A Stability Analysis of a Semi-implicit Runge–Kutta Scheme for a Nonlinear System

A. Chertock and T. Izgin and P. Öffner

Abstract The semi-implicit Runge–Kutta (SI-RK) methods proposed in [2] have been successfully applied as time-integration methods in the context of the shallow water equations. The schemes preserve the sign of the analytical solution and are of higher order (> 1). As a result, they do not belong to the class of general linear methods [1], which implies that their stability analysis becomes more complex. In this short note, we investigate the stability properties of the second-order SI-RK method when applied to a nonlinear test problem. For this analysis, we take advantage of recently developed criteria capturing the Lyapunov stability properties of non-hyperbolic fixed points. Thereby, a stability function is derived and analyzed. In particular, we derive time step restrictions related to the SI-RK scheme's stability properties. Finally, we validate our theoretical findings with numerical tests.

1 Introduction

Numerical time-stepping methods applied to differential equations

$$\mathbf{y}'(t) = \mathbf{f}(\mathbf{y}(t)) \tag{1}$$

with an initial condition $\mathbf{y}(0) = \mathbf{y}^0$ aim to approximate the analytic solution if it exists. The goal is to carry out as many properties of the analytic solution at the discrete level as possible. For instance, steady-state solutions of (1) should correspond to fixed points of the numerical scheme. A numerical scheme with this property is called *steady state preserving*. Moreover, the analytic solution of (1) is called *positive*, if $\mathbf{y}(t) > \mathbf{0}$ holds for all $t > 0$ whenever $\mathbf{y}(0) > \mathbf{0}$. A numerical method discretely reproducing this property, i. e. $\mathbf{y}^n > \mathbf{0}$ for all $n \in \mathbb{N}$ and $\Delta t > 0$ whenever $\mathbf{y}^0 > \mathbf{0}$,

Alina Chertock
Department of Mathematics, North Carolina State University, Raleigh, NC 27695, USA e-mail: chertock@math.ncsu.edu

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

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

is called *unconditionally positive*. Finally, the stability properties of fixed points of a steady state preserving method applied to (1) should be identical to those of the corresponding steady state solutions. The corresponding notions of stability are given in the following definitions.

Definition 1 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 $\varepsilon > 0$, there exists a $\delta = \delta(\varepsilon) > 0$ such that $\|\mathbf{y}(0) - \mathbf{y}^*\| < \delta$ implies $\|\mathbf{y}(t) - \mathbf{y}^*\| < \varepsilon$ 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*.

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 $\varepsilon > 0$, there exists a $\delta = \delta(\varepsilon) > 0$ such that $\left\|\mathbf{y}^0 - \mathbf{y}^*\right\| < \delta$ implies $\|\mathbf{y}^n - \mathbf{y}^*\| < \varepsilon$ 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*.

We use stability instead of Lyapunov stability throughout the manuscript. For a compact notation, we also introduce a matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$ and vectors $\mathbf{n}_1, \dots, \mathbf{n}_k$ with $k \geq 1$ which form a basis of $\ker(\mathbf{A}^\mathrm{T})$ and define the matrix $\mathbf{N} \in \mathbb{R}^{k\times N}$ whose rows are $\mathbf{n}_1^\mathrm{T}, \dots, \mathbf{n}_k^\mathrm{T}$.

The following two theorems have already been used to investigate the stability properties of several schemes [4, 5, 9] and will be used in our analysis.

Theorem 1 ([11, 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* $\mathbf{D}\mathbf{g}(\mathbf{y}^*)$ *exists and denote its spectral radius by* $\varrho(\mathbf{D}\mathbf{g}(\mathbf{y}^*))$. *Then*

a) $\mathbf{y}^*$ *is asymptotically stable if* $\varrho(\mathbf{D}\mathbf{g}(\mathbf{y}^*)) < 1$.
b) $\mathbf{y}^*$ *is unstable if* $\varrho(\mathbf{D}\mathbf{g}(\mathbf{y}^*)) > 1$.

Theorem 2 ([7, Theorem 2.9]) *Let* $\mathbf{A} \in \mathbb{R}^{N\times N}$ *such that* $\ker(\mathbf{A}) = \operatorname{span}(\mathbf{v}_1, \dots, \mathbf{v}_k)$ *represents a k-dimensional subspace of* $\mathbb{R}^N$ *with* $k \geq 1$. *Also, let* $\mathbf{y}^* \in \ker(\mathbf{A})$ *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* $\ker(\mathbf{A}) \cap \mathcal{D}$ *be a fixed point of* $\mathbf{g}$ *and suppose that* $\mathbf{g}\big|_{\mathcal{D}} \in C^1$ *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* $i = 1, \dots, k$ *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 $H = \{\mathbf{y} \in \mathbb{R}^N \mid \mathbf{N}\mathbf{y} = \mathbf{N}\mathbf{y}^*\}$ *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* $\left\|\mathbf{y}^0 - \mathbf{y}^*\right\| < \delta$ *imply* $\mathbf{y}^n \to \mathbf{y}^*$ *as* $n \to \infty$.

