf{n}_i^T\mathbf{y}^0$ for all $t \geq 0$, or equivalently $\mathbf{n}_i^T\mathbf{\Lambda} = \mathbf{0}$ for $i = 1, \dots, k$. Note that the existence of k linear invariants is given if and only if $k = \dim(\ker(\mathbf{\Lambda}^T)) = \dim(\ker(\mathbf{\Lambda}))$. The presence of k linear invariants means that $\gamma_{\mathbf{\Lambda}}(0) = k$, so that we consider in the following systems of the form

$$\mathbf{y}' = \mathbf{\Lambda}\mathbf{y}, \quad \mathbf{\Lambda} \neq \mathbf{0}, \quad \mathbf{\Lambda} - \operatorname{diag}(\mathbf{\Lambda}) \geq \mathbf{0}, \quad \mu_{\mathbf{\Lambda}}(0) = \gamma_{\mathbf{\Lambda}}(0) = k, \tag{5.1}$$

together with the initial condition

$$\mathbf{y}(0) = \mathbf{y}^0 > \mathbf{0}, \tag{5.2}$$

where $\operatorname{diag}(\mathbf{\Lambda})$ denotes the diagonal of $\mathbf{\Lambda}$. In particular, $\mathbf{\Lambda} - \operatorname{diag}(\mathbf{\Lambda}) \geq \mathbf{0}$ means that $\mathbf{\Lambda}$ is a so-called *Metzler* matrix [Lue79], which is sufficient and necessary to guarantee the positivity of the analytic solution. Moreover, in the presence of linear invariants, the conditions $\mu_{\mathbf{\Lambda}}(0) = \gamma_{\mathbf{\Lambda}}(0)$ and $\sigma(\mathbf{\Lambda}) \subseteq \overline{\mathbb{C}^-}$ are necessary for the stability of steady states of $\mathbf{y}' = \mathbf{\Lambda}\mathbf{y}$, see Theorem 2.11. To give an example, the IVP

$$\mathbf{y}'(t) = \begin{pmatrix} -a & b \\ a & -b \end{pmatrix} \mathbf{y}(t), \quad \mathbf{y}(0) = \mathbf{y}^0 \in \mathbb{R}^2_{>0}, \tag{5.3}$$

with $a, b \geq 0$ and $a + b > 0$ describes all nontrivial positive and conservative linear problems in $N = 2$. To include also non-conservative systems with a linear invariant we may consider

$$\mathbf{y}'(t) = \begin{pmatrix} -a & bc \\ ac & -b \end{pmatrix} \mathbf{y}(t), \tag{5.4}$$

where $c > 0$.

We want to note that if $k = 0$, then the only steady state is $\mathbf{y}^* = \mathbf{0}$. As we discuss in the following remark, this steady state is then asymptotically stable.

Remark 5.1. First, we want to mention that at least one diagonal element of $\mathbf{\Lambda}$ is negative. Otherwise we find $\operatorname{diag}(\mathbf{\Lambda}) \geq \mathbf{0}$, and hence, $\mathbf{\Lambda} \geq \mathbf{0}$. Then, due to $\mu_{\mathbf{\Lambda}}(0) = \gamma_{\mathbf{\Lambda}}(0) = k$ and $\mathbf{\Lambda} \neq \mathbf{0}$ we find that $k < N$, and thus, there exists a nonzero eigenvalue of $\mathbf{\Lambda}$. Therefore, $\mathbf{\Lambda}$ is not similar to a strictly upper triangular matrix. Utilizing a generalization of the Perron–Frobenius Theorem [Var00, Theorem 2.20] yields that $\mathbf{\Lambda}$ possesses a positive eigenvalue contradicting $\sigma(\mathbf{\Lambda}) \subseteq \overline{\mathbb{C}^-}$. This means that $\mathbf{\Lambda}$ is a so-called proper Metzler Matrix, i. e. a Metzler matrix $\mathbf{\Lambda}$ with at least one negative diagonal element. Consequently, [BF04, Theorem 10, Corollary 11] yields

$$\sigma(\mathbf{\Lambda}) \subseteq \mathcal{B} = \left\{ z \in \mathbb{C} \,\middle|\, |z - r| \leq |r|, r = \min_{j=1,\dots N} \lambda_{jj} \right\},$$

where $r < 0$ follows since $\mathbf{\Lambda}$ is a proper Metzler matrix. Thus, we obtain $\operatorname{Re}(\lambda) < 0$ as well as $\arg(\lambda) \in (\frac{\pi}{2}, \frac{3}{2}\pi)$ for all $0 \neq \lambda \in \sigma(\mathbf{\Lambda})$. Finally, if $k = 0$ holds in (5.1), it follows from Theorem 2.11 that $\mathbf{y}^* = \mathbf{0}$ is asymptotically stable.

Now, since we want to generalize A-stability, we may speak of stable methods rather than stating that all steady states become stable fixed points. A precise definition for positivity-preserving methods is given in the following.

Definition 5.2. Let (5.1), (5.2) with $k > 0$ fulfill the requirements for the application of a given one-step method with generating map $\mathbf{g}\colon \mathbb{R}^N_{>0} \to \mathbb{R}^N_{>0}$.

- The one-step method is called *conditionally stable*, if there exists a $c > 0$ such that any $\mathbf{y}^* \in \ker(\mathbf{\Lambda}) \cap \mathbb{R}^N_{>0}$ is a Lyapunov stable fixed point of $\mathbf{g}$ for all $0 < \Delta t < c$.
- If the method is conditionally stable and $c > 0$ can be chosen arbitrarily large, we call the method *unconditionally stable*.
- If all $\mathbf{y}^* \in \ker(\mathbf{\Lambda}) \cap \mathbb{R}^N_{>0}$ are unstable fixed points of $\mathbf{g}$, we call the method *unstable*.

Remark 5.3. In the above definition it is assumed that the one-step method can be applied to the system to (5.1), (5.2). For conservative schemes this requires $\mathbf{1} \in \ker(\mathbf{\Lambda}^T)$. It is also worth mentioning that not being conditionally stable implies instability if the method is linear. However, this does not need to be true for nonlinear one-step methods as we will discuss later.

5.2 Main Theorem for Stability

In this section we provide a theorem for the investigation of stability, defined in Section 2.4, of the numerical methods from Chapter 3 applied to stable positive linear systems (5.1) with $k > 0$.

As a consequence of the presence of linear invariants, 0 is always an eigenvalue of $\mathbf{\Lambda}$ which implies the existence of nontrivial steady state solutions $\mathbf{y}^*$. For every reasonable time integration scheme $\mathbf{y}^{n+1} = \mathbf{g}(\mathbf{y}^n)$, these steady state solutions have to be fixed points. The common way to study the stability of a fixed point $\mathbf{y}^*$ of $\mathbf{g}$ is to compute the eigenvalues of the Jacobian $\mathbf{Dg}(\mathbf{y}^*)$. It is well-known that the fixed point $\mathbf{y}^*$ is asymptotically stable if the spectral radius ρ of the Jacobian satisfies $\rho(\mathbf{Dg}(\mathbf{y}^*)) < 1$, see Theorem 2.15. Unfortunately, the existence of linear invariants leads to non-hyperbolic fixed points $\mathbf{y}^*$ of the numerical scheme, i. e. the Jacobian $\mathbf{Dg}(\mathbf{y}^*)$ has at least one eigenvalue λ with $|\lambda| = 1$.

If the time integration scheme applied to (5.1) results in a linear iteration

$$\mathbf{y}^{n+1} = \boldsymbol{R}(\Delta t, \mathbf{\Lambda})\mathbf{y}^n,$$

as is the case for Runge–Kutta schemes, the stability of the non-hyperbolic fixed point $\mathbf{y}^*$ is again fully determined by the eigenvalues of the Jacobian

$$\mathbf{Dg}(\mathbf{y}^*) = \boldsymbol{R}(\Delta t, \mathbf{\Lambda})$$

as discussed in Remark 2.14.

Unfortunately, the application of higher-order positivity-preserving schemes to the linear system (5.1) results in a nonlinear iteration of the form

$$\mathbf{y}^{n+1} = \boldsymbol{R}(\Delta t, \mathbf{\Lambda}, \mathbf{y}^n)\mathbf{y}^n,$$

see [OH17] for an illustrative example. For such iterations the stability is not fully determined by the eigenvalues of the Jacobian, see for instance Example 2.17. Hence, the stability analysis of these numerical methods requires the investigation of non-hyperbolic fixed points of a nonlinear iteration. This is significantly more demanding compared to the linear case.

One way to study the stability of non-hyperbolic fixed points of nonlinear iterations is the center manifold theory from [MM76, Car81, Ioo79], reviewed in Section 2.5. This theory states that the stability of a non-hyperbolic fixed point can be determined by studying the iteration on a lower-dimensional invariant manifold, the center manifold.

To avoid the application of the center manifold theory to each positivity-preserving scheme separately, we present a theorem which provides sufficient conditions for the stability of all such methods. Thereby, the main assumption of this new theorem published in [IKM22b] is that the fixed points of the nonlinear iteration form a linear subspace of $\mathbb{R}^N$. This is a reasonable requirement due to the fact that the steady states of the underlying differential equation (5.1) also form a linear subspace of dimension $k > 0$, whenever k linear invariants are present. The theorem contains two main statements. First, the existence of k linear invariants implies that $\lambda = 1$ is an eigenvalue of the Jacobian $\mathbf{Dg}(\mathbf{y}^*)$ of multiplicity at least k and the non-hyperbolic fixed point $\mathbf{y}^*$ is stable, if the remaining $N - k$ eigenvalues have absolute value less than one. Second, if the numerical scheme preservers all k linear invariants, then the iterates locally converge to the unique steady state of the initial value problem (5.1), (5.2). Furthermore, it is worth mentioning that the new theorem can directly be used for the stability analysis of time integration schemes in the context of nonlinear systems of differential equations as we will discuss in Remark 5.5.

In addition, we want to emphasize at this point that it is not sufficient to assess the stability of a higher-order positivity-preserving scheme in terms of a linear system of the form

$$\mathbf{y}' = \begin{pmatrix} \lambda & 0 \\ -\lambda & 0 \end{pmatrix} \mathbf{y}, \quad \mathbf{y}(0) = \mathbf{y}^0 > \mathbf{0}, \quad \lambda \in \mathbb{R}^-, \tag{5.5}$$

which can be seen as a adaptation of Dahlquist's equation

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

originally introduced in [Dah63], to linear conservative systems. One example for this fact is given in [IKM22a], where the so-called MPRK22ncs(α) schemes are investigated. These methods differ from original MPRK schemes in the non-conservative stages (ncs). To be precise, the stages are only treated with the Patankar-trick for guaranteeing unconditional positivity while the modification is only applied to the last step. The total method is still conservative and positive, however the linear systems for the stages are easier to solve. Now, these methods are proven to be L_0-stable in the following sense. Applied to the conservative system (5.5) the state variable y_1^n satisfies $y_1^{n+1} = R(\Delta t \lambda) y_1^n$ with

$$R(z) = \frac{(1-\alpha z)^{1-\frac{1}{\alpha}}}{(1-\alpha z)^{1-\frac{1}{\alpha}} - z\big(1 - (\alpha - \frac{1}{2})z\big)},$$

so that $\lim_{z\to-\infty} R(z) = 0$ and $|R(z)| \leq 1$ for all $z \leq 0$ and $\alpha \geq \frac{1}{2}$. In total this means that the first component represents the behavior of the numerical scheme applied to the Dahlquist equation for $\lambda \in \mathbb{R}^-$ and satisfies all conditions for a scheme to be L_0-stable, see [TGA96]. Nevertheless, in [IKM22a] it is proved that MPRK22ncs(α) face severe time step restrictions for $\alpha < 1$ in order to be stable when applied to a general two–dimensional linear positive and conservative system (5.3), which was also used in [IKM21] for studying the linearization of MPRK22 schemes. Hence, to understand the stability behavior of such nonlinear schemes, one should directly investigate the system (5.1).

Main Result for Stability

In this subsection we make use of the center manifold theory to investigate the stability of fixed points $\mathbf{y}^*$ of a numerical scheme $\mathbf{y}^{n+1} = \mathbf{g}(\mathbf{y}^n)$ with $\mathbf{g}\colon D \to D$ and $D \subseteq \mathbb{R}^N$. To that end, we assume that there exists a neighborhood $\mathcal{D} \subseteq D$ of $\mathbf{y}^*$ such that $\mathbf{g}\big|_{\mathcal{D}} \in \mathcal{C}^1$ has first derivatives that are Lipschitz continuous on $\mathcal{D}$, so we can apply Theorem 2.18. Based on this assumption, Theorem 5.4 below yields a sufficient condition for the Lyapunov stability of $\mathbf{y}^*$ based on the eigenvalues of the corresponding Jacobian $\mathbf{Dg}(\mathbf{y}^*)$. If $\mathbf{g}$ in addition conserves all linear invariants of $\mathbf{\Lambda}$ from (5.1), i. e. $\mathbf{n}^T\mathbf{g}(\mathbf{y}) = \mathbf{n}^T\mathbf{y}$ for all $\mathbf{y} \in D$ whenever $\mathbf{n}^T\mathbf{\Lambda} = \mathbf{0}$, then Theorem 5.4 also states that the numerical scheme locally converges towards the unique steady state $\mathbf{y}^*$ of (5.1), (5.2).

For a compact notation we introduce the matrix

$$\mathbf{N} = \begin{pmatrix} \mathbf{n}_1^T \\ \vdots \\ \mathbf{n}_k^T \end{pmatrix} \in \mathbb{R}^{k\times N} \tag{5.6}$$

with $\mathbf{n}_1, \ldots, \mathbf{n}_k$ being a basis of $\ker(\mathbf{\Lambda}^T)$ as well as the set

$$H = \{\mathbf{y} \in \mathbb{R}^N \mid \mathbf{N}\mathbf{y} = \mathbf{N}\mathbf{y}^*\} \tag{5.7}$$

and point out that for $\mathbf{y} \in H \cap D$ we have $\mathbf{g}(\mathbf{y}) \in H \cap D$, if and only if $\mathbf{g}$ conserves all linear invariants.

Theorem 5.4. Let $\mathbf{\Lambda} \in \mathbb{R}^{N\times N}$ be such that $\ker(\mathbf{\Lambda}) = \operatorname{span}(\mathbf{v}_1, \ldots, \mathbf{v}_k)$ represents a k-dimensional subspace of $\mathbb{R}^N$ with $k > 0$. Also, let $\mathbf{y}^* \in \ker(\mathbf{\Lambda})$ be a fixed point of $\mathbf{g}\colon D \to D$ where $D \subseteq \mathbb{R}^N$ contains a neighborhood $\mathcal{D}$ of $\mathbf{y}^*$. Moreover, let any element of $C = \ker(\mathbf{\Lambda}) \cap \mathcal{D}$ be a fixed point of $\mathbf{g}$ and suppose that $\mathbf{g}\big|_{\mathcal{D}} \in \mathcal{C}^1$ as well as that the first derivatives of $\mathbf{g}$ are Lipschitz continuous on $\mathcal{D}$. Then $\mathbf{D}\mathbf{g}(\mathbf{y}^*)\mathbf{v}_i = \mathbf{v}_i$ for $i = 1, \ldots, k$ and the following statements hold.

a) If the remaining $N - k$ eigenvalues of $\mathbf{D}\mathbf{g}(\mathbf{y}^*)$ have absolute values smaller than 1, then $\mathbf{y}^*$ is stable.

b) Let H be defined by (5.7) and $\mathbf{g}$ conserve all linear invariants, which means that $\mathbf{g}(\mathbf{y}) \in H \cap D$ for all $\mathbf{y} \in H \cap D$. If additionally the assumption of a) is satisfied, then there exists a $\delta > 0$ such that $\mathbf{y}^0 \in H \cap D$ and $\|\mathbf{y}^0 - \mathbf{y}^*\| < \delta$ imply $\mathbf{y}^n \to \mathbf{y}^*$ as $n \to \infty$.

Before we prove the above theorem we want to emphasize in the next remark that its application is not restricted to linear systems of differential equations (5.1).

Remark 5.5. Let us consider a general system of autonomous ordinary differential equations $\mathbf{y}' = \mathbf{f}(\mathbf{y}) \in \mathbb{R}^N$ with $k > 0$ linear invariants determined by $\mathbf{n}_1, \ldots, \mathbf{n}_k$ and a k–dimensional subspace $\mathcal{V} = \operatorname{span}(\mathbf{v}_1, \ldots, \mathbf{v}_k) \subseteq \{\mathbf{y} \in \mathbb{R}^N \mid \mathbf{f}(\mathbf{y}) = \mathbf{0}\}$. In the following, we construct a matrix $\mathbf{\Lambda}$ such that $\ker(\mathbf{\Lambda}) = \mathcal{V}$ as well as $\ker(\mathbf{\Lambda}^T) = \operatorname{span}(\mathbf{n}_1, \ldots, \mathbf{n}_k)$, and thus are in the position to apply Theorem 5.4.

As $\mathbf{\Lambda}$ is uniquely determined by its operation on a basis of $\mathbb{R}^N$ we first set $\mathbf{\Lambda}\mathbf{v}_i = \mathbf{0}$ for $i = 1, \ldots, k$ so that $\ker(\mathbf{\Lambda}) = \mathcal{V}$ is satisfied. To find an expression for $\operatorname{Im}(\mathbf{\Lambda})$ we make use of $\operatorname{Im}(\mathbf{\Lambda}) = (\ker(\mathbf{\Lambda}^T))^\perp = (\operatorname{span}(\mathbf{n}_1, \ldots, \mathbf{n}_k))^\perp$. Using the matrix notation (5.6), this means that $\mathbf{s} \in \operatorname{Im}(\mathbf{\Lambda})$ if and only if $\mathbf{N}\mathbf{s} = \mathbf{0}$, or equivalently $\mathbf{s} \in \ker(\mathbf{N})$. Since $\dim(\ker(\mathbf{N})) = N - k$, there exist linearly independent vectors $\mathbf{s}_1, \ldots, \mathbf{s}_{N-k}$ with $\operatorname{Im}(\mathbf{\Lambda}) = \operatorname{span}(\mathbf{s}_1, \ldots, \mathbf{s}_{N-k})$, and hence, there exist linearly independent vectors $\mathbf{w}_1, \ldots, \mathbf{w}_{N-k}$ such that $\mathbf{\Lambda}\mathbf{w}_i = \mathbf{s}_i$ for $i = 1, \ldots, N - k$. As a consequence, setting $\mathcal{W} = \operatorname{span}(\mathbf{w}_1, \ldots, \mathbf{w}_{N-k}) \subseteq \mathbb{R}^N$ yields $\mathcal{V} \oplus \mathcal{W} = \mathbb{R}^N$. Altogether, $\mathcal{V}$ and $\mathcal{W}$ uniquely determine the matrix $\mathbf{\Lambda}$ satisfying $\ker(\mathbf{\Lambda}) = \mathcal{V}$ and $\ker(\mathbf{\Lambda}^T) = \operatorname{span}(\mathbf{n}_1, \ldots, \mathbf{n}_k)$. Hence, Theorem 5.4 is not restricted to linear systems.

Proof of Theorem 5.4. First, we show $\mathbf{D}\mathbf{g}(\mathbf{y}^*)\mathbf{v}_i = \mathbf{v}_i$ for $i = 1, \ldots, k$. Since $\mathbf{g}$ is differentiable in $\mathbf{y}^* \in \mathcal{D}$ the directional derivatives $\partial_{\mathbf{v}}\mathbf{g}(\mathbf{y}^*) = \mathbf{D}\mathbf{g}(\mathbf{y}^*)\mathbf{v}$ exist for all directions $\mathbf{v} \in \mathbb{R}^N$ and for $i = 1, \ldots, k$ we find

$$\mathbf{D}\mathbf{g}(\mathbf{y}^*)\mathbf{v}_i = \partial_{\mathbf{v}_i}\mathbf{g}(\mathbf{y}^*) = \lim_{h\to 0} \frac{1}{h}\big(\mathbf{g}(\mathbf{y}^* + h\mathbf{v}_i) - \mathbf{g}(\mathbf{y}^*)\big).$$

For $|h|$ small enough, we see that $\mathbf{y}^* + h\mathbf{v}_i \in C$ because of the following. First of all $\mathbf{y}^* + h\mathbf{v}_i \in \ker(\mathbf{\Lambda})$ holds for all $h \in \mathbb{R}$, so that we have to show that $\mathbf{y}^* + h\mathbf{v}_i \in \mathcal{D}$ for $|h|$ small enough. Since $\mathbf{y}^* \in \mathcal{D}$, there exists a $\gamma > 0$ such that the open ball

$B_\gamma(\mathbf{y}^*)$ with center $\mathbf{y}^*$ and radius γ satisfies $B_\gamma(\mathbf{y}^*) \subseteq \mathcal{D}$. Choosing $|h| < \frac{\gamma}{\|\mathbf{v}_i\|}$ we find

$$\|\mathbf{y}^* + h\mathbf{v}_i - \mathbf{y}^*\| \leq |h|\|\mathbf{v}_i\| < \gamma,$$

such that $\mathbf{y}^* + h\mathbf{v}_i \in \ker(\boldsymbol{\Lambda}) \cap B_\gamma(\mathbf{y}^*) \subseteq \ker(\boldsymbol{\Lambda}) \cap \mathcal{D} = C$ is a fixed point of $\mathbf{g}$. Hence,

$$\mathbf{Dg}(\mathbf{y}^*)\mathbf{v}_i = \lim_{h\to 0} \frac{1}{h}\big(\mathbf{y}^* + h\mathbf{v}_i - \mathbf{y}^*\big) = \mathbf{v}_i,$$

which shows that $\mathbf{v}_i$ is an eigenvector of $\mathbf{Dg}(\mathbf{y}^*)$ with associated eigenvalue 1. Thus, the spectrum of $\mathbf{Dg}(\mathbf{y}^*)$ contains the eigenvalue 1 with a multiplicity of at least k.

a) We now assume that the remaining $N-k$ eigenvalues of $\mathbf{Dg}(\mathbf{y}^*)$ have absolute values smaller than 1. Next we introduce the matrix of generalized eigenvectors $\mathbf{S}$ where the first k columns are given by the basis vectors $\mathbf{v}_1, \dots, \mathbf{v}_k$ of $\ker(\boldsymbol{\Lambda})$. Thus, we obtain

$$\mathbf{S}^{-1}\mathbf{Dg}(\mathbf{y}^*)\mathbf{S} = \mathbf{J} \tag{5.8}$$

with the Jordan normal form $\mathbf{J}$ of $\mathbf{Dg}(\mathbf{y}^*)$. We want to point out that the upper left $k \times k$ block of $\mathbf{J}$ is the identity matrix, since the k basis vectors $\mathbf{v}_1, \dots, \mathbf{v}_k$ of $\ker(\boldsymbol{\Lambda})$ are eigenvectors with associated eigenvalue 1.

We want to use the Theorem 2.18 a) in combination with Theorem 2.19 to conclude that $\mathbf{y}^*$ is a stable fixed point. The theorems require a map $\mathbf{G}$ of form (2.18), which shall be obtained from $\mathbf{g}$ by means of an affine linear transformation. We consider the affine transformation

$$\mathbf{T}\colon \mathbb{R}^N \to \mathbb{R}^N, \quad \mathbf{y} \mapsto \mathbf{w} = \mathbf{T}(\mathbf{y}) = \mathbf{S}^{-1}(\mathbf{y} - \mathbf{y}^*),$$

where the inverse transformation $\mathbf{T}^{-1}$ is given by $\mathbf{T}^{-1}(\mathbf{w}) = \mathbf{S}\mathbf{w} + \mathbf{y}^*$. By construction, $\ker(\boldsymbol{\Lambda})$ is mapped onto the subspace spanned by the first k unit vectors $\mathbf{e}_1, \dots, \mathbf{e}_k$ of $\mathbb{R}^N$, as for $\mathbf{y}^* = \sum_{i=1}^k t_i\mathbf{v}_i \in \ker(\boldsymbol{\Lambda})$ we find

$$\begin{aligned}\mathbf{T}\left(\sum_{i=1}^k r_i\mathbf{v}_i\right) &= \mathbf{S}^{-1}\left(\sum_{i=1}^k r_i\mathbf{v}_i - \mathbf{y}^*\right) = \mathbf{S}^{-1}\left(\sum_{i=1}^k (r_i - t_i)\mathbf{v}_i\right)\\ &= \sum_{i=1}^k (r_i - t_i)\mathbf{S}^{-1}\mathbf{v}_i = \sum_{i=1}^k (r_i - t_i)\mathbf{e}_i\end{aligned}$$

for arbitrary choices of $r_1, \dots, r_k \in \mathbb{R}$. In particular, $\mathbf{y}^*$ is mapped to the origin.

In order to use Theorem 2.18, we have to define an appropriate $\mathcal{C}^1$-map $\mathbf{G}\colon \mathcal{M} \to \mathbb{R}^N$. Therefore we define $\mathcal{M} = \mathbf{T}(\mathcal{D})$ which is a neighborhood of the origin since $\mathbf{T}$ is an invertible affine linear map. In particular, we use

$$\mathbf{G}\colon \mathbf{T}(\mathcal{D}) \to \mathbb{R}^N, \quad \mathbf{G}(\mathbf{w}) = \mathbf{T}(\mathbf{g}(\mathbf{T}^{-1}(\mathbf{w}))) \tag{5.9}$$

and observe that the origin is a fixed point of $\mathbf{G}$. To represent $\mathbf{G}$ in the form (2.18), we use $\mathbf{g}(\mathbf{y}^*) = \mathbf{y}^*$ and write $\mathbf{g}$ as

$$\begin{aligned}\mathbf{g}(\mathbf{y}) &= \mathbf{g}(\mathbf{y}^*) + \mathbf{Dg}(\mathbf{y}^*)(\mathbf{y} - \mathbf{y}^*) + \mathbf{R}(\mathbf{y})\\ &= \mathbf{y}^* + \mathbf{Dg}(\mathbf{y}^*)(\mathbf{y} - \mathbf{y}^*) + \mathbf{R}(\mathbf{y}),\end{aligned} \tag{5.10}$$

where the remainder $\mathbf{R}(\mathbf{y})$ can be written as

$$\mathbf{R}(\mathbf{y}) = \mathbf{g}(\mathbf{y}) - \mathbf{y}^* - \mathbf{Dg}(\mathbf{y}^*)(\mathbf{y} - \mathbf{y}^*).$$

In particular, we have

$$\mathbf{R}(\mathbf{y}^*) = \mathbf{0}, \quad \mathbf{DR}(\mathbf{y}^*) = \mathbf{0}. \tag{5.11}$$

By inserting (5.10) in (5.9) we obtain

$$\begin{aligned}\mathbf{G}(\mathbf{w}) &= \mathbf{S}^{-1}\big(\mathbf{Dg}(\mathbf{y}^*)(\mathbf{T}^{-1}(\mathbf{w}) - \mathbf{y}^*) + \mathbf{R}(\mathbf{T}^{-1}(\mathbf{w}))\big)\\ &= \mathbf{S}^{-1}\mathbf{Dg}(\mathbf{y}^*)\mathbf{S}\mathbf{w} + \mathbf{S}^{-1}\mathbf{R}(\mathbf{T}^{-1}(\mathbf{w}))\end{aligned}$$

and using (5.8) yields

$$\mathbf{G}(\mathbf{w}) = \mathbf{J}\mathbf{w} + \mathbf{S}^{-1}\mathbf{R}(\mathbf{T}^{-1}(\mathbf{w})) = \begin{pmatrix}\mathbf{I} & \\ & \boldsymbol{R}\end{pmatrix}\mathbf{w} + \mathbf{S}^{-1}\mathbf{R}(\mathbf{T}^{-1}(\mathbf{w})), \tag{5.12}$$

where $\mathbf{I} \in \mathbb{R}^{k\times k}$ and $\boldsymbol{R} \in \mathbb{R}^{(N-k)\times(N-k)}$ and $\rho(\boldsymbol{R}) < 1$ as $N-k$ eigenvalues of $\mathbf{Dg}(\mathbf{y}^*)$ have absolute values smaller than 1. Setting $\mathbf{w} = (\mathbf{w}_1, \mathbf{w}_2)^T$ with $\mathbf{w}_1 \in \mathbb{R}^k$, $\mathbf{w}_2 \in \mathbb{R}^{N-k}$ and $(\mathbf{w}_1, \mathbf{w}_2) \in \mathbf{T}(\mathcal{D})$, (5.12) can be rewritten as

$$\mathbf{G}(\mathbf{w}_1, \mathbf{w}_2) = \begin{pmatrix}\mathbf{U}\mathbf{w}_1 + \mathbf{u}(\mathbf{w}_1, \mathbf{w}_2)\\ \mathbf{V}\mathbf{w}_2 + \mathbf{v}(\mathbf{w}_1, \mathbf{w}_2)\end{pmatrix} \tag{5.13}$$

with

$$\begin{aligned}\mathbf{U} &= \mathbf{I}, \quad \mathbf{u}(\mathbf{w}_1, \mathbf{w}_2) = \big(\mathbf{S}^{-1}\mathbf{R}(\mathbf{T}^{-1}(\mathbf{w}_1, \mathbf{w}_2))\big)_{1:k},\\ \mathbf{V} &= \boldsymbol{R}, \quad \mathbf{v}(\mathbf{w}_1, \mathbf{w}_2) = \big(\mathbf{S}^{-1}\mathbf{R}(\mathbf{T}^{-1}(\mathbf{w}_1, \mathbf{w}_2))\big)_{k+1:N,}\end{aligned} \tag{5.14}$$

where we defined $\mathbf{v}_{l:m} = (v_l, \dots, v_m)^T$ for a vector $\mathbf{v}$ and $l \leq m$. Each eigenvalue of $\mathbf{U}$ has absolute value 1 and those of $\mathbf{V}$ have absolute values smaller than 1. Furthermore, utilizing $\mathbf{T}^{-1}(\mathbf{0}, \mathbf{0}) = \mathbf{y}^*$ we conclude from (5.11) that $\mathbf{u}(\mathbf{0}, \mathbf{0}) = \mathbf{v}(\mathbf{0}, \mathbf{0}) = \mathbf{0}$, since $\mathbf{R}(\mathbf{y}^*) = \mathbf{0}$. Moreover, we have $\mathbf{Du}(\mathbf{0}, \mathbf{0}) = \mathbf{Dv}(\mathbf{0}, \mathbf{0}) = \mathbf{0}$, since $\mathbf{DR}(\mathbf{y}^*) = \mathbf{0}$. Altogether this demonstrates that (5.12) is of form (2.18), which is necessary for applying the center manifold theory.

Now, the center manifold theorem 2.18 a) states that for some $\epsilon > 0$ there exists a $\mathcal{C}^1$ function $\mathbf{h}\colon \mathbb{R}^k \to \mathbb{R}^{N-k}$ with $\mathbf{h}(\mathbf{0}) = \mathbf{0}$ and $\mathbf{Dh}(\mathbf{0}) = \mathbf{0}$, such that $(\mathbf{w}_1^1, \mathbf{w}_2^1)^T = \mathbf{G}(\mathbf{w}_1^0, \mathbf{h}(\mathbf{w}_1^0))$ implies $\mathbf{w}_2^1 = \mathbf{h}(\mathbf{w}_1^1)$ for $\|\mathbf{w}_1^0\|, \|\mathbf{w}_1^1\| < \epsilon$.

In the following we make use of the fact that the center manifold is given by

$$\{(\mathbf{w}_1, \mathbf{w}_2) \in \mathbb{R}^N \mid \mathbf{w}_2 = \mathbf{0},\ \|\mathbf{w}_1\| < \epsilon\}, \tag{5.15}$$

i. e. $\mathbf{h}(\mathbf{w}_1) = \mathbf{0}$, for a sufficiently small $\epsilon > 0$, which can be shown with Theorem 2.20. The function $\mathbf{\Phi}\colon \mathbb{R}^k \to \mathbb{R}^{N-k}$, $\mathbf{\Phi}(\mathbf{w}_1) = \mathbf{0}$ satisfies $\mathbf{\Phi}(\mathbf{0}) = \mathbf{0}$ and $\mathbf{D\Phi}(\mathbf{0}) = \mathbf{0}$. In order to compute $\mathbf{h}$ we first prove that all points $(\mathbf{w}_1, \mathbf{0}) \in \mathbf{T}(\mathcal{D})$ are fixed points of $\mathbf{G}$. Note, that points $(\mathbf{w}_1, \mathbf{0}) \in \mathbf{T}(\mathcal{D})$ even satisfy

$$\begin{aligned}\mathbf{T}^{-1}(\mathbf{w}_1, \mathbf{0}) = \mathbf{T}^{-1}\left(\sum_{i=1}^{k} (\mathbf{w}_1)_i \mathbf{e}_i\right) &= \sum_{i=1}^{k} (\mathbf{w}_1)_i \mathbf{S}\mathbf{e}_i + \mathbf{y}^* \\ &= \sum_{i=1}^{k} (\mathbf{w}_1)_i \mathbf{v}_i + \mathbf{y}^* \in \mathcal{D} \cap \ker(\boldsymbol{\Lambda}) = C.\end{aligned}$$

Hence, we find

$$\mathbf{G}(\mathbf{w}_1, \mathbf{0}) = \mathbf{T}\left(\mathbf{g}\left(\mathbf{T}^{-1}\left(\mathbf{w}_1, \mathbf{0}\right)\right)\right) = \mathbf{T}\left(\mathbf{T}^{-1}\left(\mathbf{w}_1, \mathbf{0}\right)\right) = (\mathbf{w}_1, \mathbf{0})^T. \quad (5.16)$$

Thus, it follows that

$$\begin{aligned}&\boldsymbol{\Phi}(\mathbf{U}\mathbf{w}_1 + \mathbf{u}(\mathbf{w}_1, \boldsymbol{\Phi}(\mathbf{w}_1)) - (\mathbf{V}\boldsymbol{\Phi}(\mathbf{w}_1) + \mathbf{v}(\mathbf{w}_1, \boldsymbol{\Phi}(\mathbf{w}_1))) \\ &\qquad \overset{(5.13)}{=} -(\mathbf{G}(\mathbf{w}_1, \mathbf{0}))_{k+1:N} = \mathbf{0}.\end{aligned}$$

By Theorem 2.20, $\boldsymbol{\Phi}$ is an approximation of $\mathbf{h}$ for any order $q > 1$. Thus,

$$\mathbf{h}(\mathbf{w}_1) = \boldsymbol{\Phi}(\mathbf{w}_1) = \mathbf{0} \text{ for } \|\mathbf{w}_1\| < \epsilon.$$

To investigate the stability of $\mathbf{y}^*$, we can now consider the map

$$\mathcal{G}(\mathbf{w}_1) = \mathbf{U}\mathbf{w}_1 + \mathbf{u}(\mathbf{w}_1, \mathbf{h}(\mathbf{w}_1)) = \mathbf{U}\mathbf{w}_1 + \mathbf{u}(\mathbf{w}_1, \mathbf{0})$$

for $\|\mathbf{w}_1\| < \epsilon$, where $\mathbf{U}$ and $\mathbf{u}$ are given in (5.14). According to Theorem 2.19, the fixed point $\mathbf{0} \in \mathbb{R}^N$ of $\mathbf{G}$ is stable, if the fixed point $\mathbf{0} \in \mathbb{R}^k$ is a stable fixed point of $\mathcal{G}$. From (5.16) we see

$$\mathcal{G}(\mathbf{w}_1) = (\mathbf{G}(\mathbf{w}_1, \mathbf{0}))_{1:k} = \mathbf{w}_1,$$

which implies $\mathbf{w}_1^n = \mathcal{G}(\mathbf{w}_1^{n-1}) = \mathbf{w}_1^0$ for all $n \in \mathbb{N}$ and every $\mathbf{w}_1^0$ with $\|\mathbf{w}_1^0\| < \epsilon$. Consequently, for every $\widetilde{\epsilon} > 0$ we define $\widetilde{\delta} = \min\{\widetilde{\epsilon}, \epsilon\}$ to obtain that $\|\mathbf{w}_1^0\| < \widetilde{\delta}$ implies $\|\mathbf{w}_1^n\| = \|\mathbf{w}_1^0\| < \widetilde{\delta} \leq \widetilde{\epsilon}$. Thus, $\mathbf{0} \in \mathbb{R}^k$ is a stable fixed point of $\mathcal{G}$ in the sense of Definition 2.12 a). Furthermore, by Theorem 2.19 the fixed point $\mathbf{0} \in \mathbb{R}^N$ of $\mathbf{G}$ is stable as well.

As a last step, we show that the above conclusions imply 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 $\epsilon > 0$ be arbitrary, we define $\epsilon_w = \epsilon/\|\mathbf{S}\|$ and $\delta = \delta_w/\|\mathbf{S}^{-1}\|$. Hence, 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 scheme $\mathbf{y}^{n+1} = \mathbf{g}(\mathbf{y}^n)$.

b) Recall from (5.7) that $H = \{\mathbf{y} \in \mathbb{R}^N \mid \mathbf{N}\mathbf{y} = \mathbf{N}\mathbf{y}^*\}$ and let $\mathbf{y}^0 \in H \cap D$, where $\mathbf{N}$ is given by (5.6). Note, that $\dim(H) = N - k$ as $\mathbf{N}$ has rank k, and $\mathbf{y}^n \in H$ for all $n \in \mathbb{N}_0$ since $\mathbf{g}(\mathbf{y}) \in H$ for all $\mathbf{y} \in H \cap D$. Moreover, for all $\mathbf{y} \in H$ we find

$$(\mathbf{y} - \mathbf{y}^*) \perp \ker(\mathbf{\Lambda}^T) = \operatorname{span}(\mathbf{n}_1, \dots, \mathbf{n}_k)$$

since $\mathbf{N}(\mathbf{y} - \mathbf{y}^*) = \mathbf{N}\mathbf{y}^* - \mathbf{N}\mathbf{y}^* = \mathbf{0}$. Hence $\mathbf{y}^n - \mathbf{y}^* \in (\ker(\mathbf{\Lambda}^T))^\perp = \operatorname{Im}(\mathbf{\Lambda})$ for all $n \in \mathbb{N}_0$. We now want to show that the last $N - k$ column vectors of the invertible matrix $\mathbf{S} = (\mathbf{v}_1 \dots \mathbf{v}_k \mathbf{v}_{k+1} \dots \mathbf{v}_N)$ of generalized eigenvectors associated with $\mathbf{Dg}(\mathbf{y}^*)$, see (5.8), form a basis of $\operatorname{Im}(\mathbf{\Lambda})$. Since $\mathbf{g}$ conserves all linear invariants we observe

$$\begin{aligned} \mathbf{n}_i^T \mathbf{Dg}(\mathbf{y}^*)\mathbf{v} &= \lim_{h \to 0} \frac{1}{h}\Big(\mathbf{n}_i^T \mathbf{g}(\mathbf{y}^* + h\mathbf{v}) - \mathbf{n}_i^T \mathbf{g}(\mathbf{y}^*)\Big) \\ &= \lim_{h \to 0} \frac{1}{h}\Big(\mathbf{n}_i^T (\mathbf{y}^* + h\mathbf{v}) - \mathbf{n}_i^T \mathbf{y}^*\Big) = \mathbf{n}_i^T \mathbf{v} \end{aligned}$$

for all $\mathbf{v} \in \mathbb{R}^N$, and in particular we find

$$\mathbf{n}_i^T (\mathbf{Dg}(\mathbf{y}^*) - \lambda \mathbf{I})\mathbf{v} = \mathbf{n}_i^T \mathbf{Dg}(\mathbf{y}^*)\mathbf{v} - \lambda \mathbf{n}_i^T \mathbf{v} = (1 - \lambda)\mathbf{n}_i^T \mathbf{v}. \tag{5.17}$$

If $\mathbf{v}$ is a generalized eigenvector of $\mathbf{Dg}(\mathbf{y}^*)$ corresponding to an eigenvalue $\lambda \neq 1$, so that

$$(\mathbf{Dg}(\mathbf{y}^*) - \lambda \mathbf{I})^m \mathbf{v} = \mathbf{0}$$

is satisfied for some $m \in \mathbb{N}$, it follows from (5.17) that

$$0 = \mathbf{n}_i^T (\mathbf{Dg}(\mathbf{y}^*) - \lambda \mathbf{I})^m \mathbf{v} = (1 - \lambda)\mathbf{n}_i^T (\mathbf{Dg}(\mathbf{y}^*) - \lambda \mathbf{I})^{m-1} \mathbf{v} = (1 - \lambda)^m \mathbf{n}_i^T \mathbf{v},$$

which implies $\mathbf{n}_i^T \mathbf{v} = 0$ as $\lambda \neq 1$. Hence, all generalized eigenvectors $\mathbf{v}$ corresponding to an eigenvalue $\lambda \neq 1$ are elements of $(\ker(\mathbf{\Lambda}^T))^\perp = \operatorname{Im}(\mathbf{\Lambda})$. Now note that $\mathbf{v}_{k+1}, \dots, \mathbf{v}_N$ are $N - k$ generalized eigenvectors corresponding to eigenvalues of absolute value smaller than 1. Finally, since

$$\dim(\operatorname{Im}(\mathbf{\Lambda})) = N - \dim(\ker(\mathbf{\Lambda})) = N - k,$$

the vectors $\mathbf{v}_{k+1}, \dots, \mathbf{v}_N$ form a basis of $\operatorname{Im}(\mathbf{\Lambda})$. Since

$$\mathbf{y}^n - \mathbf{y}^* \in \operatorname{Im}(\mathbf{\Lambda}) = \operatorname{span}(\mathbf{v}_{k+1}, \dots, \mathbf{v}_N),$$

there exist coefficients $\gamma_i^n \in \mathbb{R}$ such that for all $n \in \mathbb{N}_0$ we can write

$$\mathbf{y}^n - \mathbf{y}^* = \sum_{i=k+1}^{N} \gamma_i^n \mathbf{v}_i. \tag{5.18}$$

In order to prove the local convergence of the iterates $\mathbf{y}^n$ to $\mathbf{y}^*$ we investigate the local convergence of $\mathbf{w}^n$ to the origin. According to Theorem 2.18 b) the distance of the iterates $\mathbf{w}^n \in \mathbb{R}^N$ from a) to the center manifold given in (5.15) tends to zero for $n \to \infty$, if the iterates stay within a certain neighborhood of the origin. More precisely, this means that the sequence $(\mathbf{w}^n)_{n \in \mathbb{N}_0}$ approaches

$$\{(\mathbf{w}_1, \mathbf{w}_2) \in \mathbb{R}^N \mid \|\mathbf{w}_1\| < \epsilon, \mathbf{w}_2 = \mathbf{0}\} = \operatorname{span}(\mathbf{e}_1, \dots, \mathbf{e}_k) \cap B_\epsilon(\mathbf{0})$$

for $n \to \infty$, if $\|\mathbf{w}^n\| < \epsilon$ for $i = 1, \dots, N$ and all $n \in \mathbb{N}_0$, where $\epsilon > 0$ is sufficiently small. Now, since the origin is a stable fixed point of $\mathbf{G}$, as shown in a), there exists $\widetilde{\delta} > 0$ such that $\|\mathbf{w}^0\| < \widetilde{\delta}$ implies $\|\mathbf{w}^n\| < \epsilon$ for all $n \in \mathbb{N}_0$. Assuming $\|\mathbf{w}^0\| < \widetilde{\delta}$, we can conclude

$$\lim_{n\to\infty} \mathbf{w}^n \in \operatorname{span}(\mathbf{e}_1, \dots, \mathbf{e}_k). \tag{5.19}$$

Furthermore, from (5.18) it follows

$$\begin{aligned}\mathbf{w}^n = \mathbf{T}(\mathbf{y}^n) = \mathbf{S}^{-1}(\mathbf{y}^n - \mathbf{y}^*) &= \mathbf{S}^{-1}\Bigg(\sum_{i=k+1}^{N} \gamma_i^n \mathbf{v}_i\Bigg)\\ &= \sum_{i=k+1}^{N} \gamma_i^n \mathbf{S}^{-1}\mathbf{v}_i = \sum_{i=k+1}^{N} \gamma_i^n \mathbf{e}_i.\end{aligned}$$

In particular this means $\mathbf{w}^n \in \operatorname{span}(\mathbf{e}_{k+1}, \dots, \mathbf{e}_N)$, and hence, in combination with (5.19) one obtains

$$\lim_{n\to\infty} \mathbf{w}^n \in \operatorname{span}(\mathbf{e}_1, \dots, \mathbf{e}_k) \cap \operatorname{span}(\mathbf{e}_{k+1}, \dots, \mathbf{e}_N) = \{\mathbf{0}\},$$

i. e. $\lim_{n\to\infty} \mathbf{w}^n = \mathbf{0}$. Due to the transformation $\mathbf{T}$ this is equivalent to $\lim_{n\to\infty} \mathbf{y}^n = \mathbf{y}^*$ for $\mathbf{y}^0 \in H \cap D$ satisfying $\|\mathbf{y}^0 - \mathbf{y}^*\| < \delta = \widetilde{\delta}/\|\mathbf{S}^{-1}\|$ since then

$$\|\mathbf{w}^0\| = \|\mathbf{T}(\mathbf{y}^0)\| = \|\mathbf{S}^{-1}(\mathbf{y}^0 - \mathbf{y}^*)\| \leq \|\mathbf{S}^{-1}\|\|\mathbf{y}^0 - \mathbf{y}^*\| < \widetilde{\delta}$$

follows.

□

Remark 5.6. The novel theorem presented here is a generalization of [IKM22a, Theorem 2.9] and improves its statements considerably. First of all, [IKM22a, Theorem 2.9] is restricted to systems of size 2 × 2, whereas here we consider the general $N \times N$ case. Second, [IKM22a, Theorem 2.9] is restricted to conservative numerical schemes, whereas the novel theorem can be applied to general iteration maps $\mathbf{g}: D \to D$. Third, [IKM22a, Theorem 2.9] does not make clear that the stability of the non-hyperbolic fixed point requires less assumptions than the local convergence towards it. In the theorem presented above, on the other hand, it becomes evident that the preservation of linear invariants is not at all necessary to guarantee the stability of the fixed point. Therefore, the theorem can be applied to study the stability of methods that do not preserve all linear invariants. Moreover, the new theorem is formulated with less restrictive assumptions on the regularity of the map generating the numerical approximations.

Remark 5.7. As a final remark, we note that if $\mathbf{g} \in \mathcal{C}^2$, which is also assumed in [IKM22a], we may choose $\mathcal{D} \subseteq D$ in such a way that $\overline{\mathcal{D}} \subseteq D$. As a result the second derivatives are bounded on the compact set $\overline{\mathcal{D}}$, so that the first derivatives are Lipschitz continuous due to the mean value theorem. Therefore, $\mathbf{g}$ restricted to $\mathcal{D}$ is a $\mathcal{C}^1$-map with Lipschitz continuous derivatives. For more details, see for example [AE08, Remark 8.12 (b)].

Moreover, we want to mention that Theorem 5.4 recently was applied to analyze the stability properties of MPRK22(α) when applied to a nonlinear systems of ordinary differential equations, see [IKM23c].

5.3 A Necessary Condition for Non-Oscillatory Schemes

In this section, we investigate the connection between oscillations [TÖR22] and the stability theory above for $N = 2$. To that end, we first rewrite all 2–dimensional linear systems of ODEs that are positive and conservative, i. e. (5.3), with a change of variables, as the following IVP

$$\begin{cases}\mathbf{y}'(t) = \mathbf{\Lambda}_\theta \mathbf{y}(t),\\ \mathbf{y}(0) = \mathbf{y}^0 > \mathbf{0},\end{cases} \qquad \mathbf{\Lambda}_\theta = \begin{pmatrix} -\theta & 1-\theta \\ \theta & -(1-\theta)\end{pmatrix}, \quad \theta \in (0,1), \tag{5.20}$$

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

We first describe oscillations for 2–dimensional linear ODEs through the solution and the steady state. It is known that the exact solution does not overshoot the steady state, so that we require the same from the numerical approximation.

Definition 5.8. a) A method is *not overshooting* the steady state of (5.20) if $y_2^1 < \theta$ and $y_1^1 > 1-\theta$ for any given initial state $\mathbf{y}^0 = (1-\epsilon, \epsilon)^T$ with $\epsilon < \theta$, while when $\epsilon > \theta$ the method is *not overshooting* the steady state if $y_2^1 > \theta$ and $y_1^1 < 1-\theta$.

b) Otherwise the method is said to be *overshooting* the steady state of (5.20).

The following theorem extends the results from [IKM22a] to statements regarding oscillatory behavior. To apply the corresponding theory, we assume $\mathbf{g}$ to have the same properties as in [IKM22a, Theorem 2.9].

Theorem 5.9. Let any positive steady state of (5.20) be a fixed point of a map $\mathbf{g} \in \mathcal{C}^2(\mathbb{R}^2_{>0})$. In addition, let the iterates generated by $\mathbf{y}^{n+1} = \mathbf{g}(\mathbf{y}^n)$ satisfy $\|\mathbf{y}^{n+1}\|_1 = \|\mathbf{y}^n\|_1$ for all $n \in \mathbb{N}_0$. Finally, let $\mathbf{y}^*$ be the unique positive steady state of (5.20).

Then, the spectrum of the Jacobian $\mathbf{Dg}(\mathbf{y}^*)$ is $\sigma(\mathbf{Dg}(\mathbf{y}^*)) = \{1, R\}$ with $R \in \mathbb{R}$. Furthermore, if $R < 0$, then the method generated by $\mathbf{g}$ is overshooting the steady state of (5.20).

Proof. Throughout this proof, we use $\mathbf{e}_1 = (1,0)^T$, $\mathbf{e}_2 = (0,1)^T$ to denote the standard unit vectors as well as the notation $\bar{\mathbf{y}} = (1,-1)^T$. In the proof of [IKM22a, Theorem 2.9], it is shown that $\mathbf{Dg}(\mathbf{y}^*)\mathbf{y}^* = \mathbf{y}^*$ and $\mathbf{Dg}(\mathbf{y}^*)\bar{\mathbf{y}} = R\bar{\mathbf{y}}$ with $R \in \mathbb{R}$, which means that the matrix of eigenvectors

$$\mathbf{S} = (\mathbf{y}^* \ \ \bar{\mathbf{y}}) \tag{5.21}$$

is invertible since $\bar{\mathbf{y}}$ cannot be a multiple of the positive vector $\mathbf{y}^*$. Along the lines of Theorem 5.4, we construct a map

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

by means of a transformation $\mathbf{T}(\mathbf{y}) = \mathbf{S}^{-1}(\mathbf{y}-\mathbf{y}^*)$. To see that the method defined by $\mathbf{g}$ is overshooting $\mathbf{y}^*$, we show that the transformed method given by the map $\mathbf{G}$ is overshooting the transformed steady state which is $\mathbf{w}^* = \mathbf{0}$. As demonstrated in [IKM22a, Theorem 2.9], $\mathbf{y}^0$ is transformed onto the w_2-axis and due to the conservation of the map $\mathbf{g}$, it is proven that $\mathbf{G}(\mathbf{w}^0) \in \operatorname{span}(\mathbf{w}^0)$ for $\mathbf{w}^0 = (0, w_2^0)^T$. Moreover,

$$\mathbf{G}(\mathbf{w}) = \operatorname{diag}(1, R)\mathbf{w} + \mathbf{S}^{-1}\bar{\mathbf{R}}(\mathbf{T}^{-1}(\mathbf{w})) \tag{5.22}$$

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

$$(\bar{\boldsymbol{R}}(\mathbf{y}))_i = \frac{1}{2}(\mathbf{y} - \mathbf{y}^*)^T \mathbf{H}g_i(\mathbf{y}^* + c_i(\mathbf{y} - \mathbf{y}^*))(\mathbf{y} - \mathbf{y}^*), \quad i = 1, 2 \tag{5.23}$$

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

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

Here, using $\mathbf{S}^{-1} = (\widetilde{s}_{ij})_{i,j=1,2}$ and $w_1^n = 0$ it follows from (5.22) that

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

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

$$\begin{aligned}
(\mathbf{S}^{-1}\bar{\boldsymbol{R}}(\mathbf{T}^{-1}(\mathbf{w}^0)))_2 =& \frac{1}{2}\sum_{i=1}^{2} \widetilde{s}_{2i}(\mathbf{T}^{-1}(\mathbf{w}^0) - \mathbf{y}^*)^T \mathbf{H}g_i(\xi_i^0)(\mathbf{T}^{-1}(\mathbf{w}^0) - \mathbf{y}^*) \\
=& \frac{1}{2}\sum_{i=1}^{2} \widetilde{s}_{2i}(w_2^0 \mathbf{S}\mathbf{e}_2)^T \mathbf{H}g_i(\xi_i^0)(w_2^0 \mathbf{S}\mathbf{e}_2) \\
=& \frac{1}{2}\sum_{i=1}^{2} \widetilde{s}_{2i}(w_2^0 \bar{\mathbf{y}})^T \mathbf{H}g_i(\xi_i^0)(w_2^0 \bar{\mathbf{y}}) \\
=& C(\xi_1^0, \xi_2^0) \cdot (w_2^0)^2,
\end{aligned} \tag{5.24}$$

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

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

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

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

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

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

for $i = 1, 2$. Now, according to (5.24), we have

$$w_2^1 = R w_2^0 + C(\xi_1^0, \xi_2^0) \cdot (w_2^0)^2 = (R + C(\xi_1^0, \xi_2^0) w_2^0) w_2^0 \tag{5.25}$$

as well as

$$C(\xi_1^0, \xi_2^0) w_2^0 \leq |C(\xi_1^0, \xi_2^0)||w_2^0| < |C(\xi_1^0, \xi_2^0)| \frac{|R|}{K} \leq |R|. \tag{5.26}$$

Because of $R<0$, the inequality (5.26) turns into the statement

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

and thus, $\operatorname{sgn}(w_2^1)\neq\operatorname{sgn}(w_2^0)$ due to (5.25). This proves that the method defined by $\mathbf{G}$ is overshooting $\mathbf{w}^*$ and consequently, the method with iterates given by the map $\mathbf{g}$ is overshooting $\mathbf{y}^*$. □

5.4 Lyapunov Stability Analysis

This section is devoted to the investigation of the numerical methods from Chapter 3 by means of the stability Theorem 2.15 and Theorem 5.4. We note that all schemes from Chapter 3 preserve positive steady states with the same arguments as in [HIK$^+$23] or [TÖR22, Proposition 2.3]. Thus, in order to apply Theorem 2.15 and part a) of Theorem 5.4, we need to prove a certain regularity and compute the eigenvalues of the Jacobian of $\mathbf{g}$ evaluated at some steady state $\mathbf{y}^*\in\ker(\mathbf{\Lambda})\cap D^\circ$ according to (5.30). However, to use also part b) of Theorem 5.4, we need to prove that $\mathbf{g}$ additionally conserves all linear invariants.

