ERROR ESTIMATION AND EVALUATION OF MATRIX FUNCTIONS VIA THE FABER TRANSFORM

BERNHARD BECKERMANN* AND LOTHAR REICHEL†

Abstract. The need to evaluate expressions of the form $f(A)$ or $f(A)b$, where f is a nonlinear function, A is a large sparse $n \times n$ matrix, and b is an n-vector, arises in many applications. This paper describes how the Faber transform applied to the field of values of A can be used to determine improved error bounds for popular polynomial approximation methods based on the Arnoldi process. Applications of the Faber transform to rational approximation methods and, in particular, to the rational Arnoldi process, also are discussed.

Key words. matrix function, polynomial approximation, rational approximation, Arnoldi process, rational Arnoldi process, error bound

AMS subject classifications. 30E10, 65E05, 65F30

1. Introduction.

Many applications in science and engineering require the evaluation of expressions of the form

$$f(A) \quad \text{or} \quad f(A)b, \qquad \text{where} \qquad A \in \mathbb{C}^{n\times n}, \quad b \in \mathbb{C}^n, \tag{1.1}$$

and f is a nonlinear function. The expressions (1.1) can be defined in terms of the Jordan canonical form of A, the minimal polynomial of A, or by a Cauchy-type integral. The latter definition requires f to be analytic in an open set containing the spectrum of A, with the path of integration in this set. Detailed discussions on these definitions and their requirements on f are provided by Golub and Van Loan [42, Chapter 11], Higham [50], and Horn and Johnson [52, Chapter 6]. Of particular interest are the entire functions

$$f(t) = \exp(t), \quad f(t) = (1 - \exp(t))/t, \quad f(t) = \cos(t), \quad f(t) = \sin(t),$$

with applications to the solution of ordinary and partial differential equations [2, 12, 30, 32, 40, 46, 49, 51, 58, 59, 63, 70, 74, 76, 81] as well as to inverse problems [13, 14]. Other functions of interest include Markov functions, such as $f(t) = \sqrt{t}$, which arises in the solution of systems of stochastic differential equations [3, 9, 29]. The function $f(t) = \log(t)$ is a modification of a Markov function and also can be treated with the methods of the present paper; see [15, 48, 50] for applications.

When the matrix A is small to medium-sized, the expressions (1.1) can be evaluated by determining a suitable factorization of A, e.g., in combination with a rational approximation of f; algorithms that factor A are described and analyzed in several of the above references as well as in [9, 15, 21, 38, 61, 76].

The present paper is concerned with the approximation of the expressions (1.1) when f is an entire or Markov function and the matrix A is large, sparse, and nonnormal. The methods described also apply when A is a normal matrix and simplify in this case, in particular, when A is Hermitian or skew-Hermitian. A convex compact set $\mathbb{E}$, which contains the field of values of A, defined by

$$\mathbb{W}(A) := \left\{ \frac{(Ay, y)}{(y, y)} : y \in \mathbb{C}^n \setminus \{0\} \right\}$$

* Laboratoire Painlevé UMR 8524 (ANO-EDP), UFR Mathématiques – M3, UST Lille, F-59655 Villeneuve d'Ascq CEDEX, France. E-mail: `bbecker@math.univ-lille1.fr`.

† Department of Mathematical Sciences, Kent State University, P.O. Box 5190, Kent, OH 44242. E-mail: `reichel@math.kent.edu`.

is assumed to be explicitly known. Since the field of values is convex, it is natural to choose $\mathbb{E}$ to be convex as well. Here and throughout this paper $(\cdot,\cdot)$ denotes the usual inner product in $\mathbb{C}^n$ and $||\cdot||$ is the induced Euclidean vector or spectral matrix norm; however, the results discussed extend to more abstract finite- or infinite-dimensional Hilbert spaces. For convenience, we sometimes will assume that, besides being convex and compact, the set $\mathbb{E}$ also is symmetric with respect to the real axis. Note that when A is transformed by multiplication by a scalar or by addition of a scalar multiple of the identity, the field of values, and thus $\mathbb{E}$, are transformed in a similar fashion.

This paper discusses polynomial and rational approximation methods. The polynomial methods are based on the Arnoldi process, which simplifies to the (Hermitian) Lanczos process when A is Hermitian. We would like to approximate $f(A)b$ by $p(A)b$, where p is a polynomial of fairly low degree and therefore investigate how well f can be approximated by polynomials on $\mathbb{E}$. In Section 2, we introduce the Faber transform and show that it often suffices to consider the situation when $\mathbb{E}$ is the closed unit disk $\mathbb{D}$. The Faber transform is in Section 3 used to derive new error bounds for polynomial approximants determined via the Arnoldi process. Section 4 applies these results to the approximation of the exponential function and improves bounds reported by Druskin and Knizhnerman [26, 27, 28, 54], and Hochbruck and Lubich [51].

It is well known that some functions, such as the logarithm or fractional powers, can be approximated much better by rational functions than by polynomials on convex sets $\mathbb{E}$ close to the origin. For instance, Kenney and Laub [53] proposed to use Padé approximants at the origin for the computation of $f(A)$, when $f(z) = \log(1-z)$, $\mathbb{E} = \{z \in \mathbb{C} : |z| \le ||A||\}$, and $||A|| < 1$; see also Higham [48, 50] and Davies and Higham [21]. This paper discusses error bounds for rational approximation with preassigned poles. The Faber transform allows us to consider equivalent rational approximation problems on the unit disk, and obtain error bounds in this manner. Section 5 considers application of the rational Arnoldi process, first considered by Ruhe [68], for the determination of rational approximants. Section 6 is concerned with rational approximation of Markov functions. New upper and lower bounds for the approximation error are derived. The smallest error bounds are obtained for rational approximants with carefully chosen distinct poles. Each pole, z_j, requires the solution of a linear system of equations with the matrix $z_jI - A$. If these systems are solved by LU-factorization, then the use of rational approximants with few distinct poles of fairly high multiplicity can be advantageous. We derive error bounds for this situation. The use of rational approximants with multiple poles at the origin and infinity, has been discussed by Druskin and Knizhnerman [29] for the situation when the matrix A is symmetric and positive definite, and by Knizhnerman and Simoncini [55] for more general matrices. Section 7 contains concluding remarks. Other approaches to derive error bounds for certain functions have recently been discussed by Diele et al. [23] and Moret [62]. A careful comparison of these methods with those of the present paper is presently being carried out.

In the remainder of this section, we introduce notation used throughout the paper. Thus, $\mathbb{E}$ denotes a connected compact set in the complex plane $\mathbb{C}$ and is assumed to contain at least two points. The extended complex plane is denoted by $\overline{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$ and ϕ is the Riemann mapping that maps $\overline{\mathbb{C}} \setminus \mathbb{E}$ conformally onto $\overline{\mathbb{C}} \setminus \mathbb{D}$ with the normalization $\phi(\infty) = \infty$ and $\phi'(\infty) > 0$. The inverse map is denoted by ψ, i.e.,

$\psi = \phi^{-1}$. Both ϕ and ψ have similar Laurent expansions at infinity,

$$\begin{array}{rcl} \phi(z) & = & dz + d_0 + d_{-1}z^{-1} + \dots\ , \\ \psi(w) & = & cw + c_0 + c_{-1}w^{-1} + \dots\ , \end{array} \tag{1.2}$$

with $c > 0$ and $d = 1/c$. The coefficient c is commonly referred to as the logarithmic capacity of $\mathbb{E}$ and is denoted by cap $(\mathbb{E})$.

For any $\rho \geq 1$, the set $\mathbb{E}_\rho$ is defined via its complement $\mathbb{E}_\rho^c := \{z \in \mathbb{C}\setminus\mathbb{E} : |\phi(z)| > \rho\}$, i.e., $\mathbb{E}_\rho = \mathbb{C} \setminus \mathbb{E}_\rho^c$. In particular, $\mathbb{E}_1 = \mathbb{E}$. The nth Faber polynomial $F_n = F_n^{\mathbb{E}}$ for the set $\mathbb{E}$ is defined as the polynomial part of the Laurent expansion at infinity of ϕ^n for $n = 0, 1, \dots$; cf. (1.2). Faber polynomials are discussed further below; surveys of their properties are provided, e.g., by Gaier [39, Chapter 1] and Suetin [78].

Example 1.1. Let $\mathbb{E}$ be the closed unit disk. Then $\phi(z) = z$ and the Faber polynomials are given by $F_n^{\mathbb{E}}(z) = z^n$, $n = 0, 1, \dots$. Thus, the $F_n^{\mathbb{E}}$ are Chebyshev polynomials for $\mathbb{E}$. More generally, for $\mathbb{E} = \{z \in \mathbb{C} : |z - z_0| \leq r\}$, we obtain the shifted monomials $F_n^{\mathbb{E}}(z) = (z - z_0)^n/r^n$. □

Example 1.2. Let $\mathbb{E} = [-1, 1]$. The mapping

$$\psi(w) = \frac{1}{2}(w + w^{-1})$$

is known as the Joukowski map. The Faber polynomials $F_n^{[-1,1]}$, $n = 1, 2, \dots$, are twice the Chebyshev polynomials T_n of the first kind, and $F_0^{[-1,1]} = 1$. This follows from the property

$$w^n - F_n^{[-1,1]}(\psi(w)) = \mathcal{O}(1/w), \qquad |w| \to \infty, \tag{1.3}$$

see, e.g., Gaier [39, p. 43], and the fact that

$$T_n\left(\frac{1}{2}(w + w^{-1})\right) = \frac{1}{2}\left(w^n + w^{-n}\right).$$

More generally, when $\mathbb{E}$ is an ellipse, the Faber polynomials $F_n^{\mathbb{E}}$ are Chebyshev polynomials for $\mathbb{E}$ up to a scaling factor. When the foci coalesce, $\mathbb{E}$ becomes a disk, cf. Example 1.1. Details when $\mathbb{E}$ is an ellipse, as well as further examples, can be found in [20, 78]. □

We are interested in polynomial approximation of entire functions and rational approximation of Markov functions. The latter are functions of the form

$$f(z) = \int_\alpha^\beta \frac{d\mu(x)}{z - x}, \tag{1.4}$$

where μ is a positive measure with supp$(\mu) \subset [\alpha, \beta]$, $-\infty \leq \alpha < \beta < \infty$. Thus, f is analytic in $\mathbb{C} \setminus \text{supp}(\mu)$; in particular, f is analytic in $\overline{\mathbb{C}} \setminus [\alpha, \beta]$.

Example 1.3. The function

$$f(z) = \frac{\log(1 + z)}{z}$$

has the representation

$$f(z) = \int_{-\infty}^{-1} \frac{(-1/x)dx}{z - x},$$

and therefore is a Markov function. Moreover,

$$\log(1+z) = zf(z) \tag{1.5}$$

is a simple modification of a Markov function. □

Example 1.4. Let $-1 < \gamma < 0$ and $z \in \mathbb{C} \setminus \mathbb{R}_-$, where $\mathbb{R}_- = \{z \in \mathbb{R} : z \leq 0\}$. Let C be an integration path in $\mathbb{C} \setminus \mathbb{R}_-$ surrounding z. The principal branch of z^γ can be represented by the Cauchy integral,

$$z^\gamma = \frac{1}{2\pi i} \int_C \frac{t^\gamma}{t-z} dt, \qquad i = \sqrt{-1}.$$

Moving the path C towards $\mathbb{R}_-$ yields

$$z^\gamma = \frac{\sin(\pi\gamma)}{\pi} \int_{-\infty}^0 \frac{|t|^\gamma}{t-z} dt, \tag{1.6}$$

which shows that z^γ is a Markov function. The integral in (1.6) exists because the integrand has a singularity of order $\gamma > -1$ at the origin and a zero of order $1+|\gamma| > 1$ at infinity. Fractional powers z^α, for $0 < \alpha < 1$, can be represented by multiplying z^γ by z, similarly as in (1.5). □

It is possible to represent certain meromorphic functions as Markov functions with respect to a discrete measure.

Example 1.5. We obtain from the product representation of the sine function, see, e.g., [64, Section 13.5], that

$$\frac{1}{\sqrt{z}\,\tanh(\sqrt{z})} = \int \frac{d\mu(x)}{z-x}, \qquad \mu = \delta_0 + 2\sum_{j=1}^{\infty} \delta_{j^2\pi^2},$$

with δ_x the Dirac measure at the point x. The error bounds of Section 6, however, are sharp only if the support of μ is the whole interval $[\alpha, \beta]$. □

2. The Faber transform. Let $\mathbb{A}(\mathbb{E})$ denote the Banach algebra of functions analytic in the interior, $\mathrm{Int}(\mathbb{E})$, of $\mathbb{E}$ and continuous on $\mathbb{E}$, equipped with the uniform norm $||\cdot||_{L_\infty(\mathbb{E})}$ on $\mathbb{E}$. Moreover, let $\mathbb{P}_k$ denote the set of polynomials of degree at most k, and $\mathbb{P}_k(\mathbb{E})$ the set of polynomials on $\mathbb{E}$ of degree at most k equipped with the norm $||\cdot||_{L_\infty(\mathbb{E})}$.

The Faber transform $\mathcal{F}$ maps the polynomial

$$p(w) = a_0 w^0 + a_1 w^1 + \ldots + a_k w^k, \qquad a_j \in \mathbb{C},$$

to the polynomial

$$\mathcal{F}(p)(z) = a_0 F_0(z) + a_1 F_1(z) + \ldots + a_k F_k(z),$$

where $F_j = F_j^{\mathbb{E}}$ is the Faber polynomial of degree j for $\mathbb{E}$. Thus, for $p(w) := w^j$, we have $\mathcal{F}(p)(z) = F_j(z)$.

The mapping $\mathcal{F}$ is a bijection from $\mathbb{P}_k(\mathbb{D})$ to $\mathbb{P}_k(\mathbb{E})$ with inverse

$$\mathcal{F}^{-1}(p)(w) = \frac{1}{2\pi i} \int_{\partial\mathbb{D}} p(\psi(w')) \frac{dw'}{w'-w}, \qquad |w| < 1, \qquad p \in \mathbb{P}_k(\mathbb{E}). \tag{2.1}$$

The representation (2.1) follows from the Cauchy formula and the observation that

$$F_n(\psi(w)) = w^n + \mathcal{O}(1/w), \qquad |w| \to \infty;$$

see, e.g., [39, p. 43] for a discussion of the latter; equation (1.3) is a special case.

A set $\mathbb{E}$ is said to be a *Faber set* if there is a constant d, such that for all polynomials p,

$$\|\mathcal{F}(p)\|_{L_\infty(\mathbb{E})} \leq d\|p\|_{L_\infty(\mathbb{D})}. \tag{2.2}$$

For instance, sets $\mathbb{E}$ with a piecewise smooth boundary $\partial\mathbb{E}$ without cusps are Faber sets; see, e.g., Gaier [39, Chapter 1] or Ganelius [41]. The constant d depends on the total rotation of $\partial\mathbb{E}$. Let $\partial\mathbb{E}$ be a rectifiable Jordan curve of bounded total rotation V. Then we may choose

$$d = 1 + 2\frac{V}{\pi}; \tag{2.3}$$

see, e.g., Gaier [39, Theorem 2, pp. 48–49]. For convex sets $\mathbb{E}$, we have $V = 2\pi$, i.e., $\|\mathcal{F}\| \leq 5$. In particular, finite intervals are Faber sets.

A Faber set $\mathbb{E}$ is said to be an *inverse Faber set* if there is a constant d', such that for all polynomials p,

$$\|\mathcal{F}^{-1}(p)\|_{L_\infty(\mathbb{D})} \leq d'\|p\|_{L_\infty(\mathbb{E})}. \tag{2.4}$$

Since the set of polynomials is dense in $\mathbb{A}(\mathbb{E})$, it follows from (2.2) that if $\mathbb{E}$ is a Faber set, then $\mathcal{F}$ admits a unique extension that is continuous from $\mathbb{A}(\mathbb{D})$ to $\mathbb{A}(\mathbb{E})$. We also denote this extension by $\mathcal{F}$. Analogously, if $\mathbb{E}$ is an inverse Faber set, then the inequality (2.4) shows that $\mathcal{F}^{-1}$ can be extended in a unique way to a continuous mapping from $\mathbb{A}(\mathbb{E})$ to $\mathbb{A}(\mathbb{D})$. This extension is also denoted by $\mathcal{F}^{-1}$.

Anderson and Clunie [4, Theorem 2] show that if $\mathbb{E}$ is the closure of a Jordan domain with nonempty interior, whose boundary $\partial\mathbb{E}$ is rectifiable, of bounded boundary rotation, and has no cusps, then $\mathbb{E}$ is an inverse Faber set. Thus, in this situation, $\mathcal{F}$ is a bijection from $\mathbb{A}(\mathbb{D})$ to $\mathbb{A}(\mathbb{E})$ with bounded inverse $\mathcal{F}^{-1}$, and

$$\|\mathcal{F}\| \leq d, \qquad \|\mathcal{F}^{-1}\| \leq d',$$

where d and d' are the constants in (2.2) and (2.4), respectively.

