Stefano Maset*

Relative error analysis of matrix exponential approximations for numerical integration

Abstract: In this paper, we study the relative error in the numerical solution of a linear ordinary differential equation $y'(t) = Ay(t)$, $t \geqslant 0$, where A is a normal matrix. The numerical solution is obtained by using at any step an approximation of the matrix exponential, e.g., a polynomial or a rational approximation. The error of the numerical solution with respect to the exact solution is due to this approximation as well as to a possible perturbation in the initial value. For an unperturbed initial value, we have found: (1) unlike the absolute error, the relative error always grows linearly in time; (2) in the long-time, the contributions to the relative error relevant to non-rightmost eigenvalues of A disappear.

Keywords: relative error, numerical integration, approximation of the matrix exponential

1 Introduction

The paper [9] considered the Ordinary Differential Equation (ODE):

$$\begin{cases} y'(t) = Ay(t), & t \geqslant 0 \\ y(0) = y_0 \end{cases} \tag{1.1}$$

where $A \in \mathbb{R}^{d\times d}$ and $y(t) \in \mathbb{R}^d$, and studied the propagation along the solution y of a perturbation in the initial value y_0, by considering relative errors rather than absolute errors. Whereas the propagation of the absolute error of such a perturbation is a well-known subject (the absolute error is propagated at the time t by the matrix exponential e^{tA} and bounds for it can be found, e.g., in [4]), the propagation of the relative error has not been considered and [9] tried to fill this gap. It is important to consider relative errors since they are scale-independent indicators of the quality of an approximation: unlike absolute errors, relative errors are dimensionless.

By assuming that in (1.1) the initial value $y_0 \neq 0$ is perturbed to $\widetilde{y}_0$ and then the solution y is perturbed to $\widetilde{y}$, the paper [9] studied the relation between the relative error

$$\varepsilon = \frac{\|\widetilde{y}_0 - y_0\|_2}{\|y_0\|_2} \tag{1.2}$$

of the perturbed initial value $\widetilde{y}_0$ and the relative error

$$\delta(t) = \frac{\|\widetilde{y}(t) - y(t)\|_2}{\|y(t)\|_2}, \quad t \geqslant 0 \tag{1.3}$$

of the perturbed solution $\widetilde{y}$, in case of a normal matrix A.

In [9], the following theorem was stated.

Theorem 1.1. *Let A be a normal matrix. Partition the spectrum*

$$\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_p\}$$

***Corresponding author: Stefano Maset,** Università di Trieste, Dipartimento di Matematica e Geoscienze, Via Valerio 12/A, 34127 Trieste, Italy. Email: maset@units.it

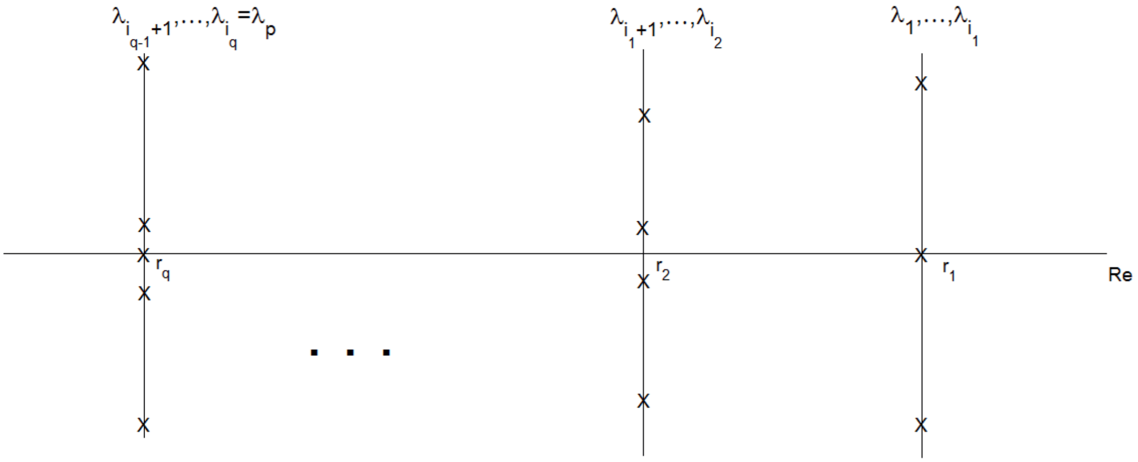
$\lambda_{i_{q-1}+1},\dots,\lambda_{i_q}=\lambda_p$
$\lambda_{i_1+1},\dots,\lambda_{i_2}$
$\lambda_1,\dots,\lambda_{i_1}$
r_q
r_2
r_1
Re
. . .

Fig. 1: Spectrum of A partitioned by decreasing real part.

of matrix A, where $\lambda_1, \lambda_2, \dots, \lambda_p$ are the distinct eigenvalues of A, by decreasing real part in the subsets $\Lambda_1, \Lambda_2, \dots, \Lambda_q$: we have

$$\begin{aligned} \Lambda_j &= \{\lambda_{i_{j-1}+1}, \lambda_{i_{j-1}+2}, \dots, \lambda_{i_j}\} \\ \operatorname{Re}\left(\lambda_{i_{j-1}+1}\right) &= \operatorname{Re}\left(\lambda_{i_{j-1}+2}\right) = \cdots = \operatorname{Re}\left(\lambda_{i_j}\right) = r_j, \quad j = 1, 2, \dots, q \end{aligned}$$

with $0 = i_0 < i_1 < \cdots < i_q = p$ ($i_1, \dots, i_q$ are the final indices i_j in the sets Λ_j) and

$$r_1 > r_2 > \cdots > r_q$$

(see Fig. 1).

Suppose that the initial value $y_0 \neq 0$ of (1.1) *is perturbed to $\widetilde{y}_0$. We have*

$$\delta(t) = \frac{\sqrt{\sum\limits_{j=1}^{q} \left(\mathrm{e}^{(r_j - r_1)t} \left\|Q_j \widehat{z}_0\right\|_2\right)^2}}{\sqrt{\sum\limits_{j=1}^{q} \left(\mathrm{e}^{(r_j - r_1)t} \left\|Q_j \widehat{y}_0\right\|_2\right)^2}} \cdot \varepsilon, \quad t \geqslant 0 \tag{1.4}$$

where

$$\widehat{z}_0 = \frac{\widetilde{y}_0 - y_0}{\|\widetilde{y}_0 - y_0\|_2}$$

is the direction of the perturbation,

$$\widehat{y}_0 = \frac{y_0}{\|y_0\|_2}$$

and

$$Q_j = \sum_{\lambda_i \in \Lambda_j} P_i, \quad j = 1, \dots, q \tag{1.5}$$

with P_i the orthogonal projection on the eigenspace of λ_i.

In (1.5), $\lambda_i \in \Lambda_j$ means $i = i_{j-1}+1, i_{j-1}+2, \dots, i_j$. Observe that there is a one-to-one correspondence between indices in $\{1, \dots, p\}$ and eigenvalues in $\Lambda = \{\lambda_1, \dots, \lambda_p\}$.

Throughout the paper we use the notations introduced in Theorem 1.1.

Remark 1.1. In practical situations of uncertainty, we do not know the direction $\widehat{z}_0$ of the perturbation but only the order of magnitude of ε. In this case, we can only say that

$$\delta(t) \leqslant K(t, A, y_0) \cdot \varepsilon, \quad t \geqslant 0 \tag{1.6}$$

where

$$K(t,A,y_0)=\max_{\substack{\widehat{z}_0\in\mathbb{R}^n\\ \|\widehat{z}_0\|_2=1}}\frac{\sqrt{\sum\limits_{j=1}^{q}\left(\mathrm{e}^{(r_j-r_1)t}\,\|Q_j\widehat{z}_0\|_2\right)^2}}{\sqrt{\sum\limits_{j=1}^{q}\left(\mathrm{e}^{(r_j-r_1)t}\,\|Q_j\widehat{y}_0\|_2\right)^2}}=\frac{1}{\sqrt{\sum\limits_{j=1}^{q}\left(\mathrm{e}^{(r_j-r_1)t}\,\|Q_j\widehat{y}_0\|_2\right)^2}}$$

with '$\leqslant$' replaced by '$=$' for all $t\geqslant 0$ if $Q_1\widehat{z}_0=\widehat{z_0}$ (see [9]). The number $K(t,A,y_0)$ is the normwise condition number (in the euclidean norm) for the problem $y_0\mapsto \mathrm{e}^{tA}y_0$ (see [1]).

If the initial value y_0 is also not known, we can only say

$$\delta(t)\leqslant K(t,A)\cdot\varepsilon,\quad t\geqslant 0 \tag{1.7}$$

where

$$K(t,A)=\max_{y_0\in\mathbb{R}^n\setminus\{0\}}K(t,A,y_0)=\mathrm{e}^{(r_1-r_q)t}$$

with '$\leqslant$' replaced by '$=$' for all $t\geqslant 0$ if $Q_1\widehat{z}_0=\widehat{z}_0$ and $Q_q y_0=y_0$. The number $K(t,A)$ is the condition number (in the spectral norm) of the matrix exponential e^{tA}.

The bound (1.7) is too pessimistic, since it holds with equality in a non-generic situation for y_0. For this reason, the bound (1.6) involving y_0 should be preferred.

In our relative error analysis presented below, the direction $\widehat{z}_0$ of the perturbation will be always involved. In situation of uncertainty about $\widehat{z}_0$, one can consider the worst case, or the average case, for $\widehat{z}_0$.

Theorem 1.1 describes the propagation of the relative error when the initial value of (1.1) is perturbed and (1.1) is exactly solved. Aim of this paper is to study the propagation of the relative error when the initial value of (1.1) is perturbed and (1.1) is numerically solved rather than exactly solved.

So, we suppose that the initial value $y_0\neq 0$ of (1.1) is perturbed to $\widetilde{y}_0$ and the relevant perturbed solution $\widetilde{y}$ is numerically computed over the mesh

$$t_n=nh,\quad n=0,1,2,\ldots \tag{1.8}$$

of constant stepsize $h>0$, with numerical solution $\widetilde{y}_0,\widetilde{y}_1,\widetilde{y}_2,\ldots$ given by

$$\widetilde{y}_{n+1}=R(hA)\,\widetilde{y}_n,\quad n=0,1,2,\ldots \tag{1.9}$$

and then

$$\widetilde{y}_n=R(hA)^n\,\widetilde{y}_0,\quad n=0,1,2,\ldots$$

where $R(z)$, $z\in\mathcal{D}\subseteq\mathbb{C}$, is an analytic approximant of the exponential e^z, $z\in\mathbb{C}$, e.g., a polynomial or a rational approximant. Here, $\mathcal{D}$ is the domain of R. When a Runge–Kutta (RK) method is applied over the mesh (1.8), we obtain a numerical solution (1.9) with R the stability function of the RK method (see [7]), which is a polynomial or a rational function. We assume that the approximant R has order l, where l is a positive integer, i.e., we have

$$R(z)-\mathrm{e}^z=Cz^{l+1}+O(z^{l+2}),\quad z\to 0 \tag{1.10}$$

for some constant $C\neq 0$. Moreover, we assume $h\lambda_i\in\mathcal{D}$, $i=1,\ldots,p$.

In this paper, we study the relative error

$$\delta_n=\frac{\|\widetilde{y}_n-y(t_n)\|_2}{\|y(t_n)\|_2},\quad n=0,1,2,\ldots \tag{1.11}$$

of the perturbed numerical solution $\widetilde{y}_n$ with respect to the exact solution $y(t_n)$. This error is due to the perturbed initial value and to the approximant. Let

$$\gamma_n=\frac{\|y_n-y(t_n)\|_2}{\|y(t_n)\|_2},\quad n=0,1,2,\ldots \tag{1.12}$$

where

$$y_n = R\,(h\lambda)^n\, y_0$$

be the relative error δ_n for an unperturbed initial value.

As in [9], we suppose A normal. The assumption that A is normal is not too much restrictive, since the family of normal matrices includes important types of matrices as symmetric, skew-symmetric, shifted skew-symmetric, and orthogonal matrices. Moreover, the test problem (1.1) with A normal is worthwhile to be investigated in this context of the relative error analysis, since it shows new and unexplored situations about the numerical integration of ODEs. Finally, it seems adequate to consider normal matrices in a first paper on the subject, as the present paper is.

The linear ODE (1.1) with A normal can be seen as d uncoupled linear scalar ODEs, once we consider components in the orthonormal base of eigenvectors. So, we could assume from the beginning that A is diagonal, but we do not do this because it does not simplify the exposition: for example, a formula like (1.4) is not simplified by knowing that A is diagonal. However, it is clear that the net and simple form that this formula has is due to the decoupling of the ODE.

We remark that a relative error analysis of numerical solutions of ODEs is not yet accomplished in literature. Indeed, it is tradition in numerical ODEs to measure errors by absolute errors, not by relative errors, perhaps because one assumes that the solution does not become neither small nor large and so absolute and relative errors have the same order of magnitude. Clearly, absolute error and relative error have the same order of convergence to zero with respect to h, but the fundamental point is that the two errors behave differently with respect to time.

The only situation (however important) where relative errors are considered is in the error control for variable stepsize integrators. Such integrators produce a sequence $\{y_n\}$ of approximations for the values $y(t_n)$ of the solution y of a d-dimensional ODE by selecting stepsizes $h_{n+1} = t_{n+1} - t_n$, $n = 0, 1, 2, \ldots$, such that

$$\left|y_{n+1,i} - z_{n+1,i}(t_{n+1})\right| \leqslant \mathrm{ATOL}_i + \mathrm{RTOL}_i \cdot \left|y_{n,i}\right|$$

holds for each component $y_{n+1,i}$, $i = 1, \ldots, d$, of the numerical solution y_{n+1}. Here, z_{n+1} of components $z_{n+1,i}$, $i = 1, \ldots, d$, is the solution exiting from (t_n, y_n) and ATOL_i and RTOL_i, $i = 1, \ldots, d$, are fixed tolerances on absolute and relative errors, respectively, see [15]. In this context, the relative errors are considered componentwise, not normwise as in the present paper.

We recall that in literature both componentwise and normwise approaches are considered in studying relative errors (see [1]). Moreover, it is worthwhile to remark that normwise relative error and componentwise relative errors are strictly related (see the introduction in [9]): we have

$$\begin{aligned}
\varepsilon &\leqslant \max_{i=1,\ldots,d} \frac{\left|\widetilde{y}_{0,i} - y_{0,i}\right|}{\left|y_{0,i}\right|}\\
\delta_n &\leqslant \max_{i=1,\ldots,d} \frac{\left|\widetilde{y}_{n,i} - y_i\,(t_n)\right|}{\left|y_i\,(t_n)\right|}, \qquad n = 0, 1, 2, \ldots\\
\frac{\left|\widetilde{y}_{n,i} - y_i\,(t_n)\right|}{\left|y_i\,(t_n)\right|} &\leqslant \frac{\|y(t_n)\|_2}{|y_i(t_n)|}\delta_n, \qquad n = 0, 1, 2, \ldots, \quad i = 1, \ldots, d
\end{aligned}$$

where $\widetilde{y}_{0,i}$, $y_{0,i}$, $\widetilde{y}_{n,i}$, and $y_i(t_n)$, $i = 1, \ldots, d$, are the components of $\widetilde{y}_0$, y_0, $\widetilde{y}_n$, and $y(t_n)$, respectively. So, if all the components of y_0 are perturbed with relative errors within a tolerance TOL, then $\varepsilon \leqslant \mathrm{TOL}$ and if δ_n is known to be large, then some component of $y(t)$ has a large relative error.

The plan of the paper is as follows: In Section 2, we give a formula for the relative error δ_n defined in (1.11), on which our relative error analysis of the numerical solution is based on. In Section 3, we define the errors introduced by the approximant. The relative error analysis is presented in Section 4: the main finding of this section is that the relative error γ_n defined in (1.12) and relevant to an unperturbed initial value grows linearly in time. A long-time relative error analysis is given in Section 5: the main finding of this section is that, in the long-time, the contributions to the error γ_n coming from non-rightmost eigenvalues of A vanish. Examples with numerical experiments are considered in Section 6 and Section 7 and conclusions are draft in Section 8.

Observe that the classical ‘scaling and squaring method’ (see [5]) for computing the matrix exponential is related to our study. In fact, this method computes e^{A} as $R(hA)^{n}$, where $h = 1/2^{s}$ and $n = 2^{s}$ for a suitable positive integer s, with R a polynomial or rational approximant of the exponential. In literature several relative error analyses of the ‘scaling and squaring method’ can be found (see [2, 5, 6, 13, 21]), but they are different from the analysis presented here. In fact, in these analyses the interest is on the computation of the matrix exponential, not on the computation of the matrix exponential times a vector: in other words, the interest is on the relative error

$$\frac{\left\| R(hA)^{n} - \mathrm{e}^{A} \right\|_{2}}{\left\| \mathrm{e}^{A} \right\|_{2}}$$

not on the relative error

$$\gamma_{n} = \frac{\left\| \left(R(hA)^{n} - \mathrm{e}^{A} \right) y_{0} \right\|_{2}}{\left\| \mathrm{e}^{A} y_{0} \right\|_{2}}$$

considered in the present paper.

We conclude this introduction by precising the meaning of the relations $a \approx b$ and $a \lessapprox b$, which are used throughout the paper. For $a, b \in \mathbb{C}$ or $a, b \in \mathbb{R}$, $a \approx b$ means

$$a = b\,(1 + e)$$

with $|e| \ll 1$. For $a, b \in \mathbb{R}$, $a \lessapprox b$ means $a \leqslant c$ for some $c \in \mathbb{R}$ such that $c \approx b$. For $a, b \in \mathbb{R}^{d}$, $a \approx b$ means

$$\frac{\|a - b\|_{2}}{\|b\|_{2}} \ll 1.$$

In the present paper, for $a \geqslant 0$, ‘a small’ is the same as ‘$a \ll 1$’ and ‘a large’ is the same as ‘$a \gg 1$’.

2 A formula for the error δ_n

We introduce the *relative approximant* $S : \mathcal{D} \to \mathbb{C}$ given by

$$S\,(z) = \mathrm{e}^{-z} R\,(z)\,, \quad z \in \mathcal{D}.$$

The following theorem provides an useful formula for the relative error δ_{n} defined in (1.11), in case of a normal matrix A.

Theorem 2.1. *Let A be a normal matrix. For the relative error δ_{n}, we have*

$$\delta_{n} = \frac{\sqrt{\sum\limits_{j=1}^{q} \left(\mathrm{e}^{(r_{j}-r_{1})t_{n}} \varepsilon_{n,j} \right)^{2}}}{\sqrt{\sum\limits_{j=1}^{q} \left(\mathrm{e}^{(r_{j}-r_{1})t_{n}} \left\| Q_{j} \widehat{y}_{0} \right\|_{2} \right)^{2}}}, \quad n = 0, 1, 2, \ldots \tag{2.1}$$

where

$$\varepsilon_{n,j} := \left\| \sum_{\lambda_{i} \in \Lambda_{j}} \left(\left(S\,(h\lambda_{i})^{n} - 1 \right) P_{i} \widehat{y}_{0} + \varepsilon S\,(h\lambda_{i})^{n} P_{i} \widehat{z}_{0} \right) \right\|_{2}. \tag{2.2}$$

Proof. Fix $n = 0, 1, 2, \ldots$ We have

$$\widetilde{y}_{n} = \mathrm{e}^{t_{n}A} \mathrm{e}^{-t_{n}A} R\,(hA)^{n}\, \widetilde{y}_{0} = \mathrm{e}^{t_{n}A} \widetilde{Y}_{n}$$

i.e., $\widetilde{y}_{n}$ is the perturbed exact solution of (1.1) at t_{n}, when the initial value y_{0} is perturbed to

$$\widetilde{Y}_{n} = \mathrm{e}^{-t_{n}A} R\,(hA)^{n}\, \widetilde{y}_{0} = S(hA)^{n} \widetilde{y}_{0}.$$

This is backward error analysis.

