Note : high08_FM Chapter 6

Zhengbo Zhou*

January 27, 2023

Contents

Abstract

This is a note for [5], and this file by no mean serve as a correct reference for knowledge. This is only for myself convenience.

1 Introduction

In this chapter, we will always consider the principal square root, denoted as $A^{1/2}$. Recall that if $A \in \mathbb{C}^{n\times n}$ with no eigenvalues on $\mathbb{R}^-$, $A^{1/2}$ is the unique square root X of A whose spectrum lies in the open right half-plane. We will denote $\sqrt{A}$ be any arbitrary, possibly non-principal square roots. The matrix (principal) square root also has an integral expression:

$$A^{1/2} = \frac{2}{\pi} A \int_0^\infty (t^2 I + A)^{-1} \, \mathrm{d}t. \tag{1.1}$$

The layout of this chapter will be

*Department of Mathematics, University of Manchester, Manchester, M13 9PL, England (zhengbo.zhou@postgrad.manchester.ac.uk).

1. Sensitivity and conditioning: Analysis of the conditioning of the matrix square root, and the sensitivity of the relative residual.
2. Schur method: A Schur method and a version working entirely in real arithmetic are described.
3. Stability and limiting accuracy: Newton's method and several variants follow, with a stability analysis revealing that the variants do not suffer the instability that vitiates the Newton iteration.
4. Scaling the Newton iteration.
5. Numerical Experiments.
6. Iteration via matrix sign function: A class of coupled iterations obtained via iterations for the matrix sign function are derived and their stability proved.
7. Special matrices: Linearly convergent iterations for matrices that are "almost diagonal", as well as for M-matrices, are analyzed, and a preferred iteration for Hermitian posi- tive definite matrices is given.
8. Computing small-normed square roots: The issue of choosing from among the many square roots of a given matrix is addressed by considering how to compute a small-normed square root.
9. Comparison of methods.
10. Involutory matrices.

2 Sensitivity and Conditioning

2.1 Condition number

Theorem 2.1 (Theorem 3.5 in [5]). *Let f and f^{-1} both exist and be continuous in an open neighbourhood of X and $f(X)$. Assume L_f exists at the neighbourhood and nonsingular at X. Then, $L_{f^{-1}}$ exists at $f(X)$, and*

$$L_f\left(X, L_{f^{-1}}(f(X), E)\right) = E.$$

Example 2.2. Suppose $f(X) = X^2$, recall that $L_{x^2}(X, E) = XE + EX$. Obviously, $f^{-1}(X) = \sqrt{X}$, and let $A = f(X) = X^2$. Then, by 2.1, we have

$$L_{x^2}(X, L_{x^{1/2}}(A, E)) = E, \quad \Longrightarrow \quad L_{x^{1/2}}(A, E)X + XL_{x^{1/2}}(A, E) = E.$$

Formally, suppose $g(A) = \sqrt{A}$, then we have $L_g(A, E)$ defined by the following matrix equation ♡(Sylvester equation?)

$$L_g(A, E)A^{1/2} + A^{1/2}L_g(A, E) = E.$$

Using $\mathrm{vec}(AXB) = (B^T \otimes A)\mathrm{vec}(X)$, and define $L_g(A, E) \equiv L$, we have

$$(X^T \otimes I + I \otimes X)\mathrm{vec}(L) = \mathrm{vec}(E),$$

and then we can deduce that

$$\|L\|_F = \|(X^T \otimes I + I \otimes X)^{-1}\|_2.$$

Hence, we have the Frobenius norm (relative) condition number of the matrix square root at A is

$$\kappa_{\mathrm{sqrt}}(X) = \frac{\|(X^T \otimes I + I \otimes X)^{-1}\|_2 \|A\|_F}{\|X\|_F}.$$

If follows that

$$\kappa_{\mathrm{sqrt}}(X) \geq \frac{1}{\min_{i,j=1:n}|\mu_i + \mu_j|} \frac{\|A\|_F}{\|X\|_F} \tag{2.1}$$

where μ_j are the eigenvalues of $X = \sqrt{A}$.

This inequality reveals the situation that the κ_{sqrt} must be large:

1. When A (hence X) has an eigenvalue of small modulus.
2. When the square root is the principal square root and a real A has a pair of complex conjugate eigenvalues close to the negative real axis. Suppose $\lambda = r\mathrm{e}^{\mathrm{i}(\pi-\epsilon)}$ $(0 < \epsilon \ll 1)$ and $\overline{\lambda} = r\mathrm{e}^{\mathrm{i}(\epsilon-\pi)}$. Then

$$|\lambda^{1/2} + \overline{\lambda}^{1/2}| = r^{1/2}|\mathrm{e}^{\mathrm{i}(\pi-\epsilon)/2} - \mathrm{e}^{-\mathrm{i}(\pi-\epsilon)/2}| = r^{1/2}O(\epsilon). \tag{2.2}$$

If A is normal, then X is normal, by using the fact that a matrix is normal if and only if it has a spectral decomposition. Then, we have the equality in (2.2).