We note for future reference that for sets $\mathbb{E}$ with nonempty interior,

$$\mathcal{F}(p)(z) = \frac{1}{2\pi i}\int_{\partial\mathbb{E}} p(\phi(\zeta))\frac{d\zeta}{\zeta - z}, \qquad z \in \mathrm{Int}(\mathbb{E}). \tag{2.5}$$

Our interest in explicit bounds for the norms of $\mathcal{F}$ and $\mathcal{F}^{-1}$ is motivated by our desire to bound the errors for best uniform polynomial and rational approximation with fixed poles of functions in $\mathbb{A}(\mathbb{E})$. For some polynomial $q(z) = \prod_{j=1}^{m}(z - z_j)$, $z_j \notin \mathbb{E}$, let

$$\eta_k^q(f, \mathbb{E}) := \min\{\|f - \frac{p}{q}\|_{L_\infty(\mathbb{E})} : p \in \mathbb{P}_k\}. \tag{2.6}$$

The residue theorem and (2.1) show that, for any $\hat{w} \in \mathbb{C} \setminus \mathbb{D}$ and $z \in \mathrm{Int}(\mathbb{E})$, we have

$$\mathcal{F}(\frac{1}{w - \hat{w}})(z) = \frac{\psi'(\hat{w})}{z - \psi(\hat{w})}. \tag{2.7}$$

Let

$$\widetilde{q}(w) = (w - w_1)(w - w_2)\ldots(w - w_m), \qquad w_j = \phi(z_j) \notin \mathbb{D}. \tag{2.8}$$

Then the operator $\mathcal{F}$ is a bijection from $\mathbb{P}_k/\widetilde{q}$ onto $\mathbb{P}_k/q$ for $k \geq m-1$; see Ellacott [34], Ganelius [41], or Suetin [79, p. 1324]. It follows that for all $f \in \mathbb{A}(\mathbb{E})$ and $k \geq m-1$, we obtain the bounds

$$\frac{1}{||\mathcal{F}^{-1}||}\,\eta_k^{\widetilde{q}}(\mathcal{F}^{-1}(f),\mathbb{D}) \leq \eta_k^{q}(f,\mathbb{E}) \leq \|\mathcal{F}\|\,\eta_k^{\widetilde{q}}(\mathcal{F}^{-1}(f),\mathbb{D}); \tag{2.9}$$

see, e.g., [4, Theorem 1] or [35] for details. The above inequalities show that it generally suffices to consider best uniform polynomial and rational approximation with fixed poles on $\mathbb{D}$.

In particular cases it is possible to improve the right-hand side bound in (2.9) by considering, instead of $\mathcal{F}$, the modified Faber operators $\mathcal{F}_\pm$, defined for $g \in \mathbb{A}(\mathbb{D})$ by

$$\mathcal{F}_-(g)(z) := \mathcal{F}(g)(z) - g(0), \qquad \mathcal{F}_+(g)(z) := \mathcal{F}(g)(z) + g(0). \tag{2.10}$$

For convex $\mathbb{E}$, it is known that $\|\mathcal{F}_-\| \leq 2$. This bound can be established, e.g., by modifying the proof of [39, Theorem 2, p. 49]. Moreover, it is shown implicitly by Kővari and Pommerenke [56] that $\|\mathcal{F}_+\| \leq 2$. An explicit proof of the latter inequality is given in Theorem 2.1 below. Thus, for convex $\mathbb{E}$ and $k \geq m-1$, we may replace the quantity $\|\mathcal{F}\|$ in (2.9) by 2. We remark that no simple explicit bound for $||\mathcal{F}^{-1}||$ appears to be available.

Theorem 2.1 below generalizes the bound $\|\mathcal{F}_+(g)\| \leq 2\|g\|_{L_\infty(\mathbb{D})}$ for the modified Faber transform to matrix arguments. This enables us to bound the error in matrix function approximations. This generalization is implicitly contained in the double layer potential representation of $f(A)$ discussed by Badea et al. [5, Section 4], but these authors do not establish a connection to the modified Faber transform. Our proof follows fairly closely ideas of Crouzeix and his collaborators on norms of functions of matrices and operators [5, 16, 17, 18, 19, 22]. In particular, Crouzeix [17] shows that for any set $\mathbb{E} \subset \mathbb{C}$ and any matrix or Hilbert space operator A with $\mathbb{W}(A) \subset \mathbb{E}$, the bound

$$\|f(A)\| \leq K\|f\|_{L_\infty(\mathbb{E})}, \qquad \forall f \in \mathbb{A}(\mathbb{E}), \tag{2.11}$$

holds for the universal constant

$$K = 11.08. \tag{2.12}$$

Crouzeix conjectures that the bound (2.11) holds for $K = 2$.

Theorem 2.1. *Let the set $\mathbb{E}$ be convex and consist of more than one point. Then the operator $\mathcal{F}_+$ defined by (2.10) satisfies*

$$\|\mathcal{F}_+\| \leq 2. \tag{2.13}$$

Let $\mathbb{W}(A) \subset \mathbb{E}$. Then, for $\tilde{f} \in \mathbb{A}(\mathbb{D})$, we have

$$||\mathcal{F}_+(\widetilde{f})(A)|| \leq 2\,||\widetilde{f}||_{L_\infty(\mathbb{D})}. \tag{2.14}$$

Proof. We first show (2.13) under the assumption that $\mathrm{Int}(\mathbb{E}) \neq \emptyset$. Let $z \in \mathrm{Int}(\mathbb{E})$. Then the function

$$g(w) := \overline{\widetilde{f}(1/\overline{w})},$$

where the bar denotes complex conjugation, is analytic in $\overline{\mathbb{C}} \setminus \mathbb{D}$ (where, as usual, $\overline{\mathbb{C}}$ denotes the extended complex plane) and continuous on $|w| \geq 1$, with $g(w) = \overline{\widetilde{f}(w)}$ for $|w| = 1$. Thus,

$$\frac{1}{2\pi i}\int_{\partial\mathbb{D}} \overline{\widetilde{f}(w)}\frac{\psi'(w)dw}{\psi(w)-z} = \frac{1}{2\pi i}\int_{\partial\mathbb{D}} g(w)\frac{\psi'(w)dw}{\psi(w)-z} = g(\infty) = \overline{\widetilde{f}(0)},$$

where the second equality follows from the residue theorem applied in the closed complement of $\mathbb{D}$. Adding the conjugate of the above equation to the relation

$$\mathcal{F}(\widetilde{f})(z) = \frac{1}{2\pi i}\int_{\partial\mathbb{D}} \widetilde{f}(w)\frac{\psi'(w)dw}{\psi(w)-z},$$

which is obtained by substituting $\zeta = \psi(w)$ into (2.5), yields

$$\mathcal{F}_+(\widetilde{f})(z) = \frac{1}{\pi}\int_{\partial\mathbb{D}} \widetilde{f}(w)\kappa(w)|dw| \tag{2.15}$$

with

$$\kappa(w) := \frac{1}{2i}\left(\frac{\psi'(w)}{\psi(w)-z}\frac{dw}{|dw|} - \overline{\frac{\psi'(w)}{\psi(w)-z}\frac{dw}{|dw|}}\right). \tag{2.16}$$

Let $\zeta := \psi(w) \in \partial\mathbb{E}$ for $|w| = 1$. Then $\alpha(\zeta) := \arg(\psi'(w)\,dw/|dw|)$ exists for almost all $|w| = 1$, with $e^{i\alpha(\zeta)}$ being the the tangent to $\partial\mathbb{E}$ at ζ. The convexity of $\mathbb{E}$ yields

$$\frac{1}{2i}\left(e^{-i\alpha(\zeta)}(z-\zeta) - \overline{e^{-i\alpha(\zeta)}(z-\zeta)}\right) > 0. \tag{2.17}$$

It follows that $\kappa(w) > 0$ for all $|w| = 1$. We obtain from (2.15) that

$$\begin{aligned} |\mathcal{F}_+(\widetilde{f})(z)| &\leq \frac{1}{\pi}\int_{\partial\mathbb{D}} |\widetilde{f}(w)|\kappa(w)|dw| \leq \|\widetilde{f}\|_{L_\infty(\mathbb{D})}\frac{1}{\pi}\int_{\partial\mathbb{D}} \kappa(w)|dw| \\ &= \|\widetilde{f}\|_{L_\infty(\mathbb{D})}\mathcal{F}_+(1)(z) = 2\|\widetilde{f}\|_{L_\infty(\mathbb{D})}. \end{aligned} \tag{2.18}$$

This establishes (2.13) for $z \in \mathrm{Int}(\mathbb{E})$. Furthermore, since the boundary can be neglected in the L_∞-norm, the inequality (2.18) holds for all $z \in \mathbb{E}$.

Finally, if $z \in \partial\mathbb{E}$ and $\mathbb{E}$ has no interior points, then $\mathbb{E}$ is an interval. In this situation $\partial\mathbb{E}$ is traversed twice (once in each direction) as w traverses the unit circle. The tangent vectors vanish at the endpoints of the interval. The bound (2.18) also holds in this situation. This completes the proof of (2.13).

We turn to the proof of (2.14), and first assume that $\mathbb{W}(A)$ is contained in the interior of $\mathbb{E}$. In order to derive a matrix-valued analog of the expression (2.16), we observe that

$$\mathcal{F}(\widetilde{f})(A) = \frac{1}{2\pi i}\int_{\partial\mathbb{D}} \widetilde{f}(w)(\psi(w)I - A)^{-1}\psi'(w)dw.$$

Moreover, since the matrix A and its transpose, A^T, have the same eigenvalues, it follows that

$$\frac{1}{2\pi i}\int_{\partial\mathbb{D}} g(w)(\psi(w)I - A^T)^{-1}\psi'(w)dw = \overline{\widetilde{f}(0)}I.$$

Adding the conjugate of the latter expression to the former yields

$$\mathcal{F}_+(\widetilde{f})(A) = \mathcal{F}(\widetilde{f})(A) + \widetilde{f}(0)I = \frac{1}{\pi}\int_{\partial\mathbb{D}} \widetilde{f}(w)K(w)\,|dw|,$$

where

$$K(w) := \frac{1}{2i}\left(\psi'(w)(\psi(w)I - A)^{-1}\frac{dw}{|dw|} - \overline{\psi'(w)}(\overline{\psi(w)}I - A^*)^{-1}\frac{\overline{dw}}{|dw|}\right)$$

and A^* denotes the conjugate transpose of A.

We would like to show that the Hermitian matrix $K(w)$ is positive definite for all $|w| = 1$. This is equivalent to establishing that the matrix

$$G(w) := (\psi(w)I - A)K(w)(\overline{\psi(w)}I - A^*)$$

is positive definite. With ζ and $\alpha(\zeta)$ defined as above, we have

$$G(w) = \frac{1}{2i}\left(e^{i\alpha(\zeta)}(\overline{\zeta}I - A^*) - e^{-i\alpha(\zeta)}(\zeta I - A)\right).$$

Let $v \in \mathbb{C}^n$ be a unit vector. Then $a := v^*Av$ lives in $\mathbb{W}(A)$ and we obtain

$$v^*G(w)v = \frac{1}{2i}\left(e^{i\alpha(\zeta)}(\overline{\zeta} - \overline{a}) - e^{-i\alpha(\zeta)}(\zeta - a)\right) > 0,$$

where the inequality follows similarly as (2.17). Application of the Cauchy-Schwarz inequality now yields

$$\begin{aligned}
||\mathcal{F}_+(\widetilde{f})(A)|| &\le \sup_{||u||,||v||\le 1} \frac{1}{\pi}\int_{\partial\mathbb{D}} |\widetilde{f}(w)v^*K(w)u|\,|dw| \\
&\le ||\widetilde{f}||_{L_\infty(\mathbb{D})} \sup_{||u||,||v||\le 1}\left[\frac{1}{\pi}\int_{\partial\mathbb{D}} u^*K(w)u\,|dw|\frac{1}{\pi}\int_{\partial\mathbb{D}} v^*K(w')v\,|dw'|\right]^{1/2} \\
&= ||\widetilde{f}||_{L_\infty(\mathbb{D})} \sup_{||u||,||v||\le 1}\left[(\mathcal{F}_+(1)(A)u,u)(\mathcal{F}_+(1)(A)v,v)\right]^{1/2} = 2\,||\widetilde{f}||_{L_\infty(\mathbb{D})},
\end{aligned}$$

in agreement with (2.14).

We turn to the situation when $\mathbb{W}(A)$ is not contained in the interior of $\mathbb{E}$. If $\mathbb{E}$ has no interior point, then both $\mathbb{E}$ and $\mathbb{W}(A)$ are intervals. In particular, the matrix A is normal, and it follows that

$$||\mathcal{F}_+(\widetilde{f})(A)|| \le \max_{z\in\sigma(A)} |\mathcal{F}_+(\widetilde{f})(z)| \le \max_{z\in\mathbb{E}} |\mathcal{F}_+(\widetilde{f})(z)| = \|\mathcal{F}_+(\widetilde{f})\|_{L_\infty(\mathbb{E})} \le 2\,||\widetilde{f}||_{L_\infty(\mathbb{D})}$$

by (2.13), where $\sigma(A)$ denotes the spectrum of A. We therefore may assume that $\mathbb{E}$ has an interior point z_0, and let $\epsilon \in (0,1)$. Then the field of values of the matrix $A_\epsilon := \epsilon z_0 I + (1-\epsilon)A$, given by $\mathbb{W}(A_\epsilon) = z_0 + (1-\epsilon)(\mathbb{W}(A) - z_0)$, is in the interior of $\mathbb{E}$. Hence, for all $\epsilon \in (0,1)$,

$$||\mathcal{F}_+(\widetilde{f})(A_\epsilon)|| \le 2\,||\widetilde{f}||_{L_\infty(\mathbb{D})}.$$

The bound (2.14) follows by letting $\epsilon \searrow 0$. □

REMARK 2.2. *Let* $\widetilde{f}(w) := w^n$ *for* $n = 1, 2, \ldots$. *Then* $\mathcal{F}_+(\widetilde{f}) = \mathcal{F}(\widetilde{f}) = F_n^{\mathbb{E}}$, $n = 1, 2, \ldots$, *are Faber polynomials for* $\mathbb{E}$. *We obtain from (2.14) that*

$$\|F_n^{\mathbb{E}}(A)\| \leq 2, \qquad n = 1, 2, \ldots . \tag{2.19}$$

This inequality recently has been shown in [7, Theorem 1] in a similar manner. We note that a for convex set $\mathbb{E}$ *it follows from [56, Theorem 2] that* $||F_n^{\mathbb{E}}||_{L_\infty(\mathbb{E})} \leq 2$.

Consider the solution of the linear system of equations $Ax = b$ *by the GMRES iterative method, described, e.g., in [71, Chapter 6] and [72]. Let* x_0 *be an initial approximate solution and let, for* $k = 1, 2, \ldots$, x_k *denote the kth iterate generated by the method. Define the associated residual errors* $r_k := b - Ax_k$, $k = 0, 1, \ldots$. *The inequality (2.19) can be used to derive bounds for the* r_k *in terms of the field of values of* A *when* $0 \notin \mathbb{E}$. *Specifically, one can show that*

$$\frac{||r_k||}{||r_0||} \leq \min\left\{\frac{2}{1 - |\phi(0)|^{-k-1}}, 2 + |\phi(0)|^{-1}\right\} |\phi(0)|^{-k};$$

see [7] for details. This inequality improves bounds reported in [8, 31, 33] and [44, Chapter 3]. □

The following result, which is a consequence of Theorem 2.1, is applied in the remainder of this paper.

COROLLARY 2.3. *Assume that* $\mathbb{E}$ *and* $\mathbb{W}(A)$ *satisfy the conditions of Theorem 2.1. Let* $g \in \mathbb{A}(\mathbb{E})$, $\widetilde{g} := \mathcal{F}^{-1}(g)$, $\widetilde{r} \in \mathbb{A}(\mathbb{D})$, *and*

$$r(z) := \mathcal{F}_+(\widetilde{r})(z) + (\mathcal{F} - \mathcal{F}_+)(\widetilde{g})(z) = \mathcal{F}_+(\widetilde{r})(z) - \widetilde{g}(0).$$

Then

$$||g(A) - r(A)|| \leq 2\, ||\widetilde{g} - \widetilde{r}||_{L_\infty(\mathbb{D})}.$$

3. Polynomial approximation via the Arnoldi process. In this section, we assume that the matrix $A \in \mathbb{C}^{n \times n}$ in (1.1) is large and sparse, and that the vector $b \in \mathbb{C}^n$ is of unit length. The Arnoldi process applied to A with initial vector b yields, after m steps, the decomposition

$$AV_m = V_m H_m + h_m e_m^T, \tag{3.1}$$

where $V_m = [v_1, v_2, \ldots, v_m] \in \mathbb{C}^{n \times m}$ and $h_m \in \mathbb{C}^n$ satisfy $v_1 = b$, $V_m^* V_m = I$, and $V_m^* h_m = 0$. Throughout this paper e_j denotes the jth axis vector of appropriate dimension. The matrix $H_m \in \mathbb{C}^{m \times m}$ is of upper Hessenberg form, and

$$\text{range}\,(V_m) = \mathcal{K}_m(A, b),$$

where

$$\mathcal{K}_m(A, b) := \text{span}\,\{b, Ab, \ldots, A^{m-1}b\}$$

is a Krylov subspace. In particular, $v_j \in \mathcal{K}_j(A, b)$, i.e., there is a polynomial $p_{j-1} \in \mathbb{P}_{j-1}$, such that

$$v_j = p_{j-1}(A)b, \qquad j = 1, 2, \ldots, m; \tag{3.2}$$

see, e.g., [42, Chapter 9] for details on the Arnoldi process. We refer to (3.1) as an Arnoldi decomposition.

We remark that if $h_m = 0$, then range (V_m) is an invariant subspace, and it follows that $f(A)b = V_m f(H_m)e_1$. We therefore henceforth will assume that $h_m \neq 0$. When A is Hermitian, the Arnoldi process simplifies to the Hermitian Lanczos process and the matrix H_m in the decomposition (3.1) is Hermitian and tridiagonal.

The columns v_j of V_m are generated for increasing values of j; the computation of v_j requires the evaluation of $j-1$ matrix-vector product with A and orthogonalization against all the already computed columns $v_1, v_2, \ldots, v_{j-1}$. One would like to keep m in Arnoldi decompositions (3.1) used in applications fairly small, because the computational effort and storage required to generate the Arnoldi decomposition increases with m. Moreover, instead of computing $f(A)b$, we will evaluate $f(H_m)e_1$, and the computational effort required for the latter typically grows rapidly with m.