Theorem 1.1 says that

$$\delta_n = \frac{\sqrt{\sum\limits_{j=1}^{q} \left(e^{(r_j - r_1)t_n} \left\| Q_j \widehat{Z}_n \right\|_2 \right)^2}}{\sqrt{\sum\limits_{j=1}^{q} \left(e^{(r_j - r_1)t_n} \left\| Q_j \widehat{y}_0 \right\|_2 \right)^2}} \cdot E_n = \frac{\sqrt{\sum\limits_{j=1}^{q} \left(e^{(r_j - r_1)t_n} \left\| Q_j E_n \widehat{Z}_n \right\|_2 \right)^2}}{\sqrt{\sum\limits_{j=1}^{q} \left(e^{(r_j - r_1)t_n} \left\| Q_j \widehat{y}_0 \right\|_2 \right)^2}}$$

where

$$\widehat{Z}_n = \frac{\widetilde{Y}_n - y_0}{\left\| \widetilde{Y}_n - y_0 \right\|_2}, \qquad E_n = \frac{\left\| \widetilde{Y}_n - y_0 \right\|_2}{\| y_0 \|_2}.$$

By writing the perturbed initial value $\widetilde{y}_0$ as

$$\widetilde{y}_0 = y_0 + \varepsilon \, \| y_0 \|_2 \, \widehat{z}_0$$

we have

$$\widetilde{Y}_n - y_0 = S\,(hA)^n \, \widetilde{y}_0 - y_0 = \left(S\,(hA)^n - I \right) y_0 + \varepsilon \, \| y_0 \|_2 \, S\,(hA)^n \, \widehat{z}_0.$$

So

$$\begin{aligned} E_n \widehat{Z}_n &= \left(S(hA)^n - I \right) \widehat{y}_0 + \varepsilon S\,(hA)^n \, \widehat{z}_0 \\ &= \sum_{i=1}^{p} \left(\left(S\,(h\lambda_i)^n - 1 \right) P_i \widehat{y}_0 + \varepsilon S\,(h\lambda_i)^n \, P_i \widehat{z}_0 \right) \end{aligned}$$

and then, for $j = 1, 2, \ldots, q$,

$$Q_j E_n \widehat{Z}_n = \sum_{\lambda_i \in \Lambda_j} \left(\left(S\,(h\lambda_i)^n - 1 \right) P_i \widehat{y}_0 + \varepsilon S\,(h\lambda_i)^n \, P_i \widehat{z}_0 \right).$$

This completes the proof. □

The errors $\varepsilon_{n,j}$ in (2.1) are defined in (2.2) in terms of two sources of error: the number ε, which takes into account the fact that the initial data is perturbed, and the function S, which takes into account the fact that we are numerically integrating (1.1) and not solving it exactly.

When $\varepsilon = 0$ the initial data is not perturbed and when

$$S(z) = 1, \quad z \in \mathcal{D} = \mathbb{C} \tag{2.3}$$

the ODE (1.1) is exactly solved.

Of course, (2.3) never holds when we are numerically integrating (1.1) and it holds when $R(z) = \mathrm{e}^z$, $z \in \mathcal{D} = \mathbb{C}$, i.e., when we use

$$\widetilde{y}_{n+1} = \mathrm{e}^{hA} \widetilde{y}_n, \quad n = 0, 1, 2, \ldots$$

instead of (1.9), namely we use the matrix exponential e^{hA} instead of an its approximation. We can think that this is implemented in MATLAB by the matrix exponential function *expm*. In the numerical tests of Section 6 we compute exact solutions in this manner.

The situation $\varepsilon = 0$, where the initial value is not perturbed and the sole source of error is the approximant of the exponential, is of particular interest: we have already observed above that a systematic analysis of the relative error of numerical solutions of ODEs has not been developed in literature. In this situation, the error δ_n becomes the error γ_n defined in (1.12).

The situation (2.3), where the ODE (1.1) is exactly solved and the sole source of error is the perturbation in the initial data, has been already studied in [9]. In this situation, the error δ_n becomes the error $\delta(t_n)$ defined in (1.3).

3 The errors of the approximant

We can rewrite the error $\varepsilon_{n,j}$ in (2.2) as

$$\varepsilon_{n,j} := \left\| \sum_{\lambda_i \in \Lambda_j} \left(\varphi\left(n\sigma_i\right) P_i \widehat{y}_0 + \varepsilon\left(1 + \varphi\left(n\sigma_i\right)\right) P_i \widehat{z}_0 \right) \right\|_2 \tag{3.1}$$

where

$$\varphi\left(n\sigma_i\right) := \mathrm{e}^{n\sigma_i} - 1$$

and

$$\sigma_i := \log S(h\lambda_i). \tag{3.2}$$

Here, we are considering the principal value of the complex logarithm log, which is the branch defined by the Mercator series

$$\log z = \sum_{i=1}^{\infty} \frac{(-1)^i}{i!} (z-1)^i, \quad |z-1| < 1.$$

The complex numbers σ_i, $i = 1, \dots, p$, defined in (3.2) are considered as the errors introduced by the approximant R: observe that in (2.2) we have $S(h\lambda_i) = 1$ if and only if $\sigma_i = 0$.

The next proposition says how small is the error σ_i when $h\lambda_i$ is small.

Proposition 3.1. For $i = 1, \dots, p$, we have

$$\sigma_i = C\left(h\lambda_i\right)^{l+1}\left(1 + O\left(h\lambda_i\right)\right), \quad h\lambda_i \to 0 \tag{3.3}$$

where l is the order of the approximant and C is the constant in (1.10).

Proof. By (1.10) we obtain

$$S(z) = 1 + Cz^{l+1} + O\left(z^{l+2}\right), \quad z \to 0$$

and then

$$\log S(z) = \log\left(1 + Cz^{l+1} + O\left(z^{l+2}\right)\right) = Cz^{l+1} + O\left(z^{l+2}\right), \quad z \to 0.$$

This completes the proof. □

Next remark contains important observations or consequences of Proposition 3.1. In the following, a similar remark is presented for each proposition or theorem.

Remark 3.1.

1. If $\lambda_i = 0$, then $\sigma_i = 0$.
2. If the approximant R has real coefficients, then, for a pair λ_i, λ_k of complex conjugate eigenvalues of A, σ_i and σ_k are complex conjugate and so $|\sigma_i| = |\sigma_k|$.
3. Let Γ be a nonempty subset of the spectrum Λ of A and let

$$\rho_\Gamma := \max_{\lambda_i \in \Gamma} |\lambda_i|$$
$$\mu_\Gamma := \min_{\lambda_i \in \Gamma} |\lambda_i|.$$

We have

$$\max_{\lambda_i \in \Gamma} |\sigma_i| = |C|\left(h\rho_\Gamma\right)^{l+1}\left(1 + O\left(h\rho_\Gamma\right)\right)$$
$$\min_{\lambda_i \in \Gamma} |\sigma_i| = |C|\left(h\mu_\Gamma\right)^{l+1}\left(1 + O\left(h\rho_\Gamma\right)\right)$$

as $h\rho_\Gamma \to 0$.

Let Γ and Δ nonempty subsets of Λ and let

$$K_{\Gamma\Delta} := \frac{\max\limits_{\lambda_i \in \Gamma} |\sigma_i|}{\min\limits_{\lambda_i \in \Delta} |\sigma_i|}. \tag{3.4}$$

We have

$$K_{\Gamma\Delta} = \left(\frac{\rho_\Gamma}{\mu_\Delta}\right)^{l+1} \left(1 + O\left(h\rho_{\Gamma\cup\Delta}\right)\right), \quad h\rho_{\Gamma\cup\Delta} \to 0.$$

4. In the previous point 3, we prefer to write $h\rho_\Gamma \to 0$ and $h\rho_{\Gamma\cup\Delta} \to 0$ rather than $h \to 0$. We use the dimensionless stepsizes $h\rho_\Gamma$ and $h\rho_{\Gamma\cup\Delta}$ rather than the stepsize h, because $h\rho_\Gamma$ and $h\rho_{\Gamma\cup\Delta}$ are small or large independently of the particular unit used for the time t.

The errors σ_i appear in the errors $\varepsilon_{n,j}$ by means of $\varphi\left(n\sigma_i\right)$ (see (3.1)). Next proposition says something about this.

Proposition 3.2. Let $i = 1, \ldots, p$ and let $c \geqslant 0$. If $n\left|\sigma_i\right| \leqslant c$, then

$$\varphi(n\sigma_i) = n\sigma_i\left(1 + w_i\right)$$

where

$$|w_i| \leqslant g(c) := \frac{e^c - 1 - c}{c}.$$

Proof. Let $z \in \mathbb{C}$. For

$$\varphi(z) = e^z - 1$$

we have

$$\varphi(z) = z\left(1 + w\right)$$

where

$$|w| \leqslant \frac{e^c - 1 - c}{c}$$

whenever $|z| \leqslant c$. In fact, we have

$$\varphi(z) = z\left(1 + \frac{z}{2!} + \frac{z^2}{3!} + \frac{z^3}{4!} + \cdots\right)$$

with

$$\left|\frac{z}{2!} + \frac{z^2}{3!} + \frac{z^3}{4!} + \cdots\right| \leqslant \frac{e^c - 1 - c}{c}.$$

This completes the proof. □

Remark 3.2.

1. The function

$$g(c) = \frac{e^c - 1 - c}{c}, \quad c \geqslant 0$$

is increasing and we have

$$g(c) = \frac{c}{2} + O(c^2), \quad c \to 0.$$

Since the function g is often used throughout the paper, some reference values of g are collected in Table 1.

2. If $n|\sigma_i| \ll 1$, then $\varphi(n\sigma_i) \approx n\sigma_i$.
3. Remind point 3 in Remark 3.1. Let Γ be a nonempty subset of Λ. We can write

$$n \max_{\lambda_i \in \Gamma} |\sigma_i| = t_n \rho_\Gamma E_\Gamma$$
$$n \min_{\lambda_i \in \Gamma} |\sigma_i| = t_n \rho_\Gamma F_\Gamma$$

c	$g(c)$
0.5	0.29744
1	0.71828
1.5	1.3211
2	2.1945

Tab. 1: Reference values of g. Value $g(c) = 1$ is for $c = 1.2564$.

where

$$
\begin{aligned}
E_\Gamma &:= \frac{\max\limits_{\lambda_i \in \Gamma} |\sigma_i|}{h\rho_\Gamma} = |C| \left(h\rho_\Gamma\right)^l \left(1 + O\left(h\rho_\Gamma\right)\right) \\
F_\Gamma &:= \frac{\min\limits_{\lambda_i \in \Gamma} |\sigma_i|}{h\rho_\Gamma} = \frac{1}{K_{\Gamma\Gamma}} E_\Gamma = \left(\frac{\mu_\Gamma}{\rho_\Gamma}\right)^{l+1} |C| \left(h\rho_\Gamma\right)^l \left(1 + O\left(h\rho_\Gamma\right)\right)
\end{aligned}
$$

as $h\rho_\Gamma \to 0$.

4. Remind point 4 in Remark 3.1. In the previous point 3, we use the dimensionless times $t_n\rho_\Gamma$ rather than the time t_n, because $t_n\rho_\Gamma$ is small or large independently of the particular unit used for the time.

4 Analysis of the error δ_n

In the present section, we study how the relative error δ_n in (1.11) grows with the index n. We consider separately the situation of an unperturbed initial value, i.e., $\varepsilon = 0$, and of a perturbed initial value.

The following notation is introduced. Let

$$
\begin{aligned}
\Lambda^* &:= \{\lambda_i \in \Lambda : P_i y_0 \neq 0\} \\
\Lambda^{**} &:= \left\{\lambda_i \in \Lambda : P_i \widehat{z}_0 \neq 0\right\}
\end{aligned}
$$

and, for $j = 1, \ldots, q$,

$$
\begin{aligned}
\Lambda_j^* &:= \left\{\lambda_i \in \Lambda_j : P_i y_0 \neq 0\right\} \\
\Lambda_j^{**} &:= \left\{\lambda_i \in \Lambda_j : P_i \widehat{z}_0 \neq 0\right\}.
\end{aligned}
$$

Moreover, let

$$
\begin{aligned}
j^* &:= \min\left\{j \in \{1, \ldots, q\} : \Lambda_j^* \neq \emptyset\right\} \\
j^{**} &:= \min\left\{j \in \{1, \ldots, q\} : \Lambda_j^{**} \neq \emptyset\right\}.
\end{aligned}
$$

The generic situation is $\Lambda^* = \Lambda^{**} = \Lambda$, $\Lambda_j^* = \Lambda_j^{**} = \Lambda_j$ for $j = 1, \ldots, q$ and $j^* = j^{**} = 1$.

4.1 Unperturbed initial value

Next theorem gives lower and upper bounds for the relative error γ_n given in (1.12) and relevant to an unperturbed initial value.

Theorem 4.1. *Fix $c \geqslant 0$. For an index n such that*

$$
n \max_{\lambda_i \in \Lambda^*} |\sigma_i| \leqslant c \tag{4.1}
$$

we have

$$
n \min_{\lambda_i \in \Lambda^*} |\sigma_i| \left(1 - g(c)\right) \leqslant \gamma_n \leqslant n \max_{\lambda_i \in \Lambda^*} |\sigma_i| \left(1 + g(c)\right). \tag{4.2}
$$

Proof. In the situation $\varepsilon = 0$, for $n = 0, 1, 2, \ldots$ and $j = 1, \ldots, q$, we have

$$\varepsilon_{n,j} = \left\| \sum_{\lambda_i \in \Lambda_j} \varphi\left(n\sigma_i\right) P_i \widehat{y}_0 \right\|_2 = \sqrt{\sum_{\lambda_i \in \Lambda_j} \left|\varphi\left(n\sigma_i\right)\right|^2 \left\|P_i \widehat{y}_0\right\|_2^2}. \tag{4.3}$$

By (2.1), (4.3), (4.1) and Proposition 3.2, we easily obtain (4.2). □

The theorem says that *the error* γ_n *grows linearly with the index* n*, i.e., it grows linearly in time*: see point 3 in Remark 4.1 below.

Remark 4.1.

1. For an index n such that
$$n \max_{\lambda_i \in \Lambda^*} |\sigma_i| \ll 1$$
we have
$$n \min_{\lambda_i \in \Lambda^*} |\sigma_i| \lessapprox \gamma_n \lessapprox n \max_{\lambda_i \in \Lambda^*} |\sigma_i| .$$
2. Recall point 3 in Remark 3.2. In (4.1) and (4.2) we have
$$n \max_{\lambda_i \in \Lambda^*} |\sigma_i| = t_n \rho_{\Lambda^*} E_{\Lambda^*}$$
$$n \min_{\lambda_i \in \Lambda^*} |\sigma_i| = t_n \rho_{\Lambda^*} F_{\Lambda^*}$$
where
$$E_{\Lambda^*} = |C| \left(h\rho_{\Lambda^*}\right)^l \left(1 + O\left(h\rho_{\Lambda^*}\right)\right)$$
$$F_{\Lambda^*} = \left(\frac{\mu_{\Lambda^*}}{\rho_{\Lambda^*}}\right)^{l+1} |C| \left(h\rho_{\Lambda^*}\right)^l \left(1 + O\left(h\rho_{\Lambda^*}\right)\right)$$
as $h\rho_{\Lambda^*} \to 0$.
3. Fix $c > 0$. The Theorem 4.1 and point 2 above say that for a dimensionless time $t_n\rho_{\Lambda^*}$ such that
$$t_n \rho_{\Lambda^*} \leqslant \tau$$
where
$$\tau := \frac{c}{E_{\Lambda^*}} \to +\infty, \qquad \frac{1}{\tau} = O\left(\left(h\rho_{\Lambda^*}\right)^l\right) \quad \text{as} \;\; h\rho_{\Lambda^*} \to 0 \tag{4.4}$$
we have
$$b t_n \rho_{\Lambda^*} \leqslant \gamma_n \leqslant a t_n \rho_{\Lambda^*} \tag{4.5}$$
with a and b independent of n and
$$a = O(\left(h\rho_{\Lambda^*}\right)^l), \quad b = O\left(\left(h\rho_{\Lambda^*}\right)^l\right), \quad h\rho_{\Lambda^*} \to 0.$$
Hence, *the relative error* γ_n *grows linearly in time and this linear growth holds for any normal matrix* A*, for any approximant* R *and for any stepsize* h (such that $h\lambda_i \in \mathcal{D}$, $i = 1, \ldots, p$). Of course, the absolute error
$$\left\|y_n - y\left(t_n\right)\right\|_2 = \gamma_n \left\|y\left(t_n\right)\right\|_2$$
has a completely different behavior, due to the exponential growth or decrease in time of $\left\|y\left(t_n\right)\right\|_2$. Observe that the linear growth of γ_n is not a true linear growth, since (4.5) is not guaranteed to be valid for all dimensionless times $t_n\rho_{\Lambda^*}$, but only for $t_n\rho_{\Lambda^*} \leqslant \tau$. However, asymptotically as $h\rho_{\Lambda^*} \to 0$, it is valid for all dimensionless times $t_n\rho_{\Lambda^*}$: we have $\tau \to +\infty$ as $h\rho_{\Lambda^*} \to 0$.
4. When $\Lambda^* = \{0\}$, we have
$$\gamma_n = 0, \quad n = 0, 1, 2, \ldots$$

5. When $\Lambda^* \neq \{0\}$, Theorem 4.1 can be improved as follows. Fix $c \geqslant 0$. For an index n such that

$$n \max_{\lambda_i \in \Lambda^*} |\sigma_i| \leqslant c$$

we have

$$n \min_{\lambda_i \in \Lambda^* \setminus \{0\}} |\sigma_i| \left(1 - g(c)\right) B_n \leqslant \gamma_n \leqslant n \max_{\lambda_i \in \Lambda^*} |\sigma_i| \left(1 + g(c)\right) B_n$$

where

$$B_n = \frac{\sqrt{\sum\limits_{j=j^*}^{q} \left(\mathrm{e}^{(r_j - r_1)t_n} \sqrt{\sum\limits_{\lambda_i \in \Lambda_j^* \setminus \{0\}} \left\| P_i \widehat{y}_0 \right\|_2^2} \right)^2}}{\sqrt{\sum\limits_{j=j^*}^{q} \left(\mathrm{e}^{(r_j - r_1)t_n} \left\| Q_j \widehat{y}_0 \right\|_2 \right)^2}}$$

and

$$\mathrm{e}^{\left(r_{j_1^*} - r_{j^*}\right)t_n} \sqrt{\sum_{\lambda_i \in \Lambda_{j_1}^* \setminus \{0\}} \left\| P_i \widehat{y}_0 \right\|_2^2} \leqslant B_n \leqslant 1$$

holds with

$$j_1^* := \min \left\{ j \in \{j^*, \ldots, q\} : \Lambda_j^* \setminus \{0\} \neq \varnothing \right\}. \tag{4.6}$$

This result, unlike Theorem 4.1, gives a nonzero lower bound for γ_n in the case $0 \in \Lambda^*$ and $\Lambda^* \neq \{0\}$.

6. The bounds (4.2) are valid for indices n satisfying (4.1). Bounds valid for all indices n are

$$\min_{\lambda_i \in \Lambda^*} \left| \varphi\left(n\sigma_i\right) \right| \leqslant \gamma_n \leqslant \max_{\lambda_i \in \Lambda^*} \left| \varphi\left(n\sigma_i\right) \right|.$$

However, it is not interesting consider bounds valid for all indices n because, for a sufficiently large n, the error γ_n becomes not small (although it could become again small or even zero for a very large n if all the complex numbers σ_i, $\lambda_i \in \Lambda^*$, are imaginary).

4.2 Perturbed initial value

Recall that γ_n (see (1.12) and the previous subsection) is the relative error due to the sole approximant and that $\delta\left(t\right)$ (see (1.3) and (1.4)) is the relative error due to the sole perturbation in the initial value. For a perturbed initial value, next theorem describes the growth of the relative error δ_n (see (1.11)) due to the approximant and the perturbation in the initial value, in terms of the two relative errors γ_n and $\delta\left(t_n\right)$.

Theorem 4.2. *For $n = 0, 1, 2, \ldots$, we have*

$$\left|\delta_n - \gamma_n\right| \leqslant \delta\left(t_n\right) + \beta_n \delta\left(t_n\right) \tag{4.7}$$

and

$$\left|\delta_n - \delta\left(t_n\right)\right| \leqslant \gamma_n + \beta_n \delta\left(t_n\right) \tag{4.8}$$

where

$$\beta_n := \frac{\sqrt{\sum\limits_{j=1}^{q} \left(\mathrm{e}^{(r_j - r_1)t_n} \left\| \sum\limits_{\lambda_i \in \Lambda_j} \varphi\left(n\sigma_i\right) P_i \widehat{z}_0 \right\|_2 \right)^2}}{\sqrt{\sum\limits_{j=1}^{q} \left(\mathrm{e}^{(r_j - r_1)t_n} \left\| Q_j \widehat{z}_0 \right\|_2 \right)^2}} \tag{4.9}$$

is the error γ_n when the initial value y_0 of (1.1) is such that $\widehat{y}_0 = \widehat{z}_0$.