The formula for κ_{sqrt} allows us to identify the best conditioned square root of a Hermitian positive definite matrix. Define $\kappa(A) = \|A\|\|A^{-1}\|$.

Lemma 2.3. *If $A \in \mathbb{C}^{n\times n}$ is Hermitian positive definite, and X is any primary square root of A, then*

$$\kappa_{\mathrm{sqrt}}(A^{1/2}) = \frac{\|A^{-1}\|_2^{1/2}}{2} \frac{\|A\|_F}{\|A^{1/2}\|_F} \leq \kappa_{\mathrm{sqrt}}(X).$$

Moreover,

$$\frac{1}{2n^{3/2}}\kappa_F(A^{1/2}) \leq \kappa_{\mathrm{sqrt}}(A^{1/2}) \leq \frac{1}{2}\kappa_F(A^{1/2}).$$

Proof. Since A is positive definite matrix, therefore we can assume that the eigenvalues of A satisfies: $0 < \lambda_n \leq \lambda_{n-1} \leq \cdots \leq \lambda_2 \leq \lambda_1$. Also, since A is positive definite, hence it is normal, and correspondingly X is normal as well. Therefore we can use the equality (2.1),

$$\kappa_{\mathrm{sqrt}} = \frac{1}{2\sqrt{\lambda_n}} \frac{\|A\|_F}{\|X\|_F} = \frac{\|A^{-1}\|_2^{1/2}}{2} \frac{\|A\|_F}{\|A^{1/2}\|_F}.$$

Any other primary square root X has eigenvalues μ_j with moduli $\sqrt{\lambda_j}$, so the upper bound of $\kappa_{\mathrm{sqrt}}(X)$ follows from

$$\min_{i,j=1:n} |\mu_i + \mu_j| \leq \min_{i,j=1:n} |\mu_i| + |\mu_j| \leq 2\sqrt{\lambda_n}$$

together with the fact that $\|X\|_F^2 = \sum_{i=1}^n \lambda_i$ is the same for all primary square roots of A. The upper bound and the lower bound of $\kappa_{\mathrm{sqrt}}(A^{1/2})$ are comes from standard norm inequalities. [4, 2002, Chap. 6]. □

The next results shows an elegant bound for the difference between the principal square roots of two matrices.

Theorem 2.4. *If $A, B \in \mathbb{C}^{n\times n}$ are Hermitian positive definite, then for any unitarily invariant norm*

$$\|A^{1/2} + B^{1/2}\| \leq \frac{1}{\lambda_{\min}(A)^{1/2} + \lambda_{\min}(B)^{1/2}} \|A - B\|,$$

where $\lambda_{\min}$ denotes the smallest eigenvalue.

Proof. This is a special case of [7, 1980, Prop. 3.2]. □

2.2 Perturbation

Suppose $\widetilde{X} = X + E$ be an approximation to a square root X of $A \in \mathbb{C}^{n\times n}$, where $\|E\| \leq \epsilon\|X\|$. Then $\widetilde{X}^2 = A + XE + EX + E^2$, and this leads to the relative residual bound:

$$\frac{\|A - \widetilde{X}^2\|}{\|A\|} \leq (2\epsilon + \epsilon^2)\alpha(X),$$

where

$$\alpha(X) = \frac{\|X\|^2}{\|A\|} = \frac{\|X\|^2}{\|X^2\|} \geq 1. \tag{2.3}$$

The quantity $\alpha(X)$ can be regarded as a condition number for the relative residual of X. If it is large, then *a small perturbation of X (such as $fl(X)$, which is the rounded square root) can have a relative residual much larger than the size of the relative perturbation.* Therefore the conclusion is that, we cannot expect a numerical method to do better than provide a computed square root $\widehat{X}$ with relative residual of order $\alpha(\widehat{X})u$, where u is the unit roundoff.

It is easy to show that

$$\frac{\kappa(X)}{\kappa(A)} \leq \alpha(X) \leq \kappa(X).$$

Proof. Notice that $\|X^{-1}\| = \|X^{-1}X^{-1}X\| \leq \|A^{-1}\|\|X\|$. Hence, we have

$$\frac{\|X^{-1}\|}{\|A^{-1}\|} \leq \|X\| \quad \Longrightarrow \quad \frac{\|X\|\|X^{-1}\|}{\|A\|\|A^{-1}\|} \leq \frac{\|X\|^2}{\|A\|},$$

this is equivalent as $\kappa(X)/\kappa(A) \leq \alpha(X)$.

