Constructive Approximation (2021) 54:1–34
https://doi.org/10.1007/s00365-020-09504-3

CONSTRUCTIVE APPROXIMATION

Rational Minimax Iterations for Computing the Matrix *p*th Root

Evan S. Gawlik[1]

Published online: 12 May 2020

Abstract
In a previous paper by the author, a family of iterations for computing the matrix square root was constructed by exploiting a recursion obeyed by Zolotarev's rational minimax approximants of the function $z^{1/2}$. The present paper generalizes this construction by deriving rational minimax iterations for the matrix pth root, where $p \geq 2$ is an integer. The analysis of these iterations is considerably different from the case $p = 2$, owing to the fact that when $p > 2$, rational minimax approximants of the function $z^{1/p}$ do not obey a recursion. Nevertheless, we show that several of the salient features of the Zolotarev iterations for the matrix square root, including equioscillatory error, order of convergence, and stability, carry over to the case $p > 2$. A key role in the analysis is played by the asymptotic behavior of rational minimax approximants on short intervals. Numerical examples are presented to illustrate the predictions of the theory.

Keywords Matrix root · Matrix power · Rational approximation · Minimax · Uniform approximation · Matrix iteration · Chebyshev approximation · Padé approximation · Newton iteration · Zolotarev

Mathematics Subject Classification 65F30 · 65F60 · 41A20 · 49K35

1 Introduction

In recent years, a growing body of literature has highlighted the usefulness of *rational minimax iterations* for computing functions of matrices [4,7,9,28,29]. In these studies, $f(A)$ is approximated by a rational function r of A possessing two properties:

Communicated by Wolfgang Dahmen.

✉ Evan S. Gawlik
egawlik@hawaii.edu

[1] Department of Mathematics, University of Hawaii at Manoa, Honolulu, USA

r closely (and often optimally) approximates f in the uniform norm over a subset of the real line, and r can be generated from a recursion. A prominent example of such an iteration was introduced by Nakatsukasa and Freund [29], who observed that rational minimax approximants of the function $\operatorname{sign}(z) = z/(z^2)^{1/2}$ obey a recursion, allowing one to rapidly compute $\operatorname{sign}(A)$ and related decompositions such as the polar decomposition, symmetric eigendecomposition, SVD, and, in subsequent work, the CS decomposition [9]. An analogous recursion for rational minimax approximants of $z^{1/2}$ has recently been used to construct iterations for the matrix square root [7], building upon ideas of Beckermann [2]. There, the iterations are referred to as *Zolotarev iterations*, owing to the role played by explicit formulas for rational minimax approximants of $\operatorname{sign}(z)$ and $z^{1/2}$ derived by Zolotarev [36].

The aim of this paper is to introduce a family of rational minimax iterations for computing the principal pth root of a square matrix A, where $p \geq 2$ is an integer. Recall that the principal pth root of a square matrix A having no nonpositive real eigenvalues is the unique solution of $X^p = A$ whose eigenvalues are contained in $\{z \in \mathbb{C} \mid -\pi/p < \arg z < \pi/p\}$ [16, Theorem 7.2]. The iterations we propose reduce to the Zolotarev iterations for the matrix square root [7] when $p = 2$, but when $p > 2$, they differ from the Zolotarev iterations in several important ways. Notably, for all integers $p \geq 2$, the iterations generate a rational function r of A which has the property that for scalar inputs, the relative error $e(z) = (r(z) - z^{1/p})/z^{1/p}$ equioscillates on a certain interval $[a, b]$ (see Sect. 2 for our terminology). Remarkably, when $p = 2$, $e(z)$ equioscillates often enough to render $\max_{a \leq z \leq b} |e(z)|$ minimal among all choices of r with a fixed numerator and denominator degree [7]. This optimality property is the hallmark of the Zolotarev iterations, and it allows one to appeal to classical results from rational approximation theory to estimate the maximum relative error. When $p > 2$, no such optimality property holds. Much of this paper is devoted to showing that the rational minimax iterations for the pth root still enjoy many of the same desirable features as the Zolotarev iterations for the square root, despite the absence of optimality in the case $p > 2$. We take care to present our results in such a way that when $p = 2$, the salient features of the Zolotarev iterations are recovered as special cases.

There are a number of connections between the iterations we derive and existing iterations from the literature on the matrix pth root. We have already mentioned that they reduce to the Zolotarev iterations when $p = 2$. For arbitrary $p \geq 2$, the two lowest order versions of our rational minimax iterations are scaled variants of the Newton iteration and the inverse Newton iteration [16, Chapter 6], [3, Section 6], [20]. In another limiting case, our iterations reduce to the Padé iterations [24, Section 5]. Relative to these iterations, the rational minimax iterations offer advantages primarily when the matrix A has eigenvalues with widely varying magnitudes. As an extreme example, if $p = 3$ and A is Hermitian positive definite with condition number $\leq 10^{16}$, convergence is achieved in double-precision arithmetic after just 2 iterations when using our type-$(6, 6)$ rational minimax iteration. In contrast, up to 5 iterations are needed when using the type-$(6, 6)$ Padé iteration. Our numerical experiments indicate that the situation is similar, but less dramatic, for non-normal matrices with eigenvalues away from the positive real axis.

This paper is organized as follows. In Sect. 2, we review the Zolotarev iterations for the matrix square root by summarizing the contents of [7]. In Sect. 3, we introduce rational minimax iterations for the matrix pth root and present our main results: Theorems 1, 2, and their corollaries. Proofs of these results are provided separately in Sect. 4. Finally, Sect. 5 presents numerical experiments that illustrate the predictions of the theory.

2 Background: Zolotarev Iterations for the Matrix Square Root

Let us summarize the Zolotarev iterations for the matrix square root and their key properties [7]. Let $\mathcal{R}_{m,\ell}$ denote the set of all rational functions of type (m, ℓ)—ratios of polynomials of degree at most m to polynomials of degree at most ℓ. We say that a function $r(z) = g(z)/h(z)$ in $\mathcal{R}_{m,\ell}$ has *exact type* (m', ℓ') if, after canceling common factors, $g(z)$ and $h(z)$ have degree exactly $m' \le m$ and $\ell' \le \ell$, respectively. The number $d = \min\{m - m', \ell - \ell'\}$ is called the *defect* of r in $\mathcal{R}_{m,\ell}$. In most of what follows, z is a real variable; we use the letter z since the behavior of r on $\mathbb{C}$ will play an important role later in the paper.

Given a continuous, increasing bijection $f\colon [0, 1] \to [0, 1]$ and a number $\alpha \in (0, 1)$, let $r_{m,\ell}(z, \alpha, f)$ denote the best type-(m, ℓ) rational approximant of $f(z)$ on $[f^{-1}(\alpha), 1]$:

$$r_{m,\ell}(\cdot, \alpha, f) = \operatorname*{arg\,min}_{r \in \mathcal{R}_{m,\ell}} \max_{z \in [f^{-1}(\alpha), 1]} \left| \frac{r(z) - f(z)}{f(z)} \right|. \tag{1}$$

It is well-known that the minimization problem above has a unique solution [1, p. 55]. Furthermore, explicit formulas for $r_{m,\ell}(\cdot, \alpha, \sqrt{\cdot})$ are known for $\ell \in \{m - 1, m\}$ [36]. Let $\hat{r}_{m,\ell}(z, \alpha, f)$ denote the unique scalar multiple of $r_{m,\ell}(z, \alpha, f)$ with the property that

$$\min_{z \in [f^{-1}(\alpha), 1]} \frac{\hat{r}_{m,\ell}(z, \alpha, f) - f(z)}{f(z)} = 0. \tag{2}$$

For $m \in \mathbb{N}$ and $\ell \in \{m - 1, m\}$, the Zolotarev iteration of type (m, ℓ) for computing the square root of a square matrix A reads

$$X_{k+1} = X_k \hat{r}_{m,\ell}\left(X_k^{-2} A, \alpha_k, \sqrt{\cdot}\right), \qquad X_0 = I, \tag{3}$$

$$\alpha_{k+1} = \frac{\alpha_k}{\hat{r}_{m,\ell}(\alpha_k^2, \alpha_k, \sqrt{\cdot})}, \qquad \alpha_0 = \alpha. \tag{4}$$

It is proved in [7] that in exact arithmetic, $X_k \to A^{1/2}$ and $\alpha_k \to 1$ with order of convergence $m + \ell + 1$ for any A with no nonpositive real eigenvalues. In floating point arithmetic, it is necessary to reformulate the iteration to ensure its stability; we detail the stable reformulation of (3–4) later on.

The iteration (3–4) has the remarkable property that it generates an optimal rational approximation of $A^{1/2}$ of high degree. Namely, $\widetilde{X}_k := 2\alpha_k X_k/(1+\alpha_k) = r_{m_k,\ell_k}(A, \alpha, \sqrt{\cdot})$, where

$$(m_k, \ell_k) = \begin{cases} \left(\frac{1}{2}(2m)^k, \frac{1}{2}(2m)^k - 1\right), & \text{if } \ell = m-1, \\ \left(\frac{1}{2}((2m+1)^k - 1), \frac{1}{2}((2m+1)^k - 1)\right), & \text{if } \ell = m. \end{cases} \tag{5}$$

A simple consequence of this is that if A is Hermitian positive definite with eigenvalues in $[\alpha^2, 1]$, then

$$\|(\widetilde{X}_k - A^{1/2})A^{-1/2}\|_2 \le E_{m_k,\ell_k}(\sqrt{\cdot}, [\alpha^2, 1]),$$

where

$$E_{m,\ell}(f, S) = \min_{r\in\mathcal{R}_{m,\ell}} \max_{z\in S} \left|\frac{r(z) - f(z)}{f(z)}\right|.$$

For more detailed error estimates, including error estimates for non-normal A with eigenvalues in $\mathbb{C}\backslash(-\infty, 0]$, see [7]. Note that by definition,

$$E_{m,\ell}(f, [f^{-1}(\alpha), 1]) = \max_{z\in[f^{-1}(\alpha),1]} \left|\frac{r_{m,\ell}(z, \alpha, f) - f(z)}{f(z)}\right|.$$

3 Minimax Iterations for the Matrix *p*th Root

In this paper, we propose an iteration for computing pth roots of matrices that generalizes (3–4). Given $\alpha \in (0, 1)$, $m, \ell \in \mathbb{N}_0$, and an integer $p \ge 2$, the iteration reads

$$X_{k+1} = X_k \hat{r}_{m,\ell}\left(X_k^{-p} A, \alpha_k, \sqrt[p]{\cdot}\right), \qquad X_0 = I, \tag{6}$$

$$\alpha_{k+1} = \frac{\alpha_k}{\hat{r}_{m,\ell}(\alpha_k^p, \alpha_k, \sqrt[p]{\cdot})}, \qquad \alpha_0 = \alpha. \tag{7}$$

The Zolotarev iterations (3–4) correspond to the cases $\{(m, \ell, p) \mid m \in \mathbb{N}, \ell \in \{m-1, m\}, p = 2\}$ in (6–7). (Note that we colloquially referred to these cases as "the case $p = 2$" in Sect. 1). Since X_k is a rational function of A for each k, it commutes with A.

With the exception of the cases $\{(m, \ell, p) \mid m \in \mathbb{N}, \ell \in \{m-1, m\}, p = 2\}$ and $\{(m, \ell, p) \mid (m, \ell) \in \{(0, 0), (1, 0), (0, 1)\}, p \ge 2\}$, explicit formulas for $\hat{r}_{m,\ell}(z, \alpha, \sqrt[p]{\cdot})$ are not known. However, the coefficients of the numerator and denominator of $\hat{r}_{m,\ell}(z, \alpha, \sqrt[p]{\cdot})$ can be computed numerically; see Sect. 5 for details. Note that the cost of computing $\hat{r}_{m,\ell}(z, \alpha, \sqrt[p]{\cdot})$ is independent of the dimension of A, so it is expected to be negligible for problems involving large matrices.

As with the square root iteration (3–4), it is necessary to reformulate the pth root iteration (6–7) to ensure its stability. This is accomplished by considering the iteration for $Y_k = X_k^{1-p} A$ and $Z_k = X_k^{-1}$ implied by (6–7). Exploiting commutativity, we have

$$Y_{k+1} = Y_k h_{\ell,m,p}\left(Z_k Y_k, \alpha_k\right)^{p-1}, \qquad Y_0 = A, \tag{8}$$

$$Z_{k+1} = h_{\ell,m,p}\left(Z_k Y_k, \alpha_k\right) Z_k, \qquad Z_0 = I, \tag{9}$$

$$\alpha_{k+1} = \alpha_k h_{\ell,m,p}(\alpha_k^p, \alpha_k), \qquad \alpha_0 = \alpha, \tag{10}$$

where $h_{\ell,m,p}(z,\alpha) = r_{m,\ell}(z,\alpha,\sqrt[p]{\cdot})^{-1}$. (We swapped the order of the first two indices to emphasize that $h_{\ell,m,p}(z,\alpha)$ is a rational function of type (ℓ, m), not (m,ℓ).)

The remainder of this section presents a series of results about the behavior of the iteration (6–7) and its counterpart (8–10). Proofs of these results are given in Sect. 4.

3.1 Functional Iteration

A great deal of information about the behavior of the iteration (6–7) (and hence (8–10)) can be gleaned from a study of the functional iteration

$$f_{k+1}(z) = f_k(z)\hat{r}_{m,\ell}\left(\frac{z}{f_k(z)^p}, \alpha_k, \sqrt[p]{\cdot}\right), \qquad f_0(z) = 1, \tag{11}$$

$$\alpha_{k+1} = \frac{\alpha_k}{\hat{r}_{m,\ell}(\alpha_k^p, \alpha_k, \sqrt[p]{\cdot})}, \qquad \alpha_0 = \alpha. \tag{12}$$

Indeed, we have $X_k = f_k(A)$ in (6–7), and $Y_k = f_k(A)^{1-p}A$ and $Z_k = f_k(A)^{-1}$ in (8–10).

