On the rational approximation of Markov functions, with applications to the computation of Markov functions of Toeplitz matrices

Bernhard Beckermann, Joanna Bisch, Robert Luce

On the rational approximation of Markov functions, with applications to the computation of Markov functions of Toeplitz matrices[‡]

by Bernhard Beckermann[*], Joanna Bisch[*], and Robert Luce[§]

Abstract

We investigate the problem of approximating the matrix function $f(A)$ by $r(A)$, with f a Markov function, r a rational interpolant of f, and A a symmetric Toeplitz matrix. In a first step, we obtain a new upper bound for the relative interpolation error $1 - r/f$ on the spectral interval of A. By minimizing this upper bound over all interpolation points, we obtain a new, simple and sharp a priori bound for the relative interpolation error. We then consider three different approaches of representing and computing the rational interpolant r. Theoretical and numerical evidence is given that any of these methods for a scalar argument allows to achieve high precision, even in the presence of finite precision arithmetic. We finally investigate the problem of efficiently evaluating $r(A)$, where it turns out that the relative error for a matrix argument is only small if we use a partial fraction decomposition for r following Antoulas and Mayo. An important role is played by a new stopping criterion which ensures to automatically find the degree of r leading to a small error, even in presence of finite precision arithmetic.

1 Introduction and statement of the results

The need for computing matrix functions $f(A)$ for some square matrix $A \in \mathbb{R}^{n\times n}$ and some function being analytic on some neighborhood of the spectrum of A arises in a variety of applications, including network analysis [BB20, EH10], signal processing [SNF$^+$13], machine learning [Sto20], and differential equations [HO10]. We refer the reader to [Hig08] and the references therein for a detailed account on computing matrix functions for various functions f. In the present paper we are interested in the particular case of Markov functions, that is, the Cauchy transform of a positive measure μ with support $\mathrm{supp}(\mu) \subset \mathbb{R}$, and more precisely

$$f^{[\mu]}(z) = \int \frac{d\mu(x)}{z-x}, \quad \text{with infinite } \mathrm{supp}(\mu) \subset [\alpha,\beta] \text{ for suitable } -\infty \le \alpha < \beta < +\infty. \tag{1.1}$$

This includes the functions $f^{[\mu]}(z) = \frac{\log(z)}{z-1}$ or $f^{[\mu]}(z) = z^\gamma$ for $\gamma \in (-1,0)$ and $\mathrm{supp}(\mu) = (-\infty, 0]$, but also many other elementary functions, see for instance [Hen77]. In particular, elementary computations show that[2]

$$f^{[\nu]}(z) = \frac{\sqrt{|\alpha|}}{\sqrt{(z-\alpha)(z-\beta)}}, \quad \text{with density} \quad \frac{d\nu}{dx}(x) = \frac{\sqrt{|\alpha|}}{\pi\sqrt{(x-\alpha)(\beta-x)}} \quad \text{on } \mathrm{supp}(\nu) = [\alpha,\beta]. \tag{1.2}$$

The main reason for restricting ourselves to Markov functions is that many results about best rational approximants and rational interpolants are known, see for instance the first paragraph in §2 and the references therein. In addition, for evaluating $r(A)$ for a rational function r we can fully exploit the structure of A: if A is a Toeplitz matrix

$$A = \begin{pmatrix} t_0 & t_{-1} & \dots & \dots & t_{-n+1} \\ t_1 & t_0 & t_{-1} & \dots & t_{-n+2} \\ \vdots & t_1 & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & t_{-1} \\ t_{n-1} & \dots & \dots & t_1 & t_0 \end{pmatrix} \tag{1.3}$$

[‡]AMS subject classifications : 15A16, 30E10, 41A20, 65D15, 65F55, 65F60. Key words: matrix function, Toeplitz matrices, Markov function, rational interpolation, positive Thiele continued fractions.

[*]Laboratoire Paul Painlevé UMR 8524, Département de Mathématiques, Université de Lille, F-59655 Villeneuve d'Ascq, France. The work has been supported in part by the Labex CEMPI (ANR-11-LABX-0007-01). Corresponding author: Bernhard.Beckermann@univ-lille.fr

[§]Gurobi Optimization, LLC., 9450 SW Gemini Dr. #90729 Beaverton, Oregon, USA. luce@gurobi.com

[2]This includes the limiting case $\frac{d\nu}{dx}(x) = \frac{1}{\pi\sqrt{z-\beta}}$ and $f^{[\nu]}(z) = 1/\sqrt{z-\beta}$ for $\alpha \to -\infty$.

then using the concept of displacement rank we just need $\mathcal{O}(n\log^2(n))$ operations and $\mathcal{O}(n)$ memory requirements (the hidden constant depending on the degree), and a similar property seems to be true for matrices with hierarchical rank structure, see §4 and the references therein.

To be more precise, denote by $\mathcal{R}_{m,n}$ the set of rational functions, with numerator degree $\leq m$, and denominator degree $\leq n$. Given $z_1, ..., z_{2m}$ called interpolation points in $\mathbb{C} \setminus [\alpha, \beta]$, a *rational interpolant* (also sometimes called multi-point Padé approximant) $r_m^{[\mu]}$ of $f^{[\mu]}$ of type $[m-1|m]$ is a rational function in $\mathcal{R}_{m-1,m}$ which interpolates $f^{[\mu]}$ at $z_1, ..., z_{2m}$ (in the sense of Hermite if some of the interpolation points occur with multiplicity > 1). For Markov functions it is known that, provided that the non-real interpolation points occur in conjugate pairs, there is one and only one rational interpolant of type $[m-1|m]$, and this interpolant has m simple poles in (α, β), and positive residuals. For instance, Padé approximants matching Taylor expansions at z_1 are a special case of rational interpolants, with $z_1 = ... = z_{2m}$. Also, from equi-oscillation we know that a best rational approximant in $\mathcal{R}_{m-1,m}$ with respect to maximum norm on some interval $[c, d] \subset (\beta, +\infty)$ is a rational interpolant of $f^{[\mu]}$. In contrast, approximating $f(A)b$ by projection on rational Krylov spaces [BR09] gives raise to an expression $r(A)b$ where $r \in \mathcal{R}_{m-1,m}$ has prescribed poles, and satisfies only m interpolation conditions, a so-called Padé type rational approximant.

Inspired by several authors [Hig08], in the present paper we will approximate $f^{[\mu]}(A)$ by $r_m^{[\mu]}(A)$ for suitable interpolation points. For instance, the authors in [HL13] use Padé approximants (represented as convergents of a Stieltjes continued fraction) for fractional powers (1.2), combined with scaling and squaring techniques. The computation of rational interpolants $r_m^{[\mu]}$ of Markov functions $f^{[\mu]}$ is known to be delicate on a computer with finite precision arithmetic: we get instead a rational function $\widetilde{r}_m$ which might be far from $r_m^{[\mu]}$, depending on how to represent and to compute the interpolant. In addition, on a computer we will obtain a matrix R instead of $\widetilde{r}_m(A)$, again due to finite precision arithmetic. We are still far from a full understanding how these different errors accumulate, and thus will be interested in the present paper mainly in the case of symmetric matrices A. Denote by $\mathbb{E} \subset (\beta, +\infty)$ a closed set containing all eigenvalues of A, for instance the spectral interval spanned by the smallest and largest eigenvalue of A. We thus are interested in the three relative errors

$$\left\| 1 - r_m^{[\mu]}/f^{[\mu]} \right\|_{L^\infty(\mathbb{E})}, \quad \left\| 1 - \widetilde{r}_m/f^{[\mu]} \right\|_{L^\infty(\mathbb{E})}, \quad \left\| I - R\, f^{[\mu]}(A)^{-1} \right\|, \tag{1.4}$$

the first two being upper bounds of $\|I - r_m^{[\mu]}(A) f^{[\mu]}(A)^{-1}\|$, and $\|I - \widetilde{r}_m(A) f^{[\mu]}(A)^{-1}\|$, respectively. These three quantities will be discussed in §2, §3, and §5, respectively.

Let us highlight the main theoretical contributions of this paper. In Theorem 2.1 we suggest new error bounds in terms of Blaschke products of order $2m$ for the relative[3] interpolation error $\|1 - r_m^{[\mu]}/f^{[\mu]}\|_{L^\infty(\mathbb{E})}$ on subsets $\mathbb{E}$ of the real line for quite arbitrary $m, \alpha, \beta, z_1, ..., z_{2m}$. It follows that, up to some modest constant, there is a worst case measure for the relative error given by the measure ν of (1.2). As a consequence of Theorem 2.1, we derive in Corollary 2.2 new residual and a posteriori upper bounds for $\|I - r_m^{[\mu]}(A) f^{[\mu]}(A)^{-1}\|$ which do not require to know $f^{[\mu]}(A)$. Restricting ourselves to intervals $\mathbb{E} = [c, d]$, this allows us in Remark 2.3 for the Padé case $z_1 = ... = z_{2m}$ to find an optimal z_1, and in Corollary 2.4 the quasi optimal interpolation points which minimize our upper bound of Theorem 2.1. In the latter case, we deduce a very simple a priori error bound of asymptotic form $8\rho^{2m}$ in terms of the logarithmic capacity of the underlying condenser, which is sharp and again seems to be new. Based on these results, we suggest in Remark 2.5 a stopping criterion allowing to find automatically the m leading to a small interpolation error, even in the presence of rounding errors.

In Theorem 2.6 we estimate the absolute interpolation error on the closed unit disk, which is combined with the Faber operator techniques of [BR09] in order to construct rational functions $r \in \mathcal{R}_{m-1,m}$ with an explicit bound for the error $\|f - r\|_{L^\infty(\mathbb{E})}$ for compact and convex sets $\mathbb{E}$, such as the field of values of A in case where A is not symmetric. Again, optimizing the interpolation points, our bounds improve results of Knizhnerman on Faber-Padé approximants [Kni09]. Finally, this paper also contains two new results on interpolating Thiele continued fractions: in Theorem 3.1 we show that (reciprocal) Markov functions give raise to an interpolating continued fraction with positive parameters, which allows us to show in Theorem 3.3 the backward stability of positive Thiele fractions, improving [GM80, Theorem 4.1] of Graves-Morris.

[3] In most papers in the literature, the authors estimate the absolute interpolation error. However, by considering the relative error we may monitor also the error of rational approximants of functions f being a product of a Markov function and a rational function such as the logarithm or the square root, see Examples 5.1 and 5.3.

We conclude this introduction by summarizing the structure of the paper. In the first paragraph of §2 we recall several results scattered in the literature on upper bounds for rational interpolants and best rational approximants of Markov functions. We then state and prove our new bounds for the interpolation error on subsets of the real line in §2.1, and on the unit disk in §2.2. In order to monitor the second term in (1.4), we will discuss in §3 three different ways of representing and computing $r_m^{[\mu]}$, namely in §3.2 a partial fraction decomposition, in §3.3 an interpolating barycentric representation of $r_m^{[\mu]}$, and in §3.4 a Thiele interpolating continued fraction, which generalizes the above-mentioned Stieltjes continued fraction to arbitrary distinct and real interpolation points. We give in Figures 3.1 and 5.1–5.3 numerical evidence that we may reach nearly machine precision for the error $\|1-\widetilde{r}_m/f^{[\mu]}\|_{L^\infty(\mathbb{E})}$ in (1.4) for any of the three representations $\widetilde{r}_m$ of $r_m^{[\mu]}$, if we use the stopping criterion of Remark 2.5.

In §4 we provide more information how to evaluate $\widetilde{r}_m(A)$ for a (symmetric) Toeplitz matrix A, using the concept of small displacement rank. In particular, we show in Theorem 4.1 the above claimed complexity $\mathcal{O}(n\log^2(n))$ and memory requirements $\mathcal{O}(n)$. Finally, in §5 we give numerical experiments, and investigate also possible improvements through a combination with scaling and squaring for particular functions. We conclude that only a representation of $r_m^{[\mu]}$ as a partial fraction decomposition allows to attain small errors, if we want to exploit the Toeplitz structure of A.

2 The error of rational interpolants of Markov functions

The aim of this section is to estimate the error of rational interpolants $r_m^{[\mu]}$ of type $[m-1|m]$ of a Markov function $f^{[\mu]}$ of a measure μ with support in $[\alpha,\beta]$, both on a real set $\mathbb{E}\subset\mathbb{R}\setminus[\alpha,\beta]$ as for instance a real interval in our Theorem 2.1, and on the unit disk in our Theorem 2.6. We are less interested in asymptotic results on the error, and refer the interested reader to the work of Gonchar [Gon78a] and the book of Stahl and Totik [ST92] for mth root asymptotics, the work of López Lagomasino [Lag86, Lag87] on ratio asymptotics, and the work of Stahl [Sta00] on strong asymptotics, though some of the tools in these papers are also of help for deriving upper bounds. In this paper we want to derive upper bounds of the form $C\rho^{2m}$, where the constants C,ρ only depend on $[\alpha,\beta]$ and the interpolation points, but not on the regularity of μ. Previous work on this subject include error estimates for Padé approximants of Markov functions, see, e.g., the book of Baker and Graves-Morris [BGM96, Thm 5.2.6 and Thm 5.4.4]. We are only aware of work of Ganelius [Gan82, Chap. 4] and Braess [Bra87, Thm 2.1] on upper bounds of the form $C\rho^{2m}$. These authors look for particularly well-chosen interpolation points which allow to make the link with best rational approximants. Also we should mention the more recent work of Knizhnerman [Kni09, Part.2, Section 3.1] on Faber Padé approximants, who does not give an explicit value of C. Our aim is to improve the constant C in all these findings. Also, we want to prove the claim in [BR09] that for rational interpolants with free poles we should get the square of the bounds of [BR09] and [MR21] in terms of (minimal) Blaschke products obtained for rational interpolants with prescribed poles.

2.1 Estimates on the real line

We start with the interval case, where we allow for more general real or complex conjugate interpolation points $z_1,...,z_{2m}\in\mathbb{C}\setminus[\alpha,\beta]$ and estimate the relative interpolation error. The comparison principle [Bra86, Lemma V.3.8 and Thm V.3.9] allows to relate absolute errors for rational interpolants of Markov functions for two measures $\mu\le\nu$. The situation is different for relative errors since, as we show in the next theorem, there is (up to a factor 2 or 3) a worst case measure given by the scaled equilibrium measure of the interval $[\alpha,\beta]$ (and thus neither depending on the choice of $\mathbb{E}$ nor on the interpolation points). We also give upper bounds in terms of Blaschke products.

Theorem 2.1. *Let $-\infty\le\alpha<\beta<\infty$, and let the Markov functions $f^{[\mu]}$ and $f^{[\nu]}$ be as in* (1.1) *and* (1.2). *Furthermore, let $\mathbb{E}\subset\mathbb{R}\setminus[\alpha,\beta]$, and consider interpolation points $z_1,...,z_{2m}\in\mathbb{C}\setminus[\alpha,\beta]$ where we suppose that non-real points only occur in conjugate pairs. We refer to the positive case if the real interpolation points have even multiplicity.*[4] *Then for the interpolant $r_m^{[\mu]}$ of type $[m-1|m]$ of $f^{[\mu]}$ we*

[4] If $\mathbb{E}$ is a finite union of closed intervals, it is sufficient to suppose that interpolation points in $Int(\mathbb{E})$ only have even multiplicity, and there is an even number of interpolation points in any subinterval of $\mathbb{R}\setminus Int(\mathbb{E})$.

may bound the relative error as follows

$$\|\frac{f^{[\mu]}-r_m^{[\mu]}}{f^{[\mu]}}\|_{L^\infty(\mathbb{E})} \leqslant \begin{cases} 2\|\frac{f^{[\nu]}-r_m^{[\nu]}}{f^{[\nu]}}\|_{L^\infty(\mathbb{E})} \leq 4\eta_{2m} & \text{in the positive case,} \\ \|1-\left(\frac{r_m^{[\nu]}}{f^{[\nu]}}\right)^2\|_{L^\infty(\mathbb{E})} \leq 4\frac{\eta_{2m}}{(1-\eta_{2m})^2} & \text{in the general case,} \end{cases} \tag{2.1}$$

where

$$\eta_{2m} = \max_{z\in\mathbb{E}} |G_{2m}(z)|, \quad G_{2m}(z) = \prod_{j=1}^{2m} \frac{\varphi(z)-\varphi(z_j)}{1-\varphi(z)\varphi(z_j)}$$

with φ mapping conformally $\overline{\mathbb{C}}\backslash[\alpha,\beta]$ onto the complement of the closed unit disk.

