USING THE RD RATIONAL ARNOLDI METHOD FOR EXPONENTIAL INTEGRATORS

PAOLO NOVATI*

Abstract. In this paper we investigate some practical aspects concerning the use of the Restricted-Denominator (RD) rational Arnoldi method for the computation of the core functions of exponential integrators for parabolic problems. We derive some useful a-posteriori bounds together with some hints for a suitable implementation inside the integrators. Numerical experiments arising from the discretization of sectorial operators are presented.

Key words. Rational Arnoldi method, matrix functions, exponential integrators

AMS subject classifications. 65F60, 65L05

1. Introduction.

For the solution of large stiff problems of the type

$$u'(t) = f(y(t)) = Lu(t) + N(u(t)), \tag{1.1}$$

where $L \in \mathbb{R}^{M\times M}$ arises from the discretization of unbounded sectorial operators and N is a nonlinear function, in recent years much work has been done on the construction of exponential integrators that might represent a promising alternative to classical solvers (see e.g. [20] or [15] for a comprehensive survey). As well known the computation of the matrix exponential or related functions of matrices is at the core of this kind of integrators. The main idea is to damp the stiffness of the problem (assumed to be contained in L) on these computations so that the integrator can be explicit.

Under the hypothesis that the functions of matrices involved are exactly evaluated, the linear stability can be trivially achieved for both Runge-Kutta and multistep based exponential integrators and hence highly accurate and stable integrators can be constructed. On the other hand, the main problem with this class of integrators is just the efficient computation of such functions of matrices, so that, very few reliable codes have been written (we remember the Rosenbrock type exponential integrators presented in [4], [14], [26]). For this reason many authors are still doubtful about the potential of exponential integrators with respect to classical implicit solvers even for semilinear problem of type (1.1).

An exponential integrator requires at each time step the evaluation of a certain number (depending on the accuracy) of functions of matrices of the type $\varphi_k(hL)v$, where

$$\begin{aligned} \varphi_0(h\lambda) &= \exp(h\lambda), \\ \varphi_{k+1}(h\lambda) &= \frac{\varphi_k(h\lambda) - \frac{1}{k!}}{h\lambda}, \text{ for } k = 0, 1, 2, \dots , \end{aligned} \tag{1.2}$$

being h the time step. Actually this represents the general situation for the Exponential Time Differencing methods, that is, the methods based on the variation-of-constants formula; for Lawson's type method (also called Integrating Factor methods) only the matrix exponential is involved. We refer again to [20] and the reference therein for a background.

*Department of Mathematics, University of Padova, Italy (novati@math.unipd.it).

Among the existing techniques for the computation of functions of matrices (we quote here the recent book of Higham [12] for a survey), in this context the Restricted-Denominator (RD) Rational Arnoldi algorithm introduced independently in [33] and [22] for the computation of the matrix exponential seems to be an reliable approach. It is based on the use of the so called RD rational forms, studied in [25] for the exponential function,

$$R_{i,j}(\lambda) = \frac{q_i(\lambda)}{(1-\delta\lambda)^j}, \quad \delta \in \mathbb{R},$$

where q_i is a polynomial of degree $\leq i$. We refer again to [22] for the basic references about the properties and the use of such rational forms. While in the matrix case, the use of these approximants requires the solution of linear systems with the matrix $(I - \delta L)$, as shown in [26] in the context of the solution of (1.1) when L is sectorial so typically sparse and well structured this linear algebra drawback can be almost completely overtaken organizing suitably the step-size control strategy and exploiting the properties of the RD Arnoldi method concerning the choice of the parameter δ. In other words the number of linear systems to be solved can be drastically reduced with respect to the total number of computations of functions of matrices required by the integrator. Therefore the mesh independence property of the method, that leads to a very fast convergence with respect to a standard polynomial approach (see again [22]), can be fully exploited for the construction of competitive integrators.

A problem still open is that inside the integrator the rational Arnoldi algorithm (responsible for most of the computational cost) have to be supported by a robust and sharp error estimator. In the self-adjoint case the problem has been treated in [23] where the author presents effective a-posteriori error estimates, even in absence of information on the location of the spectrum of L. Anyway, in the general case, when (1.1) arises for instance from the discretization of parabolic problems with advection terms and/or non-zero boundary conditions the numerical range of L, that we denote by $F(L)$, may not reduce to a line segment. In this sense the basic aim of this paper is to fill this gap providing error estimates for the non-symmetric case using as few as possible information about the location of $F(L)$. It is necessary to keep in mind that a competitive code for (1.1) should also be able to update L (interpreted as the Jacobian of f, [31], [4]) so that $F(L)$ is may be not fixed during the integration, and so it is important to reduce as much as possible any pre-processing technique to estimate $F(L)$. In particular assuming that $F(L) \subseteq \mathbb{C}^-$we shall provide a-posteriori error estimates for the RD Arnoldi process using only information about the angle of the sector containing $F(L)$, angle that is typically independent of the sharpness of the discretization and hence computable working in small dimension.

The paper is organized as follows. In Section 2 we present the basic idea of the RD rational Arnoldi method and in Section 3 we derive some first general error bounds based on the standard approaches. In Section 4, exploiting the relation between the derivatives of the function $e^{1/z}$ and the Laguerre polynomials extended to the complex plane, we derive some a-posteriori error bounds. The problem of defining reliable a-priori bounds is investigated in Section 5. Section 6 is devoted to the analysis of the generalized residual as error estimator, that can be used to obtain information about the choice of the parameter δ for the rational approximation. In Section 7 we present some numerical examples arising from the discretization of a one-dimensional advection-diffusion model. In Section 8 we provide some hints about the use of the RD rational Arnoldi method inside an exponential integrator with the aim of reducing

as much as possible the number of implicit computations of $(I-\delta L)^{-1}$. Finally, in Section 9 we furnish a deeper analysis concerning the fast rate of convergence of the method, that will provide further information about the optimal choice of the parameter δ.

2. The RD rational Arnoldi method. In what follows we denote by $\|\cdot\|$ the Euclidean vector norm and its induced matrix norm. As already mentioned, the notation $F(L)$ indicates the *numerical range* of L, that is,

$$F(L) := \left\{ \frac{x^H L x}{x^H x}, x \in \mathbb{C}^M \setminus \{0\} \right\},$$

while the spectrum of L is denoted by $\sigma(L)$. The notation Π_m indicates the space of the algebraic polynomials of degree $\leq m$.

Given $0 \leq \theta < \frac{\pi}{2}$, let

$$S_\theta = \{\lambda : |\arg(-\lambda)| \leq \theta\} \subset \mathbb{C}^- \tag{2.1}$$

be the unbounded sector of the left half complex plane, symmetric with respect to the real axis with vertex in 0 and semiangle θ. Let moreover Γ_θ be the boundary of S_θ. Throughout the paper we assume that $F(L) \subset int(S_\theta)$, the interior of S_θ. Accordingly, L is a so-called sectorial operator (see e.g. [16] Chap. V, for a background).

Given a vector $v \in \mathbb{R}^M$, with $\|v\| = 1$, consider the problem of computing

$$y^{(k)} = \varphi_k(hL)v, \tag{2.2}$$

where φ_k is defined by (1.2). The RD rational approach seeks for approximations to $\varphi_k(h\lambda)$ of the type

$$R_{m-1,m-1}(\lambda) = \frac{p_{k,m-1}(\lambda)}{(1-\delta\lambda)^{m-1}}, \quad p_{k,m-1}(\lambda) \in \Pi_{m-1}, \quad m \geq 1,$$

where $\delta > 0$ is a suitable parameter. Turning to the matrix case, $y^{(k)}$ is approximated by elements of the Krylov subspaces

$$K_m(Z, v) = span\left\{v, Zv, Z^2v, ..., Z^{m-1}v\right\}, \quad m \geq 1,$$

with respect to v and the matrix Z defined by the transform

$$Z = (I - \delta L)^{-1}.$$

In this sense the idea is to use a polynomial method to compute $y^{(k)} = f_k(Z)v$, where

$$f_k(z) := \varphi_k\left(\frac{h}{\delta}\left(1 - \frac{1}{z}\right)\right)$$

is singular at 0.

For the construction of the subspaces $K_m(Z, v)$ we employ the classical Arnoldi method. As is well known it generates an orthonormal sequence $\{v_j\}_{j\geq 1}$, with $v_1 = v$, such that $K_m(Z, v) = span\{v_1, v_2, ..., v_m\}$. Moreover, for every m,

$$ZV_m = V_m H_m + h_{m+1,m} v_{m+1} e_m^H, \tag{2.3}$$

where $V_m = [v_1, v_2, ..., v_m]$, H_m is upper Hessenberg matrix with entries $h_{i,j} = v_i^H Z v_j$ and e_j is the j-th vector of the canonical basis of $\mathbb{R}^m$.