The right-hand side can be viewed by the following inequality:

$$\alpha(X) = \frac{\|X\|^2\|X^{-1}\|}{\|A\|\|X^{-1}\|} = \frac{\kappa(X)\|X\|}{\|A\|\|X^{-1}\|} \leq \frac{\kappa(X)\|X\|}{\|X\|} = \kappa(X),$$

where the inequality comes from $\|X\| = \|AX^{-1}\| \leq \|A\|\|X^{-1}\|$. Thus a large value of $\alpha(X)$ implies that X is ill-conditioned, and if A is well-conditioned, then $\alpha(X) \approx \kappa(X)$. If X is normal, then $\alpha(X) = 1$ in 2-norm. □

3 Schur Method

Theorem 3.1 (Schur decomposition). *Let $A \in \mathbb{C}^{n\times n}$. Then there exists a unitary matrix U and an upper triangular matrix T such that $U^*AU = T$, that is, $A = UTU^*$.*

Theorem 3.2 (real Schur decomposition). *Let $A \in \mathbb{R}^{n\times n}$. Then there exists an orthogonal matrix U and an upper quasi-triangular matrix T such that $U^TAU = T$. Here,*

$$T = \begin{bmatrix} T_{11} & T_{12} & \cdots & T_{1m} \\ 0 & T_{22} & \cdots & T_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & T_{mm} \end{bmatrix},$$

where each T_{ii} is either 1×1 or 2×2 complex conjugate eigenvalues.

3.1 General Schur Decomposition

Let $A \in \mathbb{C}^{n\times n}$ be nonsingular, and let $f(A)$ denotes any primary square root of A. Given a Schur decomposition $A = QTQ^*$, where Q is unitary and T is upper triangular, and $f(A) = Qf(T)Q^*$. Hence, computing the square root of A reduces to computing the square roots $U = f(T)$ of upper triangular T. The (i,i) and $(i,j)(j>i)$ elements of the equation $U^2 = T$ can be written as

$$\begin{aligned} u_{ii}^2 &= t_{ii}, \\ (u_{ii}+u_{jj})u_{ij} &= t_{ij} - \sum_{k=i+1}^{j-1} u_{ik}u_{kj}. \end{aligned} \tag{3.1}$$

We can compute the diagonal of U and then solve for the u_{ij} either a superdiagonal at a time or a column at a time. We have the algorithm 1.

Algorithm 1 (Schur Method). Given a nonsingular $A \in \mathbb{C}^{n\times n}$, this algorithm computes $X = \sqrt{A}$ via a Schur decomposition, where $\sqrt{\cdot}$ denotes any primary square root.

1: Compute a (complex) Schur decomposition $A = QTQ^*$.
2: $u_{ii} = \sqrt{t_{ii}},\ i = 1:n$
3: **for** $j = 2:n$ **do**
4: **for** $i = j-1:-1:1$ **do**
5: $u_{ij} = \dfrac{t_{ij} - \sum_{k=i+1}^{j-1} u_{ik}u_{kj}}{u_{ii}+u_{jj}}$
6: **end for**
7: **end for**
8: $X = QUQ^*$

The *cost*: $25n^3$ flops for the Schur decomposition plus $n^3/3$ for U and $3n^3$ to form X, which gives $28\frac{1}{3}n^3$ flops in total.

Algorithm 1 generates all the primary square roots of A as different choices of sign in $u_{ii} = \sqrt{t_{ii}} = \pm t_{ii}^{1/2}$ are used.

3.2 Real Schur Decomposition

If A is real but has some nonreal eigenvalues, then Algorithm 1 uses complex arithmetic. This is *undesirable*, because (i) complex arithmetic is more expensive than real arithmetic, and (ii) rounding errors may cause a computed result to be produced with nonzero imaginary part. We can use the real Schur decomposition instead to avoid complex arithmetic.