Proof. For $n = 0, 1, 2, \ldots$, we have

$$\begin{aligned}
|\delta_n - \gamma_n| \sqrt{\sum_{j=1}^{q} \left(\mathrm{e}^{(r_j-r_1)t_n} \left\|Q_j\widehat{y}_0\right\|_2\right)^2} &= \left| \sqrt{\sum_{j=1}^{q} \left(\mathrm{e}^{(r_j-r_1)t_n} \varepsilon_{n,j}\right)^2} - \sqrt{\sum_{j=1}^{q} \left(\mathrm{e}^{(r_j-r_1)t_n} \left\| \sum_{\lambda_i \in \Lambda_j} \varphi\left(n\sigma_i\right) P_i \widehat{y}_0 \right\|_2\right)^2} \right| \\
&\leqslant \sqrt{\sum_{j=1}^{q} \left(\mathrm{e}^{(r_j-r_1)t_n} \left(\varepsilon_{n,j} - \left\| \sum_{\lambda_i \in \Lambda_j} \varphi\left(n\sigma_i\right) P_i \widehat{y}_0 \right\|_2\right)\right)^2} \\
&\leqslant \sqrt{\sum_{j=1}^{q} \left(\mathrm{e}^{(r_j-r_1)t_n} \left\| \sum_{\lambda_i \in \Lambda_j} \varepsilon\left(1 + \varphi\left(n\sigma_i\right)\right) P_i \widehat{z}_0 \right\|_2\right)^2} \\
&\leqslant \sqrt{\sum_{j=1}^{q} \left(\mathrm{e}^{(r_j-r_1)t_n} \left(\left\| \sum_{\lambda_i \in \Lambda_j} \varepsilon P_i \widehat{z}_0 \right\|_2 + \left\| \sum_{\lambda_i \in \Lambda_j} \varepsilon \varphi\left(n\sigma_i\right) P_i \widehat{z}_0 \right\|_2\right)\right)^2} \\
&\leqslant \sqrt{\sum_{j=1}^{q} \left(\mathrm{e}^{(r_j-r_1)t_n} \left\| \sum_{\lambda_i \in \Lambda_j} \varepsilon P_i \widehat{z}_0 \right\|_2\right)^2} \\
&\quad + \sqrt{\sum_{j=1}^{q} \left(\mathrm{e}^{(r_j-r_1)t_n} \left\| \sum_{\lambda_i \in \Lambda_j} \varepsilon \varphi\left(n\sigma_i\right) P_i \widehat{z}_0 \right\|_2\right)^2}.
\end{aligned}$$

The bound (4.7) now easily follows.

For the other bound (4.8), the proof proceeds similarly: for $n = 0, 1, 2, \ldots$, we have

$$\begin{aligned}
\left|\delta_n - \delta\left(t_n\right)\right| \sqrt{\sum_{j=1}^{q} \left(\mathrm{e}^{(r_j-r_1)t_n} \left\|Q_j\widehat{y}_0\right\|_2\right)^2} &= \left| \sqrt{\sum_{j=1}^{q} \left(\mathrm{e}^{(r_j-r_1)t_n} \varepsilon_{n,j}\right)^2} - \sqrt{\sum_{j=1}^{q} \left(\mathrm{e}^{(r_j-r_1)t_n} \left\|\varepsilon Q_j\widehat{z}_0\right\|_2\right)^2} \right| \\
&\leqslant \sqrt{\sum_{j=1}^{q} \left(\mathrm{e}^{(r_j-r_1)t_n} \left\| \sum_{\lambda_i \in \Lambda_j} \left(\varphi\left(n\sigma_i\right) P_i \widehat{y}_0 + \varepsilon \varphi\left(n\sigma_i\right) P_i \widehat{z}_0\right) \right\|_2\right)^2} \\
&\leqslant \sqrt{\sum_{j=1}^{q} \left(\mathrm{e}^{(r_j-r_1)t_n} \left\| \sum_{\lambda_i \in \Lambda_j} \varphi\left(n\sigma_i\right) P_i \widehat{y}_0 \right\|_2\right)^2} \\
&\quad + \sqrt{\sum_{j=1}^{q} \left(\mathrm{e}^{(r_j-r_1)t_n} \left\| \sum_{\lambda_i \in \Lambda_j} \varepsilon \varphi\left(n\sigma_i\right) P_i \widehat{z}_0 \right\|_2\right)^2}.
\end{aligned}$$

This completes the proof. □

The bound (4.7) shows how a perturbation in the initial value affects the error γ_n and the bound (4.8) shows how the numerical integration affects the error $\delta(t_n)$.

Remark 4.2.

1. Fix $c > 0$. For an index n such that

   $$n \max_{\lambda_i \in \Lambda^{**}} |\sigma_i| \leqslant c$$

   we have

   $$n \min_{\lambda_i \in \Lambda^{**}} |\sigma_i| \left(1 - g(c)\right) \leqslant \beta_n \leqslant n \max_{\lambda_i \in \Lambda^{**}} |\sigma_i| \left(1 + g(c)\right).$$

   Moreover, for an index n such that

   $$n \max_{\lambda_i \in \Lambda^{**}} |\sigma_i| \ll 1$$

we have

$$n \min_{\lambda_i \in \Lambda^{**}} |\sigma_i| \lessapprox \beta_n \lessapprox n \max_{\lambda_i \in \Lambda^{**}} |\sigma_i| \,.$$

This follows by Theorem 4.1.

2. For an index n such that

$$\beta_n \ll 1 \tag{4.10}$$

we have

$$\left|\delta_n - \gamma_n\right| \lessapprox \delta\left(t_n\right)$$

by (4.7). The condition (4.10) holds when

$$n \max_{\lambda_i \in \Lambda^{**}} |\sigma_i| \ll 1.$$

3. Assume $\Lambda^* \neq \{0\}$. For an index n such that

$$\frac{\beta_n \delta\left(t_n\right)}{\gamma_n} = \frac{\sqrt{\sum\limits_{j=1}^{q} \left(\mathrm{e}^{\left(r_j - r_1\right)t_n} \left\| \sum\limits_{\lambda_i \in \Lambda_j} \varphi\left(n\sigma_i\right) P_i \widehat{z}_0 \right\|_2 \right)^2}}{\sqrt{\sum\limits_{j=1}^{q} \left(\mathrm{e}^{\left(r_j - r_1\right)t_n} \left\| \sum\limits_{\lambda_i \in \Lambda_j} \varphi\left(n\sigma_i\right) P_i \widehat{y}_0 \right\|_2 \right)^2}} \varepsilon \ll 1 \tag{4.11}$$

we have

$$\left|\delta_n - \delta\left(t_n\right)\right| \lessapprox \gamma_n$$

by (4.8). The condition (4.11) holds when

$$\Lambda^{**} \subseteq \Lambda^*, \quad \max_{\lambda_i \in \Lambda^{**}} \frac{\left\|P_i \widehat{z}_0\right\|_2}{\left\|P_i \widehat{y}_0\right\|_2} \cdot \varepsilon \ll 1$$

or

$$0 \notin \Lambda^*_{j^*}, \quad n \max\nolimits_{\lambda_i \in \Lambda^*_{j^*}} |\sigma_i| \ll 1, \quad n \max\nolimits_{\lambda_i \in \Lambda^{**}} |\sigma_i| \ll 1$$
$$K_{\Lambda^{**}\Lambda^*_{j^*}} \tfrac{\mathrm{e}^{\left(r_{j^{**}} - r_{j^*}\right)t_n}}{\|Q_{j^*} \widehat{y}_0\|_2} \varepsilon \ll 1$$

where $K_{\Lambda^{**}\Lambda^*_{j^*}}$ is defined in (3.4).

4. Assume $\Lambda^{**} \neq \{0\}$. For an index n such that

$$\frac{\gamma_n}{\beta_n \delta\left(t_n\right)} = \frac{\sqrt{\sum\limits_{j=1}^{q} \left(\mathrm{e}^{\left(r_j - r_1\right)t_n} \left\| \sum\limits_{\lambda_i \in \Lambda_j} \varphi\left(n\sigma_i\right) P_i \widehat{y}_0 \right\|_2 \right)^2}}{\sqrt{\sum\limits_{j=1}^{q} \left(\mathrm{e}^{\left(r_j - r_1\right)t_n} \left\| \sum\limits_{\lambda_i \in \Lambda_j} \varphi\left(n\sigma_i\right) P_i \widehat{z}_0 \right\|_2 \right)^2} \varepsilon} \ll 1 \tag{4.12}$$

we have

$$\frac{\left|\delta_n - \delta\left(t_n\right)\right|}{\delta\left(t_n\right)} \lessapprox \beta_n$$

by (4.8). The condition (4.12) holds when

$$0 \notin \Lambda^{**}_{j^{**}}, \quad n \max\nolimits_{\lambda_i \in \Lambda^*} |\sigma_i| \ll 1, \quad n \max\nolimits_{\lambda_i \in \Lambda^{**}} |\sigma_i| \ll 1$$
$$\frac{K_{\Lambda^*\Lambda^{**}_{j^{**}}}}{\mathrm{e}^{\left(r_{j^{**}} - r_{j^*}\right)t_n} \|Q_{j^{**}} \widehat{z}_0\|_2 \varepsilon} \ll 1.$$

5. Recall point 4. For an index n such that (4.10) and (4.12) hold we have $\delta_n \approx \delta(t_n)$.

6. By (4.7) or (4.8), we have

$$\delta_n \leqslant \gamma_n + \delta\left(t_n\right) + \beta_n \delta\left(t_n\right), \quad n = 0, 1, 2, \ldots$$

For an index n such that (4.10) holds we have

$$\delta_n \lessapprox \gamma_n + \delta\left(t_n\right).$$

7. Recall point 6. If $\Lambda^* = \{\lambda_i\}$, $\varphi(n\sigma_i) \geqslant 0$, and $\widehat{z}_0 = \widehat{y}_0$, then

$$\delta_n = \gamma_n + \delta\left(t_n\right) + \beta_n \delta\left(t_n\right), \quad n = 0, 1, 2, \ldots$$

5 Long-time behavior

In the present section, we are interested in the behavior of the error δ_n for large indices n. As in the previous section, we consider separately the situations of an unperturbed initial value and of a perturbed initial value.

5.1 Unperturbed initial value

5.1.1 The long-time solution y^{long}

Let y^{long} be the solution of (1.1) with initial value $Q_{j^*} y_0$ rather than y_0. This solution is a *long-time solution* as stated in the following theorem.

Theorem 5.1. *For a time* $t \geqslant 0$ *such that*

$$\sqrt{\sum_{j=j^*+1}^{q} \left(\mathrm{e}^{(r_j - r_{j^*})t} \frac{\left\| Q_j \widehat{y}_0 \right\|_2}{\left\| Q_{j^*} \widehat{y}_0 \right\|_2} \right)^2} \ll 1 \tag{5.1}$$

we have

$$y(t) \approx y^{\mathrm{long}}(t).$$

Proof. For $t \geqslant 0$, we have

$$y\left(t\right) = \sum_{j=j^*}^{q} \sum_{\lambda_i \in \Lambda_j} \mathrm{e}^{\lambda_i t} P_i y_0$$

and

$$y^{\mathrm{long}}\left(t\right) = \sum_{\lambda_i \in \Lambda_{j^*}} \mathrm{e}^{\lambda_i t} P_i y_0.$$

Then

$$\frac{\left\| y\left(t\right) - y^{\mathrm{long}}\left(t\right) \right\|_2^2}{\left\| y^{\mathrm{long}}\left(t\right) \right\|_2^2} = \frac{\left\| \sum\limits_{j=j^*+1}^{q} \sum\limits_{\lambda_i \in \Lambda_j} \mathrm{e}^{\lambda_i t} P_i y_0 \right\|_2^2}{\left\| \sum\limits_{\lambda_i \in \Lambda_{j^*}} \mathrm{e}^{\lambda_i t} P_i y_0 \right\|_2^2} = \frac{\sum\limits_{j=j^*+1}^{q} \left(\mathrm{e}^{r_j t} \left\| Q_j y_0 \right\|_2 \right)^2}{\left(\mathrm{e}^{r_{j^*} t} \left\| Q_{j^*} y_0 \right\|_2 \right)^2} = \sum_{j=j^*+1}^{q} \left(\mathrm{e}^{(r_j - r_{j^*})t} \frac{\left\| Q_j y_0 \right\|_2}{\left\| Q_{j^*} y_0 \right\|_2} \right)^2.$$

This completes the proof. □

Remark 5.1.

1. Theorem 5.1 says that

$$\frac{\left\| y\left(t\right) - y^{\mathrm{long}}\left(t\right) \right\|_2}{\left\| y^{\mathrm{long}}\left(t\right) \right\|_2} \ll 1 \quad \text{for a sufficiently large } t.$$

In addition, by looking at the proof of Theorem 5.1, we see that

$$\frac{\left\| y\left(t\right) - y^{\mathrm{long}}\left(t\right) \right\|_2}{\left\| y^{\mathrm{long}}\left(t\right) \right\|_2} \to 0, \quad t \to +\infty.$$

2. For $\Lambda^* \setminus \Lambda^*_{j^*} \neq \varnothing$, the condition (5.1) holds when

$$\mathrm{e}^{(r_{m^*}-r_{j^*})t_n} \frac{\sqrt{1-\|Q_{j^*}\widehat{y}_0\|_2^2}}{\|Q_{j^*}\widehat{y}_0\|_2} \ll 1$$

where

$$m^* := \min\{j \in \{j^*+1,\ldots,q\} : \Lambda^*_j \neq \varnothing\}. \tag{5.2}$$

5.1.2 The error γ_n^{long}

Let γ_n^{long} be the error γ_n relevant to the long-time solution y^{long}. By considering the formula (2.1) with initial value $Q_{j^*}y_0$ rather than y_0, we obtain

$$\gamma_n^{\mathrm{long}} = \frac{\varepsilon_{n,j^*}}{\|Q_{j^*}\widehat{y}_0\|_2}, \quad n = 0,1,2,\ldots$$

with

$$\varepsilon_{n,j^*} = \left\| \sum_{\lambda_i \in \Lambda^*_{j^*}} \varphi(n\sigma_i) P_i \widehat{y}_0 \right\|_2 .$$

Observe that $\varepsilon_{n,j^*}/\|Q_{j^*}\widehat{y}_0\|_2$ is also obtained by considering, in the formula (2.1) with initial value y_0, only the leading exponential terms $\mathrm{e}^{(r_{j^*}-r_1)t_n}\varepsilon_{n,j^*}$ at the numerator and $\mathrm{e}^{(r_{j^*}-r_1)t_n}\|Q_{j^*}\widehat{y}_0\|_2$ at the denominator.

Next theorem gives lower and upper bounds for the error γ_n^{long}.

Theorem 5.2. *Fix $c \geqslant 0$. For an index n such that*

$$n \max_{\lambda_i \in \Lambda^*_{j^*}} |\sigma_i| \leqslant c \tag{5.3}$$

we have

$$n \min_{\lambda_i \in \Lambda^*_{j^*}} |\sigma_i| (1-g(c)) \leqslant \gamma_n^{\mathrm{long}} \leqslant n \max_{\lambda_i \in \Lambda^*_{j^*}} |\sigma_i| (1+g(c)) . \tag{5.4}$$

Theorem 5.2 is identical to Theorem 4.1 excepts for Λ^* replaced with its subset $\Lambda^*_{j^*}$ constituted by the rightmost eigenvalue in Λ^*. The proof proceeds similarly to the proof of Theorem 4.1.

Remark 5.2.

1. For an index n such that

$$n \max_{\lambda_i \in \Lambda^*_{j^*}} |\sigma_i| \ll 1$$

we have

$$n \min_{\lambda_i \in \Lambda^*_{j^*}} |\sigma_i| \lessapprox \gamma_n^{\mathrm{long}} \lessapprox n \max_{\lambda_i \in \Lambda^*_{j^*}} |\sigma_i| .$$

2. Recall point 2 in Remark 3.1. If $\Lambda^*_{j^*}$ is constituted by a single real eigenvalue λ_i or by a single pair $\lambda_i, \overline{\lambda}_i$ of complex conjugate eigenvalues and the approximant R has real coefficients, then

$$n|\sigma_i|(1-g(c)) \leqslant \gamma_n^{\mathrm{long}} \leqslant n|\sigma_i|(1+g(c))$$

whenever

$$n|\sigma_i| \leqslant c$$

and

$$\gamma_n^{\mathrm{long}} \approx n|\sigma_i|$$

whenever

$$n|\sigma_i| \ll 1.$$

3. Recall point 3 in Remark 3.2. In (5.3) and (5.4) we can write

$$n \max_{\lambda_i \in \Lambda^*_{j^*}} |\sigma_i| = t_n \rho_{\Lambda^*_{j^*}} E_{\Lambda^*_{j^*}}$$

$$n \min_{\lambda_i \in \Lambda^*_{j^*}} |\sigma_i| = t_n \rho_{\Lambda^*_{j^*}} F_{\Lambda^*_{j^*}}$$

   where

$$E_{\Lambda^*_{j^*}} = |C| \left(h\rho_{\Lambda^*_{j^*}} \right)^l \left(1 + O\left(h\rho_{\Lambda^*_{j^*}} \right) \right)$$

$$F_{\Lambda^*_{j^*}} = \left(\frac{\mu_{\Lambda^*_{j^*}}}{\rho_{\Lambda^*_{j^*}}} \right)^{l+1} |C| \left(h\rho_{\Lambda^*_{j^*}} \right)^l \left(1 + O\left(h\rho_{\Lambda^*_{j^*}} \right) \right)$$

   as $h\rho_{\Lambda^*_{j^*}} \to 0$.

4. If $\Lambda^*_{j^*} = \{0\}$, then

$$\gamma_n^{\text{long}} = 0, \quad n = 0, 1, 2, \ldots$$

5. When $\Lambda^*_{j^*} \neq \{0\}$, Theorem 5.2 can be improved as follows. Fix $c \geqslant 0$. For an index n such that

$$n \max_{\lambda_i \in \Lambda^*_{j^*}} |\sigma_i| \leqslant c$$

   we have

$$n \min_{\lambda_i \in \Lambda^*_{j^*} \setminus \{0\}} |\sigma_i| \left(1 - g(c)\right) B^{\text{long}} \leqslant \gamma_n^{\text{long}} \leqslant n \max_{\lambda_i \in \Lambda^*_{j^*}} |\sigma_i| \left(1 + g(c)\right) B^{\text{long}}$$

   where

$$B^{\text{long}} = \frac{\sqrt{\sum\limits_{\lambda_i \in \Lambda^*_{j^*} \setminus \{0\}} \left\| P_i \widehat{y}_0 \right\|_2^2}}{\left\| Q_{j^*} \widehat{y}_0 \right\|_2} \leqslant 1.$$

   This result, unlike Theorem 5.2, gives a nonzero lower bound for γ_n^{long} in the case $0 \in \Lambda^*_{j^*}$ and $\Lambda^*_{j^*} \neq \{0\}$.

5.1.3 Long-time behavior of the error γ_n

Since $y(t) \approx y^{\text{long}}(t)$ holds in the long-time and the error γ_n for y^{long} is γ_n^{long}, an important question is:

$$\text{Does } \gamma_n \approx \gamma_n^{\text{long}} \text{ hold in the long-time?} \tag{5.5}$$

Regarding the meaning of 'long-time', observe that it is not of interest to consider what happens asymptotically as $n \to \infty$, since γ_n becomes not small for a sufficiently large n: we are interested to have $\gamma_n \approx \gamma_n^{\text{long}}$ when γ_n is small. See also point 6 in Remark 4.1.

About the question (5.5), we have the following result.