We note for future reference that since $H_m = V_m^* A V_m$,

$$\mathbb{W}(H_m) = \left\{ \frac{(Ax, x)}{(x, x)} : x = V_m y,\ y \in \mathbb{C}^m \setminus \{0\} \right\} \subset \mathbb{W}(A) \subset \mathbb{E}. \tag{3.3}$$

One easily verifies by induction that for any $p \in \mathbb{P}_{m-1}$, we have

$$p(A)b = V_m p(H_m) V_m^* b = V_m p(H_m) e_1; \tag{3.4}$$

see, e.g., [26, 70]. This motivates the use of the polynomial approximation

$$V_m f(H_m) e_1 \approx f(A)b, \tag{3.5}$$

where in view of (3.2), the left-hand side can be written as $p(A)b$ for some $p \in \mathbb{P}_{m-1}$. The Crouzeix bound (2.11) with the constant (2.12) yields an immediate bound for the approximation error in (3.5) in terms best polynomial approximation of f on $\mathbb{E}$; cf. (2.6) with $q \equiv 1$.

Proposition 3.1. *Let $\mathbb{E}$ be a convex compact set, such that $\mathbb{W}(\mathbb{A}) \subset \mathbb{E}$. Assume that $f \in \mathbb{A}(\mathbb{E})$. Then, for all $m \geq 1$,*

$$\|f(A)b - V_m f(H_m)e_1\| \leq 23\, \eta_{m-1}^1(f, \mathbb{E}). \tag{3.6}$$

Proof. It follows from (3.4) that for any $p \in \mathbb{P}_{m-1}$, we have

$$\begin{aligned} \|f(A)b - V_m f(H_m)e_1\| &= \|(f-p)(A)b - V_m (f-p)(H_m)e_1\| \\ &\leq \|(f-p)(A)\| + \|(f-p)(H_m)\|, \end{aligned}$$

where we have used that $\|b\| = 1$. The inequality (3.6) now is a consequence of (2.11), (2.12), (3.3), and the fact that $\mathbb{W}(\mathbb{A}) \subset \mathbb{E}$. □

The following theorem connects polynomial approximation of $f(A)b$ with polynomial approximation of $\mathcal{F}^{-1}(f)$ on $\mathbb{D}$.

Theorem 3.2. *Let $\mathbb{E}$ be a convex and compact set, such that $\mathbb{W}(A) \subset \mathbb{E}$. Assume that $f \in \mathbb{A}(\mathbb{E})$. Then, for all $m \geq 1$,*

$$\|f(A)b - V_m f(H_m)e_1\| \leq 4\, \eta_{m-1}^1(\mathcal{F}^{-1}(f), \mathbb{D}). \tag{3.7}$$

More generally, for $g(z) = q_1(z) + q_2(z) f(z)$ with $q_1 \in \mathbb{P}_{m+s-1}$, $q_2 \in \mathbb{P}_s$, there holds

$$\|g(A)b - V_{m+s} g(H_{m+s})e_1\| \leq 4\, \|q_2(A)b\|\, \eta_{m-1}^1(\mathcal{F}^{-1}(f), \mathbb{D}). \tag{3.8}$$

Finally, we have the bounds

$$|f_m| \le \eta^1_{m-1}(\mathcal{F}^{-1}(f), \mathbb{D}) \le \sum_{j=m}^{\infty} |f_j| \tag{3.9}$$

in terms of the coefficients in the Faber series expansion of f,

$$f_j := \frac{1}{2\pi i}\int_{\partial\mathbb{D}} f(\psi(w))\frac{dw}{w^{j+1}}, \qquad j = 0, 1, \ldots\ . \tag{3.10}$$

Proof. Since $\|b\| = 1$, the bound (3.7) follows from (3.8) by taking $q_1(z) = 0$, $s = 0$, and $q_2(z) = 1$. In order to show the latter bound, we choose an extremal polynomial $\widetilde{p} \in \mathbb{P}_{m-1}$, such that $||\mathcal{F}^{-1}(f) - \widetilde{p}||_{L_\infty(\mathbb{D})} = \eta^1_{m-1}(\mathcal{F}^{-1}(f), \mathbb{D})$. Similarly as in Corollary 2.3, we define

$$\widetilde{f}(w) := \mathcal{F}^{-1}(f)(w), \qquad p(z) := \mathcal{F}_+(\widetilde{p})(z) - \widetilde{f}(0), \qquad q(z) := q_1(z) + q_2(z)p(z).$$

Then $q \in \mathbb{P}_{m+s-1}$ and, according to (3.4),

$$\begin{aligned} &\|g(A)b - V_{m+s}g(H_{m+s})e_1\| = \|(g-q)(A)b - V_{m+s}(g-q)(H_{m+s})e_1\| \\ &\le \|(f-p)((A)q_2(A)b\| + \|V_{m+s}(f-p)(H_{m+s})q_2(H_{m+s})e_1\| \\ &\le \|f(A) - p(A)\|\,\|q_2(A)b\| + \|f(H_{m+s}) - p(H_{m+s})\|\,\|q_2(H_{m+s})e_1\|, \end{aligned}$$

where, again by (3.4), we have $\|q_2(H_{m+s})e_1\| = \|q_2(A)b\|$. Corollary 2.3 yields

$$\|f(A) - \hat{p}(A)\| \le 2\,\|\widetilde{f} - \widetilde{p}\|_{L_\infty(\mathbb{D})} = 2\,\eta^1_{m-1}(\mathcal{F}^{-1}(f), \mathbb{D}),$$

and (3.3) combined with Corollary 2.3 gives

$$\|f(H_m) - \hat{p}(H_m)\| \le 2\,||\widetilde{f} - \widetilde{p}||_{L_\infty(\mathbb{D})} = 2\,\eta^1_{m-1}(\mathcal{F}^{-1}(f), \mathbb{D}).$$

This establishes the inequalities (3.7) and (3.8). Comparing (3.10) to (2.1), we observe that f_m is the mth coefficient in the Taylor expansion of $F^{-1}(f)$ at the origin. Hence, with the extremal $\widetilde{p} \in \mathbb{P}_{m-1}$ as above,

$$f_m = \frac{1}{2\pi i}\int_{|w|=1} \frac{\mathcal{F}^{-1}(f)(w)}{w^{m+1}}dw = \frac{1}{2\pi i}\int_{|w|=1} \frac{\mathcal{F}^{-1}(f)(w) - \widetilde{p}(w)}{w^{m+1}}dw,$$

the absolute value being bounded above by $||\mathcal{F}^{-1}(f) - \widetilde{p}||_{L_\infty(\mathbb{D})} = \eta^1_{m-1}(\mathcal{F}^{-1}(f), \mathbb{D})$. Finally, if $\sum_{j=m}^{\infty} |f_j| < \infty$, then

$$f(z) = \sum_{j=0}^{\infty} f_j F_j(z), \qquad \mathcal{F}^{-1}(f)(w) = \sum_{j=0}^{\infty} f_j w^j,$$

because both series are absolutely convergent. The upper bound (3.9) now follows by approximating $\mathcal{F}^{-1}(f)$ by its Taylor sum $\sum_{j=0}^{m-1} f_j w^j$. □

Remark 3.3. *Let $f \in \mathbb{A}(\mathbb{E}_\rho)$ for some $\rho > 1$ and change the path of integration from $\partial\mathbb{D}$ to $\{w \in \mathbb{C} : |w| = \rho\}$ in the definition of the Faber coefficients (3.10). Then one easily verifies that*

$$\sum_{j=m}^{\infty} |f_j| \le ||f||_{L_\infty(\mathbb{E}_\rho)} \frac{\rho^{-m}}{(1-\rho^{-1})},$$

where the factor ρ^{-m} *corresponds to the classical rate of best polynomial approximation on* $\mathbb{E}$ *of functions in* $\mathbb{A}(\mathbb{E}_\rho)$*; see, e.g., [83, Theorem IV.5]. In particular, the lower and upper bounds in (3.9) differ only by a term that decreases geometrically, or even faster when* f *is an entire function, such as the exponential function; see below.*

In order to compare Proposition 3.1 and Theorem 3.2, one may either use (2.9), or apply the bounds

$$|f_m| \le \eta^1_{m-1}(f,\mathbb{E}) \le 2\sum_{j=m}^{\infty} |f_j|. \tag{3.11}$$

The lower bound can be shown similarly as in the proof of (3.9), and the upper bound by using a partial Faber sum, as well as the fact that $\|F_j^{\mathbb{E}}\|_{L_\infty(\mathbb{E})} \le 2$*; see Remark 2.2.* □

REMARK 3.4. *Let us compare Theorem 3.2 with bounds reported by Druskin and Knizhnerman [26, 27, 28, 54]. Knizhnerman [54, Theorem 1] shows that there are positive constants* C *and* α*, which depend on the shape of* $\mathbb{E} := \mathbb{W}(A)$*, such that*

$$||f(A)b - V_m f(H_m)e_1|| \le C\sum_{k=m}^{\infty} |f_k| k^\alpha.$$

When $\mathbb{E} = [-1,1]$*, the Faber polynomials* $F_j^{\mathbb{E}}$*, for* $j \ge 1$*, are twice the Chebyshev polynomials of the first kind; cf. Example 1.2. The observation that in this case*

$$||f(A)b - V_m f(H_m)e_1|| \le 4\sum_{k=m}^{\infty} |f_k|$$

is (at least implicitly) included in [27, Proof of Theorem 1]. For the exponential function and $\mathbb{E} = [-1,1]$ *further improvements and more explicit bounds are derived in [26, 28] by using the fact that the Faber coefficients are explicitly known in terms of Bessel functions.* □

REMARK 3.5. *Hochbruck and Lubich [51] derive error bounds for analytic functions* f *in terms of integral formulas and exploit the latter to obtain bounds for the error in polynomial approximations of* $\exp(\tau A)b$*,* $\tau > 0$*, determined by the Arnoldi process with* $\mathbb{W}(A)$ *contained in various convex compact sets* $\mathbb{E}$*.*

Let $\mathbb{E}$ *be a convex compact set containing* $\mathbb{W}(A)$ *in its interior, and let* $\mathbb{E}'$ *be a bounded set that contains* $\mathbb{E}$*. The boundary* Γ *of* $\mathbb{E}'$ *is assumed to be a piecewise smooth Jordan curve. Let the function* f *be analytic in the interior of* Γ *and continuous on the closure of* $\mathbb{E}'$*. Then Hochbruck and Lubich [51, Lemma 1] show that*

$$||f(A)b - V_m f(H_m)e_1|| \le C \cdot \min_{p\in\mathbb{P}_{m-1}} \frac{1}{2\pi}\int_\Gamma |f(z)-p(z)|\frac{|dz|}{|\phi(z)|^m} \tag{3.12}$$

for

$$C := \frac{\mathit{length}(\partial\mathbb{E})}{\mathit{dist}(\partial\mathbb{E},\mathbb{W}(A))\ \mathit{dist}(\Gamma,\mathbb{W}(A))}.$$

We would like to compare this bound to Theorem 3.2 and will use the inequalities

$$\frac{\mathit{dist}(z,\mathbb{E})\,|\phi'(z)|}{1-|\phi(z)|^{-1}} \le 1 + |\phi(z)| \le 2|\phi(z)|, \qquad z \in \mathbb{C}\setminus\mathbb{E}, \tag{3.13}$$

which follow from [80, Theorem 3.1] and its proof. Let $p \in \mathbb{P}_{m-1}$ minimize the right-hand side of (3.12). Then, for all $j \geq m$,

$$f_j = \frac{1}{2\pi i}\int_\Gamma f(z)\frac{\phi'(z)\,dz}{\phi(z)^{j+1}} = \frac{1}{2\pi i}\int_\Gamma [f(z)-p(z)]\frac{\phi'(z)\,dz}{\phi(z)^{j+1}},$$

and, using (3.13), the right-hand side of (3.7) can be bounded according to

$$\begin{aligned} 4\sum_{j=m}^{\infty}|f_j| &\leq \frac{2}{\pi}\int_\Gamma |f(z)-p(z)|\,\frac{|\phi'(z)|\,|dz|}{|\phi(z)|^{m+1}(1-1/|\phi(z)|)} \\ &\leq \frac{4}{\pi}\int_\Gamma |f(z)-p(z)|\,\frac{|dz|}{dist(z,\mathbb{E})\,|\phi(z)|^m} \\ &\leq C'\cdot\frac{1}{2\pi}\int_\Gamma |f(z)-p(z)|\,\frac{|dz|}{|\phi(z)|^m}, \qquad C' := \frac{8}{dist(\Gamma,\mathbb{E})}, \end{aligned}$$

where we note that the bound in each step may be quite crude. Nevertheless, the ratio C'/C can be made arbitrarily small by choosing $\partial\mathbb{E}$ close to $\mathbb{W}(A)$. We would expect the bound of Theorem 3.2 to be most accurate in this situation. Thus, Theorem 3.2 may provide useful bounds when (3.12) does not. The next section discusses applications of Theorem 3.2. □

4. Approximation of the exponential function. The following result provides bounds for the Faber coefficients of the exponential function, and thereby also for the approximation error achieved by the Arnoldi process, via Theorem 3.2, and for best polynomial approximation, via (3.11). The bounds of Corollary 4.1 below only depend on the logarithmic capacity of $\mathbb{E}$ for large values of m, whereas Corollary 4.2 discusses the dependence on the outer angle at the right-most boundary point of $\mathbb{E}$ for small values of m.

Corollary 4.1. *Let $f(z) = \exp(\tau z)$, where $\tau > 0$ is an arbitrary parameter. Let the set $\mathbb{E}$ be compact, convex, and symmetric with respect to the real axis, with capacity $c = cap\,(\mathbb{E}) = \psi'(\infty)$. Then, for $r \geq 1$, the Faber coefficients satisfy*

$$|f_m| \leq \frac{e^{\tau\psi(r)}}{r^m} \tag{4.1}$$

and

$$\eta^1_{m-1}(\mathcal{F}^{-1}(f),\mathbb{D}) \leq \frac{e^{\tau\psi(r)}}{r^m(1-r^{-1})}. \tag{4.2}$$

The minimum of the right-hand side of (4.1) is attained for $r = 1$ if $\tau\psi'(1) \geq m$, and otherwise at the unique solution of the equation $r\psi'(r) = m/\tau$. In particular, if $m \geq 2c\tau$, then

$$|f_m| \leq \frac{7}{2}e^{\tau\psi(1)}\frac{(\tau c)^m}{m!}, \qquad \eta^1_{m-1}(\mathcal{F}^{-1}(f),\mathbb{D}) \leq 7e^{\tau\psi(1)}\frac{(\tau c)^m}{m!}. \tag{4.3}$$

Proof. According to (3.10), we obtain for any $r \geq 1$ the simple upper bound

$$|f_m| \leq \frac{1}{2\pi}\int_{-\pi}^{\pi}\left|\frac{e^{\tau\psi(re^{it})}}{r^m}\right|dt = r^{-m}\exp\left(\tau\max_{t\in[-\pi,\pi]}\Re\psi(re^{it})\right),$$

where by symmetry of $\mathbb{E}$, the maximum of the right-hand side is attained at the right-most point of $\mathbb{E}$, i.e., at $t = 0$. This shows (4.1). The bound (4.2) is obtained by using the fact that

$$\eta^1_{m-1}(\mathcal{F}^{-1}(f), \mathbb{D}) \le \sum_{j=m}^{\infty} |f_j|.$$

It remains to be shown that the equation $\tau r \psi'(r) = m$ has at most one solution. First notice that $(1, +\infty) \ni r \mapsto \psi'(r)$ is real by symmetry of $\mathbb{E}$, does not change sign, and tends to $c > 0$ as $r \to \infty$. Hence, $\psi'(r)$ is strictly positive in $(1, +\infty)$. It follows from convexity that $\Re(1 + \frac{w\psi''(w)}{\psi'(w)}) > 0$ for all $|w| > 1$. Therefore, $r \mapsto r\psi'(r)$ is increasing.

