Calcolo (2022) 59: 23
https://doi.org/10.1007/s10092-022-00466-5

Relative error long-time behavior in matrix exponential approximations for numerical integration: the stiff situation

S. Maset[1]

Received: 11 October 2020 / Revised: 15 October 2021 / Accepted: 7 April 2022 /
Published online: 23 May 2022

Abstract
In the stiff situation, we consider the long-time behavior of the relative error γ_n in the numerical integration of a linear ordinary differential equation $y'(t) = Ay(t), \quad t \geq 0$, 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. We study the long-time behavior of γ_n by comparing it to the relative error γ_n^{long} in the numerical integration of the long-time solution, i.e. the projection of the solution on the eigenspace of the rightmost eigenvalues. The error γ_n^{long} grows linearly in time, it is small and it remains small in the long-time. We give a condition under which $\gamma_n \approx \gamma_n^{\text{long}}$, i.e. $\frac{\gamma_n}{\gamma_n^{\text{long}}} \approx 1$, in the long-time. When this condition does not hold, the ratio $\frac{\gamma_n}{\gamma_n^{\text{long}}}$ is large for all time. These results describe the long-time behavior of the relative error γ_n in the stiff situation.

Keywords Relative error · Linear ordinary differential equations · Numerical integration · Approximation of the matrix exponential · Stiff problems · Long-time behavior

Mathematics Subject Classification 65F60 · 65L04 · 65L05 · 65L06 · 65L20 · 65L70

✉ S. Maset
maset@units.it

[1] Dipartimento di Matematica e Geoscienze, Università di Trieste, Trieste, Italy

1 Introduction

Consider the ordinary differential equation (ODE)

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

where $A \in \mathbb{R}^{d\times d}$ and $y(t) \in \mathbb{R}^d$, and consider, over the mesh

$$t_n = nh, \ n = 0, 1, 2, \dots,$$

of constant stepsize $h > 0$, a numerical solution of (1.1) given by

$$y_n = R(hA)^n y_0, \quad n = 0, 1, 2, \dots, \tag{1.2}$$

where $R : \mathscr{D} \subseteq \mathbb{C} \to \mathbb{C}$ is a analytic approximant of the exponential e^z, $z \in \mathbb{C}$. When the numerical solution is obtained by a Runge–Kutta (RK) method, the approximant R is the stability function of the RK method and it is a polynomial or a rational function.

The paper [9] analyzed in the non-stiff situation the time behavior of the normwise relative error

$$\gamma_n = \frac{\left\| y_n - y(t_n) \right\|_2}{\left\| y(t_n) \right\|_2}, \ n = 0, 1, 2, \dots, \tag{1.3}$$

in case of a normal matrix A. It seems to be the first paper in literature dealing in detail with the relative error time behavior of numerical solutions of ODEs. This is quite surprising because relative errors are generally considered better than absolute errors as quality measures of approximations. Indeed, componentwise relative errors are involved in the stepsize control mechanism (see [12]).

The present paper continues to analyze, in case of A normal, the error γ_n by considering its long-time behavior in the stiff situation. Next subsection, with all its subsubsections, contains the basic material for facing such an analysis. Part of this material was introduced in [9].

1.1 Fundamental notations and notions

1.1.1 Small and large

We set that, for $a \geq 0$, "a is small" is the same as "$a \ll 1$ " and "a is large" is the same as "$a \gg 1$".

For $b \geq 0$ and $c > 0$, $b \ll c$ means $\frac{b}{c} \ll 1$.

1.1.2 The notation $\approx$

For $a, b \in \mathbb{R}$, $a \approx b$ means

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

with $|e| \ll 1$. We say $a \approx b$ *with degree* ε, where $\varepsilon > 0$, if $|e| \leq \varepsilon$.

Moreover, $a \lessapprox b$ means $a \leq c$ and $c \approx b$ for some $c \in \mathbb{R}$.

For $a, b \in \mathbb{R}^d$, $a \approx b$ means

$$\frac{\|a - b\|_2}{\|b\|_2} \ll 1.$$

We say $a \approx b$ with degree ε, where $\varepsilon > 0$, if $\frac{\|a-b\|_2}{\|b\|_2} \leq \varepsilon$.

1.1.3 The meaning of "it is expected"

In the paper, we often say "it is expected S", where S is a statement, with the meaning that the statement not S is "unlikely" or "unusual" or "extreme".

Sentences of this form can seem vague, although they are able to convey significant information. However, they are never used in definitions or theorems, which are stated in a precise manner without any such type of vagueness. The sentences are used for a better understanding of technical notions and results.

By introducing probability measures on data, we could made "it is expected S" mathematically precise, but this is out of the scope of the present paper.

1.1.4 The spectrum of *A*

The spectrum

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

of the normal matrix A, where $\lambda_1, \lambda_2, \ldots, \lambda_p$ are the distinct eigenvalues of A, is partitioned by decreasing real part in the subsets $\Lambda_1, \Lambda_2, \ldots, \Lambda_q$ (see Fig. 1): we have

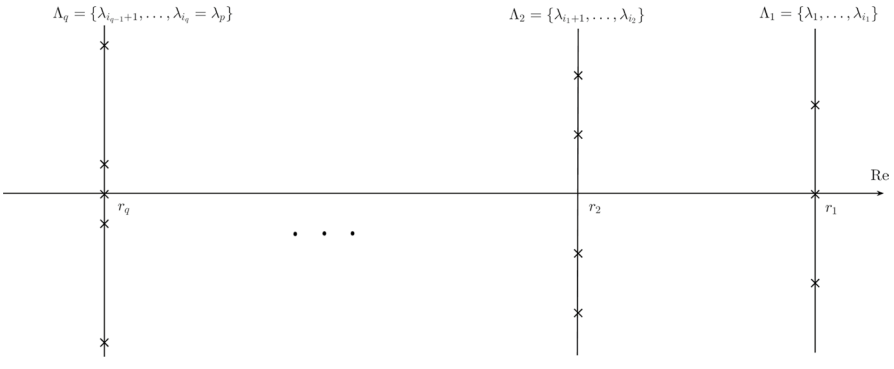

Fig. 1 Spectrum of A partitioned by decreasing real parts

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

with $0 = i_0 < i_1 < \cdots < i_q = p$, and

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

For $i = 1, \dots, p$, let P_i be the orthogonal projection on the eigenspace of λ_i. For $j = 1, \dots, q$, let

$$Q_j := \sum_{\lambda_i \in \Lambda_j} P_i.$$

For a nonempty subset Γ of Λ, let

$$\rho_\Gamma := \max_{\lambda_i \in \Gamma} |\lambda_i| \quad \text{and} \quad \mu_\Gamma := \min_{\lambda_i \in \Gamma} |\lambda_i|. \tag{1.4}$$

1.1.5 The initial value y_0

We assume $y_0 \neq 0$. Thus $y(t) = \mathrm{e}^{tA} y_0 \neq 0$, for any t, and the relative error γ_n is defined for any n. Let

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

be the *normalized initial value*.

Let

$$\begin{aligned}\Lambda^* &:= \big\{\lambda_i \in \Lambda : P_i y_0 \neq 0\big\}\\ \Lambda_j^* &:= \Lambda_j \cap \Lambda^*,\ j = 1, \dots, q.\end{aligned}$$

The generic situation for the initial value y_0 is $\Lambda^* = \Lambda$. In order to use simpler notations, we assume this generic situation.

If it does not hold, then below we have to see $\Lambda_1, \dots, \Lambda_q$ as $\Lambda_1^*, \dots, \Lambda_q^*$ without the sets Λ_j^* that are empty. In other words, we see Λ_1 as $\Lambda_{j_1^*}$ where

$$j_1^* := \min\{j \in \{1, \dots, q\} : \Lambda_j^* \neq \emptyset\},$$

Λ_2 as $\Lambda_{j_2^*}$ where

$$j_2^* := \min\{j \in \{j_1^* + 1, \dots, q\} : \Lambda_j^* \neq \emptyset\},$$

and so on. Of course, when we do this, the number q of sets in $\Lambda_1, \dots, \Lambda_q$ is no longer equal to the number of possible real parts in the spectrum Λ, but it is equal to the number of possible real parts in Λ^*.

1.1.6 Rightmost and non-rightmost eigenvalues

The set Λ_1 is the set of the *rightmost eigenvalues*. The set

$$\Lambda^- := \Lambda \backslash \Lambda_1 = \bigcup_{j=2}^{q} \Lambda_j$$

is the set of the *non-rightmost eigenvalues*. We assume $q > 1$ in order to have $\Lambda^- \neq \emptyset$.

By recalling the definitions (1.4), we set

$$\begin{aligned} \rho &:= \rho_\Lambda \quad \text{and} \quad \mu := \mu_\Lambda \\ \rho_1 &:= \rho_{\Lambda_1} \quad \text{and} \quad \mu_1 := \mu_{\Lambda_1} \\ \rho^- &:= \rho_{\Lambda^-} \quad \text{and} \quad \mu^- := \mu_{\Lambda^-}. \end{aligned} \tag{1.5}$$

1.1.7 The numbers β_j

For $j = 2, \dots, q$, i.e. for any non-rightmost real part, let

$$\beta_j := \frac{r_j - r_1}{\rho_1}.$$

Observe that

$$0 > \beta_2 > \cdots > \beta_q.$$

It is expected $|\beta_2|$ non-small.

1.1.8 Dimensionless quantities

We use the dimensionless stepsize $h\rho_1$, or $h\rho$, and the dimensionless time $t\rho_1$, or $t\rho$, rather than the stepsize h and the time t, respectively, because they are small or large independently of the unit used for time.

In this paper, when we say that a certain quantity is small or large, this quantity is always dimensionless.

The numbers β_j defined above are dimensionless, as well as the errors σ_i now introduced.

1.1.9 The errors σ_i

We assume that the approximant R has order l, where l is a positive integer. This means

$$R(z) - \mathrm{e}^z = Cz^{l+1} + O\big(z^{l+2}\big),\ z \to 0, \tag{1.6}$$

with $C \neq 0$. It is assumed that the domain $\mathscr{D}$ of R includes a neighborhood of zero. Moreover, we assume $h\lambda_i \in \mathscr{D}$, $i = 1, \dots, p$.

We introduce the complex numbers

$$\sigma_i := \log S\big(h\lambda_i\big),\ i = 1, \dots, p, \tag{1.7}$$

where

$$S(z) = \mathrm{e}^{-z}R(z),\ z \in \mathscr{D}, \tag{1.8}$$

is the *relative approximant*. The numbers σ_i are logarithmic errors of R as an approximant of the exponential, since

$$\sigma_i = \log R\big(h\lambda_i\big) - \log \mathrm{e}^{h\lambda_i},\ i = 1, \dots, p.$$

For a nonempty subset Γ of Λ, we have

$$\begin{aligned} \max_{\lambda_i \in \Gamma} \big|\sigma_i\big| &= |C|\big(h\rho_\Gamma\big)^{l+1}\big(1 + O\big(h\rho_\Gamma\big)\big) \\ \min_{\lambda_i \in \Gamma} \big|\sigma_i\big| &= |C|\big(h\mu_\Gamma\big)^{l+1}\big(1 + O\big(h\rho_\Gamma\big)\big). \end{aligned} \tag{1.9}$$

as $h\rho_\Gamma \to 0$, where ρ_Γ and μ_Γ are defined in (1.4).

1.1.10 Local relative errors and global relative errors

As particular cases of (1.9), we obtain

$$\max_{\lambda_i \in \Lambda_1} \big|\sigma_i\big| = |C|\big(h\rho_1\big)^{l+1}\big(1 + O\big(h\rho_1\big)\big),\ h\rho_1 \to 0, \tag{1.10}$$

and

$$\max_{\lambda_i \in \Lambda} \big|\sigma_i\big| = |C|(h\rho)^{l+1}(1 + O(h\rho)), h\rho \to 0. \tag{1.11}$$

We introduce

$$E_1 := \frac{\max\limits_{\lambda_i \in \Lambda_1} |\sigma_i|}{h\rho_1}$$

and

$$E := \frac{\max\limits_{\lambda_i \in \Lambda} |\sigma_i|}{h\rho}$$

We can consider $\max_{\lambda_i \in \Lambda_1} |\sigma_i|$ and $\max_{\lambda_i \in \Lambda} |\sigma_i|$ as *local relative errors*, and E_1 and E as *global relative errors*, of the numerical integration. An explanation for this is given below at points 2 of Remarks 1.1 and 1.2.

1.1.11 The ratios K_1 and K

When $0 \notin \Lambda_1$, let

$$K_1 := \frac{\max_{\lambda_i \in \Lambda_1} |\sigma_i|}{\min_{\lambda_i \in \Lambda_1} |\sigma_i|} = \left(\frac{\rho_1}{\mu_1}\right)^{l+1} \big(1 + O(h\rho_1)\big),\ h\rho_1 \to 0, \tag{1.12}$$

The right-hand side follows by (1.9). Observe that the generic situation for the matrix A is to have Λ_1 constituted by a real eigenvalue or by a unique pair of complex conjugate eigenvalues. In this generic situation, we have $K_1 = 1$.

When $0 \notin \Lambda$, let

$$K := \frac{\max_{\lambda_i \in \Lambda} |\sigma_i|}{\min_{\lambda_i \in \Lambda} |\sigma_i|} = \left(\frac{\rho}{\mu}\right)^{l+1} (1 + O(h\rho)),\ h\rho \to 0.$$

1.1.12 The ratios M_i and M

For $\lambda_i \in \Lambda^-$, i.e. λ_i is a non-rightmost eigenvalue, let

$$M_i := \frac{|\sigma_i|}{\max_{\lambda_k \in \Lambda_1} |\sigma_k|} = \left(\frac{|\lambda_i|}{\rho_1}\right)^{l+1} (1 + O(h\rho)),\ h\rho \to 0. \tag{1.13}$$

Moreover, let

$$M := \max_{\lambda_i \in \Lambda^-} M_i = \frac{\max_{\lambda_i \in \Lambda^-} |\sigma_i|}{\max_{\lambda_k \in \Lambda_1} |\sigma_k|} = \left(\frac{\rho^-}{\rho_1}\right)^{l+1} (1 + O(h\rho)),\ h\rho \to 0. \tag{1.14}$$

1.1.13 The base situation

We call *base situation* the situation where $\max_{\lambda_i \in \Lambda_1} |\sigma_i|$ is small.

Here are some observations about the base situation.

– In the base situation, it is expected E_1 small, i.e. $\max_{\lambda_i \in \Lambda_1} |\sigma_i| \ll h\rho_1$, and $h\rho_1$ non-large. Look at (1.10).

– We do not say that in the base situation it is expected $h\rho_1$ small. In fact, we do not see the case where $\max_{\lambda_i \in \Lambda_1} |\sigma_i|$ is small and $h\rho_1$ is not small as "unusual", when R is an high order approximant.

1.1.14 The non-stiff situation and the stiff situation

The base situation is partitioned in two disjoint sub-situations: the non-stiff situation and the stiff situation.

We call *non-stiff situation* (*stiff situation*) the sub-situation of the base situation where $\max_{\lambda_i \in \Lambda} |\sigma_i|$ is small ($\max_{\lambda_i \in \Lambda} |\sigma_i|$ is not small), equivalently $\max_{\lambda_i \in \Lambda^-} |\sigma_i|$ is small ($\max_{\lambda_i \in \Lambda^-} |\sigma_i|$ is not small).

The non-stiff situation and the stiff situation correspond to what is meant as non-stiff and stiff in the traditional terminology of numerical ODEs. The explanation is given below at point 3 of Remark 1.2.

Here are some observations about the non-stiff and stiff situations.

– In the non-stiff situation, it is expected E small, i.e. $\max_{\lambda_i \in \Lambda} |\sigma_i| \ll h\rho$, and $h\rho$ non-large. Look at (1.11).
– In the non-stiff situation, it is expected

$$\frac{\max_{\lambda_i \in \Lambda^-} |\sigma_i|}{h\rho_1} = ME_1 \tag{1.15}$$

small. In fact, it is expected E_1 small and then to have both $\max_{\lambda_i \in \Lambda^-} |\sigma_i|$ and $\max_{\lambda_i \in \Lambda_1} |\sigma_i|$ small with their ratio M not satisfying $M \ll \frac{1}{E_1}$ appears to be an "extreme" case.
– In the stiff situation, it is expected $h\rho$ non-small. In fact, to have $\max_{\lambda_i \in \Lambda} |\sigma_i|$ non-small with $h\rho$ small appear to be "unlikely".
– In the stiff situation, M is large since it the ratio between a non-small number and a small number.

1.1.15 The function *g*

Let

$$g(c) := \frac{\mathrm{e}^c - 1 - c}{c}, \ c \geq 0.$$

The function g is increasing with $g(0) = 0$. We have $g(c) \approx \frac{c}{2}$ for c small, $g(1) = 0.71828$ and $g(c) = 1$ for $c = 1.2564$.

1.2 Analysis of the error γ_n

After having introduced the basic material in the previous subsection, we can proceed with our analysis of the error γ_n.

Next theorem (it is Theorem 4.1 in [9] stated with E instead of $\max_{\lambda_i\in\Lambda}|\sigma_i|$) describes how the error γ_n grows in time.

Theorem 1.1 *Assume* $0\notin\Lambda$. *Fix* $c>0$. *For* $t_n\rho\le\frac{c}{E}$, *we have*

$$\frac{t_n\rho E}{K}(1-g(c))\le\gamma_n\le t_n\rho E(1+g(c)).$$

The theorem with $c=1$ reads

$$0.28172\cdot\frac{t_n\rho E}{K}\le\gamma_n\le 1.7183\cdot t_n\rho E \tag{1.16}$$

for $t_n\rho\le\frac{1}{E}$.

If $E\ll 1$, then (1.16) says that γ_n is small and grows linearly in time up to large times $t_n\rho$, precisely up to the large time $\frac{1}{E}$. This result is useful in the non-stiff situation, where it is expected $E\ll 1$.

Remark 1.1

1. By taking a small c in the previous theorem, we have

$$\frac{t_n\rho E}{K}\lessapprox\gamma_n\lessapprox t_n\rho E.$$

To be more precise, this holds for times $t_n\rho\le x$, where $x>0$ is such that $xE\ll 1$.

2. After one step ($n=1$), we have

$$\frac{\max_{\lambda_i\in\Lambda}|\sigma_i|}{K}(1-g(c))\le\gamma_1\le\max_{\lambda_i\in\Lambda}|\sigma_i|(1+g(c)).$$

This explains because $\max_{\lambda_i\in\Lambda}|\sigma_i|$ can be considered as local relative error in the numerical integration of the solution. At $t_n\rho=1$, we have

$$\frac{E}{K}(1-g(c))\le\gamma_n\le E(1+g(c)).$$

This explains because E can be considered as global relative error in the numerical integration of the solution.

3. The theorem assumes $0\notin\Lambda$. If $\Lambda=\{0\}$, we have $\gamma_n=0$ for any n. For the case $0\in\Lambda$ and $\Lambda\neq\{0\}$, see point 5 of Remark 4.1 in [9].

1.2.1 The long-time solution

Let y^{long} be the solution of (1.1) with initial value $Q_1 y_0$ instead of y_0.

The solution y^{long} is the *long-time solution* of (1.1), since we have $y(t) \approx y^{\text{long}}(t)$ for $t\rho_1$ large. In particular, we have $y(t) \approx y^{\text{long}}(t)$ with degree ε, where $\varepsilon > 0$, if

$$\sqrt{\sum_{j=2}^{q} \left(\mathrm{e}^{(r_j - r_1)t} \frac{\left\| Q_j \widehat{y}_0 \right\|_2}{\left\| Q_1 \widehat{y}_0 \right\|_2} \right)^2} = \sqrt{\sum_{j=2}^{q} \left(\mathrm{e}^{\beta_j t \rho_1} \frac{\left\| Q_j \widehat{y}_0 \right\|_2}{\left\| Q_1 \widehat{y}_0 \right\|_2} \right)^2} \leq \varepsilon \tag{1.17}$$

(see Theorem 5.1 in [9]). Observe that the left-hand side of (1.17) goes to zero as $t\rho_1 \to +\infty$.

1.2.2 The error γ_n^{long}

Let γ_n^{long} be the error γ_n of the long-time solution y^{long}. Next theorem (it is Theorem 5.2 in [9] stated with E_1 instead of $\max_{\lambda_i \in \Lambda_1} |\sigma_i|$) describes how the error γ_n^{long} grows in time.

Theorem 1.2 *Assume* $0 \notin \Lambda_1$. *Fix* $c > 0$. *For* $t_n \rho_1 \leq \frac{c}{E_1}$, *we have*

$$\frac{t_n \rho_1 E_1}{K_1} (1 - g(c)) \leq \gamma_n^{\text{long}} \leq t_n \rho_1 E_1 (1 + g(c)).$$

The theorem with $c = 1$ reads

$$0.28172 \cdot \frac{t_n \rho_1 E_1}{K_1} \leq \gamma_n^{\text{long}} \leq 0.28172 \cdot t_n \rho_1 E_1 \tag{1.18}$$

for $t_n \rho \leq \frac{1}{E_1}$.