Theorem 5.3. *Assume $\Lambda^* \setminus \Lambda^*_{j^*} \neq \varnothing$ and $0 \notin \Lambda^*_{j^*}$. Fix $c_{j^*} \geqslant 0$ with c_{j^*} such that $g\left(c_{j^*}\right) < 1$, i.e., $c_{j^*} < 1.2564$ (remind point 1 in Remark 3.2). Fix $c_j \geqslant 0$ for $j = j^* + 1, \ldots, q$ such that $\Lambda^*_j \neq \varnothing$.*

For an index n such that

$$\begin{gathered}
n \max_{\lambda_i \in \Lambda^*_{j^*}} |\sigma_i| \leqslant c_{j^*} \\
n \max_{\lambda_i \in \Lambda^*_j} |\sigma_i| \leqslant c_j, \quad j = j^* + 1, \ldots, q, \quad \Lambda^*_j \neq \varnothing \\
\sum_{j=j^*+1}^{q} \left(\mathrm{e}^{\left(r_j - r_{j^*}\right)t_n} \frac{\left\| Q_j \widehat{y}_0 \right\|_2}{\left\| Q_{j^*} \widehat{y}_0 \right\|_2} \right)^2 \ll 1 \\
\sum_{\substack{j=j^*+1 \\ \Lambda^*_j \neq \varnothing}}^{q} \left(\mathrm{e}^{\left(r_j - r_{j^*}\right)t_n} \frac{1 + g(c_j)}{1 - g(c_{j^*})} K_{\Lambda^*_j \Lambda^*_{j^*}} \frac{\left\| Q_j \widehat{y}_0 \right\|_2}{\left\| Q_{j^*} \widehat{y}_0 \right\|_2} \right)^2 \ll 1
\end{gathered} \tag{5.6}$$

where $K_{\Lambda_j^*\Lambda_{j^*}^*}$ *is defined in* (3.4) *and*

$$K_{\Lambda_j^*\Lambda_{j^*}^*} = \left(\frac{\rho_{\Lambda_j^*}}{\mu_{\Lambda_{j^*}^*}}\right)^{l+1}\left(1 + O\left(h\rho_{\Lambda^*}\right)\right), \quad h\rho_{\Lambda^*} \to 0$$

holds, we have

$$\gamma_n \approx \gamma_n^{\mathrm{long}}.$$

Proof. For $n = 0, 1, 2, \ldots$, we write

$$\gamma_n = \gamma_n^{\mathrm{long}} \frac{\sqrt{1 + \sum\limits_{j=j^*+1}^{q}\left(\mathrm{e}^{(r_j - r_{j^*})t_n}\frac{\varepsilon_{n,j}}{\varepsilon_{n,j^*}}\right)^2}}{\sqrt{1 + \sum\limits_{j=j^*+1}^{q}\left(\mathrm{e}^{(r_j - r_{j^*})t_n}\frac{\left\|Q_j\widehat{y}_0\right\|_2}{\left\|Q_{j^*}\widehat{y}_0\right\|_2}\right)^2}} = \gamma_n^{\mathrm{long}}\left(1 + e_n\right)$$

where

$$\begin{aligned} |e_n| &= \left|\frac{\sqrt{1 + \sum\limits_{j=j^*+1}^{q}\left(\mathrm{e}^{(r_j - r_{j^*})t_n}\frac{\varepsilon_{n,j}}{\varepsilon_{n,j^*}}\right)^2}}{\sqrt{1 + \sum\limits_{j=j^*+1}^{q}\left(\mathrm{e}^{(r_j - r_{j^*})t_n}\frac{\left\|Q_j\widehat{y}_0\right\|_2}{\left\|Q_{j^*}\widehat{y}_0\right\|_2}\right)^2}} - 1\right| \\ &\leqslant \frac{1}{2}\max\left\{\sum_{j=j^*+1}^{q}\left(\mathrm{e}^{(r_j - r_{j^*})t_n}\frac{\varepsilon_{n,j}}{\varepsilon_{n,j^*}}\right)^2, \sum_{j=j^*+1}^{q}\left(\mathrm{e}^{(r_j - r_{j^*})t_n}\frac{\left\|Q_j\widehat{y}_0\right\|_2}{\left\|Q_{j^*}\widehat{y}_0\right\|_2}\right)^2\right\}. \end{aligned}$$

Now, for $j = j^* + 1, \ldots, q$ such that $\Lambda_j^* \neq \varnothing$, we have

$$\frac{\varepsilon_{n,j}}{\varepsilon_{n,j^*}} = \frac{\left\|\sum\limits_{i\in\Lambda_j^*}\varphi\left(n\sigma_i\right)P_i\widehat{y}_0\right\|_2}{\left\|\sum\limits_{i\in\Lambda_{j^*}^*}\varphi\left(n\sigma_i\right)P_i\widehat{y}_0\right\|_2} \leqslant \frac{\max\limits_{\lambda_i\in\Lambda_j^*}\left|\varphi\left(n\sigma_i\right)\right|\left\|Q_j\widehat{y}_0\right\|_2}{\min\limits_{\lambda_i\in\Lambda_{j^*}^*}\left|\varphi\left(n\sigma_i\right)\right|\left\|Q_{j^*}\widehat{y}_0\right\|_2} \leqslant \frac{1 + g(c_j)}{1 - g(c_{j^*})}K_{\Lambda_j^*\Lambda_{j^*}^*}\frac{\left\|Q_j\widehat{y}_0\right\|_2}{\left\|Q_{j^*}\widehat{y}_0\right\|_2}.$$

This completes the proof. □

In the case $c_{j^*} = c_j = c$, where $c > 0$ is such that $g(c) < 1$, the condition (5.6) becomes

$$\begin{gathered} n\max\nolimits_{\lambda_i\in\Lambda^*}\left|\sigma_i\right| \leqslant c \\ \sum_{j=j^*+1}^{q}\left(\mathrm{e}^{(r_j - r_{j^*})t_n}\frac{\left\|Q_j\widehat{y}_0\right\|_2}{\left\|Q_{j^*}\widehat{y}_0\right\|_2}\right)^2 \ll 1 \\ \left(\frac{1 + g(c)}{1 - g(c)}\right)^2\sum_{\substack{j=j^*+1 \\ \Lambda_j^*\neq\varnothing}}^{q}\left(\mathrm{e}^{(r_j - r_{j^*})t_n}K_{\Lambda_j^*\Lambda_{j^*}^*}\frac{\left\|Q_j\widehat{y}_0\right\|_2}{\left\|Q_{j^*}\widehat{y}_0\right\|_2}\right)^2 \ll 1. \end{gathered} \tag{5.7}$$

Theorem 5.3 says that $\gamma_n \approx \gamma_n^{\mathrm{long}}$ holds in the long-time: see points 3 and 4 in Remark 5.3 below. In other words, *the contributions to the error* γ_n *coming from non-rightmost eigenvalues in* Λ^* *vanish in the long-time.*

Remark 5.3.

1. For an index n such that

$$\begin{gathered} n\max_{\lambda_i\in\Lambda^*}\left|\sigma_i\right| \ll 1 \\ \sum_{j=j^*+1}^{q}\left(\mathrm{e}^{(r_j - r_{j^*})t_n}\frac{\left\|Q_j\widehat{y}_0\right\|_2}{\left\|Q_{j^*}\widehat{y}_0\right\|_2}\right)^2 \ll 1 \\ \sum_{\substack{j=j^*+1 \\ \Lambda_j^*\neq\varnothing}}^{q}\left(\mathrm{e}^{(r_j - r_{j^*})t_n}K_{\Lambda_j^*\Lambda_{j^*}^*}\frac{\left\|Q_j\widehat{y}_0\right\|_2}{\left\|Q_{j^*}\widehat{y}_0\right\|_2}\right)^2 \ll 1 \end{gathered} \tag{5.8}$$

we have $\gamma_n \approx \gamma_n^{\text{long}}$.

2. For an index n such that (5.7) holds, we have $\gamma_n \approx \gamma_n^{\text{long}}$ and

$$n \min_{\lambda_i \in \Lambda_{j^*}^*} |\sigma_i| \, (1 - g(c)) \leqslant \gamma_n^{\text{long}} \leqslant n \max_{\lambda_i \in \Lambda_{j^*}^*} |\sigma_i| \, (1 + g(c)) \, .$$

   Hence

$$n \min_{\lambda_i \in \Lambda_{j^*}^*} |\sigma_i| \, (1 - g\,(c)) \lessapprox \gamma_n \lessapprox n \max_{\lambda_i \in \Lambda_{j^*}^*} |\sigma_i| \, (1 + g\,(c))$$

   and the lower and upper bounds here are tighter than the lower and upper bounds in (4.2): Λ^* is replaced with its subset $\Lambda_{j^*}^*$ constituted by the rightmost eigenvalues in Λ^*.
   For an index n such that (5.8) holds, we have $\gamma_n \approx \gamma_n^{\text{long}}$ and

$$n \min_{\lambda_i \in \Lambda_{j^*}^*} |\sigma_i| \lessapprox \gamma_n^{\text{long}} \lessapprox n \max_{\lambda_i \in \Lambda_{j^*}^*} |\sigma_i| \, .$$

   Hence

$$n \min_{\lambda_i \in \Lambda_{j^*}^*} |\sigma_i| \lessapprox \gamma_n \lessapprox n \max_{\lambda_i \in \Lambda_{j^*}^*} |\sigma_i| \, .$$

3. In (5.7) the index n should be sufficiently small in order to have the first condition satisfied but sufficiently large in order to have the second and third conditions satisfied. For indices n such that

$$kc \leqslant n \max_{\lambda_i \in \Lambda^*} |\sigma_i| \leqslant c$$

   where $0 < k < 1$, the exponentials terms $\mathrm{e}^{(r_j - r_{j^*})t_n}$ in the second and third conditions satisfy

$$\mathrm{e}^{\,(r_j - r_{j^*})t_n} \leqslant \mathrm{e}^{kc\,h(r_j - r_{j^*})/\max_{\lambda_i \in \Lambda^*} |\sigma_i|}$$

   and for the exponent in the right-hand side, we have

$$kc \frac{h(r_j - r_{j^*})}{\max\limits_{\lambda_i \in \Lambda^*} |\sigma_i|} = \frac{r_j - r_{j^*}}{\rho_{\Lambda^*}} k\tau$$

   where τ is given in (4.4). Hence, if

$$kc \leqslant n \max_{\lambda_i \in \Lambda^*} |\sigma_i| \leqslant c$$
$$\sum_{j=j^*+1}^{q} \left(\mathrm{e}^{(r_j - r_{j^*})/(\rho_{\Lambda^*})k\tau} \frac{\left\| Q_j \widehat{y}_0 \right\|_2}{\left\| Q_{j^*} \widehat{y}_0 \right\|_2} \right)^2 \ll 1$$
$$\left(\frac{1 + g(c)}{1 - g(c)} \right)^2 \sum_{\substack{j=j^*+1 \\ \Lambda_j^* \neq \varnothing}}^{q} \left(\mathrm{e}^{(r_j - r_{j^*})/(\rho_{\Lambda^*})k\tau} K_{\Lambda_j^* \Lambda_{j^*}^*} \frac{\left\| Q_j \widehat{y}_0 \right\|_2}{\left\| Q_{j^*} \widehat{y}_0 \right\|_2} \right)^2 \ll 1$$

   then $\gamma_n \approx \gamma_n^{\text{long}}$. By assuming $k\tau \gg 1$ (remind $\tau \to +\infty$, as $h\rho_{\Lambda^*} \to 0$), it is expected to have in (5.7) the second and the third conditions satisfied for large indices satisfying the first condition.
   In the next point 4, we provide an interval I such that $\gamma_n \approx \gamma_n^{\text{long}}$ for $t_n \rho_{\Lambda^*} \in I$.

4. Suppose $h\rho_{\Lambda^*}$ sufficiently small so that

$$K_{\Lambda_j^* \Lambda_{j^*}^*} \leqslant 2 \left(\frac{\rho_{\Lambda_j^*}}{\mu_{\Lambda_{j^*}^*}} \right)^{l+1}, \quad j = j^* + 1, \ldots, q \ \text{ such that } \Lambda_j^* \neq \varnothing$$

   holds. Fix $c > 0$ such that $g(c) < 1$. Theorem 5.3 says that for

$$\tau_0 \leqslant t_n \rho_{\Lambda^*} \leqslant \tau$$

where τ is given in (4.4) and $\tau_0 \geqslant 0$ is such that

$$\sum_{j=j^*+1}^{q} \left(\mathrm{e}^{(r_j - r_{j^*})/(\rho_{\Lambda^*})\tau_0} \frac{\left\|Q_j \widehat{y}_0\right\|_2}{\left\|Q_{j^*} \widehat{y}_0\right\|_2} \right)^2 \ll 1$$

$$\left(2\frac{1+g\left(c\right)}{1-g\left(c\right)} \right)^2 \sum_{\substack{j=j^*+1 \\ \Lambda_j^* \neq \varnothing}}^{q} \left(\mathrm{e}^{(r_j - r_{j^*})/(\rho_{\Lambda^*})\tau_0} \left(\frac{\rho_{\Lambda_j^*}}{\mu_{\Lambda_{j^*}^*}} \right)^{l+1} \frac{\left\|Q_j \widehat{y}_0\right\|_2}{\left\|Q_{j^*} \widehat{y}_0\right\|_2} \right)^2 \ll 1$$

we have $\gamma_n \approx \gamma_n^{\mathrm{long}}$. Observe that $\tau \to +\infty$, as $h\rho_{\Lambda^*} \to 0$, and τ_0 can be chosen independently of $h\rho_{\Lambda^*}$. We conclude that, *for a dimensionless time $t_n\rho_{\Lambda^*}$ in the interval $[\tau_0, \tau]$, the contributions to the relative error γ_n given by errors σ_i with $\lambda_i \in \Lambda^* \setminus \Lambda_{j^*}^*$, i.e., the contributions given by non-rightmost eigenvalues in Λ^*, vanish.* Of course, this is coherent with the fact that, in the long-time, the solution y becomes the long-time solution y^{long}, which depends only on the rightmost eigenvalues of Λ^*, namely the eigenvalues in $\Lambda_{j^*}^*$.

5. The second and third conditions in (5.7) hold when

$$\left(\mathrm{e}^{\left(r_{m^*} - r_{j^*}\right)t_n} \frac{\sqrt{1 - \left\|Q_{j^*} \widehat{y}_0\right\|_2^2}}{\left\|Q_{j^*} \widehat{y}_0\right\|_2} \right)^2 \ll 1$$

$$\left(\frac{1+g\left(c\right)}{1-g(c)} \right)^2 \left(\mathrm{e}^{\left(r_{m^*} - r_{j^*}\right)t_n} K_{\Lambda^* \setminus \Lambda_{j^*}^* \Lambda_{j^*}^*} \frac{\sqrt{1 - \left\|Q_{j^*} \widehat{y}_0\right\|_2^2}}{\left\|Q_{j^*} \widehat{y}_0\right\|_2} \right)^2 \ll 1$$

where m^* is given in (5.2) and $K_{\Lambda^* \setminus \Lambda_{j^*}^* \Lambda_{j^*}^*}$ is defined in (3.4) and

$$K_{\Lambda^* \setminus \Lambda_{j^*}^* \Lambda_{j^*}^*} = \left(\frac{\rho_{\Lambda^* \setminus \Lambda_{j^*}^*}}{\mu_{\Lambda_{j^*}^*}} \right)^{l+1} \left(1 + O\left(h\rho_{\Lambda^*}\right)\right), \quad h\rho_{\Lambda^*} \to 0$$

holds.

6. Theorem 5.3 assumes $\Lambda^* \setminus \Lambda_{j^*}^* \neq \varnothing$ and $0 \notin \Lambda_{j^*}^*$. If $\Lambda^* \setminus \Lambda_{j^*}^* = \varnothing$, then

$$\gamma_n = \gamma_n^{\mathrm{long}}, \quad n = 0, 1, 2, \ldots$$

If $\Lambda^* \setminus \Lambda_{j^*}^* \neq \varnothing$ and $\Lambda_{j^*}^* = \{0\}$, then

$$\gamma_n^{\mathrm{long}} = 0, \quad n = 0, 1, 2, \ldots$$

In this case, by considering in the formula (2.1) only the leading exponential terms $\mathrm{e}^{(r_{j_1^*} - r_1)t_n}\varepsilon_{n,j_1^*}$ at the numerator (j_1^* defined in (4.6)) and $\mathrm{e}^{(r_{j^*} - r_1)t_n} \cdot \left\|Q_{j^*} \widehat{y}_0\right\|_2$ at the denominator, we obtain

$$\frac{\mathrm{e}^{\left(r_{j_1^*} - r_{j^*}\right)t_n}\varepsilon_{n,j_1^*}}{\left\|Q_{j^*} \widehat{y}_0\right\|_2}, \quad n = 0, 1, 2, \ldots$$

which is exponentially decreasing in time.
For $\Lambda^* \setminus \Lambda_{j^*}^* \neq \varnothing$ and $\Lambda_{j^*}^* \neq \{0\}$, an improved version of theorem holds with the fourth condition in (5.6) replaced by

$$\sum_{\substack{j=j^*+1 \\ \Lambda_j^* \neq \varnothing}}^{q} \left(\mathrm{e}^{\left(r_j - r_{j^*}\right)t_n} \frac{1+g(c_j)}{1-g(c_{j^*})} K_{\Lambda_j^* \Lambda_{j^*}^* \setminus \{0\}} \frac{\sqrt{\sum\limits_{\lambda_i \in \Lambda_j^* \setminus \{0\}} \left\|P_i \widehat{y}_0\right\|_2^2}}{\sqrt{\sum\limits_{\lambda_i \in \Lambda_{j^*}^* \setminus \{0\}} \left\|P_i \widehat{y}_0\right\|_2^2}} \right)^2 \ll 1$$

where $K_{\Lambda_j^* \Lambda_{j^*}^* \setminus \{0\}}$ is defined in (3.4) and

$$K_{\Lambda_j^* \Lambda_{j^*}^* \setminus \{0\}} = \left(\frac{\rho_{\Lambda_j^*}}{\mu_{\Lambda_{j^*}^* \setminus \{0\}}} \right)^{l+1} \left(1 + O\left(h\rho_{\Lambda^*}\right)\right), \quad h\rho_{\Lambda^*} \to 0$$

holds. This result can be used for the case $0 \in \Lambda_{j^*}^*$ and $\Lambda_{j^*}^* \neq \{0\}$.

5.2 Perturbed initial value

In the previous subsection we have described the long-time behavior of the relative error γ_n due to the sole approximant. Now, we describe the long-time behavior of the relative error $\delta(t)$ due to the sole perturbation in the initial value.

5.2.1 Long-time behavior of the error $\boldsymbol{\delta}(t)$

By considering in the formula (1.4) only the leading exponential terms $e^{(r_{j^{**}}-r_1)t_n} \left\|Q_{j^{**}}\widehat{z}_0\right\|_2$ at the numerator and $e^{(r_{j^*}-r_1)t_n} \left\|Q_{j^*}\widehat{y}_0\right\|_2$ at the denominator, we define

$$\delta^{\text{long}}(t) := e^{(r_{j^{**}}-r_{j^*})t} \frac{\left\|Q_{j^{**}}\widehat{z}_0\right\|_2}{\left\|Q_{j^*}\widehat{y}_0\right\|_2} \varepsilon, \quad t \geqslant 0.$$

The error $\delta^{\text{long}}(t)$ is exponentially decreasing in t when $j^* < j^{**}$, exponentially increasing when $j^* > j^{**}$ and constant

$$\delta^{\text{long}}(t) = \delta^{\text{long}} := \frac{\left\|Q_{j^*}\widehat{z}_0\right\|_2}{\left\|Q_{j^*}\widehat{y}_0\right\|_2} \varepsilon, \quad t \geqslant 0$$

when $j^* = j^{**}$. In the generic situation $j^* = j^{**} = 1$, we have

$$\delta^{\text{long}}(t) = \delta^{\text{long}} = \frac{\left\|Q_1\widehat{z}_0\right\|_2}{\left\|Q_1\widehat{y}_0\right\|_2} \varepsilon, \quad t \geqslant 0.$$

Next theorem describe the long-time behavior of the relative error $\delta(t)$.