Let $A \in \mathbb{R}^{n\times n}$ have the real Schur decomposition $A = QRQ^T$, where Q is orthogonal and R is upper quasi-triangular with 1×1 and 2×2 diagonal blocks. Then $f(A) = Qf(R)Q^T$, where $U = f(R)$ is upper quasi-triangular with the same block structure as R. The equation $U^2 = R$ can be written as in an analogous way as (3.1):

$$\begin{aligned} U_{ii}^2 &= R_{ii} \\ U_{ii}U_{ij} + U_{ij}U_{jj} &= R_{ij} - \sum_{k=i+1}^{j-1} U_{ik}U_{kj}. \end{aligned} \tag{3.2}$$

Once the diagonal blocks U_{ii} are computed, we can use (3.2) to compute the remaining blocks U_{ij} a block superdiagonal or a block column at a time. The condition for the Sylvester equation (3.2) to have a unique solution U_{ij} is that *U_{ii} and $-U_{jj}$ have no eigenvalue in common, and this is guaranteed for any primary square root when A is nonsingular.* ♡(why?) When neither U_{ii} nor U_{jj} is a scalar, (3.2) can solved by writing in the form

$$(I \otimes U_{ii} + U_{jj}^T \otimes I)\mathrm{vec}(U_{ij}) = \mathrm{vec}(R_{ij} - \sum_{k=i+1}^{j-1} U_{ik}U_{kj}),$$

which is a linear system $Ax = b$ of order 4 that can be solved by Gaussian elimination with partial pivoting.

We now consider the computation of $\sqrt{R_{ii}}$ for 2×2 blocks R_{ii}, which is necessarily have distinct complex conjugate eigenvalues. (refer to the properties of the Schur decomposition).

Lemma 3.3. *Let $A \in \mathbb{R}^{2\times 2}$ have distinct complex conjugate eigenvalues. Then A has four square roots, and all of them are primary functions of A. Two of them are real, with complex conjugate eigenvalues; two are pure imaginary, with eigenvalues that are not complex conjugates.*

Proof. Since A has distinct eigenvalues $\theta \pm \mathrm{i}\mu$, then A has four square roots, all of them are functions of A [5, 2008, Thm 1.26]. Remain to construct them. Consider the following matrix construction:

$$Z^{-1}AZ = \mathrm{diag}(\lambda, \overline{\lambda}) = \theta I + \mathrm{i}\mu K, \quad K = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

Then $A = \theta I + \mu W$, where $W = \mathrm{i}ZKZ^{-1}$, and since $\theta, \mu \in \mathbb{R}$, it follows that $W \in \mathbb{R}^{2\times 2}$. Suppose $(\alpha + \mathrm{i}\beta)^2 = \theta + \mathrm{i}\mu$, then all four square roots of A are given by $X = ZDZ^{-1}$, where $D = \pm\,\mathrm{diag}(\alpha + \mathrm{i}\beta, \pm(\alpha - \mathrm{i}\beta))$. Using previous notation, we can categorise the diagonal matrix D into two distinct cases:

$$D_1 = \pm(\alpha I + \mathrm{i}\beta K), \qquad D_2 = \pm\mathrm{i}(\beta I - \mathrm{i}\alpha K).$$

Thus, we can reconstruct the matrix square root X in the following two ways: Real matrix with complex conjugate eigenvalues $\lambda(X_1) = \pm(\alpha + \mathrm{i}\beta, \alpha - \mathrm{i}\beta)$.

$$X_1 = ZD_1Z^{-1} = \pm(\alpha I + \beta \mathrm{i} ZKZ^{-1}) = \pm(\alpha I + \beta W) \in \mathbb{R}^{2\times 2}.$$

Pure imaginary matrix with non complex conjugate eigenvalues $\lambda(X_2) = \pm(\alpha+\mathrm{i}\beta, -\alpha+\mathrm{i}\beta)$.

$$X_2 = ZD_2Z^{-1} = \pm\mathrm{i}(\beta I - \alpha \mathrm{i} ZKZ^{-1}) = \pm\mathrm{i}(\beta I - \alpha W).$$

□

The proof gives a way to construct $R_{ii}^{1/2}$, writting

$$R_{ii} = \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix}$$

the eigenvalues of R_{ii} are $\theta \pm \mathrm{i}\mu$, where

$$\theta = \frac{1}{2}(r_{11} + r_{22}), \qquad \mu = \frac{1}{2}\left(-(r_{11} - r_{22})^2 - 4r_{21}r_{12}\right)^{1/2}.$$

The former equation is constructed using $\mathrm{trace}(A) = \sum_i \lambda_i(A)$. The latter equation is constructed using $(\theta + \mathrm{i}\mu)(\theta - \mathrm{i}\mu) = \theta^2 + \mu^2$.

We now require α and β such that $(\alpha + \mathrm{i}\beta)^2 = \theta + \mathrm{i}\mu$. A stable way to compute them is as follows:

Algorithm 2 This algorithm computes the square root $\alpha + \mathrm{i}\beta$ of $\theta + \mathrm{i}\mu$ with $\alpha \geq 0$.