If $E_1 \ll 1$, then (1.18) says that γ_n^{long} is small and grows linearly in time up to large times $t_n \rho_1$, precisely up to the large time $\frac{1}{E_1}$. This result is useful in the base situation, where it is expected $E_1 \ll 1$.

Remark 1.2

1. By taking a small c in the previous theorem, we have

$$\frac{t_n \rho_1 E_1}{K_1} \lessapprox \gamma_n \lessapprox t_n \rho_1 E_1.$$

This holds for times $t_n \rho_1 \leq x$, where $x > 0$ is such that $x E_1 \ll 1$. If Λ_1 is constituted by a real eigenvalue or by a complex conjugate pair of eigenvalues (the generic situation for the matrix A), we have $K_1 = 1$ and then

$$\gamma_n^{\text{long}} \approx t_n \rho_1 E_1. \tag{1.19}$$

2. Similarly to the point 1 of Remark 1.1, we can explain because $\max_{\lambda_i \in \Lambda_1} |\sigma_i|$ and E_1 can be considered as local relative error and global relative error, respectively, in the numerical integration of the long-time solution.
3. Since $\max_{\lambda_i \in \Lambda} |\sigma_i|$ and $\max_{\lambda_i \in \Lambda_1} |\sigma_i|$ can be considered as local relative errors in the numerical integration of the solution and the long-time solution, respectively, we can say that in the non-stiff situation the local relative error of the solution is small, whereas in the stiff situation the local relative error of the solution is not small, but the local relative error of the long-time solution is small. This agrees with the traditional concepts of non-stiff and stiff.
4. The theorem assumes $0 \notin \Lambda_1$. If $\Lambda_1 = \{0\}$, we have $\gamma_n^{\text{long}} = 0$ for any n. For the case $0 \in \Lambda_1$ and $\Lambda_1 \neq \{0\}$, see point 5 of Remark 5.2 in [9].

1.2.3 Long-time behavior of γ_n

We want to study the long-time behavior of the error γ_n. This is done by comparing it to the error γ_n^{long}.

Since in the long-time the solution y becomes the solution y^{long} whose error γ_n is just γ_n^{long}, it is quite reasonable to have $\gamma_n \approx \gamma_n^{\text{long}}$ in the long-time.

Indeed, at point 4 of Remark 5.3 in [9], it is stated the following result.

Theorem 1.3 *Assume $q > 1$ and $0 \notin \Lambda_1$. Fix $c > 0$ such that $g(c) < 1$, i.e. $c < 1.2564$. For any $\varepsilon > 0$, there exist $H_0 > 0$ (independent of ε) and $s \geq 0$ (dependent on ε) such that, for $h\rho \leq H_0$ and $s \leq t_n \rho \leq \frac{c}{E}$, we have $\gamma_n \approx \gamma_n^{\text{long}}$ with degree ε.*

Remark 1.3 The theorem assumes $q > 1$. If $q = 1$, then $\gamma_n = \gamma_n^{\text{long}}$ for any n. In addition, it assumes $0 \notin \Lambda_1$. If $q > 1$ and $\Lambda_1 = \{0\}$, then $\gamma_n^{\text{long}} = 0$ for any n and it does not make sense look at $\gamma_n \approx \gamma_n^{\text{long}}$, since this implies $\gamma_n = 0$. For the case $q > 1$, $0 \in \Lambda_1$ and $\Lambda_1 \neq \{0\}$, see point 6 of Remark 5.3 in [9].

The previous theorem is of interest in the non-stiff situation, where the condition $h\rho \leq H_0$ is not restrictive. In fact, in the non-stiff situation it is expected $h\rho$ non-large.

On the other hand, the result is not useful in the stiff situation, since the condition $h\rho \leq H_0$ is restrictive. In fact, in the stiff situation it is expected $h\rho$ non-small.

1.3 The contents of this paper

The present paper wants to study the long-time behavior of the relative error γ_n in the stiff situation. As above, this is done by comparing it to γ_n^{long}.

In the stiff situation, it is important to have $\gamma_n \approx \gamma_n^{\text{long}}$ in the long-time. In fact, if this happens, since γ_n^{long} is small up to large times $t_n \rho_1$, we have the very surprising fact that the error γ_n is small in the long-time, although the stepsize h is

tuned only for having a small local relative error of the long-time solution and, because of this, the local relative error of the solution is not small.

In other words, when we are interested in the numerical integration of the solution in the long-time, we can start from the beginning with a stepsize suitable for integrating with a small local relative error the long-time solution, larger than the stepsize suitable for integrating with a small local relative error the solution, and in the long-time we will have a small error γ_n.

As in [9], we confine our attention to normal matrices. This should be not considered as a limitation, since the class of the normal matrices is sufficiently large to include important types of matrices and, moreover, the test problem (1.1) with A normal shows unexplored and interesting situations in numerical ODEs.

The plan of the paper is as follows.

- Section 2 shows two examples of stiff situation where we can fail to get $\gamma_n \approx \gamma_n^{\text{long}}$ in the long-time with γ_n non-small and growing unboundedly.
- Section 3 introduces the definition of "$\gamma_n \approx \gamma_n^{\text{long}}$ in the long-time" of our interest.
- Section 4 gives the condition for having, in the stiff situation, $\gamma_n \approx \gamma_n^{\text{long}}$ in the long-time.
- Section 5 show that when this condition does not hold, we have, in the stiff situation, $\frac{\gamma_n}{\gamma_n^{\text{long}}} \gg 1$ for all time.
- Section 6 revises the examples of Sect. 2 in the light of the results of Sects. 4 and 5.
- Section 7 studies when the condition for having $\gamma_n \approx \gamma_n^{\text{long}}$ holds independently of the specific non-rightmost spectrum.
- Conclusions are draft in Sect. 8.

1.4 Replies to general questions or criticisms

This final subsection includes replies to general questions or criticisms which could be issued about the contents of this paper.

- **Question**. *What is the motivation of this paper?*

  Reply. This paper studies the relative error of numerical approximations of ODEs, although confined to linear systems with normal matrix. Of course, the absolute error and the relative error of the numerical approximations have the same order of convergence with respect to the stepsize h, but they have a different time behavior in the numerical integration of a solution spanning over various orders of magnitude.

  The motivation for studying the relative error time behavior in numerical ODEs, as the present paper is doing, comes from the following two facts:

  - as it is widely recognized, the relative error is an important measure for the quality of an approximation, often better than the absolute error;

– there has been no attention in the numerical ODEs community about the relative error time behavior of numerical approximations.

Anyway, the fact that in the numerical ODEs field the relative error is considered important is attested by the numerical solvers, which accept as an input argument a tolerance on the componentwise relative error. Thus, this paper (similarly to [9] and [10]) try to fill this gap between theory, where there are not studies on the relative error, and practice, where the relative error is used.

– **Question**. *What is the relevance of the results achieved?*

Reply. For the numerical ODEs community, it should be of interest to know the relative error time behavior of numerical approximations of the ODE (1.1) with A normal. The results achieved describes this time behavior and their relevance is that they give a new prospective on the numerical integration errors. We can summarize this new perspective in the following points.

– In the non-stiff situation, the relative error is small and it grows linearly in time. Moreover, this linear growth is determined in the long-time only by the rightmost eigenvalues.
– In the stiff situation, the relative error is not small at the beginning of the numerical integration and it is not guaranteed that in the long-time it will become small, with a linear growth determined only by the rightmost eigenvalues. This happens if and only if a certain condition is satisfied and this condition is a novelty in the numerical ODE theory.
– Gauss RK methods, despite they are considered stable in the classic numerical linear stability theory (they are A-stable methods), are not suitable to have the above condition satisfied. On the other hand, Radau and Lobatto IIIC RK methods are suitable to have this condition satisfied.

– **Criticism**. *Componentwise relative errors*

$$\frac{|y_{n,i} - y_i(t_n)|}{|y_i(t_n)|}, \ i = 1, \dots, d,$$

where $y_{n,i}$ *and* $y_i(t_n)$, $i = 1, \dots, d$, *are the components of* y_n *and* $y(t_n)$, *should be considered (as in the numerical ODE solvers), not the normwise relative error* (1.3).

Reply. In literature both normwise relative errors and componentwise relative errors are considered as quality measures of vector approximations (see [2]). The componentwise approach has the advantage that it gives information on the precision of the components, but it has the drawback that the components must be nonzero (when some component becomes zero, we need to switch to the absolute error). On the other hand, the normwise approach can give anyway information about the componentwise relative errors (for example, a large normwise relative error implies that some component has a large relative error) and it works also when some component becomes zero.

– **Criticism**. *Relative errors should be not considered in situations where the exact solution approaches zero, as those studied in this paper. A rule of thumb in numerical analysis says that one should switch to the absolute error in this situation*.

Reply. In mathematical modeling and numerical analysis there is a threshold in the order of magnitude of quantities (scalars or vectors) under which they are considered zero. Under the threshold, it is important to use the absolute error for approximations, since they are considered approximations of zero. But, in case of a solution of (1.1) which is going to zero, and so it is spanning over several orders of magnitude, it could be of interest to compute with a good precision this solution for the orders of magnitude larger than the threshold. In this situation, the relative error is important.

Of course, the numerical analyst's point of view is that the threshold is the order of magnitude of the machine epsilon, but in applications this threshold can be larger.

As an example, we can consider the radioactivity decay of radionuclides, where the activity $a(t)$ (measured in becquerel (Bq) by a Geiger counter) of a given amount of radionuclide satisfies $a'(t) = -\lambda a(t)$ with $\lambda > 0$. For a decay chain, we have $a'(t) = Aa(t)$, where A is a lower bi-diagonal matrix, the so-called Bateman equation. The threshold could be the order of magnitude 10^2Bq/kg of the background radiation. Of course, this threshold becomes a much smaller ten power by using an unit larger than the becquerel, e.g. the curie. It could be interesting to numerically compute with a good precision a solution $a(t)$ whose initial value has order of magnitude 10^6Bq/kg (like in a nuclear plant accident). Since the solution becomes small compared to the initial value, using the relative error for the approximations of the solution is better than using the absolute error when the solution is not yet considered as zero.

Another example could be a space discretization of the heat equation, with homogeneous Dirichlet boundary condition, by the method of lines. In this case, the space discrete temperature approaches zero (the border temperature) and under a given threshold in the order of magnitude, say 10^{-2} °C, it can be considered zero. But, over this order of magnitude, the temperature is not zero and it becomes important to use the relative error for time-space approximations, especially when the solution spans over several orders of magnitude due to an initial value with order of magnitude larger than the threshold, for example 10^2 °C.

We remark that the analysis in this paper also consider the situation where the solution, instead of approaching to zero, grows up to large values with respect to the initial value. Also in this situation the relative error is important.

– **Criticism**. *The paper considers ODEs* (1.1) *with matrix A normal. Such problems can be diagonalized with a unitary transformation and then one can assume without loss of generality that A is diagonal*.

Reply. In the paper, we do not assume from the beginning that A is diagonal because this does not simplify the exposition. In fact, the analysis presented starts from the fundamental relation (4.6) given below for the relative error,

which maintains the same form when A is diagonal. We have such a net expression for the relative error precisely for the possibility to reduce to the diagonal case by a unitary transformation. Hence, the assumption that A is diagonal is already implicitly done when one decides to deal with a normal matrix.

– **Criticism**. *Since it is possible to reduce to the diagonal case, it would be sufficient to study the behavior of the numerical scheme at a scalar problem, which is really trivial.*

  Reply. Although we can reduce to a linear systems of uncoupled scalar differential equations, this does not mean that they are fully uncoupled in the numerical scheme, since we are using the same stepsize h in all scalar equations. This reflects the fact that the numerical scheme is applied to an ODE (1.1) with a matrix A in general non-diagonal, without thinking to diagonalize it in advance. Moreover, the analysis of the present paper requires to have rightmost and non-rightmost eigenvalues. In other words, we need eigenvalues with different real parts, i.e. an ODE (1.1) with different time scales. The case of a sole scalar equation is not considered. Anyway, we can observe that in the base situation for a scalar equation, the relative error $\gamma_n = \gamma_n^{\text{long}}$ is expected to be small and linearly growing in time up to large times.

2 Examples

In this section, we give two examples of stiff situations where the error γ_n is not small from the beginning of the numerical integration and it grows without to approach in the long-time to the small error γ_n^{long}.

We remind that the *stability region* of the approximant R (see [5]) is the set

$$\mathscr{R} := \{z \in \mathscr{D} : |R(z)| \leq 1\}$$

and the *order star* of R (see [5–7, 13]) is the set

$$\mathscr{S} := \{z \in \mathscr{D} : |S(z)| > 1\},$$

where S is the relative approximant given in (1.8). The complementary set of $\mathscr{S}$ is

$$\mathscr{S}^c = \mathscr{D} \backslash \mathscr{S} = \{z \in \mathscr{D} : |S(z)| \leq 1\}.$$

2.1 Same approximant with different ODEs

As first example, we consider the ODE (1.1) with the symmetric matrix

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

whose eigenvalues are a and b with relevant eigenvectors $(1, 1)$ and $(1, -1)$, respectively. We consider $a = -1$ and the following three possibilities for b:

(P1) $b = -11$;
(P2) $b = -13.5$;
(P3) $b = -16$.

The initial value is $y_0 = (2, -1)$, for which we have

$$\|P_1\widehat{y}_0\|_2 = \frac{1}{\sqrt{10}} \quad \text{and} \quad \|P_2\widehat{y}_0\|_2 = \frac{3}{\sqrt{10}}.$$

The solution y quickly approaches to the long-time solution y^{long}: we have $y(t) \approx y^{\text{long}}(t)$ if

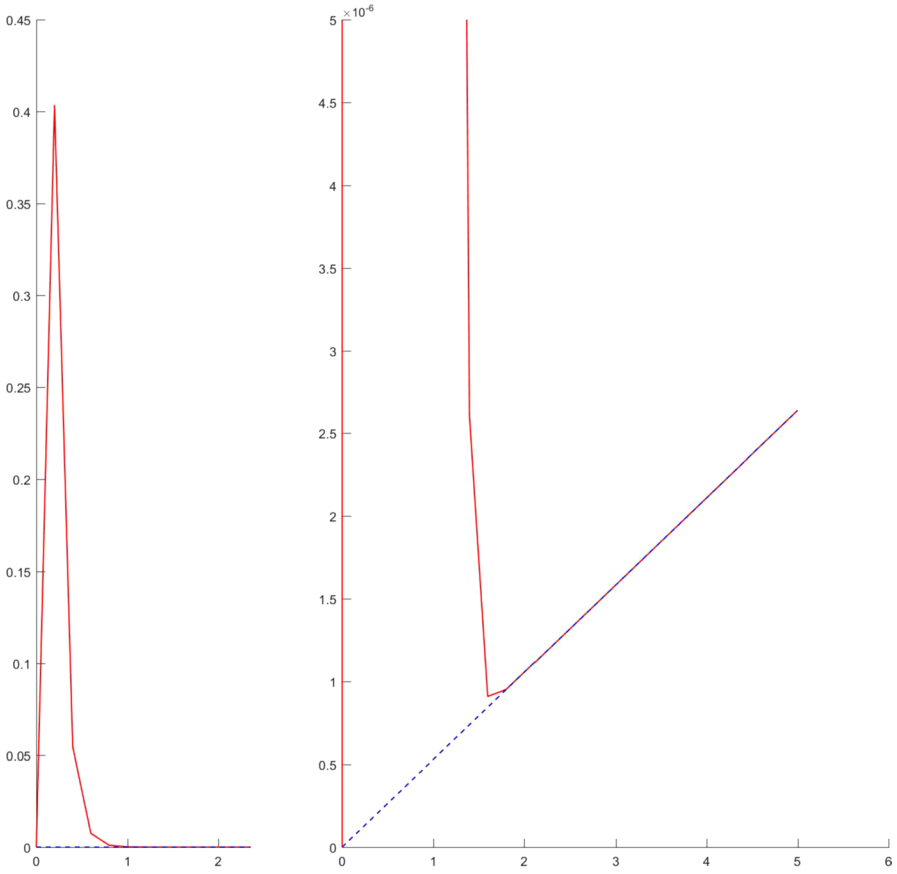

Fig. 2 Possibility P1) with initial value $y_0 = (2, -1)$. Errors γ_n (solid red line) and γ_n^{long} (dash blue line). The abscissas are the times $t_n = nh$, $n = 0, 1, 2, \ldots, N$

$$e^{(b-a)t}\frac{\|P_2\widehat{y}_0\|_2}{\|P_1\widehat{y}_0\|_2} = 3e^{(b+1)t} \ll 1$$

(look at (1.17)).

For the numerical integration, we use the Taylor approximant of order five

$$R(z) = 1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \frac{z^4}{24} + \frac{z^5}{120},\ z \in \mathbb{C},$$

with stepsize $h = \frac{1}{5}$ over $N = 25$ steps up to $t_N = Nh = 5$.

We have:

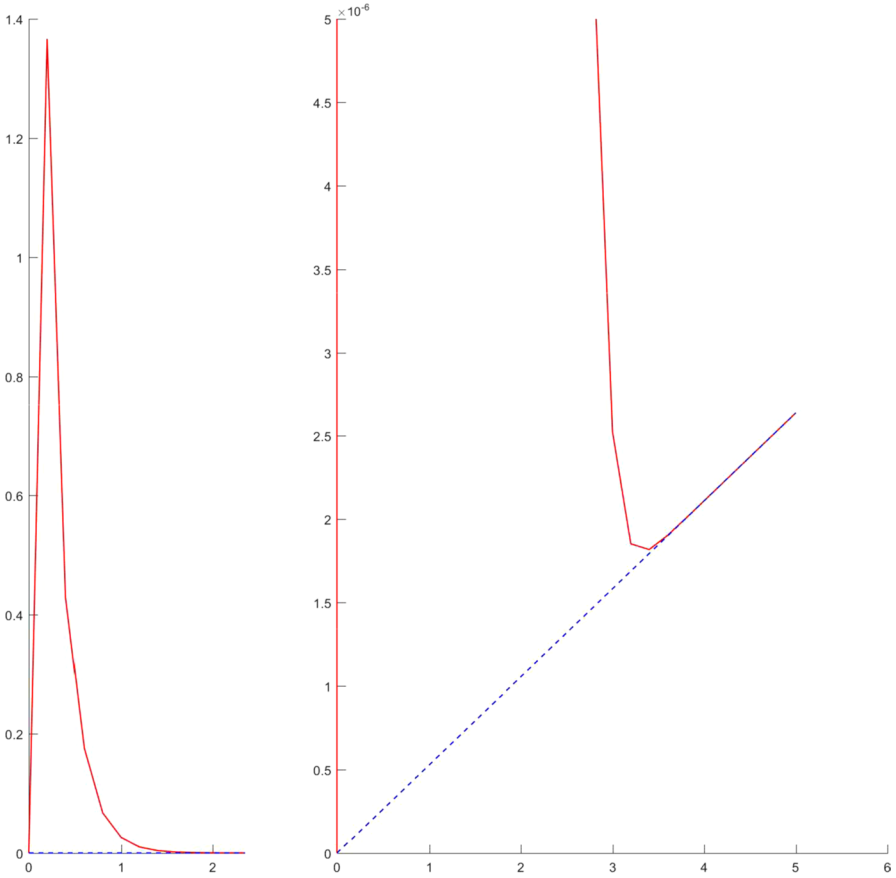

Fig. 3 Possibility P2) with initial value $y_0 = (2, -1)$. Errors γ_n (solid red line) and γ_n^{long} (dash blue line). The abscissas are the times $t_n = nh$, $n = 0, 1, 2, \ldots, N$