The m-th RD-rational Arnoldi approximation to $y^{(k)}$ is defined as (see [17])

$$y_m^{(k)} = V_m f_k(H_m) e_1. \tag{2.4}$$

It can be seen that

$$y_m^{(k)} = \overline{p}_{k,m-1}(Z) v, \tag{2.5}$$

where $\overline{p}_{k,m-1} \in \Pi_{m-1}$ interpolates, in the Hermite sense, the function $f_k(z)$ in the eigenvalues of H_m (see [28]).

As mentioned in the Introduction this technique has been introduced independently in [33] and [22]. Anyway, the idea of using rational Krylov approximations to matrix functions was originally introduced in [8]. More recently this approach has been extended to the case of multiple poles and is commonly referred to as RKS (Rational Krylov Subspace) approximation (see [18], [27], [3]).

3. General error bounds. Before stating a general error bound for the method, we need to locate $F(Z)$. Consider the function $\chi(\lambda) = (1 - \delta\lambda)^{-1}$. Denoting by $D_{1/2,1/2}$ the disk centered in $1/2$ with radius $1/2$, let

$$G_\theta = \{z : z = \chi(\lambda), \lambda \in S_\theta\} \subseteq D_{1/2,1/2}. \tag{3.1}$$

Its boundary, Σ_θ, is made by two circular arcs meeting with angle 2θ at 0 and 1. Regarding the field of values of Z, $F(Z)$, we can state the following result that will be used frequently throughout the paper.

PROPOSITION 3.1. *If $F(L) \subset int(S_\theta)$ then $F(Z) \subset int(G_\theta)$.*

Proof. Obviously $\sigma(Z) = \chi(\sigma(L))$, so $F(Z)$ cannot lie entirely outside G_θ. Now assume that there exists $\lambda \in \Gamma_\theta$ such that $\chi(\lambda) \in F(Z)$, that is, $F(Z) \cap \Sigma_\theta \neq \varnothing$. Hence, there exists $y \in \mathbb{C}^M$, $\|y\| = 1$, such that

$$y^H (I - \delta L)^{-1} y = \frac{1}{1 - \delta\lambda}. \tag{3.2}$$

Defining $x := (I - \delta L)^{-1} y$ we easily obtain

$$x^H \left(I - \delta L^T\right) x = \frac{1}{1 - \delta\lambda},$$

and hence

$$1 - \delta \frac{x^H L^T x}{x^H x} = \frac{1}{(1 - \delta\lambda) \|x\|^2}.$$

By (3.2) we have

$$\|x\| \, |1 - \delta\lambda| \geq 1. \tag{3.3}$$

Now let us define $\mu := \frac{x^H L^T x}{x^H x} \in F(L)$. We have

$$\|x\|^2 = (1 - \delta\lambda)^{-1} (1 - \delta\mu)^{-1}, \tag{3.4}$$

and hence

$$\mathrm{Im}\left((1-\delta\lambda)^{-1}(1-\delta\mu)^{-1}\right)=0,$$

that implies

$$\frac{\mathrm{Im}\left((1-\delta\lambda)^{-1}\right)}{\mathrm{Re}\left((1-\delta\lambda)^{-1}\right)}=-\frac{\mathrm{Im}\left((1-\delta\mu)^{-1}\right)}{\mathrm{Re}\left((1-\delta\mu)^{-1}\right)}. \tag{3.5}$$

Now since $(1-\delta\mu)^{-1}\in int(G_\theta)$ and $(1-\delta\lambda)^{-1}\in\Sigma_\theta$, by (3.5) it must be $\mathrm{Re}\left((1-\delta\lambda)^{-1}\right)>\mathrm{Re}\left((1-\delta\mu)^{-1}\right)$ and $\left|\mathrm{Im}\left((1-\delta\lambda)^{-1}\right)\right|>\left|\mathrm{Im}\left((1-\delta\mu)^{-1}\right)\right|$ so that

$$|1-\delta\mu|^{-1}<|1-\delta\lambda|^{-1}.$$

Using this relation, by (3.4) we finally have

$$\|x\|^2|1-\delta\lambda|^2<1,$$

that contradicts (3.3). Since the field of values is connected the proof is complete. □

REMARK 3.2. *In order to provide information about the geometry of $F(Z)$, it is worth referring to [35] Theorem 5.2 in which the author proves that if L is an invertible matrix then*

$$\lim_{s\to\infty}\left(\frac{1}{F((L-sI)^{-1})}+s\right)=F(L).$$

Taking $\delta=1/s$, we have that for small values of δ

$$F((I-\delta L)^{-1})\approx\frac{1}{1-\delta F(L)}.$$

Going back to our method, the corresponding error $E_{k,m}:=y^{(k)}-y_m^{(k)}$ can be expressed and bounded in many ways (we quote here the recent papers [3] and [7] for a background on the error estimates for both polynomial and rational Arnoldi approximation to matrix functions). The following proposition states a general result.

PROPOSITION 3.3. *Let $G\subseteq D_{1/2,1/2}$ be a compact such that $F(Z)\subset int(G)$ and whose boundary Σ is a rectifiable Jordan curve. For every $p_{m-1}\in\Pi_{m-1}$*

$$\|E_{k,m}\|\le\frac{1}{2\pi}\int_\Sigma\frac{|f_k(z)-p_{m-1}(z)|}{dist(z,F(Z))}\left\|v-(zI-Z)V_m(zI-H_m)^{-1}e_1\right\||dz|. \tag{3.6}$$

Proof. Using the properties of the Arnoldi algorithm we know that for every $p_{m-1}\in\Pi_{m-1}$,

$$V_m p_{m-1}(H_m)e_1=p_{m-1}(Z)v.$$

Hence from this identity it follows that, for $m\ge1$

$$E_{k,m}=f_k(Z)v-p_{m-1}(Z)v-V_m(f_k(H_m)-p_{m-1}(H_m))e_1. \tag{3.7}$$

Now since $F(H_m) \subseteq F(Z)$ we can write (3.7) in the Dunford-Taylor integral form

$$E_{k,m} = \frac{1}{2\pi i} \int_{\Sigma} \left(f_k(z) - p_{m-1}(z)\right) \left[(zI - Z)^{-1} v - V_m \left(zI - H_m\right)^{-1} e_1 \right] dz.$$

Collecting $(zI - Z)^{-1}$ and using (see [30])

$$\left\| (zI - Z)^{-1} \right\| \leq \frac{1}{dist(z, F(Z))},$$

we prove (3.6). □

Now since

$$v - (zI - Z) V_m \left(zI - H_m\right)^{-1} e_1 = \frac{q_m(Z)v}{q_m(z)},$$

where

$$q_m(z) = \det(zI - H_m),$$

(see [21]), any bound for $\|q_m(Z)v\| / |q_m(z)|$ and any choice for G and p_{m-1} leads to a bound for $\|E_{k,m}\|$. This technique has been used for instance in [22] and [13]. In particular in [22] the authors use the relation

$$\|q_m(Z)v\| = \prod_{j=1}^{m} h_{j+1,j}, \tag{3.8}$$

and the inequality

$$|q_m(z)| \geq dist(z, F(Z))^m. \tag{3.9}$$

Going back to our situation, the main problem is that if we simply assume that $F(L) \subset S_\theta$ (in other words $F(L)$ arbitrarily large) we have that $dist(z, F(Z)) \to 0$ as $z \to 0$ ($\operatorname{Re}\lambda \to -\infty$) because we have to consider the singularity of f_k at 0. Therefore using a lower bound like (3.9) (but the situation remains true even for other approaches (cf. [13])) terms of the type $f_k(z)/z^{m+1}$ would appear in (3.6). In the exponential case ($k = 0$) this is not a problem because $f_0(z)/z^{m+1} \to 0$ for $z \to 0$, but for $k > 0$ the situation changes completely since

$$\frac{f_k(z)}{z^{m+1}} \approx \frac{\delta}{h(k-1)!} \frac{1}{z^m}$$

for $z \to 0$.

