Linear Algebra and its Applications 432 (2010) 1905–1922

Contents lists available at ScienceDirect

Linear Algebra and its Applications

journal homepage: www.elsevier.com/locate/laa

On Newton's method and Halley's method for the principal pth root of a matrix

Chun-Hua Guo [1]

Department of Mathematics and Statistics, University of Regina, Regina, Saskatchewan, Canada S4S 0A2

ARTICLE INFO

Article history:
Received 24 July 2008
Accepted 23 February 2009
Available online 28 March 2009

Submitted by D. Kressner

AMS classification:
65F30
41A58
15A48

Keywords:
Matrix pth root
Newton's method
Halley's method
Convergence
Series expansion
M-matrix
H-matrix

ABSTRACT

If A is a matrix with no negative real eigenvalues and all zero eigenvalues of A are semisimple, the principal pth root of A can be computed by Newton's method or Halley's method, with a preprocessing procedure if necessary. We prove a new convergence result for Newton's method, and discover an interesting property of Newton's method and Halley's method in terms of series expansions. We explain how the convergence of Newton's method and Halley's method can be improved when the eigenvalues of A are known or when A is a singular matrix. We also prove new results on pth roots of M-matrices and H-matrices, and consider the application of Newton's method and Halley's method to find the principal pth roots of these special matrices.

1. Introduction

Let $p \geqslant 2$ be an integer. Suppose that $A \in \mathbb{C}^{n\times n}$ has no negative real eigenvalues and all zero eigenvalues of A are semisimple. Let the Jordan canonical form of A be

$$Z^{-1}AZ = \operatorname{diag}(J_1, J_2, \ldots, J_q).$$

Then the principal pth root of A is

$$A^{1/p} = Z \operatorname{diag}\left(J_1^{1/p}, J_2^{1/p}, \ldots, J_q^{1/p}\right) Z^{-1}.$$

E-mail address: chguo@math.uregina.ca

[1] Supported in part by a grant from the Natural Sciences and Engineering Research Council of Canada.

0024-3795/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.laa.2009.02.030

Here for the $m_k \times m_k$ Jordan block $J_k = J_k(\lambda_k), k = 1, \ldots, q$,

$$J_k^{1/p} = \begin{bmatrix} f(\lambda_k) & f'(\lambda_k) & \cdots & \frac{f^{(m_k-1)}(\lambda_k)}{(m_k-1)!} \\ & f(\lambda_k) & \ddots & \vdots \\ & & \ddots & f'(\lambda_k) \\ & & & f(\lambda_k) \end{bmatrix},$$

where $f(z) = z^{1/p}$ is the principal pth root of the complex number z. It follows that the eigenvalues of $A^{1/p}$ are either 0 or in the segment $\{z \in \mathbb{C}\backslash\{0\} : -\pi/p < \arg(z) < \pi/p\}$.

A practical method for computing the pth root of A is the Schur method given in [15]. Inverse Newton iteration [2,5,12,16], Newton's method [9,10], and Halley's method [10] are good alternatives to the Schur method. In this paper we will present some new results on Newton's method and Halley's method.

2. Newton's method and Halley's method

If we take $x_0 = 1$ and apply Newton's method to the scalar equation $x^p - a = 0$, we get the iteration

$$x_{k+1} = \frac{1}{p}\left((p-1)x_k + a x_k^{1-p}\right), \quad x_0 = 1.$$

In the matrix case, Newton's method for finding $A^{1/p}$ is the following:

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

Similarly, if we take $x_0 = 1$ and apply Halley's method to the scalar equation $x^p - a = 0$, we get the iteration

$$x_{k+1} = x_k \frac{(p-1)x_k^p + (p+1)a}{(p+1)x_k^p + (p-1)a}, \quad x_0 = 1.$$

Halley's method for finding $A^{1/p}$ is given by

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

Due to the special choice of X_0, we have $X_k A = A X_k$ for Newton's method and Halley's method whenever X_k is defined. In the sequel, we will use this commutativity and its consequences freely.

Of course, some conditions will have to be imposed on the matrix A to guarantee the convergence (to $A^{1/p}$) of the sequence $\{X_k\}$ generated by Newton's method or Halley's method. For the inverse Newton's method (Newton's method applied to $X^{-p} - A = 0$), a result about convergence (to $A^{-1/p}$) is proved in [5] using a residual relation, which can be found in [2] for example. In the next section, we will derive residual relations for Newton's method and Halley's method. These residual relations will turn out to be very useful, just as for the inverse Newton's method.

3. Residual relations

For Newton's method or Halley's method, we define the residual by

$$R(X_k) = I - A X_k^{-p}.$$

Lemma 1. *Assume that $\rho(I - A) \leqslant 1$, where $\rho(\cdot)$ denotes the spectral radius. Then the Newton sequence is well defined by (1) and*

$$R(X_{k+1}) = \sum_{i=2}^{\infty} c_i (R(X_k))^i,$$

where $c_i > 0$ for $i \geqslant 2$ and $\sum_{i=2}^{\infty} c_i = 1$. Moreover, if $\|R(X_0)\| = \|I - A\| \leqslant 1$ for a sub-multiplicative matrix norm $\|\cdot\|$, then for each $k \geqslant 0$

$$\|R(X_k)\| \leqslant \left\|(R(X_0))^{2^k}\right\| \leqslant \|R(X_0)\|^{2^k}.$$

Proof. We prove by induction that for each $k \geqslant 0$

$$X_k \text{ is nonsingular and } \rho(R(X_k)) \leqslant 1. \tag{3}$$

For $k = 0$ the statement (3) is true by assumption. Assume that X_k is nonsingular and $\rho(R(X_k)) \leqslant 1$. Then $X_{k+1} = \frac{1}{p}\left((p-1)X_k + AX_k^{1-p}\right) = X_k\left(\left(1 - \frac{1}{p}\right)I + \frac{1}{p}AX_k^{-p}\right) = X_k\left(I - \frac{1}{p}R(X_k)\right)$ is nonsingular, and

$$R(X_{k+1}) = I - \left(I - \frac{1}{p}R(X_k)\right)^{-p} AX_k^{-p} = I - \left(I - \frac{1}{p}R(X_k)\right)^{-p}(I - R(X_k)). \tag{4}$$

By Taylor expansion we have

$$\left(I - \frac{1}{p}R(X_k)\right)^{-p} = \sum_{i=0}^{\infty} d_i (R(X_k))^i,$$

where

$$d_0 = 1, \quad d_1 = 1, \quad d_i = \frac{(p+1)(p+2)\cdots(p+i-1)}{i!\, p^{i-1}}, \quad i \geqslant 2.$$

It then follows from (4) that

$$R(X_{k+1}) = \sum_{i=2}^{\infty} c_i (R(X_k))^i, \tag{5}$$

where $c_i = d_{i-1} - d_i > 0$ for $i \geqslant 2$ and $\sum_{i=2}^{\infty} c_i = 1$. Thus $\rho(R(X_{k+1})) \leqslant \sum_{i=2}^{\infty} c_i (\rho(R(X_k)))^i \leqslant 1$. This proves (3) for all $k \geqslant 0$. If $\|R(X_0)\| \leqslant 1$, then $\|R(X_k)\| \leqslant 1$ for all $k \geqslant 0$ by (5). Moreover,

$$\begin{aligned}
\|R(X_k)\| &\leqslant \left\|(R(X_{k-1}))^2\right\| \left\|\sum_{i=2}^{\infty} c_i (R(X_{k-1}))^{i-2}\right\| \\
&\leqslant \left\|(R(X_{k-1}))^2\right\| \\
&\leqslant \left\|(R(X_{k-2}))^{2^2}\right\| \left\|\left(\sum_{i=2}^{\infty} c_i (R(X_{k-2}))^{i-2}\right)^2\right\| \\
&\leqslant \left\|(R(X_{k-2}))^{2^2}\right\| \\
&\leqslant \cdots \\
&\leqslant \left\|(R(X_0))^{2^k}\right\| \\
&\leqslant \|R(X_0)\|^{2^k}.
\end{aligned}$$

This completes the proof. □

The above result shows how the residual error is reduced right from the beginning if $\|R(X_0)\| < 1$. It also has interesting applications in the next two sections.

We now consider the Halley iteration and assume that $\sigma(A) \subset \mathbb{C}_+$ (all eigenvalues of A are in the open right half plane). In this case it is shown in [10] that the Halley iteration (2) is well defined, and

the iterates X_k are nonsingular and converge to $A^{1/p}$. We will establish a residual relation for Halley's method.