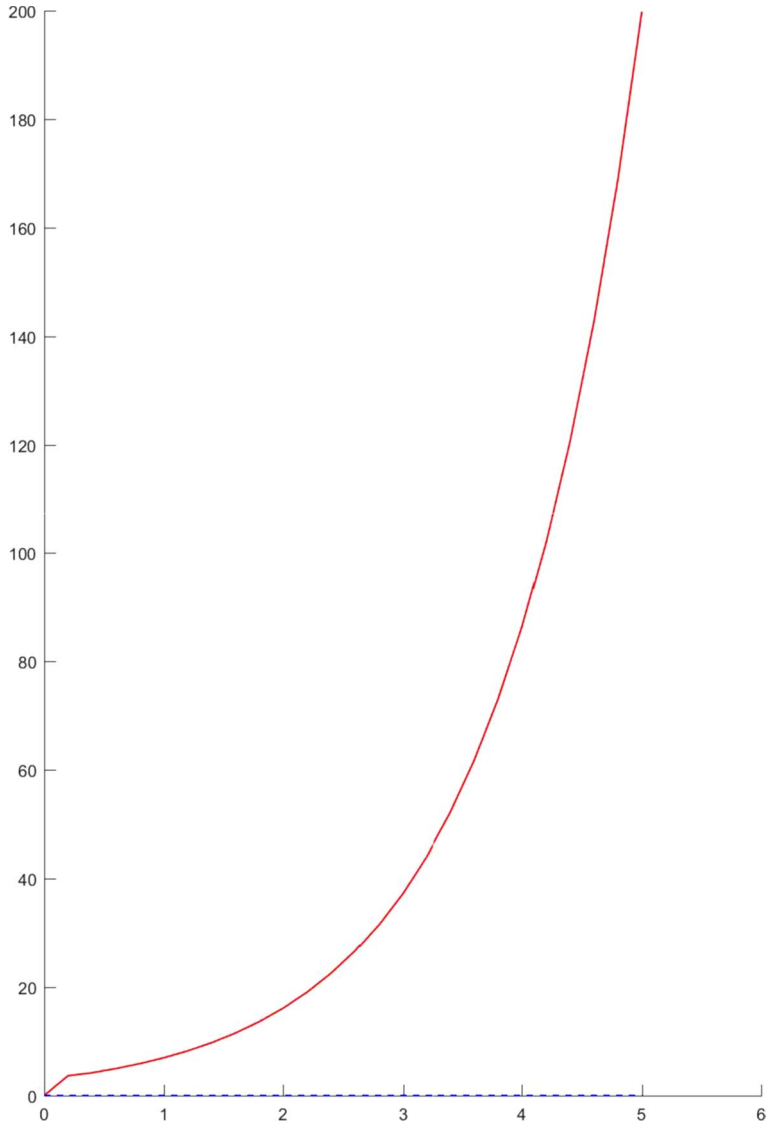

Fig. 4 Possibility P3) with initial value $y_0 = (2, -1)$. Errors γ_n (solid red line) and γ_n^{long} (dash blue line). The abscissas are the times $t_n = nh$, $n = 0, 1, 2, \dots, N$

$$h\rho_1 = 0.2,\ |\sigma_1| = 1.06 \cdot 10^{-7} \text{ and } E_1 = \frac{|\sigma_1|}{h\rho_1} = 5.28 \cdot 10^{-7}.$$

We are in the stiff situation: in the three possibilities for b, we have

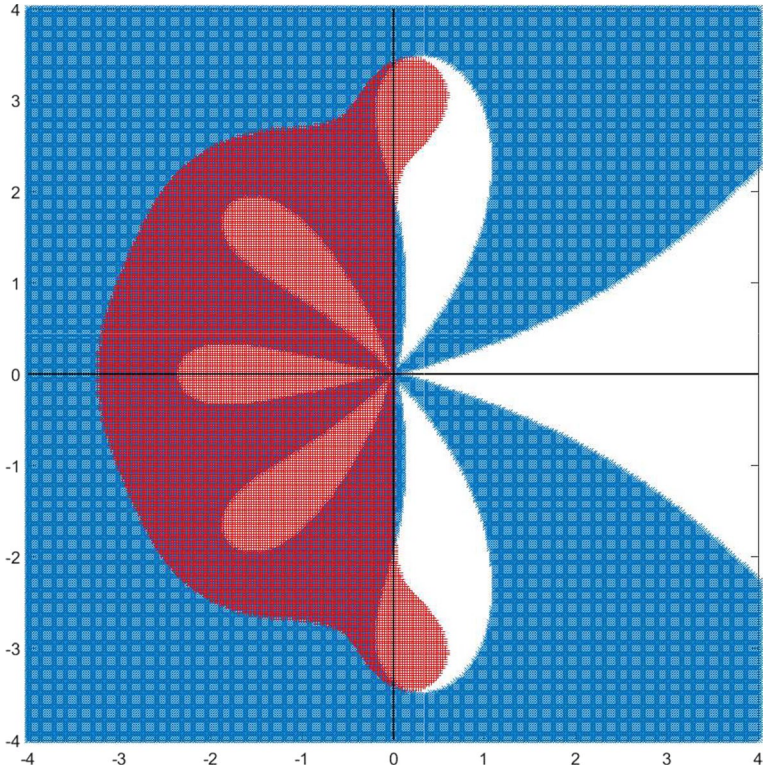

Fig. 5 Order star (in blue) and complementary set (in white) of the five order Taylor approximant with the stability region (in red) overlapped

	$\lvert\sigma_2\rvert$
P1)	4.09
P2)	3.50
P3)	4.46

Since

$$t_N \rho_1 E_1 = 2.64 \cdot 10^{-6} \ll 1,$$

by (1.19) we obtain

$$\gamma_n^{\text{long}} \approx t_n \rho_1 E_1 = t_n \cdot 5.28 \cdot 10^{-7}$$

for $t_n \leq t_N$.

For the possibility P1), we see in Fig. 2, for $n = 0, 1, 2, \ldots, N$, the relative errors γ_n (solid red line) and γ_n^{long} (dash blue line).

Starting from a non-small γ_1 (remind that $\gamma_0 = 0$), the error γ_n goes down to the small error γ_n^{long}. In the long-time, we have small errors γ_n although the stepsize is tuned only for having a small σ_1, without any concern about σ_2.

For the possibility P2), we see in Fig. 3 the same as in Fig. 2. As in P1), starting from a non-small γ_1, the error γ_n goes down to γ_n^{long}, although γ_n^{long} is reached at a larger time with respect to P1).

Finally, for the possibility P3), we see in Fig. 4 the same as in Figs. 2 and 3. Unlike P1) and P2), the error γ_n does not go down to γ_n^{long}, but it continues to grow.

2.1.1 Order star and stability region

Fixed $a = -1$, we are interested in understanding for which b, with $b < a$, we have $\gamma_n \approx \gamma_n^{\text{long}}$ in the long-time. This happens in P1) and P2), but not in P3).

Order star and stability region for the Taylor approximant of order five are depicted in Fig. 5.

The values of $|S(hb)|$ and $|R(hb)|$ are:

	$\lvert S(hb)\rvert$	$\lvert R(hb)\rvert$
P1)	0.0728	0.00807
P2)	4.72	0.317
P3)	23.8	0.968

Observe that $hb \in \mathscr{R}$ for all three possibilities and $hb \in \mathscr{S}^{\text{e}}$ only in P1). In other words, by looking at the negative real axis of Fig 5, hb lies in the red region for all three possibilities and hb lies in the white finger only in P1).

In the next Sect. 4, we will see a condition on hb for having $\gamma_n \approx \gamma_n^{\text{long}}$ in the long-time. When it does not hold, we have $\frac{\gamma_n}{\gamma_n^{\text{long}}} \gg 1$ for all time. The condition is something between $hb \in \mathscr{S}^{\text{e}}$ (i.e. to stay in the white finger) and $hb \in \mathscr{R}$ (i.e. to stay in the red region). Indeed, to have $hb \in \mathscr{S}^{\text{e}}$ is sufficient, but not necessary, for this condition on hb and to have $hb \in \mathscr{R}$ is necessary, but not sufficient.

2.2 Same ODE with different approximants

As second example, we consider the ODE (1.1) with the normal matrix

$$A = QDQ^H$$

with

$$D = \operatorname{diag}\left(\lambda_1, \overline{\lambda_1}, \lambda_3, \overline{\lambda_3}\right) = \operatorname{diag}(-1+i, -1-i, -3+1000i, -3-1000i)$$

and Q with orthonormal columns $u_1, \overline{u_1}, u_3, \overline{u_3}$, where

$$
\begin{aligned}
u_1 =& v_1 + iv_2,\ u_3 = v_3 + iv_4,\\
v_1 =& \frac{1}{2\sqrt{2}}(1,1,1,1), \quad v_2 = \frac{1}{2\sqrt{2}}(1,1,-1,-1),\\
v_3 =& \frac{1}{2\sqrt{2}}(1,-1,1,-1), \quad v_4 = \frac{1}{2\sqrt{2}}(-1,1,1,-1).
\end{aligned}
$$

Consider the initial value $y_0 = (3,3,3,-2)$ for which we have

$$\|P_1\widehat{y}_0\|_2 = \|P_2\widehat{y}_0\|_2 = 0.5462 \text{ and } \|P_3\widehat{y}_0\|_2 = \|P_4\widehat{y}_0\|_2 = 0.4490.$$

The solution y consists of two decaying oscillations: the fast oscillation $y - y^{\text{long}}$ decays faster than the slow oscillation y^{long} and, in the long-time, only the slow oscillation is present. We have $y(t) \approx y^{\text{long}}(t)$ if

$$\mathrm{e}^{(r_2-r_1)t}\frac{\|Q_2\widehat{y}_0\|_2}{\|Q_1\widehat{y}_0\|_2} = 0.82199 \cdot \mathrm{e}^{-2t} \ll 1$$

(look at (1.17)).

Assume that the numerical integration of the ODE is accomplished by the fourth order two-stage Gauss RK method, corresponding to the $(2,2)$–Padé approximant

$$R(z) = \frac{1 + \frac{z}{2} + \frac{z^2}{12}}{1 - \frac{z}{2} + \frac{z^2}{12}},\ z \in \mathbb{C}\backslash\left\{3 \pm i\sqrt{3}\right\},$$

and by the third order two-stage Radau RK method, corresponding to the $(1,2)$–Padé approximant

$$R(z) = \frac{1 + \frac{z}{3}}{1 - \frac{2z}{3} + \frac{z^2}{6}},\ z \in \mathbb{C}\backslash\left\{2 \pm i\sqrt{2}\right\}.$$

Both methods are applied with stepsize $h = \frac{1}{10}$ over $N = 100$ steps up to $t_N = Nh = 10$. Observe that such a stepsize is not suitable for approximating the fast oscillation.

We are in the stiff situation:

	$h\rho_1$	$\lvert\sigma_1\rvert = \lvert\sigma_2\rvert$	E_1	$\lvert\sigma_3\rvert = \lvert\sigma_4\rvert$
Gauss RK method	0.141	$7.86 \cdot 10^{-8}$	$5.56 \cdot 10^{-7}$	0.51
Radau RK method	0.141	$5.41 \cdot 10^{-6}$	$3.82 \cdot 10^{-5}$	3.78

Since

$$t_N\rho_1 E_1 = \begin{cases} 7.86 \cdot 10^{-6} & \text{for the Gauss RK method} \\ 5.41 \cdot 10^{-4} & \text{for the Radau RK method} \end{cases} \ll 1,$$

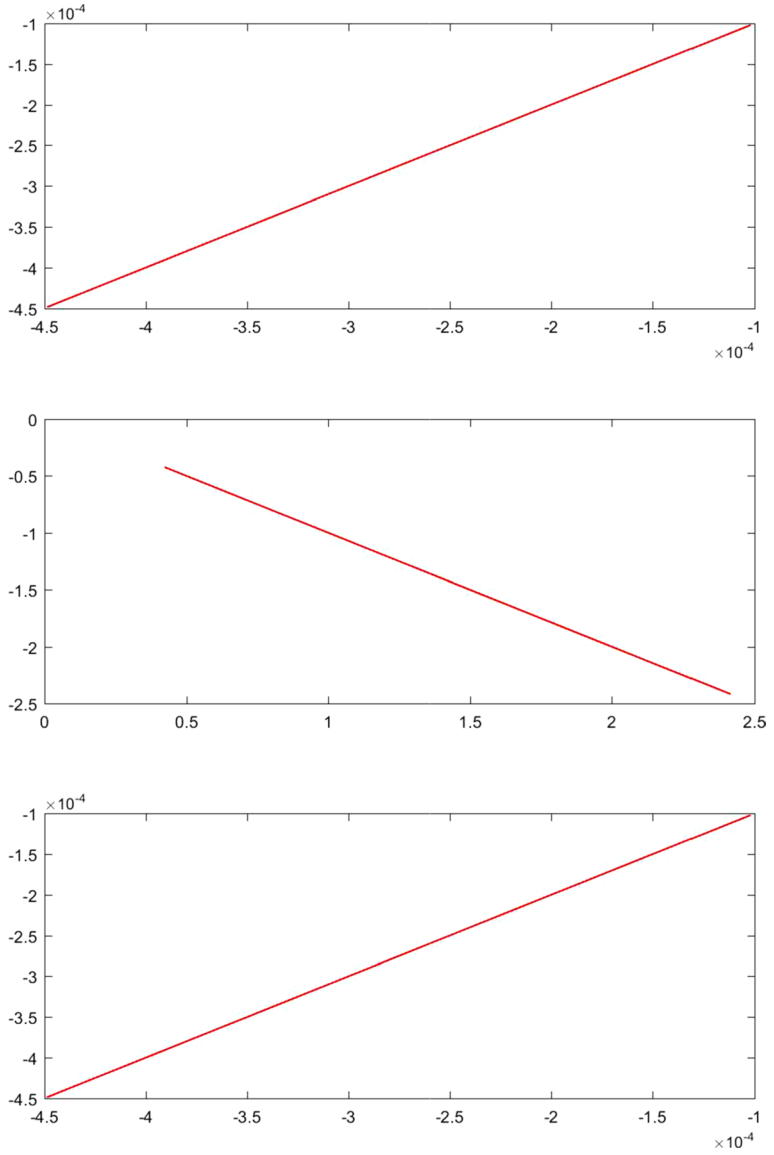

Fig. 6 Upper part: trajectory $\big(y_1(t_n), y_2(t_n)\big)$, $t_n \in [8, 10]$, for the exact solution. Middle part: trajectory $\big(y_{n,1}, y_{n,2}\big)$, $t_n \in [8, 10]$, for the Gauss RK method solution. Lower part: trajectory $\big(y_{n,1}, y_{n,2}\big)$, $t_n \in [8, 10]$, for the Radau RK method solution

by (1.19) we have

$$\gamma_n^{\text{long}} \approx t_n \rho_1 E_1 = \begin{cases} t_n \cdot 7.86 \cdot 10^{-7} & \text{for the Gauss RK method} \\ t_n \cdot 5.41 \cdot 10^{-5} & \text{for the Radau RK method} \end{cases}$$

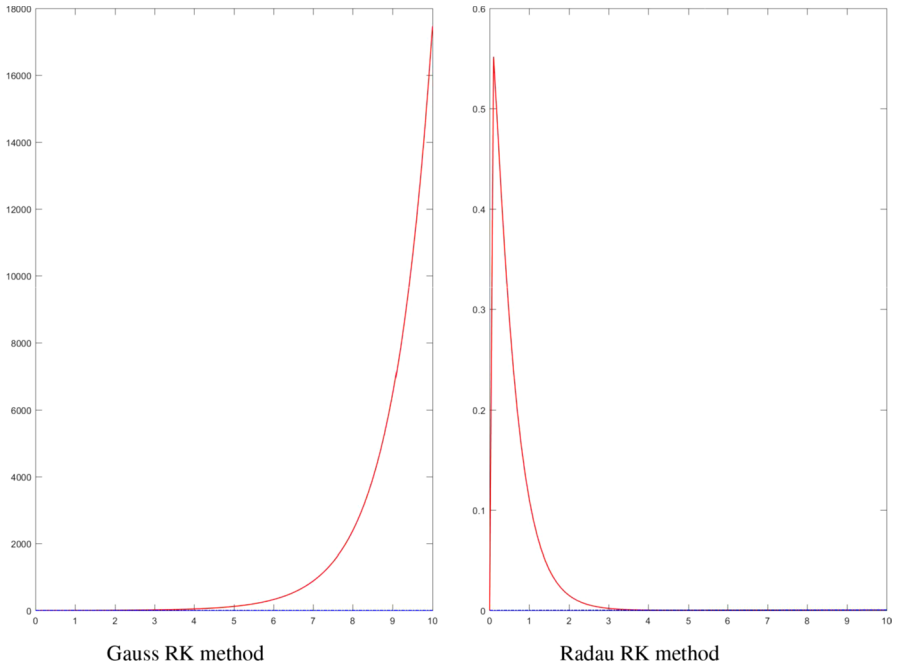

Fig. 7 Errors γ_n (solid red line) and γ_n^{long} (dash blue line). The abscissas are the times t_n, $n = 0, 1, \ldots, N$

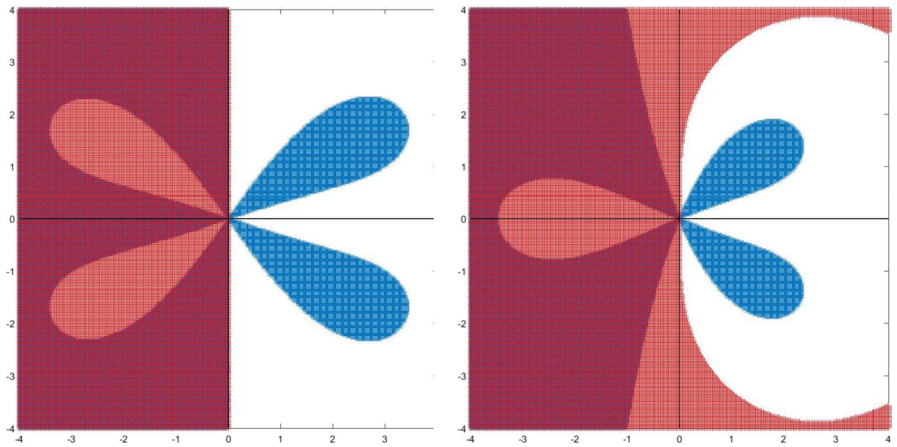

Fig. 8 Order star (in blue) and complementary set (in white) with the stability region (in red) overlapped for the Gauss RK method (left) and the Radau RK method (right)

for $t_n \leq t_N$.

In the upper part of Fig. 6, we see the trajectory $t_n \mapsto \big(y_1(t_n), y_2(t_n)\big)$ in the plane $\mathbb{R}^2$ for the first two components of the exact solution $y(t_n)$, when $t_n \in [8, 10]$. In the middle and lower parts, we see the trajectory $t_n \mapsto \big(y_{n,1}, y_{n,2}\big)$ for the first two components of

the numerical solution y_n, when $t_n \in [8, 10]$. Middle part for the Gauss RK method and lower part for the Radau RK method.

For the long-time $t_n \in [8, 10]$, where only the slow oscillation y^{long} is present, the exact components $y_1(t_n)$ and $y_2(t_n)$ are equal and have order of magnitude 10^{-4}. The Gauss RK method exhibits numerical components $y_{n,1}$ and $y_{n,2}$ of order of magnitude 10^0. On the other hand, the Radau RK method exhibits accurate numerical components $y_{n,1}$ and $y_{n,2}$, although the stepsize is not suitable for approximating the fast oscillation.

In Fig. 7, we see the error γ_n, for $n = 0, 1, \dots, N$, for both approximants: for the Gauss RK method the error continues to grow and for the Radau RK method it goes down to γ_n^{long}.

2.2.1 Order star and stability region

Fixed $\lambda_1 = -1 + i$ and $\lambda_3 = -3 + 1000i$, we are interested in understanding for which approximants we have $\gamma_n \approx \gamma_n^{\text{long}}$ in the long-time. In our situation, this happens for the Radau RK method, but not for the Gauss RK method.

Order star and stability region for such approximants are shown in Fig. 8.

We have $h\lambda_3 \in \mathscr{R}$ for both methods, since they are A-stable. On the other hand, we have $h\lambda_3 \in \mathscr{S}$ only for the Radau RK method:

$$\left|S\left(h\lambda_3\right)\right| = \begin{cases} 1.3494 & \text{for the Gauss RK method} \\ 0.0270 & \text{for the Radau RK method.} \end{cases}$$

In Sect. 4, we will see a condition on the approximant for having $\gamma_n \approx \gamma_n^{\text{long}}$ in the long-time. When the condition does not hold, we have $\frac{\gamma_n}{\gamma_n^{\text{long}}} \gg 1$ for all time. To have $h\lambda_3 \in \mathscr{S}$ (i.e., with reference Fig. 8, the white region of the approximant contains $h\lambda_3$) is sufficient, but not necessary, for this condition on the approximant and to have $h\lambda_3 \in \mathscr{R}$ (i.e. the red region of the approximant contains $h\lambda_3$) is necessary, but not sufficient.

3 The appropriate definition of $\gamma_n \approx \gamma_n^{\text{long}}$ in the long-time

In the following, we assume to be in the base situation. Then, it is expected E_1 small and $h\rho_1$ non-large. To make easier the exposition, we assume E_1 small and $h\rho_1$ non-large.

Since E_1 is small, the error γ_n^{long} grows linearly in time and it is small up to large times $t_n\rho_1$.

We also fix a number $c > 0$ and let

$$\tau := \frac{c}{E_1}.$$

(The number c plays a role similar the number c appearing in Theorems 1.1, 1.2 and 1.3). As a reference value for c, one can take $c = 1$. As a matter of generality, we do not confine c only to this value. In all theorems below, it is stated for which $c > 0$

they are valid. However, when the theorems are applied, c is considered non-small, so we have

$$\tau \gg 1,$$

and such that $g(c) < 1$, i.e. $c < 1.2564$, with $1 - g(c)$ non-small.

