A NEW SCALING FOR NEWTON'S ITERATION FOR THE POLAR DECOMPOSITION AND ITS BACKWARD STABILITY

RALPH BYERS*†‡ AND HONGGUO XU*§

Abstract. We propose a scaling scheme for Newton's iteration for calculating the polar decomposition. The scaling factors are generated by a simple scalar iteration in which the initial value depends only on estimates of the extreme singular values of the original matrix, which can for example be the Frobenius norms of the matrix and its inverse. In exact arithmetic, for matrices with condition number no greater than 10^{16}, with this scaling scheme, no more than 9 iterations are needed for convergence to the unitary polar factor with a convergence tolerance roughly equal to 10^{-16}. It is proved that if matrix inverses computed in finite precision arithmetic satisfy a backward-forward error model then the numerical method is backward stable. It is also proved that Newton's method with Higham's scaling or with Frobenius norm scaling is backward stable.

Key words. matrix sign function, polar decomposition, singular value decomposition (SVD), Newton's method, numerical stability, scaling

AMS subject classifications. 65F05, 65G05

1. Introduction. Every matrix $A \in \mathbb{C}^{n\times n}$ has a polar decomposition $A = QH$, where $H = H^* \in \mathbb{C}^{n\times n}$ is Hermitian positive semi-definite and $Q \in \mathbb{C}^{n\times n}$ is unitary, i.e., $Q^*Q = I$. The polar decomposition is unique with positive definite symmetric factor H if and only if A is nonsingular. Its applications include unitary approximation and distance calculations [8, 9, 12]. The polar decomposition generalizes to rectangular matrices; see, for example, [15]. We consider only the square matrix case here, because numerical methods for computing the polar decomposition typically begin by reducing to the problem down to the square matrix case using, for example, a QR factorization [5, 8]. (An algorithm that works directly with rectangular matrices appears in [6].)

The polar decomposition may be easily constructed from a singular value decomposition (SVD) of A. However, the SVD is a substantial calculation that displays much more of the structure of A than does the polar decomposition. Constructing the polar decomposition from the SVD destroys this extra information and wastes the arithmetic work used to compute it. It is intuitively more appealing to use the polar decomposition as a preliminary step in the computation of the SVD as in [12].

When A is nonsingular one way to compute the polar decomposition is through Newton's iteration

$$Q_{k+1} = \frac{1}{2}\left(\zeta_k Q_k + (\zeta_k Q_k)^{-*}\right), \qquad Q_0 = A, \tag{1.1}$$

where $\zeta_k = \zeta(Q_k) > 0$ is a positive scalar function of Q_k chosen to accelerate convergence [8]. Each iterate Q_k has polar decomposition $Q_k = QH_k$ where Q is the unitary polar factor of A, $H_0 = H$ is the Hermitian polar factor of A, and

*Department of Mathematics, University of Kansas, Lawrence, Kansas 66045, USA. xu@math.ku.edu.

†Deceased.

‡This material is based upon work partially supported by the University of Kansas General Research Fund allocations 2301062-003 and 2301054-003 and by the National Science Foundation awards 0098150, 0112375, and 9977352.

§This author is partially supported by NSF under Grant No.EPS-9874732 with matching support from the State of Kansas and the University of Kansas General Research Fund allocation 2301717-003.

$H_{k+1} = (\zeta_k H_k + (\zeta_k H_k)^{-1})/2$, $k \geq 0$. For appropriately chosen acceleration parameters ζ_k, $\lim_{k\to\infty} H_k = I$. Hence, the unitary polar factor is $Q = \lim_{k\to\infty} Q_k$ and the Hermitian polar factor is $H = \lim_{k\to\infty} Q_k^* A$.

Iteration (1.1) was first proposed in [8] and studied further in [5, 6, 14]. It is called "Newton's iteration" because it can be derived from Newton's method applied to the equation $X^*X = I$. It is closely related to Newton's iteration for the matrix sign function [17, 22].

Simplicity is an attractive feature of (1.1). Apart from the computation of ζ_k, each iteration needs only one matrix inversion and one matrix-matrix addition. The simplicity allows implementations of (1.1) to take advantage of the hierarchical memory and parallelism [1, 2, 13]. Many authors have studied choices of the acceleration parameters ζ_k [3, 4, 8, 16, 17, 22]. If $\zeta_k \equiv 1$, then the iterates Q_k converge quadratically to the unitary polar factor Q [8]. Convergence is also quadratic if $\zeta = \zeta(U)$ is a smooth function of $U \in \mathbb{C}^{n\times n}$ and $\zeta(U) = 1$ whenever U is unitary.

The choice

$$\zeta_k^{(2)} = \sqrt{\frac{\|Q_k^{-1}\|_2}{\|Q_k\|_2}} \tag{1.2}$$

where $\|\cdot\|_2$ is the spectral norm is proposed in [8]. This scale factor is optimal in the sense that, given Q_k, (1.2) minimizes the next error $\|Q_{k+1} - Q\|_2$. With this scale factor, for the matrices Q_k generated by (1.1), the error sequence $\|Q_k - Q\|_2$ converges monotonically to zero. Unfortunately, to determine the scale factor (1.2), one needs to compute two extreme singular values of Q_k at each iteration. In order to preserve rapid convergence of (1.1) with scaling (1.2), highly accurate values of these extreme singular values are required to guarantee $\zeta_k^{(2)} \to 1$. This is expensive enough to make scale factor (1.2) unattractive.

To save the cost of computing the extreme singular values, one might approximate (1.2). A commonly used scale factor is the $(1,\infty)$-scaling

$$\zeta_k^{(1,\infty)} = \left(\frac{\|Q_k^{-1}\|_1 \|Q_k^{-1}\|_\infty}{\|Q_k\|_1 \|Q_k\|_\infty}\right)^{\frac{1}{4}}, \tag{1.3}$$

(where $\|\cdot\|_1$ and $\|\cdot\|_\infty$ are the 1-norm and ∞-norm, respectively) which was proposed by Higham in [8]. The factor $\zeta_k^{(1,\infty)}$ is within a constant factor of $\zeta_k^{(2)}$. It adds a negligible amount of arithmetic work compared to the cost of Q_k^{-1} which is needed at each iteration anyway.

The scale factor

$$\zeta_k^{(F)} = \|Q_k^{-1}\|_F^{1/2} \|Q_k\|_F^{-1/2} \tag{1.4}$$

(where $\|\cdot\|_F$ is the Frobenius norm) is discussed in [5, 8, 16]. It can also be computed at a negligible cost. It is optimal in the sense that, given Q_k, it minimizes $\|Q_{k+1}\|_F$, and causes the sequence $\|Q_k\|_F$ to converge monotonically [5].

