SIAM J. Matrix Anal. Appl.
Vol. 28, No. 2, pp. 503–523

ON THE NEWTON METHOD FOR THE MATRIX *P*TH ROOT*

BRUNO IANNAZZO[†]

Abstract. Stable versions of Newton's iteration for computing the principal matrix pth root $A^{1/p}$ of an $n \times n$ matrix A are provided. In the case in which X_0 is the identity matrix, it is proved that the method converges for any matrix A having eigenvalues with modulus less than 1 and with positive real parts. Based on these results we provide a general algorithm for computing the principal pth root of any matrix A having no nonpositive real eigenvalues. The algorithm has quadratic convergence, is stable in a neighborhood of the solution, and has a cost of $O(n^3 \log p)$ operations per step. Numerical experiments and comparisons are performed.

Key words. matrix pth root, Newton's method, numerical stability, matrix iterations, rational iterations, Julia set

AMS subject classifications. 15A24, 65F30

DOI. 10.1137/050624790

1. Introduction.

A useful tool for solving nonlinear equations is the Newton method,

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$$

for an initial value x_0. For the algebraic equation $x^p - a = 0$, $a \in \mathbb{C}$, it turns into

$$x_{k+1} = \frac{(p-1)x_k + a x_k^{1-p}}{p}. \tag{1.1}$$

As pointed out by Cayley in 1879 [4], the study of the convergence of this iteration is very hard for $p > 2$. In fact the set of initial values such that the iteration (1.1) converges to a specific root is a *beautiful* but *complicated* set, and its boundary is the so-called Julia set of the iteration.

For $A \in \mathbb{C}^{n\times n}$, one can consider the matrix iteration

$$X_{k+1} = \frac{(p-1)X_k + A X_k^{1-p}}{p} \tag{1.2}$$

for solving the matrix equation

$$X^p - A = 0. \tag{1.3}$$

One of the most interesting solutions of (1.3) is the *principal pth root* $A^{1/p}$ of A whose eigenvalues lie in the sector

$$\mathcal{S}_p = \{z \in \mathbb{C}\backslash\{0\}\,,\ -\pi/p < \arg z < \pi/p\}. \tag{1.4}$$

[*]Received by the editors February 19, 2005; accepted for publication (in revised form) by A. Frommer January 26, 2006; published electronically June 21, 2006. This work was supported by MIUR grant 2002014121.

http://www.siam.org/journals/simax/28-2/62479.html

[†]Dipartimento di Matematica "L.Tonelli," Università di Pisa, Largo B. Pontecorvo 5, 56127 Pisa, Italy (iannazzo@mail.dm.unipi.it).

If A has no nonpositive real eigenvalues, then there exists a unique principal pth root. Here and hereafter we refer to (1.2) as the *simplified Newton iteration.*

The main applications of the matrix pth root are for the computation of the logarithm of a matrix and the sector function; for other applications, see [7, 11].

Convergence and stability properties of (1.2) are important issues which play a fundamental role in the design of an algorithm for the matrix pth root. Hoskins and Walton [12] and Smith [18] take as initial value the matrix A. Unfortunately, as discussed in [18], this choice leads to a convergence region not nice enough to design a simple global convergent method.

Concerning stability, Higham [8] and Smith [18] have shown that the simplified Newton iteration is unstable. That is, a small perturbation Δ in X_k, say the one generated by roundoff, may lead to divergence of the sequence obtained by replacing X_k by $X_k + \Delta$. Thus divergence may occur even though the computation of X_k is performed with a numerically stable algorithm. This makes the iteration of almost no practical use.

In this paper we present a suitable modification of the simplified Newton iteration which guarantees stability. Moreover we prove that, choosing $X_0 = I$, convergence occurs for any A having eigenvalues in the set $\mathcal{D} = \{z \in \mathbb{C} : \ |z| = 1, \operatorname{Re} z > 0\}$. This restriction can be relaxed by means of a suitable scaling, and we provide an algorithm which converges for any A for which $A^{1/p}$ is defined. The iteration that we obtain in this way has quadratic convergence and a cost per step of $O(n^3 \log p)$ arithmetic operations (ops).

Regarding available algorithms, an efficient numerical method for the principal pth root uses the Schur form and was originally proposed by Björck and Hammarling [3] for the square root, then extended by Higham [9] who suggested using the real Schur form for real matrices, and generalized by Smith [18] to the matrix pth root. This method, implemented in the MATLAB toolbox [10], is numerically stable and requires $O(n^3 p)$ ops [11]. The p factor in the operation count is a drawback for large p, and it is desirable to have methods whose cost grows more slowly with p. An interesting analysis of computing the principal matrix pth root has been performed in [2], where the problem is investigated in terms of structured matrix computations and where the Newton iteration for the equation $X^p - A^{-1}$ is proposed. Other methods can be designed based on the identities $A^{1/p} = \exp(\frac{1}{p} \log A)$, where the functions $\log(\cdot)$ and $\exp(\cdot)$ are the matrix generalizations of the customary log and exp functions, respectively [16].

The paper is organized in the following way. In section 2 we show that for $X_0 = I$, Newton's iteration converges for any matrix A with eigenvalues in $\mathcal{D}$. In section 3 we discuss instability issues and propose new variants of (1.2) which, while keeping the same cost of $O(n^3 \log p)$ ops, are proved to be stable in a neighborhood of the solution. In section 4 we describe our general algorithm and discuss some related computational issues. Finally in section 5 we present some numerical experiments and compare our method with the Schur method and with the method based on logarithm and exponential. These results confirm the numerical stability and the overall good performance of the new algorithms.

In the rest of the paper we use the notation $\pi/2p$ instead of $\pi/(2p)$ for the sake of readability.

Remark 1. It was observed in [12, 18] that if A has no nonpositive real eigenvalues and if X_0 commutes with A, then the iterates generated by (1.2) coincide with the ones generated by the Newton method in the Banach algebra of the matrices $n \times n$

for the equation $F(X) = X^p - A = 0$; that is

$$X_{k+1} = X_k - F'^{-1}_{X_k}\left(F(X_k)\right), \tag{1.5}$$

provided that the X_k are well defined. The symbol F'_{X_k} here denotes the Fréchet derivative computed at the point X_k. Unfortunately, even if the Fréchet derivative is nonsingular in a neighborhood of $A^{1/p}$, for some choice of A and X_0 the Newton method (1.5) may break down while the simplified one (1.2) still can be applied. See, for instance, [11].

For this reason we will not consider the general theory of the Newton method in Banach algebras, but only the theory of rational iterations. In fact, this approach is easily generalizable to root-finding algorithms different from the Newton method.

2. Convergence. For $p > 2$ rational iterations such as (1.1) have a complicated behavior [14], and it is very difficult to describe the set of initial values for which the iteration converges to a root. The matrix case has a similar behavior; indeed it can be reduced to the scalar one.

Our goal is to determine the set of $A \in \mathbb{C}^{n\times n}$ for which Newton's iteration converges to $A^{1/p}$ for an initial value X_0. The *usual* choice $X_0 = A$ [12, 18] gives a complicated convergence region; here we show that with $X_0 = I$ the convergence region is more suitable for designing a globally convergent algorithm.

First, we consider A diagonalizable, i.e., $A = M^{-1}DM$ with D diagonal and M nonsingular. The general case has similar behavior and will be discussed later. Since $X_0 = I$, all the iterates are diagonalizable and we may define $D_k = MX_kM^{-1}$ so that (1.2) becomes

$$D_{k+1} = \frac{(p-1)D_k + DD_k^{1-p}}{p}, \tag{2.1}$$

which involves only diagonal matrices, and is essentially n uncoupled scalar iterations of the type

$$\begin{cases} x_{k+1} = \dfrac{(p-1)x_k + \lambda x_k^{1-p}}{p}, \\ x_0 = 1, \end{cases} \tag{2.2}$$

with λ being an eigenvalue of A.

Thus our main problem is to determine the set $\mathcal{B}_p$ of λ such that the iteration (2.2) with $x_0 = 1$ is well defined and converges to the principal pth root $\lambda^{1/p}$, i.e., a pth root of λ whose argument lies in the sector $\mathcal{S}_p$ of (1.4).

For any diagonalizable matrix A having eigenvalues in $\mathcal{B}_p$, the Newton iteration, with $X_0 = I$, converges to $A^{1/p}$. It is not surprising that the sets $\mathcal{B}_p$, for $p > 2$, are bounded by fractals similar to the Julia set of Newton's iteration.

Some of these sets are sketched in Figure 2.1, in which we made a grid of 400×400 points corresponding to a discretization $\widehat{Q}$ of the square

$$Q = \{z \in \mathbb{C}, -3 \le \operatorname{Re} z \le 3, -3 \le \operatorname{Im} z \le 3\}$$

and computed some steps of the Newton sequence (2.2) for $\lambda \in \widehat{Q}$. We plotted in light gray the points λ for which the sequence x_k *converges* to the principal pth root of λ, and in dark gray the others.