In the case where the mapping $\mathbf{g}$ satisfying $\mathbf{y}^{n+1}=\mathbf{g}(\mathbf{y}^n)$ is implicitly given we compute $\mathbf{Dg}(\mathbf{y}^*)$ as described in [IKM22a, HIK$^+$23, IÖ23] by introducing functions $\mathbf{\Phi}_i$ and several auxiliary Jacobians. The functions $\mathbf{\Phi}_i$ arise from rearranging the equations for the s stages and the updating step of the numerical method leading to

$$\begin{aligned}\mathbf{0}&=\mathbf{\Phi}_i(\mathbf{y}^n,\mathbf{y}^{(1)}(\mathbf{y}^n),\dots,\mathbf{y}^{(i)}(\mathbf{y}^n)),\quad i=1,\dots,s\\ \mathbf{0}&=\mathbf{\Phi}_{n+1}(\mathbf{y}^n,\mathbf{y}^{(1)}(\mathbf{y}^n),\dots,\mathbf{y}^{(s)}(\mathbf{y}^n),\mathbf{g}(\mathbf{y}^n)).\end{aligned}\tag{5.27}$$

Note that $\mathbf{\Phi}_i(\mathbf{x}_0,\dots,\mathbf{x}_i)$ is a function of $i+1$ vector-valued variables while $\mathbf{\Phi}_{n+1}(\mathbf{x}_0,\dots,\mathbf{x}_s,\mathbf{y})$ depends on $s+2$ variables. We will find that $\mathbf{\Phi}_k,\mathbf{\Phi}_{n+1}$ are in $\mathcal{C}^1$ for all schemes from Chapter 3, so that we may define

$$\mathbf{D}_n\mathbf{\Phi}_i=\frac{\partial}{\partial\mathbf{x}_0}\mathbf{\Phi}_i,\quad \mathbf{D}_l\mathbf{\Phi}_i=\frac{\partial}{\partial\mathbf{x}_l}\mathbf{\Phi}_i,\tag{5.28}$$

for $i,l=1,\dots,s$ with $l\leq i$, and

$$\mathbf{D}_n\mathbf{\Phi}_{n+1}=\frac{\partial}{\partial\mathbf{x}_0}\mathbf{\Phi}_{n+1},\quad \mathbf{D}_l\mathbf{\Phi}_{n+1}=\frac{\partial}{\partial\mathbf{x}_l}\mathbf{\Phi}_{n+1},\quad \mathbf{D}_{n+1}\mathbf{\Phi}_{n+1}=\frac{\partial}{\partial\mathbf{y}}\mathbf{\Phi}_{n+1}\tag{5.29}$$

for $l=1,\dots,s$. Besides for GeCo and gBBKS, we will even be able to show that $\mathbf{\Phi}_i,\mathbf{\Phi}_{n+1}$ are in $\mathcal{C}^2$, and by means of the implicit function theorem, $\mathbf{g}\in\mathcal{C}^1$ has locally Lipschitz first derivatives. For GeCo and gBBKS more effort is needed to justify the application of Theorem 5.4.

Moreover, we introduce operators $\mathbf{D}_k^*$ indicating the evaluation of the corresponding auxiliary Jacobian at $\mathbf{y}^*,\mathbf{y}^{(1)}(\mathbf{y}^*)$ et cetera, e. g.

$$\mathbf{D}_n^*\mathbf{\Phi}_2=\mathbf{D}_n\mathbf{\Phi}_2(\mathbf{y}^*,\mathbf{y}^{(1)}(\mathbf{y}^*),\mathbf{y}^{(2)}(\mathbf{y}^*)).$$

As we interpret $\mathbf{y}^{(i)}=\mathbf{y}^{(i)}(\mathbf{y}^n)$ we also introduce the Jacobian

$$\mathbf{D}^*\mathbf{y}^{(i)}=\mathbf{D}\mathbf{y}^{(i)}(\mathbf{y}^*).$$

With that we can derive a formula for computing $\mathbf{Dg}(\mathbf{y}^*)$, where $\mathbf{y}^{n+1}=\mathbf{g}(\mathbf{y}^n)$ is the unique solution to (5.27). The chain rule yields

$$\mathbf{0}=\mathbf{D}_n^*\mathbf{\Phi}_i+\sum_{l=1}^{i}\mathbf{D}_l^*\mathbf{\Phi}_i\mathbf{D}^*\mathbf{y}^{(l)},\quad i=1,\dots,s,$$

$$\mathbf{0}=\mathbf{D}_n^*\mathbf{\Phi}_{n+1}+\sum_{l=1}^{s}\mathbf{D}_l^*\mathbf{\Phi}_{n+1}\mathbf{D}^*\mathbf{y}^{(l)}+\mathbf{D}_{n+1}^*\mathbf{\Phi}_{n+1}\mathbf{D}\mathbf{g}(\mathbf{y}^*),$$

which can be rewritten to

$$\begin{aligned}\mathbf{D}^*\mathbf{y}^{(i)}&=-\left(\mathbf{D}_i^*\mathbf{\Phi}_i\right)^{-1}\left(\mathbf{D}_n^*\mathbf{\Phi}_i+\sum_{l=1}^{i-1}\mathbf{D}_l^*\mathbf{\Phi}_i\mathbf{D}^*\mathbf{y}^{(l)}\right),\quad j=1,\dots,s,\\ \mathbf{D}\mathbf{g}(\mathbf{y}^*)&=-\left(\mathbf{D}_{n+1}^*\mathbf{\Phi}_{n+1}\right)^{-1}\left(\mathbf{D}_n^*\mathbf{\Phi}_{n+1}+\sum_{l=1}^{s}\mathbf{D}_l^*\mathbf{\Phi}_{n+1}\mathbf{D}^*\mathbf{y}^{(l)}\right),\end{aligned}\tag{5.30}$$

if all occurring inverses exist. Also, in order to avoid long formulas in the following, we may omit to write the functions $\mathbf{\Phi}_i$ together with all their arguments.

Since we already discussed the case of Runge–Kutta methods in Section 2.4 we start analyzing MPRK schemes. To that end, we will use the notation of MPRK as an NSARK method. In contrast, all other MP methods presented in Chapter 3 will be analyzed directly because of the following. First, the NS weights of gBBKS and GeCo are not in $\mathcal{C}^1$. Also, as discussed in Remark 3.10, MPDeC methods can be understood as NSARK methods with potentially negative Butcher tableau entries. This is why we will focus in this work on the ansatz followed in [IÖ23]. Moreover, SSPMPRK methods do not fit into the form of an NSARK method. Nonetheless, their analysis using ARK methods in Shu–Osher form will be part of my future research.

5.4.1 Modified Patankar–Runge–Kutta

It turns out to be convenient to derive the stability properties of MPRK methods using the notation of NSARK schemes. However, we thereby restrict to non-negative Butcher tableaux, i. e. we use the vector notation (3.2). Moreover, we derive the Jacobian of the NSARK method in a more general context since Theorem 5.4 is not restricted only to linear systems. In particular, let us consider $\mathbf{y}'=\mathbf{f}(\mathbf{y})=\mathbf{F}(\mathbf{y})\mathbf{y}$, where $\mathbf{F}(\mathbf{y})\in\mathbb{R}^{N\times N}$ consists of the columns $\mathbf{F}^1(\mathbf{y}),\dots,\mathbf{F}^N(\mathbf{y})$. Hence, $\mathbf{F}=\sum_{\nu=1}^N\mathbf{F}^\nu\mathbf{e}_\nu^T$, where $\mathbf{e}_\nu$ is the νth column unit vector in $\mathbb{R}^N$. Moreover, this gives $\mathbf{f}(\mathbf{y})=\sum_{\nu=1}^N\mathbf{F}^\nu(\mathbf{y})y_\nu=\sum_{\nu=1}^N\mathbf{f}^{[\nu]}(\mathbf{y})$ with

$$\mathbf{f}^{[\nu]}(\mathbf{y})=\mathbf{F}^\nu(\mathbf{y})y_\nu.$$

Note that in the case of the linear system (5.1), we have $\mathbf{F}(\mathbf{y})=\mathbf{\Lambda}$. We restrict to conservative problems, which means that we will assume that $\mathbf{1}^T\mathbf{f}(\mathbf{y})=\mathbf{0}$ for all $\mathbf{y}$ in the domain of $\mathbf{f}$. With that we reproduce the results from the literature [IKM22a, IKM22b, IÖ23]. The generalization to non-conservative problems is then straightforward.

Since MPRK methods are linear implicit and based on explicit RK schemes, the stage equation for $\mathbf{y}^{(i)}$ depends only $\mathbf{y}^n,\dots,\mathbf{y}^{(i)}$. even more, we have $\mathbf{y}^n=\mathbf{y}^{(1)}$, however, in order to keep the notation, we will not substitute this directly into the stage equations, $\mathbf{\Phi}_i$ or $\mathbf{\Phi}_{n+1}$.

In [AGKM21] it was assumed that the PWDs only depend on the νth component of the stages, i. e.

$$\pi_\nu^{(i)} = \pi_\nu^{(i)}(y_\nu^n, y_\nu^{(1)}, \dots, y_\nu^{(i-1)}) \quad \text{and} \quad \sigma_\nu = \sigma_\nu(y_\nu^n, y_\nu^{(1)}, \dots, y_\nu^{(s)}), \tag{5.31}$$

which includes the PWDs presented in [KM18a, KM18b]. Thus we will assume this as well for our analysis. Moreover, the NS weights

$$\gamma_\nu^{[i]} = \frac{y_\nu^{(i)}}{\pi_\nu^{(i)}} \quad \text{and} \quad \delta_\nu = \frac{y_\nu^{n+1}}{\sigma_\nu},$$

see (3.6), will be understood as functions of the stages in the following. Furthermore, we will assume that

$$\pi_\nu^{(i)}(y_\nu^*, y_\nu^*, \dots, y_\nu^*) = y_\nu^* \quad \text{and} \quad \sigma_\nu(y_\nu^*, y_\nu^*, \dots, y_\nu^*) = y_\nu^*, \tag{5.32}$$

for any steady state $\mathbf{y}^*$ of the ODE, which is also fulfilled by the MPRK schemes presented so far. Therefore,

$$\gamma_\nu^{[i]}(y_\nu^*, y_\nu^*, \dots, y_\nu^*, y_\nu^*) = 1 \quad \text{and} \quad \delta_\nu(y_\nu^*, y_\nu^*, \dots, y_\nu^*, y_\nu^*) = 1. \tag{5.33}$$

Altogether, the mappings $\mathbf{\Phi}_i, \mathbf{\Phi}_{n+1}$ of the NSARK method (NSARK) are

$$\begin{aligned} \mathbf{\Phi}_i &= \mathbf{y}^n + \Delta t \sum_{j=1}^{i-1} \sum_{\nu=1}^{N} a_{ij} \gamma_\nu^{[i]}(y_\nu^n, y_\nu^{(1)}, \dots, y_\nu^{(i)}) \mathbf{f}^{[\nu]}(\mathbf{y}^{(j)}) - \mathbf{y}^{(i)}, \\ \mathbf{\Phi}_{n+1} &= \mathbf{y}^n + \Delta t \sum_{j=1}^{s} \sum_{\nu=1}^{N} b_j \delta_\nu(y_\nu^n, y_\nu^{(1)}, \dots, y_\nu^{(s)}, y_\nu^{n+1}) \mathbf{f}^{[\nu]}(\mathbf{y}^{(j)}) - \mathbf{y}^{n+1}. \end{aligned} \tag{5.34}$$

Now, in the special case of $\mathbf{f}(\mathbf{y}) = \mathbf{\Lambda}\mathbf{y}$, we note that

$$\sum_{\nu=1}^{N} x_\nu \mathbf{f}^{[\nu]}(\mathbf{y}^{(j)}) = \mathbf{F}(\mathbf{y}^{(j)}) \cdot \begin{pmatrix} x_1 y_1^{(j)} \\ \vdots \\ x_N y_N^{(j)} \end{pmatrix} = \mathbf{\Lambda} \cdot \begin{pmatrix} x_1 y_1^{(j)} \\ \vdots \\ x_N y_N^{(j)} \end{pmatrix}$$

for any values $x_1, \dots, x_N$. Substituting this information into (5.34), we observe that MPRK schemes preserve all linear invariants. Moreover, due to (5.33) we see that the MPRK methods are steady state preserving as already mentioned.

Moreover, the maps $\mathbf{\Phi}_i$ and $\mathbf{\Phi}_{n+1}$ are in $\mathcal{C}^2$ for positive arguments, and as defined in (5.34), vanish for the argument $(\mathbf{y}^n, \mathbf{y}^{(1)}(\mathbf{y}^n), \dots, \mathbf{y}^{(i)}(\mathbf{y}^n))$, and $(\mathbf{y}^n, \mathbf{y}^{(1)}(\mathbf{y}^n), \dots, \mathbf{y}^{(s)}(\mathbf{y}^n), \mathbf{g}(\mathbf{y}^n))$, respectively. And since the computation of $\mathbf{y}^{n+1}$ requires only the solution of linear systems which possess always a unique solution for any $\mathbf{y}^n > \mathbf{0}$, the function $\mathbf{g}$ is also a C^2-map. According to Remark 5.7 and Theorem 5.4 we thus only need to compute the eigenvalues of the Jacobian of $\mathbf{g}$ to investigate the stability of MPRK schemes. The upcoming lemma is a first step towards this goal.

Lemma 5.10. Assume $\mathbf{1}^T\mathbf{f} = \mathbf{0}$, and that (5.31) and (5.32) hold with $\pi_\nu^{(i)}, \sigma_\nu \in \mathcal{C}^1$, and let $\mathbf{f}^{[\nu]} \in \mathcal{C}^1$. Then $\mathbf{y}^n = \mathbf{y}^*$ implies $\mathbf{y}^{(i)} = \mathbf{y}^*$ and $\mathbf{y}^{n+1} = \mathbf{y}^*$ for any positive steady state $\mathbf{y}^*$, and the maps $\mathbf{\Phi}_i$ and $\mathbf{\Phi}_{n+1}$ from (5.34) satisfy

$$\mathbf{D}_k^*\mathbf{\Phi}_i = \begin{cases} \mathbf{I} - \Delta t c_i \mathbf{F}(\mathbf{y}^*)\mathbf{D}_n^*\boldsymbol{\pi}^{(i)}, & k = n,\\ -\Delta t c_i \mathbf{F}(\mathbf{y}^*)\mathbf{D}_k^*\boldsymbol{\pi}^{(i)} + \Delta t a_{ik}\mathbf{D}\mathbf{f}(\mathbf{y}^*), & k = 1,\dots,i-1,\\ \Delta t c_i \mathbf{F}(\mathbf{y}^*) - \mathbf{I}, & k = i, \end{cases}$$

$$\mathbf{D}_l^*\mathbf{\Phi}_{n+1} = \begin{cases} \mathbf{I} - \Delta t \mathbf{F}(\mathbf{y}^*)\mathbf{D}_n^*\boldsymbol{\sigma}, & l = n,\\ -\Delta t \mathbf{F}(\mathbf{y}^*)\mathbf{D}_l^*\boldsymbol{\sigma} + \Delta t b_l\mathbf{D}\mathbf{f}(\mathbf{y}^*), & l = 1,\dots,s,\\ \Delta t \mathbf{F}(\mathbf{y}^*) - \mathbf{I}, & l = n+1, \end{cases}$$

where $\boldsymbol{\pi}^{(i)} = (\pi_1^{(i)}, \dots, \pi_N^{(i)})^T$ and $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_N)^T$.

Proof. Let $\delta_{m,l}$ denote the Kronecker delta. For $i = 1,\dots,s$, straightforward calculations yield

$$\mathbf{D}_k^*\mathbf{\Phi}_i = \begin{cases} \mathbf{I} + \Delta t \sum_{j=1}^{i-1}\sum_{\nu=1}^N a_{ij}\mathbf{f}^{[\nu]}(\mathbf{y}^*)\nabla_k^*\gamma_\nu^{[i]}, & k = n,\\ \Delta t \sum_{j=1}^{i-1}\sum_{\nu=1}^N a_{ij}\mathbf{f}^{[\nu]}(\mathbf{y}^*)\nabla_k^*\gamma_\nu^{[i]} + \Delta t a_{ik}\mathbf{D}\mathbf{f}(\mathbf{y}^*), & k = 1,\dots,i-1,\\ \Delta t \sum_{j=1}^{i-1}\sum_{\nu=1}^N a_{ij}\mathbf{f}^{[\nu]}(\mathbf{y}^*)\nabla_k^*\gamma_\nu^{[i]} - \mathbf{I}, & k = i, \end{cases}$$

$$\mathbf{D}_l^*\mathbf{\Phi}_{n+1} = \begin{cases} \mathbf{I} + \Delta t \sum_{j=1}^{s}\sum_{\nu=1}^N b_j\mathbf{f}^{[\nu]}(\mathbf{y}^*)\nabla_l^*\delta_\nu, & l = n,\\ \Delta t \sum_{j=1}^{s}\sum_{\nu=1}^N b_j\mathbf{f}^{[\nu]}(\mathbf{y}^*)\nabla_l^*\delta_\nu + \Delta t b_l\mathbf{D}\mathbf{f}(\mathbf{y}^*), & l = 1,\dots,s,\\ \Delta t \sum_{j=1}^{s}\sum_{\nu=1}^N b_j\mathbf{f}^{[\nu]}(\mathbf{y}^*)\nabla_l^*\delta_\nu - \mathbf{I}, & l = n+1, \end{cases}$$

where

$$\nabla_k^*\gamma_\nu^{[i]} = \begin{cases} \frac{1}{y_\nu^*}\mathbf{e}_\nu^T, & k = i,\\ -\frac{1}{y_\nu^*}\nabla_k^*\pi_\nu^{(i)}, & k \neq i, \end{cases}$$

$$\nabla_l^*\delta_\nu = \begin{cases} \frac{1}{y_\nu^*}\mathbf{e}_\nu^T, & l = n+1,\\ -\frac{1}{y_\nu^*}\nabla_l^*\sigma_\nu, & l \neq n+1. \end{cases}$$

Using

$$\sum_{\nu=1}^N \mathbf{f}^{[\nu]}(\mathbf{y}^*)\frac{1}{y_\nu^*}\mathbf{e}_\nu^T = \sum_{\nu=1}^N \mathbf{F}^\nu(\mathbf{y}^*)\mathbf{e}_\nu^T = \mathbf{F}(\mathbf{y}^*),$$

we end up with

$$\begin{aligned} \mathbf{D}_k^*\mathbf{\Phi}_i &= \begin{cases} \mathbf{I} - \Delta t \sum_{j=1}^{i-1}\sum_{\nu=1}^N a_{ij}\mathbf{f}^{[\nu]}(\mathbf{y}^*)\frac{1}{y_\nu^*}\nabla_n^*\pi_\nu^{(i)}, & k = n,\\ -\Delta t \sum_{j=1}^{i-1}\sum_{\nu=1}^N a_{ij}\mathbf{f}^{[\nu]}(\mathbf{y}^*)\frac{1}{y_\nu^*}\nabla_k^*\pi_\nu^{(i)} + \Delta t a_{ik}\mathbf{D}\mathbf{f}(\mathbf{y}^*), & k = 1,\dots,i-1,\\ \Delta t \sum_{j=1}^{i-1}\sum_{\nu=1}^N a_{ij}\mathbf{f}^{[\nu]}(\mathbf{y}^*)\frac{1}{y_\nu^*}\mathbf{e}_\nu^T - \mathbf{I}, & k = i, \end{cases}\\ &= \begin{cases} \mathbf{I} - \Delta t c_i \mathbf{F}(\mathbf{y}^*)\mathbf{D}_n^*\boldsymbol{\pi}^{(i)}, & k = n,\\ -\Delta t c_i \mathbf{F}(\mathbf{y}^*)\mathbf{D}_k^*\boldsymbol{\pi}^{(i)} + \Delta t a_{ik}\mathbf{D}\mathbf{f}(\mathbf{y}^*), & k = 1,\dots,i-1,\\ \Delta t c_i \mathbf{F}(\mathbf{y}^*) - \mathbf{I}, & k = i. \end{cases} \end{aligned}$$

Analogously, we obtain

$$\mathbf{D}_l^*\mathbf{\Phi}_{n+1} = \begin{cases} \mathbf{I} - \Delta t \mathbf{F}(\mathbf{y}^*)\mathbf{D}_n^*\boldsymbol{\sigma}, & l = n,\\ -\Delta t \mathbf{F}(\mathbf{y}^*)\mathbf{D}_l^*\boldsymbol{\sigma} + \Delta t b_l\mathbf{D}\mathbf{f}(\mathbf{y}^*), & l = 1,\dots,s,\\ \Delta t \mathbf{F}(\mathbf{y}^*) - \mathbf{I}, & l = n+1. \end{cases} \qquad \square$$

For linear conservative systems we thus obtain the following from (5.30).

Theorem 5.11. In the situation of Lemma 5.10, the Jacobian of the generating map $\mathbf{g}$ of (MPRK) applied to a conservative problem $\mathbf{y}' = \mathbf{\Lambda}\mathbf{y}$ with $\sigma(\mathbf{\Lambda}) \subseteq \overline{\mathbb{C}^-}$ reads

$$\begin{aligned}\mathbf{D}^*\mathbf{y}^{(i)} &= (\mathbf{I}-\Delta t c_i\mathbf{\Lambda})^{-1}\left(\mathbf{I}-\Delta t c_i\mathbf{\Lambda}\mathbf{D}_n^*\boldsymbol{\pi}^{(i)} - \Delta t\mathbf{\Lambda}\sum_{l=1}^{i-1}\left(c_i\mathbf{D}_l^*\boldsymbol{\pi}^{(i)} - a_{il}\mathbf{I}\right)\mathbf{D}^*\mathbf{y}^{(l)}\right),\\ \mathbf{D}\mathbf{g}(\mathbf{y}^*) &= (\mathbf{I}-\Delta t\mathbf{\Lambda})^{-1}\left(\mathbf{I}-\Delta t\mathbf{\Lambda}\mathbf{D}_n^*\boldsymbol{\sigma} - \Delta t\mathbf{\Lambda}\sum_{l=1}^{s}\left(\mathbf{D}_l^*\boldsymbol{\sigma} - b_l\mathbf{I}\right)\mathbf{D}^*\mathbf{y}^{(l)}\right),\end{aligned} \tag{5.35}$$

where $i = 1, \dots, s$.

Proof. The inverses exist since $\sigma(\Delta t c_i\mathbf{\Lambda} - \mathbf{I}) \subseteq \mathbb{C}^-$ for all $c_i \geq 0$. The rest follows from (5.30) and Lemma 5.10. □

Remark 5.12. Our framework opens the door to a comprehensive approach of investigating even PDRS, since negative rest terms $\mathbf{r}^d$ are weighted like destruction terms and positive rest terms $\mathbf{r}^p$ are not modified. Hence, already at this point we may also consider PDRS with $\mathbf{r}^p = \mathbf{0}$ and $\mathbf{r}^d > \mathbf{0}$ and investigate the asymptotic stability of the origin using the same stability function as for $\mathbf{r} = \mathbf{0}$. In the case of $\mathbf{r}^p > \mathbf{0}$, one may revisit the proof of Lemma 5.10 adjusting the appearing Jacobians of the PWDs. The analysis of Patankar–Runge–Kutta methods would then also be available since production terms can formally be treated as positive rest terms. However, this together with the corresponding analyses of the stability functions and numerical experiments is beyond this work.

In the following we replicate the results from [IKM22a, IKM22b, IÖ23] using this new framework.

MPE The MPE method for conservative and autonomous PDS can be found in (MPE). Since the first stage equals $\mathbf{y}^n$ and we also have $\boldsymbol{\sigma} = \mathbf{y}^n$, we find from (5.35) that

$$R(z) = \frac{1-z+z}{1-z} = \frac{1}{1-z}.$$

Hence, the MPE method has the same stability function as the implicit Euler scheme. As a consequence of that and Theorem 5.4 we obtain the following results.

Corollary 5.13. The MPE method is unconditionally stable in the sense of Definition 5.2.

Corollary 5.14. Let $\mathbf{y}^*$ be the unique steady state of the initial value problem (5.1), (5.2) with $\mathbf{1} \in \ker(\mathbf{\Lambda}^T)$. Then there exists a $\delta > 0$ such that $\|\mathbf{y}^0 - \mathbf{y}^*\| < \delta$ implies the convergence of the iterates of MPE towards $\mathbf{y}^*$ as $n \to \infty$ for all $\Delta t > 0$.

MPRK22(α) The second order MPRK method for a conservative and autonomous PDS is given in (MPRK22). Here, we have $\mathbf{y}^{(1)} = \mathbf{y}^n$, $\boldsymbol{\pi}^{(2)} = \mathbf{y}^n$ and $\sigma_\nu = (y_\nu^{(2)})^{\frac{1}{\alpha}}(y_\nu^n)^{1-\frac{1}{\alpha}}$. Hence, due to $c_1 = 0$, we find $\mathbf{D}^*\mathbf{y}^{(1)} = \mathbf{I}$ and

$$\mathbf{D}_n^*\boldsymbol{\pi}^{(2)} = \mathbf{I}, \quad \mathbf{D}_1^*\boldsymbol{\pi}^{(2)} = \mathbf{0}, \quad \mathbf{D}_n^*\boldsymbol{\sigma} = \left(1-\frac{1}{\alpha}\right)\mathbf{I}, \quad \mathbf{D}_1^*\boldsymbol{\sigma} = \mathbf{0}, \quad \mathbf{D}_2^*\boldsymbol{\sigma} = \frac{1}{\alpha}\mathbf{I}.$$

Since $\mathbf{Dg}(\mathbf{y}^*)$ is a rational function of $\mathbf{\Lambda}$ and the identity matrix $\mathbf{I}$, any eigenvector of $\mathbf{\Lambda}$ with the eigenvalue λ is consequently an eigenvector of $\mathbf{Dg}(\mathbf{y}^*)$. Therefore, using (5.35) we see $\sigma(\mathbf{Dg}(\mathbf{y}^*)) = \{R(\Delta t\lambda) \mid \lambda \in \sigma(\mathbf{\Lambda})\}$, where

$$\begin{aligned} R(z) &= \frac{1 - z\left(1-\frac{1}{\alpha}\right) - z\left(0 - b_1 + (\frac{1}{\alpha} - b_2)\frac{1-c_2z-z(0+a_{21})}{1-c_2z}\right)}{1-z} \\ &= \frac{1 - z\left(1-\frac{1}{\alpha}\right) - z\left(-1 + \frac{1}{2\alpha} + (\frac{1}{\alpha} - \frac{1}{2\alpha})\frac{1-2\alpha z}{1-\alpha z}\right)}{1-z} = \frac{-z^2-2\alpha z+2}{2(1-\alpha z)(1-z)}. \end{aligned} \tag{5.36}$$

Proposition 5.15. The stability function $R(z) = \frac{-z^2-2\alpha z+2}{2(1-\alpha z)(1-z)}$ from (5.36) with $\alpha > \frac{1}{2}$ satisfies $R(0) = 1$ and $|R(z)| < 1$ for all $z \in \overline{\mathbb{C}^-} \setminus \{0\}$. For $\alpha = \frac{1}{2}$ we have $|R(z)| < 1$ for all z with $\operatorname{Re}(z) < 0$ and $|R(z)| = 1$, if $\operatorname{Re}(z) = 0$.

Proof. We first investigate $|R(z)|$ for $z = \mathrm{i}y$ and $y \in \mathbb{R}$. A small calculation reveals that the numerator of $|R(z)|^2$ can be written as

$$|-z^2-2\alpha z+2|^2 = |y^2+2+(-2\alpha y)\mathrm{i}|^2 = (y^2+2)^2+4\alpha^2y^2 = y^4+4y^2(1+\alpha^2)+4. \tag{5.37}$$

Performing a similar calculation for the denominator of $|R(z)|^2$ we find

$$\begin{aligned} |2(1-\alpha z)(1-z)|^2 &= |2\alpha z^2 - 2z(1+\alpha)+2|^2 = |-2\alpha y^2+2+(-2y(1+\alpha))\mathrm{i}|^2 \\ &= (-2\alpha y^2+2)^2+4y^2(1+\alpha)^2 = 4\alpha^2y^4+4y^2(1+\alpha^2)+4. \end{aligned}$$

Using (5.37) and $\alpha = \frac{1}{2}$ we see that $|R(z)| = 1$ on the imaginary axis, and if $\alpha > \frac{1}{2}$ we find $|R(\mathrm{i}y)| < 1$ for all $y \neq 0$.

Next we note that R is a holomorphic function which is defined for all $z \in \overline{\mathbb{C}^-}$. Since R is rational we can apply the Phragmén–Lindelöf principle [SS03, Tit39] on the union of the origin and $\mathbb{C}^-$ and conclude that $|R(z)| \leq 1$ for all $z \in \overline{\mathbb{C}^-}$. Furthermore, since R is not constant, we conclude from the maximum modulus principle that there exist no $z_0 \in \mathbb{C}^-$ with $|R(z_0)| = 1$, or equivalently, $|R(z_0)| < 1$ holds for all z_0 with $\operatorname{Re}(z_0) < 0$.

□

As a direct consequence of the application of Theorem 5.4 in combination with Proposition 5.15 we obtain the following two corollaries, whereby we note that all nonzero eigenvalues of $\mathbf{\Lambda}$ from (5.1) have a negative real part, see Remark 5.1.

Corollary 5.16. The MPRK22(α) scheme is unconditionally stable for all $\alpha \geq \frac{1}{2}$.

Corollary 5.17. Let $\mathbf{y}^*$ be the unique steady state of the initial value problem (5.1), (5.2) with $\mathbf{1} \in \ker(\mathbf{\Lambda}^T)$. Then there exists a $\delta > 0$ such that $\|\mathbf{y}^0 - \mathbf{y}^*\| < \delta$ implies the convergence of the iterates of MPRK22(α) towards $\mathbf{y}^*$ as $n \to \infty$ for all $\Delta t > 0$ and $\alpha \geq \frac{1}{2}$.

Remark 5.18. We note that as long as the stability function $R(z) = \frac{N(z)}{D(z)}$ with polynomials N, D satisfying $\deg(N) \leq \deg(D)$ and $D(z) \neq 0$ for all $z \in \overline{\mathbb{C}^-}$, we can conclude $|R(z)| < 1$ for $\operatorname{Re}(z) < 0$ whenever $|R(z)| \leq 1$ holds on the imaginary axis with the same reasoning as in the proof of Proposition 5.15.

Moreover, we point out that the Phragmén–Lindelöf principle can also be applied to different sectors $S_{(\varphi_1,\varphi_2)} = \left\{ z \in \overline{\mathbb{C}^-} \mid \arg(z) \in (\varphi_1, \varphi_2) \right\}$ of $\overline{\mathbb{C}^-}$.

MPRK43 We consider the two families of third order MPRK schemes presented in Chapter 3. The PWDs can be found in (3.9), where $\boldsymbol{\sigma}$ is given only implicitly. However, interpreting $\boldsymbol{\sigma} = \mathbf{y}^{(4)}$ and introducing $a_{41} = \beta_1$ and $a_{42} = \beta_2$ as well as $\pi_\nu^{(4)} = (y_\nu^{(2)})^{\frac{1}{a_{21}}}(y_\nu^n)^{1-\frac{1}{a_{21}}}$, we can use the derived formula from Theorem 5.11. To that end, we note that the nonzero auxiliary Jacobians are

$$\mathbf{D}_n^*\boldsymbol{\pi}^{(2)} = \mathbf{I}, \quad \mathbf{D}_n^*\boldsymbol{\pi}^{(3)} = \left(1-\tfrac{1}{p}\right)\mathbf{I}, \quad \mathbf{D}_2^*\boldsymbol{\pi}^{(3)} = \tfrac{1}{p}\mathbf{I},$$
$$\mathbf{D}_n^*\boldsymbol{\pi}^{(4)} = \left(1-\tfrac{1}{a_{21}}\right)\mathbf{I}, \quad \mathbf{D}_2^*\boldsymbol{\pi}^{(4)} = \tfrac{1}{a_{21}}\mathbf{I}, \quad \mathbf{D}_4^*\boldsymbol{\sigma} = \mathbf{I}.$$

As a result of (5.35), the stability function is

$$\begin{aligned} R(z) =& \frac{1 - z\left(-b_1 - b_2\frac{1-zc_2+za_{21}}{1-zc_2} - b_3\frac{1-zc_3\left(1-\frac{1}{p}\right)-z\left(-a_{31}+\left(c_3\frac{1}{p}-a_{32}\right)\frac{1-zc_2+za_{21}}{1-zc_2}\right)}{1-zc_3}\right)}{1-z} \\ &+ \frac{-z\frac{1-zc_4\left(1-\frac{1}{a_{21}}\right)-z\left(-a_{41}+\left(\frac{c_4}{a_{21}}-a_{42}\right)\frac{1-zc_2+za_{21}}{1-zc_2}\right)}{1-zc_4}}{1-z} \\ =& \frac{1 + z\left(b_1 + \frac{b_2}{1-zc_2} + b_3\frac{1+zc_3\left(\frac{1}{p}-1\right)+z\left(a_{31}+\frac{a_{32}-\frac{c_3}{p}}{1-zc_2}\right)}{1-zc_3}\right)}{1-z} \\ &- \frac{z\frac{1+zc_4\left(\frac{1}{a_{21}}-1\right)+z\left(a_{41}+\frac{a_{42}-\frac{c_4}{a_{21}}}{1-zc_2}\right)}{1-zc_4}}{1-z}. \end{aligned} \tag{5.38}$$

MPRK43(α, β) In [IÖ23] the stability function of MPRK43(α, β) was computed using a different approach. Unfortunately, there is a typo in the stability function on page 2328: Instead of writing "$-\frac{\beta}{p}$" as suggested in equation (33) on the same page, it is written "$-\frac{\beta}{q}$". This typo undermines all claims that are based on it. The stability function actually is

$$\begin{aligned} R(z) =& \frac{1 + b_1 z + \frac{b_2 z}{1-\alpha z} + \frac{b_3 z\left(1+z\left(\left(\frac{1}{p}-1\right)\beta+a_{31}\right)+\frac{z\left(-\frac{\boldsymbol{\beta}}{\boldsymbol{p}}+a_{32}\right)}{1-\alpha z}\right)}{1-\beta z}}{1-z} \\ &- \frac{\frac{z\left(1+z\left(\frac{1}{q}-1+\beta_1\right)+\frac{z(-\frac{1}{q}+\beta_2)}{1-\alpha z}\right)}{1-z}}{1-z} \\ =& \frac{((\frac{1}{2}-\beta)\alpha-\frac{1}{6})z^4 + ((\frac{1}{2}-\beta)\alpha+\frac{1}{2}\beta+\frac{1}{6})z^3 + ((\beta+1)\alpha+\beta-\frac{1}{2})z^2}{(z-1)^2(z\beta-1)(\alpha z-1)} \\ &+ \frac{-(1+\alpha+\beta)z+1}{(z-1)^2(z\beta-1)(\alpha z-1)}, \end{aligned} \tag{5.39}$$

which can also be obtained within our framework by substituting (3.10) and (3.12) into (5.38). It is thus the purpose of this subsection to correct and to extend the results from [IÖ23] concerning this stability analysis.

Since different cases for different (α, β) pairs need to be distinguished, see

(3.11), the analysis of the stability function (5.39) is more involved. Moreover, we will find out that the method is not unconditionally stable for all feasible parameters. In order to give an insight in the stability properties, we investigate the stability function numerically. We first rewrite (5.39) to

$$R(z) = \frac{\sum_{j=0}^{4} n_j z^j}{\sum_{j=0}^{4} d_j z^j},$$

where

$$\begin{aligned} n_0 &= 1, & d_0 &= 1,\\ n_1 &= -(1+\alpha+\beta), & d_1 &= -(\alpha+\beta+2),\\ n_2 &= (\beta+1)\alpha+\beta-\tfrac{1}{2}, & d_2 &= (\beta+2)\alpha+2\beta+1,\\ n_3 &= (\tfrac{1}{2}-\beta)\alpha+\tfrac{1}{2}\beta+\tfrac{1}{6}, & d_3 &= -(2\beta+1)\alpha-\beta,\\ n_4 &= (\tfrac{1}{2}-\beta)\alpha-\tfrac{1}{6}, & d_4 &= \alpha\beta. \end{aligned}$$

In what follows, we investigate the polynomial $p_{\alpha,\beta,\varphi}(r)$ from Lemma A.1 satisfying

$$|R(re^{\mathrm{i}\varphi})| < 1 \iff p_{\alpha,\beta,\varphi}(r) < 0 \quad \text{and} \quad |R(re^{\mathrm{i}\varphi})| > 1 \iff p_{\alpha,\beta,\varphi}(r) > 0,$$

which means that MPRK43(α,β) is unconditionally stable, if $p_{\alpha,\beta,\frac{\pi}{2}}(r) < 0$ for all $r > 0$. Moreover, if $p_{\alpha,\beta,\frac{\pi}{2}}(r) > 0$ for some $r > 0$, then the method cannot be unconditionally stable. For instance, observing $p_{\frac{1}{3},\frac{2}{3},\frac{\pi}{2}}(r) = -\frac{11}{81}r^6 - \frac{11}{36}r^4 < 0$ for all $r > 0$, we find that MPRK$(\frac{1}{3},\frac{2}{3})$ is unconditionally stable.

We want to note that for any other feasible pair (α,β), see Figure 3.1, the degree of $p_{\alpha,\beta,\varphi}$ is 8. To see this, we point out that the leading coefficient

$$n_4^2 - d_4^2 = -\alpha^2\beta^2 + (-\alpha\beta + \tfrac{1}{2}\alpha - \tfrac{1}{6})^2 = -(\alpha - \tfrac{1}{3})\alpha\beta + (\tfrac{\alpha}{2} - \tfrac{1}{6})^2 \tag{5.40}$$

vanishes for $(\alpha,\beta) = (\frac{1}{3},\frac{2}{3})$ and $(\alpha, \frac{3\alpha-1}{12\alpha})$. However, since $\frac{3\alpha-1}{12\alpha} < \frac{3\alpha}{12\alpha} = \frac{1}{4}$ and all feasible values of β lie in $[0.25, 0.75]$, the pair $(\alpha, \frac{3\alpha-1}{12\alpha})$ does not lie in the feasible domain.

Altogether, the question of unconditional stability can be answered if the polynomial $p_{\alpha,\beta,\frac{\pi}{2}}$ has no positive root because of the following. Suppose that all roots are non-positive. Since the coefficient of β in the leading coefficient (5.40) of $p_{\alpha,\beta,\varphi}(r)$ is negative for $\alpha > \frac{1}{3}$, we find due to $\beta \in [0.25, 0.75]$, that

$$n_4^2 - d_4^2 \le -(\alpha - \tfrac{1}{3})\alpha\tfrac{1}{4} + (\tfrac{\alpha}{2} - \tfrac{1}{6})^2 = \tfrac{1}{36} - \tfrac{\alpha}{12} < 0.$$

This means that $\lim_{r\to\infty} p_{\alpha,\beta,\varphi}(r) = -\infty$. Finally, as we assumed that there are no positive roots, this implies that the polynomial is negative for all $r > 0$.

For the investigation of the remaining parameter combinations, we create a grid for $(\alpha,\beta) \in [\frac{1}{3}, 2] \times [0.25, 0.75]$ with a resolution of 102^2 points in a unit square $[0,1]^2$. We chose this resolution to sample the domain for β of length 0.5 with about 100 points. In particular, we use 102 points so that the set of sampled pairs includes the combinations $(0.5, 0.75)$ and $(1, 0.5)$ which are used in the literature [KM18b]. Given a pair (α,β) from the grid, we also sample $\varphi_j = \frac{j}{1000}\pi$, $j = 0, 1, \ldots, 1000$ and determine the smallest value of j such that $p_{\alpha,\beta,\varphi_j}$ has no positive root by using Sturm's Theorem [Coh03, Theorem 8.8.14]. The corresponding value

$$\theta_{\mathrm{num}} = \min_{j=0,1,\ldots,1000} \{2(\pi - \varphi_j) \mid p_{\alpha,\beta,\varphi_j}(r) = 0 \Longrightarrow r \le 0\} \tag{5.41}$$

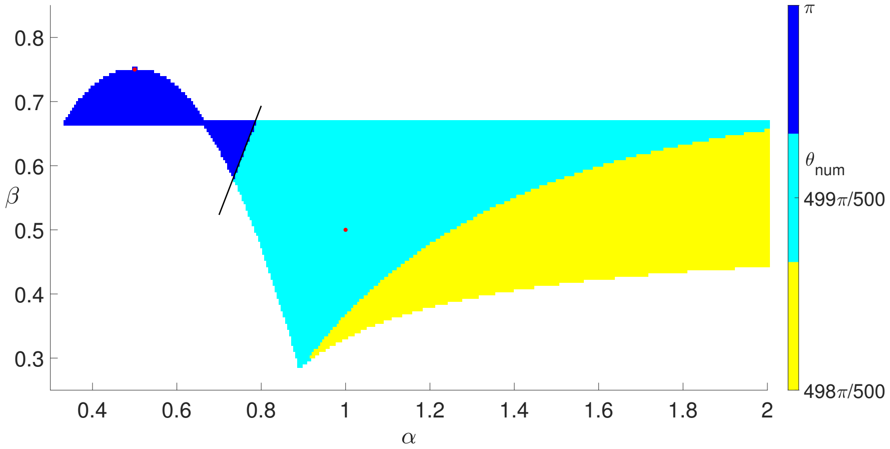


Figure 5.1: Plot of θ_{num}, see (5.41), estimating the opening angle of the stability domain of MPRK43(α, β). Pairs (α, β) colored in dark blue belong to $\theta_{\text{num}} = \pi$ and yield an unconditionally stable MPRK43(α, β) method. The black line is given by $\beta = \frac{17}{10}\alpha - \frac{2}{3}$ and its intersection with the feasible domain lies in the dark blue segment. The red markers indicate the position of the pairs $(0.5, 0.75)$ with an opening angle of at least $\theta_{\text{num}} = \pi$, and $(1, 0.5)$ with an opening angle of at least $\theta_{\text{num}} = \pi - \frac{\pi}{500} = \frac{499}{500}\pi$.

represents a lower bound for the opening angle θ of the stability domain of MPRK43(α, β). Moreover, if there exist a simple positive root, or with odd multiplicity, we know that the polynomial $p_{\alpha,\beta,\frac{\pi}{2}}$ will become positive within a neighborhood of that root, and hence, the method cannot be unconditional stable. In this case we have the the error estimate $\theta_{\text{num}} \leq \theta < \theta_{\text{num}} + \frac{\pi}{500}$ since

$$2(\pi - \varphi_j) - 2(\pi - \varphi_{j+1}) = \frac{\pi}{500}.$$

Note that in the case of $\theta_{\text{num}} \geq \pi$ the related method is unconditionally stable. The plot of θ_{num} can be found in Figure 5.1, noting that the MPRK43(α, β) scheme is not defined for $\alpha = \beta$.

As one can see, several parameter combinations (α, β) are unconditionally stable and the smallest θ_{num} observed is $\frac{498}{500}\pi$. In particular the pair corresponding to $\alpha = \frac{1}{2}$ and $\beta = \frac{3}{4}$ already used in the literature [KM18b] is now proved to be unconditionally stable. This cannot be said with certainty about the combination $\alpha = 1$ and $\beta = \frac{1}{2}$ as the corresponding value of θ_{num} is $\frac{499}{500}\pi$. Indeed, computing the roots of $p_{1,0.5,\frac{\pi}{2}}$ we find a simple positive root, proving that the corresponding method is not unconditionally stable.

MPRK43(γ) Substituting (3.13) and (3.14) into (5.38), we obtain the stability function

$$R(z) = \frac{-5z^4 + 7z^3 + 23z^2 - 42z + 18}{2(2z-3)^2(z-1)^2}. \tag{5.42}$$

Note that the stability function is independent of the parameter γ, so that the following investigation is valid for all $\frac{3}{8} \leq \gamma \leq \frac{3}{4}$.

Proposition 5.19. Let R be defined by (5.42). Then $|R(z)|<1$ holds true for all $z\in\overline{\mathbb{C}^-}\setminus\{0\}$ and $R(0)=1$.

Proof. A straightforward calculation yields

$$R(z)=\frac{-\frac{5}{18}z^4+\frac{7}{18}z^3+\frac{23}{18}z^2-\frac{42}{18}z+1}{(\frac{2}{3}z-1)^2(z-1)^2}=\frac{\sum_{j=0}^4 n_jz^j}{\sum_{j=0}^4 d_jz^j},$$

where

$$\begin{aligned}
&n_0=1, && n_1=-\frac{7}{3}, && n_2=\frac{23}{18}, && n_3=\frac{7}{18}, && n_4=-\frac{5}{18},\\
&d_0=1, && d_1=-\frac{10}{3}, && d_2=\frac{37}{9}, && d_3=-\frac{20}{9}, && d_4=\frac{4}{9}.
\end{aligned}$$

Hence $R(0)=1$ and the polynomial $p_\varphi(r)$ from Lemma A.1 satisfies

$$p_{\frac{\pi}{2}}(r)=-\frac{13}{108}r^8-\frac{137}{324}r^6-\frac{1}{12}r^4<0$$

for all $r>0$. Hence, it follows that $|R(\mathrm{i}y)|<1$ for all $y\neq 0$. The claim then follows from Remark 5.18. □

From this result, we can conclude as a direct consequence of Theorem 5.4 the following statements.

Corollary 5.20. a) The MPRK(γ) method is unconditionally stable for all $\frac{3}{8}\leq\gamma\leq\frac{3}{4}$.

b) If $\mathbf{y}^*$ is the unique steady state of the initial value problem (5.1), (5.2) with $\mathbf{1}\in\ker(\mathbf{\Lambda}^T)$, then there exists a $\delta>0$ such that $\|\mathbf{y}^0-\mathbf{y}^*\|<\delta$ implies the convergence of the iterates of MPRK(γ) towards $\mathbf{y}^*$ for all $\Delta t>0$ and $\frac{3}{8}\leq\gamma\leq\frac{3}{4}$.

5.4.2 Strong-Stability Preserving Modified Patankar–Runge–Kutta

As SSPMPRK schemes from [HS19, HZS19] are only constructed for positive and conservative PDS, we assume that the linear test equation (5.1) is conservative, i. e. $\mathbf{1}\in\ker(\mathbf{\Lambda}^T)$. Since $\mathbf{\Lambda}$ is a Metzler matrix, the test equation can be rewritten as a positive and conservative PDS with $p_{ij}(\mathbf{y})=d_{ji}(\mathbf{y})=\lambda_{ij}y_j$ for $i\neq j$ and $p_{ii}=d_{ii}=0$. Moreover, from $\mathbf{1}\in\ker(\mathbf{\Lambda}^T)$, one can easily derive $\sum_{j=1}^N\lambda_{ji}=0$ and thus obtain

$$-\sum_{j=1}^N d_{ij}(\mathbf{y})=-\sum_{\substack{j=1\\ j\neq i}}^N\lambda_{ji}y_i=\lambda_{ii}y_i, \tag{5.43}$$

which will be used in the following to write the SSPMPRK schemes in the matrix-vector notation.

SSPMPRK2(α,β) When applied to a conservative system (5.1), the terms p_{ij} and d_{ij} fulfill (5.43). As a consequence, the scheme (SSPMPRK2) can be rewritten as

$$\begin{aligned}\mathbf{0} =&\mathbf{\Phi}_1(\mathbf{y}^n,\mathbf{y}^{(1)}) = \mathbf{y}^n+\beta\Delta t\mathbf{\Lambda}\mathbf{y}^{(1)}-\mathbf{y}^{(1)},\\ \mathbf{0} =&\mathbf{\Phi}_{n+1}(\mathbf{y}^n,\mathbf{y}^{(1)},\mathbf{y}^{n+1}) = (1-\alpha)\mathbf{y}^n+\alpha\mathbf{y}^{(1)}\\ &+\Delta t\mathbf{\Lambda}\operatorname{diag}(\mathbf{y}^{n+1})(\operatorname{diag}(\mathbf{y}^{(1)}))^{-s}(\operatorname{diag}(\mathbf{y}^n))^{s-1}(\beta_{20}\mathbf{y}^n+\beta_{21}\mathbf{y}^{(1)})-\mathbf{y}^{n+1},\end{aligned}\tag{5.44}$$

where we use the notation $(\operatorname{diag}(\mathbf{y}))_{ij}=\delta_{ij}y_i$ with the Kronecker delta δ_{ij} as well as $((\operatorname{diag}(\mathbf{y}))^x)_{ij}=\delta_{ij}y_i^x$ for $x\in\mathbb{R}$. Furthermore, $\mathbf{y}^{(1)}=\mathbf{y}^{(1)}(\mathbf{y}^n)$ and $\mathbf{y}^{n+1}=\mathbf{g}(\mathbf{y}^n)$ defined by (5.44) are functions of $\mathbf{y}^n$. In order to apply Theorem 2.15 and Theorem 5.4, we have to investigate the map $\mathbf{g}$ with respect to its smoothness as well as steady state and linear invariants preservation.

First of all, we show that $\mathbf{g}\in\mathcal{C}^2$ and then use Remark 5.7 in order to see that the first derivatives are Lipschitz continuous on an appropriately chosen neighborhood $\mathcal{D}$ of $\mathbf{y}^*$.

Indeed, the maps $\mathbf{\Phi}_1\colon\mathbb{R}^N_{>0}\times\mathbb{R}^N_{>0}\to\mathbb{R}^N$ and $\mathbf{\Phi}_{n+1}\colon\mathbb{R}^N_{>0}\times\mathbb{R}^N_{>0}\times\mathbb{R}^N_{>0}\to\mathbb{R}^N$ are in $\mathcal{C}^2$ for the same reasons as for MPRK schemes, which means that $\mathbf{g}$ is also a $\mathcal{C}^2$-map.

Next, we show that any positive steady state of (5.1) is a fixed point of $\mathbf{g}$. To see this, we want to mention that $\mathbf{y}^n=\mathbf{y}^{(1)}=\mathbf{y}^{n+1}=\mathbf{y}^*$ is a solution to the system of equations (5.44) due to $\mathbf{\Lambda}\mathbf{y}^*=\mathbf{0}$. Since the solution for given $\mathbf{y}^n$ is unique, we conclude that $\mathbf{y}^n=\mathbf{y}^*$ implies $\mathbf{y}^{(1)}=\mathbf{y}^{n+1}=\mathbf{y}^*$, i. e. $\mathbf{g}(\mathbf{y}^*)=\mathbf{y}^*$.

Moreover, $\mathbf{g}$ conserves all linear invariants since $\mathbf{n}^T\mathbf{\Lambda}=\mathbf{0}$ and (5.44) imply

$$\begin{aligned}\mathbf{n}^T\mathbf{g}(\mathbf{y}^n)=\mathbf{n}^T\mathbf{y}^{n+1}&=(1-\alpha)\mathbf{n}^T\mathbf{y}^n+\alpha\mathbf{n}^T\mathbf{y}^{(1)}+\mathbf{0}\\ &=(1-\alpha)\mathbf{n}^T\mathbf{y}^n+\alpha\mathbf{n}^T(\mathbf{y}^n+\beta\Delta t\mathbf{\Lambda}\mathbf{y}^{(1)})=\mathbf{n}^T\mathbf{y}^n.\end{aligned}$$

Therefore, the map $\mathbf{g}\colon\mathbb{R}^N_{>0}\to\mathbb{R}^N_{>0}$ meets the assumptions of Theorem 5.4, so that we now focus on computing the Jacobian $\mathbf{Dg}(\mathbf{y}^*)$ according to (5.30). In particular, we have

$$\mathbf{Dg}(\mathbf{y}^*)=-(\mathbf{D}^*_{n+1}\mathbf{\Phi}_{n+1})^{-1}\left(\mathbf{D}^*_n\mathbf{\Phi}_{n+1}+\mathbf{D}^*_1\mathbf{\Phi}_{n+1}\mathbf{D}^*\mathbf{y}^{(1)}\right),\tag{5.45}$$

where

$$\mathbf{D}^*\mathbf{y}^{(1)}=-\left(\mathbf{D}^*_1\mathbf{\Phi}_1\right)^{-1}\mathbf{D}^*_n\mathbf{\Phi}_1,\tag{5.46}$$

if $\mathbf{D}^*_1\mathbf{\Phi}_1$ is invertible. Hence, we have to compute several auxiliary Jacobians in order to calculate $\mathbf{Dg}(\mathbf{y}^*)$ and we start with

$$\mathbf{D}^*_n\mathbf{\Phi}_1=\mathbf{I}\quad\text{and}\quad\mathbf{D}^*_1\mathbf{\Phi}_1=\beta\Delta t\mathbf{\Lambda}-\mathbf{I}.$$

Note that $\beta>0$ and $\sigma(\mathbf{\Lambda})\subseteq\overline{\mathbb{C}^-}$, which implies that $\mathbf{D}^*_1\mathbf{\Phi}_1$ is nonsingular. Thus, we can use (5.46) and find

$$\mathbf{D}^*\mathbf{y}^{(1)}=-(\beta\Delta t\mathbf{\Lambda}-\mathbf{I})^{-1}\cdot\mathbf{I}=(\mathbf{I}-\beta\Delta t\mathbf{\Lambda})^{-1}.$$

Next, we compute $\mathbf{D}^*_n\mathbf{\Phi}_{n+1}$ and $\mathbf{D}^*_1\mathbf{\Phi}_{n+1}$. To that end, we first define

$$\mathbf{f}(\mathbf{y}^n,\mathbf{y}^{(1)})=\operatorname{diag}(\mathbf{y}^n)^k(\beta_{20}\mathbf{y}^n+\beta_{21}\mathbf{y}^{(1)})$$

for some $k \in \mathbb{R}$ and get

$$\begin{aligned}(\mathbf{D}_n^*\mathbf{f})_{ij} &= \partial_{y_j^n}\left((y_i^n)^k(\beta_{20}y_i^n + \beta_{21}y_i^{(1)}))\right)\Big|_{\mathbf{y}^n=\mathbf{y}^*} \\ &= \delta_{ij}\left(k(y_i^*)^{k-1}(\beta_{20}+\beta_{21})y_i^* + (y_i^*)^k\beta_{20}\right) \\ &= \left(\operatorname{diag}(\mathbf{y}^*)^k\right)_{ij}(k(\beta_{20}+\beta_{21})+\beta_{20}),\end{aligned} \tag{5.47}$$

where we have used the fact that $\mathbf{y}^{(1)}(\mathbf{y}^*) = \mathbf{y}^*$. Similarly, defining

$$\mathbf{u}(\mathbf{y}^n, \mathbf{y}^{(1)}) = \operatorname{diag}(\mathbf{y}^{(1)})^k(\beta_{20}\mathbf{y}^n + \beta_{21}\mathbf{y}^{(1)}),$$

we obtain

$$\mathbf{D}_1^*\mathbf{u} = \operatorname{diag}(\mathbf{y}^*)^k(k(\beta_{20}+\beta_{21})+\beta_{21}). \tag{5.48}$$

In order to apply the formulae (5.47) and (5.48) to compute $\mathbf{D}_n^*\mathbf{\Phi}_{n+1}$ and $\mathbf{D}_1^*\mathbf{\Phi}_{n+1}$, we also make use of the fact that diagonal matrices commute, so that we end up with

$$\begin{aligned}\mathbf{D}_n^*\mathbf{\Phi}_{n+1} &= (1-\alpha)\mathbf{I} + \Delta t\mathbf{\Lambda}((s-1)(1-\alpha\beta)+\beta_{20}), \\ \mathbf{D}_1^*\mathbf{\Phi}_{n+1} &= \alpha\mathbf{I} + \Delta t\mathbf{\Lambda}(-s(1-\alpha\beta)+\beta_{21}),\end{aligned}$$

where we have exploited $\beta_{20}+\beta_{21} = 1-\alpha\beta$. Finally, to compute $\mathbf{D}_{n+1}^*\mathbf{\Phi}_{n+1}$ we rewrite (5.44) utilizing $\operatorname{diag}(\mathbf{v})\mathbf{w} = \operatorname{diag}(\mathbf{w})\mathbf{v}$ to get

$$\begin{aligned}\mathbf{\Phi}_{n+1} =&(1-\alpha)\mathbf{y}^n + \alpha\mathbf{y}^{(1)} \\ &+ \Delta t\mathbf{\Lambda}\operatorname{diag}(\beta_{20}\mathbf{y}^n + \beta_{21}\mathbf{y}^{(1)})(\operatorname{diag}(\mathbf{y}^{(1)}))^{-s}(\operatorname{diag}(\mathbf{y}^n))^{s-1}\mathbf{y}^{n+1} - \mathbf{y}^{n+1}.\end{aligned} \tag{5.49}$$

From this, it is easy to see that

$$\mathbf{D}_{n+1}^*\mathbf{\Phi}_{n+1} = (1-\alpha\beta)\Delta t\mathbf{\Lambda} - \mathbf{I}$$

which is a nonsingular matrix since $\sigma(\mathbf{\Lambda}) \subseteq \overline{\mathbb{C}^-}$ and $1-\alpha\beta \geq \frac{1}{2\beta} > 0$, see (3.15). Finally, we introduce the expressions for the auxiliary Jacobians into the formula (5.45) resulting in

$$\begin{aligned}\mathbf{Dg}(\mathbf{y}^*) = (\mathbf{I} - (1-\alpha\beta)\Delta t\mathbf{\Lambda})^{-1}\Big(&(1-\alpha)\mathbf{I} + \Delta t\mathbf{\Lambda}((s-1)(1-\alpha\beta)+\beta_{20}) \\ &+ (\alpha\mathbf{I} + \Delta t\mathbf{\Lambda}(-s(1-\alpha\beta)+\beta_{21}))(\mathbf{I} - \beta\Delta t\mathbf{\Lambda})^{-1}\Big).\end{aligned}$$

Since $\mathbf{Dg}(\mathbf{y}^*)$ is a rational function of $\mathbf{\Lambda}$ and the identity matrix $\mathbf{I}$, we find $\sigma(\mathbf{Dg}(\mathbf{y}^*)) = \{R(\Delta t\lambda) \mid \lambda \in \sigma(\mathbf{\Lambda})\}$, where

$$R(z) = \frac{1-\alpha+z((s-1)(1-\alpha\beta)+\beta_{20}) + \frac{\alpha+z(-s(1-\alpha\beta)+\beta_{21})}{1-\beta z}}{1-(1-\alpha\beta)z}.$$

From

$$\beta_{20} = 1 - \frac{1}{2\beta} - \alpha\beta, \quad \beta_{21} = \frac{1}{2\beta} \quad \text{and} \quad s = \frac{\alpha\beta^2 - \alpha\beta + 1}{\beta(1-\alpha\beta)}$$

elementary computations lead to

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

In summary, we obtain the following proposition.