The choice $r = \frac{m}{ct}$ in (4.1) and (4.2) leads to (4.3). However, there is a missing factor $1/\sqrt{m}$, which requires refinement of our bounds. We first show that for $|w| \ge 1$,

$$|\psi(w) - \psi(r) - c(w - r)| \le c \left| \frac{1}{w} - \frac{1}{r} \right|. \tag{4.4}$$

By convexity of $\mathbb{E}$, we have the inequality

$$|\psi'(w) - c| \le \frac{c}{|w|^2}, \qquad |w| \ge 1,$$

due to Grötzsch and Golusin; see [56, Section 2]. Hence,

$$|\psi(w) - \psi(r) - c(w - r)| \le \int_R^w |\psi'(\zeta) - c| \, |d\zeta| \le c \int_R^w \frac{|d\zeta|}{|\zeta|^2},$$

and we obtain the inequality (4.4) by taking as path of integration the circular arc $[0, 1] \ni t \mapsto \frac{1}{1/r + t(1/w - 1/r)}$, staying outside of the unit disk. Notice that (4.4) for $w = 1$ implies that

$$\psi(r) - \psi(1) \le c(r - 1) + c\left(1 - \frac{1}{r}\right) = c\left(r - \frac{1}{r}\right). \tag{4.5}$$

Applying our inequality for $w = re^{it}$, $|t| \le \pi$, and using again the symmetry of $\mathbb{E}$, we obtain

$$\begin{aligned}
\Re\psi(re^{it}) &\le \Re\left(\psi(r) + c(w - r)\right) + c\left|\frac{1}{w} - \frac{1}{r}\right| \\
&= \psi(r) + cr(\cos(t) - 1) + \frac{c}{r}|e^{-it} - 1| \\
&= \psi(r) - 2cr\sin^2(|t/2|) + \frac{2c}{r}\sin(|t/2|) \\
&\le \psi(r) - 2\frac{cr}{\pi^2}t^2 + \frac{c}{r}|t| = \psi(r) - 2\frac{cr}{\pi^2}\left(|t| - \frac{\pi^2}{4r^2}\right)^2 + \frac{c\pi^2}{8r^3},
\end{aligned}$$

which yields

$$\begin{aligned}
|f_m| &\le \frac{1}{2\pi r^m} \int_{-\pi}^{\pi} \exp\left(\tau\psi(r) - 2\frac{\tau c r}{\pi^2}\left(|t| - \frac{\pi^2}{4r^2}\right)^2 + \frac{\tau c \pi^2}{8r^3}\right) dt \\
&\le \frac{e^{\psi(r) + \frac{\tau c\pi^2}{8r^3}}}{\pi r^m} \int_0^{\pi} \exp\left(-2\frac{\tau c r}{\pi^2}\left(|t| - \frac{\pi^2}{4r^2}\right)^2\right) dt \\
&= \sqrt{\pi}\frac{\exp(\tau\psi(r) + \frac{\tau c\pi^2}{8r^3})}{r^m\sqrt{2\tau c r}} \le \pi\frac{\exp(\tau\psi(1) + \tau c r - \frac{\tau c}{r} + \frac{\tau c \pi^2}{8r^3})}{r^m\sqrt{2\pi\tau c r}},
\end{aligned}$$

where, in the last inequality, we applied (4.5). Now the choice $r = \frac{m}{\tau c} \geq 2$ gives $-\frac{\tau c}{r} + \frac{\tau c \pi^2}{8r^3} \leq 0$ and

$$|f_m| \leq \pi \exp(\tau\psi(1)) \frac{(\tau c)^m}{\sqrt{2m}(m/e)^m} \leq \frac{7}{2} \exp(\tau\psi(1)) \frac{(\tau c)^m}{m!},$$

as claimed in the first inequality of (4.3). The second inequality of (4.3) follows by observing that $1/(1-1/r) \leq 2$. □

Let $\mathbb{E} = \{z \in \mathbb{C} : |z - z_0| \leq c\}$ for some constants $z_0 \in \mathbb{R}$ and $c > 0$. Then $\psi(w) = cw + z_0$ and the Faber coefficients are given by $f_m = e^{\tau z_0} \frac{(\tau c)^m}{m!}$. Hence, the bound (4.3) for $|f_m|$ is sharp up to the factor $\frac{7}{2} \exp(\tau c)$, independently of m, whereas the bound (4.1) with optimal parameter $r = \frac{m}{\tau c} \geq 1$ is sharp up to a factor $\sqrt{2\pi m}$. Further, when $\mathbb{E}$ is an interval on the real or imaginary axis, explicit formulas for the Faber coefficients f_m can be given in terms of Bessel functions [12, 26]. These formulas show the bound (4.3) for $|f_m|$ to be asymptotically sharp as $m \to \infty$ up to a constant independent of m.

Hochbruck and Lubich [51, Theorems 2,4-6] apply the bound (3.12) to the exponential function when $\mathbb{W}(A) \subset \mathbb{E}$ for i) $\mathbb{E} = [-4c, 0]$ an interval on the negative real axis and $\psi(w) = c(w + 1/w - 2)$, ii) $\mathbb{E} = [-2ic, 2ic]$ an interval on the imaginary axis and $\psi(w) = c(w - 1/w)$, iii) $\mathbb{E}$ a disk, such that $\psi(w) = cw + c_0$, and iv) $\mathbb{E}$ a drop-shaped region, for which $\psi(w) = cw(1 - 1/w)^\alpha$ with $\alpha > 1$. The latter set has an outer angle $\alpha\pi$ at the vertex $\psi(1) = 0$. The bounds for large m given in [51, Theorems 2,4-6] essentially coincide with (4.1) for $r = \frac{m}{\tau c}$, though the absolute constants in [51] are somewhat larger.

When $\mathbb{E}$ is the interval $[-4c, 0]$ or the drop-shaped region, Hochbruck and Lubich [51] also provide upper bounds for the situation when $m \leq 2\tau c$. These sets $\mathbb{E}$ have an outer angle $\alpha\pi > \pi$ at the right-most boundary point, which is the pre-image of $w = 1$ under ψ. Therefore, $\psi'(1) = 0$ and inequality (4.1) indicates that $|f_m|$ may be smaller than $e^{\tau\psi(1)}$ also for $m \leq 2\tau c$. It is possible to give a bound depending only on this outer angle.

COROLLARY 4.2. *Under the assumptions of Corollary 4.1, suppose in addition that $\mathbb{E}$ has an outer angle $\alpha\pi$ with $\alpha > 1$ at its right-most boundary point $\psi(1)$. Then, for $m \leq 2\tau c$, we have*

$$|f_m| \leq \exp\left(\tau\psi(1) - \frac{\alpha - 1}{7} m \left(\frac{m}{3\tau c}\right)^{\frac{1}{\alpha-1}}\right), \tag{4.6}$$

$$\eta^1_{m-1}(\mathcal{F}^{-1}(f), \mathbb{D}) \leq 3 \left(\frac{m}{3\tau c}\right)^{-\frac{1}{\alpha-1}} \exp\left(\tau\psi(1) - \frac{\alpha - 1}{7} m \left(\frac{m}{3\tau c}\right)^{\frac{1}{\alpha-1}}\right). \tag{4.7}$$

Proof. In the first part of the proof we show the improvement of (4.5),

$$\psi(r) - \psi(1) \leq cr \left(1 - \frac{1}{r}\right)^\alpha \left(1 + \frac{1}{r}\right)^{2-\alpha}. \tag{4.8}$$

Introduce the generating function for the Faber polynomials,

$$\frac{r\psi'(r)}{\psi(r) - \psi(1)} = 1 + \sum_{n=1}^{\infty} \frac{F_n(\psi(1))}{r^n},$$

where we use the representation of the Faber polynomials from [66, Lemma 1], here for convex $\mathbb{E}$, for $n \geq 1$,

$$F_n(\psi(1)) = \frac{1}{\pi}\int_{-\pi}^{\pi} e^{ins} d_s \arg\Big(\psi(e^{is}) - \psi(1)\Big), \quad \int_{-\pi}^{\pi}\Big|d_s \arg\Big(\psi(e^{is}) - \psi(1)\Big)\Big| \leq 2\pi.$$

Since the Stieltjes integral has a jump of $\pi\alpha$ at $s = 0$ and elsewhere the argument is increasing, we obtain, by taking out the jump and using the symmetry, that

$$F_n(\psi(1)) = \alpha + \frac{1}{\pi}\int_{0+}^{\pi}(e^{ins} + e^{-ins}) d_s \arg\Big(\psi(e^{is}) - \psi(1)\Big),$$

and, therefore,

$$\begin{aligned}\frac{r\psi'(r)}{\psi(r) - \psi(1)} - 1 - \alpha\frac{1}{r-1} &= \frac{2}{\pi}\int_{0+}^{\pi} \Re\Big(\frac{e^{is}}{r - e^{is}}\Big) d_s \arg\Big(\psi(e^{is}) - \psi(1)\Big) \\ &= \frac{2}{\pi}\int_{0+}^{\pi} \frac{r\cos(s) - 1}{r^2 + 1 - 2r\cos(s)} d_s \arg\Big(\psi(e^{is}) - \psi(1)\Big) \\ &\geq -\frac{1}{r+1}\frac{2}{\pi}\int_{0+}^{\pi} d_s \arg\Big(\psi(e^{is}) - \psi(1)\Big) = -\frac{2-\alpha}{r+1}.\end{aligned}$$

Integrating this inequality from r to ∞ gives (4.8).

Since $\alpha > 1$, we may choose $r \geq 1$, such that

$$\left(r - \frac{1}{r}\right)^{\alpha - 1} = \frac{m}{3\tau c} \leq 1. \tag{4.9}$$

Hence, $r \leq r^* = (1 + \sqrt{5})/2$, and we obtain from (4.8) that

$$\begin{aligned}\tau(\psi(r) - \psi(1)) - m\log(r) &\leq \tau c\left(\frac{r}{r+1}\right)^{\alpha-1}\left(1 - \frac{1}{r}\right)\left(r - \frac{1}{r}\right)^{\alpha-1}(r+1)^{2-\alpha} + m\log\left(\frac{1}{r}\right) \\ &\leq m\left(1 - \frac{1}{r+1}\right)^{\alpha-1}\left(1 - \frac{1}{r}\right)\frac{r^*+1}{3} + m\left(\frac{1}{r} - 1\right).\end{aligned}$$

Since $y \mapsto y^{\alpha-1}$ is concave, we deduce that

$$\begin{aligned}\tau(\psi(r) - \psi(1)) - m\log(r) &\leq m\left(1 - \frac{\alpha-1}{r+1}\right)\left(1 - \frac{1}{r}\right) + m\left(\frac{1}{r} - 1\right) = -m(\alpha-1)\frac{r - \frac{1}{r}}{(r+1)^2} \\ &\leq -m(\alpha-1)\frac{r - \frac{1}{r}}{(r^*+1)^2} \leq -m(\alpha-1)\frac{r - \frac{1}{r}}{7}.\end{aligned}$$

Inserting (4.9) gives (4.6). The bound (4.7) follows by observing that $1/(1 - 1/r) \leq (r^* + 1)/(r - 1/r)$. □

We conclude this section with three further illustrations/extensions of Corollary 4.1.

Example 4.1. If the matrix A has a negative semi-definite real part, then a simple set containing $\mathbb{W}(A)$ is given by

$$\mathbb{E} = (\rho\mathbb{D}) \cap \left\{w \in \mathbb{C} : \Re(w) \leq \lambda_{\max}\left(\frac{A + A^*}{2}\right)\right\},$$

provided that $\rho > 0$ is large enough. For instance, ρ can be chosen to be the norm of A or the numerical radius, $\max\{|z| : z \in \mathbb{W}(A)\}$. Define the angle $\beta \in [0, \pi/2)$ by

$$\cos(\beta) = -\frac{\lambda_{\max}(\frac{A+A^*}{2})}{\rho} \geq 0.$$

In order to apply the bounds (4.3), we only require the value $\psi(1) = -\rho\cos(\beta) = \lambda_{\max}(\frac{A+A^*}{2})$ and the capacity of $\mathbb{E}$, which given by

$$c = \rho\,\frac{\pi}{2\pi - \beta}\,\frac{\sin(\beta)}{\cos(\frac{\beta}{4-2\beta/\pi})}.$$

The latter can be seen by constructing the conformal mapping ϕ, cf. (1.2), which can be expressed as the composition $\phi := \frac{1}{\rho}T_3 \circ T_2 \circ T_1$, where

$$T_1(z) := e^{i\beta/2}\frac{z + e^{-i\beta}}{z + e^{i\beta}},\;\; T_2(z) := z^{\frac{\pi}{2\pi-\beta}},\;\; T_3(z) := \frac{\gamma}{i}\frac{z + \overline{\gamma}}{z - \gamma},\;\; \gamma := \exp\left(i\frac{\beta}{4 - 2\beta/\pi}\right);$$

see, e.g., [47, 60] for discussions on the construction of conformal mappings. □

The symmetry of $\mathbb{E}$ with respect to the real axis is not essential for showing bounds of the form (4.3). A bound valid for nonsymmetric sets $\mathbb{E}$ can be obtained by replacing $\psi(1)$ by $\Re(\psi(1))$ in (4.3). The essential ingredient in the proof is the property that $|f(z)| \leq \exp(\tau\Re(z))$. Similar properties hold for hyperbolic and trigonometric functions.

Example 4.2. Let $f(z) = \sinh(\tau z)$ or $f(z) = \cosh(\tau z)$ with $\tau > 0$. Then $|f(z)| \leq \exp(\tau|\Re(z)|)$ and

$$\max\Big\{2|f_m|, \eta^1_{m-1}(\mathcal{F}^{-1}(f), \mathbb{D})\Big\} \leq 7\Big(e^{\tau\Re\psi(1)} + e^{-\tau\Re\psi(-1)}\Big)\frac{(\tau c)^m}{m!}, \qquad m \geq 2c\tau.$$

In order to show this bound, it suffices to slightly modify the proof of (4.3): Let $-\pi \leq \theta_1 \leq \theta_2 \leq \pi$ be such that $\Re\psi(re^{it}) \geq 0$ if and only if $t \in [\theta_1, \theta_2]$. In this interval, we obtain as above

$$|\Re\psi(re^{it})| \leq \Re\psi(1) + c\left(r - \frac{1}{r}\right) + cr(\cos(t) - 1) + \frac{c}{r}\left|e^{it} - 1\right|,$$

as required for our conclusion. For $t \in [\theta_2, 2\pi + \theta_1]$ we have from (4.4), with r replaced by $-r$, that

$$\begin{aligned}|\Re\psi(re^{it})| &= -\Re(\psi(re^{it})) \leq -\psi(-r) - cr(\cos(t) + 1) + \frac{c}{r}\left|e^{it} - 1\right| \\ &\leq -\Re\psi(-1) + c\left(r - \frac{1}{r}\right) + cr(\cos(t - \pi) - 1) + \frac{c}{r}2\sin\left(\frac{|t - \pi|}{2}\right),\end{aligned}$$

and the second part of the integral can be bounded as before. In particular, replacing A by iA yields for $f(z) = \sin(\tau z)$ or $f(z) = \cos(\tau z)$ that

$$\max\Big\{2|f_m|, \eta^1_{m-1}(\mathcal{F}^{-1}(f), \mathbb{D})\Big\} \leq 14\,\exp\Big(\tau\max_{z\in\mathbb{E}}|\Im(z)|\Big)\frac{(\tau c)^m}{m!}, \qquad m \geq 2\tau c.$$

□

Example 4.3. Define, for integers $\ell \geq 1$, the functions

$$\phi_\ell(z) := \frac{1}{z^\ell}\Big(e^z - \sum_{j=0}^{\ell-1} \frac{z^j}{j!}\Big) = \int_0^1 e^{z(1-u)} \frac{u^{\ell-1}}{(\ell-1)!} du,$$

which are of interest in connection with exponential integrators.

Let $f(z) = \phi_\ell(\tau z)$ for some $\tau > 0$ and fixed integer $\ell \geq 1$, and let $\mathbb{E}$ be a subset of the left half-plane. Then, for $m \geq 2\tau c$,

$$\begin{aligned} |f_m| &\leq \int_0^1 \frac{1}{2\pi} \int_{|w|=\frac{m}{(1-u)\tau c}} \left| \frac{e^{\tau(1-u)\psi(w)}}{w^{m+1}} \right| |dw| \frac{u^{\ell-1}}{(\ell-1)!} du \\ &\leq \frac{7}{2} \frac{(\tau c)^m}{m!} \int_0^1 e^{\tau(1-u)\Re(\psi(1)}(1-u)^m \frac{u^{\ell-1}}{(\ell-1)!} du \\ &\leq \frac{7}{2} \frac{(\tau c)^m}{m!} \int_0^1 (1-u)^m \frac{u^{\ell-1}}{(\ell-1)!} du = \frac{7}{2} \frac{(\tau c)^m}{(m+\ell)!}. \end{aligned}$$

The same upper bound holds for $\eta^1_{m-1}(\mathcal{F}^{-1}(f), \mathbb{D})/2$. □

Druskin et al. [30] provide a nice discussion on rational approximation of the matrix exponential for symmetric matrices. We consider rational approximation in the following sections.

5. Rational approximation and the rational Arnoldi process. We consider the approximation of f on $\mathbb{E}$ by a rational function r, determined by approximating $\widetilde{f} := \mathcal{F}^{-1}(f)$ on $\mathbb{D}$ by a rational function $\widetilde{r} = \widetilde{p}/\widetilde{q}$, where $\widetilde{p}, \widetilde{q} \in \mathbb{P}_m$, and the polynomial $\widetilde{q}$ is monic with zeros $w_j = \psi(z_j) \notin \mathbb{D}$. Let $\widetilde{r}$ be such a rational function. Then

$$r(z) = \mathcal{F}_+(\widetilde{p}/\widetilde{q})(z) - \widetilde{f}(0) \tag{5.1}$$

is a rational function of the form $r = p/q$ with $p, q \in \mathbb{P}_m$. The monic polynomial q has the zeros $z_j = \phi(w_j)$ of the same multiplicity as the corresponding zeros w_j of $\widetilde{q}$, i.e., $\widetilde{q}$ and q are related as in (2.8). It follows from Corollary 2.3 that

$$\|f(A) - r(A)\| \leq 2\, \|\widetilde{f} - \widetilde{p}/\widetilde{q}\|_{L_\infty(\mathbb{D})}. \tag{5.2}$$

We therefore are interested in results on the approximation of $\widetilde{f}$ on $\widetilde{D}$ by rational functions with prescribed poles. The case when f is a Markov function is discussed in Section 6, where we also consider the choice of suitable poles w_j. In this section, we are concerned with the evaluation of $r(A)b$, either for a given rational function $\widetilde{r}$, or by using the rational Arnoldi process. The latter approach determines the numerator p for a user-specified denominator q.