Proof. Define

$$\omega(z) = \pm\prod_{j=1}^{2m}(z-z_j). \tag{2.2}$$

where by assumption on $z_1,...,z_{2m}$ the function ω is real-valued on the real axis and different from zero in $[\alpha,\beta]$, we hence may fix the sign such that $\omega(z) > 0$ for $z \in [\alpha,\beta]$. It was probably Gonchar in [Gon78a] who observed first that the denominator Q_m of the interpolant $r_m^{[\mu]}$ of our Markov function $f^{[\mu]}$ is necessarily a scalar multiple of an mth orthonormal polynomial with respect to the measure $d\mu/\omega$, in particular the rational interpolant $r_m^{[\mu]}$ exists, is unique, and has m simple poles in (α,β), with positive residuals, see also [ST92, Lemma 6.1.2]. Gonchar also gave the integral formula

$$\forall z \in \mathbb{C}\setminus[\alpha,\beta]: \quad f^{[\mu]}(z) - r_m^{[\mu]}(z) = \frac{\omega(z)}{Q_m^2(z)}\int \frac{Q_m^2(x)}{\omega(x)}\frac{d\mu(x)}{z-x}. \tag{2.3}$$

In our approach we use two polynomial extremal problems: we claim that [5]

$$\forall z \in \mathbb{R}\setminus[\alpha,\beta]: \quad |f^{[\mu]}(z) - r_m^{[\mu]}(z)| = \min_{\deg Q\leq m} \frac{|\omega(z)|}{Q^2(z)}\int \frac{Q^2(x)}{\omega(x)}\frac{d\mu(x)}{|z-x|} \tag{2.4}$$

$$\leq |f^{[\mu]}(z)| \min_{\deg Q\leq m} \frac{|\omega(z)|}{Q^2(z)}\|\frac{Q^2}{\omega}\|_{L^\infty([\alpha,\beta])}. \tag{2.5}$$

Indeed, for any $z \in \mathbb{R}\setminus\mathrm{supp}(\mu)$ it follows from [Sze75, Thm 3.1.3 and 3.1.4] that the denominator Q_m is extremal for the extremal problem on the right-hand side of (2.4), and hence the claimed equality (2.4) follows from (2.3), whereas inequality (2.5) is a trivial consequence of (2.4) and of the fact that $\mathrm{supp}(\mu) \subset [\alpha,\beta]$ by (1.1). It remains to solve the L^∞ extremal problem on the right-hand side of (2.5), and again we will see that we get the same extremal polynomial for all $z \in \mathbb{R}\setminus[\alpha,\beta]$, namely the weighted Chebyshev polynomial. By the theory of best approximation and the Chebyshev theorem [Mei67, Section 4.1 and 4.4], all we have to do is to find a polynomial P of degree $\leq m$ such that $P(x)/\sqrt{\omega(x)}$ is of modulus ≤ 1 on $[\alpha,\beta]$, and takes $m+1$ times in $[\alpha,\beta]$ alternately the values 1 and -1.

We first show that it is sufficient to consider the interval $[-1,1]$. Let $T \in \mathcal{R}_{1,1}$ be a Moebius transform with $T([-1,1]) = [\alpha,\beta]$ and $T(\mathbb{R}) = \mathbb{R}$, and define $w = \varphi(z)$ by the formula $z = T(\frac{1}{2}(w+\frac{1}{w}))$, then φ is a conformal bijection of the exterior of the interval $[\alpha,\beta]$ onto the exterior of the unit disk.[6] For any polynomial p of degree $\leq m$ we find a polynomial P of degree $\leq m$ such that

$$\frac{P(T(y))}{\sqrt{\omega(T(y))}} = \frac{p(y)}{\sqrt{\rho(y)}}, \quad \text{where} \quad \rho(y) = \pm\prod_{j=1}^{2m}(y-T^{-1}(z_j)) \tag{2.6}$$

and, as in (2.2), the sign is chosen such that $\rho > 0$ in $[-1,1]$. Especially, with p, also P has the desired oscillatory behavior. It remains to construct p, which is explained in [Mei67, Section 4.4] if ω is a square of a polynomial (the case of points of even multiplicity) but easily extends to our more general setting. We may factorize

$$\rho(\frac{1}{2}(w+\frac{1}{w})) = H(w)H(\frac{1}{w}), \quad H(w) = \sum_{k=0}^{2m} H_k w^k = H_{2m}\prod_{j=1}^{2m}(w-w_j), \quad |w_j| > 1$$

[5] In other approaches like in [ST92, Section 6.1] the authors eliminate the term $1/(z-x)$ in the integral leading to bounds for the absolute error.

[6] The interested reader may observe that we do not impose a normalization condition, and hence neither T nor φ are unique, though the function G_{2m} can be shown to be unique, see also Remark 2.3.

where $\frac{1}{2}(w_j+\frac{1}{w_j}) = T^{-1}(z_j)$, in other words, $w_j = \varphi(z_j)$ for $j = 1, ..., 2m$. Since $w^\ell + w^{-\ell}$ is a polynomial of degree ℓ of $w + w^{-1}$ for $\ell = 0, ..., m$, we conclude that p defined by

$$p(\frac{1}{2}(w+\frac{1}{w}) = \frac{1}{2}(w^{-m}H(w) + w^m H(\frac{1}{w}))$$

is a polynomial of degree $\leq m$. Introduce the Blaschke product

$$B(w) = \frac{w^{2m}H(\frac{1}{w})}{H(w)} = \prod_{j=1}^{2m} \frac{1-w_j w}{w-w_j} = \frac{1}{G_{2m}(\varphi^{-1}(w))}$$

having all its zeros in $\mathbb{D}$, non-real zeros occurring in conjugate pairs. Then for $x = \cos(t)$ and $w = e^{it}$ we have that

$$\frac{w^{-m}H(w)}{|w^{-m}H(w)|} = e^{-is}, \quad \frac{w^m H(1/w)}{|w^m H(1/w)|} = e^{is}, \quad B(w) = e^{2is},$$

and hence $p(\cos(t))/\sqrt{\rho(\cos(t))} = \cos(s)$. However, for a Blaschke product as above, we know that with $t \in [0, \pi]$, $2s$ runs trough the interval $[0, 2m\pi]$, leading to the desired oscillatory behavior. To summarize, we have shown that

$$\min_{\deg Q \leq m} \|\frac{|\omega|}{Q^2}\|_{L^\infty(\mathbb{E})} \|\frac{Q^2}{\omega}\|_{L^\infty([\alpha,\beta])} = \max_{x\in\mathbb{E}} \min_{\deg Q\leq m} |\frac{\omega(x)}{Q^2(x)}| \, \|\frac{Q^2}{\omega}\|_{L^\infty([\alpha,\beta])} = \max_{x\in\mathbb{E}} \frac{4|G_{2m}(x)|}{(1+G_{2m}(x))^2}, \quad (2.7)$$

with $G_{2m}(x) \in (-1, 1)$ for $x \in \mathbb{E}$ in the general case. In the positive case we know in addition that $G_{2m}(\beta) = 1 > 0$, G_{2m} has an even number of sign changes in any subinterval of $\mathbb{R} \setminus Int(\mathbb{E})$, and only zeros with even multiplicities in $Int(\mathbb{E})$. Hence $G_{2m}(x) \in [0, 1)$ for $x \in \mathbb{E}$ in the positive case. We still need to show that we may express rational interpolants of the Markov function $f^{[\nu]}$ in terms of G_{2m}. With the same change of variables as above, $w = \varphi(z)$, $z = T(y)$, $y = (w + 1/w)/2$, we claim that there exist rational functions $r, R \in \mathcal{R}_{m-1,m}$ such that, for $z \notin [\alpha, \beta]$,

$$\frac{1-G_{2m}(z)}{1+G_{2m}(z)} = \sqrt{y^2-1}r(y) = \frac{R(z)}{f^{[\nu]}(z)}. \quad (2.8)$$

Here the first identity is obtained by taking the above $p(y)$ as denominator, and the second is left to the reader. Observing that the left-hand side of (2.8) equals one iff $z \in \{z_1, ..., z_{2m}\}$, we conclude that $r(y)$ is the rational interpolant of type $[m-1|m]$ of $1/\sqrt{y^2-1}$ at the nodes $y_j = T^{-1}(z_j)$, and $R(z) = r_m^{[\nu]}(z)$ is the rational interpolant of type $[m-1|m]$ of $f^{[\nu]}(z)$ at the nodes z_j. In particular, for $z \in \mathbb{E}$,

$$\frac{4|G_{2m}(z)|}{(1+G_{2m}(z))^2} = |1 - \Big(\frac{r_m^{[\nu]}(z)}{f^{[\nu]}(z)}\Big)^2| \leq \frac{4\eta_{2m}}{(1-\eta_{2m})^2} \quad \text{in the general case,}$$

$$\frac{4|G_{2m}(z)|}{(1+G_{2m}(z))^2} \leq \frac{4\,|G_{2m}(z)|}{1+G_{2m}(z)} \leq 2\,|1 - \frac{r_m^{[\nu]}(z)}{f^{[\nu]}(z)}| \leq 4\eta_{2m} \quad \text{in the positive case.}$$

Combining with (2.5) and (2.7), we arrive at the conclusion (2.1) of Theorem 2.1. □

In a later section, it will be necessary to estimate the relative error $1 - r_m^{[\mu]}/f^{[\mu]}$ at a matrix argument A, without computing explicitly $f^{[\mu]}(A)$. We suggest two bounds, the first one following directly from Theorem 2.1 and the observation that $(f^{[\nu]}(A))^{-2}$ is easy to evaluate, leading to a kind of residual for the inverse square root. The second approach, based on a generalization of Theorem 2.1, states that the relative error does not change much if one replaces $f^{[\mu]}$ by $r_{m+m'}^{[\mu]}$ for a modest value of m'. Such a trick was also applied as a heuristic in error estimates for matrix functions times a vector, especially for the case $f(z) = 1/z$ of solving systems on linear equations, see [GM10] and the references therein.

Corollary 2.2. *With the notations of Theorem 2.1, let A be a symmetric matrix with spectrum $\mathbb{E} = \sigma(A)$ out of $[\alpha, \beta]$. Then we have the residual bound*

$$\|I - r_m^{[\mu]}(A)\Big(f^{[\mu]}(A)\Big)^{-1}\| \leq \|I - r_m^{[\nu]}(A)^2 \frac{1}{|\alpha|}(A-\alpha I)(A-\beta I)\|.$$

If in addition $\eta_{2m} \leq (\sqrt{2}-1)^2$, *and*

$$\delta := \frac{4\widetilde{\eta}}{(1-\widetilde{\eta})^2} \in (0,1), \quad \widetilde{\eta} := \max_{z\in\mathbb{E}} \Big| \prod_{j=2m+1}^{2m+2m'} \frac{\varphi(z)-\varphi(z_j)}{1-\varphi(z)\varphi(z_j)} \Big|,$$

then we have the a posteriori bound

$$\|I - r_m^{[\mu]}(A)\Big(f^{[\mu]}(A)\Big)^{-1}\| \leq \frac{1+\delta}{1-\delta} \|I - r_m^{[\mu]}(A)\Big(r_{m+m'}^{[\mu]}(A)\Big)^{-1}\|.$$

Proof. The first claim is an immediate application of Theorem 2.1. For the second one we vary slightly the argument in (2.5) (with m replaced by $m+m'$): instead of taking a general polynomial Q of degree $\leq m+m'$, we take a polynomial $Q = PQ_m$, with P of degree $\leq m'$, and obtain for $z \in \mathbb{E}$ that

$$\Big| \frac{f^{[\mu]}(z) - r_{m+m'}^{[\mu]}(z)}{f^{[\mu]}(z) - r_m^{[\mu]}(z)} \Big| \leq \min_{\deg P \leq m'} \frac{|\widetilde{\omega}(z)|}{P^2(z)} \, \|\frac{P^2}{\widetilde{\omega}}\|_{L^\infty([\alpha,\beta])}, \quad \widetilde{\omega}(z) = \prod_{j=2m+1}^{2m+2m'} (z-z_j).$$

Proceeding as in the proof of Theorem 2.1 we conclude that

$$\Big| \frac{1 - r_{m+m'}^{[\mu]}(z)/f^{[\mu]}(z)}{1 - r_m^{[\mu]}(z)/f^{[\mu]}(z)} \Big| \leq \frac{4\widetilde{\eta}}{(1-\widetilde{\eta})^2} = \delta \in (0,1), \quad |1 - r_m^{[\mu]}(z)/f^{[\mu]}(z)| \leq \frac{4\eta_{2m}}{(1-\eta_{2m})^2} \leq 1,$$

the last inequality following from assumption on η_{2m}. As a consequence,

$$\Big| \frac{1 - r_m^{[\mu]}(z)/f^{[\mu]}(z)}{1 - r_m^{[\mu]}(z)/r_{m+m'}^{[\mu]}(z)} \Big| \leq \Big| \frac{r_{m+m'}^{[\mu]}(z)}{f^{[\mu]}(z)} \Big| \, \frac{|1 - r_m^{[\mu]}(z)/f^{[\mu]}(z)|}{|1 - r_m^{[\mu]}(z)/f^{[\mu]}(z)| - |1 - r_{m+m'}^{[\mu]}(z)/f^{[\mu]}(z)|} \leq \frac{1+\delta}{1-\delta},$$

as required to conclude. □

Remark 2.3. *In order to make the rate of convergence in Theorem 2.1 more explicit (which may guide us in the choice of "good" interpolation points), we need a more precise knowledge on the quantity* $\eta_{2m} = \eta_{2m}(\mathbb{E})$, *what is possible for intervals. Let* $[c,d] \subset \mathbb{R} \setminus [\alpha,\beta]$ *a closed interval containing* $\mathbb{E}$ *(and possibly containing* ∞*), such that* $\eta_{2m}(\mathbb{E}) \leq \eta_{2m}([c,d])$. *Since the composition of two Blaschke factors is a Blaschke factor, it turns out that the rate* η_{2m} *in Theorem 2.1 does not depend on the particular choice of the Moebius map* T, *that is,* φ *is not necessarily normalized at infinity, all we need is that* $T(\mathbb{R}) = \mathbb{R}$, $T(-1) = \alpha$, $T(1) = \beta$, *and* T *is increasing in* $[-1,1]$. *There exists however a unique such* T *which satisfies in addition* $T(1/\kappa) = c$ *and* $T(-1/\kappa) = d$, *where the value of* $\kappa \in (0,1)$ *is uniquely obtained by observing that the cross ratio of four co-linear reals is invariant under linear transformations, that is,*

$$T(-1) = \alpha,\ T(1) = \beta,\ T(\frac{1}{\kappa}) = c,\ T(-\frac{1}{\kappa}) = d, \quad \frac{(c-\alpha)(d-\beta)}{(c-\beta)(d-\alpha)} = \Big(\frac{1+\kappa}{1-\kappa}\Big)^2 =: \frac{1}{k^2}, \tag{2.9}$$

and in addition $T([-1,1]) = [\alpha,\beta]$, $T([\frac{1}{\kappa},-\frac{1}{\kappa}]) = [c,d]$, *where* $[\frac{1}{\kappa},-\frac{1}{\kappa}] = \overline{\mathbb{R}} \setminus (-\frac{1}{\kappa},\frac{1}{\kappa})$. *Using this Moebius map, we get the simplified expression*

$$\eta_{2m}([c,d]) = \max_{w\in[\frac{1}{\lambda},-\frac{1}{\lambda}]} |\prod_{j=1}^{2m} \frac{w-w_j}{1-w_j w}|, \quad w_j = \varphi(z_j), \quad \lambda = \frac{1}{\varphi(c)} = -\frac{1}{\varphi(d)} = \frac{1-\sqrt{k}}{1+\sqrt{k}}. \tag{2.10}$$

This also allows to find configurations of interpolation points which lead to small $\eta_{2m}([c,d])$*: for instance for a Padé approximant at* z_1 *we have that* $z_1 = ... = z_{2m}$, *with the optimal choice* $z_1 = \varphi^{-1}(\infty) \in (c,d)$, *leading to the rate* $\eta_{2m}([c,d]) = \lambda^{2m}$. *The same rate is obtained for the two-point Padé approximant with* $z_1, ..., z_m = c, z_{m+1}, ..., z_{2m} = d$, *we refer to [Bis21] for further details.*

Finally, minimizing (2.10) over all choices of w_j of modulus > 1 leads to the problem of minimal Blaschke products (after the substitution $u = 1/w$) on the interval $[-\lambda,\lambda]$, which has been recently reviewed in [NT15]. We summarize these findings in the following corollary.