Proposition 5.21. Let $\mathbf{g}\colon \mathbb{R}^N_{>0}\to\mathbb{R}^N_{>0}$ be the map given by the application of the second order SSPMPRK family (SSPMPRK2) to the differential equation (5.1) with $\mathbf{1}\in\ker(\mathbf{\Lambda}^T)$. Then any $\mathbf{y}^*\in\ker(\mathbf{\Lambda})\cap\mathbb{R}^N_{>0}$ is a fixed point of $\mathbf{g}$ and $\mathbf{g}\in\mathcal{C}^2(\mathbb{R}^N_{>0},\mathbb{R}^N_{>0})$, whereby the first derivatives of $\mathbf{g}$ are Lipschitz continuous in an appropriate neighborhood of $\mathbf{y}^*$. Moreover, all linear invariants are conserved and an eigenvalue λ of $\mathbf{\Lambda}$ corresponds to the eigenvalue $R(\Delta t\lambda)$ of the Jacobian of $\mathbf{g}$ where

$$R(z)=\frac{-2+(2\alpha\beta^2-2\alpha\beta+1)z^2-2\beta(\alpha-1)z}{2(1+(\alpha\beta-1)z)(\beta z-1)}. \tag{5.50}$$

By this proposition, the SSPMPRK2(α,β) scheme satisfies all preconditions in order to apply Theorem 5.4. Thus, we have to analyze the stability function R.

Proposition 5.22. Let R be defined by (5.50) with α,β satisfying (3.15).

a) For any $\alpha>\frac{1}{2\beta}$, the set $\{z\in\overline{\mathbb{C}^-}\mid |R(z)|\leq 1\}$ is bounded.

b) For all $\alpha<\frac{1}{2\beta}$ with $(\alpha,\beta)\neq(0,\frac12)$ we have $|R(z)|<1$ for all $z\in\overline{\mathbb{C}^-}\setminus\{0\}$.

c) For $\alpha=\frac{1}{2\beta}$ or $(\alpha,\beta)=(0,\frac12)$ the relation $|R(z)|<1$ is true for all z with $\operatorname{Re}(z)<0$, and $|R(z)|=1$ holds whenever $\operatorname{Re}(z)=0$.

Proof. For proving part a), we consider (5.50) with $z=re^{\mathrm{i}\varphi}\in\overline{\mathbb{C}^-}\setminus\{0\}$ which yields

$$\lim_{r\to\infty}R(z)=\frac{2\alpha\beta^2-2\alpha\beta+1}{2\beta(\alpha\beta-1)}=\frac{2\alpha\beta^2-2\alpha\beta+1}{2\alpha\beta^2-2\beta}.$$

Note that for $\alpha=\frac{1}{2\beta}$, we obtain $\lim_{r\to\infty}R(z)=\frac{\beta-1+1}{\beta-2\beta}=-1$. Finally, it is straightforward to verify

$$\begin{aligned}\partial_\alpha\left(\lim_{r\to\infty}R(z)\right)&=\partial_\alpha\left(\frac{2\alpha\beta^2-2\alpha\beta+1}{2\beta(\alpha\beta-1)}\right)\\&=\frac{2\beta(\beta-1)2\beta(\alpha\beta-1)-(2\alpha\beta(\beta-1)+1)2\beta^2}{4\beta^2(\alpha\beta-1)^2}\\&=\frac{1-2\beta}{2(\alpha\beta-1)^2}<0,\end{aligned}$$

since $\beta\geq\frac12$. Therefore $\lim_{r\to\infty}R(z)$ decreases with increasing α. As a result, for any $\alpha>\frac{1}{2\beta}$, we find $\lim_{r\to\infty}R(z)<-1$ and thus, there exists $z^*\in\overline{\mathbb{C}^-}$ so that $|R(z^*)|>1$. Indeed, the set $\{z\in\overline{\mathbb{C}^-}\mid|R(z)|\leq1\}$ is bounded, as we find $|R(z^*)|>1$ for any $z^*\in\overline{\mathbb{C}^-}$ with $|z^*|$ large enough.

We now focus on the derivation of the remaining statements, we investigate $|R(z)|$ first on the imaginary axis. A technical but elementary computation for $z=\mathrm{i}b$, with $b\in\mathbb{R}$, yields

$$|R(\mathrm{i}b)|^2=\frac{1+b^4(\alpha\beta^2-\alpha\beta+\frac12)^2+b^2(1+(\alpha^2+1)\beta^2-2\alpha\beta)}{(1+(\alpha\beta-1)^2b^2)(\beta^2b^2+1)}.$$

Subtracting the denominator from the numerator leads to the expression

$$-(2\alpha\beta - 2\beta - 1)(2\alpha\beta - 1)(2\beta - 1)b^4. \tag{5.51}$$

With respect to statement b), we consider $\alpha < \frac{1}{2\beta}$ and $\beta > \frac{1}{2}$, as $\beta = \frac{1}{2}$ implies $\alpha = 0$ due to equation (3.15). It follows that $2\beta - 1 > 0$. Due to $\alpha < \frac{1}{2\beta}$, we see $2\alpha\beta < 1$ and $2\alpha\beta - 2\beta - 1 < 1 - 2\beta - 1 < 0$, so that the whole product (5.51) becomes negative, whenever $z = \mathrm{i}b \neq 0$. This is equivalent to $|R(z)| < 1$ on the imaginary axis without the origin. Using Remark 5.18, we see that $|R(z)| < 1$ holds for all $z \in \overline{\mathbb{C}^-} \setminus \{0\}$.

The assertion c) can be proved in a similar way using (5.51). Indeed, in the case of $\alpha = \frac{1}{2\beta}$ or $(\alpha, \beta) = (0, \frac{1}{2})$, the product (5.51) vanishes proving $|R(z)| = 1$ on the imaginary axis. Once again taking advantage of the Phragmén–Lindelöf principle one can conclude $|R(z)| < 1$ in $\mathbb{C}^-$. □

As a result we obtain the following corollaries that are a direct consequence of the application of Theorem 2.15 and Theorem 5.4, as well as Remark 5.1.

Corollary 5.23. Let $\mathbf{y}^*$ be a positive steady state of the differential equation (5.1) with $\mathbf{1} \in \ker(\mathbf{\Lambda}^T)$. Then $\mathbf{y}^*$ is a fixed point of the SSPMPRK2(α, β) scheme and the following holds:

a) For any $\alpha > \frac{1}{2\beta}$, the stability region of the SSPMPRK2(α, β) method is bounded.

b) For all $\alpha \leq \frac{1}{2\beta}$, the SPPMPRK22$(\alpha, \beta)$ scheme is unconditionally stable.

Corollary 5.24. Let the unique steady state $\mathbf{y}^*$ of the initial value problem (5.1), (5.2) be positive and $\mathbf{1} \in \ker(\mathbf{\Lambda}^T)$. Then there exists a $\delta > 0$ such that $\|\mathbf{y}^0 - \mathbf{y}^*\| < \delta$ implies the convergence of the iterates of the SSPMPRK2(α, β) scheme towards $\mathbf{y}^*$ for all $\Delta t > 0$, if $\alpha \leq \frac{1}{2\beta}$. For $\alpha > \frac{1}{2\beta}$, the method is conditionally stable.

In order to illustrate the consequences of Corollary 5.23 consider Figure 5.2, where due to (3.15) all permitted pairs of (α, β) with $\beta \leq 5$ lie between the β-axis and the black curve. The blue graph is determined by $\alpha = \frac{1}{2\beta}$, and thus, separates pairs of parameters associated with unconditionally stable methods from those with bounded stability domains. As an example, here we will consider the red rectangular with vertices $(0.2, 3)$, $(0.2, 3.5)$, $(0.24, 3)$ and $(0.24, 3.5)$, which is located in that critical region, so that we further analyze the corresponding choices of parameters with the help of Figure 5.3, where we plot the corresponding stability regions. One can observe that the chosen pairs of parameters from Figure 5.2 that are closer to the blue graph are associated with a larger stability domain. The smallest stability region among the examples from Figure 5.3 are associated with the (α, β) pair at the top right corner of the red rectangular from Figure 5.2.

SSPMPRK3(η_2) As the first step, we apply (SSPMPRK3) to the linear test problem (5.1), assuming conservativity, and rewrite it in the matrix-vector notation. For this, we again make use of equation (5.43) and the fact that the production and destruction terms are linear, which results in

$$\begin{aligned}
\mathbf{0} =& \mathbf{\Phi}_1(\mathbf{y}^n, \mathbf{y}^{(1)}) = \alpha_{10}\mathbf{y}^n + \beta_{10}\Delta t \mathbf{\Lambda} \mathbf{y}^{(1)} - \mathbf{y}^{(1)}, \\
\mathbf{0} =& \mathbf{\Phi}_{\boldsymbol{\rho}}(\mathbf{y}^n, \mathbf{y}^{(1)}, \boldsymbol{\rho}) = n_1 \mathbf{y}^{(1)} + n_2 (\operatorname{diag}(\mathbf{y}^{(1)}))^2 (\operatorname{diag}(\mathbf{y}^n))^{-1} \mathbf{1} - \boldsymbol{\rho},
\end{aligned}$$

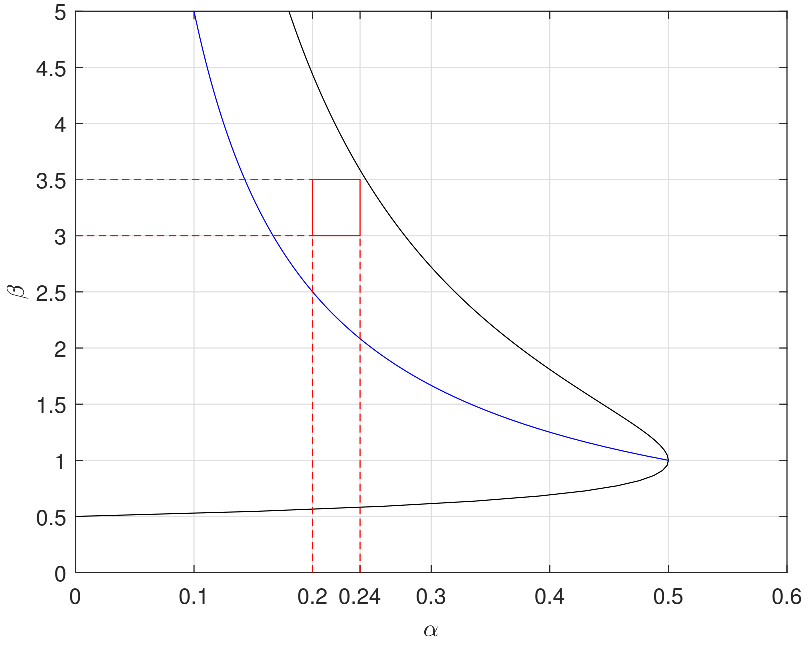


Figure 5.2: The black curve is implicitly given by the function $\alpha(\beta) = \frac{1-\frac{1}{2\beta}}{\beta}$. The blue graph is determined by the equation $\alpha = \frac{1}{2\beta}$ and the red rectangular possesses the vertices (α,β) with $(0.2, 3)$, $(0.2, 3.5)$, $(0.24, 3)$ and $(0.24, 3.5)$ which lie between the black and blue curve.

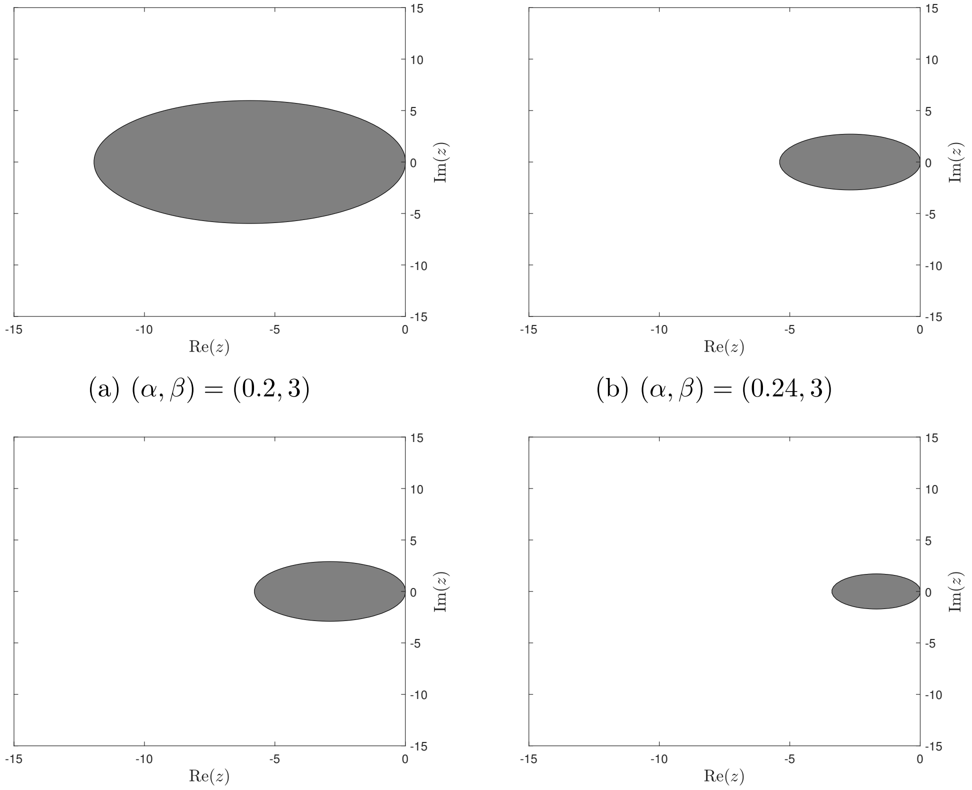
$$\begin{aligned}
\mathbf{0} =&\mathbf{\Phi}_2(\mathbf{y}^n,\mathbf{y}^{(1)},\boldsymbol{\rho},\mathbf{y}^{(2)}) = \alpha_{20}\mathbf{y}^n + \alpha_{21}\mathbf{y}^{(1)}\\
&+ \Delta t\mathbf{\Lambda}\operatorname{diag}(\mathbf{y}^{(2)})(\operatorname{diag}(\boldsymbol{\rho}))^{-1}(\beta_{20}\mathbf{y}^n + \beta_{21}\mathbf{y}^{(1)}) - \mathbf{y}^{(2)},\\
\mathbf{0} =&\mathbf{\Phi}_{\boldsymbol{\gamma}}(\mathbf{y}^n,\mathbf{y}^{(1)},\boldsymbol{\gamma}) = \eta_1\mathbf{y}^n + \eta_2\mathbf{y}^{(1)}\\
&+ \Delta t\mathbf{\Lambda}\operatorname{diag}(\boldsymbol{\gamma})(\operatorname{diag}(\mathbf{y}^n))^{s-1}(\operatorname{diag}(\mathbf{y}^{(1)}))^{-s}(\eta_3\mathbf{y}^n + \eta_4\mathbf{y}^{(1)}) - \boldsymbol{\gamma},\\
\mathbf{0} =&\mathbf{\Phi}_{\boldsymbol{\sigma}}(\mathbf{y}^n,\mathbf{y}^{(2)},\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{\sigma}) = \boldsymbol{\gamma} + \zeta(\operatorname{diag}(\mathbf{y}^{(2)}))(\operatorname{diag}(\boldsymbol{\rho}))^{-1}\mathbf{y}^n - \boldsymbol{\sigma},\\
\mathbf{0} =&\mathbf{\Phi}_{n+1}(\mathbf{y}^n,\mathbf{y}^{(1)},\boldsymbol{\rho},\mathbf{y}^{(2)},\mathbf{y}^{n+1}) = \alpha_{30}\mathbf{y}^n + \alpha_{31}\mathbf{y}^{(1)}\\
&+ \alpha_{32}\mathbf{y}^{(2)} + \Delta t\mathbf{\Lambda}\operatorname{diag}(\mathbf{y}^{n+1})(\operatorname{diag}(\boldsymbol{\sigma})^{-1}(\beta_{30}\mathbf{y}^n + \beta_{31}\mathbf{y}^{(1)} + \beta_{32}\mathbf{y}^{(2)}) - \mathbf{y}^{n+1},
\end{aligned} \tag{5.52}$$

where we omitted to write the arguments as functions of $\mathbf{y}^n$. Moreover, the parameter s is determined by [HZS19, Eq. (3.19)], also see [HIK$^+$22] for the details of the computation.

Now, we could introduce three stages $\mathbf{y}^{(3)}$, $\mathbf{y}^{(4)}$ and $\mathbf{y}^{(5)}$ for quantities $\boldsymbol{\rho}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\sigma}$ in order to keep the notation from (5.30) as we did for MPRK43 schemes. However, this might be more confusing at this point. We instead introduce auxiliary Jacobians $\mathbf{D}_{\boldsymbol{\sigma}}$ etc. in the same manner as for (5.30).

We want to point out that all functions from above are $\mathcal{C}^2$-maps for positive arguments. Thus, the map $\mathbf{g}$, which is determined by solving linear systems, is in $\mathcal{C}^2$. Due to Remark 5.7, the first derivatives are Lipschitz continuous for a sufficiently small neighborhood of $\mathbf{y}^*$.

Also, we can prove $\mathbf{g}(\mathbf{v}) = \mathbf{v}$ for all $\mathbf{v} \in \ker(\mathbf{\Lambda}) \cap \mathbb{R}^N_{>0}$ as follows. We know that $\mathbf{\Phi}_1(\mathbf{y}^*,\mathbf{y}^*) = \mathbf{0}$, and hence, $\mathbf{y}^n = \mathbf{y}^*$ implies $\mathbf{y}^{(1)} = \mathbf{y}^*$ as $\mathbf{y}^{(1)}$ is uniquely determined by $\mathbf{y}^n$. Analogously, we conclude $\boldsymbol{\rho}(\mathbf{y}^*) = \mathbf{y}^*$ as $n_1 + n_2 = 1$.



(a) $(\alpha, \beta) = (0.2, 3)$

(b) $(\alpha, \beta) = (0.24, 3)$

(c) $(\alpha, \beta) = (0.2, 3.5)$

(d) $(\alpha, \beta) = (0.24, 3.5)$

Figure 5.3: Different stability domains of the SSPMPRK2(α, β) method are plotted for (α, β) associated with the corners of the red rectangular from Figure 5.2.

As a consequence, we conclude from $a_{20} + a_{21} = 1$ at double precision that also $\mathbf{y}^{(2)}(\mathbf{y}^*) = \mathbf{y}^*$. However, $\boldsymbol{\gamma}(\mathbf{y}^*) = (\eta_1 + \eta_2)\mathbf{y}^*$, from which it follows that $\boldsymbol{\sigma}(\mathbf{y}^*) = (\eta_1+\eta_2)\mathbf{y}^*+\zeta\mathbf{y}^* = \mathbf{y}^*$ since $\eta_1+\eta_2 = 1-\zeta$. Finally $\mathbf{y}^{n+1}(\mathbf{y}^*) = \mathbf{g}(\mathbf{y}^*) = \mathbf{y}^*$ follows because $\sum_{i=0}^2 \alpha_{3i} = 1$ is true at double precision.

In the following we use $a_{20} + a_{21} = 1$, $\sum_{i=0}^2 \alpha_{3i} = 1$ and $\alpha_{10} = 1$ as well as the values of the functions evaluated at $\mathbf{y}^*$ without further notice.

Moreover, we can observe that $\mathbf{g}$ conserves all linear invariants as follows. First, $\mathbf{n}^T\boldsymbol{\Lambda} = \mathbf{0}$ implies

$$\mathbf{n}^T\mathbf{y}^{(1)} = \alpha_{10}\mathbf{n}^T\mathbf{y}^n + \beta_{10}\Delta t\mathbf{n}^T\boldsymbol{\Lambda}\mathbf{y}^{(1)} = \mathbf{n}^T\mathbf{y}^n.$$

As a consequence, we obtain

$$\mathbf{n}^T\mathbf{y}^{(2)} = \alpha_{20}\mathbf{n}^T\mathbf{y}^n + \alpha_{21}\mathbf{n}^T\mathbf{y}^{(1)} + \mathbf{0} = (\alpha_{20} + \alpha_{21})\mathbf{n}^T\mathbf{y}^n = \mathbf{n}^T\mathbf{y}^n.$$

Altogether, we find that $\mathbf{g}$ is linear invariants preserving due to

$$\mathbf{n}^T\mathbf{g}(\mathbf{y}^n) = \mathbf{n}^T\mathbf{y}^{n+1} = \sum_{i=0}^{2} \alpha_{3i}\mathbf{n}^T\mathbf{y}^n + \mathbf{0} = \mathbf{n}^T\mathbf{y}^n.$$

Hence, also in the third order case, the map $\mathbf{g}$ satisfies all conditions for applying Theorem 2.15 and Theorem 5.4. Therefore, we are now interested in computing the Jacobian of $\mathbf{g}$, which can be done by using the same techniques as for the second order SSPMPRK scheme. Since we use a slightly different notation, let us recall the formula for $\mathbf{Dg}(\mathbf{y}^*)$. Using the chain rule for the last equation of (5.52) and solving for $\mathbf{Dg}(\mathbf{y}^*)$ yield

$$\begin{aligned}\mathbf{Dg}(\mathbf{y}^*) = -(\mathbf{D}^*_{n+1}\mathbf{\Phi}_{n+1})^{-1}(\mathbf{D}^*_n\mathbf{\Phi}_{n+1} + \mathbf{D}^*_1\mathbf{\Phi}_{n+1}\mathbf{D}^*\mathbf{y}^{(1)} + \mathbf{D}^*_2\mathbf{\Phi}_{n+1}\mathbf{D}^*\mathbf{y}^{(2)}\\ + \mathbf{D}^*_\sigma\mathbf{\Phi}_{n+1}\mathbf{D}^*\boldsymbol{\sigma}),\end{aligned} \tag{5.53}$$

if $(\mathbf{D}^*_{n+1}\mathbf{\Phi}_{n+1})^{-1}$ exists. Hence, we need formulae for $\mathbf{D}^*\mathbf{y}^{(1)}, \mathbf{D}^*\mathbf{y}^{(2)}$ and $\mathbf{D}^*\boldsymbol{\sigma}$. We use the same strategies as for the second order scheme and obtain by means of the chain rule of the corresponding equation in (5.52) the formulae

$$\begin{aligned}\mathbf{D}^*\mathbf{y}^{(1)} &= -(\mathbf{D}^*_1\mathbf{\Phi}_1)^{-1}\mathbf{D}^*_n\mathbf{\Phi}_1,\\ \mathbf{D}^*\mathbf{y}^{(2)} &= -(\mathbf{D}^*_2\mathbf{\Phi}_2)^{-1}(\mathbf{D}^*_n\mathbf{\Phi}_2 + \mathbf{D}^*_1\mathbf{\Phi}_2\mathbf{D}^*\mathbf{y}^{(1)} + \mathbf{D}^*_\rho\mathbf{\Phi}_2\mathbf{D}^*\boldsymbol{\rho}),\\ \mathbf{D}^*\boldsymbol{\sigma} &= -(\mathbf{D}^*_\sigma\mathbf{\Phi}_{\boldsymbol{\sigma}})^{-1}(\mathbf{D}^*_n\mathbf{\Phi}_{\boldsymbol{\sigma}} + \mathbf{D}^*_2\mathbf{\Phi}_{\boldsymbol{\sigma}}\mathbf{D}^*\mathbf{y}^{(2)} + \mathbf{D}^*_\rho\mathbf{\Phi}_{\boldsymbol{\sigma}}\mathbf{D}^*\boldsymbol{\rho} + \mathbf{D}^*_{\boldsymbol{\gamma}}\mathbf{\Phi}_{\boldsymbol{\sigma}}\mathbf{D}^*\boldsymbol{\gamma}),\end{aligned} \tag{5.54}$$

provided that the inverses exist. However, to compute the last two Jacobians, we now require to have knowledge about $\mathbf{D}^*\boldsymbol{\rho}$ and $\mathbf{D}^*\boldsymbol{\gamma}$. These Jacobians can be obtained by

$$\begin{aligned}\mathbf{D}^*\boldsymbol{\rho} &= -(\mathbf{D}^*_{\boldsymbol{\rho}}\Phi_{\boldsymbol{\rho}})^{-1}(\mathbf{D}^*_n\mathbf{\Phi}_{\boldsymbol{\rho}} + \mathbf{D}^*_1\mathbf{\Phi}_{\boldsymbol{\rho}}\mathbf{D}^*\mathbf{y}^{(1)}),\\ \mathbf{D}^*\boldsymbol{\gamma} &= -(\mathbf{D}^*_{\boldsymbol{\gamma}}\mathbf{\Phi}_{\boldsymbol{\gamma}})^{-1}(\mathbf{D}^*_n\mathbf{\Phi}_{\boldsymbol{\gamma}} + \mathbf{D}^*_1\mathbf{\Phi}_{\boldsymbol{\gamma}}\mathbf{D}^*\mathbf{y}^{(1)}),\end{aligned} \tag{5.55}$$

if the expressions are defined. Starting off with the calculation of $\mathbf{D}^*\mathbf{y}^{(1)}$, we obtain

$$\mathbf{D}^*_n\mathbf{\Phi}_1 = \alpha_{10}\mathbf{I}, \quad \mathbf{D}^*_1\mathbf{\Phi}_1 = \beta_{10}\Delta t\mathbf{\Lambda} - \mathbf{I}.$$

Since $\beta_{10} > 0$ we can use (5.54) to conclude that

$$\mathbf{D}^*\mathbf{y}^{(1)} = -(\beta_{10}\Delta t\mathbf{\Lambda} - \mathbf{I})^{-1} \cdot \alpha_{10}\mathbf{I} = (\mathbf{I} - \beta_{10}\Delta t\mathbf{\Lambda})^{-1}$$

is defined. Next we focus on $\mathbf{D}^*\boldsymbol{\rho}$ so that we can compute $\mathbf{D}^*\mathbf{y}^{(2)}$ afterwards. For this, we use again that diagonal matrices commute and that $\operatorname{diag}(\mathbf{v})\mathbf{w} = \operatorname{diag}(\mathbf{w})\mathbf{v}$ holds. Hence, we find

$$\mathbf{D}^*_n\mathbf{\Phi}_{\boldsymbol{\rho}} = -n_2\mathbf{I}, \quad \mathbf{D}^*_1\mathbf{\Phi}_{\boldsymbol{\rho}} = (n_1 + 2n_2)\mathbf{I}, \quad \mathbf{D}^*_{\boldsymbol{\rho}}\Phi_{\boldsymbol{\rho}} = -\mathbf{I},$$

and due to (5.55),

$$\mathbf{D}^*\boldsymbol{\rho} = -n_2\mathbf{I} + (n_1 + 2n_2)(\mathbf{I} - \beta_{10}\Delta t\mathbf{\Lambda})^{-1}.$$

The computation of the following Jacobians requires the same technique as described in equations (5.47) and (5.49), from which we get

$$\begin{aligned}\mathbf{D}^*_n\mathbf{\Phi}_2 &= \alpha_{20}\mathbf{I} + \beta_{20}\Delta t\mathbf{\Lambda}, & \mathbf{D}^*_1\mathbf{\Phi}_2 &= \alpha_{21}\mathbf{I} + \beta_{21}\Delta t\mathbf{\Lambda},\\ \mathbf{D}^*_\rho\mathbf{\Phi}_2 &= -(\beta_{20} + \beta_{21})\Delta t\mathbf{\Lambda}, & \mathbf{D}^*_2\mathbf{\Phi}_2 &= (\beta_{20} + \beta_{21})\Delta t\mathbf{\Lambda} - \mathbf{I},\end{aligned}$$

respectively. Since $\beta_{20} + \beta_{21} > 0$ the inverse of $\mathbf{D}^*_2\mathbf{\Phi}_2$ exists, and thus, $\mathbf{D}^*\mathbf{y}^{(2)}$ is formally given by (5.54).

Next, we need $\mathbf{D}^*\boldsymbol{\gamma}$ in order to find $\mathbf{D}^*\boldsymbol{\sigma}$. Exploiting once again the ideas from (5.47) and (5.49), we obtain with $\boldsymbol{\gamma}(\mathbf{y}^*) = (\eta_1 + \eta_2)\mathbf{y}^*$ the Jacobians

$$\mathbf{D}^*_n\mathbf{\Phi}_{\boldsymbol{\gamma}} = \eta_1\mathbf{I} + (\eta_1 + \eta_2)\Delta t\mathbf{\Lambda}((s-1)(\eta_3 + \eta_4) + \eta_3),$$

$$\begin{aligned}
\mathbf{D}_1^*\boldsymbol{\Phi}_{\boldsymbol{\gamma}} &= \eta_2\mathbf{I} + (\eta_1+\eta_2)\Delta t\boldsymbol{\Lambda}(-s(\eta_3+\eta_4)+\eta_4),\\
\mathbf{D}_{\boldsymbol{\gamma}}^*\boldsymbol{\Phi}_{\boldsymbol{\gamma}} &= (\eta_3+\eta_4)\Delta t\boldsymbol{\Lambda} - \mathbf{I},
\end{aligned}$$

where $\mathbf{D}_{\boldsymbol{\gamma}}^*\boldsymbol{\Phi}_{\boldsymbol{\gamma}}$ is nonsingular since $\eta_3+\eta_4>0$. Hence, with (5.55) even the Jacobian $\mathbf{D}^*\boldsymbol{\gamma}$ can be determined.

Computing

$$\mathbf{D}_n^*\boldsymbol{\Phi}_{\boldsymbol{\sigma}} = \zeta\mathbf{I},\quad \mathbf{D}_2^*\boldsymbol{\Phi}_{\boldsymbol{\sigma}} = \zeta\mathbf{I},\quad \mathbf{D}_{\rho}^*\boldsymbol{\Phi}_{\boldsymbol{\sigma}} = -\zeta\mathbf{I},\quad \mathbf{D}_{\boldsymbol{\gamma}}^*\boldsymbol{\Phi}_{\boldsymbol{\sigma}} = \mathbf{I},\quad \mathbf{D}_{\boldsymbol{\sigma}}^*\boldsymbol{\Phi}_{\boldsymbol{\sigma}} = -\mathbf{I},$$

we are able to obtain $\mathbf{D}^*\boldsymbol{\sigma}$ from (5.54). Finally, the remaining Jacobians are given by

$$\begin{aligned}
\mathbf{D}_n^*\boldsymbol{\Phi}_{n+1} &= \alpha_{30}\mathbf{I} + \beta_{30}\Delta t\boldsymbol{\Lambda}, \quad \mathbf{D}_1^*\boldsymbol{\Phi}_{n+1} = \alpha_{31}\mathbf{I} + \beta_{31}\Delta t\boldsymbol{\Lambda},\\
\mathbf{D}_2^*\boldsymbol{\Phi}_{n+1} &= \alpha_{32}\mathbf{I} + \beta_{32}\Delta t\boldsymbol{\Lambda}, \quad \mathbf{D}_{\boldsymbol{\sigma}}^*\boldsymbol{\Phi}_{n+1} = -\Delta t\boldsymbol{\Lambda}\sum_{i=0}^{2}\beta_{3i},\\
\mathbf{D}_{n+1}^*\boldsymbol{\Phi}_{n+1} &= \Delta t\boldsymbol{\Lambda}\sum_{i=0}^{2}\beta_{3i} - \mathbf{I}
\end{aligned}$$

with $\sum_{i=0}^2\beta_{3i}>0$, so that we are now in the position to compute $\mathbf{Dg}(\mathbf{y}^*)$ using (5.53). As all the matrices occurring within the expressions of the Jacobians above are either the identity matrix $\mathbf{I}$ or the system matrix $\boldsymbol{\Lambda}$ from (5.1), the stability function for the third order SSPMPRK scheme can easily be computed by calculating $\mathbf{Dg}(\mathbf{y}^*)$ and substituting $\Delta t\boldsymbol{\Lambda}$ by $\Delta t\lambda = z$, so that we end up with the stability function $R(\Delta t\lambda) = R(z)$ that reads

$$\begin{aligned}
R(z) =& \frac{1}{1-z\sum_{i=0}^2\beta_{3i}}\Bigg[\alpha_{30}+\beta_{30}z+\frac{\alpha_{31}+\beta_{31}z}{1-\beta_{10}z}+(\alpha_{32}+\beta_{32}z)P(z)\\
&-z\sum_{i=0}^{2}\beta_{3i}\Bigg(\zeta+\zeta P(z)-\zeta\left(\frac{n_1+2n_2}{1-\beta_{10}z}-n_2\right)\\
&+\frac{1}{1-(\eta_3+\eta_4)z}\Bigg(\eta_1+(\eta_1+\eta_2)z\Big((s-1)(\eta_3+\eta_4)+\eta_3\Big) \qquad (5.56)\\
&+\frac{\eta_2+(\eta_1+\eta_2)z\big(-s(\eta_3+\eta_4)+\eta_4\big)}{1-\beta_{10}z}\Bigg)\Bigg)\Bigg],\\
P(z) =& \frac{\alpha_{20}+\beta_{20}z+\frac{\alpha_{21}+\beta_{21}z}{1-\beta_{10}z}-(\beta_{20}+\beta_{21})z\left(\frac{n_1+2n_2}{1-\beta_{10}z}-n_2\right)}{1-(\beta_{20}+\beta_{21})z}.
\end{aligned}$$

Before a detailed investigation of the stability function R, we summarize the above derived results by means of the following proposition.

Proposition 5.25. Let $\mathbf{g}\colon \mathbb{R}^N_{>0}\to\mathbb{R}^N_{>0}$ be the generating map of SSPMPRK3(η_2) when applied to the differential equation (5.1) with $\mathbf{1}\in\ker(\boldsymbol{\Lambda}^T)$. Then any $\mathbf{y}^*\in\ker(\boldsymbol{\Lambda})\cap\mathbb{R}^N_{>0}$ is a fixed point of $\mathbf{g}\in\mathcal{C}^2(\mathbb{R}^N_{>0},\mathbb{R}^N_{>0})$, whereby the first derivatives of $\mathbf{g}$ are Lipschitz continuous in an appropriate neighborhood of $\mathbf{y}^*$. Moreover, all linear invariants are conserved and an eigenvalue λ of $\boldsymbol{\Lambda}$ corresponds to the eigenvalue $R(\Delta t\lambda)$ of the Jacobian of $\mathbf{g}$ where R is defined in (5.56) and the parameters are given in (3.18).

Next, we will prove that the third order SSPMPRK scheme possesses stable fixed points for all $\eta_2\in[0,r_1]$ when applied to the test equation.

Proposition 5.26. The stability function $R(z)$ of the third order SSPMPRK scheme satisfies $R(0) = 1$ and $|R(z)| < 1$ for all $z \in \overline{\mathbb{C}^-} \setminus \{0\}$ up to double precision.

Proof. It is straightforward to see that $R(0) = \alpha_{30} + \alpha_{31} + \alpha_{32}(\alpha_{20} + \alpha_{21})$ holds true. Up to double precision, we obtain $\alpha_{20} + \alpha_{21} = 1$ and $\alpha_{30} + \alpha_{31} + \alpha_{32} = 1$, so that $R(0) = 1$. Also, as $\alpha_{ij}, \beta_{ij}, \eta_3 + \eta_4 > 0$, see (3.18), no poles of R are located in $\overline{\mathbb{C}^-}$. Furthermore, by a technical calculation we can rewrite R to receive

$$R(z) = \frac{\sum_{j=0}^{4} n_j z^j}{\sum_{j=0}^{4} d_j z^j},$$

where, for $\eta_2 \in [0, r_1] \subseteq [0, \frac{1}{2})$ the coefficients are given by

$$\begin{aligned}
n_0 =& \frac{0.47620819268131705757\eta_2 - 1.0537480911094115481}{0.47620819268131703\eta_2 - 1.0537480911094114871},\\
n_1 =& \frac{-3.1507612671062001337\eta_2 + 3.9798736646158920698}{0.47620819268131703\eta_2 - 1.0537480911094114871}\\
&+ \frac{0.61107641837494959323\eta_2^2}{0.47620819268131703\eta_2 - 1.0537480911094114871},\\
n_2 =& \frac{2.4343280828365809236\eta_2 - 2.5818776483048969774}{0.47620819268131703\eta_2 - 1.0537480911094114871}\\
&+ \frac{-0.57282016379130601724\eta_2^2}{0.47620819268131703\eta_2 - 1.0537480911094114871},\\
n_3 =& \frac{0.65368695844177787153\eta_2 - 0.81355615989342266462}{0.47620819268131703\eta_2 - 1.0537480911094114871}\\
&+ \frac{-0.12926039115580354457\eta_2^2}{0.47620819268131703\eta_2 - 1.0537480911094114871},\\
n_4 =& \frac{-0.59499575916146815582\eta_2 + 0.63887056198975790458}{0.47620819268131703\eta_2 - 1.0537480911094114871}\\
&+ \frac{0.13841284380675759\,36\eta_2^2}{0.47620819268131703\eta_2 - 1.0537480911094114871},\\
d_0 =&1,\\
d_1 =& - 4.7768739020212929733 + 1.2832127371313151768\eta_2,\\
d_2 =&6.7270587897458664634 - 2.4860903284764154151\eta_2,\\
d_3 =& - 3.7332290665687486456 + 1.5730472371819288192\eta_2,\\
d_4 =&0.71670702950202557445 - 0.32389312216150656420\eta_2,
\end{aligned}$$

where $n_0 = 1$ at double precision, see [HIK+22]. We want to mention here, that these values were computed with *Maple 2021* and Digits $= 20$, which means that 20 digits were used when making calculations with software floating-point numbers.

We investigate the polynomial $p_{\frac{\pi}{2}}(r)$ from Lemma A.1 with η_2 being a parameter. At double precision, we obtain $n_0 = 1$, so that $n_0^2 - 1 = 0$, i. e. $p_{\frac{\pi}{2}}(0) = 0$. Next, our strategy is to prove that all nonzero coefficients of r^k, in the following denoted by c_k are negative.

For $\eta_2 \leq r_1 < \frac{1}{2}$, it suffices for our argument to round to three decimal places in the following expressions, which can be reproduced using the Maple repository

[HIK+22] and read

$$
\begin{aligned}
c_8 &\approx \frac{4.410(-0.168\eta_2^2 + 0.271\eta_2 + 0.046\eta_2^3 - 0.162 - 0.005\eta_2^4)}{(\eta_2 - 2.213)^2},\\
c_6 &\approx \frac{4.410(-0.790\eta_2^2 + 1.310\eta_2 + 0.210\eta_2^3 - 0.808 - 0.021\eta_2^4)}{(\eta_2 - 2.213)^2},\\
c_4 &\approx \frac{4.410(0.556\eta_2 - 0.442 + 0.032\eta_2^3 - 0.232\eta_2^2)}{(\eta_2 - 2.231)^2},\\
10^{14}c_2 &\approx \frac{\eta_2(\eta_2 - 1 - 0.2\eta_2^2 + 0.03\eta_2^3)}{(0.476\eta_2 - 1.054)^2}.
\end{aligned}
$$

First of all, the denominators occurring in any of the above c_k are positive. Also, positive terms in the numerator are multiplied with powers of $\eta_2 < \frac{1}{2}$ and thus are smaller than the absolute value of the corresponding constant, which is always negative. This holds true even if the rounding error is taken into account, i. e. after adding 10^{-2} to positive terms and subtracting it from negative expressions. This proves that $c_k < 0$, and thus, $|R(\mathrm{i}y)| < 1$ for all $y \in \mathbb{R} \setminus \{0\}$.

Finally, we can conclude even $|R(z)| < 1$ for all $z \in \overline{\mathbb{C}^-} \setminus \{0\}$ by means of Remark 5.18. □

As an immediate consequence of this proposition in combination with Theorem 2.15 and Theorem 5.4, we obtain the following results.

Corollary 5.27. The SSPMPRK3(η_2) scheme is unconditionally stable for all $\eta_2 \in [0, r_1]$, where $r_1 \approx 0.37$.

Corollary 5.28. Let $\mathbf{y}^*$ be the unique steady state of the initial value problem (5.1), (5.2) with $\mathbf{1} \in \ker(\mathbf{\Lambda}^T)$. Then there exists a $\delta > 0$ such that $\|\mathbf{y}^0 - \mathbf{y}^*\| < \delta$ implies the convergence of the iterates of of SSPMPRK3(η_2) towards $\mathbf{y}^*$ for all $\Delta t > 0$ and $\eta_2 \in [0, r_1]$.

5.4.3 Modified Patankar Deferred Correction

In this subsection we investigate (MPDeC). Since the index function γ depends on the sign of θ_r^m, we introduce the nonnegative part $\theta_{m,+} = \max\{0, \theta_r^m\}$ and nonpositive part $\theta_{m,-} = \min\{0, \theta_r^m\}$. It is worth mentioning that

$$\theta_{r,\pm}^m = \frac{\theta_r^m \pm |\theta_r^m|}{2}$$

and

$$\theta_r^m = \begin{cases} \theta_{r,-}^m, & \theta_r^m < 0, \\ \theta_{r,+}^m, & \theta_r^m \geq 0 \end{cases}$$

as well as $\theta_{r,-}^m + \theta_{r,+}^m = \theta_r^m$. With that, we split the sum appearing in (MPDeC) into two sums containing $\theta_{r,+}^m$ and $\theta_{r,-}^m$, respectively. For the separated sums, we know the value of $\gamma(j,i,\theta_r^m)$ so that we introduce the positive part

$$\mathbf{p}^{r,(k)}(\mathbf{y}^{r,(k-1)}, \mathbf{y}^{m,(k-1)}, \mathbf{y}^{m,(k)}) = \mathbf{\Lambda} \operatorname{diag}(\mathbf{y}^{m,(k)}) \left(\operatorname{diag}(\mathbf{y}^{m,(k-1)})\right)^{-1} \mathbf{y}^{r,(k-1)} \tag{5.57}$$

analogously as we did for SSPMPRK, as well as the negative part $\mathbf{n}^{r,(k)}$ given by

$$\begin{aligned} n_i^{r,(k)}(\mathbf{y}^{r,(k-1)}, \mathbf{y}^{m,(k-1)}, \mathbf{y}^{m,(k)}) = \sum_{j=1}^N \Bigg(& p_{ij}(\mathbf{y}^{r,(k-1)}) \frac{y_i^{m,(k)}}{y_i^{m,(k-1)}} \\ & - d_{ij}(\mathbf{y}^{r,(k-1)}) \frac{y_j^{m,(k)}}{y_j^{m,(k-1)}} \Bigg) \end{aligned} \tag{5.58}$$

for $i = 1, \dots, N$, $r = 0, \dots, M$ and $k = 1, \dots, K$. Using $p_{ij}(\mathbf{y}) = d_{ji}(\mathbf{y}) = \lambda_{ij} y_j$ for $i \neq j$ and $p_{ii}(\mathbf{y}) = d_{ii}(\mathbf{y}) = 0$ this can be rewritten as

$$n_i^{r,(k)}(\mathbf{y}^{r,(k-1)}, \mathbf{y}^{m,(k-1)}, \mathbf{y}^{m,(k)}) = \frac{y_i^{m,(k)}}{y_i^{m,(k-1)}} \sum_{\substack{j=1 \\ j\neq i}}^N \lambda_{ij} y_j^{r,(k-1)} - y_i^{r,(k-1)} \sum_{\substack{j=1 \\ j\neq i}}^N \lambda_{ji} \frac{y_j^{m,(k)}}{y_j^{m,(k-1)}}. \tag{5.59}$$

Utilizing these vector fields, the iterates from (MPDeC) satisfy

$$\begin{aligned} \mathbf{0} &= \mathbf{\Phi}_k^m(\mathbf{y}^n, \mathbf{y}^{1,(k-1)}, \dots, \mathbf{y}^{M,(k-1)}, \mathbf{y}^{m,(k)}) \\ \mathbf{\Phi}_k^m &= \mathbf{y}^{m,(k)} - \mathbf{y}^n - \sum_{r=0}^M \theta_{r,+}^m \Delta t \mathbf{p}^{r,(k)}(\mathbf{y}^{r,(k-1)}, \mathbf{y}^{m,(k-1)}, \mathbf{y}^{m,(k)}) \\ &\quad - \sum_{r=0}^M \theta_{r,-}^m \Delta t \mathbf{n}^{r,(k)}(\mathbf{y}^{r,(k-1)}, \mathbf{y}^{m,(k-1)}, \mathbf{y}^{m,(k)}) \end{aligned} \tag{5.60}$$

for $k = 1, \dots, K$ and $m = 1, \dots, M$. Furthermore, analogously to the auxiliary Jacobians introduced in (5.28) and (5.29), we write $\mathbf{D}_x^* \mathbf{\Phi}_k^m$ to represent the Jacobian with respect to the entries of the vector $\mathbf{y}^x$ for some x, evaluated at $(\mathbf{y}^*, \dots, \mathbf{y}^{m,(k)}(\mathbf{y}^*))$. Finally, we introduce similar notations for the auxiliary Jacobians of $\mathbf{p}^{r,(k)}$ and $\mathbf{n}^{r,(k)}$ with respect to $\mathbf{y}^x$.

Also note that MPDeC schemes are steady state preserving as plugging in $\mathbf{y}^{m,(k)} = \mathbf{y}^n = \mathbf{y}^* \in \ker(\mathbf{\Lambda})$ into (MPDeC) yields a true statement. Hence, $\mathbf{y}^{m,(k)}(\mathbf{y}^*) = \mathbf{y}^*$ for all $k = 1, \dots, K$ and $m = 1, \dots, M$.

The next theorem summarizes further properties of the method and provides us a formula for the computation of $\mathbf{Dg}(\mathbf{y}^*)$.

Theorem 5.29. Let $\mathbf{g}:\mathbb{R}^N_{>0}\to\mathbb{R}^N_{>0}$, implicitly given by the solution of (5.60), be the generating map of the MPDeC iterates when applied to (5.1) with $\mathbf{1}\in\ker(\mathbf{\Lambda}^T)$. Furthermore, let $\mathbf{y}^*\in\ker(\mathbf{\Lambda})\cap\mathbb{R}^N_{>0}$ be a steady state of (5.1).

Then, $\mathbf{g}\in\mathcal{C}^2$ and the Jacobian of $\mathbf{g}$ evaluated at $\mathbf{y}^*$ is given by

$$\mathbf{Dg}(\mathbf{y}^*)=\mathbf{D}^*_n\mathbf{y}^{M,(K)},$$

$$\mathbf{D}^*_n\mathbf{y}^{m,(k)}=-(\mathbf{D}^*_{m,(k)}\mathbf{\Phi}^m_k)^{-1}\left(\mathbf{D}^*_n\mathbf{\Phi}^m_k+(1-\delta_{k1})\sum_{r=1}^M\mathbf{D}^*_{r,(k-1)}\mathbf{\Phi}^m_k\mathbf{D}^*_n\mathbf{y}^{r,(k-1)}\right) \tag{5.61}$$

for $m=1,\dots,M$ and $k=1,\dots,K$. Thereby, δ_{ij} is the Kronecker delta and

$$\mathbf{D}^*_n\mathbf{\Phi}^m_k=\begin{cases}-\left(\mathbf{I}+\Delta t(\mathbf{\Lambda}+\operatorname{diag}(\mathbf{y}^*)\mathbf{\Lambda}^T(\operatorname{diag}(\mathbf{y}^*))^{-1})\sum_{r=0}^M\theta^m_{r,-}\right), & k=1,\\ -(\mathbf{I}+\theta^m_0\Delta t\mathbf{\Lambda}), & k>1,\end{cases} \tag{5.62}$$

as well as

$$\mathbf{D}^*_{l,(s)}\mathbf{\Phi}^m_k=\begin{cases}-\theta^m_l\Delta t\mathbf{\Lambda},\\ \sum_{r=0}^M\theta^m_{r,+}\Delta t\mathbf{\Lambda}-\sum_{r=0}^M\theta^m_{r,-}\Delta t(\operatorname{diag}(\mathbf{y}^*)\mathbf{\Lambda}^T(\operatorname{diag}(\mathbf{y}^*))^{-1})-\theta^m_m\Delta t\mathbf{\Lambda},\\ \mathbf{I}-\sum_{r=0}^M\theta^m_{r,+}\Delta t\mathbf{\Lambda}+\sum_{r=0}^M\theta^m_{r,-}\Delta t\operatorname{diag}(\mathbf{y}^*)\mathbf{\Lambda}^T(\operatorname{diag}(\mathbf{y}^*))^{-1},.\end{cases} \tag{5.63}$$

for

$$s=\begin{cases}k-1>0,\ 0<l\neq m,\\ k-1>0,\ l=m,\\ s=k,\ l=m,\end{cases}$$

respectively.

Proof. Since the θ^m_r are fixed for a given scheme, the functions $\mathbf{\Phi}^m_k$ are in $\mathcal{C}^2$ and as a consequence of solving only linear systems, the map $\mathbf{g}$ is also in $\mathcal{C}^2$. Furthermore, the formula (5.61) follows analogously to (5.30), whereby we want to point out that the sum appearing in (5.61) is multiplied with 0 for $k=1$ since $\mathbf{y}^{r,(k-1)}=\mathbf{y}^n$ in this case. Hence, we only have to prove the formulae (5.62) and (5.63). For this, we compute the Jacobians of each addend of the sums in (5.60) separately by considering (5.57) and (5.59).

Let us start proving (5.62), first considering $k=1$. From (5.57) and $\mathbf{y}^{s,(0)}=\mathbf{y}^n$ for all $s=0,\dots,M$ it follows that

$$\mathbf{p}^{r,(1)}(\mathbf{y}^n,\mathbf{y}^n,\mathbf{y}^{m,(1)})=\mathbf{p}^{r,(1)}(\mathbf{y}^n,\mathbf{y}^{m,(1)})=\mathbf{\Lambda}\mathbf{y}^{m,(1)}$$

and hence, $\mathbf{D}^*_n\mathbf{p}^{r,(1)}=\mathbf{0}$. Moreover, (5.59) for $k=1$ yields

$$n^{r,(1)}_i(\mathbf{y}^n,\mathbf{y}^n,\mathbf{y}^{m,(1)})=n^{r,(1)}_i(\mathbf{y}^n,\mathbf{y}^{m,(1)})=\frac{y^{m,(1)}_i}{y^n_i}\sum_{\substack{j=1\\j\neq i}}^N\lambda_{ij}y^n_j-y^n_i\sum_{\substack{j=1\\j\neq i}}^N\lambda_{ji}\frac{y^{m,(1)}_j}{y^n_j}.$$

Hence, using $\mathbf{1} \in \ker(\mathbf{\Lambda}^T)$, we obtain

$$\begin{aligned} \frac{\partial}{\partial y_i^n} n_i^{r,(1)}(\mathbf{y}^*, \mathbf{y}^*) &= -\frac{1}{y_i^*} \sum_{\substack{j=1 \\ j \neq i}}^N \lambda_{ij} y_j^* - \sum_{\substack{j=1 \\ j \neq i}}^N \lambda_{ji} = \frac{1}{y_i^*} \left(- \sum_{\substack{j=1 \\ j \neq i}}^N \lambda_{ij} y_j^* + \lambda_{ii} y_i^* \right) \\ &= \frac{1}{y_i^*} \left(\underbrace{- \sum_{j=1}^N \lambda_{ij} y_j^*}_{(\mathbf{\Lambda y}^*)_i = 0} + 2\lambda_{ii} y_i^* \right) = 2\lambda_{ii}, \end{aligned} \tag{5.64}$$

and for $q \neq i$ we find

$$\frac{\partial}{\partial y_q^n} n_i^{r,(1)}(\mathbf{y}^*, \mathbf{y}^*) = \lambda_{iq} + \lambda_{qi} \frac{y_i^*}{y_q^*}.$$

Altogether, we obtain

$$\mathbf{D}_n^* \mathbf{n}^{r,(1)} = \mathbf{\Lambda} + \begin{pmatrix} \lambda_{11} & \lambda_{21} \frac{y_1^*}{y_2^*} & \dots & \lambda_{N1} \frac{y_1^*}{y_N^*} \\ \lambda_{12} \frac{y_2^*}{y_1^*} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \lambda_{1N} \frac{y_N^*}{y_1^*} & \dots & \dots & \lambda_{NN} \end{pmatrix} = \mathbf{\Lambda} + \operatorname{diag}(\mathbf{y}^*) \mathbf{\Lambda}^T (\operatorname{diag}(\mathbf{y}^*))^{-1}, \tag{5.65}$$

and thus,

$$\mathbf{D}_n^* \mathbf{\Phi}_1^m = - \left(\mathbf{I} + (\mathbf{\Lambda} + \operatorname{diag}(\mathbf{y}^*) \mathbf{\Lambda}^T (\operatorname{diag}(\mathbf{y}^*))^{-1}) \Delta t \sum_{r=0}^M \theta_{r,-}^m \right).$$

Next, for $k > 1$ it follows from (5.57) that

$$\mathbf{D}_n^* \mathbf{p}^{r,(k)} = \delta_{r0} \mathbf{\Lambda}.$$

Similarly, $\mathbf{D}_n^* \mathbf{n}^{r,(k)} = \mathbf{0}$ if $r \neq 0$. Furthermore,

$$n_i^{0,(k)}(\mathbf{y}^n, \mathbf{y}^{m,(k-1)}, \mathbf{y}^{m,(k)}) = \frac{y_i^{m,(k)}}{y_i^{m,(k-1)}} \sum_{\substack{j=1 \\ j \neq i}}^N \lambda_{ij} y_j^n - y_i^n \sum_{\substack{j=1 \\ j \neq i}}^N \lambda_{ji} \frac{y_j^{m,(k)}}{y_j^{m,(k-1)}}$$

yields

$$\begin{aligned} \frac{\partial}{\partial y_i^n} n_i^{0,(k)}(\mathbf{y}^*, \mathbf{y}^*, \mathbf{y}^*) &= - \sum_{\substack{j=1 \\ j \neq i}}^N \lambda_{ji} = \lambda_{ii} \\ \frac{\partial}{\partial y_q^n} n_i^{0,(k)}(\mathbf{y}^*, \mathbf{y}^*, \mathbf{y}^*) &= \lambda_{iq}, \quad i \neq q, \end{aligned}$$

so that $\mathbf{D}_n^* \mathbf{n}^{r,(k)} = \delta_{r0} \mathbf{\Lambda}$. This results in

$$\mathbf{D}_n^* \mathbf{\Phi}_k^m = - \big(\mathbf{I} + (\theta_{0,-}^m + \theta_{0,+}^m) \Delta t \mathbf{\Lambda} \big) = - \left(\mathbf{I} + \theta_0^m \Delta t \mathbf{\Lambda} \right),$$

proving (5.62).