For any $n \times n$ matrices X and Y with $XY = YX$ and Y nonsingular, we will use $\frac{X}{Y}$ to denote $Y^{-1}X$ (which is the same as XY^{-1}). So for the Halley iteration, we have

$$\begin{aligned} X_{k+1} &= X_k \frac{(p-1)X_k^p + (p+1)A}{(p+1)X_k^p + (p-1)A} \\ &= X_k \frac{(p-1)I + (p+1)AX_k^{-p}}{(p+1)I + (p-1)AX_k^{-p}} \\ &= X_k \frac{(p-1)I + (p+1)(I - R(X_k))}{(p+1)I + (p-1)(I - R(X_k))} \\ &= X_k \frac{I - \frac{p+1}{2p}R(X_k)}{I - \frac{p-1}{2p}R(X_k)}. \end{aligned}$$

Now

$$\begin{aligned} R(X_{k+1}) &= I - \left(\frac{I - \frac{p-1}{2p}R(X_k)}{I - \frac{p+1}{2p}R(X_k)}\right)^p AX_k^{-p} \\ &= I - \left(\frac{I - \frac{p-1}{2p}R(X_k)}{I - \frac{p+1}{2p}R(X_k)}\right)^p (I - R(X_k)). \end{aligned} \tag{6}$$

Let

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

Then $f(t)$ has the Taylor expansion

$$f(t) = \sum_{i=3}^{\infty} c_i t^i, \quad |t| < \frac{2p}{p+1}, \tag{7}$$

where $\sum_{i=3}^{\infty} c_i = 1$. It is easy to find that $c_3 = (p^2-1)/(12p^2) > 0$ and to show that $c_4 > 0$. Experiments suggest that $c_i > 0$ for all $i \geqslant 3$. Let q be the unique positive number such that

$$\sum_{i=3}^{\infty} |c_i| q^{i-3} = 1. \tag{8}$$

Note that $q \leqslant 1$ and moreover $q = 1$ if we indeed have $c_i > 0$ for all $i \geqslant 3$.

Lemma 2. *Assume that $\sigma(A) \subset \mathbb{C}_+ \cup \{0\}$. Then the Halley sequence $\{X_k\}$ is well defined with X_k nonsingular, and when $\rho(R(X_k)) < 2p/(p+1)$*

$$R(X_{k+1}) = \sum_{i=3}^{\infty} c_i (R(X_k))^i, \tag{9}$$

where c_i are as in (7). Moreover, if $\|R(X_0)\| = \|I - A\| \leqslant q$ for a matrix norm $\|\cdot\|$, where q is given in (8), then for each $k \geqslant 0$

$$\|R(X_k)\| \leqslant \left\|(R(X_0))^{3^k}\right\| \leqslant \|R(X_0)\|^{3^k}. \tag{10}$$

Proof. By the Jordan canonical form of A, to show X_k is well defined and nonsingular we only need to show this when the Halley iteration is applied to each Jordan block of A. For Jordan blocks corresponding to zero eigenvalues each X_k is defined and has a single eigenvalue $((p-1)/(p+1))^k$. For Jordan blocks corresponding to nonzero eigenvalues the result is proved in [10]. When $\rho(R(X_k)) < 2p/(p+1)$ we have (9) in view of (6) and (7). Now suppose $\|R(X_0)\| \leqslant q$. Then $\|R(X_1)\| \leqslant \|R(X_0)\|^3 \leqslant q^3 \leqslant q$ by (9) and (8). It then follows that $\|R(X_k)\| \leqslant q$ for all $k \geqslant 0$. Again by (9) and (8) we have for each $k \geqslant 1$ that $\|R(X_k)\| \leqslant \left\|(R(X_{k-1}))^3\right\|$. We then proceed as in the proof of Lemma 1 to get (10). □

4. Convergence results

The convergence of Newton's method and Halley's method for a matrix A follows from the convergence of scalar Newton's method and Halley's method applied to the eigenvalues of A (see [8, Theorem 4.15]).

For the scalar Newton's method, the following result is proved in [9], after the proof of seven technical lemmas.

Theorem 3. *Let λ be any complex number in $\{z : \mathrm{Re}\, z > 0, |z| \leqslant 1\}$. Then Newton's method with $x_0 = 1$, applied to the equation $x^p - \lambda = 0$, converges to $\lambda^{1/p}$.*

This convergence region allows one to compute the pth root of any matrix with no nonpositive real eigenvalues, by performing one square root computation and a proper scaling [9]. Later, the inverse Newton's method is studied in [5] and it is suggested there that a few more matrix square roots be performed so that the inverse Newton's method applied to the new matrix will have faster convergence and better numerical stability. The same suggestion applies to Newton's method and Halley's method as well. Ideally, the matrix after preprocessing should have all eigenvalues close to 1. The strategy is natural since we start the iterations with $X_0 = I$.

In view of this, the region in Theorem 3 becomes insufficient since 1 is on the boundary of the region. Then the following result is proved in [10] (see Theorem 6.1 and Corollary 6.2 there), on the basis of Theorem 3 and its proof.

Theorem 4. *Let λ be any complex number in $\{z : 0 < |z| \leqslant 2, |\arg(z)| < \pi/4\}$. Then Newton's method with $x_0 = 1$, applied to the equation $x^p - \lambda = 0$, converges to $\lambda^{1/p}$.*

However, the region in the next theorem is the most natural one. Part of this region is not covered by any of the two regions in Theorems 3 and 4. The three regions in Theorems 3–5 are depicted in Fig. 1 by solid lines.

We remark that, as noted by one referee, a re-examination of the proof of [10, Theorem 6.1] shows that the condition $|\arg(z)| < \pi/4$ in Theorem 4 can be relaxed to $|\arg(z)| \leqslant \pi/3$, with only some small changes in that proof. In fact, one can change $\frac{\sqrt{\alpha_0^2 - \pi^2/16}}{p} > 0 > \frac{\log 2}{p}$ (which contains a casual error) in the last line of [10, p. 1461] to $\frac{\sqrt{\alpha_0^2 - \pi^2/16}}{p} > \frac{\log 2}{p}$, and then simply replace 4 by 3 and replace 16 by 9 in that proof, in a total of 13 places. The union of the region in Theorem 3 and the region in the strengthened Theorem 4 will then cover the region E in Theorem 5 except the point 0, for which the convergence is easy to prove. See Fig. 1 again, where the dotted lines indicate the extension in the strengthened Theorem 4.

A direct proof of Theorem 5 will be based on Lemma 1 and the result that the basin of attraction for any attractive fixed point of a rational iteration is an open set. The author thanks Bruno Iannazzo for bringing this result to his attention. Without using this result, we would have produced a much longer proof, similar to that of Lemma 2.5 in [5].

Theorem 5. *Let λ be any complex number in $E = \{z : |z - 1| \leqslant 1\}$. Then Newton's method with $x_0 = 1$, applied to the equation $x^p - \lambda = 0$, converges to $\lambda^{1/p}$.*

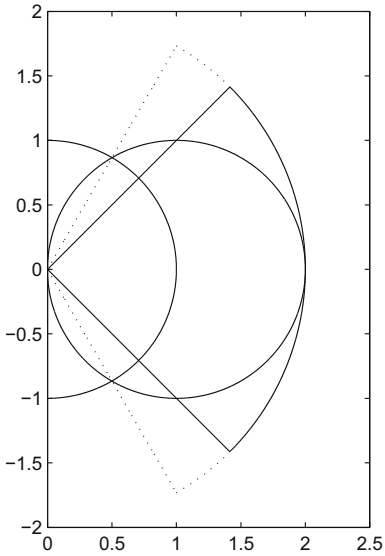

Fig. 1. Convergence regions in Theorems 3–5.