Corollary 2.4. *Let $\alpha, \beta, c, d, k, \lambda, \varphi$ with $\mathbb{E} \subset [c,d]$ be as in Remark 2.3. Then the optimal nodes minimizing $\eta_{2m}([c,d])$ are given in terms of Jacobi elliptic functions $sn(\cdot,\cdot)$ [AS64] and the complete elliptic integral $K(\cdot)$ [AS64] by*

$$\frac{1}{\varphi(z_j)} = \frac{1}{w_j} = \lambda\, sn\Big(K(\lambda^2)\big(-1+\frac{2j-1}{2m}\big), \lambda^2\Big) \in (-\lambda, \lambda) \tag{2.11}$$

for $j = 1, 2, ..., 2m$, leading to the a priori bound

$$\|\frac{f^{[\mu]} - r_m^{[\mu]}}{f^{[\mu]}}\|_{L^\infty([c,d])} \leq 8\rho^{2m}/(1-2\rho^{2m})^2, \quad \rho := \exp(\frac{-1}{\mathrm{cap}([\alpha,\beta],[c,d])}), \tag{2.12}$$

provided that $2\rho^{2m} < 1$.

Proof. In [NT15, Problem D, p. 112], the authors recall the following link with the third Zolotarev problem

$$\eta_{2m}([c,d]) = \min_{B \text{ Blaschke of order } 2m} \|B\|_{L^\infty([-\lambda,\lambda])} = \sqrt{\min_{R \in \mathcal{R}_{2m,2m}} \|R\|_{L^\infty([-\lambda,\lambda])}\, \|\frac{1}{R}\|_{L^\infty([1/\lambda,-1/\lambda])}},$$

and give explicitly in [NT15, Section 3.2, p.109] the roots (2.11) of an optimal Blaschke product. Here [BT19, Corollary 3.2] gives us asymptotically sharp upper bound for this Zolotarev number and thus $\eta_{2m}([c,d])$ for optimal z_j as in (2.11), namely

$$\eta_{2m}([c,d]) \leq 2\, \exp(-\frac{m}{\mathrm{cap}([-\lambda,\lambda],[1/\lambda,-1/\lambda])}) = 2\,\rho^{2m}, \tag{2.13}$$

the last equality following by symmetry and by the fact that the logarithmic capacity is invariant under conformal mappings of the underlying doubly connected domain. Combining with Theorem 2.1 and using $\eta_{2m}(\mathbb{E}) \leq \eta_{2m}([c,d])$), we get the claimed inequality (2.12). □

Remark 2.5. *With the notations of Theorem 2.1, we may even slightly improve the statement of Corollary 2.4: combining Theorem 2.1 and (2.13) we find that the relative error for $r_m^{[\mu]}$ is bounded above by the residual error for $r_m^{[\nu]}$, which itself is bounded above by the a priori bound given in (2.12). Given a symmetric matrix A with spectrum $\sigma(A) \subset [c,d]$, we will report in Figure 3.1 and §5 about numerical experiments showing that, due to finite precision, the computed rational interpolants, evaluated on a computer at a matrix argument A, do no longer respect these inequalities, More precisely, the relative error for $r_m^{[\mu]}(A)$ has a quite erratic behavior once the error is no longer smaller than the a priori bound. In order to find the index m corresponding to smallest error, we suggest to compute $r_m^{[\mu]}(A)$ and $r_m^{[\nu]}(A)$ for $m = 1, 2, ..$ and stop one index before the first m where the residual error is larger than five times the a priori bound, that is, where*

$$\|I - r_m^{[\nu]}(A)\frac{1}{|\alpha|}(A-\alpha I)(A-\beta I) r_m^{[\nu]}(A)\| \geq 40\rho^{2m}/(1-2\rho^{2m})^2. \tag{2.14}$$

For the relative error curves presented in Figure 3.1 we have displayed all indices m verifying (2.14) by markers. It turns out that this heuristic stopping criterion, which implicitly assumes that the floating point errors for $r_m^{[\mu]}(A)$ and $r_m^{[\nu]}(A)$ are about the same, does not require the a priori knowledge of $f^{[\mu]}(A)$, and seems to work very well in practice.

Notice that the interpolation points (2.11) are just optimal for our upper bound, but not necessarily for the relative interpolation error. Indeed, one expects sharper bounds to hold if the support of $\mathrm{supp}(\mu)$ consisting for instance of two intervals is a proper subset of $[\alpha,\beta]$, or $\mathbb{E}$ is a proper subset of $[c,d]$. However, if $\mathrm{supp}(\mu) = [\alpha,\beta]$ and μ is regular in the sense of [ST92], then Gonchar [Gon78b, Theorem 1] (see also [ST92, Theorem 6.2.2]) showed that the $2m$th root of the error of best approximation of f on $[c,d]$ in $R_{m-1,m}$ tends to ρ for $m \to \infty$. More precisely, for the particular Markov function $f^{[\mu]}(z) = 1/\sqrt{z}$ with $[\alpha,\beta] = [-\infty, 0]$, inequalities for the relative error of best approximation of $f^{[\mu]}$ in $R_{m-1,m}$ are known since the work of Zolotarev[Zol77], who expressed several extremal problems in terms of Zolotarev numbers, see also [Ach90, p. 147] and the Appendix of [BT19], from which it follows that the relative error is $\leq 4\rho^{2m}$, and behaves like $4\rho^{2m}(1 + o(\rho^{2m})_{m\to\infty})$, see also, e.g., [Bra86, Theorem V.5.5]. Hence, if we want interpolation points which work for any Markov function, the interpolation points (2.11) are optimal up to a factor at most 2.

2.2 The disk case

With interpolation points $z_1, ..., z_{2m}$ as before, we still keep the integral representation (2.3) for the error for complex z, but (2.4) and (2.5) are no longer true, and need to be adapted following the technique of the Freud Lemma [Fre71, section III.7]. This result has also been exploited in the work of Ganelius [Gan82] and of Braess [Bra87, Theorem 2.1]. The work of these authors on interpolation points with even multiplicity did inspire us in the proof of the following theorem, but we had to correct an erroneous application of the Freud Lemma in [Bra87, Eqn. (2.7)] where an additional $\beta - \alpha$ factor should occur, and could improve [Bra87, Eqn. (2.7)] by a factor 2. We also give an estimate for interpolation points of arbitrary multiplicity which to our knowledge is new.

Theorem 2.6. *Let $-\infty \le \alpha < \beta < -1$, and let the Markov function $f^{[\mu]}$ be as in (1.1). Consider interpolation points $z_1, ..., z_{2m} \in \mathbb{C} \setminus [\alpha, \beta]$ where we suppose that non-real points only occur in conjugate pairs. If all interpolation points have even multiplicity, then*[7]

$$\|f^{[\mu]} - r_m^{[\mu]}\|_{L^\infty(\overline{\mathbb{D}})} \le C \max_{z\in[\alpha,\beta]} \Big|\prod_{j=1}^{2m} \frac{1 - zz_j}{z - z_j}\Big|, \quad C = \frac{1-\beta}{-1-\beta}\, f(-1).$$

In the general case,

$$\|f^{[\mu]} - r_m^{[\mu]}\|_{L^\infty(\overline{\mathbb{D}})} \le C\, \frac{4\eta'_{2m}}{(1-\eta'_{2m})^2}, \quad \eta'_{2m} = \max_{z\in\overline{\mathbb{D}}} G_{2m}(z)$$

where C is as before, and G_{2m} as in Theorem 2.1.

Proof. Define ω as in (2.2). Let Q be any polynomial of degree at most m with real coefficients, then $x \mapsto \frac{Q(x)/Q(z)-1}{z-x}$ is a polynomial of degree at most $m-1$, and thus orthogonal to Q_m with respect to the measure μ/ω. This leads to the well-known fact that

$$f^{[\mu]}(z) - r_m^{[\mu]}(z) = \frac{\omega(z)}{Q_m^2(z)} \int \frac{Q_m^2(x)}{\omega(x)} \frac{d\mu(x)}{z-x} = \frac{\omega(z)}{Q_m(z)} \int \frac{Q_m(x)}{\omega(x)} \frac{d\mu(x)}{z-x} = \frac{\omega(z)}{Q_m(z)Q(z)} \int \frac{Q_m(x)Q(x)}{\omega(x)} \frac{d\mu(x)}{z-x}.$$

We now observe that

$$\text{for } |z| = 1 \text{ and } x \le \beta < -1: \quad \frac{1}{1-x} \le \mathrm{Re}\Big(\frac{1}{z-x}\Big) \le \frac{1}{|z-x|} \le \frac{1}{-1-x} \le \frac{1}{1-x}\frac{1-\beta}{-1-\beta}, \tag{2.15}$$

and apply the Cauchy-Schwarz inequality in the last integral in order to obtain

$$\begin{aligned} |f^{[\mu]}(z) - r_m^{[\mu]}(z)|^2 &\le \Big|\frac{\omega(z)}{Q_m^2(z)}\Big| \int \frac{Q_m^2(x)}{\omega(x)} \frac{d\mu(x)}{|z-x|} \Big|\frac{\omega(z)}{Q^2(z)}\Big| \int \frac{Q^2(x)}{\omega(x)} \frac{d\mu(x)}{|z-x|}. \\ &\le \frac{1-\beta}{-1-\beta} |f^{[\mu]}(z) - r_m^{[\mu]}(z)| \Big|\frac{\omega(z)}{Q^2(z)}\Big| \int \frac{Q^2(x)}{\omega(x)} \frac{d\mu(x)}{|z-x|}, \end{aligned}$$

the second inequality following from (2.15). Thus we get for $|z| = 1$ the following upper bounds for the absolute and relative interpolation errors

$$\begin{aligned} &|f^{[\mu]}(z) - r_m^{[\mu]}(z)| \le C \min_{\deg Q \le m} \|\frac{\omega}{Q^2}\|_{L^\infty(\partial\mathbb{D})} \|\frac{Q^2}{\omega}\|_{L^\infty([\alpha,\beta])}, \\ &\Big|1 - \frac{r_m^{[\mu]}(z)}{f^{[\mu]}(z)}\Big| \le (\frac{1-\beta}{-1-\beta})^2 \min_{\deg Q \le m} \|\frac{\omega}{Q^2}\|_{L^\infty(\partial\mathbb{D})} \|\frac{Q^2}{\omega}\|_{L^\infty([\alpha,\beta])}, \end{aligned} \tag{2.16}$$

the second following from the first since $C = \frac{1-\beta}{-1-\beta}\, f^{[\mu]}(-1) \le (\frac{1-\beta}{-1-\beta})^2\, \mathrm{Re} f^{[\mu]}(z) \le (\frac{1-\beta}{-1-\beta})^2\, |f^{[\mu]}(z)|$, again by (2.15).

In the case of interpolation points of even multiplicity, say, $z_{m+j} = z_j$ for $j = 1, ..., m$, we get the upper bound claimed in Theorem 2.6 by taking $Q(x) = (1 - z_1 z)...(1 - z_m z)$, which can be shown using the maximum principle for analytic functions to be the extremal polynomial in (2.16) in this special case. Finally, in the general case we use the same polynomial for the interval $[\alpha, \beta]$ as in the proof of Theorem 2.1, which can be shown to be optimal for (2.16) up to the factor $4/(1-\eta'_{2m})^2$. □

[7] Notice that for the critical case β close to -1 and μ a probability measure, the above constant C behaves at worst as $1/\mathrm{dist}(\overline{\mathbb{D}}, [\alpha, \beta]) = 1/(-1-\beta)$, a term which also occurs in the works of Ganelius [Gan82] and Braess [Bra87].

Remark 2.7. *As in the previous chapter, one may ask for a single optimal interpolation point $z_1 = ... = z_{2m}$, or for a configuration of distinct points minimizing η'_{2m}. Here the results of the previous chapter remain valid, by choosing $[c,d] = [1/\beta, 1/\alpha]$ in (2.9), we refer the reader to [Bis21] for further details. As a rule of thumb, "good" interpolation points are in $[1/\beta, 1/\alpha]$.*

In contrast, in his study of Faber-Padé approximants [Kni09], Knizhnerman considered the special case $z_1 = ... = z_{2m} = 0$ and hence $\eta'_{2m} = (1/\beta)^{2m}$. In this or in the more general case $z_j \in [1/\alpha, 0]$, the error analysis simplifies considerably: since the maximum of the error $|f^{[\mu]} - r_m^{[\mu]}|$ on the unit circle can be shown to be attained at $z = -1$, and the same is true for the statement of Theorem 2.6. In addition, the factor $\frac{1-\beta}{-1-\beta}$ can be dropped.

Remark 2.8. *Braess [Bra87] used the Carathéodory-Fejér method to derive $L^\infty([-1,1])$ estimates from $L^\infty(\overline{\mathbb{D}})$ estimates for the interpolation error. Comparing our Theorems 2.1 and 2.6, our interval estimates seem to be sharper.*

More generally, for a general convex compact set $\mathbb{E}$ being symmetric with respect to the real axis, one may use the Faber map $\mathcal{F}$ (see [Gai87] or [BR09]) and its modification $\mathcal{F}_+(h) = \mathcal{F}(h) + h(0)$ to get good rational approximants on $\mathbb{E}$ from those on $\mathbb{D}$. More precisely, Ellacott [Ell83, Theorem 1.1] showed that $r \in \mathcal{R}_{m-1,m}$ iff $\mathcal{F}(r) \in \mathcal{R}_{m-1,m}$ and simultaneously [Kni09] and [BR09] found out that the Faber pre-image of a Markov function is a Markov function, with explicit formulas for the measure. Thus one may use the inequality

$$\|\mathcal{F}_+(f^{[\mu]}) - \mathcal{F}_+(r_m^{[\mu]})\|_{L^\infty(\mathbb{E})} \leq 2\,\|f^{[\mu]} - r_m^{[\mu]}\|_{L^\infty(\overline{\mathbb{D}})}$$

shown in [Gai87, Theorem 2] together with our findings in Theorem 2.6 to find good rational approximants on $\mathbb{E}$ for Markov functions. In the context of Markov functions of matrices, we should also mention the result [BR09, Theorem 2.1] that

$$\|\mathcal{F}_+(f^{[\mu]})(A) - \mathcal{F}_+(r_m^{[\mu]})(A)\| \leq 2\,\|f^{[\mu]} - r_m^{[\mu]}\|_{L^\infty(\overline{\mathbb{D}})} \tag{2.17}$$

provided that the field of values of the square matrix A is a subset of $\mathbb{E}$.

To summarize, in §2 we presented a detailed study on the relative error (in exact arithmetic) obtained by approaching $f^{[\mu]}$ by $r_m^{[\mu]}$ with $r_m^{[\mu]}$ a rational interpolant of $f^{[\mu]}$ of type $[m-1|m]$ at quite arbitrary interpolation points. This allowed us to find quasi-optimal interpolation points which minimize our upper bound for the error, together with the explicit and simple a priori bound (2.12). Also, several new a posteriori error bounds are provided, which allowed to derive in Remark 2.5 a new heuristic stopping criterion for finding the degree m leading to a small relative error even in the context of floating point arithmetic. However, for a successful implementation we need to discuss how to represent and compute our interpolants.