2 Preliminaries

In this paper, we investigate the stability properties of the second-order semi-implicit Runge–Kutta (SI-RK2) scheme developed in [2],

$$\textbf{SI-RK2:}\qquad \begin{aligned} y_i^{(1)} &= \frac{y_i^n + \Delta t (y_j^n)^2}{1+\Delta t y_i^n}, & y_i^{(2)} &= \frac{1}{2} y_i^n + \frac{1}{2}\frac{y_i^{(1)} + \Delta t (y_j^{(1)})^2}{1+\Delta t y_i^{(1)}}, \\ y_i^{n+1} &= \frac{y_i^{(2)} + \Delta t^2 (y_j^{(2)})^2 y_i^{(2)}}{1+\Delta t^2 (y_i^{(2)})^2} & &\text{for } i,j\in\{1,2\} \text{ and } i\neq j. \end{aligned} \tag{2}$$

To this end, we focus on a test initial value problem (also considered in [8])

$$\mathbf{y}'(t) = \mathbf{f}(\mathbf{y}) = \left(\begin{smallmatrix} y_2^2 - y_1^2 \\ y_1^2 - y_2^2 \end{smallmatrix}\right), \quad \mathbf{y}(0) = \mathbf{y}^0 = \left(\begin{smallmatrix} y_1^0 \\ y_2^0 \end{smallmatrix}\right) > \mathbf{0}. \tag{3}$$

The analytical solution of (3) is given by

$$\mathbf{y}(t) = \frac{1}{2}(y_1^0 + y_2^0)\mathbf{1} + \frac{1}{2}\Big(\mathbf{y}^0 - \left(\begin{smallmatrix} y_2^0 \\ y_1^0 \end{smallmatrix}\right)\Big) e^{-2(y_1^0 + y_2^0)t}, \tag{4}$$

where $\mathbf{1} = (1,1)^{\mathrm{T}}$. It is straightforward to see that the exponential term in (4) vanishes for positive initial conditions as $t \longrightarrow \infty$ and one can describe the set of positive steady states of (3) by the intersection $\operatorname{span}(\mathbf{1}) \cap \mathbb{R}^2_{>0}$:

$$\mathbf{y}^* = c\mathbf{1}, \quad c = \frac{1}{2}(y_1^0 + y_2^0) > 0. \tag{5}$$

It can be proven that all positive steady states are stable in the sense of Definition 1 choosing $\delta = \varepsilon$, see [8]. Note that no $\mathbf{y}^*$ in (5) is asymptotically stable since the positive steady states lie on a curve in $\mathbb{R}^2$. Hence, there are infinitely many steady states in every neighborhood of any steady state. Altogether, we see that $\mathbf{y}^*$ is a stable steady state solution of the test problem (3) but $\mathbf{y}^*$ is not asymptotically stable. We hence seek to come across similar stability properties when analyzing the corresponding fixed points of the SI-RK2 method (2).

3 Main Result

In this section, we present our main result where we use the theorems above to analyze the stability of the SI-RK2 method (2) when applied to (3). We note that in our case the matrix $\mathbf{A} = \left(\begin{smallmatrix} 1 & -1 \\ -1 & 1 \end{smallmatrix}\right) = \mathbf{A}^{\mathrm{T}}$ satisfies $\ker(\mathbf{A}) = \operatorname{span}(\mathbf{1}) \subseteq \{\mathbf{y} \in \mathbb{R}^N \mid \mathbf{f}(\mathbf{y}) = \mathbf{0}\}$ and it is sufficient to prove $\mathbf{g} \in \mathcal{C}^2$ rather than that $\mathbf{g} \in \mathcal{C}^1$ has locally Lipschitz first derivatives, [6, 7]. As the result of [7, Remark 2.10] it remains to check, if the function $\mathbf{g}$ generating the SI-RK2 for problem (2)

(i) is in C^2,
(ii) satisfies $\mathbf{g}(\mathbf{y}^*) = \mathbf{y}^*$, where $\mathbf{y}^* = c\mathbf{1}$ for all $c > 0$, and in particular for c as in (5),
(iii) has a Jacobian with a spectrum $\sigma(\mathbf{Dg}(\mathbf{y}^*)) = \{1, R\}$ with $|R| < 1$ (in order to apply part a) Theorem 2, or $|R| > 1$ for part b) of Theorem 1),
(iv) satisfies $\mathbf{1}^\mathrm{T}\mathbf{g}(\mathbf{y}^n) = \mathbf{1}^\mathrm{T}\mathbf{y}^n$ for all $\mathbf{y}^n > \mathbf{0}$ (in order to apply part b) of Theorem 2).