The following theorem summarizes the properties of the functional iteration (11–12). In the interest of generality, it focuses on a slight generalization of (11–12) that reduces to (11–12) when the function f appearing below is $f(z) = z^{1/p}$. The theorem makes use of the following terminology. A continuous function $g(z)$ is said to *equioscillate* m times on an interval $[a,b]$ if there exist m points $a \le z_0 < z_1 < \cdots < z_{m-1} \le b$ at which

$$g(z_j) = \sigma(-1)^j \max_{z\in[a,b]} |g(z)|, \quad j = 0, 1, \ldots, m-1.$$

for some $\sigma \in \{-1, 1\}$. It is well known that the minimax approximants (1) are uniquely characterized by the property that $\frac{r_{m,\ell}(z,\alpha,f)-f(z)}{f(z)}$ equioscillates at least $m+\ell+2-d$ times on $[f^{-1}(\alpha), 1]$, where d is the defect of $r_{m,\ell}(z,\alpha,f)$ in $\mathcal{R}_{m,\ell}$ [32, Theorem 24.1]. We will be particularly interested in those functions f for which:

(3.A) For every $\alpha \in (0,1)$ and $m, \ell \in \mathbb{N}_0$, $r_{m,\ell}(z,\alpha,f)$ has exact type (m,ℓ). Furthermore, $\frac{r_{m,\ell}(z,\alpha,f)-f(z)}{f(z)}$ equioscillates exactly $m+\ell+2$ times on $[f^{-1}(\alpha), 1]$, achieves its maximum at $z = f^{-1}(\alpha)$, and achieves an extremum at $z = 1$.

The function $f(z) = z^{1/p}$ satisfies this hypothesis; see Lemma 5 for a proof.

Theorem 1 *Let $f:[0,1]\to[0,1]$ be a continuous, increasing bijection satisfying* (3.A). *Let $\alpha\in(0,1)$ and $m,\ell\in\mathbb{N}_0$, and define $f_k(z)$ recursively by*

$$f_{k+1}(z)=f_k(z)\hat{r}_{m,\ell}\left(f^{-1}\left(\frac{f(z)}{f_k(z)}\right),\alpha_k,f\right),\qquad f_0(z)=1, \tag{13}$$

$$\alpha_{k+1}=\frac{\alpha_k}{\hat{r}_{m,\ell}(f^{-1}(\alpha_k),\alpha_k,f)},\qquad \alpha_0=\alpha. \tag{14}$$

Then, with $\widetilde{f}_k(z)=\frac{2\alpha_k}{1+\alpha_k}f_k(z)$ and $\varepsilon_k=\max_{z\in[f^{-1}(\alpha),1]}\left|\frac{\widetilde{f}_k(z)-f(z)}{f(z)}\right|$, we have:

(3.i) *For every $k\ge 0$,*

$$\alpha_k=\frac{1-\varepsilon_k}{1+\varepsilon_k} \tag{15}$$

and

$$\varepsilon_{k+1}=E_{m,\ell}(f,[f^{-1}(\alpha_k),1]). \tag{16}$$

(3.ii) *For every $k\ge 0$, the relative error $\frac{\widetilde{f}_k(z)-f(z)}{f(z)}$ equioscillates $(m+\ell+1)^k+1$ times on $[f^{-1}(\alpha),1]$, and it achieves its extrema at the endpoints.*

(3.iii) *If $f\in C^{m+\ell+1}([\alpha,1])$, f^{-1} is Lipschitz on $[\alpha,1]$, and $(m,\ell)\neq(0,0)$, then $\varepsilon_k\to 0$ monotonically with order of convergence $m+\ell+1$ as $k\to\infty$.*

Let us discuss the meaning of this theorem. It states that the iteration (13–14) generates a function $\widetilde{f}_k(z)\approx f(z)$ with the following curious property: The maximum relative error in $\widetilde{f}_k(z)$ on the interval $[f^{-1}(\alpha),1]$ is equal to the maximum relative error in the best rational approximant of $f(z)$ *on a much smaller interval* $[f^{-1}(\alpha_{k-1}),1]$. Indeed, as k increases, the length of $[f^{-1}(\alpha),1]$ remains constant, whereas the length of $[f^{-1}(\alpha_{k-1}),1]=[f^{-1}(\alpha_{k-1}),f^{-1}(1)]$ is $O(1-\alpha_{k-1})=O(\varepsilon_{k-1})$ by (15), assuming f^{-1} is Lipschitz near $z=1$. Since rational functions of type (m,ℓ) can approximate analytic functions on intervals of length $O(\varepsilon_{k-1})$ with (generically) accuracy $O(\varepsilon_{k-1}^{m+\ell+1})$ [32, Theorem 27.1], we see from (16) that $\varepsilon_k=O(\varepsilon_{k-1}^{m+\ell+1})$, assuming f is smooth enough near $z=1$. That is, $\varepsilon_k\to 0$ with order of convergence $m+\ell+1$.

For most functions f, the iteration (13–14) is not useful, as it (rather circularly) uses f (and f^{-1}) to generate an approximation of f. Furthermore, the approximation it generates need not be a rational function of z. The function $f(z)=z^{1/p}$, however, is exceptional, in that the iteration (13–14)—which reduces to (11–12) for this f—generates a rational function $f_k(z)$ without requiring the evaluation of any pth roots.

The following theorem specializes Theorem 1 to the case $f(z)=z^{1/p}$ and gives precise information about the constants implicit in the convergence result (3.iii). In it, we use the notation $(\beta)_m$ for the rising factorial (the Pochhammer symbol): $(\beta)_m=\beta(\beta+1)(\beta+2)\cdots(\beta+m-1)$.

Theorem 2 *Let $\alpha \in (0,1)$, $m, \ell \in \mathbb{N}_0$, and $p \in \mathbb{N}$ with $p \geq 2$ and $(m,\ell) \neq (0,0)$. Let $f_k(z)$ and α_k be defined by the iteration (11–12), and let $\widetilde{f}_k(z) = \frac{2\alpha_k}{1+\alpha_k} f_k(z)$ and $\varepsilon_k = \max_{z\in[\alpha^p,1]} \left| \frac{\widetilde{f}_k(z)-z^{1/p}}{z^{1/p}} \right|$. Then the conclusions (3.i) and (3.ii) hold with $f(z) = z^{1/p}$. Furthermore, as $k \to \infty$, $\varepsilon_k \to 0$ monotonically with*

$$\varepsilon_{k+1} = C(m,\ell,p)\varepsilon_k^{m+\ell+1} + o(\varepsilon_k^{m+\ell+1}), \tag{17}$$

where

$$C(m,\ell,p) = \frac{p^{m+\ell+1} m!\ell!(1/p)_{\ell+1}(1-1/p)_m}{2^{m+\ell}(m+\ell+1)!(m+\ell)!}. \tag{18}$$

Note that when $p = 2$ and $\ell \in \{m-1, m\}$, (18) simplifies to $C(m,\ell,2) = 4^{-(m+\ell)}$. This is consistent with the results of [7], where it is shown that for these m, ℓ, and p, an asymptotically sharp bound of the form $\varepsilon_k \leq 4\rho^{-(m+\ell+1)^k}$ holds with ρ a constant depending on α.

Let m_k and ℓ_k be the degrees of the polynomials in the numerator and denominator, respectively, of $\widetilde{f}_k$. Since the relative error $\frac{\widetilde{f}_k(z)-z^{1/p}}{z^{1/p}}$ equioscillates $(m+\ell+1)^k+1$ times on $[\alpha^p,1]$, it is natural to wonder how the number $(m+\ell+1)^k+1$ compares with $m_k+\ell_k+2$, the number of equioscillations achieved by $\operatorname{argmin}_{r\in\mathcal{R}_{m_k,\ell_k}} \max_{z\in[\alpha^p,1]} \left| \frac{r(z)-z^{1/p}}{z^{1/p}} \right|$ (which has defect 0 in $\mathcal{R}_{m_k,\ell_k}$; see Lemma 5). We address this question below.

Proposition 1 *Let α, m, ℓ, p, and $\widetilde{f}_k$ be as in Theorem 2. Then, for each $k \in \mathbb{N}$, $\widetilde{f}_k$ is a rational function of type (m_k, ℓ_k), where*

$$m_k = \begin{cases} \frac{1}{p}(pm)^k, & \text{if } \ell < m, \\ \frac{1}{p}\left[(p\ell+1)^k - (p(\ell-m)+1)^k\right], & \text{if } \ell \geq m, \end{cases}$$

$$\ell_k = \begin{cases} \frac{1}{p}(pm)^k - (m-\ell), & \text{if } \ell < m, \\ \frac{1}{p}\left[(p\ell+1)^k - 1\right], & \text{if } \ell \geq m. \end{cases}$$

As $k \to \infty$, the asymptotic relation

$$\frac{(m+\ell+1)^k+1}{m_k+\ell_k+2} \sim \begin{cases} \frac{p}{2}\left(\frac{m+\ell+1}{pm}\right)^k, & \text{if } \ell < m, \\ \frac{p}{2}\left(\frac{m+\ell+1}{p\ell+1}\right)^k, & \text{if } \ell \geq m \neq 0, \\ p\left(\frac{\ell+1}{p\ell+1}\right)^k, & \text{if } \ell > m = 0 \end{cases} \tag{19}$$

holds.

When $p = 2$ and $\ell \in \{m-1, m\}$, the asymptotic relation (19) is an equality: $\frac{(m+\ell+1)^k+1}{m_k+\ell_k+2} = 1$ for every k.

3.2 Convergence of the Matrix Iteration

An immediate consequence of Theorem 2 is that the iteration (6–7) converges when A is Hermitian positive definite with eigenvalues in $[\alpha^p, 1]$.

Corollary 1 *Let $\alpha \in (0,1)$, $m, \ell \in \mathbb{N}_0$, and $p, n \in \mathbb{N}$ with $p \geq 2$ and $(m,\ell) \neq (0,0)$. Let $A \in \mathbb{C}^{n\times n}$ be Hermitian positive definite. If the eigenvalues of A lie in $[\alpha^p, 1]$, then the iteration (6–7) generates a sequence $\widetilde{X}_k = 2\alpha_k X_k/(1+\alpha_k)$ that converges to $A^{1/p}$ with order $m+\ell+1$. In particular, we have*

$$\|\widetilde{X}_k A^{-1/p} - I\|_2 \leq \varepsilon_k,$$

for every $k \geq 0$, where ε_k obeys the recursion

$$\begin{aligned} \varepsilon_{k+1} &= E_{m,\ell}\left(\sqrt[p]{\cdot}, \left[\left(\frac{1-\varepsilon_k}{1+\varepsilon_k}\right)^p, 1\right]\right) = C(m,\ell,p)\varepsilon_k^{m+\ell+1} + o(\varepsilon_k^{m+\ell+1}), \\ \varepsilon_0 &= \frac{1-\alpha}{1+\alpha}, \end{aligned} \tag{20}$$

and $C(m,\ell,p)$ is given by (18).

A similar result holds for the coupled iteration (8–10).

Corollary 2 *Let α, m, ℓ, p, n, and A be as in Corollary 1. Then the coupled iteration (8–10) generates sequences $\widetilde{Y}_k = (1+\alpha_k)^{p-1} Y_k/(2\alpha_k)^{p-1}$ and $\widetilde{Z}_k = (1+\alpha_k) Z_k/(2\alpha_k)$ that converge to $A^{1/p}$ and $A^{-1/p}$ respectively, with order $m+\ell+1$. In particular, we have*

$$\begin{aligned} \|\widetilde{Y}_k A^{-1/p} - I\|_2 &\leq \frac{(1+\varepsilon_k)^{p-1} - 1}{(1-\varepsilon_k)^{p-1}}, \\ \|\widetilde{Z}_k A^{1/p} - I\|_2 &\leq \frac{\varepsilon_k}{1-\varepsilon_k}, \end{aligned}$$

for every $k \geq 0$, where ε_k obeys the recursion (20).

Note that the bounds above imply corresponding bounds on the relative errors $\|\widetilde{X}_k - A^{1/p}\|_2/\|A^{1/p}\|_2$, $\|\widetilde{Y}_k - A^{1/p}\|_2/\|A^{1/p}\|_2$, and $\|\widetilde{Z}_k - A^{-1/p}\|_2/\|A^{-1/p}\|_2$. For instance,

$$\frac{\|\widetilde{X}_k - A^{1/p}\|_2}{\|A^{1/p}\|_2} = \frac{\|(\widetilde{X}_k A^{-1/p} - I)A^{1/p}\|_2}{\|A^{1/p}\|_2} \leq \|\widetilde{X}_k A^{-1/p} - I\|_2 \leq \varepsilon_k.$$

When A is non-normal and/or has eigenvalues away from the positive real axis, the behavior of the matrix iteration (6–7) (and hence (8–10)) is dictated by the behavior of the scalar iteration (11–12) on complex inputs z. This has been analyzed in detail for the case $p = 2$ in [7], but for $p > 2$, numerical experiments indicate that the scalar iteration converges in a subset of the complex plane with fractal structure, a

typical feature of iterations for the pth root. We study this behavior numerically in Sect. 5. It remains an open problem to determine theoretically the convergence region $\{z \in \mathbb{C} \mid \lim_{k\to\infty} f_k(z) = z^{1/p}\}$ for the iteration (11–12).

3.3 Special Cases

For certain values of m, ℓ, and p, the theory above recovers some known results from the literature. We discuss these situations below.

3.3.1 Square Roots

When $p = 2, m \in \mathbb{N}$, and $\ell \in \{m-1, m\}$, a remarkable phenomenon occurs, allowing us to draw the connection between Theorem 1 and the results of [7] that we alluded to earlier. For these p, m, and ℓ, the function $\widetilde{f}_k(z)$ is a rational function of type (m_k, ℓ_k), where (m_k, ℓ_k) is given by (5). In both the case $\ell = m - 1$ and the case $\ell = m$, we have

$$m_k + \ell_k = (m + \ell + 1)^k - 1,$$

so (3.ii) implies that $\frac{\widetilde{f}_k(z) - f(z)}{f(z)}$ equioscillates $m_k + \ell_k + 2$ times on $[f^{-1}(\alpha), 1]$. It follows from the theory of rational minimax approximation that $\widetilde{f}_k(z)$ is the best rational approximant of $\sqrt{z}$ of type (m_k, ℓ_k) on $[\alpha^2, 1]$:

$$\widetilde{f}_k(z) = r_{m_k,\ell_k}(z, \alpha, \sqrt{\cdot}), \text{ if } p = 2 \text{ and } \ell \in \{m-1, m\}.$$

In particular,

$$\varepsilon_k = E_{m,\ell}(\sqrt{\cdot}, [\alpha_k^2, 1]) = E_{m_k,\ell_k}(\sqrt{\cdot}, [\alpha^2, 1]), \text{ if } p = 2 \text{ and } \ell \in \{m-1, m\},$$

for every $k \geq 1$. This shows that Theorem 1 includes [7, Theorem 1] as a special case.