Because of the difficulties just explained, our approach for deriving error bounds is not based on the use of the Cauchy integral formula. Exploiting the interpolatory nature of the standard Arnoldi method, we notice, as pointed out also in [9], that the error can be expressed in the form

$$E_{k,m} = g_{k,m}(Z) q_m(Z) v, \tag{3.10}$$

where (cf. (2.5))

$$g_{k,m}(z) := \frac{f_k(z) - \overline{p}_{k,m-1}(z)}{\det(zI - H_m)}. \tag{3.11}$$

In [9] this relationship is used as the basis for the construction of restarted methods for the computation of matrix functions.

We can state the following basic result that will be used throughout the paper and that allows to overcome the difficulties of working with formula (3.6).

PROPOSITION 3.4. *Let $F(L) \subset S_\theta$ and let $\tau := h/\delta$. Then*

$$\|E_{k,m}\| \leq K \frac{1}{\tau^k (m+k)!} \max_{z \in G_\theta} \left| \frac{d^{m+k}}{dz^{m+k}} f_0(z) z^k \right| \prod_{i=1}^{m} h_{i+1,i}, \tag{3.12}$$

where $2 \leq K \leq 11.08$. In the symmetric case we can take $K = 1$.

Proof. By [5] we know that

$$\|g_{k,m}(Z)\| \leq K \max_{z \in F(Z)} |g_{k,m}(z)|,$$

and hence by (3.8) and (3.10)

$$\|E_{k,m}\| \leq K \max_{z \in F(Z)} |g_{k,m}(z)| \prod_{i=1}^{m} h_{i+1,i}.$$

Now, by induction one proves that for $k \geq 1$

$$f_k(z) = \frac{f_0(z) z^k - s_{k-1}(z) z}{\tau^k (z-1)^k}, \tag{3.13}$$

where $s_0(z) = 1$ and

$$s_k(z) = s_{k-1}(z) z + \frac{\tau^k (z-1)^k}{k!} \in \Pi_k \text{ for } k \geq 1.$$

Putting (3.13) in (3.11) we obtain

$$g_{k,m}(z) = \frac{f_0(z) z^k - s_{k-1}(z) z - \tau^k (z-1)^k \overline{p}_{k,m-1}(z)}{\tau^k (z-1)^k \det(zI - H_m)}.$$

Now, the polynomial $\tau^k (z-1)^k \overline{p}_{k,m-1}(z) \in \prod_{m+k-1}$ interpolates in the Hermite sense the function $f_0(z) z^k - s_{k-1}(z) z$ in the eigenvalues of H_m and in $z = 1$. Henceforth $g_{k,m}(z)$ is a divided difference that can be bounded using the Hermite-Genocchi formula (see e.g. [6]), so that

$$|g_{k,m}(z)| \leq \frac{1}{\tau^k (m+k)!} \max_{\xi \in co(\{z, \sigma(H_m), 1\})} \left| \frac{d^{m+k}}{d\xi^{m+k}} f_0(\xi) \xi^k \right|,$$

where $co(\{z, \sigma(H_m), 1\}$ denotes the convex hull of the point set given by z, $\sigma(H_m)$ and 1. Since $\sigma(H_m) \subset F(Z)$, and $F(Z) \subset G_\theta$ by Proposition 3.1, the result follows. ∎

4. A posteriori error estimates. By (3.12), in order to provide a-posteriori error estimates we just need to study the derivatives of the function $f_0(z) z^k$. We need to introduce the generalized Laguerre polynomials, defined by

$$L_n^{(\alpha)}(z) = \sum_{j=0}^{n} (-1)^j \binom{n+\alpha}{n-j} \frac{z^j}{j!}.$$

We can state the following result.

LEMMA 4.1. *Let* $\tau = \frac{h}{\delta}$. *For* $m \geq 1$

$$\frac{1}{\tau^k (m+k)!} \frac{d^{m+k}}{dz^{m+k}} f_0(z) z^k = \frac{(-1)^{m+1}\tau}{z^{m+k+1}} f_0(z) \frac{(m-1)!}{(m+k)!} L_{m-1}^{(k+1)}(\frac{\tau}{z}). \tag{4.1}$$

Proof. First of all remember that $f_0(z) = e^{\tau} e^{-\tau/z}$. Defining $\omega = z/\tau$ and using Rodrigues' formula for Laguerre polynomials (see [1] p.101) we obtain

$$\begin{aligned} \frac{d^{m+k}}{dz^{m+k}} \exp(-\frac{\tau}{z}) z^k &= \frac{1}{\tau^m} \frac{d^{m+k}}{d\omega^{m+k}} \exp\left(-\omega^{-1}\right) \left(\omega^{-1}\right)^{-k}, \\ &= \frac{1}{\tau^m} (-1)^{m+k} (m+k)! \exp(-\omega^{-1}) \omega^{-m} L_{m+k}^{(-1-k)}(\omega^{-1}). \end{aligned}$$

The result arises from the relation (see [19] p.240)

$$L_{m+k}^{(-1-k)}(\frac{\tau}{z}) = (-1)^{k+1} (\frac{\tau}{z})^{k+1} \frac{(m-1)!}{(m+k)!} L_{m-1}^{(k+1)}(\frac{\tau}{z}).$$

□

Before stating the main result we need to remember the following properties of the generalized Laguerre polynomials, that can be found in [1] pp. 785-786.

L1

$$L_n^{(\alpha+\beta+1)}(z_1 + z_2) = \sum_{j=0}^{n} L_j^{(\alpha)}(z_1) L_{n-j}^{(\beta)}(z_2).$$

L2

$$L_n^{(\alpha)}(z_1 z_2) = \sum_{j=0}^{n} \binom{n+\alpha}{j} L_j^{(\alpha)}(z_1) z_2^j (1-z_2)^{n-j}.$$

L3

$$\exp(\frac{-x}{2}) \left| L_n^{(\alpha)}(x) \right| \leq \frac{\Gamma(n+\alpha+1)}{n!\Gamma(\alpha+1)}, \quad \text{for } x \geq 0.$$

PROPOSITION 4.2. *Given* $r \geq 0$, *let* $z = \left(1 + \delta r e^{i\theta}\right)^{-1} \in \Sigma_\theta$. *Let moreover*

$$c_j(\theta) := \left(1 + \sqrt{2(1-\cos\theta)}\right)^j. \tag{4.2}$$

Then

$$\left| L_{m-1}^{(k+1)}(\frac{\tau}{z}) \right| \leq e^{\frac{hr}{2}} \sum_{j=0}^{m-1} \left| L_{m-1-j}^{(k)}(\tau) \right| c_j(\theta), \tag{4.3}$$

$$\leq e^{\frac{\tau+hr}{2}} \sum\nolimits_{j=0}^{m-1} \binom{m+k-j-1}{k} c_j(\theta). \tag{4.4}$$

Proof. For $z = \left(1 + \delta r e^{i\theta}\right)^{-1}$

$$\frac{\tau}{z} = \tau + hre^{i\theta}, \quad r \geq 0.$$

Using L1 with $\alpha = k$, $\beta = 0$, $z_1 = \tau$ and $z_2 = hre^{i\theta}$, and then L2 with $z_1 = hr$ and $z_2 = e^{i\theta}$, we have

$$\left|L_{m-1}^{(k+1)}(\frac{\tau}{z})\right| = \left|\sum_{j=0}^{m-1} L_{m-j-1}^{(k)}(\tau) L_j^{(0)}(hre^{i\theta})\right|, \tag{4.5}$$

$$\leq \sum_{j=0}^{m-1} \left|L_{m-j-1}^{(k)}(\tau)\right| \sum_{s=0}^{j} \left|L_s^{(0)}(hr)\right| \left|\binom{j}{s} e^{is\theta}(1-e^{i\theta})^{j-s}\right|.$$

Since

$$\sum_{s=0}^{j} \left|\binom{j}{s} e^{is\theta}(1-e^{i\theta})^{j-s}\right| = c_j(\theta),$$

formulas (4.3) and (4.4) are obtained applying L3 to $L_s^{(0)}(hr)$ and then to $L_{m-j-1}^{(k)}(\tau)$. □

THEOREM 4.3. *Assume that $F(L) \subset S_\theta$, with $\theta < \frac{\pi}{3}$. Then*

$$\|E_{k,m}\| \leq K \frac{e^{\tau\left(\cos\theta - \frac{1}{2}\right) - m - k - 1}}{\tau^{m+k}} \left(\frac{2(m+k+1)}{2\cos\theta - 1}\right)^{m+k+1} C_{k,m}(\tau,\theta) \prod_{i=1}^{m} h_{i+1,i}, \tag{4.6}$$

$$\leq K \frac{e^{\tau\cos\theta - m - k - 1}}{\tau^{m+k}} \left(\frac{2(m+k+1)}{2\cos\theta - 1}\right)^{m+k+1} C'_{k,m}(\theta) \prod_{i=1}^{m} h_{i+1,i}, \tag{4.7}$$

where

$$C_{k,m}(\tau,\theta) := \frac{(m-1)!}{(m+k)!} \sum_{j=0}^{m-1} \left|L_{m-1-j}^{(k)}(\tau)\right| c_j(\theta), \tag{4.8}$$

$$C'_{k,m}(\theta) := \frac{(m-1)!}{(m+k)!} \sum\nolimits_{j=0}^{m-1} \binom{m+k-j-1}{k} c_j(\theta), \tag{4.9}$$

and K defined as in Proposition 3.4.

