SIAM J. MATRIX ANAL. APPL.
Vol. 40, No. 2, pp. 696–719

ZOLOTAREV ITERATIONS FOR THE MATRIX SQUARE ROOT*

EVAN S. GAWLIK†

Abstract. We construct a family of iterations for computing the principal square root of a square matrix A using Zolotarev's rational minimax approximants of the square root function. We show that these rational functions obey a recursion, allowing one to iteratively generate optimal rational approximants of $\sqrt{z}$ of high degree using compositions and products of low-degree rational functions. The corresponding iterations for the matrix square root converge to $A^{1/2}$ for any input matrix A having no nonpositive real eigenvalues. In special limiting cases, these iterations reduce to known iterations for the matrix square root: the lowest-order version is an optimally scaled Newton iteration, and for certain parameter choices, the principal family of Padé iterations is recovered. Theoretical results and numerical experiments indicate that the iterations perform especially well on matrices having eigenvalues with widely varying magnitudes.

Key words. matrix square root, rational approximation, Zolotarev, minimax, matrix iteration, Chebyshev approximation, Padé approximation, Newton iteration, Denman–Beavers iteration

AMS subject classifications. 65F30, 65F60, 41A20, 49K35

DOI. 10.1137/18M1178529

1. Introduction.

A well-known method for computing the square root of an $n \times n$ matrix A with no nonpositive real eigenvalues is the Newton iteration [14]

$$X_{k+1} = \frac{1}{2}(X_k + X_k^{-1}A), \quad X_0 = A. \tag{1}$$

In exact arithmetic, the matrix X_k converges quadratically to $A^{1/2}$, the principal square root of A [16, Theorem 6.9]. (In floating point arithmetic, mathematically equivalent reformulations of (1), such as the Denman–Beavers (DB) iteration [7], are preferred for stability reasons [16, section 6.4].)

If A is diagonalizable, then each eigenvalue $\lambda_k^{(i)}$ of X_k, $i = 1, 2, \ldots, n$, obeys a recursion of the form

$$\lambda_{k+1}^{(i)} = \frac{1}{2}\left(\lambda_k^{(i)} + \frac{\lambda_0^{(i)}}{\lambda_k^{(i)}}\right),$$

which is the Newton iteration for computing a root of $z^2 - \lambda_0^{(i)} = 0$. One can thus think of (1) as an iteration that, in the limit as $k \to \infty$, implicitly maps a collection of scalars $\lambda_0^{(i)}$, $i = 1, 2, \ldots, n$, to $\sqrt{\lambda_0^{(i)}}$ for each i. In order for each scalar to converge rapidly, it is necessary that the rational function $f_k(z)$ defined recursively by

$$f_{k+1}(z) = \frac{1}{2}\left(f_k(z) + \frac{z}{f_k(z)}\right), \quad f_0(z) = z, \tag{2}$$

converges rapidly to $f(z) = \sqrt{z}$ on the set $\bigcup_{i=1}^n \{\lambda_0^{(i)}\} \subset \mathbb{C}$.

*Received by the editors April 2, 2018; accepted for publication (in revised form) by A. Frommer April 16, 2019; published electronically June 4, 2019.

http://www.siam.org/journals/simax/40-2/M117852.html

Funding: The work of the author was partially supported by the National Science Foundation under grant DMS-1703719.

†Department of Mathematics, University of Hawaii at Manoa, Honolulu, HI 96822 (egawlik@hawaii.edu).

To generalize and improve the Newton iteration, it is natural to study other recursive constructions of rational functions, with the aim of approximating $\sqrt{z}$ on a subset $S \subset \mathbb{C}$ containing the spectrum of A. Of particular interest are rational functions that minimize the maximum relative error

$$\max_{z \in S} \left| (r(z) - \sqrt{z})/\sqrt{z} \right| \tag{3}$$

among all rational functions $r(z)$ of a given type (m, ℓ). By type (m, ℓ), we mean $r(z) = p(z)/q(z)$ is a ratio of polynomials p and q of degree at most m and ℓ, respectively. We denote the set of rational functions of type (m, ℓ) by $\mathcal{R}_{m,\ell}$.

On a positive real interval S, explicit formulas for the minimizers $r \in \mathcal{R}_{m,m-1}$ and $r \in \mathcal{R}_{m,m}$ of (3) are known for each m. The formulas, derived by Zolotarev [32], are summarized in section 3.1. We show in this paper that, remarkably, the minimizers obey a recursion analogous to (2). This fact is intimately connected to (and indeed follows from) an analogous recursion for rational minimax approximations of the function $\mathrm{sign}(z) = z/\sqrt{z^2}$ recently discovered by Nakatsukasa and Freund [24].

The lowest-order version of the recursion for square root approximants has been known for several decades [27, 25] [5, section V.5.C]. Beckermann [3] recently studied its application to matrices, and Wachspress [30] performed a similar study many years earlier, focusing on positive definite matrices [28, p. 219]. In this paper, we generalize these ideas by constructing a family of iterations for computing the matrix square root, one for each pair of integers (m, ℓ) with $\ell \in \{m-1, m\}$. We prove that these *Zolotarev iterations* are stable and globally convergent with (R-)order of convergence $m + \ell + 1$. By writing Zolotarev's rational functions in partial fraction form, the resulting algorithms are highly parallelizable. Numerical examples demonstrate that the iterations exhibit good forward stability.

The Zolotarev iterations for the matrix square root bear several similarities to the Padé iterations studied in [15, pp. 231–233], [17, section 6], and [16, section 6.7]. In fact, the Padé iterations can be viewed as a limiting case of the Zolotarev iterations; see Proposition 4. One of the messages we hope to convey in this paper is that the Zolotarev iterations are often preferable to the Padé iterations when the eigenvalues of A have widely varying magnitudes. Roughly, this can be understood by noting that the Padé approximants of $\sqrt{z}$ are designed to be good approximations of $\sqrt{z}$ near a point, whereas Zolotarev's minimax approximants are designed to be good approximations of $\sqrt{z}$ over an entire interval. For more details, particularly with regard to how these arguments carry over to the complex plane, see section 5.1.

This paper builds upon a stream of research that, in recent years, has sparked renewed interest in the applications of Zolotarev's work on rational approximation to numerical linear algebra. These applications include algorithms for the SVD, the symmetric eigendecomposition, and the polar decomposition [24]; algorithms for the CS decomposition [10]; bounds on the singular values of matrices with displacement structure [4]; computation of spectral projectors [20, 11, 22]; and the selection of optimal parameters for the alternating direction implicit method [31, 21]. Zolotarev's functions have even been used to compute the matrix square root [12]; however, there is an important distinction between that work and ours: In [12], Zolotarev's functions are not used as the basis of an iterative method. Rather, a rational function of A is evaluated once and for all to approximate $A^{1/2}$. As we argue below, recursive constructions of Zolotarev's functions offer significant advantages over this strategy. (Note, however, that if the goal is to compute a product $A^{1/2}b$ with b a vector, then recursive constructions lose their utility.)

This paper is organized as follows. In section 2, we state our main results without proof. In section 3, we prove these results. In section 4, we discuss the implementation of the Zolotarev iterations and how they compare with other known iterations. In section 5, we evaluate the performance of the Zolotarev iterations on numerical examples.

2. Statement of results. In this section, we state our main results and discuss some of their implications. Proofs are presented in section 3.

Recursion for rational approximations of $\sqrt{z}$. We begin by introducing a recursion satisfied by Zolotarev's best rational approximants of the square root function. For each $m, \ell \in \mathbb{N}_0$ and $\alpha \in (0,1)$, let $r_{m,\ell}(z,\alpha)$ denote the rational function of type (m,ℓ) that minimizes (3) on $S = [\alpha^2, 1]$. Let $\hat{r}_{m,\ell}(z,\alpha)$ be the unique scalar multiple of $r_{m,\ell}(z,\alpha)$ with the property that

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

The following theorem, which is closely related to [24, Corollary 4] and includes [3, Lemma 1] as a special case, will be proved in section 3.

THEOREM 1. *Let $m \in \mathbb{N}$ and $\alpha \in (0,1)$. Define $f_k(z)$ recursively by*

$$f_{k+1}(z) = f_k(z)\hat{r}_{m,m-1}\left(\frac{z}{f_k(z)^2}, \alpha_k\right), \qquad f_0(z) = 1, \tag{4}$$

$$\alpha_{k+1} = \frac{\alpha_k}{\hat{r}_{m,m-1}(\alpha_k^2, \alpha_k)}, \qquad \alpha_0 = \alpha. \tag{5}$$

Then, for every $k \geq 1$,

$$f_k(z) = \hat{r}_{p,p-1}(z,\alpha) = \frac{1+\alpha_k}{2\alpha_k} r_{p,p-1}(z,\alpha), \quad p = \frac{1}{2}(2m)^k. \tag{6}$$

If instead

$$f_{k+1}(z) = f_k(z)\hat{r}_{m,m}\left(\frac{z}{f_k(z)^2}, \alpha_k\right), \qquad f_0(z) = 1, \tag{7}$$

$$\alpha_{k+1} = \frac{\alpha_k}{\hat{r}_{m,m}(\alpha_k^2, \alpha_k)}, \qquad \alpha_0 = \alpha, \tag{8}$$

then, for every $k \geq 1$,

$$f_k(z) = \hat{r}_{p,p}(z,\alpha) = \frac{1+\alpha_k}{2\alpha_k} r_{p,p}(z,\alpha), \quad p = \frac{1}{2}((2m+1)^k - 1). \tag{9}$$

The remarkable nature of these recursions is worth emphasizing with an example. When $m = 7$, three iterations of (4)–(5) generate (up to rescaling) the best rational approximation of $\sqrt{z}$ of type $(1372, 1371)$ on the interval $[\alpha^2, 1]$. Not only is this an efficient way of computing $r_{1372,1371}(z,\alpha)$, but it also defies intuition that an iteration involving so few parameters could deliver the solution to an optimization problem (the minimization of (3) over $\mathcal{R}_{1372,1371}$) with thousands of degrees of freedom.

Zolotarev iterations for the matrix square root. Theorem 1 leads to a family of iterations for computing the square root of an $n \times n$ matrix A, namely,

$$X_{k+1} = X_k \hat{r}_{m,\ell}(X_k^{-2}A, \alpha_k), \qquad X_0 = I, \tag{10}$$

$$\alpha_{k+1} = \frac{\alpha_k}{\hat{r}_{m,\ell}(\alpha_k^2, \alpha_k)}, \qquad \alpha_0 = \alpha, \tag{11}$$

where m is a positive integer and $\ell \in \{m-1, m\}$. We will refer to each of these iterations as a *Zolotarev iteration* of type (m, ℓ). (Like the Newton iteration, these iterations are ill-suited for numerical implementation in their present form, but a reformulation renders them numerically stable; see the end of this section.) At first glance, these iterations would appear to be suitable only for Hermitian positive definite matrices (or, more generally, diagonalizable matrices with positive real eigenvalues) that have been scaled so that their eigenvalues lie in the interval $[\alpha^2, 1]$, but in fact they converge for any $A \in \mathbb{C}^{n\times n}$ with no nonpositive real eigenvalues. This is made precise in the forthcoming theorem, which is a generalization of [3, Theorem 4] and is related to [12, Theorem 4.1].

To state the theorem, we introduce some notation, following [3]. A compact set $S \subseteq \mathbb{C}$ is called L-spectral for $A \in \mathbb{C}^{n\times n}$ if

$$\|f(A)\|_2 \le L \sup_{z\in S} |f(z)|$$

for every function f analytic in S [18, Chapter 37]. For instance, the spectrum of A is 1-spectral for every normal matrix A, and the closure of the pseudospectrum $\Lambda_\epsilon(A) = \{z \in \mathbb{C} \mid \|(A - zI)^{-1}\|_2 > 1/\epsilon\}$ is C_ϵ-spectral with $C_\epsilon = \text{length}(\partial\Lambda_\epsilon(A))/(2\pi\epsilon)$ for every A [18, Fact 23.3.5].