3.3.2 Low-Order Iterations

When $p \geq 2$ is an integer and $(m, \ell) = (1, 0)$ or $(0, 1)$, we recover variants of another family of iterations.

Proposition 2 *Let $p \geq 2$ be an integer and $\alpha \in (0, 1)$. We have*

$$\hat{r}_{1,0}(z, \alpha, \sqrt[p]{\cdot}) = \frac{1}{p}\left((p-1)\mu + \frac{z}{\mu^{p-1}}\right), \quad \mu = \left(\frac{\alpha - \alpha^p}{(p-1)(1-\alpha)}\right)^{1/p}. \tag{21}$$

and

$$\hat{r}_{0,1}(z, \alpha, \sqrt[p]{\cdot}) = \frac{p}{(p+1)\nu - \nu^{p+1}z}, \quad \nu = \left(\frac{(p+1)(1-\alpha)}{1-\alpha^{p+1}}\right)^{1/p}. \tag{22}$$

Note that the formula (21) for $\hat{r}_{1,0}(z, \alpha, \sqrt[p]{\cdot})$ appears in [27, Theorem 2] and [23]; see also [14, Lemma 3.2] for a related result. (When comparing (21) with [27, Theorem 2], one must bear in mind that $r_{1,0}(z, \alpha, \sqrt[p]{\cdot})$ and $\hat{r}_{1,0}(z, \alpha, \sqrt[p]{\cdot})$ differ by a factor of $\frac{2}{1+\hat{r}_{1,0}(1,\alpha,\sqrt[p]{\cdot})} = \frac{2\mu^p p}{\mu+\mu^p(\mu(p-1)+p)}$.)

The preceding proposition shows that when $(m, \ell) = (1, 0)$, the iteration (6–7) reads

$$\begin{aligned} X_{k+1} &= \frac{1}{p}\left((p-1)\mu_k X_k + (\mu_k X_k)^{1-p} A\right), && X_0 = I, \\ \alpha_{k+1} &= \frac{p\alpha_k}{(p-1)\mu_k + \mu_k^{1-p}\alpha_k^p}, && \alpha_0 = \alpha, \end{aligned}$$

where

$$\mu_k = \left(\frac{\alpha_k - \alpha_k^p}{(p-1)(1-\alpha_k)}\right)^{1/p}. \tag{23}$$

This is a scaled variant of the popular Newton iteration [16, Equation 7.5] for the matrix pth root. The scaling heuristic above is reminiscent of one proposed by Hoskins and Walton [19], but theirs is based on type-$(1, 0)$ rational minimax approximants of $z^{(p-1)/p}$.

On the other hand, when $(m, \ell) = (0, 1)$, the iteration (6–7) reads

$$\begin{aligned} X_{k+1} &= pX_k\left((p+1)\nu_k I - \nu_k^{p+1} X_k^{-p} A)\right)^{-1}, && X_0 = I, \\ \alpha_{k+1} &= \frac{1}{p}\alpha_k\left((p+1)\nu_k - \nu_k^{p+1}\alpha_k^p\right), && \alpha_0 = \alpha, \end{aligned}$$

where

$$\nu_k = \left(\frac{(p+1)(1-\alpha_k)}{1-\alpha_k^{p+1}}\right)^{1/p}. \tag{24}$$

In terms of the matrix $Z_k = X_k^{-1}$, the iteration for X_k becomes

$$Z_{k+1} = \frac{1}{p}\left((p+1)\nu_k Z_k - (\nu_k Z_k)^{p+1} A\right), \qquad Z_0 = I,$$

which is a scaled variant of the inverse Newton iteration [16, Equation (7.12)] for computing $A^{-1/p}$.

3.3.3 Padé Iterations

We recover one more family of iterations by considering the limit as $\alpha \uparrow 1$ in (6–7).

Below, we say that a family of rational functions $\{r_\alpha \in \mathcal{R}_{m,\ell} \mid \alpha \in (0,1)\}$ converges coefficientwise to $r_1 \in \mathcal{R}_{m,\ell}$ as $\alpha \uparrow 1$ if the coefficients of the polynomials in the numerator and denominator of r_α, appropriately normalized, approach those of r_1 as $\alpha \uparrow 1$.

Proposition 3 *As $\alpha \uparrow 1$, $\hat{r}_{m,\ell}(z, \alpha, \sqrt[p]{\cdot})$ converges coefficientwise to the type-(m,ℓ) Padé approximant $P_{m,\ell,p}(z)$ of $z^{1/p}$ at $z = 1$:*

$$P_{m,\ell,p}(z) = \sum_{j=0}^{m} \frac{(-m)_j(-1/p-\ell)_j}{j!(-\ell-m)_j}(1-z)^j \Bigg/ \sum_{j=0}^{\ell} \frac{(1/p)_j(1/p-m)_m(j-\ell-m)_m}{j!(-\ell-m)_m(j+1/p-m)_m}(1-z)^j. \tag{25}$$

It follows that the iteration (6–7) reduces formally to

$$X_{k+1} = X_k P_{m,\ell,p}\left(X_k^{-p} A\right), \quad X_0 = I \tag{26}$$

as $\alpha \uparrow 1$. This is precisely the Padé iteration for the matrix pth root studied by Laszkiewicz and Ziętak [24, Equation (36)]. When $(m,\ell) = (1,1)$, it is the Halley iteration [21, p. 11], [13]. In terms of $Y_k = X_k^{1-p}A$ and $Z_k = X_k^{-1}$, the iteration (26) reads

$$Y_{k+1} = Y_k Q_{\ell,m,p}\left(Z_k Y_k\right)^{p-1}, \qquad Y_0 = A, \tag{27}$$

$$Z_{k+1} = Q_{\ell,m,p}\left(Z_k Y_k\right) Z_k, \qquad Z_0 = I, \tag{28}$$

where $Q_{\ell,m,p}(z) = P_{m,\ell,p}(z)^{-1}$.

For later use, it will be convenient to define

$$\hat{r}_{m,\ell}(z, 1, \sqrt[p]{\cdot}) := P_{m,\ell,p}(z),$$
$$h_{\ell,m,p}(z, 1) := Q_{\ell,m,p}(z).$$

The Padé iterations (26) and (27–28) are then simply the iterations obtained by setting $\alpha = 1$ in the minimax iterations (6–7) and (8–10), respectively.

3.4 Stability of the Coupled Matrix Iteration

As alluded to earlier, the uncoupled matrix iteration (6–7) exhibits numerical instability, whereas the coupled iteration (8–10) does not. We justify the latter claim below.

We recall the following definition. A matrix iteration $X_{k+1} = g(X_k)$ with fixed point X_* is said to be *stable* in a neighborhood of X_* if the Fréchet derivative of g at X_* has bounded powers [16, Definition 4.17]. That is, if $L_g(A,E)$ denotes the Fréchet derivative of g at $A \in \mathbb{C}^{n\times n}$ in a direction $E \in \mathbb{C}^{n\times n}$, then there exists a

constant $c > 0$ such that $\|G^j(E)\| \le c\|E\|$ for every j and every $E \in \mathbb{C}^{n\times n}$, where $G(E) = L_g(X_*, E)$.

We first address the stability of the coupled Padé iteration (27–28).

Proposition 4 *Let $m, \ell \in \mathbb{N}_0$ and $p, n \in \mathbb{N}$ with $(m, \ell) \neq (0, 0)$ and $p \ge 2$. The Padé iteration (27–28) is stable in a neighborhood of (B, B^{-1}) for any $B \in \mathbb{C}^{n\times n}$. In particular, with $g(Y, Z) = (YQ_{\ell,m,p}(ZY)^{p-1}, Q_{\ell,m,p}(ZY)Z)$, we have*

$$L_g(B, B^{-1}; E, F) = \frac{1}{p}\left(E - (p-1)BFB,\ (p-1)F - B^{-1}EB^{-1}\right)$$

for any $E, F \in \mathbb{C}^{n\times n}$, and $L_g(B, B^{-1}; \cdot, \cdot)$ is idempotent.

Consider now the coupled minimax iteration (8–10). Theorem 1 established that α_k converges to 1 in (10). We argue in Sect. 5 that when α_k is close to 1, it is numerically prudent to set α_k (and all subsequent iterates) equal to 1, thereby reverting to the Padé iteration (27–28). Since the latter iteration is stable, it follows that the aforementioned modification of (8–10) is stable as well.

4 Proofs

In this section, we prove Theorems 1 and 2, Corollaries 1 and 2, and Propositions 1, 2, 3, and 4.

4.1 Proof of Theorem 1

4.1.1 Equioscillation

To prove the claims (3.i) and (3.ii) in Theorem 1, we use an inductive argument. When $k = 0$, (3.ii) holds since the relative error $\frac{\widetilde{f}_0(z)-f(z)}{f(z)} = \frac{2\alpha}{f(z)(1+\alpha)} - 1$ decreases monotonically from $\frac{1-\alpha}{1+\alpha}$ to $-\frac{1-\alpha}{1+\alpha}$ as z runs from $f^{-1}(\alpha)$ to 1. This shows also that $\varepsilon_0 = \frac{1-\alpha}{1+\alpha}$, so (15) holds when $k = 0$. Next, we prove two lemmas in preparation for the inductive step.

Lemma 1 *Let $f\colon [0, 1] \to [0, 1]$ be a continuous, increasing bijection satisfying* (3.A). *Then the recurrence* (14) *is equivalent to*

$$\alpha_{k+1} = \frac{1 - E_{m,\ell}(f, [f^{-1}(\alpha_k), 1])}{1 + E_{m,\ell}(f, [f^{-1}(\alpha_k), 1])}, \quad \alpha_0 = \alpha. \tag{29}$$

Proof Since

$$\min_{z\in[f^{-1}(\alpha),1]} \frac{r_{m,\ell}(z, \alpha, f)}{f(z)} = 1 - E_{m,\ell}(f, [f^{-1}(\alpha), 1]),$$

the defining property (2) of $\hat{r}_{m,\ell}(z, \alpha, f)$ implies that

$$\hat{r}_{m,\ell}(z, \alpha, f) = \frac{r_{m,\ell}(z, \alpha, f)}{1 - E_{m,\ell}(f, [f^{-1}(\alpha), 1])}.$$

Also, the assumption (3.A) implies that

$$\frac{r_{m,\ell}(f^{-1}(\alpha), \alpha, f)}{f(f^{-1}(\alpha))} = \max_{z \in [f^{-1}(\alpha), 1]} \frac{r_{m,\ell}(z, \alpha, f)}{f(z)} = 1 + E_{m,\ell}(f, [f^{-1}(\alpha), 1]),$$

so

$$\frac{\alpha}{\hat{r}_{m,\ell}(f^{-1}(\alpha), \alpha, f)} = \frac{1 - E_{m,\ell}(f, [f^{-1}(\alpha), 1])}{1 + E_{m,\ell}(f, [f^{-1}(\alpha), 1])}.$$

Since this holds for any $\alpha \in (0, 1)$, it follows that the recurrence (14) is equivalent to (29). □

Lemma 2 *Let $f : [0, 1] \to [0, 1]$ be a continuous, increasing bijection satisfying* (3.A). *Let $\alpha \in (0, 1)$ and $m, \ell \in \mathbb{N}_0$. Let $\widetilde{F}(z)$ be any continuous function on $[f^{-1}(\alpha), 1]$ with the property that $\frac{\widetilde{F}(z) - f(z)}{f(z)}$ equioscillates q times on $[f^{-1}(\alpha), 1]$ and achieves its extrema $\pm\varepsilon$ at the endpoints, where $q \geq 2$ and $0 < \varepsilon < 1$. Define*

$$\begin{aligned}
\alpha' &= \frac{1 - \varepsilon}{1 + \varepsilon}, \\
\alpha'' &= \frac{1 - E_{m,\ell}(f, [f^{-1}(\alpha'), 1])}{1 + E_{m,\ell}(f, [f^{-1}(\alpha'), 1])}, \\
F(z) &= \frac{1 + \alpha'}{2\alpha'} \widetilde{F}(z), \\
H(z) &= \frac{2\alpha''}{1 + \alpha''} F(z) \hat{r}_{m,\ell}\left(f^{-1}\left(\frac{f(z)}{F(z)}\right), \alpha', f\right).
\end{aligned}$$

Then $\frac{H(z) - f(z)}{f(z)}$ equioscillates $(m + \ell + 1)(q - 1) + 1$ times on $[f^{-1}(\alpha), 1]$ with extrema $\pm E_{m,\ell}(f, [f^{-1}(\alpha'), 1])$, and it achieves its extrema at the endpoints.

Proof The assumed equioscillation of $\frac{\widetilde{F}(z)}{f(z)} - 1$ on $[f^{-1}(\alpha), 1]$ implies that the function $\frac{\widetilde{F}(f^{-1}(z))}{z} - 1$ equioscillates q times on $[\alpha, 1]$ with extrema $\pm\varepsilon$. If we now define

$$S(z) = \frac{z(1 - \varepsilon^2)}{\widetilde{F}(f^{-1}(z))},$$

then we conclude that $S(z) - 1$ equioscillates q times on $[\alpha, 1]$ with extrema $\frac{1 - \varepsilon^2}{1 \pm \varepsilon} - 1 = \mp\varepsilon$. Moreover, it achieves its extrema at the endpoints by our assumptions on $\widetilde{F}$.