Here and in the remainder of this paper, we assume the set $\mathbb{E}$ to be symmetric with respect to the real axis, and that f satisfies $f(\overline{z}) = \overline{f(z)}$. The Faber pre-image $\widetilde{f}$ of f also has the latter property. Therefore, it suffices to consider rational approximants $\widetilde{r}$ with real or complex conjugate poles and residues. In order to fix ideas, suppose that $\widetilde{r}$ has m simple finite poles,

$$\widetilde{r}(w) = \widetilde{r}(\infty) + \sum_{j=1}^m \frac{c_j}{w_j - w}.$$

Then by (5.1) and (2.7), we obtain

$$r(z) = r(\infty) + \sum_{j=1}^{m} \frac{\psi'(w_j)c_j}{z_j - z}, \qquad r(\infty) = \widetilde{r}(\infty) + \widetilde{r}(0) - \widetilde{f}(0), \qquad z_j = \psi(w_j).$$

The evaluation of $r(A)b$ requires the solution of m shifted linear systems of equations

$$(z_j I - A)x_j = b. \tag{5.3}$$

The approximation of $f(A)b$ by $r(A)b$ is meaningful when A is a large sparse matrix, such that the shifted systems (5.3) can be solved efficiently by a sparse direct method, but solution by Krylov subspace methods or by Schur reduction to triangular form are impractical. For example, discretization of the two-dimensional Laplace operator on a square, using the standard 5-point finite difference stencil, gives rise to such a matrix.

REMARK 5.1. *The matrices in (5.3) generate the same Krylov subspaces $\mathcal{K}_j(A, b)$, $j = 1, 2, \ldots, m$. This makes it possible to solve the m linear systems of equations simultaneously by an iterate method that uses the same Krylov subspace; see, e.g., [37, 82]. However, solving these shifted systems in this manner, e.g., by the GMRES iterative method, implies that we determine a polynomial approximant of f. It may be possible to compute more accurate polynomial approximants of f for the same computational effort by using the approach described in Section 3.* □

There are situations when it suffices to solve fewer than m shifted systems of equations. For instance, when all poles are distinct and the poles and coefficients c_j appear in complex conjugate pairs, say, $z_{m+1-j} = \overline{z_j}$, and $c_{m+1-j} = \overline{c_j}$, we obtain

$$r(A)b = r(\infty) + 2\Re\left(\sum_{j=1}^{m/2} \psi'(w_j)c_j(z_j I - A)^{-1}b\right),$$

where we have taken into account that A and b have real entries. Thus, only $m/2$ shifted systems of equations have to be solved. In case of multiple poles, one has to solve several linear systems of equations with the same matrix $z_j I - A$, but with different right-hand sides. The number of LU-factorizations required is the number of distinct poles with nonnegative imaginary part.

For an efficient implementation of our approach, we need to compute the inverse Faber image of f and the Faber image of a rational function. This poses no difficulty if ψ is known in closed form or is a Schwarz-Christoffel mapping; see, e.g., Driscoll and Trefethen [25] or Henrici [47, Chapter 5] for discussions of the latter; software for computing Schwarz-Christoffel mappings is made available by Driscoll [24].

The above approach requires knowledge of a suitable rational approximant $\widetilde{r}$, not only its poles. The rational Krylov method, introduced by Ruhe [68, 69], only requires the poles to be specified, and gives an error, which similarly to (3.7), is bounded by $4\eta_m^{\widetilde{q}}(\widetilde{f}, \mathbb{D})$; see Theorem 5.2 below. Thus, in view of (5.2), the rational Krylov method is quasi-optimal (up to a factor 2). For the sake of completeness, we shortly describe this method. The introduction of an artificial pole $z_{m+1} := \infty$ leads to a slight simplifications compared to the presentation in [68, 69]. Given complex poles $z_1, z_2, \ldots, z_m$, including the case of a pole $z_j = \infty$, we let q be the product of the linear factors corresponding to finite poles. Let $z_0 \in \mathbb{C}$ be sufficiently far away from the poles z_j, but otherwise arbitrary. We compute by an Arnoldi-type process an orthonormal basis $\{v_j\}_{j=1}^{m+1}$ of the rational Krylov subspace $q(A)^{-1}\mathrm{span}\{b, Ab, \ldots, A^m b\}$ in the following manner. Let $v_1 = b$ and determine v_{j+1} by orthogonalizing

$$(z_j - z_0)(z_j I - A)^{-1}(A - z_0 I)v_j$$

against the available vectors $v_1, v_2, \ldots, v_j$, followed by normalization. If $z_j = \infty$, then we orthogonalize $(A - z_0 I)v_j$. The vectors v_j satisfy for suitable scalars $h_{k,j}$ the recursion formula

$$h_{j+1,j}v_{j+1} = (z_j - z_0)(z_j I - A)^{-1}(A - z_0 I)v_j - h_{1,j}v_1 - \ldots - h_{j,j}v_j, \quad j = 1, 2, \ldots, m,$$

with $v_1 = b$.

Let $V_{m+1} = [v_1, v_2, \ldots, v_{m+1}]$ and define the upper Hessenberg matrix $H_{m+1} = [h_{j,k}]_{j,k=1,\ldots,m+1}$. The formula for the projection $A_{m+1} := V_{m+1}^* A V_{m+1}$ is more complicated than for the standard Arnoldi process. Introduce

$$D_{m+1} = \mathrm{diag}\left[\frac{1}{z_1 - z_0}, \ldots, \frac{1}{z_{m+1} - z_0}\right].$$

Then, for $j \leq m$,

$$\begin{aligned}(A - z_0 I)v_j = (A - z_0 I)V_{m+1}e_j &= \left(I - \frac{1}{z_j - z_0}(A - z_0 I)\right)\sum_{k=1}^{j+1} h_{k,j}v_k \\ &= (V_{m+1}H_{m+1} - (A - z_0 I)V_{m+1}H_{m+1}D_{m+1})e_j. \end{aligned} \tag{5.4}$$

When $j = m + 1$, we have to include the additional term

$$h_{m+2,m+1}\left(I - \frac{1}{z_{m+1} - z_0}(A - z_0 I)\right)v_{m+2} = h_{m+2,m+1}v_{m+2}$$

in the right-hand side of (5.4), where the equality follows from the choice $z_{m+1} = \infty$. We obtain from (5.4) that

$$(A - z_0 I)V_{m+1}(I + H_{m+1}D_{m+1}) = V_{m+1}H_{m+1} + h_{m+2,m+1}v_{m+2}e_{m+1}^T.$$

In view of that $V_{m+1}^* v_{m+2} = 0$ and $V_{m+1}^* V_{m+1} = I$, this leads to the formula

$$A_{m+1} = V_{m+1}^* A V_{m+1} = z_0 I + H_{m+1}(I + H_{m+1}D_{m+1})^{-1}. \tag{5.5}$$

Notice that the choices $z_0 = 0$ and $z_1 = \ldots = z_m = \infty$ yield the standard Arnoldi process, with $A_{m+1} = H_{m+1}$ determined by (3.1) with m replaced by $m+1$. A bound analogous to (3.7) for the standard Arnoldi method also holds for the rational Arnoldi method.

THEOREM 5.2. *Let $\mathbb{E}$ be a compact convex set, such that $\mathbb{W}(A) \subset \mathbb{E}$. Assume that $f \in \mathbb{A}(\mathbb{E})$, and let $z_1, z_2, \ldots, z_m \notin \mathbb{E}$, $z_{m+1} = \infty$. Then, for all $m \geq 1$,*

$$\|f(A)b - V_{m+1}f(A_{m+1})e_1\| \leq 4\,\eta_m^{\widetilde{q}}(\mathcal{F}^{-1}(f), \mathbb{D}) \tag{5.6}$$

with $\widetilde{q}$ as in (2.8). More generally, for $g(z) = q_1(z) + q_2(z)f(z)$ with $q_1 \in \mathbb{P}_{m+s}$ and $q_2 \in \mathbb{P}_s$, there holds with $z_{m+1} = z_{m+2} = \ldots = z_{m+s+1} = \infty$,

$$\|g(A)b - V_{m+s+1}g(A_{m+s+1})e_1\| \leq 4\,\|q_2(A)b\|\,\eta_m^{\widetilde{q}}(\mathcal{F}^{-1}(f), \mathbb{D}). \tag{5.7}$$

Proof. Since $v_j \in q(A)^{-1}\mathcal{K}_{m+1}(A,b)$, there exists $p_j \in \mathbb{P}_m$, such that

$$v_j = q(A)^{-1}p_j(A)b, \qquad j = 1,2,\ldots,m+1.$$

With $v_1, v_2, \ldots, v_{m+1}$ being a basis of the rational Krylov subspace, the polynomials $p_0, p_1, \ldots, p_m$ form a basis of $\mathbb{P}_m$. Since $v_1 = b$, we have $p_0 = q$. We now show that

$$e_j = \widetilde{v}_j := q(A_{m+1})^{-1}p_{j-1}(A_{m+1})e_1, \qquad j = 1,2,\ldots,m+1. \tag{5.8}$$

This is trivially true for $j = 1$, and the general case follows by induction. By definition of v_j and p_{j-1}, the vectors $\widetilde{v}_j$ satisfy

$$h_{j+1,j}\widetilde{v}_{j+1} = (z_j - z_0)(z_jI - A_{m+1})^{-1}(A_{m+1} - z_0I)\widetilde{v}_j - h_{1,j}\widetilde{v}_1 - \ldots - h_{j,j}\widetilde{v}_j.$$

It only remains to observe that, by (5.5),

$$\begin{aligned}
&(z_j - z_0)(z_jI - A_{m+1})^{-1}(A_{m+1} - z_0I)e_j \\
&= (z_j - z_0)(A_{m+1} - z_0I)\left((z_j - z_0)I - (A_{m+1} - z_0I)\right)^{-1} e_j \\
&= (z_j - z_0)H_{m+1}\left((z_j - z_0)I + H_{m+1}((z_j - z_0)D_{m+1} - I)\right)^{-1} e_j \\
&= H_{m+1}e_j,
\end{aligned}$$

since $((z_j - z_0)D_{m+1} - I)e_j = 0$. This shows (5.8). Any $p \in \mathbb{P}_m$ may be written as $p = c_1p_0 + \ldots + c_{m+1}p_m$. Therefore, the Arnoldi approximation is exact for $g = p/q$,

$$\begin{aligned}
q(A)^{-1}p(A)b &= \sum_{j=1}^{m+1} c_jq(A)^{-1}p_{j-1}(A)b = \sum_{j=1}^{m+1} c_jv_j = V_{m+1}\sum_{j=1}^{m+1} c_je_j \\
&= \sum_{j=1}^{m+1} c_jq(A_{m+1})^{-1}p_{j-1}(A_{m+1})b = V_{m+1}q(A_{m+1})^{-1}p(A_{m+1})e_1.
\end{aligned}$$

As a consequence, we obtain similarly as in the proof of Theorem 3.2, that

$$\begin{aligned}
\|f(A)b - V_{m+1}f(A_{m+1})e_1\| &\le \min_{p\in\mathbb{P}_m} \left\|(f - \frac{p}{q})(A)b\right\| + \left\|(f - \frac{p}{q})(A_{m+1})e_1\right\| \\
&\le 4 \min_{\widetilde{p}\in\mathbb{P}_m} \left\|\widetilde{f} - \frac{\widetilde{p}}{\widetilde{q}}\right\|_{L_\infty(\mathbb{D})} = 4\eta_m^{\widetilde{q}}(\widetilde{f},\mathbb{D}),
\end{aligned}$$

since $\mathbb{W}(A_{m+1}) \subset \mathbb{W}(A) \subset \mathbb{E}$. This yields (5.6). The bound (5.7) can be shown in a similar way as in Theorem 3.2. We therefore omit the details. □

A bound similar to (5.6) for the case when the matrix A is symmetric recently has been shown independently by Druskin et al. [30]. Concerning the implementation of the rational Arnoldi process, we have to solve shifted linear systems of equations

$$(z_jI - A)x_j = v_j.$$

Complex conjugation of z_j does not correspond to complex conjugation of v_j. In situations when it is feasible to compute LU-factorizations of the matrices $z_jI - A$, only factorizations for distinct finite nonnegative z_j have to be determined. In particular, we just need to compute one LU-factorization of A if $z_{2j-1} = 0$ and $z_{2j} = \infty$, $j = 1,2,\ldots$. This kind of rational approximant is discussed in [29, 55].

The derivation of an analogue of Theorem 5.2 for the approximation of entire functions, such as the exponential function, and the application of (5.2) to such functions, is beyond the scope of this paper; see, e.g., Ganelius [39] for a discussion on the rate of convergence of rational approximants of such functions.

6. Rational approximation of Markov functions. This section applies the error bounds of Theorems 3.2 and 5.2 for the standard and rational Arnoldi processes, respectively, to Markov functions f, given by (1.4), and to simple modifications of Markov functions, such as those discussed in Examples 1.3 and 1.4.

As far as we know, only asymptotic results are known for rational interpolants with free poles, see, e.g., [77, Section 6] or [11], and a posteriori error bounds are available for rational approximants obtained by balanced truncation and AAK theory; see, e.g., [10]. The present section derives explicit sharp upper and lower bounds for the error of best approximation $\eta_m^{\widetilde{q}}(\mathcal{F}^{-1}(f),\mathbb{D})$ for rational approximants with prescribed denominator $\widetilde{q}$ of degree at most m. These bounds are believed to be new. We also construct nearly optimal approximants $\widetilde{r}$, which can be used for explicit evaluation as explained in the previous section. Since, by (5.1) and (2.7),

$$\eta_m^{q}(f,\mathbb{E}) \le 2\,\eta_m^{\widetilde{q}}(\mathcal{F}^{-1}(f),\mathbb{D}),$$

we obtain explicit upper bounds for rational approximants of Markov functions on $\mathbb{E}$.

The following theorem establishes our main results for Markov functions. It discusses properties of the Blaschke product

$$B(w) = \frac{w^m \widetilde{q}(1/w)}{\widetilde{q}(w)} = \prod_{j=1}^{m} \frac{1-w_j w}{w-w_j}, \tag{6.1}$$

whose poles w_j are assumed to be real or occur in complex conjugate pairs and satisfy $1 < |w_j| \le \infty$. It follows that $B(w)$ has real coefficients when expressed in terms of positive and negative powers of w and, moreover, $B(1/w) = 1/B(w)$.

THEOREM 6.1. *Let the set $\mathbb{E}$ be compact, convex, and symmetric with respect to the real axis. Let f be a Markov function (1.4), and assume that $-\infty \le \alpha < \beta < \gamma := \min\{\Re(z) : z \in \mathbb{E}\}$.*
(a) *Then $\widetilde{f} = \mathcal{F}^{-1}(f)$ is a Markov function,*

$$\widetilde{f}(w) = \int_\alpha^\beta \frac{\phi'(x)\,d\mu(x)}{w-\phi(x)} =: \int \frac{d\widetilde{\mu}(x)}{w-x}. \tag{6.2}$$

(b) *Let $R = P/\widetilde{q}$ with $P \in \mathbb{P}_{m-1}$ be the rational interpolant of $\widetilde{f}$ with prescribed poles w_j at the reflected points $1/\overline{w_j}$ for $j = 1, 2, \ldots, m$ (counting multiplicities), and let*

$$\widetilde{r}(w) = R(w) + B(w)\Big(\frac{\widetilde{f}(1)-R(1)}{2B(1)} + \frac{\widetilde{f}(-1)-R(-1)}{2B(-1)}\Big).$$

Then $\widetilde{r} \in \mathbb{P}_m/\widetilde{q}$ and

$$\eta_m^{\widetilde{q}}(\mathcal{F}^{-1}(f),\mathbb{D}) \le \|\widetilde{f}-\widetilde{r}\|_{L_\infty(\mathbb{D})} \le \int_\alpha^\beta \frac{1}{|B(\phi(x))|}\frac{|\phi'(x)|\,d\mu(x)}{|\phi(x)|^2-1} \tag{6.3}$$

$$\le \frac{\|f\|_{L_\infty(\mathbb{E})}}{|\phi(\beta)|} \max_{y\in\phi([\alpha,\beta])} \frac{1}{|B(y)|}. \tag{6.4}$$

(c) *If, in addition, the poles $w_j \in (\phi(\alpha),\phi(\beta))$ have even multiplicity, then*

$$\eta_m^{\widetilde{q}}(\mathcal{F}^{-1}(f),\mathbb{D}) \ge \int_\alpha^\beta \frac{1}{|B(\phi(x))|}\frac{|\phi'(x)|\,d\mu(x)}{|\phi(x)|^2-1/|B(\phi(x))|}, \tag{6.5}$$

and for the approximant $\widetilde{r}$ of part (b), we also have the a posteriori bound

$$\|\widetilde{f} - \widetilde{r}\|_{L_\infty(\mathbb{D})} = |\widetilde{f}(-1) - \widetilde{r}(-1)| = \int_\alpha^\beta \frac{1}{|B(\phi(x))|} \frac{|\phi'(x)|\, d\mu(x)}{|\phi(x)|^2 - 1}.$$

We comment on the bounds before showing the theorem.

