MATHEMATICS OF COMPUTATION
VOLUME 46, NUMBER 174
APRIL 1986, PAGES 537–549

Newton's Method for the Matrix Square Root*

By Nicholas J. Higham

Abstract. One approach to computing a square root of a matrix A is to apply Newton's method to the quadratic matrix equation $F(X) \equiv X^2 - A = 0$. Two widely-quoted matrix square root iterations obtained by rewriting this Newton iteration are shown to have excellent mathematical convergence properties. However, by means of a perturbation analysis and supportive numerical examples, it is shown that these simplified iterations are numerically unstable. A further variant of Newton's method for the matrix square root, recently proposed in the literature, is shown to be, for practical purposes, numerically stable.

1. Introduction. A square root of an $n \times n$ matrix A with complex elements, $A \in \mathbf{C}^{n\times n}$, is a solution $X \in \mathbf{C}^{n\times n}$ of the quadratic matrix equation

$$F(X) \equiv X^2 - A = 0. \tag{1.1}$$

A natural approach to computing a square root of A is to apply Newton's method to (1.1). For a general function G: $\mathbf{C}^{n\times n} \to \mathbf{C}^{n\times n}$, Newton's method for the solution of $G(X) = 0$ is specified by an initial approximation X_0 and the recurrence (see [14, p. 140], for example)

$$X_{k+1} = X_k - G'(X_k)^{-1}G(X_k), \qquad k = 0,1,2,\ldots, \tag{1.2}$$

where G' denotes the Fréchet derivative of G. Identifying

$$F(X + H) = X^2 - A + (XH + HX) + H^2$$

with the Taylor series for F we see that $F'(X)$ is a linear operator, $F'(X)$: $\mathbf{C}^{n\times n} \to \mathbf{C}^{n\times n}$, defined by

$$F'(X)H = XH + HX.$$

Thus Newton's method for the matrix square root can be written

$$X_0 \text{ given},$$

$$\text{(N):} \quad \left.\begin{aligned} X_kH_k + H_kX_k &= A - X_k^2 \quad (1.3)\\ X_{k+1} &= X_k + H_k \quad (1.4) \end{aligned}\right\}, \qquad k = 0,1,2,\ldots.$$

Applying the standard local convergence theorem for Newton's method [14, p. 148], we deduce that the Newton iteration (N) converges quadratically to a square root X of A if $\|X - X_0\|$ is sufficiently small and if the linear transformation $F'(X)$ is nonsingular. However, the most stable and efficient methods for solving Eq. (1.3),

Received October 22, 1984; revised July 30, 1985.
1980 *Mathematics Subject Classification*. Primary 65F30, 65H10.
Key words and phrases. Matrix square root, Newton's method, numerical stability.
* This work was carried out with the support of a SERC Research Studentship.

0025-5718/86 $1.00 + $.25 per page

[1], [6], require the computation of a Schur decomposition of X_k, assuming X_k is full. Since a square root of A can be obtained directly and at little extra cost once a single Schur decomposition, that of A, is known, [2], [9], we see that in general Newton's method for the matrix square root, in the form (N), is computationally expensive.

It is therefore natural to attempt to "simplify" iteration (N). Since X commutes with $A = X^2$, a reasonable assumption (which we will justify in Theorem 1) is that the commutativity relation

$$X_k H_k = H_k X_k$$

holds, in which case (1.3) may be written

$$2X_k H_k = 2H_k X_k = A - X_k^2,$$

and we obtain from (N) two new iterations

$$\text{(I):} \quad Y_{k+1} = \tfrac{1}{2}\left(Y_k + Y_k^{-1}A\right), \tag{1.5}$$

$$\text{(II):} \quad Z_{k+1} = \tfrac{1}{2}\left(Z_k + AZ_k^{-1}\right). \tag{1.6}$$

These iterations are well-known; see for example [2], [7, p. 395], [11], [12], [13].

Consider the following numerical example. Using iteration (I) on a machine with approximately nine decimal digit accuracy, we attempted to compute a square root of the symmetric positive definite Wilson matrix [16, pp. 93, 123]

$$W = \begin{bmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{bmatrix},$$

for which the 2-norm condition number $\kappa_2(W) = \|W\|_2\|W^{-1}\|_2 \approx 2984$. Two implementations of iteration (I) were employed (for the details see Section 5). The first is designed to deal with general matrices, while the second is for the case where A is positive definite and takes full advantage of the fact that all iterates are (theoretically) positive definite (see Corollary 1). In both cases we took $Y_0 = I$; as we will prove in Theorem 2, for this starting value iteration (I) should converge quadratically to $W^{1/2}$, the unique symmetric positive definite square root of W.

Denoting the computed iterates by $\hat{Y}_k$, the results obtained were as in Table 1. Both implementations failed to converge; in the first, $\hat{Y}_{20}$ was unsymmetric and indefinite. In contrast, a further variant of the Newton iteration, to be defined in Section 4, converged to $W^{1/2}$ in nine iterations.