3 The computation of rational interpolants for distinct real interpolation nodes

The computation of a rational interpolant $r_m = P_m/Q_m$ of type $[m-1|m]$ or $[m|m]$ of f or its evaluation at some argument $z \in [c,d]$ is strongly connected to the way how we represent our rational interpolant. We will suppose in what follows our (finite or infinite sequence of) interpolation points z_j for $j \geq 1$ are distinct, real, and ordered such that

$$\beta < c \leq z_1 < z_2 < ... \leq d. \tag{3.1}$$

3.1 Computing separately numerator and denominator

A first perhaps naive approach would be to represent both numerator and denominator in the same polynomial basis, and then solve for the coefficients in this basis by writing a homogeneous system of $2m$ equations and $2m+1$ unknowns translating the interpolation conditions $f(z_j)Q_m(z_j) - P_m(z_j) = 0$ for $j = 1, ..., 2m$. Using the basis of monomials, interesting complexity results for evaluating $P_m(A)$ and $Q_m(A)$ are given in [Fas19]. However, in practice the underlying matrix of coefficients turns out to be quite often very ill-conditioned, there is a rule of thumb for Padé approximants using monomials [BGM96, Section 2.1] that we might loose at least m decimal digits of precision in solving such systems. Of course,

other (scaled) polynomial bases, like Chebyshev polynomials scaled to the interval $[c,d]$ or a Newton basis corresponding to a suitable ordering of the interpolation points might lead to better conditioning, but a "good" basis should not only depend on $[c,d]$ but also on the function f to be interpolated. Thus we have not implemented such an approach.

3.2 Computing poles and residuals in a partial fraction decomposition

Since for Markov functions f the interpolant r_m has m simple poles $x_1, ..., x_m$ in (α, β), we may look directly for the partial fraction decomposition

$$r_m(z) = \frac{a_1}{z - x_1} + ... + \frac{a_m}{z - x_m}. \tag{3.2}$$

By the work of Mayo and Antoulas [MA07] nicely summarized in the recent paper [EI19, Section 2], r_m may be represented as a transfer function of a SISO dynamical system with help of a matrix pencil: we have that $r_m(z) = W(\mathbb{L}_s - z\mathbb{L})^{-1}V^T$ with the row vectors $W = (f(z_{2j-1}))_{j=1,...,m}$, $V = (f(z_{2j}))_{j=1,...,m}$ and the Loewner matrices

$$\mathbb{L} = \left(\frac{f(z_{2j}) - f(z_{2k-1})}{z_{2j} - z_{2k-1}}\right)_{\substack{j=1,...,m \\ k=1,...,m}} \quad \text{and} \quad \mathbb{L}_s = \left(\frac{z_{2j}f(z_{2j}) - z_{2k-1}f(z_{2k-1})}{z_{2j} - z_{2k-1}}\right)_{\substack{j=1,...,m \\ k=1,...,m}}.$$

Thus the poles are the eigenvalues of the Loewner matrix pencil $\mathbb{L}_s - z\mathbb{L}$ and can be computed with standard software, we used the Matlab function `eig`. We then compute the residuals by a least square fitting,[8] see Algorithm 3.1.

Algorithm 3.1: Given a function f and interpolation points $z_1, ..., z_{2m}$, compute poles z_j and residuals a_j of the partial fraction decomposition (3.2) of the rational interpolant of f of type $[m-1|m]$.

Result: Poles $x_1, ..., x_m$ and residuals $a_1, ..., a_m$ in (3.2).
begin
Find the eigenvalues $x_1, ..., x_m$ of the Loewner matrix pencil $\mathbb{L}_s - z\mathbb{L}$ of Mayo and Antoulas;
Compute the solution $y = (a_1, ..., a_m)^T$ of the least square problem of minimizing

$$\| \left(\frac{1}{z_j - x_k}\right)_{j=1,...,2m, k=1,...,m} y - \left(f(z_j)\right)_{j=1,...,2m} \|$$

.
end

3.3 Barycentric rational functions

The barycentric representation [BBM05] of a rational function $r \in \mathcal{R}_{m,m}$ with distinct support points $t_0, ..., t_m$ is given by

$$r(z) = \sum_{j=0}^{m} \frac{\alpha_j}{z - t_j} \Bigg/ \sum_{j=0}^{m} \frac{\beta_j}{z - t_j}, \tag{3.3}$$

we refer the reader to [DH04, p.551] and [SC08, Proposition 2.4.3] and the discussion in [FNTB18, Section 2.3] for backward and forward stability results on evaluating such rational functions. For constructing rational interpolants of type $[m|m]$ of f, one typically choses $\alpha_j = f(t_j)\beta_j$ ensuring that r interpolates f at these support points. Such interpolating rational functions with well-chosen support points have been the building block for a new implementation *minimax* of the rational Remez algorithm [FNTB18], which allows to compute best rational approximants of type $[m'|m]$ of Markov functions for $m', m \leq 40$ to machine precision in double precision arithmetic where previous implementations required high precision arithmetic to achieve this goal.

[8] However, the underlying rectangular Cauchy matrix $(\frac{1}{z_j - x_k})$ might be quite ill-conditioned, even after row or column scaling [BT19, Cor 4.2]. A closer analysis seems to show that the given bound for the inverse condition number is related to the rate of best rational approximants of our f on the interval $[c,d]$.

For computing the rational interpolant of type $[m|m]$, we choose $t_j = z_{2j+1}$ for $j = 0, ..., m$, and it remains to solve a homogeneous linear system for $\beta_0, ..., \beta_m$ in order to get also interpolation at the points z_{2k} for $k = 1, ..., m$, see Algorithm 3.2. Notice that the underlying matrix of coefficients is the transposed of the Loewner matrix $\mathbb{L}$ seen in §3.2, bordered with one additional column. For ensuring stability,

Algorithm 3.2: Given a function f and interpolation points $z_1, ..., z_{2m+1}$, compute support points t_j and weights β_j, $\alpha_j = f(t_j)\beta_j$ of the rational interpolant (3.3) of f of type $[m|m]$.

Result: For $j = 0, 1, ..., m$: Support points t_j and weights β_j, $\alpha_j = f(t_j)\beta_j$ in (3.3).
begin
Define support points $t_j = z_{2j+1}$ for $j = 0, 1, ..., m$;
Compute a solution $y = (\beta_0, ..., \beta_m)^T$ of the homogeneous system of linear equations

$$\Big(\frac{f(z_{2k}) - f(t_j)}{z_{2k} - t_j}\Big)_{k=1,2,...,m,j=0,1,...,m} y = 0.$$

end

numerical experiments show that it is mandatory that the support points and the other interpolation points interlace, see (3.1). This is supported in [FNTB18, Cor 4.5] saying that, after suitable explicit column and row scaling, the Cauchy matrix $(\frac{1}{z_{2j} - z_{2k-1}})_{j,k}$ is unitary provided that we have the interlacing (3.1).

Things become slightly more technical for the interpolant of type $[m-1|m]$, here we have taken the support points $t_0 = z_1$ and $t_j = z_{2j}$ for $j = 1, ..., m$ ensuring nearly interlacing with the remaining interpolation points. Again we have to solve a homogeneous linear system for $\beta_0, ..., \beta_m$ to impose interpolation at z_{2k-1} for $k = 2, ..., m$, where the additional equation $f(t_0)\beta_0 + ... + f(t_m)\beta_m = 0$ ensures that the degrees are correct.

3.4 Thiele continued fractions

We finally turn to a representation of rational interpolants through continued fractions. Following [BGM96, Section 7.1], for given interpolation points $z_1, z_2, ...$ and parameters $f_1^{(1)}, f_2^{(2)}, ... \in \mathbb{C}$, the Mth convergent of a Thiele continued fraction is the rational function

$$R_M^{(1)}(z) = f_1^{(1)} + \cfrac{z - z_1}{\Big|\, f_2^{(2)}} + ... + \cfrac{z - z_{M-1}}{\Big|\, f_M^{(M)}}. \tag{3.4}$$

We refer to a positive Thiele fraction if all $f_j^{(j)}$ are strictly positive, and (3.1) holds. Given a function $f^{(1)}$, we define its reciprocal differences by

$$\forall 1 \le k \le M: \quad f_k^{(1)} = f^{(1)}(z_k), \quad \forall 1 \le j < k \le M: \qquad f_k^{(j+1)} = \frac{z_k - z_j}{f_k^{(j)} - f_j^{(j)}}, \tag{3.5}$$

where we tacitly suppose that there is no breakdown (that is, no division by 0). Then $R_M^{(1)}(z_k) = f^{(1)}(z_k)$ for $k = 1, ..., M$, more precisely, $R_{2m+1}^{(1)}$ is the rational interpolant of type $[m|m]$ of $f^{(1)}$ at the interpolation points $z_1, ..., z_{2m+1}$, and $R_{2m}^{(1)}$ is the rational interpolant of type $[m|m-1]$ of $f^{(1)}$ at the interpolation points $z_1, ..., z_{2m}$. Setting $f^{(1)}(z) = 1/f(z)$, we conclude that $1/R_{2m}^{(1)}$ is the desired rational interpolant of type $[m-1|m]$ of f. This interpolation property becomes immediate by introducing the families of functions

$$\forall 1 \le j < k \le M: \quad f^{(j+1)}(z) = \frac{z - z_j}{f^{(j)}(z) - f^{(j)}(z_j)}, \quad R_M^{(j+1)}(z) = \frac{z - z_j}{R_M^{(j)}(z) - R_M^{(j)}(z_j)}, \tag{3.6}$$

since then $f_k^{(j)} = f^{(j)}(z_k) = R_M^{(j)}(z_k)$ for $1 \le j \le k \le M$, and

$$R_M^{(j)}(z) = f_j^{(j)} + \cfrac{z - z_j}{\Big|\, f_{j+1}^{(j+1)}} + ... + \cfrac{z - z_{M-1}}{\Big|\, f_M^{(M)}}.$$

In particular, $R_M^{(j)}$ is a rational interpolant of $f^{(j)}$ at the interpolation points $z_j, z_{j+1}, ..., z_M$. The backward evaluation scheme at a fixed argument z of a Thiele continued fraction given the parameters $f_j^{(j)}$ is given by

$$R_M^{(M)}(z) = f_M^{(M)}, \quad \text{and for } \; j = M-1, M-2, ..., 1: \quad R_M^{(j)}(z) = f_j^{(j)} + \frac{z - z_j}{R_M^{(j+1)}(z)}. \tag{3.7}$$

In his stability analysis of this scheme, Graves-Morris [GM81] observed that, before computing $f_k^{(j+1)}$ for $k = j+1, ..., M$ via (3.5), it is important to reorder the couples $(f_k^{(j)}, z_k)$ for $k = j, j+1, ..., M$ such that, after reordering,

$$|f_j^{(j)}| = \min\{|f_k^{(j)}| : k = j, j+1, ..., M\}, \tag{3.8}$$

reminding of partial pivoting in Gaussian elimination. A combination of (3.5), (3.8), and (3.7) gives the modified Thacher-Tukey algorithm of [GM81] and [BGM96, Section 7.1], which we have simplified a bit by omitting the case of breakdown in (3.5), see Algorithm 3.3.

Algorithm 3.3: Given a function $f^{(1)}$ and interpolation points $z_1, ..., z_M$, compute and evaluate at $z \in \mathbb{C}$ via the modified Thacher-Tukey algorithm of [GM81] the Thiele continued fraction representation (3.4) of the rational interpolant of $f^{(1)}$ of type $[m|m-1]$ (if $M = 2m$) or of type $[m|m]$ (if $M = 2m+1$).

Result: Coefficients $f_1^{(1)}, ..., f_M^{(M)}$ in (3.4) and value $R_M^{(1)}(z)$ of the interpolant.
begin
  for $k = 1, 2, ..., M$ **do**
    initialize $f_k^{(1)} = f^{(1)}(z_k)$;
  end
  for $j = 1, 2, ..., M-1$ **do**
    Permute $(f_k^{(j)}, z_k)$ for $k = j, j+1, ..., M$ such that, after reordering, (3.8) holds;
    for $k = j+1, j+2, ..., M$ **do**
      $f_k^{(j+1)} = (z_k - z_j)/(f_k^{(j)} - f_j^{(j)})$;
    end
  end
  initialize $R_M^{(M)}(z) = f_M^{(M)}$;
  for $j = M-1, M-2, ..., 1$ **do**
    $R_M^{(j)}(z) = f_j^{(j)} + \frac{z - z_j}{R_M^{(j+1)}(z)}$;
  end
end

Notice that if $R_M^{(1)}$ is a positive continued fraction then by recurrence on $k - j$ using (3.1) and (3.5) one shows that $0 < f_j^{(j)} < f_k^{(j)}$ for $1 \le j < k \le M$, that is, there is no breakdown in (3.5), and we obtain (3.8) without pivoting. However, we are not aware of results in the literature on classes of functions where the interpolating Thiele continued fraction is positive. Such a class is given in our first main result, the proof is presented later.

Theorem 3.1. *If this is true for $1/f^{(1)}$, then all functions $1/f^{(j)}$ defined in* (3.6) *are Markov functions with a measure $\mu^{(j)}$ having an infinite support $\subset [\alpha, \beta]$.*

Since a Markov function as in Theorem 3.1 is positive and decreasing in $(\beta, +\infty)$, we conclude with (3.1) that $f^{(j)}(z_k) > f^{(j)}(z_j) > 0$ for $k > j$, that is, the interpolating Thiele continued fraction of $f^{(1)}$ is positive.

Example 3.2. *Take $f^{(1)}(z) = \sqrt{z}$ such that $1/f^{(1)}(z)$ is a Markov function with support $[\alpha, \beta] = (-\infty, 0]$, a limiting case of* (1.2). *Then the reader easily verifies by recurrence that $f_k^{(1)} = \sqrt{z_k}$ and, for $j \ge 2$,*

$$f^{(j)}(z) = \sqrt{z} + \sqrt{z_{j-1}}, \quad f_k^{(j)} = \sqrt{z_k} + \sqrt{z_{j-1}} > 0.$$

In particular, also $1/f^{(j)}(z)$ is a Markov function with support $[\alpha,\beta]=(-\infty,0]$, and the interpolating Thiele continued fraction

$$\sqrt{z}=\sqrt{z_1}+\cfrac{z-z_1}{\sqrt{z_2}-\sqrt{z_1}}+\cfrac{z-z_2}{\sqrt{z_3}-\sqrt{z_2}}+...$$

is positive. We have not seen before such an explicit formula for the interpolating Thiele continued fraction, only the limiting case of Padé approximants, see, e.g., [Hig08, Theorem 5.9].

We now state and prove our second main result of this subsection on the backward stability of the modified Thacher-Tukey algorithm: an error in finite precision in (3.5) gives parameters of a continued fraction with exact values at z_k not far from the desired values $f^{(1)}(z_k)$, provided that we use the standard model [Hig02, Eqn. (2.4)] for finite precision arithmetic between real machine numbers. In our proof of this result we have been inspired by a similar result [GM80, Theorem 4.1] of Graves-Morris, who considered non necessarily positive Thiele continued fractions with pivoting (3.8), made a first order error analysis and got an additional growth factor 2^k for the error which we are able to eliminate.