REMARK 6.2. *For $x \in [\alpha, \beta]$, we have $|B(\phi(x))| > 1$ and*

$$\frac{1}{|\phi(x)|^2 - 1} \le \frac{1}{1 - |\phi(\beta)|^{-2}} \frac{1}{|\phi(x)|^2 - 1/|B(\phi(x))|},$$

independently of the choice of poles w_j and of their number m. Therefore, for all poles w_j and m, the lower bound (6.5) is bounded below by the factor $1 - |\phi(\beta)|^{-2}$ times the upper bound (6.3). The upper and lower bounds give a quite precise idea of the accuracy of the best approximation of $\widetilde{f}$ in $\mathbb{P}_m/\widetilde{q}$ on the unit circle. Concerning part (b), we also should mention that our quasi-optimal approximant $\widetilde{r}$ is obtained by a simple modification of the interpolant R, where R is known to be the best approximant of $\widetilde{f}$ in $\mathbb{P}_{m-1}/\widetilde{q}$ with respect to the 2–norm on the unit circle. □

REMARK 6.3. *Consider the polynomial case when $w_1 = \ldots = w_m = \infty$ and, hence, $B(w) = w^m$ and $\widetilde{q}(w) = 1$. From Theorem 6.1(a) and its proof, we see that the Faber coefficients of f satisfy*

$$|f_j| = \int_\alpha^\beta \frac{|\phi'(x)|\, d\mu(x)}{|\phi(x)|^{j+1}}.$$

It is not difficult to verify that for this special case, the two bounds of Theorem 6.1(b) and (c) take the form

$$\sum_{j=0}^{\infty} |f_{m+j(m+1)}| \le \eta^1_{m-1}(\widetilde{f}, \mathbb{D}) \le \sum_{j=0}^{\infty} |f_{m+2j}| \le \frac{\|f\|_{L_\infty(\mathbb{E})}}{|\phi(\beta)|^m}. \tag{6.6}$$

These bounds improve on the inequalities (3.9). Moreover, the quantity ρ in Remark 3.3 is at most $|\phi(\beta)|$; ρ has to be chosen smaller if f is not continuous at β. Hence, (6.6) also is an improvement of the bound furnished by Remark 3.3. □

The proof of Theorem 6.1 is divided into three parts. Following the proof, we discuss some configurations of poles obtained by minimizing the bound (6.4). This enables us to compare our approach with the shifted Arnoldi process, see, e.g., [81], and the use of Talbot quadratures rules discussed in [45].

Proof. Theorem 6.1(a): The Faber coefficients of the Markov function f of (1.4) satisfy, by the Fubini theorem and the Cauchy formula,

$$\begin{aligned} f_j &= \frac{1}{2\pi i} \int_{\partial\mathbb{E}} \frac{\phi'(z)dz}{\phi(z)^{j+1}} \int_\alpha^\beta \frac{d\mu(x)}{z - x} \\ &= \int_\alpha^\beta d\mu(x)\, \frac{1}{2\pi i} \int_{\partial\mathbb{E}} \frac{\phi'(z)dz}{\phi(z)^{j+1}(z - x)} = -\int_\alpha^\beta \frac{\phi'(x)\, d\mu(x)}{\phi(x)^{j+1}}. \end{aligned}$$

The last identity is obtained by deforming the path of integration $\partial\mathbb{E}$ in $\overline{\mathbb{C}} \setminus \mathbb{E}$ in order to obtain a circle around x with mathematically negative orientation. Recalling that $\mathcal{F}^{-1}(f)(w) = f_0 + f_1 w + \ldots$, we conclude that $\widetilde{f} = \mathcal{F}^{-1}(f)$ is of the form (6.2). By the symmetry of $\mathbb{E}$, it follows that the function $|\phi(x)| = -\phi(x)$ is decreasing for

$x \in (-\infty, \gamma)$ and that $\phi'(x)$ is positive for $x \in (-\infty, \gamma)$. Thus, $\widetilde{\mu}$ is a positive measure and $\widetilde{f}$ is a Markov function.

Theorem 6.1(b): We first establish a well-known integral formula for rational interpolants with prescribed poles of Markov functions; see, e.g., [83, Theorem VIII.2]. The numerator P is the interpolation polynomial of $\widetilde{q}\widetilde{f}$ at the points $1/w_1, 1/w_2, \ldots, 1/w_m$, and, therefore,

$$\begin{aligned}
\widetilde{f}(w) - R(w) &= \frac{(w-1/w_1)\ldots(w-1/w_m)}{\widetilde{q}(w)}[1/w_1, 1/w_2, \ldots, 1/w_m, w](\widetilde{q}\widetilde{f}) \\
&= \frac{(w-1/w_1)\ldots(w-1/w_m)}{\widetilde{q}(w)}[1/w_1, 1/w_2, \ldots, 1/w_m, w]_t \int \widetilde{q}(x)\frac{d\widetilde{\mu}(x)}{t-x} \\
&= \frac{(w-1/w_1)\ldots(w-1/w_m)}{\widetilde{q}(w)} \int \frac{\widetilde{q}(x)}{(x-1/w_1)\ldots(x-1/w_m)}\frac{d\widetilde{\mu}(x)}{w-x} \\
&= B(w)\int \frac{1}{B(x)}\frac{d\widetilde{\mu}(x)}{w-x}.
\end{aligned}$$

In particular, we find for the modified approximant that

$$\begin{aligned}
\frac{\widetilde{f}(w)-\widetilde{r}(w)}{B(w)} &= \int \frac{1}{B(x)}\Big(\frac{1}{w-x} - \frac{1/2}{1-x} - \frac{1/2}{-1-x}\Big)d\widetilde{\mu}(x) \\
&= \int \frac{1}{B(x)}\Big(\frac{1}{w-x} + \frac{x}{x^2-1}\Big)d\widetilde{\mu}(x) = -\int \frac{1}{B(x)}\frac{1-wx}{w-x}\frac{d\widetilde{\mu}(x)}{x^2-1}.
\end{aligned}$$

Taking into account that Blaschke factors are of unit modulus on the unit circle, and proceeding similarly as in the proof of (6.2), gives the upper bound (6.3).

It follows from (3.13) that

$$\int_\alpha^\beta \frac{1}{|B(\phi(x))|}\frac{|\phi'(x)|\,d\mu(x)}{|\phi(x)|^2-1} \le \int_\alpha^\beta \frac{1}{|\phi(x)|\,|B(\phi(x))|}\frac{d\mu(x)}{\mathrm{dist}(x,\mathbb{E})}.$$

The distance is achieved for $\gamma \in \mathbb{E}$ for all $x \in [\alpha, \beta]$, and $x \mapsto 1/|\phi(x)|$ is increasing in $[\alpha, \beta]$. This shows (6.4).

Theorem 6.1(c): We first recall that $\phi'(x) > 0$ for $x \in [\alpha, \beta]$. Moreover, by assumption, $\widetilde{B}(w) := w^2 B(w)$ is a rational function with real coefficients, having all its $m+2$ roots in the open unit disk, and its poles in the interval $\phi([\alpha, \beta])$ have even multiplicity. Hence, $\widetilde{B}$ is of constant sign on $\phi([\alpha, \beta])$. Thus, $\epsilon\widetilde{B}(\phi(x)) > 0$ for $x \in [\alpha, \beta]$ with $\epsilon^2 = 1$. Let

$$\mathbb{D}_{m+2} = \{x_0, x_1, \ldots, x_{m+1}\} := \{w \in \mathbb{C} : \epsilon\widetilde{B}(w) = 1\},$$

with the x_j being distinct points on the unit circle, ordered according to increasing argument. Theorem 6.1(c) will follow by showing that

$$\eta_m^{\widetilde{q}}(\widetilde{f}, \mathbb{D}_{m+2}) = \min_{p \in \mathbb{P}_m} \|\widetilde{f} - p/\widetilde{q}\|_{L_\infty(\mathbb{D}_{m+2})} = \int_\alpha^\beta \frac{\phi'(x)\,d\mu(x)}{\epsilon\widetilde{B}(\phi(x)) - 1} =: \delta. \tag{6.7}$$

Let $\widetilde{R} \in \mathbb{P}_{m+1}/\widetilde{q}$ be the rational interpolant of $\widetilde{f}$ at the points in $\mathbb{D}_{m+2}$, and with prescribed denominator $\widetilde{q}$. Denote the coefficient for w^{m+1} of $\widetilde{R}(w)\widetilde{q}(w)$ by a. Since $wB(w)\widetilde{q}(w) - \widetilde{q}(0)w^{m+1} \in \mathbb{P}_m$, we obtain that

$$\frac{p^*(w)}{\widetilde{q}(w)} := \widetilde{R}(w) - \frac{wB(w)}{\widetilde{q}(0)}a \in \mathbb{P}_m/\widetilde{q}.$$

Elementary computations give, for $j = 0, 1, \dots, m+1$, that

$$\begin{aligned}\delta_j := \widetilde{f}(x_j) - \frac{p^*(x_j)}{\widetilde{q}(x_j)} &= \widetilde{R}(x_j) - \frac{p^*(x_j)}{\widetilde{q}(x_j)} = \frac{x_j B(x_j)}{\widetilde{q}(0)} a = \frac{\widetilde{B}(x_j)}{x_j \widetilde{q}(0)} a \\ &= \frac{1}{\epsilon x_j \widetilde{q}(0)} [x_0, x_1, \dots, x_{m+1}](\widetilde{f}\widetilde{q}) \\ &= -\frac{1}{\epsilon x_j \widetilde{q}(0)} \int \frac{\widetilde{q}(x) d\widetilde{\mu}(x)}{(x - x_0) \dots (x - x_{m+1})} \\ &= -\frac{1}{x_j} \int \frac{d\widetilde{\mu}(x)}{\epsilon \widetilde{B}(x) - 1} = -\frac{\delta}{x_j},\end{aligned}$$

where $[x_0, x_1, \dots, x_{m+1}]$ denotes the divided-difference operator defined by the nodes $x_0, x_1, \dots, x_{m+1}$. Thus, $\|\widetilde{f} - p^*/\widetilde{q}\|_{L_\infty(\mathbb{D}_{m+2})} = \delta$. In order to show (6.7), we will establish that

$$\eta_m^{\widetilde{q}}(\widetilde{f}, \mathbb{D}_{m+2}) = \|\widetilde{f} - p^*/\widetilde{q}\|_{L_\infty(\mathbb{D}_{m+2})},$$

i.e., that $p^*/\widetilde{q}$ is the best approximant with respect to the uniform norm on $\mathbb{D}_{m+2}$. According to the Kolmogorov Theorem (see, e.g., [75, Satz 6.2]) it suffices to show the existence of positive coefficients $\alpha_0, \alpha_1, \dots, \alpha_{m+1}$, such that, for all $p \in \mathbb{P}_m$,

$$\sum_{j=0}^{m+1} \alpha_j \overline{\delta_j} \frac{p(x_j)}{\widetilde{q}(x_j)} = 0.$$

We notice that, since $p \in \mathbb{P}_m$,

$$0 = [x_0, x_1, \dots, x_{m+1}] p = \sum_{j=0}^{m+1} \frac{p(x_j)}{\prod_{\ell \neq j} (x_j - x_\ell)} = \sum_{j=0}^{m+1} \frac{p(x_j)}{\widetilde{q}(x_j)} \frac{\widetilde{q}(0)}{\widetilde{B}'(x_j)},$$

and $\overline{\delta_j} = -x_j \delta$. Hence, for our assertion it is sufficient to show that $\epsilon x_j \widetilde{B}'(x_j) > 0$ for all j. Since all poles of the Blaschke product $\widetilde{B}$ are outside of the closed unit disk, we may write $\epsilon \widetilde{B}(e^{it}) = e^{i\alpha(t)}$, with $\alpha(t)$ a real-valued and strictly increasing function. By definition, $x_j = e^{it_j}$ with $e^{i\alpha(t_j)} = \epsilon \widetilde{B}(x_j) = 1$, and, therefore,

$$\epsilon x_j \widetilde{B}'(x_j) = \frac{1}{i} \frac{d}{dt} e^{i\alpha(t)}|_{t=t_j} = \alpha'(t_j) e^{i\alpha(t_j)} = \alpha'(t_j) > 0.$$

We conclude that (6.7) holds, and this implies (6.5). Finally, the a posteriori estimate for $\widetilde{r}$ of Theorem 6.1(b) is an immediate consequence of the error formula given in the proof of Theorem 6.1(b) and of the fact that also B does not change sign in the interval $(\phi(\alpha), \phi(\beta))$. □

6.1. Rational approximation with one or two multiple poles. The upper bound (6.4) of Theorem 6.1 suggests that we should choose the poles w_j in the Blaschke product (6.1), such that $|B|$ is as large as possible on the interval $[\phi(\alpha), \phi(\beta)]$. This subsection discusses the cases of a single pole $w_1 \in \mathbb{R} \cup \{\infty\}$ of multiplicity m, and of two distinct poles, $w_1, w_2 \in \mathbb{R} \cup \{\infty\}$, each of multiplicity $m/2$. The following results are immediate consequences of (6.4) and of the monotonicity of B. We therefore omit the proofs.

Corollary 6.4. **(a)** *Let the conditions of Theorem 6.1(b) hold and let the prescribed denominator be given by* $\widetilde{q}(w) = (w - w_1)^m$, $w_1 \in \mathbb{R} \setminus [-1, 1]$. *Then*

$$\eta_m^{\widetilde{q}}(\mathcal{F}^{-1}(f), \mathbb{D}) \le \|\widetilde{f} - \widetilde{r}\|_{L_\infty(\mathbb{D})} \le \frac{\|f\|_{L_\infty(\mathbb{E})}}{|\phi(\beta)|} \max\left\{ \left| \frac{w_1 - \phi(\alpha)}{1 - w_1\phi(\alpha)} \right|, \left| \frac{w_1 - \phi(\beta)}{1 - w_1\phi(\beta)} \right| \right\}^m.$$

The right-hand side is minimal and decreases with the geometric rate $|y_{opt}|^{-m}$ *for the pole*

$$w_1 = \frac{1 + \phi(\alpha) y_{opt}}{\phi(\alpha) + y_{opt}}, \qquad y_{opt} = -\frac{1}{\kappa} - \sqrt{\frac{1}{\kappa^2} - 1}, \qquad \kappa = \frac{\phi(\beta) - \phi(\alpha)}{\phi(\beta)\phi(\alpha) - 1} \in (0, 1).$$

(b) *For* m *even and the poles* $w_1 = \phi(\alpha)$ *and* $w_2 = \phi(\beta)$, *each of multiplicity* $m/2$, *we obtain the same geometric rate of decrease of the error bound,*

$$\eta_m^{\widetilde{q}}(\mathcal{F}^{-1}(f), \mathbb{D}) \le \|\widetilde{f} - \widetilde{r}\|_{L_\infty(\mathbb{D})} \le \frac{\|f\|_{L_\infty(\mathbb{E})}}{|\phi(\beta)|} |y_{opt}|^{-m}.$$

For the polynomial case, i.e., when we have the single pole $w_1 = \infty$, Corollary 6.4(a) yields the same bound as Remark 6.3.

The advantage of using one finite pole of (high) multiplicity m, compared to the use of m simple poles, is that only one LU-factorization, of $\psi(w_1)I - A$, has to be computed. This holds for the rational Arnoldi process with pole $\psi(w_1)$, or, equivalently, for the standard Arnoldi process applied to the matrix $(\psi(w_1)I - A)^{-1}$ and vector b, as well as for the evaluation of $r(A)b$ with r defined by (5.1). In the latter case, $r(A)$ is a polynomial in $(\psi(w_1)I - A)^{-1}$, and therefore $r(A)b$ can be evaluated efficiently by a Horner scheme.

When $(\alpha, \beta] = (-\infty, 0]$, the choice of the poles

$$z_1 = \psi(w_1) = \alpha = -\infty, \qquad z_2 = \psi(w_2) = \beta = 0, \tag{6.8}$$

each of multiplicity $m/2$, has been discussed by Druskin and Knizhnerman [29] for symmetric positive definite matrices A and $\mathbb{E} = [\lambda_{\min}, \lambda_{\max}]$. Here $\phi(z) = \zeta + \sqrt{\zeta^2 - 1}$ and

$$\zeta = \frac{2z - \lambda_{\min} - \lambda_{\max}}{\lambda_{\max} - \lambda_{\min}}, \qquad \kappa = \frac{1}{|\phi(0)|} = \frac{\sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} - 1}{\sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} + 1}, \qquad \frac{1}{|y_{opt}|} = \frac{\sqrt[4]{\frac{\lambda_{\max}}{\lambda_{\min}}} - 1}{\sqrt[4]{\frac{\lambda_{\max}}{\lambda_{\min}}} + 1}.$$

Thus, replacing the standard Arnoldi process by the rational Arnoldi process, with either one optimally allocated multiple pole, or with two multiple poles at the endpoints of the support of the Markov function, the factor $\sqrt{\lambda_{\min}/\lambda_{\max}}$ is replaced by its square root in the convergence bound,

$$\frac{1}{|\phi(0)|^m} \approx \exp\Big(-2m\sqrt{\frac{\lambda_{\min}}{\lambda_{\max}}}\Big), \qquad \frac{1}{|y_{opt}|^m} \approx \exp\Big(-2m\sqrt[4]{\frac{\lambda_{\min}}{\lambda_{\max}}}\Big).$$

Since the matrix A is symmetric positive definite, the quotient $\lambda_{\max}/\lambda_{\min}$ is its condition number. The error bounds are seen to decrease with the condition number.

The use of one distinct pole when the matrix A is nonsymmtric has recently been considered by Knizhnerman and Simoncini [55], but their error bounds include nonexplicit constants.