Another relatively inexpensive scale factor is [4]

$$\zeta_k^{(d)} = |\det(Q_k)|^{-1/n}. \tag{1.5}$$

The complex modulus of the determinant is very inexpensively obtained from the same matrix factorization used to calculate Q_k^{-1}. This scaling is optimal in the sense that

for a given iterate Q_k, it minimizes $D(Q_{k+1}) = \sum_{j=1}^{n}(\ln(\sigma_j^{(k+1)}))^2$ where $\sigma_j^{(k+1)}$ is the j-th singular value of Q_{k+1}. The function $D(Q_{k+1})$ is a measure of the departure of Q_{k+1} from the unitary matrices.

This paper considers the *sub-optimal* scaling strategy

$$\zeta_0 = 1/\sqrt{ab}, \quad \zeta_1 = \sqrt{\frac{2\sqrt{ab}}{a+b}}, \quad \zeta_k = 1/\sqrt{\rho(\zeta_{k-1})}, \quad k = 2, 3, \ldots, \tag{1.6}$$

where $\rho(x) = (x + x^{-1})/2$ and a and b are any numbers such that $0 < a \le \|A^{-1}\|_2^{-1} \le \|A\|_2 \le b$. Apart from estimating the extreme singular values of the initial matrix $Q_0 = A$, the scale factor costs only several floating point operations per iteration. Moreover, only rough estimates of $\|A\|_2$ and $\|A^{-1}\|_2^{-1}$ are needed. One may simply choose $a = \|A^{-1}\|_F^{-1}$ and $b = \|A\|_F$. From Table 2.1 with such choices for any matrices with condition number no greater than 10^{16} and size no greater than 10^{11}, at most nine iterations of (1.1) with scaling (1.6) are necessary to approximate the unitary polar factor Q to within 2-norm distance less than 10^{-16}.

We show below that in the presence of rounding error, (1.1) with (1.6) is numerically stable assuming that matrix inverses are calculated with small forward-backward error. This is the case, for example, when matrix inverses are computed using the bidiagonal reduction [7, Page 252]. We also prove the numerical stability of Newton's iteration with any of the scalings (1.2) - (1.4).

Commenting on an early draft of this paper, Krystyna Ziętak pointed out that the suboptimal (quasi-optimal) scaling parameters were discovered independently by Andrzej Kiełbasiński but not published in the open literature. They were presented by Krystyna Ziętak at the 1999 Householder meeting at Whistler and the 1999 ILAS conference at Barcelona. In Section 5 of their recent paper [19], Kiełbasiński *et al* mention the quasi-optimal scaling parameters. In [18], these authors gave an error analysis of Higham's method [8] with the same mixed backward-forward stability assumption for matrix inversion. They gave numerical experiments in [23].

In the following $A \in \mathbb{C}^{n\times n}$ is always nonsingular. $A = U\Sigma V^*$ is the SVD of A, where U and V are unitary, $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$ is diagonal, and $\sigma_1 \ge \ldots \ge \sigma_n \ge 0$ are the singular values of A. The set of the singular values is denoted by $\sigma(A)$. The condition number with respect to the spectral norm of A is denoted by $\kappa_2(A) = \sigma_1/\sigma_n$. Following [7, Page 18], a flop is the computational work of a floating point addition, subtraction, multiplication or division together with the associated subscripting and indexing overhead. It takes two flops to execute the Fortran statement `A(I,J) = A(I,J) + C*A(K,J)`.

2. Scaling and convergence. Let $A \in \mathbb{C}^{n\times n}$ be nonsingular with the SVD $A = U\Sigma V^*$ with $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$. Each Newton iterate Q_k in (1.1) has singular value decomposition $Q_k = U\Sigma_k V^*$ where

$$\Sigma_{k+1} = \left(\zeta_k\Sigma_k + (\zeta_k\Sigma_k)^{-1}\right)/2, \qquad \Sigma_0 = \Sigma. \tag{2.1}$$

In particular, Q_k has singular values $\sigma_1^{(k)}$, $\sigma_2^{(k)}$, $\ldots$ $\sigma_n^{(k)}$ (in no particular order when $k > 0$) that obey

$$\sigma_j^{(0)} = \sigma_j, \quad \sigma_j^{(k+1)} = \frac{1}{2}\left(\zeta_k\sigma_j^{(k)} + \frac{1}{\zeta_k\sigma_j^{(k)}}\right) = \rho(\zeta_k\sigma_j^{(k)}), \quad k = 0,\ 1,\ 2,\ \ldots, \tag{2.2}$$

where $\rho(x) = (x + x^{-1})/2$. For appropriately chosen ζ_k, $\lim_{k\to\infty} \sigma_j^{(k)} = 1$ and $\lim_{k\to\infty} Q_k = UV^* = Q$. Consequently, the convergence properties of (1.1) derive directly from the n scalar sequences $\sigma_j^{(k)}$ determined by (2.2). To attain good convergence behavior in (1.1), the acceleration parameters ζ_k must interact well with $\rho(x) = (x + x^{-1})/2$.

The following two lemmas list some easily verified elementary properties of $\rho(x) = (x + x^{-1})/2$.

LEMMA 2.1. *If $x > 0$, then*

1. $\rho(\frac{1}{x}) = \rho(x)$,
2. $1 \le \rho(x) \le \max(x, x^{-1})$ *with either equality iff* $x = 1$.
3. $\rho(x)$ *is decreasing on* $x \in (0, 1]$ *and increasing on* $x \in [1, \infty)$.

LEMMA 2.2. *Suppose that $0 < a \le b$. Define $\alpha_\zeta = \max\{(\zeta a)^{-1}, \zeta b\}$ and $\zeta_{opt} = (ab)^{-1/2}$. Then we have the following properties.*

1. *For any $\zeta > 0$, $1 \le \max_{a<x<b} \rho(\zeta x) = \rho(\alpha_\zeta)$, and $1 = \max_{a<x<b} \rho(\zeta x)$ iff $\alpha_\zeta = 1$.*
2. *For any $\zeta > 0$, $1 \le \min_{a<x<b} \rho(\zeta x)$, and $1 = \min_{a<x<b} \rho(\zeta x)$ iff $\zeta a \le 1 \le \zeta b$.*
3. $\min_{\zeta>0} \alpha_\zeta = \alpha_{\zeta_{opt}} = \sqrt{b/a}$, *and $\zeta = \zeta_{opt}$ is the only minimizer.*
4. $\min_{\zeta>0} \max_{a\le x\le b} \rho(\zeta x) = \min_{\zeta>0} \rho(\alpha_\zeta) = \rho\left(\alpha_{\zeta_{opt}}\right) = \rho\left(\sqrt{b/a}\right)$.