Proof. The Newton sequence $\{x_k\}$ is well defined by Lemma 1, and the residual is now $r(x_k) = 1 - \lambda x_k^{-p}$. If $\lambda = 0$, then $x_k = ((p-1)/p)^k$, converging to 0 linearly. If $\lambda \neq 0$ and $|r(x_0)| = |\lambda - 1| \leqslant 1$, then by Lemma 1

$$|r(x_1)| \leqslant |c_2 + c_3 r(x_0)| + \sum_{i=4}^{\infty} c_i < \sum_{i=2}^{\infty} c_i.$$

The second inequality is strict since equality would hold only when $|r(x_0)| = 1$ and $r(x_0)$ is a positive real number (in other words, only when $\lambda = 0$). So $|r(x_1)| < 1$ and $|r(x_k)| \leqslant |r(x_1)|^{2^{k-1}}$ for $k \geqslant 1$. Since $x_k^p = \lambda/(1 - r(x_k))$, the sequence $\{x_k\}$ is bounded. It then follows from $|x_{k+1} - x_k| = \frac{1}{p}|x_k||r(x_k)|$ that $\{x_k\}$ is a Cauchy sequence and hence converges. The limit must be a pth root of λ since $r(x_k)$ converges to 0. We now prove that the limit is always the principal pth root $\lambda^{1/p}$. By [9, Proposition 2.2] and its proof, x_k converges to the pth root $\lambda^{1/p} e^{i2\pi l/p}$ $(l = 0, 1, \ldots, p-1)$ if and only if the sequence $\{z_k\}$ defined by the rational iteration

$$z_{k+1} = \frac{1}{p}\left((p-1)z_k + z_k^{1-p}\right), \quad z_0 = \lambda^{-1/p}$$

converges to the attractive fixed point $e^{i2\pi l/p}$. Let $\widehat{E} = f(E \setminus \{0\})$, where $f(z) = z^{-1/p}$. Since f is continuous and $E \setminus \{0\}$ is connected, $\widehat{E}$ is also connected. For $z_0 = 1 \in \widehat{E}$, z_k converges to 1. If for some $z_0 \in \widehat{E}$, z_k converges to $e^{i2\pi l/p}$ for some $l \neq 0$, then the connected set $\widehat{E}$ would be the union of more than one disjoint nonempty sets that are open in $\widehat{E}$, which is impossible. So z_k converges to 1 for all $z_0 \in \widehat{E}$. In other words, x_k converges to $\lambda^{1/p}$ for all $\lambda \in E\backslash\{0\}$. □

In view of [8, Theorem 4.15], we have the following result.

Theorem 6. *If all eigenvalues of A are in $\{z : |z - 1| \leqslant 1\}$ and all zero eigenvalues of A (if any) are semisimple, then the Newton sequence, with $X_0 = I$, converges to $A^{1/p}$.*

The convergence is quadratic if A has no zero eigenvalues. This can be proved by using a procedure similar to the one in the proof of Theorem 7 below. If A has semisimple zero eigenvalues, the con-

vergence is linear with rate $(p-1)/p$, which is the rate of convergence of the scalar Newton method applied to the zero eigenvalue.

We also have the following convergence result for Halley's method.

Theorem 7. *If all eigenvalues of A are in $\mathbb{C}_+$, then the Halley sequence $\{X_k\}$, with $X_0 = I$, converges to $A^{1/p}$ cubically.*

Proof. The convergence of X_k to $A^{1/p}$ has been proved in [10]. We just need to prove that the rate of convergence is cubic. Since $\|R(X_k)\| \to 0$, we have $\|R(X_k)\| \leqslant q$ for all $k \geqslant k_0$, where q is given in (8). It follows from (9) and (8) that

$$\|R(X_{k+1})\| \leqslant \|R(X_k)\|^3 \tag{11}$$

for all $k \geqslant k_0$. Let $S = A^{1/p}$. We will show that $R'(S)$, the Fréchet derivative of $R(X) = I - AX^{-p}$ at S is an invertible linear operator from $\mathbb{C}^{n\times n}$ into itself. Direct computation shows that $R'(S)$ is given by

$$R'(S)(E) = (S^{p-1}E + S^{p-2}ES + S^{p-3}ES^2 + \ldots + ES^{p-1})A^{-1}.$$

So we only need to show that the linear operator L given by

$$L(E) = S^{p-1}E + S^{p-2}ES + S^{p-3}ES^2 + \ldots + ES^{p-1}$$

is invertible. The n^2 eigenvalues of L are given by $\sum_{k=0}^{p-1} \lambda_i^k \lambda_j^{p-1-k}$ for $i, j = 1, \ldots, n$, where λ_i are the eigenvalues of S. Since $\sigma(A) \subset \mathbb{C}_+$, we have $\arg(\lambda_i) \in (-\pi/(2p), \pi/(2p))$. It follows that $\lambda_i^k \lambda_j^{p-1-k} \in \mathbb{C}_+$ for each i, j, k. So all eigenvalues of L are in $\mathbb{C}_+$ as well. We have thus proved that $R'(S)$ is invertible. By increasing k_0 if necessary, we know that there are constants $c_1, c_2 > 0$ such that for all $k \geqslant k_0$, $\|R(X_k)\| = \|R(X_k) - R(S)\|$ satisfies

$$c_1\|X_k - S\| \leqslant \|R(X_k)\| \leqslant c_2\|X_k - S\|.$$

It then follows from (11) that $\|X_{k+1} - S\| \leqslant (c_2^3/c_1)\|X_k - S\|^3$ for all $k \geqslant k_0$. So the convergence is cubic. □

We can also allow A to have semisimple zero eigenvalues. In that case, the convergence is linear with rate $(p-1)/(p+1)$, which is the rate of convergence of the scalar Halley's method applied to the zero eigenvalue.

The convergence in Theorems 6 and 7 may fail to materialize in finite precision arithmetic, since the Newton iteration (1) and the Halley iteration (2) are usually numerically unstable.

A stable version of (1) has been given in [9]:

$$\begin{aligned} &X_0 = I, \quad N_0 = A, \\ &X_{k+1} = X_k\left(\frac{(p-1)I + N_k}{p}\right), \\ &N_{k+1} = \left(\frac{(p-1)I + N_k}{p}\right)^{-p} N_k, \end{aligned} \tag{12}$$

where $N_k \to I$ and $X_k \to A^{1/p}$.

Also, a stable version of (2) has been given in [10]:

$$\begin{aligned} &X_0 = I, \quad N_0 = A, \\ &X_{k+1} = X_k((p+1)I + (p-1)N_k)^{-1}((p-1)I + (p+1)N_k), \\ &N_{k+1} = N_k\left(((p+1)I + (p-1)N_k)^{-1}((p-1)I + (p+1)N_k)\right)^{-p}, \end{aligned} \tag{13}$$

where $N_k \to I$ and $X_k \to A^{1/p}$.

The sequences $\{X_k\}$ produced by the stable versions are the same as the sequences $\{X_k\}$ produced by the original iterations in exact arithmetic. Therefore, the theoretical properties of the sequences $\{X_k\}$ produced by (1) and (2) will be exhibited on the stable versions. We should use the stable versions for actual computations, but will continue to use (1) and (2) for theoretical analysis.

5. Connection with binomial expansion

In the binomial expansion

$$(1-z)^{1/p} = \sum_{i=0}^{\infty} b_i z^i, \quad |z| < 1,$$

we have

$$b_0 = 1, \quad b_i = (-1)^i \frac{\frac{1}{p}\left(\frac{1}{p}-1\right)\cdots\left(\frac{1}{p}-i+1\right)}{i!} < 0, \quad i \geqslant 1.$$

We first consider the scalar Newton iteration for finding $(1-z)^{1/p}$ with $|z| \leqslant 1$:

$$x_{k+1} = \frac{1}{p}\left((p-1)x_k + (1-z)x_k^{1-p}\right), \quad x_0 = 1.$$

To emphasize the dependence of x_k on z, we will write $x_k(z)$ for x_k. By Lemma 1, $x_k(z) = p_k(z)/q_k(z)$ with polynomials $p_k(z)$ and $q_k(z)$ having no zeros in the closed unit disk. It follows that each $x_k(z)$ has a power series expansion

$$x_k(z) = \sum_{i=0}^{\infty} c_{k,i} z^i, \quad |z| \leqslant 1. \tag{14}$$

As we shall see later, it is of interest to study the sign pattern of the coefficients $c_{k,i}$. Since $x_k(0) = 1$ for all $k \geqslant 0$, $c_{k,0} = 1$ for all $k \geqslant 0$. Also, $c_{0,i} = 0$ for all $i \geqslant 1$, and $c_{1,1} = -1/p, c_{1,i} = 0$ for all $i \geqslant 2$. We are able to find the expressions for all $c_{2,i}$ as well. Indeed,

$$\begin{aligned} x_2(z) &= \frac{1}{p}\left((p-1)\left(1-\frac{1}{p}z\right) + (1-z)\left(1-\frac{1}{p}z\right)^{1-p}\right) \\ &= \frac{1}{p}\left(1-\frac{1}{p}z\right)\left(p-1+(1-z)\left(1-\frac{1}{p}z\right)^{-p}\right) \\ &= \frac{1}{p}\left(1-\frac{1}{p}z\right)\left(p-\sum_{i=2}^{\infty} c_i z^i\right), \end{aligned}$$

where we have used (5) for the last equality. We then find that in (14)

$$c_{2,1} = -\frac{1}{p}, \quad c_{2,i} = -((i-1)p-(i-2))\frac{(p-1)p(p+1)\cdots(p+i-3)}{i!\,p^{i+1}}, \quad i \geqslant 2.$$

So we have $c_{2,i} < 0$ for all $i \geqslant 1$. However, it seems hopeless to determine the sign pattern of $c_{k,i}$ in this way for larger k. Experiments do suggest that $c_{k,i} < 0$ $(i \geqslant 1)$ also for $k \geqslant 3$.