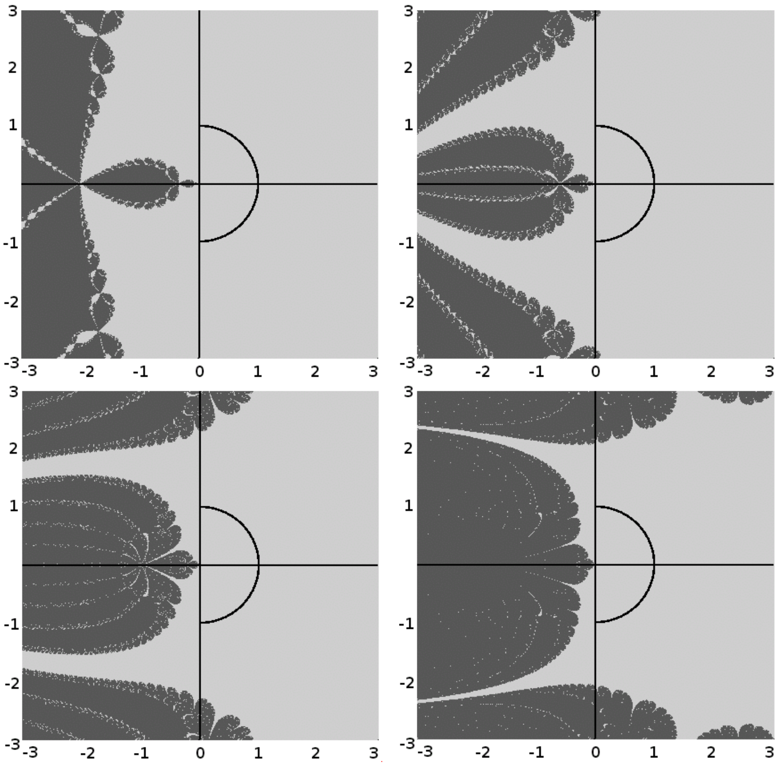

FIG. 2.1. *A sketch of the region of convergence for* $p = 3, 5, 10, 100$ *and the set* $\mathcal{D}$ *of* (2.3)*. In light gray the points for which iteration* (2.2) *converges to their principal pth root.*

It is easy to show, by means of standard arguments on the real Newton method, that the positive real axis belongs to $\mathcal{B}_p$ for every p.

The following theorem synthesizes our main convergence result.

THEOREM 2.1. *The set* $\mathcal{B}_p$ *contains the set*

$$\mathcal{D} = \{z \in \mathbb{C}\,, \operatorname{Re} z > 0, |z| \leq 1\} \tag{2.3}$$

for every $p > 1$.

Consequently if A has its eigenvalues in the set $\mathcal{D}$, then the iteration (1.2), with initial value $X_0 = I$, is well defined and converges to $A^{1/p}$.

For a general matrix A with no nonpositive real eigenvalues, the normalized matrix square root $B = A^{1/2}/\|A^{1/2}\|$, where $\|\cdot\|$ is a generic matrix operator norm, has eigenvalues in the set $\mathcal{D}$. In fact for the spectral radius of B one has $\rho(B) \leq \|B\| = 1$ and since the spectrum of $A^{1/2}$ belongs to the right half-plane, the spectrum of B belongs to the set $\mathcal{D}$. Thus the Newton method applied to the matrix equation $X^p - B = 0$, starting with $X_0 = I$, converges to $B^{1/p}$. Moreover, it is possible to recover $A^{1/p} = (B^{1/p})^2$.

To prove Theorem 2.1, we use the following property.

PROPOSITION 2.2. *Let* λ *be a complex number with no nonpositive real part.*

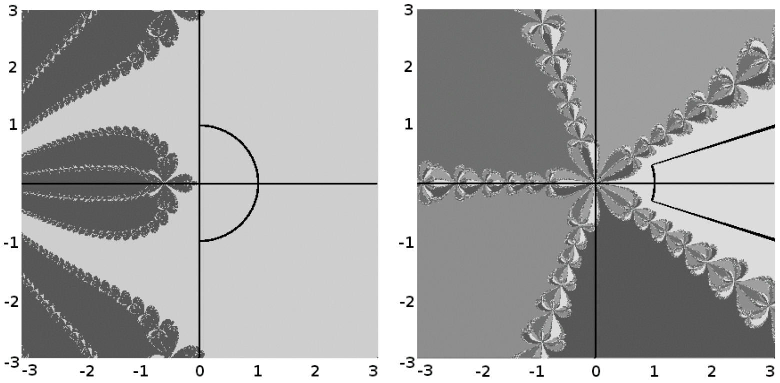

FIG. 2.2. *For $p = 5$ the set $\mathcal{D}$ of Theorem* 2.1 *(left) and the set $\mathcal{D}_5$ of Theorem* 2.3 *(right).*

Then the sequence (2.2) *converges to $\lambda^{1/p}$ if and only if the sequence*

$$\begin{cases} z_{k+1} = \dfrac{(p-1)z_k + z_k^{1-p}}{p}, \\ z_0 = \lambda^{-1/p} \end{cases} \tag{2.4}$$

converges to 1.

Proof. The proof follows from the equation $z_k = x_k \lambda^{-1/p}$, which can be proved by induction. □

Observe that (2.4) is the Newton method applied to the equation $x^p - 1 = 0$. A similar *trick* was used in [2]. The above property provides a connection between the set $\mathcal{B}_p$ and the basin of attraction of the root $x = 1$, which we denote by $\mathcal{A}_p(1)$. In fact, a complex number $a \neq 0$ belongs to $\mathcal{B}_p$ if and only if $a^{-1/p}$ belongs to $\mathcal{A}_p(1) \cap \mathcal{S}_p$.

In this way, we can restate Theorem 2.1 in the following form.

THEOREM 2.3. *The set $\mathcal{A}_p(1)$ contains the set $\mathcal{D}_p = \{z \in \mathcal{S}_{2p}, |z| \geq 1\}$ for every $p > 1$, where $\mathcal{S}_p$ is defined in* (1.4).

A graphical example of the swap between the two theorems is given in Figure 2.2.

2.1. Proof of Theorem 2.3.

Define

$$N_p(z) = \frac{(p-1)z^p + 1}{pz^{p-1}} \tag{2.5}$$

for the Newton step and denote by $N_p^{(k)}$ the k-fold composition $N \circ N \circ \cdots \circ N$. Observe also that the function $N_p(z)$ is well defined in $\mathcal{D}_p$.

The proof can be divided into two stages. First, we show that Theorem 2.3 holds if two inequalities are satisfied. Second, we show the validity of such inequalities.

We consider three sets depending on the positive values ξ_p and R_p (see Figure 2.3):

1. a *disk* $E_p = \{z \in \mathbb{C}, |z - 1| < R_p\}$ of center 1 and radius R_p;
2. a *mincing knife*, $F_p = \{z \in \mathbb{C}, 1 \leq |z| < \xi_p, |\arg(z)| \leq \pi/2p\}$;
3. a *blunt wedge*, $G_p = \{z \in \mathbb{C}, |z| \geq \xi_p, |\arg(z)| \leq \pi/2p\}$.

We provide an algebraic equation with real solution $s_p = 1 - R_p$ and such that the disk E_p is contained in $\mathcal{A}_p(1)$; then we provide a second algebraic equation with real solution ξ_p and such that each point of the set G_p is transformed by $N_p^{(k)}$ into a point in F_p for some $k \geq 1$. Finally we show that given a point z in F_p, the supremum

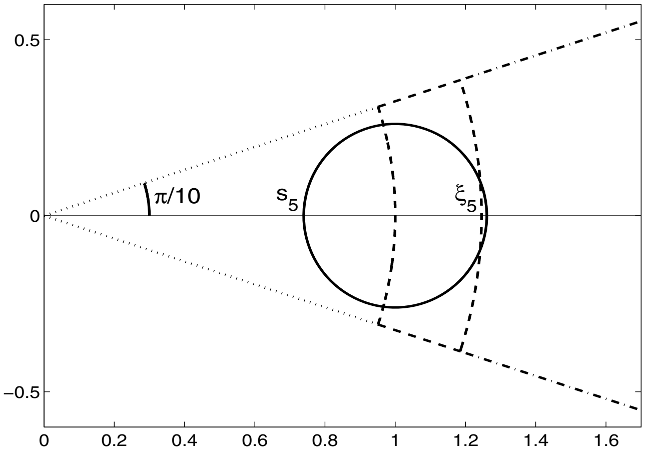

FIG. 2.3. *The three sets used in the proof of Theorem* 2.3 *for the case* $p = 5$*: the circle of radius* $1 - s_5$*, the mincing knife (dash contour), and the blunt wedge (dash-dot contour).*

of the distance of $N_p(z)$ from 1 is reached in the corners of the mincing knife. So, in order to prove that the points in F_p are transformed into points in E_p, it is enough to compute $|N_p(z)-1|$ in the corners of F_p and prove that these values are less than R_p. These are the desired inequalities. In fact, by verifying such inequalities for a specific value of p, one can easily verify that $\mathcal{D}_p \subset F_p \cup G_p$ is a subset of $\mathcal{A}_p(1)$, which is the statement of Theorem 2.3.