Clearly, iteration (I) is in some sense "numerically unstable". This instability was noted by Laasonen [13] who, in a paper apparently unknown to recent workers in this area, stated without proof that for a matrix with real, positive eigenvalues iteration (I) "if carried out indefinitely, is not stable whenever the ratio of the largest to the smallest eigenvalue of A exceeds the value 9". We wish to draw attention to this important and surprising fact. In Section 3 we provide a rigorous proof of Laasonen's claim. We show that the original Newton method (N) does not suffer from this numerical instability and we identify in Section 4 an iteration, proposed in [4], which has the computational simplicity of iteration (I) and yet does not suffer from the instability which impairs the practical performance of (I).

TABLE 1

	Implementation 1,	Implementation 2
k	$\|W^{1/2} - \hat{Y}_k\|_1$	$\|W^{1/2} - \hat{Y}_k\|_1$
0	4.9	4.9
1	1.1×10^{1}	1.1×10^{1}
2	3.6	3.6
3	6.7×10^{-1}	6.7×10^{-1}
4	3.3×10^{-2}	3.3×10^{-2}
5	4.3×10^{-4}	4.3×10^{-4}
6	3.4×10^{-5}	6.7×10^{-7}
7	9.3×10^{-4}	1.4×10^{-6}
8	2.5×10^{-2}	1.6×10^{-5}
9	6.7×10^{-1}	2.0×10^{-4}
10	1.8×10^{1}	2.4×10^{-3}
11	4.8×10^{2}	2.8×10^{-2}
12	1.3×10^{4}	3.2×10^{-1}
13	3.4×10^{5}	Error: $\hat{Y}_k$ not positive definite
20	1.2×10^{6}	

We begin by analyzing the mathematical convergence properties of the Newton iteration.

2. Convergence of Newton's Method. In this section we derive conditions which ensure the convergence of Newton's method for the matrix square root and we establish to which square root the method converges for a particular set of starting values. (For a classification of the set $\{X\colon X^2 = A\}$ see, for example, [9].)

First, we investigate the relationship between the Newton iteration (N) and its offshoots (I) and (II). To begin, note that the Newton iterates X_k are well-defined if and only if, for each k, Eq. (1.3) has a unique solution, that is, the linear transformation $F'(X_k)$ is nonsingular. This is so if and only if X_k and $-X_k$ have no eigenvalue in common [7, p. 194], which requires in particular that X_k be nonsingular.

THEOREM 1. *Consider the iterations* (N), (I) *and* (II). *Suppose* $X_0 = Y_0 = Z_0$ *commutes with* A *and that all the Newton iterates* X_k *are well-defined. Then*

(i) X_k *commutes with* A *for all* k,

(ii) $X_k = Y_k = Z_k$ *for all* k.

Proof. We sketch an inductive proof of parts (i) and (ii) together. The case $k = 0$ is given. Assume the results hold for k. From the remarks preceding the theorem we see that both the linear transformation $F'(X_k)$, and the matrix X_k, are nonsingular. Define

$$G_k = \tfrac{1}{2}\left(X_k^{-1}A - X_k\right).$$

Using $X_k A = A X_k$ we have, from (1.3),

$$F'(X_k)G_k = F'(X_k)H_k.$$

Thus $H_k = G_k$, and so from (1.4),

$$X_{k+1} = X_k + G_k = \tfrac{1}{2}\left(X_k + X_k^{-1}A\right), \tag{2.1}$$

which commutes with A. It follows easily from (2.1) that $X_{k+1} = Y_{k+1} = Z_{k+1}$. □

Thus, provided the initial approximation $X_0 = Y_0 = Z_0$ commutes with A and the correction equation (1.3) is nonsingular at each stage, the Newton iteration (N) and its variants (I) and (II) yield the same sequence of iterates. We now examine the convergence of this sequence, concentrating for simplicity on iteration (I) with starting value a multiple of the identity matrix. Note that the starting values $Y_0 = I$ and $Y_0 = A$ lead to the same sequence $Y_1 = \frac{1}{2}(I + A)$, $Y_2, \ldots$.

For our analysis we assume that A is diagonalizable, that is, there exists a nonsingular matrix Z such that

$$Z^{-1}AZ = \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n), \tag{2.2}$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A. The convenience of this assumption is that it enables us to diagonalize the iteration. For, defining

$$D_k = Z^{-1}Y_kZ \tag{2.3}$$

we have from (1.5),

$$D_{k+1} = \tfrac{1}{2}\left(Z^{-1}Y_kZ + \left(Z^{-1}Y_kZ\right)^{-1}Z^{-1}AZ\right) = \tfrac{1}{2}\left(D_k + D_k^{-1}\Lambda\right), \tag{2.4}$$

so that if D_0 is diagonal, then by induction all the successive transformed iterates D_k are diagonal too.

THEOREM 2. *Let $A \in \mathbf{C}^{n\times n}$ be nonsingular and diagonalizable, and suppose that none of A's eigenvalues is real and negative. Let*

$$Y_0 = mI, \qquad m > 0.$$

Then, provided the iterates $\{Y_k\}$ in (1.5) are defined,

$$\lim_{k\to\infty} Y_k = X$$

and

$$\|Y_{k+1} - X\| \leqslant \tfrac{1}{2}\|Y_k^{-1}\|\,\|Y_k - X\|^2, \tag{2.5}$$

where X is the unique square root of A for which every eigenvalue has positive real part.