For each $\alpha \in (0, 1)$, define

$$\varphi(z, \alpha) = \exp\left(\frac{\pi \text{sn}^{-1}(\sqrt{z}/\alpha; \alpha)}{K(\alpha')}\right), \tag{12}$$

where $\text{sn}(\cdot;\alpha)$, $\text{cn}(\cdot;\alpha)$, and $\text{dn}(\cdot;\alpha)$ denote Jacobi's elliptic functions with modulus α, $K(\alpha) = \int_0^{\pi/2}(1 - \alpha^2 \sin^2\theta)^{-1/2}\, d\theta$ is the complete elliptic integral of the first kind, and $\alpha' = \sqrt{1-\alpha^2}$ is the complementary modulus to α. Note that the function $\varphi(z, \alpha)$ supplies a conformal map from $\mathbb{C} \setminus ((-\infty, 0] \cup [\alpha^2, 1])$ to the annulus $\{z \in \mathbb{C} : 1 < |z| < \rho(\alpha)\}$ [2, pp. 138–140], where

$$\rho(\alpha) = \exp\left(\frac{\pi K(\alpha)}{K(\alpha')}\right). \tag{13}$$

THEOREM 2. *Let $A \in \mathbb{C}^{n\times n}$ have no nonpositive real eigenvalues. Suppose that $S \subseteq \mathbb{C} \setminus (-\infty, 0]$ is L-spectral for A. Let $m \in \mathbb{N}$, $\ell \in \{m-1, m\}$, $\alpha \in (0, 1)$, and $\gamma = \inf_{z\in S} |\varphi(z, \alpha)|$. For every $k \ge 1$ such that $2\gamma^{-(m+\ell+1)^k} < 1$, the matrix X_k defined by (10)–(11) satisfies*

$$\left\| \left(\frac{2\alpha_k}{1+\alpha_k}\right) X_k A^{-1/2} - I \right\|_2 \le \frac{4L\gamma^{-(m+\ell+1)^k}}{1 - 2\gamma^{-(m+\ell+1)^k}}. \tag{14}$$

If $S \subseteq [\alpha^2, 1]$, then the sharper estimate

$$\left\| \left(\frac{2\alpha_k}{1+\alpha_k}\right) X_k A^{-1/2} - I \right\|_2 \le 4L\rho(\alpha)^{-(m+\ell+1)^k}$$

holds for every $k \ge 1$.

COROLLARY 3. *Let $A \in \mathbb{C}^{n\times n}$ be Hermitian positive definite. If the eigenvalues of A lie in the interval $[\alpha^2, 1]$, then*

$$\left\| \left(\frac{2\alpha_k}{1+\alpha_k}\right) X_k A^{-1/2} - I \right\|_2 \le 4\rho(\alpha)^{-(m+\ell+1)^k}$$

for every $k \ge 1$.

Note that the error estimates above imply estimates for the relative error in the computed square root $\widetilde{X}_k := 2\alpha_k X_k/(1+\alpha_k)$, since

$$\frac{\|\widetilde{X}_k - A^{1/2}\|_2}{\|A^{1/2}\|_2} = \frac{\|(\widetilde{X}_k A^{-1/2} - I)A^{1/2}\|_2}{\|A^{1/2}\|_2} \leq \|\widetilde{X}_k A^{-1/2} - I\|_2.$$

Connections with existing iterations. It is instructive to examine the lowest-order realization of the iteration (10)–(11). When $(m,\ell) = (1,0)$, one checks (using either elementary calculations or the explicit formulas in section 3.1) that

$$\hat{r}_{1,0}(z,\alpha) = \frac{1}{2}(\alpha^{1/2} + \alpha^{-1/2}z),$$

so that the iteration (10)–(11) reduces to

$$\begin{aligned} X_{k+1} &= \frac{1}{2}(\alpha_k^{1/2} X_k + \alpha_k^{-1/2} X_k^{-1} A), & X_0 &= I, \\ \alpha_{k+1} &= \frac{2}{\alpha_k^{1/2} + \alpha_k^{-1/2}}, & \alpha_0 &= \alpha. \end{aligned}$$

Equivalently, in terms of $\mu_k := \alpha_k^{1/2}$,

$$X_{k+1} = \frac{1}{2}(\mu_k X_k + \mu_k^{-1} X_k^{-1} A), \qquad X_0 = I, \tag{15}$$

$$\mu_{k+1} = \sqrt{\frac{2}{\mu_k + \mu_k^{-1}}}, \qquad \mu_0 = \alpha^{1/2}. \tag{16}$$

This is precisely the scaled Newton iteration with a scaling heuristic studied in [3]. (In [3], starting values $X_0 = A$ and $\mu_0 = \alpha^{-1/2}$ are used, but it easy to check that this generates the same sequences $\{X_k\}_{k=1}^{\infty}$ and $\{\mu_k\}_{k=1}^{\infty}$ as (15)–(16).) This iteration has its roots in early work on rational approximation of the square root [27, 25], and it is closely linked to the scaled Newton iteration for the polar decomposition introduced in [6]. As with the unscaled Newton iteration, reformulating (15)–(16) (e.g., as a scaled DB iteration) is necessary to ensure its numerical stability.

Another class of known iterations for the matrix square root is recovered if one examines the limit as $\alpha \uparrow 1$. Below, we say that a family of functions $\{r(\cdot,\alpha) \in \mathcal{R}_{m,\ell} : \alpha \in (0,1)\}$ converges *coefficientwise* to a function $p \in \mathcal{R}_{m,\ell}$ as $\alpha \uparrow 1$ if the coefficients of the polynomials in the numerator and denominator of $r(z,\alpha)$, appropriately normalized, approach the corresponding coefficients in $p(z)$ as $\alpha \uparrow 1$.

PROPOSITION 4. *Let $m \in \mathbb{N}$ and $\ell \in \{m-1, m\}$. As $\alpha \uparrow 1$, $\hat{r}_{m,\ell}(z,\alpha)$ converges coefficientwise to $p_{m,\ell}(z)$, the type (m,ℓ) Padé approximant of $\sqrt{z}$ at $z = 1$.*

Since $p_{m,\ell}(1) = 1$, the iteration (10)–(11) formally reduces to

$$X_{k+1} = X_k p_{m,\ell}(X_k^{-2} A), \quad X_0 = I, \tag{17}$$

as $\alpha \uparrow 1$. To relate this to an existing iteration from the literature, define $Y_k = X_k^{-1} A$ and $Z_k = X_k^{-1}$. Then, using the mutual commutativity of X_k, Y_k, Z_k, and A, we arrive at the iteration

$$Y_{k+1} = Y_k q_{\ell,m}(Z_k Y_k), \qquad Y_0 = A, \tag{18}$$

$$Z_{k+1} = q_{\ell,m}(Z_k Y_k) Z_k, \qquad Z_0 = I, \tag{19}$$

where $q_{\ell,m}(z) = p_{m,\ell}(z)^{-1}$. Since $q_{\ell,m}(z)$ is the type (ℓ, m) Padé approximant of $z^{-1/2}$ at $z = 1$, this iteration is precisely the Padé iteration studied in [17, section 6], [15, p. 232], and [16, section 6.7]. There, it is shown that $Y_k \to A^{1/2}$ and $Z_k \to A^{-1/2}$ with order of convergence $m + \ell + 1$ for any A with no nonpositive real eigenvalues. Moreover, the iteration (18)–(19) is stable [16, Theorem 6.12] in the sense of [16, Definition 4.17].

Stable reformulation of the Zolotarev iterations. In view of the well-established stability theory for iterations of the form (18)–(19), we will focus in this paper on the following reformulation of the Zolotarev iteration (10)–(11):

$$Y_{k+1} = Y_k h_{\ell,m}(Z_k Y_k, \alpha_k), \qquad Y_0 = A, \tag{20}$$

$$Z_{k+1} = h_{\ell,m}(Z_k Y_k, \alpha_k) Z_k, \qquad Z_0 = I, \tag{21}$$

$$\alpha_{k+1} = \alpha_k h_{\ell,m}(\alpha_k^2, \alpha_k), \qquad \alpha_0 = \alpha, \tag{22}$$

where $h_{\ell,m}(z, \alpha) = \hat{r}_{m,\ell}(z, \alpha)^{-1}$ and $h_{\ell,m}(z, 1) = q_{\ell,m}(z)$. In exact arithmetic, Y_k and Z_k are related to X_k from (10)–(11) via $Y_k = X_k^{-1}A$, $Z_k = X_k^{-1}$. We remark that $h_{\ell,m}(z, \alpha)$ is, up to a rescaling, an optimal rational approximant of $1/\sqrt{z}$, in the sense that it minimizes the maximum relative error on $[\alpha^2, 1]$ among rational functions of type (ℓ, m). We elaborate on this point in section 3.1.

The following theorem will summarize the properties of the iteration (20)–(22). We first clarify our terminology. A sequence of scalars or matrices X_k is said to converge to X_* with Q-order of convergence p if $X_k \to X_*$ and there exist constants c and k_0 such that $\|X_{k+1} - X_*\| \le c\|X_k - X_*\|^p$ for every $k \ge k_0$. It is said to converge with R-order of convergence p if $\|X_k - X_*\| \le \varepsilon_k$ for some sequence of scalars ε_k that converge to zero with Q-order p [26, p. 620]. If X_k is generated from an iteration $X_{k+1} = f(X_k)$, then we say the iteration is stable if the Fréchet derivative of f at X_* has bounded powers [16, Definition 4.17]. That is, with $g(E) = L_f(X_*, E)$ denoting the Fréchet derivative of f at X_* in a direction E, there exists a constant c such that $\|g^j(E)\| \le c\|E\|$ for every j and every E. Note that in the coupled iteration (20)–(22), the map f under scrutiny is, strictly speaking, $f(Y, Z, \alpha) = (Y h_{\ell,m}(ZY, \alpha), h_{\ell,m}(ZY, \alpha)Z, \alpha h_{\ell,m}(\alpha^2, \alpha))$. However, we will argue in section 4 that it is numerically prudent to set α_k (and all subsequent iterates) equal to 1 once α_k exceeds a suitable threshold $1 - \epsilon$, $\epsilon \ll 1$. We will therefore say that (20)–(22) is stable if the Fréchet derivative of $f(Y, Z) = (Y h_{\ell,m}(ZY, 1), h_{\ell,m}(ZY, 1)Z)$ has bounded powers at $(Y, Z) = (A^{1/2}, A^{-1/2})$.

THEOREM 5. *Let $m \in \mathbb{N}$, $\ell \in \{m - 1, m\}$, and $\alpha \in (0, 1)$. For any $A \in \mathbb{C}^{n\times n}$ with no nonpositive real eigenvalues, the iteration* (20)–(22) *is stable, and $Y_k \to A^{1/2}$, $Z_k \to A^{-1/2}$, and $\alpha_k \to 1$ with R-order of convergence $m + \ell + 1$.*

Note that although Theorem 5 places no restrictions on the spectral radius of A nor the choice of $\alpha \in (0, 1)$, it should be clear that it is preferable to scale A so that its spectral radius is 1 (or approximately 1), and set $\alpha = \sqrt{|\lambda_{\min}/\lambda_{\max}|}$ (or an estimate thereof), where $\lambda_{\max}$ and $\lambda_{\min}$ are the eigenvalues of A with the largest and smallest magnitudes, respectively. See section 5.1 for more details.

3. Proofs. In this section, we present proofs of Theorem 1, Theorem 2, Proposition 4, and Theorem 5.

3.1. Background. We begin by reviewing a few facts from the theory of rational minimax approximation. For a thorough presentation of this material, see, for example, [1, Chapter II] and [2, Chapter 9].

Rational minimax approximation. Let $S = [a, b]$ be a finite interval. A continuous function $g(z)$ is said to *equioscillate* between N extreme points on S if there exist N points $z_1 < z_2 < \cdots < z_N$ in S at which

$$g(z_j) = \sigma(-1)^j \max_{z\in S} |g(z)|, \quad j = 1, 2, \ldots, N,$$

for some $\sigma \in \{-1, 1\}$.

Let f and w be continuous, real-valued functions on S with $w \neq 0$ on S. Consider the problem of finding a rational function $r \in \mathcal{R}_{p,q}$ that minimizes

$$\max_{z\in S} |(r(z) - f(z))w(z)|$$

among all rational functions of type (p, q). It is well-known that this problem admits a unique solution r^* [1, p. 55]. Furthermore, the following are sufficient conditions guaranteeing optimality: If $r \in \mathcal{R}_{p,q}$ has the property that $(r(z)-f(z))w(z)$ equioscillates between $p + q + 2$ extreme points on S, then $r = r^*$ [1, p. 55]. (If S is a union of two disjoint intervals and $w \equiv 1$, then this statement holds with $p + q + 2$ replaced by $p + q + 3$ [24, Lemma 2].) If $f \neq 0$ on S, then we denote

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

Rational approximation of the sign function. Our analysis will make use of a connection between rational minimax approximants of $\sqrt{z}$ and rational minimax approximants of the function $\operatorname{sign}(z) = z/\sqrt{z^2}$. For each $p \in \mathbb{N}$ and $0 < a < b < \infty$, let $s_p(z, [a, b])$ denote the unique minimizer for $E_{p,p}(\operatorname{sign}, [-b, -a] \cup [a, b]) =: e_p(a/b)$. Explicit formulas for $s_p(z, [a, b])$ are known thanks to the seminal work of Zolotarev [32]. They have the form

$$s_p(z, [a, b]) = s_p(z/b, [a/b, 1]) = zR_p(z^2, [a^2, b^2]), \tag{23}$$

where, for each $m \in \mathbb{N}_0$ and $\ell \in \{m - 1, m\} \cap \mathbb{N}_0$, $R_{m+\ell+1}(z, [a, b])$ is a rational function of type (ℓ, m).