1: If $\theta = 0$ and $\mu = 0$, then $\alpha = \beta = 0$, quit, end.
2: $t = \left(\dfrac{|\theta| + (\theta^2 + \mu^2)^{1/2}}{2}\right)^{1/2}$
3: if $\theta \geq 0$
4: $\quad \alpha = t, \beta = \mu/(2\alpha)$
5: else
6: $\quad \beta = t, \alpha = \mu/(2\beta)$
7: end

Finally, the real square roots of R_{ii} are obtained from

$$\begin{aligned} U_{ii} &= \pm(\alpha I + \frac{1}{2\alpha}(R_{ii} - \theta I)) \\ &= \pm \begin{bmatrix} \alpha + \frac{1}{4\alpha}(r_{11} - r_{22}) & \frac{1}{2\alpha} r_{12} \\ \frac{1}{2\alpha} r_{21} & \alpha - \frac{1}{4\alpha}(r_{11} - r_{22}) \end{bmatrix}. \end{aligned} \tag{3.3}$$

Before present the algorithm, let us discuss the number of real square roots for arbitrary $A \in \mathbb{R}^{n\times n}$.

Theorem 3.4. *Let $A \in \mathbb{R}^{n\times n}$ be nonsingular. If A has a real negative eigenvalue, then A has no real square root that is primary function of A.*

If A has no real negative eigenvalues, then there are precisely 2^{r+c} real primary square roots of A, where r is then number of distinct real eigenvalues and c is the number of distinct complex conjugate eigenvalue pairs.

Proof. Let A has a real Schur decomposition. Since $f(A) = Qf(R)Q^T$, $f(A)$ is real if and only if $f(R)$ is real. If A has a real negative eigenvalue, $R_i = (r_{ii})$ say, then $f(R_i)$ is necessarily nonreal, and this gives the first part of the proof.

If A has no real negative eigenvalues, consider the 2^s primary square roots of A described in Theorem 1.26. We have $s = r + 2c$, i.e. A has 2^r real primary square roots from real, distinct eigenvalues; also, a half of 2^{2c} primary square roots from complex conjugate eigenvalues are *real* according to Lemma 3.3, hence precisely 2^{r+c} of its primary square roots are real. □

Algorithm 3 (real Schur method). Given $A \in \mathbb{R}^{n\times n}$ with no eigenvalues on $\mathbb{R}^-$, this algorithm computes $X = \sqrt{A}$ via a Schur decomposition, where $\sqrt{\cdot}$ denotes any real primary square root.

Compute a real Schur decomposition, $A = QRQ^T$, where R is block $m \times m$.
Compute $U_{ii} = \sqrt{R_{ii}}$, $i = 1 : m$, using (3.3) whenever R_{ii} is 2×2.
for $j = 1 : m$ **do**
  for $i = j - 1 : -1 : 1$ **do**
    Solve $U_{ii}U_{ij} + U_{ij}U_{jj} = R_{ij} - \sum_{k=i+1}^{j-1} U_{ik}U_{kj}$ for U_{ij}.
  end for
end for
$X = QUQ^T$.

Remark 3.5.

1. The principal square root is computed if the principal square root is taken at line 2, which for 2×2 blocks means taking the positive sign in (3.3).

2. Second, as for 1, it is necessary that whenever R_{ii} and R_{jj} have the same eigenvalues, we take the same square root.

3.3 Numerical stability

Now we consider the numerical stability of Algorithm 1 and 3. A straightforward rounding error analysis shows that the computed square root $\widehat{U}$ of T in Algorithm 1 satisfies

$$\widehat{U}^2 = T + \Delta T, \quad |\Delta T| \le \tilde{\gamma}_n |\widehat{U}|^2$$

where the inequality is to be interpreted elementwise. Computation of the Schur decomposition by QR algorithm is a backward stable process [3, 2013, Sec. 7.5.6] and standard error analysis leads to the overall result

$$\widehat{X}^2 = A + \Delta A, \quad \|\Delta A\|_F \le \tilde{\gamma}_{n^3} \|\widehat{X}\|_F^2$$

which can be expressed as

$$\frac{\|A - \widehat{X}^2\|_F}{\|A\|_F} \le \tilde{\gamma}_{n^3} \alpha_F(\widehat{X}) \tag{3.4}$$

where α is defined in (2.3). The same conclusion holds for Algorithm 3, which can be shown to satisfy the same error bound (3.4).