Proof. We will use the notation (2.2). In view of (2.3) and (2.4) it suffices to analyze the convergence of the sequence $\{D_k\}$. $D_0 = mI$ is diagonal, so D_k is diagonal for each k. Writing $D_k = \operatorname{diag}(d_i^{(k)})$ we see from (2.4) that

$$d_i^{(k+1)} = \tfrac{1}{2}\left(d_i^{(k)} + \lambda_i/d_i^{(k)}\right), \qquad 1 \leqslant i \leqslant n,$$

that is, (2.4) is essentially n uncoupled scalar Newton iterations for the square roots $\sqrt{\lambda_i}$, $1 \leqslant i \leqslant n$.

Consider therefore the scalar iteration

$$z_{k+1} = \tfrac{1}{2}\left(z_k + a/z_k\right).$$

We will use the relations [17, p. 84]

$$z_{k+1} \pm \sqrt{a} = \left(z_k \pm \sqrt{a}\right)^2/(2z_k), \tag{2.6}$$

$$\frac{z_{k+1} - \sqrt{a}}{z_{k+1} + \sqrt{a}} = \left(\frac{z_0 - \sqrt{a}}{z_0 + \sqrt{a}}\right)^{2^{k+1}} \equiv \gamma^{2^{k+1}}. \tag{2.7}$$

If a does not lie on the nonpositive real axis then we can choose $\sqrt{a}$ to have positive real part, in which case it is easy to see that for real $z_0 > 0$, $|\gamma| < 1$. Consequently, for a and z_0 of the specified form we have from (2.7), provided that the sequence $\{z_k\}$ is defined,

$$\lim_{k\to\infty} z_k = \sqrt{a}, \qquad \operatorname{Re}\sqrt{a} > 0.$$

Since the eigenvalues λ_i and the starting values $d_i^{(0)} = m > 0$ are of the form of a and z_0, respectively, then

$$\lim_{k\to\infty} D_k = \Lambda^{1/2} = \operatorname{diag}\left(\lambda_i^{1/2}\right), \qquad \operatorname{Re}\lambda_i^{1/2} > 0, \tag{2.8}$$

and thus

$$\lim_{k\to\infty} Y_k = Z\Lambda^{1/2}Z^{-1} = X$$

(provided the iterates $\{Y_k\}$ are defined), which is clearly a square root of A whose eigenvalues have positive real part. The uniqueness of X follows from Theorem 4 in [9].

Finally, we can use (2.6), with the minus sign, to deduce that

$$D_{k+1} - \Lambda^{1/2} = \tfrac{1}{2}D_k^{-1}\left(D_k - \Lambda^{1/2}\right)^2;$$

performing a similarity transformation by Z gives

$$Y_{k+1} - X = \tfrac{1}{2}Y_k^{-1}(Y_k - X)^2,$$

from which (2.5) follows on taking norms. □

Theorem 2 shows, then, that under the stated hypotheses on A iterations (N), (I) and (II) with starting value a multiple of the identity matrix, when defined, will indeed converge: quadratically, to a particular square root of A the form of whose spectrum is known a priori.

Several comments are worth making. First, we can use Theorem 4 in [9] to deduce that the square root X in Theorem 2 is indeed a function of A, in the sense defined in [5, p. 96]. (Essentially, B is a function of A if B can be expressed as a polynomial in A.) Next, note that the proof of Theorem 2 relies on the fact that the matrix which diagonalizes A also diagonalizes each iterate Y_k. This property is maintained for Y_0 an arbitrary function of A, and under suitable conditions convergence can still be proved, but the spectrum $\{\pm\sqrt{\lambda_1}, \ldots, \pm\sqrt{\lambda_n}\}$ of the limit matrix, if it exists, will depend on Y_0. Finally, we remark that Theorem 2 can be proved without the assumption that A is diagonalizable, using, for example, the technique in [13].

We conclude this section with a corollary which applies to the important case where A is Hermitian positive definite.

COROLLARY 1. *Let $A \in \mathbf{C}^{n\times n}$ be Hermitian positive definite. If $Y_0 = mI$, $m > 0$, then the iterates $\{Y_k\}$ in* (1.5) *are all Hermitian positive definite,* $\lim_{k\to\infty} Y_k = X$, *where X is the unique Hermitian positive definite square root of A, and* (2.5) *holds.*

3. Stability Analysis. We now consider the behavior of Newton's method for the matrix square root, and its variants (I) and (II), when the iterates are subject to perturbations. We will regard these perturbations as arising from rounding errors sustained during the evaluation of an iteration formula, though our analysis is quite general.