In the following, for ease of notation let $\tau(x)$ be the function

$$\tau(x) = \rho(\sqrt{x}) = \frac{1}{2}\left(\sqrt{x} + \frac{1}{\sqrt{x}}\right). \tag{2.3}$$

The k-fold composition of $\tau(x)$ with itself is written $\tau^k(x)$, i.e., $\tau^0(x) = x$, $\tau^1(x) = \tau(x)$, and for $k > 1$, $\tau^{k+1}(x) = \tau(\tau^k(x))$. Similarly $\rho^k(x)$ is the k-fold composition of $\rho(x) = (x + x^{-1})/2$ with itself.

Suppose that $0 < a \le \sigma_n \le \sigma_1 \le b$. Consider the sequence of intervals generated by Newton's iteration: $[a_0, b_0] = [a, b]$, $[a_1, b_1] = \rho(\zeta_0[a_0, b_0])$, $[a_2, b_2] = \rho(\zeta_1[a_1, b_1])$, ... It follows from (2.2) that $\sigma_j^{(k)} \in [a_k, b_k]$, $j = 1, 2, \ldots, n$, $k = 0, 1, \ldots$. Note that $\min_{x>0} \rho(x) = 1$, so, for $k \ge 1$, $[a_k, b_k] \subseteq [1, b_k]$ and

$$1 \le \sigma_j^{(k)} \le b_k. \tag{2.4}$$

It is intuitively satisfying to choose the sequence of acceleration parameters ζ_k in (1.1) to minimize the sequence b_k.

From Lemma 2.2, the initial optimal scaling factor is $\zeta_0 = (ab)^{-1/2}$. The initial interval is scaled to be $\zeta_0[a, b] = [\sqrt{a/b}, \sqrt{b/a}]$ which contains 1. The next interval is

$$[a_1, b_1] = \rho(\zeta_0[a, b]) = [1, \rho(\sqrt{b/a})] = [1, \tau(b/a)]$$

where $\tau(x)$ is given by (2.3). The left endpoint is $a_1 = 1$, so the optimal scaling factor for the next iteration is $\zeta_1 = b_1^{-1/2} = 1/\sqrt{\tau(b/a)}$. The next interval is

$$[a_2, b_2] = \rho(\zeta_1[a_1, b_1]) = [1, \rho(\sqrt{\tau(b/a)})] = [1, \tau^2(b/a)].$$

An easy induction shows that the sequence of intervals is $[a_0, b_0] = [a, b]$ and for $k \ge 1$, $[a_k, b_k] = [1, \tau^k(b/a)]$, and the sequence of optimal scaling factors is

$$\zeta_0 = 1/\sqrt{ab}, \quad \zeta_k = 1/\sqrt{\tau^k(b/a)}, \quad k = 1, 2, 3, \ldots \tag{2.5}$$

which is equivalent to (1.6).

Since $\tau(x) = \rho(\sqrt{x}) \le \rho(x)$ for $x \ge 1$,

$$\tau(b/a) \le \rho(b/a).$$

By induction we have

$$\tau^k(b/a) \le \rho^k(b/a),$$

for all $k \ge 1$. The sequence $b/a, \rho(b/a), \rho^2(b/a), \ldots$, is generated by Newton's iteration $x_{k+1} = \rho(x_k)$ with $x_0 = b/a$. It converges to 1 quadratically. Obviously $\tau^k(b/a) \ge 1$ for $k \ge 0$. So $b/a, \tau(b/a), \tau^2(b/a), \ldots$ also converges to 1 at least quadratically. It is not difficult to show that $1 \le \tau(x) \le x$ for any $x \ge 1$. We have

$$b_k = \tau^k(b/a) = \tau(\tau^{k-1}(b/a)) \le \tau^{k-1}(b/a) = b_{k-1}.$$

Hence, after the first step, the sequence of intervals satisfies

$$[a_1, b_1] \supseteq [a_2, b_2] \supseteq \ldots \supseteq [a_k, b_k] \supseteq \ldots,$$

and it converges to the single point 1 quadratically. (Note $a_k = 1$ for all $k \ge 1$.) The initial interval, $[a_0, b_0] = [a, b]$ is an exception, because, in general, $[a, b] \not\supseteq [1, \tau(b/a)]$.

Based on this fact and (2.4), the convergence properties of (1.1) with (1.6) are clear, and we summarize them in the following theorem.

THEOREM 2.3. *If*

$$0 < a \le \|A^{-1}\|_2^{-1} \le \|A\|_2 \le b, \tag{2.6}$$

and Q_k is obtained from the Newton iteration (1.1) with scaling (1.6) then

$$\|Q_k - Q\| \le \tau^k(b/a) - 1 \le \rho^k(b/a) - 1, \quad k = 1, 2, \ldots. \tag{2.7}$$

In fact the convergence properties are highly satisfactory even in case b/a is large. Table 2.1 uses Theorem 2.3 to list the number of Newton's iteration (1.1) with scaling (1.6) (and exact arithmetic) required to guarantee selected absolute errors $\delta > \|Q_k - Q\|_2$ and values of b/a. The table demonstrates that Newton's iteration (1.1) with scaling (1.6) typically needs no more than nine iterations to attain typical floating point precision accuracy. The table also demonstrates that convergence is insensitive to the choice of a and b—widely differing values of b/a need similar numbers of iterations to attain similar accuracy. In particular the easy-to-compute choices $a = \|A^{-1}\|_F^{-1}$ and $b = \|A\|_F$ satisfy (2.6) and are unlikely to lead to even one more iteration than the optimum choices of $a = \|A^{-1}\|_2^{-1}$ and $b = \|A\|_2$, particularly for ill-conditioned matrices. For instance, for any $A \in \mathbb{C}^{n\times n}$ with $\kappa_2(A) = 10^{16}$, for $a = \|A^{-1}\|_F^{-1}$ and $b = \|A\|_F$ we have $b/a \le n\kappa(A) = n10^{16}$. Then $b/a \le 10^{27}$ for any $n \le 10^{11}$, and the number of iterations is 9, same as with the optimum choices.

In Theorem 2.3, smaller values of b/a give smaller values of $\tau^k(b/a)$ and hence better error bounds. Inequality (2.6) implies that $b/a \ge \kappa_2(A) = \|A^{-1}\|_2\|A\|_2$ and equality can be achieved only with $a = \|A^{-1}\|_2^{-1} = \sigma_n$ and $b = \|A\|_2 = \sigma_1$. With $a = \sigma_n$ and $b = \sigma_1$ the scaling factors (1.6) are

$$\zeta_0 = 1/\sqrt{\sigma_1\sigma_n}, \qquad \zeta_k = 1/\sqrt{\tau^k(\kappa_2(A))}, \quad (k \ge 1), \tag{2.8}$$

TABLE 2.1
Number of Newton iterations (1.1) with scaling (1.6) (and exact arithmetic) required to guarantee absolute error $\|Q_k - Q\|_2 < \delta$ for selected values of δ and b/a such that $0 < a \le \|A^{-1}\|_2^{-1} \le \|A\|_2 \le b$. See Theorem 2.3.