In order to describe the long-time behavior of the error γ_n, we compare it to γ_n^{long} and we are interested in whether or not $\gamma_n \approx \gamma_n^{\text{long}}$ in the long-time.

Here, "in the long-time" does not mean $t_n\rho_1 \to +\infty$. In fact, it is not of great interest to consider what happens for $t_n\rho_1 \to +\infty$, since γ_n^{long} becomes non-small for a sufficiently large $t_n\rho_1$. It is of interest to have $\gamma_n \approx \gamma_n^{\text{long}}$ starting from times $t_n\rho_1$ such that γ_n^{long} is still small.

So, we introduce the following definition.

Definition 3.1 We say that $\gamma_n \approx \gamma_n^{\text{long}}$ *in the long-time* if, for some $s \in [0, \tau)$, $\gamma_n \approx \gamma_n^{\text{long}}$ for $t_n\rho_1$ in the interval $[s, \tau]$ and $\gamma_n^{\text{long}} \ll 1$ for $t_n\rho_1$ up to the beginning of this interval, i.e. for $t_n\rho_1 \in [0, \kappa s]$ and $\kappa \geq 1$ non-large.

In the definition, we consider times $t_n\rho_1$ up to τ. Observe that if K_1 is not large (remind (1.12) and remind that $K_1 = 1$ is the generic situation for the matrix A), then the error γ_n^{long} is not small for $t_n\rho_1$ at the end of the interval $[0, \tau]$.

In fact, for $t_n\rho_1 \in [\kappa\tau, \tau]$, where $\kappa \in (0, 1]$ is not small, by Theorem 1.2 we have

$$\gamma_n^{\text{long}} \geq t_n\rho_1 \frac{E_1}{K_1}(1 - g(c)) \geq \frac{\kappa c(1 - g(c))}{K_1}$$

and the right-hand side in this inequality is not small.

3.1 The definition of $\gamma_n \approx \gamma_n^{\text{long}}$ in the long-time with monitor function

We can make the previous definition more precise by a *monitor function*.

Definition 3.2 Let $s : (0, +\infty) \times (0, +\infty) \to [0, +\infty)$ be a function such that

$$\lim_{x \to +\infty} \frac{s(\varepsilon, x)}{x} = 0 \quad \text{for any } \varepsilon > 0. \tag{3.1}$$

We say that $\gamma_n \approx \gamma_n^{\text{long}}$ in the long-time with monitor function s if, for any $\varepsilon > 0$, we have

$$|e_n| \leq \varepsilon \text{ for } t_n\rho_1 \in [s(\varepsilon, \tau), \tau], \tag{3.2}$$

where e_n is such that $\gamma_n = \gamma_n^{\text{long}}(1 + e_n)$.

Remark 3.1 In the previous definition, we also allow monitor functions $s : (0, a] \times [b, +\infty) \to [0, +\infty)$, where $0 < a, b < +\infty$. In this case, we have to specify that (3.1) holds for $\varepsilon \in (0, a]$ and (3.2) holds for $\varepsilon \in (0, a]$ and $\tau \geq b$.

3.2 What does the definition with monitor function mean?

Suppose $\gamma_n \approx \gamma_n^{\text{long}}$ in the long-time with monitor function s.

Let $\varepsilon > 0$. By (3.2), we see that $\gamma_n \approx \gamma_n^{\text{long}}$ with degree ε for $t_n\rho_1 \in [s(\varepsilon,\tau),\tau]$. Moreover, by Theorem 1.2, we see that if $\frac{s(\varepsilon,\tau)}{\tau} \ll 1$, then

$$\gamma_n^{\text{long}} \leq t_n\rho_1 E_1(1+g(c)) = \kappa\frac{s(\varepsilon,\tau)}{\tau}c(1+g(c)) \ll 1 \tag{3.3}$$

for $t_n\rho_1 \in [0,\kappa s(\varepsilon,\tau)]$, where $\kappa \geq 1$ is not large, i.e. $\gamma_n^{\text{long}} \ll 1$ for $t_n\rho_1$ up to the beginning of the interval $[s(\varepsilon,\tau),\tau]$.

In summary:

> If $\gamma_n \approx \gamma_n^{\text{long}}$ in the long-time with monitor function s and $\dfrac{s(\varepsilon,\tau)}{\tau} \ll 1$ for some small $\varepsilon > 0$, then $\gamma_n \approx \gamma_n^{\text{long}}$ in the long-time. In particular, we have $\gamma_n \approx \gamma_n^{\text{long}}$ with degree ε for $t_n\rho_1 \in [s(\varepsilon,\tau),\tau]$.

Regarding the satisfiability of $\frac{s(\varepsilon,\tau)}{\tau} \ll 1$, observe that s satisfies (3.1) and we have $\tau \gg 1$.

4 Analysis of the long-time behavior of γ_n

In the paper [9], it was presented an analysis of the long-time behavior of the error γ_n important for the non-stiff situation. In the present paper, it is developed another type of analysis important for the stiff situation. In this new analysis, the complex numbers w_i and α_i introduced below are important.

4.1 The numbers w_i

For any $\lambda_i \in \Lambda^-$, i.e. for any non-rightmost eigenvalue, we introduce the complex number

$$w_i := \mathrm{e}^{h(r_j-r_1)}S(h\lambda_i),$$

where $j = 2,\dots,q$ is such that $\lambda_i \in \Lambda_j$. We set

$$W := \max_{\lambda_i\in\Lambda^-} |w_i|. \tag{4.1}$$

4.2 The numbers α_i

For any $\lambda_i \in \Lambda^-$, we introduce the complex number

$$\alpha_i := \frac{\log w_i}{h\rho_1}.$$

We set

$$\alpha := \max_{\lambda_i \in \Lambda^-} \mathrm{Re}(\alpha_i) = \frac{\log W}{h\rho_1}. \tag{4.2}$$

Since

$$\sigma_i = \log S(h\lambda_i),$$

we have

$$\alpha_i = \beta_j + \frac{\sigma_i}{h\rho_1}, \tag{4.3}$$

where $j = 2, \dots, q$ is such that $\lambda_i \in \Lambda_j$.

As a consequence we obtain

$$|\alpha - \beta_2| \le \frac{\max_{\lambda_i \in \Lambda^-} |\sigma_i|}{h\rho_1}. \tag{4.4}$$

Here are some observations about α.

- It is expected $|\alpha|$ non-small. In fact, let $\lambda_i \in \Lambda_j$, with $j = 2, \dots, q$, be a non-right-most eigenvalue such that

$$\alpha = \mathrm{Re}(\alpha_i) = \beta_j + \frac{\mathrm{Re}(\sigma_i)}{h\rho_1}.$$

  The case where $|\alpha|$ is small, i.e.

$$h\rho_1 = \frac{\mathrm{Re}(\sigma_i)}{e + |\beta_j|},$$

  with $|e| \ll 1$, is "unlikely". Observe that it is expected $|\beta_j|$ non-small.
- In the non-stiff situation, it is expected α negative non-small. In fact, it is expected $|\beta_2|$ non-small and, in the non-stiff situation, it is expected that the right-hand side of (4.4) is small and then it is expected $|\alpha - \beta_2|$ small.

4.3 The basic theorem

The next theorem is, in our new analysis, the analog of Theorem 5.3 in [9] (which was suitable for studying the long-time behavior of γ_n in the non-stiff situation).

Theorem 4.1 *Assume $q > 1$ and $0 \notin \Lambda_1$. Fix $c > 0$ such that $g(c) < 1$, i.e. $c < 1.2564$.*

For $t_n\rho_1 \le \tau$, we have

$$\gamma_n = \gamma_n^{\text{long}}\big(1 + e_n\big),$$

where

$$|e_n| \le \frac{1}{2}\max\left\{\sum_{j=2}^{q}\left(\mathrm{e}^{\beta_j t_n \rho_1}\frac{\left\|Q_j\widehat{y}_0\right\|_2}{\left\|Q_1\widehat{y}_0\right\|_2}\right)^2,\right.$$
$$\left.\sum_{j=2}^{q}\sum_{\lambda_i\in\Lambda_j}\left(\frac{|\mathrm{e}^{\alpha_i t_n\rho_1} - \mathrm{e}^{\beta_j t_n\rho_1}|\tau}{t_n\rho_1}\cdot\frac{K_1}{c(1-g(c))}\cdot\frac{\left\|P_i\widehat{y}_0\right\|_2}{\left\|Q_1\widehat{y}_0\right\|_2}\right)^2\right\}. \tag{4.5}$$

Proof For $n = 0, 1, 2, \ldots$, the error γ_n is given by

$$\gamma_n = \frac{\sqrt{\sum_{j=1}^{q}\left(\mathrm{e}^{(r_j - r_1)t_n}\epsilon_{n,j}\right)^2}}{\sqrt{\sum_{j=1}^{q}\left(\mathrm{e}^{(r_j-r_1)t_n}\left\|Q_j\widehat{y}_0\right\|_2\right)^2}}, \tag{4.6}$$

where

$$\epsilon_{n,j} := \left\|\sum_{\lambda_i\in\Lambda_j}\big(S(h\lambda_i)^n - 1\big)P_i\widehat{y}_0\right\|_2,\ j = 1, \ldots, q,$$

(see Theorem 2.1 in [9]). By (4.6), as applied with the initial value $Q_1 y_0$ instead of y_0, we obtain

$$\gamma_n^{\text{long}} = \frac{\epsilon_{n,1}}{\left\|Q_1\widehat{y}_0\right\|_2}. \tag{4.7}$$

By (4.6) and (4.7), we can write

$$\gamma_n = \gamma_n^{\text{long}}\big(1 + e_n\big),$$

where

$$
\begin{aligned}
|e_n| &= \left| \frac{\sqrt{1+\sum_{j=2}^{q}\left(\mathrm{e}^{(r_j-r_1)t_n}\frac{\epsilon_{n,j}}{\epsilon_{n,1}}\right)^2}}{\sqrt{1+\sum_{j=2}^{q}\left(\mathrm{e}^{(r_j-r_1)t_n}\frac{\|Q_j\widehat{y}_0\|_2}{\|Q_1\widehat{y}_0\|_2}\right)^2}} - 1 \right| \\
&\leq \frac{1}{2}\max\left\{ \sum_{j=2}^{q}\left(\mathrm{e}^{(r_j-r_1)t_n}\frac{\|Q_j\widehat{y}_0\|_2}{\|Q_1\widehat{y}_0\|_2}\right)^2, \sum_{j=2}^{q}\left(\mathrm{e}^{(r_j-r_1)t_n}\frac{\epsilon_{n,j}}{\epsilon_{n,1}}\right)^2 \right\}.
\end{aligned} \tag{4.8}
$$

By Theorem 1.2 and (4.7), we have

$$
\epsilon_{n,1} \geq \frac{t_n\rho_1 E_1}{K_1}(1-g(c))\|Q_1\widehat{y}_0\|_2 = \frac{t_n\rho_1}{\tau}\cdot\frac{c(1-g(c))}{K_1}\|Q_1\widehat{y}_0\|_2.
$$

(The assumption $g(c) < 1$ implies that the right-hand side is positive). Moreover, for $j = 2, \ldots, q$, we have

$$
\begin{aligned}
\mathrm{e}^{(r_j-r_1)t_n}\epsilon_{n,j} &= \left\| \sum_{\lambda_i\in\Lambda_j}\left(\mathrm{e}^{(r_j-r_1)t_n}S(h\lambda_i)^n - \mathrm{e}^{(r_j-r_1)t_n}\right)P_i\widehat{y}_0 \right\|_2 \\
&= \left\| \sum_{\lambda_i\in\Lambda_j}\left(w_i^n - \mathrm{e}^{(r_j-r_1)t_n}\right)P_i\widehat{y}_0 \right\|_2 \\
&= \sqrt{\sum_{\lambda_i\in\Lambda_j}\left|w_i^n - \mathrm{e}^{(r_j-r_1)t_n}\right|^2\|P_i\widehat{y}_0\|_2^2} \\
&= \sqrt{\sum_{\lambda_i\in\Lambda_j}\left|\mathrm{e}^{\alpha_i t_n\rho_1} - \mathrm{e}^{\beta_j t_n\rho_1}\right|^2\|P_i\widehat{y}_0\|_2^2},
\end{aligned} \tag{4.9}
$$

where the third equality follows by A normal, which implies the orthogonality of the eigenspaces. So, in (4.8) we have

$$
\sum_{j=2}^{q}\left(\mathrm{e}^{(r_j-r_1)t_n}\frac{\epsilon_{n,j}}{\epsilon_{n,1}}\right)^2 \leq \sum_{j=2}^{q}\sum_{\lambda_i\in\Lambda_j}\left(\frac{|\mathrm{e}^{\alpha_i t_n\rho_1} - \mathrm{e}^{\beta_j t_n\rho_1}|\tau}{t_n\rho_1}\cdot\frac{K_1}{c(1-g(c))}\cdot\frac{\|P_i\widehat{y}_0\|_2}{\|Q_1\widehat{y}_0\|_2}\right)^2
$$

and (4.5) now follows. □

Remark 4.1 In the case $0 \in \Lambda_1$ and $\Lambda_1 \neq \{0\}$, the theorem still holds holds with (4.5) replaced by

$$|e_n| \leq \frac{1}{2} \max \left\{ \sum_{j=2}^{q} \left(\mathrm{e}^{\beta_j t_n \rho_1} \frac{\left\| Q_j \widehat{y}_0 \right\|_2}{\left\| Q_1 \widehat{y}_0 \right\|_2} \right)^2, \right.$$
$$\left. \sum_{j=2}^{q} \sum_{\lambda_i \in \Lambda_j} \left(\frac{|\mathrm{e}^{\alpha_i t_n \rho_1} - \mathrm{e}^{\beta_j t_n \rho}| \tau}{t_n \rho_1} \cdot \frac{K_1^+}{c(1-g(c))} \cdot \frac{\left\| P_i \widehat{y}_0 \right\|_2}{\sqrt{\sum_{\lambda_k \in \Lambda_1 \setminus \{0\}} \left\| P_k \widehat{y}_0 \right\|_2^2}} \right)^2 \right\},$$

where

$$K_1^+ = \frac{\max\limits_{\lambda_i \in \Lambda_1} |\sigma_i|}{\min\limits_{\lambda_i \in \Lambda_1 \setminus \{0\}} |\sigma_i|} = \left(\frac{\rho_1}{\mu_{\Lambda_1 \setminus \{0\}}} \right)^{l+1} \left(1 + O\big(h\rho_1\big)\right), \ h\rho_1 \to 0.$$

In the following, we continue to assume $0 \notin \Lambda_1$, but all our conclusions are valid (with easy adaptations) also for the case $0 \in \Lambda_1$ and $\Lambda_1 \neq \{0\}$.

4.4 A first result

We give a first theorem about $\gamma_n \approx \gamma_n^{\text{long}}$ in the long-time with a monitor function.

Theorem 4.2 *Assume $q > 1$ and $0 \notin \Lambda_1$. Fix $c > 0$ such that $g(c) < 1$, i.e. $c < 1.2564$.*

We have $\gamma_n \approx \gamma_n^{\text{long}}$ in the long-time with monitor function

$$s(\epsilon, x) = s(\epsilon) = \frac{1}{|\beta_2|} \left(\max \left\{ 0, \log\left(\mathrm{e}^{Mc} - 1\right) + \log K_1 + \log \frac{1}{c(1-g(c))} \right\} \right.$$
$$\left. + \log \frac{\sqrt{1 - \left\| Q_1 \widehat{y}_0 \right\|_2^2}}{\left\| Q_1 \widehat{y}_0 \right\|_2} + \frac{1}{2} \log \frac{1}{\epsilon} - \frac{1}{2} \log 2 \right), \tag{4.10}$$

defined for $\epsilon > 0$.

Proof Let $s \in [0, \tau]$. For $t_n \rho_1 \in [s, \tau]$, in (4.5) of Theorem 4.1 we have

$$\mathrm{e}^{\beta_j t_n \rho_1} \leq \mathrm{e}^{\beta_j s}$$

and

$$\begin{aligned}\frac{|\mathrm{e}^{\alpha_i t_n \rho_1} - \mathrm{e}^{\beta_j t_n \rho_1}|\tau}{t_n \rho_1} =& \mathrm{e}^{\beta_j t_n \rho_1} \frac{|\mathrm{e}^{(\alpha_i - \beta_j) t_n \rho_1} - 1|\tau}{t_n \rho_1} \leq \mathrm{e}^{\beta_j t_n \rho_1} \frac{\left(\mathrm{e}^{|\alpha_i - \beta_j| t_n \rho_1} - 1\right)\tau}{t_n \rho_1} \\ \leq& \mathrm{e}^{\beta_j t_n \rho_1}\left(\mathrm{e}^{|\alpha_i - \beta_j|\tau} - 1\right) \\ =& \mathrm{e}^{\beta_j t_n \rho_1}\left(\mathrm{e}^{M_i c} - 1\right) \leq \mathrm{e}^{\beta_j s}\left(\mathrm{e}^{M_i c} - 1\right),\end{aligned}$$

where the last equality follows by (4.3) (remind (1.13)).

So, for $t_n \rho_1 \in [s, \tau]$, we obtain

$$\begin{aligned}|e_n| \leq& \frac{1}{2} \max\left\{ \sum_{j=2}^{q} \left(\mathrm{e}^{\beta_j s} \frac{\left\|Q_j \widehat{y}_0\right\|_2}{\left\|Q_1 \widehat{y}_0\right\|_2} \right)^2, \right. \\ & \left. \sum_{j=2}^{q} \sum_{\lambda_i \in \Lambda_j} \left(\mathrm{e}^{\beta_j s}\left(\mathrm{e}^{M_i c} - 1\right) \frac{K_1}{c(1 - g(c))} \cdot \frac{\left\|P_i \widehat{y}_0\right\|_2}{\left\|Q_1 \widehat{y}_0\right\|_2} \right)^2 \right\}. \\ \leq& \frac{1}{2} \left(\mathrm{e}^{\beta_2 s} \max\left\{ 1, (\mathrm{e}^{Mc} - 1) \frac{K_1}{c(1 - g(c))} \right\} \frac{\sqrt{1 - \left\|Q_1 \widehat{y}_0\right\|_2^2}}{\left\|Q_1 \widehat{y}_0\right\|_2} \right)^2 = f(s)\end{aligned}$$

By inverting the function f, we obtain the monitor function (4.10). □

Remark 4.2

1. For $\|Q_1 \widehat{y}_0\|_2$ sufficiently close to 1, we have $s(\varepsilon) < 0$. There are two ways for dealing with this. One is to redefine $s(\varepsilon)$ as 0 when $s(\varepsilon) < 0$. The other is to use $(0, a]$ as domain of s, where $s(a) = 0$. So, we have $s(\varepsilon) \geq 0$ for $\varepsilon \in (0, a]$.
2. By (4.10), (1.14) and (1.12), one can easily prove Theorem 1.3.

The previous theorem with $c = 1$ gives the following results.

Theorem 4.3 *Let $\tau = \frac{1}{E_1}$, let $k > 0$ and let*

$$s = \frac{1}{|\beta_2|}\left(M + \log K_1 + \log \frac{\sqrt{1 - \left\|Q_1 \widehat{y}_0\right\|_2^2}}{\left\|Q_1 \widehat{y}_0\right\|_2} + \frac{k}{2} + 0.9203 \right). \tag{4.11}$$

If $\varepsilon = \mathrm{e}^{-k} \ll 1$ and

$$\frac{s}{\tau} \ll 1,$$

then $\gamma_n \approx \gamma_n^{\mathrm{long}}$ for $t_n \rho_1$ in the interval $[s, \tau]$ and $\gamma_n^{\mathrm{long}} \ll 1$ for $t_n \rho_1$ up to the beginning of this interval. In particular, we have $\gamma_n \approx \gamma_n^{\mathrm{long}}$ with degree ε for $t_n \rho_1 \in [s, \tau]$ and

$$\gamma_n^{\text{long}} \le 1.7183\kappa\frac{s}{\tau} \ll 1 \tag{4.12}$$

for $t_n\rho_1 \in [0, \kappa s]$, where $\kappa \ge 1$ is not large.