Consider first iteration (I) with $Y_0 = mI$, $m > 0$, and make the same assumptions as in Theorem 2. Let $\hat{Y}_k$ denote the computed kth iterate, $\hat{Y}_k \approx Y_k$, and define

$$\Delta_k = \hat{Y}_k - Y_k.$$

Our aim is to analyze how the error matrix Δ_k propagates at the $(k+1)$st stage (note the distinction between Δ_k and the "true" error matrix $\hat{Y}_k - X$). To simplify the analysis we assume that no rounding errors are committed when computing $\hat{Y}_{k+1}$, so that

$$\hat{Y}_{k+1} = \tfrac{1}{2}\big(\hat{Y}_k + \hat{Y}_k^{-1}A\big) = \tfrac{1}{2}\big(Y_k + \Delta_k + (Y_k + \Delta_k)^{-1}A\big). \tag{3.1}$$

Using the perturbation result [15, p. 188 ff.]

$$(A + E)^{-1} = A^{-1} - A^{-1}EA^{-1} + O\big(\|E\|^2\big), \tag{3.2}$$

we obtain

$$\hat{Y}_{k+1} = \tfrac{1}{2}\big(Y_k + \Delta_k + Y_k^{-1}A - Y_k^{-1}\Delta_k Y_k^{-1}A\big) + O\big(\|\Delta_k\|^2\big).$$

Subtracting (1.5) yields

$$\Delta_{k+1} = \tfrac{1}{2}\big(\Delta_k - Y_k^{-1}\Delta_k Y_k^{-1}A\big) + O\big(\|\Delta_k\|^2\big). \tag{3.3}$$

Using the notation (2.2) and (2.3), let

$$\tilde{\Delta}_k = Z^{-1}\Delta_k Z, \tag{3.4}$$

and transform (3.3) to obtain

$$\tilde{\Delta}_{k+1} = \tfrac{1}{2}\big(\tilde{\Delta}_k - D_k^{-1}\tilde{\Delta}_k D_k^{-1}\Lambda\big) + O\big(\|\tilde{\Delta}_k\|^2\big). \tag{3.5}$$

From the proof of Theorem 2,

$$D_k = \operatorname{diag}\big(d_i^{(k)}\big), \tag{3.6}$$

so with

$$\tilde{\Delta}_k = \big(\tilde{\delta}_{ij}^{(k)}\big), \tag{3.7}$$

Eq. (3.5) can be written elementwise as

$$\tilde{\delta}_{ij}^{(k+1)} = \pi_{ij}^{(k)}\tilde{\delta}_{ij}^{(k)} + O\big(\|\tilde{\Delta}_k\|^2\big), \qquad 1 \leqslant i, j \leqslant n,$$

where

$$\pi_{ij}^{(k)} = \tfrac{1}{2}\big(1 - \lambda_j/\big(d_i^{(k)}d_j^{(k)}\big)\big).$$

Since $D_k \to \Lambda^{1/2}$ as $k \to \infty$ (see (2.8)) we can write

$$d_i^{(k)} = \lambda_i^{1/2} + \varepsilon_i^{(k)}, \tag{3.8}$$

where $\varepsilon_i^{(k)} \to 0$ as $k \to \infty$. Then,

$$\pi_{ij}^{(k)} = \tfrac{1}{2}\big(1 - (\lambda_j/\lambda_i)^{1/2}\big) + O(\varepsilon^{(k)}), \tag{3.9}$$

where

$$\varepsilon^{(k)} = \max_i \big|\varepsilon_i^{(k)}\big|. \tag{3.10}$$

To ensure the numerical stability of the iteration we require that the error amplification factors $\pi_{ij}^{(k)}$ be bounded in modulus by 1; hence we require, in particular, that

$$\tfrac{1}{2}\left|1-\left(\lambda_j/\lambda_i\right)^{1/2}\right| \leq 1, \qquad 1 \leq i,\ j \leq n. \tag{3.11}$$

This is a severe restriction on the matrix A. For example, if A is Hermitian positive definite the condition is equivalent to (cf. [13])

$$\kappa_2(A) \leq 9, \tag{3.12}$$

where the condition number $\kappa_2(A) = \|A\|_2\|A^{-1}\|_2$.

To clarify the above analysis it is helpful to consider a particular example. Suppose A is Hermitian positive definite, so that in (2.2) we can take $Z = Q$ where $Q = (q_1, \ldots, q_n)$ is unitary. Thus,

$$Q^*AQ = \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n), \qquad Q^*Q = I \tag{3.13}$$

and (cf. (2.3))

$$Q^*Y_kQ = D_k = \operatorname{diag}\left(d_i^{(k)}\right). \tag{3.14}$$

Consider the special (unsymmetric) rank-one perturbation

$$\Delta_k = \varepsilon q_i q_j^*, \qquad i \neq j; \qquad \|\Delta_k\|_2 = \varepsilon > 0.$$

For this Δ_k the Sherman-Morrison formula [7, p. 3] gives

$$\left(Y_k + \Delta_k\right)^{-1} = Y_k^{-1} - Y_k^{-1}\Delta_k Y_k^{-1}.$$

Using this identity in (3.1) we obtain, on subtracting (1.5),

$$\Delta_{k+1} = \tfrac{1}{2}\left(\Delta_k - Y_k^{-1}\Delta_k Y_k^{-1}A\right), \tag{3.15}$$

that is, (3.3) with the order term zero. Using (3.13) and (3.14) in (3.15), we have

$$\begin{aligned}
\Delta_{k+1} &= \tfrac{1}{2}\left(\Delta_k - \left(QD_k^{-1}Q^*\right)\left(\varepsilon q_i q_j^*\right)\left(QD_k^{-1}Q^*\right)\left(Q\Lambda Q^*\right)\right) \\
&= \tfrac{1}{2}\left(\Delta_k - \varepsilon QD_k^{-1}e_i e_j^* D_k^{-1}\Lambda Q^*\right) \\
&= \frac{1}{2}\left(\Delta_k - \varepsilon\left(\frac{1}{d_i^{(k)}}q_i\right)\left(\frac{\lambda_j}{d_j^{(k)}}q_j^*\right)\right) \\
&= \frac{1}{2}\left(1 - \frac{\lambda_j}{d_i^{(k)}d_j^{(k)}}\right)\Delta_k.
\end{aligned}$$