We start by giving a way to find a disk centered at the point $z = 1$, such that Newton's iteration converges if x_0 is in this disk.

LEMMA 2.4. *The equation*

$$(2p-1)s^p - 2ps^{p-1} + 1 = 0 \tag{2.6}$$

has a unique real solution s_p *in the open interval* $(0, 1)$ *and for every* z *such that* $|z-1| < R_p = 1 - s_p$ *and* $z \neq 1$*, it holds that* $|N_p(z) - 1| < |z - 1|$*.*

Proof. Let us start from inequality $|N_p(z) - 1| < |z - 1|$, namely

$$\left|\frac{(p-1)z^p - pz^{p-1} + 1}{pz^{p-1}}\right| < |z-1|. \tag{2.7}$$

The polynomial $\phi_p(z) = (p-1)z^p - pz^{p-1} + 1$ can be factorized as $\phi_p(z) = ((p-1)z^{p-2} + \cdots + 3z^2 + 2z + 1)(z-1)^2$, and the inequality (2.7) becomes

$$\frac{|(p-1)z^{p-2} + \cdots + 3z^2 + 2z + 1||z-1|}{|pz^{p-1}|} < 1, \quad z \neq 1. \tag{2.8}$$

Now,

$$\frac{|(p-1)z^{p-2} + \cdots + 2z + 1||z-1|}{|pz^{p-1}|} \leq \frac{1}{p}\left(\frac{p-1}{|z|} + \cdots + \frac{2}{|z|^{p-2}} + \frac{1}{|z|^{p-1}}\right)|z-1|.$$

If $|z-1| < 1-s$, then $|z|^n > s^n$ for every n and the inequality (2.7) holds if

$$\frac{1}{p}\left(\frac{p-1}{s} + \cdots + \frac{2}{s^{p-2}} + \frac{1}{s^{p-1}}\right)(1-s) - 1 \leq 0. \tag{2.9}$$

Multiplying both sides of the above inequality by $ps^{p-1}(1-s)$, with $0 < s < 1$, yields $\phi_p(s) - ps^{p-1}(1-s) \leq 0$, that is,

$$(2p-1)s^p - 2ps^{p-1} + 1 \leq 0. \tag{2.10}$$

It is not difficult to show that the function $f_p(s) = (2p-1)s^p - 2ps^{p-1} + 1$ has the following properties: $f_p(0) > 0$, $f_p(1) = 0$, $f_p'(1) > 0$, and f_p has only a relative minimum in the interval $(0, 1)$. All these facts guarantee that the equation $f_p(s) = 0$ has a unique solution s_p in the interval $(0, 1)$ and that the inequality (2.10) holds for every $s_p \leq s \leq 1$.

To conclude, we recall that $|z-1| < 1-s$, and so for $0 < |z-1| < R_p = 1 - s_p$, it holds that $|N_p(z) - 1| < |z-1|$, which was what we wanted to show. Moreover, we have a constructive way to find R_p by solving the polynomial equation of (2.10) in the interval $(0, 1)$. □

This lemma guarantees that for each positive real value $R < R_p$ the closed disk of center 1 and radius R belongs to $\mathcal{A}_p(1)$. Moreover, it holds that $|N_p(s_p) - 1| = |s_p - 1|$ and then Lemma 2.4 is not true for any value $R > R_p$.

In order to prove that the set $\mathcal{D}_p$ is a subset of $\mathcal{A}_p(1)$, we split $\mathcal{D}_p$ into two subsets. The former is sent by N_p into the disk of convergence found above, and the latter is sent into the former after some iterations. First, we give a technical lemma that states that any point of a blunt wedge is transformed by N_p into a point of the wedge. This will be used to show that a point of modulus greater than 1 gets closer to 1, after some iterations, but still remains in the sector.

Lemma 2.5. *If $|z| > 1$ and $z \in \mathcal{S}_{2p}$, then $|N_p(z)| < |z|$ and $|\arg(N_p(z))| \leq |\arg(z)|$.*

Proof. For the first statement, it is easy to show that $|z| > 1$ yields

$$|N_p(z)| = \left| \frac{(p-1)}{p} z + \frac{1}{pz^{p-1}} \right| \leq \frac{(p-1)}{p}|z| + \frac{1}{p|z|^{p-1}} < \frac{(p-1)}{p}|z| + \frac{1}{p} < |z|.$$

For the second statement, let $z = re^{i\theta}$ with $r > 1$ and $0 < |\theta| < \pi/2p$; moreover, let $N(z) = r_1 e^{i\theta_1}$. Our goal is to prove that $|\theta_1| \leq |\theta|$, which is equivalent to $|\tan\theta_1| \leq |\tan\theta|$. From the definition of N_p it can be shown that

$$\tan\theta_1 = \frac{r^p(p-1)\sin\theta - \sin((p-1)\theta)}{r^p(p-1)\cos\theta + \cos\left((p-1)\theta\right)}$$

so that inequality $|\tan\theta_1| \leq |\tan\theta|$, for $\theta > 0$, becomes

$$-\frac{\sin\theta}{\cos\theta} \leq \frac{r^p(p-1)\sin\theta - \sin((p-1)\theta)}{r^p(p-1)\cos\theta + \cos((p-1)\theta)} \leq \frac{\sin\theta}{\cos\theta}. \tag{2.11}$$

By means of trigonometric identities, the second inequality of (2.11) is equivalent to $\sin(p\theta) \geq 0$, which is true because $0 < \theta < \pi/2p$.

The first inequality (2.11) is equivalent to

$$r \geq \sqrt[p]{\frac{\sin((p-2)\theta)}{(p-1)\sin(2\theta)}},$$

which is true in the region we have considered, where $r > 1$, and

$$\frac{\sin((p-2)\theta)}{(p-1)\sin(2\theta)} < 1.$$

The case $\theta < 0$ is analogous by symmetry, and the case $\theta = 0$ is trivial. □

Even if a point of the sector having modulus greater than a real number $R > 1$ is transformed by the Newton step N_p into another point in the sector, we need to cut the wedge enough so that each point of the blunt wedge is transformed by N_p into a point of modulus greater than 1. In the next lemma, we find a real value ξ_p that satisfies this condition and is the least in modulus.

Lemma 2.6. *The equation*

$$(p-1)^2 s^{2p} - p^2 s^{2p-2} + 1 = 0 \tag{2.12}$$

has a unique real solution $1 < \xi_p < 2$. *For every* $z \in \mathcal{S}_{2p}$ *such that* $|z| > \xi_p$, *it holds that* $|N(z)| > 1$.

Proof. Let $R \geq 1$ and let us consider the set $K = \{z \in \mathbb{C},\ |z| \geq R, |\arg(z)| \leq \pi/2p\}$. The minimum of $|N_p(z)|$ on the set K is attained at the point $z_0 = R\mathrm{e}^{i\pi/2p}$. In order to prove this, let $z = r\mathrm{e}^{i\theta}$ and consider $|N_p(z)|$ as a function of θ. We have

$$f_r(\theta) = |N_p(z)| = \left|\frac{(p-1)z^p + 1}{pz^{p-1}}\right| = \frac{1}{pr^{p-1}}|(p-1)r^p\mathrm{e}^{ip\theta} + 1|.$$

Observe that $f_r(\theta)$ is minimum for $\theta = \pi/2p$. Moreover, since

$$g(r) = |N_p(r\mathrm{e}^{i\pi/2p})|^2 = \frac{(p-1)^2 r^{2p} + 1}{(pr^{p-1})^2}$$

is increasing for $r > 1$, we deduce that the minimum of $|N_p(z)|$ is attained at the corners of K and in particular at the point z_0. Now, in order to prove that $|N_p(z)| \geq 1$, we solve the equation $|N_p(s\mathrm{e}^{i\pi/2p})| = 1$, that is,

$$\frac{\sqrt{(p-1)^2 s^{2p} + 1}}{ps^{p-1}} = 1,$$

which yields (2.12). Now, it is not difficult to show that the function $g_p(s) = (p-1)^2 s^{2p} - p^2 s^{2p-2} + 1$ in (2.12) has the following properties: $g_p(1) < 0$, $g_p'(1) < 0$, $g_p(2) > 0$, and g_p has only a critical point (a minimum) in the interval $(1, 2)$. All these facts guarantee that (2.12) has a unique solution ξ_p in the interval $(1, 2)$. Therefore, if $|z| \geq \xi_p$, it holds that $|N_p(z)| \geq |N_p(\xi_p\mathrm{e}^{i\pi/2p})| = 1$ and this completes the proof. □

From Lemmas 2.5 and 2.6 we can conclude that a point of the set $\mathcal{D}_p$ having modulus greater than ξ_p is sent, after some iterations, into a point of $\mathcal{D}_p$ having modulus less than ξ_p.

Now, if the mincing knife

$$F_{p,R} = \{z \in \mathbb{C},\ 1 \leq |z| \leq R,\ |\mathrm{arg} z| \leq \pi/2p\}, \tag{2.13}$$

with $R = \xi_p$, is sent into the ball $|N_p(z) - 1| < R_p$, then the theorem is true.