To derive (5.63) consider first the case $s = k - 1 > 0$ and $0 < l \neq m$. From

(5.57) it follows immediately that

$$\mathbf{D}^*_{l,(k-1)}\mathbf{p}^{r,(k)} = \delta_{rl}\mathbf{\Lambda}.$$

Moreover, (5.59) yields

$$\begin{aligned}\frac{\partial}{\partial y_i^{l,(k-1)}} n_i^{r,(k)}(\mathbf{y}^*,\mathbf{y}^*,\mathbf{y}^*) &= -\delta_{rl}\sum_{\substack{j=1\\ j\neq i}}^N \lambda_{ji} = \delta_{rl}\lambda_{ii},\\ \frac{\partial}{\partial y_q^{l,(k-1)}} n_i^{r,(k)}(\mathbf{y}^*,\mathbf{y}^*,\mathbf{y}^*) &= \delta_{rl}\lambda_{iq}, \quad i\neq q,\end{aligned}$$

which means that $\mathbf{D}^*_{l,(k-1)}\mathbf{n}^{r,(k)} = \delta_{rl}\mathbf{\Lambda}$ for $l\neq m$. In total (5.60) gives us

$$\mathbf{D}^*_{l,(k-1)}\mathbf{\Phi}_k^m = -\Delta t(\theta^m_{l,+}+\theta^m_{l,-})\mathbf{\Lambda} = -\Delta t\theta_l^m\mathbf{\Lambda}.$$

Next, we investigate the case of $s=k-1>0$ and $l=m$. Using once again $\operatorname{diag}(\mathbf{v})\mathbf{w} = \operatorname{diag}(\mathbf{w})\mathbf{v}$ and (5.57), we obtain

$$\begin{aligned}\mathbf{D}^*_{m,(k-1)}\mathbf{p}^{r,(k)} &= \mathbf{D}^*_{m,(k-1)}\left(\mathbf{\Lambda}\operatorname{diag}(\mathbf{y}^{m,(k)})\operatorname{diag}\left(\mathbf{y}^{m,(k-1)}\right)^{-1}\mathbf{y}^{r,(k-1)}\right)\\ &= -(1-\delta_{rm})\mathbf{\Lambda}.\end{aligned}$$

Furthermore, recalling (5.59), i. e.

$$n_i^{r,(k)}(\mathbf{y}^{r,(k-1)},\mathbf{y}^{m,(k-1)},\mathbf{y}^{m,(k)}) = \frac{y_i^{m,(k)}}{y_i^{m,(k-1)}}\sum_{\substack{j=1\\ j\neq i}}^N \lambda_{ij}y_j^{r,(k-1)} - y_i^{r,(k-1)}\sum_{\substack{j=1\\ j\neq i}}^N \lambda_{ji}\frac{y_j^{m,(k)}}{y_j^{m,(k-1)}},$$

we also distinguish between $r=m$ and $r\neq m$. In the first case we observe $n_i^{m,(k)} = n_i^{m,(k)}(\mathbf{y}^{m,(k-1)},\mathbf{y}^{m,(k)})$ and

$$\begin{aligned}\frac{\partial}{\partial y_i^{m,(k-1)}} n_i^{m,(k)}(\mathbf{y}^*,\mathbf{y}^*) &= -\frac{1}{y_i^*}\sum_{\substack{j=1\\ j\neq i}}^N \lambda_{ij}y_j^* - \sum_{\substack{j=1\\ j\neq i}}^N \lambda_{ji} \overset{(5.64)}{=} 2\lambda_{ii},\\ \frac{\partial}{\partial y_q^{m,(k-1)}} n_i^{m,(k)}(\mathbf{y}^*,\mathbf{y}^*) &= \lambda_{iq} + \lambda_{qi}\frac{y_i^*}{y_q^*} \overset{(5.65)}{=} (\mathbf{\Lambda} + \operatorname{diag}(\mathbf{y}^*)\mathbf{\Lambda}^T(\operatorname{diag}(\mathbf{y}^*))^{-1})_{iq},\end{aligned}$$

for $i\neq q$, which means that $\mathbf{D}^*_{m,(k-1)}\mathbf{n}^{m,(k)} = \mathbf{\Lambda} + \operatorname{diag}(\mathbf{y}^*)\mathbf{\Lambda}^T(\operatorname{diag}(\mathbf{y}^*))^{-1}$. Turning to the case $r\neq m$, we find

$$\begin{aligned}\frac{\partial}{\partial y_i^{m,(k-1)}} n_i^{r,(k)}(\mathbf{y}^*,\mathbf{y}^*,\mathbf{y}^*) &= -\frac{1}{y_i^*}\sum_{\substack{j=1\\ j\neq i}}^N \lambda_{ij}y_j^* \overset{(5.64)}{=} \lambda_{ii},\\ \frac{\partial}{\partial y_q^{m,(k-1)}} n_i^{r,(k)}(\mathbf{y}^*,\mathbf{y}^*,\mathbf{y}^*) &= \lambda_{qi}\frac{y_i^*}{y_q^*} \overset{(5.65)}{=} (\operatorname{diag}(\mathbf{y}^*)\mathbf{\Lambda}^T(\operatorname{diag}(\mathbf{y}^*))^{-1})_{iq}, \quad i\neq q,\end{aligned}$$

resulting in $\mathbf{D}^*_{m,(k-1)}\mathbf{n}^{r,(k)} = \operatorname{diag}(\mathbf{y}^*)\mathbf{\Lambda}^T(\operatorname{diag}(\mathbf{y}^*))^{-1}$ for $r\neq m$. Altogether, we thus end up with

$$\mathbf{D}^*_{m,(k-1)}\mathbf{\Phi}^m_k = \sum_{r=0}^{M} \theta^m_{r,+}\Delta t\mathbf{\Lambda} - \sum_{r=0}^{M} \theta^m_{r,-}\Delta t(\operatorname{diag}(\mathbf{y}^*)\mathbf{\Lambda}^T(\operatorname{diag}(\mathbf{y}^*))^{-1}) - \theta^m_m\Delta t\mathbf{\Lambda}.$$

Finally, we have to consider the case $s = k$ and $l = m$, i. e. we have to compute $\mathbf{D}^*_{m,(k)}\mathbf{\Phi}^m_k$. Using $\operatorname{diag}(\mathbf{v})\mathbf{w} = \operatorname{diag}(\mathbf{w})\mathbf{v}$ and (5.57) once again we see that

$$\mathbf{D}^*_{m,(k)}\mathbf{p}^{r,(k)} = \mathbf{\Lambda}.$$

Furthermore, we obtain

$$\frac{\partial}{\partial y_i^{m,(k)}} n_i^{r,(k)}(\mathbf{y}^*,\mathbf{y}^*,\mathbf{y}^*) = \frac{1}{y_i^*}\sum_{\substack{j=1\\ j\neq i}}^{N} \lambda_{ij}y_j^* \overset{(5.64)}{=} -\lambda_{ii},$$
$$\frac{\partial}{\partial y_q^{m,(k)}} n_i^{r,(k)}(\mathbf{y}^*,\mathbf{y}^*,\mathbf{y}^*) = -\lambda_{qi}\frac{y_i^*}{y_q^*} \overset{(5.65)}{=} -(\operatorname{diag}(\mathbf{y}^*)\mathbf{\Lambda}^T(\operatorname{diag}(\mathbf{y}^*))^{-1})_{iq}, \quad i \neq q,$$

resulting in

$$\mathbf{D}^*_{m,(k)}\mathbf{\Phi}^m_k = \mathbf{I} - \sum_{r=0}^{M} \theta^m_{r,+}\Delta t\mathbf{\Lambda} + \sum_{r=0}^{M} \theta^m_{r,-}\Delta t\,\operatorname{diag}(\mathbf{y}^*)\mathbf{\Lambda}^T(\operatorname{diag}(\mathbf{y}^*))^{-1}.$$

With this, we have finally proven Theorem 5.29.

□

Focusing on Gauss–Lobatto nodes, a higher-order quadrature rule is applied[1]. Here, we use $M = \lceil \frac{K}{2} \rceil$ subintervals and K=p corrections. Recall that we denoted the pth order MPDeC method by MPDeC(p) and indicated GL and EQ nodes by using MPDeCGL and MPDeCEQ, respectively. Note that MPDeC(1) is equivalent to the modified Patankar–Euler scheme and MPDeC(2) is equivalent to MPRK22(1) for both, GL and EQ nodes.

Due to $\mathbf{y}^{n+1} = \mathbf{y}^{M,(K)}$, MPDeC conserves all linear invariants, if $\theta_r^M \geq 0$ for all $r = 0, \dots, M$ since in this case the index function yields $\gamma(j, i, \theta_r^M) = j$ and (MPDeC) can be written as

$$\mathbf{y}^{n+1} - \mathbf{y}^n - \sum_{r=0}^{M} \theta_r^M \Delta t\mathbf{\Lambda}\,\operatorname{diag}(\mathbf{y}^{n+1})(\operatorname{diag}(\mathbf{y}^{M,(K-1)})^{-1}\mathbf{y}^{r,(K-1)} = \mathbf{0},$$

which means that $\mathbf{n}^T\mathbf{y}^{n+1} = \mathbf{n}^T\mathbf{y}^n$ for all $\mathbf{n} \in \ker(\mathbf{\Lambda}^T)$. Indeed, for equispaced nodes, θ_r^M with $r = 0, \dots, M$ are the weights of the closed Newton–Cotes formulas for integrals over $I = [0, 1]$. Hence, a negative θ_r^M occurs for the first time at $M = 7$, i. e. with MPDeCEQ(8). In this case, we also have to consider $\mathbf{n}^{r,(K)}(\mathbf{y}^{r,(K-1)}, \mathbf{y}^{M,(K-1)}, \mathbf{y}^{n+1})$ given in (5.58), resulting in

$$\mathbf{n}^T\mathbf{n}^{r,(K)} = \sum_{i,j=1}^{N} n_i \frac{y_i^{n+1}}{y_i^{M,(K-1)}} p_{ij}(\mathbf{y}^{r,(K-1)}) - \sum_{i,j=1}^{N} n_j \frac{y_i^{n+1}}{y_i^{M,(K-1)}} p_{ij}(\mathbf{y}^{r,(K-1)}),$$

[1] The L^2 operator inside the DeC framework is based on a collocation method with Lobatto nodes (also known as the RK Lobatto III A method).

where we switched indices and used $d_{ij} = p_{ji}$. We observe that $\mathbf{n}^T\mathbf{n}^{r,(k)}$ does not need to vanish for $\mathbf{n} \notin \operatorname{span}(\mathbf{1})$, so that the preservation of all linear invariants can not be guaranteed anymore for arbitrary systems $\mathbf{y}' = \mathbf{\Lambda}\mathbf{y}$ and MPDeC(p) with equispaced nodes with $p \geq 8$. However, as the system matrix of (5.1) satisfies additional properties, additional research is required to answer this question.

Moreover, in the case of Gauss–Lobatto nodes, the values $2\theta_r^M$ for $r = 0, \dots, M$ equal the weights of the corresponding Gauss–Lobatto quadrature, which are always positive. This gives us the following result.

Lemma 5.30. The MPDeCGL methods conserve all linear invariants when applied to (5.1).

Remark 5.31. From Theorem 5.29, we see that the Jacobian in general depends on $\mathbf{y}^*$, if there exist negative correction weights θ_r^m, i. e. not being conditional stable does not necessarily result in instability in this case. For equispaced or Gauss–Lobatto points negative correction weights already occur for $K > 2$. Hence, to study the stability of MPDeC schemes applied to general linear systems, one needs to locate the eigenvalues of the Jacobian, which possibly depend on $\mathbf{y}^*$ themselves. Such an analysis is outside the scope of this work, which is why we will focus on the following class of problems.

If $\mathbf{\Lambda}$ is normal, then $\mathbf{\Lambda}$ and $\mathbf{\Lambda}^T$ share the same eigenvectors and the corresponding eigenvalues are the complex conjugate of each other. Since $\mathbf{1} \in \ker(\mathbf{\Lambda}^T)$ this means that even $\mathbf{1} \in \ker(\mathbf{\Lambda})$. Hence, we may discuss the stability of $\mathbf{y}^* = \mathbf{1}$. Then, we find

$$\sigma\left(r_1(\mathbf{\Lambda}) + r_2\left(\operatorname{diag}(\mathbf{y}^*)\mathbf{\Lambda}^T(\operatorname{diag}(\mathbf{y}^*))^{-1}\right)\right) = \{r_1(\lambda) + r_2(\bar{\lambda}) \mid \lambda \in \sigma(\mathbf{\Lambda})\}$$

for any rational maps r_1, r_2, which means that the spectrum of the Jacobian of the map $\mathbf{g}$ generating the MPDeC iterates can be written only in terms of the eigenvalues of $\mathbf{\Lambda}$. Using (5.61), (5.62) and (5.63), the stability function R_p of MPDeC(p) for normal matrices $\mathbf{\Lambda}$ and $\mathbf{y}^* = \mathbf{1}$ can be computed recursively by

$$\begin{aligned}
R^{m,(1)}(z) =& \frac{1 + (z + \bar{z})\sum_{r=0}^M \theta_{r,-}^m}{1 - \left(z\sum_{r=0}^M \theta_{r,+}^m - \bar{z}\sum_{r=0}^M \theta_{r,-}^m\right)},\\
R^{m,(\hat{k})}(z) =& \frac{1 + \theta_0^m z + z \displaystyle\sum_{\substack{r=1\\ r\neq m}}^{M} \theta_r^m R^{r,(\hat{k}-1)}(z)}{1 - \left(z\sum_{r=0}^M \theta_{r,+}^m - \bar{z}\sum_{r=0}^M \theta_{r,-}^m\right)}\\
&- \frac{\left(z\sum_{r=0}^M \theta_{r,+}^m - \bar{z}\sum_{r=0}^M \theta_{r,-}^m - z\theta_m^m\right) R^{m,(\hat{k}-1)}(z)}{1 - \left(z\sum_{r=0}^M \theta_{r,+}^m - \bar{z}\sum_{r=0}^M \theta_{r,-}^m\right)},\\
R_p(z) =& R^{M,(K)}(z),
\end{aligned} \tag{5.66}$$

for $\hat{k} = 2, \dots, K$ and $m = 1, \dots, M$. Note that if $\mathbf{\Lambda}$ is symmetric it is also normal and we obtain $\sigma(\mathbf{\Lambda}) \subseteq \mathbb{R}$, so that one can further simplify (5.66) using

$\theta^m_{r,+} + \theta^m_{r,-} = \theta^m_r$ to receive

$$\begin{aligned}
R^{m,(1)}(z) &= \frac{1 + 2z\sum_{r=0}^M \theta^m_{r,-}}{1 - z\sum_{r=0}^M |\theta^m_r|},\\
R^{m,(\hat{k})}(z) &= \frac{1 + \theta^m_0 z + z \displaystyle\sum_{\substack{r=1\\ r\neq m}}^{M} \theta^m_r R^{r,(\hat{k}-1)}(z) - z\left(\sum_{r=0}^{M} |\theta^m_r| - \theta^m_m\right) R^{m,(\hat{k}-1)}(z)}{1 - z\sum_{r=0}^M |\theta^m_r|},\\
R_p(z) &= R^{M,(K)}(z).
\end{aligned} \tag{5.67}$$

It is also worth mentioning that for the system matrix

$$\mathbf{\Lambda} = \begin{pmatrix} -a & b \\ a & -b \end{pmatrix} \tag{5.68}$$

with $a, b > 0$, used in [IKM22a, TÖR22], we find that $\ker(\mathbf{\Lambda}) = \operatorname{span}(\mathbf{y}^*)$ with $\frac{y_1^*}{y_2^*} = \frac{b}{a}$, and thus

$$\operatorname{diag}(\mathbf{y}^*)\mathbf{\Lambda}^T(\operatorname{diag}(\mathbf{y}^*))^{-1} = \begin{pmatrix} -a & a\frac{b}{a} \\ b\frac{a}{b} & -b \end{pmatrix} = \mathbf{\Lambda},$$

so that the stability function R_p in this case is also given by (5.67).

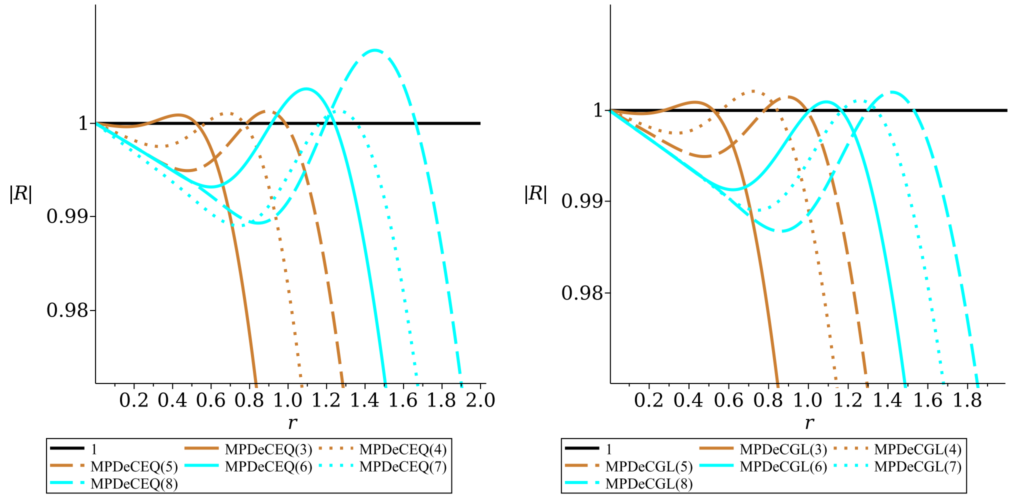
Deferred Correction schemes are described by an iterative process which can be compared with classical RK schemes with more stages [ALMÖT22]. As MPDeC and DeC share the same amount of stages, we thus know that MPDeCEQ(3) contains 5 stages. Furthermore, MPDeCEQ(4) has already 10 stages inside, resulting in rational function with polynomial degree 10 in the numerator and denominator. Using Gauss–Lobatto nodes decreases the number of stages. For an MPDeCGL(4) we would end up with seven stages.

Similarly as we did for MPRK43(α, β) we can estimate the stability region for MPDeC schemes. It turns out that MPDeCEQ(p) and MPDeCGL(p) are unconditionally stable for $p = 1, 2$, as they coincide with MPE and MPRK22(1), respectively. For $p = 3, \dots, 8$ we collect the lower bounds θ_{num} for the opening angle θ of the stability domain for normal system matrices $\mathbf{\Lambda}$ in Table 5.1. Thereby we show that θ_{num} actually satisfies the error bound $\theta_{\text{num}} \leq \theta < \theta_{\text{num}} + \frac{\pi}{500}$ by adding $\frac{\pi}{500}$ to θ_{num} and demonstrating that the absolute value of the stability function then exceeds 1 for some $r = |z|$, see Figure 5.4.

To give a first insight in the stability properties of MPDeC methods of order higher than 8, we analyze the reduced stability function (5.67). In both cases described in Remark 5.31, the eigenvalues of $\mathbf{\Lambda}$ leading to (5.67) are real, which is why we present the absolute value of the stability function over real z in Figure 5.5. To obtain a stable scheme, the absolute value of $R(z)$ has to be always smaller than one (the black line). In Figure 5.5b, we investigate MPDeC from order 4 to 14 using Gauss–Lobatto points. As can be recognized all MPDeC methods are stable using Gauss–Lobatto points. In Figure 5.5a, the stability functions of MPDeC schemes from 4th to 14th order are depicted for equispaced nodes. Here, we recognize that MPDeCEQ(12) and MPDeCEQ(14) are unstable but MPDeCEQ(13) is stable. However, this is not surprising since already for classical DeC methods using equidistant points has been problematic for high-order methods,

p	MPDeCEQ(p)	MPDeCGL(p)
1	π	π
2	π	π
3	$\frac{498}{500}\pi$	$\frac{498}{500}\pi$
4	$\frac{496}{500}\pi$	$\frac{496}{500}\pi$
5	$\frac{495}{500}\pi$	$\frac{495}{500}\pi$
6	$\frac{495}{500}\pi$	$\frac{494}{500}\pi$
7	$\frac{494}{500}\pi$	$\frac{494}{500}\pi$
8	$\frac{495}{500}\pi$	$\frac{494}{500}\pi$

Table 5.1: Estimate θ_{num} from (5.41) of the opening angle θ of the stability domain of MPDeC(p) for problems (5.1) with a normal system matrix. We have $\theta_{\text{num}} \leq \theta < \theta_{\text{num}} + \frac{\pi}{500}$, see Figure 5.4.

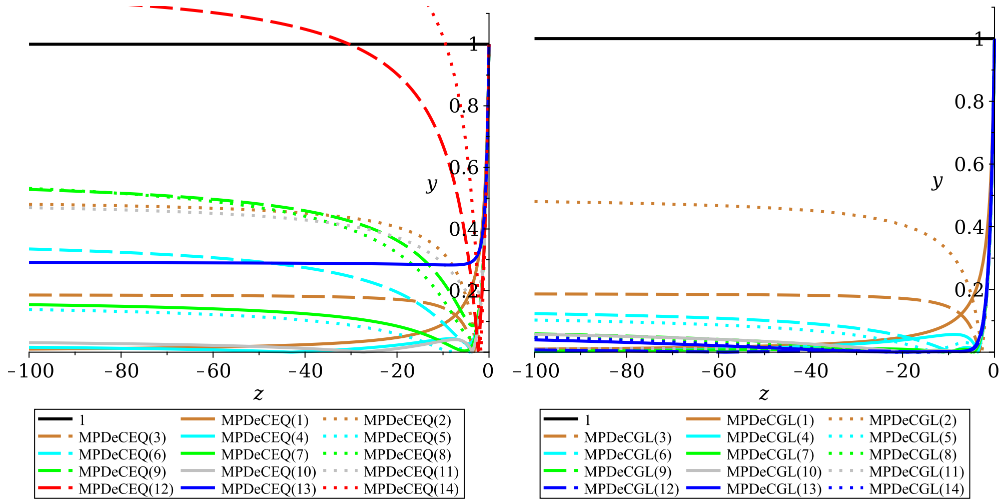


(a) Equispaced nodes

(b) Gauss–Lobatto nodes

Figure 5.4: Plots of $|R(re^{\mathrm{i}\varphi})|$ over r, where $\varphi = \pi - \frac{\theta_{\text{num}} + \frac{\pi}{500}}{2}$ corresponds to the opening angle $\theta_{\text{num}} + \frac{\pi}{500}$ with θ_{num} from Table 5.1.

cf. [DGR00, HÖT21, ÖT20, TÖR22] and references therein. The reason for this is related with classical interpolation theory where it is known that equidistant points may lead to Runge's phenomenon. However, we would like to point out that our investigation also supports the numerical investigation in [TÖR22] where problems in MPDeCEQ have been recognized.



(a) Equispaced points

(b) Gauss–Lobatto points

Figure 5.5: Absolute value of the stability function over $z \leq 0$.

5.4.4 Geometric Conservative

We will start analyzing GeCo1 applied to a general positive linear test problem with stable steady states. Turning to GeCo2, we prove that already for the 2×2 system (5.4), (5.2) the stability domain of GeCo2 is bounded. Approaches for the analysis of GeCo2 for $N \times N$ systems will be discussed at the end of the respective subsection.

Stability of GeCo1

In this subsection, we investigate the stability properties of GeCo1, see (GeCo1),

To that end, we first rewrite $\mathbf{y}' = \mathbf{\Lambda}\mathbf{y}$ as a bio chemical system of the form

$$\mathbf{y}' = \mathbf{\Lambda}\mathbf{y} = \mathbf{f}^{[P]}(\mathbf{y}) - \mathbf{f}^{[D]}(\mathbf{y}) = \mathbf{S}^+\mathbf{y} - \mathbf{S}^-\mathbf{y} \tag{5.69}$$

with $\mathbf{S}^+, \mathbf{S}^- \geq \mathbf{0}$. Since $\mathbf{\Lambda}$ is a Metzler matrix, $\mathbf{S}^- = (s^-_{ij})_{i,j=1,\dots,N}$ is a diagonal matrix and $f_i^{[D]}(\mathbf{y}) = s^-_{ii} y_i$. Moreover, Remark 5.1 states that at least one diagonal element of $\mathbf{\Lambda}$ is negative, which results in

$$\operatorname{trace}(\mathbf{S}^-) > 0. \tag{5.70}$$

With this in mind, let us recall the function φ from (3.22), that is

$$\varphi(x) = \begin{cases} \frac{1-e^{-x}}{x}, & x > 0, \\ 1, & x = 0 \end{cases}$$

and write the GeCo1 scheme (GeCo1) applied to (5.69) as

$$\begin{aligned} \mathbf{g}(\mathbf{y}^n) = \mathbf{y}^{n+1} &= \mathbf{y}^n + \Delta t \varphi\left(\Delta t \sum_{i=1}^N \frac{f_i^{[D]}(\mathbf{y})}{y_i^n} \right) \mathbf{\Lambda}\mathbf{y}^n \\ &= \mathbf{y}^n + \Delta t \varphi(\Delta t \operatorname{trace}(\mathbf{S}^-))\mathbf{\Lambda}\mathbf{y}^n. \end{aligned}$$

Due to (5.70), the GeCo1 scheme can be rewritten as

$$\mathbf{y}^{n+1}=\mathbf{g}(\mathbf{y}^n)=(\mathbf{I}+\Phi(\Delta t)\mathbf{\Lambda})\mathbf{y}^n,\quad \Phi(\Delta t)=\Delta t\varphi(\Delta t\,\mathrm{trace}(\mathbf{S}^-))=\tfrac{1-e^{-\Delta t\,\mathrm{trace}(\mathbf{S}^-)}}{\mathrm{trace}(\mathbf{S}^-)}. \tag{5.71}$$

It is worth mentioning that this reasoning holds for all $k=\dim(\ker(\mathbf{\Lambda}))\geq 0$. Also note that steady states of (5.1) become fixed points of $\mathbf{g}$, and that $\mathbf{g}\in\mathcal{C}^\infty$ conserves all linear invariants, if there are any. Hence, we are in the position to apply Theorem 5.4. It is worth noting that the eigenvalues of the Jacobian of the GeCo1 map $\mathbf{g}$ in general not only depend on $\Delta t\lambda$, but also on the trace of $\mathbf{S}^-$. Nevertheless, we are able to prove that in the case of GeCo1, the remaining $N-k$ eigenvalues of $\mathbf{Dg}(\mathbf{y}^*)$ lie inside the unit circle, resulting in the following theorem.

Theorem 5.32. If $k>0$ then the steady state $\mathbf{y}^*$ of (5.1), (5.2) is a stable fixed point of GeCo1 for all $\Delta t>0$. Furthermore, there exists a $\delta>0$ such that $\|\mathbf{y}^0-\mathbf{y}^*\|<\delta$ implies the convergence of the iterates towards $\mathbf{y}^*$ for all $\Delta t>0$. If $k=0$, then $\mathbf{y}^*=\mathbf{0}$ is an asymptotically stable fixed point of GeCo1 for all $\Delta t>0$.

Proof. The Jacobian $\mathbf{Dg}(\mathbf{y}^*)$ reads

$$\mathbf{Dg}(\mathbf{y}^*)=\mathbf{I}+\Phi(\Delta t)\mathbf{\Lambda}=\mathbf{I}+\frac{1-e^{-\Delta t\,\mathrm{trace}(\mathbf{S}^-)}}{\mathrm{trace}(\mathbf{S}^-)}\mathbf{\Lambda}$$

and its eigenvalues are

$$\mu=1+\Phi(\Delta t)\lambda$$

with $\lambda\in\sigma(\mathbf{\Lambda})$. Hereby, we see that in the case of $k>0$, any $\mathbf{v}\in\ker(\mathbf{\Lambda})\setminus\{\mathbf{0}\}$ is an eigenvector of the Jacobian $\mathbf{Dg}(\mathbf{y}^*)$ with an associated eigenvalue of 1.

In order to investigate the location of the remaining $N-k$ eigenvalues of the Jacobian for $k\geq 0$, we first numerate the distinct and nonzero eigenvalues of $\mathbf{\Lambda}$ from (5.1) by $\lambda_1,\dots,\lambda_m$. Now, the corresponding eigenvalues $\mu_i=1+\Phi(\Delta t)\lambda_i$ with $i=1,\dots,m$ lie inside the unit circle if and only if

$$|1+\Phi(\Delta t)\lambda_i|^2<1,\quad i=1,\dots,m,$$

which can be written as

$$(\mathrm{Re}(\Phi(\Delta t)\lambda_i)+1)^2+\mathrm{Im}(\Phi(\Delta t)\lambda_i)^2<1,\quad i=1,\dots,m,$$

or equivalently,

$$2\Phi(\Delta t)\,\mathrm{Re}(\lambda_i)+\Phi(\Delta t)^2|\lambda_i|^2<0,\quad i=1,\dots,m.$$

Dividing by $\Phi(\Delta t)>0$ and exploiting $\sigma(\mathbf{\Lambda})\subseteq\overline{\mathbb{C}^-}$ gives

$$\Phi(\Delta t)<-\frac{2\,\mathrm{Re}(\lambda_i)}{|\lambda_i|^2}=\frac{2|\mathrm{Re}(\lambda_i)|}{|\lambda_i|^2},\quad i=1,\dots,m.$$

Introducing

$$M=\min_{i=1,\dots,m}\left\{2\tfrac{|\mathrm{Re}(\lambda_i)|}{|\lambda_i|^2}\right\}, \tag{5.72}$$

we end up with the equivalent condition

$$\Phi(\Delta t)<M.$$

Hence, after plugging in $\Phi(\Delta t) = \frac{1-e^{-\Delta t\,\mathrm{trace}(\mathbf{S}^-)}}{\mathrm{trace}(\mathbf{S}^-)}$, we multiply with its denominator $\mathrm{trace}(\mathbf{S}^-) > 0$, see (5.70), to get

$$|\mu_i|^2 < 1, \quad i = 1, \dots, m$$

if and only if

$$1 - e^{-\Delta t\,\mathrm{trace}(\mathbf{S}^-)} < M\,\mathrm{trace}(\mathbf{S}^-).$$

Now, if $M\,\mathrm{trace}(\mathbf{S}^-) \geq 1$, then

$$M\,\mathrm{trace}(\mathbf{S}^-) \geq 1 > 1 - e^{-\Delta t\,\mathrm{trace}(\mathbf{S}^-)}$$

is true for all $\Delta t > 0$, and hence, the remaining eigenvalues of $\mathbf{Dg}(\mathbf{y}^*)$ associated with nonzero eigenvalues of $\mathbf{\Lambda}$ lie inside the unit circle. We now aim to prove that

$$M\,\mathrm{trace}(\mathbf{S}^-) \geq 1$$

is indeed the case.

Due to [BF04, Theorem 10, Corollary 11] it holds that

$$\sigma(\mathbf{\Lambda}) \subseteq \mathcal{B} = \left\{ z \in \mathbb{C} \,\middle|\, |z - r| \leq |r|, r = \min_{j=1,\dots,N} \lambda_{jj} \right\}.$$

Since $\mathbf{\Lambda}$ is a proper Metzler matrix, see Remark 5.1, there exists an $l \in \{1, \dots, N\}$ such that

$$r = \min_{j=1,\dots,N} \lambda_{jj} = \lambda_{ll} < 0. \tag{5.73}$$

Even more, we know $\mathrm{Re}(\lambda) < 0$ as well as $\arg(\lambda) \in (\frac{\pi}{2}, \frac{3}{2}\pi)$ for all $0 \neq \lambda \in \sigma(\mathbf{\Lambda})$. For any given $\lambda \in \sigma(\mathbf{\Lambda}) \setminus \{0\}$, we define $\alpha = \pi - \arg(\lambda) \in (-\frac{\pi}{2}, \frac{\pi}{2})$, so that

$$\cos(\alpha) = \frac{|\mathrm{Re}(\lambda)|}{|\lambda|} \neq 0.$$

Next, we choose $\theta < 0$ satisfying

$$|\lambda| = \cos(\alpha)|\theta|.$$

A sketch of this geometry can be found in Figure 5.6. With this, equation (5.72)

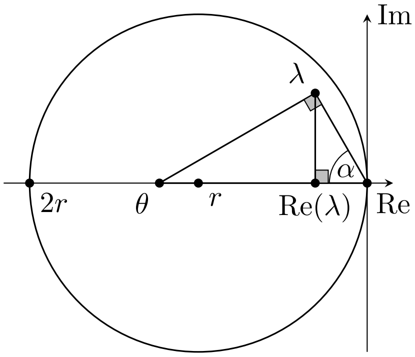


Figure 5.6: Sketch of the geometric setup for $\arg(\lambda) \in (\frac{\pi}{2}, \pi)$.

becomes

$$M = \min_{i=1,\dots,m}\left\{2\tfrac{|\mathrm{Re}(\lambda_i)|}{|\lambda_i|^2}\right\} = \min_{i=1,\dots,m}\left\{2\tfrac{\cos(\alpha_i)}{|\lambda_i|}\right\} = \min_{i=1,\dots,m}\left\{\tfrac{2}{|\theta_i|}\right\}.$$

Moreover, with Thales's Theorem we can conclude that even $\theta_i \in \mathbb{R}^-$ is contained in $\mathcal{B}$, and thus, satisfies

$$|\theta_i| \leq 2|\min_{j=1,\dots,N} \lambda_{jj}|.$$

From (5.73) it thus follows that

$$M = \min_{i=1,\dots,m}\left\{\tfrac{2}{|\theta_i|}\right\} \geq \frac{2}{2|\min_{j=1,\dots,N}\lambda_{jj}|} = \frac{1}{|\lambda_{ll}|}.$$

Additionally, setting $S = \{j \in \{1,\dots,N\} \mid \lambda_{jj} < 0\}$ we find

$$\mathrm{trace}(\mathbf{S}^-) = -\sum_{j\in S}^{N}\lambda_{jj} = \sum_{j\in S}^{N}|\lambda_{jj}| \geq |\lambda_{ll}|,$$

and thus

$$M\,\mathrm{trace}(\mathbf{S}^-) \geq \frac{1}{|\lambda_{ll}|}|\lambda_{ll}| = 1,$$

which finishes the proof as we have also proven that $\rho(\mathbf{Dg}(\mathbf{0})) < 1$ in the case of $k = 0$. □

With this theorem, a stability result for the GeCo1 scheme is provided for the first time. Having proved the unconditional stability of all fixed points of GeCo1 associated with steady states of the general $N \times N$ system of differential equations (5.1), we can conclude that GeCo1 mimics the stability behavior of the analytic solution close to a steady state solution for any chosen time step size $\Delta t > 0$. Whether or not this already suggests that the explicit GeCo1 scheme is even capable of solving stiff problems will be discussed in Section 5.6.7.

Stability of GeCo2

In this subsection we aim to prove that (GeCo2) applied to (5.4) can be described by a $\mathcal{C}^1$-map $\mathbf{g}$ using Lemma A.2 from the appendix, and to compute the spectrum of the corresponding Jacobian. To prove that the partial derivatives are even locally Lipschitz continuous, we use Lemma A.3 from the appendix. However, we will also prove that $\mathbf{g} \notin \mathcal{C}^2$ for any neighborhood of $\mathbf{y}^*$, extending the work [IKMM23]. This underlines the benefits discussed in Remark 5.6 on the stability theorems from [IKM22a] and Theorem 5.4, published in [IKM22b].

Let us investigate the GeCo2 scheme applied to (5.2), (5.4) with

$$\mathbf{\Lambda} = \underbrace{\begin{pmatrix} 0 & bc \\ a & 0 \end{pmatrix}}_{=\mathbf{S}^+} - \underbrace{\begin{pmatrix} ac & 0 \\ 0 & b \end{pmatrix}}_{=\mathbf{S}^-}$$

and $\mathbf{r}(\mathbf{y}) = \mathbf{y}$. This means that $\mathbf{f}^{[D]}(\mathbf{y}) = \mathbf{S}^-\mathbf{r}(\mathbf{y}) = \left(\begin{smallmatrix} acy_1 \\ by_2 \end{smallmatrix}\right)$, and hence the GeCo2 scheme (GeCo2) reads

$$\begin{aligned}
\mathbf{y}^{(2)} &= \mathbf{y}^n + \Delta t\varphi(\Delta t\,\mathrm{trace}(\mathbf{S}^-))\mathbf{\Lambda}\mathbf{y}^n\\
\mathbf{y}^{n+1} &= \mathbf{y}^n + \frac{1}{2}\Delta t\varphi\left(\Delta t\left(\frac{w_1^+(\mathbf{y}^n)}{y_1^n}+\frac{w_2^+(\mathbf{y}^n)}{y_2^n}\right)\right)\mathbf{\Lambda}\left(\mathbf{y}^n+\mathbf{y}^{(2)}\right)\\
&= \mathbf{y}^n + \frac{1}{2}\Delta t\varphi\left(\Delta t\left(\frac{w_1^+(\mathbf{y}^n)}{y_1^n}+\frac{w_2^+(\mathbf{y}^n)}{y_2^n}\right)\right)\mathbf{\Lambda}\left(2\mathbf{y}^n+\Delta t\varphi(\Delta t\,\mathrm{trace}(\mathbf{S}^-))\mathbf{\Lambda}\mathbf{y}^n\right),
\end{aligned} \tag{5.74}$$

where $w_i^+(\mathbf{y}^n)=\max(0,w_i(\mathbf{y}^n))$ for $i=1,2$ and

$$\begin{aligned}
\mathbf{w}(\mathbf{y}^n) &= 2\varphi(\Delta t\,\mathrm{trace}(\mathbf{S}^-))\mathbf{\Lambda}\mathbf{y}^n - \mathbf{\Lambda}\mathbf{y}^n - \mathbf{\Lambda}\mathbf{y}^{(2)}\\
&= \left(2\varphi(\Delta t\,\mathrm{trace}(\mathbf{S}^-))\mathbf{\Lambda} - 2\mathbf{\Lambda} - \mathbf{\Lambda}^2\Delta t\varphi(\Delta t\,\mathrm{trace}(\mathbf{S}^-))\right)\mathbf{y}^n.
\end{aligned} \tag{5.75}$$

We formulate a helpful lemma to understand some properties of $\mathbf{w}$ and to express equation (5.74) with $\mathbf{w}$ rather than w_1^+ and w_2^+.

Lemma 5.33. The map $\mathbf{w}$ from (5.75) with $\mathbf{\Lambda}$ from (5.4) satisfies $w_1=-\frac{1}{c}w_2$, and we have

$$w_1(\mathbf{y}^n)\begin{cases} >0, & y_1^n > \frac{b}{a}y_2^n,\\ =0, & y_1^n = \frac{b}{a}y_2^n,\\ <0, & y_1^n < \frac{b}{a}y_2^n.\end{cases}$$

Proof. First note that $y_1^n=\frac{b}{a}y_2^n$ is equivalent to $\mathbf{y}^n\in\ker(\mathbf{\Lambda})$, and thus (5.75) yields $\mathbf{w}(\mathbf{y}^n)=\mathbf{0}$.

Next, we focus on finding conditions for $\mathbf{y}^n$ so that $w_1(\mathbf{y}^n)>0$. For this, it is worth mentioning that for every $\mathbf{y}^n>\mathbf{0}$, there exists a unique $\mathbf{y}^*\in\ker(\mathbf{\Lambda})\cap\mathbb{R}^2_{>0}$ satisfying $\mathbf{n}^T\mathbf{y}^n=\mathbf{n}^T\mathbf{y}^*$ with $\mathbf{n}=(1,c)^T$, see [IKM22a, Lemma 2.8]. Hence, since $\mathbf{y}^*>\mathbf{0}$ and $\bar{\mathbf{y}}=(1,-\frac{1}{c})^T$ are linearly independent, there exists a unique $s^n\in\mathbb{R}$ such that $\mathbf{y}^n=\mathbf{y}^*+s^n\bar{\mathbf{y}}$. Also note that $\mathbf{\Lambda}\bar{\mathbf{y}}=\lambda\bar{\mathbf{y}}$ with $\lambda=-(ac+b)<0$ and

$$s^n\begin{cases} >0, & y_1^n > \frac{b}{a}y_2^n,\\ =0, & y_1^n = \frac{b}{a}y_2^n,\\ <0, & y_1^n < \frac{b}{a}y_2^n.\end{cases} \tag{5.76}$$

Thus, the linearity of $\mathbf{w}$ and (3.22) lead to

$$\begin{aligned}
\mathbf{w}(\mathbf{y}^n) = \mathbf{w}(\mathbf{y}^*)+\mathbf{w}(s^n\bar{\mathbf{y}}) &= \left(2\varphi(-\Delta t\lambda)\lambda - 2\lambda - \lambda^2\Delta t\varphi(-\Delta t\lambda)\right)s^n\bar{\mathbf{y}}\\
&= \frac{1}{\Delta t}\left(2(1-e^{\Delta t\lambda}) - 2\Delta t\lambda - \Delta t\lambda(1-e^{\Delta t\lambda})\right)s^n\bar{\mathbf{y}}\\
&= \frac{1}{\Delta t}\left(2-3\Delta t\lambda + e^{\Delta t\lambda}(\Delta t\lambda-2)\right)s^n\bar{\mathbf{y}}.
\end{aligned} \tag{5.77}$$

Furthermore, introducing the function

$$p(z) = -3z+2-e^z(2-z),$$

we can rewrite (5.77) to get

$$\mathbf{w}(\mathbf{y}^n) = \frac{1}{\Delta t}p(\Delta t\lambda)s^n\bar{\mathbf{y}}. \tag{5.78}$$

Now, the first derivative of p satisfies

$$p'(z) = -(3 + e^z(1-z)) < 0$$

for all $z \leq 0$. Hence, the function p is strictly decreasing for $z \leq 0$ and satisfies $p(0) = 0$ proving that $p(\lambda\Delta t) > 0$ for all $\Delta t > 0$. Therefore, with (5.76) it follows that $w_1(\mathbf{y}^n) > 0$ if $y_1^n > \frac{b}{a}y_2^n$. Similarly, $w_1(\mathbf{y}^n) < 0$ holds if $y_1^n < \frac{b}{a}y_2^n$. Finally, note that (5.78) implies $w_1(\mathbf{y}^n) = -\frac{1}{c}w_2(\mathbf{y}^n)$. □

As a consequence of this lemma we simplify (5.74) by introducing

$$H\colon \mathbb{R}^2_{>0} \to \mathbb{R}_{>0}$$

with

$$H(\mathbf{x}) = \Delta t\varphi\left(\Delta t\left(\frac{w_1^+(\mathbf{x})}{x_1} + \frac{w_2^+(\mathbf{x})}{x_2}\right)\right) = \begin{cases} \widetilde{H}_1(\mathbf{x}), & x_1 > \frac{b}{a}x_2, \\ \Delta t, & x_1 = \frac{b}{a}x_2, \\ \widetilde{H}_2(\mathbf{x}), & x_1 < \frac{b}{a}x_2, \end{cases}$$

$$\widetilde{H}_i(\mathbf{x}) = \frac{1 - e^{-\Delta t\frac{w_i(\mathbf{x})}{x_i}}}{\frac{w_i(\mathbf{x})}{x_i}}, \quad i = 1, 2$$

and point out that H is continuous, since φ from (3.22) is in $\mathcal{C}^2$ and $\mathbf{w} \in \mathcal{C}^\infty$. As a result of Lemma 5.33, we even know that

$$\widetilde{H}_i \in \mathcal{C}^\infty(\mathbb{R}^2_{>0} \setminus \ker(\mathbf{\Lambda}))$$

for $i = 1, 2$.

The map $\mathbf{g}$ defining the iterates of the GeCo2 scheme when applied to (5.2), (5.4) is given by (5.74) and can be written as

$$\mathbf{g}(\mathbf{x}) = \mathbf{x} + \frac{1}{2}H(\mathbf{x})\mathbf{\Lambda}\left(2\mathbf{x} + \Delta t\varphi(\Delta\operatorname{trace}(\mathbf{S}^-))\mathbf{\Lambda}\mathbf{x}\right).$$

Introducing $\mathbf{G}(\mathbf{x}) = \mathbf{\Lambda}\mathbf{x}H(\mathbf{x})$ we obtain

$$\mathbf{g}(\mathbf{x}) = \mathbf{x} + \mathbf{G}(\mathbf{x}) + \frac{1}{2}\Delta t\varphi(\Delta t\operatorname{trace}(\mathbf{S}^-))\mathbf{\Lambda}\mathbf{G}(\mathbf{x}). \tag{5.79}$$

The following theorem uses this representation of $\mathbf{g}$ to analyze the stability properties of GeCo2.

Theorem 5.34. Let $\mathbf{g}$, given by (5.79), be the generating function of the GeCo2 iterates $\mathbf{y}^n$ when applied to (5.4), (5.2). Further, let $\mathbf{y}^* > \mathbf{0}$ be a steady state solution of (5.4).

a) The map $\mathbf{g} \in \mathcal{C}^1(\mathcal{D})$ has Lipschitz continuous derivatives on a sufficiently small neighborhood $\mathcal{D}$ of $\mathbf{y}^*$. Moreover, the stability function reads

$$R(z) = 1 + z + \frac{1}{2}z^2\varphi(\Delta t\,\mathrm{trace}(\mathbf{S}^-)). \tag{5.80}$$

If $|R(-\Delta t(ac+b))| < 1$, then $\mathbf{y}^*$ is stable and there exists a $\delta > 0$ such that $\left(\frac{1}{c}\right)^T\mathbf{y}^0 = \left(\frac{1}{c}\right)^T\mathbf{y}^*$ and $\|\mathbf{y}^0 - \mathbf{y}^*\| < \delta$ imply $\lim_{n\to\infty}\mathbf{y}^n = \mathbf{y}^*$. If $|R(-\Delta t(ac+b))| > 1$, then $\mathbf{y}^*$ is an unstable fixed point of GeCo2.

b) There holds $\mathbf{g} \notin \mathcal{C}^2$ in any neighborhood of $\mathbf{y}^*$.

Proof. a) We demonstrate that all assumptions of Theorem 5.4 and Theorem 2.15 are fulfilled.

From part a) of Lemma A.2 from the appendix with $\mathbf{\Phi}(\mathbf{x}) = \mathbf{\Lambda}\mathbf{x}$ and $\Psi(\mathbf{x}) = H(\mathbf{x})$, it follows that the partial derivatives of $\mathbf{G}(\mathbf{x}) = \mathbf{\Lambda}\mathbf{x}H(\mathbf{x})$ on $\ker(\mathbf{\Lambda})$ exist and that $\mathbf{DG}(\mathbf{x}_0) = \Psi(\mathbf{x}_0)\mathbf{\Lambda} = \Delta t\mathbf{\Lambda}$ holds for $i = 1, 2$ and all $\mathbf{x}_0 \in \ker(\mathbf{\Lambda})$. As a result of (5.79) we obtain

$$\mathbf{Dg}(\mathbf{y}^*) = \mathbf{I} + \Delta t\mathbf{\Lambda} + \frac{1}{2}(\Delta t)^2\varphi(\Delta t\,\mathrm{trace}(\mathbf{S}^-))\mathbf{\Lambda}^2$$

and the eigenvalues are given by 1 and $R(-\Delta t(ac+b))$, where

$$R(z) = 1 + z + \frac{1}{2}z^2\varphi(-\Delta t\,\mathrm{trace}(\mathbf{S}^-)).$$

In total, we can write

$$\mathbf{DG}(\mathbf{x}) = \mathbf{B}(\mathbf{x}) + \mathbf{C}(\mathbf{x}) \tag{5.81}$$

with

$$\mathbf{B}(\mathbf{x}) = \mathbf{\Lambda}H(\mathbf{x}) \quad \text{and} \quad \mathbf{C}(\mathbf{x}) = \mathbf{\Lambda}\mathbf{x}\cdot\begin{cases}\nabla\widetilde{H}_1(\mathbf{x}), & x_1 > \frac{b}{a}x_2,\\ \mathbf{0}^T, & x_1 = \frac{b}{a}x_2,\\ \nabla\widetilde{H}_2(\mathbf{x}), & x_1 < \frac{b}{a}x_2.\end{cases}$$

Note that, if each entry of $\mathbf{B} = (b_{ij})_{i,j=1,2}$ and $\mathbf{C} = (c_{ij})_{i,j=1,2}$ satisfies the assumptions of Lemma A.3 from the appendix, we can conclude that $\mathbf{G} \in \mathcal{C}^1(\mathcal{D})$ in a sufficiently small neighborhood $\mathcal{D}$ of $\mathbf{y}^*$ and that the first derivatives are Lipschitz continuous on $\mathcal{D}$. As a direct consequence of (5.79), the same would then hold true for $\mathbf{g}$.

Now we show that the entries b_{ij} and c_{ij} of of the matrices $\mathbf{B}$ and $\mathbf{C}$ satisfy the assumptions of Lemma A.3 from the appendix, that is

a) b_{ij} and c_{ij} are continuous on $\mathbb{R}^2_{>0}$,

b) b_{ij} and c_{ij} are constant on $\ker(\mathbf{\Lambda})$,

c) b_{ij} and c_{ij} are in $\mathcal{C}^1$ on $\mathbb{R}^2_{>0}\setminus\ker(\mathbf{\Lambda})$ and

d) $\lim_{\mathbf{x}\to\mathbf{x}_0}\nabla b_{ij}(\mathbf{x})$ and $\lim_{\mathbf{x}\to\mathbf{x}_0}\nabla c_{ij}(\mathbf{x})$ exist for all $\mathbf{x}_0 \in \ker(\mathbf{\Lambda})\cap\mathbb{R}^2_{>0}$

for $i, j \in \{1, 2\}$. First, note that $\mathbf{B}$ and $\mathbf{C}$ are constant on $\ker(\mathbf{\Lambda})$, and due to $\widetilde{H}_k \in \mathcal{C}^2$ for $k = 1, 2$, we find that each entry of the two matrices is continuously differentiable on $\mathbb{R}^2_{>0} \backslash \ker(\mathbf{\Lambda})$. Even more, since H is continuous we know that $b_{ij} \in \mathcal{C}(\mathbb{R}^2_{>0})$ for $i, j \in \{1, 2\}$.