Theorem 3.3. *Let $1/f^{(1)}$ be a Markov function as before, and suppose that the quantities $\widetilde{f}_k^{(j)}$ for $1\le j\le k\le M$ are computed via (3.5) using finite precision arithmetic with machine precision ε. Denote by $\widetilde{R}_M^{(1)}$ the (exact) continued fraction constructed with the (inexact) parameters $\widetilde{f}_1^{(1)},...,\widetilde{f}_M^{(M)}$ which are supposed to be >0 (despite finite precision, see Remark 3.4). Then*

$$k=1,...,M:\quad |\widetilde{R}_M^{(1)}(z_k)-f^{(1)}(z_k)|\le\frac{3k\varepsilon}{1-3k^2\varepsilon}.|\widetilde{R}_M^{(1)}(z_k)|.$$

Proof. The standard model for finite precision arithmetic of [Hig02, Eqn.(2.4)] gives the following finite precision counterpart of (3.5): for $k>j$

$$f^{(1)}(z_k)=\widetilde{f}_k^{(1)}(1+\epsilon_{1,k}),\quad \widetilde{f}_k^{(j+1)}=\frac{z_k-z_j}{\widetilde{f}_k^{(j)}-\widetilde{f}_j^{(j)}}(1+\epsilon_{j+1,k}),$$

where $\epsilon_{1,k}$ comes from rounding $f^{(1)}(z_k)$, the term $\epsilon_{j+1,k}$ translates errors in the two subtractions and the division, and $|\epsilon_{j,k}|\le\frac{3\varepsilon}{1-3\varepsilon}$ by [Hig02, Lemma 3.1]. In accordance to (3.7), we consider the rational functions defined by

$$\widetilde{R}_M^{(M)}(z)=\widetilde{f}_M^{(M)},\quad \text{and for}\ \ j=M-1,M-2,...,1:\quad \widetilde{R}_M^{(j)}(z)=\widetilde{f}_j^{(j)}+\frac{z-z_j}{\widetilde{R}_M^{(j+1)}(z)},$$

and claim that

$$\widetilde{f}_k^{(j)}=(1+\delta_{j,k})\widetilde{R}_M^{(j)}(z_k),\quad |\delta_{j,k}|\le\gamma_{k-j},\quad \gamma_\ell=\frac{3\ell\varepsilon}{1-3\ell^2\varepsilon}. \tag{3.9}$$

We argue by recurrence on $k-j$ and notice that the case $k=j$ is trivial since $\widetilde{R}_M^{(j)}(z_j)=\widetilde{f}_j^{(j)}$ by definition. In case $k>j$ we may write

$$\begin{aligned}&\widetilde{f}_k^{(j)}-\widetilde{R}_M^{(j)}(z_k)=\widetilde{f}_k^{(j)}-\widetilde{f}_j^{(j)}-(\widetilde{R}_M^{(j)}(z_k)-\widetilde{f}_j^{(j)})\\&=\frac{z_k-z_j}{\widetilde{f}_k^{(j+1)}}(1+\epsilon_{j+1,k})-(\widetilde{R}_M^{(j)}(z_k)-\widetilde{f}_j^{(j)})=\Big(\frac{1+\epsilon_{j+1,k}}{1+\delta_{j+1,k}}-1\Big)(\widetilde{R}_M^{(j)}(z_k)-\widetilde{f}_j^{(j)}).\end{aligned}$$

Our claim (3.9) then follows by observing[9] that $|\widetilde{R}_M^{(j)}(z_k)-\widetilde{f}_j^{(j)}|=\widetilde{R}_M^{(j)}(z_k)-\widetilde{R}_M^{(j)}(z_j)\le\widetilde{R}_M^{(j)}(z_k)$ by assumption $\widetilde{f}_j^{(j)}>0$ for $j=1,...,M$, and by the inequality

$$|\frac{1+\epsilon_{j+1,k}}{1+\delta_{j+1,k}}-1|\le\frac{|\epsilon_{j+1,k}|+\gamma_{k-j-1}}{1-\gamma_{k-j-1}}\le\gamma_{k-j}.$$

In a similar manner, we deduce the assertion of the Theorem from (3.9) for $j=1$. □

[9] Without this positivity assumption, [GM80, Theorem 4.1] observed with (3.8) that, up to $\mathcal{O}(\varepsilon)$, we have that $|\widetilde{R}_M^{(j)}(z_k)-\widetilde{f}_j^{(j)}|\approx|\widetilde{f}_k^{(j)}-\widetilde{f}_j^{(j)}|\le|\widetilde{f}_k^{(j)}|+|\widetilde{f}_j^{(j)}|\le 2|\widetilde{f}_k^{(j)}|\approx 2\,|\widetilde{R}_M^{(j)}(z_k)|$, leading to some exponentially increasing growth factor.

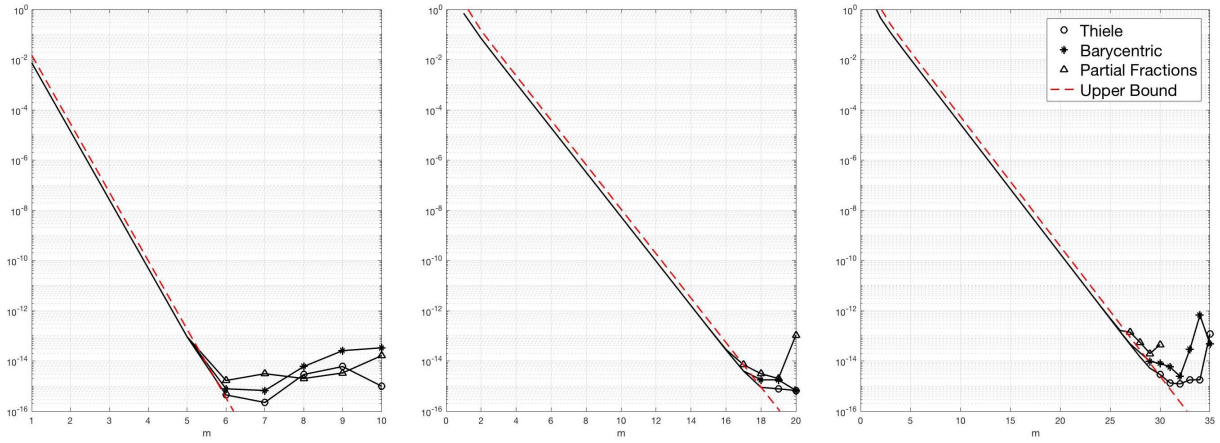

Figure 3.1: Relative L^∞ error on the interval $[c,d]$ of rational interpolants of type $[m-1|m]$ of the Markov function $f(z)=1/\sqrt{z}$, with $\alpha=-\infty,\beta=0$, $d=1$ and $c\in\{1/2,10^{-3},10^{-6}\}$ (from the left to the right). For each c and m, we take the quasi-optimal interpolation points of (2.11) (depending on m and α,β,c,d), and show the relative error of the same rational interpolant (black solid line), computed with three different methods: the partial fraction decomposition of §3.2 (triangle markers), the barycentric representation of §3.3 (star markers) and finally the Thiele interpolating continued fraction of §3.4 (circle markers). The fourth graph (red dashed) gives the a priori upper bound (2.12) of Corollary 2.4.

Remark 3.4. *Extensive numerical experiments showed us that the parameters $\widetilde{f}_k^{(j+1)}$ of Theorem 3.3 only fail to be positive if the error $R_j^{(1)}(z)-f^{(1)}(z)$ for $z\in[c,d]$ is already close to machine precision. To prove such a statement, one requires a (rough) forward stability result on $\widetilde{f}_k^{(j)}-f_k^{(j)}$ which seems to be possible but quite involved, we omit details.*

Example 3.5. *In Figure 3.1 we represent the relative $L^\infty([c,d])$ error of the same interpolants r_m for the same Markov function $f(z)=1/\sqrt{z}$ and interpolation points (2.11) depending on m, computed with the three different methods discussed so far. Here we have discretized $[c,d]$ by 500 cosine points, the entries of some diagonal matrix A. Recall that, in exact arithmetic, all curves should have identical behavior, and stay below the a priori upper bound (2.12). However, in finite precision arithmetic we observe that, once the method and the value of c is fixed, the corresponding error polygon crosses the upper bound once and, afterwards, does hardly decrease, and sometimes even increases. We use markers on the error curves for indices m which have been rejected by our stopping criterion of Remark 2.5. If we denote by m' the index such that $m'+1$ is the first rejected index, the error of m' is below the a priori bound, and the crossing happens between the indices m' and $m'+1$. Also, we observe without theoretical evidence that the error for any $m>m'$ is never smaller than $1/10$ times the error for m'. This confirms that our stopping criterion works well in practice. Notice that, for any of the three methods, the final relative error is about the same size (not far from machine precision), and increases only modestly with d/c. In Section 5 we will see that such a behavior is no longer true if we evaluate our interpolants at general matrix arguments instead of scalar arguments.*

It still remains to present a proof of Theorem 3.1 which will be based on the following Lemma which is partly known from the classical Stieltjes moment problem up to a change of variables, see for instance [BGM96, Sections 5.2 et 5.3] or [Bra86, Thm V.4.4]. In the remainder of this section we suppose that $\alpha<\beta<z_0$. For a function g analytic in some neighborhood of z_0 the Hankel matrices are defined with help of the Taylor coefficients of g at z_0

$$\mathcal{H}_n^{(\ell)}(g)=\begin{bmatrix} g_\ell & g_{\ell+1} & \cdots & g_{n+\ell} \\ g_{\ell+1} & g_{\ell+2} & \cdots & g_{n+\ell+1} \\ \vdots & \vdots & & \vdots \\ g_{n+\ell} & g_{n+\ell+1} & \cdots & g_{2n+\ell} \end{bmatrix}, \quad g(z)=\sum_{j=0}^{\infty} g_j(z-z_0)^j. \tag{3.10}$$

The following lemma will be applied for $z_0\in\{z_1,z_2,...\}$ with the $z_1,z_2,..$ as in (3.1).

Lemma 3.6. *If f is a Markov function with measure μ having an infinite support included in $[\alpha,\beta]$ then, for all $n \geq 0$, the Hankel matrices $\mathcal{H}_n^{(0)}(f)$ are positive definite, and the Hankel matrices $\mathcal{H}_n^{(1)}(f)$ are negative definite.*
Conversely, if f is analytic in $\mathbb{C} \setminus [\alpha,\beta]$, with Hankel matrices $\mathcal{H}_n^{(0)}(f)$ positive definite and $\mathcal{H}_n^{(1)}(f)$ negative definite for all $n \geq 0$, then f is a Markov function with measure μ having an infinite support included in $[\alpha,\beta]$.

Proof. A proof of the first part is elementary, noticing that the Taylor coefficients are moments of μ given by $f_j = \int \frac{d\mu(x)}{(z_0-x)(x-z_0)^j}$, leading to an integral expression of $y^T \mathcal{H}_n^{(\ell)}(f) y$ for any $y \in \mathbb{R}^{n+1}$ with a unique sign depending only on the parity of ℓ provided that $y \neq 0$, for details see [Bis21].
To show the converse implication, one considers $r_m = p_m/q_m$ being the Padé approximant of type $[m-1|m]$ of f at z_0. The sign assumption on the Hankel determinants allows to conclude that q_m has a determinant representation $q_m(z) = \det\left(\mathcal{H}_{m-1}^{(0)}(f) - (z-z_0)\mathcal{H}_{m-1}^{(1)}(f)\right)$. Moreover, there is a three term recurrence between three consecutive denominators with the sign of the coefficients being known. An additional Sturm sequence argument allows us to conclude that $r_m = p_m/q_m$ has m distinct poles $x_{1,m}, \ldots x_{m,m} \in (-\infty, z_0)$ and positive residuals $a_{j,m}$, that is

$$r_m(z) = \frac{p_m(z)}{q_m(z)} = \int \frac{d\mu_m(x)}{z-x}, \quad \mu_m = \sum_{j=1}^{m} a_{j,m}\delta_{x_{j,m}}. \tag{3.11}$$

As in [Bra86, Proof of Thm. V.4.4], there exists a subsequence $(\mu_{m_\ell})_\ell$ of $(\mu_m)_m$ having the weak-star limit $\widetilde{\mu}$, $\mathrm{supp}(\widetilde{\mu}) \subset (-\infty, z_0]$, and for all $k \geq 0$

$$\lim_{\ell\to\infty} \int \frac{d\mu_{m_\ell}(x)}{(z_0-x)^{k+1}} = \int \frac{d\widetilde{\mu}(x)}{(z_0-x)^{k+1}} = (-1)^k \frac{g^{(k)}(z_0)}{k!},$$

with the Markov function $g(z) = \int \frac{d\widetilde{\mu}(x)}{z-x}$. From the interpolation conditions of a Padé approximant of f at z_0 we also know that, for $k \leq 2m$,

$$\int \frac{d\mu_m(x)}{(z_0-x)^{k+1}} = (-1)^k \frac{r_m^{(k)}(z_0)}{k!} = (-1)^k \frac{f^{(k)}(z_0)}{k!}.$$

Combining these two relations we find that $g^{(k)}(z_0)$ is finite, and $g^{(k)}(z_0) = f^{(k)}(z_0)$ for all $k \geq 0$. In particular, with f also g is analytic in a neighborhood U of z_0, and $f(z) = g(z)$ for all $z \in U$. Recalling that by assumption f is analytic in $\mathbb{C} \setminus [\alpha,\beta]$, we see that our Markov function g for the measure $\widetilde{\mu}$ has an analytic continuation f in $\mathbb{C} \setminus [\alpha,\beta]$, and hence $\mathrm{supp}(\widetilde{\mu}) \subset [\alpha,\beta]$. Thus also the converse implication is true. □

Proof of Theorem 3.1. We only need to show this statement for $j = 1$. Let $f^{(1)} = 1/f$, with f a Markov function with measure μ having an infinite support included in $[\alpha,\beta]$. Since $f(z) \neq 0$ for $z \notin [\alpha,\beta]$, we conclude that $f^{(1)}$ is analytic in $\mathbb{C} \setminus [\alpha,\beta]$, and that the same is true for

$$g(z) = \frac{f^{(1)}(z) - f^{(1)}(z_1)}{z - z_1}.$$

Moreover, since f is non-real in $\mathbb{C} \setminus \mathbb{R}$ and strictly decreasing in $\mathbb{R} \setminus [\alpha,\beta]$, we also observe using (3.6) that $f^{(2)} = 1/g$ is analytic in $\mathbb{C} \setminus [\alpha,\beta]$. As a consequence of the first part of Lemma 3.6 applied to f, the Hankel matrices

$$\mathcal{H}_n^{(0)}\left(\frac{1}{f^{(1)}}\right) = \mathcal{H}_n^{(0)}(f), \quad \text{and} \quad -\mathcal{H}_n^{(1)}\left(\frac{1}{f^{(1)}}\right) = -\mathcal{H}_n^{(1)}(f)$$

are positive definite for all $n \geq 0$. According to the second part of Lemma 3.6 applied to $g = 1/f^{(2)}$, it only remains to show that the Hankel matrices

$$\mathcal{H}_n^{(0)}\left(\frac{1}{f^{(2)}}\right) = \mathcal{H}_n^{(1)}(f^{(1)}), \quad \text{and} \quad -\mathcal{H}_n^{(1)}\left(\frac{1}{f^{(2)}}\right) = -\mathcal{H}_n^{(2)}(f^{(1)})$$

are positive definite for all $n \geq 0$. The latter is a consequence of the Hadamard bigradient identity [BGM96, Thm 2.4.1]:

$$\det H_n^{(m)}(f^{(1)}) = (-1)^{(n+1)+(m-1)(m-2)/2} f^{(1)}(z_1)^{m+2n+1} \det H_{m+n-1}^{(2-m)}(\frac{1}{f^{(1)}}),$$

since then

$$\det \mathcal{H}_n^{(0)}(\frac{1}{f^{(2)}}) = \det \mathcal{H}_n^{(1)}(f^{(1)}) = (-1)^{n+1} f^{(1)}(z_1)^{2n+2} \det H_n^{(1)}(\frac{1}{f^{(1)}}) > 0,$$
$$(-1)^{n+1} \det \mathcal{H}_n^{(1)}(\frac{1}{f^{(2)}}) = (-1)^{n+1} \det \mathcal{H}_n^{(2)}(f^{(1)}) = f^{(1)}(z_1)^{2n+3} \det H_{n+1}^{(0)}(\frac{1}{f^{(1)}}) > 0,$$

as required to conclude. □

To summarize, in Section 3 we addressed the question how to represent and compute the rational interpolant r_m for given real interpolation nodes in finite precision arithmetic (since, as reported in [BGM96, Sections 2.1], a naive implementation might lead to a loss of at least m decimal digits of precision). Here we compare three approaches: firstly the partial fraction decomposition of §3.2 promoted by Mayo and Antoulas [MA07] where the poles are the eigenvalues of a Loewner matrix pencil and the residuals are found through a least square problem. Secondly we analyze the barycentric representation of §3.3 which recently [FNTB18] has been used quite successfully for stabilizing the rational Remez algorithm, and for which backward and forward results are known for evaluating such rational functions in finite precision arithmetic. Finally we consider the Thiele interpolating continued fraction in §3.4 which generalizes the concept of Stieljes continued fraction representation of Padé approximants of Markov functions. Our main original contributions in this section are Theorem 3.1 showing that parameters of the Thiele interpolating continued fraction of a Markov function are positive, and Theorem 3.3 where we provide a proof of backward stability of Thiele interpolating continued fractions improving a result of Graves-Morris [GM80, Theorem 4.1]. Numerical experiences presented in Figure 3.1 show that any of these three methods combined with our stopping criterion of Remark 2.5 allows to attain nearly machine precision for scalar arguments, but this will be no longer true for matrix arguments.