On the interval $[a, b]$, the function $s_p(z, [a, b])$ achieves its extremal values $1 \pm e_p(a/b)$ exactly $p + 1$ times in an alternating fashion [1, p. 286], with

$$s_p(a, [a, b]) = 1 - e_p(a/b), \quad s_p(b, [a, b]) = 1 - (-1)^p e_p(a/b). \tag{24}$$

This confirms the optimality of $s_p(z, [a, b])$. To see this, write $p = m + \ell + 1$ with $m \in \mathbb{N}_0$ and $\ell \in \{m-1, m\}\cap\mathbb{N}_0$. The minimizer for $E_{p,p}(\operatorname{sign}, [-b, -a]\cup[a, b])$ must be an odd function of z and therefore must have type $(2\ell+1, 2m)$, so the equioscillation of $s_p(z, [a, b]) - \operatorname{sign}(z)$ between $2(p + 1) = (2\ell + 1) + 2m + 3$ extreme points on $[-b, -a] \cup [a, b]$ renders $s_p(z, [a, b])$ optimal.

The notation used above is intentionally suggestive: The function

$$R_{m+\ell+1}(z, [a, b]) = \frac{s_{m+\ell+1}(\sqrt{z}, [\sqrt{a}, \sqrt{b}])}{\sqrt{z}}$$

is itself extremal for $E_{\ell,m}(1/\sqrt{z}, [a, b])$. Indeed, as z runs from a to b,

$$\frac{R_{m+\ell+1}(z, [a, b]) - 1/\sqrt{z}}{1/\sqrt{z}} = s_{m+\ell+1}(\sqrt{z}, [\sqrt{a}, \sqrt{b}]) - 1$$

equioscillates $m+\ell+2$ times with extrema $\pm e_{m+\ell+1}(\sqrt{a/b})$. In particular,

$$E_{\ell,m}(1/\sqrt{z},[a,b]) = e_{m+\ell+1}(\sqrt{a/b})$$

for each $m \in \mathbb{N}_0$ and $\ell \in \{m-1, m\} \cap \mathbb{N}_0$.

A similar observation holds for the function

$$r_{m+\ell+1}(z,[a,b]) = \frac{(1-e_{m+\ell+1}(\sqrt{a/b})^2)\sqrt{z}}{s_{m+\ell+1}(\sqrt{z},[\sqrt{a},\sqrt{b}])}, \quad m \in \mathbb{N}_0,\ \ell \in \{m-1,m\} \cap \mathbb{N}_0. \tag{25}$$

It is extremal for $E_{m,\ell}(\sqrt{z},[a,b])$, since

$$\frac{r_{m+\ell+1}(z,[a,b]) - \sqrt{z}}{\sqrt{z}} = \frac{1-e_{m+\ell+1}(\sqrt{a/b})^2}{s_{m+\ell+1}(\sqrt{z},[\sqrt{a},\sqrt{b}])} - 1$$

equioscillates $m+\ell+2$ times on $[a,b]$ with extrema $\frac{1-e_{m+\ell+1}(\sqrt{a/b})^2}{1\pm e_{m+\ell+1}(\sqrt{a/b})} - 1 = \mp e_{m+\ell+1}(\sqrt{a/b})$. In particular,

$$E_{m,\ell}(\sqrt{z},[a,b]) = e_{m+\ell+1}(\sqrt{a/b})$$

for each $m \in \mathbb{N}_0$ and $\ell \in \{m-1,m\} \cap \mathbb{N}_0$. Note that in the notation of section 2, $r_{m+\ell+1}(z,[\alpha^2,1]) = r_{m,\ell}(z,\alpha)$.

Error estimates. The errors $e_p(\alpha) = E_{p,p}(\mathrm{sign},[-1,-\alpha]\cup[\alpha,1])$ are known to satisfy

$$e_p(\alpha) = \frac{2\sqrt{\mathcal{Z}_p(\alpha)}}{1+\mathcal{Z}_p(\alpha)}$$

for each $p \in \mathbb{N}$, where

$$\mathcal{Z}_p(\alpha) = \inf_{r\in\mathcal{R}_{p,p}} \frac{\sup_{z\in[\alpha,1]}|r(z)|}{\inf_{z\in[-1,-\alpha]}|r(z)|}$$

is the *Zolotarev number* of the sets $[-1,-\alpha]$ and $[\alpha,1]$ [4, p. 9]. An explicit formula for $\mathcal{Z}_p(\alpha)$ is given in [4, Theorem 3.1]. For our purposes, it is enough to know that $\mathcal{Z}_p(\alpha)$ obeys an asymptotically sharp inequality [4, Corollary 3.2]

$$\mathcal{Z}_p(\alpha) \le 4\rho(\alpha)^{-2p},$$

where $\rho(\alpha)$ is given by (13). This shows that for each $m \in \mathbb{N}_0$ and $\ell \in \{m-1,m\}\cap\mathbb{N}_0$,

$$E_{\ell,m}(1/\sqrt{z},[\alpha^2,1]) = e_{m+\ell+1}(\alpha) \le 2\sqrt{\mathcal{Z}_{m+\ell+1}(\alpha)} \le 4\rho(\alpha)^{-(m+\ell+1)}, \tag{26}$$

$$E_{m,\ell}(\sqrt{z},[\alpha^2,1]) = e_{m+\ell+1}(\alpha) \le 2\sqrt{\mathcal{Z}_{m+\ell+1}(\alpha)} \le 4\rho(\alpha)^{-(m+\ell+1)}, \tag{27}$$

and these bounds are asymptotically sharp. (The upper bound for $E_{m,m-1}(\sqrt{z},[\alpha^2,1])$ also appears in [5, Theorem 5.5, p. 151].)

Explicit formulas. For later use, we record here explicit formulas for $s_p(z,[a,b])$. For $m \in \mathbb{N}_0$, $\ell \in \{m-1,m\}\cap\mathbb{N}_0$ and $\alpha \in (0,1)$, we have [1, p. 286]

$$s_{m+\ell+1}(z,[\alpha,1]) = M(\alpha) z \frac{\prod_{j=1}^{\ell}(z^2+c_{2j}(\alpha))}{\prod_{j=1}^{m}(z^2+c_{2j-1}(\alpha))},$$

where

$$c_j(\alpha) = \alpha^2 \frac{\mathrm{sn}^2\left(\frac{jK(\alpha')}{m+\ell+1};\alpha'\right)}{\mathrm{cn}^2\left(\frac{jK(\alpha')}{m+\ell+1};\alpha'\right)}, \tag{28}$$

and $M(\alpha)$ is a scalar uniquely defined by the condition that

$$\min_{z\in[\alpha,1]} \left(s_{m+\ell+1}(z,[\alpha,1]) - 1\right) = -\max_{z\in[\alpha,1]} \left(s_{m+\ell+1}(z,[\alpha,1]) - 1\right).$$

The extrema of $s_{m+\ell+1}(z,[\alpha,1])$ on $[\alpha,1]$ occur at the points $\alpha = z_0 < z_1 < \cdots < z_{m+\ell+1} = 1$ given by [1, p. 286]:

$$z_j = \alpha \Big/ \mathrm{dn}\left(\frac{jK(\alpha')}{m+\ell+1};\alpha'\right),\ j = 0,1,\ldots,m+\ell+1. \tag{29}$$

3.2. Composition of rational approximants. The functions $s_p(z,[a,b])$, $p \in \mathbb{N}$, obey a beautiful relationship under composition. This relationship has been studied in [24] for the case in which p is odd, but, as mentioned there, their argument extends easily to general parity.

In detail, let $a_p = 1 - e_p(a/b)$ and $b_p = 1 + e_p(a/b)$. For any $q \in \mathbb{N}$, the function $s_q(s_p(z,[a,b]),[a_p,b_p]) - 1$ equioscillates $qp+1$ times on $[a,b]$, owing to two facts. The values of $s_p(z,[a,b])$ run from/to a_p to/from b_p a total of p times as z runs from a to b, and $s_q(z,[a_p,b_p]) - 1$ equioscillates $q+1$ times on $[a_p,b_p]$, achieving its extremal values at the endpoints. Since $s_{qp}(z,[a,b])$ is the unique odd rational function of type (qp,qp) with the property that $s_{qp}(z,[a,b]) - 1$ equioscillates $qp+1$ times on $[a,b]$, it follows that

$$s_{qp}(z,[a,b]) = s_q(s_p(z,[a,b]),[a_p,b_p]), \quad a_p = 1 - e_p(a/b),\ b_p = 1 + e_p(a/b). \tag{30}$$

From this we will obtain composition formulas for $r_p(z,[a,b])$. For convenience, we focus on intervals of the form $[\alpha^2,1]$, $\alpha \in (0,1)$, and denote

$$\hat{r}_p(z,\alpha) = \frac{r_p(z,[\alpha^2,1])}{1 - e_p(\alpha)} = \frac{(1+e_p(\alpha))\sqrt{z}}{s_p(\sqrt{z},[\alpha,1])}. \tag{31}$$

Note that in the notation of section 2, $\hat{r}_{2m}(z,\alpha) = \hat{r}_{m,m-1}(z,\alpha)$ and $\hat{r}_{2m+1}(z,\alpha) = \hat{r}_{m,m}(z,\alpha)$; hence, $R_{2m}(z,[\alpha^2,1]) = h_{m-1,m}(z,\alpha)(1+e_{2m}(\alpha))$ and $R_{2m+1}(z,[\alpha^2,1]) = h_{m,m}(z,\alpha)(1+e_{2m+1}(\alpha))$. We may summarize these identities as follows: For $m \in \mathbb{N}_0$ and $\ell \in \{m-1,m\} \cap \mathbb{N}_0$,

$$h_{\ell,m}(z,\alpha) = \frac{1}{\hat{r}_{m,\ell}(z,\alpha)} = \frac{1 - e_{m+\ell+1}(\alpha)}{r_{m,\ell}(z,\alpha)} = \frac{R_{m+\ell+1}(z,[\alpha^2,1])}{1 + e_{m+\ell+1}(\alpha)}.$$

LEMMA 6. *For any $p,q \in \mathbb{N}$ and $\alpha \in (0,1)$,*

$$\hat{r}_{qp}(z,\alpha) = \hat{r}_p(z,\alpha)\hat{r}_q\left(\frac{z}{\hat{r}_p(z,\alpha)^2},\beta_p\right), \quad \beta_p = \frac{1 - e_p(\alpha)}{1 + e_p(\alpha)}. \tag{32}$$

Proof. Using (31), (23), and the composition formula (30), we have

$$\begin{aligned}\frac{(1+e_{qp}(\alpha))\sqrt{z}}{\hat{r}_{qp}(z,\alpha)} &= s_{qp}(\sqrt{z},[\alpha,1])\\ &= s_q(s_p(\sqrt{z},[\alpha,1]),[1-e_p(\alpha),1+e_p(\alpha)])\\ &= s_q\left(\frac{s_p(\sqrt{z},[\alpha,1])}{1+e_p(\alpha)},[\beta_p,1]\right)\\ &= s_q\left(\frac{\sqrt{z}}{\hat{r}_p(z,\alpha)},[\beta_p,1]\right)\\ &= \frac{(1+e_q(\beta_p))\sqrt{z}/\hat{r}_p(z,\alpha)}{\hat{r}_q(z/\hat{r}_p(z,\alpha)^2,\beta_p)}.\end{aligned}$$

The result follows upon noting that

$$\begin{aligned}e_{qp}(\alpha) &= 1-s_{qp}(\alpha,[\alpha,1])\\ &= 1-s_q(s_p(\alpha,[\alpha,1]),[1-e_p(\alpha),1+e_p(\alpha)])\\ &= 1-s_q(1-e_p(\alpha),[1-e_p(\alpha),1+e_p(\alpha)])\\ &= 1-s_q(\beta_p,[\beta_p,1])\\ &= e_q(\beta_p),\end{aligned} \tag{33}$$

where we have used (24) in the first and third lines above. ∎

3.3. Proofs of results from section 2. We now prove the results claimed in section 2. We begin with Theorem 1, which, when written more compactly, states the following. If $\alpha \in (0,1)$, $q \in \{2,3,4,\ldots\}$, and

$$f_{k+1}(z) = f_k(z)\hat{r}_q\left(\frac{z}{f_k(z)^2},\alpha_k\right), \qquad f_0(z)=1, \tag{34}$$

$$\alpha_{k+1} = \frac{\alpha_k}{\hat{r}_q(\alpha_k^2,\alpha_k)}, \qquad \alpha_0=\alpha, \tag{35}$$

then

$$f_k(z) = \hat{r}_{q^k}(z,\alpha) = \frac{1+\alpha_k}{2\alpha_k} r_{q^k}(z,[\alpha^2,1]) \tag{36}$$

for every $k \geq 1$. Indeed, the relations (34)–(36) with $q=2m$ and $q=2m+1$, respectively, are equivalent to the relations (4)–(6) and (7)–(9), respectively. Note that the case $q=1$ corresponds to the trivial iteration $f_{k+1}(z)=f_k(z)$, $\alpha_{k+1}=\alpha_k$.