We want to point out that if $\lim_{\mathbf{x}\to\mathbf{x}_0} \nabla \widetilde{H}_k(\mathbf{x})$ exists, this proves the continuity of c_{ij} as well as that $\lim_{\mathbf{x}\to\mathbf{x}_0} \nabla b_{ij}(\mathbf{x})$ exists for all $i, j \in \{1, 2\}$. Furthermore, for $\mathbf{x} \notin \ker(\mathbf{\Lambda})$ we find

$$\begin{aligned}\nabla c_{ij}(\mathbf{x}) &= \nabla(\mathbf{\Lambda x} \nabla \widetilde{H}_k(\mathbf{x}))_{ij} = \nabla((\mathbf{\Lambda x})_i \tfrac{\partial}{\partial x_j} \widetilde{H}_k(\mathbf{x})) \\ &= (\mathbf{\Lambda e}_i)^T \tfrac{\partial}{\partial x_j} \widetilde{H}_k(\mathbf{x}) + (\mathbf{\Lambda x})_i \nabla(\tfrac{\partial}{\partial x_j} \widetilde{H}_k(\mathbf{x})).\end{aligned}$$

Thus, $\lim_{\mathbf{x}\to\mathbf{x}_0} \nabla c_{ij}(\mathbf{x})$ exists, if both limits, $\lim_{\mathbf{x}\to\mathbf{x}_0} \nabla \widetilde{H}_k(\mathbf{x})$ as well as $\lim_{\mathbf{x}\to\mathbf{x}_0} \nabla(\frac{\partial}{\partial x_j} \widetilde{H}_k(\mathbf{x}))$ exist for $i, j, k \in \{1, 2\}$. To see that both limits exist for $\mathbf{x}_0 \in \ker(\mathbf{\Lambda}) \cap \mathbb{R}^2_{>0}$, we introduce $\Phi(z) = \frac{1-e^{-\Delta t z}}{z}$, so that

$$\widetilde{H}_k(\mathbf{x}) = \Phi(\tfrac{w_k(\mathbf{x})}{x_k}).$$

Hence, we have $\Phi \in \mathcal{C}^2(\mathbb{R} \setminus \{0\})$ and

$$\begin{aligned}\nabla \widetilde{H}_k(\mathbf{x}) =& \Phi'(\tfrac{w_k(\mathbf{x})}{x_k}) \left(\frac{\nabla w_k(\mathbf{x})}{x_k} + w_k(\mathbf{x}) \nabla \left(\frac{1}{x_k} \right) \right), \\ \nabla \frac{\partial \widetilde{H}_k(\mathbf{x})}{\partial x_j} =& \Phi''(\tfrac{w_k(\mathbf{x})}{x_k}) \left(\frac{\nabla w_k(\mathbf{x})}{x_k} + w_k(\mathbf{x}) \nabla \left(\frac{1}{x_k} \right) \right) \left(\frac{\frac{\partial w_k(\mathbf{x})}{\partial x_j}}{x_k} + w_k(\mathbf{x}) \frac{\partial \left(\frac{1}{x_k} \right)}{\partial x_j} \right) \\ &+ \Phi'(\tfrac{w_k(\mathbf{x})}{x_k}) \nabla \left(\frac{\frac{\partial}{\partial x_j} w_k(\mathbf{x})}{x_k} + w_k(\mathbf{x}) \tfrac{\partial}{\partial x_j} \left(\frac{1}{x_k} \right) \right).\end{aligned} \tag{5.82}$$

As $\lim_{\mathbf{x}\to\mathbf{x}_0} \frac{w_k(\mathbf{x})}{x_k} = 0$ for $\mathbf{x}_0 \in \ker(\mathbf{\Lambda}) \cap \mathbb{R}^2_{>0}$, see (5.75), we are interested in the limits of the first two derivatives of Φ at $z = 0$. By l'Hospital's rule, a straightforward calculation yields

$$\lim_{z\to 0} \Phi'(z) = -\frac{(\Delta t)^2}{2} \quad \text{and} \quad \lim_{z\to 0} \Phi''(z) = \frac{(\Delta t)^3}{3}. \tag{5.83}$$

In addition, due to (5.75), we know that ∇w_k is a constant function for $k = 1, 2$, which means that

$$\begin{aligned}\nabla \left(\frac{\frac{\partial}{\partial x_j} w_k(\mathbf{x})}{x_k} + w_k(\mathbf{x}) \frac{\partial}{\partial x_j} \left(\frac{1}{x_k} \right) \right) =& \frac{\partial}{\partial x_j} w_k(\mathbf{x}) \nabla \left(\frac{1}{x_k} \right) \\ &+ \nabla w_k(\mathbf{x}) \frac{\partial}{\partial x_j} \left(\frac{1}{x_k} \right) \\ &+ w_k(\mathbf{x}) \nabla \left(\frac{\partial}{\partial x_j} \left(\frac{1}{x_k} \right) \right).\end{aligned}$$

It thus follows from (5.82) and (5.83) that $\lim_{\mathbf{x}\to\mathbf{x}_0} \nabla \widetilde{H}_k(\mathbf{x})$ as well as $\lim_{\mathbf{x}\to\mathbf{x}_0} \nabla(\frac{\partial}{\partial x_j} \widetilde{H}_k(\mathbf{x}))$ exist for all $i, j, k \in \{1, 2\}$ and each $\mathbf{x}_0 \in \ker(\mathbf{\Lambda}) \cap \mathbb{R}^2_{>0}$. This finishes this part of the proof.

b) First note that (5.79) implies that

$$\mathbf{g}(\mathbf{x}) = \mathbf{x} + \mathbf{B}\mathbf{G}(\mathbf{x})$$

with

$$\mathbf{B} = \mathbf{I} + \frac{1}{2}\Delta t\varphi(\Delta t(ac+b))\mathbf{\Lambda}.$$

Hence, $\mathbf{g} \in \mathcal{C}^2$ if and only if $\mathbf{BG} \in \mathcal{C}^2$. In general, (5.81) can be expressed by

$$\mathbf{DG}(\mathbf{x}) = \mathbf{\Lambda} H(\mathbf{x}) + \mathbf{\Lambda x}\begin{cases}\nabla\widetilde{H}_1(\mathbf{x}), & x_1 > \frac{b}{a}x_2,\\ \mathbf{c}(\mathbf{x}), & x_1 = \frac{b}{a}x_2,\\ \nabla\widetilde{H}_2(\mathbf{x}), & x_1 < \frac{b}{a}x_2\end{cases}$$

with an arbitrary function $\mathbf{c}\colon \mathbb{R}^2_{>0} \to \mathbb{R}^2_{>0}$. Our strategy is to use Lemma A.2 to conclude that the first partial derivative of the first column of

$$\mathbf{B\Lambda x}\begin{cases}\nabla\widetilde{H}_1(\mathbf{x}), & x_1 > \frac{b}{a}x_2,\\ \mathbf{c}(\mathbf{x}), & x_1 = \frac{b}{a}x_2,\\ \nabla\widetilde{H}_2(\mathbf{x}), & x_1 < \frac{b}{a}x_2\end{cases}$$

does not exist at $\mathbf{y}^*$. To that end, we prove that $\mathbf{T}(\mathbf{x}) = \mathbf{x}$ satisfies

$$\partial_j\mathbf{T}(\mathbf{x}) = \mathbf{e}_j \notin \ker(\mathbf{B\Lambda}),$$

and that

$$\lim_{\mathbf{x}\to\mathbf{y}^*}\nabla\widetilde{H}_1(\mathbf{x}) \neq \lim_{\mathbf{x}\to\mathbf{y}^*}\nabla\widetilde{H}_2(\mathbf{x}), \tag{5.84}$$

which shows the claim independently of $\mathbf{c}(\mathbf{y}^*)$. Indeed, the matrix $\mathbf{B}$ satisfies $\mathbf{B}\mathbf{y}^* = \mathbf{y}^*$ and $\mathbf{B}\bar{\mathbf{y}} = \mu\bar{\mathbf{y}}$, where $\bar{\mathbf{y}} = (1,-1)^T$ and

$$\mu = 1 - \frac{1}{2}\varphi(\Delta t(ac+b))\Delta t(ac+b).$$

Using $z = -\Delta t(ac+b) < 0$ we have

$$\mu = 1 + \frac{1}{2}(1-e^z) = \frac{3}{2} - \frac{1}{2}e^z > 0,$$

which means that $\mathbf{B}$ is invertible, and hence $\ker(\mathbf{B\Lambda}) = \ker(\mathbf{\Lambda})$. Therefore, we obtain $\partial_j\mathbf{T}(\mathbf{x}) \notin \ker(\mathbf{\Lambda}) = \ker(\mathbf{B\Lambda})$.

For proving (5.84) we use (5.83), (5.82) and $\mathbf{w}(\mathbf{y}^*) = \mathbf{0}$ to find

$$\lim_{\mathbf{x}\to\mathbf{y}^*}\nabla\widetilde{H}_i(\mathbf{x}) = -\frac{(\Delta t)^2}{2}\frac{\nabla w_i(\mathbf{y}^*)}{y_i^*}.$$

Let us now suppose (5.84) is not satisfied and recall that $w_1 = -\frac{1}{c}w_2$ holds true because of Lemma 5.33. Hence, we would have

$$\frac{\nabla w_2(\mathbf{y}^*)}{y_2^*} = -\frac{\nabla w_2(\mathbf{y}^*)}{cy_1^*}. \tag{5.85}$$

From (5.75) it follows that ∇w_2 is constant. We observe that $\nabla w_2(\mathbf{x}) \neq \mathbf{0}$ for all $\mathbf{x}$ as otherwise even $\mathbf{Dw}(\mathbf{x}) = \mathbf{0}$ for all $\mathbf{x}$, which contradicts Lemma 5.33

as (5.75) would imply that $\mathbf{w}(\mathbf{y}^n)=\mathbf{0}$ for all $\mathbf{y}^n$. Hence, without loss of generality, assume $\partial_1 w_2(\mathbf{x})\neq 0$. Then (5.85) implies

$$\partial_1 w_2(\mathbf{x})(cy_1^*+y_2^*)=0,$$

which is not true since $\partial_1 w_2(\mathbf{x})\neq 0$ and $cy_1^*+y_2^*>0$. Hence, (5.84) is true, so that Lemma A.2 implies $\mathbf{BG}\notin\mathcal{C}^2$ in any neighborhood of $\mathbf{y}^*$, and thus, the same holds for $\mathbf{g}$.

□

Note that part b) of Theorem 5.34 means that the assumptions of [IKM22a, Theorem 2.9] are not fulfilled, while those of the generalization, Theorem 5.4 are satisfied.

Remark 5.35. A numerical calculation shows that the stability function R from (5.80) with $\Delta t\,\mathrm{trace}(\mathbf{S}^-)=-z$ satisfies $|R(z)|<1$ for $z\in(z^*,0]$ with $-3.9924\leq z^*\leq -3.9923$. Hence, the stability region of GeCo2 when applied to (5.2), (5.4) is almost twice as big as the one of the underlying Heun scheme which is $(-2,0]$.

To investigate $N\times N$ systems one needs to generalize Lemma 5.33 and Lemma A.3 from the appendix, which is outside the scope of the present work.

5.4.5 Generalized BBKS

When it comes to the analysis of gBBKS schemes, we face similar obstacles as for GeCo2. The aim of this work is to present results from [IKMM23] giving a first insight into the stability properties of these schemes. As done for GeCo2 we will discuss at the end of this section an ansatz to generalize the following analysis.

Stability of first order gBBKS Schemes

When applied to the system of differential equations (5.4), i. e. $\mathbf{y}'=\mathbf{\Lambda}\mathbf{y}$ with $\mathbf{\Lambda}=\left(\begin{smallmatrix}-a & bc\\ ac & -b\end{smallmatrix}\right)$, the first order gBBKS schemes (gBBKS1) are given by

$$\mathbf{y}^{n+1}=\mathbf{y}^n+\Delta t\mathbf{\Lambda}\mathbf{y}^n\left(\prod_{m\in M^n}\frac{y_m^{n+1}}{\sigma_m^n}\right)^{r^n},\quad i=1,2,\tag{5.86}$$

where

$$M^n=\{m\in\{1,2\}\mid(\mathbf{\Lambda}\mathbf{y}^n)_m<0\}.$$

In this section we investigate the stability properties of gBBKS schemes by first proving that the assumptions of Theorem 5.4 are met. The existence and uniqueness of a function $\mathbf{g}$ generating the iterates from (5.86), i. e. $\mathbf{y}^{n+1}=\mathbf{g}(\mathbf{y}^n)$, is already proven in [AKM20]. Thereby, $\mathbf{g}$ is given by the unique solution to some equation

$$\mathbf{F}(\mathbf{x},\mathbf{g}(\mathbf{x}))=\mathbf{0},$$

where $\mathbf{F}\colon\mathbb{R}^2_{>0}\times\mathbb{R}^2_{>0}\to\mathbb{R}^2$ with $(\mathbf{x},\mathbf{y})\mapsto\mathbf{F}(\mathbf{x},\mathbf{y})$. In the following we denote by

$$\mathbf{D_x F}(\mathbf{x},\mathbf{y})=\frac{\partial\mathbf{F}}{\partial\mathbf{x}}(\mathbf{x},\mathbf{y}),$$
$$\mathbf{D_y F}(\mathbf{x},\mathbf{y})=\frac{\partial\mathbf{F}}{\partial\mathbf{y}}(\mathbf{x},\mathbf{y})$$

the Jacobians of $\mathbf{F}$ with respect to $\mathbf{x}$ and $\mathbf{y}$, respectively.

An intuitive way of proving $\mathbf{g} \in \mathcal{C}^1(\mathcal{D})$, where $\mathcal{D}$ is a neighborhood of a fixed point $\mathbf{y}^*$ of $\mathbf{g}$, is to use the implicit function theorem. Unfortunately, we will see in the following that in our case $\mathbf{F}$ is not differentiable on $\mathcal{D} \times \mathcal{D}$. Since the existence and uniqueness of the map $\mathbf{g}$ is already known here, the differentiability of $\mathbf{g}$ can be obtained by weaker assumptions on $\mathbf{F}$ as the next theorem states.

Theorem 5.36 ([LS14, Theorem 11.1])**.** Let $D \subseteq \mathbb{R}^2$ be open and $\mathbf{g}\colon D \to D$ be continuous in $\mathbf{x}_0$. Furthermore, let $\mathbf{F}\colon D \times D \to \mathbb{R}^2$ with $(\mathbf{x}, \mathbf{y}) \mapsto \mathbf{F}(\mathbf{x}, \mathbf{y})$ be differentiable in $(\mathbf{x}_0, \mathbf{g}(\mathbf{x}_0))^T$ and $\mathbf{D_y F}(\mathbf{x}_0, \mathbf{g}(\mathbf{x}_0))$ be invertible. Suppose that $\mathbf{F}(\mathbf{x}, \mathbf{g}(\mathbf{x})) = \mathbf{0}$ for all $\mathbf{x} \in D$, then also $\mathbf{g}$ is differentiable in $\mathbf{x}_0$ and

$$\mathbf{Dg}(\mathbf{x}_0) = -(\mathbf{D_y F}(\mathbf{x}_0, \mathbf{g}(\mathbf{x}_0)))^{-1} \mathbf{D_x F}(\mathbf{x}_0, \mathbf{g}(\mathbf{x}_0)).$$

Before we formulate the stability theorem for gBBKS1, we introduce some assumptions on the exponent r^n as well as σ_m^n from (5.86). In particular, $r^n > 0$ and $\sigma_m^n > 0$ may depend on $\mathbf{y}^n$ and hence will be interpreted as functions $r^n = r(\mathbf{y}^n)$ and $\sigma_m^n = \sigma_m(\mathbf{y}^n)$. For the analysis of the gBBKS1 schemes we do not further specify the expressions for the functions r or σ_m. Instead, we assume some reasonable properties such as that r and σ_m are positive for all $\Delta t \geq 0$. Furthermore, we require $\sigma_m(\mathbf{v}) = v_m$ whenever $\mathbf{v} \in \ker(\mathbf{\Lambda}) \cap \mathbb{R}^2_{>0}$ which is in agreement with the literature [AKM20, BBKS07, BRBM08]. To guarantee the regularity of the map generating the iterates $\mathbf{y}^n$, we also assume that r, σ_1 and σ_2 are in $\mathcal{C}^2$. In total, we prove the following theorem.

Theorem 5.37. Let $\mathbf{y}^* > \mathbf{0}$ be a steady state solution of (5.4), and assume $\sigma_1, \sigma_2, r \in \mathcal{C}^2(\mathbb{R}^2_{>0}, \mathbb{R}_{>0})$. Further, let $\mathcal{D}$ be a sufficiently small neighborhood of $\mathbf{y}^*$ and suppose that $\boldsymbol{\sigma}(\mathbf{v}) = \mathbf{v}$ for all $\mathbf{v} \in C = \ker(\mathbf{\Lambda}) \cap \mathcal{D}$. Then the map $\mathbf{g}$ generating the iterates of the gBBKS1 family, implicitly given by (5.86), satisfies $\mathbf{g}(\mathbf{v}) = \mathbf{v}$ for all steady states $\mathbf{v} \in C$ and the following statements hold.

a) The map $\mathbf{g}$ satisfies $\mathbf{g} \in \mathcal{C}^1(\mathcal{D})$ and $\mathbf{Dg}(\mathbf{y}^*) = \mathbf{I} + \Delta t \mathbf{\Lambda}$.

b) The first derivatives of $\mathbf{g}$ are bounded and Lipschitz continuous on $\mathcal{D}$.

c) The map $\mathbf{g}$ does not belong to $\mathcal{C}^2$ for any open neighborhood of $\mathbf{y}^*$, if $\Delta t \neq (ac + b)^{-1}$.

Proof. Before we start the proof of a), we make some preparatory considerations.

Since $(\mathbf{\Lambda y}^n)_1 = c(-a y_1^n + b y_2^n)$ and $c(\mathbf{\Lambda y}^n)_2 = -(\mathbf{\Lambda y}^n)_1$ we find

$$M^n = \begin{cases} \{1\}, & y_1^n > \frac{b}{a} y_2^n, \\ \emptyset, & y_1^n = \frac{b}{a} y_2^n, \\ \{2\}, & y_1^n < \frac{b}{a} y_2^n. \end{cases}$$

Hence, when applied to (5.4), (5.2) the scheme (5.86) turns into

$$\mathbf{y}^{n+1} = \mathbf{y}^n + \Delta t \mathbf{\Lambda y}^n \begin{cases} \left(\frac{y_1^{n+1}}{\sigma_1^n}\right)^{r^n}, & y_1^n > \frac{b}{a} y_2^n, \\ 1, & y_1^n = \frac{b}{a} y_2^n, \\ \left(\frac{y_2^{n+1}}{\sigma_2^n}\right)^{r^n}, & y_1^n < \frac{b}{a} y_2^n, \end{cases} \tag{5.87}$$

where $(\frac{b}{a} y_2^n, y_2^n)^T \in \ker(\mathbf{\Lambda})$ is a steady state solution of (5.4).

Recall that the map $\mathbf{g}$ generates the iterates $\mathbf{y}^n$, that is $\mathbf{y}^{n+1} = \mathbf{g}(\mathbf{y}^n)$. Hence, inserting $\mathbf{y}^n = \mathbf{v} \in C$ into equation (5.87) yields $\mathbf{y}^{n+1} = \mathbf{g}(\mathbf{v})$ on the left and $\mathbf{v}$ on the right, and thus $\mathbf{g}(\mathbf{v}) = \mathbf{v}$. Furthermore, we introduce the function $\mathbf{F}$ defined by

$$\begin{aligned}
\mathbf{F}&\colon \mathbb{R}^2_{>0} \times \mathbb{R}^2_{>0} \to \mathbb{R}^2,\\
\mathbf{F}(\mathbf{x},\mathbf{y}) &= \mathbf{y} - \mathbf{x} - \Delta t \boldsymbol{\Lambda} \mathbf{x} H(\mathbf{x},\mathbf{y}),\\
H(\mathbf{x},\mathbf{y}) &= \begin{cases} \widetilde{H}_1(\mathbf{x},\mathbf{y}), & x_1 > \frac{b}{a}x_2,\\ 1, & x_1 = \frac{b}{a}x_2,\\ \widetilde{H}_2(\mathbf{x},\mathbf{y}), & x_1 < \frac{b}{a}x_2, \end{cases}\\
\widetilde{H}_i(\mathbf{x},\mathbf{y}) &= \left(\frac{y_i}{\sigma_i(\mathbf{x})}\right)^{r(\mathbf{x})}, \quad i = 1,2,
\end{aligned} \tag{5.88}$$

which satisfies $\mathbf{F}(\mathbf{x},\mathbf{g}(\mathbf{x})) = \mathbf{0}$ for all $\mathbf{x} > \mathbf{0}$.

a) We first show that $\mathbf{F}$ is not differentiable on $\mathcal{D} \times \mathcal{D}$. For this, we choose $\mathbf{x}_0 \in C$ as well as $\mathbf{y}_0 > \mathbf{0}$ with $\frac{(\mathbf{y}_0)_1}{(\mathbf{x}_0)_1} \neq \frac{(\mathbf{y}_0)_2}{(\mathbf{x}_0)_2}$ and define $\Psi(\mathbf{x}) = H(\mathbf{x},\mathbf{y}_0)$. As a result of $\boldsymbol{\sigma}(\mathbf{x}_0) = \mathbf{x}_0$ and $r, \boldsymbol{\sigma} \in \mathcal{C}$ it follows that

$$\lim_{h\searrow 0} \Psi(\mathbf{x}_0 + h\mathbf{e}_1) = \lim_{h\searrow 0} \widetilde{H}_1(\mathbf{x}_0 + h\mathbf{e}_1, \mathbf{y}_0) = \left(\frac{(\mathbf{y}_0)_1}{\sigma_1(\mathbf{x}_0)}\right)^{r(\mathbf{x}_0)} = \left(\frac{(\mathbf{y}_0)_1}{(\mathbf{x}_0)_1}\right)^{r(\mathbf{x}_0)}.$$

Analogously, we obtain

$$\lim_{h\nearrow 0} \Psi(\mathbf{x}_0 + h\mathbf{e}_1) = \lim_{h\nearrow 0} \widetilde{H}_2(\mathbf{x}_0 + h\mathbf{e}_1, \mathbf{y}_0) = \left(\frac{(\mathbf{y}_0)_2}{\sigma_2(\mathbf{x}_0)}\right)^{r(\mathbf{x}_0)} = \left(\frac{(\mathbf{y}_0)_2}{(\mathbf{x}_0)_2}\right)^{r(\mathbf{x}_0)},$$

which shows that $\Psi(\mathbf{x}_0 + h\mathbf{e}_1)$ possesses several accumulation points as $h \to 0$, and hence, part b) of Lemma A.2 from the appendix with $\boldsymbol{\Phi}(\mathbf{x}) = \boldsymbol{\Lambda}\mathbf{x}$ implies that the 1st partial derivative of $\mathbf{F}$ does not exist.

As mentioned above, this means that we can not apply the implicit function theorem to $\mathbf{F}$ on $\mathcal{D} \times \mathcal{D}$ in order to prove that $\mathbf{g} \in \mathcal{C}^1(\mathcal{D})$. Nevertheless, $\mathbf{F}$ is differentiable in $(\mathbf{x},\mathbf{y}) \in E = \mathcal{D} \setminus \ker(\boldsymbol{\Lambda}) \times \mathcal{D}$, since in this case we have

$$\mathbf{F}(\mathbf{x},\mathbf{y}) = \mathbf{y} - \mathbf{x} - \Delta t \boldsymbol{\Lambda}\mathbf{x}\left(\frac{y_i}{\sigma_i(\mathbf{x})}\right)^{r(\mathbf{x})}, \quad i = \begin{cases} 1, & x_1 > \frac{b}{a}x_2,\\ 2, & x_1 < \frac{b}{a}x_2 \end{cases} \tag{5.89}$$

with $\sigma_1, \sigma_2, r \in \mathcal{C}^2(\mathbb{R}^2_{>0}, \mathbb{R}_{>0})$. In order to show that $\mathbf{g} \in \mathcal{C}^1(\mathcal{D} \setminus \ker(\boldsymbol{\Lambda}))$, we first show that the inverse of $\mathbf{D_y F}(\mathbf{x},\mathbf{g}(\mathbf{x}))$ exists for all $\mathbf{x} \in \mathcal{D} \setminus \ker(\boldsymbol{\Lambda})$. It is straightforward to verify that

$$\mathbf{D_y F}(\mathbf{x},\mathbf{g}(\mathbf{x})) = \mathbf{I} - \Delta t \boldsymbol{\Lambda}\mathbf{x} \nabla_{\mathbf{y}} \widetilde{H}_i(\mathbf{x},\mathbf{g}(\mathbf{x})) = \mathbf{I} - \Delta t \boldsymbol{\Lambda}\mathbf{x}\mathbf{e}_i^T \frac{r(\mathbf{x})}{\sigma_i(\mathbf{x})}\left(\frac{g_i(\mathbf{x})}{\sigma_i(\mathbf{x})}\right)^{r(\mathbf{x})-1}$$

holds for $\mathbf{x} \notin C$. Introducing the vectors

$$\mathbf{v}^{(i)}(\mathbf{x}) = \Delta t \boldsymbol{\Lambda}\mathbf{x}\frac{r(\mathbf{x})}{\sigma_i(\mathbf{x})}\left(\frac{g_i(\mathbf{x})}{\sigma_i(\mathbf{x})}\right)^{r(\mathbf{x})-1} = \Delta t \boldsymbol{\Lambda}\mathbf{x}\frac{r(\mathbf{x})}{g_i(\mathbf{x})}\left(\frac{g_i(\mathbf{x})}{\sigma_i(\mathbf{x})}\right)^{r(\mathbf{x})} \tag{5.90}$$

for i from (5.89), we can write the Jacobian in the compact form

$$\mathbf{D_y F}(\mathbf{x},\mathbf{g}(\mathbf{x})) = \mathbf{I} - \mathbf{v}^{(i)}(\mathbf{x})\mathbf{e}_i^T. \tag{5.91}$$

Note that due to (5.91), the Jacobian of $\mathbf{F}$ with respect to $\mathbf{y}$ is a triangular matrix, depending on i from (5.89). Nevertheless, in either case we find

$$\det(\mathbf{D_y F}(\mathbf{x},\mathbf{g}(\mathbf{x}))) = 1 - v_i^{(i)}(\mathbf{x}). \tag{5.92}$$

Now, we know that $(\mathbf{\Lambda x})_i < 0$ for i form (5.89) by construction of the gBBKS schemes, which in particular means that

$$v_i^{(i)}(\mathbf{x}) \neq 1. \tag{5.93}$$

As a result of (5.92), (5.93) the inverse of $\mathbf{D_y F}(\mathbf{x},\mathbf{g}(\mathbf{x}))$ exists.

Considering a zero $(\mathbf{x}_0,\mathbf{g}(\mathbf{x}_0)) \in E$ of $\mathbf{F}$, the implicit function theorem thus provides the existence of a unique $\mathcal{C}^1$-map $\widetilde{\mathbf{g}}$ satisfying $\mathbf{F}(\mathbf{x},\widetilde{\mathbf{g}}(\mathbf{x})) = \mathbf{0}$ in a sufficiently small neighborhood of $(\mathbf{x}_0,\mathbf{g}(\mathbf{x}_0))$. Since $\mathbf{g}$ and $\widetilde{\mathbf{g}}$ are unique, we find $\mathbf{g} = \widetilde{\mathbf{g}}$, and since $\mathbf{x}_0$ was arbitrary, we have shown that $\mathbf{g} \in \mathcal{C}^1$ on $\mathcal{D} \setminus \ker(\mathbf{\Lambda})$, and in particular

$$\mathbf{Dg}(\mathbf{x}) = -(\mathbf{D_y F}(\mathbf{x},\mathbf{g}(\mathbf{x})))^{-1}\mathbf{D_x F}(\mathbf{x},\mathbf{g}(\mathbf{x})) \tag{5.94}$$

for $\mathbf{x} \in \mathcal{D} \setminus \ker(\mathbf{\Lambda})$. It thus remains to show that $\mathbf{g}\colon D \to D$ is also differentiable in any $\mathbf{x} \in \ker(\mathbf{\Lambda}) \cap \mathcal{D} = C$ and that the first derivatives are continuous in any $\mathbf{x} \in C$.

To prove the differentiability of $\mathbf{g}$ in any $\mathbf{x} \in C$ we make use of Theorem 5.36, and hence we have to prove the following.

1. The map $\mathbf{g}$ is continuous in any $\mathbf{x} \in C$.
2. The map $\mathbf{F}$ is differentiable in $(\mathbf{x},\mathbf{g}(\mathbf{x}))$ for all $\mathbf{x} \in C$.
3. The Jacobian $\mathbf{D_y F}(\mathbf{x},\mathbf{g}(\mathbf{x}))$ with respect to $\mathbf{y}$ is invertible for all $\mathbf{x} \in C$.

If we have shown these properties, then Theorem 5.36 together with the considerations above implies that (5.94) even holds for all $\mathbf{x} \in \mathcal{D}$.

We first prove that $\mathbf{g}$ is continuous on C. Since gBBKS schemes are positive and conserve all linear invariants, we find from (5.4) that

$$\min\{1,c\}\|\mathbf{g}(\mathbf{x})\|_1 \leq g_1(\mathbf{x}) + cg_2(\mathbf{x}) = x_1 + cx_2 \leq \max\{1,c\}\|\mathbf{x}\|_1. \tag{5.95}$$

Now, $\|\mathbf{x}\|_1$ is bounded on a sufficiently small neighborhood $\mathcal{D}$ of $\mathbf{y}^*$ as we can make sure that the closure of $\mathcal{D}$ is contained in the domain of $\mathbf{g}$. And since norms on $\mathbb{R}^2$ are equivalent, we even find from (5.95) that $\|\mathbf{g}\|$ is bounded on C. As a result, $H(\cdot,\mathbf{g}(\cdot))$ is bounded on $\mathcal{D}$ since the reciprocal of $\boldsymbol{\sigma} \in \mathcal{C}^2$ as well as $r \in \mathcal{C}^2$ are bounded on a sufficiently small $\mathcal{D}$. It thus follows that $\mathbf{\Lambda x}H(\mathbf{x},\mathbf{g}(\mathbf{x}))$ tends to $\mathbf{0}$ as $\mathbf{x} \to \mathbf{y}^*$. From (5.88) with $\mathbf{y} = \mathbf{g}(\mathbf{x})$ we therefore obtain

$$\lim_{\mathbf{x}\to\mathbf{y}^*} \mathbf{g}(\mathbf{x}) = \mathbf{y}^* = \mathbf{g}(\mathbf{y}^*),$$

which means that $\mathbf{g}\colon D \to D$ is continuous in all $\mathbf{x} \in C$.

Next, we show that $\mathbf{F}$ is differentiable in $(\mathbf{x},\mathbf{g}(\mathbf{x}))$ for all $\mathbf{x} \in C$. For this consider an $\mathbf{x}_0 \in C$ and set $\mathbf{y}_0 = \mathbf{g}(\mathbf{x}_0)$. Note that $\Psi = H(\cdot,\mathbf{y}_0)$ is continuous in $\mathbf{x}_0 \in C$ with $\Psi(\mathbf{x}_0) = 1$ since $\mathbf{g}(\mathbf{x}_0) = \boldsymbol{\sigma}(\mathbf{x}_0) = \mathbf{x}_0$.

In this case, part a) of Lemma A.2 from the appendix with $\mathbf{\Phi}(\mathbf{x}) = \mathbf{\Lambda x}$

yields

$$\mathbf{D_x F}(\mathbf{x}_0, \mathbf{g}(\mathbf{x}_0)) = -\mathbf{I} - \Delta t \boldsymbol{\Lambda}. \tag{5.96}$$

Furthermore, as $\boldsymbol{\Lambda}\mathbf{x}\widetilde{H}_i(\mathbf{x}, \mathbf{y}) = \mathbf{0}$ for all $\mathbf{x} \in C$ and $\mathbf{y} \in \mathbb{R}^2_{>0}$, it follows immediately that

$$\mathbf{D_y F}(\mathbf{x}_0, \mathbf{g}(\mathbf{x}_0)) = \mathbf{I}, \tag{5.97}$$

which shows that $\mathbf{F}$ is partially differentiable in $(\mathbf{x}_0, \mathbf{g}(\mathbf{x}_0))$. To prove that $\mathbf{F}$ is differentiable in $(\mathbf{x}_0, \mathbf{g}(\mathbf{x}_0))$, we show that the partial derivatives are continuous in $(\mathbf{x}_0, \mathbf{g}(\mathbf{x}_0))$. Therefore, we consider the case $\mathbf{x} \notin C$ and differentiate $\mathbf{F}$ from (5.88) with respect to $\mathbf{x}$ and $\mathbf{y}$. We have

$$\mathbf{D_x F}(\mathbf{x}, \mathbf{g}(\mathbf{x})) = -\mathbf{I} - \Delta t \left(\boldsymbol{\Lambda}\widetilde{H}_i(\mathbf{x}, \mathbf{g}(\mathbf{x})) + \boldsymbol{\Lambda}\mathbf{x}\nabla_{\mathbf{x}}\widetilde{H}_i(\mathbf{x}, \mathbf{g}(\mathbf{x})) \right), \tag{5.98}$$

where the gradient denotes a row vector and

$$i = \begin{cases} 1, & x_1 > \frac{b}{a}x_2, \\ 2, & x_1 < \frac{b}{a}x_2. \end{cases}$$

Now, since $\mathbf{g}, \boldsymbol{\sigma} > \mathbf{0}$ we can write $\widetilde{H}_i(\mathbf{x}, \mathbf{g}(\mathbf{x})) = e^{r(\mathbf{x})\ln\left(\frac{g_i(\mathbf{x})}{\sigma_i(\mathbf{x})}\right)}$, from which it follows that

$$\nabla_{\mathbf{x}}\widetilde{H}_i(\mathbf{x}, \mathbf{g}(\mathbf{x})) = \widetilde{H}_i(\mathbf{x}, \mathbf{g}(\mathbf{x})) \left(\nabla_{\mathbf{x}} r(\mathbf{x}) \ln\left(\frac{g_i(\mathbf{x})}{\sigma_i(\mathbf{x})}\right) - r(\mathbf{x})\frac{\nabla_{\mathbf{x}}\sigma_i(\mathbf{x})}{\sigma_i(\mathbf{x})} \right) \tag{5.99}$$

since

$$\nabla_{\mathbf{x}} \ln\left(\frac{y_i}{\sigma_i(\mathbf{x})}\right) = \nabla_{\mathbf{x}} \ln(y_i) - \nabla_{\mathbf{x}} \ln(\sigma_i(\mathbf{x})) = -\frac{\nabla_{\mathbf{x}}\sigma_i(\mathbf{x})}{\sigma_i(\mathbf{x})}.$$

Plugging (5.99) into (5.98), we find

$$\begin{aligned}\mathbf{D_x F}(\mathbf{x}, \mathbf{g}(\mathbf{x})) = -\mathbf{I} - \Delta t \widetilde{H}_i(\mathbf{x}, \mathbf{g}(\mathbf{x})) \Bigg(\boldsymbol{\Lambda} + \boldsymbol{\Lambda}\mathbf{x} \Bigg(&\nabla_{\mathbf{x}} r(\mathbf{x}) \ln\left(\frac{g_i(\mathbf{x})}{\sigma_i(\mathbf{x})}\right) \\ &- r(\mathbf{x})\frac{\nabla_{\mathbf{x}}\sigma_i(\mathbf{x})}{\sigma_i(\mathbf{x})} \Bigg) \Bigg).\end{aligned} \tag{5.100}$$

Furthermore, $\boldsymbol{\Lambda}\mathbf{x}_0 = \mathbf{0}$ for $\mathbf{x}_0 \in C$ together with $\sigma_1, \sigma_2, r \in \mathcal{C}^2$ as well as equation (5.100) yield

$$\lim_{\mathbf{x}\to\mathbf{x}_0} \mathbf{D_x F}(\mathbf{x}, \mathbf{g}(\mathbf{x})) = -\mathbf{I} - \Delta t \lim_{\mathbf{x}\to\mathbf{x}_0} \widetilde{H}_i(\mathbf{x}, \mathbf{g}(\mathbf{x}))\boldsymbol{\Lambda} = -\mathbf{I} - \Delta t \boldsymbol{\Lambda}. \tag{5.101}$$

Moreover, due to (5.91) and since $\mathbf{v}^{(i)}$ is continuous with $\mathbf{v}^{(i)}(\mathbf{x}_0) = \mathbf{0}$ for $\mathbf{x}_0 \in C$, we find

$$\lim_{\mathbf{x}\to\mathbf{x}_0} \mathbf{D_y F}(\mathbf{x}, \mathbf{g}(\mathbf{x})) = \mathbf{I} - \mathbf{v}^{(i)}(\mathbf{x}_0)\mathbf{e}_i^T = \mathbf{I}. \tag{5.102}$$

As a result of (5.96), (5.101) and (5.97), (5.102), we thus know that all partial first derivatives of $\mathbf{F}$ are continuous in $(\mathbf{x}_0, \mathbf{g}(\mathbf{x}_0))$ for all $\mathbf{x}_0 \in C$, which implies that $\mathbf{F}$ is differentiable in $(\mathbf{x}_0, \mathbf{g}(\mathbf{x}_0))$ for all $\mathbf{x}_0 \in C$.

Finally, due to (5.97) we know that $\mathbf{D}_{\mathbf{y}}\mathbf{F}(\mathbf{x}_0, \mathbf{g}(\mathbf{x}_0))$ is invertible for all $\mathbf{x}_0 \in C$.

Altogether, all requirements of Theorem 5.36 are fulfilled, which implies that $\mathbf{g}$ is differentiable on C and that

$$\mathbf{D}\mathbf{g}(\mathbf{x}_0) = -(\mathbf{D}_{\mathbf{y}}\mathbf{F}(\mathbf{x}_0, \mathbf{g}(\mathbf{x}_0)))^{-1}\mathbf{D}_{\mathbf{x}}\mathbf{F}(\mathbf{x}_0, \mathbf{g}(\mathbf{x}_0)) \tag{5.103}$$

holds for all $\mathbf{x}_0 \in C$. Moreover, all entries of the inverse of $\mathbf{D}_{\mathbf{y}}\mathbf{F}(\mathbf{x}, \mathbf{g}(\mathbf{x}))$ are continuous functions of $\mathbf{x}$, which proves that $\mathbf{g} \in \mathcal{C}^1(\mathcal{D})$. Finally, (5.96) and (5.97) yield

$$\mathbf{D}\mathbf{g}(\mathbf{y}^*) = \mathbf{I} + \Delta t \mathbf{\Lambda}.$$

b) In this part, we use the equations (5.100) and

$$(\mathbf{D}_{\mathbf{y}}\mathbf{F}(\mathbf{x}, \mathbf{g}(\mathbf{x})))^{-1} = \begin{cases} \frac{1}{1-v_1^{(1)}(\mathbf{x})}\begin{pmatrix} 1 & 0 \\ v_2^{(1)}(\mathbf{x}) & 1-v_1^{(1)}(\mathbf{x}) \end{pmatrix}, & x_1 > \frac{b}{a}x_2, \\ \frac{1}{1-v_2^{(2)}(\mathbf{x})}\begin{pmatrix} 1-v_2^{(2)}(\mathbf{x}) & v_1^{(2)}(\mathbf{x}) \\ 0 & 1 \end{pmatrix}, & x_1 < \frac{b}{a}x_2 \end{cases} \tag{5.104}$$

to show that the first derivatives of $\mathbf{g}$ are Lipschitz continuous on a sufficiently small neighborhood $\mathcal{D}$ of $\mathbf{y}^*$. For this, we make use of the fact that the set of bounded Lipschitz continuous functions is closed under summation, multiplication and composition. Hence, all we need to prove is that each entry in the matrices (5.100) and (5.104) is bounded and Lipschitz continuous on $\mathcal{D}$, and to use the fact that the natural logarithm and each exponential function are locally Lipschitz continuous.

To bound the corresponding functions, we choose $\mathcal{D}$ in such a way that g_i, σ_i and $1 - v_i^{(i)}$ have an upper bound $C_i > 0$ and lower bound $c_i > 0$. This is possible by choosing $\overline{\mathcal{D}} \subseteq D$ since these functions are continuous at $\mathbf{y}^*$ and satisfy $\mathbf{g}(\mathbf{y}^*) = \boldsymbol{\sigma}(\mathbf{y}^*) = \mathbf{y}^* > \mathbf{0}$ as well as $1 - v_i^{(i)}(\mathbf{y}^*) = 1$. As a result, even the first two derivatives of $\boldsymbol{\sigma}$ and r are bounded on $\mathcal{D}$. This way, we can compute the Lipschitz constants of $\boldsymbol{\sigma}$, its first derivatives and its reciprocal by using the mean value theorem, see [AE08, Remark 8.12 (b)] for the details. Analogously, $\mathbf{g}$ as well as $\frac{1}{g_i}$ are bounded Lipschitz continuous functions for $i = 1, 2$ as their first derivatives are bounded on $\mathcal{D}$. By this reasoning, it is straightforward to verify that each matrix entry in (5.100) and (5.104) is a bounded Lipschitz continuous function.

c) Assume that $\mathbf{g} \in \mathcal{C}^2$ for some appropriate neighborhood of $\mathbf{y}^*$. Introducing

$$d_{jk}(\mathbf{x}) = (\mathbf{D}_{\mathbf{x}}\mathbf{F}(\mathbf{x}, \mathbf{g}(\mathbf{x})))_{jk},$$

equations (5.104) and (5.103) yield

$$\partial_1 g_2(\mathbf{x}) = -\begin{cases} \frac{v_2^{(1)}(\mathbf{x})}{1-v_1^{(1)}(\mathbf{x})} d_{11}(\mathbf{x}) + d_{21}(\mathbf{x}) & , x_1 \geq \frac{b}{a}x_2 \\ \frac{1}{1-v_2^{(2)}(\mathbf{x})} d_{21}(\mathbf{x}) & , x_1 \leq \frac{b}{a}x_2 \end{cases}. \tag{5.105}$$

Our strategy is to compute $\partial_2\partial_1 g_2(\mathbf{y}^*)$ and derive $1 = \Delta t(ca + b)$ from it.

Using $\mathbf{v}^{(i)}(\mathbf{y}^*)=\mathbf{0}$ we get from (5.105)

$$-\partial_2\partial_1 g_2(\mathbf{y}^*)=\partial_2 v_2^{(1)}(\mathbf{y}^*)d_{11}(\mathbf{y}^*)+\partial_2 d_{21}(\mathbf{y}^*)$$

as well as

$$-\partial_2\partial_1 g_2(\mathbf{y}^*)=\partial_2 d_{21}(\mathbf{y}^*)+\partial_2 v_2^{(2)}(\mathbf{y}^*)d_{21}(\mathbf{y}^*).$$

As a result, we obtain

$$\partial_2 v_2^{(1)}(\mathbf{y}^*)d_{11}(\mathbf{y}^*)=\partial_2 v_2^{(2)}(\mathbf{y}^*)d_{21}(\mathbf{y}^*). \tag{5.106}$$

Using (5.90), we find that

$$\partial_j v_k^{(i)}(\mathbf{y}^*)=\Delta t\lambda_{kj}\frac{r(\mathbf{y}^*)}{y_i^*},$$

and from (5.96), we know that $d_{jk}(\mathbf{y}^*)=-(\mathbf{I}+\Delta t\mathbf{\Lambda})_{jk}$, so that (5.106) reads

$$-\Delta t\lambda_{22}\frac{r(\mathbf{y}^*)}{y_1^*}(\mathbf{I}+\Delta t\mathbf{\Lambda})_{11}=-\Delta t\lambda_{22}\frac{r(\mathbf{y}^*)}{y_2^*}(\mathbf{I}+\Delta t\mathbf{\Lambda})_{21}.$$

Using the fact that $r>0$ and $\lambda_{22}\neq 0$, this equation reduces to

$$(\mathbf{I}+\Delta t\mathbf{\Lambda})_{11}=\frac{y_1^*}{y_2^*}(\mathbf{I}+\Delta t\mathbf{\Lambda})_{21}=\frac{y_1^*}{y_2^*}(\Delta t\mathbf{\Lambda})_{21},$$

or equivalently,

$$1=\Delta t\left(\tfrac{y_1^*}{y_2^*}\lambda_{21}-\lambda_{11}\right)\overset{(5.4)}{=}\Delta t(ca+b),$$

which finishes also this part of the proof. □

It is worth mentioning that part c) of the above theorem demonstrates, that in general $\mathbf{g}\notin\mathcal{C}^2$. As a result we could not apply [IKM22a, Theorem 2.9], however, the generalization Theorem 5.4 can be applied, which gives us the following statements due to $\mathbf{Dg}(\mathbf{y}^*)=\mathbf{I}+\Delta t\mathbf{\Lambda}$.

Corollary 5.38. Let $\mathbf{y}^*>\mathbf{0}$ be an arbitrary steady state of (5.4). Under the assumptions of Theorem 5.37, the gBBKS1 schemes have the same stability function as the underlying Runge–Kutta method, i. e. $R(z)=1+z$ and the following holds.

a) If $|R(-(ac{+}b)\Delta t)|<1$, then $\mathbf{y}^*$ is a stable fixed point of each gBBKS1 scheme and there exists a $\delta>0$, such that $\left(\begin{smallmatrix}1\\c\end{smallmatrix}\right)^T\mathbf{y}^0=\left(\begin{smallmatrix}1\\c\end{smallmatrix}\right)^T\mathbf{y}^*$ and $\|\mathbf{y}^0-\mathbf{y}^*\|<\delta$ imply $\mathbf{y}^n\to\mathbf{y}^*$ as $n\to\infty$.

b) If $|R(-(ac+b)\Delta t)|>1$, then $\mathbf{y}^*$ is an unstable fixed point of each gBBKS1 scheme.

Stability of second order gBBKS schemes

In this subsection we investigate the gBBKS2(α) schemes (gBBKS2) applied to (5.4), (5.2), which can be written in the form

$$\mathbf{y}^{(2)} = \mathbf{y}^n + \alpha\Delta t\mathbf{\Lambda}\mathbf{y}^n\left(\prod_{j\in J^n}\frac{y_j^{(2)}}{\pi_j^n}\right)^{q^n}, \tag{5.107a}$$

$$\mathbf{y}^{n+1} = \mathbf{y}^n + \Delta t\left(\left(1-\frac{1}{2\alpha}\right)\mathbf{\Lambda}\mathbf{y}^n + \frac{1}{2\alpha}\mathbf{\Lambda}\mathbf{y}^{(2)}\right)\left(\prod_{m\in M^n}\frac{y_m^{n+1}}{\sigma_m^n}\right)^{r^n}, \tag{5.107b}$$

for $i=1,2$, $\alpha\geq\frac{1}{2}$ and

$$\begin{aligned} J^n &= \{j\in\{1,2\}\mid (\mathbf{\Lambda}\mathbf{y}^n)_j<0\},\\ M^n &= \left\{m\in\{1,2\}\;\middle|\;\left(1-\frac{1}{2\alpha}\right)(\mathbf{\Lambda}\mathbf{y}^n)_m+\frac{1}{2\alpha}(\mathbf{\Lambda}\mathbf{y}^{(2)})_m<0\right\}. \end{aligned}$$

Similarly to the gBBKS1 case, we introduce functions $r, q, \boldsymbol{\pi}$ and $\boldsymbol{\sigma}$ to describe the dependence of the parameters on $\mathbf{y}^n$. Note that $\boldsymbol{\sigma}$ can depend on $\mathbf{y}^n$ as well as $\mathbf{y}^{(2)}$, see [AKM20, BBKS07, BRBM08], and thus will be described by a map $\boldsymbol{\sigma}\colon \mathbb{R}^2_{>0}\times\mathbb{R}^2_{>0}\to\mathbb{R}^2_{>0}$.

Theorem 5.39. Let $\pi_1,\pi_2,r,q\in\mathcal{C}^2(\mathbb{R}^2_{>0},\mathbb{R}_{>0})$, $\boldsymbol{\sigma}\in\mathcal{C}^2(\mathbb{R}^2_{>0}\times\mathbb{R}^2_{>0},\mathbb{R}^2_{>0})$ and $\mathbf{y}^*>\mathbf{0}$ be a steady state solution of (5.4). Also, let $\mathcal{D}$ be a sufficiently small neighborhood of $\mathbf{y}^*$ and suppose that $\boldsymbol{\sigma}(\mathbf{v},\mathbf{v})=\boldsymbol{\pi}(\mathbf{v})=\mathbf{v}$ is fulfilled for all $\mathbf{v}\in C=\ker(\mathbf{\Lambda})\cap\mathcal{D}$. Then the map $\mathbf{g}$ generating the iterates of the gBBKS2(α) family satisfies $\mathbf{g}(\mathbf{v})=\mathbf{v}$ for all steady states $\mathbf{v}\in C$ and the following statements are true.

a) The map $\mathbf{g}$ satisfies $\mathbf{g}\in\mathcal{C}^1(\mathcal{D})$ and $\mathbf{D}\mathbf{g}(\mathbf{y}^*)=\mathbf{I}+\Delta t\mathbf{\Lambda}+\frac{(\Delta t)^2}{2}\mathbf{\Lambda}^2$.

b) The first derivatives of $\mathbf{g}$ are bounded and Lipschitz continuous on $\mathcal{D}$.

c) The map $\mathbf{g}$ does not belong to $\mathcal{C}^2$ for any open neighborhood of $\mathbf{y}^*$, if $\Delta t\neq(ac+b)^{-1}$.

Proof. Our main strategy is to follow the ideas used in the proof of Theorem 5.37. For this, we first compute the sets J^n and M^n in the case of the linear test problem (5.4). Using (5.107a), we obtain

$$\left(1-\frac{1}{2\alpha}\right)(\mathbf{\Lambda}\mathbf{y}^n)_m+\frac{1}{2\alpha}(\mathbf{\Lambda}\mathbf{y}^{(2)})_m=(\mathbf{\Lambda}\mathbf{y}^n)_m\left(1+\alpha\Delta t\left(\prod_{j\in J^n}\frac{y_j^{(2)}}{\pi_j^n}\right)^{q^n}\right),$$

so that

$$M^n=J^n=\begin{cases}\{1\}, & y_1^n>\frac{b}{a}y_2^n,\\ \emptyset, & y_1^n=\frac{b}{a}y_2^n,\\ \{2\}, & y_1^n<\frac{b}{a}y_2^n\end{cases}$$

follows as in the case of gBBKS1. Next, we define

$$\mathbf{y}^{(2)}(\mathbf{x})=\mathbf{x}-\Delta t\alpha\boldsymbol{\Lambda}\mathbf{x}\begin{cases}\left(\frac{y_1^{(2)}(\mathbf{x})}{\pi_1(\mathbf{x})}\right)^{q(\mathbf{x})}, & x_1>\frac{b}{a}x_2,\\ 1, & x_1=\frac{b}{a}x_2,\\ \left(\frac{y_2^{(2)}(\mathbf{x})}{\pi_2(\mathbf{x})}\right)^{q(\mathbf{x})}, & x_1<\frac{b}{a}x_2\end{cases}\tag{5.108}$$

and

$$\mathbf{F}(\mathbf{x},\mathbf{y})=\mathbf{y}-\mathbf{x}-\Delta t\left(\left(1-\frac{1}{2\alpha}\right)\boldsymbol{\Lambda}\mathbf{x}+\frac{1}{2\alpha}\boldsymbol{\Lambda}\mathbf{y}^{(2)}(\mathbf{x})\right)H(\mathbf{x},\mathbf{y}),\tag{5.109}$$

where

$$H(\mathbf{x},\mathbf{y})=\begin{cases}\widetilde{H}_1(\mathbf{x},\mathbf{y}), & x_1>\frac{b}{a}x_2,\\ 1, & x_1=\frac{b}{a}x_2,\\ \widetilde{H}_2(\mathbf{x},\mathbf{y}), & x_1<\frac{b}{a}x_2\end{cases}$$

as well as

$$\widetilde{H}_i(\mathbf{x},\mathbf{y})=\left(\frac{y_i}{\sigma_i(\mathbf{x},\mathbf{y}^{(2)}(\mathbf{x}))}\right)^{r(\mathbf{x})},\quad i=1,2,\tag{5.110}$$

and point out that the function $\mathbf{g}$ generating the gBBKS2(α) iterates is the unique solution to

$$\mathbf{0}=\mathbf{F}(\mathbf{x},\mathbf{g}(\mathbf{x})).\tag{5.111}$$

Note that equation (5.108) represents the gBBKS1 schemes applied to (5.4) with a time step size of $\Delta t\alpha$. Hence, Theorem 5.37 implies that the function $\mathbf{y}^{(2)}$ is a $\mathcal{C}^1$-map on $\mathcal{D}$ with Lipschitz continuous first derivatives and

$$\mathbf{D}\mathbf{y}^{(2)}(\mathbf{y}^*)=\mathbf{I}+\Delta t\alpha\boldsymbol{\Lambda}.$$

Furthermore, $\mathbf{v}\in\ker(\boldsymbol{\Lambda})$ implies $\mathbf{y}^{(2)}(\mathbf{v})=\mathbf{v}$, and thus, inserting $\mathbf{x}=\mathbf{v}$ into (5.109), (5.111) yields $\mathbf{g}(\mathbf{v})=\mathbf{v}$.

a) Along the same lines as in the proof of Theorem 5.37 we see that the map $\mathbf{F}$ is not differentiable on $\mathcal{D}\times\mathcal{D}$ since $\boldsymbol{\sigma}(\mathbf{x}_0,\mathbf{y}^{(2)}(\mathbf{x}_0))=\mathbf{x}_0$ holds for all $\mathbf{x}_0\in C$. However, $\mathbf{F}$ is differentiable in $(\mathbf{x},\mathbf{y})\in E=\mathcal{D}\setminus\ker(\boldsymbol{\Lambda})\times\mathcal{D}$ since

$$\mathbf{F}(\mathbf{x},\mathbf{y})=\mathbf{y}-\mathbf{x}-\Delta t\left(\left(1-\frac{1}{2\alpha}\right)\boldsymbol{\Lambda}\mathbf{x}+\frac{1}{2\alpha}\boldsymbol{\Lambda}\mathbf{y}^{(2)}(\mathbf{x})\right)\left(\frac{y_i}{\sigma_i(\mathbf{x},\mathbf{y}^{(2)}(\mathbf{x}))}\right)^{r(\mathbf{x})}\tag{5.112}$$

for

$$i=\begin{cases}1 & ,x_1>\frac{b}{a}x_2,\\ 2 & ,x_1<\frac{b}{a}x_2\end{cases}\tag{5.113}$$

and $r\in\mathcal{C}^2(\mathbb{R}^2_{>0},\mathbb{R}_{>0})$, $\boldsymbol{\sigma}\in\mathcal{C}^2(\mathbb{R}^2_{>0}\times\mathbb{R}^2_{>0},\mathbb{R}^2_{>0})$ as well as $\mathbf{y}^{(2)}\in\mathcal{C}^1(\mathcal{D})$. Following the proof of Theorem 5.37, we show that $\mathbf{D_y F}(\mathbf{x},\mathbf{g}(\mathbf{x}))$ is nonsingular in order to show that $\mathbf{g}\in\mathcal{C}^1$ on $\mathcal{D}\setminus\ker(\boldsymbol{\Lambda})$. First note that for $\mathbf{x}\notin C$ we have

$$\mathbf{D_y F}(\mathbf{x},\mathbf{g}(\mathbf{x}))=\mathbf{I}-\Delta t\left(\left(1-\frac{1}{2\alpha}\right)\boldsymbol{\Lambda}\mathbf{x}+\frac{1}{2\alpha}\boldsymbol{\Lambda}\mathbf{y}^{(2)}(\mathbf{x})\right)\nabla_{\mathbf{y}}\widetilde{H}_i(\mathbf{x},\mathbf{g}(\mathbf{x}))\tag{5.114}$$

for i from (5.113). Now, (5.110) yields

$$\nabla_{\mathbf{y}}\widetilde{H}_i(\mathbf{x},\mathbf{g}(\mathbf{x})) = \frac{r(\mathbf{x})}{g_i(\mathbf{x})}\widetilde{H}_i(\mathbf{x},\mathbf{g}(\mathbf{x}))\mathbf{e}_i^T$$

with the ith unit vector $\mathbf{e}_i \in \mathbb{R}^2$ as in the proof of Theorem 5.37. In order to see that $\mathbf{D_y F}(\mathbf{x},\mathbf{g}(\mathbf{x}))$ is invertible, we introduce

$$\mathbf{v}^{(i)}(\mathbf{x}) = \Delta t\left(\left(1-\frac{1}{2\alpha}\right)\mathbf{\Lambda x} + \frac{1}{2\alpha}\mathbf{\Lambda y}^{(2)}(\mathbf{x})\right)\frac{r(\mathbf{x})}{g_i(\mathbf{x})}\widetilde{H}_i(\mathbf{x},\mathbf{g}(\mathbf{x}))$$

and rewrite (5.114) as

$$\mathbf{D_y F}(\mathbf{x},\mathbf{g}(\mathbf{x})) = \mathbf{I} - \mathbf{v}^{(i)}(\mathbf{x})\mathbf{e}_i^T. \tag{5.115}$$

Hence, we obtain

$$\det(\mathbf{D_y F}(\mathbf{x},\mathbf{g}(\mathbf{x}))) = 1 - v_i^{(i)}(\mathbf{x}).$$

Using (5.108), we see that

$$\mathbf{v}^{(i)}(\mathbf{x}) = \Delta t\mathbf{\Lambda x}\left(1+\alpha\Delta t\left(\frac{y_i^{(2)}(\mathbf{x})}{\pi_i(\mathbf{x})}\right)^{q(\mathbf{x})}\right)\frac{r(\mathbf{x})}{g_i(\mathbf{x})}\widetilde{H}_i(\mathbf{x},\mathbf{g}(\mathbf{x})),$$

where $(\mathbf{\Lambda x})_i < 0$ by definition of the gBBKS2(α) schemes. As a result we know $v_i^{(i)}(\mathbf{x}) < 0$, and hence $\det(\mathbf{D_y F}(\mathbf{x},\mathbf{g}(\mathbf{x}))) \neq 0$ proving that $\mathbf{D_y F}(\mathbf{x},\mathbf{g}(\mathbf{x}))$ is invertible. This together with the corresponding arguments of Theorem 5.37 implies that $\mathbf{g} \in \mathcal{C}^1$ on $\mathcal{D}\setminus\ker(\mathbf{\Lambda})$ and

$$\mathbf{Dg}(\mathbf{x}) = -(\mathbf{D_y F}(\mathbf{x},\mathbf{g}(\mathbf{x})))^{-1}\mathbf{D_x F}(\mathbf{x},\mathbf{g}(\mathbf{x})) \tag{5.116}$$

for $\mathbf{x} \in \mathcal{D}\setminus\ker(\mathbf{\Lambda})$. To apply Theorem 5.36, we proceed as in the proof of Theorem 5.37, i. e. we have to show that

1. the map $\mathbf{g}$ is continuous in any $\mathbf{x} \in C$.
2. the map $\mathbf{F}$ is differentiable in $(\mathbf{x},\mathbf{g}(\mathbf{x}))$ for all $\mathbf{x} \in C$.
3. the Jacobian $\mathbf{D_y F}(\mathbf{x},\mathbf{g}(\mathbf{x}))$ with respect to $\mathbf{y}$ is invertible for all $\mathbf{x} \in C$.