$\delta \setminus b/a$	10	10^5	10^{10}	10^{15}	10^{20}	10^{25}	10^{27}
10^{-1}	2	4	5	6	6	6	6
10^{-4}	4	5	6	7	7	8	8
10^{-7}	4	6	7	8	8	8	8
10^{-10}	5	7	7	8	8	9	9
10^{-13}	5	7	8	8	9	9	9
10^{-16}	5	7	8	9	9	9	9
10^{-19}	6	7	8	9	9	10	10

and the corresponding intervals are $[\sigma_n, \sigma_1]$, and $[1, \tau^k(\kappa_2(A))]$ for $k \ge 1$. Let Σ_k be the matrices generated by (2.1) with $a = \sigma_n$ and $b = \sigma_1$. It is easy to verify that in this case, the right endpoint of the k-th interval is a singular value of Σ_k. This in turn implies that inequality (2.7) is an equality, i.e.,

$$\|Q_k - Q\|_2 = \|\Sigma_k - I\|_2 = \tau^k(\kappa_2(A)) - 1.$$

The number sequence b_k was also derived in [16] in order to show the convergence behavior of Newton's method with the optimal scale factors. It is shown that when $a = \sigma_n$ and $b = \sigma_1$, for Q_k generated with $\zeta_{k-1}^{(2)}$ defined in (1.2), one has $\|Q_k\|_2 \le b_k$ [16, 11]. Due to this fact we call ζ_k defined in (2.5) *sub-optimal* scale factors. Note that b_k is derived based on different interpretations here. It is the right end point of the kth interval generated by applying Newton's iteration to the initial interval $[a, b]$. For this interval iteration, ζ_k is the scale factor that minimizes $b_{k+1} - 1$ (i.e., it makes $[a_{k+1}, b_{k+1}]$ as close to 1 as possible).

3. The Algorithm. The Newton's method (1.1) with scaling scheme (1.6) is implemented by the following algorithm.

Algorithm 3.1 (Newton's method (1.1) with scaling (1.6))
Input: Nonsingular matrix $A \in \mathbb{C}^{n\times n}$ and a stopping criterion $\delta > 0$
Output: The polar decomposition $A = QH$.

Step 0: **a.** Set $Q_0 = A$; Compute Q_0^{-*}
b. Choose $a \le \|Q_0^{-1}\|_2^{-1}$ and $b \ge \|Q_0\|_2$; $\zeta_0 = 1/\sqrt{ab}$
c. Set $k = 0$
Step 1: While $\|Q_k - Q_k^{-*}\|_F \ge \delta$
a. $Q_{k+1} = (\zeta_k Q_k + \zeta_k^{-1} Q_k^{-*})/2$
b.
If $k = 0$, $\zeta_1 = \sqrt{\frac{2}{\sqrt{b/a}+\sqrt{a/b}}}$
Else $\zeta_{k+1} = \sqrt{\frac{2}{\zeta_k^{-1}+\zeta_k}}$
End if
c. Compute Q_{k+1}^{-*}
d. $k = k + 1$
End while
Step 2: $Q = (Q_k + Q_k^{-*})/2$; $H = \frac{1}{2}(Q^*A + (Q^*A)^*)$

Here are some remarks.

1. The matrix $A^{-1} = Q_0^{-1}$ needs to be computed in the first iteration anyway. Hence, power iterations on A and A^{-1} may be used evaluate the extreme singular values σ_1 and σ_n^{-1}, respectively using only $O(n^2)$ extra flops per iteration. These estimates may then serve as the sub-optimal scaling factors b and a^{-1}, respectively. Since highly accurate estimates of σ_1 and σ_n are unnecessary, a few power iterations should suffice. Alternatively, $\|A\|_F$ and $\|A^{-1}\|_F^{-1}$ may be used for b and a.

2. The stopping criterion $\|Q_k - Q_k^{-*}\|_F < \delta$ is essentially equivalent to $\|Q_{k+1} - Q_k\|_F < \delta$, which is used in [11, Section 8.9]. This follows from the fact that when $\zeta_k \approx 1$ (which is usually the case for a small δ),

$$Q_{k+1} - Q_k = \frac{1}{2}(\zeta_k^{-1} Q_k^{-*} - (2 - \zeta_k)Q_k) \approx \frac{1}{2}(Q_k^{-*} - Q_k).$$

In practice, in order for the computed Q_k to be within $O(\varepsilon)$ of a unitary matrix, where ε is the machine epsilon, it is sufficient to choose $\delta = O(\sqrt{\varepsilon})$. See (4.16).

3. Commonly the matrix inversion method used in the algorithm is an LU factorization-based method such as the Gaussian elimination with partial pivoting or complete pivoting [8]. Such an inversion method usually works well in practice [18]. In order to guarantee the algorithm to be numerically backward stable, one may use the more expensive bidiagonal reduction-based matrix inversion method provided in Appendix A.1. So the computed matrix inverses satisfy the backward-forward error model. See Assumption 4.1 in Section 4 below.

4. Estimating σ_1 and σ_n usually uses $O(n^2)$ flops. Each iteration uses $2n^3$ flops for the matrix inverse by an LU factorization-based method and $O(n^2)$ flops for matrix addition. Computing H uses $2n^3$ flops. If p is the number of iterations for convergence, then the algorithm uses a total of roughly $2(p+1)n^3$ flops [8]. When $p = 9$ it is about $20n^3$ flops, which is less than the QR-like SVD method (which takes $22n^3$ to $26n^3$ flops for the SVD and $4n^3$ for Q and H). If the bidiagonal reduction-based matrix inversion method is used, the total cost will be $2(3p + 1)n^3$ flops.

5. In order to reduce the cost while maintaining numerical stability, one may first use the bidiagonal reduction-based method for a few iterations. When $\kappa_2(Q_k)$ is not too large, say 100, one shifts to an LU factorization-based inversion method for the subsequent iterations. The matrix inverses essentially satisfy the backward-forward error model in the latter case [10, Section 14]. Also, it takes only a few iterations for the condition number to drop below 100. In the case when $\kappa_2(A) = 10^{-16}$, with $a = \sigma_n$ and $b = \sigma_1$, then $\|Q_3\|_2 = \tau^3(10^{-16}) \approx 42$. Since this is usually the worst case in practice, the bidiagonal reduction-based method is required in no more than 3 iterations. With this strategy, the maximum cost (with $p = 9$) is $3 \cdot 6n^3 + (9 - 3) \cdot 2n^3 + 2n^3 = 32n^3$ flops.