By the same reasoning as above, the function

$$s_{m,\ell}(z,\alpha',f)=\frac{z(1-(\varepsilon')^2)}{r_{m,\ell}(f^{-1}(z),\alpha',f)},\quad \varepsilon'=E_{m,\ell}(f,[f^{-1}(\alpha'),1]),$$

has the property that $s_{m,\ell}(z,\alpha',f)-1$ equioscillates $m+\ell+2$ times on $[\alpha',1]$ with extrema $\pm\varepsilon'$, and it achieves its extrema at the endpoints by the assumption (3.A).

Consider now the function

$$g(z)=s_{m,\ell}\left(\frac{S(z)}{1+\varepsilon},\alpha',f\right). \tag{30}$$

We claim that $g(z)-1$ equioscillates on $[\alpha,1]$ with extrema $\pm\varepsilon'$. To see this, we make two observations. First, as z runs from α to 1, $\frac{S(z)}{1+\varepsilon}$ runs from/to $\frac{1-\varepsilon}{1+\varepsilon}=\alpha'$ to/from $\frac{1+\varepsilon}{1+\varepsilon}=1$ a total of $q-1$ times, achieving its extrema at the endpoints each time. Second, each time $y=\frac{S(z)}{1+\varepsilon}$ runs from/to α' to/from 1, $s_{m,\ell}(y,\alpha',f)-1$ equioscillates $m+\ell+2$ times with extrema $\pm\varepsilon'$. By counting extrema, we conclude that the composition (30) (minus 1) equioscillates

$$(m+\ell+2)(q-1)-(q-2)=(m+\ell+1)(q-1)+1$$

times on $[\alpha,1]$ with extrema $\pm\varepsilon'$.

Finally, consider the function

$$h(z)=\frac{(1-(\varepsilon')^2)}{g(f(z))}.$$

In view of the equioscillation of (30), the function $h(z)-1$ equioscillates $(m+\ell+1)(q-1)+1$ times on $[f^{-1}(\alpha),1]$ with extrema $\frac{1-(\varepsilon')^2}{1\pm\varepsilon'}-1=\mp\varepsilon'$, and it achieves its extrema at the endpoints. We will complete the proof by showing that $h(z)=\frac{H(z)}{f(z)}$. Using the fact that $1-\varepsilon'=\frac{2\alpha''}{1+\alpha''}$, $\widetilde{F}(z)=(1-\varepsilon)F(z)$, and $r_{m,\ell}(z,\alpha',f)=(1-\varepsilon')\hat{r}_{m,\ell}(z,\alpha',f)$, we have

$$\begin{aligned}
h(z)&=\frac{(1-(\varepsilon')^2)}{s_{m,\ell}\left(\frac{S(f(z))}{1+\varepsilon},\alpha',f\right)}\\
&=\frac{r_{m,\ell}\left(f^{-1}\left(\frac{S(f(z))}{1+\varepsilon}\right),\alpha',f\right)}{\frac{S(f(z))}{1+\varepsilon}}\\
&=\frac{r_{m,\ell}\left(f^{-1}\left(\frac{f(z)(1-\varepsilon)}{\widetilde{F}(z)}\right),\alpha',f\right)}{\frac{f(z)(1-\varepsilon)}{\widetilde{F}(z)}}\\
&=(1-\varepsilon')\frac{F(z)\hat{r}_{m,\ell}\left(f^{-1}\left(\frac{f(z)}{F(z)}\right),\alpha',f\right)}{f(z)}
\end{aligned}$$

$$= \frac{H(z)}{f(z)}.$$

□

Remark 1 When $f(z) = z^{1/p}$, the function

$$s_{m,\ell}(z, \alpha', \sqrt[p]{\cdot}) = \frac{z(1-(\varepsilon')^2)}{r_{m,\ell}(z^p, \alpha', \sqrt[p]{\cdot})}$$

appearing in the proof above is a rational approximant of the *sector function* $\mathrm{sect}_p(z) = z/(z^p)^{1/p}$; see Fig. 1. In fact, the proof above reveals that on each of the segments $\{z \in \mathbb{C} \mid e^{-2\pi ij/p} z \in [\alpha', 1]\}$, $j = 0, 1, 2, \ldots, p-1$, the relative error

$$\frac{s_{m,\ell}(z, \alpha', \sqrt[p]{\cdot}) - \mathrm{sect}_p(z)}{\mathrm{sect}_p(z)} = e^{-2\pi ij/p} s_{m,\ell}(z, \alpha', \sqrt[p]{\cdot}) - 1$$

is real-valued and equioscillates $m + \ell + 2$ times with extrema $\pm\varepsilon'$. In particular, for $\ell \in \{m-1, m\}$, $s_{m,\ell}(z, \alpha', \sqrt{\cdot})$ is Zolotarev's type-$(2\ell+1, 2m)$ best rational approximant of the sign function $\mathrm{sign}(z) = z/(z^2)^{1/2}$ on $[-1, -\alpha'] \cup [\alpha', 1]$ [29].

We are now ready to prove (3.i–3.ii). Suppose (3.ii) and (15) hold at step k in the iteration (11–12). Then Lemma 2 (applied with $\widetilde{F} = \widetilde{f}_k$, $\varepsilon = \varepsilon_k$, and $q = (m+\ell+1)^k + 1$, so that $\alpha' = \alpha_k$ and $\alpha'' = \alpha_{k+1}$) implies that (3.ii) and (15) hold at step $k+1$, so in fact they hold for all k. It now follows immediately that (16) is equivalent to (29), which, in turn, is equivalent to (14) by Lemma 1. This completes the proof of (3.i–3.ii).

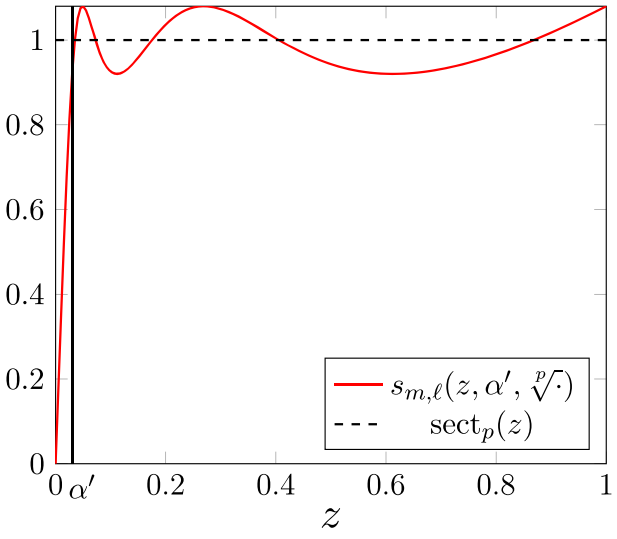

Fig. 1 Plots of $s_{m,\ell}(z, \alpha', \sqrt[p]{\cdot})$ and $\mathrm{sect}_p(z)$ with $m = 2$, $\ell = 2$, $p = 3$, and $\alpha' = 0.03$

4.1.2 Convergence

We now address the last claim (3.iii) of Theorem 1, which concerns the convergence of ε_k to 0 in the iteration

$$\varepsilon_{k+1} = G(\varepsilon_k), \quad \varepsilon_0 = \frac{1-\alpha}{1+\alpha}, \tag{31}$$

with $\alpha \in (0, 1)$,

$$G(\varepsilon) = E_{m,\ell}\left(f, \left[f^{-1}\left(\frac{1-\varepsilon}{1+\varepsilon}\right), 1\right]\right), \tag{32}$$

and $(m, \ell) \neq (0, 0)$.

Lemma 3 *Let $m, \ell \in \mathbb{N}_0$, and let $f:[0,1] \to [0,1]$ be a continuous, increasing bijection satisfying* (3.A). *If $(m, \ell) \neq (0, 0)$, then G is continuous, nonnegative, and nondecreasing on* $(0, 1)$. *Furthermore, $G(\varepsilon) < \varepsilon$ for every $\varepsilon \in (0, 1)$.*

Proof It is obvious that G is nonnegative and nondecreasing. To show that $G(\varepsilon) < \varepsilon$ for every $\varepsilon \in (0, 1)$, note that (32) is no larger than the uniform relative error committed by the constant function $g(z) = 1 - \varepsilon$:

$$-\varepsilon = \frac{1-\varepsilon-f(1)}{f(1)} \leq \frac{g(z)-f(z)}{f(z)} \leq \frac{1-\varepsilon-f\left(f^{-1}\left(\frac{1-\varepsilon}{1+\varepsilon}\right)\right)}{f\left(f^{-1}\left(\frac{1-\varepsilon}{1+\varepsilon}\right)\right)} = \varepsilon$$

for every $z \in \left[f^{-1}\left(\frac{1-\varepsilon}{1+\varepsilon}\right), 1\right]$. This establishes that $G(\varepsilon) \leq \varepsilon$. The inequality is in fact strict since we assumed (3.A), which implies that the minimizer of the relative error is not a constant function when $(m, \ell) \neq (0, 0)$. It remains to show that G is continuous on $(0, 1)$. We assumed in (3.A) that the minimizer for $E_{m,\ell}(f, [f^{-1}(\alpha), 1])$ has defect 0 in $\mathcal{R}_{m,\ell}$ for each $\alpha \in (0, 1)$, so, for each fixed $\alpha \in (0, 1)$, the map $g \mapsto r_{m,\ell}(\cdot, \alpha, g)$ is continuous with respect to the uniform norm at $g = f$ [26]. By considering functions g obtained by scaling and translating the input to f, we deduce that $r_{m,\ell}(\cdot, \alpha, f)$ depends continuously on $\alpha \in (0, 1)$, again with respect to the uniform norm. Hence, the map $\alpha \mapsto E_{m,\ell}(f, [f^{-1}(\alpha), 1])$ is continuous on $(0, 1)$, and so too is G. □

It follows from the above properties of G that $\varepsilon_k \to 0$ monotonically in the iteration $\varepsilon_{k+1} = G(\varepsilon_k)$ for every $\varepsilon_0 \in (0, 1)$.

4.1.3 Rate of Convergence

It remains to show that the order of convergence of ε_k to 0 is $m+\ell+1$. As we explained in the paragraph below Theorem 1, it suffices to note that when f is $C^{m+\ell+1}$ in a neighborhood of 1,

$$E_{m,\ell}(f, [a, 1]) = O((1-a)^{m+\ell+1}), \quad \text{as } a \to 1.$$

Indeed, this, together with (16), gives

$$\varepsilon_{k+1} = O\left(\left(1 - f^{-1}\left(\frac{1-\varepsilon_k}{1+\varepsilon_k}\right)\right)^{m+\ell+1}\right) = O(\varepsilon_k^{m+\ell+1}), \tag{33}$$

assuming that f^{-1} is Lipschitz near 1 and $f^{-1}(1) = 1$. Below, we give more precise information about the constant implicit in (33). We begin with a lemma that shows, in essence, that the uniform error in the best type-(m, ℓ) rational approximant of a function $g(z)$ on a small interval $[-\delta, \delta]$ is about $2^{m+\ell}$ times smaller than the uniform error in the type-(m, ℓ) Padé approximant of $g(z)$. (Note that this does not contradict Proposition 3; the difference between the two aforementioned uniform errors still tends to 0 as $\delta \to 0$.)

Lemma 4 *Let $g(z)$ be $C^{m+\ell+1}$ and positive in a neighborhood of 0. Assume that the type-(m, ℓ) Padé approximant $p(z)$ of $g(z)$ about 0 has defect 0 in $\mathcal{R}_{m,\ell}$, and*

$$p(z) - g(z) = c_g z^{m+\ell+1} + o(z^{m+\ell+1}),$$

where $c_g \in \mathbb{R}$. For each $\delta > 0$, let

$$r_\delta = \operatorname*{arg\,min}_{r \in \mathcal{R}_{m,\ell}} \max_{-\delta \le z \le \delta} \left| \frac{r(z) - g(z)}{g(z)} \right|.$$

Then, as $\delta \to 0$,

$$\max_{-\delta \le z \le \delta} \left| \frac{r_\delta(z) - g(z)}{g(z)} \right| = \frac{2|c_g|}{g(0)} \left(\frac{\delta}{2}\right)^{m+\ell+1} + o(\delta^{m+\ell+1}).$$

Proof Let

$$q = \operatorname*{arg\,min}_{r \in \mathcal{R}_{m+\ell,0}} \max_{-\delta \le z \le \delta} |r(z) - z^{m+\ell+1}|. \tag{34}$$

Among polynomials of degree $m + \ell + 1$ with unit leading coefficient, the polynomial $z^{m+\ell+1} - q(z)$ is the one that deviates least from 0 on $[-\delta, \delta]$. Up to a rescaling, this is precisely the degree-$(m + \ell + 1)$ Chebyshev polynomial of the first kind $T_{m+\ell+1}(z)$:

$$z^{m+\ell+1} - q(z) = 2\left(\frac{\delta}{2}\right)^{m+\ell+1} T_{m+\ell+1}\left(\frac{z}{\delta}\right).$$

Now let $R(z)$ be the type-(m, ℓ) Padé approximant of

$$\bar{g}(z) = g(z) - c_g q(z).$$

Since we assumed that the Padé approximant of $g(z)$ has defect 0 in $\mathcal{R}_{m,\ell}$, the Taylor coefficients of $R(z)$ approach those of $p(z)$ as $\delta \to 0$ [34, Corollary of Theorem 2a]. It follows that for each $\delta > 0$ sufficiently small,

$$R(z) - \bar{g}(z) = \bar{c}_g z^{m+\ell+1} + o(z^{m+\ell+1}),$$

for some $\bar{c}_g$ with $\bar{c}_g - c_g = o(1)$ as $\delta \to 0$. Thus, for each $\delta > 0$ sufficiently small,

$$\begin{aligned} R(z) - g(z) &= R(z) - \bar{g}(z) - c_g q(z) \\ &= \bar{c}_g z^{m+\ell+1} - c_g z^{m+\ell+1} + 2c_g \left(\frac{\delta}{2}\right)^{m+\ell+1} T_{m+\ell+1}\left(\frac{z}{\delta}\right) + o(z^{m+\ell+1}). \end{aligned}$$