REMARK 6.5. *The rate of convergence with two distinct multiple poles can be increased by allocating the finite pole more carefully than (6.8). Let, similarly as above, m be even and the multiplicities of w_1 and w_2 be $m/2$. Choose $w_1 = \phi(\alpha) = \infty$ and let d be the unique solution in the interval* $(0, 1/|\phi(\beta)|)$ *of*

$$-\sqrt{\frac{2d^2}{1+d^4}} = \frac{1/\phi(\beta)+d}{1+d/\phi(\beta)}.$$

The pole $w_2 := -(d^2+1)/(2d)$, which is slightly smaller than the choice of (6.8), yields the rate of convergence d^m, which is faster than the geometric rate of convergence achieved with the poles (6.8). □

6.2. Rational approximation with quasi-optimal poles. This subsection discusses the choice of poles $w_1, w_2, \ldots, w_m$ of the Blaschke product B, defined by (6.1), with the aim of making $|B|$ as large as possible on the interval $[\phi(\alpha), \phi(\beta)]$ and, thereby, obtaining a small bound (6.4) for the approximation error. Since $B(w) = 1/B(1/w)$, we, equivalently, may choose the poles to make $|B|$ as small as possible on the interval $[1/\phi(\beta), 1/\phi(\alpha)]$. This kind of minimization problem has received considerable attention in complex approximation theory; see, e.g., [6, 36]. From [73, Theorem VIII.3.1], we obtain, for any Blaschke product of the form (6.1), that

$$\|B\|_{L_\infty([1/\phi(\beta),1/\phi(\alpha)])} \geq R([\alpha,\beta],\mathbb{E})^{-m} = \exp\Big(-\frac{m}{\mathrm{cap}\,(\mathbb{E},\mathbb{F})}\Big), \tag{6.9}$$

where $\mathrm{cap}\,(\mathbb{E},\mathbb{F})$ denotes the logarithmic capacity of a two-dimensional condenser with plates $\mathbb{E}$ and $\mathbb{F}$; see, e.g., [73, eq. (VIII.3.9)]. Theorem 6.6 below shows that the minimal Blaschke product achieves this bound within a factor 2. Indeed, by the work of Zolotarev, minimizing Blaschke products are explicitly known and can be expressed in terms of conformal mappings for doubly connected domains or with Jacobi elliptic functions; see [1]. We discuss the construction of minimal Blaschke products in the proof of Theorem 6.6.

Before stating our main result of this subsection, we introduce some notation for doubly connected domains. For disjoint closed sets $\mathbb{E}, \mathbb{F} \subset \mathbb{C}$ with simply connected complements, there is a conformal invariant $R = R(\mathbb{E},\mathbb{F}) = R(\mathbb{F},\mathbb{E}) > 1$, occurring already in (6.9), and a conformal bijective map

$$\chi_{\mathbb{E},\mathbb{F}} : \{\zeta \in \mathbb{C} : 1 < |\zeta| < R\} \mapsto \overline{\mathbb{C}} \setminus (\mathbb{E} \cup \mathbb{F})$$

with boundary behavior

$$\chi_{\mathbb{E},\mathbb{F}}(\{|\zeta| = 1\}) = \partial\mathbb{E}, \quad \chi_{\mathbb{E},\mathbb{F}}(\{|\zeta| = R\}) = \partial\mathbb{F}.$$

This map is uniquely determined by a suitable normalization. For instance, when $\mathbb{E}$ is a real interval, we may fix $\chi_{\mathbb{E},\mathbb{F}}(1)$ to be the right endpoint of this interval. We note that

$$\chi_{[\alpha,\beta],\mathbb{E}}(\zeta) = \psi(\chi_{[\phi(\alpha),\phi(\beta)],\mathbb{D}}(\zeta)),$$

where $\chi_{[\phi(\alpha),\phi(\beta)],\mathbb{D}}$ can be expressed in terms of Jacobi elliptic functions; see [1, §49, Example 3] or the proof of Theorem 6.6 below. In particular,

$$R([\alpha,\beta],\mathbb{E}) = R([\phi(\alpha),\phi(\beta)],\mathbb{D}) = R([\frac{1}{\kappa},+\infty),\mathbb{D}) = e^{\nu(\kappa)},$$

$$\text{where}\quad \kappa := \frac{\phi(\alpha)-\phi(\beta)}{1-\phi(\alpha)\phi(\beta)} \in (0,1), \quad \nu(\kappa) := \frac{\pi}{2}\frac{K'(\kappa)}{K(\kappa)} \tag{6.10}$$
$$\text{and}\quad K(\kappa) = \int_0^1 \frac{dt}{\sqrt{(1-t^2)(1-\kappa^2 t^2)}}, \quad K'(\kappa) = K(\sqrt{1-\kappa^2}).$$

The following result is based on (6.4).

THEOREM 6.6. *Let the conditions of Theorem 6.1 hold, and define the poles*

$$w_j = \chi_{[\phi(\alpha),\phi(\beta)],\mathbb{D}}\left(\exp\left(2\pi i \frac{2j-1}{4m}\right)\right) \in [\phi(\alpha),\phi(\beta)], \quad j=1,2,\ldots,m, \tag{6.11}$$

where we use the normalization $\chi_{[\phi(\alpha),\phi(\beta)],\mathbb{D}}(1) = \phi(\beta)$. *Then*

$$\eta_m^{\widetilde{q}}(\mathcal{F}^{-1}(f),\mathbb{D}) \le \|\widetilde{f}-\widetilde{r}\|_{L_\infty(\mathbb{D})} \le \frac{2\,\|f\|_{L_\infty(\mathbb{E})}}{|\phi(\beta)|}\, R([\alpha,\beta],\mathbb{E})^{-m},$$

where $\widetilde{q}(w) = (w-w_1)(w-w_2)\ldots(w-w_m)$. *Furthermore,*

$$R([\alpha,\beta],\mathbb{E})^{-m} \le \exp\left(-m\frac{\pi^2/4}{\log(4/\sqrt{1-\kappa^2})}\right). \tag{6.12}$$

Proof. We will show that the Blaschke product (6.1) determined by the poles (6.11) satisfies

$$\|B\|_{L_\infty([1/\phi(\beta),1/\phi(\alpha)])} = \|1/B\|_{L_\infty([\phi(\alpha),\phi(\beta)])} \le 2\,R([\alpha,\beta],\mathbb{E})^{-m}, \tag{6.13}$$

and that $1/B$ is a minimal Blaschke product for the interval $[\phi(\alpha),\phi(\beta)]$. The theorem then follows from (6.4).

It is useful to recall the third Zolotarev problem for closed sets $\mathbb{E}_0$ and $\mathbb{F}_0$. It entails solving

$$Z_m(\mathbb{E}_0,\mathbb{F}_0) = \min_{r\in\mathbb{Q}_{m,m}} \|r\|_{L_\infty(\mathbb{E}_0)}\,\|1/r\|_{L_\infty(\mathbb{F}_0)},$$

where $\mathbb{Q}_{m,m}$ denotes the set of rational functions of numerator and denominator degrees at most m. Following Zolotarev, the extremal rational function can be constructed explicitly for the segments

$$\mathbb{E}_0 = [-\sqrt{k},\sqrt{k}], \qquad \mathbb{F}_0 = [1/\sqrt{k},+\infty)\cup(-\infty,-1/\sqrt{k}] = 1/\mathbb{E}_0,$$

with the aid of the Jacobi elliptic function $\mathrm{sn}(u;k)$, $0<k<1$.

Consider the function $\nu : (0,1) \mapsto (0,+\infty)$ defined in (6.10). This function, known as the Grötsch modulus, is strictly increasing and bijective. Define k_m by $\nu(k_m) = m\nu(k)$, and let

$$r(z) = \sqrt{k_m}\,\mathrm{sn}\left(m\frac{K(k_m)}{K(k)}u + (m+1)K(k_m); k_m\right), \qquad z = \sqrt{k}\,\mathrm{sn}(u;k).$$

Then $r \in \mathbb{Q}_{m,m}$; see [1, Table XXII]. The symmetry property

$$1/r(z) = r(1/z)$$

follows from the identity

$$k\,\mathrm{sn}(u+iK'(k);k) = 1/\mathrm{sn}(u;k).$$

If z runs once through $\mathbb{E}_0$ (or $\mathbb{F}_0$), then $u = u(z)$ runs once through the segment $[-K(k), K(k)]$ (or $[K(k)+iK'(k), -K(k)+iK'(k)]$). This implies that $v = m\frac{K(k_m)}{K(k)}u + (m+1)K(k_m)$ runs once through $[K(k_m), (1+2m)K(k_m)]$ (or $[(1+2m)K(k_m) + iK'(k_m), K(k_m) + iK'(k_m)]$). It follows from well-known properties of $\mathrm{sn}(u,k)$, see, e.g., [1, p. 207], that

$$\|r\|_{L_\infty(\mathbb{E}_0)} = \sqrt{k_m} = (-1)^j r\left(\sqrt{k}\,\mathrm{sn}\big(\frac{2j-m}{m}K(k), k\big)\right), \quad j = 0, 1, \ldots, m.$$

Thus, r attains the values $\pm\|r\|_{L_\infty(\mathbb{E}_0)}$ at $m+1$ points in $[-\sqrt{k}, \sqrt{k}]$. This alternation property, together with the above symmetry property, allows us to conclude that r is extremal for the Zolotarev problem,

$$Z_m(\mathbb{E}_0, \mathbb{F}_0) = \|r\|^2_{L_\infty(\mathbb{E}_0)} = k_m;$$

see, e.g., [1, §50] for details. In addition, since the roots of r are real, the symmetry property of r implies that $(-1)^m r$ is a Blaschke product and, in particular, the Blaschke product of minimal L_∞–norm on $\mathbb{E}_0$.

Recall from, e.g., [1, §49], that

$$\chi_{\mathbb{E}_0,\mathbb{F}_0}(e^{iv}) = \chi_{\mathbb{E}_0,\{|w|\geq 1\}}(e^{iv}) = \sqrt{k}\,\mathrm{sn}\left(\frac{2K(k)}{\pi}v + K(k); k\right), \qquad \chi_{\mathbb{E}_0,\mathbb{F}_0}(1) = \sqrt{k}.$$

Therefore, $R(\mathbb{E}_0, \{|w| \geq 1\}) = \sqrt{R(\mathbb{E}_0, \mathbb{F}_0)} = e^{\nu(k)/2}$. It follows from the above explicit formula that the zeros of r are given by

$$\begin{aligned}\widehat{w}_j &= \sqrt{k}\,\mathrm{sn}\left(K(k)\frac{m+1-2j}{m}; k\right)\\ &= \chi_{\mathbb{E}_0,\{|w|\geq 1\}}\left(\exp\left(2\pi i\frac{2j-1}{4m}\right)\right), \qquad j = 1, 2, \ldots, m.\end{aligned}$$

Thus, the zeros of r are the images of the first $(4m)$th roots of unity, that are not $(2m)$th roots of unity.

In order to relate a minimal Blaschke product on $[-\sqrt{k}, \sqrt{k}]$ to a minimal Blaschke product on $[\phi(\alpha), \phi(\beta)] \subset [-\infty, -1)$, we introduce the transformation

$$\zeta_2 = T(w) = (T_2 \circ T_1)(w), \qquad \zeta_1 = T_1(w) = \frac{1 - \phi(\beta)w}{w - \phi(\beta)}, \qquad T_2(\zeta_1) = \frac{\sqrt{k}\zeta_1 - 1}{\zeta_1 - \sqrt{k}},$$

with

$$T([\phi(\alpha), \phi(\beta)]) = [-\sqrt{k}, \sqrt{k}], \qquad T(\mathbb{D}) = \{|w_2| \geq 1\}, \qquad T(\phi(\beta)) = \sqrt{k},$$

where $\kappa \in (0,1)$ is as in (6.10), and $k \in (0,1)$ satisfies $\kappa = \frac{2\sqrt{k}}{1+k}$. Then $\pm r \circ T$ is a Blaschke product of minimal norm on $[\phi(\alpha), \phi(\beta)]$. It follows from

$$\chi_{[\phi(\alpha),\phi(\beta)],\mathbb{D}}(u) = T^{-1}\Big(\chi_{\mathbb{E}_0,\{\|w|\geq 1\}}(u)\Big), \qquad \chi_{[\phi(\alpha),\phi(\beta),\mathbb{D}]}(1) = T^{-1}(\sqrt{k}) = \phi(\beta),$$

that the zeros of $r \circ T$, given by $T^{-1}(\widehat{w}_j)$, can be determined via (6.11). Hence, the Blaschke product $1/B$, with B as defined in the beginning of the proof, indeed, is of minimal norm on $[\phi(\alpha), \phi(\beta)]$. In order to bound this norm, we use the inequality

$$\nu(\rho) \leq \log(4/\rho), \tag{6.14}$$

which is asymptotically sharp as ρ approaches zero. We have

$$2\log\left(\frac{\|1/B\|_{L_\infty([\phi(\alpha),\phi(\beta)])}}{2}\right) = 2\log\left(\frac{\|r\|_{L_\infty([-\sqrt{k},\sqrt{k}])}}{2}\right)$$
$$= \log(\frac{k_m}{4}) \leq -\nu(k_m) = -m\nu(k) = -2m\nu(\kappa),$$

where for the last equality, we applied the Gauss transform [1, Table XXI]. Comparison with (6.10) gives (6.13).

It remains to establish the bound (6.12). We obtain, in view of (6.10) and (6.14), that

$$R([\alpha,\beta],\mathbb{E})^{-m} = \exp(-m\nu(\kappa)) = \exp\left(-m\frac{(\pi/2)^2}{\nu(\sqrt{1-\kappa^2})}\right) \leq \exp\left(-m\frac{(\pi/2)^2}{\log(\frac{4}{\sqrt{1-\kappa^2}})}\right).$$

□

In actual computations, the ordering of the poles can be important. We propose Leja ordering; see [57, 67], where the latter reference illustrates Leja ordering of a finite point set.

Consider the case, discussed at the end of Section 6.1, when $[\alpha,\beta] = [-\infty,0]$, $\mathbb{E} = [\lambda_{\min},\lambda_{\max}]$, and the matrix A is positive definite. Note that the rate of convergence achieved with the poles of Theorem 6.6 depends on the logarithm of the quotient $\lambda_{\max}/\lambda_{\min}$, not on a root of the quotient. We have

$$\frac{4}{\sqrt{1-\kappa^2}} \leq 2\sqrt[4]{\lambda_{\max}/\lambda_{\min}},$$

and it follows from (6.12) that

$$R([\alpha,\beta],\mathbb{E})^{-m} \leq \exp\left(-m\frac{\pi^2}{\log(16\lambda_{\max}/\lambda_{\min})}\right).$$

Remark 6.7. *Theorem 6.6 provides m simple poles on $[\phi(\alpha),\phi(\beta)]$. Therefore, both for rational Arnoldi and for direct evaluation of $r(A)b$, we have to solve m shifted linear systems of equations. Assume this is done by LU-factorization. If we use each pole ℓ times, then we may determine a rational approximant of order $m\ell$ without computing additional LU-factorizations. In this case, $\widetilde{q}(w) = (w-w_1)^\ell(w-w_2)^\ell\ldots(w-w_m)^\ell$, and, according to (6.4) and (6.13), we have the bound*

$$\eta^{\widetilde{q}}_{\ell m}(\mathcal{F}^{-1}(f),\mathbb{D}) \leq \|\widetilde{f}-\widetilde{r}\|_{L_\infty(\mathbb{D})} \leq \frac{2^\ell\,\|f\|_{L_\infty(\mathbb{E})}}{|\phi(\beta)|} R([\alpha,\beta],\mathbb{E})^{-\ell m},$$

i.e., we loose a factor $2^{\ell-1}$ compared to Theorem 6.6. As long as ℓ is modest this may be acceptable. Recall that for ℓ even, Theorem 6.1(c) yields an error bound, which is sharp up to the factor $1-|\phi(\beta)|^{-2}$. In case $m=1$, we recover the pole of Corollary 6.4(a), though the error bound stated there is sharper. □

Gonchar [43] conjectured, and subsequently Parfenov [65] proved, that for a large class of functions f analytic in $\overline{\mathbb{C}}\setminus[\alpha,\beta]$, the error of best rational approximation with m free poles satisfies

$$\limsup_{m\to\infty}\left(\min_{q\in\mathbb{P}_m}\eta^q_m(f,\mathbb{E})\right)^{\frac{1}{m}} = R([\alpha,\beta],\mathbb{E})^{-2}. \tag{6.15}$$

Stahl and Totik [77, Theorem 6.2.2] show (6.15) for Markov functions (1.4) under weak regularity assumptions on the measure μ. In view of (2.9), Theorem 6.6 only gives us the geometric rate $R([\alpha,\beta],\mathbb{E})^{-1}$. This is a classical dilemma in rational approximation with prescribed poles.

We conclude this section by relating our rational approximants to those obtained by Hale et al. [45] via Talbot quadrature formulas. Let $\mathbb{F}$ be a closed set in the complex plane with connected complement, and let f be analytic in $\overline{\mathbb{C}}\setminus\mathbb{F}$. Let the matrix A satisfy $\mathbb{W}(A)\subset\mathbb{E}$, where $\mathbb{E}\subset\mathbb{C}$ satisfies $\mathbb{E}\cap\mathbb{F}=\emptyset$. Following Hale et al. [45], we seek to approximate $f(A)$ by approximating the contour integral

$$f(A)=\frac{1}{2\pi i}\int_{\mathcal{C}} f(z)(zI-A)^{-1}dz$$

by a quadrature rule. Here the contour $\mathcal{C}\subset\overline{\mathbb{C}}\setminus(\mathbb{E}\cup\mathbb{F})$ encircles $\mathbb{E}$ once. Error bounds for this approach depend on the choice of the contour, as well as on the quadrature rule used. We choose the "central" level curve of the underlying conformal mapping, namely the contour $z=\chi_{\mathbb{F},\mathbb{E}}(\rho e^{it})$, $-\pi\le t\le\pi$, where $\rho=\sqrt{R(\mathbb{F},\mathbb{E})}$, and apply the composite $2m$-point midpoint rule in the variable t. This is a Gauss-Szegő quadrature rule. We obtain the approximation

$$r(A)=\frac{1}{2m}\sum_{j=1}^{2m} f(\chi_{\mathbb{F},\mathbb{E}}(\zeta_j))(\chi_{\mathbb{F},\mathbb{E}}(\zeta_j)I-A)^{-1}\chi'_{\mathbb{F},\mathbb{E}}(\zeta_j)\zeta_j,\quad \zeta_j:=\rho e^{\pi i\frac{2j-1-2m}{2m}}.$$

Hale et al. [45, Theorem 1] show that for symmetric A and all $\widetilde{\rho}\in(1,\rho)$, there holds[1]

$$\|f(A)-r(A)\|=\|f-r\|_{L_\infty(\mathbb{E})}=O(1/\widetilde{\rho}^{2m}).$$

Their proof is based on the observations that, for any $z\in\mathbb{E}$, the function $\zeta\mapsto f(\chi_{\mathbb{F},\mathbb{E}}(\zeta))(\chi_{\mathbb{F},\mathbb{E}}(\zeta)-z)^{-1}\chi'_{\mathbb{F},\mathbb{E}}(\zeta)\zeta$ is analytic in $1/\widetilde{\rho}\le|\zeta|\le\widetilde{\rho}$, and that the composite $2m$-point midpoint rule integrates $\int_{-\pi}^{\pi}e^{i\ell t}\,dt$ exactly for $-2m<\ell<2m$. Computed examples for nonnormal matrices are presented in [45], but no explicit error bounds are provided.