6. Although the cost for computing the scale factors (1.3) - (1.5) is negligible in Newton's iteration, computing the sub-optimal scale factors is essentially costless. Also, the use of sub-optimal scaling simplifies the algorithm, since the "shifting scale factor to 1" strategy, which is used for the $(1, \infty)$-scaling ([8]), is not needed. Finally, with sub-optimal scaling, in general, the number of iterations is no greater than 9, and it can be obtained by simply computing $\tau^k(b/a) - 1$. It is still not clear how to predict the number of iterations with other scalings, although in practice it is observed that the $(1, \infty)$-scaling and the sub-optimal scaling essentially have the same convergence rate.

4. Stability and Rounding Error Analysis. In this section, a first order error analysis establishes that Newton's method (1.1) with scaling (1.6) can be implemented

in a backward stable way. The same conclusion is drawn for the scalings (1.2), (1.3), and (1.4). In outline the approach is to estimate the residual $\|A - \widehat{Q}\widehat{H}\|_2$ for the rounding-error-perturbed unitary factor $\widehat{Q}$ and Hermitian factor $\widehat{H}$ produced by finite precision arithmetic in the algorithm in Section 3. The method here is to first estimate the forward errors $\widehat{Q} - Q$ and $\widehat{H} - H$ and then use them to estimate the residual.

For the error analysis, we employ the standard model of floating point arithmetic with machine epsilon ε [10, Section 2.2].

We also need the following assumptions.

ASSUMPTION 4.1. *If a nonsingular matrix $A \in \mathbb{C}^{n\times n}$ is inverted using finite precision arithmetic with machine epsilon ε to obtain a "computed inverse" X, then*

$$X = (A + E)^{-1} + F$$

where E, $F \in \mathbb{C}^{n\times n}$ are perturbation matrices satisfying

$$\|E\|_2 \le c_1(n)\varepsilon\|A\|_2, \qquad \|F\|_2 \le c_2(n)\varepsilon\|A^{-1}\|_2$$

and $c_i(n)$ $(i = 1, 2)$ are some low degree polynomials of n.

In Newton's method it is typical to use Gaussian elimination with partial or complete pivoting for computing matrix inverses. Although it works well in practice, the computed matrix inverses may not satisfy Assumption 4.1 [10, Section 14.1]. We show in Appendix A.1 that Assumption 4.1 is satisfied by a matrix inversion algorithm that uses the bidiagonal reduction method.

ASSUMPTION 4.2.

$$c_3(n)\kappa_2(A)\varepsilon < 1,$$

where $c_3(n)$ is a low degree polynomial of n.

Note that $1/\kappa_2(A)$ is the measure of the relative distance of a nonsingular A to the nearest singular matrices [7, Page 73], i.e.,

$$\frac{1}{\kappa_2(A)} = \min_{\det(A+E)=0} \frac{\|E\|_2}{\|A\|_2}.$$

If such a condition doesn't hold then matrices like $A + E$ in Assumption 4.1 can be singular. So this is a condition about numerical nonsingularity of A. It is essential in the subsequent first order error analysis, although it won't be explicitly stated.

We now begin the error analysis. In practice, rounding errors perturb Newton's method recurrence (1.1). Under Assumptions 4.1, if $\widehat{Q}_k$ is the computed version of Q_k, then

$$\begin{aligned}
\widehat{Q}_{k+1} &= \frac{\zeta_k}{2}\widehat{Q}_k + F_{k,1} + \frac{1}{2\zeta_k}\left((\widehat{Q}_k + F_{k,2})^{-*} + F_{k,3}\right) \\
&= \frac{\zeta_k}{2}(\widehat{Q}_k + F_{k,2}) + \frac{1}{2\zeta_k}(\widehat{Q}_k + F_{k,2})^{-*} + \left(F_{k,1} + \frac{1}{2\zeta_k}F_{k,3} - \frac{\zeta_k}{2}F_{k,2}\right) \\
&=: \frac{\zeta_k}{2}(\widehat{Q}_k + F_{kb}) + \frac{1}{2\zeta_k}(\widehat{Q}_k + F_{kb})^{-*} + F_{kf} \qquad (4.1)
\end{aligned}$$

where $F_{kb} = F_{k,2}$ and $F_{kf} = \left(F_{k,1} + \frac{1}{2\zeta_k}F_{k,3} - \frac{\zeta_k}{2}F_{k,2}\right)$. The perturbation matrix $F_{k,1}$ represents rounding errors introduced by floating point matrix addition and scalar

multiplication, and the perturbation matrices $F_{k,2}$ and $F_{k,3}$ represent rounding errors introduced by matrix inversion under Assumption 4.1. The F's obey the bounds

$$\begin{aligned}
\|F_{k,1}\|_2 &\le d_1\varepsilon \max\left(\|\zeta_k\widehat{Q}_k\|_2, \|\zeta_k^{-1}\widehat{Q}_k^{-*}\|_2\right)\\
\|F_{k,2}\|_2 &\le d_2\varepsilon\|\widehat{Q}_k\|_2\\
\|F_{k,3}\|_2 &\le d_3\varepsilon\|\widehat{Q}_k^{-*}\|_2\\
\|F_{kb}\|_2 &\le d_b\varepsilon\|\widehat{Q}_k\|_2\\
\|F_{kf}\|_2 &\le d_f\frac{\varepsilon}{2} \max\left(\|\zeta_k\widehat{Q}_k\|_2, \|\zeta_k^{-1}\widehat{Q}_k^{-*}\|_2\right),
\end{aligned}$$

where ε is the machine epsilon and d_1, $d_2 = d_b$, d_3 and d_f are some modest constants that may depend on n, the details of the arithmetic and the inversion algorithm but depend neither on Q_k nor $\widehat{Q}_k$. Each Q_k is a smooth function of Q_{k-1} and each $F_{k,j} = O(\varepsilon)$, so, by induction,

$$\widehat{Q}_k = Q_k + O(\varepsilon). \tag{4.2}$$

Hence, the bounds above may be loosely expressed in terms of Q_k as

$$\begin{aligned}
\|F_{kb}\|_2 &\le d_b\varepsilon\|Q_k\|_2 + O(\varepsilon^2) && (4.3)\\
\|F_{kf}\|_2 &\le d_f\frac{\varepsilon}{2} \max\left(\|\zeta_k Q_k\|_2, \|\zeta_k^{-1}Q_k^{-*}\|_2\right) + O(\varepsilon^2)\\
&\le d_f\varepsilon\|Q_{k+1}\|_2 + O(\varepsilon^2). && (4.4)
\end{aligned}$$

Inequality (4.4) is a consequence of (2.2).

We need the following lemma for continuing our analysis.