Proof. For $z \in \Sigma_\theta$

$$\frac{1}{z} = 1 + \delta r e^{i\theta}, \quad r \geq 0,$$

and

$$f_0(z) = e^{\tau - \frac{\tau}{z}} = e^{-hre^{i\theta}}.$$

Hence, using (3.12), (4.1) and (4.3) we obtain

$$\|E_{k,m}\| \leq K \max_{r\geq 0} \left|e^{-hr(\cos\theta - \frac{1}{2})} \left(1 + \delta r e^{i\theta}\right)^{m+k+1}\right| \tau$$

$$\times \frac{(m-1)!}{(m+k)!} \sum_{j=0}^{m-1} \left|L_{m-1-j}^{(k)}(\tau)\right| c_j(\theta) \prod_{i=1}^{m} h_{i+1,i}. \tag{4.10}$$

Since for $\theta < \pi/3$

$$e^{-hr\left(\cos\theta - \frac{1}{2}\right)} (1+\delta r)^{m+k+1} \leq \frac{e^{\tau\left(\cos\theta - \frac{1}{2}\right) - m - k - 1}}{\tau^{m+k+1}} \left(\frac{2(m+k+1)}{2\cos\theta - 1}\right)^{m+k+1},$$

(looking for the maximum with respect to r), we immediately obtain (4.6). Using again (3.12) and (4.1) but now with (4.4) we arrive at the coarser bound (4.7). □

REMARK 4.4. *While formulas (4.6) and (4.7) theoretically hold for $\theta < \frac{\pi}{3}$ since $h_{m+1,m} = 0$ for $m \leq M$, it is necessary to point out that for $\theta \approx \frac{\pi}{3}$ we may observe a rapid growth of the term*

$$\left(\frac{1}{2\cos\theta - 1}\right)^{m+k+1} \prod_{i=1}^{m} h_{i+1,i},$$

depending of course on the problem, so that the bounds may be useless. This situation is caused by the bound (4.3) that leads the quantity $2\cos\theta - 1$ at the denominator. Working in inexact arithmetic the situation is even more difficult because of the loss of orthogonality of the vectors v_j of the Arnoldi algorithm and hence the accumulation of errors on the entries $h_{i+1,i}$. For these reasons, in practice, formulas (4.6) and (4.7) should be used only for θ not much close to $\frac{\pi}{3}$.

REMARK 4.5. *For the exponential case ($k = 0$) we have*

$$C'_{0,m}(\theta) = \frac{1}{m} \sum\nolimits_{j=0}^{m-1} c_j(\theta),$$

and hence by (4.7)

$$\|E_{0,m}\| \leq K \frac{e^{\tau\cos\theta - m - 1}}{m\tau^m} \left(\frac{2(m+1)}{2\cos\theta - 1}\right)^{m+1} \sum\nolimits_{j=0}^{m-1} c_j(\theta) \prod_{i=1}^{m} h_{i+1,i}. \tag{4.11}$$

REMARK 4.6. *In the self-adjoint case $\theta = 0$ we have $c_j(\theta) = 1$ and formula (4.7) simplifies to*

$$\|E_{k,m}\| \leq K \frac{e^{\tau - m - k - 1}}{\tau^{m+k}} \frac{(2(m+k+1))^{m+k+1}}{(k+1)!} \prod_{i=1}^{m} h_{i+1,i}.$$

The reason for which we consider two bounds in Theorem 4.3 is that the second one (4.7) allows us to define suitably the parameter τ (and then δ) while the first one (4.6) should be used whenever τ has been defined. Indeed, assuming $\prod_{i=1}^{m} h_{i+1,i}$ independent of δ and then of τ (actually this is not true as we explain in Section 9) by (4.7), looking for the minimum of $e^{\tau\cos\theta}\tau^{-(m+k)}$ we easily find that the optimal value for τ is given by

$$\tau = \frac{m+k}{\cos\theta}. \tag{4.12}$$

The following result considers the bound (4.7) at the iteration m that defines τ in (4.12).

COROLLARY 4.7. *Assume that $F(L) \subset S_\theta$, with $\theta < \frac{\pi}{3}$. Taking $\tau = \frac{m+k}{\cos\theta}$ we have*

$$\|E_{k,m}\| \leq K \left(\frac{2\cos\theta}{2\cos\theta - 1}\right)^{m+k+1} \left(1 + \sqrt{2(1-\cos\theta)}\right)^{m} \frac{1}{k!} \prod_{i=1}^{m} h_{i+1,i}. \tag{4.13}$$

Proof. By the definitions (4.2) and (4.9), it is rather easy to show that

$$C'_{k,m}(\theta) = \frac{(m-1)!}{(m+k)!}\sum\nolimits_{j=0}^{m-1}\binom{m+k-j-1}{k}c_j(\theta),$$
$$\leq \frac{1}{k!(m+k)}\left(\frac{m-1}{m+k-1}\right)^m c_m(\theta). \tag{4.14}$$

Substituting (4.14) in (4.7) we obtain the result. □

The above corollary is quite important since it allows to understand what happens at an iteration number that is expected to be close to the convergence. In particular, the bound (4.13) clearly shows the dependence on the angle θ and the difficult of working with $\theta \approx \frac{\pi}{3}$, since in this case

$$\left(\frac{2\cos\theta}{2\cos\theta-1}\right)^{m+k+1}\left(1+\sqrt{2(1-\cos\theta)}\right)^m \approx \frac{2^m}{\left[\sqrt{3}(\frac{\pi}{3}-\theta)\right]^{m+k+1}}.$$

In practice, in order to determine θ and use the a-posteriori bounds provided by Theorem 4.3 one may compute the boundary of $F(L)$ using the standard codes available in literature (as for instance the Matlab code `fv.m` by Higham [11]). It is important to observe that θ is generally independent of the discretization so that one can work in smaller dimension.

While the hypothesis $F(L) \subset S_\theta$ of Theorem 4.3 is extremely general, the underlying assumption is that L represents an arbitrary sharp discretization of an unbounded operator. On the other side, if it is known that $F(L)$ is contained in a bounded sector then Proposition 3.3 can be used to derive sharper error estimates. In general we may refer again to [3] and the references therein for a background on the most used techniques based on the use of the integral representation of the error.

Anyway, here we want also to show how to adapt our approach in presence of more information on $F(L)$. Let $D_{0,R}$ be the disk centered at 0 with radius hR, and assume that $F(L) \subset S_\theta \cap D_{0,hR}$. Using again (3.12) and (4.1), we arrive at the bound

$$\|E_{k,m}\| \leq K \max_{0\leq s\leq hR}\left|e^{-s\cos\theta}\left(1+\frac{s}{\tau}\right)^{m+k+1} L_{m-1}^{(k+1)}(\tau+se^{i\theta})\right|\tau$$
$$\times\frac{(m-1)!}{(m+k)!}\prod_{i=1}^{m}h_{i+1,i}. \tag{4.15}$$

In order to define a suitable value for τ, we just need to bound the Laguerre polynomials as in (4.4), so that the optimal value is obtained looking for the minimum of

$$\tau\left(1+\frac{s}{\tau}\right)^{m+k+1} e^{\frac{\tau}{2}}.$$

A good approximation for this minimum is given by

$$\tau = \sqrt{2hR(m+k+1)}, \tag{4.16}$$

that is obtained considering the bound

$$\left(1+\frac{s}{\tau}\right)^{m+k+1} \leq \exp\left((m+k+1)\frac{hR}{\tau}\right).$$

Using this value of τ we can derive practical error bounds seeking for the maximum of the function $\left|e^{-s\cos\theta}\left(1+\frac{s}{\tau}\right)^{m+k+1}L_j^{(0)}(se^{i\theta})\right|$ (cf. (4.5)) in the interval $[0,hR]$.