Let $Y_k = A^{1/2}$ (the Hermitian positive definite square root of A), so that $D_k = \Lambda^{1/2}$, and choose i, j so that $\lambda_j/\lambda_i = \kappa_2(A)$. Then

$$\Delta_{k+1} = \tfrac{1}{2}\left(1 - \kappa_2(A)^{1/2}\right)\Delta_k.$$

Assuming that $\hat{Y}_{k+2}, \hat{Y}_{k+3}, \ldots$, like $\hat{Y}_{k+1}$, are computed exactly from the preceding iterates, it follows that

$$\hat{Y}_{k+r} = A^{1/2} + \left[\tfrac{1}{2}\left(1 - \kappa_2(A)^{1/2}\right)\right]^r\Delta_k, \qquad r \geq 0.$$

In this example, $\hat{Y}_k$ is an arbitrary distance $\varepsilon > 0$ away from $A^{1/2}$ in the 2-norm, yet if $\kappa_2(A) > 9$ the subsequent iterates diverge, growing unboundedly.

Consider now the Newton iteration (N) with $X_0 = mI$, $m > 0$, so that by Theorem 1, $X_k \equiv Y_k$, and make the same assumptions as in Theorem 2. Then

$$X_k - X \to 0 \quad \text{as } k \to \infty \tag{3.16}$$

(quadratically), where X is the square root of A defined in Theorem 2. Let X_k be perturbed to $\hat{X}_k = X_k + \Delta_k$ and denote the corresponding perturbed sequence of iterates (computed exactly from $\hat{X}_k$) by $\{\hat{X}_{k+r}\}_{r \geq 0}$. The standard local convergence theorem for Newton's method implies that, for $\|\hat{X}_k - X\|$ sufficiently small, that is, for k sufficiently large and $\|\Delta_k\|$ sufficiently small,

$$\hat{X}_{k+r} - X \to 0 \quad \text{as } r \to \infty \tag{3.17}$$

(quadratically). From (3.16) and (3.17) it follows that

$$\Delta_{k+r} = \hat{X}_{k+r} - X_{k+r} \to 0 \quad \text{as } r \to \infty.$$

Thus, unlike iteration (I), the Newton iteration (N) has the property that once convergence is approached, a suitable norm of the error matrix $\Delta_k = \hat{X}_k - X_k$ is not magnified, but rather decreased, in succeeding iterations.

To summarize, for iterations (N) and (I) with initial approximation mI ($m > 0$), our analysis shows how a small perturbation Δ_k in the kth iterate is propagated at the $(k+1)$st stage. For iteration (I), depending on the eigenvalues of A, a small perturbation Δ_k in Y_k may induce perturbations of increasing norm in succeeding iterates, and the sequence $\{\hat{Y}_k\}$ may "diverge" from the sequence of true iterates $\{Y_k\}$. The same conclusion applies to iteration (II) for which a similar analysis holds. In contrast, for large k, the Newton iteration (N) damps a small perturbation Δ_k in X_k.

Our conclusion, then, is that in simplifying Newton's method to produce the ostensibly attractive formulae (1.5) and (1.6), one sacrifices numerical stability of the method.

4. A Further Newton Variant. The following matrix square root iteration is derived in [4] using the matrix sign function:

$$P_0 = A, \quad Q_0 = I,$$

$$\text{(III):} \qquad \left.\begin{aligned} P_{k+1} &= \tfrac{1}{2}\left(P_k + Q_k^{-1}\right) \quad (4.1) \\ Q_{k+1} &= \tfrac{1}{2}\left(Q_k + P_k^{-1}\right) \quad (4.2) \end{aligned}\right\}, \qquad k = 0, 1, 2, \ldots.$$

It is easy to prove by induction (using Theorem 1) that if $\{Y_k\}$ is the sequence computed from (1.5) with $Y_0 = I$, then

$$\left.\begin{aligned} P_k &= Y_k \quad (4.3) \\ Q_k &= A^{-1}Y_k \quad (4.4) \end{aligned}\right\}, \qquad k = 1, 2, \ldots.$$

Thus if A satisfies the conditions of Theorem 2 and the sequence $\{P_k, Q_k\}$ is defined, then

$$\lim_{k \to \infty} P_k = X, \qquad \lim_{k \to \infty} Q_k = X^{-1},$$

where X is the square root of A defined in Theorem 2.

At first sight, iteration (III) appears to have no advantage over iteration (I). It is in general no less computationally expensive; it computes simultaneously approximations to X and X^{-1}, when probably only X is required; and intuitively the fact that

A is present only in the initial conditions, and not in the iteration formulae, is displeasing. However, as we will now show, this "coupled" iteration does not suffer from the numerical instability which vitiates iteration (I).