In the next lemma, we show that the maximum of the function $|N_p(z) - 1|$ on the mincing knife is attained at one of the corners and so it is enough to check if these four points are sent into the ball of convergence (for the symmetry of the problem we need to check only two of them).

Lemma 2.7. *Given a real number* $R > 1$, *the function* $f(z) = |N_p(z) - 1|$ *defined on the set* $F = F_{p,R}$ *of* (2.13) *takes its maximum at one of the corners of* F.

Proof. Let $z = re^{i\theta}$ be a point of F. Observing that $N(\overline{z}) = \overline{N(z)}$, it is enough to consider the case $\theta \geq 0$.

We show that, restricted to the circle of radius $r \geq 1$, the function f is nondecreasing with respect to θ (nonnegative), and hence the maximum lies on the segment corresponding to $\theta = \pi/2p$, $1 \leq r \leq R$. Then, we show that the function is convex in this segment and then takes its maximum at one of the two vertices, which are the top corners.

To simplify the problem, consider the function

$$\begin{aligned}&\hat{f}(r,\theta) = p^2|N_p(z) - 1|^2 - p^2\\ &= (p-1)^2 r^2 + \frac{1}{r^{2p-2}} + \frac{2(p-1)}{r^{p-2}}\cos(p\theta) - 2(p^2-p)r\cos(\theta) - \frac{2p}{r^{p-1}}\cos((p-1)\theta),\end{aligned}$$

which has the same point of maximum of $|N_p(z) - 1|$ and is simpler.

First, consider the restriction of $\hat{f}$ to an arc relative to a fixed value of r and study the behavior with respect to θ.

Define $g_r(\theta) = \hat{f}(r,\theta)$. We prove that $g_r(\theta)$ is nondecreasing by showing that its derivative,

$$g_r'(\theta) = \frac{2p(p-1)}{r^{p-1}}\left(\sin((p-1)\theta) - r\sin(p\theta) + r^p\sin(\theta)\right),$$

is nonnegative for $0 \leq \theta \leq \pi/2p$. From the sine addition formula, one has

$$\begin{aligned}(2.14)\quad &\sin((p-1)\theta) - r\sin(p\theta) + r^p\sin(\theta)\\ &\qquad = \sin((p-1)\theta)(1 - r\cos(\theta)) + r\sin(\theta)(-\cos((p-1)\theta) + r^{p-1}) \geq 0,\end{aligned}$$

and the last inequality follows from

$$\frac{r\cos(\theta) - 1}{r(r^{p-1} - \cos((p-1)\theta))} \leq \frac{r-1}{r(r^{p-1}-1)} \leq \frac{1}{r\sum_{k=0}^{p-2} r^k} \leq \frac{1}{p-1} \leq \frac{\sin(\theta)}{\sin((p-1)\theta))},$$

where we used the fact that $r \geq 1$ and that the inequality $\sin(n\theta) \leq n\sin(\theta)$ holds for any positive integer n and $0 \leq \theta \leq \pi/2p$.

The inequality (2.14) implies that $g_r(\theta)$ is nondecreasing for any $r \geq 1$, and then the maximum of $\hat{f}$ (and of f) is assumed on the segment of F corresponding to $\theta = \pi/2p$.

Consider the function $\varphi(r) = f(r, \pi/2p)$ on the interval $[1, R]$. We claim that $\varphi(r)$ is a convex function, namely $\varphi''(r) \geq 0$. Since $\cos(p\frac{\pi}{2p}) = 0$ and $\cos((p-1)\frac{\pi}{2p}) = \sin(\pi/2p)$, it holds that

$$\varphi(r) = (p-1)^2 r^2 + \frac{1}{r^{2p-2}} - \frac{2p}{r^{p-1}}\sin(\pi/2p) - 2p(p-1)r\cos(\pi/2p).$$

For its second derivative it holds that

$$\begin{aligned}\varphi''(r) &= 2(p-1)^2 + \frac{2(p-1)(2p-1)}{r^{2p}} - \frac{2p^2(p-1)}{r^{p+1}}\sin(\pi/2p)\\ &\geq 2(p-1)^2 + \frac{2(p-1)(2p-1)}{r^{2p}} - \frac{2(p-1)(2p-1)}{r^{p+1}} = \widetilde{h}(r).\end{aligned}$$

The inequality follows from $p^2\sin(\pi/2p) \leq (2p-1)$ for $p \geq 2$.

Now, $\widetilde{h}(r)$ is positive, and in fact can be rewritten as

$$\frac{2(p-1)}{r^{2p}}\left(r^{p-1}(r^{p+1}(p-1)-2p+1)+2p-1\right) \geq \frac{2(p-1)}{r^{2p}}\left(r^{p+1}(p-1)\right) \geq 0,$$

since $r^{p-1} \geq 1$. The positivity of $\widetilde{h}$ implies that the second derivative of $\varphi(r)$ is positive as well; then the function $\varphi(r)$ is convex so that, restricted to any $[a,b] \subset [1,+\infty)$, it takes its maximum at one of the edges a or b, and the proof is completed. □

Finally we have a procedure to prove that for a value $p > 1$, Theorem 2.3 is true.

- Compute an approximation of R_p and ξ_p by means of some zero-finder method.
- Check if $|N_p(\xi_p \mathrm{e}^{i\pi/2p}) - 1| < R_p$ and $|N_p(\mathrm{e}^{i\pi/2p}) - 1| < R_p$.

To conclude, it is enough to prove that the two inequalities are true for every $p \geq 3$ (the case $p = 2$ is relatively easy and was treated in [8]). We find an explicit expression for a sequence $b_p \leq R_p$ and a sequence $a_p \geq \xi_p$, and then we prove that

$$|N_p(\mathrm{e}^{i\pi/2p}) - 1| < b_p, \quad |N_p(a_p\mathrm{e}^{i\pi/2p}) - 1| < b_p.$$

This is enough; in fact by Lemma 2.7 applied to the set F_{p,a_p}, it holds that

$$|N_p(\xi_p\mathrm{e}^{i\pi/2p}) - 1| \leq |N_p(a_p\mathrm{e}^{i\pi/2p}) - 1| < b_p \leq R_p.$$

We start with a lemma that gives explicitly values for a_p and b_p.

LEMMA 2.8. *The equation* $\mathrm{e}^{-\alpha}(1+2\alpha) - 1 = 0$ *has a unique positive solution* α_0*, and it holds that*

$$\xi_p \leq a_p = \frac{p}{p-1}, \qquad R_p \geq b_p = \frac{\alpha}{p}$$

for every $0 < \alpha \leq \alpha_0$.

Proof. ξ_p is the solution greater than 1 of $g_p(s) = 0$, where $g_p = (p-1)^2 s^{2p} - p^2 s^{2p-2} + 1 = s^{2p-2}((p-1)^2 s^2 - p^2) + 1$. Now, $g_p(\frac{p}{p-1}) = 1 > 0$ and from the arguments in the proof of Lemma 2.6, it follows that $a_p > \xi_p$.

Concerning R_p, let us consider the polynomial $f_p = (2p-1)s^p - 2ps^{p-1} + 1$. The number $s_p = 1 - R_p$ is the solution of the equation $f_p = 0$ and $0 < s_p < 1$. From the proof of Lemma 2.4, proving that $f_p(1 - b_p) < 0$ means that $1 - b_p \geq s_p$ and then $b_p \leq R_p$.

To find a lower bound to R_p of the type α/p, let us consider a generic $0 < \alpha < 3$ that yields

$$f_p\left(1 - \frac{\alpha}{p}\right) = \left(\frac{p-\alpha}{p}\right)^{p-1}\left(\frac{\alpha - (2\alpha+1)p}{p}\right) + 1.$$

In order to have $f_p(1 - \alpha/p) < 0$, it is enough to prove that for every $p > 2$ the sequence

$$d_p = \left(\frac{p-\alpha}{p}\right)^{p-1}\left(\frac{(2\alpha+1)p - \alpha}{p}\right)$$

is greater than 1. This sequence is decreasing for $\alpha > 0$, as we will prove in Lemma 2.9, and its limit is $e^{-\alpha}(1+2\alpha)$. Therefore, $f_p(1-\alpha/p) > 1$ if $e^{-\alpha}(1+2\alpha) > 1$ and this holds for each $0 < \alpha \leq \alpha_0$, where α_0 is the solution in $(0,3)$ of the equation $e^{-\alpha}(1+2\alpha) = 1$. It is easy to prove that this solution exists and is unique and that $\alpha_0 > 1.256$. □

LEMMA 2.9. *The sequence d_p of Lemma* 2.8 *is decreasing.*

Proof. It is sufficient to prove that the function

$$f(x) = \left(\frac{x-\alpha}{x}\right)^{x-1}\left(\frac{(2\alpha+1)x-\alpha}{x}\right)$$

is decreasing for $x \geq 3$. For this purpose we prove that $f'(x)$ is negative. We have $f'(x) = g(x)h(x)$ with $h(x)$ trivially positive and

$$g(x) = \log\left(\frac{x-\alpha}{x}\right) + \frac{(x-1)\alpha}{x(x-\alpha)} + \frac{\alpha}{x((2\alpha+1)x-\alpha)}$$

is negative; in fact it is increasing and its limit to infinity is 0. To prove that $g(x)$ is increasing, it is enough to show that its derivative is positive, which holds by a direct inspection. □

Now we can finally complete the proof of our main theorem by means of the following lemma.