LEMMA 4.3. *Let $A \in \mathbb{C}^{n\times n}$ be a nonsingular matrix with polar decomposition $A = QH$ with $Q \in \mathbb{C}^{n\times n}$ unitary and $H \in \mathbb{C}^{n\times n}$ Hermitian positive definite. If $F \in \mathbb{C}^{n\times n}$ with $\|F\|_2 = 1$ and $t \ge 0$, then when t is sufficiently small $A + tF$ has the polar decomposition*

$$A + tF = Q\left((I + tE) + O(t^2)\right)\left(H + tG + O(t^2)\right)$$

where $G \in \mathbb{C}^{n\times n}$ is the unique Hermitian solution to

$$F^*A + A^*F = GH + HG, \tag{4.5}$$

and $E \in \mathbb{C}^{n\times n}$ is the unique skew-Hermitian solution to

$$Q^*F - F^*Q = EH + HE. \tag{4.6}$$

Also, E is given by

$$E = (Q^*F - G)H^{-1}. \tag{4.7}$$

Proof. See proof in Appendix A.2. □

At each of the perturbed Newton iteration (4.1) rounding errors are equivalent to perturbing $\widehat{Q}_k$ to $\widehat{Q}_k + F_{kb}$, taking one Newton step (1.1), then perturbing the result by adding F_{kf}. Let $\widehat{Q}_k = W_k\widehat{H}_k$ and $\widehat{Q}_k + F_{kb} = \widetilde{W}_k\widetilde{H}_k$ be the polar decompositions of $\widehat{Q}_k$ and $\widehat{Q}_k + F_{kb}$, respectively. By Lemma 4.3,

$$\widetilde{W}_k = W_k(I + E_{kb}) + O(\varepsilon^2), \tag{4.8}$$

where E_{kb} satisfies

$$E_{kb}\widehat{H}_k + \widehat{H}_k E_{kb} = W_k^* F_{kb} - F_{kb}^* W_k + O(\varepsilon^2). \tag{4.9}$$

From (4.1),

$$\frac{\zeta_k}{2}(\widehat{Q}_k + F_{kb}) + \frac{1}{2\zeta_k}(\widehat{Q}_k + F_{kb})^{-*} = \widehat{Q}_{k+1} - F_{kf}.$$

Since the unitary factor in the polar decomposition of the left-hand side matrix is $\widetilde{W}_k$, applying Lemma 4.3 to $\widehat{Q}_{k+1}$, we have

$$\widetilde{W}_k = W_{k+1}(I - E_{kf}) + O(\varepsilon^2),$$

or equivalently,

$$W_{k+1} = \widetilde{W}_k(I + E_{kf}) + O(\varepsilon^2), \tag{4.10}$$

where E_{kf} satisfies

$$E_{kf}\widehat{H}_{k+1} + \widehat{H}_{k+1}E_{kf} = W_{k+1}^* F_{kf} - F_{kf}^* W_{k+1} + O(\varepsilon^2). \tag{4.11}$$

Since Q_k has the polar decomposition $Q_k = QH_k$ and $\widehat{Q}_k$ satisfies (4.2), by Lemma 4.3 one also has $\widehat{H}_k = H_k + O(\varepsilon)$, $W_k = Q + O(\varepsilon)$. Based on these first-order error results, (4.9) and (4.11) can be expressed as

$$E_{kb}H_k + H_k E_{kb} = Q^* F_{kb} - F_{kb}^* Q + O(\varepsilon^2) \tag{4.12}$$

$$E_{kf}H_{k+1} + H_{k+1}E_{kf} = Q^* F_{kf} - F_{kf}^* Q + O(\varepsilon^2). \tag{4.13}$$

Combining (4.8) and (4.10), one has

$$W_{k+1} = W_k(I + E_{kb} + E_{kf}) + O(\varepsilon^2).$$

It follows by induction (with $W_0 = Q$),

$$W_j = Q(I + E_j) + O(\varepsilon^2), \quad j > 0,$$

where

$$E_j = \sum_{k=0}^{j-1}(E_{kb} + E_{kf}). \tag{4.14}$$

Suppose that Algorithm 3.1 applied to a nonsingular matrix $A \in \mathbb{C}^{n\times n}$ with polar decomposition $A = QH$ completes after p iterations in Step one. We obtain $\widehat{Q}_p$ that satisfies $\|\widehat{Q}_p - \widehat{Q}_p^{-*}\|_2 \le \|\widehat{Q}_p - \widehat{Q}_p^{-*}\|_F < \delta$. With the polar decomposition $\widehat{Q}_p = W_p\widehat{H}_p$ and (2.2), we have (see the proof in Appendix A.3)

$$\frac{1}{2}(\widehat{Q}_p + \widehat{Q}_p^{-*}) = W_p + \Delta\widehat{Q}_p, \tag{4.15}$$

where

$$\|\Delta\widehat{Q}_p\|_2 < \delta^2/8.$$

Suppose δ is small. Step two of the algorithm produces approximate polar factors

$$\widehat{Q} = \frac{1}{2}(\widehat{Q}_p + \widehat{Q}_p^{-*}) + F, \quad \widehat{H} = \frac{1}{2}\left(\widehat{Q}^* A + A^* \widehat{Q} + K\right),$$

where F accounts for rounding error forming $\widehat{Q}$ from $\widehat{Q}_p$ and K for rounding error forming $\widehat{H}$ from A and $\widehat{Q}$ obeying

$$\|F\| \le d_F \varepsilon, \qquad \|K\| \le d_K \varepsilon \|A\|_2,$$

with modest constants d_F and d_K. So we have

$$\|\widehat{Q} - W_p\|_2 \le d_F \varepsilon + \delta^2/8. \tag{4.16}$$

Since $W_p = Q(I+E)$ with $E := E_p$ defined in (4.14), by (4.15),

$$\widehat{Q} = Q(I+E) + \Delta\widehat{Q}_p + F = Q(I+E+L), \tag{4.17}$$

where $L = Q^*(\Delta\widehat{Q}_p + F)$ satisfies

$$\|L\|_2 \le d_L \max(\varepsilon, \delta^2)$$

for some modest constant d_L, which combines the rounding error F and the effect of stopping criterion. Both d_L and d_K may depend on n and the details of the finite precision arithmetic and computational algorithm but not on A, $\widehat{Q}$ or $\widehat{H}$. With (4.17) and the fact that E is skew-Hermitian,

$$\begin{aligned}
\widehat{H} &= \frac{1}{2}\left((I+E+L)^* Q^* A + A^* Q (I+E+L) + K\right) \\
&= \frac{1}{2}\left(1 - E + L^*\right) H + H(I+E+L) + K) \\
&= \frac{1}{2}(2H - EH + HE + L^* H + HL + K). \qquad (4.18)
\end{aligned}$$

So by (4.17) and (4.18), to the first order,

$$\begin{aligned}
\widehat{Q}\widehat{H} - A &= \frac{1}{2} Q(I+E+L)(2H - EH + HE + L^*H + HL + K) - A \\
&= \frac{1}{2} Q\left(2H - EH + HE + L^*H + HL + 2EH + 2LH + K\right) - A \\
&\quad + O(\|L\|_2^2) + O(\varepsilon\|L\|_2) + O(\varepsilon^2) \\
&= \frac{1}{2} Q\left(EH + HE + L^*H + HL + 2LH + K\right) + O(\max(\varepsilon^2, \varepsilon\delta^2, \delta^4)).
\end{aligned}$$

From (4.14) this expression can be written

$$\begin{aligned}
\widehat{Q}\widehat{H} - A &= \frac{1}{2} Q \sum_{k=0}^{p-1} (E_{kb} H + H E_{kb} + E_{kf} H + H E_{kf}) \\
&\quad + \frac{1}{2} Q(L^* H + HL + 2LH + K) + O(\max(\varepsilon^2, \varepsilon\delta^2, \delta^4)). \qquad (4.19)
\end{aligned}$$

Note that so far the sub-optimal scale factors have not play a role. In order to continue the analysis we need the following lemma which involves the sub-optimal scaling.