4 Functions of Toeplitz-like matrices

In the last years, several authors tried to take advantage of structure in a square matrix A in order to speed up the approximate computation of matrix functions $f(A)$. One possible approach is to consider algebras of structured matrices as for instance hierarchical matrices in HODLR or HSS format which are closed under addition, multiplication with a scalar and inversion, and contain the identity, see for instance the recent paper [MRK20] and the references therein. The hierarchical rank k of $A \in \mathbb{R}^{n\times n}$ gives a complexity parameter and, roughly speaking, the above matrix operations can be carried out in $\mathcal{O}(k^2 n)$ or $\mathcal{O}(k^2 n \log(n))$ operations, see [MRK20, Table in §4.3]. Replacing f by a rational function r of type $[m-1|m]$ and evaluating $r(A)$ within the hierarchical algebra following the operations described in §3.2, §3.3, or §3.4, requires to compute about $2m$ shifted inverses, and in the worst case might increase the hierarchical rank from k (for A) to $2mk$ (for $r(A)$). It is therefore important to know that the above operations are combined with a compression procedure of the same complexity, in order to keep the hierarchical rank as small as possible.

Another structural property allowing for the approximate computation of $f(A)$ in low complexity is *displacement structure*, which was discovered independently by Heinig and Rost [HR84, HR89], and in a series of works by Kailath and others, e.g., [KKM79, CK91, KS95, Pan93]. Our inspiration to use this kind of structure was a work of Kressner and Luce [KL18], who used displacement structure for the fast computation of the matrix exponential for a Toeplitz matrix.

The displacement operator that we will use here is the so-called *Sylvester* displacement operator

$S : \mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}$, defined by

$$S(A) = Z_1 A - A Z_{-1}, \quad \text{where } Z_\theta = \begin{pmatrix} 0 & \dots & \dots & 0 & \theta \\ 1 & \ddots & & & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 1 & 0 \end{pmatrix}. \tag{4.1}$$

The number $\tau = \tau(A) := \operatorname{rank}(S(A))$ is called the *displacement rank of* A, and any pair of matrices $G, B \in \mathbb{C}^{n\times\tau}$ such that $S(A) = GB^*$ is called *a generator* for $S(A)$. It is readily verified that the displacement rank of a Toeplitz matrix (1.3) is at most two, and the same is true for the shifted matrix $A - zI$ (being a Toeplitz matrix itself), and also for the resolvent $(zI - A)^{-1}$. It is customary to say that a matrix $A \in \mathbb{C}^{n\times n}$ is *Toeplitz-like*, if its displacement rank is "small", i.e., if $\tau = \operatorname{rank}(S(A)) \ll n$.

We will now briefly review how the displacement rank behaves under elementary operations. Consider two (Toeplitz-like) matrices A_1, A_2 having displacement ranks τ_1 and τ_2, respectively.

- The identity matrix has displacement rank one.
- For a scalar $0 \neq s \in \mathbb{C}$ we have $\tau(sA_1) = \tau_1$.
- For sums of Toeplitz-like matrices we have $\tau(A_1 + A_2) \leq \tau_1 + \tau_2$.
- For products of Toeplitz-like matrices we have $\tau(A_1 A_2) \leq \tau_1 + \tau_2 + 1$.

It is interesting to note that for many operations it is possible to efficiently compute a generator of the result directly from the generators of the operands. For example, for a Toeplitz-like matrix A with generator (G, B) one finds that $S(A^{-*}) = -(Z_1 A^{-*} B)(Z_{-1}^* A^{-1} G)^*$, implying that a generator for A^{-*} can be obtained by solving two linear system with A^* and A, respectively (and that the displacement rank does not increase here either). An exhaustive description and discussion of operations among Toeplitz-like matrices, and the effects on the displacement rank is given in [Bis21].

In order to understand the arithmetic complexity of evaluating a rational function $r(A)$ of a Toeplitz-like matrix A, we will need to evaluate matrix-vector products with A, and solve linear systems of equations with A. A matrix-vector product with a A (or A^*) can be computed in $\mathcal{O}(\tau n \log(n))$, using the FFT. The currently best asymptotic complexity for solving linear systems of equations with a Toeplitz-like matrix is in $\mathcal{O}(\tau^2 n \log^2(n))$[10], plus a compression procedure based on a singular value decomposition of $S(A)$ in complexity $\mathcal{O}(\tau^2 n)$ where we drop contributions from singular values below the machine precision. Again, we refer to the thesis [Bis21], where the details and further references are spelled out.

In our numerical experiments reported in §5, we use the "TLComp" Matlab toolbox[11], which offers an automatic dispatch of operations with Toeplitz-like matrices to implementations that work directly with generators (as opposed to the full, unstructured matrix). Note that in contrast to the best possible asymptotic complexity pointed out above, this toolbox solves linear systems of equations using the GKO algorithm [GKO95], which (only) has a complexity in $\mathcal{O}(\tau n^2)$. Extensive numerical experiments have shown, however, that for the practical range of dimension $n \in [10^3, 10^5]$ under consideration here, the classical GKO algorithm turns out to be much faster, which is why we stick to it in our numerical experiments. Of course, we still phrase complexity results in Theorem 4.1 below with respect to the better asymptotic bound. This toolbox also allows the fast computation (or approximation) of various norms of Toeplitz-like matrices, and the reconstruction of the full matrix A from its generators in (optimal) complexity $\mathcal{O}(\tau n^2)$. A complete description of "TLComp" and its functionality will be subject of a future publication.

In all our experiments we considered only real symmetric Toeplitz and Toeplitz-like matrices with spectrum in a given interval $[c, d]$, where the different operations even simplify, and we have the error estimates

$$\|f(A) - r_m(A)\| \leq \|f - r_m\|_{L^\infty([c,d])}, \quad \|I - r_m(A) f(A)^{-1}\| \leq \|\frac{f - r_m}{f}\|_{L^\infty([c,d])}.$$

[10] This algorithm is based on transforming via FFT a Toeplitz-like matrix into a Cauchy-like matrix being a hierarchical matrix of hierarchical rank $k = \mathcal{O}(\tau \log(n))$, and then to use the hierarchical solver of `hm-toolbox`, see [MRK20, §5.1 and Table in §3.4] and [XXG12].

[11] `https://github.com/rluce/tlcomp`

Notice that a priori there is no reason to expect that $f(A)$ has a small displacement rank, even for our special case of a Markov function $f = f^{[\mu]}$. However, for a rational function $r_m^{[\mu]}$ of type $[m-1|m]$, the displacement rank of $r_m^{[\mu]}(A)$ is at most $\mathcal{O}(m(\tau+1))$. Also, we will always choose rational interpolants $r_m^{[\mu]}$ with interpolation nodes given by (2.11), which according to (2.12) allows us to achieve precision $\delta > 0$ for $m = \mathcal{O}(\log(1/\delta))$, with the hidden constant depending only on the cross ratio of α, β, c, d. Indeed, in all experiments reported in §5, the displacement rank of $r_m^{[\mu]}(A)$ (after compression) grows at most linearly with m, and sometimes even less if the precision increases.

We summarize our findings in the following theorem, in its proof we also discuss the different implementations of our three approaches of §3 in the algebra of Toeplitz-like matrices.

Theorem 4.1. *Let $f^{[\mu]} : z \mapsto \int \frac{d\mu(x)}{z-x}$ be a Markov function with $\mathrm{supp}(\mu) \subset [\alpha, \beta]$, $\delta > 0$, $m \geq 1$, $A \in \mathbb{R}^{n\times n}$ a symmetric Toeplitz-like matrix with displacement rank τ, and spectrum included in the real interval $[c, d]$, with $c > \beta$. Furthermore, denote by $r_m^{[\mu]}$ the rational interpolant of $f^{[\mu]}$ of type $[m-1|m]$ (in exact arithmetic) at the interpolation nodes (2.11) (depending only on m and α, β, c, d). Then for $m = \mathcal{O}(\log(1/\delta))$, $r_m^{[\mu]}(A)$ of displacement rank $\mathcal{O}(m\tau)$ is an approximation of $f^{[\mu]}(A)$ of (relative) precision $\mathcal{O}(\delta)$.*
Furthermore, computing the generators of $r_m^{[\mu]}(A)$ through the techniques of §3.2, §3.3, and §3.4 within the algebra of Toeplitz-like matrices has complexity $\mathcal{O}(m\tau^3 n \log^2(n))$ for the first two approaches, and $\mathcal{O}(m^2\tau^3 n \log^2(n))$ for the Thiele continued fraction.

Proof. It only remains to show the last part, where we ignore the cost of computing poles/residuals or other parameters which is of complexity $O(m^3)$. The partial fraction decomposition (3.2) in §3.2 seems to be the easiest approach to compute the generators of $r_m^{[\mu]}(A)$: we just have to compute the generators of each resolvent $(A - x_j I)^{-1}$ (with displacement rank bounded by $\tau + 1$), combine and compress. Here the essential work is to compute m times the generator of a resolvent, by solving at most $2m(\tau+1)$ systems of Toeplitz-like matrices, leading to the claimed complexity.

The barycentric representation §3.3 requires to compute separately the generators of

$$P(A) = \sum_{j=0}^{m} f(t_j)\beta_j (A - t_j I)^{-1}, \; Q(A) = \sum_{j=0}^{m} \beta_j (A - t_j I)^{-1}$$

of displacement rank at most $(m+1)(\tau+1)$, then those of $Q(A)^{-1}$ and finally those of $P(A)Q(A)^{-1}$ with a cost being about 4 times the one discussed before.

Finally, for insuring stability in §3.4, we use the backward evaluation of $R_{2m}^{(1)}(A)$ via (3.7), leading to $R_{2m}^{(2m)}(A) = f_{2m}^{(2m)} I$, and $R_{2m}^{(j)}(A) = f_j^{(j)} I + (A - z_j I) R_{2m}^{(j+1)}(A)^{-1}$ for $j = 2m-1, 2m-2, ..., 1$. Here the cost is dominated by finding the generators of the inverse of $R_{2m}^{(j+1)}(A)$, of displacement rank at most $(\tau+1)(2m+2-j)/2$. □

To summarize, in §4 we have reported about how to efficiently evaluate $r_m^{[\mu]}(A)$ for a Toeplitz matrix $A \in \mathbb{R}^{n\times n}$ in complexity $\mathcal{O}(n \log^2(n))$ (with the hidden constant depending on m and the desired precision), by exploiting the additional Toeplitz structure. This part has been strongly inspired by previous work of Kressner and Luce [KL18], see also [MRK20], who exploited the theory of small displacement rank and Toeplitz-like structured matrices. Let us illustrate these findings by some numerical experiments.

5 Numerical experiments

In this section, we illustrate our findings for Toeplitz matrices by reporting several numerical experiments. In all figures to follow, for a fixed symmetric positive definite Toeplitz matrix A of order 500 with extremal eigenvalues $\lambda_{\min}, \lambda_{\max}$ and a fixed Markov function $f^{[\mu]}$ recalled in the caption, we present four different cases (displayed from the left to the right)

$$\begin{array}{ll} (i) & \text{Toeplitz-like arithmetic, } [c,d] = [\lambda_{\min}, \lambda_{\max}], \\ (ii) & \text{Toeplitz-like arithmetic, } [c,d] = [\frac{1}{2}\lambda_{\min}, 2\lambda_{\max}], \\ (iii) & \text{without Toeplitz-like arithmetic, } [c,d] \text{ as in } (i), \\ (iv) & [c,d] \text{ as in } (i), \text{ diagonal matrix of order 500 with entries being cosine points in } [c,d]. \end{array} \tag{5.1}$$

Hence, the impact of an enlarged spectral interval can be measured in comparing cases (i) and (ii), the impact of our particular Toeplitz-like arithmetic by comparing cases (i) and (iii), and finally the impact of a matrix-valued argument instead of a scalar argument (again due to finite precision arithmetic) by comparing the cases $(i)--(iii)$ with (iv). For each of the four cases, we show the relative error of the same rational interpolant of type $[m-1|m]$ with the quasi-optimal interpolation points of (2.11) depending on m (black solid line). Due to finite precision arithmetic, we obtain three error curves: the partial fraction decomposition of §3.2 (triangle markers), the barycentric representation of §3.3 (star markers) and finally the Thiele interpolating continued fraction of §3.4 (circle markers). As in Figure 3.1, we use a marker for index m if (2.14) holds, in other words, this index is rejected by our stopping criterion of Remark 2.5. The forth graph (red dashed) gives the a priori upper bound (2.12) of Corollary 2.4. Notice that computing the relative error requires to evaluate $f(A)$ via built-in matrix functions of Matlab (such as `logm()`), which explains why we consider only matrices of moderate size.

Example 5.1. *We present a first numerical example where we approach* $\log(A)$*, with* A *being a symmetric positive definite Toeplitz matrix of order* 500*, with extremal eigenvalues* $\lambda_{\min} = 25, \lambda_{\max} = 139.2$ *and a condition number* 5.568*, created by using random generators and shifting the spectrum. Notice that* $f(z) = \log(z)/(z-1)$ *is a Markov function with* $\alpha = -\infty$ *and* $\beta = 0$*. Denoting by* r_m *a rational interpolant of type* $[m-1|m]$ *of* f*, we will thus approach* $\log(A)$ *by* $(A-I)r_m(A)$*, leading to the same relative error as approaching* $f(A)$ *by* $r_m(A)$*. The four different cases (from left to the right) are those explained above in* (5.1)*. Some observations are in place:*

(a) In any of the four cases and 3 methods, the stopping criterion of Remark 2.5 works perfectly well: all accepted indices give errors below our a priori bound being nicely decreasing for increasing m*. Also, all rejected indices correspond to errors above our a priori bound, and these errors are never much smaller than the error at the last accepted index.*

(b) In the case (iv) *of diagonal matrices* A *(or, equivalently, for scalar arguments), all three methods for evaluating* $r_m(A)$ *are equivalent, this confirms similar observations in Figure 3.1.*

(c) In any of the cases $(i)-(iii)$*, that is, for full matrices* A*, the barycentric representation of* $r_m(A)$ *leads to much larger errors, in particular if one uses Toeplitz-like arithmetic.*

(d) The partial fraction decomposition and the Thiele continued fraction approaches have a similar behavior, and lead to a small error of order 10^{-12}*.*

(e) Enlarging the spectral interval as in case (ii) *does not lead to a smaller error, but might require to compute interpolants of higher degree for achieving the same error.*

(f) For measuring the complexity, it is also interesting to observe that the displacement rank of $r_m(A)$ *in the cases* (i) *and* (ii) *is increasing in* m*, it first increases linearly, and then stabilizes around* 22 *(for Thiele and partial fractions, the double for barycentric), once a good precision is reached.*

The observations (a)–(f) obtained in Example 5.1 for a particular matrix can also be made for other symmetric positive definite Toeplitz matrices, as long as the condition number remains modest, say, below 10. Notice that, with $\alpha = -\infty$ and $\beta = 0$, the condition number of A equals the cross ratio in (2.9) (for the cases (i), (iii) and (iv)), and hence determines the asymptotic rate of convergence (2.12), essentially the slope of our a priori upper bound.

However, observation (d) is no longer true for condition numbers larger than 10: we present in Figure 5.2 two other examples for $\log(A)$ and two symmetric positive definite Toeplitz matrices A, with condition number 121.7 (on the top) and $1.35 \cdot 10^5$ (on the bottom).[12] Here not only the barycentric representation but also the Thiele continued fraction approach fails completely to give acceptable relative errors, in particular for the two cases (i) and (ii) of Toeplitz-like arithmetic. In contrast, the partial fraction decomposition gives a relative error of order 10^{-11} for the top case, and 10^{-9} for the bottom case, which probably is acceptable for most applications.

An alternative to deal with ill-conditioned matrices A is the classical inverse scaling and squaring technique, which can be applied for the logarithm and for the fractional power $x \mapsto x^\gamma$, $\gamma \in \mathbb{R}$,

$$\log(A) = 2^\ell \log(A^{\frac{1}{2^\ell}}), \quad A^\gamma = \left((A^{\frac{1}{2^\ell}})^\gamma\right)^{2^\ell}, \tag{5.2}$$

[12] Further explanations on Figure 5.2 are given in Example 5.2 below.