LEMMA 2.10. *The two points of the corners of F_{p,a_p} are sent by the Newton iteration N_p into points in the ball of center* 1 *and radius b_p, i.e.,*

$$\left|N_p(\mathrm{e}^{i\pi/2p}) - 1\right| < \frac{\alpha_0}{p}, \qquad \left|N_p\left(\frac{p}{p-1}\mathrm{e}^{i\pi/2p}\right) - 1\right| < \frac{\alpha_0}{p}.$$

Proof. For the point $z = \mathrm{e}^{i\pi/2p}$ we have

$$|N_p(z)-1|^2 = \frac{1}{p^2}\left(2p^2 - 2p - 2 - 2p(p-1)\cos\left(\frac{\pi}{2p}\right) - 2p\sin\left(\frac{\pi}{2p}\right)\right).$$

Since $p > 2$ and $\cos(x) \geq 1 - x^2/2$ and $\sin(x) \geq x - x^3/6$ for $0 < x < \pi/2$,

$$p^2|N_p(z)-1|^2 \leq 2p^2 - 2p + 2 - (2p^2-2p)\left(1 - \frac{\pi^2}{8p^2}\right) - 2p\left(\frac{\pi}{2p} - \frac{\pi^3}{48p^3}\right)$$
$$= 2 + \frac{\pi^2}{4} - \pi + \left(\frac{\pi^3}{24p^2} - \frac{\pi^2}{4p}\right) \leq 2 + \frac{\pi^2}{4} - \pi + \frac{\pi^3}{24\cdot 9} < 1.47 < 1.57 < \alpha_0^2 = p^2 b_p,$$

which is what we wanted to prove.

For the point $z = a_p\mathrm{e}^{i\pi/2p}$, setting $\gamma_p = (\frac{p-1}{p})^{p-1} = a_p^{1-p}$, one has

$$|N_p(z)-1|^2 = \frac{1}{p^2}\left(2p^2 + \gamma_p^2 - 2p^2\cos\left(\frac{\pi}{2p}\right) - 2p\gamma_p\sin\left(\frac{\pi}{2p}\right)\right).$$

It is possible to prove as in Lemma 2.9 that γ_p is a decreasing sequence that tends to $1/\mathrm{e}$; thus it holds that $1/\mathrm{e} = \gamma_\infty \leq \gamma_p \leq \gamma_3 = 4/9$.

Finally we have

$$p^2|N_p(z)-1|^2 \leq 2p^2 + \gamma_3^2 - 2p^2\left(1 - \frac{\pi^2}{8p^2}\right) - 2p\gamma_\infty\left(\frac{\pi}{2p} - \frac{\pi^3}{48p^3}\right)$$
$$= \left(\frac{4}{9}\right)^2 + \frac{\pi^2}{4} - \frac{\pi}{\mathrm{e}} + \frac{\pi^3}{24\cdot 9\mathrm{e}} < 1.563 < 1.57 < \alpha_0^2 = p^2 b_p.$$

This completes the proof. □

A consequence of this proof is the applicability of the scalar Newton method because the sequence z_k of (2.4) never reaches zero in $\mathcal{D}_p$, and so the sequence (2.2) never reaches zero in $\mathcal{D}$.

2.2. Matrix convergence. We have shown that if the matrix A is diagonalizable, then the iteration can be reduced to uncoupled scalar iterations, one for each of the eigenvalues. In the general case, by means of the Jordan canonical form of A, we may restrict our attention to the case where $A \in \mathbb{C}^{n\times n}$ is a Jordan block, $J(\lambda, n)$, and λ belongs to the region $\mathcal{D}$ defined in (2.3).

In this case, define the functions $g_k(\lambda)$ as the kth iterate x_k of the sequence (2.2) and $f_k(z_0)$ as the kth iterate z_k of (2.4) and let $\phi(\lambda) = \lambda^{-1/p}$ be defined on the set $\mathbb{C}\backslash(-\infty, 0]$. From Proposition 2.2 it follows that for any $z \in \mathbb{C}\backslash(-\infty, 0]$, $g_k(z) = (f_k \circ \phi)(z)z^{1/p}$. Observe that for the matrix iteration (1.2) with initial value $X_0 = I$, it holds that $X_k = g_k(J)$. We aim to prove that $g_k(J)$ converges to $J^{1/p}$ and that the convergence is quadratic.

Let us recall that a function applied to a Jordan block is defined as [16, p. 311]

$$f(J) = \begin{bmatrix} f(\lambda) & f'(\lambda) & \dots & \frac{f^{(n-1)}(\lambda)}{(n-1)!} \\ & \ddots & \ddots & \vdots \\ & & f(\lambda) & f'(\lambda) \\ 0 & & & f(\lambda) \end{bmatrix}.$$

Then to prove Jordan block convergence from scalar convergence it is sufficient to prove that

$$\frac{g_k^{(n)}(\lambda)}{n!} \longrightarrow \frac{1}{n!}\frac{d^n}{dz^n} z^{1/p}\bigg|_{z=\lambda}, \qquad n = 1, 2, 3, \dots .$$

We prove this fact in two steps. First, we show that the sequence $g_k(z)$ converges uniformly on any compact subset of an open neighborhood of any point z belonging to the set $\mathcal{D}$ of (2.3). Then, we show that the derivatives of g_k evaluated at λ converge to the derivative of the pth root function evaluated at λ and that the convergence of $g_k(J)$ to $J^{1/p}$ is dominated by a quadratically convergent sequence.

We use the notation $\|f(x)\|_K = \sup_K |f(x)|$.

LEMMA 2.11. *The sequence $g_k(z)$ converges uniformly to $z^{1/p}$ in any compact subset of the set $\mathcal{G} = \{z \in \mathbb{C}, \operatorname{Re} z > 0, |z| < 1 + \varepsilon\}$ for some $\varepsilon > 0$.*

Proof. By the proof of Theorem 2.3, the set $\{z \in \mathbb{C}, |z| \geq 1, |\arg z| \leq \pi/2p\}$ is a subset of the immediate basin of attraction $\mathcal{F}$ for the fixed point 1 of the rational iteration f_k, which is open; thus the compact arc $\{|z| = 1, |\arg z| \leq \pi/2\}$ admits a finite open covering belonging to $\mathcal{F}$ and then there exists δ such that $\mathcal{G}_p = \{z \in \mathbb{C}, |z| > 1 - \delta, |\arg z| < \pi/2\}$ is a subset of $\mathcal{F}$ and, from the properties of the Fatou set [14, 1], the set $\{f_k\}$ is a normal family on $\mathcal{G}_p$, and, by an easy argument it can be shown that the sequence f_k converges uniformly to 1 for any compact subset of $\mathcal{G}_p$ (see [1, Thm. 6.3.1]).

Now, consider a compact set $\widetilde{K} \subset \mathcal{G} = \{z \in \mathbb{C}, \operatorname{Re}\ z > 0, |z| < (1 - \delta)^{-p}\}$, since $\phi(z) = z^{-1/p}$ is a continuous map from the set $\mathcal{G}$ to the set $\mathcal{G}_p$, $\phi(\widetilde{K}) = K$ is a compact subset of $\mathcal{G}_p$, and, from what we said above, $\|f_k(z) - 1\|_K \to 0$. If we set $\|z^{1/p}\|_{\widetilde{K}} = M$, then

$$\|g_k(z) - z^{1/p}\|_{\widetilde{K}} = \|z^{1/p}((f_k \circ \phi)(z) - 1)\|_{\widetilde{K}} \leq M\|(f_k \circ \phi)(z) - 1\|_{\widetilde{K}} = M\|f_k(z) - 1\|_K,$$

and, since the last term tends to zero, the proof is thus achieved by choosing $\varepsilon = (1 - \delta)^{-p} - 1$. $\square$

To conclude, consider a compact neighborhood $K \subset \mathcal{D}$ of λ and a circle γ of radius R, centered in λ and fully contained in K. The Cauchy formula yields

$$\left| \frac{g_n^{(k)}(\lambda)}{k!} - \frac{1}{k!}\frac{d^k}{dz^k} z^{1/p}\Big|_{z=\lambda} \right| = \left| \frac{1}{2\pi \mathrm{i}} \oint_\gamma \frac{g_n(z) - z^{1/p}}{(z-\lambda)^{k+1}} dz \right| \leq \frac{1}{R^k} \|g_n(z) - z^{1/p}\|_K \to 0,$$

and then $g_n(J)$ converges to $J^{1/p}$. Moreover, $\|g_n(J) - J^{1/p}\|_\infty \leq \alpha \|f_n(z) - 1\|_{\phi(K)}$ for some constant α, and the sequence $f_n(z)$ converges to 1 in any compact subset of $\mathcal{D}_p$ and the convergence is quadratic (since it converges uniformly and in a neighborhood of 1, it converges quadratically).