4 Newton's Method and Its Variants

4.1 Newton and Commutativity

Newton method for solving $X^2 = A$ can be derived as follows. Let Y be an approximate solution, and set $Y + E = X$, where E is to be determined. Then $A = (Y + E)^2 = Y^2 + EY + YE + E^2$. Dropping the second order term in E leads to the Newton's method:

$$\left.\begin{aligned} X_0 \text{ given} \\ \text{Solve } X_k E_k + E_k X_k = A - X_k^2 \\ X_{k+1} = X_k + E_k \end{aligned}\right\} \quad k = 0, 1, 2, \ldots, \tag{4.1}$$

At each iteration, a Sylvester equation mush be solved for E_k. The standard way of solving the Sylvester equation is via Schur decomposition of the coefficient matrices, which in this case are both X_k. But the Schur method of the previous section can compute a square root with just one Schur decomposition, so Newton's method is unduly expensive in the form (4.1).

The following lemma enable us to reduce the cost. Note that the E_k in (4.1) is well defined, that is, the Sylvester equation is nonsingular, if and only if X_k and $-X_k$ have no eigenvalues in common.

Lemma 4.1. *Suppose that in the Newton iteration (4.1), X_0 commutes with A and all the iterates are well-defined. Then for all k, X_k commutes with A and $X_{k+1} = \frac{1}{2}(X_k + X_k^{-1}A)$.*

Proof. We prove this by induction. Firstly, we notice that if X_k and A are commute, then there is a trivial solution for the Sylvester equation $X_k E_k + E_k X_k = A - X_k^2$, which is $E_k = \frac{1}{2}(X_k^{-1}A - X_k)$ and this solution is clearly commute with A. Hence, we will stick to this solution and start for induction.
Induction Statement. For $k = 1, 2, \ldots$, X_k commute with A and $X_{k+1} = \frac{1}{2}(X_k + X_k^{-1}A)$.
Base case. For $k = 1$, we have $X_1 = X_0 + E_0$, where $E_0 = \frac{1}{2}(X_0^{-1}A - X_0)$, and

$$X_1 = X_0 + \frac{1}{2}(X_0^{-1}A - X_0) = \frac{1}{2}(X_0^{-1}A + X_0).$$

Here, X_1 is commute with A, and E_0 is commute with A.
Inductive step. Suppose the statement is true for $k = n-1$, i.e. X_{n-1} commute with A and $X_n = \frac{1}{2}(X_{n-1}^{-1}A + X_{n-1})$ which is also commute with A by $AX_{n-1} = X_{n-1}A$. By solution of the Sylvester equation, we have $E_n = \frac{1}{2}(X_n^{-1}A - X_n)$, and this is clearly commute with A by the commutativity between X_n and A. Moreover, $X_{n+1} = \frac{1}{2}(X_n^{-1}A + X_n)$ which is also commute with A by A and X_n are commute. The proof is then complete. □

This lemma shows that if X_0 is chosen to commute with A, then all the X_k and E_k in (4.1) commute with A, permitting a good simplification of the iteration. The most common choice of X_0 is A, giving the Newton iteration

Newton iteration (matrix square root)

$$X_{k+1} = \frac{1}{2}(X_k + X_k^{-1}A), \qquad X_0 = A. \tag{4.2}$$

4.2 Link to the Matrix Sign Function

If A is nonsingular, standard convergence theory for Newton's method allows us to deduce quadratic convergence of (4.1) to a primary square root for X_0 sufficiently close to square root, since the Fréchet derivative of $F(X) = X^2 - A$ is nonsingular at a primary square root. The next result shows the unconditional quadratic convergence of (4.2) to the *principal* square root. Moreover, it shows that (4.2) is equivalent to the Newton sign iteration (Sign)

Newton iteration (matrix sign function)

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

Theorem 4.2 (Convergence of Newton square root iteration)**.** *Let $A \in \mathbb{C}^{n\times n}$ has no eigenvalues on $\mathbb{R}^-$. The Newton square root iterates X_k from (4.2) with any X_0 that commutes with A are related to the Newton sign iterates*

$$S_{k+1} = \frac{1}{2}(S_k + S_k^{-1}), \quad S_0 = A^{-1/2}X_0,$$

by $X_k \equiv A^{1/2}S_k$. Hence provided that $A^{-1/2}X_0$ has no pure imaginary eigenvalues, the X_k are defined and X_k converges quadratically to $A^{1/2}\mathrm{sign}(A^{-1/2}X_0)$.

In particular, if the spectrum of $A^{-1/2}X_0$ lies in the right half-plane then X_k converges quadratically to $A^{1/2}$ and, for any consistent norm,

$$\|X_{k+1} - A^{1/2}\| \le \frac{1}{2}\|X_k^{-1}\|\|X_k - A^{1/2}\|^2 \tag{4.3}$$

Proof. We first note that any matrix that commutes with A commutes with $A^{\pm 1/2}$ since it's a polynomial of A. We have $X_0 = A^{1/2}S_0$, hence S_0 commute with A.