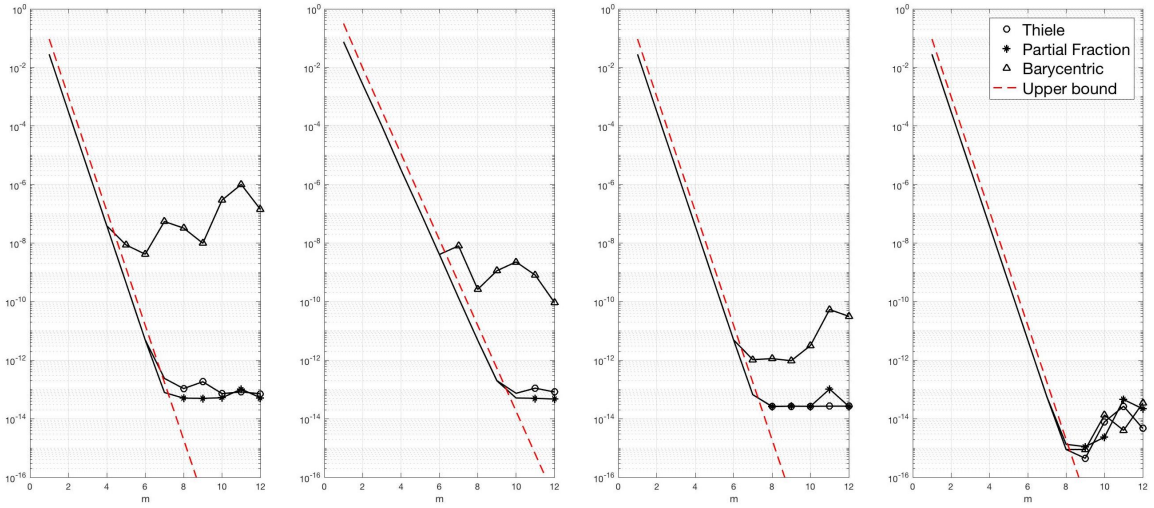

Figure 5.1: Relative errors for approaching $f(A)$ for the Markov function $f(z) = \log(z)/(z-1)$, and a symmetric positive definite Toeplitz matrix of order 500, with extremal eigenvalues $\lambda_{\min} = 25, \lambda_{\max} = 139.2$ and a condition number 5.568. Further explanations on the legend and the four different cases $(i) - (iv)$ (displayed from the left to the right) are given in the first paragraph of §5.

see for instance [Hig08, Chapter 11.5] and [HL13]. In other words, we start by computing ℓ square roots $A_0 = A$, and $A_j = (A_{j-1})^{1/2}$ for $j = 1, ..., \ell$, with the integer ℓ chosen such that

$$\mathrm{cond}(A^{\frac{1}{2^\ell}}) \leq (d/c)^{\frac{1}{2^\ell}} \leq 10. \tag{5.3}$$

Here each square root is obtained by a scaled Newton method, and more precisely the product form of the scaled DB iteration[13] given in Algorithm 5.1. Notice that $M_k - I = X_k B^{-1} X_k - I$ is what we have

Algorithm 5.1: Product form of the scaled DB iteration for approximating the matrix square root $B^{1/2}$ for a symmetric positive definite matrix B with spectrum in $[c, d]$ and parameters $\mu_0, \mu_1, ...$

Result: Return X_k approximation of $B^{1/2}$
begin
  $M_0 = X_0 = B$, $k = 0$;
  while $\|I - M_k\| > tol$ **do**
    $M_{k+1} = \frac{1}{4}\left(2I + \mu_k^2 M_k + \frac{1}{\mu_k^2} M_k^{-1}\right)$;
    $X_{k+1} = \frac{1}{2}\mu_k (I + \mu_k^{-2} M_k^{-1}) X_k$;
    $k \leftarrow k + 1$;
  end
end

called in Corollary 2.2 the residual of the square root $B^{1/2}$. A suitable choice of parameters allows to speed up the first iterations of the Newton method. The following parameters have been suggested for scalar arguments by Rutishäuser [Rut63], and discussed for matrix arguments by Beckermann [Bec13], see also Zietak & Zielinski [KZZ07] and Byers & Xu [BX08],

$$\mu_0 = \frac{1}{\sqrt[4]{cd}}, \quad \mu_1 = \sqrt{\frac{2\sqrt[4]{cd}}{\sqrt{c} + \sqrt{d}}}, \quad \mu_{k+1} = \sqrt{\frac{2\mu_k}{1 + \mu_k^2}} \tag{5.4}$$

[13] In the original formulation [DB76], Denman and Beavers gave a coupled two-term recurrence for X_k and $Y_k := A^{-1} X_k$. The product form was obtained later in [CHKL01], where Y_k is replaced by $M_k = X_k Y_k$. These authors suggest the recurrence $X_{k+1} = \frac{1}{2}\mu_k X_k (I + \mu_k^{-2} M_k^{-1})$, but our recurrence seems to be more suitable in case where the matrices do no longer commute due to finite precision arithmetic. It seems that this slight modification has no impact on the analysis of stability and limiting accuracy, and even hardly no impact on the (relative) error.

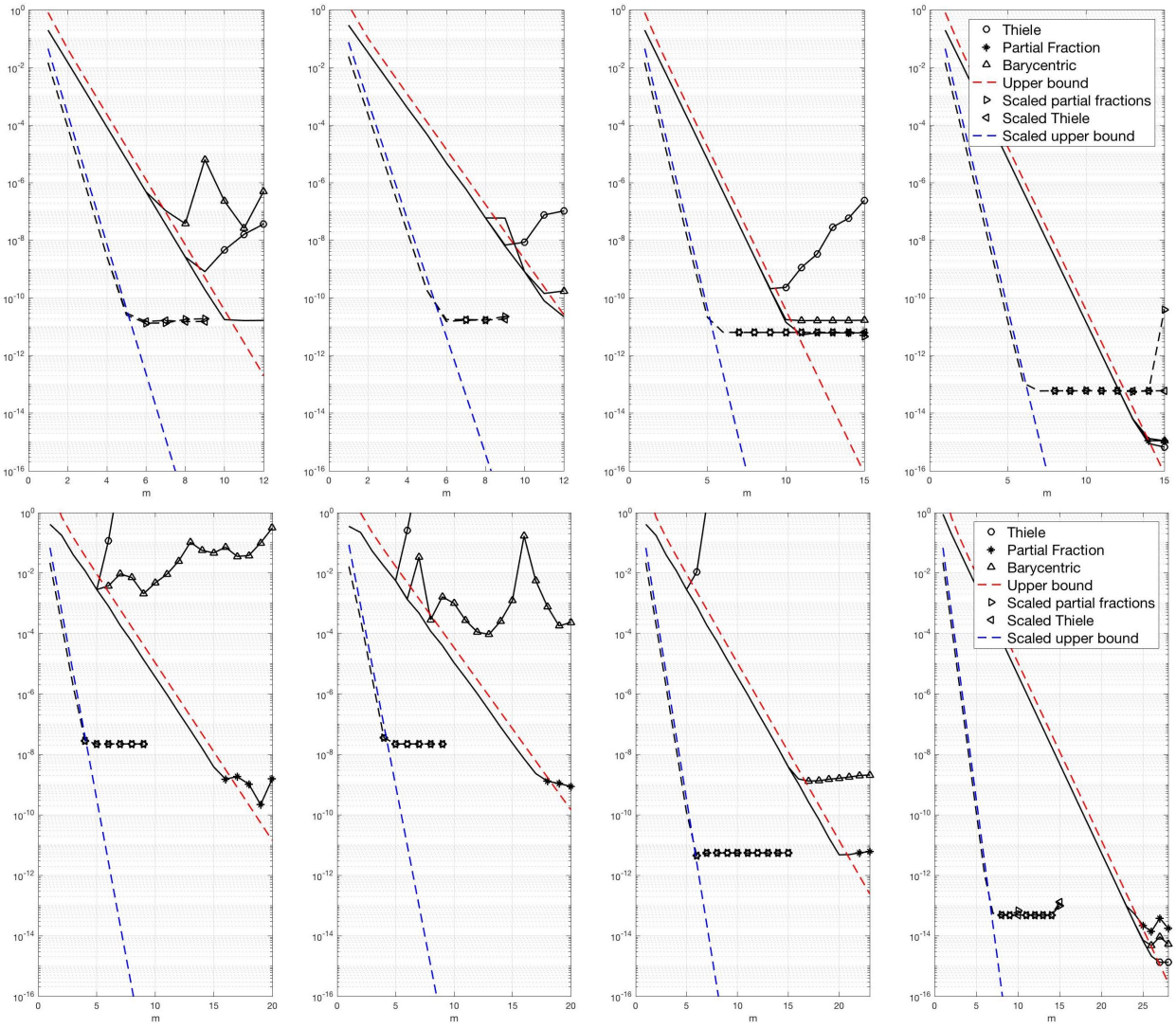

Figure 5.2: Relative errors for approaching $f(A)$ for the Markov function $f(z) = \log(z)/(z-1)$, and two symmetric positive definite Toeplitz matrix of order 500, with extremal eigenvalues $\lambda_{\min} = 0.918, \lambda_{\max} = 111.7$ and condition number 121.7 for the first matrix (on the top), and $\lambda_{\min} = 0.001, \lambda_{\max} = 135$ and condition number $1.35 \cdot 10^5$ for the second matrix (on the bottom).

for $k \geq 1$. For this choice of parameters (which are $\in (0,1)$ and tend quickly to 1), one shows by recurrence that, for $k \geq 1$,

$$\sigma(X_k B^{-1/2}) \subset [1, \frac{1}{\mu_k^2}], \quad \sigma(M_k) \subset [1, \frac{1}{\mu_k^4}], \quad \|M_k - I\| \leq \frac{1-\mu_k^4}{\mu_k^4}.$$

In order to keep stability and limiting accuracy shown for parameters $\mu_k = 1$ in [Hig08, §6.4], we suggest to proceed in two phases: in the first phase we apply Newton with parameters as in (5.4) until $\frac{1-\mu_K^4}{\mu_K^4} \leq 10^{-3}$ (for instance $K \leq 5$ for $\text{cond}(A) \leq 10^6$). For $k \geq K$, we then choose $\mu_k = 1$ and thus

$$\|M_{k+1} - I\| \leq \frac{1}{4}\|(M_k - I)^2 M_k^{-1}\| \leq \frac{1}{3}\|M_k - I\|^2.$$

According to this quadratic convergence, 3 Newton steps in the second phase should lead to high precision even in finite precision. After having computed $A^{\frac{1}{2^\ell}}$, we then evaluate rational interpolants of our particular Markov function f at $A^{\frac{1}{2^\ell}}$, and finally perform the squaring or renormalization in order to

approximate $f(A)$. In the cases $(i)-(ii)$ we implemented Newton and the squaring within the algebra of Toeplitz-like matrices, in order to speed up computation time. Notice that the cost of evaluating the interpolants for various values of m is much higher than the cost for scaling or squaring, at least for $\text{cond}(A) \leq 10^6$ where we have to compute $\ell \leq 3$ square roots, and we have at most 8 Newton steps for each square root.

Example 5.2. *Reconsider the problem of approaching* $\log(A)$ *for the two symmetric positive definite Toeplitz matrices A of Figure 5.2. Beside the error curves described in the first paragraph of §5, we have added in Figure 5.2 the a priori upper bound for the matrix* $A^{\frac{1}{2^\ell}}$ *(in blue dashed), as well as the relative error (black dashed) obtained by evaluating at* $A^{\frac{1}{2^\ell}}$ *the interpolant of* $f(x) = \log(x)/(x-1)$ *via a partial fraction decomposition (triangles pointing to the right), or via a Thiele continued fraction (triangles pointing to the left). As we have seen before, both approaches have a very similar behavior according to* (5.3)*. On the top, with a matrix of condition number* 121.7 *(and hence* $\ell = 2$*), we observe that this inverse scaling and squaring technique combined with Toeplitz-like arithmetic gives about the same relative error as partial fraction decomposition applied directly to A. On the bottom, with a matrix of condition number* $1.35 \cdot 10^5$ *and hence* $\ell = 3$*, the conclusion is different: here our inverse scaling and squaring technique combined with Toeplitz-like arithmetic gives a relative error about 10 times larger than that for partial fraction decomposition applied directly to A.*

Detailed information about the rate of convergence and the final precision of each Newton iteration for computing $A_j = (A_{j-1})^{1/2}$ *for* $j = 1, ..., \ell$*,* $A_0 = A$ *are given in [Bis21], we only report here that the final relative error for scaled Thiele or scaled partial fraction decomposition is dominated by the relative error in computing the square root* $A_1 = A^{1/2}$*, somehow as expected since this matrix has the worst condition number among the matrices* A_j*. Also, we tried other equivalent formulations of the Newton method, and obtained similar conclusions.*

Example 5.3. *In our final example we study the fractional power* $x \mapsto x^{-1/3}$ *of two symmetric positive definite Toeplitz matrices, displayed in Figure 5.3. We first modify slightly the approach described in* (5.2)*, since* $x \mapsto x^\gamma$ *is only a Markov function provided that* $\gamma \in [-1, 0)$*. Also, preliminary numerical experiments not reported here indicate that squaring* ℓ *times seems to increase the relative error. For* $\gamma \in \mathbb{R}$*, we thus write* $2^\ell \gamma = k + \gamma'$ *with* $k \in \mathbb{Z}$ *and* $\gamma' \in [-1, 0)$*, such that* $A^\gamma = g(A^{\frac{1}{2^\ell}})(A^{\frac{1}{2^\ell}})^k$ *with a Markov function* $g(x) = x^{\gamma'}$*, which is approached by* $r_m(A^{\frac{1}{2^\ell}})(A^{\frac{1}{2^\ell}})^k$ *with* r_m *an interpolant of g. As in Example 5.1, we thus may apply our bounds for the relative error.*[14]

For the matrix on the top of Figure 5.3 with condition number 121.7 *we find that* $\ell = 2$ *and hence* $k = -1$*,* $\gamma' = -1/3$*, whereas for the matrix on the bottom with condition number* $1.01 \cdot 10^5$ *we find* $\ell = 3$ *and* $k = -2$*,* $\gamma' = -2/3$*. Limiting ourselves to cases* $(i)-(ii)$ *using Toeplitz-like arithmetic, we obtain relative errors for unscaled Thiele of about* 10^{-7} *on the top and only* 10^{-2} *on the bottom. Both scaled Thiele or scaled partial fractions allow to achieve relative errors of about* 10^{-12} *on the top and* 10^{-8} *on the bottom. However, the smallest relative error is obtained for an unscaled partial fraction decomposition, namely* 10^{-13} *on the top and* 10^{-11} *on the bottom.*

To summarize, in §5 we have presented numerical results for two Markov functions and several symmetric positive definite Toeplitz matrices A, which show that our stopping criterion of Remark 2.5 works surprisingly well in practice. Also, exploiting the Toeplitz structure gives an interesting complexity for large n, but in general also increases the error. If we exploit the Toeplitz structure, we should avoid the barycentric representation of §3.3 and the Thiele interpolating continued fraction of §3.4, since the smallest relative error is obtained by the partial fraction decomposition of §3.2, especially for larger condition numbers of A. Finally, for the functions considered in (5.2), one might also want to combine the partial fraction decomposition of §3.2 with inverse scaling and squaring, which seems to increase the error, but has lower complexity since the involved rational functions have lower degree.

6 Conclusion

In this paper we presented a detailed study of how to efficiently and reliably approximate $f(A)$ by $r_m(A)$, with f a Markov function, A a symmetric Toeplitz matrix, and r_m a suitable rational interpolant of f.

[14] It is interesting to compare our findings to those in [HL13] where the authors approach A^γ after scaling by evaluating Padé approximants at the single interpolation point $z_1 = ... = z_{2m} = 1$ using Stieltjes continued fractions, somehow a confluent counterpart of our approach.