Hence, as $\delta \to 0$,

$$R(z) - g(z) = 2c_g \left(\frac{\delta}{2}\right)^{m+\ell+1} T_{m+\ell+1}\left(\frac{z}{\delta}\right) + o(\delta^{m+\ell+1})$$

for every $z \in [-\delta, \delta]$, uniformly in z. Multiplying by $\frac{1}{g(z)} = \frac{1}{g(0)} + o(1)$, we conclude that

$$\frac{R(z) - g(z)}{g(z)} = \frac{2c_g}{g(0)} \left(\frac{\delta}{2}\right)^{m+\ell+1} T_{m+\ell+1}\left(\frac{z}{\delta}\right) + o(\delta^{m+\ell+1}) \tag{35}$$

for every $z \in [-\delta, \delta]$, uniformly in z. Finally, by the definition of r_δ,

$$\max_{-\delta \le z \le \delta} \left| \frac{r_\delta(z) - g(z)}{g(z)} \right| \le \max_{-\delta \le z \le \delta} \left| \frac{R(z) - g(z)}{g(z)} \right| = \frac{2c_g}{g(0)} \left(\frac{\delta}{2}\right)^{m+\ell+1} + o(\delta^{m+\ell+1}).$$

In fact, this bound is sharp, for the following reason. The relation (35) shows that for δ sufficiently small, $\frac{R(z)-g(z)}{g(z)}$ approximately equioscillates, in the sense that there exist $m + \ell + 2$ points $-\delta \le z_0 \le z_1 \le \cdots \le z_{m+\ell+1} \le \delta$ at which $\frac{R(z)-g(z)}{g(z)}$ alternates in sign and satisfies

$$\left| \frac{R(z_j) - g(z_j)}{g(z_j)} \right| \ge \frac{2|c_g|}{g(0)} \left(\frac{\delta}{2}\right)^{m+\ell+1} - \gamma, \quad j = 0, 1, \ldots, m + \ell + 1,$$

where $\gamma = o(\delta^{m+\ell+1})$. The de la Vallée Poussin lower bound [32, Exercise 24.5] then implies that

$$\max_{-\delta \le z \le \delta} \left| \frac{r_\delta(z) - g(z)}{g(z)} \right| \ge \frac{2|c_g|}{g(0)} \left(\frac{\delta}{2}\right)^{m+\ell+1} - \gamma.$$

□

Remark 2 The proof above suggests a heuristic for constructing near-best rational minimax approximants on short intervals $[-\delta, \delta]$: one computes the Padé approximant of $\bar{g}(z) = g(z) - c_g z^{m+\ell+1} + 2c_g(\delta/2)^{m+\ell+1} T_{m+\ell+1}(z/\delta)$ rather than $g(z)$. In view of (35), this heuristic is closely related to Chebyshev–Padé approximation [35].

Remark 3 The near equioscillation of R in the proof above can be used to show that R is close to r_δ: $R(z) - r_\delta(z) = o(\delta^{m+\ell+1})$, uniformly in $z \in [-\delta, \delta]$ as $\delta \to 0$. The argument is essentially the same as the one used in [33, pp. 429–430] to show that Charathéodory–Fejér approximants are close to minimax approximants on small intervals. See Fig. 2 for an illustration.

It is now a simple matter to estimate the constant implicit in (33). As $\varepsilon \to 0$, the above lemma gives

$$\begin{aligned} G(\varepsilon) &= E_{m,\ell}\left(f, \left[f^{-1}\left(\frac{1-\varepsilon}{1+\varepsilon}\right), 1\right]\right) \\ &= \max_{f^{-1}\left(\frac{1-\varepsilon}{1+\varepsilon}\right) \le z \le 1} \left|\frac{r_{m,\ell}(z, \alpha, f) - f(z)}{f(z)}\right| \\ &= \frac{2|c_{f,\delta}|}{f(1-\delta)}\left(\frac{\delta}{2}\right)^{m+\ell+1} + o(\delta^{m+\ell+1}), \end{aligned}$$

where

$$\delta = \frac{1}{2}\left(1 - f^{-1}(\alpha)\right), \quad \alpha = \frac{1-\varepsilon}{1+\varepsilon},$$

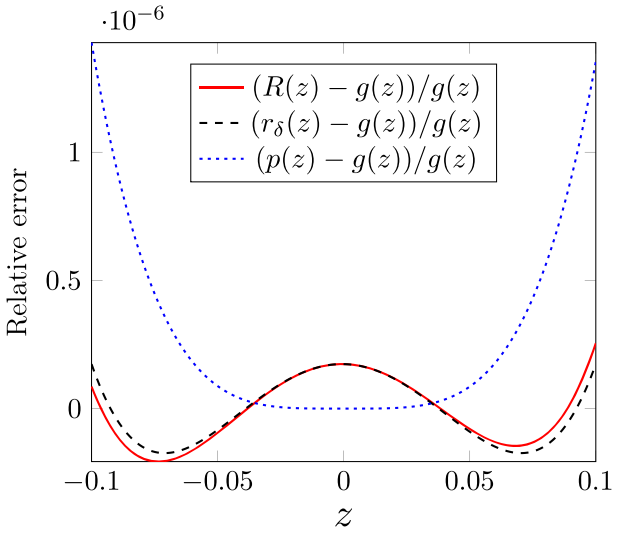

Fig. 2 Relative errors in $R(z)$, $r_\delta(z)$, and the type-(m, ℓ) Padé approximant $p(z)$ of $g(z) = e^z$ with $m = 2$, $\ell = 1$, and $\delta = 0.1$

and $c_{f,\delta}$ is the Taylor coefficient of $(z-1+\delta)^{m+\ell+1}$ in the difference between $f(z)$ and its type-(m,ℓ) Padé approximant about $z=1-\delta$. Since $\frac{1}{f(1-\delta)} = \frac{1}{f(1)} + o(1) = 1 + o(1)$, we have

$$G(\varepsilon) = 2|c_{f,\delta}| \left(\frac{\delta}{2}\right)^{m+\ell+1} + o(\delta^{m+\ell+1}).$$

A short calculation shows that $\delta = \varepsilon (f^{-1})'(1) + o(\varepsilon) = \varepsilon / f'(1) + o(\varepsilon)$ and $c_f := c_{f,0} = c_{f,\delta} + o(1)$, so

$$G(\varepsilon) = \frac{|c_f|}{2^{m+\ell} f'(1)^{m+\ell+1}} \varepsilon^{m+\ell+1} + o(\varepsilon^{m+\ell+1}).$$

It follows that in the iteration (31), we have

$$\varepsilon_{k+1} = \frac{|c_f|}{2^{m+\ell} f'(1)^{m+\ell+1}} \varepsilon_k^{m+\ell+1} + o(\varepsilon_k^{m+\ell+1}). \tag{36}$$

4.2 Proof of Theorem 2

Having proved Theorem 1, we now verify that the function $f(z) = z^{1/p}$ satisfies the hypothesis (3.A), and we prove Theorem 2.

We begin by establishing a few properties of the minimax approximants $r_{m,\ell}(z, \alpha, \sqrt[p]{\cdot})$. The proof of the following lemma is similar to that of [31, Lemma 2], which studies rational functions of type $(\ell+1, \ell)$ that minimize the maximum *absolute* error on $[0, 1]$ rather than the maximum *relative* error on $[\alpha, 1]$, $\alpha > 0$. The proof makes use of the following terminology. A *Chebyshev system* of dimension N on an interval $I \subseteq \mathbb{R}$ is a linearly independent set $\{g_j(z)\}_{j=1}^N$ of continuous functions on I with the property that any nontrivial linear combination $\sum_{j=1}^N c_j g_j(z)$ has at most $N-1$ (distinct) roots in I.

Lemma 5 *Let $m, \ell \in \mathbb{N}_0$, $0 < a < b < \infty$, and $p \in \mathbb{N}$, $p \geq 2$. If $r \in \mathcal{R}_{m,\ell}$ minimizes*

$$\max_{z\in[a,b]} |e(z)|, \quad e(z) = \frac{r(z) - z^{1/p}}{z^{1/p}},$$

then r has exact type (m,ℓ), $e(z)$ equioscillates exactly $m+\ell+2$ times on $[a,b]$, and

$$e(a) = \max_{z\in[a,b]} |e(z)|, \tag{37}$$

$$e(b) = (-1)^{m+\ell+1} \max_{z\in[a,b]} |e(z)|. \tag{38}$$

Proof Suppose that $r(z) = g(z)/h(z)$, where $g(z)$ and $h(z)$ are polynomials of exact degree $m' \leq m$ and $\ell' \leq \ell$, respectively. Observe that the function

$$z^{1/p} h(z) e(z) = g(z) - z^{1/p} h(z)$$

belongs to the space W spanned by

$$\{1, z, z^2, \ldots, z^{m'}, z^{1/p}, z^{1+1/p}, z^{2+1/p}, \ldots, z^{\ell'+1/p}\},$$

which is a Chebyshev system on $[a, b]$ of dimension $m' + \ell' + 2$ [22, p. 9, Example 1]. Thus, $z^{1/p}h(z)e(z)$ has at most $m' + \ell' + 1$ zeros on $[a, b]$. In particular, $e(z)$ has at most $m' + \ell' + 1$ zeros on $[a, b]$, so it equioscillates at most $m' + \ell' + 2$ times on $[a, b]$. But $e(z)$ equioscillates at least $m + \ell + 2 - d$ times on $[a, b]$, where $d = \min\{m - m', \ell - \ell'\} \geq 0$. It follows that

$$m' + \ell' + 2 \geq m + \ell + 2 - d,$$

so

$$d \geq (m - m') + (\ell - \ell') \geq 2d.$$

From this we conclude that $d = 0$, $m' = m$, $\ell' = \ell$, and $e(z)$ equioscillates exactly $m + \ell + 2$ times on $[a, b]$.

Let $a \leq z_0 < z_1 < \cdots < z_{m+\ell+1} \leq b$ be the points at which $e(z)$ achieves its extrema on $[a, b]$. Suppose that $z_0 > a$ or $z_{m+\ell+1} < b$. By considering the graph of $e(z)$, one easily deduces that there exists $c \in \mathbb{R}$ such that $e(z) - c$ has at least $m + \ell + 2$ roots in $[a, b]$. But

$$z^{1/p}h(z)(e(z) - c) = z^{1/p}h(z)e(z) - cz^{1/p}h(z) \in W,$$

so $z^{1/p}h(z)(e(z) - c)$ has at most $m' + \ell' + 1 = m + \ell + 1$ roots in $[a, b]$. In particular, $e(z) - c$ has at most $m + \ell + 1$ roots in $[a, b]$, a contradiction. It follows that $z_0 = a$ and $z_{m+\ell+1} = b$.

It remains to verify that the signs in (37–38) are correct. Consider the dependence of $e(z)$ on the parameters a and b. Denote this dependence by $e(z; a, b)$. By an argument similar to the one made in the proof of Lemma 3, the maps $a \mapsto e(a; a, b)$ and $b \mapsto e(a; a, b)$ are continuous on $(0, b)$ and (a, ∞), respectively. These maps also have no zeros, since $e(z; a, b)$ has a nonzero extremum at $z = a$ for every $0 < a < b < \infty$. Now, for small $\delta > 0$, the proof of Lemma 4 shows that for $z \in [1 - \delta, 1 + \delta]$,

$$e(z; 1 - \delta, 1 + \delta) = 2c_f \left(\frac{\delta}{2}\right)^{m+\ell+1} T_{m+\ell+1}\left(\frac{z - 1}{\delta}\right) + o(\delta^{m+\ell+1}),$$

where c_f is the coefficient of $(z-1)^{m+\ell+1}$ in the Taylor expansion of $P_{m,\ell,p}(z) - z^{1/p}$ about $z = 1$. In particular, $e(1-\delta; 1-\delta, 1+\delta)$ has the same sign as $c_f T_{m+\ell+1}(-1) = (-1)^{m+\ell+1}c_f$ for δ close to 0, which, as we verify below in (40), is positive. By continuity, $e(a; a, b) > 0$ for every $0 < a < b < \infty$, and (37–38) follow. □

The preceding lemma shows that the function $f(z) = z^{1/p}$ satisfies the hypothesis (3.A), so Theorem 2 will follow if we can show that the constant $C(m, \ell, p)$ in the

estimate (17) is given by (18). In view of the general estimate (36), it suffices to determine the coefficient c_f of the leading-order term $c_f(z-1)^{m+\ell+1}$ in $P_{m,\ell,p}(z) - z^{1/p}$, where $P_{m,\ell,p}(z)$ is the Padé approximant (25) of $z^{1/p}$ about $z = 1$. This is given by [11, Lemma 3.12]

$$c_f = (-1)^{m+\ell+1} \frac{m!\ell!(1/p)_{\ell+1}(1-1/p)_m}{(m+\ell+1)!(m+\ell)!}. \tag{39}$$

Inserting this into (36) and noting that $f'(1) = \frac{1}{p}$ and

$$|c_f| = (-1)^{m+\ell+1} c_f, \tag{40}$$

we obtain (18).