This approach can be generalized without any effort to any rational iteration applied to a matrix.

3. Stable variants of the Newton method.

Two stable iterations for the matrix square root, that is, the Denman and Beavers iteration [6, 8]

$$\begin{cases} X_0 = A, \quad Y_0 = I, \\ X_{k+1} = \dfrac{1}{2}\,(X_k + Y_k^{-1}), \quad Y_{k+1} = \dfrac{1}{2}\,(Y_k + X_k^{-1}), \quad k = 0, 1, \ldots, \end{cases} \tag{3.1}$$

and the Meini iteration [17]

$$\begin{cases} Y_0 = I - A, \quad Z_0 = 2(I + A), \\ Y_{k+1} = -Y_k Z_k^{-1} Y_k, \quad Z_{k+1} = Z_k - 2Y_k Z_k^{-1} Y_k, \quad k = 0, 1, \ldots, \end{cases} \tag{3.2}$$

are variants of the Newton iteration. In particular the latter can be rewritten as an iteration for the increment [13]

$$\begin{cases} X_0 = A, \quad H_0 = \dfrac{1}{2}(I - A), \\ X_{k+1} = X_k + H_k, \quad H_{k+1} = -\dfrac{1}{2} H_k X_{k+1}^{-1} H_k. \end{cases} \tag{3.3}$$

In fact, the instability of the simplified Newton iterations $X_{k+1} = (X_k + AX_k^{-1})/2$ and $X_{k+1} = (X_k + X_k^{-1}A)/2$, shown by Higham [8], is mainly due to the pre- or post-multiplication of X_k^{-1} by A. On the other hand, since X_k commutes with A (see [13]), the iteration can be rewritten as

$$X_{k+1} = \frac{X_k + A^{1/2} X_k^{-1} A^{1/2}}{2} \tag{3.4}$$

and also is stable in this new form, as one can see by a particular case of the analysis made in section 3.1. Obviously (3.4) is useless since it involves the square root of A, but it helps us to stabilize the iteration by introducing the variable $Y_k = A^{-1/2} X_k A^{-1/2} = A^{-1} X_k = X_k A^{-1}$. The resulting iteration is that of Denman and Beavers; we refer the reader to [13] for more details on this subject.

Repeating these arguments for the pth root, one has the simplified Newton iteration $X_{k+1} = \frac{1}{p}\big((p-1)X_k + AX_k^{1-p}\big)$ or $X_{k+1} = \frac{1}{p}\big((p-1)X_k + X_k^{1-p}A\big)$, which are unstable as shown in [18]. We show in section 3.1 that, since the instability is due to the one-sided multiplication by A, the modified equation

$$X_{k+1} = \frac{(p-1)X_k + (A^{1/p} X^{-1})^{p-1} A^{1/p}}{p} \tag{3.5}$$

provides in principle an iteration with optimal stability.

Now, with the square root in mind, we introduce the auxiliary variable $N_k = AX_k^{-p}$. It can be shown by induction that with the initial values $X_0 = I$ and $N_0 = A$, each of X_k, N_k, and A commutes with the others. This provides the following variant of the simplified Newton iteration:

$$\begin{cases} X_0 = I, \quad N_0 = A, \\ X_{k+1} = X_k \left(\dfrac{(p-1)I + N_k}{p} \right), \\ N_{k+1} = \left(\dfrac{(p-1)I + N_k}{p} \right)^{-p} N_k. \end{cases} \tag{3.6}$$

Observe that the matrix A does not explicitly appear in the iteration. We denote with the acronym HWA (handled without A) iterations having this feature. Observe that, while X_k converges to $A^{1/p}$, the sequence N_k converges to the identity matrix.

On the other hand, one can introduce the increment

$$H_k = \frac{AX_k^{1-p} - X_k}{p} = -\frac{X_k^{1-p}}{p}\left(X_k^p - A\right), \tag{3.7}$$

where $X_k^p - A$ is the *residual* at the step k. Note that H_k commutes with A and X_k. From (3.7) we obtain $A = (X_k + pH_k)X_k^{p-1}$, which allows us to write

$$H_{k+1} = -\frac{X_{k+1}^{1-p}}{p}\left(X_{k+1}^p - A\right) = -\frac{X_{k+1}^{1-p}}{p}\left(X_{k+1}^p - (X_k + pH_k)X_k^{p-1}\right).$$

Now, because $X_{k+1} = X_k + H_k$ we obtain

$$\begin{aligned} H_{k+1} &= -\frac{X_{k+1}^{1-p}}{p}\left(X_{k+1}^p - (pX_{k+1} - (p-1)X_k)X_k^{p-1}\right) \\ &= -\frac{X_{k+1}X_{k+1}^{-p}}{p}\left(X_{k+1}^p - pX_{k+1}X_k^{p-1} + (p-1)X_k^p\right) \\ &= -\frac{X_{k+1}}{p}\left(I - pX_{k+1}^{1-p}X_k^{p-1} + (p-1)X_k^pX_{k+1}^{-p}\right). \end{aligned} \tag{3.8}$$

Setting $F_k = X_kX_{k+1}^{-1}$ we can write an iteration for the increment of the Newton iteration

$$\begin{cases} X_0 = I, \quad H_0 = \dfrac{(A-I)}{p}, \\ X_{k+1} = X_k + H_k, \quad F_k = X_kX_{k+1}^{-1}, \\ H_{k+1} = -X_{k+1}\left(\dfrac{I - F_k^p}{p} + F_k^{p-1}(F_k - I) \right), \end{cases} \tag{3.9}$$

where the expression for H_{k+1} has been written in a form that reduces the phenomenon of numerical cancellation.

Unfortunately, the iteration (3.9) does not reduce to (3.3) in the case of the square root. A nicer form that generalizes (3.3) is

(3.10)
$$\begin{cases} X_0 = I, \quad H_0 = \dfrac{(A-I)}{p}, \\ X_{k+1} = X_k + H_k, \quad F_k = X_k X_{k+1}^{-1} \\ H_{k+1} = -\dfrac{1}{p} H_k (X_{k+1}^{-1} I + 2X_{k+1}^{-1} F_k + 3X_{k+1}^{-1} F_k^2 + \cdots + (p-1) X_{k+1}^{-1} F_k^{p-2}) H_k. \end{cases}$$

We call it incremental Newton (IN). Even if the form (3.10) is more symmetric than (3.9), its computational cost is higher; in fact the computation of H_{k+1} in the iteration (3.10) can be performed in $O(n^3 p)$ ops, and in the iteration (3.9), it can be performed in $O(n^3 \log p)$ ops.

3.1. Stability analysis. In accordance with [5] we define an iteration $X_{k+1} = f(X_k)$ to be *stable in a neighborhood of a solution* $X = f(X)$ if the error matrices $E_k = X_k - X$ satisfy

$$E_{k+1} = L(E_k) + O(\|E_k\|^2),$$

where L is a linear operator that has bounded powers; that is, there exists a constant $c > 0$ such that for all $n > 0$ and arbitrary E of unit norm, $L^n(E) < c$. This means that a small perturbation introduced in a certain step will not be amplified in the subsequent iterations.

Note that this definition of stability is an asymptotic property and is different from the usual concept of numerical stability, which concerns the global error propagation, aiming to bound the minimum relative error over the computed iterates.

First, we show that the iteration (3.5) has *optimal stability*; i.e., the operator L coincides with the null operator. Then we show that the iterations (3.6) and (3.9) are stable.

With $E_k = X_k - A^{1/p}$, we have

$$E_{k+1} = X_{k+1} - A^{1/p} = \frac{p-1}{p} X_k + \frac{A^{1/p} X_k^{-1} \cdots A^{1/p} X_k^{-1} A^{1/p}}{p} - A^{1/p}. \tag{3.11}$$

Now

$$X_k^{-1} = (A^{1/p} + E_k)^{-1} = A^{-1/p} - A^{-1/p} E_k A^{-1/p} + O(\|E_k\|^2).$$

From this relation we obtain that

$$A^{1/p} X_k^{-1} \cdots A^{1/p} X_k^{-1} A^{1/p} = A^{1/p} - (p-1) E_k + O(\|E_k\|^2). \tag{3.12}$$

Finally combining (3.11) and (3.12) yields

$$E_{k+1} = \frac{p-1}{p}(A^{1/p} + E_k) + \frac{A^{1/p} - (p-1)E_k}{p} - A^{1/p} + O(\|E_k\|^2) = O(\|E_k\|^2), \tag{3.13}$$

which means that this iteration is stable, and the most stable possible according to our definition because $L = 0$.

Now we consider the iteration (3.6) and introduce the error matrices $E_k = X_k - A^{1/p}$ and $F_k = N_k - I$. For the sake of simplicity, we perform a *first order error analysis*; that is, we omit all the terms that are quadratic in the errors. Equality up to second order terms is denoted with the symbol $\doteq$.