Theorem 5.4. *For $t \geqslant 0$ such that*

$$\begin{aligned} \sum_{j=j^*+1}^{q} \left(e^{(r_j - r_{j^*})t} \frac{\left\|Q_j\widehat{y}_0\right\|_2}{\left\|Q_{j^*}\widehat{y}_0\right\|_2} \right)^2 &\ll 1 \\ \sum_{j=j^{**}+1}^{q} \left(e^{(r_j - r_{j^{**}})t} \frac{\left\|Q_j\widehat{z}_0\right\|_2}{\left\|Q_{j^{**}}\widehat{z}_0\right\|_2} \right)^2 &\ll 1 \end{aligned} \tag{5.9}$$

we have

$$\delta(t) \approx \delta^{\text{long}}(t).$$

Proof. For $t \geqslant 0$, we have

$$\begin{aligned} \delta(t) &= \delta^{\text{long}}(t) \frac{\sqrt{1 + \sum\limits_{j=j^{**}+1}^{q} \left(e^{(r_j - r_{j^{**}})t} \frac{\left\|Q_j\widehat{z}_0\right\|_2}{\left\|Q_{j^{**}}\widehat{z}_0\right\|_2} \right)^2}}{\sqrt{1 + \sum\limits_{j=j^*+1}^{q} \left(e^{(r_j - r_{j^*})t} \frac{\left\|Q_j\widehat{y}_0\right\|_2}{\left\|Q_{j^*}\widehat{y}_0\right\|_2} \right)^2}} \\ &= \delta^{\text{long}}(t)\left(1 + e(t)\right) \end{aligned}$$

with

$$|e(t)| \leqslant \frac{1}{2} \max \left\{ \sum_{j=j^{**}+1}^{q} \left(e^{(r_j - r_{j^*})t} \frac{\left\|Q_j\widehat{z}_0\right\|_2}{\left\|Q_{j^{**}}\widehat{z}_0\right\|_2} \right)^2, \sum_{j=j^*+1}^{q} \left(e^{(r_j - r_{j^*})t} \frac{\left\|Q_j\widehat{y}_0\right\|_2}{\left\|Q_{j^*}\widehat{y}_0\right\|_2} \right)^2 \right\}.$$

This completes the proof. □

Remark 5.4.

1. Theorem 5.4 says that
$$\delta(t) \approx \delta^{\text{long}}(t) \quad \text{for a sufficiently large } t.$$
In addition, by looking at the proof of the theorem, we see that
$$\frac{\delta(t)}{\delta^{\text{long}}(t)} \to 1, \quad t \to +\infty.$$
On the other hand, Theorem 5.3 about the error γ_n says neither
$$\frac{\gamma_n}{\gamma_n^{\text{long}}} \to 1, \quad n \to \infty$$
nor
$$\gamma_n \approx \gamma_n^{\text{long}} \quad \text{for a sufficiently large index } n.$$
In fact, Theorem 5.3 cannot give asymptotic results as $n \to \infty$ because of the first and second conditions in (5.6), which require to take n sufficiently small.
2. For $\Lambda^* \setminus \Lambda^*_{j^*} \neq \varnothing$ and $\Lambda^{**} \setminus \Lambda^{**}_{j^{**}} \neq \varnothing$, the condition (5.9) holds when
$$\left(\mathrm{e}^{(r_{m^*} - r_{j^*})t_n} \frac{\sqrt{1 - \|Q_{j^*}\widehat{y}_0\|_2^2}}{\|Q_{j^*}\widehat{y}_0\|_2} \right)^2 \ll 1$$
$$\left(\mathrm{e}^{(r_{m^{**}} - r_{j^{**}})t_n} \frac{\sqrt{1 - \|Q_{j^{**}}\widehat{z}_0\|_2^2}}{\|Q_{j^{**}}\widehat{z}_0\|_2} \right)^2 \ll 1$$
where m^* is defined in (5.2) and similarly we have
$$m^{**} = \min\{ j \in \{j^{**} + 1, \ldots, q\} : \Lambda_j^{**} \neq 0 \}.$$

5.2.2 Long-time behavior of the error δ_n

Theorem 4.2 and the consequent points 2, 3, 4, and 6 in Remark 4.2 describe the growth of the error δ_n in terms of the errors γ_n, β_n, and $\delta(t_n)$. The long-time behaviors of the errors γ_n and β_n are described in Theorem 5.3 (remind that β_n in (4.9) is the error γ_n for $\widehat{y}_0 = \widehat{z}_0$) and the long-time behavior of the error $\delta(t_n)$ is described in Theorem 5.4.

In sight of this, we can have information about the long-time behavior of δ_n. Below, in the next remark, we show what can be said by means of points 2, 3, and 4 in Remark 4.2.

Remark 5.5. Assume $\Lambda^* \setminus \Lambda^*_{j^*} \neq \varnothing$, $0 \notin \Lambda^*_{j^*}$, $\Lambda^{**} \setminus \Lambda^*_{j^{**}} \neq \varnothing$, and $0 \notin \Lambda^{**}_{j^{**}}$. Consider an index n such that
$$n \max_{\lambda_i \in \Lambda^*} |\sigma_i| \ll 1, \qquad n \max_{\lambda_i \in \Lambda^{**}} |\sigma_i| \ll 1$$
$$\sum_{j=j^*+1}^{q} \left(\mathrm{e}^{(r_j - r_{j^*})t_n} \frac{\|Q_j\widehat{y}_0\|_2}{\|Q_{j^*}\widehat{y}_0\|_2} \right)^2 \ll 1$$
$$\sum_{\substack{j=j^*+1 \\ \Lambda_j^* \neq \varnothing}}^{q} \left(\mathrm{e}^{(r_j - r_{j^*})t_n} K_{\Lambda_j^* \Lambda_{j^*}^*} \frac{\|Q_j\widehat{y}_0\|_2}{\|Q_{j^*}\widehat{y}_0\|_2} \right)^2 \ll 1$$
$$\sum_{j=j^{**}+1}^{q} \left(\mathrm{e}^{(r_j - r_{j^{**}})t_n} \frac{\|Q_j\widehat{z}_0\|_2}{\|Q_{j^{**}}\widehat{z}_0\|_2} \right)^2 \ll 1$$
$$\sum_{\substack{j=j^{**}+1 \\ \Lambda_j^{**} \neq \varnothing}}^{q} \left(\mathrm{e}^{(r_j - r_{j^{**}})t_n} K_{\Lambda_j^{**} \Lambda_{j^{**}}^{**}} \frac{\|Q_j\widehat{z}_0\|_2}{\|Q_{j^*}\widehat{z}_0\|_2} \right)^2 \ll 1$$

By point 1 in Remark 5.3 and Theorem 5.4, we have

$$\delta\left(t_n\right) \approx \delta^{\text{long}}\left(t_n\right), \qquad \gamma_n \approx \gamma_n^{\text{long}}, \quad \beta_n \approx \beta_n^{\text{long}}.$$

The following points A, B, and C can be stated.

A. Recall point 2 in Remark 4.2. We have

$$\left|\delta_n - \gamma_n\right| \lessapprox \delta^{\text{long}}\left(t_n\right).$$

B. Recall point 3 in Remark 4.2. If

$$K_{\Lambda_{j^{**}}\Lambda_{j^*}}\, \delta^{\text{long}}\left(t_n\right) \ll 1$$

then

$$\left|\delta_n - \delta\left(t_n\right)\right| \lessapprox \gamma_n^{\text{long}}.$$

C. Recall point 4 in Remark 4.2. If

$$\frac{K_{\Lambda_{j^*}\Lambda_{j^{**}}}}{\delta^{\text{long}}\left(t_n\right)} \ll 1$$

then

$$\frac{\left|\delta_n - \delta\left(t_n\right)\right|}{\delta\left(t_n\right)} \lessapprox \beta_n^{\text{long}}.$$

A more detailed analysis of the long-time behavior of the error δ_n is presented in [11].

6 A first example

As an example, we consider an ODE (1.1) with $A \in \mathbb{R}^{2\times 2}$ symmetric and non-singular. Let λ_1 and λ_2 be the non-zero eigenvalues of A with $\lambda_1 > \lambda_2$.

Regarding the errors σ_1 and σ_2 of the approximant, we have

$$\begin{gathered}
\left|\sigma_1\right| = C\left(h\left|\lambda_1\right|\right)^{l+1}\left(1 + O\left(h\rho\right)\right) \\
\left|\sigma_2\right| = C\left(h\left|\lambda_2\right|\right)^{l+1}\left(1 + O\left(h\rho\right)\right) \\
\max\left\{\sigma_1, \sigma_2\right\} = \left(h\rho\right)^{l+1}\left(1 + O\left(h\rho\right)\right) \\
\min\left\{\sigma_1, \sigma_2\right\} = \left(h\mu\right)^{l+1}\left(1 + O\left(h\rho\right)\right)
\end{gathered}$$

as $h\rho \to 0$, where l is the order of the approximant, $\rho := \max\left\{\left|\lambda_1\right|, \left|\lambda_2\right|\right\}$ and $\mu := \min\left\{\left|\lambda_1\right|, \left|\lambda_2\right|\right\}$.

All information given in the next Subsections 6.2, 6.3, and 6.4 derives from the theory developed in the previous sections. We consider indices n such that

$$n \max\left\{\left|\sigma_1\right|, \left|\sigma_2\right|\right\} \ll 1.$$

Moreover, we set

$$K := \frac{\left|\sigma_2\right|}{\left|\sigma_1\right|} = \left(\frac{\left|\lambda_2\right|}{\left|\lambda_1\right|}\right)^{l+1}\left(1 + O\left(h\rho\right)\right), \quad h\rho \to 0.$$

6.1 Numerical experiments

We accomplish numerical experiments with the particular 2 × 2 symmetric and non-singular matrix

$$A = \frac{1}{2}\begin{bmatrix} a+b & a-b \\ a-b & a+b \end{bmatrix}$$

where the eigenvalues are $\lambda_1 = a$ with relevant eigenvector $(1, 1)$ and $\lambda_2 = b$ with relevant eigenvector $(1, -1)$. We consider two possibilities:

	l	Σ_1	Σ_2	$K = \Sigma_2/\Sigma_1 = \lvert\sigma_2\rvert\,/\,\lvert\sigma_1\rvert$
(A1)	1	2.21e-01	2.48e-02	0.113
	2	2.20e-03	8.27e-05	0.0376
	3	1.65e-05	2.07e-07	0.0126
(A2)	1	2.52e-02	2.21e-01	8.77
	2	8.40e-05	2.20e-03	26.2
	3	2.10e-07	1.65e-05	78.4

Tab. 2: Two possibilities for matrix A.

(A1) $a = 3$ and $b = 1$;
(A2) $a = -1$ and $b = -3$.

We use the Taylor approximants of the exponential

$$z \mapsto 1 + z, \quad z \mapsto 1 + z + \frac{z^2}{2}, \quad z \mapsto 1 + z + \frac{z^2}{2} + \frac{z^3}{6}, \quad z \in \mathbb{C}$$

of orders $l = 1, 2, 3$, respectively, for the numerical integration. This numerical integration is accomplished with stepsize $h = 1/100$ over $N = 500$ steps up to $t_N = Nh = 5$. The numbers $\Sigma_1 = N\,|\sigma_1|$ and $\Sigma_2 = N\,|\sigma_2|$ in possibilities (A1) and (A2) are listed in Table 5.2.

Observe that

$$\max\{\Sigma_1, \Sigma_2\} \ll 1$$

holds (but weakly for the order one approximant) and then we have

$$n \max\{|\sigma_1|, |\sigma_2|\} \ll 1$$

for all indices $n = 0, 1, \ldots, N$. Moreover, observe that

$$K = \frac{\Sigma_2}{\Sigma_1} = \frac{|\sigma_2|}{|\sigma_1|} \approx \left(\frac{|b|}{|a|}\right)^{l+1}.$$

6.2 Unperturbed initial value

Assume the generic situation $\Lambda^* = \{\lambda_1, \lambda_2\}$.

Solutions y and y^{long}. We have

$$\begin{aligned} y(t) &= e^{\lambda_1 t} P_1 y_0 + e^{\lambda_2 t} P_2 y_0, \quad t \geqslant 0 \\ y^{\text{long}}(t) &= e^{\lambda_1 t} P_1 y_0, \quad t \geqslant 0. \end{aligned}$$

Long-time behavior: $y(t) \approx y^{\text{long}}(t)$ for t such that

$$e^{(\lambda_2 - \lambda_1)t} \frac{\|P_2 \widehat{y}_0\|_2}{\|P_1 \widehat{y}_0\|_2} \ll 1.$$

Errors γ_n and γ_n^{long}. We have

$$\begin{aligned} n \min\{|\sigma_1|, |\sigma_2|\} \lessapprox \gamma_n &\lessapprox n \max\{|\sigma_1|, |\sigma_2|\} \\ \gamma_n^{\text{long}} &\approx n\,|\sigma_1|. \end{aligned}$$

Long-time behavior: $\gamma_n \approx \gamma_n^{\text{long}}$ for n such that

$$\left(e^{(\lambda_2 - \lambda_1)t_n} \max\{1, K\} \frac{\|P_2 \widehat{y}_0\|_2}{\|P_1 \widehat{y}_0\|_2}\right)^2 \ll 1.$$

6.2.1 Numerical experiments

We numerically test the possibilities (A1) and (A2) with the initial value $y_0 = (2, -1)$, for which

$$P_1 y_0 = \frac{1}{2}\,(1, 1)\,, \quad P_2 y_0 = \frac{3}{2}(1, -1)$$

and

$$\left\|P_1 \widehat{y}_0\right\|_2 = \frac{1}{\sqrt{10}}, \quad \left\|P_2 \widehat{y}_0\right\|_2 = \frac{3}{\sqrt{10}}.$$

Solutions y and y^{long}. We have

$$\begin{aligned} y(t) &= \frac{1}{2}\mathrm{e}^{at}(1, 1) + \frac{3}{2}\mathrm{e}^{bt}(1, -1), \quad t \geqslant 0 \\ y^{\text{long}}(t) &= \frac{1}{2}\mathrm{e}^{at}(1, 1), \quad t \geqslant 0. \end{aligned}$$

Long-time behavior: $y(t) \approx y^{\text{long}}(t)$ for t such that $3\mathrm{e}^{-2t} \ll 1$.

Errors γ_n and γ_n^{long}. We have

$$\begin{gathered} \frac{n}{N}\min\{\Sigma_1, \Sigma_2\} \lessapprox \gamma_n \lessapprox \frac{n}{N}\max\{\Sigma_1, \Sigma_2\} \\ \gamma_n^{\text{long}} \approx \frac{n}{N}\Sigma_1 \end{gathered}$$

for all $n = 0, 1, \ldots, N$. Long-time behavior: $\gamma_n \approx \gamma_n^{\text{long}}$ for $n = 0, 1, \ldots, N$ such that

$$\left(3\max\{1, K\}\mathrm{e}^{-2t_n}\right)^2 \ll 1$$

In Figs. 2 and 3, for possibilities (A1) and (A2), respectively, we see the error γ_n (solid red line) along with $\frac{n}{N}\Sigma_1$ (dashed blue line) and $\frac{n}{N}\Sigma_2$ (dash-dotted green line) for $n = 0, 1, \ldots, N$.

6.3 Perturbed initial value, I

Assume the generic situation $\Lambda^* = \Lambda^{**} = \{\lambda_1, \lambda_2\}$.

Solutions y and y^{long}: as in Subsection 6.2.

Errors $\delta(t)$ and $\delta^{\text{long}}(t)$. We have

$$\begin{gathered} \delta\,(t) = \frac{\sqrt{\left\|P_1 \widehat{z}_0\right\|_2^2 + \left(\mathrm{e}^{(\lambda_2-\lambda_1)t}\left\|P_2 \widehat{z}_0\right\|_2\right)^2}}{\sqrt{\left\|P_1 \widehat{y}_0\right\|_2^2 + \left(\mathrm{e}^{(\lambda_2-\lambda_1)t}\left\|P_2 \widehat{y}_0\right\|_2\right)^2}}\varepsilon, \quad t \geqslant 0 \\ \delta^{\text{long}}(t) = \delta^{\text{long}} = \frac{\left\|P_1 \widehat{z}_0\right\|}{\left\|P_1 \widehat{y}_0\right\|_2}\varepsilon, \quad t \geqslant 0. \end{gathered}$$

Long-time behavior: $\delta\,(t) \approx \delta^{\text{long}}$ for t such that

$$\begin{gathered} \left(\mathrm{e}^{(\lambda_2-\lambda_1)t}\frac{\left\|P_2 \widehat{y}_0\right\|_2}{\left\|P_1 \widehat{y}_0\right\|_2}\right)^2 \ll 1 \\ \left(\mathrm{e}^{(\lambda_2-\lambda_1)t}\frac{\left\|P_2 \widehat{z}_0\right\|_2}{\left\|P_1 \widehat{z}_0\right\|_2}\right)^2 \ll 1. \end{gathered}$$

Errors γ_n and γ_n^{long}: as in Subsection 6.2.

Error δ_n. We have

$$\left|\delta_n - \gamma_n\right| \lessapprox \delta\,(t_n)$$

and

$$\left|\delta_n - \delta\,(t_n)\right| \lessapprox \gamma_n$$

when

$$\max\left\{\frac{\left\|P_1 \widehat{z}_0\right\|_2}{\left\|P_1 \widehat{y}_0\right\|_2}, \frac{\left\|P_2 \widehat{z}_0\right\|_2}{\left\|P_2 \widehat{y}_0\right\|_2}\right\}\varepsilon \ll 1.$$

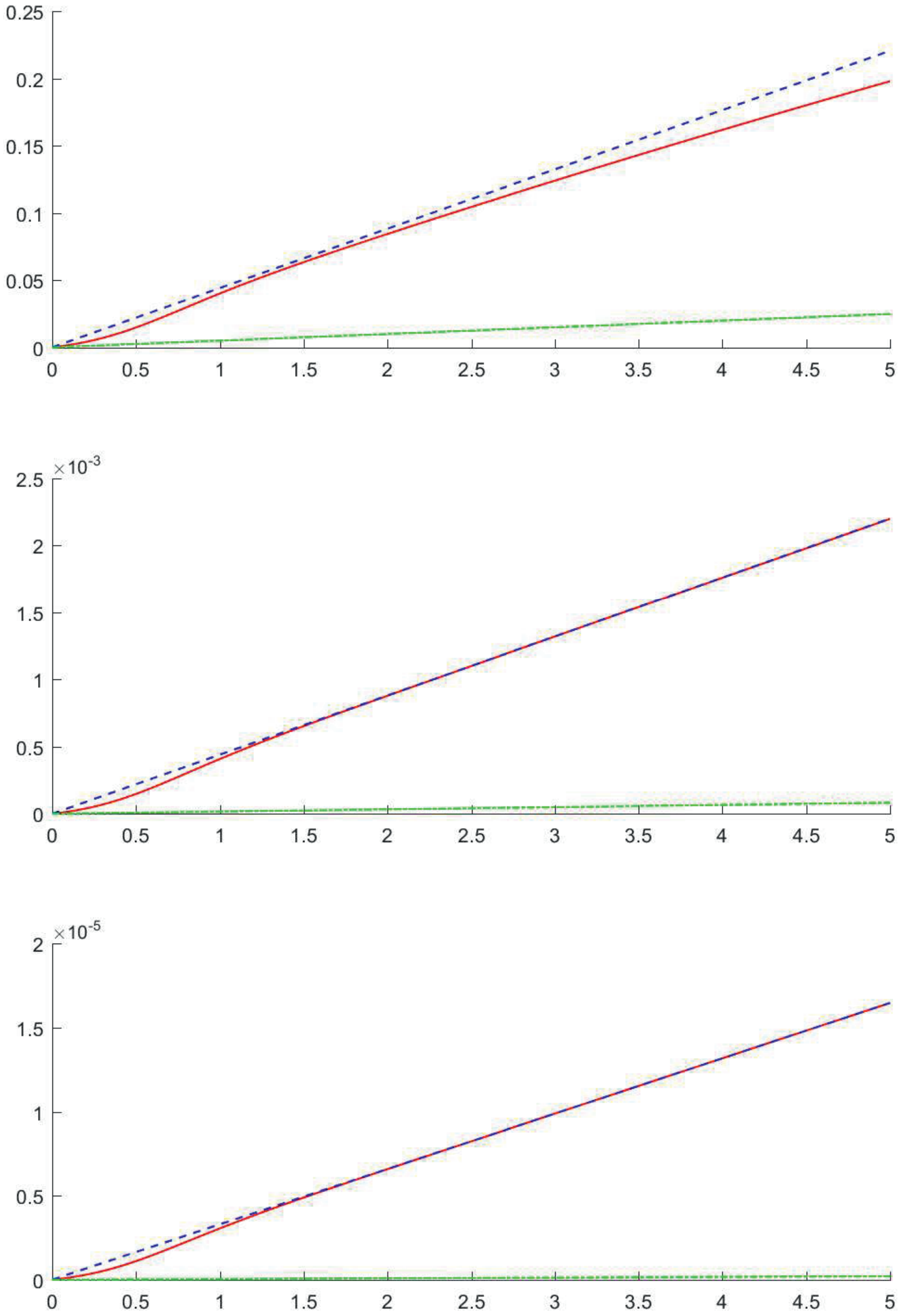

Fig. 2: Possibility (A1) with initial value $y_0 = (2, -1)$ unperturbed. The l-th row, $l = 1, 2, 3$, corresponds to the approximant of order l. Error γ_n (solid red line) along with $\frac{n}{N}\Sigma_1$ (dashed blue line) and $\frac{n}{N}\Sigma_2$ (dash-dotted green line). The abscissas are the times $t_n = nh$, $n = 0, 1, 2, \ldots, N$.