4.3 Proof of Proposition 1

To prove Proposition 1, it suffices to analyze f_k, which is a rational function of the same type as $\widetilde{f_k}$. Write $f_k(z) = \frac{u_k(z)}{v_k(z)}$ with u_k and v_k polynomials of degree m_k and ℓ_k, respectively. Since

$$\hat{r}_{m,\ell}(z, \alpha, \sqrt[p]{\cdot}) = \frac{a_m z^m + a_{m-1} z^{m-1} + \cdots + a_0}{b_\ell z^\ell + b_{\ell-1} z^{\ell-1} + \cdots + b_0}$$

for some coefficients a_j and b_j depending on α, m, ℓ, and p, we have

$$\begin{aligned} f_{k+1}(z) &= f_k(z) \hat{r}_{m,\ell}\left(\frac{z}{f_k(z)^p}, \alpha_k, \sqrt[p]{\cdot}\right) \\ &= \frac{u_k(z)}{v_k(z)} \left(\frac{a_m \left(\frac{z v_k(z)^p}{u_k(z)^p}\right)^m + a_{m-1}\left(\frac{z v_k(z)^p}{u_k(z)^p}\right)^{m-1} + \cdots + a_0}{b_\ell \left(\frac{z v_k(z)^p}{u_k(z)^p}\right)^\ell + b_{\ell-1}\left(\frac{z v_k(z)^p}{u_k(z)^p}\right)^{\ell-1} + \cdots + b_0} \right), \end{aligned}$$

where the coefficients a_j and b_j vary with the iteration number k. In the case where $\ell < m$, we can write this as a ratio of two polynomials,

$$\begin{aligned} &f_{k+1}(z) \\ &= \frac{a_m z^m v_k(z)^{pm} + a_{m-1} z^{m-1} u_k(z)^p v_k(z)^{p(m-1)} + \cdots + a_0 u_k(z)^{pm}}{b_\ell z^\ell u_k(z)^{p(m-\ell)-1} v_k(z)^{p\ell+1} + b_{\ell-1} z^{\ell-1} u_k(z)^{p(m-\ell+1)-1} v_k(z)^{p(\ell-1)+1} + \cdots + b_0 u_k(z)^{pm-1} v_k(z)}. \end{aligned}$$

An inductive argument shows that the terms $a_0 u_k(z)^{pm}$ and $b_0 u_k(z)^{pm-1} v_k(z)$ have the highest degree among terms in the numerator and denominator, respectively. Hence, f_{k+1} has type (m_{k+1}, ℓ_{k+1}), where

$$\begin{aligned} m_{k+1} &= pmm_k, & m_1 &= m, \\ \ell_{k+1} &= (pm-1)m_k + \ell_k, & \ell_1 &= \ell. \end{aligned}$$

Solving this recursion gives

$$\begin{aligned} m_k &= \frac{1}{p}(pm)^k, \\ \ell_k &= \frac{1}{p}(pm)^k - (m-\ell). \end{aligned}$$

The case in which $\ell \geq m$ is similar. This time we write

$$\begin{aligned} &f_{k+1}(z) \\ &= \frac{a_m z^m u_k(z)^{p(\ell-m)+1} v_k(z)^{pm} + a_{m-1} z^{m-1} u_k(z)^{p(\ell-m+1)+1} v_k(z)^{p(m-1)} + \cdots + a_0 u_k(z)^{p\ell+1}}{b_\ell z^\ell v_k(z)^{p\ell+1} + b_{\ell-1} z^{\ell-1} u_k(z)^p v_k(z)^{p(\ell-1)+1} + \cdots + b_0 u_k(z)^{p\ell} v_k(z)}, \end{aligned}$$

leading to the recursion

$$\begin{aligned} m_{k+1} &= m + (p(\ell-m)+1)m_k + pm\ell_k, & m_1 &= m, \\ \ell_{k+1} &= \ell + (p\ell+1)\ell_k, & \ell_1 &= \ell, \end{aligned}$$

with solution

$$\begin{aligned} m_k &= \frac{1}{p}\left[(p\ell+1)^k - (p(\ell-m)+1)^k\right], \\ \ell_k &= \frac{1}{p}\left[(p\ell+1)^k - 1\right]. \end{aligned}$$

The asymptotic relation (19) follows easily.

4.4 Proof of Corollaries 1 and 2

To prove Corollaries 1 and 2, let $e_k(z) = \frac{\widetilde{f}_k(z) - z^{1/p}}{z^{1/p}}$. Since $X_k = f_k(A) = \frac{1+\alpha_k}{2\alpha_k}\widetilde{f}_k(A)$, $Y_k = X_k^{1-p}A$, and $Z_k = X_k^{-1}$ in (6), (8), and (9), we have

$$\begin{aligned} \widetilde{X}_k A^{-1/p} - I &= e_k(A), \\ \widetilde{Y}_k A^{-1/p} - I &= \widetilde{X}_k^{-(p-1)} A^{(p-1)/p} - I \\ &= (I + e_k(A))^{-(p-1)} \left(I - (I + e_k(A))^{p-1}\right), \end{aligned}$$

and

$$\begin{aligned} \widetilde{Z}_k A^{1/p} - I &= \widetilde{X}_k^{-1} A^{1/p} - I \\ &= -(I + e_k(A))^{-1} e_k(A). \end{aligned}$$

The results follow from the above equalities and the bounds

$$\|e_k(A)\|_2 \le \max_{\alpha^p \le z \le 1} |e_k(z)| = \varepsilon_k,$$
$$\|(I + e_k(A))^{-1}\|_2 \le \frac{1}{1 - \|e_k(A)\|_2} \le \frac{1}{1 - \varepsilon_k},$$

and

$$\begin{aligned}\|I - (I + e_k(A))^{p-1}\|_2 &= \left\| -\sum_{j=1}^{p-1} \binom{p-1}{j} e_k(A)^j \right\|_2 \\ &\le \sum_{j=1}^{p-1} \binom{p-1}{j} \varepsilon_k^j \\ &= (1 + \varepsilon_k)^{p-1} - 1.\end{aligned}$$

4.5 Proof of Proposition 2

To prove the formula (21) for $\hat{r}_{1,0}(z, \alpha, \sqrt[p]{\cdot})$, it suffices to show that the function

$$\hat{e}(z) := \frac{\hat{r}_{1,0}(z, \alpha, \sqrt[p]{\cdot}) - z^{1/p}}{z^{1/p}}$$

achieves its global maximum on $[\alpha^p, 1]$ at both endpoints and has global minimum 0 on $[\alpha^p, 1]$. Indeed, if this is the case, then the rescaled function

$$\frac{2}{2 + \hat{e}(1)} \hat{r}_{1,0}(z, \alpha, \sqrt[p]{\cdot})$$

has relative error which equioscillates three times on $[\alpha^p, 1]$, and so must be the minimizer for $E_{1,0}(\sqrt[p]{\cdot}, [\alpha^p, 1])$. A calculation verifies that $\hat{e}(z)$ has a critical point at $z = \mu^p$, $\hat{e}(\mu^p) = 0$, $\hat{e}(\alpha^p) = \hat{e}(1)$, $\hat{e}(z)$ is decreasing on (α^p, μ^p), and $\hat{e}(z)$ is increasing on $(\mu^p, 1)$.

The proof of (22) is similar. In this case, a calculation verifies that the function

$$\hat{e}(z) := \frac{\hat{r}_{0,1}(z, \alpha, \sqrt[p]{\cdot}) - z^{1/p}}{z^{1/p}}$$

has a critical point at $z = 1/\nu^p$, $\hat{e}(1/\nu^p) = 0$, $\hat{e}(\alpha^p) = \hat{e}(1)$, $\hat{e}(z)$ is decreasing on $(\alpha^p, 1/\nu^p)$, and $\hat{e}(z)$ is increasing on $(1/\nu^p, 1)$.

4.6 Proof of Proposition 3

Trefethen and Gutknecht [34, Theorem 3b] have shown that for any function f analytic in a neighborhood of 1, $\operatorname{argmin}_{r \in \mathcal{R}_{m,\ell}} \max_{z \in [1-\delta, 1]} |r(z) - f(z)|$ converges coefficien-

twise as $\delta \to 0$ to the type-(m, ℓ) Padé approximant of f about $z = 1$, provided that the Padé approximant has defect 0 in $\mathcal{R}_{m,\ell}$. Their proof carries over easily to minimizers of the relative error $|(r(z) - f(z))/f(z)|$, assuming $f(1) \neq 0$. Since $P_{m,\ell,p}(z)$ has defect 0 in $\mathcal{R}_{m,\ell}$ [10], Proposition 3 follows. The explicit formula (25) for $P_{m,\ell,p}(z)$ is from [24, p. 954].

4.7 Proof of Proposition 4

Since $Q_{\ell,m,p}(z)^{-1} = P_{m,\ell,p}(z)$ is a Padé approximant of $f(z) = z^{1/p}$ about $z = 1$ of type $(m, \ell) \neq (0, 0)$, we have $Q_{\ell,m,p}(1) = 1$ and

$$-Q'_{\ell,m,p}(1) = \frac{-Q'_{\ell,m,p}(1)}{Q_{\ell,m,p}(1)^2} = P'_{m,\ell,p}(1) = f'(1) = \frac{1}{p}.$$

Hence, $Q_{\ell,m,p}(I) = I$, $L_{Q_{\ell,m,p}}(I, E) = -\frac{1}{p}E$, and $L_{Q^{p-1}_{\ell,m,p}}(I, E) = -\frac{p-1}{p}E$ for any $E \in \mathbb{C}^{n\times n}$. Thus, with $g(Y, Z) = (YQ_{\ell,m,p}(ZY)^{p-1}, Q_{\ell,m,p}(ZY)Z)$, we obtain

$$\begin{aligned} L_g(B, B^{-1}; E, F) &= \left(E - B\left(\frac{p-1}{p}\right)(FB + B^{-1}E),\ F - \frac{1}{p}(FB + B^{-1}E)B^{-1}\right) \\ &= \frac{1}{p}\left(E - (p-1)BFB,\ (p-1)F - B^{-1}EB^{-1}\right). \end{aligned}$$

Setting $\widetilde{E} = \frac{1}{p}(E - (p-1)BFB)$ and $\widetilde{F} = \frac{1}{p}((p-1)F - B^{-1}EB^{-1})$, we find that $L_g(B, B^{-1}; \widetilde{E}, \widetilde{F}) = L_g(B, B^{-1}; E, F)$, so $L_g(B, B^{-1}; \cdot, \cdot)$ is idempotent.

5 Numerical Examples

In this section, we present numerical examples and discuss the implementation of the rational minimax iteration (8–10).

5.1 Implementation

Implementing the rational minimax iteration (8–10) requires evaluating the rational function $h_{\ell,m,p}(z, \alpha_k) = \hat{r}_{m,\ell}(z, \alpha_k, \sqrt[p]{\cdot})^{-1}$ at a matrix argument Z_kY_k. With the exception of the special cases detailed in Sect. 3.3, explicit formulas for this function are not available. Nevertheless, $\hat{r}_{m,\ell}(z, \alpha_k, \sqrt[p]{\cdot})$ (or, more precisely, its unscaled counterpart $r_{m,\ell}(z, \alpha_k, \sqrt[p]{\cdot})$) can be determined numerically using, for instance, the function `MiniMaxApproximation` from Mathematica's `FunctionApproximations` package. This function uses the Remez exchange algorithm to determine rational minimax approximants on real intervals. We used this function along with `Apart` to compute $h_{\ell,m,p}(z, \alpha_k)$ in partial fraction form. For α_k close to 1, the computation of $h_{\ell,m,p}(z, \alpha_k)$ poses numerical difficulties, so we rounded α_k to 1 (thereby reverting to the Padé iteration (27–28)) whenever $\alpha_k > 0.99$. We also observed that for α_k close to

0 and $\ell = m$, accuracy improved if $r_{m,m}(z, \alpha_k, \sqrt[p]{\cdot})$ was computed as $R(1/z)$, where $R = \text{argmin}_{r \in \mathcal{R}_{m,m}} \max_{1 \le z \le \alpha_k^{-p}} |(r(z) - z^{-1/p})/z^{-1/p}|$. The time taken to determine $r_{m,\ell}(z, \alpha_k, \sqrt[p]{\cdot})$ with `MiniMaxApproximation` ranged from about 0.01 seconds (for $(m, \ell) = (1, 1)$ and α far from 0) to about 1 second (for $(m, \ell) = (8, 8)$ and α close to 0), with little dependence on p.

Note that a more robust option for computing minimizers of the maximum *absolute* error $|r(z) - f(z)|$ is the Chebfun function `minimax` [6]. However, Chebfun currently does not support minimization of the maximum *relative* error $|(r(z) - f(z))/f(z)|$.

Algorithm 1 summarizes the implementation of the rational minimax iteration (8–10). For simplicity, it focuses on the type-(m, m) iteration. The type-(m, ℓ) iteration with $\ell \ne m$ is similar, but the form of the partial fraction expansion of $h_{\ell,m,p}(z, \alpha)$ varies with ℓ. In the algorithm, the eigenvalues of A with the smallest and largest magnitudes are denoted $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$, respectively.

Algorithm 1 Type-(m, m) rational minimax iteration for the matrix pth root

$\tau = |\lambda_{\max}(A)|$
$\alpha_0 = |\lambda_{\min}(A)/\lambda_{\max}(A)|^{1/p}$
$Y_0 = A/\tau$
$Z_0 = I$
$k = 0$
while not converged **do**
  if $\alpha_k > 0.99$ **then** $\alpha_k = 1$ **end if**
  Compute $h_{m,m,p}(z, \alpha_k)$ and its partial fraction expansion

$$h_{m,m,p}(z, \alpha_k) = a_0 + \sum_{j=1}^{m} \frac{a_j}{z + b_j}.$$

  $W = \sum_{j=1}^{m} a_j (Z_k Y_k + b_j I)^{-1}$
  $Y_{k+1} = Y_k (a_0 I + W)^{p-1}$
  $Z_{k+1} = a_0 Z_k + W Z_k$
  $\alpha_{k+1} = \alpha_k h_{m,m,p}(\alpha_k^p, \alpha_k)$
  $k = k + 1$
end while
$\widetilde{Y}_k = \tau^{1/p} (1 + \alpha_k)^{p-1} Y_k / (2\alpha_k)^{p-1}$
$\widetilde{Z}_k = \tau^{-1/p} (1 + \alpha_k) Z_k / (2\alpha_k)$
return $\widetilde{Y}_k \approx A^{1/p}$, $\widetilde{Z}_k \approx A^{-1/p}$

The choices of α_0 and τ used in the algorithm are motivated by Corollary 2: they ensure that the spectrum of A/τ is contained in the annulus $\{z \in \mathbb{C} \mid \alpha_0^p \le |z| \le 1\}$. In particular, if A is Hermitian positive definite, then the spectrum of A/τ is contained in $[\alpha_0^p, 1]$, and Corollary 2 is directly applicable. Neither $\lambda_{\min}(A)$ nor $\lambda_{\max}(A)$ need to be computed accurately; our experience suggests that estimates can be used without significantly degrading the algorithm's performance.

As a termination criterion, we terminated the iterations when

$$\|\widetilde{Z}_{k-1} \widetilde{Y}_{k-1} - I\|_\infty \le p \left(\frac{\Delta}{(p-1) C(m, \ell, p)} \right)^{1/(m+\ell+1)},$$

where $\Delta = 10^{-15}$ is a relative error tolerance. This is a generalization to arbitrary p of the termination criterion described in [7, Section 4.3].