Proof By putting $c = 1$ in (4.10) and by observing that $\log(\mathrm{e}^M - 1) \le M$, we obtain

$$\begin{aligned} s(\epsilon) \le s &= \frac{1}{|\beta_2|}\left(\max\left\{0, M + \log K_1 + \log\frac{1}{1-g(1)}\right\}\right. \\ &\quad \left. + \log\frac{\sqrt{1-\left\|Q_1\widehat{y}_0\right\|_2^2}}{\left\|Q_1\widehat{y}_0\right\|_2} + \frac{1}{2}\log\frac{1}{\epsilon} - \frac{1}{2}\log 2\right), \\ &= \frac{1}{|\beta_2|}\left(M + \log K_1 + \log\frac{1}{1-g(1)}\right. \\ &\quad \left. + \log\frac{\sqrt{1-\left\|Q_1\widehat{y}_0\right\|_2^2}}{\left\|Q_1\widehat{y}_0\right\|_2} + \frac{1}{2}\log\frac{1}{\epsilon} - \frac{1}{2}\log 2\right) \end{aligned}$$

since $K_1 \ge 1$ and $1 - g(1) < 1$. Now, (4.11) follows since

$$\log\frac{1}{1-g(1)} - \frac{1}{2}\log 2 = 0.9203.$$

About (4.12), take $c = 1$ in (3.3). □

Theorem 4.4 *If*

$$\begin{aligned} &\frac{1}{|\beta_2|}\cdot\frac{\max\limits_{\lambda_i\in\Lambda^-}|\sigma_i|}{h\rho_1} \ll 1 \\ &\frac{1}{|\beta_2|}\log K_1 \cdot E_1 \ll 1 \\ &\frac{1}{|\beta_2|}\log\frac{\sqrt{1-\left\|Q_1\widehat{y}_0\right\|_2^2}}{\left\|Q_1\widehat{y}_0\right\|_2}\cdot E_1 \ll 1 \\ &\frac{1}{|\beta_2|}\left(\frac{1}{2}\log\frac{1}{E_1} + 0.9203\right)\cdot E_1 \ll 1, \end{aligned} \tag{4.13}$$

then $\gamma_n \approx \gamma_n^{\text{long}}$ in the long-time.

Proof Use the previous theorem with $\epsilon = \frac{1}{\tau} = E_1 \ll 1$ and observe that

$$\frac{M}{\tau} = ME_1 = \frac{\max_{\lambda_i\in\Lambda^-}|\sigma_i|}{h\rho_1}.$$

□

It is expected that the last three conditions in (4.13) are satisfied. In fact, since $E_1 \ll 1$, they are not satisfied only in "extreme" cases. Moreover, in the non-stiff situation, it is expected that the first condition is satisfied. In fact, since in the non-stiff situation it is expected $\frac{\max_{\lambda_i \in \Lambda^-} |\sigma_i|}{h\rho_1} \ll 1$, the first condition is not satisfied only in "extreme" cases.

So, we can state the following important conclusion.

Conclusion 4.5 Suppose to be in the non-stiff situation. It is expected $\gamma_n \approx \gamma_n^{\text{long}}$ in the long-time.

4.5 The condition A

We introduce the condition

$$\text{A:} \quad W < 1, \quad \text{equivalently} \quad \alpha < 0,$$

where W and α are defined in (4.1) and (4.2), respectively.

Next theorem shows that, under the condition A, we have $\gamma_n \approx \gamma_n^{\text{long}}$ in the long-time with a new monitor function different from (4.10).

Theorem 4.6 *Assume $q > 1$ and $0 \notin \Lambda_1$. Fix $c > 0$ such that $g(c) < 1$, i.e. $c < 1.2564$.*

If A holds, then $\gamma_n \approx \gamma_n^{\text{long}}$ in the long-time with monitor function

$$s(\epsilon, x) = \frac{1}{\min\{|\alpha|, |\beta_2|\}} \left(\log x + \log K_1 + \log \frac{\sqrt{1 - \|Q_1 \widehat{y}_0\|_2^2}}{\|Q_1 \widehat{y}_0\|_2} + \frac{1}{2} \log \frac{1}{\epsilon} + \log \frac{1}{c(1 - g(c))} + \frac{1}{2} \log 2 \right) \tag{4.14}$$

defined for $x \geq 1$ and for $\epsilon > 0$ such that the right-hand side of (4.14) with $x = 1$ is greater than or equal to 1, so we have $s(\epsilon, x) \geq 1$ for $x \geq 1$ and such ϵ.

Proof Let A holds. Let $\tau \geq 1$ and let $s \in [1, \tau]$. For $t_n \rho_1 \in [s, \tau]$, in (4.5) of Theorem 4.1 we have

$$e^{\beta_j t_n \rho_1} \leq e^{\beta_j s}$$

and

$$\frac{|e^{\alpha_i t_n \rho_1} - e^{\beta_j t_n \rho_1}|}{t_n \rho_1} \leq \frac{e^{\text{Re}(\alpha_i) t_n \rho_1} + e^{\beta_j t_n \rho_1}}{t_n \rho_1} \leq \frac{e^{\text{Re}(\alpha_i) s} + e^{\beta_j s}}{s} \leq e^{\text{Re}(\alpha_i) s} + e^{\beta_j s},$$

where the last equality follows by $s \geq 1$. So, for $t_n \rho \in [s, \tau]$, we obtain

$$
\begin{aligned}
|e_n| \leq & \frac{1}{2} \max \left\{ \sum_{j=2}^{q} \left(\mathrm{e}^{\beta_j s} \frac{\left\| Q_j \widehat{y}_0 \right\|_2}{\left\| Q_1 \widehat{y}_0 \right\|_2} \right)^2, \right. \\
& \left. \sum_{j=2}^{q} \sum_{\lambda_i \in \Lambda_j} \left(\left(\mathrm{e}^{\mathrm{Re}(\alpha_i) s} + \mathrm{e}^{\beta_j s} \right) \tau \frac{K_1}{c(1-g(c))} \cdot \frac{\left\| P_i \widehat{y}_0 \right\|_2}{\left\| Q_1 \widehat{y}_0 \right\|_2} \right)^2 \right\} \\
\leq & \frac{1}{2} \left(\mathrm{e}^{\max\{\alpha, \beta_2\} s} \max \left\{ 1, 2\tau \frac{K_1}{c(1-g(c))} \right\} \frac{\sqrt{1 - \left\| Q_1 \widehat{y}_0 \right\|_2^2}}{\left\| Q_1 \widehat{y}_0 \right\|_2} \right)^2 \\
= & \frac{1}{2} \left(\mathrm{e}^{\max\{\alpha, \beta_2\} s} 2\tau \frac{K_1}{c(1-g(c))} \cdot \frac{\sqrt{1 - \left\| Q_1 \widehat{y}_0 \right\|_2^2}}{\left\| Q_1 \widehat{y}_0 \right\|_2} \right)^2 = f(s, \tau),
\end{aligned}
$$

where the last equality follows by

$$
\tau \frac{K_1}{c(1-g(c))} \geq \frac{1}{c(1-g(c))} > 1.
$$

By inverting the function f with respect to s, we obtain the monitor function (4.14). □

Remark 4.3

1. The monitor function (4.14) is defined for $\varepsilon \in (0, a]$, where

$$
a = \frac{1}{2} \left(\mathrm{e}^{-\min\{|\alpha|, |\beta_2|\}} 2 \frac{K_1}{c(1-g(c))} \cdot \frac{\sqrt{1 - \left\| Q_1 \widehat{y}_0 \right\|_2^2}}{\left\| Q_1 \widehat{y}_0 \right\|_2} \right)^2.
$$

2. By looking at the proof of the previous theorem, we see that there is also a monitor function $s(\varepsilon, x)$ defined for all $\varepsilon > 0$ and $x > 0$. It is obtained by inverting with respect to s the upper bound

$$
\frac{1}{2} \left(\frac{\mathrm{e}^{-\min\{|\alpha|, |\beta_2|\} s}}{s} 2\tau \frac{K_1}{c(1-g(c))} \cdot \frac{\sqrt{1 - \left\| Q_1 \widehat{y}_0 \right\|_2^2}}{\left\| Q_1 \widehat{y}_0 \right\|_2} \right)^2 = f(s, \tau).
$$

of $|e_n|$, where e_n is given in Theorem 4.1. Observe that the inverse exists since $\frac{\mathrm{e}^{-\min\{|\alpha|, |\beta_2|\} s}}{s}$ is a strictly decreasing function of s. This new monitor function has

the advantage that it is no longer necessary to suppose $\varepsilon \le a$ (where is a given at point 1) above) and $\tau \ge 1$. However, we prefer to use the old monitor function (4.14) because it has an explicit expression.

The previous theorem with $c = 1$ gives the next important results.

Theorem 4.7 *Suppose A holds. Let* $\tau = \frac{1}{E_1}$, *let* $k > 0$ *and let*

$$s = \frac{1}{\min\{|\alpha|, |\beta_2|\}} \left(\log \tau + \log K_1 + \log \frac{\sqrt{1 - \left\| Q_1 \widehat{y}_0 \right\|_2^2}}{\left\| Q_1 \widehat{y}_0 \right\|_2} + \frac{k}{2} + 1.6134 \right). \tag{4.15}$$

If $\varepsilon = \mathrm{e}^{-k} \ll 1$ *and*

$$\frac{\max\{1, s\}}{\tau} \ll 1,$$

then $\gamma_n \approx \gamma_n^{\text{long}}$ *for* $t_n \rho_1$ *in the interval* $[\max\{1, s\}, \tau]$ *and* $\gamma_n^{\text{long}} \ll 1$ *for* $t_n \rho_1$ *up to the beginning of this interval. In particular, we have* $\gamma_n \approx \gamma_n^{\text{long}}$ *with degree* ε *for* $t_n \rho_1 \in [\max\{1, s\}, \tau]$ *and*

$$\gamma_n^{\text{long}} \le 1.7183 \kappa \frac{\max\{1, s\}}{\tau} \ll 1$$

for $t_n \rho_1 \in [0, \kappa \max\{1, s\}]$, *where* $\kappa \ge 1$ *is not large.*

Proof We use Theorem 4.6 with $c = 1$. If $s \ge 1$, then $s(\varepsilon, \tau) = s$. Observe that in (4.14) with $c = 1$ we have

$$\log \frac{1}{1 - g(1)} + \frac{1}{2} \log 2 = 1.6134.$$

Theorem 4.6 says that $\gamma_n \approx \gamma_n^{\text{long}}$ with degree ε for $t_n \rho_1 \in [s, \tau]$. If $s < 1$, consider $\overline{k}$, with $\overline{k} > k$, such that

$$\frac{1}{\min\{|\alpha|, |\beta_2|\}} \left(\log \tau + \log K_1 + \log \frac{\sqrt{1 - \left\| Q_1 \widehat{y}_0 \right\|_2^2}}{\left\| Q_1 \widehat{y}_0 \right\|_2} + \frac{\overline{k}}{2} + 1.6134 \right) = 1.$$

We have $s(\overline{\varepsilon}, \tau) = 1$, where $\overline{\varepsilon} = \mathrm{e}^{-\overline{k}}$, with $\overline{\varepsilon} < \varepsilon$. Theorem 4.6 says that $\gamma_n \approx \gamma_n^{\text{long}}$ with degree $\overline{\varepsilon}$, and then with degree ε, for $t_n \rho_1 \in [1, \tau]$. □

Theorem 4.8 *Suppose A holds. If*

$$\begin{aligned}
&\frac{1}{\min\{|\alpha|,|\beta_2|\}}\log K_1\cdot E_1 \ll 1\\
&\frac{1}{\min\{|\alpha|,|\beta_2|\}}\log\frac{\sqrt{1-\left\|Q_1\widehat{y}_0\right\|_2^2}}{\left\|Q_1\widehat{y}_0\right\|_2}\cdot E_1 \ll 1\\
&\frac{1}{\min\{|\alpha|,|\beta_2|\}}\left(\frac{3}{2}\log\frac{1}{E_1}+1.6134\right)\cdot E_1 \ll 1,
\end{aligned} \tag{4.16}$$

then $\gamma_n \approx \gamma_n^{\text{long}}$ *in the long-time.*

Proof Use the previous theorem with $\varepsilon = \frac{1}{\tau} = E_1 \ll 1$. □

It is expected that if A holds, then the three conditions in (4.16) are satisfied. In fact, since $E_1 \ll 1$, they are not satisfied only in "extreme" cases. So, it is expected that if A holds then $\gamma_n \approx \gamma_n^{\text{long}}$ in the long-time. Of course, we already know that in the non-stiff situation it is expected $\gamma_n \approx \gamma_n^{\text{long}}$ in the long-time, independently of the condition A, as well as we know that it expected that A holds in the non-stiff situation.

So, what is really important is the following conclusion.

Conclusion 4.9 Suppose to be in the stiff situation. It is expected that if A holds, then $\gamma_n \approx \gamma_n^{\text{long}}$ in the long-time.

4.5.1 Order star and stability region

The condition A can be related to the order star and the stability region of the approximant (recall the beginning of Sect. 2).

Theorem 4.10 *Let* $\mathscr{S}^c$ *be the complementary set of the order star of the approximant. The condition*

$$h\lambda_i \in \mathscr{S}^c \text{ for any } \lambda_i \in \Lambda^-$$

implies A.

Proof Let $\lambda_i \in \Lambda^-$. If $h\lambda_i \in \mathscr{S}^c$, then $|w_i| < 1$. In fact, if $h\lambda_i \in \mathscr{S}^c$, then

$$|w_i| = \mathrm{e}^{h(r_j - r_1)}\left|S\left(h\lambda_i\right)\right| < \left|S\left(h\lambda_i\right)\right| \le 1,$$

where $j = 2, \ldots, q$ is such that $\lambda_i \in \Lambda_j$. □

Theorem 4.11 *Let*

$$\overset{\circ}{\mathscr{R}} = \{z \in \mathscr{D} : |R(z)| < 1\}$$

be the interior of the stability region $\mathscr{R}$ of the approximant. If $r_1 \leq 0$, then A implies the condition

$$h\lambda_i \in \overset{\circ}{\mathscr{R}} \text{ for any } \lambda_i \in \Lambda^-. \tag{4.17}$$

If $r_1 \geq 0$, then the condition (4.17) implies A.

Proof Let $\lambda_i \in \Lambda^-$. If $r_1 \leq 0$ and $|w_i| < 1$, then

$$|R(h\lambda_i)| = \mathrm{e}^{hr_1}|w_i| \leq |w_i| < 1.$$

If $r_1 \geq 0$ and $h\lambda_i \in \overset{\circ}{\mathscr{R}}$, then

$$|w_i| = \mathrm{e}^{-hr_1}|R(h\lambda_i)| \leq |R(h\lambda_i)| < 1.$$

□

These theorems agree with what has been observed in the examples of Sect. 2 about order stars and stability regions.

4.5.2 The region $\mathscr{R}_x$

For any $x \in \mathbb{R}$, let

$$\mathscr{R}_x := \{z \in \mathscr{D} : |R(z)| < \mathrm{e}^x\}. \tag{4.18}$$

We have $\mathscr{R}_0 = \overset{\circ}{\mathscr{R}}$.

For $\lambda_i \in \Lambda^-$, we have

$$|w_i| = \mathrm{e}^{-hr_1}|R(h\lambda_i)|.$$

Thus, the condition A can be restated as

$$h\lambda_i \in \mathscr{R}_{hr_1} \text{ for any } \lambda_i \in \Lambda^-.$$

In the case $r_1 = 0$, it becomes

$$h\lambda_i \in \overset{\circ}{\mathscr{R}} \text{ for any } \lambda_i \in \Lambda^-$$

according to Theorem 4.11 about the cases $r_1 \leq 0$ and $r_1 \geq 0$.

4.5.3 The conditions B and C

When A does not hold, we have B or C, where

$$\text{B:} \quad W > 1, \quad \text{equivalently} \quad \alpha > 0,$$

and

$$\text{C:}\quad W = 1, \quad \text{equivalently} \quad \alpha = 0.$$

In the next section we study, what happens when A does not holds. Of course, when A does not hold, it is expected that B holds.

5 When the condition A does not hold

Next theorem helps to say when $\frac{\gamma_n}{\gamma_n^{\text{long}}} \gg 1$. We exclude the time $t_n\rho_1 = 0$, i.e. the index $n = 0$, since the ratio $\frac{\gamma_n}{\gamma_n^{\text{long}}}$ for $n = 0$ is indeterminate $\frac{0}{0}$.

Theorem 5.1 *Assume $q > 1$. Fix $c > 0$.*

For $t_n\rho_1 \in (0, \tau]$, we have

$$\frac{\gamma_n}{\gamma_n^{\text{long}}} \geq \max_{\lambda_i \in \Lambda^-} \frac{|e^{\alpha_i t_n \rho_1} - e^{\beta_j t_n \rho_1}|\tau}{t_n \rho_1} \cdot \frac{1}{c(1 + g(c))} \left\| P_i \widehat{y}_0 \right\|_2, \tag{5.1}$$

where $j = 2, \ldots, q$ is such that $\lambda_i \in \Lambda_j$.

Proof Recall (4.6) and (4.7). For any $j = 2, \ldots, q$ and $\lambda_i \in \Lambda_j$ we have, for $n = 0, 1, 2, \ldots$,

$$\begin{aligned}
\frac{\gamma_n}{\gamma_n^{\text{long}}} &= \frac{\sqrt{1 + \sum_{m=2}^{q} \left(e^{(r_m - r_1) t_n} \frac{\epsilon_{n,m}}{\epsilon_{n,1}} \right)^2}}{\sqrt{1 + \sum_{m=2}^{q} \left(e^{(r_m - r_1) t_n} \frac{\left\| Q_m \widehat{y}_0 \right\|_2}{\left\| Q_1 \widehat{y}_0 \right\|_2} \right)^2}} \\
&= \frac{\sqrt{1 + \sum_{m=2}^{q} \sum_{\lambda_k \in \Lambda_m} \left(\frac{|e^{\alpha_k t_n \rho_1} - e^{\beta_m t_n \rho_1}| \left\| P_k \widehat{y}_0 \right\|_2}{\epsilon_{n,1}} \right)^2}}{\sqrt{1 + \sum_{m=2}^{q} \left(e^{\beta_m t_n \rho_1} \frac{\left\| Q_m \widehat{y}_0 \right\|_2}{\left\| Q_1 \widehat{y}_0 \right\|_2} \right)^2}} \\
&\geq \frac{\left| e^{\alpha_i t_n \rho_1} - e^{\beta_j t_n \rho_1} \right|}{\epsilon_{n,1}} \left\| P_i \widehat{y}_0 \right\|_2 \left\| Q_1 \widehat{y}_0 \right\|_2,
\end{aligned}$$

where the second equality follows by (4.9).

By Theorem 1.2, we have

$$\epsilon_{n,1} \leq t_n \rho_1 E_1 (1 + g(c)) \left\| Q_1 \widehat{y}_0 \right\|_2$$

and then

$$\frac{\gamma_n}{\gamma_n^{\text{long}}} \geq \frac{|e^{\alpha_i t_n \rho_1} - e^{\beta_j t_n \rho_1}|\tau}{t_n \rho_1} \cdot \frac{1}{c(1 + g(c))} \left\| P_i \widehat{y}_0 \right\|_2.$$

The inequality (5.1) now follows. □

5.1 Definition of $\frac{\gamma_n}{\gamma_n^{\text{long}}} \gg 1$ for all time

Here, "for all time" we do not mean for all times $t_n\rho_1$, since in our analysis we consider $t_n\rho_1$ up to τ. So, we introduce the following definition

Definition 5.1 We say that $\frac{\gamma_n}{\gamma_n^{\text{long}}} \gg 1$ *for all time* if $\frac{\gamma_n}{\gamma_n^{\text{long}}} \gg 1$ for $t_n\rho_1 \in (0, \tau]$.

This definition is made more precise by using a monitor function.

Definition 5.2 Let $F : (0, +\infty) \to (0, +\infty)$ such that

$$\lim_{x \to +\infty} F(x) = +\infty. \tag{5.2}$$

We say that $\frac{\gamma_n}{\gamma_n^{\text{long}}} \gg 1$ for all time with monitor function F if

$$\frac{\gamma_n}{\gamma_n^{\text{long}}} \geq F(\tau) \text{ for } t_n\rho_1 \in (0, \tau].$$

Remark 5.1 In the previous definition, we also allow monitor functions $F : [b, +\infty) \to [0, +\infty)$, where $0 < b < +\infty$.

Thus:

if $\frac{\gamma_n}{\gamma_n^{\text{long}}} \gg 1$ for all time with monitor function F and $F(\tau) \gg 1$,
then $\frac{\gamma_n}{\gamma_n^{\text{long}}} \gg 1$ for all time. In particular, we have $\frac{\gamma_n}{\gamma_n^{\text{long}}} \gg F(\tau)$
for $t_n\rho_1 \in (0, \tau]$.

Regarding the satisfiability of $F(\tau) \gg 1$, observe that F satisfies (5.2) and we have $\tau \gg 1$.

5.2 The condition B

The next theorem explains what happens under the condition B.

Theorem 5.2 *Assume $q > 1$. Fix $c > 0$.*

If B holds, then $\frac{\gamma_n}{\gamma_n^{\text{long}}} \gg 1$ for all time with monitor function

$$F(x) = \frac{1}{c(1+g(c))}\alpha \min_{\substack{\lambda_i \in \Lambda^- \\ \mathrm{Re}(\alpha_i) > 0}} \left\|P_i \widehat{y}_0\right\|_2 \cdot x \tag{5.3}$$

defined for $x > 0$.