LEMMA 4.4. *Let $A \in \mathbb{C}^{n\times n}$ be a nonsingular matrix with singular value decomposition $U\Sigma V^*$, $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$ and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$. Consider Newton's iteration (2.1) with scaling (1.6) and initial iterate $\Sigma_0 = \Sigma$. Let $\sigma_j^{(k)}$ be the j-th diagonal entry of Σ_k, and let $\sigma_{\max}^{(k)} = \max_{1\leq j\leq n} \sigma_j^{(k)}$. If a, b satisfy $0 < a \leq \sigma_n$ and $b \geq \sigma_1$, then, for all $k \geq 0$ and $1 \leq i, j \leq n$,*

$$\frac{\sigma_{\max}^{(k)}}{\sigma_i^{(k)} + \sigma_j^{(k)}} \leq \frac{b}{\sigma_i + \sigma_j}.$$

Proof. See Appendix A.4. □

Recall that E_{kb}, E_{kf} satisfy (4.12) and (4.13), respectively. Let

$$H_k = V\Sigma_k V^*$$

be an eigen-decomposition where V is unitary and Σ_k is diagonal obeying (2.1). (Recall that throughout the algorithm the singular vectors of the Q_k's are the singular vectors of A. In particular, for all k, the unitary matrix of right singular vectors of A is also a unitary matrix of eigenvectors of H_k.) In this notation, (4.12) and (4.13) can be written

$$\begin{aligned}\widetilde{E}_{kb}\Sigma_k + \Sigma_k\widetilde{E}_{kb} &= \widetilde{F}_{kb} - \widetilde{F}_{kb}^* + O(\varepsilon^2)\\ \widetilde{E}_{kf}\Sigma_{k+1} + \Sigma_{k+1}\widetilde{E}_{kf} &= \widetilde{F}_{kf} - \widetilde{F}_{kf}^* + O(\varepsilon^2)\end{aligned}$$

where

$$\widetilde{E}_{kb} = V^*E_{kb}V,\ \widetilde{E}_{kf} = V^*E_{kf}V,\ \widetilde{F}_{kb} = V^*Q^*F_{kb}V,\ \widetilde{F}_{kf} = V^*Q^*F_{kf}V. \tag{4.20}$$

So, the (i,j)-th entries of $\widetilde{E}_{kb}$ and $\widetilde{E}_{kf}$ are

$$\widetilde{e}_{ij,kb} = \frac{\widetilde{f}_{ij,kb} - \overline{\widetilde{f}_{ji,kb}}}{\sigma_i^{(k)} + \sigma_j^{(k)}} + O(\varepsilon^2), \qquad \widetilde{e}_{ij,kf} = \frac{\widetilde{f}_{ij,kf} - \overline{\widetilde{f}_{ji,kf}}}{\sigma_i^{(k+1)} + \sigma_j^{(k+1)}} + O(\varepsilon^2). \tag{4.21}$$

Note that $\|\widetilde{E}_{kj}\|_2 = \|E_{kj}\|_2$ and $\|\widetilde{F}_{kj}\|_2 = \|F_{kj}\|_2$ for $j = b, f$, because V and Q are unitary.

Multiplying (4.19) on the left by V^* and on the right by V gives

$$\begin{aligned}V^*(\widehat{Q}\widehat{H} - A)V = \frac{1}{2}V^*QV\sum_{k=0}^{p-1}\left(\widetilde{E}_{kb}\Sigma + \Sigma\widetilde{E}_{kb} + \widetilde{E}_{kf}\Sigma + \Sigma\widetilde{E}_{kf}\right)\\ + \frac{1}{2}V^*\left(L^*H + HL + 2LH + K\right)V + O(\max(\varepsilon^2, \varepsilon\delta^2, \delta^4))\end{aligned}$$

where $\widetilde{E}_{kb}$ and $\widetilde{E}_{kf}$ are given by (4.20). From (4.21), the (i,j)-th entry of the sum

$\sum_{k=0}^{p-1}\left(\widetilde{E}_{kb}\Sigma+\Sigma\widetilde{E}_{kb}+\widetilde{E}_{kf}\Sigma+\Sigma\widetilde{E}_{kf}\right)$ is

$$\begin{aligned}
&\sum_{k=0}^{p-1}(\widetilde{e}_{ij,kb}+\widetilde{e}_{ij,kf})(\sigma_i+\sigma_j)\\
&=\sum_{k=0}^{p-1}\left(\frac{\widetilde{f}_{ij,kb}-\overline{\widetilde{f}_{ji,kb}}}{\sigma_i^{(k)}+\sigma_j^{(k)}}+\frac{\widetilde{f}_{ij,kf}-\overline{\widetilde{f}_{ji,kf}}}{\sigma_i^{(k+1)}+\sigma_j^{(k+1)}}\right)(\sigma_i+\sigma_j)+O(\varepsilon^2)\\
&=\sum_{k=0}^{p-1}\left(\frac{\widetilde{f}_{ij,kb}-\overline{\widetilde{f}_{ji,kb}}}{\sigma_{\max}^{(k)}}\,\frac{(\sigma_i+\sigma_j)\sigma_{\max}^{(k)}}{\sigma_i^{(k)}+\sigma_j^{(k)}}+\frac{\widetilde{f}_{ij,kf}-\overline{\widetilde{f}_{ji,kf}}}{\sigma_{\max}^{(k+1)}}\,\frac{(\sigma_i+\sigma_j)\sigma_{\max}^{(k+1)}}{\sigma_i^{(k+1)}+\sigma_j^{(k+1)}}\right)\\
&\quad+O(\varepsilon^2).
\end{aligned}$$