Proof of Theorem 1. Taking $p=q^k$ in Lemma 6 gives

$$\hat{r}_{q^{k+1}}(z,\alpha) = \hat{r}_{q^k}(z,\alpha)\hat{r}_q\left(\frac{z}{\hat{r}_{q^k}(z,\alpha)^2},\beta_{q^k}\right), \quad \beta_{q^k} = \frac{1-e_{q^k}(\alpha)}{1+e_{q^k}(\alpha)},$$

so the proof will be complete if we can show that α_k in (35) satisfies

$$\alpha_k = \beta_{q^k}, \quad k=0,1,2,\ldots, \tag{37}$$

and (comparing (36) with (31))

$$\frac{1+\alpha_k}{2\alpha_k} = \frac{1}{1-e_{q^k}(\alpha)}. \tag{38}$$

A direct calculation shows that $e_1(\alpha) = E_{1,1}(\text{sign}, [-1,-\alpha]\cup[\alpha,1]) = \frac{1-\alpha}{1+\alpha}$, so $\beta_1 = \frac{1-e_1(\alpha)}{1+e_1(\alpha)} = \alpha = \alpha_0$. We also have, by (33), (24), and (31),

$$\beta_{q^{k+1}} = \frac{1-e_{q^{k+1}}(\alpha)}{1+e_{q^{k+1}}(\alpha)} = \frac{1-e_q(\beta_{q^k})}{1+e_q(\beta_{q^k})} = \frac{s_q(\beta_{q^k},[\beta_{q^k},1])}{1+e_q(\beta_{q^k})} = \frac{\beta_{q^k}}{\hat{r}_q(\beta_{q_k}^2,\beta_{q^k})},$$

so (37) holds for all k. The relation (38) now follows immediately from (37).

Proof of Theorem 2. Theorem 2 is a consequence of the following lemma, which we will prove in nearly the same way that Beckermann proves [3, Theorem 4].

LEMMA 7. *Let $m \in \mathbb{N}$, $\ell \in \{m-1, m\}$, $\alpha \in (0,1)$, and $z \in \mathbb{C}\setminus((-\infty,0]\cup[\alpha^2,1])$. If $2|\varphi(z,\alpha)|^{-(m+\ell+1)} < 1$, then*

$$|r_{m,\ell}(z,\alpha)/\sqrt{z} - 1| \le \frac{4|\varphi(z,\alpha)|^{-(m+\ell+1)}}{1-2|\varphi(z,\alpha)|^{-(m+\ell+1)}},$$

where $\varphi(z,\alpha)$ and $\rho(\alpha)$ are given by (12) *and* (13).

Remark 8. When $z \in [\alpha^2,1]$, the slightly sharper bound

$$|r_{m,\ell}(z,\alpha)/\sqrt{z} - 1| \le 4\rho(\alpha)^{-(m+\ell+1)}$$

holds in view of (27).

Proof. With $\mathcal{Z} := \mathcal{Z}_{m+\ell+1}(\alpha)$, let

$$Q(z) = \frac{1-\left(\frac{1+\mathcal{Z}}{1-\mathcal{Z}}\right)s_{m+\ell+1}(z,[\alpha,1])}{1+\left(\frac{1+\mathcal{Z}}{1-\mathcal{Z}}\right)s_{m+\ell+1}(z,[\alpha,1])}. \tag{39}$$

Since $s_{m+\ell+1}(z,[\alpha,1])$ takes values in $[1-2\sqrt{\mathcal{Z}}/(1+\mathcal{Z}), 1+2\sqrt{\mathcal{Z}}/(1+\mathcal{Z})]$ on the interval $[\alpha,1]$, $Q(z)$ takes values in $[-\sqrt{\mathcal{Z}},\sqrt{\mathcal{Z}}]$ on $[\alpha,1]$. On the other hand, since $s_{m+\ell+1}(z,[\alpha,1])$ is purely imaginary for $z \in i\mathbb{R}$, $|Q(z)| = 1$ for $z \in i\mathbb{R}$.

Recall that $\varphi(z,\alpha)$ supplies a conformal map from $\mathbb{C}\setminus((-\infty,0]\cup[\alpha^2,1])$ to the annulus $\{z\in\mathbb{C} : 1 < |z| < \rho(\alpha)\}$. Thus, by the maximum principle,

$$\begin{aligned}\sup_{z\in\mathbb{C}\setminus((-\infty,0]\cup[\alpha^2,1])} |\varphi(z,\alpha)|^{m+\ell+1}|Q(\sqrt{z})| &= \sup_{z\in\mathbb{C}\setminus((-\infty,0]\cup[\alpha^2,1])} |\varphi(z,\alpha)^{m+\ell+1}Q(\sqrt{z})| \\ &= \sup_{w\in i\mathbb{R}\cup[\alpha,1]} |\varphi(w^2,\alpha)^{m+\ell+1}Q(w)| \\ &\le \max\{1, \rho(\alpha)^{m+\ell+1}\sqrt{\mathcal{Z}}\}.\end{aligned}$$

Since

$$\mathcal{Z} = \mathcal{Z}_{m+\ell+1}(\alpha) \le 4\rho(\alpha)^{-2(m+\ell+1)}, \tag{40}$$

it follows that

$$|Q(\sqrt{z})| \le 2|\varphi(z,\alpha)|^{-(m+\ell+1)} \tag{41}$$

for every $z \in \mathbb{C}\setminus((-\infty,0]\cup[\alpha^2,1])$.

Now observe that by (25) and (39),

$$r_{m,\ell}(z,\alpha) = \left(1-\frac{4\mathcal{Z}}{(1+\mathcal{Z})^2}\right)\frac{\sqrt{z}}{s_{m+\ell+1}(\sqrt{z},[\alpha,1])} = \left(\frac{1-\mathcal{Z}}{1+\mathcal{Z}}\right)\left(\frac{1+Q(\sqrt{z})}{1-Q(\sqrt{z})}\right)\sqrt{z},$$

so

$$r_{m,\ell}(z,\alpha)/\sqrt{z} - 1 = \frac{2(Q(\sqrt{z}) - \mathcal{Z})}{(1+\mathcal{Z})(1-Q(\sqrt{z}))}.$$

Since $|Q(\sqrt{z})| \le \mathcal{Y} := 2|\varphi(z,\alpha)|^{-(m+\ell+1)}$ and $\mathcal{Z} \le \mathcal{Y}^2$, it follows that

$$|r_{m,\ell}(z,\alpha)/\sqrt{z} - 1| \le \frac{2}{1+\mathcal{Z}} \max\left\{\frac{\mathcal{Y}-\mathcal{Z}}{1-\mathcal{Y}}, \frac{\mathcal{Y}+\mathcal{Z}}{1+\mathcal{Y}}\right\} \le \frac{2\mathcal{Y}}{1-\mathcal{Y}}$$

if $\mathcal{Y} < 1$. ∎

Proof of Proposition 4. It is straightforward to deduce from [16, Theorem 5.9] the following explicit formula for the type (m,ℓ) Padé approximant of $\sqrt{z}$ at $z = 1$ for $\ell \in \{m-1, m\}$:

$$p_{m,\ell}(z) = \sqrt{z}\frac{(1+\sqrt{z})^{m+\ell+1} + (1-\sqrt{z})^{m+\ell+1}}{(1+\sqrt{z})^{m+\ell+1} - (1-\sqrt{z})^{m+\ell+1}}.$$

It is then easy to check by direct substitution that the roots and poles of $p_{m,\ell}(z)$ are $\{-\tan^2\,(\frac{(2j-1)\pi}{2(m+\ell+1)})\}_{j=1}^m$ and $\{-\tan^2\,(\frac{j\pi}{m+\ell+1})\}_{j=1}^\ell$, respectively.

On the other hand, the roots and poles of

$$\hat{r}_{m,\ell}(z,\alpha) = \frac{1+e_{m+\ell+1}(\alpha)}{M(\alpha)}\frac{\prod_{j=1}^m (z+c_{2j-1}(\alpha))}{\prod_{j=1}^\ell (z+c_{2j}(\alpha))}$$

are $\{-c_{2j-1}(\alpha)\}_{j=1}^m$ and $\{-c_{2j}(\alpha)\}_{j=1}^\ell$, respectively, where $c_j(\alpha)$ is given by (28). These approach the roots and poles of $p_{m,\ell}(z)$, since the identities $K(0) = \pi/2$, $\mathrm{sn}(z,0) = \sin z$, and $\mathrm{cn}(z,0) = \cos z$ [8, Table 22.5.3] imply that

$$\lim_{\alpha\uparrow 1} c_j(\alpha) = \tan^2\left(\frac{j\pi}{2(m+\ell+1)}\right).$$

The proof is completed by noting that $\hat{r}_{m,\ell}$ is scaled in such a way that $\lim_{\alpha\uparrow 1}\hat{r}_{m,\ell}(1,\alpha) = 1 = p_{m,\ell}(1)$.

Remark 9. Proposition 4 is related to a general result concerning the convergence of minimax approximants to Padé approximants [29]. We showed above that $p_{m,\ell}(z)$ is nondegenerate (it has exactly m roots and ℓ poles), so Theorem 3b of [29] implies that $\arg\min_{r\in\mathcal{R}_{m,\ell}} \max_{z\in[\alpha^2,1]} |r(z)-\sqrt{z}|$ converges coefficientwise to $p_{m,\ell}(z)$ as $\alpha \uparrow 1$. Proposition 4 implies that the same is true for minimizers of the maximum *relative* error $|(r(z)-\sqrt{z})/\sqrt{z}|$.

Proof of Theorem 5. Let $q = m+\ell+1$, and let $\rho(\alpha)$ be as in (13). We see from (37) and (26)–(27) that

$$1 - \alpha_k = 1 - \frac{1-e_{q^k}(\alpha)}{1+e_{q^k}(\alpha)} = \frac{2e_{q^k}(\alpha)}{1+e_{q^k}(\alpha)} \le 2e_{q^k}(\alpha) \le 8\rho(\alpha)^{-q^k},$$

so $\alpha_k \to 1$ with R-order of convergence $m+\ell+1$. Now let $A \in \mathbb{C}^{n\times n}$ have no nonpositive real eigenvalues. For $\epsilon > 0$ sufficiently small, the pseudospectrum $\Lambda_\epsilon(A)$ is compactly contained in $\mathbb{C}\setminus(\infty,0]$, so $1 < \sup_{z\in\overline{\Lambda_\epsilon(A)}} |\varphi(z,\alpha)| =: \gamma$. Since $\overline{\Lambda_\epsilon(A)}$ is

a spectral set for A, we conclude from Theorem 2 that there exists a constant L such that the matrix $\widetilde{X}_k := 2\alpha_k X_k/(1+\alpha_k)$ in the iteration (10)–(11) satisfies

$$\|\widetilde{X}_k A^{-1/2} - I\|_2 \le \frac{4L\gamma^{-(m+\ell+1)^k}}{1-2\gamma^{(m+\ell+1)^k}}$$

for every k sufficiently large. Considering now the coupled iteration (20)–(22), let $\widetilde{Y}_k = (1+\alpha_k)Y_k/(2\alpha_k)$ and $\widetilde{Z}_k = (1+\alpha_k)Z_k/(2\alpha_k)$. Since $X_k = f_k(A)$ with f_k the rational function defined by (34)–(35) with $q = m+\ell+1$, we have

$$\begin{aligned}
\widetilde{X}_k &= \widetilde{f}_k(A), & \widetilde{f}_k(z) &= \frac{2\alpha_k}{1+\alpha_k} f_k(z),\\
\widetilde{Y}_k &= \widetilde{X}_k^{-1} A = \widetilde{u}_k(A), & \widetilde{u}_k(z) &= \frac{z}{\widetilde{f}_k(z)},\\
\widetilde{Z}_k &= \widetilde{X}_k^{-1} = \widetilde{v}_k(A), & \widetilde{v}_k(z) &= \frac{1}{\widetilde{f}_k(z)}.
\end{aligned}$$