Floating Point Operations If A is $n \times n$ and $(a_0 I + W)^{p-1}$ is computed with binary powering in Line 10 of Algorithm 1, then the cost of each iteration in Algorithm 1 is about $(6 + 2m + \beta \log_2(p-1))n^3$ flops, where $\beta \in [1, 2]$ [16, p. 72]. In the first iteration, the cost reduces to $(2 + 2m + \beta \log_2(p-1))n^3$ flops since $Z_0 = I$. If parallelism is exploited, then the m matrix inversions in Line 9 can be performed simultaneously, as can Lines 10–11. The effective cost (i.e., the span/depth) of such a parallel implementation is $(4 + \beta \log_2(p-1))n^3$ flops in the first iteration and $(6 + \beta \log_2(p-1))n^3$ flops in each remaining iteration. Further savings in computational costs can be achieved when $p = 2$; see [7, Section 4.2] for details.

In the vast majority of our numerical experiments with, for instance, the type-$(8, 8)$ minimax iteration, convergence was achieved in two iterations (see Table 3), yielding an effective parallel cost of $(10 + 2\beta \log_2(p-1))n^3$ flops. For small to moderate p, this cost compares favorably against Schur-based algorithms for the matrix pth root, which are not easy to parallelize and typically cost at least $28n^3$ flops [14,17,18,30].

5.2 Scalar Iteration

Asymptotic Convergence Rates To verify the asymptotic convergence rates predicted by Theorem 2, we computed $\varepsilon_k = \frac{1-\alpha_k}{1+\alpha_k}$, $k = 1, 2, 3$, for various choices of m, ℓ, p, and ε_0. Table 1 reports the results for three such choices. (We selected values of m, ℓ, p, and ε_0 so that the asymptotic regime was reached before convergence to machine precision occurred.) The table demonstrates that the ratios $\varepsilon_k / \varepsilon_{k-1}^{m+\ell+1}$ approach the constant $C(m, \ell, p)$ given by (18). Note that the entry in the row $k = 3$ of the last column has been omitted, since ε_3 was below machine precision in that instance.

Complex Inputs To study the behavior of the rational function $\widetilde{f}_k(z)$ generated by the type-(m, ℓ) iteration (11–12), we numerically computed the sets

$$\mathcal{S}(k) = \mathcal{S}(k; \delta, \alpha, m, \ell, p) = \left\{ z \in \mathbb{C} : \left| \frac{\widetilde{f}_k(z) - z^{1/p}}{z^{1/p}} \right| \leq \delta \right\}$$

for various choices of δ, α, m, ℓ, and p. The boundaries of these sets are plotted in Fig. 3. They are plotted in the $(\log_{10}|z|, \arg z)$ coordinate plane rather than the usual $(\mathrm{Re}\, z, \mathrm{Im}\, z)$ coordinate plane to facilitate viewing. The shaded regions in the plots correspond to points $z \in \mathbb{C}$ for which $\lim_{k\to\infty} \widetilde{f}_k(z) \neq z^{1/p}$. Numerical evidence indicates that at these points, $\lim_{k\to\infty} \widetilde{f}_k(z) \in \{e^{2\pi i j/p} z^{1/p} \mid j \in \{1, 2, \ldots, p-1\}\}$. Furthermore, the shaded regions have a fractal structure. Both of these phenomena are typical features of iterations for the pth root when $p > 2$ [5].

Figure 3 gives valuable insight into the behavior of the matrix iteration (6–7) (and, of course, its coupled counterpart (8–10)). Indeed, if A is a normal matrix with eigenvalues in $\mathcal{S}(k)$, then the iteration (6–7) converges in at most k iterations with a relative tolerance δ in the 2-norm. As an example, the plot in row 3, column 2 of Fig. 3

Table 1 Values of $\{\varepsilon_k\}_{k=1}^{3}$ generated by the iteration (31) with $f(z) = z^{1/p}$ for various choices of m, ℓ, p, and ε_0. In each instance, the ratios $\varepsilon_k/\varepsilon_{k-1}^{m+\ell+1}$ approach the constant $C(m, \ell, p)$ given by (18), whose value is recorded in the last row of the table for reference

	$(m, \ell, p) = (1, 1, 13)$		$(m, \ell, p) = (2, 2, 3)$		$(m, \ell, p) = (3, 3, 5)$	
k	ε_k	$\varepsilon_k/\varepsilon_{k-1}^{m+\ell+1}$	ε_k	$\varepsilon_k/\varepsilon_{k-1}^{m+\ell+1}$	ε_k	$\varepsilon_k/\varepsilon_{k-1}^{m+\ell+1}$
0	$5.0000 \cdot 10^{-1}$		$9.9999 \cdot 10^{-1}$		$9.0000 \cdot 10^{-1}$	
1	$1.4864 \cdot 10^{-1}$	$1.19 \cdot 10^{0}$	$7.8215 \cdot 10^{-1}$	$7.82 \cdot 10^{-1}$	$4.2647 \cdot 10^{-2}$	$8.92 \cdot 10^{-2}$
2	$9.5361 \cdot 10^{-3}$	$2.90 \cdot 10^{0}$	$1.4269 \cdot 10^{-2}$	$4.87 \cdot 10^{-2}$	$2.1116 \cdot 10^{-11}$	$8.23 \cdot 10^{-2}$
3	$3.0325 \cdot 10^{-6}$	$3.50 \cdot 10^{0}$	$1.4379 \cdot 10^{-11}$	$2.43 \cdot 10^{-2}$		
		$3.50 \cdot 10^{0}$		$2.43 \cdot 10^{-2}$		$8.25 \cdot 10^{-2}$

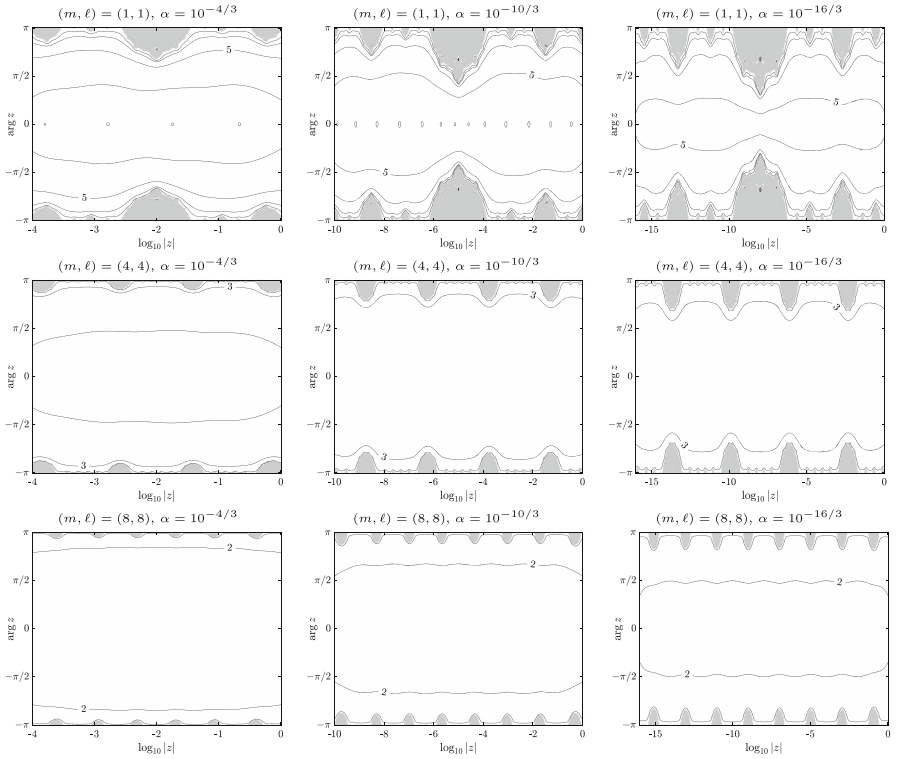

Fig. 3 Boundaries of the sets $\mathcal{S}(k;\delta,\alpha,m,\ell,p)$ with $\delta = 10^{-14}$, $p = 3$, $(m,\ell) = (1,1)$ (first row), $(m,\ell) = (4,4)$ (second row), $(m,\ell) = (8,8)$ (third row), $\alpha = 10^{-4/3}$ (first column), $\alpha = 10^{-10/3}$ (second column), and $\alpha = 10^{-16/3}$ (third column). In each plot, one of the boundaries has been selected arbitrarily and labelled with its index k. Each unlabelled boundary has an index which differs by $+1$ from that of its nearest inner neighbor. Shaded regions correspond to points z for which $\lim_{k\to\infty} \widetilde{f}_k(z) \neq z^{1/p}$

demonstrates that $\mathcal{S}(2)$ contains the set

$$\{z \in \mathbb{C} \mid \log_{10}|z| \in [-10, 0],\ \arg z \in [-\pi/2, \pi/2]\}$$

when $(m,\ell) = (8,8)$, $p = 3$, and $\alpha = 10^{-10/3}$. It follows that the type-$(8,8)$ iteration (6–7) converges to $A^{1/3}$ in at most 2 iterations for any normal matrix A with spectrum in the right half plane and $|\lambda_{\max}(A)/\lambda_{\min}(A)| \leq 10^{10}$.

For comparison, Fig. 4 shows the boundaries of the sets

$$\mathcal{T}(k) = \mathcal{T}(k;\delta,\alpha,m,\ell,p) = \left\{ z \in \mathbb{C}:\ \left| \frac{\widetilde{f}_k(z/\alpha^{p/2}) - (z/\alpha^{p/2})^{1/p}}{(z/\alpha^{p/2})^{1/p}} \right| \leq \delta \right\},$$

where this time $\widetilde{f}_k(z)$ is the rational function generated by (11–12) with the initial condition $\alpha_0 = \alpha$ replaced by $\alpha_0 = 1$. By Proposition 3, the sets $\mathcal{T}(k)$ characterize the convergence behavior of the Padé iteration (26) (and its coupled counterpart (27–28))

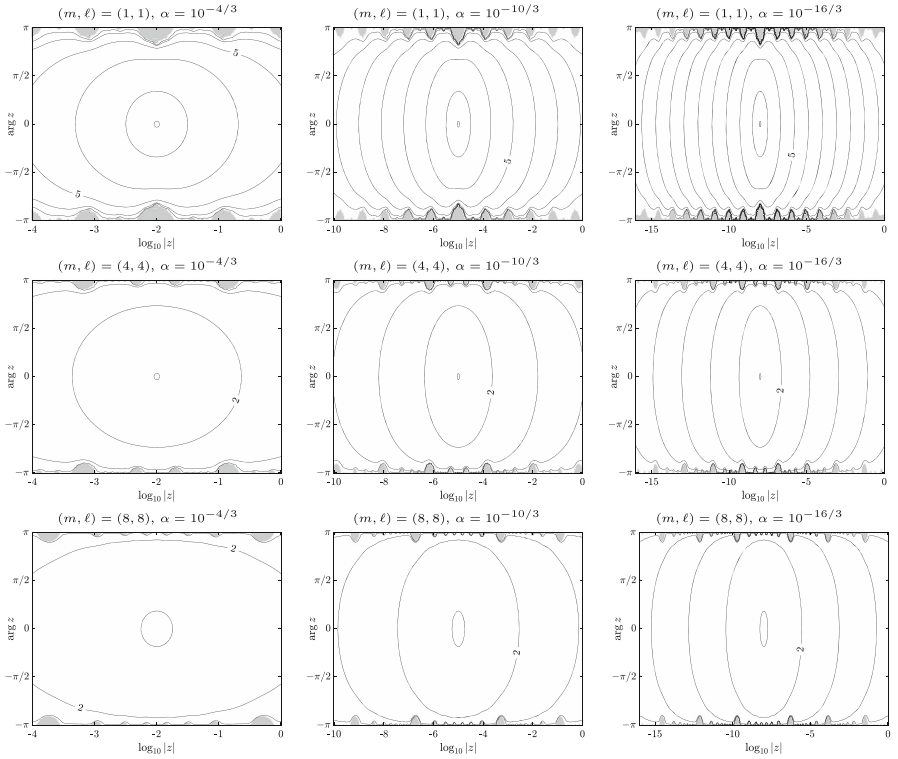

Fig. 4 Boundaries of the sets $\mathcal{T}(k; \delta, \alpha, m, \ell, p)$ with the same parameters as in Fig. 3

with the initial iterate scaled by $1/\alpha^{p/2}$. (Scaling by $1/\alpha^{p/2}$ facilitates the comparison with Fig. 3 by centering the Padé contours around $z = \alpha^{p/2}$.)

Notice that for small α (the two rightmost columns of Fig. 4), the sets $\mathcal{T}(k)$ do not contain scalars with extreme magnitudes ($|z| = \alpha^p$ and $|z| = 1$) unless k is relatively large. Comparing, for instance, the bottom right plots in Figs. 3 and 4, we see that if A is Hermitian positive definite with spectrum in $[10^{-16}, 1]$, then the type-(8, 8) rational minimax iteration (11–12) converges in at most 2 iterations, whereas the type-(8, 8) Padé iteration (25) converges in at most 5. The same observation holds, in fact, for the type-(6, 6) and type-(7, 7) iterations, which are not shown in Figs. 3 and 4. This is entirely analagous to the behavior observed in the case $p = 2$ in [7, Section 5.1]. In fact, with the exception of the low-order iterations, Figs. 3 and 4 bear a rather strong resemblance to Figs. 1-2 of [7].

It is worth noting that for the low-order iterations, the sets $\{z \in \mathbb{C} \mid \lim_{k\to\infty} \widetilde{f}_k(z) \neq z^{1/p}\}$ occupy more of the complex plane when $\widetilde{f}_k(z)$ is generated from the rational minimax iteration than when $\widetilde{f}_k(z)$ is generated from the Padé iteration (see the shaded regions in row 1 of Figs. 3 and 4). This appears to be a drawback of the low-order rational minimax iterations. The moderate-order and high-order iterations do not suffer as much from this issue; compare the shaded regions in the bottom two rows of Figs. 3 and 4, which occupy only a small neighborhood of the nonpositive real axis ($|\arg z| =$

π). The latter observation suggests that for moderate to high m and ℓ, it is safe to apply Algorithm 1 to matrices with spectrum contained in $\{z \in \mathbb{C}\colon |\arg z| \le \Theta\}$, where $\Theta < \pi$ is close to π. For matrices with eigenvalues that lie very near but not on the nonpositive real axis, a simple workaround is to compute $A^{1/2}$ using any algorithm for the matrix square root, and then compute $((A^{1/2})^{1/p})^2$. One can also compute $((A^{1/2^s})^{1/p})^{2^s}$ with $s > 1$, as in [14,17], but the advantages of minimax approximation over Padé approximation become less pronounced as s increases, since $A^{1/2^s}$ has eigenvalues clustered near 1 for large s.