We derive the stability function of the SI-RK2 method, by first introducing the functions $\mathbf{\Phi}_1 \colon \mathbb{R}^2_{>0} \to \mathbb{R}^2_{>0}$, $\mathbf{\Phi}_2 \colon \mathbb{R}^2_{>0} \times \mathbb{R}^2_{>0} \to \mathbb{R}^2_{>0}$ and $\mathbf{\Phi}_{n+1} \colon \mathbb{R}^2_{>0} \to \mathbb{R}^2_{>0}$, whose components are defined by

$$\begin{aligned} (\mathbf{\Phi}_1(\mathbf{x}))_i &= \frac{x_i + \Delta t (x_j)^2}{1 + \Delta t x_i}, & (\mathbf{\Phi}_2(\mathbf{x}, \mathbf{y}))_i &= \frac{1}{2} x_i + \frac{1}{2} (\mathbf{\Phi}_1(\mathbf{y}))_i, \\ (\mathbf{\Phi}_{n+1}(\mathbf{x}))_i &= \frac{x_i + \Delta t^2 (x_j)^2 x_i}{1 + \Delta t^2 (x_i)^2} & &\text{for } i, j \in \{1, 2\} \text{ and } i \neq j, \end{aligned} \tag{6}$$

so that $\mathbf{y}^{(1)} = \mathbf{\Phi}_1(\mathbf{y}^n)$, $\mathbf{y}^{(2)} = \mathbf{\Phi}_2(\mathbf{y}^n, \mathbf{\Phi}_1(\mathbf{y}^n))$ and

$$\mathbf{y}^{n+1} = \mathbf{\Phi}_{n+1}(\mathbf{\Phi}_2(\mathbf{y}^n, \mathbf{\Phi}_1(\mathbf{y}^n))) = \mathbf{g}(\mathbf{y}^n) \tag{7}$$

follow from (2). With that in mind, we see that (i) is satisfied since $\mathbf{g}$ is a composition of C^2-maps for positive arguments.

Secondly for proving (ii), we substitute a steady state $\mathbf{y}^n = \mathbf{y}^* = c\mathbf{1}$ into (2) and find $y^n_j = y^n_i$ for $i, j \in \{1, 2\}$. As a result, we obtain

$$\begin{aligned} (\mathbf{\Phi}_1(\mathbf{y}^*))_i = y^*_i, \qquad (\mathbf{\Phi}_2(\mathbf{y}^*, \mathbf{\Phi}_1(\mathbf{y}^*)))_i = (\mathbf{\Phi}_2(\mathbf{y}^*, \mathbf{y}^*))_i = \frac{1}{2} y^*_i + \frac{1}{2} (\mathbf{\Phi}_1(\mathbf{y}^*))_i = y^*_i, \\ g_i(\mathbf{y}^*) = (\mathbf{\Phi}_{n+1}(\mathbf{\Phi}_2(\mathbf{y}^*, \mathbf{\Phi}_1(\mathbf{y}^*))))_i = (\mathbf{\Phi}_{n+1}(\mathbf{y}^*))_i = y^*_i. \end{aligned} \tag{8}$$

Next, we note that (iv) is not satisfied. To see this, we choose $\mathbf{y}^n = (1, 3)^\mathrm{T}$ and $\Delta t = 1$ which implies $\mathbf{1}^\mathrm{T}\mathbf{y}^n = 4$ (and thus $\mathbf{y}^* = 2 \cdot \mathbf{1}$). A small calculation reveals that $\mathbf{y}^{(1)} = (5, 1)^\mathrm{T}$, $\mathbf{y}^{(2)} = (1, 8)^\mathrm{T}$, and thus $\mathbf{g}(\mathbf{y}^n) = (32.5, 8)^\mathrm{T}$, which implies that $\mathbf{1}^\mathrm{T}\mathbf{g}(\mathbf{y}^n) = 40.5 \neq \mathbf{1}^\mathrm{T}\mathbf{y}^n$. Moreover, these values for $\mathbf{g}$ suggest that we will come across time step restrictions for guaranteeing stability.

Now since (iv) is not satisfied, we cannot apply part b) of Theorem 2 to show the local convergence to the correct steady state solution as $\mathbf{1}^\mathrm{T}\mathbf{y}^* = \mathbf{1}^\mathrm{T}\mathbf{y}^0$ follows from (5) implying that $\mathbf{g}$ does not conserve all linear invariants. Nevertheless, we may apply part a) of Theorem 2 or part b) of Theorem 1 to investigate the stability properties of $\mathbf{y}^*$ as a fixed point of the mapping $\mathbf{g}$. We have to compute the spectrum of $\mathbf{Dg}(\mathbf{y}^*)$. Note that $\mathbf{\Phi}_2 = \mathbf{\Phi}_2(\mathbf{x}, \mathbf{y})$ is a function of two vectors so that we can introduce $\mathbf{D}_\mathbf{x}\mathbf{\Phi}(\mathbf{x}_0, \mathbf{y}_0) = \frac{\partial \mathbf{\Phi}}{\partial \mathbf{x}}(\mathbf{x}_0, \mathbf{y}_0)$, $\mathbf{D}_\mathbf{y}\mathbf{\Phi}(\mathbf{x}_0, \mathbf{y}_0) = \frac{\partial \mathbf{\Phi}}{\partial \mathbf{y}}(\mathbf{x}_0, \mathbf{y}_0)$. Hence, it follows from (7), (8) and the chain rule that