The continuity of $\mathbf{g}$ follows along the same lines as in the case of gBBKS1, where we additionally use $\mathbf{y}^{(2)} \in \mathcal{C}^1(\mathcal{D})$ for bounding $H(\cdot,\mathbf{g}(\cdot))$.

For proving the differentiability of $\mathbf{F}$ in $(\mathbf{x},\mathbf{g}(\mathbf{x}))$ for all $\mathbf{x} \in C$ we consider an arbitrary element $\mathbf{x}_0 \in C$. Note that $\Psi(\mathbf{x}) = H(\mathbf{x},\mathbf{g}(\mathbf{x}_0))$ is continuous in $\mathbf{x}_0$ with $\Psi(\mathbf{x}_0) = 1$. Furthermore,

$$\mathbf{\Phi}(\mathbf{x}) = \left(1-\frac{1}{2\alpha}\right)\mathbf{\Lambda x} + \frac{1}{2\alpha}\mathbf{\Lambda y}^{(2)}(\mathbf{x})$$

satisfies $\mathbf{\Phi}(\mathbf{x}_0) = \mathbf{0}$, which means that part a) of Lemma A.2 from the appendix together with $\mathbf{D}\mathbf{y}^{(2)}(\mathbf{x}_0) = \mathbf{I} + \Delta t\alpha\mathbf{\Lambda}$ yields

$$\begin{aligned}\mathbf{D}_{\mathbf{x}}\mathbf{F}(\mathbf{x}_0,\mathbf{g}(\mathbf{x}_0)) &= -\mathbf{I} - \Delta t\left(\left(1-\frac{1}{2\alpha}\right)\mathbf{\Lambda} + \frac{1}{2\alpha}\mathbf{\Lambda}\mathbf{D}\mathbf{y}^{(2)}(\mathbf{x}_0)\right)\\ &= -\mathbf{I} - \Delta t\left(\mathbf{\Lambda} + \frac{\Delta t}{2}\mathbf{\Lambda}^2\right).\end{aligned} \tag{5.117}$$

Also, since $\mathbf{\Phi}(\mathbf{x}_0)H(\mathbf{x}_0,\mathbf{y}) = \mathbf{0}$ for all $\mathbf{y} \in \mathbb{R}^2_{>0}$, we find

$$\mathbf{D}_{\mathbf{y}}\mathbf{F}(\mathbf{x}_0,\mathbf{g}(\mathbf{x}_0)) = \mathbf{I}, \tag{5.118}$$

which shows that $\mathbf{F}$ is partially differentiable in $(\mathbf{x}_0,\mathbf{g}(\mathbf{x}_0))$. We now prove that the partial derivatives of $\mathbf{F}$ are also continuous in $(\mathbf{x}_0,\mathbf{g}(\mathbf{x}_0))$, which shows the differentiability of $\mathbf{F}$ in $(\mathbf{x},\mathbf{g}(\mathbf{x}))$ for all $\mathbf{x} \in C$. To that end, we consider $\mathbf{x} \notin C$ and differentiate $\mathbf{F}$ from (5.112) with respect to $\mathbf{x}$ and $\mathbf{y}$. First, due to (5.114) and since $\mathbf{v}^{(i)}$ is continuous with $\mathbf{v}^{(i)}(\mathbf{x}_0) = \mathbf{0}$ for $\mathbf{x}_0 \in C$, we find

$$\lim_{\mathbf{x}\to\mathbf{x}_0} \mathbf{D}_{\mathbf{y}}\mathbf{F}(\mathbf{x},\mathbf{g}(\mathbf{x})) = \mathbf{I} - \mathbf{v}^{(i)}(\mathbf{x}_0)\mathbf{e}_i^T = \mathbf{I}$$

proving the continuity of the partial derivatives in $(\mathbf{x}_0,\mathbf{g}(\mathbf{x}_0))$ with respect to $\mathbf{y}$. Furthermore, we have

$$\begin{aligned}\mathbf{D}_{\mathbf{x}}\mathbf{F}(\mathbf{x},\mathbf{g}(\mathbf{x})) =& -\mathbf{I} - \Delta t\left(\left(1-\frac{1}{2\alpha}\right)\mathbf{\Lambda} + \frac{1}{2\alpha}\mathbf{\Lambda}\mathbf{D}\mathbf{y}^{(2)}(\mathbf{x})\right)\widetilde{H}_i(\mathbf{x},\mathbf{g}(\mathbf{x}))\\ &- \Delta t\left(\left(1-\frac{1}{2\alpha}\right)\mathbf{\Lambda}\mathbf{x} + \frac{1}{2\alpha}\mathbf{\Lambda}\mathbf{y}^{(2)}(\mathbf{x})\right)\nabla_{\mathbf{x}}\widetilde{H}_i(\mathbf{x},\mathbf{g}(\mathbf{x})),\end{aligned}$$

whose entries converge to those of $\mathbf{D}_{\mathbf{x}}\mathbf{F}(\mathbf{x}_0,\mathbf{g}(\mathbf{x}_0))$ from (5.117) because of the following. First, we have $\mathbf{y}^{(2)}(\mathbf{x}_0) = \mathbf{x}_0$ and $\widetilde{H}_i \in \mathcal{C}^1(\mathcal{D}\times\mathcal{D})$, which means that the last addend disappears as $\mathbf{x} \to \mathbf{x}_0 \in C$. Additionally, inserting $\mathbf{D}\mathbf{y}^{(2)}(\mathbf{x}_0) = \mathbf{I} + \alpha\Delta t\mathbf{\Lambda}$ and $\lim_{\mathbf{x}\to\mathbf{x}_0} \widetilde{H}_i(\mathbf{x},\mathbf{g}(\mathbf{x})) = 1$ yield (5.117).

Finally, it follows from (5.118) that the Jacobian $\mathbf{D}_{\mathbf{y}}\mathbf{F}(\mathbf{x},\mathbf{g}(\mathbf{x}))$ with respect to $\mathbf{y}$ is invertible for all $\mathbf{x} \in C$. Hence, Theorem 5.36 together with the considerations above proves that $\mathbf{g}$ is differentiable in all $\mathbf{x} \in \mathcal{D}$.

Moreover, since $\mathbf{F}$ is continuously differentiable in $(\mathbf{x},\mathbf{g}(\mathbf{x}))$, we find due to (5.116) that even $\mathbf{g} \in \mathcal{C}^1(\mathcal{D})$ holds true. Furthermore, inserting (5.117) and (5.118) into formula (5.116) yields

$$\mathbf{D}\mathbf{g}(\mathbf{x}) = -(\mathbf{D}_{\mathbf{y}}\mathbf{F}(\mathbf{x},\mathbf{g}(\mathbf{x})))^{-1}\mathbf{D}_{\mathbf{x}}\mathbf{F}(\mathbf{x},\mathbf{g}(\mathbf{x})) = \mathbf{I} + \Delta t\mathbf{\Lambda} + \frac{(\Delta t)^2}{2}\mathbf{\Lambda}^2.$$

b) We know that $\mathbf{y}^{(2)} \in \mathcal{C}^1$ has Lipschitz continuous first derivatives on $\mathcal{D}$ and that $\boldsymbol{\sigma} \in \mathcal{C}^2$. Hence, with

$$\begin{aligned}\nabla_{\mathbf{x}}\widetilde{H}_i(\mathbf{x},\mathbf{g}(\mathbf{x})) =& \widetilde{H}_i(\mathbf{x},\mathbf{g}(\mathbf{x}))\Bigg(\nabla_{\mathbf{x}}r(\mathbf{x})\ln\left(\frac{g_i(\mathbf{x})}{\sigma_i(\mathbf{x},\mathbf{y}^{(2)}(\mathbf{x}))}\right)\\ &- r(\mathbf{x})\frac{\nabla_{\mathbf{x}}\sigma_i(\mathbf{x},\mathbf{y}^{(2)}(\mathbf{x})) + \nabla_{\mathbf{y}}\sigma_i(\mathbf{x},\mathbf{y}^{(2)}(\mathbf{x}))\mathbf{D}\mathbf{y}^{(2)}(\mathbf{x})}{\sigma_i(\mathbf{x})}\Bigg),\end{aligned}$$

which is analogous to (5.99), this part can be proven along the same lines as in the proof of part b) of Theorem 5.37.

c) Since (5.115) and (5.116) are of the form (5.104), (5.103), this part is proven along the same lines as part c) of Theorem 5.37.

□

This theorem together with Theorem 5.4 and Theorem 2.15 allows us to conclude the following statements from $\mathbf{Dg}(\mathbf{y}^*) = \mathbf{I} + \Delta t\mathbf{\Lambda} + \frac{1}{2}(\Delta t\mathbf{\Lambda})^2$.

Corollary 5.40. Let $\mathbf{y}^* > \mathbf{0}$ be an arbitrary steady state of (5.4). Under the assumptions of Theorem 5.39, the gBBKS2(α) schemes have the same stability function as the underlying Runge–Kutta method, i. e. $R(z) = 1 + z + \frac{z^2}{2}$ and the following holds.

a) If $|R(-(ac+b)\Delta t)| < 1$, then $\mathbf{y}^*$ is a stable fixed point of each gBBKS2(α) scheme and there exists a $\delta > 0$, such that $\mathbf{y}^n \to \mathbf{y}^*$ as $n \to \infty$ for all $\mathbf{y}^0$ satisfying $\left(\begin{smallmatrix}1\\c\end{smallmatrix}\right)^T\mathbf{y}^0 = \left(\begin{smallmatrix}1\\c\end{smallmatrix}\right)^T\mathbf{y}^*$ and $\|\mathbf{y}^0 - \mathbf{y}^*\| < \delta$.

b) If $|R(-(ac+b)\Delta t)| > 1$, then $\mathbf{y}^*$ is an unstable fixed point of each gBBKS2(α) scheme.

To summarize the presented analysis of gBBKS schemes, we conclude that the first and second order gBBKS schemes preserve the stability domain of the underlying Runge–Kutta method while preserving positivity. To generalize these results to $N \times N$ systems we need to exploit more properties of the particular choices of $r, q, \boldsymbol{\pi}$ and $\boldsymbol{\sigma}$ from the literature [AKM20].

5.5 Summary of Stability Properties

In the previous section we investigated several Patankar-type methods with respect to their stability. The purpose of this rather short section is to summarize our findings, see Table 5.2. We also recall that MPDeC(1) corresponds to MPE and MPDeC(2) equals MPRK22(1). Also, for more insights on the stability properties of MPRK43(α, β) we refer to Figure 5.1, where a lower bound for the maximal opening angle of the stability domain is depicted for (α, β) pairs in the feasible domain in $[0, 2] \times [0, \frac{3}{4}]$ with a resolution of 102^2 pairs per unit square. The opening angle estimate for MPDeC up to order 8 can be found in Table 5.1.

Method	Parameter Specification	Unconditionally Stable?
MPE	–	$\surd$
MPRK22(α)	$\alpha \geq \frac{1}{2}$	$\surd$
MPRK43(α, β)	$(\alpha, \beta) = (0.5, 0.75)$	$\surd$
MPRK43(α, β)	$(\alpha, \beta) = (1, 0.5)$	$\times$
MPRK43(γ)	$\frac{3}{8} \leq \gamma \leq \frac{3}{4}$	$\surd$
SSPMPRK2(α, β)	$\alpha \leq \frac{1}{2\beta}$	$\surd$
SSPMPRK2(α, β)	$\alpha > \frac{1}{2\beta}$	$\times$
SSPMPRK3(η_2)	$0 \leq \eta_2 \leq 0.37$	$\surd$
MPDeC(p)	$p \in \{1, 2\}$	$\surd$
MPDeC(p)	$p = 3, \dots, 8$	$\times$
MPDeCGL(p)	$p = 9, \dots, 14$	$(\surd)$
MPDeCEQ(p)	$p = 9, 10, 11, 13$	$(\surd)$
MPDeCEQ(p)	$p \in \{12, 14\}$	$\times$
GeCo1	–	$\surd$
GeCo2	–	$\times$
gBBKS1	–	$\times$
gBBKS2(α)	$\alpha \geq \frac{1}{2}$	$\times$

Table 5.2: Overview on the stability properties of several Patankar-type methods from Chapter 3. Here, “$(\surd)$” means that the unconditional stability is investigated only numerically and only for normal system matrices with real eigenvalues.

In the upcoming section we introduce several linear problems for testing the predictions contained in Table 5.2 together with the corresponding numerical experiments. For a deeper insight into numerical experiments concerning oscillatory behavior we refer to [IÖT22].

5.6 Numerical Experiments

As mentioned in the preceding section, this part of the thesis is dedicated to the numerical validation of the theoretical claims concerning stability, parts of which are summarized in Table 5.2. Thereby, we also incorporate the hypothesis mentioned and tested in [IKMS23] stating that the claimed properties of stability and convergence towards the steady state solution of (5.1) are even of global nature for MPRK schemes that are based on a non-negative Butcher tableau. In fact, so far the only cases of MP methods where the stability properties were observed to be non-global are MPRK22(α) with $\alpha < \frac{1}{2}$ and MPDeCEQ(p) for some values $p \geq 8$. In all cases the schemes can be understood as MP methods based on RK schemes with a Butcher array containing also negative entries, see [IÖ23, IKMS23]. In particular, we present the numerical experiments with MPDeCEQ(8) in this work to give an example of this phenomenon.

For the numerical validation different test cases are of interest, which we will discuss in the following subsection.

5.6.1 Test Problems

In the following we only consider conservative problems, i. e. the systems matrices we are going to introduce have an eigenvalue $\lambda = 0$. Furthermore, most of the following test cases are chosen in such a way that all nonzero eigenvalues either lie in $\mathbb{R}^-$ or in $\mathbb{C}^- \setminus \mathbb{R}^-$. Moreover, as we are interested in testing part b) of Theorem 5.4, we also consider a test problem with two linear invariants. It is beneficial to consider these test cases rather than a single one with a spectrum in $\overline{\mathbb{C}^-}$ because this way we can test the stability domain of conditional stable methods at two distinct spots of the stability domain. Nevertheless, we will also include a test problem with real as well as complex eigenvalues for testing unconditionally stable schemes.

Test problem with exclusively real eigenvalues

The linear initial value problem

$$\mathbf{y}' = 100 \begin{pmatrix} -2 & 1 & 1 \\ 1 & -4 & 1 \\ 1 & 3 & -2 \end{pmatrix} \mathbf{y}, \quad \mathbf{y}(0) = \begin{pmatrix} 1 \\ 9 \\ 5 \end{pmatrix} \tag{5.119}$$

contains a system matrix, which has only positive off-diagonal elements and is therefore a Metzler matrix. Due to the positive initial values, this ensures that each component of the solution of the initial value problem is positive for all times. By a straightforward calculation of the eigenvalues $\lambda_1 = 0$, $\lambda_2 = -300$ and $\lambda_3 = -500$ of the system matrix as well as their associated eigenvectors, the solution reads

$$\mathbf{y}(t) = c_1 \begin{pmatrix} 5 \\ 3 \\ 7 \end{pmatrix} + c_2 e^{-300t} \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} + c_3 e^{-500t} \begin{pmatrix} 0 \\ -1 \\ 1 \end{pmatrix} \tag{5.120}$$

with coefficients $c_1 = 1$, $c_2 = 4$ and $c_3 = -6$ determined by the initial condition. Since only non-positive eigenvalues are present and the absolute values of the negative eigenvalues are large, there is a fast convergence to the equilibrium state

$$\mathbf{y}^* = \lim_{t\to\infty} \mathbf{y}(t) = \begin{pmatrix} 5 \\ 3 \\ 7 \end{pmatrix}$$

as depicted in Figure 5.7. Furthermore the zero eigenvalue is simple, and hence there exists exactly one linear invariant, which is given by $\mathbf{1}^T\mathbf{y}$ due to the fact that the sum of the elements in each column of the system matrix is always vanishing. This conservativity can also be observed in Figure 5.7.

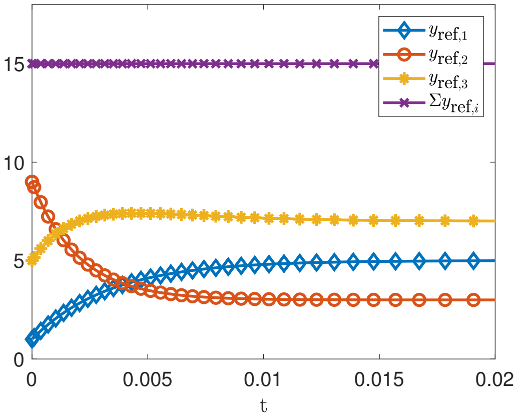


Figure 5.7: Exact solution (5.120) of the initial value problem (5.119) and the linear invariant $\mathbf{1}^T\mathbf{y}$.

Test problem with complex eigenvalues

As a second test case, we consider the conservative system

$$\mathbf{y}' = 100 \begin{pmatrix} -4 & 3 & 1 \\ 2 & -4 & 3 \\ 2 & 1 & -4 \end{pmatrix} \mathbf{y}, \quad \mathbf{y}(0) = \begin{pmatrix} 9 \\ 20 \\ 8 \end{pmatrix}. \tag{5.121}$$

Again, the system matrix is a Metzler matrix, so that the solution of the initial value problem is always positive due to the positive initial conditions. Considering the eigenvalues $\lambda_1 = 0$, $\lambda_2 = 100(-6+\mathrm{i})$ and $\lambda_3 = \overline{\lambda_2}$ as well as the corresponding eigenvectors of the system matrix, the solution can be written in the form

$$\begin{aligned} \mathbf{y}(t) = \begin{pmatrix} 13 \\ 14 \\ 10 \end{pmatrix} &- 2e^{-600t} \left(\cos(100t) \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} - \sin(100t) \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} \right) \\ &- 6e^{-600t} \left(\cos(100t) \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} + \sin(100t) \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} \right). \end{aligned} \tag{5.122}$$

The nonzero complex eigenvalues have a negative real part with a large absolute value. Hence, one can expect a rapid convergence of the solution to the steady state given by

$$\mathbf{y}^* = \lim_{t\to\infty} \mathbf{y}(t) = \begin{pmatrix} 13 \\ 14 \\ 10 \end{pmatrix}.$$

Analogous to the first test case, the only linear invariant is $\mathbf{1}^T\mathbf{y}$, which is presented together with the exact solution in Figure 5.8.

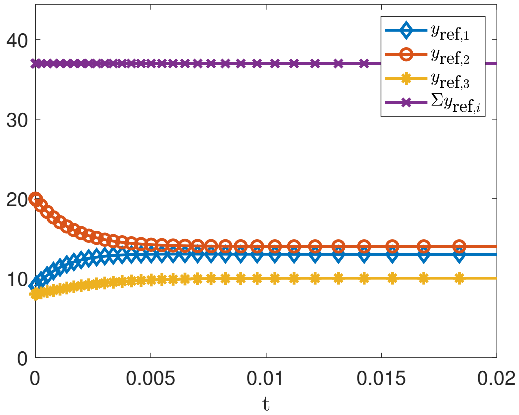


Figure 5.8: The exact solution (5.122) of the initial value problem (5.121) and the linear invariant $\mathbf{1}^T\mathbf{y}$.

Test problem with double zero eigenvalue

Considering the linear initial value problem

$$\mathbf{y}' = 100 \begin{pmatrix} -2 & 0 & 0 & 1 \\ 0 & -4 & 3 & 0 \\ 0 & 4 & -3 & 0 \\ 2 & 0 & 0 & -1 \end{pmatrix} \mathbf{y}, \quad \mathbf{y}(0) = \begin{pmatrix} 4 \\ 1 \\ 9 \\ 1 \end{pmatrix}, \tag{5.123}$$

we are faced with a Metzler matrix including a double zero eigenvalue $\lambda_1 = \lambda_2 = 0$. Therefore, besides $\mathbf{1}^T\mathbf{y}$, a second linear invariant $\mathbf{n}^T\mathbf{y}$ with $\mathbf{n} = (1,2,2,1)^T$ is present. Due to the remaining eigenvalues $\lambda_3 = -300$ and $\lambda_3 = -700$ and the associated eigenvectors of all eigenvalues, the solution of the initial value problem writes

$$\mathbf{y}(t) = c_1 \begin{pmatrix} 0 \\ 1 \\ \frac{4}{3} \\ 0 \end{pmatrix} + c_2 \begin{pmatrix} 1 \\ 0 \\ 0 \\ 2 \end{pmatrix} + c_3 e^{-700t} \begin{pmatrix} 0 \\ 1 \\ -1 \\ 0 \end{pmatrix} + c_4 e^{-300t} \begin{pmatrix} 1 \\ 0 \\ 0 \\ -1 \end{pmatrix} \tag{5.124}$$

with coefficients

$$c_1 = \frac{30}{7}, \quad c_2 = \frac{5}{3}, \quad c_3 = -\frac{23}{7} \quad \text{and} \quad c_4 = \frac{7}{3}.$$

Once again, a fast convergence to the equilibrium state

$$\mathbf{y}^* = \lim_{t\to\infty} \mathbf{y}(t) = c_1 \begin{pmatrix} 0 \\ 1 \\ \frac{4}{3} \\ 0 \end{pmatrix} + c_2 \begin{pmatrix} 1 \\ 0 \\ 0 \\ 2 \end{pmatrix} = \frac{1}{21} \begin{pmatrix} 7 \\ 90 \\ 120 \\ 70 \end{pmatrix}$$

takes place. The course of the solution together with the two linear invariants are shown in Figure 5.9.

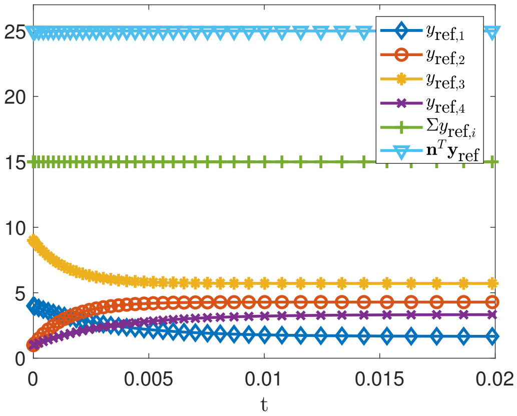


Figure 5.9: The exact solution (5.124) of the initial value problem (5.123) and the associated two linear invariants $\mathbf{1}^T\mathbf{y}$ and $\mathbf{n}^T\mathbf{y}$ with $\mathbf{n}^T = (1, 2, 2, 1)$.

At this point we want to note that the presented test cases represent stiff problems due to the occurrence of large absolute values of the corresponding eigenvalues. Hence, it is not surprising that the exact solution satisfies the inequality $\|\mathbf{y}(t) - \mathbf{y}^*\|_2 < 2 \cdot 10^{-2}$ at time $t = 0.02$ for all of three problems.

Test Problem with mixed Eigenvalues

Finally, we consider the initial value problem

$$\mathbf{y}' = \mathbf{\Lambda}\mathbf{y}, \quad \mathbf{y}(0) = \mathbf{y}^0 = (0, 3, 3, 3, 4)^T, \tag{5.125}$$

where $\mathbf{\Lambda}$ is the 5×5 Metzler matrix

$$\mathbf{\Lambda} = \begin{pmatrix} -4 & 2 & 1 & 2 & 2 \\ 1 & -4 & 1 & 0 & 2 \\ 0 & 0 & -4 & 2 & 0 \\ 2 & 2 & 2 & -4 & 0 \\ 1 & 0 & 0 & 0 & -4 \end{pmatrix}. \tag{5.126}$$

The spectrum of $\mathbf{\Lambda}$ is given by $\sigma(\mathbf{\Lambda}) = \{0, -5-\sqrt{3}, -5+\sqrt{3}, -5-\mathrm{i}, -5+\mathrm{i}\} \subseteq \overline{\mathbb{C}^-}$ including real as well as non-real eigenvalues. Furthermore, the kernel of $\mathbf{\Lambda}^T$ is given by $\ker(\mathbf{\Lambda}^T) = \operatorname{span}(\mathbf{n})$ with $\mathbf{n} = (1, 1, 1, 1, 1)^T$. Hence, the total mass $\mathbf{n}^T\mathbf{y}(t) = \mathbf{n}^T\mathbf{y}^0 = 13$ is a linear invariant for the system, in correspondence of the initial value $\mathbf{y}(0) = \mathbf{y}^0$. The reference solution of the problem is depicted in Figure 5.10 and satisfies $\|\mathbf{y}(t) - \mathbf{y}^*\|_2 < 10^{-2}$ at time $t = 1.61$.

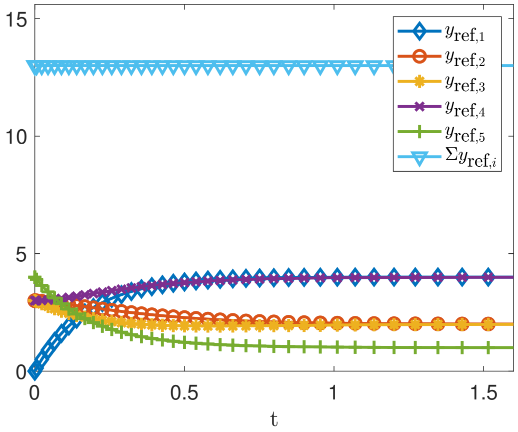


Figure 5.10: The reference solution of the initial value problem (5.125).

We want to note that even though the stability functions of gBBKS and GeCo2 were obtained by analyzing a 2×2 system, we will see that the corresponding stability results are well reflected also for a larger system.

5.6.2 Investigation of MPRK Schemes

As in Chapter 5, we consider here MPRK schemes up to order three. The stability analysis and numerical experiments for the fourth order MPRK method are left for future work. In particular, the numerical experiments will be performed with MPE, MPRK22(α) for $\alpha \in \{0.5, 1, 5\}$, MPRK43(0.5, 0.75) and MPRK43(0.563), all of which are proven to be unconditionally stable and locally converging towards the steady state solution. Hence, we consider the problem (5.125) using a comparably large time step size of $\Delta t = 5$.

MPE

The results for MPE can be seen in Figure 5.11. As one can see, the method is stable and converging using the initial condition from (5.125). An error of around 10^{-14} is already obtained after $t = 100$, that is after 20 steps using $\Delta t = 5$. Note again that this is a comparably large Δt as the analytic solution satisfies $\|\mathbf{y}(t) - \mathbf{y}^*\|_2 < 2 \cdot 10^{-2}$ at time $t = 1.61$. In Figure 5.12 on can see that the second linear invariant is also preserved.

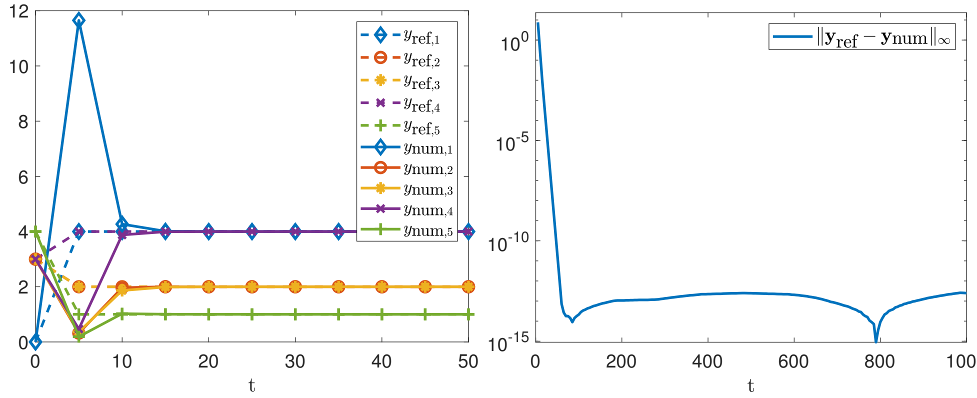


Figure 5.11: Numerical solution of (5.125) and error plot using MPE. The dashed lines represent the reference solution.

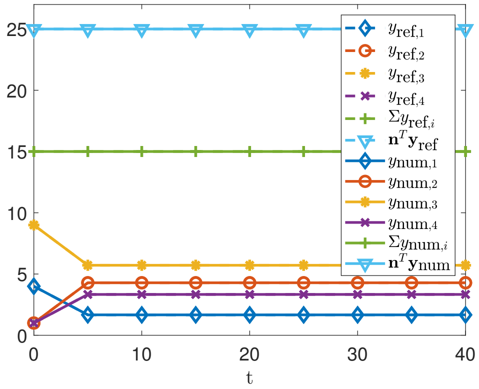


Figure 5.12: Numerical approximation of (5.123) using MPE. The dashed lines represent the exact solution (5.124) and coincide for this example with the numerical solution. The second linear invariant is determined by $\mathbf{n} = (1, 2, 2, 1)^T$.

MPRK22(α)

In the Figures 5.13 and 5.14, we compare the MPRK22(α) schemes for $\alpha \in \left\{\frac{1}{2}, 1, 5\right\}$ and find that for $\alpha = 1$ or $\alpha = 5$ the methods produce errors near machine precision at $t = 200$, i. e. after around 40 steps, whereas for $\alpha = \frac{1}{2}$ we cannot observe the convergence of the iterates towards $\mathbf{y}^*$ within $t \in [0, 50]$. Nevertheless, the results depicted on the top right show that even for the case $\alpha = \frac{1}{2}$, the stability and convergence proved in Corollaries 5.16 and 5.17 can be confirmed numerically by extending the observation period. Moreover, the second linear invariant is also preserved, see Figure 5.14.

MPRK43($0.5, 0.75$)

Similarly as before, all theoretical claims for MPRK43($0.5, 0.75$) are well reflected in the numerical approximation of (5.125), see Figure 5.15 and Figure 5.16.

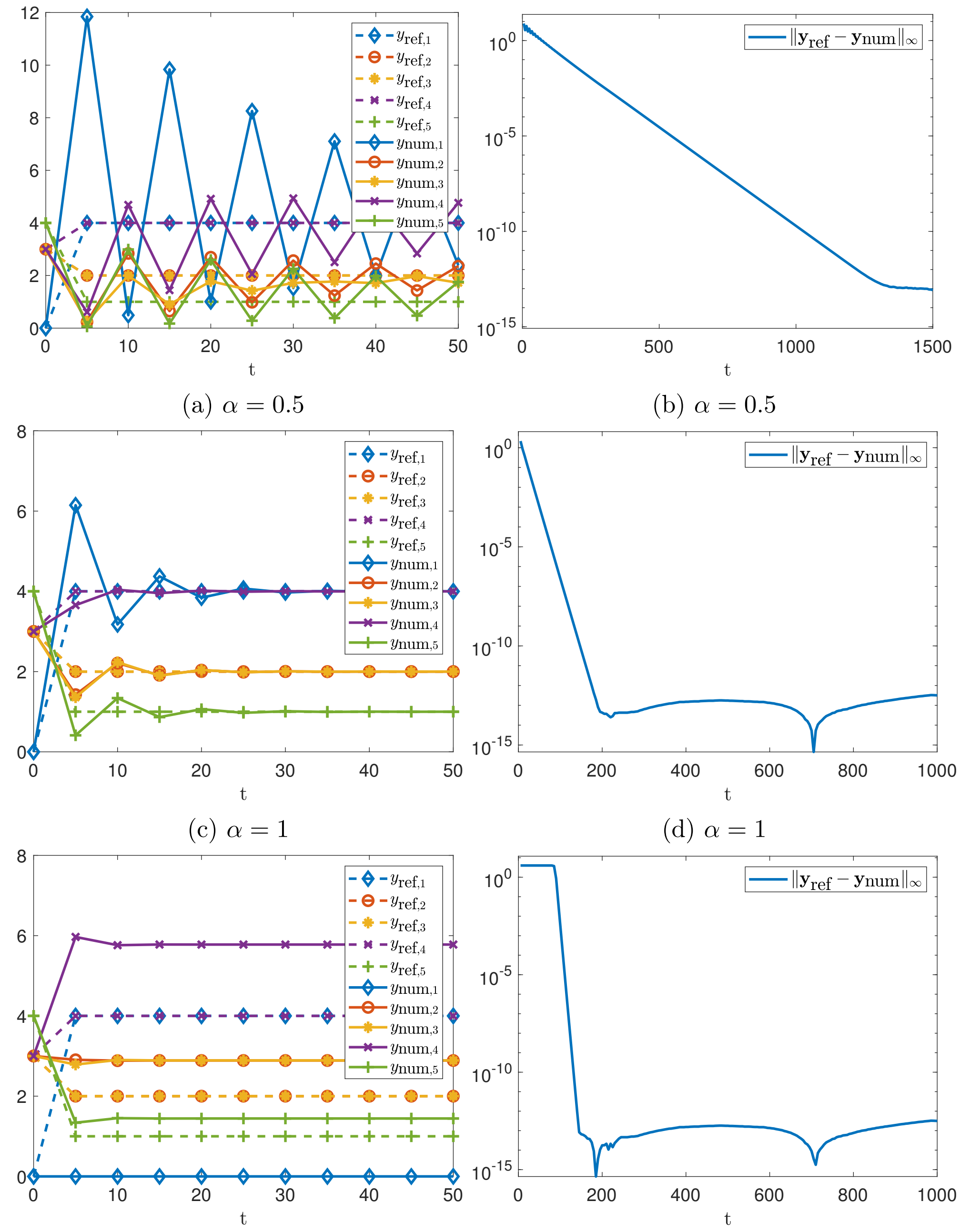
MPRK43(0.563)

According to the investigation in [IÖT22], MPRK43(γ) has the largest Δt bound for fulfilling the necessary condition for avoiding oscillations, if $\gamma \approx 0.563$. This is why we restrict to this method hereafter. Since this method is also proven to be unconditionally stable, we proceed as for the previously discussed methods. The results can be found in Figure 5.17 and Figure 5.18 and reflect well our theoretical claims from Corollary 5.20.

5.6.3 Investigation of SSPMPRK Schemes

Hereafter, we confirm numerically that SSPMPRK schemes are stable as claimed in Corollary 5.23 and Corollary 5.27. Furthermore, we investigate the local convergence to the steady state solution as stated in Corollary 5.24 and Corollary 5.28 by choosing $\mathbf{y}^0 = \mathbf{y}(0)$ and $\Delta t = 5$, if not stated otherwise. Indeed, in all experiments below the convergence in the stable case can be observed even for $\mathbf{y}^0 = \mathbf{y}(0)$.

In particular, we are interested in the properties of SSPMPRK3($\frac{1}{3}$) which is the preferred scheme presented in [HZS19]. Moreover, we investigate SSPMPRK2(α, β) for three different pairs (α, β) covering all cases mentioned in Proposition 5.22. For the case $\alpha > \frac{1}{2\beta}$ we choose the lower left vertex of the red rectangular from



(a) $\alpha = 0.5$

(b) $\alpha = 0.5$

(c) $\alpha = 1$

(d) $\alpha = 1$

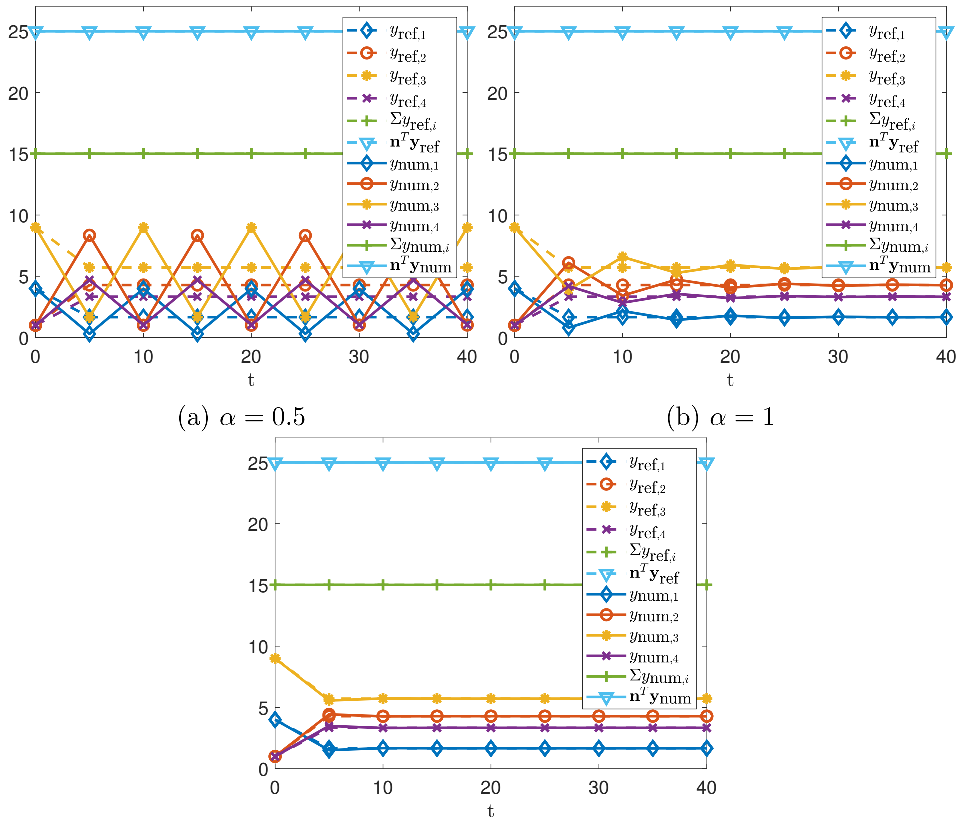
(e) $\alpha = 5$

(f) $\alpha = 5$

Figure 5.13: Numerical solution of (5.125) and error plots using MPRK22(α) schemes. The dashed lines represent the reference solution.

Figure 5.2, i. e. $(\alpha, \beta) = (0.2, 3)$. In this case, we choose different time steps to demonstrate that the computed stability regions are correct. At this point we want to note that the eigenvalues of the system matrices from the test problems lie on the red or blue line depicted in Figure 5.19. We scale the time step size Δt in such a way that $\Delta t \rho(\mathbf{Dg}(\mathbf{y}^*)) = z_i$ for $i \in \{1, 2, 3, 4\}$, respectively, so that for all test cases we consider the cases of stable as well as unstable fixed points.

As a representative for the case $\alpha = \frac{1}{2\beta}$ we use $(\alpha, \beta) = (\frac{1}{2}, 1)$ which is the preferred choice presented in [HS19]. Finally, we choose $(\alpha, \beta) = (0.1, 1)$ satisfying $\alpha < \frac{1}{2\beta}$.



(a) $\alpha = 0.5$

(b) $\alpha = 1$

(c) $\alpha = 5$

Figure 5.14: Numerical approximations of (5.123) using MPRK22(α) schemes. The dashed lines indicate the exact solution (5.124) and $\mathbf{n} = (1, 2, 2, 1)^T$.

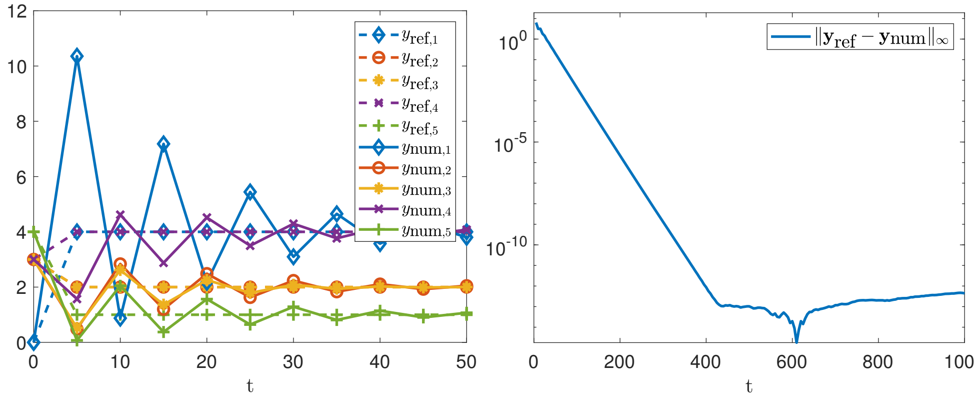


Figure 5.15: Numerical solution of (5.125) and error plot using the third order MPRK43(0.5, 0.75) method. The dashed lines represent the reference solution.

SSPMPRK2(α, β)

In the subsequent figures, SSPMPRK2(α, β) schemes are used to solve the test problems. In all four figures 5.20, 5.21, 5.22 and 5.23, we can observe the same qualitative behavior. In Figure 5.20 and Figure 5.22, the preferred choice of $(\alpha, \beta) = (\frac{1}{2}, 1)$ seems to be less damping than $(\alpha, \beta) = (0.1, 1)$. However, in both cases a convergence towards the steady state solution can be observed. In Figure 5.21 and Figure 5.23, the pair (α, β) lies in the critical region where the stability domain is bounded. If Δt is chosen in such a way that $\Delta t\rho(\mathbf{Dg}(\mathbf{y}^*)) = z_i$

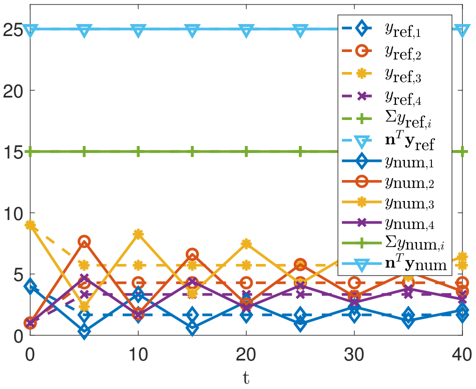


Figure 5.16: Numerical approximation of (5.123) using MPRK43(0.5, 0.75). The dashed lines represent the exact solution (5.124) and $\mathbf{n} = (1, 2, 2, 1)^T$.

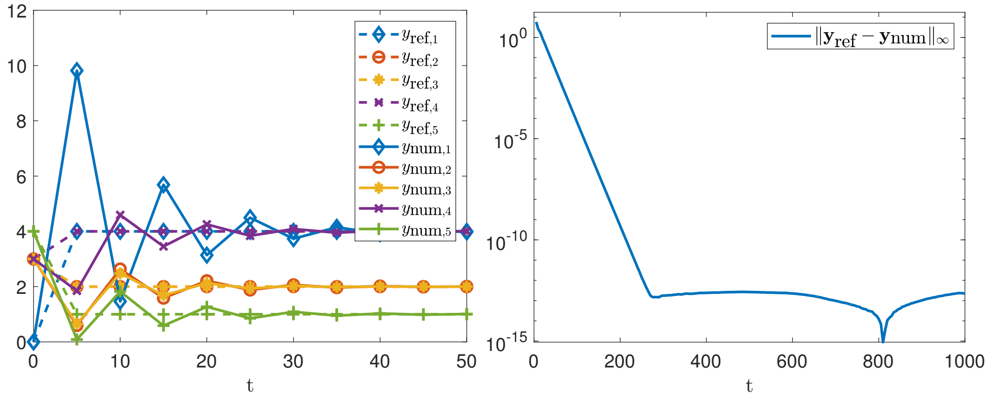


Figure 5.17: Numerical solution of (5.125) and error plot using MPRK43(0.563). The dashed lines represent the reference solution.

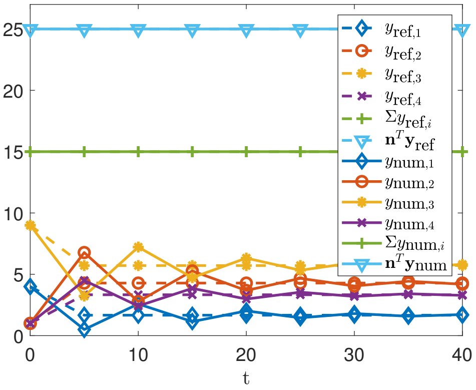


Figure 5.18: Numerical approximation of (5.123) using MPRK43(0.563). The dashed lines represent the exact solution (5.124) and $\mathbf{n} = (1, 2, 2, 1)^T$.

for $i = 2$ or $i = 4$, respectively, see Figure 5.19, the numerical approximations behave as expected converging towards the corresponding steady state which is a stable fixed point of the method. However, increasing Δt by approximately $2 \cdot 10^{-3}$, we find that $\Delta t \rho(\mathbf{Dg}(\mathbf{y}^*)) = z_i$ for $i = 1$ or $i = 3$, respectively. As a result, even when we modify the starting vector to be $\mathbf{y}^0 = \mathbf{y}^* + 10^{-5}\mathbf{v}$ with $\mathbf{v} = (1, -2, 1)^T$, the numerical approximation diverges from the steady state as predicted by the presented theory, can be observed. All parameters however lead to a scheme that

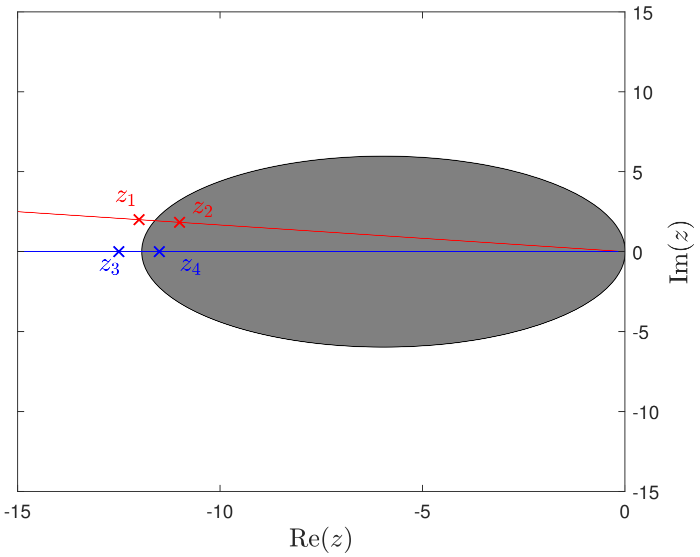


Figure 5.19: The stability region for SSPMPRK2$(0.2, 3)$. The red line is the set $\{a(-6+\mathrm{i}) \mid 2.5 \leq a \leq 0\}$. In particular, the red marked complex numbers are $z_1 = 2(-6+\mathrm{i})$ and $z_2 = \frac{11}{6}(-6+\mathrm{i})$. The blue line is the interval $[-15, 0]$. In particular, the blue marked numbers are $z_3 = -12.5$ and $z_4 = -11.5$.

also preserve the second linear invariant as Figure 5.24 suggests.

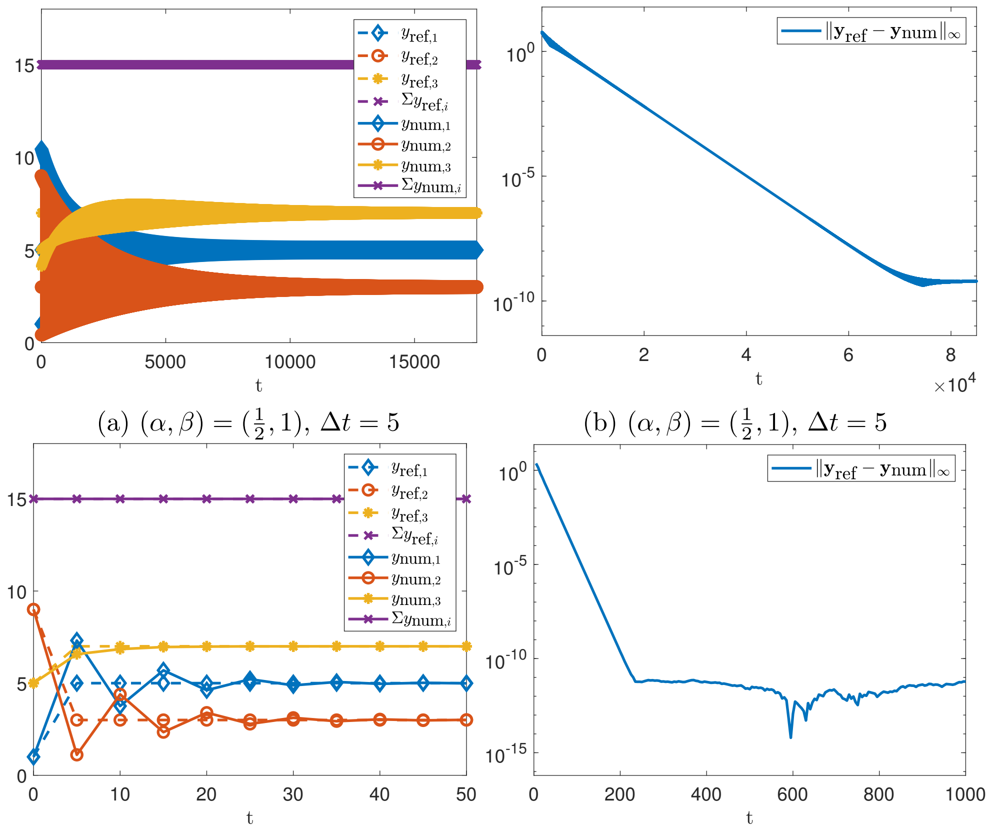
SSPMPRK3($\frac{1}{3}$)

In Figure 5.25, the SSPMPRK3($\frac{1}{3}$) scheme is used to integrate the test problems (5.125) with mixed eigenvalues and (5.123) with a second linear invariant. The numerical experiments support the theoretical claims, i. e. the fixed points seem to be stable and locally attracting. Moreover, all linear invariants are conserved by the method.

Altogether, the numerical experiments support very well the theoretical results from Section 5.4.2 on SSPMPRK methods.

5.6.4 Investigation of MPDeC Schemes

In this section we restrict to the investigation of MPDeC schemes with equidistant nodes and refer to [IÖ23] for the numerical experiments concerning MPDeCGL methods. As we have discovered in Figure 5.5, MPDeCEQ(p) for $p = 12$ and $p = 14$ have a bounded stability domain for problems with exclusively real eigenvalues. Moreover, it was observed in [TÖR22, Figure B.9] that the iterates of MPDeCEQ(8) only locally converge towards the steady state. This is in accordance with the presented theory, however, we did not observe this behavior within the numerical experiments of the previously discussed schemes. Nevertheless, MPDeCEQ(8) is not the only scheme with that rather unpleasant property. Indeed, in [IKMS23], which is based on the master thesis [Sch23], the authors demonstrate that this phenomenon also occurs with MPRK22(α) schemes for $\alpha < \frac{1}{2}$. The common circumstance for both schemes is that both are based on RK methods with non-positive Butcher tableau. The resulting hypothesis was tested and supported with numerical experiments in [IKMS23].



(a) $(\alpha, \beta) = (\frac{1}{2}, 1)$, $\Delta t = 5$

(b) $(\alpha, \beta) = (\frac{1}{2}, 1)$, $\Delta t = 5$

(c) $(\alpha, \beta) = (0.1, 1)$, $\Delta t = 5$

(d) $(\alpha, \beta) = (0.1, 1)$, $\Delta t = 5$

Figure 5.20: Numerical approximations of (5.119) using the SSPMPRK2 scheme. The dashed lines indicate the exact solution (5.120).

Nevertheless, we want to mention that if we violate the stability condition, we can start arbitrary close to the steady state solution, and still, the iterates will not converge to $\mathbf{y}^*$.

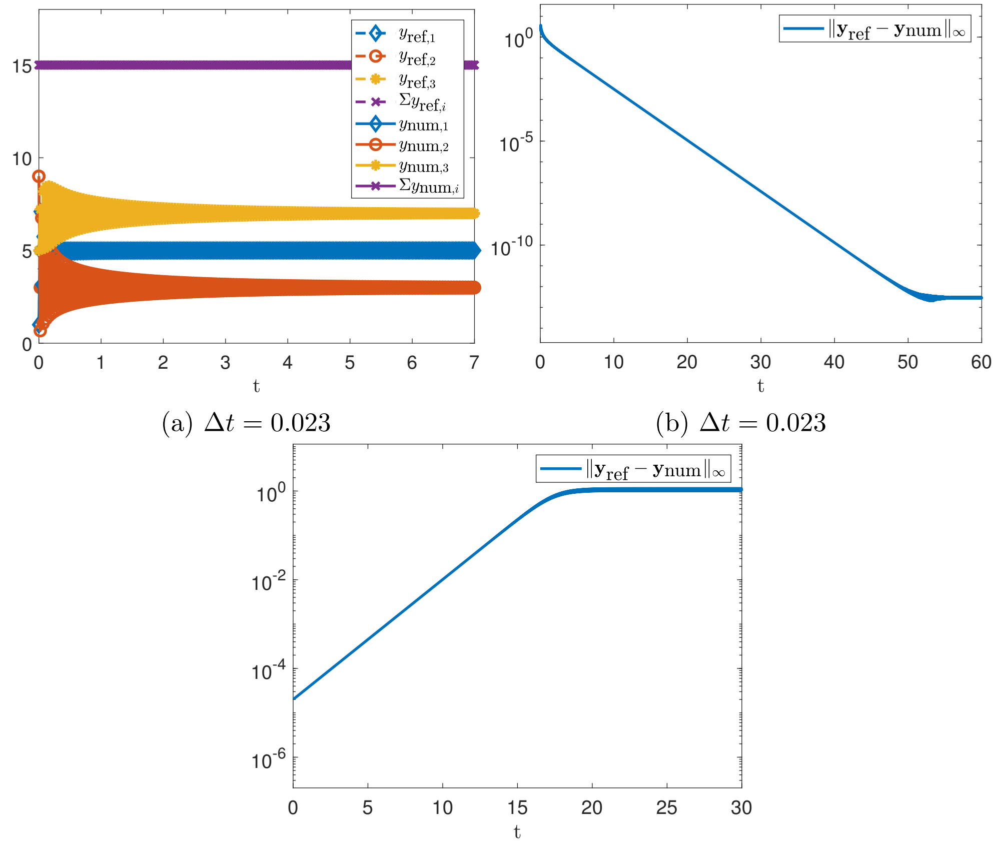
Now, we reproduce the result from [TÖR22] investigating MPDeCEQ(8), see Figure 5.26. Furthermore, we present experiments with the 12th and 14th order method when applied to (5.119), see Figure 5.27 and Figure 5.28. In both cases the largest time step size is chosen such that $|R(\Delta t^{\pm}\rho(\mathbf{\Lambda}))| = 1 \pm 0.1$ for the stable and unstable scenario, respectively, see Figure 5.5 for the graph of the stability functions. Since $\rho(\mathbf{\Lambda}) = 500$, the time step sizes for MPDeCEQ(12) are

$$\Delta t^{+}_{\mathrm{EQ}(12)} \approx \frac{59}{500} = 0.118 \quad \text{and} \quad \Delta t^{-}_{\mathrm{EQ}(12)} \approx \frac{20}{500} = 0.04.$$

In the case of MPDeCEQ(14) they are

$$\Delta t^{+}_{\mathrm{EQ}(14)} =\approx \frac{12}{500} = 0.024 \quad \text{and} \quad \Delta t^{-}_{\mathrm{EQ}(14)} \approx \frac{7.6}{500} = 0.0152.$$

Overall, the expected behavior can be observed.