Since $x_{k+1}(1) = \frac{p-1}{p}x_k(1)$ and $x_0(1) = 1$, we have $x_k(1) = ((p-1)/p)^k$ and thus $\sum_{i=0}^{\infty} c_{k,i} = ((p-1)/p)^k$ for each $k \geqslant 0$.

In summary we have the following result.

Proposition 8. *For Newton's method and the coefficients $c_{k,i}$ in (14) we have*

(a) $c_{k,0} = 1$ *for* $k \geqslant 0$.
(b) $c_{0,i} = 0$ *for* $i \geqslant 1$.

(c) $c_{1,1} = -1/p, c_{1,i} = 0$ *for* $i \geqslant 2$.
(d) $c_{2,i} < 0$ *for* $i \geqslant 1$.
(e) $\sum_{i=0}^{\infty} c_{k,i} = ((p-1)/p)^k$ *for* $k \geqslant 0$.

We now consider the scalar Halley iteration for finding $(1-z)^{1/p}$ with $|z| \leqslant 1$:

$$x_{k+1} = x_k \frac{(p-1)x_k^p + (p+1)(1-z)}{(p+1)x_k^p + (p-1)(1-z)}, \quad x_0 = 1.$$

We already know that $x_k(z)$ is defined and nonzero whenever $|z| \leqslant 1$. So $x_k(z) = p_k(z)/q_k(z)$ with polynomials $p_k(z)$ and $q_k(z)$ having no zeros in the closed unit disk, and each $x_k(z)$ has a power series expansion

$$x_k(z) = \sum_{i=0}^{\infty} c_{k,i} z^i, \quad |z| \leqslant 1. \tag{15}$$

Since $x_k(0) = 1$ for all $k \geqslant 0, c_{k,0} = 1$ for all $k \geqslant 0$. Also, $c_{0,i} = 0$ for all $i \geqslant 1$. Moreover,

$$x_1(z) = \frac{(p-1) + (p+1)(1-z)}{(p+1) + (p-1)(1-z)} = 1 - \frac{1}{p} z \frac{1}{1 - \frac{p-1}{2p} z} = 1 - \frac{1}{p} \sum_{i=1}^{\infty} \left(\frac{p-1}{2p} \right)^{i-1} z^i.$$

So we have $c_{1,i} < 0$ for all $i \geqslant 1$. Since $x_k(1) = ((p-1)/(p+1))^k$, we have $\sum_{i=0}^{\infty} c_{k,i} = ((p-1)/(p+1))^k$ for all $k \geqslant 0$.

In summary we have the following result.

Proposition 9. *For Halley's method and the coefficients* $c_{k,i}$ *in* (15) *we have*

(a) $c_{k,0} = 1$ *for* $k \geqslant 0$.
(b) $c_{0,i} = 0$ *for* $i \geqslant 1$.
(c) $c_{1,i} < 0$ *for* $i \geqslant 1$.
(d) $\sum_{i=0}^{\infty} c_{k,i} = ((p-1)/(p+1))^k$ *for* $k \geqslant 0$.

The main purpose of this section, however, is to reveal the following interesting connection between the Newton/Halley iteration and the binomial expansion.

Theorem 10. *For Newton's method,* $c_{k,i} = b_i$ *for* $k \geqslant 0$ *and* $0 \leqslant i \leqslant 2^k - 1$. *For Halley's method,* $c_{k,i} = b_i$ *for* $k \geqslant 0$ *and* $0 \leqslant i \leqslant 3^k - 1$.

Proof. We fix $k \geqslant 0$. For Newton's method, take $B = J(0)_{2^k \times 2^k}$, the $2^k \times 2^k$ Jordan block with 0's on the main diagonal, and apply the matrix Newton iteration to $A = I - B$. Since $\|R(X_0)\|_1 = \|I - A\|_1 = \|B\|_1 \leqslant 1$, we have by Lemma 1 that $\|R(X_k)\|_1 \leqslant \|R(X_0)^{2^k}\|_1 = \|B^{2^k}\|_1 = 0$. So $R(X_k) = 0$. Thus $X_k = (I - B)^{1/p}$ by Theorem 6. For our special choice of B,

$$X_k = \sum_{i=0}^{\infty} c_{k,i} B^i = \begin{bmatrix} c_{k,0} & c_{k,1} & \cdots & c_{k,2^k-1} \\ & c_{k,0} & \ddots & \vdots \\ & & \ddots & c_{k,1} \\ & & & c_{k,0} \end{bmatrix},$$

$$(I - B)^{1/p} = \sum_{i=0}^{\infty} b_i B^i = \begin{bmatrix} b_0 & b_1 & \cdots & b_{2^k-1} \\ & b_0 & \ddots & \vdots \\ & & \ddots & b_1 \\ & & & b_0 \end{bmatrix}.$$

It follows that $c_{k,i} = b_i$ for $i = 0, 1, \ldots, 2^k - 1$.

For Halley's method, take $B = J(0)_{3^k\times 3^k}$ and apply the matrix Halley iteration to $A = I - B$. Since $\rho(B) = 0$, $\|B\| \leqslant q$ for some matrix norm, where q is as in Lemma 2. Then $\|R(X_0)\| \leqslant q$ and by Lemma 2 $\|R(X_k)\| \leqslant \left\|R(X_0)^{3^k}\right\| = \left\|B^{3^k}\right\| = 0$. So $R(X_k) = 0$. Thus $X_k = (I - B)^{1/p}$ by Theorem 7. It follows that $c_{k,i} = b_i$ for $i = 0, 1, \ldots, 3^k - 1$. □

In addition to the results in Propositions 8 and 9, we have the following corollary of Theorem 10.

Corollary 11. *For Newton's method, $c_{k,i} < 0$ for $k \geqslant 3$ and $1 \leqslant i \leqslant 2^k - 1$. For Halley's method, $c_{k,i} < 0$ for $k \geqslant 2$ and $1 \leqslant i \leqslant 3^k - 1$.*

However, the proof of the following conjecture seems difficult.

Conjecture 12. *For Newton's method, $c_{k,i} < 0$ for $k \geqslant 3$ and $i \geqslant 2^k$. For Halley's method, $c_{k,i} < 0$ for $k \geqslant 2$ and $i \geqslant 3^k$.*

The following result about matrix iterations follows directly from Theorem 10.

Theorem 13. *Suppose that all eigenvalues of A are in $\{z : |z - 1| < 1\}$ and write $A = I - B$ (so $\rho(B) < 1$). Let $(I - B)^{1/p} = \sum_{i=0}^{\infty} b_i B^i$ be the binomial expansion. Then the sequence $\{X_k\}$ generated by Newton's method or by Halley's method has the Taylor expansion $X_k = \sum_{i=0}^{\infty} c_{k,i} B^i$. For Newton's method we have $c_{k,i} = b_i$ for $i = 0, 1, \ldots, 2^k - 1$, and for Halley's method we have $c_{k,i} = b_i$ for $i = 0, 1, \ldots, 3^k - 1$.*

It is known that $-1 < b_i < 0$ $(i \geqslant 1)$. If Conjecture 12 is true, we also have $-1 < c_{k,i} \leqslant 0$ $(k \geqslant 0, i \geqslant 1)$ for Newton's method and Halley's method, by Propositions 8 and 9. In that case, we have by Theorem 13 that for any matrix norm

$$\left\|X_k - A^{1/p}\right\| < \sum_{i=2^k}^{\infty} \|B^i\| \quad \text{and} \quad \left\|X_k - A^{1/p}\right\| < \sum_{i=3^k}^{\infty} \|B^i\|$$

for Newton's method and Halley's method, respectively. When $\|B\| < 1$, we have further

$$\left\|X_k - A^{1/p}\right\| < \frac{\|B\|^{2^k}}{1 - \|B\|} \quad \text{and} \quad \left\|X_k - A^{1/p}\right\| < \frac{\|B\|^{3^k}}{1 - \|B\|} \tag{16}$$

for Newton's method and Halley's method, respectively. These neat error estimates show the practical importance of Conjecture 12.

Although Conjecture 12 remains unproven, Theorem 13 is instructive in designing faster Newton iteration or Halley iteration for finding the pth root of a matrix. It shows that the actual error $X_k - A^{1/p}$ is largely determined by $(\rho(B))^{2^k}$ and $(\rho(B))^{3^k}$ for Newton's method and Halley's method, respectively, when k is not too small. Thus, in general, a given nonsingular matrix will need to go through a preprocessing step, so that the resulting matrix has the form $A = I - B$ with $\rho(B)$ significantly smaller than 1, say $\rho(B) \leqslant \frac{1}{2}$.