5. A-priori error bounds. Formula (4.12) obviously requires to know the number of iterations that are necessary to achieve a certain accuracy. In this sense we need to bound in some way $\prod_{i=1}^{m}h_{i+1,i}$. By (3.8) and since

$$\|q_m(Z)v\| \le \|p_m(Z)v\|$$

for each monic polynomial p_m of exact degree m (see [32] p. 269), a bound for $\prod_{i=1}^{m}h_{i+1,i}$ can be stated using Faber polynomials as explained in [2], that leads to

$$\prod_{i=1}^{m}h_{i+1,i} = \|q_m(Z)v\| \le 2\gamma(G)^m, \tag{5.1}$$

where $\gamma(G)$ is the logarithmic capacity of a compact G containing $F(Z)$ and where f_k is analytic.

Now consider the function

$$\rho(\theta) := \left(1+\sqrt{2(1-\cos\theta)}\right)\frac{\cos\theta}{4\cos\theta-2}\frac{\pi}{\pi-\theta}. \tag{5.2}$$

Since $1/2 \le \rho(\theta) < 1$ for $0 \le \theta < \theta^*$, where $\theta^* = 0.48124$, we can state the following result.

PROPOSITION 5.1. *Assume that $F(L) \subset S_\theta$, with $\theta < \theta^*$. Then for $\tau = (m+k)/\cos\theta$*

$$\|E_{k,m}\| \le 11K\rho(\theta)^m. \tag{5.3}$$

Proof. Since $F(Z) \subset G_\theta$ by Proposition 3.1, let us consider the compact subset $G = G_\theta$. The associated conformal mapping

$$\psi : \mathbb{C}\setminus\{w : |w| \le 1\} \to \mathbb{C}\setminus G_\theta,$$

is given by

$$\begin{aligned}\psi(w) &= \frac{(w+1)^{2-\nu}}{(w+1)^{2-\nu}-(w-1)^{2-\nu}},\\ &= \frac{1}{2(2-\nu)}w + \frac{1}{2} + \frac{1}{6}\frac{(1-\nu)(3-\nu)}{2-\nu}\frac{1}{w} + O\left(\frac{1}{w^2}\right),\end{aligned} \tag{5.4}$$

where $\nu = 2\theta/\pi$. The coefficient of the leading term of the Laurent expansion (5.4) is the logarithmic capacity, so that by (5.1) we have

$$\prod_{i=1}^{m}h_{i+1,i} \le 2\left(\frac{1}{2(2-\nu)}\right)^m. \tag{5.5}$$

Inserting this bound in (4.7) we easily obtain for $\theta < \frac{\pi}{3}$

$$\begin{aligned}\|E_{k,m}\| &\le K\frac{e^{\tau\cos\theta-m-k-1}}{\tau^{m+k}}\left(\frac{m+k+1}{2\cos\theta-1}\right)^{m+k+1}2^{-m+k+2}\left(\frac{\pi}{\pi-\theta}\right)^m C'_{k,m}(\theta),\\ &\le K\frac{m+k+1}{\cos\theta}\left(\frac{\cos\theta}{2\cos\theta-1}\right)^{m+k+1}2^{-m+k+2}\left(\frac{\pi}{\pi-\theta}\right)^m C'_{k,m}(\theta),\end{aligned}$$

where the second inequality arises from the choice $\tau = (m+k)/\cos\theta$.

Now, using the bound (4.14)

$$C'_{k,m}(\theta) \leq \frac{1}{k!(m+k)} \left(\frac{m-1}{m+k-1} \right)^m c_m(\theta),$$

we have

$$\|E_{k,m}\| \leq K \frac{e^{-k}}{k! \cos\theta} \left(\frac{\cos\theta}{2\cos\theta - 1} \right)^{k+1} 2^{k+3} \left[\rho(\theta)\right]^m. \tag{5.6}$$

Since for each $k \geq 0$

$$\frac{e^{-k}}{k! \cos\theta} \left(\frac{\cos\theta}{2\cos\theta - 1} \right)^{k+1} 2^{k+3} \leq \frac{8}{\cos\theta^*} \left(\frac{\cos\theta^*}{2\cos\theta^* - 1} \right) = 10.351$$

the proof is complete. □

REMARK 5.2. *We point out the bound (5.3) only holds for the value of m that defines τ. Whenever the right-hand side of (5.3) is less than the prescribed accuracy the corresponding m will be used to define τ (δ) and then the matrix Z.*

REMARK 5.3. *Proposition 5.1 shows the mesh-independence of the method for $\theta < \theta^*$ since the bound (5.3) is independent of the discretization of the underlying sectorial operator. By (5.6) and (5.2), in the self-adjoint case ($\theta = 0$) the bound (5.3) reads*

$$\|E_{k,m}\| \leq \frac{8}{k!} \left(\frac{2}{e} \right)^k \left(\frac{1}{2} \right)^m.$$

It is worth noting that by (3.7) for every $p_{m-1} \in \Pi_{m-1}$ we have that

$$\|E_{k,m}\| \leq 2K \max_{z \in G} |f_k(z) - p_{m-1}(z)|,$$

where we assume that $G \subset D_{1/2,1/2}$ is compact, connected, with associated conformal mapping ϕ, and such that $F(Z) \subset G$. Therefore, in principle, one could try to derive a-priori error bounds choosing suitably the polynomial sequence $\{p_{m-1}\}_{m\geq 1}$. Anyway, the classical results in complex polynomial approximation state that even taking $\{p_{m-1}\}_{m\geq 1}$ as a sequence of polynomials that asymptotically behaves as the sequence of polynomial of best uniform approximation of f_k on G (see e.g [29] for a theoretical background and examples) we have

$$\left[\max_{z \in G} |f_k(z) - p_{m-1}(z)| \right]^{1/m} \to \frac{1}{R} \quad \text{as } m \to \infty,$$

where R is such that $\phi(-R) = 0$, since f_k is singular at 0 (*maximal convergence* property, see e.g [34] Chapter IV). The main problem is that assuming L to be unbounded, $0 \in G$ and consequently $R = 1$.

For this reasons, in our opinion the only reasonable approach to derive a-priori error bounds, is to define $\{p_{m-1}\}_{m\geq 1}$ as a sequence of polynomials interpolating f_k at point belonging to G, and then to use the Hermite-Genocchi formula to bound the divided differences. Using this formula and taking for instance p_{m-1} as the sequence of interpolants at the zeros of Faber polynomials we just obtain the error bound given in Proposition 5.1 (see [21]).

6. The generalized residual. By the integral representation of function of matrices and (2.4), we know that the error can be written as

$$E_{k,m} = \frac{1}{2\pi i} \int_{\Sigma_\theta} f_k(z)[(zI - Z)^{-1}v - V_m(zI - H_m)^{-1}e_1]dz. \tag{6.1}$$

In order to monitor the approximations during the computation we can consider the so-called generalized residual [14], defined as

$$R_{k,m} = \frac{1}{2\pi i} \int_{\Gamma} f_k(z) r_m(z) dz, \tag{6.2}$$

which is obtained from (6.1) by replacing the error

$$(zI - Z)^{-1}v - V_m(zI - H_m)^{-1}e_1$$

with the corresponding residual

$$r_m(z) = v - (zI - Z)V_m(zI - H_m)^{-1}e_1.$$

Using the fundamental relation (2.3) we have immediately

$$r_m(z) = h_{m+1,m}(e_m^H(zI - H_m)^{-1}e_1)v_{m+1},$$

and inserting this relation in (6.2) we obtain

$$R_{k,m} = h_{m+1,m}(e_m^H f_k(H_m)e_1)v_{m+1},$$

so that we may assume

$$E_{k,m} \approx \|R_{k,m}\| = h_{m+1,m} \left| e_m^H f_k(H_m)e_1 \right|. \tag{6.3}$$

In order to show the reliability of this approximation let us consider the operator

$$Lu = -u'' + cu', \quad c \geq 0, \tag{6.4}$$

discretized with central differences in $[0, 1]$ with uniform mesh $h = 1/(M + 1)$, and Dirichelet boundary conditions. For our examples, we consider the computation of $\varphi_k(hL)v$ for $k = 1, 2$, with $v = (1, ..., 1)^T/\sqrt{M}$, comparing the exact error and the generalized residual. We take $M = 1000$, $h = 0.1$, and we consider the cases of $c = 2$ and $c = 4$, whose corresponding sector semiangles are respectively $\theta = 0.201$ and $\theta = 0.425$. We define $\tau = 15/\cos\theta$. The results, collected in Figure 6.1, shows the accuracy of the approximation (6.3).