$$\begin{aligned} \mathbf{Dg}(\mathbf{y}^*) &= \mathbf{D\Phi}_{n+1}(\mathbf{y}^*) \Big(\mathbf{D}_\mathbf{x}\mathbf{\Phi}_2(\mathbf{y}^*, \mathbf{y}^*), \mathbf{D}_\mathbf{y}\mathbf{\Phi}_2(\mathbf{y}^*, \mathbf{y}^*) \Big) \begin{pmatrix} \mathbf{I} \\ \mathbf{D\Phi}_1(\mathbf{y}^*) \end{pmatrix} \\ &= \mathbf{D\Phi}_{n+1}(\mathbf{y}^*) \left(\mathbf{D}_\mathbf{x}\mathbf{\Phi}_2(\mathbf{y}^*, \mathbf{y}^*) + \mathbf{D}_\mathbf{y}\mathbf{\Phi}_2(\mathbf{y}^*, \mathbf{y}^*) \mathbf{D\Phi}_1(\mathbf{y}^*) \right). \end{aligned} \tag{9}$$

Introducing the matrix $\mathbf{B} = \left(\begin{smallmatrix} -1 & 2 \\ 2 & -1 \end{smallmatrix}\right)$, we find from (6) and $\mathbf{y}^* = c\mathbf{1}$ that

$$\mathbf{D\Phi}_1(\mathbf{y}^*) = \begin{pmatrix} \frac{1+\Delta t y_1^* - y_1^*(1+\Delta t y_1^*)\Delta t}{(1+\Delta t y_1^*)^2} & \frac{2\Delta t y_2^*}{1+\Delta t y_1^*} \\ \frac{2\Delta t y_1^*}{1+\Delta t y_2^*} & \frac{1+\Delta t y_1^* - y_1^*(1+\Delta t y_1^*)\Delta t}{(1+\Delta t y_1^*)^2} \end{pmatrix} = \begin{pmatrix} \frac{1-c\Delta t}{1+c\Delta t} & \frac{2c\Delta t}{1+c\Delta t} \\ \frac{2c\Delta t}{1+c\Delta t} & \frac{1-c\Delta t}{1+c\Delta t} \end{pmatrix} = \frac{\mathbf{I} + c\Delta t\mathbf{B}}{1 + c\Delta t}.$$

Furthermore, it is straightforward to see that $\mathbf{D_x\Phi}_2(\mathbf{y}^*, \mathbf{y}^*) = \frac{1}{2}\mathbf{I}$, $\mathbf{D_y\Phi}_2(\mathbf{y}^*, \mathbf{y}^*) = \frac{1}{2}\mathbf{D\Phi}_1(\mathbf{y}^*)$. Finally, introducing $\mathbf{C} = \left(\begin{smallmatrix} -1 & 1 \\ 1 & -1 \end{smallmatrix}\right)$, we compute the Jacobian of $\mathbf{\Phi}_{n+1}$ which reads

$$\mathbf{D\Phi}_{n+1}(\mathbf{y}^*) = \begin{pmatrix} 1 - \frac{2c^2\Delta t^2}{1+c^2\Delta t^2} & \frac{2c^2\Delta t^2}{1+c^2\Delta t^2} \\ \frac{2c^2\Delta t^2}{1+c^2\Delta t^2} & 1 - \frac{2c^2\Delta t^2}{1+c^2\Delta t^2} \end{pmatrix} = \mathbf{I} + \tfrac{2c^2\Delta t^2}{1+c^2\Delta t^2}\mathbf{C}.$$

Altogether, (9) yields

$$\mathbf{Dg}(\mathbf{y}^*) = \left(\mathbf{I} + \tfrac{2c^2\Delta t^2}{1+c^2\Delta t^2}\mathbf{C}\right)\frac{1}{2}\left(\mathbf{I} + \left(\frac{1}{1 + c\Delta t}(\mathbf{I} + c\Delta t\mathbf{B})\right)^2\right) \tag{10}$$

Note that $\mathbf{By}^* = \mathbf{y}^*$ and $\mathbf{Cy}^* = \mathbf{0}$, which implies

$$\mathbf{Dg}(\mathbf{y}^*)\mathbf{y}^* = \frac{1}{2}\left(1 + \left(\frac{1}{1 + c\Delta t}(1 + c\Delta t)\right)^2\right)\mathbf{y}^* = \mathbf{y}^*.$$

Similarly, defining $\bar{\mathbf{y}} = (1, -1)^{\mathrm{T}}$, we see that $\mathbf{B}\bar{\mathbf{y}} = -3\bar{\mathbf{y}}$ and $\mathbf{C}\bar{\mathbf{y}} = -2\bar{\mathbf{y}}$. Thus,

$$\mathbf{Dg}(\mathbf{y}^*)\bar{\mathbf{y}} = \left(1 - 2\tfrac{2c^2\Delta t^2}{1+c^2\Delta t^2}\right)\frac{1}{2}\left(1 + \left(\frac{1}{1 + c\Delta t}(1 - 3c\Delta t)\right)^2\right)\bar{\mathbf{y}}$$