For a given matrix with semisimple zero eigenvalues, we can use a proper linear combination of two consecutive Newton (or Halley) iterates to recover quadratic (or cubic) convergence.

These procedures will be described in more detail in the next section. We end this section by noting that we can also prove the following analogue of Theorem 13 for the inverse Newton method.

Theorem 14. *Suppose that all eigenvalues of A are in $\{z : |z - 1| < 1\}$ and write $A = I - B$. Let $(I - B)^{-1/p} = \sum_{i=0}^{\infty} \hat{b}_i B^i$ be the binomial expansion. Then the sequence $\{X_k\}$ generated by the inverse Newton method*

$$X_{k+1} = \frac{1}{p}\left((p+1)X_k - X_k^{p+1}A\right), \quad X_0 = I \tag{17}$$

has the Taylor expansion $X_k = \sum_{i=0}^{\infty} \hat{c}_{k,i} B^i$, and $\hat{c}_{k,i} = \hat{b}_i$ for $i = 0, 1, \ldots, 2^k - 1$.

6. Convergence improvement

To compute the pth root of a general matrix with no nonpositive real eigenvalues, we proceed as in [5]. Let $p = 2^{k_0}q$ with $k_0 \geqslant 0$ and q odd. When $q = 1, A^{1/p}$ can be found by taking the square root [3,7] k_0 times. So we assume $q \geqslant 3$. Let $A = QRQ^*$ be the Schur decomposition of A, and the eigenvalues of A be ordered such that $|\lambda_1| \geqslant |\lambda_2| \geqslant \ldots \geqslant |\lambda_n|$. Let $k_1 \geqslant k_0$ be the smallest integer such that

(i) $|\lambda_1/\lambda_n|^{1/2^{k_1}} \leqslant 2$, if all λ_i are real.

(ii) $|\lambda_1/\lambda_n|^{1/2^{k_1}} \leqslant 2$ and $\arg(\lambda_i^{1/2^{k_1}}) \in [-\pi/8, \pi/8]$ for all i, if not all λ_i are real.

Then we can find $A^{1/p}$ using $A^{1/p} = Q((R^{1/2^{k_1}})^{1/q})^{2^{k_1-k_0}}Q^*$, where for $T = R^{1/2^{k_1}}$, $T^{1/q}$ is found by applying (12) or (13), after the matrix T is scaled properly.

We write $T = c(I - B)$ with $c > 0$ and $B = I - \frac{1}{c}T$. Then $T^{1/q} = c^{1/q}(I - B)^{1/q}$ and $(I - B)^{1/q}$ is computed by (12) or (13). In view of Theorem 13, we should choose c such that $\rho(B)$ is minimized.

We have two cases. If all eigenvalues of A are real, then T has real eigenvalues $\mu_1 \geqslant \mu_2 \geqslant \ldots \geqslant \mu_n > 0$ and $\mu_1/\mu_n \leqslant 2$. In this case, $\rho(B)$ is minimized when $c = (\mu_1 + \mu_n)/2$ and the minimum is

$$\rho(B) = \frac{\mu_1 - \mu_n}{\mu_1 + \mu_n} = \frac{\mu_1/\mu_n - 1}{\mu_1/\mu_n + 1} \leqslant \frac{1}{3}.$$

If not all eigenvalues of A are real, then the eigenvalues of T are arranged such that $|\mu_1| \geqslant |\mu_2| \geqslant \cdots \geqslant |\mu_n|$ and $|\mu_1/\mu_n| \leqslant 2$ and moreover $\arg(\mu_i) \in [-\pi/8, \pi/8]$. In this case, if we take $c = (|\mu_1| + |\mu_n|)/2$, then

$$\frac{2}{3} \leqslant \frac{1}{c}|\mu_i| \leqslant \frac{4}{3}, \quad i = 1, \ldots, n, \tag{18}$$

since $|\mu_1/\mu_n| \leqslant 2$, and thus $\left|1 - \frac{1}{c}\mu_i\right| \leqslant \left|1 - \frac{4}{3}e^{i\pi/8}\right|$. So $\rho(B) \leqslant \left|1 - \frac{4}{3}e^{i\pi/8}\right| \in (0.560445, 0.560446)$.

With $O(n)$ operations we can also determine c such that $\rho(B) = \max_i \left|1 - \frac{1}{c}\mu_i\right|$ is nearly minimized. We let

$$\xi_i = \frac{2\mu_i}{|\mu_1| + |\mu_n|}, \quad s = \frac{|\mu_1| + |\mu_n|}{2c}$$

and will determine s such that $f(s) = \max_i |1 - s\xi_i|$ is minimized. We use the idea in the proof of Proposition 4.5 in [6]. Let

$$D_1 = \{z : |z - 1/2| \leqslant 1/2\}, \quad D_2 = \{z : |z - 1/2| \geqslant 1/2\},$$

and for $s \in (0, \infty)$

$$f_1(s) = \max_{1 \leqslant i \leqslant n:\, s\xi_i \in D_1} |1 - s\xi_i|, \quad f_2(s) = \max_{1 \leqslant i \leqslant n:\, s\xi_i \in D_2} |1 - s\xi_i|,$$

where the maximum over an empty set is defined to be zero. Then we have the following result.

Proposition 15. *A positive number s minimizes $f(s)$ if and only if $f_1(s) = f_2(s)$.*

Proof. We can decrease s slightly to decrease $f(s)$ if $f_1(s) < f_2(s)$, and can increase s slightly to decrease $f(s)$ if $f_1(s) > f_2(s)$. So s minimizes $f(s)$ only if $f_1(s) = f_2(s)$.

If $f_1(s) = f_2(s)$, decreasing s always increases $f_1(s)$ and thus $f(s)$, while increasing s always increases $f_2(s)$ and thus $f(s)$. Thus s minimizes $f(s)$ if $f_1(s) = f_2(s)$. □

Since $\frac{2}{3} \leqslant |\xi_i| \leqslant \frac{4}{3}$ by (18), the optimal s must satisfy $\frac{2}{3}s - 1 < 0.560446$ and $1 - \frac{4}{3}s < 0.560446$. So $s \in (0.33, 2.35)$. We can then use a simple bisection procedure to find the optimal s:

(1) Let $a = 0.33$ and $b = 2.35$.
(2) Compute $r = (a + b)/2$, $f_1(r)$, $f_2(r)$.

- If $f_1(r) = f_2(r)$ then $s = r$ is optimal.
- If $f_1(r) < f_2(r)$ then $b = r$ and goto (2).
- If $f_1(r) > f_2(r)$ then $a = r$ and goto (2).

So if the optimal s is not found earlier, we know $s \in (a, b)$ with $b - a < 2 \times 10^{-6}$ after 20 bisections, and $r = (a + b)/2$ is near-optimal.

For illustration, we consider $T \in \mathbb{C}^{2\times 2}$ with $\mu_1 = 2e^{i\pi/8}$ and $\mu_2 = 1$. If we take $c = \frac{1}{2}(|\mu_1| + |\mu_2|) = 1.5$, then $\rho(B) = \rho\left(1 - \frac{1}{c}T\right) = \left|1 - \frac{2}{3}e^{i\pi/8}\right| \approx 0.560445$. A near-optimal c found by the above procedure is $c \approx 1.76937$ and for this c, $\rho(B) \approx 0.434827$. Now suppose that k Newton iterations are used to get an approximation to $(I - B)^{1/p}$ and that the error is exactly determined by $\rho(B)^{2^k}$. For $k = 5$ in the above example, the error would be 9.0×10^{-9} for $c = 1.5$ and 2.7×10^{-12} for $c = 1.76937$. So the near-optimal c provides better accuracy and may save one iteration dependent on the accuracy requirement. In practice, however, a smaller $\rho(B)$ will not always mean better convergence. Nevertheless, the search for a near-optimal c (in the sense that $\rho(B)$ is nearly minimized) is worthwhile in general since it requires only $O(n)$ flops, while one Newton iteration requires $O(n^3 \log p)$ flops.

In view of Theorem 14, the above strategy for selecting the parameter c can also be used for the Schur–Newton method in [5], where (in the notation of this paper) $T^{1/q} = c^{1/q}((I - B)^{-1/q})^{-1}$, and $(I - B)^{-1/q}$ is computed using the following stable version [9,12] of the inverse Newton iteration (17):

$$\begin{aligned} &X_0 = I, \quad N_0 = A, \\ &X_{k+1} = X_k\left(\frac{(p+1)I - N_k}{p}\right), \\ &N_{k+1} = \left(\frac{(p+1)I - N_k}{p}\right)^p N_k, \end{aligned} \tag{19}$$

where $N_k \to I$ and $X_k \to A^{-1/p}$ when $A = I - B$ with $\rho(B) < 1$.