Proof Let B holds. For $t_n\rho_1 \in (0, \tau]$, in (5.1) we have

$$\frac{\left|\mathrm{e}^{\alpha_i t_n \rho_1} - \mathrm{e}^{\beta_j t_n \rho_1}\right|}{t_n \rho_1} \geq \frac{\mathrm{e}^{\mathrm{Re}(\alpha_i) t_n \rho_1} - 1}{t_n \rho_1}.$$

If $\mathrm{Re}(\alpha_i) \leq 0$, we have

$$\frac{\mathrm{e}^{\mathrm{Re}(\alpha_i) t_n \rho_1} - 1}{t_n \rho_1} \leq 0.$$

If $\mathrm{Re}(\alpha_i) > 0$, we have

$$\frac{\mathrm{e}^{\mathrm{Re}(\alpha_i) t_n \rho_1} - 1}{t_n \rho_1} \geq \mathrm{Re}(\alpha_i).$$

So, by (5.1), we obtain

$$\begin{aligned} \frac{\gamma_n}{\gamma_n^{\mathrm{long}}} &\geq \max_{\lambda_i \in \Lambda^-} \frac{\left(\mathrm{e}^{\mathrm{Re}(\alpha_i) t_n \rho_1} - 1\right)\tau}{t_n \rho_1} \cdot \frac{1}{c(1+g(c))} \cdot \left\|P_i \widehat{y}_0\right\|_2 \\ &\geq \max_{\substack{\lambda_i \in \Lambda^- \\ \mathrm{Re}(\alpha_i) > 0}} \mathrm{Re}(\alpha_i)\tau \frac{1}{c(1+g(c))} \left\|P_i \widehat{y}_0\right\|_2 \geq F(\tau) \end{aligned}$$

for $t_n\rho_1 \in (0, \tau]$, where F is the function in (5.3). □

Remark 5.2 We can have another monitor function by substituting in (5.3) the term $\alpha \min_{\substack{\lambda_i \in \Lambda^- \\ \mathrm{Re}(\alpha_i) > 0}} \left\|P_i \widehat{y}_0\right\|_2$ by $\min_{\substack{\lambda_i \in \Lambda^- \\ \mathrm{Re}(\alpha_i) > 0}} \mathrm{Re}(\alpha_i) \cdot \max_{\substack{\lambda_i \in \Lambda^- \\ \mathrm{Re}(\alpha_i) > 0}} \left\|P_i \widehat{y}_0\right\|_2$.

The previous theorem with $c = 1$ gives the following important results.

Theorem 5.3 *Suppose B holds. Let* $\tau = \frac{1}{E_1}$. *If*

$$F(\tau) = 0.5820 \cdot \alpha \min_{\substack{\lambda_i \in \Lambda^- \\ \mathrm{Re}(\alpha_i) > 0}} \left\|P_i \widehat{y}_0\right\|_2 \tau \gg 1, \tag{5.4}$$

then $\frac{\gamma_n}{\gamma_n^{\mathrm{long}}} \gg 1$ *for* $t_n\rho_1 \in (0, \tau]$. *In particular, we have* $\frac{\gamma_n}{\gamma_n^{\mathrm{long}}} \geq F(\tau)$ *for* $t_n\rho_1 \in (0, \tau]$.

Theorem 5.4 *Suppose B holds. If*

$$\alpha \min_{\substack{\lambda_i \in \Lambda^- \\ \mathrm{Re}(\alpha_i) > 0}} \left\| P_i \widehat{y}_0 \right\|_2 \cdot \frac{1}{E_1} \gg 1, \tag{5.5}$$

then $\frac{\gamma_n}{\gamma_n^{\mathrm{long}}} \gg 1$ *for all time*.

In the stiff situation, it is expected that if B holds, then (5.5) holds. In fact, $E_1 \ll 1$ and it is expected $|\alpha|$ non-small. So, we can state the following important conclusion.

Conclusion 5.5 Suppose to be in the stiff situation. It is expected that if B holds, then $\frac{\gamma_n}{\gamma_n^{\mathrm{long}}} \gg 1$ for all time.

The all time lower bound (5.4) of the ratio $\frac{\gamma_n}{\gamma_n^{\mathrm{long}}}$ is proportional to $\alpha\tau$. At the end of the interval $[0, \tau]$ this ratio has a lower bound exponential in $\alpha\tau$.

In fact, by Theorem 5.1, we see that for $t_n\rho_1 \in [\kappa\tau, \tau]$, where $\kappa \in (0, 1]$ is not small,

$$\frac{\gamma_n}{\gamma_n^{\mathrm{long}}} \geq \frac{1}{c(1+g(c))} \min_{\substack{\lambda_i \in \Lambda^- \\ \mathrm{Re}(\alpha_i) > 0}} \left\| P_i \widehat{y}_0 \right\|_2 \frac{\mathrm{e}^{\kappa\alpha\tau} - 1}{\kappa}.$$

Moreover, by Theorem 1.2 we obtain

$$\gamma_n \geq \frac{1 - g(c)}{1 + g(c)} \cdot \frac{1}{K_1} \min_{\substack{\lambda_i \in \Lambda^- \\ \mathrm{Re}(\alpha_i) > 0}} \left\| P_i \widehat{y}_0 \right\|_2 (e^{\kappa\alpha\tau} - 1).$$

5.3 The condition C

Although it is expected that C does not hold, we study anyway the condition C since it characterizes the transition between $\gamma_n \approx \gamma_n^{\mathrm{long}}$ in the long-time and $\frac{\gamma_n}{\gamma_n^{\mathrm{long}}} \gg 1$ for all time.

For the condition C, we need a weak form of $\frac{\gamma_n}{\gamma_n^{\mathrm{long}}} \gg 1$ for all time.

Definition 5.3 Let $S : (0, +\infty) \to (0, +\infty)$ be a function such that

$$S(x) \leq x \text{ for } x > 0 \text{ and } \lim_{x \to +\infty} S(x) = +\infty.$$

We say that *S-weakly* $\frac{\gamma_n}{\gamma_n^{\mathrm{long}}} \gg 1$ *for all time* if $\frac{\gamma_n}{\gamma_n^{\mathrm{long}}} \gg 1$ for $t_n\rho_1 \in (0, S(\tau)]$.

Here is the definition with a monitor function.

Definition 5.4 Let $S, F : (0, +\infty) \to (0, +\infty)$ such that

$$S(x) \le x \text{ for } x > 0, \quad \lim_{x\to+\infty} S(x) = +\infty \text{ and } \lim_{x\to+\infty} F(x) = +\infty.$$

We say that *S-weakly* $\frac{\gamma_n}{\gamma_n^{\text{long}}} \gg 1$ *for all time with monitor function F* if

$$\frac{\gamma_n}{\gamma_n^{\text{long}}} \ge F(\tau) \text{ for } t_n\rho_1 \in (0, S(\tau)].$$

Remark 5.3 In the two previous definitions, we also allow functions $S, F : [b, +\infty) \to [0, +\infty)$, where $0 < b < +\infty$.

Thus:

if S-weakly $\frac{\gamma_n}{\gamma_n^{\text{long}}} \gg 1$ for all time with monitor function F and $F(\tau) \gg 1$, then S-weakly $\frac{\gamma_n}{\gamma_n^{\text{long}}} \gg 1$ for all time. In particular, we have $\frac{\gamma_n}{\gamma_n^{\text{long}}} \gg F(\tau)$ for $t_n\rho_1 \in (0, S(\tau)]$.

The next theorem explains what happens under the condition C.

Theorem 5.6 *Assume $q > 1$. Fix $c > 0$. Let $v \in (0, 1)$ and let*

$$S(x) = x^v, \ x \ge 1.$$

If C holds, then S-weakly $\frac{\gamma_n}{\gamma_n^{\text{long}}} \gg 1$ *for all time with monitor function*

$$F(x) = \frac{1}{c(1+g(c))}(1 - e^{\beta_2}) \max_{\substack{\lambda_i \in \Lambda^- \\ \alpha_i = 0}} \left\|P_i \widehat{y}_0\right\|_2 \cdot x^{1-v} \tag{5.6}$$

defined for $x \ge 1$.

Proof Let C holds. Let $S \in (0, \tau]$. For $t_n\rho_1 \in (0, S]$, in (5.1) we have

$$\frac{\left|e^{\alpha_i t_n\rho_1} - e^{\beta_j t_n\rho_1}\right|}{t_n\rho_1} \ge \frac{e^{\text{Re}(\alpha_i)t_n\rho_1} - e^{\beta_j t_n\rho_1}}{t_n\rho_1}$$

If $\text{Re}(\alpha_i) < 0$, we have

$$\frac{e^{\text{Re}(\alpha_i)t_n\rho_1} - e^{\beta_j t_n\rho_1}}{t_n\rho_1} \le \frac{1 - e^{\beta_j t_n\rho_1}}{t_n\rho_1}.$$

If $\text{Re}(\alpha_i) = 0$, we have

$$\frac{e^{\text{Re}(\alpha_i)t_n\rho_1} - e^{\beta_j t_n\rho_1}}{t_n\rho_1} = \frac{1 - e^{\beta_j t_n\rho_1}}{t_n\rho_1}.$$

Thus, by (5.1), we obtain

$$\begin{aligned}\frac{\gamma_n}{\gamma_n^{\text{long}}} &\geq \max_{\lambda_i \in \Lambda^-} \frac{\left(\mathrm{e}^{\mathrm{Re}(\alpha_i)t_n\rho_1} - \mathrm{e}^{\beta_j t_n \rho_1}\right)\tau}{t_n\rho_1} \cdot \frac{1}{c(1+g(c))} \cdot \left\|P_i \widehat{y}_0\right\|_2 \\ &= \max_{\substack{\lambda_i \in \Lambda^- \\ \mathrm{Re}(\alpha_i)=0}} \frac{\left(1 - \mathrm{e}^{\beta_j t_n \rho_1}\right)\tau}{t_n\rho_1} \cdot \frac{1}{c(1+g(c))} \cdot \left\|P_i \widehat{y}_0\right\|_2 \\ &\geq \max_{\substack{\lambda_i \in \Lambda^- \\ \mathrm{Re}(\alpha_i)=0}} \frac{\left(1 - \mathrm{e}^{\beta_j S}\right)\tau}{S} \cdot \frac{1}{c(1+g(c))} \cdot \left\|P_i \widehat{y}_0\right\|_2\end{aligned} \tag{5.7}$$

for $t_n\rho_1 \in (0, S]$. In particular, for $S = \tau^v$, in (5.7) we have

$$\frac{\left(1 - \mathrm{e}^{\beta_j S}\right)\tau}{S} = \left(1 - \mathrm{e}^{\beta_j \tau^v}\right)\tau^{1-v} \geq \left(1 - \mathrm{e}^{\beta_j}\right)\tau^{1-v}$$

whenever $\tau \geq 1$. Then

$$\frac{\gamma_n}{\gamma_n^{\text{long}}} \geq \max_{\substack{\lambda_i \in \Lambda^- \\ \mathrm{Re}(\alpha_i)=0}} \left(1 - \mathrm{e}^{\beta_j}\right)\tau^{1-v} \frac{1}{c(1+g(c))} \left\|P_i \widehat{y}_0\right\|_2 \geq F(\tau)$$

for $t_n\rho_1 \leq (0, \tau^v]$, where F is the functions in (5.6). □

The previous theorem with $c = 1$ gives the following results.

Theorem 5.7 *Suppose C holds. Let* $\tau = \frac{1}{E_1}$. *Let* $v \in (0, 1)$. *If*

$$F(\tau) = 0.5820 \cdot \left(1 - \mathrm{e}^{\beta_2}\right) \max_{\substack{\lambda_i \in \Lambda^- \\ \mathrm{Re}(\alpha_i)=0}} \left\|P_i \widehat{y}_0\right\|_2 \tau^{1-v} \gg 1$$

then $\frac{\gamma_n}{\gamma_n^{\text{long}}} \gg 1$ *for* $t_n\rho_1 \in (0, \tau^v]$. *In particular, we have* $\frac{\gamma_n}{\gamma_n^{\text{long}}} \geq F(\tau)$ *for* $t_n\rho_1 \in (0, \tau^v]$.

Theorem 5.8 *Suppose C holds. Let* $v \in (0, 1)$. *If*

$$\left(1 - \mathrm{e}^{\beta_2}\right) \max_{\substack{\lambda_i \in \Lambda^- \\ \mathrm{Re}(\alpha_i)=0}} \left\|P_i \widehat{y}_0\right\|_2 \cdot \frac{1}{E_1^{1-v}} \gg 1, \tag{5.8}$$

then S-weakly $\frac{\gamma_n}{\gamma_n^{\text{long}}} \gg 1$ *for all time, where* $S(x) = x^v,\ x \geq 1$.

Suppose to be in the stiff situation and suppose that C holds and $E_1^{1-v} \ll 1$. Then it is expected that (5.8) holds. So, we can state the following conclusion.

Conclusion 5.9 Let $S(x) = x^v$, $x \geq 1$, with $v \in (0,1)$ such that $E_1^{1-v} \ll 1$. Suppose to be in the stiff situation and suppose that C holds. It is expected S-weakly $\frac{\gamma_n}{\gamma_n^{\text{long}}} \gg 1$ for all time.

6 Examples revisited

Now, we look at the two examples of Sect. 2 in the light of the results of Sects. 4 and 5.

6.1 Same approximant with different ODEs

The conditions A, B and C, are $W < 1$, $W > 1$ and $W = 1$, respectively, where

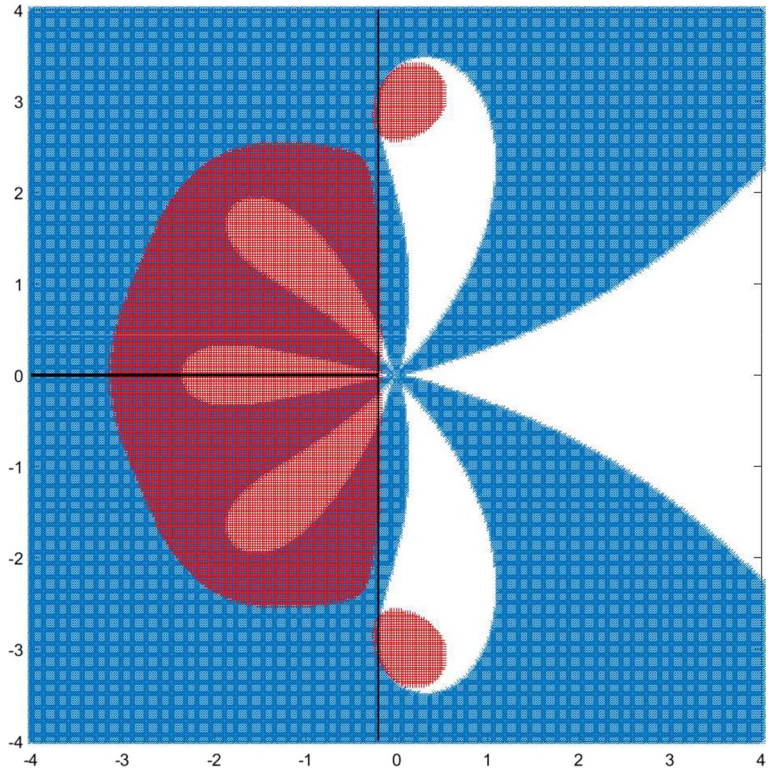

Fig. 9 Order star (in blue) and complementary set (in white) of the five order Taylor approximant with the region $\mathscr{R}_{-0.2}$ (in red) overlapped: the horizontal black half-line is $(-\infty, ha) = (-\infty, -0.2)$ and the vertical black line is the line $\text{Re}(z) = ha = -0.2$

$$W = |w_2| = \mathrm{e}^{-ha}|R(hb)|$$

with $ha = -0.2$. We have

	W	$\alpha = \mathrm{Re}(\alpha_2) = \frac{\log \lvert w_2\rvert}{h\lvert a\rvert}$	$\beta_2 = \frac{b-a}{\lvert a\rvert}$
(P1)	0.00986	− 23.1	− 10
(P2)	0.387	− 4.75	− 12.5
(P3)	1.183	1.19	− 15

in the three possibilities for b. With $c = 1$, we have

$$\tau = \frac{1}{E_1} = 1.89 \cdot 10^6.$$

In (P1) and (P2), the condition A holds. The values of s in in (4.15) relevant to $k = 3$, i.e. $\varepsilon = 4.98 \cdot 10^{-2}$, are:

$$s = \begin{cases} 1.87 \text{ in (P1)} \\ 3.93 \text{ in (P2).} \end{cases}$$

We have $\gamma_n \approx \gamma_n^{\mathrm{long}}$ with degree ε for $t_n\rho_1 = t_n \in [s, \tau]$ and

$$\gamma_n^{\mathrm{long}} \le 1.7183\kappa\frac{s}{\tau} \le \kappa s \cdot 10^{-6} \ll 1$$

for $t_n\rho_1 \in [0, \kappa s]$, where $\kappa \ge 1$ is not large. We have $\gamma_n \approx \gamma_n^{\mathrm{long}}$in the long-time. Observe that the values s agree with Figs. 2 and 3.

In (P3), the condition B holds. The value of the monitor function (5.4) is $F(\tau) = 8.79 \cdot 10^5$. We have $\frac{\gamma_n}{\gamma_n^{\mathrm{long}}} \ge F(\tau)$ for $t_n\rho_1 = t_n \in (0, \tau]$ and so $\frac{\gamma_n}{\gamma_n^{\mathrm{long}}} \gg 1$ for all time.

6.1.1 The region $\mathscr{R}_{hr_1}$

Recall Sect. 4.5.2. The condition A can be stated as

$$hb \in \mathscr{R}_{ha} = \mathscr{R}_{-0.2}$$

or, since $hb < ha$,

$$hb \in \mathscr{R}_{ha} \cap (-\infty, ha) = \mathscr{R}_{-0.2} \cap (-\infty, -0.2).$$

The region $\mathscr{R}_{-0.2}$ is shown in Fig. 9 (compare with Fig. 5 showing $\mathscr{R}_0 = \overset{\circ}{\mathscr{R}}$). The part of $\mathscr{R}_{-0.2} \cap (-\infty, -0.2)$ in the white finger corresponds to the sufficient condition $hb \in \mathscr{S}$. Out of the white finger, we have an additional range of values for hb guaranteeing the condition A. The border value for b between the conditions A and B, where the condition C holds, is $b = -15.565$. Observe that we are out of the white finger for $b < -11.887$ and out of the stability region for $b < -16.085$.

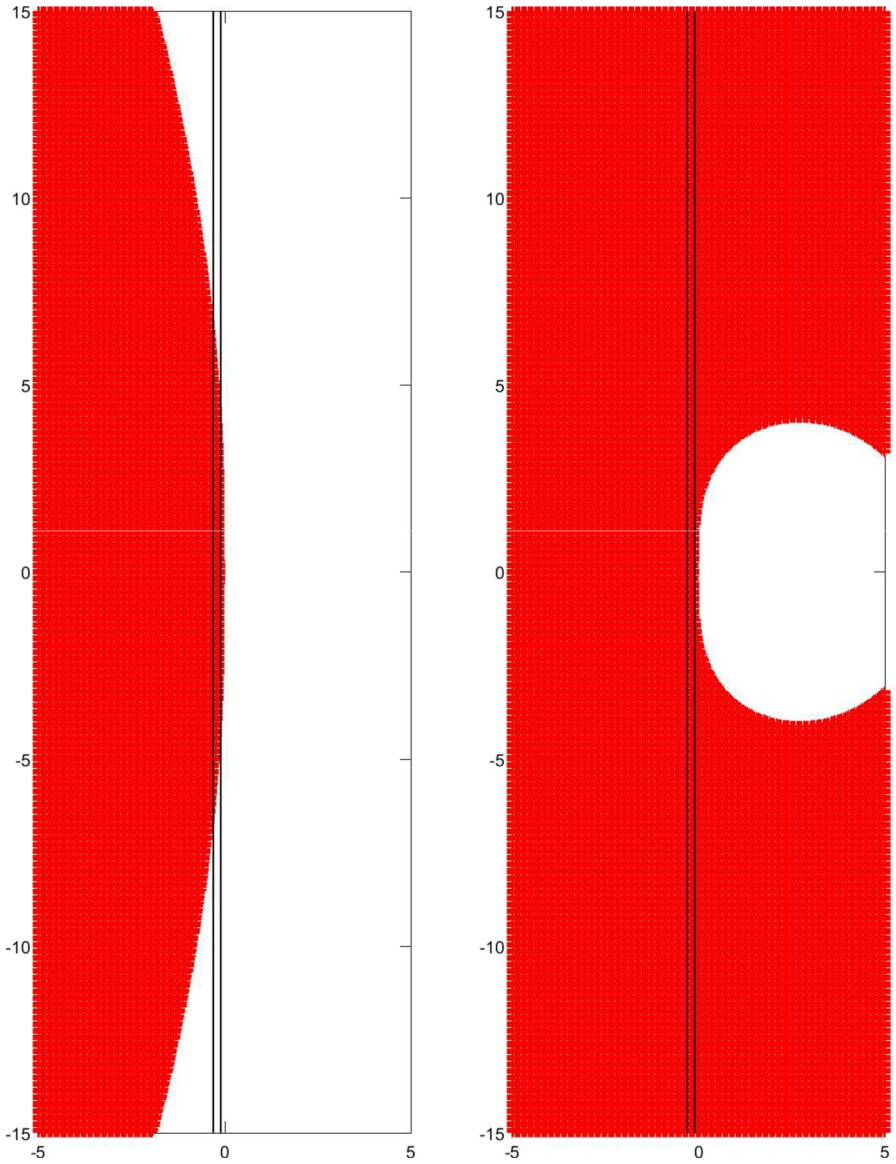

Fig. 10 Region $\mathscr{R}_{-0.1}$ (in red) for the Gauss RK method (left) and the Radau RK method (right). The two vertical black lines are the lines $\mathrm{Re}(z) = hr_1 = -0.1$ and $\mathrm{Re}(z) = hr_2 = -0.3$

6.2 Same ODE with different approximants

The conditions A, B and C, are $W < 1$, $W > 1$ and $W = 1$, respectively, where

$$W = |w_3| = \mathrm{e}^{-hr_1}\left|R\left(h\lambda_3\right)\right| < 1$$

with $hr_1 = -0.1$. We have

	$W = \lvert w_3\rvert$	$\alpha = \mathrm{Re}(\alpha_3) = \frac{\log\lvert w_3\rvert}{h\lvert\lambda_1\rvert}$	$\beta_2 = \frac{r_2 - r_1}{\lvert\lambda_1\rvert}$
Guass RK method	1.105	2.12	$-\sqrt{2}$
Radau RK method	0.0221	-27.0	$-\sqrt{2}$

and, with $c = 1$,

$$\tau = \frac{1}{E_1} = \begin{cases} 1.80 \cdot 10^6 & \text{for the Gauss RK method} \\ 2.61 \cdot 10^4 & \text{for the Radau RK method.} \end{cases}$$

For the Gauss RK method, the condition B holds. The value of the monitor function (5.4) is $F(\tau) = 3.31 \cdot 10^5$. We have $\frac{\gamma_n}{\gamma_n^{\mathrm{long}}} \geq F(\tau)$ for $t_n\rho_1 = \sqrt{2}t_n \in (0, \tau]$ and so $\frac{\gamma_n}{\gamma_n^{\mathrm{long}}} \gg 1$ for all time.