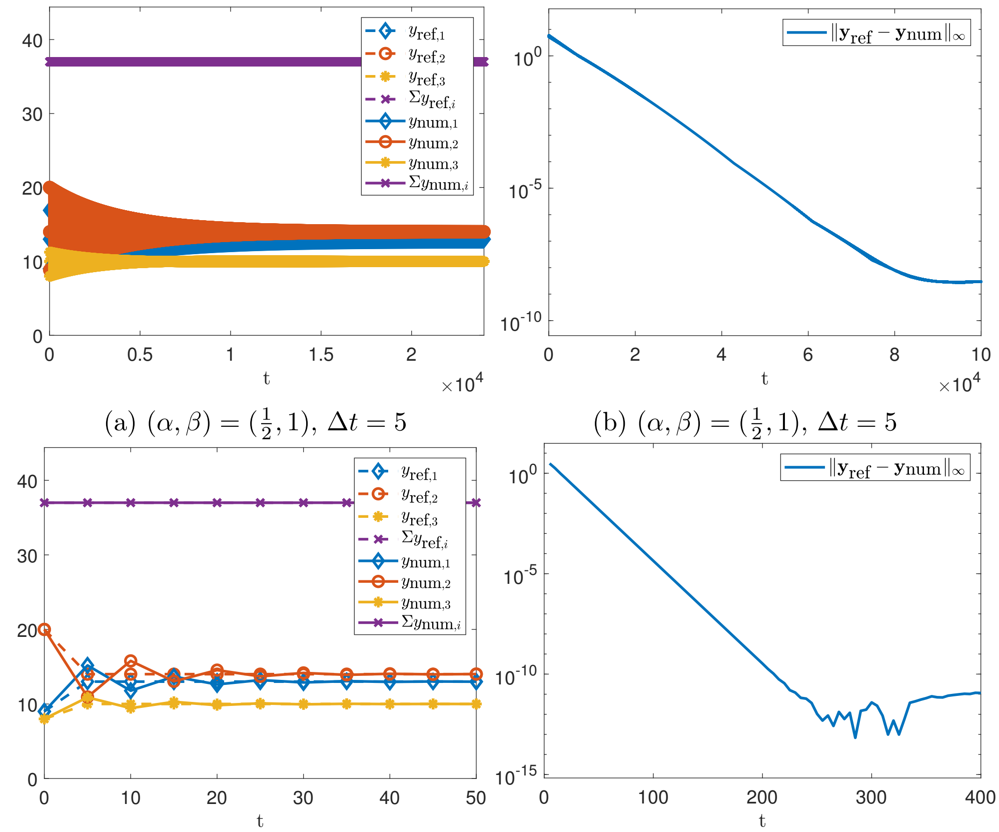


(a) $\Delta t = 0.023$

(b) $\Delta t = 0.023$

(c) $\Delta t = 0.025$

Figure 5.21: Numerical approximations of (5.119) using the SSPMPRK2(0.2, 3) scheme. The dashed lines indicate the exact solution (5.120). In 5.21c, we used $\mathbf{y}^0 = \mathbf{y}^* + 10^{-5}(1, -2, 1)^T$.



(a) $(\alpha, \beta) = (\frac{1}{2}, 1)$, $\Delta t = 5$

(b) $(\alpha, \beta) = (\frac{1}{2}, 1)$, $\Delta t = 5$

(c) $(\alpha, \beta) = (0.1, 1)$, $\Delta t = 5$

(d) $(\alpha, \beta) = (0.1, 1)$, $\Delta t = 5$

Figure 5.22: Numerical approximations of (5.121) using the second order SSPMPRK scheme. The dashed lines indicate the exact solution (5.122).

5.6.5 Investigation of GeCo Schemes

GeCo1

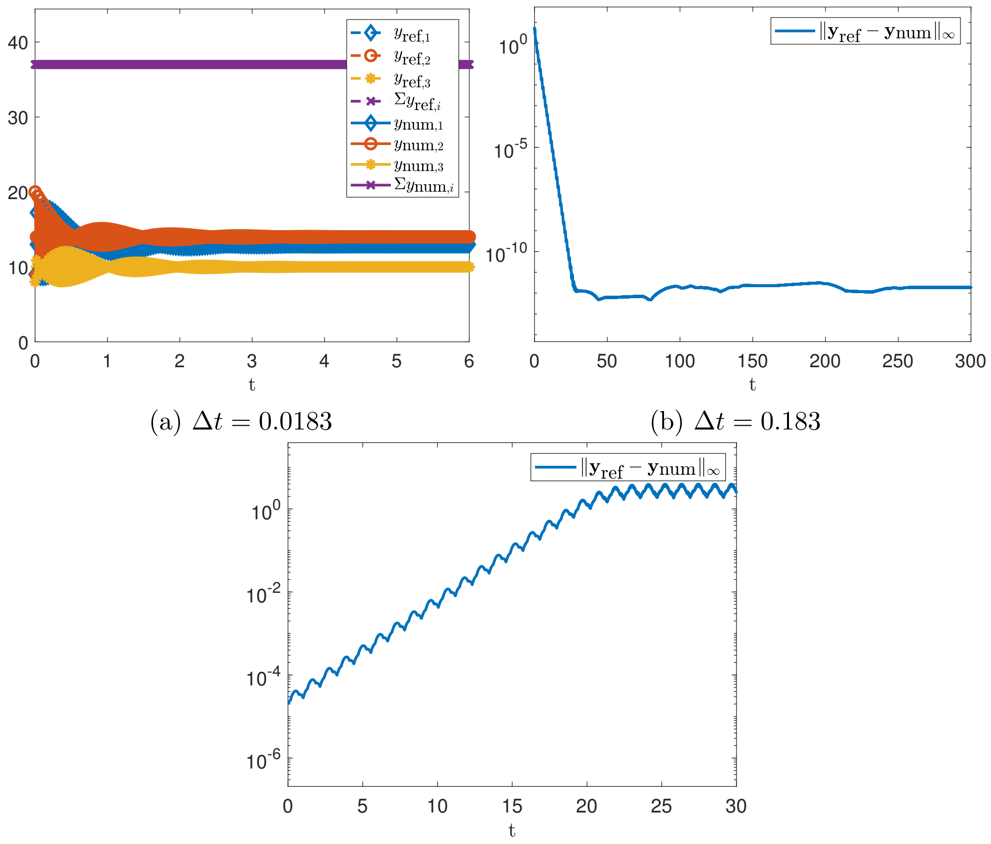
Numerical solutions obtained by GeCo1 and the corresponding error plots are shown in Figure 5.29. In error plot 5.29b, the convergence of the numerical solution to the steady state in the long run can be seen, despite the low accuracy in the short run with the comparatively large time step of $\Delta t = 5$. Hence, the result from Theorem 5.32 is well reflected here. Nevertheless, a shift of the numerical solution can be recognized for the chosen time step size. This can also be observed in Section 5.6.7, where we apply the method to increasingly stiff problems.

GeCo2

Based on the analysis for the system (5.4), we use the function

$$R(z) = 1 + z + \frac{1}{2} z^2 \varphi(\Delta t \operatorname{trace}(\mathbf{S}^-))$$

even in the context of (5.125) to determine the critical time step size $\Delta t_{\mathrm{GeCo2}}$ of GeCo2. For the system matrix (5.126), we find $\operatorname{trace}(\mathbf{S}^-) = -\operatorname{trace}(\mathbf{\Lambda}) = 20$. A numerical calculation shows that $|R(\Delta t \lambda)| < 1$ for all $\lambda \in \sigma(\mathbf{\Lambda}) \setminus \{0\}$ if $\Delta t < \Delta t_{\mathrm{GeCo2}} \approx 0.3572$, where $\Delta t_{\mathrm{GeCo2}}$ was rounded to five significant figures. Moreover, $|R(\Delta t(-5 - \sqrt{3}))| > 1$ if $\Delta t > \Delta t_{\mathrm{GeCo2}}$.



(a) $\Delta t = 0.0183$

(b) $\Delta t = 0.183$

(c) $\Delta t = 0.020$

Figure 5.23: Numerical approximations of (5.121) using the SSPMPRK22(0.2, 3) scheme. The dashed lines indicate the exact solution (5.122). In 5.23c, the initial vector $\mathbf{y}^0 = \mathbf{y}^* + 10^{-5}(1, -2, 1)^T$ is chosen.

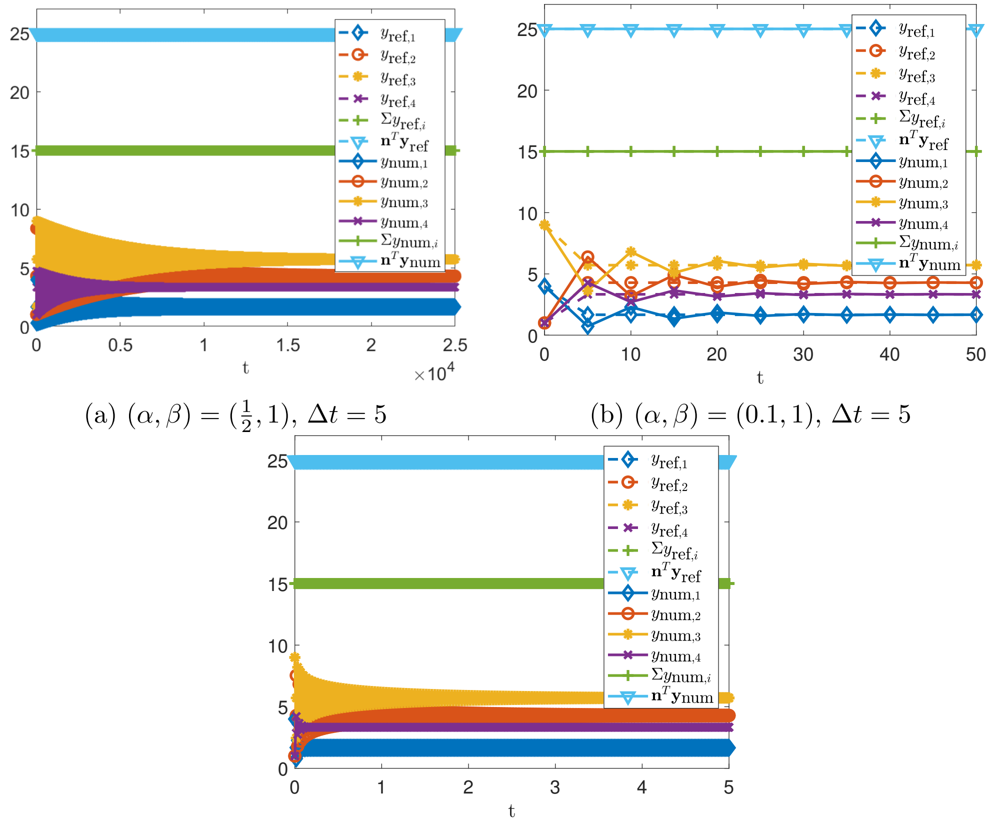
In order to numerically confirm the stability results from Theorem 5.34 even in the context of the model problem (5.125), we solve the initial value problem (5.125) using $\Delta t = \Delta t_{\text{GeCo2}} \cdot (1 - 10^{-3}) \approx 0.3569$. The expected stable behavior of GeCo2 and the convergence of the iterates can be observed in Figures 5.30a and 5.30b. In order to demonstrate the expected divergence of the iterates when $\Delta t > \Delta t_{\text{GeCo2}}$ even for starting vectors that lie within a small neighborhood of the steady state solution, we choose $\Delta t = \Delta t_{\text{GeCo2}} \cdot (1 + 10^{-3}) \approx 0.3576$ and the initial value

$$\widetilde{\mathbf{y}}^0 = \mathbf{y}^* + 10^{-5} \cdot (-2, 1, 1, -1, 1)^T.$$

In Figure 5.30c, a small decrease of the error can observed before it increases to an error of approximately 10^{-3}. Altogether, the numerical experiments reflect the expected behavior independent of $\mathbf{y}^0$, at least for the selected model problem.

5.6.6 Investigation of BBKS Schemes

The stability functions of BBKS1 and BBKS2(1) in the context of (5.4) are given by Theorem 5.37 and Theorem 5.39, respectively. We apply the schemes to the initial value problem (5.125) and test the stability for specific time step sizes. An elementary calculation reveals that the stability functions for both schemes satisfy $|R(\Delta t\lambda)| < 1$ for all $\lambda \in \sigma(\mathbf{\Lambda}) \setminus \{0\}$ if $\Delta t < \Delta t_{\text{BBKS}} = \frac{5-\sqrt{3}}{11}$, and $|R(\Delta t(-5 - \sqrt{3}))| > 1$ if $\Delta t > \Delta t_{\text{BBKS}}$. As we did for GeCo2, we investigate the



(a) $(\alpha, \beta) = (\frac{1}{2}, 1)$, $\Delta t = 5$

(b) $(\alpha, \beta) = (0.1, 1)$, $\Delta t = 5$

(c) $(\alpha, \beta) = (0.2, 3)$, $\Delta t = 0.0164$

Figure 5.24: Numerical approximations of (5.123) using the second order SSPMPRK scheme. The dashed lines indicate the exact solution (5.124), where $\mathbf{n} = (1, 2, 2, 1)^T$.

BBKS schemes by varying the time step size around Δt_{BBKS} by multiplying with 1 ± 10^{-3}, respectively. Furthermore, we also choose $\widetilde{\mathbf{y}}^0 = \mathbf{y}^* + 10^{-5} \cdot (-2, 1, 1, -1, 1)^T$ in the case $\Delta t > \Delta t_{\mathrm{BBKS}}$ in order to highlight the expected divergence of the iterates.

In Figure 5.31 the numerical solutions of (5.125) and the error plots using BBKS1 are shown. In 5.31a, corresponding to the step size

$$\Delta t = \Delta t_{\mathrm{BBKS}} \cdot (1 - 10^{-3}) \approx 0.2968,$$

all components of the numerical solution tend to the reference solution in the long run, with an error between 10^{-13} and 10^{-12}. In the unstable case, see Figure 5.31c, when

$$\Delta t = \Delta t_{\mathrm{BBKS}} \cdot (1 + 10^{-3}) \approx 0.2974,$$

the error increases almost to 10^{-3}. Similar conclusions can be deduced by looking at Figure 5.32, where the numerical solutions and the error plots of BBKS2(1) are shown, in correspondence of the same step sizes used for BBKS1.

Altogether, the stability properties shown in Figures 5.31 and 5.32 are in accordance with the stability results expected from the theory presented in Section 5.4.5.

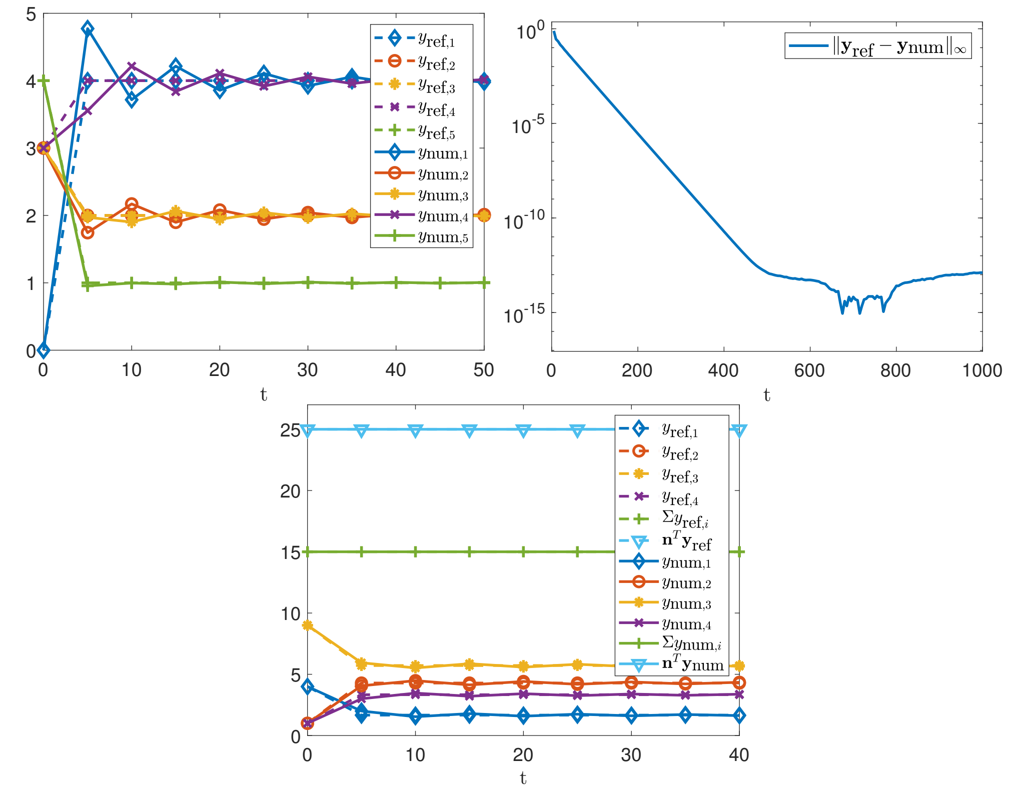
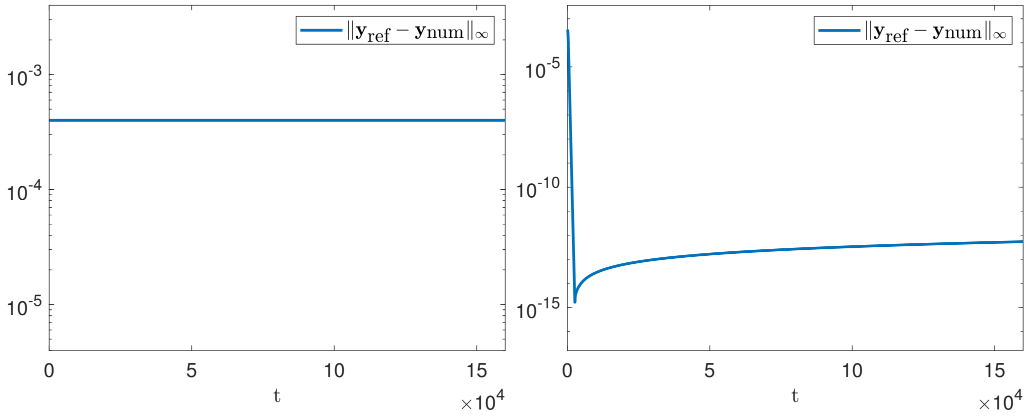


Figure 5.25: Numerical approximations of (5.125) with error plot and (5.123) using SSPMPRK3($\frac{1}{3}$) schemes and $\Delta t = 5$. The dashed lines indicate the reference solutions and $\mathbf{n} = (1, 2, 2, 1)^T$.



(a) $\epsilon = 0.1$

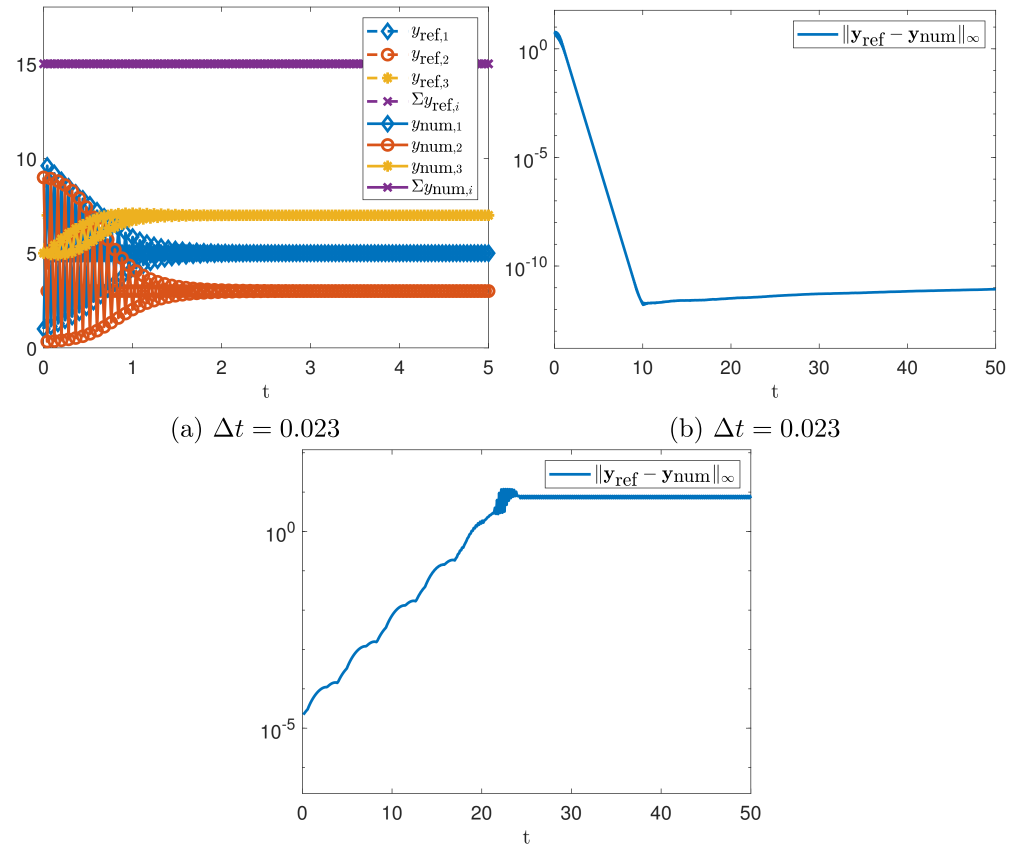
(b) $\epsilon = 0.01$

Figure 5.26: Error of the numerical solution of (5.20) using the MPDeCEQ(8) scheme and $\mathbf{y}^0 = \mathbf{y}^* + \epsilon(-1, 1)^T$.

5.6.7 Applicability of GeCo1 to Stiff Problems

Since the GeCo1 scheme is stable for arbitrary time step sizes, at least locally, this scheme might be able to solve stiff problems. Unfortunately, this is not true as demonstrated in [IKMM23]. In the following we present the investigation from Kopecz performed therein.

To assess the usability for stiff problems, Kopecz [IKMM23] proposed to



(a) $\Delta t = 0.023$

(b) $\Delta t = 0.023$

(c) $\Delta t = 0.025$

Figure 5.27: Numerical approximations of (5.119) using the MPDeCEQ(12) scheme. The dashed lines indicate the exact solution (5.120). In Figure 5.27c, we used $\mathbf{y}^0 = \mathbf{y}^* + 10^{-5}(1, -2, 1)^T$.

consider the linear initial value problem $\mathbf{y}' = \mathbf{\Lambda}\mathbf{y}$, $\mathbf{y}(0) = \mathbf{y}^0$ with

$$\mathbf{\Lambda} = \begin{pmatrix} -K & 0 & 0 \\ K & -1 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \quad \mathbf{y}^0 = \begin{pmatrix} 0.98 \\ 0.01 \\ 0.01 \end{pmatrix}. \tag{5.127}$$

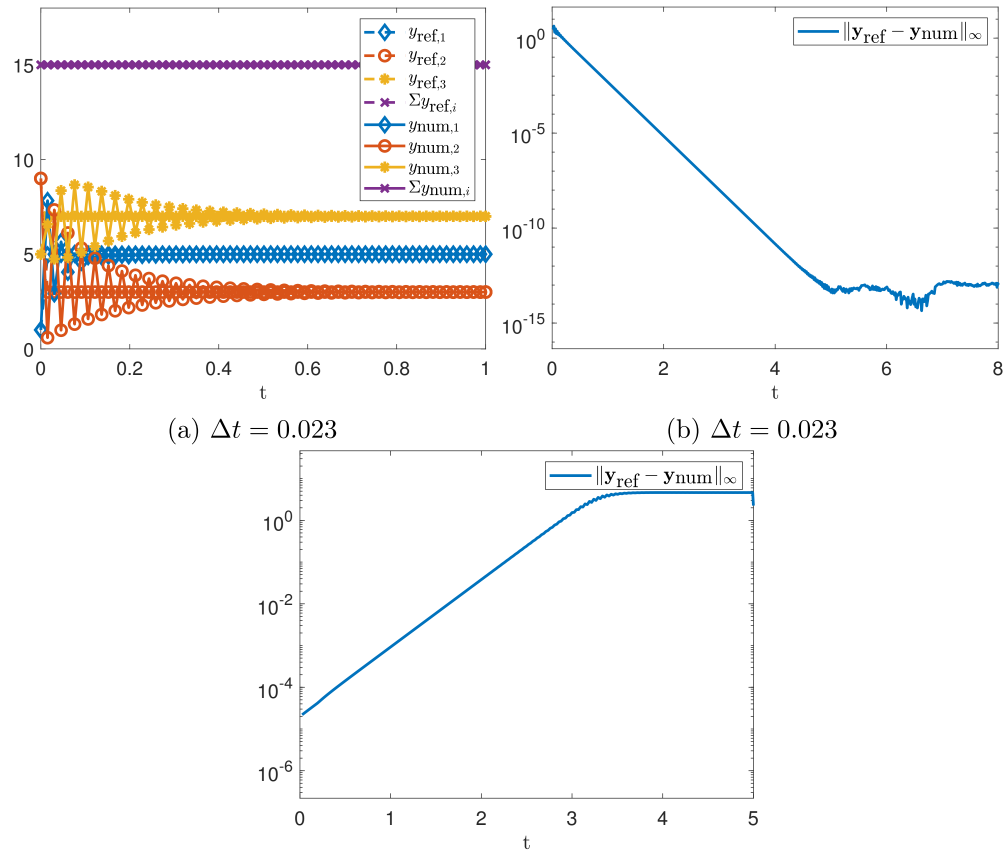
This system becomes increasingly stiff as the value of $K > 0$ is increased. For $K \neq 1$ the solution reads

$$y_1(t) = \frac{49e^{-Kt}}{50}, \quad y_2(t) = \frac{(99K-1)e^{-t}}{100(K-1)} - \frac{49Ke^{-Kt}}{50(K-1)},$$
$$y_3(t) = 1 - \frac{(99K-1)e^{-t}}{100(K-1)} + \frac{49e^{-Kt}}{50(K-1)}.$$

Defining $\hat{\mathbf{y}}(t) = \lim_{K\to\infty} \mathbf{y}(t)$ we find

$$\hat{y}_1(t) = 0, \quad \hat{y}_2(t) = \frac{99}{100}e^{-t}, \hat{y}_3(t) = 1 - \frac{99}{100}e^{-t}$$

for $t > 0$. In the limit $K \to \infty$, y_2 and y_3 should therefore be equal at approximately $t = 0.7$. In Figure 5.33, we present the plots from [IKMM23] of GeCo1 solving (5.127) for different values of K. As already observed in our previous numerical



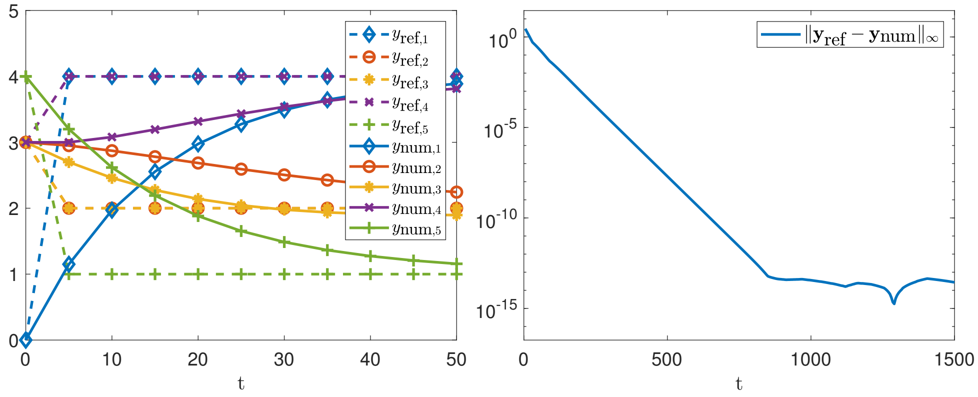
(a) $\Delta t = 0.023$

(b) $\Delta t = 0.023$

(c) $\Delta t = 0.025$

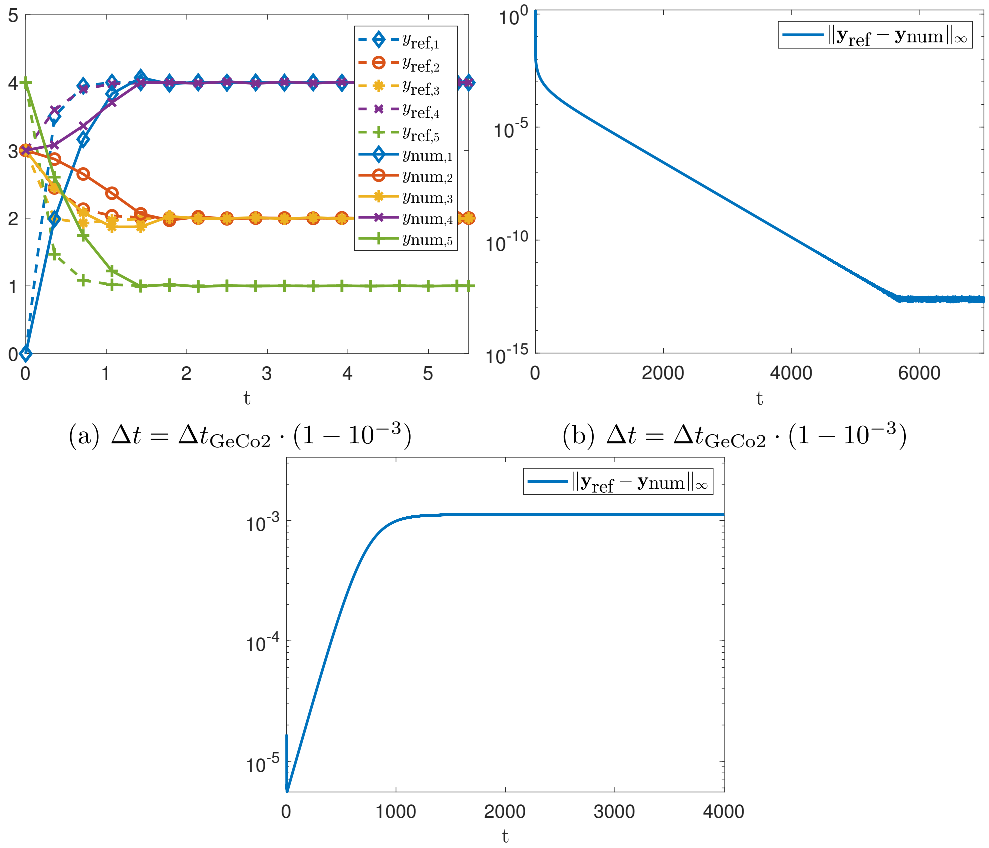
Figure 5.28: Numerical approximations of (5.119) using the MPDeCEQ14) scheme. The dashed lines indicate the exact solution (5.120). In 5.28c, the initial vector $\mathbf{y}^0 = \mathbf{y}^* + 10^{-5}(1, -2, 1)^T$ is chosen.

experiments, there is a significant phase error so that y_2 and y_3 are equal at about $t = 7$ for $K = 10$ and about $t = 70$ for $K = 100$, which is far from $t = 0.7$. Hence, for increasingly stiff problems, GeCo1 gets less accurate if the time step size is not adapted correspondingly. Altogether, this means that GeCo1 can hardly be regarded as a stiff solver.



(a) $\Delta t = 5$ (b) $\Delta t = 5$

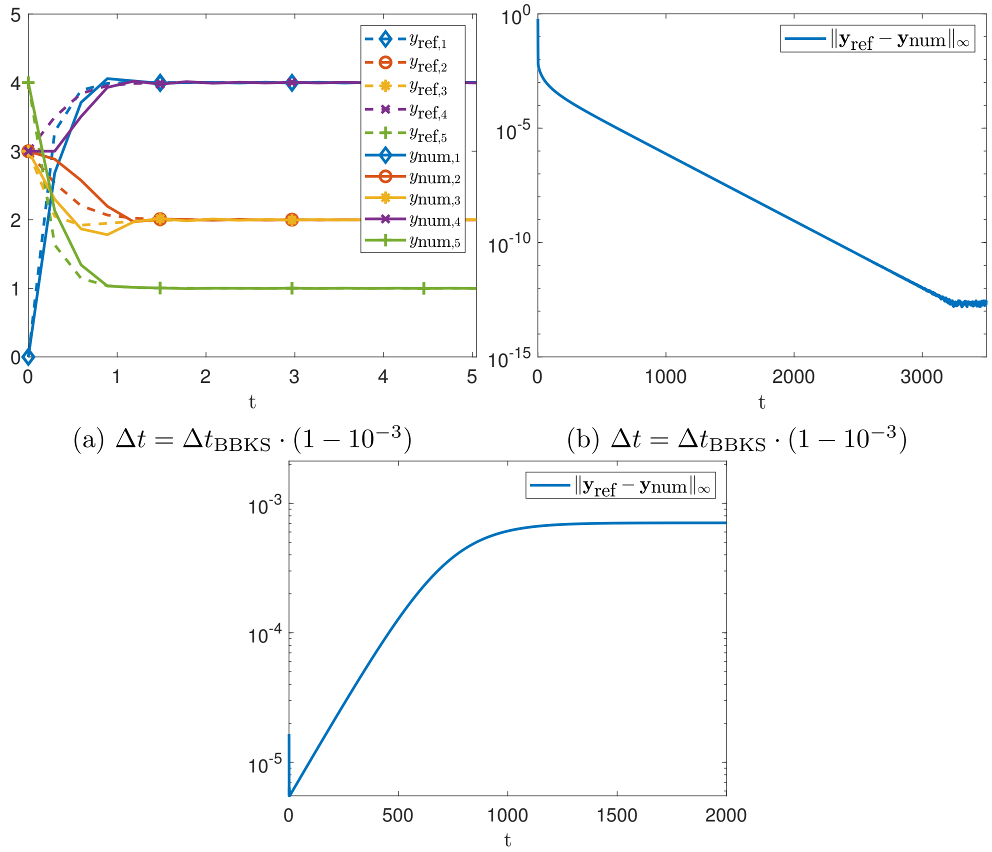
Figure 5.29: Numerical approximations of (5.125) and error plot using GeCo1.



(a) $\Delta t = \Delta t_{\mathrm{GeCo2}} \cdot (1 - 10^{-3})$ (b) $\Delta t = \Delta t_{\mathrm{GeCo2}} \cdot (1 - 10^{-3})$

(c) $\Delta t = \Delta t_{\mathrm{GeCo2}} \cdot (1 + 10^{-3})$

Figure 5.30: Numerical approximation of (5.125) and error plots using GeCo2. In 5.30c the starting vector $\widetilde{\mathbf{y}}^0 = \mathbf{y}^* + 10^{-5} \cdot (-2, 1, 1, -1, 1)^T$ was used.

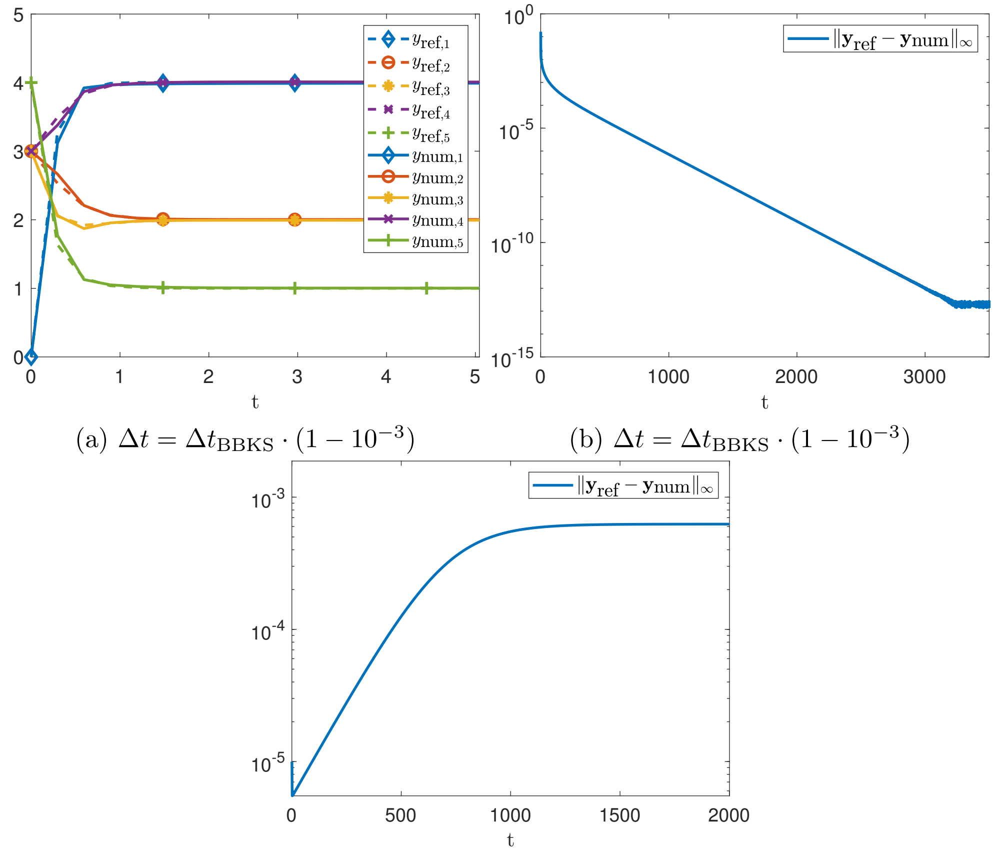


(a) $\Delta t = \Delta t_{\mathrm{BBKS}} \cdot (1 - 10^{-3})$

(b) $\Delta t = \Delta t_{\mathrm{BBKS}} \cdot (1 - 10^{-3})$

(c) $\Delta t = \Delta t_{\mathrm{BBKS}} \cdot (1 + 10^{-3})$

Figure 5.31: Numerical approximations of (5.125) and error plots using BBKS1. The starting vector $\widetilde{\mathbf{y}}^0 = \mathbf{y}^* + 10^{-5} \cdot (-2, 1, 1, -1, 1)^T$ was chosen in 5.31c.

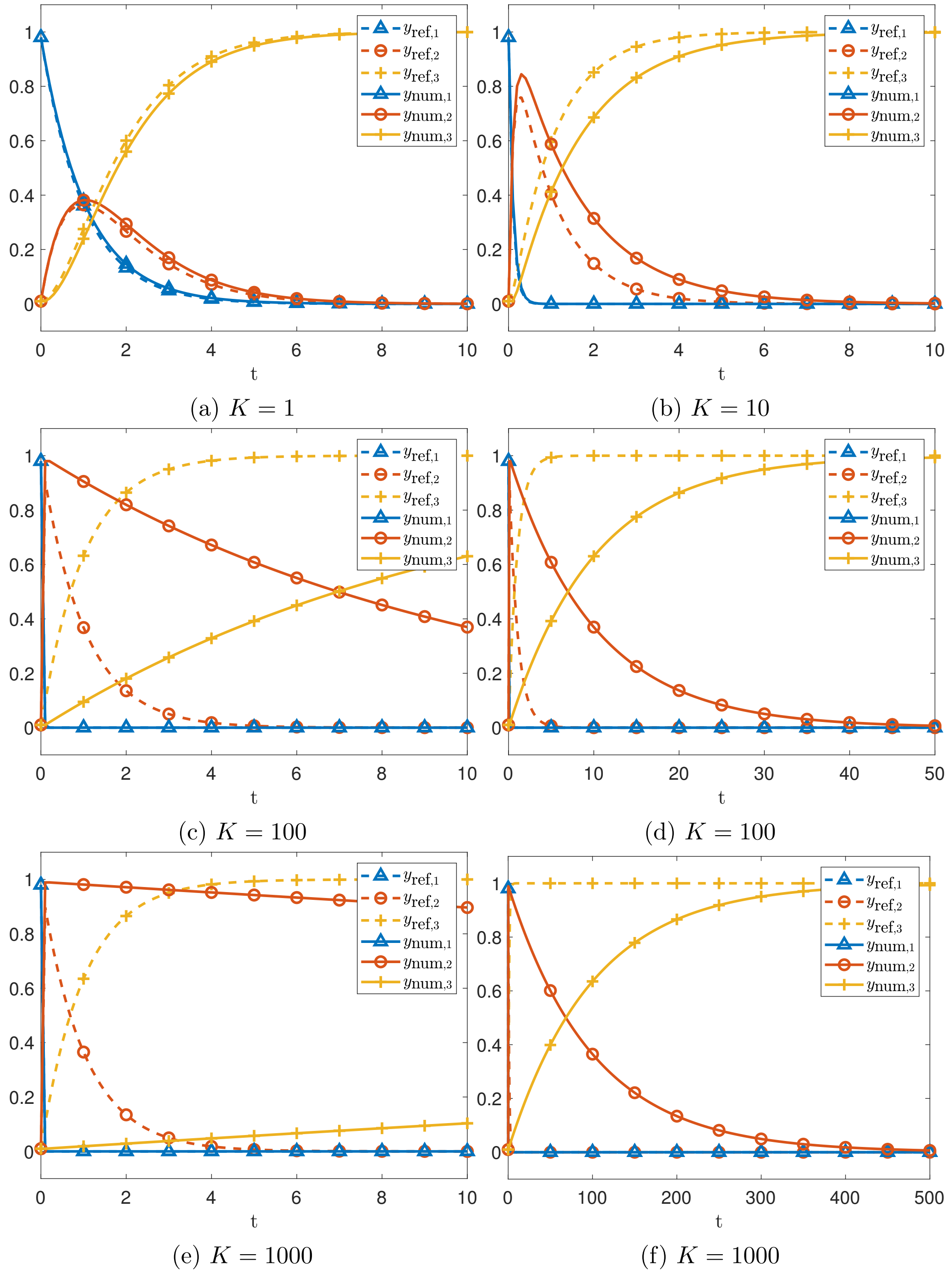


(a) $\Delta t = \Delta t_{\mathrm{BBKS}} \cdot (1 - 10^{-3})$

(b) $\Delta t = \Delta t_{\mathrm{BBKS}} \cdot (1 - 10^{-3})$

(c) $\Delta t = \Delta t_{\mathrm{BBKS}} \cdot (1 + 10^{-3})$

Figure 5.32: Numerical approximations of (5.125) and error plots using BBKS2(1). In 5.32c the starting vector $\widetilde{\mathbf{y}}^0 = \mathbf{y}^* + 10^{-5} \cdot (-2, 1, 1, -1, 1)^T$ was chosen.



(a) $K = 1$

(b) $K = 10$

(c) $K = 100$

(d) $K = 100$

(e) $K = 1000$

(f) $K = 1000$

Figure 5.33: Numerical solutions of (5.127) computed with GeCo1 for different values of K [IKMM23]. The step size used is $\Delta t = 0.1$. The dashed lines indicate the reference solution.

Chapter 6

Conclusion and Outlook

The present work dealt with two major topics concerning the numerical analysis of Runge–Kutta-like methods, namely their stability and order of convergence.

We motivated and introduced modified Patankar schemes as a subclass of Runge–Kutta-like methods and emphasized their importance. The first major part of this thesis was then dedicated to providing a tool for deriving order conditions for MP methods. The proposed approach may yields implicit order conditions, which can be rewritten in explicit form using the NB-series of the stages [IKM23b]. The obtained explicit order conditions can be further reduced using Gröbner bases computations. With the presented approach, it was possible for the first time to obtain conditions for the construction of 3rd and 4th order GeCo as well as 4th order MPRK schemes. Moreover, we constructed a new 4th order MPRK method using our theory and validated the order of convergence numerically. Future work within this topic include the adaptation of this approach for further nonlinear methods such as SSPMPRK schemes and the construction of higher order schemes. In particular, constructing 4th order MPRK methods with a minimal number of stages is of interest. Furthermore, to investigate the order of GeCo and gBBKS methods in the context of non-autonomous problems is to the authors best knowledge still an open task.

The second major part was concerned with the stability of nonlinear time integrators preserving at least one linear invariant. We discussed how the given approach generalizes the notion of A-stability. The main difficulty in the analysis comes from the presence of linear invariants, so that any steady state of the corresponding linear system of ODEs resulted in a non-hyperbolic fixed point of the steady state preserving nonlinear method. Even though the investigation of non-hyperbolic fixed points in general is a case by case study, we were able to find an exception for steady states forming a subspace, as is the case for the linear test problem we considered. As a result, we were able to prove that investigating the Jacobian of the generating map is sufficient to understand the stability of the nonlinear method in a neighborhood of the steady state. This approach allowed for the first time the investigation of several modified Patankar schemes such as MPRK, SSPMPRK, MPDeC, GeCo and gBBKS methods which was performed in [IKM22b, HIK$^+$23, IÖ23, IKMM23], also presented and extended within this work. In particular, we tackled the question of unconditional stability for all of the above mentioned methods and summarized our findings in Table 5.2. In addition to that, we demonstrated that GeCo2 and gBBKS methods are not in $\mathcal{C}^2$ and proved asymptotic stability for GeCo1 schemes and, for some PDRS, also for MPRK methods. The investigation of MPRK schemes together with the analysis for gBBKS and GeCo2 methods applied to general linear systems represents a future research topic. In the particular case of MPRK schemes, we computed the stability function for arbitrary MPRK schemes in a way that can be easily adapted to the case of PDRS while we pointed out ideas how to generalize our findings for GeCo2 and gBBKS. Finally, our findings support the numerically observed robustness of MP methods while we were able to provide sharp bounds on the time step in the case of conditional stability. Moreover, it might be interesting to apply the presented stability theory in the context of linear multistep methods.

We also connected the approach coming from dynamical systems with that of [TÖR22] concerning oscillatory behavior of nonlinear methods. Here, the zeros of the respective stability function are interlinked with a necessary condition for avoiding oscillatory behavior, which was numerically validated in [IÖT22].

Although the proven stability properties are initially local in nature, the work [IKMS23] suggests that they can be provably global if the underlying Butcher tableau contains only non-negative entries, while there are schemes with negative Butcher entries for which the stability properties are only local. To further investigate or even prove this claim is of high importance and will be part of my future research.

Also, the implications of this approach for the analysis of numerical methods in the context of partial differential equations (PDEs) is of interest. In particular, generalizing the main stability result, Theorem 5.4, to the infinite dimensional case promises interesting applications in the field of numerical analysis of PDEs, as refining the grid in space of the semi-discrete system corresponds to increasingly larger systems of ODEs.

It is also worth mentioning that there are two further tasks arsing naturally as future research topics.

First, there is not much work available concerning the efficiency of modified Patankar schemes using a time step controller. To the authors knowledge, there is only [KMP21], where standard step size controller were applied to Patankar–Runge–Kutta methods. However, also considering more general controllers from digital signal processing [Söd06, SW06, Söd02, Söd03, GLS88, Gus91, Gus94, Zon64] might result in even better performances. The exploration of such controllers is one of my future research topics.

A second aspect related to efficiency is the construction of dense output formulae for MP schemes. The major task here is to provide not only an approximation for any point in time within a given order of accuracy but to force the approximation to be also positive and conservative, or linear invariant preserving in general. Following the idea from [KLJK17], it seems to be possible to construct second order dense output formulae, i. e. for third order MP methods, using our approach of NB-series. However, in the same work the authors find a negative result, i. e. using their approach there is no third order dense output formula for MP scheme based on a Butcher tableau with non-negative entries [KLJK17, Theorem 1]. However, as we wish to use only non-negative Butcher arrays for reasons of stability, we are forced to take a different approach for constructing even third order dense output formulae for MP methods. To construct such a formula together with the above mentioned properties is to my best knowledge still an open problem, yet of high importance. If such formulae are available they might be also useful to construct higher order MPRK methods since the PWDs need to be positive approximations to classical Runge–Kutta stages, that is to the exact solution at intermediate times.

Appendix A

Intermediate Results for the Stability Analysis

In this appendix, we present results with rather technical proofs.

Lemma A.1. Let $R(z) = \frac{\sum_{j=0}^{4} n_j z^j}{\sum_{j=0}^{4} d_j z^j}$ with $d_0 = 1$. Then $|R(re^{\mathrm{i}\varphi})| < 1$ with $r > 0$ and $\varphi \in [0, 2\pi)$ is equivalent to

$$\begin{aligned} p_\varphi(r) =& (-d_4^2 + n_4^2)r^8 + 2(-d_3d_4 + n_3n_4)\cos(\varphi)r^7 \\ &+ (4(-d_2d_4 + n_2n_4)\cos(\varphi)^2 + 2d_2d_4 - d_3^2 - 2n_2n_4 + n_3^2)r^6 \\ &+ (8(-d_1d_4 + n_1n_4)\cos(\varphi)^3 + 2(3d_1d_4 - d_2d_3 - 3n_1n_4 + n_2n_3)\cos(\varphi))r^5 \\ &+ (16(n_0n_4 - d_4)\cos(\varphi)^4 + 4(-d_1d_3 - 4n_0n_4 + n_1n_3 + 4d_4)\cos(\varphi)^2 \\ &+ 2d_1d_3 - d_2^2 + 2n_0n_4 - 2n_1n_3 + n_2^2 - 2d_4)r^4 \\ &+ (8(n_0n_3 - d_3)\cos(\varphi)^3 + 2(-d_1d_2 - 3n_0n_3 + n_1n_2 + 3d_3)\cos(\varphi))r^3 \\ &+ (4(n_0n_2 - d_2)\cos(\varphi)^2 - d_1^2 - 2n_0n_2 + n_1^2 + 2d_2)r^2 \\ &+ 2(n_0n_1 - d_1)\cos(\varphi)r + n_0^2 - 1 < 0 \end{aligned}$$

Furthermore, $|R(re^{\mathrm{i}\varphi})| > 1$ is equivalent to $p_\varphi(r) > 0$.

Proof. A straightforward calculation rewriting

$$1 > |R(re^{\mathrm{i}\varphi})|^2 = \frac{\left(\sum_{j=0}^{4} r^j n_j \cos(j\varphi)\right)^2 + \left(\sum_{j=0}^{4} r^j n_j \sin(j\varphi)\right)^2}{\left(\sum_{j=0}^{4} r^j d_j \cos(j\varphi)\right)^2 + \left(\sum_{j=0}^{4} r^j d_j \sin(j\varphi)\right)^2}$$

yields the result. □

The next statement provides us conditions under which the product of a scalar continuous function and a partially differentiable vector field is partially differentiable again, and conditions under which a partial derivative of the product does not exist.

Lemma A.2. Let $D \subseteq \mathbb{R}^N$ be open and $\mathbf{e}_i$ denote the ith unit vector in $\mathbb{R}^N$. Furthermore, let $\mathbf{\Phi}\colon D \to \mathbb{R}^N$ be partially differentiable in $\mathbf{x}_0 \in D$ with $\mathbf{\Phi}(\mathbf{x}_0) = \mathbf{0}$ and let $\Psi\colon D \to \mathbb{R}$.

a) If Ψ is continuous in $\mathbf{x}_0$, then the product $\Psi \cdot \mathbf{\Phi}\colon D \to \mathbb{R}^N$ is partially differentiable in $\mathbf{x}_0$ with

$$\mathbf{D}(\Psi \cdot \mathbf{\Phi})(\mathbf{x}_0) = \Psi(\mathbf{x}_0)\mathbf{D\Phi}(\mathbf{x}_0).$$

b) If $\Psi(\mathbf{x}_0 + \mathbf{e}_i h)$ has several accumulation points as $h \to 0$ and $\partial_i \mathbf{\Phi}(\mathbf{x}_0) \neq \mathbf{0}$, then the ith partial derivative of $\Psi \cdot \mathbf{\Phi}$ does not exists.

Proof. a) Since $\mathbf{\Phi}(\mathbf{x}_0) = \mathbf{0}$ we have

$$\begin{aligned} &\frac{\Psi(\mathbf{x}_0 + h\mathbf{e}_i)\mathbf{\Phi}(\mathbf{x}_0 + h\mathbf{e}_i) - \Psi(\mathbf{x}_0)\mathbf{\Phi}(\mathbf{x}_0)}{h} \\ &\quad= \frac{\Psi(\mathbf{x}_0 + h\mathbf{e}_i)\mathbf{\Phi}(\mathbf{x}_0 + h\mathbf{e}_i) - \Psi(\mathbf{x}_0 + h\mathbf{e}_i)\mathbf{\Phi}(\mathbf{x}_0)}{h} \\ &\quad= \Psi(\mathbf{x}_0 + h\mathbf{e}_i) \cdot \frac{\mathbf{\Phi}(\mathbf{x}_0 + h\mathbf{e}_i) - \mathbf{\Phi}(\mathbf{x}_0)}{h}. \end{aligned} \tag{A.1}$$

Passing to the limit $h \to 0$ on both sides shows

$$\frac{\partial(\Psi\mathbf{\Phi})}{\partial x_i}(\mathbf{x}_0) = \Psi(\mathbf{x}_0)\frac{\partial\mathbf{\Phi}}{\partial x_i}(\mathbf{x}_0), \quad i = 1, \dots, N,$$

and hence

$$\mathbf{D}(\Psi\mathbf{\Phi})(\mathbf{x}_0) = \Psi(\mathbf{x}_0)\mathbf{D\Phi}(\mathbf{x}_0).$$

b) If $\Psi(\mathbf{x}_0 + \mathbf{e}_i h)$ possesses several accumulation points as $h \to 0$, then this is also true for

$$\Psi(\mathbf{x}_0 + h\mathbf{e}_i) \cdot \frac{\mathbf{\Phi}(\mathbf{x}_0 + h\mathbf{e}_i) - \mathbf{\Phi}(\mathbf{x}_0)}{h}$$

as $\frac{\partial\mathbf{\Phi}}{\partial x_i}(\mathbf{x}_0) \neq \mathbf{0}$. As a result of (A.1) we thus obtain that $\frac{\partial(\Psi\cdot\mathbf{\Phi})}{\partial x_i}(\mathbf{x}_0)$ does not exist.

□

The last result of this section is concerned with sufficient conditions for a map $T\colon \mathbb{R}^2_{>0} \to \mathbb{R}$ to be locally Lipschitz continuous even though it is not in $\mathcal{C}^1$ on its entire domain.

Lemma A.3. Let $\mathbf{\Lambda} \in \mathbb{R}^{2\times 2}$ be given by (5.4) and set $D_1 = \{\mathbf{x} \in \mathbb{R}^2_{>0} \mid x_1 > \frac{b}{a}x_2\}$, $D_2 = \{\mathbf{x} \in \mathbb{R}^2_{>0} \mid x_1 < \frac{b}{a}x_2\}$ and $C = \ker(\mathbf{\Lambda}) \cap \mathbb{R}^2_{>0}$. Let $T : \mathbb{R}^2_{>0} \to \mathbb{R}$ be continuous with $T|_C = \text{const}$ and $T|_{D_i} \in \mathcal{C}^1$ for $i = 1, 2$. If $\lim_{\mathbf{x}\to\mathbf{c}} \nabla T(\mathbf{x})$ exists for any $\mathbf{c} \in C$, then T is locally Lipschitz continuous.

Proof. Note that $C = \partial D_1 = \partial D_2$ and that T is locally Lipschitz on D_1 and D_2 because $T|_{D_i} \in \mathcal{C}^1$ for $i = 1, 2$. As a first step, we prove that T is also locally Lipschitz on $\overline{D_i} = D_i \cup C$. For this, we consider closed half balls

$$H_{\epsilon,i}(\mathbf{v}) = \overline{B_\epsilon(\mathbf{v})} \cap \overline{D_i},$$

where $\mathbf{v} \in C$ and $B_\epsilon(\mathbf{v})$ denotes the open ball with center $\mathbf{v}$ and radius $\epsilon > 0$.