A good scaling factor c for the matrix T is given in [5] using the eigenvalues μ_i of T. If all μ_i are real then

$$c = \begin{cases} \frac{(\mu_1/\mu_n)^{1/q}\mu_1 - \mu_n}{((\mu_1/\mu_n)^{1/q} - 1)(q+1)}, & \text{if } \mu_1 \neq \mu_n \\ \mu_1, & \text{if } \mu_1 = \mu_n, \end{cases}$$

otherwise $c = (|\mu_1| + |\mu_n|)/2$. When all μ_i are real, our strategy here is much simpler (we simply take $c = (\mu_1 + \mu_n)/2$). Moreover, our strategy here is backed by theoretical results even when not all μ_i are real. However, the advantage of the new strategy over the one in [5] is not expected to be very significant in numerical computations, since both strategies are built on the conditions (i) and (ii) at the beginning of this section.

Our new strategy is the same for Newton's method, Halley's method and the inverse Newton's method. The strategy in [10] (for Newton's method and Halley's method) is quite different. To reduce the integer k_1 appearing in $T = R^{1/2^{k_1}}$, the conditions (i) and (ii) are not enforced in [10]. Instead, for Newton's method k_1 is chosen to be the smallest nonnegative integer such that all eigenvalues of $\frac{1}{c}T$ are in the disk $\{z : |z - 6/5| \leqslant 3/4\}$ for some $c > 0$, and for Halley's method k_1 is chosen to be the smallest nonnegative integer such that all eigenvalues of $\frac{1}{c}T$ are in the disk $\{z : |z - 8/5| \leqslant 1\}$ for some $c > 0$. The two disks are determined heuristically in [10] by experiments on the scalar Newton iteration and Halley iteration for a few selected values of p. The idea there is to ensure that in the scalar case at most 5 iterations are needed for Newton's method and at most 3 iterations are needed for Halley's method. It is easy to see that the strategy in [10] reduces k_1 by 1 in most cases (by 0 or 2 in other cases), as compared to our new strategy here. However, Newton's method or Halley's method

may need more iterations if one uses the strategy in [10] instead of our strategy here, since no attempt is made in [10] to choose c properly so that the eigenvalues of $\frac{1}{c}T$ are better distributed in the two disks. Note that the computational work for one Newton or Halley iteration is significantly more than that saved by reducing k_1 by 1 or 2, particularly when p is large. Reducing k_1 by 1 saves only $2/3\,n^3$ flops, which is quite small compared to the $28n^3$ flops required by the Schur method. For the above reasons, our new strategy is to be preferred in general.

We mentioned earlier that the convergence of Newton's method and Halley's method will be linear when A has semisimple zero eigenvalues (assuming that the remaining eigenvalues of A are in $\{z : |z-1| \leqslant 1\} \setminus \{0\}$ for Newton's method and in $\mathbb{C}_+$ for Halley's method). We now explain how we can speed up convergence. Suppose A has Jordan canonical form

$$Z^{-1}AZ = \operatorname{diag}(J_1, \ldots, J_r, 0, \ldots, 0),$$

where the eigenvalues of $J_1, \ldots, J_r$ are in $\{z : |z-1| \leqslant 1\} \setminus \{0\}$ for Newton's method and in $\mathbb{C}_+$ for Halley's method.

Then the sequence $\{X_k\}$ generated by Newton's method has the form

$$X_k = Z \operatorname{diag}(X_k^{(1)}, \ldots, X_k^{(r)}, ((p-1)/p)^k, \ldots, ((p-1)/p)^k) Z^{-1},$$

where $X_k^{(i)}$ $(i = 1, \ldots, r)$ are obtained when Newton's method is applied to J_i. So $X_k^{(i)}$ converges to $J_i^{1/p}$ quadratically, but X_k converges to $A^{1/p}$ only linearly. However, a simple linear combination of two consecutive iterates will eliminate the linearly convergent terms. In fact, we can compute $Z_k = pX_{k+1} - (p-1)X_k$. Then Z_k converges to $A^{1/p}$ quadratically.

Similarly, for Halley's method we compute $Z_k = \frac{1}{2}((p+1)X_{k+1} - (p-1)X_k)$. Then Z_k converges to $A^{1/p}$ cubically.

One often uses the usual residual definition $R(Y) = Y^p - A$ to measure the accuracy of Y as an approximation to $A^{1/p}$. However, one should keep in mind the following easily verified result.

Proposition 16. *Suppose that A has semisimple zero eigenvalues. Then for Newton's method or Halley's method*

$$\|R(X_k)\| = O(\|X_k - A^{1/p}\|^p), \quad \|R(Z_k)\| = O(\|Z_k - A^{1/p}\|).$$

Thus the error $\|X_k - A^{1/p}\|$ is much larger than a small $\|R(X_k)\|$ would suggest.

7. Application to *M*-matrices and *H*-matrices

For special matrices, Newton's method and Halley's method can be applied without using the Schur decomposition. This is the case for M-matrices and H-matrices.

For any matrices $A, B \in \mathbb{R}^{m\times n}$, we write $A \geqslant B (A > B)$ if $a_{ij} \geqslant b_{ij} (a_{ij} > b_{ij})$ for all i, j. A real square matrix A is called a Z-matrix if all its off-diagonal entries are nonpositive. Any Z-matrix A can be written as $sI - B$ with $B \geqslant 0$. A Z-matrix A is called a nonsingular M-matrix if $s > \rho(B)$ and a singular M-matrix if $s = \rho(B)$. For a matrix $A = [a_{ij}] \in \mathbb{C}^{m\times n}$, its absolute value is $|A| = [|a_{ij}|]$. For a matrix $A \in \mathbb{C}^{n\times n}$, its comparison matrix is $B = [b_{ij}]$ with $b_{ii} = |a_{ii}|$ and $b_{ij} = -|a_{ij}|$ for $i \neq j$; A is called a nonsingular H-matrix if its comparison matrix B is a nonsingular M-matrix.

The next result is well known.

Lemma 17. *Let A be a nonsingular M-matrix. If $B \geqslant A$ is a Z-matrix, then B is also a nonsingular M-matrix.*

It is known [1,4,11] that $A^{1/p}$ is a nonsingular M-matrix for every nonsingular M-matrix A. We now consider the generalization of this result to H-matrices.

Theorem 18. *Let A be a nonsingular H-matrix with positive diagonal entries. Then the principal pth root of A exists and is a nonsingular H-matrix whose diagonal entries have positive real parts.*

Proof. Let B be the comparison matrix of A. Then

$$A = sI - C, \quad B = sI - D,$$

where $s > \rho(D), D \geqslant 0$, and $|C| = D$. It follows that $\rho(C) \leqslant \rho(|C|) < s$. Let $E = C/s$ and $F = D/s$. Then $\rho(E) < 1$ and $\rho(F) < 1$. The principal pth roots of A and B are given by

$$A^{1/p} = s^{1/p}(I - G), \quad B^{1/p} = s^{1/p}(I - H),$$

where

$$G = \sum_{k=1}^{\infty} c_k E^k, \quad H = \sum_{k=1}^{\infty} c_k F^k$$

with $c_k = (-1)^{k-1} \frac{\frac{1}{p}\left(\frac{1}{p}-1\right)\cdots\left(\frac{1}{p}-k+1\right)}{k!} > 0$. Since $F \geqslant 0$ and $|E| = F$, we have $H \geqslant 0$ and $|G| \leqslant H$. As noted in [11], $\rho(H) = 1 - (1 - \rho(F))^{1/p} < 1$. So $B^{1/p}$ is a nonsingular M-matrix. It follows from $|G| \leqslant H$ that $|\mathrm{Re}(g_{ii})| \leqslant |g_{ii}| \leqslant h_{ii} < 1$. Thus all diagonal entries of $A^{1/p}$ have positive real parts. Moreover, the comparison matrix T of $A^{1/p}$ satisfies $T \geqslant B^{1/p}$. So T is a nonsingular M-matrix by Lemma 17, and thus $A^{1/p}$ is a nonsingular H-matrix. □

When A is a complex nonsingular H-matrix with positive diagonal entries, the diagonal entries of $A^{1/p}$ are not necessarily real. On the other hand, $A^{1/p}$ is a real matrix when A is real (see [8]). So we have the following result, which has been proved in [13] for $p = 2$.

Corollary 19. *If A is a real nonsingular H-matrix with positive diagonal entries, then so is $A^{1/p}$.*

We also present the following result about singular M-matrices.