In terms of the functions

$$\sigma_k(z) = \frac{\widetilde{f}_k(z) - \sqrt{z}}{\sqrt{z}}, \tag{42}$$

$$\tau_k(z) = \frac{\widetilde{u}_k(z) - \sqrt{z}}{\sqrt{z}} = \frac{\widetilde{v}_k(z) - 1/\sqrt{z}}{1/\sqrt{z}} = \frac{-\sigma_k(z)}{1+\sigma_k(z)}, \tag{43}$$

we have

$$\frac{\|\widetilde{Y}_k - A^{1/2}\|_2}{\|A^{1/2}\|_2} = \frac{\|(\widetilde{Y}_k A^{-1/2} - I)A^{1/2}\|_2}{\|A^{1/2}\|_2} \le \|\widetilde{Y}_k A^{-1/2} - I\|_2 = \|\tau_k(A)\|_2,$$

$$\frac{\|\widetilde{Z}_k - A^{-1/2}\|_2}{\|A^{-1/2}\|_2} = \frac{\|(\widetilde{Z}_k A^{1/2} - I)A^{-1/2}\|_2}{\|A^{-1/2}\|_2} \le \|\widetilde{Z}_k A^{1/2} - I\|_2 = \|\tau_k(A)\|_2,$$

and

$$\|\sigma_k(A)\|_2 = \|\widetilde{X}_k A^{-1/2} - I\|_2 \le \frac{4L\gamma^{-(m+\ell+1)^k}}{1-2\gamma^{(m+\ell+1)^k}}$$

for k sufficiently large. Now

$$\|\tau_k(A)\|_2 \le \|(I+\sigma_k(A))^{-1}\|_2 \|\sigma_k(A)\|_2 \le \frac{\|\sigma_k(A)\|_2}{1-\|\sigma_k(A)\|_2} \le \frac{4L\gamma^{-(m+\ell+1)^k}}{1-(4L+2)\gamma^{-(m+\ell+1)^k}}$$

for k sufficiently large, so $\widetilde{Y}_k \to A^{1/2}$ and $\widetilde{Z}_k \to A^{-1/2}$ with R-order of convergence $m+\ell+1$. The same is true for Y_k and Z_k since

$$\|Y_k - A^{1/2}\|_2 \le \|Y_k - \widetilde{Y}_k\|_2 + \|\widetilde{Y}_k - A^{1/2}\|_2 = \frac{1-\alpha_k}{1+\alpha_k}\|\widetilde{Y}_k\|_2 + \|\widetilde{Y}_k - A^{1/2}\|_2,$$

$$\|Z_k - A^{-1/2}\|_2 \le \|Z_k - \widetilde{Z}_k\|_2 + \|\widetilde{Z}_k - A^{-1/2}\|_2 = \frac{1-\alpha_k}{1+\alpha_k}\|\widetilde{Z}_k\|_2 + \|\widetilde{Z}_k - A^{-1/2}\|_2.$$

Stability of the iteration (in the sense described above Theorem 5) follows from the fact that (20)–(22) reduces to the stable Padé iteration (18)–(19) when $\alpha_k = 1$ [16, Theorem 6.12]. Indeed, it is shown in [16, Theorem 6.12] that the map $f(Y,Z) = (Yh_{\ell,m}(ZY,1), h_{\ell,m}(ZY,1)Z) = (Yq_{\ell,m}(ZY), q_{\ell,m}(ZY)Z)$ has Fréchet derivative

$$L_f((A^{1/2}, A^{-1/2}),(E,F)) = \frac{1}{2}(E - A^{1/2}FA^{1/2}, F - A^{-1/2}EA^{-1/2})$$

at $(A^{1/2}, A^{-1/2})$. Since $L_f((A^{1/2}, A^{-1/2}),(\cdot,\cdot))$ is idempotent, it has bounded powers.

Remark 10. The preceding proof reveals that the functional analogue of the coupled Zolotarev iteration (20)–(22),

$$u_{k+1}(z) = u_k(z)h_{\ell,m}(v_k(z)u_k(z), \alpha_k), \qquad u_0(z) = z, \tag{44}$$

$$v_{k+1}(z) = h_{\ell,m}(v_k(z)u_k(z), \alpha_k)v_k(z), \qquad v_0(z) = 1, \tag{45}$$

$$\alpha_{k+1} = \alpha_k h_{\ell,m}(\alpha_k^2, \alpha_k), \qquad \alpha_0 = \alpha, \tag{46}$$

generates, up to rescaling, an optimal rational approximant of $1/\sqrt{z}$ on $[\alpha^2, 1]$. Indeed,

$$\frac{1+\alpha_k}{2\alpha_k} v_k(z) = \widetilde{v}_k(z) = \frac{1}{r_{q^k}(z, [\alpha^2, 1])} = \frac{R_{q^k}(z, [\alpha^2, 1])}{1 - e_{q^k}(\alpha)^2}, \quad q = m + \ell + 1.$$

The same cannot be said for $u_k(z)$, since $u_k(z)$ has a root at $z = 0$, whereas the optimal rational approximants of $\sqrt{z}$ on $[\alpha^2, 1]$ do not.

4. Practical considerations. In this section, we discuss the implementation of the Zolotarev iterations, strategies for terminating the iterations, and computational costs.

4.1. Implementation. To implement the Zolotarev iteration (20)–(22), we advocate the use of a partial fraction expansion of $h_{\ell,m}(\cdot, \alpha)$, since it enhances parellelizability and, in our experience, tends to improve stability. The following lemma, adapted from [24, Proposition 7], details the partial fraction expansion of $h_{\ell,m}(z, \alpha)$ for $\ell \in \{m-1, m\}$. For the reader's benefit, we recall here that

$$c_j(\alpha) = \alpha^2 \frac{\operatorname{sn}^2\left(\frac{jK(\alpha')}{m+\ell+1}; \alpha'\right)}{\operatorname{cn}^2\left(\frac{jK(\alpha')}{m+\ell+1}; \alpha'\right)}, \tag{47}$$

where $\operatorname{sn}(\cdot;\alpha)$, $\operatorname{cn}(\cdot;\alpha)$, and $\operatorname{dn}(\cdot;\alpha)$ denote Jacobi's elliptic functions with modulus α, $K(\alpha) = \int_0^{\pi/2} (1 - \alpha^2 \sin^2\theta)^{-1/2}\, d\theta$ is the complete elliptic integral of the first kind, and $\alpha' = \sqrt{1-\alpha^2}$ is the complementary modulus to α.

LEMMA 11. *We have*

$$h_{\ell,m}(z, \alpha) = \begin{cases} \hat{M}(\alpha) \displaystyle\sum_{j=1}^{m} \frac{a_j(\alpha)}{z + c_{2j-1}(\alpha)} & \text{if } \ell = m-1, \quad (48) \\ \hat{N}(\alpha) \left(1 + \displaystyle\sum_{j=1}^{m} \frac{a_j(\alpha)}{z + c_{2j-1}(\alpha)}\right) & \text{if } \ell = m, \quad (49) \end{cases}$$

where

$$a_j(\alpha) = \prod_{p=1}^{\ell} (c_{2p}(\alpha) - c_{2j-1}(\alpha)) \Big/ \prod_{\substack{p=1 \\ p \neq j}}^{m} (c_{2p-1}(\alpha) - c_{2j-1}(\alpha)), \tag{50}$$

$$\hat{M}(\alpha) = \left(\sqrt{\zeta} \sum_{j=1}^{m} \frac{a_j(\alpha)}{\zeta + c_{2j-1}(\alpha)}\right)^{-1}, \quad \zeta = \alpha^2 \Big/ \operatorname{dn}^2\left(\frac{K(\alpha')}{2m}; \alpha'\right), \tag{51}$$

$$\hat{N}(\alpha) = \left(1 + \sum_{j=1}^{m} \frac{a_j(\alpha)}{1 + c_{2j-1}(\alpha)}\right)^{-1}. \tag{52}$$

Proof. Since

$$h_{\ell,m}(z,\alpha) = \hat{r}_{m,\ell}(z,\alpha)^{-1} = \frac{M(\alpha)}{1+e_{m+\ell+1}(\alpha)} \frac{\prod_{p=1}^{\ell}(z+c_{2p}(\alpha))}{\prod_{p=1}^{m}(z+c_{2p-1}(\alpha))} \tag{53}$$

and the poles $c_{2p-1}(\alpha)$ are distinct, the partial fraction expansions of $h_{m-1,m}(z,\alpha)$ and $h_{m,m}(z,\alpha)$ must have the form (48) and (49), respectively, for some scalars $a_j(\alpha)$, $\hat{M}(\alpha)$, and $\hat{N}(\alpha)$. Multiplying (53) and (48)–(49) by $z+c_{2j-1}(\alpha)$ and setting $z = -c_{2j-1}(\alpha)$ gives (50), up to a rescaling. The scalars $\hat{M}(\alpha)$ and $\hat{N}(\alpha)$ are determined uniquely by the condition that

$$\min_{z\in[\alpha^2,1]}(h_{\ell,m}(z,\alpha)^{-1}/\sqrt{z}-1) = 0.$$

We know from (29) and (24) that on the interval $[\alpha^2,1]$, $h_{\ell,m}(z,\alpha)^{-1}/\sqrt{z} = \hat{r}_{m,\ell}(z,\alpha)/\sqrt{z} = (1+e_{m+\ell+1}(\alpha))/s_{m+\ell+1}(\sqrt{z},[\alpha,1])$ achieves its minimum at (for instance) $z = \alpha^2/\mathrm{dn}^2(\frac{K(\alpha')}{2m};\alpha')$ (when $\ell = m-1$) and $z=1$ (when $\ell = m$), so (51)–(52) follow. ∎

Note that $c_j(\alpha)$ and $a_j(\alpha)$ in (47) and (50) depend implicitly on m and ℓ. In particular, $a_j(\alpha)$ and $c_{2j-1}(\alpha)$ have different values in (48) and (51) (where $\ell = m-1$) than they do in (49) and (52) (where $\ell = m$).

Note also that accurate evaluation of $K(\alpha')$, $\mathrm{sn}(\cdot;\alpha')$, $\mathrm{cn}(\cdot;\alpha')$, and $\mathrm{dn}(\cdot;\alpha')$ in floating point arithmetic is a delicate task when $\alpha' \approx 1 \iff \alpha \approx 0$ [24, section 4.3]. Rather than using the built-in MATLAB functions `ellipj` and `ellipke` to evaluate these elliptic functions, we recommend using the code described in [24, section 4.3], which is tailored for our application.

Written in full, the Zolotarev iteration (20)–(22) of type $(m,m-1)$[1] reads

$$Y_{k+1} = \hat{M}(\alpha_k)\sum_{j=1}^{m} a_j(\alpha_k)Y_k(Z_kY_k + c_{2j-1}(\alpha_k)I)^{-1}, \qquad Y_0 = A, \tag{54}$$

$$Z_{k+1} = \hat{M}(\alpha_k)\sum_{j=1}^{m} a_j(\alpha_k)(Z_kY_k + c_{2j-1}(\alpha_k)I)^{-1}Z_k, \qquad Z_0 = I, \tag{55}$$

$$\alpha_{k+1} = \alpha_k h_{m-1,m}(\alpha_k^2,\alpha_k), \qquad \alpha_0 = \alpha, \tag{56}$$

and the Zolotarev iteration of type (m,m) reads

$$Y_{k+1} = \hat{N}(\alpha_k)\left(Y_k + \sum_{j=1}^{m} a_j(\alpha_k)Y_k(Z_kY_k + c_{2j-1}(\alpha_k)I)^{-1}\right), \qquad Y_0 = A, \tag{57}$$

$$Z_{k+1} = \hat{N}(\alpha_k)\left(Z_k + \sum_{j=1}^{m} a_j(\alpha_k)(Z_kY_k + c_{2j-1}(\alpha_k)I)^{-1}Z_k\right), \qquad Z_0 = I, \tag{58}$$

$$\alpha_{k+1} = \alpha_k h_{m,m}(\alpha_k^2,\alpha_k), \qquad \alpha_0 = \alpha. \tag{59}$$

As alluded to earlier, a suitable choice for α is $\alpha = \sqrt{|\lambda_{\min}(A)/\lambda_{\max}(A)|}$ (or an estimate thereof), and it is important to scale A so that its spectral radius is 1 (or

[1] Although $h_{m-1,m}(z,\alpha)$ is a rational function of type $(m-1,m)$, we continue to refer to this iteration as the type $(m,m-1)$ Zolotarev iteration since $\hat{r}_{m,m-1}(z,\alpha) = h_{m-1,m}(z,\alpha)^{-1}$ is of type $(m,m-1)$.

approximately 1). In addition, to eliminate the possibility that roundoff errors render $\alpha_k > 1$ for some k, we recommend setting α_k (and all subsequent iterates) equal to 1 (thereby reverting to the Padé iteration (18)–(19)) once α_k exceeds a threshold $1 - \epsilon$, $\epsilon \ll 1$.