It is necessary to point out that the use of (6.3) has the basic disadvantage that it requires the computation of $f_k(H_m)$, $m = 1, 2, ...$, and this is a computational drawback whenever a great amount of matrix functions evaluations are required to integrate a certain problem, even if m can be considered much smaller than M. Moreover, it frequently happens (as in our experiments) that the generalized residual tends to underestimate the error during the first iterations, and this can be particularly dangerous when computing $\varphi_{k+1}(hL)v$ with $\|v\| \ll 1$, as for instance in the case of the computation of the internal stages of an exponential Runge-Kutta method, in which $\|v\| = O(h)$.

On the other side, exploiting the mesh independence of the method the generalized residual can be successfully used to estimate the optimal value for the parameter τ, that is $\tau_{opt} = (m+k)/\cos\theta$. In other words, using a coarser discretization of the operator we look for the value of m such that using the corresponding τ_{opt} we obtain a certain tolerance in exactly m iterations. For the experiments reported in Figure 6.1 we considered the discretization of (6.4) with only $M = 50$ internal points, observing in both cases that $\|R_{k,m}\| \leq 1e-12$ for $m \geq 13$. For this reason we have chosen $\tau = 15/\cos\theta$.

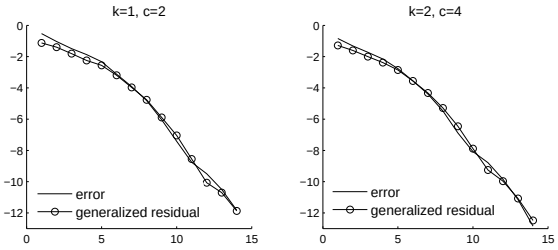

FIG. 6.1. *Comparison between the exact error and the generalized residual for problem (6.4) with $h = 0.1$. In both experiments $\tau = 15/\cos\theta$.*

7. Numerical experiments for the a-posteriori error bound. For our numerical experiments we consider again the operator (6.4), discretized as in previous section. We consider the computation of the functions $\varphi_k(hL)v$, with v as before and $k = 0, 1, 2$, for $h = 0.5$ (Figure 7.1) and $h = 0.05$ (Figure 7.2). In all examples we do not consider the symmetric case corresponding to $c = 0$ (already investigated in [23]), but only the cases $c = 2$ and $c = 4$. As before, for the choice of τ we examined the behavior of the method for the coarser discretization of the same operator with only $M = 50$ interior points, thus exploiting the mesh independence of the method. The analysis suggested to take $\tau = 8/\cos\theta$ for all experiments with $h = 0.5$ and $\tau = 15/\cos\theta$ for those with $h = 0.05$, thus independently of the function and c, using the tolerance $1e-12$. Inside the Arnoldi iterations the vectors Zv_j, $j \geq 1$ (cf. Section 2), are computed via the LU factorization of $I - \delta L$. The error bound is given by (4.6).

Comparing Figure 7.1 with Figure 7.2 we can observe that the method tends to become slower reducing h. The reason is that for small values of h, the rate of the decay of the singular values of Z becomes slower and this reduces the rate of the decay of $\prod_{i=1}^{m} h_{i+1,i}$. A deeper analysis of this behavior will be presented in Section 9.

8. Non-optimal choice of τ. Employing the RD Arnoldi method inside an exponential integrator requires some considerations. First of all, in our opinion the method can be used only if the implicit computation of Z can be obtained with a sparse factorization technique. The use of an inner-outer iteration can be too much expensive in this context. Indeed, the basic point is that organizing suitably the code one can heavily reduce the number of factorizations of $I - \delta L$ (see e.g [26]), because the method seems to be really robust with respect to the choice of τ. For this reason we want here to show what happens taking τ even quite far from the optimal one.

For simplicity (the situation is representative of what happens in general) let us assume to work with exponential function and $\theta = 0$. We assume moreover that the corresponding bound (4.11) (in which $c_j(\theta) = 1$, $j \geq 0$) is equal to a prescribed

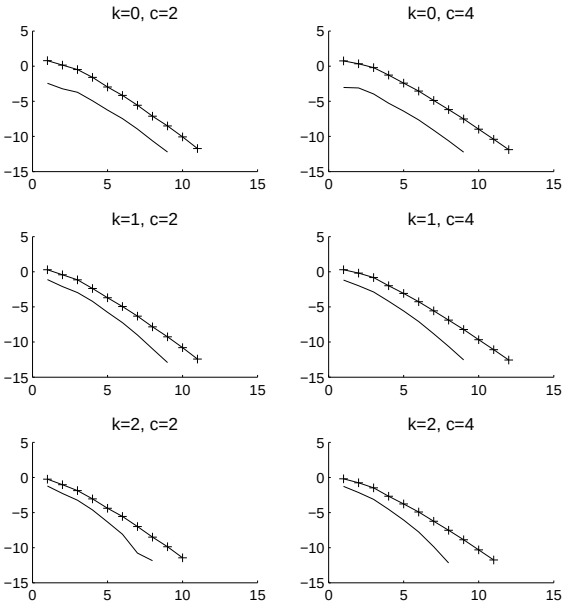

FIG. 7.1. *Error and error bound (4.6) for $k = 0, 1, 2$, $h = 0.5$, L arising from (6.4) with $c = 2$ and $c = 4$.*

tolerance for a certain m with the theoretical optimal choice $\tau_{opt} = m$. We seek for the interval $I_{m,n} = [\tau_1, \tau_2]$ such that for $\tau \in I_{m,n}$ the number of iterations necessary to achieve the same tolerance is at most n ($\geq m$). Using (4.11) and the approximation $h_{m+1,m} \approx 1/4$ ($m > 1$) that is obtained forcing the equal sign in the a-priori bound (5.5), in Figure 8.1 we can observe the result for $n = m + 1, m + 2$. For each m the corresponding extremal points τ_1 and τ_2 of the intervals $I_{m,m+1}$ and $I_{m,m+2}$ are plotted. These points are obtained solving with respect to τ the equation (cf. (4.11))

$$\frac{e^{\tau-n-1}}{\tau^n}(2(n+1))^{n+1}\prod_{i=1}^{n} h_{i+1,i} = \frac{e^{-1}}{m^m}(2(m+1))^{m+1}\prod_{i=1}^{m} h_{i+1,i},$$

for $n = m + 1, m + 2$.

We point out that the results are even a bit conservative with respect to what happens in practice, and this is due to the approximation $h_{m+1,m} \approx 1/4$. Indeed larger intervals would be obtained taking $h_{m+1,m} < 1/4$ as it occurs in practice.

In order prove the effectiveness of the above considerations let us consider again the operator (6.4) with the usual discretization. We consider the case $c = 2$, $k = 1$ for $h = 0.1$. To define τ we consider again the discretization with $M = 50$ interior points observing the generalized residual. This leads us to define $\tau = (m + k)/\cos\theta$ with $m = 14$. In Figure 8.2 we consider the behavior of the method for τ, $\tau/2$ and 2τ.

The robustness of the method with respect to the choice of τ is maybe the most important aspect concerning its use inside an exponential integrator. We want to give here some practical suggestions assuming to use a sparse factorization technique to solve the linear systems with $I - \delta L$, that, computationally, has to be considered the

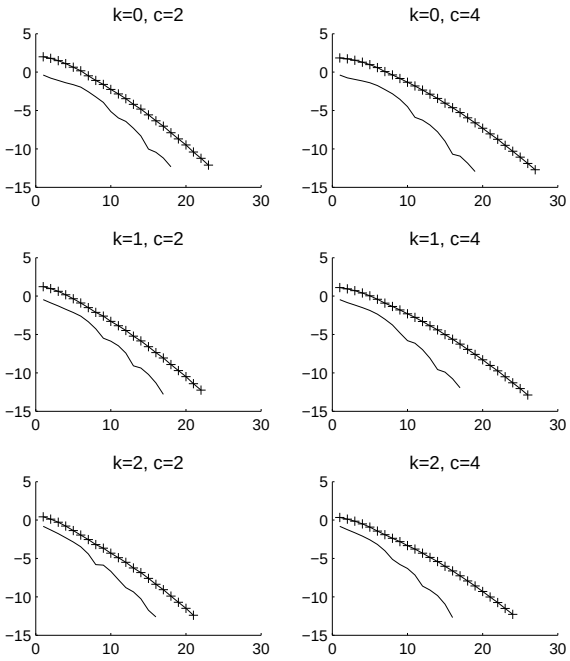

FIG. 7.2. *Error and error bound (4.6) for* $k = 0, 1, 2$, $h = 0.05$, *L arising from (6.4) with* $c = 2$ *and* $c = 4$.

heaviest part of the method.