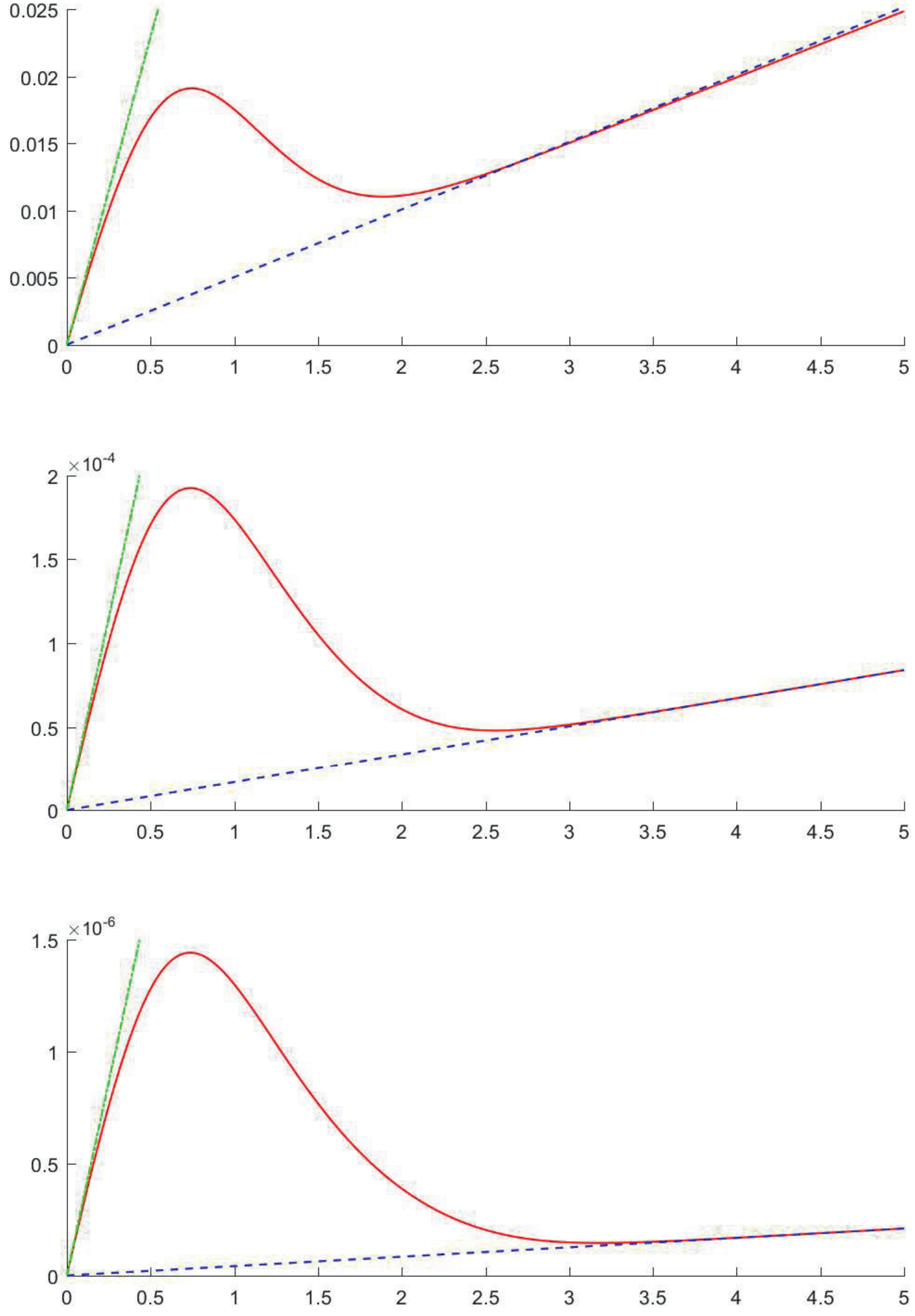

Fig. 3: Possibility (A2) with initial value $y_0 = (2, -1)$ unperturbed. The l-th row, $l = 1, 2, 3$, corresponds to the approximant of order l. Error γ_n (solid red line) along with $\frac{n}{N}\Sigma_1$ (dashed blue line) and $\frac{n}{N}\Sigma_2$ (dash-dotted green line). The abscissas are the times $t_n = nh$, $n = 0, 1, 2, \dots, N$.

6.3.1 Numerical experiments

For the numerical tests we consider the same initial value $y_0 = (2, -1)$ as in Subsubsection 6.2.1, but now it is perturbed with $\widehat{z}_0 \in \operatorname{span}(1, 2)$, for which

$$\left\|P_1\widehat{z}_0\right\|_2 = \frac{3}{\sqrt{10}}, \qquad \left\|P_2\widehat{z}_0\right\|_2 = \frac{1}{\sqrt{10}}.$$

Solutions y and y^{long}: as in Subsubsection 6.2.1.

Errors $\delta(t)$ and $\delta^{\mathrm{long}}(t)$. We have

$$\begin{aligned}\delta(t) &= \frac{\sqrt{3^2 + \left(\mathrm{e}^{-2t}\right)^2}}{\sqrt{1 + \left(3\mathrm{e}^{-2t}\right)^2}}\varepsilon, \quad t \geqslant 0\\ \delta^{\mathrm{long}}(t) &= \delta^{\mathrm{long}} = 3\varepsilon, \quad t \geqslant 0.\end{aligned}$$

Long-time behavior: $\delta(t) \approx \delta^{\mathrm{long}}$ for t such that $\left(3\mathrm{e}^{-2t}\right)^2 \ll 1$.

Errors γ_n and γ_n^{long}: as in Subsubsection 6.2.1.

Error δ_n. We have

$$\left|\delta_n - \gamma_n\right| \lessapprox \delta(t_n)$$

for all indices $n = 0, 1, \ldots, N$ and

$$\left|\delta_n - \delta(t_n)\right| \lessapprox \gamma_n$$

for all indices $n = 0, 1, \ldots, N$ when

$$3\varepsilon \ll 1.$$

In Figs. 4 and 5, for possibilities (A1) and (A2), respectively, and $\varepsilon = 10^{-2}$, we see, for $n = 0, 1, 2, \ldots, N$, in the left column the deviation $\left|\delta_n - \gamma_n\right|$ (solid red line) along with the error $\delta(t_n)$ (dashed blue line) and in the right column the deviation $\left|\delta_n - \delta(t_n)\right|$ (solid red line) along with the error γ_n (dashed blue line).

6.4 Perturbed initial value, II

Assume the non-generic situation $\Lambda^* = \{\lambda_2\}$ and $\Lambda^{**} = \{\lambda_1, \lambda_2\}$.

Solutions y and y^{long}. We have

$$y(t) = y^{\mathrm{long}}(t) = \mathrm{e}^{\lambda_2 t} y_0, \quad t \geqslant 0.$$

Errors $\delta(t)$ and $\delta^{\mathrm{long}}(t)$. We have

$$\begin{aligned}\delta(t) &= \frac{\sqrt{\left\|P_1\widehat{z}_0\right\|_2^2 + \left(\mathrm{e}^{(\lambda_2-\lambda_1)t}\left\|P_2\widehat{z}_0\right\|_2\right)^2}}{\mathrm{e}^{(\lambda_2-\lambda_1)t}}\varepsilon, \quad t \geqslant 0\\ \delta^{\mathrm{long}}(t) &= \mathrm{e}^{(\lambda_1-\lambda_2)t}\left\|P_1\widehat{z}_0\right\|\varepsilon, \quad t \geqslant 0.\end{aligned}$$

Long-time behavior: $\delta(t) \approx \delta^{\mathrm{long}}(t)$ for t such that

$$\left(\mathrm{e}^{(\lambda_2-\lambda_1)t}\frac{\left\|P_2\widehat{z}_0\right\|_2}{\left\|P_1\widehat{z}_0\right\|_2}\right)^2 \ll 1.$$

Errors β_n and β_n^{long}: as the errors γ_n and γ_n^{long} in Subsection 6.2.

Error δ_n. We have

$$\frac{\left|\delta_n - \delta(t_n)\right|}{\delta(t_n)} \lessapprox \beta_n$$

for n such that

$$\frac{K}{\mathrm{e}^{(\lambda_1-\lambda_2)t_n}\left\|P_1\widehat{z}_0\right\|_2\varepsilon} \ll 1.$$

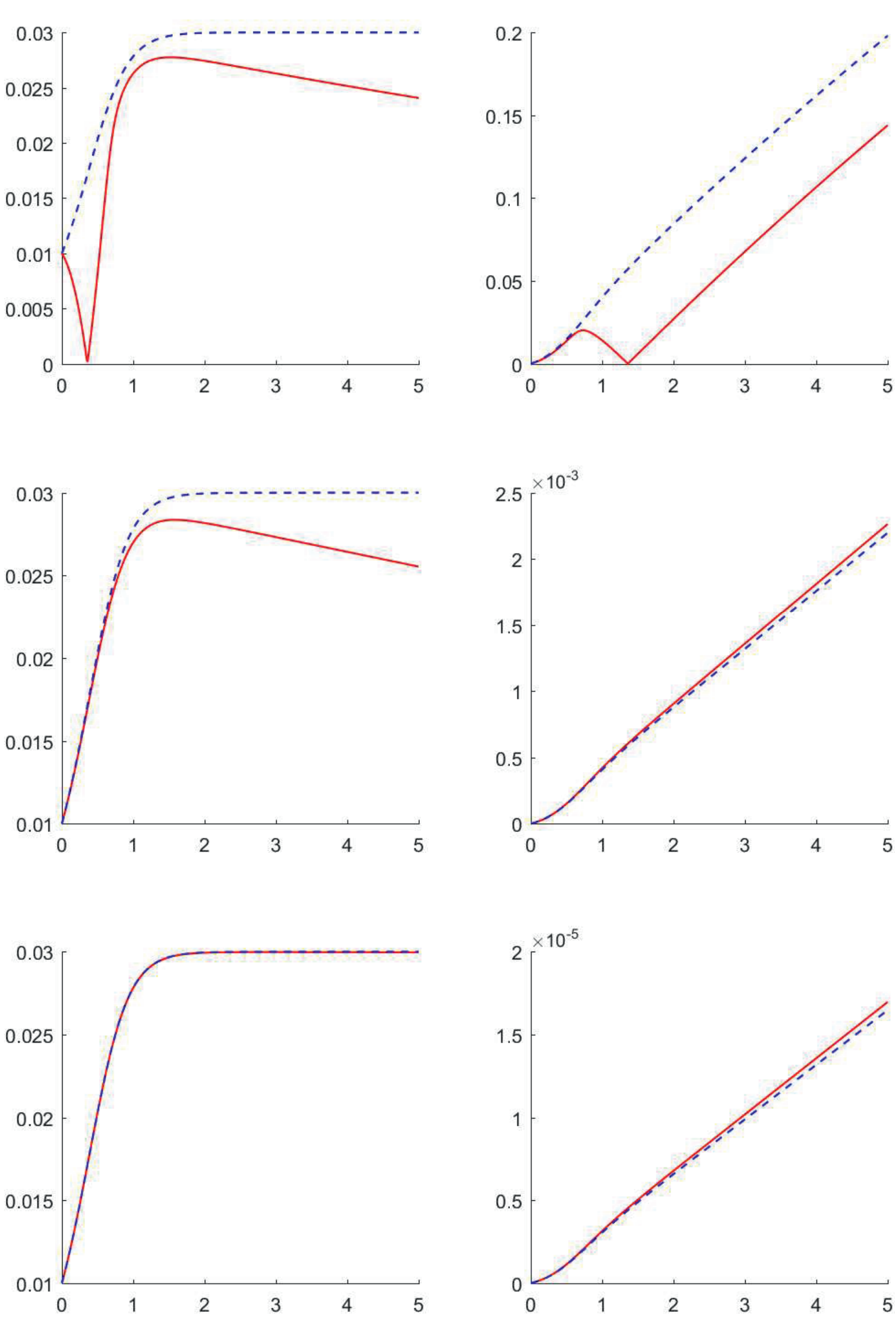

Fig. 4: Possibility (A1) with $y_0 = (2, -1)$ perturbed with $\varepsilon = 10^{-2}$ and $\widehat{z}_0 \in \operatorname{span}(1, 2)$. The l-th row, $l = 1, 2, 3$, corresponds to the approximant of order l. Left column: deviation $\left|\delta_n - \gamma_n\right|$ (solid red line) along with the error $\delta(t_n)$ (dashed blue line). Right column: deviation $|\delta_n - \delta(t_n)|$ (solid red line) along with the error γ_n (dashed blue line). The abscissas are the times $t_n = nh$, $n = 0, 1, 2, \ldots, N$.

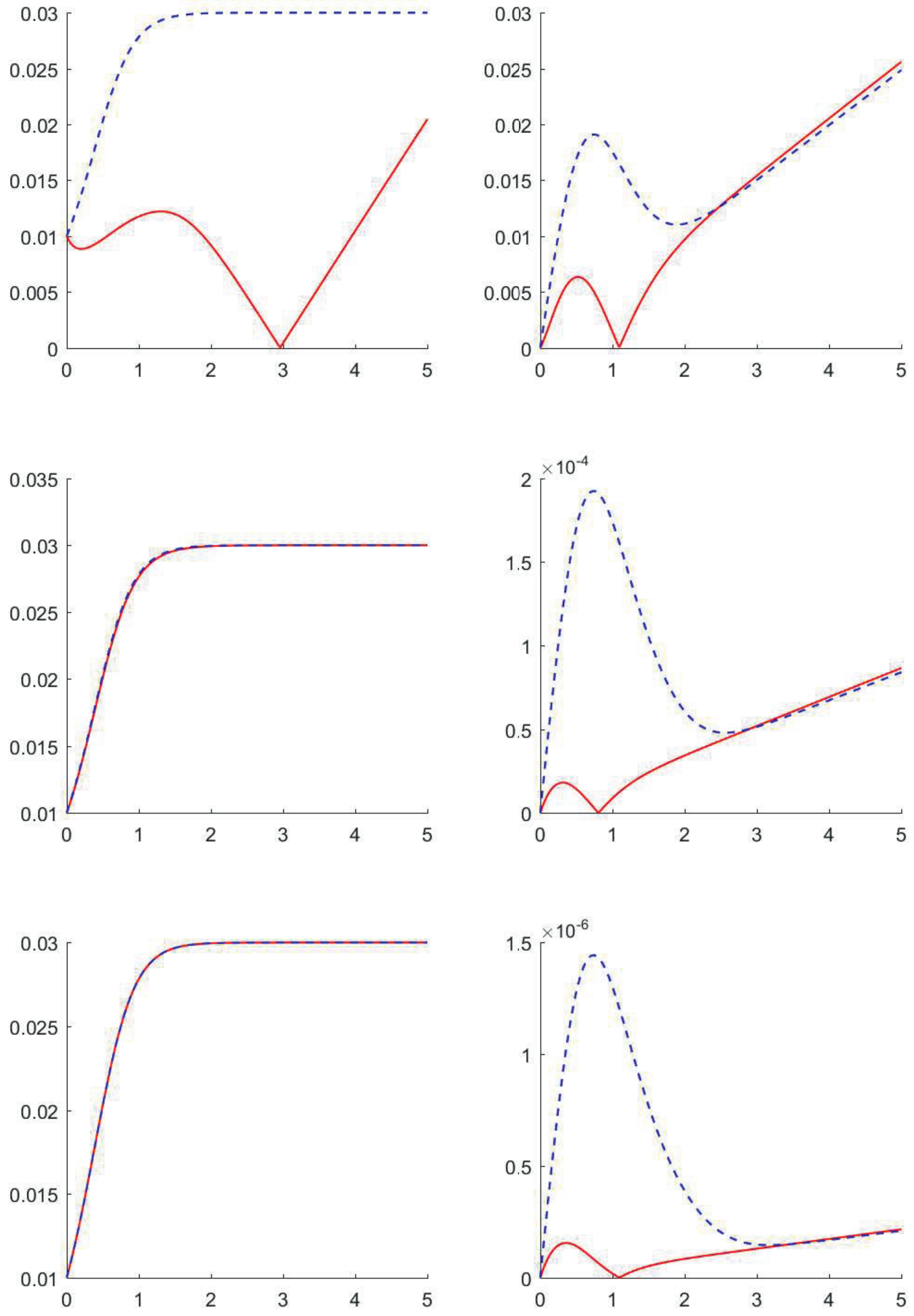

Fig. 5: Possibility (A2) with $y_0 = (2, -1)$ perturbed with $\varepsilon = 10^{-2}$ and $\widehat{z}_0 \in \operatorname{span}(1, 2)$. The l-th row, $l = 1, 2, 3$, corresponds to the approximant of order l. Left column: deviation $|\delta_n - \gamma_n|$ (solid red line) along with the error $\delta(t_n)$ (dashed blue line). Right column: deviation $|\delta_n - \delta(t_n)|$ (solid red line) along with the error γ_n (dashed blue line). The abscissas are the times $t_n = nh$, $n = 0, 1, 2, \dots, N$.

6.4.1 Numerical experiments

In the numerical tests, we consider the initial value $y_0 = (1, -1)$, for which $\varLambda^* = \{\lambda_2\}$, perturbed with $\widehat{z}_0 \in \operatorname{span}(1, 2)$ (as in Subsubsection 6.3.1).

Solutions y and y^{long}. We have

$$y(t) = y^{\text{long}}(t) = \mathrm{e}^{bt}(1, 1), \quad t \geqslant 0.$$

Errors $\delta(t)$ *and* $\delta^{\text{long}}(t)$. We have

$$\begin{aligned}\delta(t) &= \frac{1}{\sqrt{10}} \cdot \frac{\sqrt{3^2 + (\mathrm{e}^{-2t})^2}}{\mathrm{e}^{-2t}} \varepsilon, \quad t \geqslant 0 \\ \delta^{\text{long}}(t) &= \frac{3}{\sqrt{10}} \mathrm{e}^{2t} \varepsilon, \quad t \geqslant 0.\end{aligned}$$

Long-time behavior: $\delta(t) \approx \delta^{\text{long}}(t)$ for t such that $\left(\frac{1}{3}\mathrm{e}^{-2t}\right)^2 \ll 1$.

Errors β_n *and* β_n^{long}: as the errors γ_n and γ_n^{long} in Subsubsection 6.2.1.

Error δ_n. We have

$$\frac{|\delta_n - \delta(t_n)|}{\delta(t_n)} \lessapprox \beta_n$$

for $n = 0, 1, \ldots, N$ such that

$$\frac{\sqrt{10}}{3} \cdot \frac{K}{\mathrm{e}^{2t_n}\varepsilon} \ll 1.$$

In Figs. 6 and 7, for possibilities (A1) and (A2), respectively, and $\varepsilon = 10^{-2}$, we see the relative deviation $|\delta_n - \delta(t_n)| / \delta(t_n)$ (solid red line) along with the error β_n (dashed blue line) for $n = 0, 1, \ldots, N$.