Theorem 20. *If A is a singular M-matrix with semisimple zero eigenvalues, then so is $A^{1/p}$.*

Proof. We use the definition of $A^{1/p}$ through the Jordan canonical form of A. It is clear that $A^{1/p}$ is a singular matrix with semisimple zero eigenvalues. To show that $A^{1/p}$ is an M-matrix, we let $A(\epsilon) = A + \epsilon I$ with $\epsilon > 0$. So $A(\epsilon)$ is a nonsingular M-matrix and thus $A(\epsilon)^{1/p}$ is a nonsingular M-matrix. It follows from the definition of $A(\epsilon)^{1/p}$ through the Jordan canonical form of $A(\epsilon)$ that $A(\epsilon)^{1/p}$ converges to $A^{1/p}$ as $\epsilon \to 0$. Therefore $A^{1/p}$ is an M-matrix. □

Corollary 21. *If A is an irreducible singular M-matrix, then so is $A^{1/p}$.*

Proof. $A^{1/p}$ is irreducible since otherwise $A = (A^{1/p})^p$ would be reducible. The result follows since any irreducible M-matrix has a simple zero eigenvalue. □

We now consider the computation of $A^{1/p}$, where A is a nonsingular H-matrix with positive diagonal entries (including all nonsingular M-matrices) or a singular M-matrix with semisimple zero eigenvalues.

Let s be the largest diagonal entry of A. Then $A = s(I - B)$ with $\rho(B) \leqslant 1$. We compute $A^{1/p}$ through $A^{1/p} = s^{1/p}(I - B)^{1/p}$. To find $(I - B)^{1/p}$ we generate a sequence $\{X_k\}$ by Newton's method or Halley's method, with $X_0 = I$ in each case. When A is a singular M-matrix, we need to generate a new sequence $\{Z_k\}$ for faster convergence and better accuracy, as described in the previous section. For the remainder of this section, we assume $A = I - B$ is in $\mathcal{M}_1$ or $\mathcal{H}_1$, where $\mathcal{M}_1$ is the set of all nonsingular M-matrix with $0 < a_{ii} \leqslant 1$ (in this case $B \geqslant 0$), and $\mathcal{H}_1$ is the set of all nonsingular real H-matrix with $0 < a_{ii} \leqslant 1$. We would like to know whether Newton's method and Halley's method are structure preserving: are the approximations X_k all in $\mathcal{M}_1$ (or $\mathcal{H}_1$) when A is in $\mathcal{M}_1$ (or $\mathcal{H}_1$)?

Proposition 22. *For Newton's method or Halley's method, the matrix X_k is in $\mathcal{M}_1$ for all matrices A in $\mathcal{M}_1$ (of all sizes) if and only if $c_{k,i} \leqslant 0$ for all $i \geqslant 1$.*

Proof. If X_k is in $\mathcal{M}_1$ for all matrices $A = I - B$ in $\mathcal{M}_1$ (of all sizes), then it is so for $A = -J(-1)_{m\times m}$ and each $m \geqslant 1$, where $J(-1)_{m\times m}$ is the $m \times m$ Jordan block with -1's on the diagonal (and then $B = J(0)_{m\times m}$). It follows from $X_k = I + \sum_{i=1}^{\infty} c_{k,i}B^i$ that $c_{k,i} \leqslant 0$ for all $i \geqslant 1$.

Now assume that $c_{k,i} \leqslant 0$ for all $i \geqslant 1$ and A is in $\mathcal{M}_1$. Then $A = I - B$ with $B \geqslant 0$ and $\rho(B) < 1$, and $X_k = I - \sum_{i=1}^{\infty}(-c_{k,i})B^i$ is a Z-matrix. Also,

$$1 - \rho\left(\sum_{i=1}^{\infty}(-c_{k,i})B^i\right) \geqslant 1 - \sum_{i=1}^{\infty}(-c_{k,i})\rho(B)^i \geqslant 1 - \sum_{i=1}^{\infty}(-c_{k,i}) > 0,$$

where the last inequality follows from Proposition 8 for Newton's method and from Proposition 9 for Halley's method. Thus X_k is in $\mathcal{M}_1$. □

For Newton's method we know from Proposition 8 that X_1 and X_2 are in $\mathcal{M}_1$ when A is so. For Halley's method we know from Proposition 9 that X_1 is in $\mathcal{M}_1$ when A is so. In view of Corollary 11, other X_k are likely in $\mathcal{M}_1$, but a confirmation will depend on the proof of Conjecture 12.

When $p = 2$, it is shown in [14] that for Newton's method all X_k are in $\mathcal{M}_1$ when A is so. So Conjecture 12 is true for Newton's method with $p = 2$, by Proposition 22. Even for $p = 2$, it is an open problem as to whether the matrices X_k generated by Newton's method (with $X_0 = I$) are nonsingular M-matrices for every nonsingular M-matrix A (with $a_{ii} > 1$ for some i). However, this problem is of purely theoretical interest, since it is more appropriate to compute $A^{1/2}$ through $A^{1/2} = s^{1/2}(I - B)^{1/2}$, in view of Theorem 13. Here s is the largest diagonal entry of A.

The next result shows that if Newton's method and Halley's method are structure preserving in $\mathcal{M}_1$ then they are also structure preserving in $\mathcal{H}_1$.

Proposition 23. *Let A be in $\mathcal{H}_1$. If $c_{k,i} \leqslant 0$ for all $i \geqslant 1$ for Newton's method or Halley's method, then the matrix X_k from Newton's method or Halley's method is also in $\mathcal{H}_1$.*

Proof. Let $\widehat{A}$ be the comparison matrix of A. Write $A = I - B$ and $\widehat{A} = I - \widehat{B}$. So $\widehat{B} \geqslant 0$ and $|B| = \widehat{B}$. We know $\widehat{X}_k = I - \sum_{i=1}^{\infty}(-c_{k,i})\widehat{B}^i$ is in $\mathcal{M}_1$ by Proposition 22. Then $X_k = I - \sum_{i=1}^{\infty}(-c_{k,i})B^i$ is in $\mathcal{H}_1$ since $|\sum_{i=1}^{\infty}(-c_{k,i})B^i| \leqslant \sum_{i=1}^{\infty}(-c_{k,i})\widehat{B}^i$. □

More can be said about Newton's method for the square root of a matrix in $\mathcal{H}_1$.

Proposition 24. *Let $p = 2$ and $A \in \mathcal{H}_1$. Write $A = I - B$ (so $\rho(B) < 1$). Then*

(a) *The matrices X_k from Newton's method are all in $\mathcal{H}_1$.*
(b) *For any matrix norm such that $\|B\| < 1$,*

$$\|X_k - A^{1/2}\| < \frac{\|B\|^{2^k}}{1 - \|B\|}.$$

Proof. Recall that $c_{k,i} \leqslant 0 (k \geqslant 0, i \geqslant 1)$ for Newton's method when $p = 2$. The conclusion in (a) then follows from Proposition 23 and the conclusion in (b) follows from the discussions leading to (16). □

Since it is advisable to reduce a real H-matrix to a real H-matrix in $\mathcal{H}_1$ to compute the square root. Proposition 24 (a) thus solves the interesting part of a research problem stated in [8, Prob. 6.25].

8. Numerical results

In this section we present a few numerical examples to illustrate the usefulness of our strategies in Section 6 for convergence improvement.

Table 1
Notation used in Tables 2 and 3.

$\mathrm{err}(N,0)$	Relative error for Newton's method with the strategy in [10]
$\mathrm{err}(N,1)$	Relative error for Newton's method with our new strategy
$\mathrm{err}(H,0)$	Relative error for Halley's method with the strategy in [10]
$\mathrm{err}(H,1)$	Relative error for Halley's method with our new strategy
$\mathrm{err}(iN,0)$	Relative error for inverse Newton iteration with the strategy in [5]
$\mathrm{err}(iN,1)$	Relative error for inverse Newton iteration with our new strategy

Table 2
Newton, Halley and inverse Newton iterations for Example 1.

k	$\mathrm{err}(N,0)$	$\mathrm{err}(N,1)$	$\mathrm{err}(H,0)$	$\mathrm{err}(H,1)$	$\mathrm{err}(iN,0)$	$\mathrm{err}(iN,1)$
1	1.4e0	3.6e−1	2.7e−1	6.7e−3	4.7e−1	4.2e−1
2	1.5e−1	4.6e−3	1.3e−4	2.7e−8	8.6e−3	6.9e−3
3	1.9e−3	8.1e−7	2.8e−8		3.0e−6	2.1e−6
4	2.9e−7	2.8e−8			2.8e−8	2.8e−8
5	2.8e−8					