Hence, using Theorem 2 and substituting $z = \Delta t c$, we have to analyze the stability function

$$R(z) = \frac{1}{2}\left(1 - \frac{4z^2}{1 + z^2}\right)\left(1 + \frac{(1 - 3z)^2}{(1 + z)^2}\right) = \frac{-15z^4 + 6z^3 + 2z^2 - 2z + 1}{(z^2 + 1)(z + 1)^2} \tag{11}$$

for $z > 0$. Indeed, we obtain our main result stated in the following theorem.

Theorem 3 *The stability function R from* (11) *satisfies $R(0) = 1$, $R(1) = -1$ and $R'(z) < 0$ for all $z > 0$. In particular, $z = 1$ is the only positive solution to $|R(z)| = 1$. Furthermore, we have $|R(z)| < 1$ for $z \in (0, 1)$ and $|R(z)| > 1$ for $z > 1$.*

Proof $R(0) = 1$ is trivial and we find $R(1) = \frac{-15+6+2-2+1}{2\cdot 2^2} = \frac{-8}{8} = -1$. Through a small calculation, the first derivative of R can be computed according to

$$R'(z) = 4\frac{-9z^5 - 7z^4 - 12z^3 + 4z^2 + z - 1}{(z^2 + 1)^2(z + 1)^3}.$$

Since the denominator is positive for $z > 0$, we show that the numerator $p(z) = -9z^5 - 7z^4 - 12z^3 + 4z^2 + z - 1$ is negative on $\mathbb{R}^+$. To see this, we first point out that for $z = 0$, the numerator is negative; hence, this is even true for small positive values of z. By applying Sturm's theorem (cf. [3, Theorem 8.8.14]), we demonstrate that the numerator has no positive zeros. For this, we first compute the Sturm chain of p consisting of a sequence of polynomials $p_0, p_1, \dots$ such that $p_0 = p,\ p_1 = p',\ p_{i+1} = -\operatorname{rem}(p_{i-1}, p_i),\ i \geq 1,$ where $\operatorname{rem}(p_{i-1}, p_i)$ denotes the remainder of the Euclidean division of p_{i-1} by p_i. The sequence stops at a constant polynomial p_k and the polynomials might be scaled with a positive number to avoid fractions and to obtain coprime coefficients. In our case, we find $p_1(z) = p'(z) = -45z^4 - 28z^3 - 36z^2 + 8z + 1$. A technical but elementary computation yields

$$\begin{aligned} p_2(z) &= 221z^3 - 198z^2 - 31z + 58, & p_3(z) &= 22471z^2 - 2220z - 4109,\\ p_4(z) &= 1878z - 6059, & p_5(z) &= -1. \end{aligned}$$

To apply Sturm's theorem, we compute the signs of $p_i(0)$ and $\lim_{z\to\infty} p_i(z)$, corresponding to the sign of the leading coefficient, for $i = 0, \dots, 5$. The sequence of signs of $p_i(0)$ is $(-, +, +, -, -, -)$, where two sign changes occur; one from $-$ to $+$ and one from $+$ to $-$. The other sequence reads $(-, -, +, +, +, -)$, where two sign changes can also be observed. Sturm's theorem now states that the difference of sign changes of the two sequences is equal to the number of positive zeros of p, which in this case is zero. Hence, p has no positive zeros and is negative for small positive z, proving that $p(z) < 0$ for all $z > 0$. This means that $R'(z) < 0$ for all $z > 0$ showing the remaining claims of the theorem. □

Theorems 1 and 2, together with our main result, imply the following time step restriction to ensure stability.

Corollary 1 *Let* $\mathbf{y}^* = c\mathbf{1}$ *with* $c > 0$ *be a positive steady state of the differential equation* (3)*. Then* $\mathbf{y}^*$ *is a stable fixed point of the SI-RK2 scheme if* $c\Delta t < 1$*. If* $c\Delta t > 1$*, then* $\mathbf{y}^*$ *is an unstable fixed point of the SI-RK2 method.*

According to [10, Theorem 1], a necessary condition for avoiding oscillations is a strictly positive stability function, for which the following statement is helpful.

Corollary 2 *The only positive root of the stability function R given in* (11) *is* $z^* = \frac{\sqrt{3}}{3}$.