To parallel the analysis in Section 3 suppose the assumptions of Theorem 2 hold, let $\hat{P}_k$ and $\hat{Q}_k$ denote the computed iterates from iteration (III), define

$$E_k = \hat{P}_k - P_k, \qquad F_k = \hat{Q}_k - Q_k,$$

and assume that at the $(k+1)$st stage $\hat{P}_{k+1}$ and $\hat{Q}_{k+1}$ are computed exactly from $\hat{P}_k$ and $\hat{Q}_k$. Then from (4.1) and (4.2), using (3.2), we have

$$\hat{P}_{k+1} = \tfrac{1}{2}\left(P_k + E_k + Q_k^{-1} - Q_k^{-1}F_kQ_k^{-1}\right) + O\left(\|F_k\|^2\right),$$

$$\hat{Q}_{k+1} = \tfrac{1}{2}\left(Q_k + F_k + P_k^{-1} - P_k^{-1}E_kP_k^{-1}\right) + O\left(\|E_k\|^2\right).$$

Subtracting (4.1) and (4.2), respectively, gives

$$E_{k+1} = \tfrac{1}{2}\left(E_k - Q_k^{-1}F_kQ_k^{-1}\right) + O\left(g_k^2\right), \tag{4.5}$$

$$F_{k+1} = \tfrac{1}{2}\left(F_k - P_k^{-1}E_kP_k^{-1}\right) + O\left(g_k^2\right), \tag{4.6}$$

where $g_k = \max\{\|E_k\|, \|F_k\|\}$.

From (2.2), (2.3), (4.3), (4.4) and (3.6),

$$Z^{-1}P_kZ = D_k, \quad Z^{-1}Q_kZ = \Lambda^{-1}D_k, \quad D_k = \operatorname{diag}\left(d_i^{(k)}\right);$$

thus, defining

$$\tilde{E}_k = Z^{-1}E_kZ, \qquad \tilde{F}_k = Z^{-1}F_kZ,$$

we can transform (4.5) and (4.6) into

$$\tilde{E}_{k+1} = \tfrac{1}{2}\left(\tilde{E}_k - D_k^{-1}\Lambda\tilde{F}_kD_k^{-1}\Lambda\right) + O\left(g_k^2\right),$$

$$\tilde{F}_{k+1} = \tfrac{1}{2}\left(\tilde{F}_k - D_k^{-1}\tilde{E}_kD_k^{-1}\right) + O\left(g_k^2\right).$$

Written elementwise, using the notation

$$\tilde{E}_k = \left(\tilde{e}_{ij}^{(k)}\right), \qquad \tilde{F}_k = \left(\tilde{f}_{ij}^{(k)}\right),$$

these equations become

$$\tilde{e}_{ij}^{(k+1)} = \tfrac{1}{2}\left(\tilde{e}_{ij}^{(k)} - \alpha_{ij}^{(k)}\tilde{f}_{ij}^{(k)}\right) + O\left(g_k^2\right), \tag{4.7}$$

$$\tilde{f}_{ij}^{(k+1)} = \tfrac{1}{2}\left(\tilde{f}_{ij}^{(k)} - \beta_{ij}^{(k)}\tilde{e}_{ij}^{(k)}\right) + O\left(g_k^2\right), \tag{4.8}$$

where

$$\alpha_{ij}^{(k)} = \frac{\lambda_i\lambda_j}{d_i^{(k)}d_j^{(k)}} = \left(\lambda_i\lambda_j\right)^{1/2} + O\left(\varepsilon^{(k)}\right)$$

and

$$\beta_{ij}^{(k)} = \frac{1}{d_i^{(k)}d_j^{(k)}} = \frac{1}{\left(\lambda_i\lambda_j\right)^{1/2}} + O\left(\varepsilon^{(k)}\right),$$

using (3.8) and (3.10). It is convenient to write Eqs. (4.7) and (4.8) in vector form:

$$h_{ij}^{(k+1)} = M_{ij}^{(k)}h_{ij}^{(k)} + O\left(g_k^2\right), \tag{4.9}$$

where

$$h_{ij}^{(k)} = \begin{bmatrix} \tilde{e}_{ij}^{(k)} \\ \tilde{f}_{ij}^{(k)} \end{bmatrix}$$

and

$$M_{ij}^{(k)} = \frac{1}{2}\begin{bmatrix} 1 & -\alpha_{ij}^{(k)} \\ -\beta_{ij}^{(k)} & 1 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 1 & -(\lambda_i\lambda_j)^{1/2} \\ -\dfrac{1}{(\lambda_i\lambda_j)^{1/2}} & 1 \end{bmatrix} + O(\varepsilon^{(k)})$$

$$= M_{ij} + O(\varepsilon^{(k)}).$$

It is easy to verify that the eigenvalues of M_{ij} are zero and one; denote a corresponding pair of eigenvectors by x_0 and x_1 and let

$$h_{ij}^{(k)} = a_0^{(k)}x_0 + a_1^{(k)}x_1.$$

If we make a further assumption that no new errors are introduced at the $(k+2)$nd stage of the iteration onwards (so that the analysis is tracing how an isolated pair of perturbations at the kth stage is propagated), then for k large enough and g_k small, we have, by induction,

$$h_{ij}^{(k+r)} \approx M_{ij}^r h_{ij}^{(k)} = M_{ij}^r\left(a_0^{(k)}x_0 + a_1^{(k)}x_1\right) = a_1^{(k)}x_1, \qquad r > 0. \tag{4.10}$$

While $\|h_{ij}^{(k+1)}\|_1$ may exceed $\|h_{ij}^{(k)}\|_1$ by the factor $\|M_{ij}^{(k)}\|_1 \approx \|M_{ij}\|_1 \geqslant 1$ (taking norms in (4.9)), from (4.10) it is clear that the vectors $h_{ij}^{(k+1)}$, $h_{ij}^{(k+2)}, \ldots$ remain approximately constant, that is, the perturbations introduced at the kth stage have only a bounded effect on succeeding iterates.