4.2. Floating point operations. The computational costs of the Zolotarev iterations depend on the precise manner in which they are implemented. One option is to compute Z_kY_k (1 matrix multiplication), obtain $h_{\ell,m}(Z_kY_k, \alpha_k)$ by computing $(Z_kY_k + c_{2j-1}(\alpha_k)I)^{-1}$ for each j (m matrix inversions), and multiply Y_k and Z_k by $h_{\ell,m}(Z_kY_k, \alpha_k)$ (2 matrix multiplications). An alternative that is better suited for parallel computations is to compute Z_kY_k (1 matrix multiplication), compute the LU factorization $L_jU_j = Z_kY_k + c_{2j-1}(\alpha_k)I$ for each j (m LU factorizations), and perform m "right divisions by a factored matrix" $Y_k(L_jU_j)^{-1}$ and m "left divisions by a factored matrix" $(L_jU_j)^{-1}Z_k$ via forward and back substitution. In parallel, all m LU factorizations can be performed simultaneously, and all $2m$ divisions by factored matrices can be performed simultaneously, so that the effective cost per iteration is $\frac{14}{3}n^3$ flops if A is $n \times n$. In the first iteration, the cost reduces to $\frac{8}{3}n^3$ flops since $Z_0 = I$. The total effective cost for k iterations is $(\frac{8}{3} + \frac{14}{3}(k-1))n^3$ flops, which is less than the (serial) cost of a direct method, $28\frac{1}{3}n^3$ flops [16, p. 136], whenever $k \le 6$.

Yet another alternative is to write (54)–(55) in the form

$$Y_{k+1} = \hat{M}(\alpha_k)\left(\sum_{j=1}^{m} a_j(\alpha_k)Y_k(Y_k + c_{2j-1}(\alpha_k)Z_k^{-1})^{-1}\right)Z_k^{-1}, \qquad Y_0 = A, \tag{60}$$

$$Z_{k+1} = \hat{M}(\alpha_k)\sum_{j=1}^{m} a_j(\alpha_k)(Y_k + c_{2j-1}(\alpha_k)Z_k^{-1})^{-1}, \qquad Z_0 = I, \tag{61}$$

and similarly for (57)–(58). Interestingly, this form of the iteration has exhibited the best accuracy in our numerical experiments, for reasons that are not well understood. It can be parallelized by performing the m right divisions $Y_k(Y_k + c_{2j-1}(\alpha_k)Z_k^{-1})^{-1}$ and m inversions $(Y_k + c_{2j-1}(\alpha_k)Z_k^{-1})^{-1}$ simultaneously, recycling LU factorizations in the obvious way. Moreover, the final multiplication by Z_k^{-1} in (60) can be performed in parallel with the inversion of Z_{k+1}. The effective cost in such a parallel implementation is $\frac{14}{3}kn^3$ flops.

4.3. Termination criteria. We now consider the question of how to terminate the iterations. Define $\widetilde{X}_k = 2\alpha_k X_k/(1+\alpha_k)$, $\widetilde{Y}_k = (1+\alpha_k)Y_k/(2\alpha_k)$, and $\widetilde{Z}_k = (1+\alpha_k)Z_k/(2\alpha_k)$. Since $\widetilde{X}_k$, $\widetilde{Y}_k$, $\widetilde{Z}_k$, and A commute with one another, and since $\widetilde{Y}_k = \widetilde{X}_k^{-1}A$ and $\widetilde{Z}_k = \widetilde{X}_k^{-1} = \widetilde{Y}_kA^{-1}$, it is easy to verify that

$$(\widetilde{Y}_kA^{-1/2} - I) + (\widetilde{Z}_kA^{1/2} - I) = (\widetilde{Z}_k\widetilde{Y}_k - I) - (\widetilde{Z}_kA^{1/2} - I)(\widetilde{Y}_kA^{-1/2} - I)$$

and

$$\widetilde{Y}_kA^{-1/2} - I = (I - \widetilde{X}_kA^{-1/2}) + (\widetilde{Y}_kA^{-1/2} - I)(I - \widetilde{X}_kA^{-1/2}).$$

By dropping second order terms, we see that near convergence,

$$I - \widetilde{X}_kA^{-1/2}. \approx \widetilde{Y}_kA^{-1/2} - I = \widetilde{Z}_kA^{1/2} - I \approx \frac{1}{2}(\widetilde{Z}_k\widetilde{Y}_k - I). \tag{62}$$

The relative errors $\frac{\|\widetilde{Y}_k - A^{1/2}\|}{\|A^{1/2}\|} \le \|\widetilde{Y}_kA^{-1/2} - I\|$ and $\frac{\|\widetilde{Z}_k - A^{-1/2}\|}{\|A^{-1/2}\|} \le \|\widetilde{Z}_kA^{1/2} - I\|$ will therefore be (approximately) smaller than a tolerance $\delta > 0$ so long as

$$\|\widetilde{Z}_k\widetilde{Y}_k - I\| \le 2\delta. \tag{63}$$

While theoretically appealing, the criterion (63) is not ideal for computations for two reasons. It costs an extra matrix multiplication in the last iteration, and, more importantly, (63) may never be satisfied in floating point arithmetic. A cheaper, more robust option is to approximate $\|\widetilde{Z}_k\widetilde{Y}_k - I\|$ based on the value of $\|\widetilde{Z}_{k-1}\widetilde{Y}_{k-1} - I\|$ as follows. In view of (62) and Theorem 2, we have

$$\|\widetilde{Z}_k\widetilde{Y}_k - I\|_2 \lessapprox 8L\gamma^{-(m+\ell+1)^k}$$

for some constants $L \geq 1$ and $\gamma > 1$. Denoting $\varepsilon_k := 8L\gamma^{-(m+\ell+1)^k}$, we have

$$\varepsilon_k \leq 2\delta \iff \varepsilon_{k-1} \leq 8L\left(\frac{\delta}{4L}\right)^{1/(m+\ell+1)}.$$

This suggests that we terminate the iteration and accept $\widetilde{Y}_k$ and $\widetilde{Z}_k$ as soon as

$$\|\widetilde{Z}_{k-1}\widetilde{Y}_{k-1} - I\|_2 \leq 8L\left(\frac{\delta}{4L}\right)^{1/(m+\ell+1)}.$$

In practice, L is not known, and it may be preferable to use a different norm, so we advocate terminating when

$$\|\widetilde{Z}_{k-1}\widetilde{Y}_{k-1} - I\| \leq 8\left(\frac{\delta}{4}\right)^{1/(m+\ell+1)}, \tag{64}$$

where δ is a relative error tolerance with respect to a desired norm $\|\cdot\|$. Note that this test comes at no additional cost if the product $\widetilde{Z}_{k-1}\widetilde{Y}_{k-1}$ was computed at iteration $k-1$. If $\widetilde{Z}_{k-1}^{-1}$ is known but $\widetilde{Z}_{k-1}\widetilde{Y}_{k-1}$ is not (as is the case when (60)–(61) is used), then we have found the following criterion, which is inspired by [16, equation (6.31)], to be an effective alternative:

$$\|\widetilde{Y}_k - \widetilde{Y}_{k-1}\| \leq \left(\delta\frac{\|\widetilde{Y}_k\|}{\|A^{-1}\|\|\widetilde{Z}_{k-1}^{-1}\|}\right)^{1/(m+\ell+1)}. \tag{65}$$

In either case, we recommend also terminating the iteration if the relative change in $\widetilde{Y}_k$ is small but fails to decrease significantly, e.g.,

$$\frac{1}{2}\frac{\|\widetilde{Y}_{k-1} - \widetilde{Y}_{k-2}\|}{\|\widetilde{Y}_{k-1}\|} \leq \frac{\|\widetilde{Y}_k - \widetilde{Y}_{k-1}\|}{\|\widetilde{Y}_k\|} \leq 10^{-2}. \tag{66}$$

5. Numerical examples. In this section, we study the performance of the Zolotarev iterations with numerical experiments.

5.1. Scalar iteration. To gain some intuition behind the behavior of the Zolotarev iteration for the matrix square root, we begin by investigating the behavior of its scalar counterpart.

Lemma 7 shows that if $f_k(z)$ and α_k are defined as in Theorem 1, then

$$\left|\left(\frac{2\alpha_k}{1+\alpha_k}\right)f_k(z)/\sqrt{z} - 1\right| \leq 4|\varphi(z,\alpha)|^{-(m+\ell+1)^k} + O\left(|\varphi(z,\alpha)|^{-2(m+\ell+1)^k}\right), \tag{67}$$

where $\ell = m-1$ if (4)–(5) is used and $\ell = m$ if (7)–(8) is used. Thus, for a given $z \in \mathbb{C}\setminus(\infty, 0]$ and a given relative tolerance $\delta > 0$, we can estimate the smallest k for which $|2\alpha_k f_k(z)/((1+\alpha_k)\sqrt{z}) - 1| \leq \delta$: we have $k \approx \lceil\kappa(z,\alpha)\rceil$ with

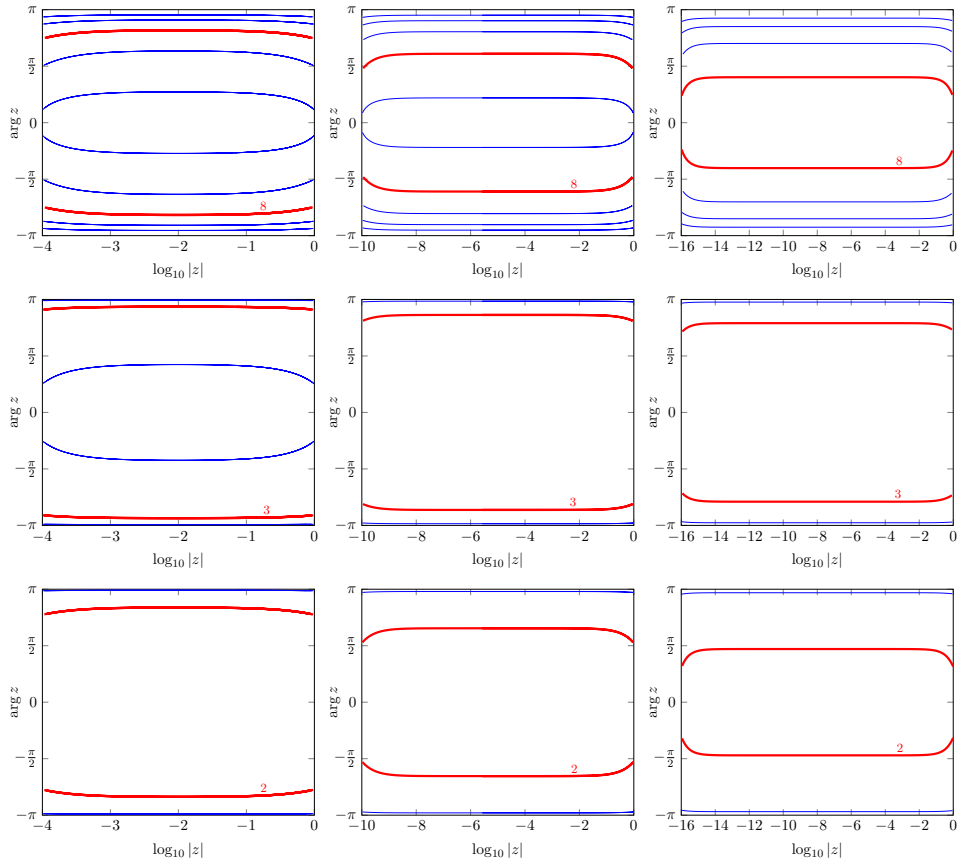

FIG. 1. *Integer level sets of $\kappa(z,\alpha)$ for $(m,\ell)=(1,0)$ (row 1), $(m,\ell)=(4,4)$ (row 2), $(m,\ell)=(8,8)$ (row 3), $\alpha=10^{-2}$ (column 1), $\alpha=10^{-5}$ (column 2), and $\alpha=10^{-8}$ (column 3). To help compare level sets within each row, we have arbitrarily selected a single level set to label and highlight in bold red. Each unlabeled level set's value differs from that of its nearest inner neighbor by $+1$.*

$$\kappa(z,\alpha)=\frac{\log\log(4/\delta)-\log\log|\varphi(z,\alpha)|}{\log(m+\ell+1)}.$$

Figure 1 plots the integer level sets of $\kappa(z,\alpha)$ for $(m,\ell)\in\{(1,0),(4,4),(8,8)\}$, $\delta=10^{-16}$, and $\alpha\in\{10^{-2},10^{-5},10^{-8}\}$ in the slit annulus $\mathcal{A}=\{z\in\mathbb{C}\mid\alpha^2\le|z|\le1,\ -\pi<\arg z<\pi\}$. To improve the clarity of the plots, we have plotted the level sets in the $(\log_{10}|z|,\arg z)$ coordinate plane rather than the usual $(\operatorname{Re}z,\operatorname{Im}z)$ coordinate plane. The level sets have the following interpretation: If $z_0\in\mathbb{C}$ lies within the region enclosed by the level set $\kappa(z,\alpha)=c\in\mathbb{N}$, then the sequence $\{2\alpha_k f_k(z_0)/(1+\alpha_k)\}_{k=0}^{\infty}$ generated by the type (m,ℓ) Zolotarev iteration from Theorem 1 converges to $\sqrt{z_0}$ in at most c iterations with a relative tolerance of $\approx10^{-16}$.