Application of this approach to the Faber pre-image $\widetilde{f}=\mathcal{F}^{-1}(f)$ and the pair of sets $\{\phi(\mathbb{F}),\mathbb{D}\}$ yields the rational approximant

$$\widetilde{r}(w)=\frac{1}{2m}\sum_{j=1}^{2m}\widetilde{f}(\chi_{\phi(\mathbb{F}),\mathbb{D}}(\zeta_j))(\chi_{\phi(\mathbb{F}),\mathbb{D}}(\zeta_j)-w)^{-1}\chi'_{\phi(\mathbb{F}),\mathbb{D}}(\zeta_j)\zeta_j. \tag{6.16}$$

According to (2.7), the Faber image of $\widetilde{r}$ is close to f; we just have to replace $\widetilde{f}\circ\psi$ by f. If we determine the rational function r via (5.1), then (5.2) allows us to bound $\|f(A)-r(A)\|$ in terms of $\|\widetilde{f}-\widetilde{r}\|_{L_\infty(\mathbb{D})}$. The latter quantity can be bounded similarly as in [45]. This yields a rate of convergence slightly slower than $\mathcal{O}(\rho^{-2m})=\mathcal{O}(R(\mathbb{F},\mathbb{E})^{-m})$.

We finally compare the rational approximant (6.16) to approximants defined according to Theorems 6.1 and 6.6 when f is a Markov function f. Let $\mathbb{F}=[\alpha,\beta]$. The rate of convergence obtained by the approximants of Theorem 6.6 is roughly the square of the rate for Tablot quadrature rules with $2m$ distinct real poles. According to Remark 6.7, the same is true for m real poles of multiplicity two, where we note

[1] [45, Theorem 1] is stated for $\mathbb{F}=[-\infty,0]$, but the proof also applies to general $\mathbb{F}$.

that the evaluation of the corresponding rational function, $r(A)$, only requires the computation of m LU-factorizations.

However, choosing an optimal rational approximant with the poles in (6.16) gives an approximation error of about the same order. Let $\widehat{B}$ denote the Blaschke product (6.1) for the poles of (6.16), and let B be the Blaschke product of Theorem 6.6 with m poles. One then can show that

$$\widehat{B}(w) = \frac{1 + B(w)^2 c^2}{B(w)^2 + c^2},$$

where $c = |B(\chi_{\phi(\mathbb{F}),\mathbb{D}}(\zeta_1))| \approx \rho^m$ and $\|1/B\|_{L_\infty([\phi(\alpha),\phi(\beta)])} = k_m \approx \rho^{-2m}$. It follows that $\|1/\widehat{B}\|_{L_\infty([\phi(\alpha),\phi(\beta)])} \approx \rho^{-2m}$, i.e., the bound (6.4) of Theorem 6.1 yields an error of the same order.

7. Conclusion. This paper discusses the approximation of analytic functions on compact sets in the complex plane by polynomials and rational functions with preselected poles. New error bounds are derived via the Faber transform, which allows the approximation problem to be studied on the unit disk. The error bounds for Markov functions provide insight into the allocation of suitable poles. The error bounds are applied to the approximation of entire and Markov functions with matrix argument. In particular, the computation of polynomial and rational approximants by standard and rational Arnoldi processes is considered. Explicit error bounds for the approximants with matrix arguments are developed in terms of the field of values of the matrix. The standard and rational Arnoldi processes are shown to yield near-optimal approximants.

Acknowledgment. Work on this paper was begun during a visit of LR to Laboratoire Painlevé. LR would like to thank Bernd Beckermann for making this visit possible and enjoyable. The authors would like to thank Michel Crouzeix, Vladimir Druskin, and Leonid Knizhnerman for comments and discussions.

REFERENCES

[1] N. I. Achieser, Elements of the Theory of Elliptic Functions, Amer. Math. Society, Providence, 1990.

[2] M. Afanasjew, M. Eiermann, O. G. Ernst, and S. Güttel, A generalization of the steepest descent method for matrix functions, Electron. Trans. Numer. Anal., 28 (2008), pp. 206–222.

[3] E. J. Allen, J. Baglama, and S. K. Boyd, Numerical approximation of the product of the square root of a matrix with a vector, Linear Algebra Appl., 310 (2000), pp. 167–181.

[4] J. M. Anderson and J. Clunie, Isomorphism of the disk algebra and inverse Faber sets, Math. Z., 188 (1985), pp. 545–558.

[5] C. Badea, M. Crouzeix, and B. Delyon, Convex domains and K-spectral sets, Math. Z., 252 (2006), pp. 345–365.

[6] L. Baratchart, V. A. Prokhorov, and E. B. Saff, On Blaschke products associated with n-widths, J. Approx. Theory, 126 (2004), pp. 40–51.

[7] B. Beckermann, Image numérique, GMRES et polynômes de Faber, C. R. Acad. Sci. Paris, Ser. I, 340 (2005), pp. 855–860.

[8] B. Beckermann, S. A. Goreinov, and E. E. Tyrtyshnikov, Some remarks on the Elman estimate for GMRES, SIAM J. Matrix Anal. Appl., 27 (2006), pp. 772–778.

[9] D. A. Bini, N. J. Higham, and B. Meini, Algorithms for the matrix pth root, Numer. Algorithms, 39 (2005), pp. 349-378.

[10] D. Braess, Rational approximation of Stieltjes functions by the Carathéodory-Fejér method, Constr. Approx., 3 (1987), pp. 43–50.

[11] F. Cala Rodríguez and G. López Lagomasino, Multipoint Pade-type approximants. Exact rate of convergence, Constr. Approx., 14 (1998), pp. 259–272.

[12] D. Calvetti and L. Reichel, Exponential integration methods for large stiff systems of differential equations, in Iterative Methods in Scientific Computing IV, eds. D. R. Kincaid and A. C. Elster, IMACS Series in Computational and Applied Mathematics, vol. 5, IMACS, New Brunswick, 1999, pp. 237–243.

[13] D. Calvetti and L. Reichel, Lanczos-based exponential filtering for discrete ill-posed problems, Numer. Algorithms, 29 (2002), pp. 45–65.

[14] D. Calvetti, L. Reichel, and Q. Zhang, Iterative exponential filtering for large discrete ill-posed problems, Numer. Math., 83 (1999), pp. 535–556.

[15] J. R. Cardoso and F. Silva Leite, Padé and Gregory error estimates for the logarithm of block triangular matrices, Appl. Numer. Math., 56 (2006), pp. 253–267.

[16] M. Crouzeix, Operators with numerical range in a parabola, Arch. Math., 82 (2004), pp. 517–527.

[17] M. Crouzeix, Numerical range and functional calculus in Hilbert space, J. Functional Anal., 244 (2007), pp. 668–690.

[18] M. Crouzeix, A functional calculus based on the numerical range. Applications, Linear Multilinear Algebra, 56 (2008), pp. 81–103.

[19] M. Crouzeix and B. Delyon, Some estimates for analytic functions of strip or sectorial operators, Arch. Math., 81 (2003), pp. 559–566.

[20] J. H. Curtiss, Faber polynomials and the Faber series, Amer. Math. Monthly, 78 (1971), pp. 577–596.

[21] P. I. Davies and N. J. Higham, Computing $f(A)b$ for matrix functions f, in QCD and Numerical Analysis III, eds. A. Borici, A. Frommer, B. Joo, A. Kennedy, and B. Pendleton, Lecture Notes in Computational Science and Engineering, vol. 47, Springer, Berlin, 2005, pp. 15–24.

[22] B. Delyon and F. Delyon, Generalization of von Neumann's spectral sets and integral representation of operators, Bull. Soc. Math. France, 1 (1999), pp. 25–42.

[23] 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.

[24] T. A. Driscoll, Improvements to the MATLAB toolbox for Schwartz-Christoffel mapping, ACM Trans. Math. Software, 31 (2005) pp. 239–251.

[25] T. A. Driscoll and L. N. Trefethen, Schwarz-Christoffel Mapping, Cambridge University Press, Cambridge, 2002.

[26] V. L Druskin and L. A. Knizhnerman, Two polynomial methods for the computation of functions of symmetric matrices, USSR Comput. Math. Math. Phys., 29 (1989) pp. 112–121.

[27] V. L. Druskin and L. A. Knizhnerman, Error estimates for the simple Lanczos process when computing functions of symmetric matrices and eigenvalues, USSR Comput. Maths. Math. Phys., 31 (1991), pp. 20–30.

[28] V. Druskin and L. Knizhnerman, Krylov subspace approximation of eigenpairs and matrix functions in exact and computer arithmetic, Numer. Linear Algebra Appl., 2 (1995), pp. 205–217.

[29] 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.

[30] V. Druskin, L. Knizhnerman, and M. Zaslavsky, Solution of large scale evolutionary problems using rational Krylov subspaces with optimized shifts, submitted for publication.

[31] M. Eiermann, Fields of values and iterative methods, Linear Algebra Appl., 180 (1993), pp. 167–197.

[32] M. Eiermann and O. Ernst, A restarted Krylov subspace method for the evaluation of matrix functions, SIAM J. Numer. Anal., 44 (2006), pp. 2481–2504.

[33] S. C. Eisenstat, H. C. Elman, and M. H. Schultz, Variational iterative methods for nonsymmetric systems of linear equations, SIAM J. Numer. Anal., 20 (1983), pp. 345–357.

[34] S. W. Ellacott, On the Faber transform and efficient numerical rational transformation, SIAM J. Numer. Anal., 20 (1983), pp. 989–1000.

[35] S. W. Ellacott and E. B. Saff, Computing with the Faber transform, in Rational Approximation and Interpolation, eds. P. R. Graves-Morris, E. B. Saff, and R. S. Varga, Lecture Notes in Mathematics # 1105, Springer, Berlin, 1984, pp. 412–418.

[36] S. D. Fisher and E. B. Saff, The asymptotic distribution of zeros of minimal Blaschke products, J. Approx. Theory, 98 (1999), pp. 104–116.

[37] R. Freund, Solution of shifted linear systems by quasi-minimal residual iterations, in Numerical Linear Algebra, eds. L. Reichel, A. Ruttan, and R. S. Varga, de Gruyter, Berlin, 1993, pp. 101–121.

[38] A. Frommer and V. Simoncini, Stopping criteria for rational matrix functions of Hermitian and symmetric matrices, SIAM J. Sci. Computing, 30 (2008), pp. 1387–1412.

[39] D. Gaier, Lectures on Complex Approximation, Birkhäuser, Basel, 1987.

[40] E. Gallopoulos and Y. Saad, Efficient solution of parabolic equations by Krylov approximation methods, SIAM J. Sci. Statist. Comput., 13 (1992), pp. 1236–1264.

[41] T. Ganelius, Degree of rational approximation, in Lectures on Approximation and Value Distribution, Les Presses de Université de Montréal, Montréal, Canada, 1982.

[42] G. H. Golub and C. F. Van Loan, Matrix Computations, 3rd ed., Johns Hopkins University Press, Baltimore, 1996.

[43] A. A. Gonchar, Rational approximation of analytic functions, J. Soviet Math. 26 (1984) no. 5, 2218-2220.

[44] A. Greenbaum, Iterative Methods for Solving Linear Systems, SIAM, Philadelphia, 1997.

[45] N. Hale, N. J. Higham, and L. N. Trefethen, Computing A^α, $\log(A)$, and related matrix functions by contour integrals, SIAM J. Numer. Anal. 46 (2008), pp. 2505–2523.

[46] G. I. Hargreaves and N. J. Higham, Efficient algorithms for the matrix cosine and sine, Numer. Algorithms, 40 (2005), pp. 383–400.

[47] P. Henrici, Applied and Computational Complex Analysis, vol. 1, Wiley, New York, 1974.

[48] N. J. Higham, Evaluating Padé approximants of the matrix logarithm, SIAM J. Matrix Anal. Appl., 22 (2001), pp. 1126–1135.

[49] N. J. Higham, The scaling and squaring method for the matrix exponential revisited, SIAM J. Matrix Anal. Appl., 26 (2005), pp. 1179–1193.

[50] N. J. Higham, Functions of Matrices: Theory and Computation, SIAM, Philadelphia, 2008.

[51] M. Hochbruck and C. Lubich, On Krylov subspace approximations to the matrix exponential operator, SIAM J. Numer. Anal., 34 (1997), pp. 1911–1925.

[52] R. A. Horn and C. R. Johnson, Topics in Matrix Analysis, Cambridge University Press, Cambridge, 1991.

[53] C. Kenney and A. J. Laub, Padé error estimates for the logarithm of a matrix, Internat. J. Control, 50 (1989), pp. 707–730.

[54] L. A. Knizhnerman, Calculation of functions of unsymmetric matrices using Arnoldi's method, USSR Comput. Maths. Math. Phys., 31(1) (1991), pp. 1–9.

[55] L. Knizhnerman and V. Simoncini, A new investigation of the extended Krylov subspace method for matrix function evaluations, Numer. Linear Algebra Appl., to appear.

[56] T. Kővari and Ch. Pommerenke, On Faber polynomials and Faber expansions, Math. Z., 99 (1967), pp. 193–206.

[57] F. Leja, Sur certaines suits liées aux ensemble plan et leur application à la representation conforme, Ann. Polon. Math., 4 (1957), pp. 8–13.

[58] L. Lopez and V. Simoncini, Analysis of projection methods for rational function approximation to the matrix exponential, SIAM J. Numer Anal., 44 (2006), pp. 613–635.

[59] U. Luther and K. Rost, Matrix exponentials and inversion of confluent Vandermonde matrices, Electron. Trans. Numer. Anal., 18 (2004), pp. 91–100.

[60] A. I. Markushevich, Theory of Functions of a Complex Variable, Chelsea, New York, 1985.

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

[62] I. Moret, Rational Lanczos approximations to the matrix square root and related functions, Numer. Linear Algebra Appl., to appear.

[63] 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.

[64] R. Nevanlinna and V. Paatero, Introduction to Complex Analysis, Amer. Math. Society, Providence, 2007.

[65] O. G. Parfenov, Estimates of the singular numbers of the Carlson imbedding operator, Math. USSR Sb., 59 (1988), pp. 497–514.

[66] Ch. Pommerenke, Konforme Abbildungen und Fekete-Punkte, Math. Z., 89 (1965), pp. 422–438.

[67] L. Reichel, Newton interpolation at Leja points, BIT, 30 (1990), pp. 332–346.

[68] A. Ruhe, Rational Krylov sequence methods for eigenvalue computations, Linear Algebra Appl., 58 (1984), pp. 391–405.

[69] A. Ruhe, Rational Krylov: a practical algorithm for large sparse nonsymmetric matrix pencils, SIAM J. Sci. Comput., 19 (1998), pp. 1535–1551.

[70] Y. Saad, Analysis of some Krylov subspace approximations to the matrix exponential operator, SIAM J. Numer. Anal., 29 (1992), pp. 209–228.

[71] Y. Saad, Iterative Methods for Sparse Linear Systems, 2nd ed., SIAM, Philadelphia, 2003.

[72] Y. Saad and M. H. Schultz, GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems, SIAM J. Sci. Stat. Comput., 7 (1986), pp. 856–869.

[73] E. B. Saff and V. Totik, Logarithmic Potentials with External Fields, Springer, Berlin, 1997.

[74] T. Schmelzer and L. N. Trefethen, Evaluating matrix functions for exponential integrators

via Carathéodory-Fejér approximation and contour integration, Electron. Trans. Numer. Anal., 29 (2007), pp. 1–18.
[75] A. Schönhage, Approximationstheorie, de Gruyter, Berlin, 1971.
[76] R. B. Sidje, Expokit: a software package for computing matrix exponentials, ACM Trans. Math. Software, 24 (1998), pp. 130–156.
[77] H. Stahl and V. Totik, General Orthogonal Polynomials, Cambridge University Press, Cambridge, 1992.
[78] P. K. Suetin, Fundamental properties of Faber polynomials, Russian Math. Surveys, 19 (1964), pp. 121–149.
[79] S. P. Suetin, On the Montessus de Ballore's theorem for nonlinear Padé approximants of orthogonal expansions and Faber series, Dokl. Akad. Nauk SSSR, 253 (1980), pp. 1322-1325; English transl. in Soviet Math. Dokl., 22 (1980).
[80] K. C. Toh and L. N. Trefethen, The Kreiss matrix theorem on a general complex domain, SIAM J. Matrix Anal. Appl., 21 (1999), pp. 145–165.
[81] J. van den Eshof and M. Hochbruck, Preconditioning Lanczos approximations to the matrix exponential, SIAM J. Sci. Comput., 27 (2006), pp. 1438–1457.
[82] J. van den Eshof and G. L. G. Sleijpen, Accurate conjugate gradient methods for families of shifted systems, Applied Numer. Math., 49 (2004), pp. 17–37.
[83] J. L. Walsh, Interpolation and Approximation by Rational Functions in the Complex Domain, 5th ed., Amer. Math. Soc., Providence, 1969.