From $N_k = I + F_k$, one has

$$\left(\frac{(p-1)I + N_k}{p}\right)^{-p} \doteq \left(I + \frac{F_k}{p}\right)^{-p} \doteq I - F_k, \tag{3.14}$$

and the relation for the errors becomes

$$\begin{bmatrix} E_{k+1} \\ F_{k+1} \end{bmatrix} \doteq \begin{bmatrix} I & \frac{1}{p}A^{1/p} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} E_k \\ F_k \end{bmatrix} = L \begin{bmatrix} E_k \\ F_k \end{bmatrix}. \tag{3.15}$$

The coefficient matrix L is idempotent ($L^2 = L$) and hence has bounded powers. Thus the iteration is stable.

For the iteration (3.9) define the error matrices $M_k = X_k - A^{1/p}$ and H_k; then

$$M_{k+1} = X_{k+1} - A^{1/p} = X_k - A^{1/p} + H_k = M_k + H_k. \tag{3.16}$$

For H_{k+1} the relation is a bit more complicated.

Using (3.16) we can write

$$X_{k+1}^{-1} = (A^{1/p} + M_k + H_k)^{-1} \doteq A^{-1/p} - A^{-1/p}M_kA^{-1/p} - A^{-1/p}H_kA^{-1/p}$$

and

$$F_k = X_kX_{k+1}^{-1} = (A^{1/p} + M_k)X_{k+1}^{-1} \doteq I - H_kA^{-1/p}.$$

The latter equation enables us to write

$$(X_kX_{k+1}^{-1})^q \doteq (I - H_kA^{-1/p})^q \doteq I - qH_kA^{-1/p}. \tag{3.17}$$

Finally we have

$$\begin{aligned} H_{k+1} &= -X_{k+1}\left(\frac{I - F_k^p}{p} + F_k^{p-1}(F_k - I)\right) \\ &\doteq -X_{k+1}\left(\frac{I - I + pH_kA^{-1/p}}{p} + (I - (p-1)H_kA^{-1/p})H_kA^{-1/p}\right) \doteq 0. \end{aligned}$$

In conclusion it holds that

$$\begin{bmatrix} M_{k+1} \\ H_{k+1} \end{bmatrix} \doteq \begin{bmatrix} I & I \\ 0 & 0 \end{bmatrix} \begin{bmatrix} M_k \\ H_k \end{bmatrix} = L \begin{bmatrix} M_k \\ H_k \end{bmatrix}. \tag{3.18}$$

Since the matrix L is idempotent, then also this iteration is stable. A similar result holds for the iteration (3.10). Observe also that, unlike in the iteration (3.6), the norm of L is independent of A.

Remark 2. The iteration analyzed in [2],

$$X_{k+1} = \frac{1}{p}\left((p+1)X_k - X_k^{p+1}A\right), \quad X_0 = I, \tag{3.19}$$

is obtained by applying Newton's iteration to the equation $X^{-p} - A = 0$, which has the same convergence as the scalar iteration $x_{k+1} = \big((p+1)x_k - x_k^{p+1}\lambda_i\big)/p$ for $x_0 = 1$, applied to any eigenvalue λ_i of the matrix A.

Like any polynomial iteration of degree greater than 2, this one has the point $x = \infty$ as (super)attractive fixed point [1], and so the basins for the roots are considerably smaller than the ones for the Newton iteration for $x^p - a = 0$. However, in [2] it is proved that the basin of attraction to 1 contains a disk of center 1 and radius 1. In the same paper, it is shown that the iteration (3.19) is unstable for general matrices. The instability of this iteration can be easily removed by applying the arguments of this section. In fact after simple manipulations we deduce the mathematically equivalent iteration

$$\begin{cases} X_0 = I, \quad N_0 = A, \\ X_{k+1} = X_k \left(\dfrac{(p+1)I - N_k}{p} \right), \\ N_{k+1} = \left(\dfrac{(p+1)I - N_k}{p} \right)^p N_k, \end{cases} \tag{3.20}$$

which is proved to be stable near the solution. One can see the similarity to the HWA method.

The iteration (3.20) was already found by Lakić [15] as the first case of a family of stable iterative methods for computing the inverse pth root.

4. The algorithm.

Here we present our algorithm for computing the principal pth root of a matrix having no nonpositive real eigenvalues. For $p = 2$ one can use the existing algorithms [6, 17, 13], so we assume that we can perform the square root.

ALGORITHM 1 (iteration for the principal pth root of a matrix A).

- Input: a matrix A, an integer $p > 2$, and an algorithm for computing the square root.
- Compute B, the principal square root of A.
- Set $C = B/\|B\|$ for a suitable norm. The eigenvalues of C belong to the set $\mathcal{D}$ of (2.3)
- By means of iteration (3.6) or (3.9)
  - If p is even, compute $S = C^{2/p}$, the $(p/2)$th root of C, and set $X = S\|B\|^{2/p}$.
  - If p is odd, compute $S = C^{1/p}$, the pth root of C, and set $X = \left(S\|B\|^{1/p}\right)^2$.

Observe that both iterations (3.6) and (3.9) can be performed in $O(n^3 \log p)$ ops per step, by means of the binary powering technique, much less than the cost of Schur method which is $O(n^3 p)$ ops. However, for small values of p, the total number of ops needed by Algorithm 1 might be larger than the number of ops needed by the Schur method.

For computing the square root, one can use the algorithm (3.3), possibly with a suitable scaling if needed, or the Schur method; this does not affect the asymptotic order of complexity with respect to p. In our numerical experiments, we have observed that the choice of the square root algorithm used in preprocessing the matrix is crucial for the accuracy of the computed solution. Using the Schur method for computing the preliminary square root and then the iteration (3.6) gives good results comparable to the ones obtained with the algorithm proposed by Smith [18]. In certain cases, it is more convenient to use an iterative method such as (3.3), to compute the preliminary square root.

More details can be given about the operation count and the number of steps needed for the numerical convergence; in fact, two matrix multiplications, one inversion, and a matrix exponentiation to the power p are sufficient to carry out one step of the HWA iteration. For computing the power X^p, with X being a matrix, one can use the binary powering technique with a cost varying from $\lfloor \log_2 p \rfloor$ to $2\lfloor \log_2 p \rfloor$ matrix multiplications. The total cost of one step is then bounded by $(3 + 2\lfloor \log_2 p \rfloor)n^3$ ops. For the IN iteration, the cost is $(p+2)n^3$ ops per step, and for the iteration (3.9) the cost is $(5 + 2\lfloor \log_2 (p-1) \rfloor)n^3$ ops per step.

As shown in section 2, the numerical convergence depends only on the localization of the eigenvalues. The closer they are to the boundary of the basin of convergence, the greater is the number of steps needed. For matrices of the form $C = A^{1/2}/\|A^{1/2}\|$, having eigenvalues in the set $\mathcal{D}$ of (2.3), the slow convergence occurs when some eigenvalue is near 0, namely, when the matrix A is ill-conditioned. For instance, if A is a symmetric positive definite matrix and we use the 2-norm, it is easy to show that the smallest eigenvalue of C is $\sqrt{1/\mu_2(A)}$, where $\mu_2(A) = \|A\|_2\|A^{-1}\|_2$ is the condition number of A. Being C diagonalizable by a unitary transform, the convergence of the matrix iteration is the same as the convergence of its smallest eigenvalue. To get an estimation of the number of steps needed by the Newton method applied to a symmetric matrix, it is enough to compute the number of steps needed by the sequence (1.1) with $a = \sqrt{1/\mu_2(A)}$ to converge.

Even though the number of steps is a growing function of p, it seems bounded from above by a constant.

Finally, it is important to point out that the algorithm we proposed works only to find the principal pth root. It is not clear if it can be used to compute any primary pth root, in particular, roots having eigenvalues in different sectors. One important advantage of the Schur method is that it can be used to compute any primary pth root, not just $A^{1/p}$.

5. Numerical experiments. We have performed several experiments in MATLAB 7. We have compared our algorithms with the simplified Newton (SN) method (1.2), with the Schur method implemented in the function `rootm` of the Matrix Computation Toolbox [10], and with the method based on the formula $A^{1/p} = \exp(\frac{1}{p}\log(A))$, using the functions `logm` and `expm` of MATLAB (this method was suggested by an anonymous referee).

For computing the square root of a matrix, we used the function `sqrtm` of MATLAB, which is based on the Schur form of A, or the iteration (3.3), and if a scaling is needed in (3.3) we used the one proposed in [13]. These algorithms have the same asymptotic cost of $O(n^3)$.

To compute the power to $-p$ in the iteration (3.6), first we compute the pth power of the matrix with the binary powering technique and then we invert the matrix. We stop the iterations when the residuals begin to grow or become NaN.

TEST 1. To illustrate the instability near the solution of the SN method (1.2) and the stability of the proposed variants, we consider the simple 3 × 3 matrix

$$A = \begin{bmatrix} 1 & 1/2 & 0 \\ 1/2 & 1 & 1/2 \\ 0 & 1/2 & 1 \end{bmatrix}$$

and compute the fourth root of the matrix A^4. In Figure 5.1 we have compared the relative residual defined as $\mathcal{R}(X) = \|X^p - A\|_F/\|A\|_F$ for the three methods: SN of equation (1.2), Newton in the version (HWA) provided by equation (3.6), and IN of equation (3.10). We denote by $\|A\|_F$ the Frobenius norm of the matrix A, i.e., $\|A\|_F = (\sum_{i,j=1}^n a_{ij}^2)^{1/2}$. As one can see, for some steps the three methods give the same residual; in fact they are analytically equivalent, but the SN method has some instability problems even after a few steps. Our methods show good stability.