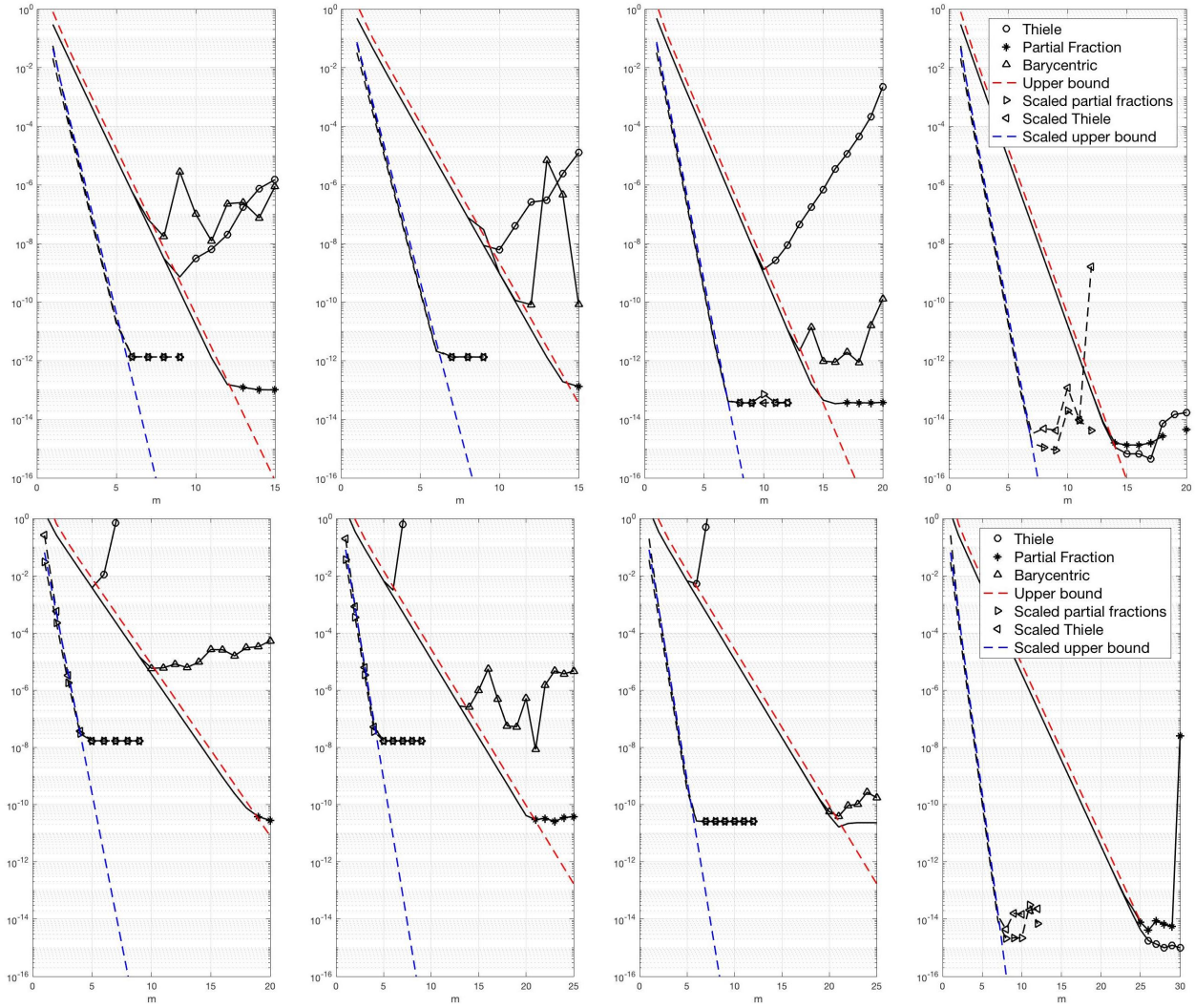

Figure 5.3: Relative errors for approaching $A^{-1/3}$ for two symmetric positive definite Toeplitz matrices: on the top we find the same matrix as in Figure 5.2 of order 500, with extremal eigenvalues $\lambda_{\min} = 0.918, \lambda_{\max} = 111.7$ and condition number 121.7, and on the bottom the $1D$ discretized Laplacian of order 499 (that is, the tridiagonal matrix containing 2 on the main diagonal and -1 on the super- and subdiagonal), with extremal eigenvalues $\lambda_{\min} = 3.95 \cdot 10^{-5}, \lambda_{\max} = 4.000$ and condition number $1.01 \cdot 10^5$.
.

Numerical evidence provided in Figure 3.1 and case (iv) on the right of Figures 5.1–5.3 shows that, for scalar arguments z, we may nearly reach machine precision for the relative error using any of these three approaches discussed in §3. The picture changes however completely for the relative error $I - r_m(A)f(A)^{-1}$ evaluated at a Toeplitz matrix argument A. Here only the partial fraction decomposition of §3.2 insures small errors, especially for larger condition numbers of A.

In this paper, we have hardly discussed the case of non necessarily symmetric (Toeplitz) matrices A, which is left as open question for further research. As explained in Remark 2.8, it is possible to construct rational approximants r_m, namely Faber images of rational interpolants, such that $f(A) - r_m(A)$ is bounded by $(1+\sqrt{2})$ times the maximum of $f - r_m$ on the field of values of A, which again can be related to the interpolation error of a Markov function on the unit disk. However, we expect such field-of-value estimates for non symmetric matrices A not to be very sharp, and maybe other K-spectral sets of A [BB13, Section 107.2] would be more suitable. Also, it is not clear for us how to represent the rational function r_m, and what kind of stability results to expect for evaluating r_m at a complex scalar argument, or at a general matrix A.

Another direction of further research could be to work with variable precision in our compression procedure of computing the numerical displacement rank, which potentially could lead to a much more efficient implementation.

Acknowledgements. The authors want to thank Stefan Güttel, Marcel Schweitzer and Leonid Knizhnerman for carefully reading a draft of this manuscript, and for their useful comments.

Conflict of interest. Partial financial support was received from the Labex CEMPI (ANR-11-LABX-0007-01). The authors declare that they have no conflict of interest.

References

[Ach90] N. I. Achieser. Elements of the theory of elliptic functions. *American Mathematical Society, Providence*, 1990.

[AS64] M. Abramowitz and I. A. Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55 of *National Bureau of Standards Applied Mathematics Series*. For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C., 1964.

[BB13] C. Badea and B. Beckermann. Spectral sets. In L. Hogben, editor, *Handb. Linear Algebr.*, chapter 37, pages 613–638. CRC Press, second edition, 2013.

[BB20] M. Benzi and P. Boito. Matrix functions in network analysis. *Gamm-Mitteilungen*, 43, 2020.

[BBM05] J.-P. Berrut, R. Baltensperger, and H. D. Mittelmann. Recent developments in barycentric rational interpolation. In *Trends and Applications in Constructive Approximation*, ISNM International Series of Numerical Mathematics, pages 27–51. Birkhäuser Basel, Basel, 2005.

[Bec13] B. Beckermann. Optimally scaled Newton iterations for the matrix square root, 2013.

[BGM96] G. A. Baker, Jr. and P. Graves-Morris. *Padé approximants*, volume 59 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, second edition, 1996.

[Bis21] J. Bisch. *Fonctions de Matrices de Toeplitz symétriques*. PhD thesis, Université de Lille (France), 2021.

[BR09] B. Beckermann and L. Reichel. Error estimates and evaluation of matrix functions via the Faber transform. *SIAM J. Numer. Anal.*, 47(5):3849–3883, 2009.

[Bra86] D. Braess. *Nonlinear approximation theory*, volume 7 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 1986.

[Bra87] D. Braess. Rational approximation of Stieltjes functions by the Carathéodory-Fejèr method. *Constr. Approx.*, 3(1):43–50, 1987.

[BT19] B. Beckermann and A. Townsend. Bounds on the singular values of matrices with displacement structure. *SIAM Rev.*, 61(2):319–344, 2019.

[BX08] R. Byers and H. Xu. A new scaling for Newton's iteration for the polar decomposition and its backward stability. *SIAM J. Matrix Anal. Appl.*, 30(2):822–843, 2008.

[CHKL01] S. Cheng, N. J. Higham, C. Kenney, and A. Laub. Approximating the logarithm of a matrix to specified accuracy. *SIAM J. Matrix Anal. Appl.*, 22:1112–1125, 2001.

[CK91] J. Chun and T. Kailath. Displacement structure for Hankel, Vandermonde, and related (derived) matrices. *Linear Algebra and its Appl.*, 151(C):199–227, 1991.

[DB76] E. D. Denman and A. N. Beavers. The matrix sign function and computations in systems. *Appl. Math. Comput.*, 2(1):63–94, 1976.

[DH04] P. I. Davies and N. J. Higham. A Schur-Parlett algorithm for computing matrix functions. *SIAM J. Matrix Anal. Appl.*, 25(2):464–485, 2004.

[EH10] E. Estrada and D. J. Higham. Network properties revealed through matrix functions. *SIAM Rev.*, 52:696–714, 2010.

[EI19] M. Embree and A. C. Ionita. Pseudospectra of Loewner matrix pencils. *CoRR*, 2019.

[Ell83] S. W. Ellacott. On the Faber transform and efficient numerical rational approximation. *SIAM J. Numer. Anal.*, 20(5):989–1000, 1983.

[Fas19] M. Fasi. Optimality of the Paterson-Stockmeyer method for evaluating matrix polynomials and rational matrix functions. *Lin. Alg. Appli.*, 574:182–200, 2019.

[FNTB18] S.-I. Filip, Y. Nakatsukasa, L. N. Trefethen, and B. Beckermann. Rational minimax approximation via adaptive barycentric representations. *SIAM Journal on Scientific Computing*, 40(4):A2427–A2455, 2018.

[Fre71] G. Freud. *Orthogonal polynomials*. Pergamon Press, Oxford New York Toronto Sydney, 1971.

[Gai87] D. Gaier. *Lectures on Complex Approximation*. Birkhäuser Boston, Boston, MA, 1st ed. 1987. edition, 1987.

[Gan82] T. Ganelius. Degree of rational approximation. In *Lectures on approximation and value distribution*, volume 79 of *Sém. Math. Sup.*, pages 9–78. Presses Univ. Montréal, Montreal, Que., 1982.

[GKO95] I. Gohberg, T. Kailath, and V. Olshevsky. Fast Gaussian elimination with partial pivoting for matrices with displacement structure. *Math. Comp.*, 64(212):1557–1576, 1995.

[GM80] P. R. Graves-Morris. Practical, reliable, rational interpolation. *J. Inst. Math. Appl.*, 25(3):267–286, 1980.

[GM81] P. R. Graves-Morris. Efficient reliable rational interpolation. In *Padé approximation and its applications (Amsterdam, 1980)*, volume 888 of *Lecture Notes in Math.*, pages 28–63. Springer, Berlin-New York, 1981.

[GM10] G. H. Golub and G. Meurant. *Matrices, Moments and Quadrature with Applications*. Princeton series in applied mathematics. Princeton University Press, Princeton (N.J.) Woodstock, 2010.

[Gon78a] A. A. Gonchar. On Markov's theorem for multipoint Padé approximants. *Math. USSR-Sb*, 34(4):449–459, 1978.

[Gon78b] A. A. Gonchar. On the speed of rational approximation of some analytic functions. *Math. USSR-Sb*, 34(2):131–145, 1978.

[Hen77] P. Henrici. *Applied and Computational Complex Analysis. Vol. 2*. Pure and Applied Mathematics. John Wiley & Sons, Inc., New York, 1977.

[Hig02] N. J. Higham. *Accuracy and Stability of Numerical Algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, second edition, 2002.

[Hig08] N. J. Higham. *Functions of matrices*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2008. Theory and computation.

[HL13] N. J. Higham and L. Lin. An improved Schur–Padé algorithm for fractional powers of a matrix and their Fréchet derivatives. *SIAM J. Matrix Anal. Appl.*, 34, 07 2013.

[HO10] M. Hochbruck and A. Ostermann. Exponential integrators. *Acta Numerica*, 19:209–286, 2010.

[HR84] G. Heinig and K. Rost. *Algebraic Methods for Toeplitz-like Matrices and Operators*. Operator Theory: Advances and Applications, 13. Birkhäuser Basel, Basel, 1984.

[HR89] G. Heinig and K. Rost. *Matrices with Displacement Structure, Generalized Bezoutians, and Moebius Transformations*, pages 203–230. Birkhäuser Basel, 1989.

[KKM79] T. Kailath, S.-Y. Kung, and M. Morf. Displacement ranks of a matrix. *Bull. Amer. Math. Soc.*, 1(5):769 – 773, 1979.

[KL18] D. Kressner and R. Luce. Fast computation of the matrix exponential for a Toeplitz matrix. *SIAM J. Matrix Anal. Appl.*, 39(1):23–47, 2018.

[Kni09] L. Knizhnerman. Padé-Faber approximation of Markov functions on real-symmetric compact sets. *Mathematical Notes*, 86(1-2):81–92, 2009.

[KS95] T. Kailath and Ali H. Sayed. Displacement structure: Theory and applications. *SIAM Rev.*, 37(3):297–386, 1995.

[KZZ07] A. Kielbasiński, P. Zieliński, and K. Ziętak. Higham's scaled method for polar decomposition and numerical matrix-inversion. *Technical report Institute of Mathematics and Computer Science Report I18/2007/P-045*, 2007.

[Lag86] G. L. Lagomasino. Szegö's Theorem for polynomials orthogonal with respect to varying measures, Orthogonal polynomials and their applications. *L. Notes in Math.*, 1329:255–260, 1986.

[Lag87] G. L. Lagomasino. On the assymptotics of the ratio of orthogonal polynomials and the convergence of multipoint Padé approximants. *Math. USSR-Sb.*, 56:216–229, 1987.

[MA07] A.J. Mayo and A.C. Antoulas. A framework for the solution of the generalized realization problem. *Linear Algebra and its Appl.*, 425(2):634 – 662, 2007. Special Issue in honor of Paul Fuhrmann.

[Mei67] G. Meinardus. *Approximation of functions: Theory and numerical methods.* Expanded translation of the German edition. Translated by Larry L. Schumaker. Springer Tracts in Natural Philosophy, Vol. 13. Springer-Verlag New York, Inc., New York, 1967.

[MR21] S. Massei and L. Robol. Rational Krylov for Stieltjes matrix functions: convergence and pole selection. *BIT Numer. Math.*, 61:237–273, 2021.

[MRK20] S. Massei, L. Robol, and D. Kressner. hm-toolbox: Matlab software for HODLR and HSS matrices, 2020. https://arxiv.org/abs/1909.07909.

[NT15] T. W. Ng and C. Y. Tsang. Chebyshev-Blaschke products: solutions to certain approximation problems and differential equations. *J. Comput. Appl. Math.*, 277:106–114, 2015.

[Pan93] V. Pan. Decreasing the displacement rank of a matrix. *SIAM J. Matrix Anal. Appl.*, 14(1):118–121, 1993.

[Rut63] H Rutishauser. Betrachtungen zur Quadratwurzeliteration. *Monatshefte für Mathematik*, 67(5):452–464, 1963.

[SC08] O. Salazar Celis. *Practical rational interpolation of exact and inexact data: theory and algorithms.* PhD thesis, Universiteit Antwerpen (Belgium), 2008.

[SNF+13] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Process. Mag.*, 30(3):83–98, 2013.

[ST92] H. Stahl and V. Totik. *General orthogonal polynomials*, volume 43 of *Encyclopedia of Mathematics and its Applications.* Cambridge University Press, Cambridge, 1992.

[Sta00] H. Stahl. Strong asymptotics for orthonormal polynomials with varying weights. *Acta. Sci. Math. (Szeged)*, 66(1-2):147–192, 2000.

[Sto20] M. Stoll. A literature survey of matrix methods for data science. *GAMM-Mitteilungen*, 43(3), 2020.

[Sze75] G. Szegö. *Orthogonal polynomials*. American Mathematical Society, Providence, R.I., fourth edition, 1975. American Mathematical Society, Colloquium Publications, Vol. XXIII.

[XXG12] J. Xia, Y. Xi, and M. Gu. A superfast structured solver for Toeplitz linear systems via randomized sampling. *SIAM J. Matrix Anal. Appl.*, 33(3):837–858, 2012.

[Zol77] E. I. Zolotarev. Application of elliptic functions to questions of functions deviating least and most from zero. *Zap. Imp. Akad. Nauk*, 30:1–59, 1877.