1. Working in much smaller dimension compute θ and use the generalized residual to estimate the initial τ_{opt}.
2. For nonlinear problems, interpreting L as the Jacobian of the system ([4], [31]), it is necessary to introduce some strategies in order to reduce as much as possible the number of updates of L during the integration, since each update would also require to update the factorization. As for exponential W-method (see [14], [26]), we suggest, whenever it is possible, to work with a time-lagged Jacobian and hence to introduce the necessary order conditions in order to preserve the theoretical order.
3. Using a quasi-constant step-size strategy (without Jacobian update) allows to keep the factorization of $I - \delta L$ constant for a certain number of steps. Whenever it is necessary to update the stepsize $h_{old} \to h_{new}$ without changing the Jacobian, if we want to keep the previous factorization of $I - \delta_{old} L$ we just need to consider the ratio $\tau = h_{new}/\delta_{old}$. If (indicatively) it is bigger than $2\tau_{opt}$ or smaller than $\tau_{opt}/2$ (cf. Figure 8.1 and 8.2), where τ_{opt} arises from a previous analysis of the generalized residual, then we need to update the factorization (cf. again [26]), otherwise we can keep the previous one. In this phase, however, one can even considers other strategies to define suitably the window of admissible values of τ around τ_{opt}, taking into account of the local accuracy required by the integrator, the norm of v, etc.

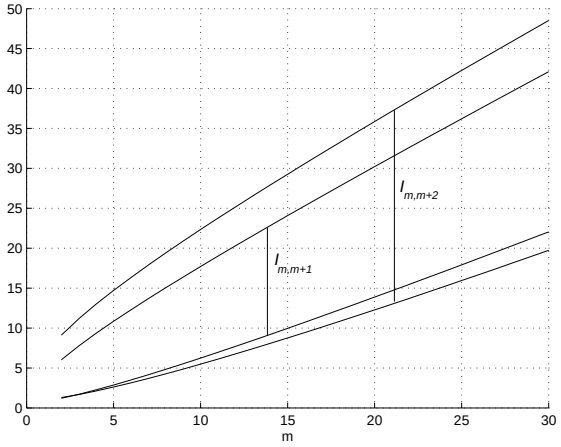

FIG. 8.1. *Boundary of the region $I_{m,m+1}$ and $I_{m,m+2}$.*

9. The superlinear decay of $\prod_{i=1}^{m} h_{i+1,i}$. Looking carefully at Figure 8.2 we notice that while the analysis in smaller dimension suggested to take $\tau = 15/\cos\theta$ for reaching the desired tolerance in exactly 14 iterations the method is unexpectedly a bit faster taking $\tau_1 = \tau/2$ (second picture). The analysis was correct because in larger dimension the method actually achieves the tolerance in 14 iterations (first picture). In order to understand the reason of this behavior, we need to remember that the definition of $\tau_{opt} = (m+k)/\cos\theta$ given at the end of Section 4 was based on the assumption that $\prod_{i=1}^{m} h_{i+1,i}$ is independent of δ, but this is not true. In what follows we try to provide a more accurate analysis studying the decay of $\prod_{i=1}^{m} h_{i+1,i}$.

We denote by σ_j, $j \geq 1$, the singular values of Z. Moreover we denote by λ_j, $j \geq 1$ the eigenvalues of Z and assume that $|\lambda_j| \geq |\lambda_{j+1}|$ for $j \geq 1$. We have the following result (cf. [24] Theorem 5.8.10).

THEOREM 9.1. *Assume that $1 \notin \sigma(Z)$ and*

$$\sum_{j\geq 1} \sigma_j^p < \infty \text{ for a certain } 0 < p \leq 1. \tag{9.1}$$

Let $p_m(z) = \prod_{i=1}^{m}(z - \lambda_i)$. Then

$$\|p_m(Z)\| \leq \Big(\frac{\eta e p}{m}\Big)^{m/p}, \tag{9.2}$$

where

$$\eta \leq \frac{1+p}{p} \sum_{j\geq 1} \sigma_j^p.$$

As already shown in Section 4

$$\prod_{i=1}^{m} h_{i+1,i} \leq \|p_m(Z)v\|$$

for each monic polynomial p_m of exact degree m (see [32] p. 269), so that Theorem 9.1 reveals that the rate of decay of $\prod_{i=1}^{m} h_{i+1,i}$ is superlinear and depends on the

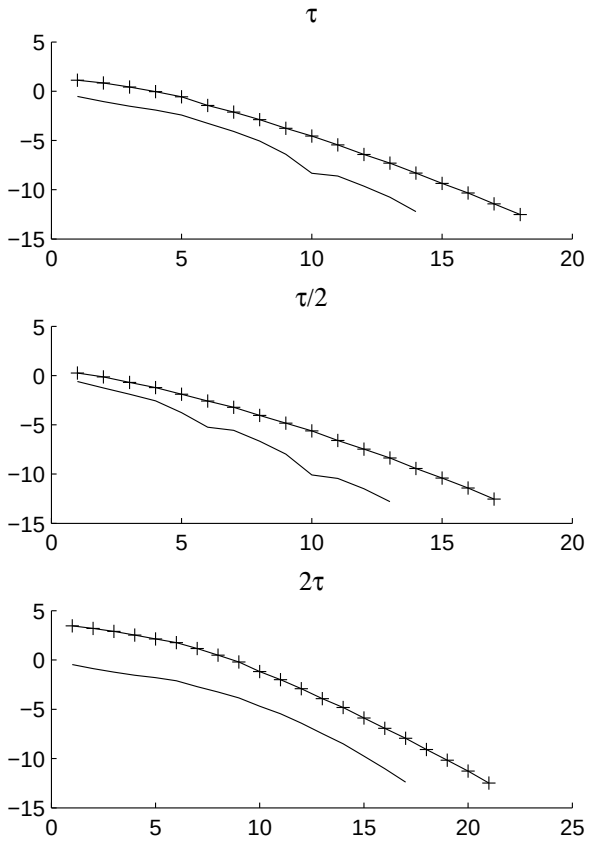

FIG. 8.2. *Error and error bound (4.6) for $k = 1$, $h = 0.1$ and L arising from (6.4) with $c = 2$. Method applied with $\tau = 15/\cos\theta$, $\tau/2$ and 2τ.*

p-summability of the singular values of Z. We remark moreover that an almost equal bound has been obtained in [10] studying the convergence of the smallest Ritz value of the Lanczos process for self-adjoint compact operators.

In practice, the use of (9.2) requires the knowledge of p and a bound for η, that is, information about the singular values of the operator Z. As a model problem we consider again the operator L defined by (6.4) with $c = 0$, whose eigenvalues are $(j\pi)^2$, $j \geq 1$, so that the eigenvalues of Z are given by $\lambda_j = 1/(1 + \delta\,(j\pi)^2)$. In this case (9.1) holds for $1/2 < p \leq 1$ so that Z can be referred to as a *trace class* operator

(see again [24]). Hence, taking for instance $p = 1$ we have

$$\begin{aligned}\sum_{i\geq 1}\sigma_i^p &\leq \frac{1}{\sqrt{\delta}}\left(\frac{1}{2}-\frac{\arctan\left(\sqrt{\delta}\pi\right)}{\pi}\right),\\ &= \frac{1}{\sqrt{\delta}\pi}\arctan\left(\frac{1}{\sqrt{\delta}\pi}\right),\\ &\leq \frac{1}{2\sqrt{\delta}},\end{aligned} \tag{9.3}$$

and so

$$\prod_{i=1}^{m} h_{i+1,i} \leq \|p_m(Z)v\| \leq \left(\frac{e}{\sqrt{\delta}m}\right)^m. \tag{9.4}$$

The bound (9.4) reveals that the rate of decay depends on the choice of δ and then on h. For large values of δ, say $\delta \geq 1$, the bound (9.3) can be heavily improved exploiting the properties of the arctan function and the convergence is extremely fast. The following proposition states a general superlinear bound that can be used when L is an elliptic differential operator of the second order, so with singular values growing like j^2. The proof is straightforward since we just require to bound $\sum_{j\geq 1}\sigma_j^p$, and apply (9.2) with $p = 1$.