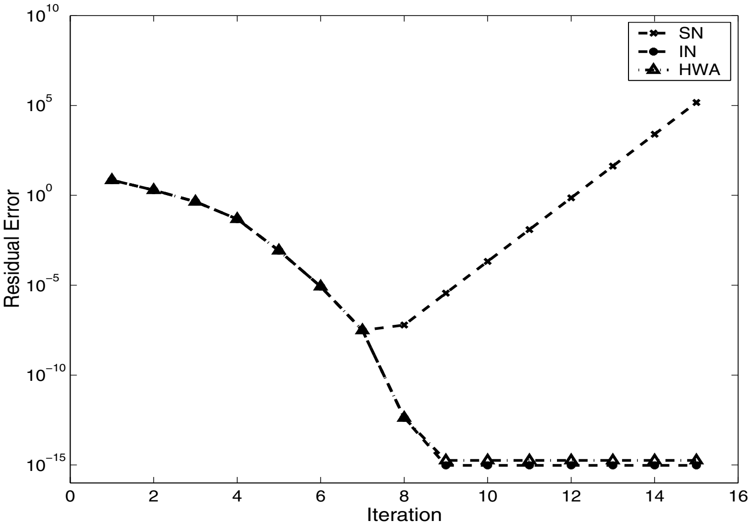

FIG. 5.1. *Comparison of the simplified Newton (*`SN`*) method, the iteration handled without A (*`HWA`*), and incremental Newton (*`IN`*).*

TEST 2. We consider some ill-conditioned matrices to compare the behavior of Algorithm 1 with the one based on the Schur form of A and with the formula $A^{1/p} = \exp(\frac{1}{p}\log(A))$. We also illustrate that the choice of the algorithm used for computing the preliminary square root on Algorithm 1 is very important for the numerical accuracy of the computed solution.

We compute pth root with four methods:

- Compute the square root with the function `sqrtm` in MATLAB and then the iteration (3.6) to compute the pth root (`sqrtm+HWA`).
- Compute the square root by means of the `IN` iteration (3.3) and then the iteration (3.6) to compute the pth root (`IN+HWA`).
- Compute the pth root with the algorithm based on the Schur form (`Smith`).
- Compute $A^{1/p} = \exp(\frac{1}{p}\log(A))$ (`explog`).

The first class of matrices we considered is the class of Hilbert matrices $H_{ij} = 1/(i+j)$ that is a traditional example of an ill-conditioned matrix. We denote by `hilb(n)` the n-dimensional Hilbert matrix. The second class is the prolate matrix, which is a symmetric ill-conditioned Toeplitz matrix whose entries are defined by the formula $A_{ii} = 1/2, A_{ij} = \sin(\pi(j-i)/2)/(\pi(j-i))$. We denote by `prolate(n)` the n-dimensional prolate matrix. The third class is the Frank matrix, an upper Hessenberg matrix with real, positive eigenvalues occurring in reciprocal pairs, half of which are ill-conditioned. We denote by `frank(n)` the n-dimensional Frank matrix. The fourth class is the companion matrix of the polynomial $x^n - 10^{-12}$, whose roots are the nth root of 10^{-12}. We denote by `compan(n)` the n-dimensional companion matrix.

In Table 5.1 we report the relative residuals and the number of iterations (for `HWA` iteration) in computing the 59th root for some of these matrices. As one can see, our algorithm, if provided with a Schur implementation for the preliminary square root, is competitive with the Smith method and provides the same results, in terms of accuracy, if tested with these ill-conditioned matrices. The `explog` algorithm gives good results, but a bit worse than our algorithm in the hardest examples.

A purely iterative algorithm (the second column) suffers from very bad condi-

TABLE 5.1
Comparison of methods for computing the 59*th root of some test matrix.*

Example	sqrtm+HWA		IN+HWA		Smith	explog
hilb(5)	$6.6 \cdot 10^{-15}$	11	$4.4 \cdot 10^{-15}$	11	$3.1 \cdot 10^{-14}$	$8.5 \cdot 10^{-15}$
hilb(10)	$1.7 \cdot 10^{-14}$	20	$1.6 \cdot 10^{-14}$	21	$2.2 \cdot 10^{-14}$	$2.7 \cdot 10^{-14}$
prolate(10)	$1.6 \cdot 10^{-14}$	14	$2.1 \cdot 10^{-14}$	12	$3.3 \cdot 10^{-14}$	$2.2 \cdot 10^{-14}$
prolate(20)	$3.1 \cdot 10^{-14}$	20	$4.3 \cdot 10^{-14}$	22	$3.4 \cdot 10^{-14}$	$4.8 \cdot 10^{-14}$
frank(10)	$2.0 \cdot 10^{-11}$	15	$7.4 \cdot 10^{-10}$	15	$3.5 \cdot 10^{-10}$	$4.5 \cdot 10^{-9}$
frank(14)	$3.5 \cdot 10^{-5}$	22	$2.6 \cdot 10^{-2}$	24	$9.8 \cdot 10^{-4}$	$8.4 \cdot 10^{-2}$
compan(5)	$1.7 \cdot 10^{-3}$	26	$8.3 \cdot 10^{-8}$	27	$5.0 \cdot 10^{-2}$	$1.5 \cdot 10^{-1}$
compan(15)	$1.4 \cdot 10^{0}$	31	$8.8 \cdot 10^{-6}$	30	$4.2 \cdot 10^{1}$	$6.0 \cdot 10^{0}$

tioning of the matrix, but it is faster and in certain cases, like the examples of the companion matrix, gives better results.

Note that when using the procedure described in section 4 for the Hilbert and prolate matrices, which are symmetric, one has a predicted number of steps that almost coincides with that of the examples.

Scaling the iteration for the preliminary square root was not necessary in these examples, but it is worth remarking that sometimes it is important to use a scaling technique in order to avoid poor results.

Acknowledgments. I would like to thank Prof. D. A. Bini for many useful discussions about almost every topic on numerical linear algebra and polynomial computation and Prof. N. J. Higham for providing very useful suggestions for improving the paper. I wish to thank two anonymous referees for their helpful comments which improved the presentation.

REFERENCES

[1] A. F. BEARDON, *Iteration of Rational Functions*, Grad. Texts in Math. 132, Springer-Verlag, New York, 1991.

[2] D. A. BINI, N. J. HIGHAM, AND B. MEINI, *Algorithms for the matrix pth root*, Numer. Algorithms, 39 (2005), pp. 349–378.

[3] Å. BJÖRCK AND S. HAMMARLING, *A Schur method for the square root of a matrix*, Linear Algebra Appl., 52/53 (1983), pp. 127–140.

[4] A. CAYLEY, *The Newton-Fourier imaginary problem*, Amer. J. Math., 2 (1879), p. 97.

[5] S. H. CHENG, N. J. HIGHAM, C. S. KENNEY, AND A. J. LAUB, *Approximating the logarithm of a matrix to specified accuracy*, SIAM J. Matrix Anal. Appl., 22 (2001), pp. 1112–1125.

[6] E. DENMAN AND A. BEAVERS, *The matrix sign function and computations in systems*, Appl. Math. Comput., 2 (1976), pp. 63–94.

[7] M. A. HASAN, A. A. HASAN, AND K. B. EJAZ, *Computation of matrix nth roots and the matrix sector function*, in Proceedings of the 40th Annual IEEE Conference on Decision and Control, 2001, pp. 4057–4062.

[8] N. J. HIGHAM, *Newton's method for the matrix square root*, Math. Comp., 46 (1986), pp. 537–549.

[9] N. J. HIGHAM, *Computing real square roots of a real matrix*, Linear Algebra Appl., 88 (1987), pp. 405–430.

[10] N. J. HIGHAM, *The Matrix Computation Toolbox*, http://www.ma.man.ac.uk/∼higham/mctoolbox.

[11] N. J. HIGHAM, *Functions of a Matrix: Theory and Computation*, in preparation.

[12] W. D. HOSKINS AND D. J. WALTON, *A faster, more stable method for computing the pth root of positive definite matrices*, Linear Algebra Appl., 26 (1979), pp. 139–163.

[13] B. IANNAZZO, *A note on computing the matrix square root*, Calcolo, 40 (2003), pp. 273–283.

[14] G. JULIA, *Mémoire sur l'itération des fonctions rationelles*, J. Math. Pures Appl. (9), 8 (1918), pp. 47–245.

[15] S. Lakić, *On the computation of the matrix kth root*, ZAMM Z. Angew. Math. Mech., 78 (1998), pp. 167–172.

[16] P. Lancaster and M. Tismenetsky, *The Theory of Matrices*, 2nd ed., Academic Press, Orlando, FL, 1985.

[17] 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.

[18] M. I. Smith, *A Schur algorithm for computing matrix pth roots*, SIAM J. Matrix Anal. Appl., 24 (2003), pp. 971–989.