Example 1 [10]. Let $A = S^{15}$, where

$$S = \begin{bmatrix} -1 & -2 & 2 \\ -4 & -6 & 6 \\ -4 & -16 & 13 \end{bmatrix}$$

has eigenvalues $1, 2, 3$. Note that the 2-norm condition number of A is $\kappa_2(A) = 1.6 \times 10^{10}$. We now compute $A^{1/15}$ using Newton's method (12), Halley's method (13) and the inverse Newton iteration (19), after a preprocessing procedure is used. For an approximation $\widetilde{X}$ to $A^{1/15}$, the relative error in Frobenius norm is $\mathrm{err} = \|\widetilde{X} - S\|_F / \|S\|_F$. Our new preprocessing strategy is the same for all three methods. The strategy in [10] produces an interval of good values of c and any point in that interval can be used. For definiteness, we take the left endpoint each time. For this example, we have $k_1 = 5$ and $c = 1.3368$ with our new strategy, $k_1 = 4$ and $c = 1.44$ with the strategy in [10] for Newton's method, $k_1 = 4$ and $c = 1.08$ with the strategy in [10] for Halley's method, $k_1 = 5$ and $c = 1.3099$ with the strategy in [5] for the inverse Newton iteration. The convergence history is shown in Table 2. The notation used in Table 2 (and later in Table 3) is listed in Table 1. Since all eigenvalues of A are real in this example, the bisection procedure in Section 6 has no role to play. From Table 2 we can see that, as compared to the strategy in [10], our new strategy may reduce the number of iterations by 1 while increasing the number of square roots by 1. So our new strategy is to be preferred. Our new strategy also appears slightly better than the strategy in [5] for the inverse Newton iteration. Indeed, with the hindsight from Theorem 14, we now know that the strategy in [5] is unnecessarily complicated and should be replaced by our new strategy in general when all eigenvalues of A are real.

Example 2. Let $A = S^5$, where

$$S = \begin{bmatrix} 0.44 & -0.88 & -0.38 & -0.50 \\ 0.68 & 2.15 & 0.48 & 0.11 \\ 0.61 & 0.77 & 2.14 & 1.04 \\ -0.16 & -0.30 & -0.67 & 1.33 \end{bmatrix}.$$

Note that $A^{1/5} = S$ and $\kappa_2(A) = 1.9 \times 10^2$. The eigenvalues of A, rounded to two decimal places, are $15.25, 0.27 \pm 16.01i, 1.10$. We now compute $A^{1/5}$ using Newton's method, Halley's method and the inverse Newton iteration, after a preprocessing procedure is used. For an approximation $\widetilde{X}$ to $A^{1/5}$, let $\mathrm{err} = \|\widetilde{X} - S\|_F / \|S\|_F$. For this example, we have $k_1 = 2$ and $c = 1.7853$ with our new strategy, $k_1 = 2$ and $c = 1.18$ with the strategy in [10] for Newton's method, $k_1 = 2$ and $c = 0.88$ with the strategy in [10] for Halley's method, $k_1 = 2$ and $c = 1.5125$ with the strategy in [5] for

Table 3
Newton, Halley and inverse Newton iterations for Example 2.

k	err$(N,0)$	err$(N,1)$	err$(H,0)$	err$(H,1)$	err$(iN,0)$	err$(iN,1)$
1	3.1e−1	9.3e−2	8.9e−2	1.1e−2	2.2e−1	1.3e−1
2	2.4e−2	3.6e−3	2.2e−5	1.1e−7	1.7e−2	1.1e−2
3	9.1e−5	5.2e−6	2.0e−15	1.5e−15	1.6e−4	5.8e−5
4	3.2e−9	1.8e−11			1.2e−8	2.5e−9
5	1.2e−15	1.3e−15			1.7e−15	2.0e−15

the inverse Newton iteration. Note that the strategy in [10] fails to reduce k_1 for this example. The convergence history is shown in Table 3. Since not all eigenvalues of A are real in this example, the bisection procedure in Section 6 is used. From Table 3 we can see that, as compared to the strategy in [10], our new strategy could possibly reduce the number of iterations by 1, depending on the accuracy requirement. So our new strategy is again to be preferred. Our new strategy also provides better results (before the limiting accuracy is achieved) than the strategy in [5] for the inverse Newton iteration. Since the cost of the bisection procedure is only $O(n)$ for an $n \times n$ matrix A, the use of this procedure is recommended in general. However, if the simplicity of the overall algorithm is important, one may continue to use the strategy in [5] when not all eigenvalues of A are real. That simpler strategy may also be used for Newton's method and Halley's method.

Example 3. Let $A = S^5$, where

$$S = \begin{bmatrix} 2 & -1 & -1 \\ -0.5 & 1.5 & -1 \\ -0.5 & -1 & 1.5 \end{bmatrix}.$$

The matrix A is an irreducible singular M-matrix and $A^{1/5} = S$. We have

$$A = 78.125 \begin{bmatrix} 1 & -0.5 & -0.5 \\ -0.25 & 0.75 & -0.5 \\ -0.25 & -0.5 & 0.75 \end{bmatrix} = 78.125\widehat{A}.$$

We then use Newton's method (12) or Halley's method (13) to get approximations $\widehat{X}_k$ to $\widehat{A}^{1/5}$, and take $X_k = (78.125)^{1/5}\widehat{X}_k$ as approximations to $A^{1/5} = S$. For any approximation Y to S, let $\mathrm{err}(Y) = \|Y - S\|_2$ and $\mathrm{res}(Y) = \|Y^5 - A\|_2$.

Newton's method converges linearly with rate 4/5. After 36 iterations, we find $\mathrm{err}(X_{36}) = 1.1 \times 10^{-3}$ and $\mathrm{res}(X_{36}) = 2.5 \times 10^{-13}$. However, for $Y_4 = 5X_5 - 4X_4$ we have $\mathrm{err}(Y_4) = 2.3 \times 10^{-15}$ and $\mathrm{res}(Y_4) = 4.1 \times 10^{-13}$. Halley's method converges linearly with rate 2/3, and we find $\mathrm{err}(X_{20}) = 1.2 \times 10^{-3}$ and $\mathrm{res}(X_{20}) = 3.4 \times 10^{-13}$. However, for $Y_3 = 3X_4 - 2X_3$ we have $\mathrm{err}(Y_3) = 1.3 \times 10^{-14}$ and $\mathrm{res}(Y_3) = 3.0 \times 10^{-13}$. The relationship between the error and the residual is as predicted by Proposition 16, and the strategy in Section 6 for convergence improvement in the singular case is very useful.

Acknowledgments

The author would like to thank Nick Higham, Bruno Iannazzo, Daniel Kressner and three anonymous referees for their helpful comments.

References

[1] T. Ando, Inequalities for M-matrices, Linear and Multilinear Algebra 8 (1980) 291–316.
[2] D.A. Bini, N.J. Higham, B. Meini, Algorithms for the matrix pth root, Numer. Algorithms 39 (2005) 349–378.
[3] Å. Björck, S. Hammarling, A Schur method for the square root of a matrix, Linear Algebra Appl. 52/53 (1983) 127–140.

[4] M. Fiedler, H. Schneider, Analytic functions of M-matrices and generalizations, Linear and Multilinear Algebra 13 (1983) 185–201.
[5] C.-H. Guo, N.J. Higham, A Schur–Newton method for the matrix pth root and its inverse, SIAM J. Matrix Anal. Appl. 28 (2006) 788–804.
[6] C.-H. Guo, N.J. Higham, Iterative solution of a nonsymmetric algebraic Riccati equation, SIAM J. Matrix Anal. Appl. 29 (2007) 396–412.
[7] N.J. Higham, Computing real square roots of a real matrix, Linear Algebra Appl. 88/89 (1987) 405–430.
[8] N.J. Higham, Functions of Matrices: Theory and Computation, SIAM, Philadelphia, 2008.
[9] B. Iannazzo, On the Newton method for the matrix pth root, SIAM J. Matrix Anal. Appl. 28 (2006) 503–523.
[10] B. Iannazzo, A family of rational iterations and its applications to the computation of the matrix pth root, SIAM J. Matrix Anal. Appl. 30 (2008) 1445–1462.
[11] C.R. Johnson, Inverse M-matrices, Linear Algebra Appl. 47 (1982) 195–216.
[12] S. Lakić, On the computation of the matrix k-th root, Z. Angew. Math. Mech. 78 (1998) 167–172.
[13] L. Lin, Z.-Y. Liu, On the square root of an H-matrix with positive diagonal elements, Ann. Oper. Res. 103 (2001) 339–350.
[14] B. Meini, The matrix square root from a new functional perspective: Theoretical results and computational issues, SIAM J. Matrix Anal. Appl. 26 (2004) 362–376.
[15] M.I. Smith, A Schur algorithm for computing matrix pth roots, SIAM J. Matrix Anal. Appl. 24 (2003) 971–989.
[16] R.A. Smith, Infinite product expansions for matrix n-th roots, J. Austral. Math. Soc. 8 (1968) 242–249.