As the limit $\lim_{\mathbf{x}\to\mathbf{c}} \nabla T(\mathbf{x})$ exists for any $\mathbf{c} \in C$, we can consider the continuous extension of ∇T to the set $H_{\epsilon,i}(\mathbf{v})$, denoted by $\widetilde{\mathbf{T}}$. Thus, the mean value theorem and the Cauchy–Schwarz inequality yield

$$|T(\mathbf{x}_1) - T(\mathbf{x}_2)| \leq \sup_{\mathbf{x}\in\overline{B_\epsilon(\mathbf{v})}\cap D_i} \|\nabla T(\mathbf{x})\|_2 \|\mathbf{x}_1 - \mathbf{x}_2\|_2 = \max_{\mathbf{x}\in H_{\epsilon,i}(\mathbf{v})} \|\widetilde{\mathbf{T}}(\mathbf{x})\|_2 \|\mathbf{x}_1 - \mathbf{x}_2\|_2 \tag{A.2}$$

for $\mathbf{x}_1, \mathbf{x}_2 \in \overline{B_\epsilon(\mathbf{v})} \cap D_i$, which means that T is Lipschitz continuous on $\overline{B_\epsilon(\mathbf{v})} \cap D_i$ for $i \in \{1, 2\}$.

Note that $T|_C = \text{const}$ implies that T is Lipschitz continuous on C. Hence, to prove the Lipschitz continuity on the closed half ball $H_{\epsilon,i}(\mathbf{v})$ it remains to consider the case $\mathbf{x}_1 \in C$ and $\mathbf{x}_2 \in \overline{B_\epsilon(\mathbf{v})} \cap D_i$ with $i \in \{1, 2\}$. For this, we introduce a sequence $(\mathbf{x}^n)_{n\in\mathbb{N}} \subseteq \overline{B_\epsilon(\mathbf{v})} \cap D_i$ with $\lim_{n\to\infty} \mathbf{x}^n = \mathbf{x}_1$. As T is continuous we therefore find $N_0 \in \mathbb{N}$ such that for all $n \geq N_0$ we have

$$|T(\mathbf{x}_1) - T(\mathbf{x}^n)| < \frac{1}{n}. \tag{A.3}$$

Altogether, using $L_i = \max_{\mathbf{x}\in H_{\epsilon,i}(\mathbf{v})} \|\widetilde{\mathbf{T}}(\mathbf{x})\|_2$ we obtain from (A.2) and (A.3)

$$|T(\mathbf{x}_1) - T(\mathbf{x}_2)| \leq |T(\mathbf{x}_1) - T(\mathbf{x}^n)| + |T(\mathbf{x}^n) - T(\mathbf{x}_2)| < \frac{1}{n} + L_i \|\mathbf{x}^n - \mathbf{x}_2\|_2,$$

and passing to the limit, we see that T is even Lipschitz continuous on the closed half ball with a Lipschitz constant L_i.

Next, we prove that for any $\mathbf{x} \in D_1$ and $\mathbf{y} \in D_2$ there exists a $\mathbf{z} \in C$ such that

$$\|\mathbf{x} - \mathbf{y}\|_2 = \|\mathbf{x} - \mathbf{z}\|_2 + \|\mathbf{z} - \mathbf{y}\|_2. \tag{A.4}$$

That is to say that $\mathbf{z}$ lies on the straight line between $\mathbf{x}$ and $\mathbf{y}$. Indeed, setting

$$\mathbf{z} = \mathbf{x} + c(\mathbf{y} - \mathbf{x}), \quad c = \frac{x_1 - \frac{b}{a}x_2}{x_1 - \frac{b}{a}x_2 + \frac{b}{a}y_2 - y_1},$$

we find $c \in (0, 1)$ as $\mathbf{x} \in D_1$ and $\mathbf{y} \in D_2$. Additionally, $\mathbf{z} \in \ker(\mathbf{\Lambda})$ since

$$\begin{aligned} z_1 - \frac{b}{a}z_2 &= x_1 + c(y_1 - x_1) - \frac{b}{a}(x_2 + c(y_2 - x_2)) \\ &= x_1 - \frac{b}{a}x_2 - c\left(x_1 - y_1 + \frac{b}{a}y_2 - \frac{b}{a}x_2\right) = 0, \end{aligned}$$

and $\mathbf{z} > \mathbf{0}$ since it is on the line between $\mathbf{x} > \mathbf{0}$ and $\mathbf{y} > \mathbf{0}$.

Let us now prove that T is Lipschitz continuous on $\overline{B_\epsilon(\mathbf{v})}$. For this, let $\mathbf{x} \in D_1$ and $\mathbf{y} \in D_2$, then choose $\mathbf{z} \in C$ such that (A.4) is satisfied. As a result we obtain

$$\begin{aligned} |T(\mathbf{x}) - T(\mathbf{y})| &\leq |T(\mathbf{x}) - T(\mathbf{z})| + |T(\mathbf{z}) - T(\mathbf{y})| \\ &\leq \max\{L_1, L_2\}(\|\mathbf{x} - \mathbf{z}\|_2 + \|\mathbf{z} - \mathbf{y}\|_2) = \max\{L_1, L_2\}\|\mathbf{x} - \mathbf{y}\|_2, \end{aligned}$$

and since $\mathbf{v} \in C$ and $\epsilon > 0$ are arbitrary, we have proven hat T is locally Lipschitz continuous. □

Appendix B

Intermediate Results for Nonstandard NB-Series

In this appendix we present and prove intermediate results that are analogous to statements in [But16]. We start by recalling Theorem 308A from [But16], for which we briefly introduce the notation.

Let $m \in \mathbb{N}$ and I be a non-decreasing and finite sequence of integers from the set $\{1, 2, \ldots, m\}$ and J_m the set of all such I, whereby we also include the empty sequence $\varnothing \in J_m$. If I contains k_j occurrences of j for each $j = 1, \ldots, m$ then we define

$$\hat{\sigma}(I) = \prod_{j=1}^{m} k_j!$$

and set $\hat{\sigma}(\varnothing) = 1$. Now let $\boldsymbol{\delta}^{(1)}, \ldots, \boldsymbol{\delta}^{(m)} \in \mathbb{R}^d$ and define for $I = (i_1, \ldots, i_l) \in J_m$ the quantity $|I| = l$ as well as

$$\boldsymbol{\delta}^I = (\boldsymbol{\delta}^{(i_1)}, \ldots, \boldsymbol{\delta}^{(i_l)}) \in (\mathbb{R}^d)^l,$$

and we set $\boldsymbol{\delta}^{\varnothing} = \varnothing$ as well as $|\varnothing| = 0$. Next, for a map $\mathbf{f} \in \mathcal{C}^{p+1}(\mathbb{R}^d, \mathbb{R}^d)$ we define $\mathbf{f}^{(0)}(\mathbf{y})\varnothing = \mathbf{f}(\mathbf{y})$ and

$$\mathbf{f}^{(l)}(\mathbf{y})\boldsymbol{\delta}^I = \sum_{j_1, \ldots, j_l = 1}^{d} \partial_{j_1 \ldots j_l} \mathbf{f}(\mathbf{y}) \delta_{j_1}^{(i_1)} \cdots \delta_{j_l}^{(i_l)}, \quad 1 \le l \le p+1,$$

which allows us to formulate [But16, Theorem 308A], where we truncate the series using the Lagrangian remainder.

Theorem B.1. Let $p \in \mathbb{N}$ and $f \in \mathcal{C}^{p+1}(\mathbb{R}^d, \mathbb{R})$ as well as $\mathbf{y}, \boldsymbol{\delta}^{(1)}, \ldots, \boldsymbol{\delta}^{(m)} \in \mathbb{R}^d$. Then

$$f\left(\mathbf{y} + \sum_{i=1}^{m} \boldsymbol{\delta}^{(i)}\right) = \sum_{\substack{I \in J_m \\ |I| \le p}} \frac{1}{\hat{\sigma}(I)} f^{(|I|)}(\mathbf{y})\boldsymbol{\delta}^I + R_p\left(\mathbf{y} + \sum_{i=1}^{m} \boldsymbol{\delta}^{(i)}, \sum_{i=1}^{m} \boldsymbol{\delta}^{(i)}\right),$$

where, using the multi index notation, we have

$$R_p(\mathbf{x}, \mathbf{a}) = \sum_{|\boldsymbol{\alpha}| = p+1} \frac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\xi})}{\boldsymbol{\alpha}!} (\mathbf{x} - \mathbf{a})^{\boldsymbol{\alpha}}$$

with $\boldsymbol{\alpha} \in \mathbb{N}_0^d$ and ξ_j between x_j and a_j.

The key observation is that this equality holds true for any values of $\boldsymbol{\delta}^{(i)}$, that is also for solution-dependent vectors $\boldsymbol{\delta}^{(i)} = \boldsymbol{\delta}^{(i)}(\mathbf{y}^n, \Delta t)$.

Our aim is to apply Theorem B.1 to each addend of the right-hand side of the differential equations (2.5). Following the idea from [But16, Lemma 310B], we prove the following result.

Lemma B.2. Let $p \in \mathbb{N}$ and $\mathbf{f}^{[\nu]} \in \mathcal{C}^{p+1}$ for $\nu = 1 \ldots, N$. Then

$$\begin{aligned}\Delta t\mathbf{f}^{[\nu]}&\left(\mathbf{y}^n + \sum_{\tau\in NT_{p-1}} \theta(\tau, \mathbf{y}^n, \Delta t)\frac{\Delta t^{|\tau|}}{\sigma(\tau)}\mathcal{F}(\tau)(\mathbf{y}^n) + \mathcal{O}(\Delta t^p)\right)\\ &= \sum_{\tau\in NT_p} \widetilde{\theta}_\nu(\tau, \mathbf{y}^n, \Delta t)\frac{\Delta t^{|\tau|}}{\sigma(\tau)}\mathcal{F}(\tau)(\mathbf{y}^n) + \mathcal{O}(\Delta t^{p+1}),\end{aligned}$$

where

$$\widetilde{\theta}_\nu(\tau, \mathbf{y}^n, \Delta t) = \begin{cases}\delta_{\nu\mu}, & \tau = \bullet^{[\mu]},\\ \delta_{\nu\mu}\prod_{i=1}^{l}\theta(\tau_i, \mathbf{y}^n, \Delta t), & \tau = [\tau_1, \ldots, \tau_l]^{[\mu]}\end{cases} \tag{B.1}$$

and $\delta_{\nu\mu}$ denotes the Kronecker delta.

Proof. Let $NT_{p-1} = \{\tau^{(i)} \mid i = 1, \ldots, |NT_{p-1}|\}$. We want to apply Theorem B.1 to each component of $\mathbf{f}^{[\nu]}$ by first writing

$$\begin{aligned}\Delta t\mathbf{f}^{[\nu]}&\left(\mathbf{y}^n + \sum_{\tau\in NT_{p-1}} \theta(\tau, \mathbf{y}^n, \Delta t)\frac{\Delta t^{|\tau|}}{\sigma(\tau)}\mathcal{F}(\tau)(\mathbf{y}^n) + \mathcal{O}(\Delta t^p)\right)\\ &= \Delta t\mathbf{f}^{[\nu]}\left(\mathbf{y}^n + \sum_{j=1}^{|NT_{p-1}|}\boldsymbol{\delta}^{(j)} + \mathcal{O}(\Delta t^p)\right)\\ &= \Delta t\mathbf{f}^{[\nu]}\left(\mathbf{y}^n + \sum_{j=1}^{|NT_{p-1}|}\boldsymbol{\delta}^{(j)}\right) + \mathcal{O}(\Delta t^{p+1})\end{aligned}$$

with

$$\boldsymbol{\delta}^{(j)} = \theta(\tau^{(j)}, \mathbf{y}^n, \Delta t)\frac{\Delta t^{|\tau^{(j)}|}}{\sigma(\tau^{(j)})}\mathcal{F}(\tau^{(j)})(\mathbf{y}^n). \tag{B.2}$$

To that end, we first introduce for $I = (i_1, \ldots, i_r) \in J_m$ with $m = |NT_{p-1}|$ the quantity $\Sigma_I = \sum_{j=1}^{r}|\tau^{(i_j)}|$ and set $\Sigma_\varnothing = 0$. With that, Theorem B.1 and (B.2) yield

$$\begin{aligned}\Delta t\mathbf{f}^{[\nu]}&\left(\mathbf{y}^n + \sum_{\tau\in NT_{p-1}} \theta(\tau, \mathbf{y}^n, \Delta t)\frac{\Delta t^{|\tau|}}{\sigma(\tau)}\mathcal{F}(\tau)(\mathbf{y}^n) + \mathcal{O}(\Delta t^p)\right)\\ &= \sum_{\substack{I\in J_m\\ \Sigma_I\leq p-1}} \frac{\Delta t}{\hat{\sigma}(I)}(\mathbf{f}^{[\nu]})^{(|I|)}(\mathbf{y}^n)\boldsymbol{\delta}^I + \mathcal{O}(\Delta t^{p+1}).\end{aligned}$$

To prove the claim, we show that

$$\sum_{\substack{I\in J_m\\ \Sigma_I\leq p-1}} \frac{\Delta t}{\hat{\sigma}(I)}(\mathbf{f}^{[\nu]})^{(|I|)}(\mathbf{y}^n)\boldsymbol{\delta}^I = \sum_{\tau\in NT_p} \widetilde{\theta}_\nu(\tau, \mathbf{y}^n, \Delta t)\frac{\Delta t^{|\tau|}}{\sigma(\tau)}\mathcal{F}(\tau)(\mathbf{y}^n)$$

by means of an induction.

If $p = 1$, we find $\frac{\Delta t}{\hat{\sigma}(\varnothing)}(\mathbf{f}^{[\nu]})^{(0)}(\mathbf{y}^n)\varnothing = \Delta t\mathbf{f}^{[\nu]}(\mathbf{y}^n)$ and

$$\sum_{\mu=1}^{N} \widetilde{\theta}_\nu(\bullet^{[\mu]}, \mathbf{y}^n, \Delta t)\frac{\Delta t^{|\tau|}}{\sigma(\bullet^{[\mu]})}\mathcal{F}(\bullet^{[\mu]})(\mathbf{y}^n) = \sum_{\mu=1}^{N} \delta_{\nu\mu}\Delta t\mathbf{f}^{[\mu]}(\mathbf{y}^n) = \Delta t\mathbf{f}^{[\nu]}(\mathbf{y}^n),$$

so that

$$\begin{aligned}\sum_{\substack{I\in J_m\\ \Sigma_I\leq 0}} \frac{\Delta t}{\hat{\sigma}(I)}(\mathbf{f}^{[\nu]})^{(|I|)}(\mathbf{y}^n)\boldsymbol{\delta}^I &= \frac{\Delta t}{\hat{\sigma}(\varnothing)}(\mathbf{f}^{[\nu]})^{(0)}(\mathbf{y}^n)\varnothing = \Delta t\mathbf{f}^{[\nu]}(\mathbf{y}^n)\\ &= \sum_{\mu=1}^{N} \widetilde{\theta}_\nu(\bullet^{[\mu]}, \mathbf{y}^n, \Delta t)\frac{\Delta t^{|\tau|}}{\sigma(\bullet^{[\mu]})}\mathcal{F}(\bullet^{[\mu]})(\mathbf{y}^n)\\ &= \sum_{\tau\in NT_1} \widetilde{\theta}(\tau, \mathbf{y}^n, \Delta t)\frac{\Delta t^{|\tau|}}{\sigma(\tau)}\mathcal{F}(\tau)(\mathbf{y}^n)\end{aligned}$$

is true. By induction we can now assume that

$$\sum_{\substack{I\in J_m\\ \Sigma_I\leq p-2}} \frac{\Delta t}{\hat{\sigma}(I)}(\mathbf{f}^{[\nu]})^{(|I|)}(\mathbf{y}^n)\boldsymbol{\delta}^I = \sum_{\tau\in NT_{p-1}} \widetilde{\theta}_\nu(\tau, \mathbf{y}^n, \Delta t)\frac{\Delta t^{|\tau|}}{\sigma(\tau)}\mathcal{F}(\tau)(\mathbf{y}^n)$$

holds true for some $p \geq 2$, so that it remains to show

$$\sum_{\substack{I\in J_m\\ \Sigma_I= p-1}} \frac{\Delta t}{\hat{\sigma}(I)}(\mathbf{f}^{[\nu]})^{(|I|)}(\mathbf{y}^n)\boldsymbol{\delta}^I = \sum_{\tau\in NT_p\setminus NT_{p-1}} \widetilde{\theta}_\nu(\tau, \mathbf{y}^n, \Delta t)\frac{\Delta t^{|\tau|}}{\sigma(\tau)}\mathcal{F}(\tau)(\mathbf{y}^n) \tag{B.3}$$

to finish the proof by induction. For this, let us consider an arbitrary element $\tau \in NT_p \setminus NT_{p-1}$, which can be written as

$$\tau = [\tau_1, \dots, \tau_l]^{[\mu]} = [(\tau^{(i_1)})^{m_1}, \dots, (\tau^{(i_j)})^{m_j}]^{[\mu]}, \quad \sum_{r=1}^{j}|\tau^{(i_r)}|m_r = p-1^{1},$$

where we point out that $\tau^{(i_r)} \in NT_{p-1}$ for $r = 1, \dots, j$. Without loss of generality, we can assume that $i_1 < i_2 < \dots < i_j$. For each such τ we can define the uniquely determined and non-decreasing sequence

$$\hat{I} = (\underbrace{i_1, \dots, i_1}_{m_1\text{times}}, \underbrace{i_2, \dots, i_2}_{m_2\text{times}}, \dots, \underbrace{i_j, \dots, i_j}_{m_j\text{times}})$$

satisfying $\hat{I} \in J_m$ and $\Sigma_{\hat{I}} = \sum_{r=1}^{l}|\tau_r| = \sum_{r=1}^{j}|\tau^{(i_r)}|m_r = p - 1$, so that equation (B.3) follows by proving

$$\frac{\Delta t}{\hat{\sigma}(\hat{I})}(\mathbf{f}^{[\nu]})^{(|\hat{I}|)}(\mathbf{y}^n)\boldsymbol{\delta}^{\hat{I}} = \sum_{\mu=1}^{N} \frac{\Delta t^p\widetilde{\theta}_\nu([\tau_1, \dots, \tau_l]^{[\mu]}, \mathbf{y}^n, \Delta t)}{\sigma([(\tau^{(i_1)})^{m_1}, \dots, (\tau^{(i_j)})^{m_j}]^{[\mu]})}\mathcal{F}([\tau_1, \dots, \tau_l]^{[\mu]})(\mathbf{y}^n),$$

since then any addend on the left-hand side of (B.3) is uniquely associated with

[1] The root node is not counted here.

the sum over the different root colors of a tree $\tau \in NT_p \setminus NT_{p-1}$. Using (B.2) and the definitions of σ, $\mathcal{F}$ and $\widetilde{\theta}_\nu$ from (2.11), (2.12) and (B.1), we indeed find

$$\begin{aligned}
&\frac{\Delta t}{\hat\sigma(\hat I)}(\mathbf{f}^{[\nu]})^{(|\hat I|)}(\mathbf{y}^n)\boldsymbol{\delta}^{\hat I}\\
&= \frac{\Delta t \prod_{r=1}^j \left(\frac{(\theta(\tau^{(i_r)},\mathbf{y}^n,\Delta t))^{m_r}\Delta t^{m_r|\tau^{(i_r)}|}}{(\sigma(\tau^{(i_r)}))^{m_r}}\right)}{\prod_{r=1}^j m_r!}(\mathbf{f}^{[\nu]})^{(l)}(\mathbf{y}^n)(\mathcal{F}(\tau_1)(\mathbf{y}^n),\dots,\mathcal{F}(\tau_l)(\mathbf{y}^n))\\
&= \frac{\Delta t^p \widetilde{\theta}_\nu([(\tau^{(i_1)})^{m_1},\dots,(\tau^{(i_j)})^{m_j}]^{[\nu]},\mathbf{y}^n,\Delta t)}{\sigma([(\tau^{(i_1)})^{m_1},\dots,(\tau^{(i_j)})^{m_j}]^{[\nu]})}\mathcal{F}([\tau_1,\dots,\tau_l]^{[\nu]})(\mathbf{y}^n)\\
&= \sum_{\mu=1}^N \frac{\Delta t^p \widetilde{\theta}_\nu([\tau_1,\dots,\tau_l]^{[\mu]},\mathbf{y}^n,\Delta t)}{\sigma([(\tau^{(i_1)})^{m_1},\dots,(\tau^{(i_j)})^{m_j}]^{[\mu]})}\mathcal{F}([\tau_1,\dots,\tau_l]^{[\mu]})(\mathbf{y}^n)
\end{aligned}$$

finishing the proof. □

With Lemma B.2 we can prove the following result, which is the analogue to Lemma 313A in [But16].

Lemma B.3. Define d_i and $g_i^{[\nu]}$ for $i = 1,\dots,s$ and $\nu = 1,\dots,N$ as in (4.1). Furthermore, let $p \in \mathbb{N}$ and $\mathbf{f}^{[\nu]} \in \mathcal{C}^{p+1}$ for $\nu = 1,\dots,N$. If

$$\mathbf{y}^{(i)} = \mathbf{y}^n + \sum_{\tau \in NT_{p-1}} \frac{\Delta t^{|\tau|}}{\sigma(\tau)} d_i(\tau,\mathbf{y}^n,\Delta t)\mathcal{F}(\tau)(\mathbf{y}^n) + \mathcal{O}(\Delta t^p)$$

then

$$\Delta t \mathbf{f}^{[\nu]}(\mathbf{y}^{(i)}) = \sum_{\tau \in NT_p} \frac{\Delta t^{|\tau|}}{\sigma(\tau)} g_i^{[\nu]}(\tau,\mathbf{y}^n,\Delta t)\mathcal{F}(\tau)(\mathbf{y}^n) + \mathcal{O}(\Delta t^{p+1}).$$

Proof. The claim follows using Lemma B.2 with $\theta(\tau,\mathbf{y}^n,\Delta t) = d_i(\tau,\mathbf{y}^n,\Delta t)$, which gives us

$$\begin{aligned}
\widetilde{\theta}_\nu(\tau,\mathbf{y}^n,\Delta t) &= \begin{cases} \delta_{\nu\mu}, & \tau = \bullet^{[\mu]},\\ \delta_{\nu\mu}\prod_{i=1}^l d_i(\tau_i,\mathbf{y}^n,\Delta t), & \tau = [\tau_1,\dots,\tau_l]^{[\mu]} \end{cases}\\
&= g_i^{[\nu]}(\tau,\mathbf{y}^n,\Delta t).
\end{aligned}$$

□

B.1 Results for Reducing Order Conditions of NSARK methods

As final intermediate results, we prove the following lemmas which are helpful to reduce the conditions for 3rd and 4th order MPRK and GeCo methods. Both families of schemes can be written in the form of an NSARK method with

$$a_{ij}^{[\nu]} = a_{ij}\gamma_\nu^{(i)}, \quad b_i^{[\mu]} = b_i\delta_\mu$$

for suitable solution-dependent functions δ_μ and $\gamma_\nu^{(i)}$, which we previously referred to as NS weights. In the following we use these general functions to reduce the order conditions (4.8) and (4.9) for 3rd and 4th order, respectively. As we assume for Theorem 4.1 that $a_{ij}^{[\nu]} = \mathcal{O}(1)$ as $\Delta t \to 0$, it suffices to prove the following results.

Lemma B.4. Let $\mathbf{A}, \mathbf{b}, \mathbf{c}$ be the coefficients of an explicit 3-stage RK scheme of order 3, and let $\gamma_\nu^{(i)} = \mathcal{O}(1)$ as $\Delta t \to 0$. Then the conditions

$$\delta_\mu = 1 + \mathcal{O}(\Delta t^3), \qquad \mu = 1, \ldots, N, \tag{B.4a}$$

$$\sum_{i=2}^{3} b_i c_i \gamma_\nu^{(i)} = \frac{1}{2} + \mathcal{O}(\Delta t^2), \qquad \nu = 1, \ldots, N, \tag{B.4b}$$

$$\sum_{i=2}^{3} b_i c_i^2 \gamma_\nu^{(i)} \gamma_\xi^{(i)} = \frac{1}{3} + \mathcal{O}(\Delta t), \qquad \nu, \xi = 1, \ldots, N, \tag{B.4c}$$

$$\sum_{i,j=2}^{3} b_i a_{ij} c_j \gamma_\nu^{(i)} \gamma_\xi^{(j)} = \frac{1}{6} + \mathcal{O}(\Delta t), \qquad \nu, \xi = 1, \ldots, N \tag{B.4d}$$

and

$$\begin{aligned} \delta_\mu &= 1 + \mathcal{O}(\Delta t^3), \quad \mu = 1, \ldots, N, \\ \sum_{i=2}^{3} b_i c_i \gamma_\nu^{(i)} &= \frac{1}{2} + \mathcal{O}(\Delta t^2), \quad \nu = 1, \ldots, N, \\ \gamma_\nu^{(i)} &= 1 + \mathcal{O}(\Delta t), \quad \nu = 1, \ldots, N, \quad i = 2, 3 \end{aligned} \tag{B.5}$$

are equivalent for any solution and step-size dependent values of δ_μ and $\gamma_\nu^{(i)}$ for $i = 2, 3$ and $\mu, \nu = 1, \ldots, N$.

Proof. It is easy to see that the conditions (B.4) are fulfilled by any solution of (B.5). To see that any solution of (B.4) must satisfy (B.5), consider the conditions from (B.4) as $\Delta t \to 0$. From $a_{ij}^{[\nu]} = \mathcal{O}(1)$, any accumulation point of $\gamma_\nu^{(i)}$ is neither ∞ nor $-\infty$. In the following, we denote by $\Gamma_\nu^{(i)}$ an arbitrary accumulation point of $\gamma_\nu^{(i)}$ as $\Delta t \to 0$. Moreover, since the underlying RK scheme is explicit with three stages, the only addend remaining on the left-hand side of (B.4d) is $b_3 a_{32} c_2 \gamma_\nu^{(3)} \gamma_\xi^{(2)} = \frac{1}{6} \gamma_\nu^{(3)} \gamma_\xi^{(2)}$. Hence, for any accumulation point $\Gamma_\nu^{(i)}$, the conditions (B.4b), (B.4c) with $\nu = \xi$, and (B.4d) together with $c_1 = 0$ imply

$$\begin{aligned} \sum_{i=2}^{3} b_i c_i \Gamma_\nu^{(i)} &= \frac{1}{2}, \qquad \nu = 1, \ldots, N, \\ \sum_{i=2}^{3} b_i c_i^2 (\Gamma_\nu^{(i)})^2 &= \frac{1}{3}, \qquad \nu = 1, \ldots, N, \\ \Gamma_\nu^{(3)} \Gamma_\xi^{(2)} &= 1, \quad \nu, \xi = 1, \ldots, N. \end{aligned}$$

This system of equations possesses for any pair (ν, ξ) the unique solution $\Gamma_\xi^{(2)} = 1$ and $\Gamma_\nu^{(3)} = 1$ for all $\xi, \nu = 1, \ldots, N$, see [KM18b, Lemma 7]. Finally, (B.4c) with $\nu = \xi$ thus implies that $\gamma_\nu^{(i)} = 1 + \mathcal{O}(\Delta t)$ proving that (B.4) and (B.5) are equivalent. □

To come up with an analogue for 4-stage RK methods of 4th order, we can follow the same ideas as in the last proof, however, this time we need to come up with a substitute for [KM18b, Lemma 7]. The precise procedure is based on Gröbner bases computations as we will see in the proof of the following lemma.

Lemma B.5. Let $\mathbf{A}, \mathbf{b}, \mathbf{c}$ be the coefficients of an explicit 4-stage RK scheme of order 4, and let $\gamma_\nu^{(i)} = \mathcal{O}(1)$ as $\Delta t \to 0$. Then the conditions

$$
\begin{aligned}
\delta_\mu &= 1 + \mathcal{O}(\Delta t^4), && \mu = 1, \ldots, N, && \text{(B.6a)}\\
\sum_{i=2}^{4} b_i c_i \gamma_\nu^{(i)} &= \frac{1}{2} + \mathcal{O}(\Delta t^3), && \nu = 1, \ldots, N, && \text{(B.6b)}\\
\sum_{i=2}^{4} b_i c_i^2 \gamma_\nu^{(i)} \gamma_\xi^{(i)} &= \frac{1}{3} + \mathcal{O}(\Delta t^2), && \nu, \xi = 1, \ldots, N, && \text{(B.6c)}\\
\sum_{i,j=2}^{4} b_i a_{ij} c_j \gamma_\nu^{(i)} \gamma_\xi^{(j)} &= \frac{1}{6} + \mathcal{O}(\Delta t^2), && \nu, \xi = 1, \ldots, N, && \text{(B.6d)}\\
\sum_{i,j=2}^{4} b_i c_i a_{ij} c_j \gamma_\nu^{(i)} \gamma_\xi^{(i)} \gamma_\eta^{(j)} &= \frac{1}{8} + \mathcal{O}(\Delta t), && \nu, \xi, \eta = 1, \ldots, N, && \text{(B.6e)}\\
\sum_{i=2}^{4} b_i c_i^3 \gamma_\nu^{(i)} \gamma_\xi^{(i)} \gamma_\eta^{(i)} &= \frac{1}{4} + \mathcal{O}(\Delta t), && \nu, \xi, \eta = 1, \ldots, N, && \text{(B.6f)}\\
\sum_{i,j,k=2}^{4} b_i a_{ij} a_{jk} c_k \gamma_\nu^{(i)} \gamma_\xi^{(j)} \gamma_\eta^{(k)} &= \frac{1}{4!} + \mathcal{O}(\Delta t), && \nu, \xi, \eta = 1, \ldots, N, && \text{(B.6g)}\\
\sum_{i,j=2}^{4} b_i a_{ij} c_j^2 \gamma_\nu^{(i)} \gamma_\xi^{(j)} \gamma_\eta^{(j)} &= \frac{1}{12} + \mathcal{O}(\Delta t), && \nu, \xi, \eta = 1, \ldots, N && \text{(B.6h)}
\end{aligned}
$$

and

$$
\begin{aligned}
\delta_\mu &= 1 + \mathcal{O}(\Delta t^4), \quad \mu = 1, \ldots, N,\\
\sum_{i=2}^{4} b_i c_i \gamma_\nu^{(i)} &= \frac{1}{2} + \mathcal{O}(\Delta t^3), \quad \nu = 1, \ldots, N, \qquad \text{(B.7)}\\
\gamma_\nu^{(i)} &= 1 + \mathcal{O}(\Delta t^2), \quad \nu = 1, \ldots, N, \quad i = 2, 3, 4
\end{aligned}
$$

are equivalent for any solution and step-size dependent values of δ_μ and $\gamma_\nu^{(i)}$ for $i = 2, 3, 4$ and $\mu, \nu = 1, \ldots, N$.

Proof. We first note that the conditions (B.6) are fulfilled by any solution of (B.7). To see that any solution of (B.6) must satisfy (B.7), consider the conditions from (B.6) as $\Delta t \to 0$. From $a_{ij}^{[\nu]} = \mathcal{O}(1)$, any accumulation point of $\gamma_\nu^{(i)}$ is neither ∞ nor $-\infty$. In the following, we denote by $\Gamma_\nu^{(i)}$ an arbitrary accumulation point of $\gamma_\nu^{(i)}$ as $\Delta t \to 0$. Moreover, since the underlying RK scheme is explicit with four stages, the only addend remaining on the left-hand side of (B.6g) is

$$
b_4 a_{43} c_3 \gamma_\nu^{(4)} \gamma_\xi^{(3)} \gamma_\eta^{(2)} = \frac{1}{4!} \gamma_\nu^{(4)} \gamma_\xi^{(3)} \gamma_\eta^{(2)}.
$$

Hence, for any accumulation point $\Gamma_\nu^{(i)}$, the conditions (B.6) together with the order conditions for the underlying RK method and $\nu = \xi = \eta$ imply

$$
\sum_{i=2}^{4} b_i c_i \left(\Gamma_\nu^{(i)} - 1 \right) = 0, \qquad \nu = 1, \ldots, N,
$$

$$\begin{aligned}
\sum_{i=2}^{4} b_i c_i^2 \left((\Gamma_\nu^{(i)})^2 - 1\right) &= 0, && \nu = 1,\dots,N,\\
\sum_{i,j=2}^{4} b_i a_{ij} c_j \left(\Gamma_\nu^{(i)}\Gamma_\nu^{(j)} - 1\right) &= 0, && \nu = 1,\dots,N,\\
\sum_{i,j=2}^{4} b_i c_i a_{ij} c_j \left((\Gamma_\nu^{(i)})^2\Gamma_\nu^{(j)} - 1\right) &= 0, && \nu = 1,\dots,N, \qquad \text{(B.8)}\\
\sum_{i=2}^{4} b_i c_i^3 \left((\Gamma_\nu^{(i)})^3 - 1\right) &= 0, && \nu = 1,\dots,N,\\
\Gamma_\nu^{(4)}\Gamma_\nu^{(3)}\Gamma_\nu^{(2)} - 1 &= 0, && \nu = 1,\dots,N,\\
\sum_{i,j=2}^{4} b_i a_{ij} c_j^2 \left(\Gamma_\nu^{(i)}(\Gamma_\nu^{(j)})^2 - 1\right) &= 0, && \nu = 1,\dots,N.
\end{aligned}$$

In what follows we fix $\nu \in \{1,\dots,N\}$. Then, we compute a reduced Gröbner basis[2] of the corresponding polynomial ideal generated by the polynomials on the left-hand sides of (B.8) in the ring $\mathbb{R}[\Gamma_\nu^{(2)},\Gamma_\nu^{(3)},\Gamma_\nu^{(4)}]$, yielding $\{\Gamma_\nu^{(2)}-1,\Gamma_\nu^{(3)}-1,\Gamma_\nu^{(4)}-1\}$. Hence, $\Gamma_\nu^{(i)} = 1$ for $\nu = 1,\dots,N$ and $i = 2,3,4$ is the unique solution to (B.8). As a result, (B.6f) with $\nu = \xi = \eta$ implies that $\gamma_\nu^{(i)} = 1 + \mathcal{O}(\Delta t)$. This already allows us to neglect the conditions (B.6e) to (B.6h) in the following as they are now fulfilled by $\gamma_\nu^{(i)} = 1 + \mathcal{O}(\Delta t)$. Substituting the ansatz[3] $\gamma_\nu^{(i)} = 1 + x_\nu^{(i)}\Delta t + \mathcal{O}(\Delta t^2)$ into the remaining conditions (B.6b) to (B.6d), the resulting coefficients of Δt must vanish, that is

$$\begin{aligned}
\sum_{i=2}^{4} b_i c_i x_\nu^{(i)} &= 0,\\
\sum_{i=2}^{4} b_i c_i^2 2 x_\nu^{(i)} &= 0, \qquad \text{(B.9)}\\
\sum_{i,j=2}^{4} b_i a_{ij} c_j (x_\nu^{(i)} + x_\nu^{(j)}) &= 0.
\end{aligned}$$

We again compute a reduced Gröbner basis of the ideal generated by the left-hand side polynomials from (B.9) in the polynomial ring $\mathbb{R}[x_\nu^{(2)},x_\nu^{(3)},x_\nu^{(4)}]$. The resulting Gröbner basis reads $\{x_\nu^{(2)},x_\nu^{(3)},x_\nu^{(4)}\}$ proving that the unique solution to the above polynomial system is given by $x_\nu^{(i)} = 0$ for $\nu = 1,\dots,N$ and $i = 2,3,4$. With that we have demonstrated that $\gamma_\nu^{(i)} = 1 + \mathcal{O}(\Delta t^2)$ which finishes the proof. □

[2] We refer to our Maple repository [IKM23a] for the computation of the Gröbner bases for this work.

[3] Formally, $x_\nu^{(i)}$ is an arbitrary accumulation point of $\frac{\gamma_\nu^{(i)}-1}{\Delta t}$ as $\Delta t \to 0$, which due to $\gamma_\nu^{(i)} = 1 + \mathcal{O}(\Delta t)$ cannot be $\pm\infty$. However, for the sake of simplicity, we refrain to introduce several $\gamma_\nu^{(i)}$ for every occurring accumulation point.

Bibliography

[AE08] H. Amann and J. Escher. *Analysis. II.* Birkhäuser Verlag, Basel, 2008. Translated from the 1999 German original by Silvio Levy and Matthew Cargo.

[AGKM21] 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:Paper No. 113350, 13, 2021.

[AKM20] A. I. Ávila, S. Kopecz, and A. Meister. A comprehensive theory on generalized BBKS schemes. *Appl. Numer. Math.*, 157:19–37, 2020.

[ALMÖT22] R. Abgrall, É. Le Mélédo, P. Öffner, and D. Torlo. Relaxation deferred correction methods and their applications to residual distribution schemes. *SMAI J. Comput. Math.*, 8:125–160, 2022.

[AMSS97] A. L. Araújo, A. Murua, and J. M. Sanz-Serna. Symplectic methods based on decompositions. *SIAM J. Numer. Anal.*, 34(5):1926–1947, 1997.

[ARS97] U. M. Ascher, S. J. Ruuth, and R. J. Spiteri. Implicit-explicit runge-kutta methods for time-dependent partial differential equations. *Applied Numerical Mathematics*, 25(2):151–167, 1997. Special Issue on Time Integration.

[BBK$^+$06] 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. *Journal of Marine Systems*, 61(3–4):180–211, 2006. Workshop on Future Directions in Modelling Physical-Biological Interactions (WKFDPBI)Workshop on Future Directions in Modelling Physical-Biological Interactions (WKFDPBI).

[BBKS07] J. Bruggeman, H. Burchard, B. W. Kooi, and B. Sommeijer. A second-order, unconditionally positive, mass-conserving integration scheme for biochemical systems. *Appl. Numer. Math.*, 57(1):36–58, 2007.

[BC78] C. Bolley and M. Crouzeix. Conservation de la positivité lors de la discrétisation des problèmes d'évolution paraboliques. *RAIRO Anal. Numér.*, 12(3):237–245, iv, 1978.

[BDM03] 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(1):1–30, 2003.

[BDM05] H. Burchard, E. Deleersnijder, and A. Meister. Application of modified Patankar schemes to stiff biogeochemical models for the water column. *Ocean Dynamics*, 55(3):326–337, 2005.

[Ber96] E. Bertolazzi. Positive and conservative schemes for mass action kinetics. *Comput. Math. Appl.*, 32(6):29–43, 1996.

[BF04] L. Benvenuti and L. Farina. Eigenvalue regions for positive systems. *Systems & Control Letters*, 51(3-4):325–330, 2004.

[BIM21] S. Blanes, A. Iserles, and S. Macnamara. Positivity–preserving methods for population models, 2021.

[BIM22] S. Blanes, A. Iserles, and S. Macnamara. Positivity-preserving methods for ordinary differential equations. *ESAIM Math. Model. Numer. Anal.*, 56(6):1843–1870, 2022.

[BMZ07] J. Benz, A. Meister, and P. Andrea 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*, volume 7, pages 2040045–2040046. Wiley Online Library, 2007.

[BMZ09] J. Benz, A. Meister, and P. A. Zardo. A conservative, positivity preserving scheme for advection-diffusion-reaction equations in biochemical applications. In E. Tadmor, J.-G. Liu, and A. Tzavaras, editors, *Hyperbolic Problems: Theory, Numerics and Applications*, volume 67.2 of *Proceedings of Symposia in Applied Mathematics*, pages 399–408. American Mathematical Society, Providence, Rhode Island, 2009.

[BRBM08] N. Broekhuizen, G. J. Rickard, J. Bruggeman, and A. Meister. An improved and generalized second order, unconditionally positive, mass conserving integration scheme for biochemical systems. *Appl. Numer. Math.*, 58(3):319–340, 2008.

[But16] J. C. Butcher. *Numerical methods for ordinary differential equations.* John Wiley & Sons, Ltd., Chichester, third edition, 2016. With a foreword by J. M. Sanz-Serna.

[Car81] J. Carr. *Applications of centre manifold theory*, volume 35 of *Applied Mathematical Sciences.* Springer-Verlag, New York, 1981.

[CD16] G. Colonna and A. D'Angola, editors. *Plasma Modeling.* 2053-2563. IOP Publishing, 2016.

[CMÖT22] M. Ciallella, L. Micalizzi, P. Öffner, and D. Torlo. An arbitrary high order and positivity preserving method for the shallow water equations. *Comput. & Fluids*, 247:Paper No. 105630, 21, 2022.

[Coh03] P. M. Cohn. *Basic algebra.* Springer-Verlag London, Ltd., London, 2003. Groups, rings and fields.

[Cro80] M. Crouzeix. Une méthode multipas implicite-explicite pour l'approximation des équations d'évolution paraboliques. *Numer. Math.*, 35(3):257–276, 1980.

[Cry73] C. W. Cryer. A new class of highly-stable methods: a_0-stable methods. *BIT Numerical Mathematics*, 13(2):153–159, 1973.

[Dah63] G. G. Dahlquist. A special stability problem for linear multistep methods. *Nordisk Tidskr. Informationsbehandling (BIT)*, 3:27–43, 1963.

[DB02] P. Deuflhard and F. Bornemann. *Scientific computing with ordinary differential equations*, volume 42 of *Texts in Applied Mathematics*. Springer-Verlag, New York, 2002. Translated from the 1994 German original by Werner C. Rheinboldt.

[DGR00] A. Dutt, L. Greengard, and V. Rokhlin. Spectral deferred correction methods for ordinary differential equations. *BIT*, 40(2):241–266, 2000.

[DK06] D. T. Dimitrov and H. V. Kojouharov. Positive and elementary stable nonstandard numerical methods with applications to predator–prey models. *Journal of Computational and Applied Mathematics*, 189(1–2):98–108, 2006. Proceedings of The 11th International Congress on Computational and Applied MathematicsThe 11th International Congress on Computational and Applied Mathematics.

[FS11a] L. Formaggia and A. Scotti. Positivity and conservation properties of some integration schemes for mass action kinetics. *SIAM J. Numer. Anal.*, 49(3):1267–1288, 2011.

[FS11b] L. Formaggia and A. Scotti. Positivity and conservation properties of some integration schemes for mass action kinetics. *SIAM Journal on Numerical Analysis*, 49(3/4):1267–1288, 2011.

[GLS88] K. Gustafsson, M. Lundh, and G. Söderlind. A PI stepsize control for the numerical solution of ordinary differential equations. *BIT Numerical Mathematics*, 28(2):270–287, 1988.

[Gre17] O. Gressel. Toward realistic simulations of magneto-thermal winds from weakly-ionized protoplanetary disks. In *Journal of Physics: Conference Series*, volume 837, page 012008. IOP Publishing, 2017.

[Gus91] K. Gustafsson. Control theoretic techniques for stepsize selection in explicit Runge-Kutta methods. *ACM Trans. Math. Software*, 17(4):533–554, 1991.

[Gus94] K. Gustafsson. Control-theoretic techniques for stepsize selection in implicit Runge-Kutta methods. *ACM Trans. Math. Software*, 20(4):496–517, 1994.

[HB10a] I. Hense and A. Beckmann. The representation of cyanobacteria life cycle processes in aquatic ecosystem models. *Ecological Modelling*, 221(19):2330–2338, 2010.

[HB10b] I. Hense and H. Burchard. Modelling cyanobacteria in shallow coastal seas. *Ecological Modelling*, 221(2):238–244, 2010.

[HIK$^+$22] J. Huang, T. Izgin, S. Kopecz, A. Meister, and C.-W. Shu. Lyapunov Stability of third order SSPMPRK schemes (code). https://github.com/IzginThomas/LyapunovSSPMPRK.git, December 2022.

[HIK$^+$23] J. Huang, T. Izgin, S. Kopecz, A. Meister, and C.-W. Shu. On the stability of strong-stability-preserving modified Patankar–Runge–Kutta schemes. *ESAIM Math. Model. Numer. Anal.*, 57(2):1063–1086, 2023.

[HNW93] E. Hairer, S. P. Nørsett, and G. Wanner. *Solving ordinary differential equations. I*, volume 8 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, second edition, 1993. Nonstiff problems.

[HÖT21] M. Han Veiga, P. Öffner, and D. Torlo. DeC and ADER: similarities, differences and a unified framework. *J. Sci. Comput.*, 87(1):35, 2021. Id/No 2.

[HS19] J. Huang and C.-W. Shu. Positivity-preserving time discretizations for production-destruction equations with applications to non-equilibrium flows. *J. Sci. Comput.*, 78(3):1811–1839, 2019.

[HW74] E. Hairer and G. Wanner. On the Butcher group and general multi-value methods. *Computing (Arch. Elektron. Rechnen)*, 13(1):1–15, 1974.

[HW10] E. Hairer and G. Wanner. *Solving ordinary differential equations. II*, volume 14 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, Berlin, 2010. Stiff and differential-algebraic problems, Second revised edition, paperback.

[HZS19] 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(2):1015–1056, 2019.

[IKM21] T. Izgin, S. Kopecz, and A. Meister. Recent developments in the field of modified patankar-runge-kutta-methods. *PAMM*, 21(1):e202100027, 2021.

[IKM22a] T. Izgin, S. Kopecz, and A. Meister. On Lyapunov stability of positive and conservative time integrators and application to second order modified Patankar–Runge–Kutta schemes. *ESAIM Math. Model. Numer. Anal.*, 56(3):1053–1080, 2022.

[IKM22b] T. Izgin, S. Kopecz, and A. Meister. On the stability of unconditionally positive and linear invariants preserving time integration schemes. *SIAM J. Numer. Anal.*, 60(6):3029–3051, 2022.

[IKM23a] T. Izgin, D. I. Ketcheson, and A. Meister. Order conditions for NSARK methods (code). https://github.com/IzginThomas/NSARK, May 2023.

[IKM23b] T. Izgin, D. I. Ketcheson, and A. Meister. Order conditions for Runge–Kutta-like methods with solution-dependent coefficients. *https://arxiv.org/abs/2305.14297*, 2023.

[IKM23c] T. Izgin, S. Kopecz, and A. Meister. A stability analysis of modified Patankar–Runge–Kutta methods for a nonlinear production–destruction system. *PAMM*, 22(1):e202200083, 2023.

[IKMM23] T. Izgin, S. Kopecz, A. Martiradonna, and A. Meister. On the dynamics of first and second order geco and gbbks schemes. *Applied Numerical Mathematics*, 193:43–66, 2023.

[IKMS23] T. Izgin, S. Kopecz, A. Meister, and Amandine Schilling. On the non-global linear stability and spurious fixed points of MPRK schemes with negative RK parameters. *https://arxiv.org/abs/2305.14297*, 2023.

[IÖ23] T. Izgin and P. Öffner. A study of the local dynamics of modified Patankar DeC and higher order modified Patankar–RK methods. *ESAIM Math. Model. Numer. Anal.*, 57(4):2319–2348, 2023.

[Ioo79] G. Iooss. *Bifurcation of maps and applications*, volume 36 of *North-Holland Mathematics Studies*. North-Holland Publishing Co., Amsterdam-New York, 1979.

[IÖT22] T. Izgin, P. Öffner, and D. Torlo. A necessary condition for non oscillatory and positivity preserving time-integration schemes. *https://arxiv.org/abs/2211.08905*, 2022.

[Jac09] Z. Jackiewicz. *General linear methods for ordinary differential equations.* John Wiley & Sons, Inc., Hoboken, New Jersey, 2009.

[KLJK17] D. I. Ketcheson, L. Lóczi, A. Jangabylova, and Adil Kusmanov. Dense output for strong stability preserving Runge-Kutta methods. *J. Sci. Comput.*, 71(3):944–958, 2017.

[KM10] J. S. Klar and J. P. Mücket. A detailed view of filaments and sheets in the warm-hot intergalactic medium. *Astronomy & Astrophysics*, 522:A114, 2010.

[KM18a] S. Kopecz and A. Meister. On order conditions for modified Patankar-Runge-Kutta schemes. *Appl. Numer. Math.*, 123:159–179, 2018.

[KM18b] S. Kopecz and A. Meister. Unconditionally positive and conservative third order modified Patankar-Runge-Kutta discretizations of production-destruction systems. *BIT*, 58(3):691–728, 2018.

[KM19a] S. Kopecz and A. Meister. A comparison of numerical methods for conservative and positive advection-diffusion-production-destruction systems. *PAMM*, 19(1):e201900209, 2019.

[KM19b] S. Kopecz and A. Meister. On the existence of three-stage third-order modified Patankar-Runge-Kutta schemes. *Numer. Algorithms*, 81(4):1473–1484, 2019.

[KMP21] S. Kopecz, A. Meister, and Helmut Podhaisky. On adaptive patankar runge–kutta methods. *PAMM*, 21(1):e202100235, 2021.

[Koo00] S. A. L. M. Kooijman. *Dynamic Energy and Mass Budgets in Biological Systems.* Cambridge University Press, 2 edition, 2000.

[KV12] B. Korte and J. Vygen. *Combinatorial optimization*, volume 21 of *Algorithms and Combinatorics*. Springer, Heidelberg, fifth edition, 2012. Theory and algorithms.

[LD21] D. Lacitignola and F. Diele. Using awareness to Z-control a SEIR model with overexposure: Insights on Covid-19 pandemic. *Chaos, Solitons & Fractals*, 150:111063, 2021.

[LS14] L. H. Loomis and S. Sternberg. *Advanced calculus*. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2014.

[Lue79] D. G. Luenberger. *Introduction to Dynamic Systems: Theory, Models, and Applications*. Wiley, 1979.

[MB10] A. Meister and J. Benz. *Phosphorus Cycles in Lakes and Rivers: Modeling, Analysis, and Simulation*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[MCD20] A. Martiradonna, G. Colonna, and F. Diele. *GeCo*: Geometric Conservative nonstandard schemes for biochemical systems. *Appl. Numer. Math.*, 155:38–57, 2020.

[Mic21] R. E. Mickens. *Nonstandard finite difference schemes—methodology and applications*. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, [2021] ©2021. Expanded second edition of [1275372].

[MM76] J. E. Marsden and M. McCracken. *The Hopf bifurcation and its applications*, volume 19 of *Applied Mathematical Sciences, Vol. 19*. Springer-Verlag, New York, 1976. With contributions by P. Chernoff, G. Childs, S. Chow, J. R. Dorroh, J. Guckenheimer, L. Howard, N. Kopell, O. Lanford, J. Mallet-Paret, G. Oster, O. Ruiz, S. Schecter, D. Schmidt and S. Smale.

[MO14] A. Meister and S. Ortleb. On unconditionally positive implicit time integration for the DG scheme applied to shallow water flows. *International Journal for Numerical Methods in Fluids*, 76(2):69–94, 2014.

[NRK21a] S. Nüsslein, H. Ranocha, and D. I. Ketcheson. Positivity-preserving adaptive Runge-Kutta methods. *Commun. Appl. Math. Comput. Sci.*, 16(2):155–179, 2021.

[NRK21b] S. Nüsslein, H. Ranocha, and D. I. Ketcheson. Positivity-preserving adaptive Runge-Kutta methods. *Commun. Appl. Math. Comput. Sci.*, 16(2):155–179, 2021.

[OH17] S. Ortleb and W. Hundsdorfer. Patankar-type Runge-Kutta schemes for linear PDEs. In *AIP Conference Proceedings*, volume 1863, page 320008. AIP Publishing LLC, 2017.

[Osi12] G. Osipenko. Center manifolds. In *Mathematics of complexity and dynamical systems. Vols. 1–3*, pages 48–62. Springer, New York, 2012.

[ÖT20] P. Öffner and D. Torlo. Arbitrary high-order, conservative and positivity preserving Patankar-type deferred correction schemes. *Appl. Numer. Math.*, 153:15–34, 2020.

[Pat80] S. V. Patankar. *Numerical heat transfer and fluid flow.* Series in computational methods in mechanics and thermal sciences. Hemisphere Pub. Corp. New York, Washington, 1980.

[San01] A. Sandu. Positive numerical integration methods for chemical kinetic systems. *J. Comput. Phys.*, 170(2):589–602, 2001.

[San02] A. Sandu. Time-stepping methods that favor positivity for atmospheric chemistry modeling. In *Atmospheric modeling (Minneapolis, MN, 2000)*, volume 130 of *IMA Vol. Math. Appl.*, pages 21–37. Springer, New York, 2002.

[Sch23] Amandine Schilling. Eigenschaften modifizierter Patankar–Runge–Kutta-Verfahren mit negativen RK-Parametern, 2023. Universität Kassel, 2023, master thesis (written in German).

[SD17] K. Semeniuk and A. Dastoor. Development of a global ocean mercury model with a methylation cycle: outstanding issues. *Global Biogeochemical Cycles*, pages n/a–n/a, 2017. 2016GB005452.

[SG15] A. Sandu and M. Günther. A generalized-structure approach to additive Runge-Kutta methods. *SIAM J. Numer. Anal.*, 53(1):17–42, 2015.

[SH98] A. Stuart and A. R. Humphries. *Dynamical systems and numerical analysis*, volume 2. Cambridge University Press, Cambridge, 1998.

[Sha86] L. F. Shampine. Conservation laws and the numerical solution of ODEs. *Comput. Math. Appl. Part B*, 12(5-6):1287–1296, 1986.

[SM03] E. Süli and D. F. Mayers. *An Introduction to Numerical Analysis.* Cambridge University Press, 2003.

[SO88] C.-W. Shu and S. Osher. Efficient implementation of essentially non-oscillatory shock-capturing schemes. *Journal of Computational Physics*, 77(2):439–471, 1988.

[Söd02] G. Söderlind. Automatic control and adaptive time-stepping. *Numer. Algorithms*, 31(1-4):281–310, 2002. Numerical methods for ordinary differential equations (Auckland, 2001).

[Söd03] G. Söderlind. Digital filters in adaptive time-stepping. *ACM Transactions on Mathematical Software (TOMS)*, 29(1):1–26, 2003.

[Söd06] G. Söderlind. Time-step selection algorithms: Adaptivity, control, and signal processing. *Applied Numerical Mathematics*, 56(3-4):488–502, 2006.

[SS03] E. M. Stein and R. Shakarchi. *Complex analysis*, volume 2 of *Princeton Lectures in Analysis.* Princeton University Press, Princeton, NJ, 2003.

[STKB05] L. F. Shampine, S. Thompson, J. A. Kierzenka, and G. D. Byrne. Non-negative solutions of ODEs. *Appl. Math. Comput.*, 170(1):556–569, 2005.

[SVV18] A. J. Steyer and E. S. Van Vleck. A Lyapunov and Sacker–Sell spectral stability theory for one-step methods. *BIT Numerical Mathematics*, 58(3):749–781, 2018.

[SW06] G. Söderlind and L. Wang. Adaptive time-stepping and computational stability. *Journal of Computational and Applied Mathematics*, 185(2):225–243, 2006.

[TGA96] E. H. Twizell, A. B. Gumel, and M. A. Arigu. Second-order, L_0-stable methods for the heat equation with time-dependent boundary conditions. *Adv. Comput. Math.*, 6(3-4):333–352 (1997), 1996. John Crank 80th birthday special issue.

[Tit39] E. C. Titchmarsh. *The theory of functions*. Oxford University Press, Oxford, second edition, 1939.

[TÖR22] D. Torlo, P. Öffner, and H. Ranocha. Issues with positivity-preserving Patankar-type schemes. *Appl. Numer. Math.*, 182:117–147, 2022.

[Var00] R. S. Varga. *Matrix iterative analysis*, volume 27 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, expanded edition, 2000.

[WHK13] A. Warns, I. Hense, and A. Kremp. Modelling the life cycle of dinoflagellates: a case study with Biecheleria baltica. *J. Plankton. Res*, 35(2):379–392, 2013.

[WS22] S. Wei and R. J. Spiteri. Qualitative property preservation of high-order operator splitting for the sir model. *Appl. Numer. Math.*, 172:332–350, 2022.

[Zon64] J. A. Zonneveld. *Automatic numerical integration*, volume 8 of *Mathematical Centre Tracts*. Mathematisch Centrum, Amsterdam, 1964.