Observe that when $(m,\ell)=(8,8)$ and z_0 lies in the right half-annulus $\{z:\operatorname{Re}z\ge0,\ \alpha^2\le|z|\le1\}$ (which corresponds to the horizontal strip $\{z:2\log_{10}\alpha\le\log_{10}|z|\le0,-\pi/2<\arg z<\pi/2\}$ in Figure 1), convergence of the scalar iteration is achieved in just two iterations whenever $\alpha\ge10^{-5}$. For nearly all other $z_0\in\mathcal{A}$, three iterations suffice.

Comparison with Padé iterations. For comparison, Figure 2 plots the integer level sets of $\kappa(z/\alpha,1)$ for the same values of (m,ℓ), δ, and α as above. In view of Proposition 4, these level sets dictate the convergence of the type (m,ℓ) Padé iteration

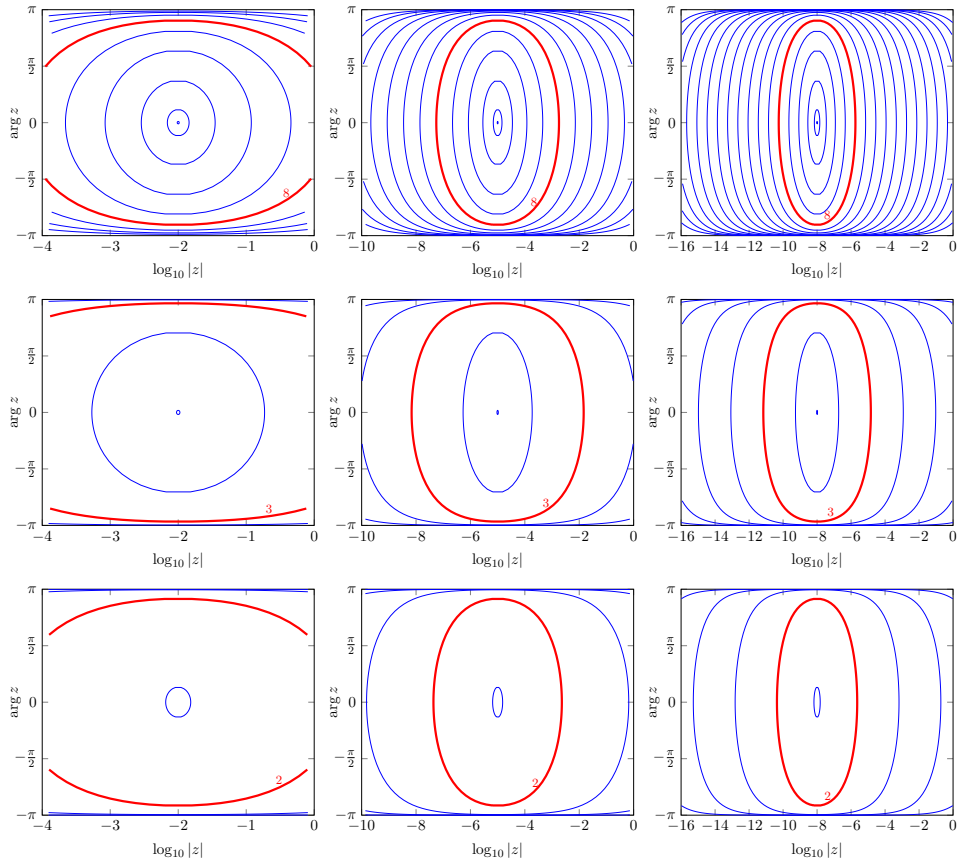

FIG. 2. *Integer level sets of* $\kappa(z/\alpha, 1)$ *for* $(m, \ell) = (1, 0)$ *(row 1),* $(m, \ell) = (4, 4)$ *(row 2),* $(m, \ell) = (8, 8)$ *(row 3),* $\alpha = 10^{-2}$ *(column 1),* $\alpha = 10^{-5}$ *(column 2), and* $\alpha = 10^{-8}$ *(column 3). To help compare level sets within each row, we have arbitrarily selected a single level set to label and highlight in bold red. Each unlabeled level set's value differs from that of its nearest inner neighbor by* $+1$.

with the initial iterate z_0 scaled by $1/\alpha$. For $\alpha = 10^{-2}$ (the leftmost column), the behavior of the Padé iteration is not significantly different from the behavior of the Zolotarev iteration. However, as α decreases, a clear pattern emerges. The level sets $\kappa(z/\alpha, 1) = c$ do not begin to enclose scalars z with extreme magnitudes ($|z| \approx \alpha^2$ and $|z| \approx 1$) until c is relatively large. For example, when $\alpha = 10^{-8}$ and $(m, \ell) = (8, 8)$, the smallest integer c for which the level set $\kappa(z/\alpha, 1) = c$ encloses both $z = \alpha^2$ and $z = 1$ is $c = 5$ (see the lower right plot of Figure 2). In contrast, for the Zolotarev iteration with the same (m, ℓ) and α, the smallest integer c for which $\kappa(z, \alpha) = c$ encloses both $z = \alpha^2$ and $z = 1$ is $c = 2$ (see the lower right plot of Figure 1). The situation is similar when $|z| \in \{\alpha^2, 1\}$ and z has nonzero imaginary part.

Implications. The preceding observations have important implications for computing the square root of a matrix $A \in \mathbb{C}^{n\times n}$ with no nonpositive real eigenvalues. Without loss of generality, we may assume that A has been scaled in such a way that its spectrum $\Lambda(A)$ is contained in the slit annulus $\{z \in \mathbb{C} \mid \alpha^2 \le |z| \le 1, -\pi < \arg z < \pi\}$ for some $\alpha \in (0, 1)$. Then, if A is normal, the number of iterations needed for the Zolotarev iteration of type (m, ℓ) to converge to $A^{1/2}$ (i.e., $\|2\alpha_k X_k A^{-1/2}/(1 + \alpha_k) - I\| \lessapprox 10^{-16}$) in exact arithmetic is given by the smallest

integer c for which the level set $\kappa(z,\alpha) = c$ encloses $\Lambda(A)$. For the Padé iteration (with A rescaled by $1/\alpha$) the same statement holds with $\kappa(z,\alpha)$ replaced by $\kappa(z/\alpha, 1)$.

We conclude from the preceding discussion that the Zolotarev iterations are often preferable when A has eigenvalues with widely varying magnitudes (assuming A is normal). For instance, if $|\lambda_{\max}(A)|/|\lambda_{\min}(A)| = \alpha^{-2} \le 10^{10}$ and the spectrum of A lies in the right half plane, then the Zolotarev iteration of type $(8,8)$ converges in at most two iterations, whereas the Padé iteration of type $(8,8)$ converges in at most four (see row 3, columns 1–2 of Figures 1–2). When considering nonnormal A and/or the effects of roundoff errors, the situation is of course more difficult to analyze, but we address this situation with numerical experiments in section 5.2.

Note that in the Padé iteration (17), it is common to scale not only the initial iterate X_0 but also subsequent iterates X_k, by $\mu_k = |\det(X_k)/\det(A)^{1/2}|^{-1/n}$. (More precisely, this is accomplished in a mathematically equivalent, numerically stabler way by scaling Y_k and Z_k by $\mu_k^{-1} = |(\det Y_k \det Z_k)^{-1/(2n)}|$ in (18)–(19) [15, equation (3.2)]). These scalars will of course depend on the distribution of the eigenvalues of A, but in the case in which $m = \ell$ and αA has real eigenvalues with logarithms uniformly distributed in $[2\log_{10}\alpha, 0]$, one finds that $\mu_k = 1$ for $k \ge 1$, showing that Figure 2 is a fair representation of the behavior of the scaled Padé iteration.

5.2. Matrix iteration. In what follows, we compare the Zolotarev iterations of type (m,ℓ) (hereafter referred to as Z-(m,ℓ)) with the following other methods: the DB iteration [16, equation (6.28)] (see also [7]), the product form of the Denman–Beavers iteration (DBp) [16, equation (6.29)], the incremental Newton (IN) iteration [16, equation (6.30)][2] (see also [23, 19]), the principal Padé iterations of type (m,ℓ) (P-(m,ℓ)) [16, equation (6.34)] (see also [15, 17]), and the MATLAB function `sqrtm`. In the Padé and Zolotarev iterations, we focus on the iterations of type $(1,0)$, $(4,4)$, and $(8,8)$ for simplicity.

In all of the iterations (except the Zolotarev iterations), we use determinantal scaling (as described in [16, section 6.5] and [15, equation (3.2)]) until the ∞-norm relative change in X_k falls below 10^{-2}. In the Zolotarev iterations, we use $\alpha = \sqrt{|\lambda_{\min}(A)/\lambda_{\max}(A)|}$, we scale A so that its spectral radius is 1, and we set α_k (and all subsequent iterates) equal to 1 as soon as α_k exceeds $1 - 10u$, where $u = 2^{-53}$ is the unit roundoff. In the Zolotarev and Padé iterations, we use the formulation (60)–(61) and its type-(m,m) counterpart, and we terminate the iterations when either (65) or (66) is satisfied in the ∞-norm with $\delta = u\sqrt{n}$. To terminate the DB and IN iterations, we use the following termination criterion [16, p. 148]: $\|X_k - X_{k-1}\|_\infty \le (\delta\|X_k\|_\infty/\|X_{k-1}^{-1}\|_\infty)^{1/2}$ or $\frac{1}{2}\|X_{k-1} - X_{k-2}\|_\infty/\|X_{k-1}\|_\infty \le \|X_k - X_{k-1}\|_\infty/\|X_k\|_\infty \le 10^{-2}$. To terminate the DBp iteration, we replace the first condition by $\|M_k - I\|_\infty \le \delta$, where M_k is the "product" matrix in [16, equation (6.29)]. We impose a maximum of 20 iterations for each method.

Four test matrices in detail. We first consider four test matrices studied previously in [16, section 6.6]:

1. $A_1 = I + wv^* \in \mathbb{R}^{8\times 8}$, where $w = \begin{pmatrix}1^2 & 2^2 & \dots & 8^2\end{pmatrix}^*$ and $v = \begin{pmatrix}0^2 & 1^2 & 2^2 & \dots & 7^2\end{pmatrix}^*$.
2. $A_2 =$ `gallery('moler',16)` $\in \mathbb{R}^{16\times 16}$.
3. $A_3 =$ `Q*rschur(8,2e2)*Q'` $\in \mathbb{R}^{8\times 8}$, where `Q=gallery('orthog',8)` and `rschur` is a function from the Matrix Computation Toolbox [13].
4. $A_4 =$ `gallery('chebvand',16)` $\in \mathbb{R}^{16\times 16}$.

[2] Note that equation (6.30) in [16] contains a typo in the last line: $E_{k+1} = -\frac{1}{2}E_k X_{k+1}^{-1} E_k$ should read $E_{k+1} = -\frac{1}{2}\widetilde{E}_k X_{k+1}^{-1}\widetilde{E}_k$.

TABLE 1
Properties of the matrices A_1, A_2, A_3, and A_4.

	A_1	A_2	A_3	A_4
$\alpha_\infty(A^{1/2})$	1.4e0	1.1e0	1.4e8	2.8e0
$\kappa_{\text{sqrt}}(A)$	4.0e1	8.3e4	5.7e7	5.2e6
$\kappa_2(A^{1/2})$	8.0e1	2.0e5	3.1e10	3.9e6

TABLE 2
Numerical results for A_1 (upper left), A_2 (upper right), A_3 (lower left), and A_4 (lower right). Each table shows the number of iterations k, relative error $\|\hat{X} - A^{1/2}\|_\infty / \|A^{1/2}\|_\infty$, and relative residual $\|\hat{X}^2 - A\|_\infty / \|A\|_\infty$ in the computed square root $\hat{X}$ of A.