Our analysis shows that iteration (III) does not suffer from the unstable error propagation which affects iteration (I) and suggests that iteration (III) is, for practical purposes, numerically stable.

In the next section we supplement the theory which has been given so far with some numerical test results.

5. Numerical Examples. In this section we give some examples of the performance in finite-precision arithmetic of iteration (I) (with $Y_0 = I$) and iteration (III).

When implementing the iterations we distinguished the case where A is symmetric positive definite; since the iterates also possess this attractive property (see Corollary 1) it is possible to use the Choleski decomposition and to work only with the "lower triangles" of the iterates.

To define our implementations, it suffices to specify our algorithm for evaluating $W = B^{-1}C$, where $B = Y_k$, $C = A$ in iteration (I), and $B = P_k$ or Q_k, $C = I$ in iteration (III). For general A we used an LU factorization of B (computed by Gaussian elimination with partial pivoting) to solve by substitution the linear systems $BW = C$. For symmetric positive definite A we first formed B^{-1}, and then computed the (symmetric) product $B^{-1}C$; B^{-1} was computed from the Choleski decomposition $B = LL^T$, by inverting L and then forming the (symmetric) product $B^{-1} = L^{-T}L^{-1}$.

The operation counts for one stage of each iteration in our implementations, measured in flops [7, p. 32] are as follows.

TABLE 2

Flops per stage: $A \in \mathbf{R}^{n\times n}$	General A	Symmetric positive definite A
Iteration (I)	$4n^3/3$	n^3
Iteration (III)	$2n^3$	n^3

The computations were performed on a Commodore 64 microcomputer with unit roundoff [7, p. 33] $u = 2^{-32} \approx 2.33 \times 10^{-10}$. In the following $\lambda(A)$ denotes the spectrum of A.

Example 1. Consider the Wilson matrix example given in Section 1. W is symmetric positive definite and $(\kappa_2(W)^{1/2} - 1)/2 \approx 27$, so the theory of Section 3 predicts that for this matrix iteration (I) may exhibit numerical instability and that for large enough k

$$(5.1)\quad \|\hat{Y}_{k+1} - W^{1/2}\|_1 \approx \|\hat{Y}_{k+1} - Y_{k+1}\|_1 \lesssim 27\|\hat{Y}_k - Y_k\|_1 \approx 27\|\hat{Y}_k - W^{1/2}\|_1.$$

Note from Table 1 that for Implementation 1 there is approximate equality throughout in (5.1) for $k \geqslant 6$; this example supports the theory well. Strictly, the analysis of Section 3 does not apply to Implementation 2, but the overall conclusion is valid (essentially, the error matrices Δ_k are forced to be symmetric, but they can still grow as k increases).

Example 2 [8].

$$A = \begin{bmatrix} 5 & 4 & 1 & 1 \\ 4 & 5 & 1 & 1 \\ 1 & 1 & 4 & 2 \\ 1 & 1 & 2 & 4 \end{bmatrix}, \qquad \lambda(A) = \{1, 2, 5, 10\},\ \kappa_2(A) = 10.$$

Iterations (I) and (III) both converged in seven iterations.

Note that condition (3.12) is not satisfied by this matrix; thus the failure of this condition to hold does not necessarily imply divergence of the computed iterates from iteration (I).

Example 3.

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & .01 & 0 & 0 \\ -1 & -1 & 100 & 100 \\ -1 & -1 & -100 & 100 \end{bmatrix},$$

$$\lambda(A) = \{.01, 1, 100 \pm 100i\}.$$

Note that the lower quasi-triangular form of A is preserved by iterations (I) and (III). Iteration (I) diverged while iteration (III) converged within ten iterations. Briefly, iteration (I) behaved as follows.

TABLE 3

k	$\|\hat{Y}_k - \hat{Y}_{k-1}\|_1$
1	9.9×10^{1}
6	2.3×10^{-1}
7	2.1×10^{-3}
8	4.0×10^{-2}
9	2.1
12	4.8×10^{5}

Example 4 [3].

$$A = \begin{bmatrix} 0 & .07 & .27 & -.33 \\ 1.31 & -.36 & 1.21 & .41 \\ 1.06 & 2.86 & 1.49 & -1.34 \\ -2.64 & -1.84 & -.24 & -2.01 \end{bmatrix}, \qquad \lambda(A) = \{.03, 3.03, -1.97 \pm i\}.$$

Iteration (I) diverged, but iteration (III) converged in eight iterations to a real square root (cf. [3] where a nonreal square root was computed).