PROPOSITION 9.2. *Let L be an elliptic differential operator of the second order. Then there exists a constant C such that*

$$\prod_{i=1}^{m} h_{i+1,i} \leq \left(\frac{C}{\sqrt{\delta}m}\right)^m. \tag{9.5}$$

This proposition can easily be generalized to operator of order $s \geq 1$, exploiting [24] Corollary 5.8.12 in which the author extends Theorem 9.1 for $p > 1$. Anyway, this is beyond the purpose of this section.

From a practical point of view, formula (9.5) is almost useless since too much information on L would be required. On the other side, it is fundamental to understand the dependence on δ. Setting as usual $\tau = h/\delta$ and putting the corresponding bound (9.5) in Theorem 4.3 (formula (4.7)), we easily find that the theoretical optimal value for τ is obtained seeking for the minimum of

$$\frac{e^{\tau\cos\theta - m - k - 1}}{\tau^{m+k}}\left(\frac{C\sqrt{\tau}}{\sqrt{h}m}\right)^m$$

with respect to τ, that is,

$$\tau_{opt} = \frac{m+2k}{2\cos\theta}.$$

This new value, less than $(m+k)/\cos\theta$, explains our considerations about Figure 8.2 given at the beginning of this section.

We need to point out that since the choice of τ_{opt} is independent of C and h, formula (9.5) is quite coarse for small values of h and not able to catch the fast decay of $\prod_{i=1}^{m} h_{i+1,i}$. In any case, if an estimate of C is available an a-priori bound for the RD Arnoldi method can be obtained taking

$$\prod_{i=1}^{m} h_{i+1,i} \leq \min\left\{\left(\frac{C\sqrt{\tau}}{\sqrt{h}m}\right)^m, 2\left(\frac{1}{2(2-\nu)}\right)^m\right\},$$

(cf. (5.5)). Consequently we argue that

$$\frac{m+2k}{2\cos\theta} \leq \tau_{opt} \leq \frac{m+k}{\cos\theta},$$

with τ_{opt} close to $(m+2k)\,/\,(2\cos\theta)$ for h large and to $(m+k)\,/\cos\theta$ for h small.

10. Conclusions. In this paper we have tried to provide all the necessary information to employ the RD Arnoldi method as a tool for solving parabolic problems with exponential integrators. The little number of codes available in literature, and consequently, the little number of comparisons with classical solvers is a source of skepticism about the practical usefulness of this kind of integrators. Indeed, with respect to the most powerful classical methods for stiff problems, the computation of a large number of matrix functions, generally performed with a polynomial method, is still representing a drawback because of the computational cost. The use of polynomial methods for these computations may even be considered inadequate whenever we assume to work with an arbitrarily sharp discretization of the operator, since this would result in a problem of polynomial approximation in arbitrarily large domains. For these reasons, the use of rational approximations as the one here presented, should be considered a valid alternative because of the fast rate of convergence and the mesh independence property, provided that we are able to exploit suitably the robustness of the method with respect to the choice of the poles, as explained in Section 8 for our case

Acknowledgement 10.1. *The author is grateful to Igor Moret and Marco Vianello for many helpful discussions and suggestions.*

REFERENCES

[1] M. Abramovitz and A. Stegun, *Handbook of Mathematical Functions*, Dover Publications Inc., New York, 1965.
[2] B. Beckermann, *Image numérique, GMRES et polynômes de Faber*, C. R. Math. Acad. Sci. Paris, 340 (2005), pp. 855–860.
[3] B. Beckermann and L. Reichel, *Error estimation and evaluation of matrix functions via the Faber transform*, SIAM J. Numer. Anal., 47 (2009), pp. 3849-3883.
[4] M. Caliari and A. Ostermann, *Implementation of exponential Rosenbrock-type integrators*, Appl. Numer. Math., 59 (2009), pp. 568–581.
[5] M. Crouzeix, *Numerical range and numerical calculus in Hilbert space*, J. Func. Anal., 244 (2007), pp. 668–690.
[6] C. de Boor, *Divided differences*, Surveys in approximation theory, 1 (2005), pp. 46–69.
[7] F. Diele, I. Moret and S. Ragni, *Error estimates for polynomial Krylov approximations to matrix functions*, SIAM J. Matrix Anal. Appl., 30 (2008), pp. 1546–1565.
[8] V. Druskin and L. Knizhnerman, *Extended Krylov subspaces: approximation of the matrix square root and related functions*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 755–771.
[9] M. Eiermann and O.G. Ernst, *A restarted Krylov subspace method for the evaluation of matrix functions*, SIAM J. Numer. Anal., 44 (2006), pp. 2481–2504.
[10] M. Hanke, *Superlinear convergence rates for the Lanczos method applied to elliptic operators*, Numer. Math., 77 (1997), pp. 487–499.
[11] N. J. Higham, *Matrix Computation Toolbox. Version 1.2*, www.mathworks.com, 2002.
[12] N. J. Higham, *Functions of matrices. Theory and computation*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2008.
[13] M. Hochbruck and C. Lubich, *On Krylov subspace approximations to the matrix exponential operator*, SIAM J. Numer. Anal., 34 (1997), pp. 1911–1925.
[14] M. Hochbruck, C. Lubich and H. Selhofer, *Exponential integrators for large systems of differential equations*, SIAM J. Sci. Comput., 19 (1998), pp. 1552–1574.
[15] M. Hochbruck and A. Ostermann, *Exponential integrators*, Acta Numerica, 19 (2010), pp. 209–286.
[16] T. Kato, *Perturbation Theory for Linear Operators*, Springer, Berlin, 1976.

[17] L. Knizhnerman, *Calculation of functions of unsymmetric matrices using Arnoldi's method*, U.S.S.R. Comput. Maths. Math. Phys., 31 (1991), pp. 1–9.
[18] L. Knizhnerman and V. Simoncini, *A new investigation of the extended Krylov subspace method for matrix function evaluations*, Numer. Linear Algebra Appl., 17 (2010), pp. 615–638.
[19] W. Magnus, F. Oberhettinger and R. P. Soni, *Formulas and theorems for the special functions of mathematical physics. Third enlarged edition*, Springer-Verlag, New York, 1966.
[20] B. V. Minchev and W. M. Wright, *A review of exponential integrators for first order semi-linear problems*, Preprint Numerics 2/2005, Norwegian University of Science and Technology, Trondheim, Norway.
[21] I. Moret and P. Novati, *An interpolatory approximation of the matrix exponential based on Faber polynomials*, J. Comput. Appl. Math., 131 (2001), pp. 361–380.
[22] I. Moret and P. Novati, *RD-rational approximations of the matrix exponential*, BIT, 44 (2004), pp. 595–615.
[23] I. Moret, *On RD-rational Krylov approximations to the core-functions of exponential integrators*, Numer. Linear Algebra Appl., 14 (2007), pp. 445–457.
[24] O. Nevanlinna, *Convergence of Iterations for Linear Equations*, Birkhäuser, Basel, 1993.
[25] S. P. Norsett, *Restricted Padé approximations to the exponential function*, SIAM J. Numer. Anal., 15 (1978), pp. 1008–1029.
[26] P. Novati, *On the construction of restricted-denominator exponential W-methods*, J. Comput. Appl. Math., 221 (2008), pp. 86–101.
[27] M. Popolizio and V. Simoncini, *Acceleration techniques for approximating the matrix exponential operator*, SIAM J. Matrix Anal. Appl., 30 (2008), pp. 657–683.
[28] Y. Saad, *Analysis of some Krylov subspace approximations to the matrix exponential operator*, SIAM J. Numer. Anal., 29 (1992), pp. 209–228.
[29] V. I. Smirnov and N. A. Lebedev, *Functions of a Complex Variable - Constructive Theory*, Iliffe Books, London, 1968.
[30] M. N. Spijker, *Numerical ranges and stability estimates*, Appl. Numer. Math., 13 (1993), pp. 241–249.
[31] M. Tokman, *Efficient integration of large stiff systems of ODEs with exponential propagation iterative (EPI) methods*, J. Comput. Phys., 213 (2006), pp. 748–776.
[32] L. N. Trefethen and D. Bau, *Numerical Linear Algebra*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1997.
[33] J. v. d. Eshof and M. Hochbruck, *Preconditioning Lanczos approximations to the matrix exponential*, SIAM J. Sci. Comput., 27 (2005), pp. 1438–1457.
[34] J. L. Walsh, *Interpolation and Approximation by Rational Functions in the Complex Domain*, AMS, Providence, 1965.
[35] P. F. Zachlin, *On the field of values of the inverse of a matrix*, PhD Thesis, Case Western Reserve University, 2008.