Proof Due to Theorem 3, there exists only one positive root. Using the factored version of R from (11), we find $1 - \frac{4(z^*)^2}{1+(z^*)^2} = 1 - \frac{\frac{12}{9}}{1+\frac{3}{9}} = 0$, which proves $R(z^*) = 0$.□

4 Numerical Experiments

In this section, we numerically validate the statements of Corollary 1 and the result [10, Theorem 1], where instead of $|R(z)| < 1$ one requires even $0 < R(z) < 1$ as a necessary condition for avoiding oscillations. We solve the test problem (3) subject to several initial condirtions using the SI-RK2 method with different time steps. To get an insight if the time step restrictions from Corollary 1 are severe, we also estimate the time t^* at which $\|\mathbf{y}(t^*) - \mathbf{y}^*\|_\infty = 10^{-15}$ holds. This condition is equivalent to

$$\left\|\frac{1}{2}\left(\mathbf{y}^0 - \begin{pmatrix} y_2^0 \\ y_1^0 \end{pmatrix}\right)e^{-2(y_1^0+y_2^0)t^*}\right\|_\infty = 10^{-15} \quad \Rightarrow \quad t^* = \frac{\ln\left(10^{15}\frac{|y_1^0-y_2^0|}{2}\right)}{2(y_1^0 + y_2^0)},$$

in the case of positive $\mathbf{y}^0 \notin \operatorname{span}(\mathbf{y}^*)$. In the following experiments, we consider the initial conditions of the form

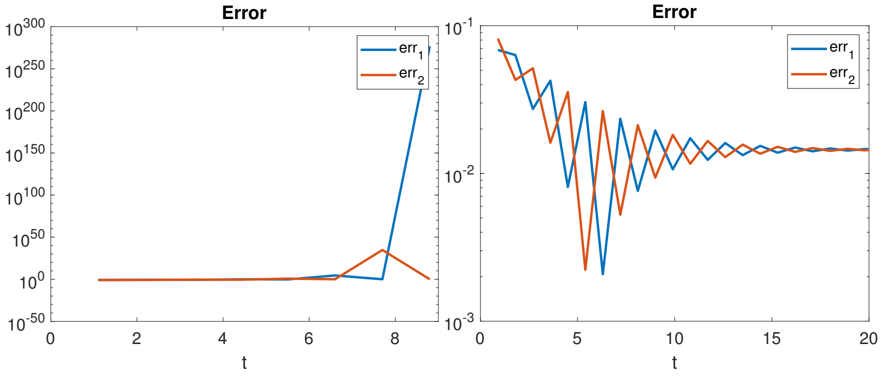
$$\mathbf{y}^0 = c\mathbf{1} + \varepsilon\bar{\mathbf{y}} = (c + \varepsilon, c - \varepsilon)^{\mathrm{T}} \tag{12}$$

with $c \in \{1,5\}$ and $\varepsilon \in \{10^{-1}, 10^{-5}\}$. As a result we can rewrite t^* in terms of c and ε obtaining $t^*(c,\varepsilon) = \frac{\ln(10^{15}\varepsilon)}{4c}$. Note that according to (5), the steady state solution is $\mathbf{y}^* = c\mathbf{1}$, and hence, Corollary 1 states that $\mathbf{y}^*$ is stable for $\Delta t < \frac{1}{c} = \Delta t^*$ while it is unstable for $\Delta t > \Delta t^*$. Unfortunately, the local convergence towards $\mathbf{y}^*$ is not guaranteed, so the error cannot be expected to approach 10^{-16}. Nevertheless, Lyapunov's stable fixed points have the property that the error is bounded, which might not be true for unstable fixed points. Furthermore, since these stability properties are local, we expect to see the impact of varying the time step size Δt^* the better, the closer we choose $\mathbf{y}^0$ to $\mathbf{y}^*$. To illustrate these stability properties, we provide for all four combinations of (c,ε) pairs mentioned above two plots using $\Delta t_{1,2} = \Delta t^* \pm \varepsilon = \frac{1}{c} \pm \varepsilon$, i. e. $\Delta t_1 > \Delta t^*$ and $\Delta t_2 < \Delta t^*$. Note that by these choices of time step sizes, we vary Δt^* less the closer $\mathbf{y}^0$ is to $\mathbf{y}^*$, i. e. the smaller ε is.

In Figures 1–4, we plot the absolute error $\mathrm{err}_i = |y_i^n - y_i(n\Delta t)|$, where y_i^n, $i = 1,2$, is the numerical solution obtained by SI-RK2 and $y_i(n\Delta t)$ is the exact solution of the initial-value problem (3), (12) as a function of time. It can be observed that the theoretical claims are well reflected as the error is bounded when $\Delta t = \Delta t_2$ and unbounded if $\Delta t = \Delta t_1$. It is also worth mentioning that in all cases, the time step is smaller than the respective $t^*(c,\varepsilon)$. Moreover, in Figures 1b, 3b, 2b and 4b, a global minimum of the error curves can be observed. This is due to the fact that the numerical approximation is close to the correct steady state at some point but eventually converges to a different vector; see, for instance, Figure 5a. However, better performance of the numerical scheme can be expected according to [10, Theorem 1] when we choose Δt such that $R(c\Delta t) > 0$. Due to $R(0) = 1$ and Corollary 2, this is the case if and only if $c\Delta t < z^* = \frac{\sqrt{3}}{3}$. In Figure 5b and 5c, one can see that the numerical approximation indeed does not show oscillatory behavior and that the error is bounded by 10^{-10} rather than 10^{-5} as it is for $\Delta t = \Delta t_2$, see Figure 3b.