In Fig. 6, it is surprising to have, at the beginning of the integration where $\delta(t_n) \ll 1$, relative deviations not much larger than β_n. This can be explained by observing that

$$\frac{|\delta_n - \delta(t_n)|}{\delta(t_n)} \leqslant \beta_n \left(1 + \frac{\gamma_n}{\beta_n \delta(t_n)}\right), \quad n = 0, 1, 2, \ldots$$

and, whenever

$$0 \notin \varLambda_{j^{**}}^{**}, \quad n \max_{\lambda_i \in \varLambda^*} |\sigma_i| \ll 1, \quad n \max_{\lambda_i \in \varLambda_{j^{**}}^{**}} |\sigma_i| \ll 1$$

we have

$$\frac{\gamma_n}{\beta_n \delta(t_n)} \lessapprox \frac{K_{\varLambda^* \varLambda_{j^{**}}^{**}}}{\mathrm{e}^{(r_{j^{**}} - r_{j^*})t_n} \|Q_{j^{**}} \widehat{z}_0\|_2 \varepsilon}$$

(this bound has been used at point 4 in Remark 4.2 for concluding that (4.12) holds whenever (4.13) holds). Then

$$\frac{\gamma_n}{\beta_n \delta(t_n)} \lessapprox \frac{K_{\varLambda^* \varLambda_{j^{**}}^{**}}}{\|Q_{j^{**}} \widehat{z}_0\|_2 \varepsilon}, \quad n = 0, 1, 2, \ldots$$

for $j^* > j^{**}$. In the possibility (A1) of Fig. 6,

$$\frac{K_{\varLambda^* \varLambda_{j^{**}}^{**}}}{\|Q_{j^{**}} \widehat{z}_0\|_2 \varepsilon} = \frac{\sqrt{10}}{3} \cdot \frac{K}{\varepsilon}$$

is not large.

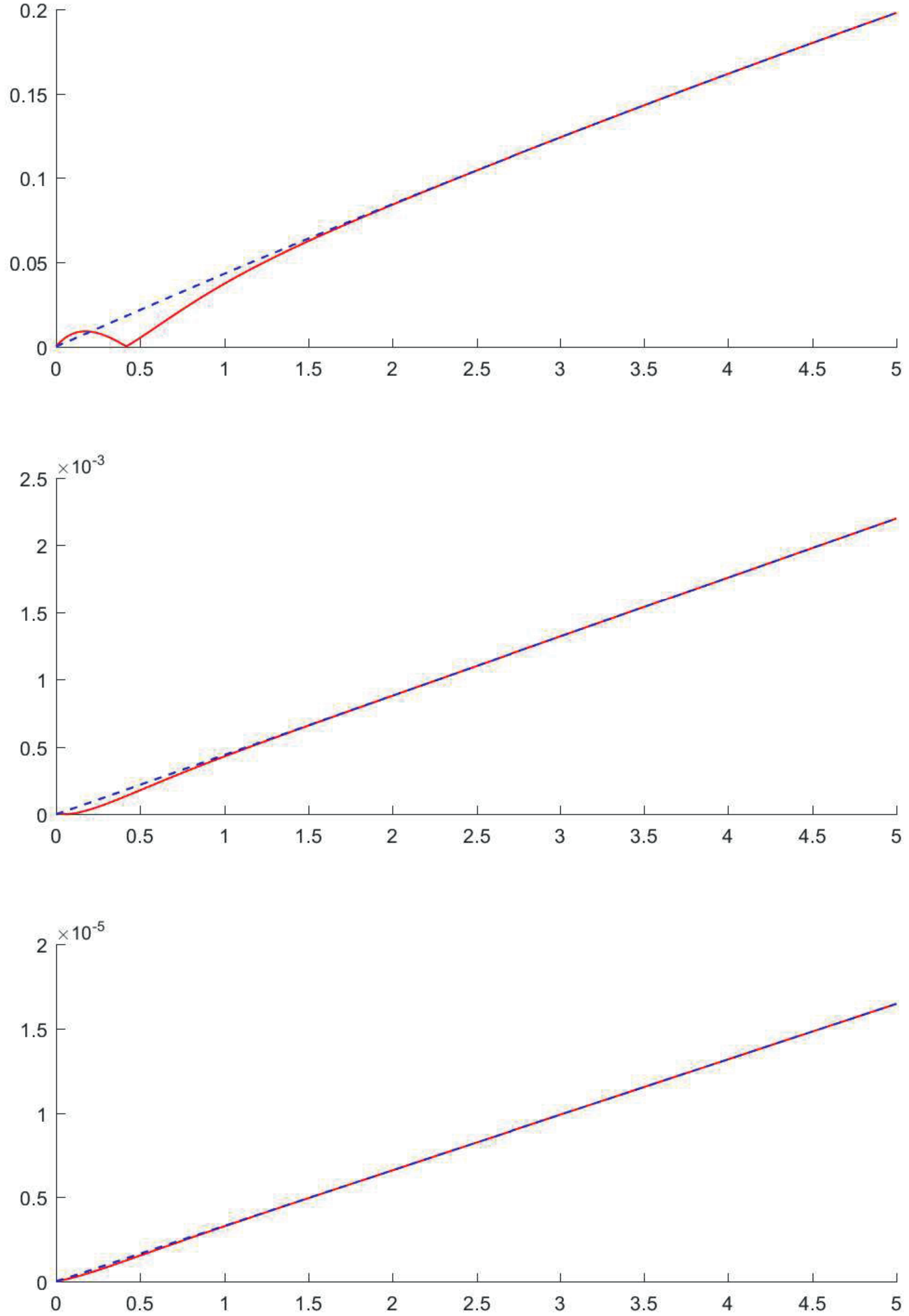

Fig. 6: Possibility (A1) with $y_0 = (1, -1)$ perturbed with $\varepsilon = 10^{-2}$ and $\widehat{z}_0 \in \operatorname{span}(1, 2)$. The l-th row, $l = 1, 2, 3$, corresponds to the approximant of order l. Relative deviations $|\bar{\delta}_n - \bar{\delta}(t_n)| / \bar{\delta}(t_n)$ (solid red line) along with the error β_n (dashed blue line). The abscissas are the times $t_n = nh$, $n = 0, 1, 2, \dots, N$.

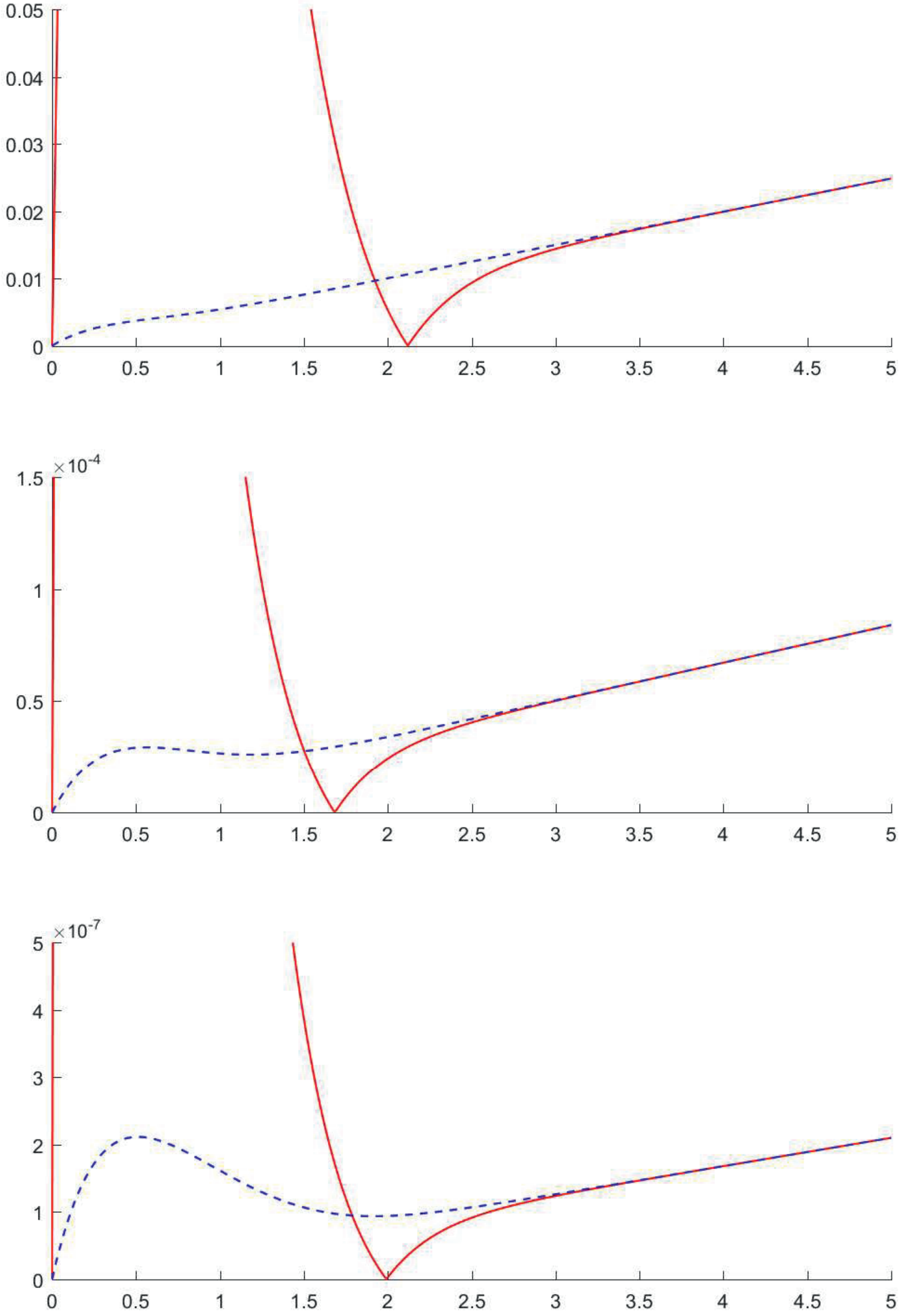

Fig. 7: Possibility (A2) with $y_0 = (1, -1)$ perturbed with $\varepsilon = 10^{-2}$ and $\widehat{z}_0 \in \operatorname{span}(1, 2)$. The l-th row, $l = 1, 2, 3$, corresponds to the approximant of order l. Relative deviations $|\delta_n - \delta(t_n)| / \delta(t_n)$ (solid red line) along with the error β_n (dashed blue line). The abscissas are the times $t_n = nh$, $n = 0, 1, 2, \ldots, N$.

7 A second example

As a second example, we consider ODEs (1.1) with A symmetric arising in consensus problems on networks modelled by graphs (see [3, 17–19]). Such ODEs also appear in describing diffusion signals on graphs (see [16]).

Given an undirected graph with d vertices, the $d \times d$ symmetric matrix A in (1.1) has off-diagonal elements

$$\begin{aligned} a_{ij} &= \begin{cases} 1 & \text{if there is an edge between vertices } i \text{ and } j \\ 0 & \text{otherwise} \end{cases} \\ i, j &= 1, \dots, d, \quad i \neq j \end{aligned} \tag{7.1}$$

and diagonal elements

$$a_{ii} = -\sum_{\substack{j=1 \\ j \neq i}}^{d} a_{ij}, \quad i = 1, \dots, d. \tag{7.2}$$

The matrix A is negative semi-definite and it has zero as rightmost eigenvalue. The solution $y(t)$ of (1.1) converges to $(\mu, \dots, \mu)$ as $t \to +\infty$, μ being the average of the components of y_0. The eigenvectors relevant to the zero eigenvalue are the equilibria $(\mu, \dots, \mu)$, $\mu \in \mathbb{R}$.

In the deformed consensus protocol (see [14]), the matrix A depends on a parameter $s \in \mathbb{R}$ regarded as an input control and it has elements

$$\begin{aligned} a_{ij} &= \begin{cases} s & \text{if there is an edge between nodes } i \text{ and } j \\ 0 & \text{otherwise} \end{cases} \\ i, j &= 1, \dots, d, \quad i \neq j \end{aligned} \tag{7.3}$$

and diagonal elements

$$a_{ii} = s^2 - 1 - s\sum_{\substack{j=1 \\ j \neq i}}^{d} a_{ij}, \quad i = 1, \dots, d. \tag{7.4}$$

When $s = 1$, we have the matrix given in (7.1)–(7.2).

7.1 Numerical experiments

We accomplish numerical experiments in the two possibilities:
(B1) The 8×8 matrix of type (7.3)–(7.4):

$$A = \begin{bmatrix} -1.75 & 0.5 & & & & & & 0.5 \\ 0.5 & -1.75 & 0.5 & & & & & \\ & 0.5 & -1.75 & 0.5 & & & & \\ & & 0.5 & -1.75 & 0.5 & & & \\ & & & 0.5 & -1.75 & 0.5 & & \\ & & & & 0.5 & -1.75 & 0.5 & \\ & & & & & 0.5 & -1.75 & 0.5 \\ 0.5 & & & & & & 0.5 & -1.75 \end{bmatrix}$$

whose eigenvalues are

$$-0.75,\ -1.0429,\ -1.0429,\ -1.75,\ -1.75,\ -2.4571,\ -2.4571,\ -2.75.$$

This possibility, taken from [14], corresponds to a cycle graph with 8 vertices with input control parameter $s = 0.5$ in the deformed consensus protocol.

Tab. 3: Values Σ_1 and Σ for possibility (B1).

l	Σ_1	Σ
1	$2.83 \cdot 10^{-2}$	$3.85 \cdot 10^{-1}$
2	$7.07 \cdot 10^{-5}$	$3.54 \cdot 10^{-3}$
3	$1.33 \cdot 10^{-7}$	$2.44 \cdot 10^{-5}$

Tab. 4: Values Σ_2 and Σ for possibility (B2).

l	Σ_2	Σ
1	$7.28 \cdot 10^{-2}$	$7.29 \cdot 10^{-1}$
2	$4.13 \cdot 10^{-4}$	$1.29 \cdot 10^{-2}$
3	$1.75 \cdot 10^{-6}$	$1.72 \cdot 10^{-4}$

(B2) The 6×6 matrix of type (7.1)–(7.2):

$$A = \begin{bmatrix} -2 & 1 & 1 & 0 & 0 & 0 \\ 1 & -4 & 1 & 1 & 1 & 0 \\ 1 & 1 & -4 & 0 & 1 & 1 \\ 0 & 1 & 0 & -2 & 1 & 0 \\ 0 & 1 & 1 & 1 & -4 & 1 \\ 0 & 0 & 1 & 0 & 1 & -2 \end{bmatrix}$$

whose eigenvalues are

$$0, \ -1.6972, \ -1.6972, \ -4.0000, \ -5.3028, \ -5.3028.$$

This possibility, taken from [17], corresponds to a graph with 6 vertices arranged as an equilateral triangular array, where the vertices 1, 4, and 6 are the vertices of the triangle and the vertices 2, 3, and 5 are the midpoints of the sides of the triangle.

As in the previous section, we consider the Taylor approximants of the exponential of order $l = 1, 2, 3$.

7.1.1 The possibility (B1)

For the possibility (B1), the numerical integration is accomplished with stepsize $h = 1/100$ over $N = 1000$ steps up to $t_N = Nh = 10$. The numbers $\Sigma_1 = N\,|\sigma_1|$ and $\Sigma = N \max_{i=1,\dots,8} |\sigma_i|$ are listed in Table 3.

We consider the initial value $y_0 = (8, 7, 6, 5, 4, 3, 2, 1)$.

For this initial value unperturbed, in Fig. 8 we see in logarithmic scale the error γ_n (solid red line) along with $\frac{n}{N}\Sigma_1$ (dashed blu line) and $\frac{n}{N}\Sigma$ (dash-dotted green line) for $n = 0, 1, \dots, N$. The same pattern of the example of Section 6 is observed: in the long-time, the error grows as $\frac{n}{N}\Sigma_1$ with n.

Now, suppose that the initial value y_0 is perturbed with $\widehat{z}_0 \in \operatorname{span}(2, -5, -1, 2, -3, 4, -1, -2)$ and $\varepsilon = 10^{-2}$. In Fig. 9, we see, for $n = 0, 1, 2, \dots, N$, in the left column the deviation $\left|\delta_n - \gamma_n\right|$ (solid red line) along with the error $\delta(t_n)$ (dashed blu line) and in the right column the deviation $\left|\delta_n - \delta(t_n)\right|$ (solid red line) along with the error γ_n (dashed blue line). As in the example of Section 6, we observe exactly what is described in points 2 and 3 of Remark 4.2.

7.1.2 The possibility (B2)

For the possibility (B2), the numerical integration is accomplished with stepsize $h = 1/100$ over $N = 500$ steps up to $t_N = Nh = 5$. Since the rightmost eigenvalue is zero, we have $\sigma_1 = 0$. The numbers $\Sigma_2 = N\,|\sigma_2|$ and $\Sigma = N \max_{i=2,\dots,6} |\sigma_i|$ are listed in Table 4.

We consider the initial value $y_0 = (3, 2, 1, -1, -2, -3)$. Since the average of the components of y_0 is equal to zero, the solution $y(t)$ of (1.1) converges to zero as $t \to +\infty$. Moreover, since y_0 is orthogonal to the equilibria, we have $j^* = 2$.

For this initial value unperturbed, in Fig. 10 we see in logarithmic scale the error γ_n (solid red line) along with $\frac{n}{N}\Sigma_2$ (dashed blue line) and $\frac{n}{N}\Sigma$ (dash-dotted green line) for $n = 0, 1, \dots, N$.

Suppose that the initial value y_0 is perturbed by $\varepsilon = 10^{-2}$ and

$$\widehat{z}_0 \in \operatorname{span}(2, -5, -1, 4, -1, -2).$$

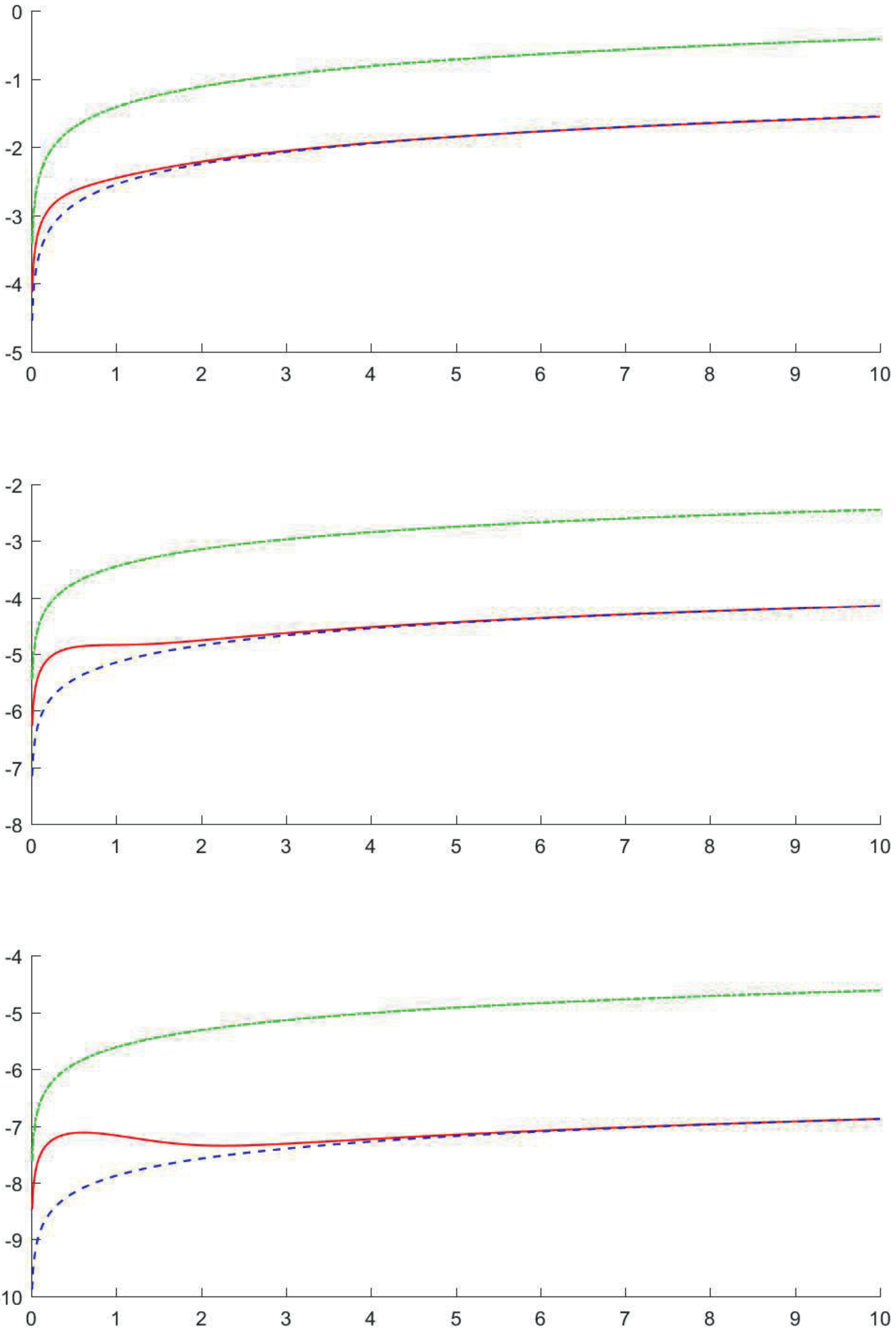

Fig. 8: Possibility (B1) with initial value $y_0 = (8, 7, 6, 5, 4, 3, 2, 1)$ unperturbed. The l-th row, $l = 1, 2, 3$, corresponds to the approximant of order l. Logarithmic error $\log_{10}(\gamma_n)$ (solid red line) along with $\log_{10}\left(\frac{n}{N}\Sigma_1\right)$ (dashed blue line) and $\log_{10}\left(\frac{n}{N}\Sigma\right)$ (dash-dotted green line). The abscissas are the times $t_n = nh$, $n = 0, 1, 2, \ldots, N$.