Example 5 [8].

$$A = \begin{bmatrix} 4 & 1 & 1 \\ 2 & 4 & 1 \\ 0 & 1 & 4 \end{bmatrix}, \qquad \lambda(A) = \{3,3,6\};\ A \text{ is defective.}$$

Both iterations converged in six steps.

We note that in Examples 3 and 4 condition (3.11) is not satisfied; the divergence of iteration (I) in these examples is "predicted" by the theory of Section 3.

6. Conclusions. When A is a full matrix, Newton's method for the matrix square root, defined in Eqs. (1.3) and (1.4), is unattractive compared to the Schur decomposition approach described in [2], [9]. Iterations (I) and (II), defined by (1.5) and (1.6), are closely related to the Newton iteration, since if the initial approximation $X_0 = Y_0 = Z_0$ commutes with A, then the sequences of iterates $\{X_k\}$, $\{Y_k\}$ and $\{Z_k\}$ are identical (see Theorem 1). In view of the relative ease with which Eqs. (1.5) and (1.6) can be evaluated, these two Newton variants appear to have superior computational merit. However, as our analysis predicts, and as the numerical examples in Section 5 illustrate, iterations (I) and (II) can suffer from numerical instability—sufficient to cause the sequence of computed iterates to diverge, even though the corresponding exact sequence of iterates is mathematically convergent. Since this happens even for well-conditioned matrices, iterations (I) and (II) must be classed as numerically unstable; they are of little practical use.

Iteration (III), defined by Eqs. (4.1) and (4.2), is also closely related to the Newton iteration and was shown in Section 4 to be numerically stable under suitable assumptions. In our practical experience (see Section 5) iteration (III) has always performed in a numerically stable manner.

As a means of computing a single square root, of the form described in Theorem 2, iteration (III) can be recommended: it is easy to code and it does not require the use of sophisticated library routines (important in a microcomputer environment, for example). In comparison, the Schur method [2], [9] is more powerful, since it yields more information about the problem and it can be used to determine a "well-conditioned" square root (see [9]); it has a similar computational cost to iteration (III) but it does require the computation of a Schur decomposition of A.

Since doing this work, we have developed a new method for computing the square root $A^{1/2}$ of a symmetric positive definite matrix A; see [10]. The method is related to iteration (I) and the techniques of this paper can be used to show that the method is numerically stable.

Acknowledgments. I am pleased to thank Dr. G. Hall and Dr. I. Gladwell for their interest in this work and for their comments on the manuscript. I also thank the referee for helpful suggestions.

Department of Mathematics
University of Manchester
Manchester M13 9PL, England

1. R. H. BARTELS & G. W. STEWART, "Solution of the matrix equation $AX + XB = C$," *Comm. ACM*, v. 15, 1972, pp. 820–826.

2. Å. BJÖRCK & S. HAMMARLING, "A Schur method for the square root of a matrix," *Linear Algebra Appl.*, v. 52/53, 1983, pp. 127–140.

3. E. D. DENMAN, "Roots of real matrices," *Linear Algebra Appl.*, v. 36, 1981, pp. 133–139.

4. E. D. DENMAN & A. N. BEAVERS, "The matrix sign function and computations in systems," *Appl. Math. Comput.*, v. 2, 1976, pp. 63–94.

5. F. R. GANTMACHER, *The Theory of Matrices*, Vol. I, Chelsea, New York, 1959.

6. G. H. GOLUB, S. NASH & C. F. VAN LOAN, "A Hessenberg-Schur method for the problem $AX + XB = C$," *IEEE Trans. Automat. Control*, v. AC-24, 1979, pp. 909–913.

7. G. H. GOLUB & C. F. VAN LOAN, *Matrix Computations*, Johns Hopkins Univ. Press, Baltimore, Maryland, 1983.

8. R. T. GREGORY & D. L. KARNEY, *A Collection of Matrices for Testing Computational Algorithms*, Wiley, New York, 1969.

9. N. J. HIGHAM, *Computing Real Square Roots of a Real Matrix*, Numerical Analysis Report No. 89, University of Manchester, 1984; *Linear Algebra Appl.* (To appear.)

10. N. J. HIGHAM, *Computing the Polar Decomposition—With Applications*, Numerical Analysis Report No. 94, University of Manchester, 1984; *SIAM J. Sci. Statist. Comput.* (To appear.)

11. W. D. HOSKINS & D. J. WALTON, "A faster method of computing the square root of a matrix," *IEEE Trans. Automat. Control*, v. AC-23, 1978, pp. 494–495.

12. W. D. HOSKINS & D. J. WALTON, "A faster, more stable method for computing the pth roots of positive definite matrices," *Linear Algebra Appl.*, v. 26, 1979, pp. 139–163.

13. P. LAASONEN, "On the iterative solution of the matrix equation $AX^2 - I = 0$," *M.T.A.C.*, v. 12, 1958, pp. 109–116.

14. J. M. ORTEGA, *Numerical Analysis: A Second Course*, Academic Press, New York, 1972.

15. G. W. STEWART, *Introduction to Matrix Computations*, Academic Press, New York, 1973.

16. C.-E. FRÖBERG, *Introduction to Numerical Analysis*, 2nd ed., Addison-Wesley, Reading, Mass., 1969.

17. P. HENRICI, *Elements of Numerical Analysis*, Wiley, New York, 1964.