For the Radau RK method, the condition A holds. The value of s in (4.15) relevant to $k = 3$ is $s = 9.25$. We have $\gamma_n \approx \gamma_n^{\mathrm{long}}$ with degree $\varepsilon = 4.98 \cdot 10^{-2}$ for $t_n\rho_1 = \sqrt{2}t_n \in [s, \tau]$ and

$$\gamma_n^{\mathrm{long}} \leq 1.7183\kappa\frac{s}{\tau} \leq \kappa s \cdot \begin{cases} 10^{-6} & \text{for the Gauss RK method} \\ 10^{-4} & \text{for the Radau RK method.} \end{cases} \ll 1$$

for $t_n\rho_1 \in [0, \kappa s]$, where $\kappa \geq 1$ is not large. We have $\gamma_n \approx \gamma_n^{\mathrm{long}}$in the long-time. With reference to Fig. 6, we have $\gamma_n \approx \gamma_n^{\mathrm{long}}$ with degree ε for $t_n \in [8, 10]$, i.e. for $t_n\rho = \sqrt{2}t_n \in [11.31, 14.14]$.

6.2.1 The region $\mathscr{R}_{hr_1}$

The condition A can be written as

$$h\lambda_3 \in \mathscr{R}_{hr_1} = \mathscr{R}_{-0.1}.$$

The region $\mathscr{R}_{-0.1}$ for the two methods is shown in Fig. 10. In the left part of the figure, we see that the region for the Gauss RK method does not cover points with large imaginary part on the line

$$\mathrm{Re}(z) = hr_2 = \mathrm{Re}(h\lambda_3) = -0.3.$$

On the other hand, in the right part, we see that the region for the Radau RK method completely includes this line.

7 Independence of the non-rightmost spectrum

In this section, we study when the condition A holds independently of the particular non-rightmost spectrum Λ^-.

Here, we consider an analytic approximant R with domain $\mathscr{D}$ such that $\{z \in \mathbb{C} : \mathrm{Re}(z) < \beta_R\} \subseteq \mathscr{D}$ for some $\beta_R \in (0, +\infty]$, i.e. $\mathscr{D}$ includes a left half-plane.

7.1 The property A(x)

We introduce the property A(x) of the approximant R.

Definition 7.1 Let $x < \beta_R$. Let

$$\mathrm{A}(x) \overset{\text{def}}{\iff} \mathrm{e}^{-x}|R(z)| < 1 \text{ for all } z \in \mathbb{C} \text{ such that } \mathrm{Re}(z) < x,$$

where $\overset{\text{def}}{\iff}$ has the meaning of "if and only if" by definition.

The property A(x) can be also written as

$$\{z \in \mathbb{C} : \mathrm{Re}(z) < x\} \subseteq \mathscr{R}_x,$$

where $\mathscr{R}_x$ is the region defined in (4.18). Observe that $A(0)$ is the A-stability property.

The property A(x) is important because A(hr_1) implies the condition A for all non-rightmost spectra Λ^-.

It is of interest to consider the property A(x) for $|x|$ non-large. In fact, $|hr_1| \leq h\rho_1$ and we are assuming $h\rho_1$ non-large.

We have the following negative result.

Theorem 7.1 *There exists $x_0 > 0$ such that, for $x < \beta_R$ with $|x| \leq x_0$ and $x \neq 0$, A(x) is not true.*

Proof Remind that l is the order of the approximant R. In the complex plane, there exists a small disk centered at the origin which consists of $l+1$ sectors of width $\frac{\pi}{l+1}$ included in the order star $\mathscr{S}$, intercalated with $l+1$ sectors of width $\frac{\pi}{l+1}$ included in $\mathscr{S}^c$. Thus, there exists $x_0 > 0$ such that, for $x < \beta_R$ with $|x| \leq x_0$ and $x \neq 0$, the line $\mathrm{Re}(z) = x$ has a non-empty intersection with the order star $\mathscr{S}$. Let w be a point in this intersection. We have $\mathrm{Re}(w) = x$ and

$$\mathrm{e}^{-x}|R(w)| > 1.$$

Then, due to the continuity of R, there exists $\varepsilon > 0$ such that, for any $z \in \mathbb{C}$ with $x - \varepsilon \leq \mathrm{Re}(z) \leq x$ and $\mathrm{Im}(z) = \mathrm{Im}(w)$, we have

$$\mathrm{e}^{-x}|R(z)| > 1.$$

□

7.2 Non-significant eigenvalues I

The previous Theorem 7.1 says that, for any $x < \beta_R$ with $|x| \le x_0$ and $x \neq 0$, there exists $z \in \mathbb{C}$ with $\mathrm{Re}(z) < x$ such that $\mathrm{e}^{-x}|R(z)| \ge 1$. So, we can have, for some normal matrix A, a situation where the rightmost real part is $r_1 = \frac{x}{h}$ and $\lambda_i = \frac{z}{h}$ is a non-rightmost eigenvalue. For this eigenvalue we have

$$|w_i| = \mathrm{e}^{-hr_1}|R(h\lambda_i)| \ge 1,$$

and then the condition A does not hold.

Definition 7.2 We say that a non-rightmost eigenvalue λ_i with $|w_i| \ge 1$ is non-significant (significant) if

$$\frac{|\sigma_i|}{h\rho_1} \ll 1 \left(\frac{|\sigma_i|}{h\rho_1} \text{ is not small } \right).$$

It is expected that any non-rightmost eigenvalue λ_i with $|w_i| \ge 1$ is significant. In fact, by (4.3) we have

$$|\beta_2| \le \frac{|\sigma_i|}{h\rho_1}.$$

and it is expected $|\beta_2|$ non-small.

So, the negative result of Theorem 7.1 is not disastrous. The theorem says that, for any rightmost real part $r_1 \neq 0$ with $|hr_1| \le x_0$, there is a situation where we have a non-rightmost eigenvalue λ_i such that $|w_i| \ge 1$. But, such eigenvalue could be non-significant and, if this is true, then it is expected that such a situation does not happen.

In Sect. 7.10 below, we will introduce a condition on the approximant under which any non-rightmost eigenvalue λ_i with $|w_i| \ge 1$ is non-significant.

7.3 The properties A(x, a) and B(x, a)

It is expected that any non-rightmost eigenvalue λ_i with $|w_i| \ge 1$ has $|h\lambda_i|$ non-small. In fact, it is expected that λ_i is significant, i.e. it is expected that

$$\frac{|\sigma_i|}{h\rho_1} = \frac{|\lambda_i|}{\rho_1}|C||h\lambda_i|^l\big(1 + O(|h\lambda_i|)\big)$$

is not small, and then it is "unlikely" to have $|h\lambda_i|$ small.

Thus, we look at condition A for a non-rightmost spectrum Λ^- with all the eigenvalues λ_i such that $|h\lambda_i|$ is not small. In this context, the following two properties of the approximant R are important.

Definition 7.3 Let $x < \beta_R$ and let $a \ge 0$. Let

$$\mathrm{A}(x,a) \overset{\text{def}}{\iff} \mathrm{e}^{-x}|R(z)| < 1 \text{ for all } z \in \mathbb{C} \text{ such that } \mathrm{Re}(z) < x \text{ and } |z| \geq a$$

$$\mathrm{B}(x,a) \overset{\text{def}}{\iff} \mathrm{e}^{-x}|R(z)| > 1 \text{ for all } z \in \mathbb{C} \text{ such that } \mathrm{Re}(z) < x \text{ and } |z| \geq a.$$

The properties $\mathrm{A}(x,a)$ and $\mathrm{B}(x,a)$ can be also written as

$$\{z \in \mathbb{C} : \mathrm{Re}(z) < x \text{ and } |z| \geq a\} \subseteq \mathscr{R}_x \text{ and } \{z \in \mathbb{C} : \mathrm{Re}(z) < x \text{ and } |z| \geq a\} \subseteq \overset{\circ}{\mathscr{R}}{}^{c}_{x},$$

respectively, where $\overset{\circ}{\mathscr{R}}{}^{c}_{x}$ is the interior of the complementary set $\mathscr{R}^{c}_{x}$ of $\mathscr{R}_x$.

Observe that $\mathrm{A}(x)$ is $\mathrm{A}(x,0)$ and

$$\mathrm{A}(x,a_1) \Rightarrow \mathrm{A}(x,a_2) \text{ and } \mathrm{B}(x,a_1) \Rightarrow \mathrm{B}(x,a_2) \text{ if } a_1 < a_2.$$

The properties $\mathrm{A}(x,a)$ and $\mathrm{B}(x,a)$ are important because $\mathrm{A}(hr_1,a)$ implies the condition A for all non-rightmost spectra Λ^- such that $h\mu^- \geq a$ and $\mathrm{B}(hr_1,a)$ implies the condition B for all non-rightmost spectra Λ^- such that $h\rho^- \geq a$. Remind that μ^- and ρ^- are defined in (1.5).

7.4 The limit L

Now, we assume that

$$L := \lim_{z\to\infty} |R(z)|$$

exists. In addition, we also assume the following.

– When $L < +\infty$:

$$|\,|R(z)| - L\,| = O\left(\frac{1}{|z|^k}\right),\ |z| \to +\infty,$$

where $k > 0$, and, for any $x < \beta_R$ and $D \geq 0$,

$$|\,|R(z)| - L\,| \leq \frac{C}{|z|^k} \text{ for } \mathrm{Re}(z) < x \text{ and } |z| \geq D, \tag{7.1}$$

where $C = C(x,D) \geq 0$.

– When $L = +\infty$:

$$\frac{1}{|\,R(z)|} = O\left(\frac{1}{|z|^k}\right),\ |z| \to +\infty,$$

where $k > 0$, and, for any $x < \beta_R$ and $D \geq 0$,

$$|R(z)| \geq C|z|^k \text{ for } \mathrm{Re}(z) < x \text{ and } |z| \geq D, \tag{7.2}$$

where $C = C(x,D) > 0$.

The next two subsections consider, for $x < \beta_R$, the cases $L > \mathrm{e}^x$ and $L < \mathrm{e}^x$.

7.5 The case $L > \mathrm{e}^x$

Theorem 7.2 *Let $x < \beta_R$. Suppose $L > \mathrm{e}^x$. For any $\theta \in (1, \mathrm{e}^{-x}L)$ there exists $a \geq 0$ such that*

$$\mathrm{e}^{-x}|R(z)| \geq \theta \text{ for all } z \in \mathbb{C} \text{ such that } \mathrm{Re}(z) < x \text{ and } |z| \geq a. \tag{7.3}$$

(*Compare with the definition of* $\mathrm{B}(x, a)$ *given above*). *We have* (7.3) *for*

$$a = \begin{cases} \inf\limits_{D \geq 0} \max \left\{ \left(\frac{C}{L - \theta \mathrm{e}^x} \right)^{\frac{1}{k}}, D \right\} \text{ if } L < +\infty \\ \\ \inf\limits_{D \geq 0} \max \left\{ \left(\frac{\theta \mathrm{e}^x}{C} \right)^{\frac{1}{k}}, D \right\} \text{ if } L = +\infty. \end{cases} \tag{7.4}$$

Proof Let $\theta \in (1, \mathrm{e}^{-x}L)$. Since

$$\lim_{z \to \infty} \mathrm{e}^{-x}|R(z)| = \mathrm{e}^{-x}L > \theta,$$

we have

$$\mathrm{e}^{-x}|R(z)| \geq \theta$$

for $|z|$ sufficiently large.

About (7.4), fix $D \geq 0$. Under the assumptions (7.1) or (7.2), we have, for $\mathrm{Re}(z) < x$ and $|z| \geq D$,

$$\mathrm{e}^{-x}|R(z)| \geq \theta$$

whenever

$$\mathrm{e}^{-x}\left(L - \frac{C}{|z|^k} \right) \geq \theta \text{ or } \mathrm{e}^{-x}C|z|^k \geq \theta,$$

i.e.

$$|z| \geq \left(\frac{C}{L - \theta \mathrm{e}^x} \right)^{\frac{1}{k}} \text{ or } |z| \geq \left(\frac{\theta \mathrm{e}^x}{C} \right)^{\frac{1}{k_1}}.$$

Now (7.4) immediately follows. □

Remark 7.1 Consider $L < +\infty$. If $C = C(x, D)$ is a decreasing function of D, and this is obtained for example by considering the "optimal"

$$C = \sup_{\substack{\mathrm{Re}(z) < x \\ |z| \geq D}} | \, |R(z)| - L \, | \cdot |z|^k,$$

then $a = \overline{D}$ in (7.4), where $\overline{D} \geq 0$ is such that

$$\left(\frac{C(x,\overline{D})}{L-\theta \mathrm{e}^x}\right)^{\frac{1}{k}} = \overline{D}.$$

A similar observation applies to the case $L = +\infty$.

Theorem 7.2 has two important consequences given in Theorems 7.3 and 7.4.

Theorem 7.3 *Let $x < \beta_R$. If $L > \mathrm{e}^x$, then $\mathrm{B}(x,a)$ for*

$$a > \begin{cases} \inf\limits_{D\geq 0} \max\left\{\left(\frac{C}{L-\mathrm{e}^x}\right)^{\frac{1}{k}}, D\right\} \text{ if } L < +\infty \\ \inf\limits_{D\geq 0} \max\left\{\left(\frac{\mathrm{e}^x}{C}\right)^{\frac{1}{k}}, D\right\} \text{ if } L = +\infty. \end{cases}$$

Proof For any $\theta \in (1, \mathrm{e}^{-x}L)$, Theorem 7.2 says that we have $\mathrm{B}(x,a)$ for

$$a \geq \begin{cases} \inf\limits_{D\geq 0} \max\left\{\left(\frac{C}{L-\theta\mathrm{e}^x}\right)^{\frac{1}{k}}, D\right\} \text{ if } L < +\infty \\ \inf\limits_{D\geq 0} \max\left\{\left(\frac{\theta\mathrm{e}^x}{C}\right)^{\frac{1}{k}}, D\right\} \text{ if } L = +\infty. \end{cases}$$

So, we have $\mathrm{B}(x,a)$ for

$$\begin{aligned} a &> \inf_{\theta\in(1,\mathrm{e}^{-x}L)} \begin{cases} \inf\limits_{D\geq 0} \max\left\{\left(\frac{C}{L-\theta\mathrm{e}^x}\right)^{\frac{1}{k}}, D\right\} & \text{if } L < +\infty \\ \inf\limits_{D\geq 0} \max\left\{\left(\frac{\theta\mathrm{e}^x}{C}\right)^{\frac{1}{k}}, D\right\} & \text{if } L = +\infty. \end{cases} \\ &= \begin{cases} \inf\limits_{D\geq 0} \max\left\{\left(\frac{C}{L-\mathrm{e}^x}\right)^{\frac{1}{k}}, D\right\} & \text{if } L < +\infty \\ \inf\limits_{D\geq 0} \max\left\{\left(\frac{\mathrm{e}^x}{C}\right)^{\frac{1}{k}}, D\right\} & \text{if } L = +\infty. \end{cases} \end{aligned}$$

□

Theorem 7.4 *If $L > \mathrm{e}^{hr_1}$, then for any $\theta \in \left(1, \mathrm{e}^{-hr_1}L\right)$ and for any non-rightmost spectrum Λ^- satisfying $h\rho^- \geq a$, where a is given in (7.4) with $x = hr_1$, the condition B holds with*

$$\alpha \geq \frac{\log \theta}{h\rho_1}.$$

Moreover,

$$\alpha \leq -\frac{r_1}{\rho_1} + \frac{\log L_{\sup}}{h\rho_1}.$$

where $L_{\sup} = \sup_{\mathrm{Re}(z)<hr_1} |R(z)|$.

Proof Suppose $h\rho^- \geq a$. For a non-rightmost eigenvalue λ_i of maximum modulus we have $|h\lambda_i| \geq a$ and then

$$|w_i| = \mathrm{e}^{-hr_1}|R(h\lambda_i)| \geq \theta$$

by Theorem 7.2. Thus

$$\alpha \geq \mathrm{Re}(\alpha_i) = \frac{\log |w_i|}{h\rho_1} \geq \frac{\log \theta}{h\rho_1}.$$

Moreover, for any non-rightmost eigenvalue λ_i, we have

$$|w_i| = \mathrm{e}^{-hr_1}|R(h\lambda_i)| \leq \mathrm{e}^{-hr_1} L_{\sup}.$$

Thus,

$$\alpha = \max_{\lambda_i \in \Lambda^-} \mathrm{Re}(\alpha_i) = \max_{\lambda_i \in \Lambda^-} \frac{\log |w_i|}{h\rho_1} \leq \frac{\log(\mathrm{e}^{-hr_1} L_{\sup})}{h\rho_1} = -\frac{r_1}{\rho_1} + \frac{\log L_{\sup}}{h\rho_1}.$$

□

Observe that, by varying θ in $\left(1, \mathrm{e}^{-hr_1} L\right)$, the lower bound $\frac{\log \theta}{h\rho_1}$ of α can be arbitrarily close from below to the positive number

$$-\frac{r_1}{\rho_1} + \frac{\log L}{h\rho_1}.$$

If, in addition, $L = L_{\sup}$, then α is not larger than this positive number and α can be arbitrarily close to it.

7.6 The case $L < \mathrm{e}^x$

Theorem 7.5 *Let $x < \beta_R$. Suppose $L < \mathrm{e}^x$. For any $\theta \in (\mathrm{e}^{-x}L, 1)$ there exists $a \geq 0$ such that*

$$\mathrm{e}^{-x}|R(z)| \leq \theta \text{ for all } z \in \mathbb{C} \text{ such that } \mathrm{Re}(z) < x \text{ and } |z| \geq a. \tag{7.5}$$

(*Compare with the definition of* $\mathrm{A}(x,a)$ *given above*). *We have* (7.5) *for*

$$a = \inf_{D\geq 0} \max\left\{ \left(\frac{C}{\theta \mathrm{e}^x - L}\right)^{\frac{1}{k}}, D \right\}. \tag{7.6}$$

Proof Let $\theta \in (\mathrm{e}^{-x}L, 1)$. Since

$$\lim_{z\to\infty} \mathrm{e}^{-x}|R(z)| = \mathrm{e}^{-x}L < \theta,$$

we have

$$\mathrm{e}^{-x}|R(z)| \leq \theta.$$

for $|z|$ sufficiently large.