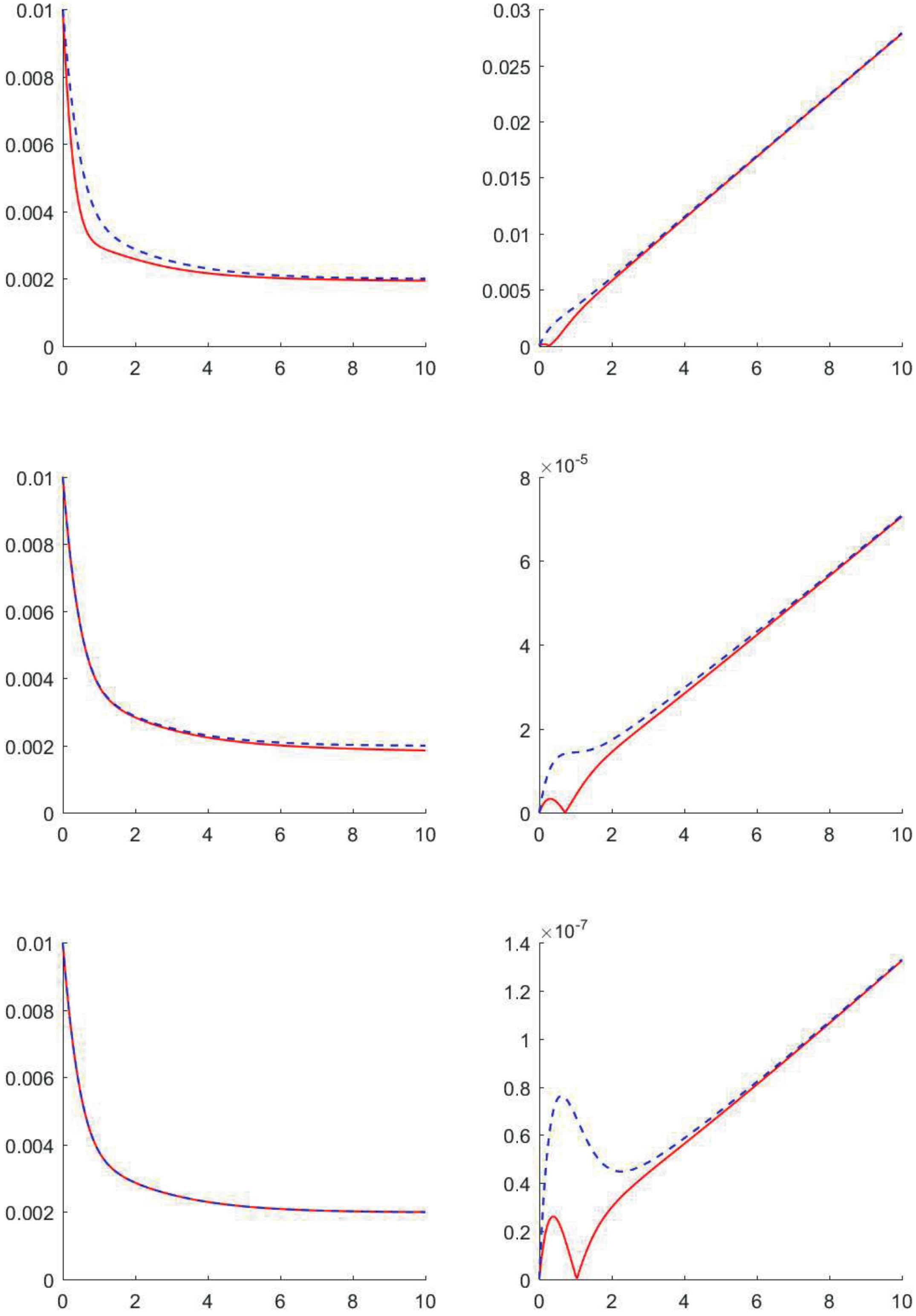

Fig. 9: Possibility (B1) with $y_0 = (8, 7, 6, 5, 4, 3, 2, 1)$ perturbed with $\varepsilon = 10^{-2}$ and $\widehat{z}_0 \in \text{span}\,(2, -5, -1, 2, -3, 4, -1, -2)$. The l-th row, $l = 1, 2, 3$, corresponds to the approximant of order l. Left column: deviation $\left|\tilde{\delta}_n - \gamma_n\right|$ (solid red line) along with the error $\delta\,(t_n)$ (dashed blue line). Right column: deviation $|\tilde{\delta}_n - \delta\,(t_n)|$ (solid red line) along with the error γ_n (dashed blue line). The abscissas are the times $t_n = nh$, $n = 0, 1, 2, \ldots, N$.

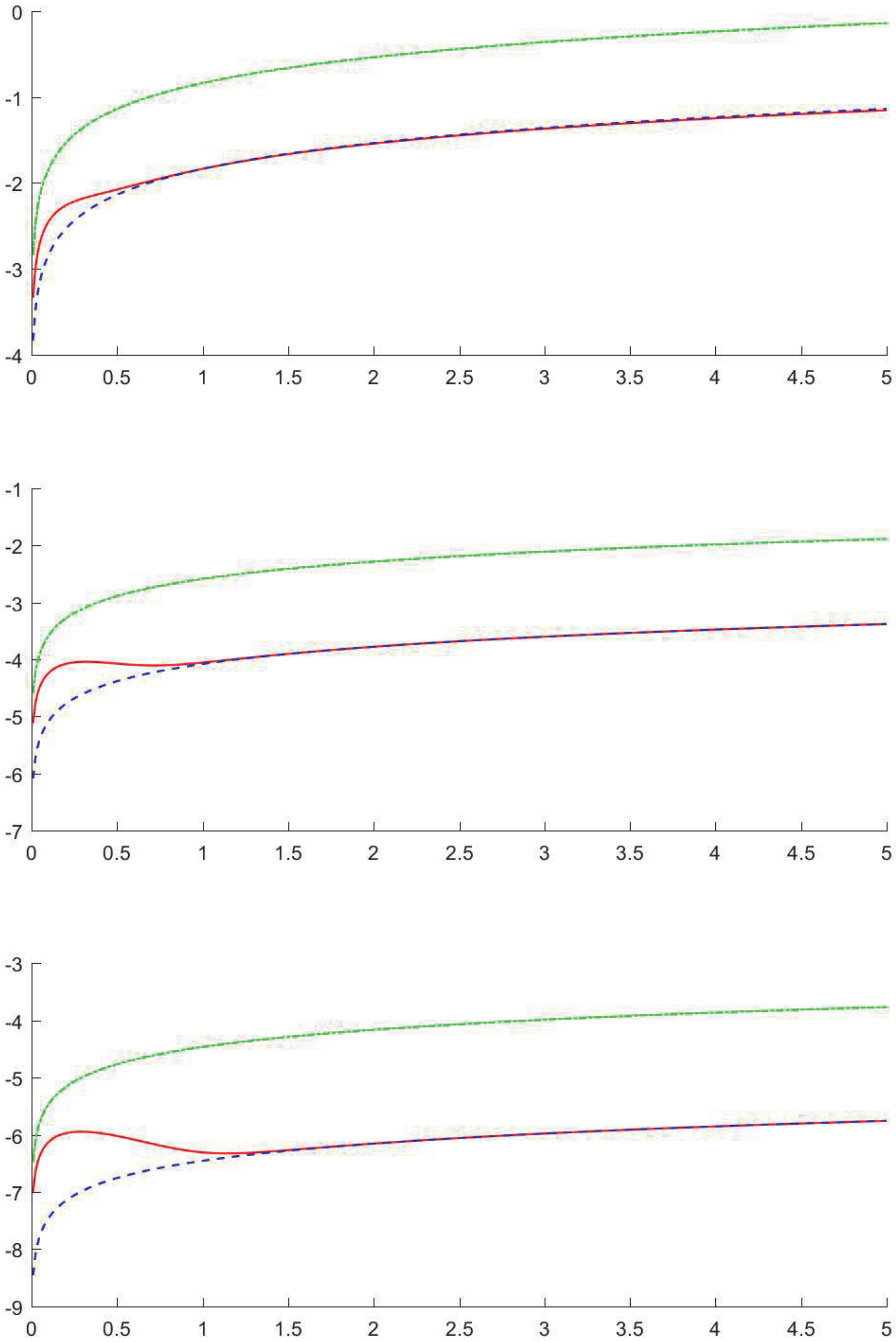

Fig. 10: Possibility (B2) with initial value $y_0 = (3, 2, 1, -1, -2, -3)$ unperturbed. The l-th row, $l = 1, 2, 3$, corresponds to the approximant of order l. Logarithmic error $\log_{10}(\gamma_n)$ (solid red line) along with $\log\left(\frac{n}{N}\Sigma_2\right)$ (dashed blue line), and $\log_{10}\left(\frac{n}{N}\Sigma\right)$ (dashed-dot green line). The abscissas are the times $t_n = nh$, $n = 0, 1, 2, \ldots, N$.

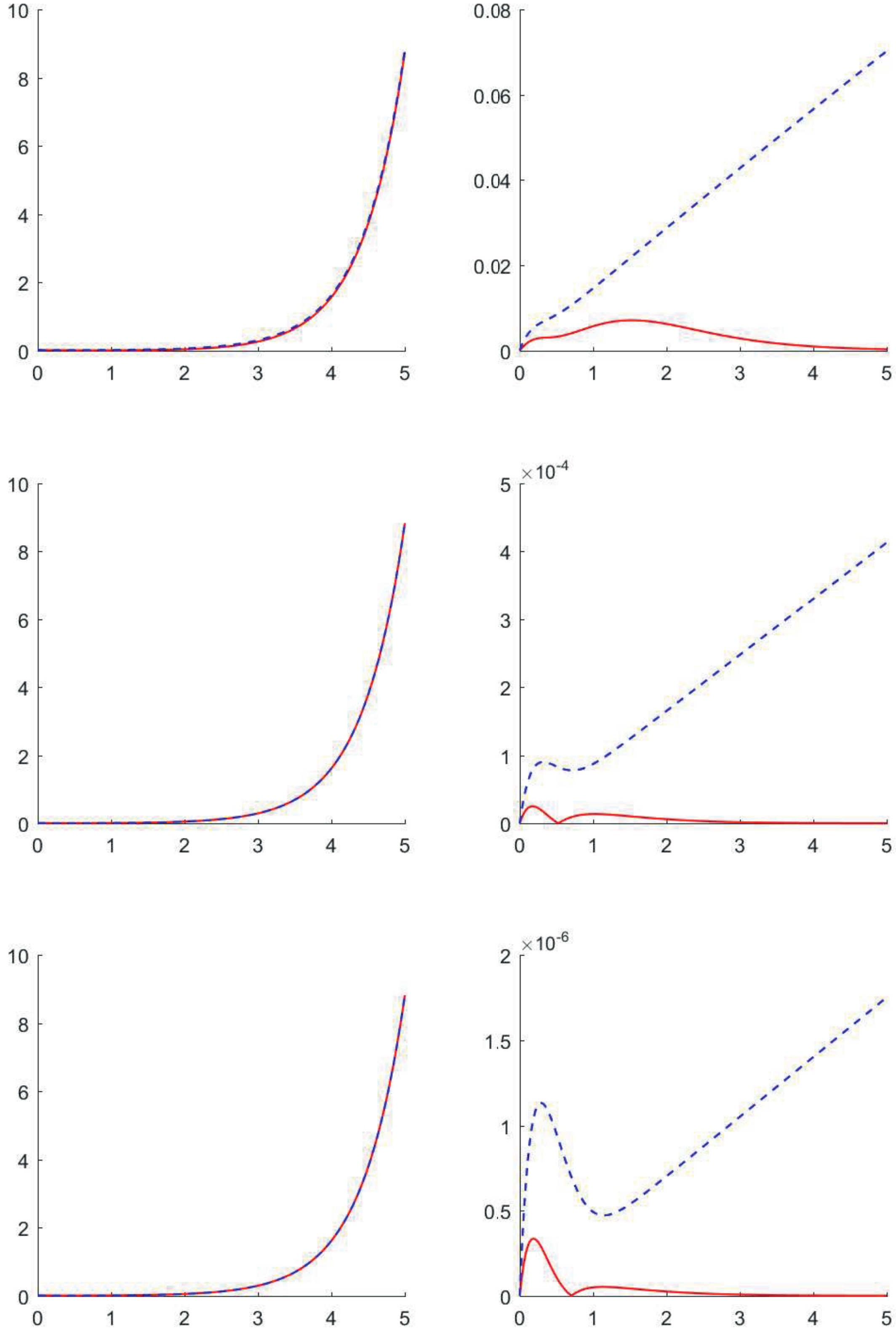

Fig. 11: Possibility (B2) with $y_0 = (3, 2, 1, -1, -2, -3)$ perturbed with $\varepsilon = 10^{-2}$ and $\widehat{z}_0 \in \operatorname{span}(2, -5, -1, 4, -1, -2)$. The l-th row, $l = 1, 2, 3$, corresponds to the approximant of order l. Left column: deviation $\left|\delta_n - \gamma_n\right|$ (solid red line) along with the error $\delta(t_n)$ (dashed blue line). Right column: deviation $|\delta_n - \delta(t_n)|$ (solid red line) along with the error γ_n (dashed blue line). The abscissas are the times $t_n = nh$, $n = 0, 1, 2, \ldots, N$.

In Fig. 11, we see, as in Fig. 8 for (B1), $|\delta_n - \gamma_n|$ along with $\delta(t_n)$ in the left side and $|\delta_n - \delta(t_n)|$ along with γ_n in the right side.

8 Conclusion

For the ODE (1.1) with A normal, we have presented a relative error analysis of the numerical solution, over a mesh $t_n = nh$, $n = 1, 2, \ldots$, of constant stepsize h, obtained by using an analytic approximant of the exponential at each step. A possible perturbation in the initial value is taking into account.

The two sources of error are the approximant, whose error is measured by the numbers σ_i defined in (3.2), and the perturbation in the initial value, whose error is measured by (1.2).

Our analysis has involved three relative errors:

- the relative error γ_n of the unperturbed numerical solution, which is defined in (1.12);
- the relative error $\delta(t)$ of the perturbed exact solution, which is defined in (1.3);
- the relative error δ_n of the perturbed numerical solution, which is defined in (1.11).

We have shown that *the relative error γ_n grows linearly in time and, in the long-time, it depends only on the errors σ_i relevant to rightmost eigenvalues*. Moreover, we have shown how the growth of δ_n is related to the growth of γ_n and $\delta(t_n)$.

Our relative error analysis covers the situation where

$$\max_{\lambda_i \in \Lambda} |\sigma_i| \ll 1 \tag{8.1}$$

with Λ the spectrum of A. We call it the *non-stiff situation*. However, our analysis does not cover the situation where (8.1) does not hold, but

$$\max_{\lambda_i \in \Lambda_1} |\sigma_i| \ll 1 \tag{8.2}$$

holds, with Λ_1 the set of the rightmost eigenvalues. We call it the *stiff situation*.

In the stiff situation, it is fundamental to understand whether, in the long-time, the relative error γ_n depends only on the errors σ_i relevant to the rightmost eigenvalues, namely the eigenvalues in Λ_1. If this happens, γ_n is small in the long-time, although the stepsize h is tuned for having (8.2) only, not (8.1). Of course, since (8.1) does not hold, the error γ_n can be not small at the beginning of the integration.

The relative error analysis of the present paper is continued in [10–12]. The paper [10] studies the long-time behavior of the error γ_n in the stiff situation. The paper [11] shows how the order stars (see [8, 20]) are involved in the relative error analysis. The paper [12] presents a relative error analysis for the numerical integration of long-time solutions to be used in the stiff situation.

Funding: The author thanks INdAM-GNCS and the University of Trieste for the financial support.

References

[1] F. Bürgisser and F. Cucker, *Condition. The Geometry of Numerical Algorithms*, Springer, 2013.

[2] L. Dieci and A. Papini, Padé approximation for the exponential of a block triangular matrix, *Linear Algebra Appl.*, **308** (2000), 183–202.

[3] L. Dieci and L. Lopez, Numerical integration of networks of differential equations. Submitted.

[4] B. Kagstrom, Bounds and perturbations for the matrix exponential, *BIT*, **17** (1977), 39–57.

[5] N. J. Higham, *Functions of Matrices. Theory and Computation*, SIAM, 2008.

[6] N. J. Higham, The scaling and squaring method for the matrix exponential revisited, *SIAM Review*, **51** (2009), 747–764.

[7] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations II. Stiff Differential-Algebraic Problems.* Springer-Verlag, Berlin–Heidelberg, 1996.

[8] A. Iserles and S. Norsett. *Order Stars: Theory and Applications*, Chapman and Hall, 1991.

[9] S. Maset, Conditioning and relative error propagation in linear autonomous ordinary differential equations, *Discrete and Continuous Dynamical Systems B*, **23** (2018), 2879–2909.

[10] S. Maset, Relative error long-time behavior in matrix exponential approximations for numerical integration: the stiff situation. Submitted.

[11] S. Maset, Further results on the relative error analysis of matrix exponential approximations for numerical integration. In preparation.

[12] S. Maset, Relative error stability and instability of matrix exponential approximations for stiff numerical integration of long-time solutions, *J. Comp. Appl. Math.*, bf390 (2021), 113387.

[13] C. Moler and C. Van Loan, Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later, *SIAM Review*, **45** (2003), 3–49.

[14] F. Morbidi, The deformed consensus protocol, *Automatica*, **49** (2013), 3049–3055.

[15] L. Shampine, I. Gladwell, and S. Thompson, *Solving ODEs with MATLAB*, Cambridge University Press, 2003.

[16] D. Thanou, X. Dong, D. Kressner, and P. Frossard, Learning heat diffusion graphs, *IEEE Trans. Signal Inform. Processing Networks*, **3** (2017), 484–499.

[17] R. Olfati-Saber and R. M. Murray, Consensus protocols for networks of dynamic agents, *Proc. American Control Conference*, **2** (2003), 951–956.

[18] R. Olfati-Saber and J. S. Shamma, Consensus filters for sensor networks and distributed sensor fusion. In: *Proc. of the 44th IEEE Conf. on Decision and Control, and the European Control Conference, CDC-ECC'05*, 2005, pp. 6698–6703.

[19] R. Olfati-Saber, J. A. Fax, and R. M. Murray, Consensus and cooperation in networked multi-agent systems. In: *Proc. of the IEEE*, 95 (2007), 215–233.

[20] G. Wanner, E. Hairer, and S. Norsett, Order stars and stability theorems, *BIT*, **18** (1978), 475–489.

[21] R. Ward, Numerical computation of the matrix exponential with accuracy estimates, *SIAM J. Numer. Anal.*, **14** (1977), 600–610.