Assume that $X_k = A^{1/2}S_k$, and S_k commute with A, then S_k commute with $A^{1/2}$, and

$$X_{k+1} = \frac{1}{2}(A^{1/2}S_k + S_k^{-1}A^{-1/2}A) = A^{1/2}\cdot\frac{1}{2}(S_k + S_k^{-1}) = A^{1/2}S_{k+1},$$

and S_{k+1} clearly commute with A. Hence $X_k \equiv A^{1/2}S_k$ by induction. Then using [5, 2008, Theorem 5.6],

$$\lim_{k\to\infty} X_k = A^{1/2}\lim_{k\to\infty} S_k = A^{1/2}\mathrm{sign}(S_0) = A^{1/2}\mathrm{sign}(A^{-1/2}X_0),$$

and the quadratic convergence of X_k follows from that of S_k.

For the last part, if $S_0 = A^{-1/2}X_0$ has spectrum in the right half-plane then $\mathrm{sign}(S_0) = I$ and hence $X_k \to A$. Using the commutativity of the iterates with A, it is easy to show that

$$X_{k+1} \pm A^{1/2} = \frac{1}{2}X_k^{-1}(X_k \pm A^{1/2})^2 \tag{4.4}$$

which, with minus sign, gives (4.3). □

Remark 4.3.

1. An implication of Theorem 4.2 of theoretical interest, which can also be deduced from the connection with the full Newton method, is that (4.2) converges to $A^{1/2}$ for any X_0 that commutes with A and is sufficiently close to $A^{1/2}$.

2. It is worth noting that the sequence X_k from (4.2) may be well-defined when that for the full Newton method (4.1) is not. No analogue of the condition in 4.2 guaranteeing that the X_k is well-defined is available for (4.1).

3. This analysis using the Newton sign iteration is more powerful than the analysis using the Jordan canonical form in [5, Section 4.9.3]. If X_0 is only known to commute with A, then X_k do not necessarily share the same Jordan block as A, so the analysis cannot break down to one single Jordan form.

4. Note that, from [5, Section 5.3], the Newton iteration for sign(A) requires more iterations if A has eigenvalues close to the imaginary axis. Theorem 4.2 therefore implies that the Newton iteration (4.2) requires more iterations if A has eigenvalues close to the negative real axis.

5. When A is positive definite, the convergence of (4.2) is monotonic from above in the positive semidefinite ordering.

6. It's interesting to consider how (4.2) behaves when X_0 does not commute with A, although commutativity is assumed in the derivation. Lack of commutativity can cause quadratic convergence, and even convergence itself, to be lost.

4.3 Other Variants of Newton Iteration

A coupled version of (4.2) can be obtained by defining $Y_k = A^{-1}X_k$. Then $X_{k+1} = \frac{1}{2}(X_k + Y_k^{-1})$ and $Y_{k+1} = A^{-1}X_{k+1} = \frac{1}{2}(Y_k + Y_k^{-1})$ on using the fact that X_k commute with A. This is the iteration of Denman and Beavers [2]

DB iteration

$$\begin{aligned} X_{k+1} &= \frac{1}{2}(X_k + Y_k^{-1}), \quad X_0 = A, \\ Y_{k+1} &= \frac{1}{2}(Y_k + X_k^{-1}), \quad Y_0 = I. \end{aligned} \tag{4.5}$$

Under the condition of 4.2

$$\lim_{k\to\infty} X_k = A^{1/2}, \qquad \lim_{k\to\infty} Y_k = A^{-1/2}. \tag{4.6}$$

Defining $M_k = X_kY_k$, we have

$$M_{k+1} = \frac{1}{2}(X_kY_k^{-1})\frac{1}{2}(Y_k + X_k^{-1}) = \frac{1}{4}(2I + M_k + M_k^{-1}),$$

gives the product form of the DB iteration, identified by Cheng, Higham, Kenney and Laub [1] in which we iterates with M_k and either X_k or Y_k:

Product form DB iteration

$$\begin{aligned} M_{k+1} &= \frac{1}{2}\left(I + \frac{M_k + M_k^{-1}}{2}\right), \quad M_0 = A, \\ X_{k+1} &= \frac{1}{2}X_k(I + M_k^{-1}), \quad X_0 = A, \\ Y_{k+1} &= \frac{1}{2}Y_k(I + M_k^{-1}), \quad Y_0 = I. \end{aligned} \tag{4.7}$$

Clearly, $\lim_{k\to\infty} M_k = I$. The product form DB iteration has the advantage in efficiency over DB iteration that it has trade one of the matrix inversions for a matrix multiplication.