About (7.6), observe that, under the assumption (7.1), we have, for $\mathrm{Re}(z) < x$ and $|z| \geq D$,

$$\mathrm{e}^{-x}|R(z)| \leq \theta$$

whenever

$$\mathrm{e}^{-x}\left(L + \frac{C}{|z|^k}\right) \leq \theta,$$

i.e.

$$|z| \geq \left(\frac{C}{\theta \mathrm{e}^x - L}\right)^{\frac{1}{k}}.$$

Now (7.6) immediately follows. □

Remark 7.2 An observation about a in (7.6), similar to the observation of Remark 7.1 can be done.

Theorem 7.5 has two important consequences given in and Theorems 7.6 and 7.7.

Theorem 7.6 *Let* $x < \beta_R$. *If* $L < \mathrm{e}^x$, *then* $\mathrm{A}(x,a)$ *for*

$$a > \inf_{D\geq 0} \max\left\{ \left(\frac{C}{\mathrm{e}^x - L}\right)^{\frac{1}{k}}, D \right\}.$$

Proof For any θ in $(\mathrm{e}^{-x}L, 1)$, Theorem 7.5 says that we have $\mathrm{A}(x,a)$ for

$$a \geq \inf_{D\geq 0} \max\left\{ \left(\frac{C}{\theta \mathrm{e}^x - L}\right)^{\frac{1}{k}}, D \right\}.$$

So, we have $\mathrm{A}(x,a)$ for

$$a > \inf_{\theta\in(\mathrm{e}^{-x}L,1)} \inf_{D\geq 0} \max\left\{\left(\frac{C}{\theta \mathrm{e}^{x}-L}\right)^{\frac{1}{k}}, D\right\} = \inf_{D\geq 0} \max\left\{\left(\frac{C}{\mathrm{e}^{x}-L}\right)^{\frac{1}{k}}, D\right\}.$$

□

Theorem 7.7 *If $L < \mathrm{e}^{hr_1}$, then for any $\theta \in \left(\mathrm{e}^{-hr_1}L, 1\right)$ and for any non-rightmost spectrum Λ^- satisfying $h\mu^- \geq a$, where a is given in (7.6) with $x = hr_1$, the condition A holds with*

$$\alpha \leq \frac{\log\theta}{h\rho_1}.$$

Moreover,

$$\alpha \geq -\frac{r_1}{\rho_1} + \frac{\log L_{\inf}}{h\rho_1}.$$

where $L_{\inf} = \inf_{\mathrm{Re}(z)<hr_1} |R(z)|$.

Proof Suppose $h\mu^- \geq a$. For a non-rightmost eigenvalues λ_i such that $\alpha = \mathrm{Re}(\alpha_i)$ we have $|h\lambda_i| \geq a$ and then

$$\left|w_i\right| = \mathrm{e}^{-hr_1}|R(h\lambda_i)| \leq \theta$$

by Theorem 7.5. Thus,

$$\alpha = \mathrm{Re}(\alpha_i) = \frac{\log|w_i|}{h\rho_1} \leq \frac{\log\theta}{h\rho_1}.$$

Moreover, for any non-rightmost eigenvalue λ_i we have

$$\left|w_i\right| = \mathrm{e}^{-hr_1}|R(h\lambda_i)| \geq \mathrm{e}^{-hr_1}L_{\inf}.$$

Thus,

$$\alpha = \max_{\lambda_i\in\Lambda^-} \mathrm{Re}(\alpha_i) = \max_{\lambda_i\in\Lambda^-} \frac{\log|w_i|}{h\rho_1} \geq \frac{\log(\mathrm{e}^{-hr_1}L_{\inf})}{h\rho_1} = -\frac{r_1}{\rho_1} + \frac{\log L_{\inf}}{h\rho_1}.$$

Observe that, by varying θ in $\left(\mathrm{e}^{-hr_1}L, 1\right)$, the upper bound $\frac{\log\theta}{h\rho_1}$ of α can be arbitrarily close from above to the negative number

$$-\frac{r_1}{\rho_1} + \frac{\log L}{h\rho_1}.$$

If, in addition, $L = L_{\inf}$, then α is not smaller than this negative number and α can be arbitrarily close to it.

7.7 Approximants with $L = +\infty$

Consider approximants with $L = +\infty$. Examples of such approximants are Taylor approximants and superdiagonal Padé approximants.

The results in Sect. 7.5 say that the condition B holds for $h\rho^-$ sufficiently away from zero, as confirmed in the first example of Sect. 2. In particular, B holds for

$$h\rho^- > \inf_{D\geq 0} \max\left\{\left(\frac{\mathrm{e}^{hr_1}}{C}\right)^{\frac{1}{k}}, D\right\}.$$

As $h\rho^- \to +\infty$, B holds with $\alpha \to +\infty$.

7.8 Approximants with $L = 0$

Consider approximants with $L = 0$. Examples of such approximants are subdiagonal Padé approximants. Radau e Lobatto IIIC RK methods correspond to the first and second subdiagonal Padé approximants, respectively.

The results in Sect. 7.6 say that the condition A holds for $h\mu^-$ sufficiently away from zero. In particular, A holds for

$$h\mu^- > \inf_{D\geq 0} \max\left\{\left(\mathrm{e}^{-hr_1}C\right)^{\frac{1}{k}}, D\right\}.$$

As $h\mu^- \to +\infty$, A holds with $\alpha \to -\infty$. Moreover, Theorem 7.1 says that, for any rightmost real part $r_1 \neq 0$ with $|hr_1| \leq x_0$, we cannot have that A holds for all $h\mu^-$.

A-stable approximants with $L = 0$ are called *L-stable* (see [5]) and they are considered particularly suitable for integrating very stiff ODEs (see [1, 3, 8, 11]). Observe that here we are also considering approximants with $L = 0$ which are not A-stable. Indeed, the A-stability property does not play a crucial role in this context. Among subdiagonal Padé approximants, only the first and second subdiagonal Padé approximants (Radau and Lobatto IIIC methods) are A-stable.

7.9 Approximants with $L = 1$

Consider approximants with $L = 1$. Examples of approximants with $L = 1$ are diagonal Pad é approximants, which are also A-stable. Gauss methods correspond to the diagonal Padé approximants.

Suppose $r_1 < 0$. The results in Sect. 7.5 say that the condition B holds for $h\rho^-$ sufficiently away from zero, as confirmed in the second example of Sect. 2. In particular, B holds for

$$h\rho^- > \inf_{D\geq 0} \max\left\{\left(\frac{C}{1-\mathrm{e}^{hr_1}}\right)^{\frac{1}{k}}, D\right\}.$$

For an A-stable approximant, B holds with $\alpha \leq -\frac{r_1}{\rho_1}$ and, as $h\rho^- \to +\infty$, $\alpha \to -\frac{r_1}{\rho_1}$.

Suppose $r_1 > 0$. The results in Sect. 7.6 say that the condition A holds for $h\mu^-$ sufficiently away from zero. In particular, A holds for

$$h\mu^- > \inf_{D\geq 0} \max\left\{ \left(\frac{C}{\mathrm{e}^{hr_1}-1}\right)^{\frac{1}{k}}, D \right\}.$$

For an A-stable approximant, A holds with $\alpha \geq -\frac{r_1}{\rho_1}$ and, as $h\mu^- \to +\infty$, $\alpha \to -\frac{r_1}{\rho_1}$.

7.10 Non-significant eigenvalues II

In this subsection we study when any non-rightmost eigenvalue λ_i with $|w_i| \geq 1$ is non-significant (see Sect. 7.2).

7.10.1 The region $\mathscr{P}_x$

For $x < \beta_R$, let

$$\mathscr{P}_x := \{z \in \mathbb{C} : \mathrm{Re}(z) < x\} \cap \mathscr{R}_x^c,$$

where $\mathscr{R}_x^c$ is the complementary set of $\mathscr{R}_x$.

We have A(x) if and only if $\mathscr{P}_x = \emptyset$. Moreover, for $a \geq 0$, we have A(x, a) if and only if the open disk of radius a centered at the origin includes $\mathscr{P}_x$.

The importance of the region $\mathscr{P}_x$ is due to the fact that, for a non-rightmost eigenvalue λ_i, we have $|w_i| \geq 1$ if and only if $h\lambda_i \in \mathscr{P}_{hr_1}$.

7.10.2 The number $a(x)$

For $x < \beta_R$, let

$$a(x) := \inf\{a \geq 0 : \mathrm{A}(x, a)\}.$$

In other words, $a(x)$ is the infimimum of the radii of open disks centered at the origin and including $\mathscr{P}_x$.

The importance of the number $a(x)$ is given by the following theorem.

Theorem 7.8 *For a non-rightmost eigenvalue λ_i such that $|w_i| \geq 1$, we have $|h\lambda_i| \leq a(hr_1)$.*

Proof The closed disk of radius $a(x)$ centered at the origin includes the region $\mathscr{P}_x$. The theorem follows by reminding that $\mathscr{P}_x$ contains the non-rightmost eigenvalues λ_i such that $|w_i| \geq 1$. □

7.10.3 The theorem on the non-significant eigenvalues

Next theorem says when any non-rightmost eigenvalue λ_i with $|w_i| \geq 1$ is non-significant. It involves the behavior of $a(x)$ as $x \to 0$.

Theorem 7.9 *Consider an approximant such that*

$$a(x) \le \eta |x|(1 + O(x)), \ x \to 0, \tag{7.7}$$

where $\eta > 0$. *If*

$$\eta^{l+1} E_1 \big(1 + O\big(h\rho_1\big)\big) \ll 1, \tag{7.8}$$

where l *is the order of the approximant, then any non-rightmost eigenvalue* λ_i *with* $|w_i| \ge 1$ *is non-significant.*

Proof Consider a non-rightmost eigenvalue λ_i with $|w_i| \ge 1$. By Theorem 7.8, we have

$$\big|h\lambda_i\big| \le a\big(hr_1\big) \le \eta |hr_1| \big(1 + O\big(hr_1\big)\big) \le \eta h \rho_1 \big(1 + O\big(h\rho_1\big)\big).$$

Recall (1.6) and (1.7). Since

$$|\log S(z)| = |C| z^{l+1} (1 + O(z)),$$

we obtain

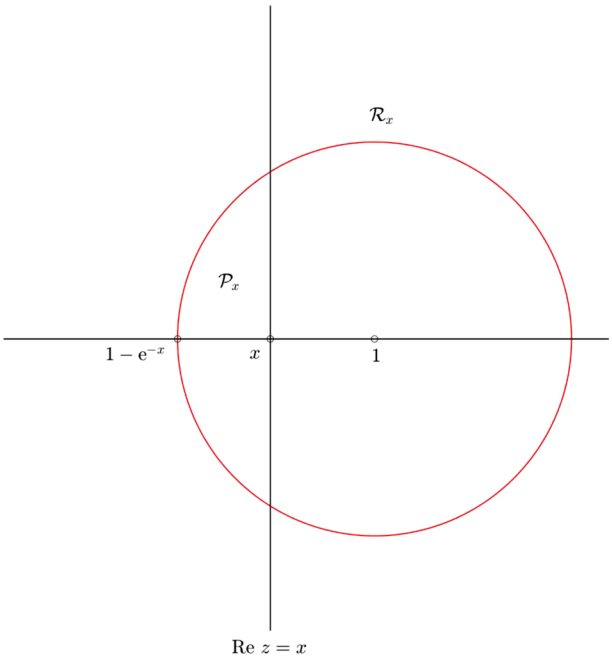

Fig. 11 Regions $\mathscr{R}_x$ and $\mathscr{P}_x$ for the implicit Euler method. $\mathscr{R}_x$ is the exterior of the red circle of center 1 and radius e^{-x}. $\mathscr{P}_x$ is the part of the closed disk at the left of the line $\mathrm{Re}(z) = x$

$$\begin{aligned}|\sigma_i| = |\log S(h\lambda_i)| &= |C|\big(|h\lambda_i|\big)^{l+1}\big(1+O\big(|h\lambda_i|\big)\big)\\ &\le |C|\big(\eta h\rho_1\big(1+O\big(h\rho_1\big)\big)\big)^{l+1}\big(1+O\big(h\rho_1\big)\big)\\ &= \eta^{l+1}|C|\big(h\rho_1\big)^{l+1}\big(1+O\big(h\rho_1\big)\big)\end{aligned}$$

and then

$$\begin{aligned}\frac{|\sigma_i|}{h\rho_1} &\le \eta^{l+1}|C|\big(h\rho_1\big)^{l}\big(1+O\big(h\rho_1\big)\big)\\ &= \eta^{l+1}\frac{E_1}{1+O(h\rho_1)}\big(1+O\big(h\rho_1\big)\big) \text{ recall (1.9)}\\ &= \eta^{l+1}E_1\big(1+O\big(h\rho_1\big)\big).\end{aligned}$$

The theorem now follows by reminding the definition of non-significant eigenvalue. □

Remark 7.3 The term $O(h\rho_1)$ in (7.8) is not larger than $Ch\rho_1$ for $h\rho_1 \le D$, where $C \ge 0$ and $D > 0$ depend only on the approximant.

By the previous theorem we obtain the following important conclusion.

Conclusion 7.10 Suppose that the approximant satisfies (7.7). It is expected that A holds.

In fact, suppose A does not hold, i.e. there is a non-rightmost eigenvalue λ_i with $|w_i| \ge 1$. It is expected that this eigenvalue is significant. On the other hand, if it is significant, then, by the previous theorem, we obtain that (7.8) does not hold and this is "unlikely".

In the next subsection, we show that the implicit Euler method satisfies (7.7).

7.11 The implicit Euler method

We examine the property A(x) and determine the number $a(x)$ for the the implicit Euler method, corresponding to the (0, 1)-Padé approximant

$$R(z) = \frac{1}{1-z},\ z \in \mathbb{C}\backslash\{1\}.$$

This approximant has $\beta_R = 1$.

The region $\mathscr{R}_x$, $x < 1$, for this approximant is the exterior of the disk of center 1 and radius e^{-x} and the region $\mathscr{P}_x$ is the part of the closed disk at the left of the line $\mathrm{Re}(z) = x$ (see Fig. 11).

Theorem 7.11 *Let $x < 1$. For the implicit Euler method, we have* A(x) *if and only $x = 0$. Moreover, we have*

$$a(x) = \sqrt{\mathrm{e}^{-2x} - 1 + 2x}. \tag{7.9}$$

Proof When $x = 0$, $A(x)$ is the A-stability. When $x \neq 0$, we have $1 - \mathrm{e}^{-x} < x$ and then

$$\mathscr{P}_x = \{z \in \mathbb{C} : \mathrm{Re}(z) < x\} \cap \mathscr{R}_x^c \neq \emptyset,$$

since the complementary set $\mathscr{R}_x^c$ of $\mathscr{R}_x$ is the closed disk of center 1 and radius e^{-x} (see Fig. 10). Thus $A(x)$ is not true.

For the second part, let $b \geq 0$. An easy computation shows that, for $z \in \mathbb{C}$ such that $|z| = b$, we have

$$z \in \mathscr{R}_x^c \;\Leftrightarrow\; |z - 1| \leq \mathrm{e}^{-x} \;\Leftrightarrow\; \mathrm{Re}(z) \geq \frac{1}{2}\left(b^2 + 1 - \mathrm{e}^{-x}\right).$$

Hence

$$\begin{aligned} \emptyset \neq \left\{z \in \mathbb{C} : z \in \mathscr{P}_x \text{ and } |z| = b\right\} &= \left\{z \in \mathbb{C} : \mathrm{Re}(z) < x \text{ and } |z| = b \text{ and } z \in \mathscr{R}_x^c\right\} \\ &= \left\{z \in \mathbb{C} : |z| = b \text{ and } \frac{1}{2}\left(b^2 + 1 - \mathrm{e}^{-x}\right) \leq \mathrm{Re}(z) < x\right\} \end{aligned}$$

if and only if

$$\frac{1}{2}\left(b^2 + 1 - \mathrm{e}^{-x}\right) < x \;\text{ and }\; x > -b \;\text{ and }\; \frac{1}{2}\left(b^2 + 1 - \mathrm{e}^{-x}\right) \leq b. \tag{7.10}$$

For $x > 0$, (7.10) is equivalent to

$$1 - \mathrm{e}^{-x} \leq b < \sqrt{\mathrm{e}^{-2x} - 1 + 2x}.$$

For $x < 0$, (7.10) is equivalent to

$$-x < b < \sqrt{\mathrm{e}^{-2x} - 1 + 2x}.$$

Now, equation (7.9) follows. □

By (7.9), we obtain

$$a(x) = \sqrt{2}|x|(1 + O(x)),\ x \to 0.$$

We can conclude that it is expected that A holds for the implicit Euler method.

8 Conclusions

In the stiff situation, we have studied the long-time behavior of the relative error in the numerical integration of the ODE (1.1) with A normal. The numerical integration is accomplished over a mesh of constant stepsize h, by using at any step of an analytic approximant R of the exponential: see (1.2). The relative error γ_n of the numerical integration is given in (1.3).

We have defined the long-time solution y^{long} as the solution of (1.1) projected on the eigenspace of the rightmost eigenvalues and we have considered the relative error γ_n^{long} of the numerical integration of y^{long}. The error γ_n^{long} grows linearly in time, it is small and it remains small in the long-time.

We have introduced the condition

$$\text{A:} \quad |R(h\lambda)| < \mathrm{e}^{hr_1} \text{ for any non-rightmost eigenvalue } \lambda \text{ of } A,$$

where r_1 is the real part of the rightmost eigenvalues of A. When A holds, we have $\gamma_n \approx \gamma_n^{\text{long}}$ in the long-time. When A does not hold, we have $\frac{\gamma_n}{\gamma_n^{\text{long}}} \gg 1$ for all time.

Let $L = \lim_{z\to\infty} |R(z)|$. In order to have the condition A satisfied, it is better to use approximant with $L = 0$ (for example Radau and Lobatto IIIC methods). Approximants with $L = 1$ (for example Gauss methods) does not work well when $r_1 < 0$.

The paper [10] analyzes the numerical integration in the stiff situation by looking to a different question. In [10], the interest is about numerical approximations (1.2) of the long-time solution starting with a perturbed initial value. The approximants are analyzed by means of their error growth function φ_R (see [4, 5]) in order to study how they propagate the initial perturbation from the relative error point of view. In this other context, we have a non-large propagation of the initial perturbation if and only if

$$\varphi_R(x) = 1 + x + o(x), \; x \to 0.$$

We have considered the case of A normal. Some numerical experiments, not included here, suggest that also for non-normal matrices we have $\gamma_n \approx \gamma_n^{\text{long}}$ in the long-time when the condition A holds and $\frac{\gamma_n}{\gamma_n^{\text{long}}} \gg 1$ for all time when A does not hold. In light of this, the results of Sect. 7 becomes more important, since they are about the condition A.

We conclude by remarking that the findings of this paper are interesting in applications involving differential models described by linear ODEs with $r_1 \neq 0$. In particular, they are interesting when we are integrating an ODE whose solution decreases to small orders of magnitude (case $r_1 < 0$), but it is not yet considered as zero, or grows up to a large orders of magnitude (case $r_1 > 0$), but it is not yet considered as infinite.

Funding Open access funding provided by Università degli Studi di Trieste within the CRUI-CARE Agreement.

References

1. Bui, T., Bui, R.: Numerical methods for extremely stiff systems of ordinary differential equations. Appl. Math. Model. **3**, 355–358 (1979)
2. Burgisser, F., Cucker, F.: Condition. The Geometry of Numerical Algorithms. Springer, Berlin (2013)
3. Gad, E., Nakhla, M., Achar, R., Zhou, Y.: A-stable and L-stable high-order integration methods for solving stiff differential equations. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **28**, 1352–1372 (2009)
4. Hairer, E., Bader, G., Lubich, C.: On the stability of semi-implicit methods for ordinary differential equations. BIT **22**, 211–232 (1982)
5. Hairer, E., Wanner, G.: Solving Ordinary Differential Equations II. Stiff Differential-Algebraic Problems, Second Revised Edition. Springer, Berlin (1996)
6. Hairer, E., Wanner, G.: Order stars and stiff integrators. J. Comput. Appl. Math. **125**, 93–105 (2000)
7. Iserles, A., Norsett, S.: Order Stars: Theory and Applications. Chapman and Hall, London (1991)
8. Logsdon, J., Biegler, L.: Accurate solution of differential-algebraic optimization problems. Ind. Eng. Chem. Res. **28**, 1628–1639 (1989)
9. Maset, S.: Relative error analysis of matrix exponential approximations for numerical integration. J. Numer. Math. **29**(2), 119–158 (2021)
10. Maset, S.: Relative error stability and instability of matrix exponential approximations for stiff numerical integration of long-time solutions. J. Comput. Appl. Math. **390**, 113387 (2021)
11. Ropp, D., Shadid, J.: Stability of operator splitting methods for systems with indefinite operators: reaction-diffusion systems. J. Comput. Phys. **203**, 449–466 (2005)
12. Shampine, L., Gladwell, I., Thompson, S.: Solving ODEs with MATLAB. Cambridge University Press, Cambridge (2003)
13. Wanner, G., Hairer, E., Norsett, S.: Order stars and stability theorems. BIT **18**, 475–489 (1978)