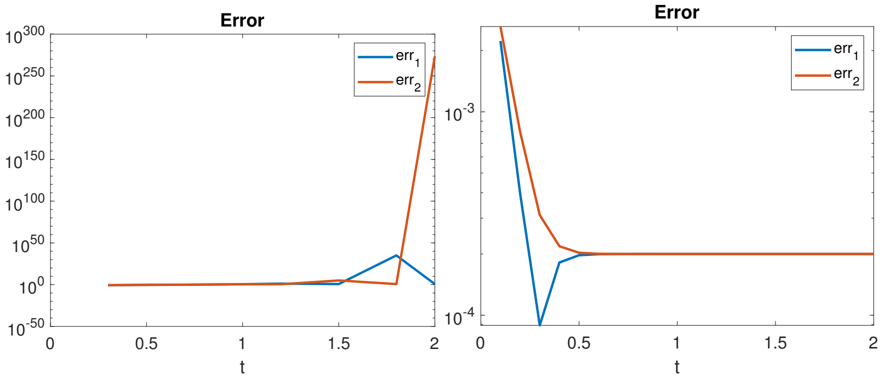
5 Summary and Conclusion

In this work, we analyzed the second-order semi-implicit Runge–Kutta method when applied to a system of two differential equations with a nonlinear right-hand side. Thereby, we discovered time step restrictions coming from the Lyapunov stability



(a) $\Delta t = \Delta t_1$ (b) $\Delta t = \Delta t_2$

Fig. 1: Error as a function of time for the case of $c = 1, \varepsilon = 10^{-1}, t^*(c, \varepsilon) \approx 8.06$.

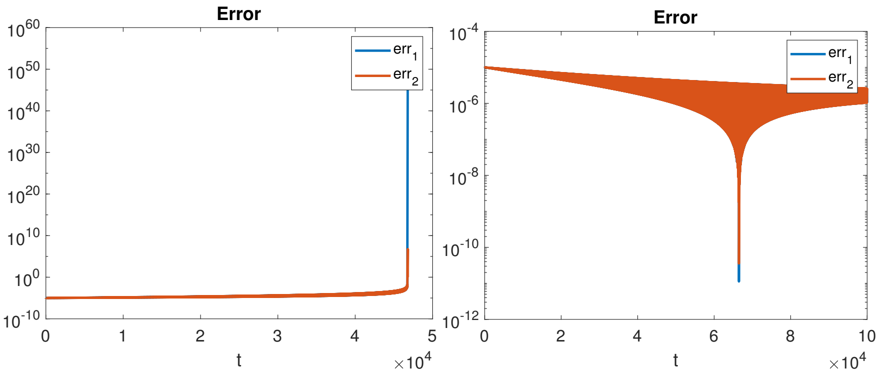


(a) $\Delta t = \Delta t_1$ (b) $\Delta t = \Delta t_2$

Fig. 2: Error as a function of time for the case of $c = 5, \varepsilon = 10^{-1}, t^*(c, \varepsilon) \approx 1.61$.

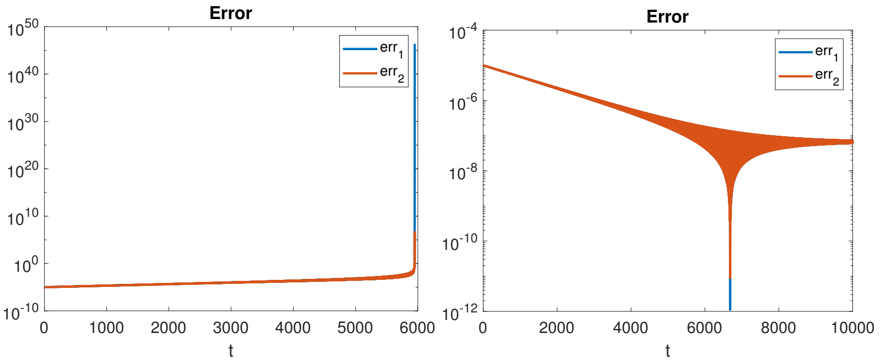
analysis of the fixed points of the method. We have performed several numerical experiments that confirm the theoretical claims. Moreover, the numerical results also suggest, in accordance with the presented theory, that the iterates do not necessarily converge towards the correct steady state solution of the underlying problem if the system of differential equations possesses linear invariants.

Future works include the investigation of higher-order semi-implicit Runge–Kutta schemes. Thereby, the derivation of Lyapunov stability properties and the time step restrictions coming from the necessary condition for avoiding oscillatory behavior are of great interest. Moreover, we aim to investigate these stability properties also in the context of hyperbolic balance laws.



(a) $\Delta t = \Delta t_1$ (b) $\Delta t = \Delta t_2$

Fig. 3: Error as a function of time for the case of $c = 1, \varepsilon = 10^{-5}, t^*(c, \varepsilon) \approx 5.76$.