Dependence on p Next, we studied the dependence of the iteration (11–12) on p. We fixed $z = \frac{1}{2}$ and $(m, \ell) = (1, 0)$, and, for various choices of p and α, we numerically determined the smallest integer k for which $|\widetilde{f}_k(z) - z^{1/p}|/|z^{1/p}| \le 10^{-14}$. The results for $2 \le p \le 10{,}000$ and $\alpha^p \in \{10^{-4}, 10^{-10}, 10^{-16}\}$ are shown in Table 2. The table indicates that the iteration count k grows with p, but does so rather slowly unless both p and α^p are small. For the higher-order iterations ($m, \ell \ge 1$), we detected little to no dependence of the iteration count on p. For instance, the iteration counts for $(m, \ell) = (1, 1)$ (not shown) were constant for $2 \le p \le 10{,}000$ (4 iterations when $\alpha^p = 10^{-4}$ and 5 iterations when $\alpha^p \in \{10^{-10}, 10^{-16}\}$).

5.3 Matrix Iteration

To test Algorithm 1, we applied it to a collection of matrices of size 10×10 from the Matrix Computation Toolbox [15]. We selected those 10×10 matrices in the toolbox with condition number $\le u^{-1}$ (where $u = 2^{-53}$ denotes the unit roundoff) and with spectrum contained in the sector $\{z \in \mathbb{C}\colon |\arg z| < 0.9\pi\}$. We also included those matrices whose spectrum could be rotated into the aforementioned sector by multiplying A by a suitable scalar $e^{i\theta}$, $\theta \in [0, 2\pi]$. A total of 41 matrices met these criteria. We carried out these tests in MATLAB, using a Wolfram Language script to call Mathematica's `MiniMaxApproximation` function in Line 8 of Algorithm 1.

Figure 5 plots the relative error $\|\widehat{X} - A^{1/p}\|_F / \|A^{1/p}\|_F$ in the computed pth root $\widehat{X}$ of A for each of the 41 matrices, where $p = 3$. The tests are sorted in order of decreasing $\kappa^{(p)}(A)$, where

Table 2 Smallest k for which $|\widetilde{f}_k(z) - z^{1/p}|/|z^{1/p}| \le 10^{-14}$. Here, $z = \frac{1}{2}$, $(m, \ell) = (1, 0)$, and results are reported for various choices of p and α

p	2	3	4	5	6	7	8	9	10	100	1000	10,000
$\alpha^p = 10^{-4}$	6	6	7	7	7	7	7	7	7	7	7	7
$\alpha^p = 10^{-10}$	7	8	9	9	9	10	10	10	10	11	11	11
$\alpha^p = 10^{-16}$	8	9	10	10	11	11	12	12	12	14	14	14

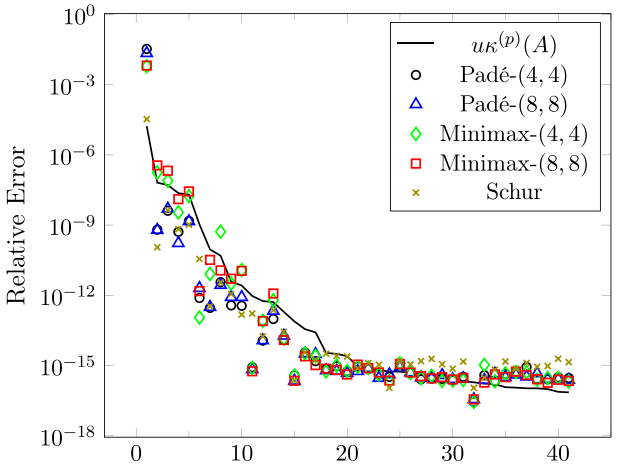

Fig. 5 Relative errors committed by the Padé iterations of type (4, 4) and (8, 8), the minimax iterations of type (4, 4) and (8, 8), and the Schur method [30]. Results are shown for 41 tests with $p = 3$, ordered by decreasing condition number $\kappa^{(p)}(A)$

Table 3 Number of iterations used by each iterative method in the tests appearing in Fig. 5

Iterations	1	2	3	4	5	≥ 6
Padé-(4, 4)	0	17	12	6	4	2
Padé-(8, 8)	0	27	7	6	1	0
Minimax-(4, 4)	0	17	20	2	1	1
Minimax-(8, 8)	0	34	6	1	0	0

$$\kappa^{(p)}(A) = \frac{\|A\|_F}{\|X\|_F} \left\| \left(\sum_{j=1}^{p} (X^{p-j})^T \otimes X^{j-1} \right)^{-1} \right\|_2$$

denotes the Frobenius-norm relative condition number of the matrix pth root X of A [16, Problem 7.4]. Results for five methods are shown: the rational minimax iterations (8–10) of type (4, 4) and (8, 8), the Padé iterations (27–28) of type (4, 4) and (8, 8), and Smith's Schur method for the matrix pth root [30]. The Padé iterations were implemented using Algorithm 1 with Lines 1–2 replaced by $\tau = 1/\sqrt{|\lambda_{\min}(A)\lambda_{\max}(A)|}$ and $\alpha_0 = 1$. The results indicate that the algorithms under consideration behave in a forward stable way, with relative errors mostly lying within a small factor of $u\kappa^{(p)}(A)$.

In Table 3, the number of iterations used by each iterative method on the 41 tests are recorded. In analogy with the results of [7], the rational minimax iterations very often converged more quickly than the Padé iterations on these tests.

6 Conclusion

This paper has constructed and analyzed a family of iterations for computing the matrix pth root using rational minimax approximants of the function $z^{1/p}$. The output of each step k of the type-(m, ℓ) iteration is a rational function r of A with the property that the scalar function $e(z) = (r(z) - z^{1/p})/z^{1/p}$ equioscillates $(m + \ell + 1)^k + 1$ times on $[\alpha^p, 1]$, where $\alpha \in (0, 1)$ is a parameter depending on A. With the exception of the Zolotarev iterations (i.e. $p = 2$ and $\ell \in \{m-1, m\}$), this equioscillatory behavior does not render $\max_{\alpha^p \le z \le 1} |e(z)|$ minimal among all choices of r with the same numerator and denominator degree. Nevertheless, we have shown that many of the desirable features of the Zolotarev iterations carry over to the general setting. A key role in the analysis was played by the asymptotic behavior of rational minimax approximants on short intervals.

Several topics mentioned in this paper are worth pursuing in more detail. Remark 1 leads naturally to a family of rational minimax iterations for the matrix sector function $\mathrm{sect}_p(A) = A(A^p)^{-1/p}$. As $\alpha \uparrow 1$, these iterations likely reduce to the Padé iterations for the sector function studied by Laszkiewicz and Ziętak [24, Section 5], so the results therein could inform an analysis of the convergence of the rational minimax iterations on matrices that are non-normal and/or have spectrum away from the positive real axis. Another topic of interest is computing the action of $A^{1/p}$ on a vector b using rational minimax iterations. Li and Yang [25] address a similar task: computing the action of a spectral filter on b using Zolotarev iterations for $\mathrm{sign}(z)$. It may be possible to construct a similar algorithm for computing $A^{1/p}b$. Finally, the functional iteration (11–12) is of interest in its own right, as it offers a method of rapidly generating rational approximants of $z^{1/p}$ with small relative error, a tool that may have applications in, for instance, numerical conformal mapping [12] and approximation theory for compositions of rational functions [8].

Acknowledgements The author was supported in part by NSF Grant DMS-1703719.

References

1. Akhiezer, N.I.: Theory of Approximation. Frederick Ungar Publishing Corporation, New York (1956)
2. Beckermann, B.: Optimally Scaled Newton Iterations for the Matrix Square Root. Advances in Matrix Functions and Matrix Equations workshop, Manchester (2013)
3. Bini, D.A., Higham, N.J., Meini, B.: Algorithms for the matrix pth root. Numer. Algorithms **39**, 349–378 (2005)
4. Byers, R., Xu, H.: A new scaling for Newton's iteration for the polar decomposition and its backward stability. SIAM J. Matrix Anal. Appl. **30**, 822–843 (2008)
5. Cardoso, J.R., Loureiro, A.F.: Iteration functions for pth roots of complex numbers. Numer. Algorithms **57**, 329–356 (2011)
6. Driscoll, T.A., Hale, N., Trefethen, L.N.: Chebfun Guide. Pafnuty Publications, Oxford (2014)
7. Gawlik, E.S.: Zolotarev iterations for the matrix square root. SIAM J. Matrix Anal. Appl. **40**, 696–719 (2019)
8. Gawlik, E.S., Nakatsukasa, Y.: Approximating the pth Root by Composite Rational Functions. arXiv preprint arXiv:1906.11326 (2019)
9. Gawlik, E.S., Nakatsukasa, Y., Sutton, B.D.: A backward stable algorithm for computing the CS decomposition via the polar decomposition. SIAM J. Matrix Anal. Appl. **39**, 1448–1469 (2018)

10. Gomilko, O., Greco, F., Ziętak, K.: A Padé family of iterations for the matrix sign function and related problems. Numer. Linear Algebra Appl. **19**, 585–605 (2012)
11. Gomilko, O., Karp, D.B., Lin, M., Ziętak, K.: Regions of convergence of a Padé family of iterations for the matrix sector function and the matrix pth root. J. Comput. Appl. Math. **236**, 4410–4420 (2012)
12. Gopal, A., Trefethen, L.N.: Representation of conformal maps by rational functions. Numer. Math. **142**, 359–382 (2019)
13. Guo, C.-H.: On Newton's method and Halley's method for the principal pth root of a matrix. Linear Algebra Appl. **432**, 1905–1922 (2010)
14. Guo, C.-H., Higham, N.J.: A Schur–Newton method for the matrix pth root and its inverse. SIAM J. Matrix Anal. Appl. **28**, 788–804 (2006)
15. Higham, N.J.: The Matrix Computation Toolbox. http://www.ma.man.ac.uk/~higham/mctoolbox Accessed 6 Nov 2018
16. Higham, N.J.: Functions of Matrices: Theory and Computation. SIAM, Philadelphia (2008)
17. Higham, N.J., Lin, L.: A Schur–Padé algorithm for fractional powers of a matrix. SIAM J. Matrix Anal. Appl. **32**, 1056–1078 (2011)
18. Higham, N.J., Lin, L.: An improved Schur–Padé algorithm for fractional powers of a matrix and their Fréchet derivatives. SIAM J. Matrix Anal. Appl. **34**, 1341–1360 (2013)
19. Hoskins, W., Walton, D.: A faster, more stable method for computing the pth roots of positive definite matrices. Linear Algebra Appl. **26**, 139–163 (1979)
20. Iannazzo, B.: On the Newton method for the matrix pth root. SIAM J. Matrix Anal. Appl. **28**, 503–523 (2006)
21. Iannazzo, B.: A family of rational iterations and its application to the computation of the matrix pth root. SIAM J. Matrix Anal. Appl. **30**, 1445–1462 (2008)
22. Karlin, S., Studden, W.: Tchebycheff Systems: With Applications in Analysis and Statistics, Pure and Applied Mathematics. Interscience Publishers, New York (1966)
23. King, R.F.: Improved Newton iteration for integral roots. Math. Comput. **25**, 299–304 (1971)
24. Laszkiewicz, B., Ziętak, K.: A Padé family of iterations for the matrix sector function and the matrix pth root. Numer. Linear Algebra Appl. **16**, 951–970 (2009)
25. Li, Y., Yang, H.: Interior Eigensolver for Sparse Hermitian Definite Matrices Based on Zolotarev's Functions. arXiv preprint arXiv:1701.08935 (2017)
26. Maehly, H., Witzgall, C.: Tschebyscheff-approximationen in kleinen Intervallen II. Numer. Math. **2**, 293–307 (1960)
27. Meinardus, G., Taylor, G.: Optimal partitioning of Newton's method for calculating roots. Math. Comput. **35**, 1221–1230 (1980)
28. Nakatsukasa, Y., Bai, Z., Gygi, F.: Optimizing Halley's iteration for computing the matrix polar decomposition. SIAM J. Matrix Anal. Appl. **31**, 2700–2720 (2010)
29. Nakatsukasa, Y., Freund, R.W.: Computing fundamental matrix decompositions accurately via the matrix sign function in two iterations: the power of Zolotarev's functions. SIAM Rev. **58**, 461–493 (2016)
30. Smith, M.I.: A Schur algorithm for computing matrix pth roots. SIAM J. Matrix Anal. Appl. **24**, 971–989 (2003)
31. Stahl, H.R.: Best uniform rational approximation of x^{α} on [0, 1]. Acta Math. **190**, 241–306 (2003)
32. Trefethen, L.N.: Approximation Theory and Approximation Practice, vol. 128. SIAM, Philadelphia (2013)
33. Trefethen, L.N., Gutknecht, M.H.: The Carathéodory–Fejér method for real rational approximation. SIAM J. Numer. Anal. **20**, 420–436 (1983)
34. Trefethen, L.N., Gutknecht, M.H.: On convergence and degeneracy in rational Padé and Chebyshev approximation. SIAM J. Math. Anal. **16**, 198–210 (1985)
35. Trefethen, L.N., Gutknecht, M.H.: Padé, stable Padé, and Chebyshev-Padé approximation. In: Mason, J.C., Cox, M.G. (eds.) Algorithms for Approximation, pp. 227–264. Clarendon Press, Oxford (1987)
36. Zolotarev, E.I.: Applications of elliptic functions to problems of functions deviating least and most from zero. Zapiski Imperatorskoj Akademii Nauk po Fiziko-Matematiceskomu Otdeleniju **30**, 1–59 (1877)