Another attraction of (4.7) is that, a convergence test can be based on the error $\|M_k - I\|$, which is available free of charge.

Yet another variant of (4.2) can be derived by noting that

$$\begin{aligned} E_{k+1} &= \frac{1}{2}(X_{k+1}^{-1}A - X_{k+1}) \\ &= \frac{1}{2}(X_{k+1}^{-1})(A - X_{k+1}^2) \\ &= \frac{1}{2}(X_{k+1}^{-1})\left(A - \frac{1}{4}(X_k + X_k^{-1}A)\right) \\ &= \frac{1}{2}X_{k+1}^{-1}\left(\frac{2A - X_k^2 - X_k^{-2}A^2}{4}\right) \\ &= -\frac{1}{2}X_{k+1}^{-1}\frac{(X_k - X_k^{-1}A)^2}{4} \\ &= -\frac{1}{2}X_{k+1}^{-1}E_k^2 = -\frac{1}{2}E_k X_{k+1}^{-1}E_k. \end{aligned}$$

Setting $Y = 2E_k$ and $Z_k = 4X_{k+1}$, we obtain the iteration

CR iteration

$$\begin{aligned} Y_{k+1} &= -Y_k Z_k^{-1} Y_k, \quad Y_0 = I - A, \\ Z_{k+1} &= Z_k + 2Y_{k+1}, \quad Z_0 = 2(I + A) \end{aligned} \tag{4.8}$$

From the derivation, we have $Y_k \to 0$ and $Z \to 4A^{1/2}$. This iteration is derived in a different way by Meini [6].

A minor variation of (4.8) is: set $X_k = Z_k/4$ and $E_k = Y_{k+1}/2$, then (4.8) becomes

IN iteration

$$\begin{aligned} X_{k+1} &= X_k + E_k, \quad X_0 = A, \\ E_{k+1} &= -\frac{1}{2}E_k X_{k+1}^{-1}E_k, \quad E_0 = \frac{1}{2}(I - A). \end{aligned} \tag{4.9}$$

Here $X_k \to A^{1/2}$ and $E_k \to 0$. This incremental form of the Newton iteration, suggested by Iannazzo, is of interest because it update X_k by a correction that is small and accurately computable.

The computational cost of the Newton iteration and its variants is compared in Table 1.

Iteration	Operations	Flops
Newton, (4.2)	D	$8n^3/3$
DB, (4.5)	$2I$	$4n^3$
Product DB, (4.7)	$M + I$	$4n^3$
CR, (4.8)	$M + D$	$14n^3/3$
IN, (4.9)	$M + D$	$14n^3/3$

Table 1: Cost per iteration of matrix square root iterations

Here M denotes a matrix multiplication, I denotes a matrix inversion, and D denotes a solution of a multiple right-hand side linear system. Clearly, the Newton iteration (4.2) is the least expensive iteration.

References

[1] Sheung Hun Cheng, Nicholas J. Higham, Charles S. Kenney, and Alan J. Laub. Approximating the logarithm of a matrix to specified accuracy. *SIAM Journal on Matrix Analysis and Applications*, 22(4):1112–1125, 2001. (Cited on p. 11.)

[2] Eugene D. Denman and Alex N. Beavers. The matrix sign function and computations in systems. *Applied Mathematics and Computation*, 2(1):63–94, 1976. (Cited on p. 11.)

[3] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins studies in the mathematical sciences. Fourth edition, Johns Hopkins University Press, Baltimore, Maryland, 2013. ISBN 978-1-4214-0794-4. (Cited on p. 8.)

[4] Nicholas J. Higham. *Accuracy and Stability of Numerical Algorithms*. Second edition, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, jan 2002. xxx+680 pp. ISBN 0-89871-521-0. (Cited on p. 3.)

[5] Nicholas J. Higham. *Functions of Matrices: Theory and Computation*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2008. xx+425 pp. ISBN 978-0-898716-46-7. (Cited on pp. 1, 2, 6, 10, and 11.)

[6] Beatrice Meini. The matrix square root from a new functional perspective: Theoretical results and computational issues. *SIAM Journal on Matrix Analysis and Applications*, 26(2):362–376, 2004. (Cited on p. 12.)

[7] J. L. van Hemmen and T. Ando. An inequality for trace ideals. *Communications in Mathematical Physics*, 76(2):143–148, 1980. (Cited on p. 4.)