Method	k	Err.	Res.
DB	9	4.6e-15	8.2e-15
DBp	9	1.1e-14	1.0e-14
IN	9	1.3e-14	2.4e-14
P-(1, 0)	9	1.3e-14	2.7e-14
P-(4, 4)	4	2.2e-15	4.7e-15
P-(8, 8)	3	3.0e-15	5.4e-15
Z-(1, 0)	6	3.2e-15	6.6e-15
Z-(4, 4)	2	1.6e-15	1.6e-15
Z-(8, 8)	2	3.0e-15	6.0e-15
sqrtm	0	2.8e-15	6.9e-16

Method	k	Err.	Res.
DB	14	8.8e-10	5.5e-10
DBp	14	1.4e-10	4.7e-11
IN	14	4.9e-14	4.0e-16
P-(1, 0)	15	7.1e-13	4.8e-13
P-(4, 4)	6	1.5e-13	1.1e-13
P-(8, 8)	5	3.2e-13	1.8e-13
Z-(1, 0)	8	3.4e-13	3.2e-13
Z-(4, 4)	3	1.8e-13	1.3e-13
Z-(8, 8)	2	7.4e-13	4.6e-13
sqrtm	0	9.3e-13	3.1e-15

Method	k	Err.	Res.
DB	7	6.6e-7	3.6e-4
DBp	6	7.6e-7	7.6e-3
IN	7	1.4e-4	4.1e-1
P-(1, 0)	8	6.3e-7	2.2e-4
P-(4, 4)	4	3.8e-7	1.2e-5
P-(8, 8)	3	2.8e-7	1.6e-6
Z-(1, 0)	7	2.4e-7	5.9e-5
Z-(4, 4)	4	2.9e-7	2.6e-5
Z-(8, 8)	3	2.8e-8	8.3e-7
sqrtm	0	1.2e-9	1.5e-8

Method	k	Err.	Res.
DB	13	9.3e-8	1.2e-7
DBp	12	5.8e-7	3.9e-7
IN	12	4.7e-12	2.5e-14
P-(1, 0)	13	1.2e-10	1.4e-10
P-(4, 4)	6	5.5e-11	6.1e-11
P-(8, 8)	5	1.1e-10	5.8e-11
Z-(1, 0)	11	1.9e-10	2.0e-10
Z-(4, 4)	4	1.9e-10	1.8e-10
Z-(8, 8)	3	2.4e-10	2.4e-10
sqrtm	0	8.9e-11	2.4e-15

Table 1 lists some basic information about these matrices, including

- the condition number of the ∞-norm relative residual of $A^{1/2}$ [16, equation (6.4)],
$$\alpha_\infty(A^{1/2}) = \frac{\|A^{1/2}\|_\infty^2}{\|A\|_\infty};$$
- the Frobenius-norm relative condition number of the matrix square root at A [16, equation (6.2)],
$$\kappa_{\text{sqrt}}(A) = \frac{\|(I \otimes A^{1/2} + (A^{1/2})^T \otimes I)^{-1}\|_2 \|A\|_F}{\|A^{1/2}\|_F};$$
- the 2-norm condition number of $A^{1/2}$,
$$\kappa_2(A^{1/2}) = \|A^{1/2}\|_2 \|A^{-1/2}\|_2.$$

Table 2 reports the number of iterations k, relative error $\|\hat{X} - A^{1/2}\|_\infty / \|A^{1/2}\|_\infty$, and relative residual $\|\hat{X}^2 - A\|_\infty / \|A\|_\infty$ in the computed square root $\hat{X}$ of A for

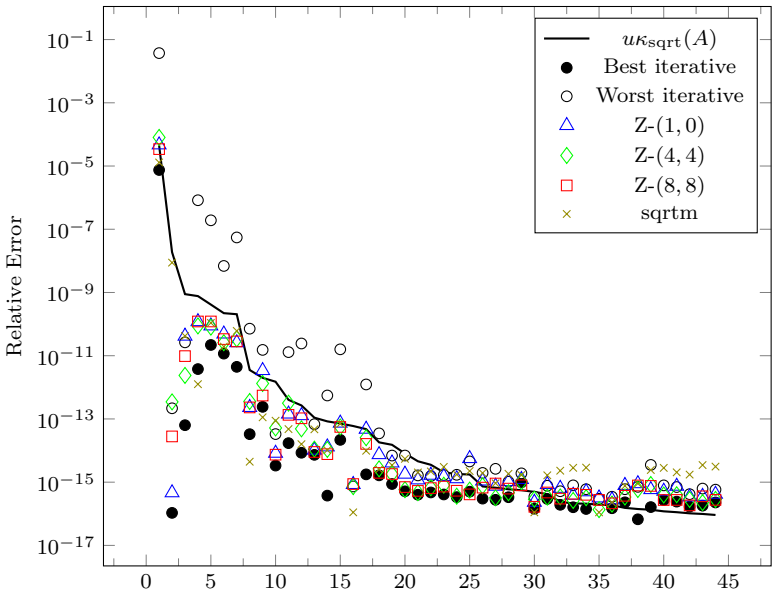

FIG. 3. *Relative errors committed by each method on* 44 *tests, ordered by decreasing condition number* $\kappa_{\mathrm{sqrt}}(A)$.

each method. (We computed the "exact" $A^{1/2}$ using variable precision arithmetic in MATLAB: `vpa(A,100)^(1/2)`.) In these tests, the Zolotarev and Padé iterations of a given type (m, ℓ) tended to produce comparable errors and residuals, but the Zolotarev iterations almost always took fewer iterations to do so. With the exception of A_3, the Zolotarev, Padé, and IN iterations achieved forward errors less than or comparable to the MATLAB function `sqrtm`. On A_3, `sqrtm` performed best, but it is interesting to note that the type $(8, 8)$ Zolotarev iteration produced the smallest forward error and smallest residual among the iterative methods.

Additional tests. We performed tests on an additional 44 matrices from the Matrix Function Toolbox [13], namely those matrices in the toolbox of size 10×10 having 2-norm condition number $\kappa_2(A) \le u^{-1}$, where $u = 2^{-53}$ is the unit roundoff. For each matrix A, we rescaled A by $e^{i\theta}$ if A had any negative real eigenvalues, with θ a random number between 0 and 2π.

Figure 3 shows the relative error $\|\hat{X} - A^{1/2}\|_\infty / \|A^{1/2}\|_\infty$ committed by each method on the 44 tests, ordered by decreasing condition number $\kappa_{\mathrm{sqrt}}(A)$. To reduce clutter, the results for the non-Zolotarev iterations (DB, DBp, IN, P-(1,0), P-(4,4), and P-(8,8)) are not plotted individually. Instead, we identified in each test the smallest and largest relative errors committed among the DB, DBp, IN, P-(1,0), P-(4,4), and P-(8,8) iterations and plotted these minima and maxima (labeled "Best iterative" and "Worst iterative" in the legend). In almost all tests, the Zolotarev iterations achieved relative errors less than or comparable to $u\kappa_{\mathrm{sqrt}}(A)$. In addition, the Zolotarev iterations tended to produce relative errors closer to the best of the non-Zolotarev iterations than the worst of the non-Zolotarev iterations.

Table 3 summarizes the number of iterations used by each method in these tests. The table reveals that on average, the Zolotarev iteration of type (m, ℓ) converged more quickly than the Padé iteration of type (m, ℓ) for each $(m, \ell) \in \{(1, 0), (4, 4), (8, 8)\}$.

TABLE 3
Number of iterations used by each method in the tests appearing in Figure 3.

Method	Mean	STD	Min	Max
DB	7.4	2.1	3	12
DBp	7.3	2.2	3	12
IN	7.7	2.8	3	20
P-(1, 0)	7.7	2.4	3	13
P-(4, 4)	3.3	1.2	2	6
P-(8, 8)	2.8	1	2	5
Z-(1, 0)	7.5	2	5	12
Z-(4, 4)	2.8	0.7	2	4
Z-(8, 8)	2.3	0.5	2	3
sqrtm	0	0	0	0

6. Conclusion. We have presented a new family of iterations for computing the matrix square root using recursive constructions of Zolotarev's rational minimax approximants of the square root function. These iterations are closely related to the Padé iterations but tend to converge more rapidly, particularly for matrices that have eigenvalues with widely varying magnitudes. The favorable behavior of the Zolotarev iterations presented here, together with the favorable behavior of their counterparts for the polar decomposition [24], suggests that other matrix functions like the matrix sign function and the matrix pth root may stand to benefit from these types of iterations [9].

Acknowledgments. I thank an anonymous reviewer for suggesting, among other things, a way to significantly improve the presentation in sections 3.1–3.2. I also wish to thank Yuji Nakatsukasa for introducing me to this topic and for sharing his code for computing the coefficients of Zolotarev's functions.

REFERENCES

[1] N. I. AKHIEZER, *Theory of Approximation*, Frederick Ungar, New York, 1956.
[2] N. I. AKHIEZER, *Elements of the Theory of Elliptic Functions*, Transl. Math. Managr. 79, AMS, Providence, RI, 1990.
[3] B. BECKERMANN, *Optimally scaled Newton iterations for the matrix square root*, presented at Advances in Matrix Functions and Matrix Equations Workshop, Manchester, UK, 2013.
[4] B. BECKERMANN AND A. TOWNSEND, *On the singular values of matrices with displacement structure*, SIAM J. Matrix Anal. Appl., 38 (2017), pp. 1227–1248.
[5] D. BRAESS, *Nonlinear Approximation Theory*, Springer Ser. Comput. Math., Springer, New York, 1986.
[6] 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 (2008), pp. 822–843.
[7] E. D. DENMAN AND A. N. BEAVERS, JR., *The matrix sign function and computations in systems*, Appl. Math. Comput., 2 (1976), pp. 63–94.
[8] F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, and B. V. Saunders, eds., *NIST Digital Library of Mathematical Functions*, https://dlmf.nist.gov/22.5#T3.
[9] E. S. GAWLIK, *Rational Minimax Iterations for Computing the Matrix pth Root*, preprint, arXiv:1903.06268, 2019.
[10] E. S. GAWLIK, Y. NAKATSUKASA, AND B. D. SUTTON, *A backward stable algorithm for computing the CS decomposition via the polar decomposition*, SIAM J. Matrix Anal. Appl., 39 (2018), pp. 1448–1469.
[11] S. GUTTEL, E. POLIZZI, P. T. P. TANG, AND G. VIAUD, *Zolotarev quadrature rules and load balancing for the FEAST eigensolver*, SIAM J. Sci. Comput., 37 (2015), pp. A2100–A2122.

[12] N. Hale, N. J. Higham, and L. N. Trefethen, *Computing A^α, $\log(A)$, and related matrix functions by contour integrals*, SIAM J. Numer. Anal., 46 (2008), pp. 2505–2523.
[13] N. J. Higham, *The Matrix Computation Toolbox*, https://www.ma.man.ac.uk/~higham/mctoolbox.
[14] N. J. Higham, *Newton's method for the matrix square root*, Math. Comput., 46 (1986), pp. 537–549.
[15] N. J. Higham, *Stable iterations for the matrix square root*, Numer. Algorithms, 15 (1997), pp. 227–242.
[16] N. J. Higham, *Functions of Matrices: Theory and Computation*, SIAM, Philadelphia, 2008.
[17] N. J. Higham, D. S. Mackey, N. Mackey, and F. Tisseur, *Functions preserving matrix groups and iterations for the matrix square root*, SIAM J. Matrix Anal. Appl., 26 (2005), pp. 849–877.
[18] L. Hogben, *Handbook of Linear Algebra*, CRC Press, Boca Raton, FL, 2016.
[19] B. Iannazzo, *A note on computing the matrix square root*, Calcolo, 40 (2003), pp. 273–283.
[20] D. Kressner and A. Susnjara, *Fast computation of spectral projectors of banded matrices*, SIAM J. Matrix Anal. Appl., 38 (2017), pp. 984–1009.
[21] B. Le Bailly and J. Thiran, *Optimal rational functions for the generalized Zolotarev problem in the complex plane*, SIAM J. Numer. Anal., 38 (2000), pp. 1409–1424.
[22] Y. Li and H. Yang, *Spectrum Slicing for Sparse Hermitian Definite Matrices Based on Zolotarev's Functions*, preprint, arXiv:1701.08935, 2017.
[23] B. Meini, *The matrix square root from a new functional perspective: Theoretical results and computational issues*, SIAM J. Matrix Anal. Appl., 26 (2004), pp. 362–376.
[24] Y. Nakatsukasa and R. W. Freund, *Computing fundamental matrix decompositions accurately via the matrix sign function in two iterations: The power of Zolotarev's functions*, SIAM Rev., 58 (2016), pp. 461–493.
[25] I. Ninomiya, *Best rational starting approximations and improved Newton iteration for the square root*, Math. Comput., 24 (1970), pp. 391–404.
[26] J. Nocedal and S. Wright, *Numerical Optimization*, Springer, New York, 2006.
[27] H. Rutishauser, *Betrachtungen zur quadratwurzeliteration*, Monat. Math., 67 (1963), pp. 452–464.
[28] J. Todd, *Applications of transformation theory: A legacy from Zolotarev (*1847–1878*)*, in Approximation Theory and Spline Functions, Springer, New York, 1984, pp. 207–245.
[29] L. N. Trefethen and M. H. Gutknecht, *On convergence and degeneracy in rational Padé and Chebyshev approximation*, SIAM J. Math. Anal., 16 (1985), pp. 198–210.
[30] E. Wachspress, *Positive Definite Square Root of a Positive Definite Square Matrix*, unpublished manuscript, 1962.
[31] E. Wachspress, *The ADI Model Problem*, Springer, New York, 2013.
[32] E. I. Zolotarev, *Applications of elliptic functions to problems of functions deviating least and most from zero*, Zap. S.-Petersburg Akad. Nauk., 30 (1877), pp. 1–59.