(a) $\Delta t = \Delta t_1$ (b) $\Delta t = \Delta t_2$

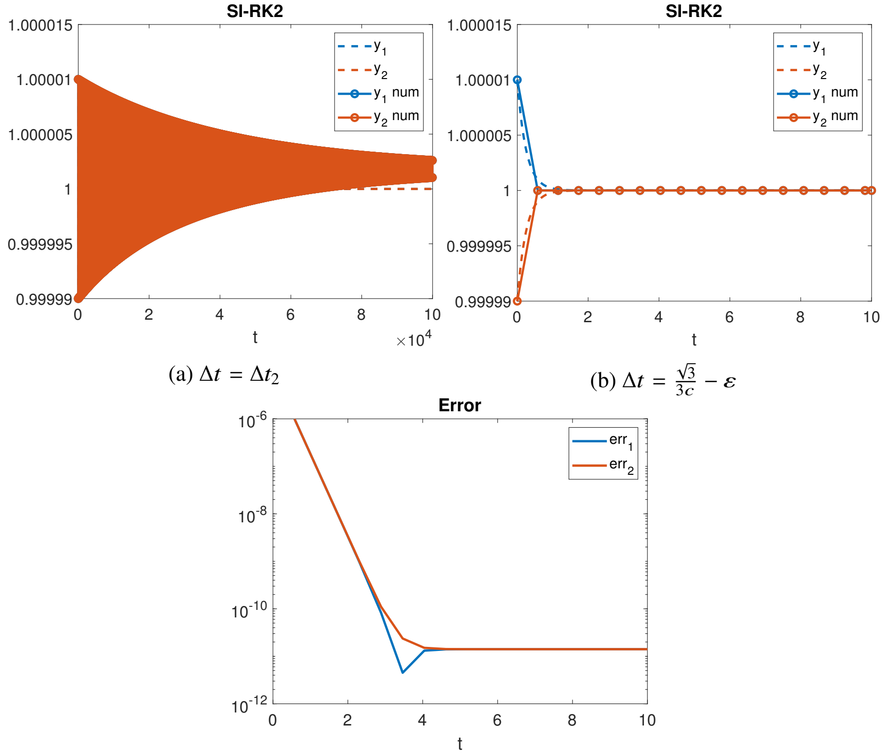
Fig. 4: Error as a function of time for the case of $c = 5, \varepsilon = 10^{-5}, t^*(c, \varepsilon) \approx 1.15$.

6 Acknowledgements

The work of A. Chertock was supported in part by NSF grants DMS-1818684 and DMS-2208438. The author T. Izgin, gratefully acknowledges the financial support by the Deutsche Forschungsgemeinschaft (DFG) through grant ME 1889/10-1. P. Öffner thanks the Gutenberg research college.

References

1. C. Bolley and M. Crouzeix, *Conservation de la positivité lors de la discrétisation des problèmes d'évolution paraboliques*, RAIRO Anal. Numér., 12 (1978), pp. 237–245, iv.
2. A. Chertock, S. Cui, A. Kurganov, and T. Wu, *Steady state and sign preserving semi-implicit Runge-Kutta methods for ODEs with stiff damping term*, SIAM J. Numer. Anal., 53 (2015), pp. 2008–2029.



(a) $\Delta t = \Delta t_2$

(b) $\Delta t = \frac{\sqrt{3}}{3c} - \varepsilon$

(c) $\mathrm{err}_i = |y_i^n - y_i(n\Delta t)|$ for $\Delta t = \frac{\sqrt{3}}{3c} - \varepsilon$

Fig. 5: Numerical approximation of (3), (12) with $c = 1, \varepsilon = 10^{-5}$ using SI-RK2.

3. P. M. Cohn, *Basic algebra*, Springer-Verlag London, Ltd., London, 2003. Groups, rings and fields.
4. J. Huang, T. Izgin, S. Kopecz, A. Meister, and C.-W. Shu, *On the stability of strong-stability-preserving modified Patankar Runge-Kutta schemes*, https://arxiv.org/abs/2205.01488, (2022).
5. T. Izgin, S. Kopecz, A. Martiradonna, and A. Meister, *Lyapunov stability of first and second order geco and gbbks schemes*, https://arxiv.org/abs/2301.10658, (2023).
6. 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 (2022), pp. 1053–1080.
7. ———, *On the stability of unconditionally positive and linear invariants preserving time integration schemes*, SIAM J Numer Anal., 60 (2022), pp. 3029–3051.
8. ———, *A stability analysis of modified Patankar-Runge-Kutta methods for a nonlinear production-destruction system*, https://arxiv.org/abs/2210.11845, (2022).
9. T. Izgin and P. Öffner, *On the stability of modified Patankar methods*, https://arxiv.org/abs/2206.07371, (2022).
10. 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).
11. A. Stuart and A. R. Humphries, *Dynamical systems and numerical analysis*, vol. 2, Cambridge University Press, 1998.