Inequalities (4.3) and (4.4) and Lemma 4.4 imply

$$\begin{aligned}
&\left|\sum_{k=0}^{p-1}(\widetilde{e}_{ij,kb}+\widetilde{e}_{ij,kf})(\sigma_i+\sigma_j)\right|\\
&\le 2\varepsilon\sum_{k=0}^{p-1}\left(d_b\frac{(\sigma_i+\sigma_j)\sigma_{\max}^{(k)}}{\sigma_i^{(k)}+\sigma_j^{(k)}}+d_f\frac{(\sigma_i+\sigma_j)\sigma_{\max}^{(k+1)}}{\sigma_i^{(k+1)}+\sigma_j^{(k+1)}}\right)+O(\varepsilon^2)\\
&\le 2p\varepsilon(d_b+d_f)b+O(\varepsilon^2).
\end{aligned}$$

Hence, the residual is bounded as

$$\begin{aligned}
\|\widehat{Q}\widehat{H}-A\|_2&\le\|Q\sum_{k=0}^{p-1}\left(E_{kb}H+HE_{kb}+E_{kf}H+HE_{kf}\right)/2\|_2\\
&\quad+\|Q(L^*H+HL+2LH+K)/2\|_2+O(\max(\varepsilon^2,\varepsilon\delta^2,\delta^4))\\
&\le np\varepsilon(d_b+d_f)b+(2d_L+d_K/2)\max(\varepsilon,\delta^2)\|A\|_2+O(\max(\varepsilon^2,\varepsilon\delta^2,\delta^4))
\end{aligned}$$

In the same way, from (4.18) we can obtain

$$\begin{aligned}
\|\widehat{H}-H\|_2&\le\|(-EH+HE)/2\|_2+\|(L^*H+HL+K)/2\|_2+O(\delta^4)\\
&\le np\varepsilon(d_b+d_f)b+(d_L+d_K/2)\max(\varepsilon,\delta^2)\|A\|_2+O(\max(\varepsilon^2,\varepsilon\delta^2,\delta^4)).
\end{aligned}$$

By applying Lemma 4.4 to (4.21) to estimate $\|E\|_2$, then from (4.17) we can derive

$$\|\widehat{Q}-Q\|_2\le\|QE\|_2+\|QL\|_2\le\theta np\varepsilon(d_b+d_f)b+d_L\max(\varepsilon,\delta^2),$$

where

$$\theta=\begin{cases}\frac{1}{\sigma_n} & A\text{ is complex}\\ \frac{2}{\sigma_{n-1}+\sigma_n} & A\text{ is real.}\end{cases}\tag{4.22}$$

(The formula in real case is based on the fact $\widetilde{e}_{ii,kb}=\widetilde{e}_{ii,kf}=0$ from (4.21).)

We present the above error analysis results as well as (4.16) in the following theorem.

THEOREM 4.5. *Suppose that $A\in\mathbb{C}^{n\times n}$ satisfies Assumption 4.2 and has the polar decomposition $A=QH$. Let $\widehat{Q}$ and $\widehat{H}$ be the matrices computed by Algorithm 3.1*

after p iterations with a matrix inversion method that satisfies the error model in Assumption 4.1. Then

$$\begin{aligned}
\|\widehat{Q}\widehat{H} - A\|_2 &\le np\varepsilon(d_b + d_f)b + (2d_L + d_K/2)\max(\varepsilon, \delta^2)\|A\|_2 + O(\max(\varepsilon^2, \varepsilon\delta^2, \delta^4)) \\
\|\widehat{H} - H\|_2 &\le np\varepsilon(d_b + d_f)b + (d_L + d_K/2)\max(\varepsilon, \delta^2)\|A\|_2 + O(\max(\varepsilon^2, \varepsilon\delta^2, \delta^4)) \\
\|\widehat{Q} - Q\|_2 &\le \theta np\varepsilon(d_b + d_f)b + d_L\max(\varepsilon, \delta^2) \\
\|\widehat{Q} - W_p\|_2 &\le d_F\varepsilon + \delta^2/8,
\end{aligned}$$

where d_b, d_f, d_L, d_K, d_F are some modest constants, W_p is a unitary matrix, and θ is defined in (4.22).

As noted in Table 2.1, in most practical situations $p \le 9$. Therefore, if b is not too much greater than $\|A\|_2$ and the algorithm uses a stopping criterion δ not too much greater than $\sqrt{\varepsilon}$, then the algorithm is backward stable.

COROLLARY 4.6. *If $b = \|A\|_2$ and $\delta = n^{\frac{1}{4}}\sqrt{\varepsilon}$ then*

$$\begin{aligned}
\|\widehat{Q}\widehat{H} - A\|_2 &\le (np(d_b + d_f) + \sqrt{n}(2d_L + d_K/2))\varepsilon\|A\|_2 + O(\varepsilon^2) \\
\|\widehat{H} - H\|_2 &\le (np(d_b + d_f) + \sqrt{n}(d_L + d_K/2))\varepsilon\|A\|_2 + O(\varepsilon^2) \\
\|\widehat{Q} - Q\|_2 &\le np(d_b + d_f)\varepsilon(\theta\|A\|_2) + \sqrt{n}d_L\varepsilon \\
\|\widehat{Q} - W_p\|_2 &\le (d_F + \sqrt{n}/8)\varepsilon.
\end{aligned}$$

Note that the error bounds for $\|\widehat{H} - H\|_2$ and $\|\widehat{Q} - Q\|_2$ coincide with the perturbation results; see, for example, [8, 21, 20] and [11, Section 8.2]. The quantity $\theta\|A\|_2$ serves as the condition number for the perturbation of Q.

REMARK 1. The same procedure can be used to give an error analysis for Newton's method with other scalings. Note that the backward stability depends on whether

$$\frac{(\sigma_i + \sigma_j)\sigma_{\max}^{(k)}}{\sigma_i^{(k)} + \sigma_j^{(k)}} = O(\|A\|_2), \tag{4.23}$$

which depends on scaling factors. From Remark 2 in Appendix A.4, (4.23) holds for the optimal scaling (1.2). Lemma A.3 in Appendix A.5 shows that (4.23) also holds for the $(1, \infty)$-scaling (1.3) and the scaling (1.4). Therefore, Newton's method with these three scalings is also backward stable under the same conditions of Theorem 4.5 and an appropriate stopping criterion, when the number of iterations is not too large. (See Remark 3 in Appendix A.5.)

We also observed that the numerical stability doesn't necessarily depend on how fast the method converges. In fact, one can show that when $\|A\|_2 \ge \|A^{-1}\|_2$, Newton's method without scaling ($a = b = 1$) computes a polar decomposition satisfying the same error bounds given in Theorem 4.5.

5. Numerical Examples. We did some numerical experiments with Newton's method (1.1) with scaling (1.6) using $a = \|A^{-1}\|_F^{-1}$, $b = \|A\|_F$ and also with the $(1, \infty)$-scaling (1.3). The main purpose is to test the numerical stability results and convergence rate, and to compare the sub-optimal scaling and $(1, \infty)$-scaling. For this reason we used the bidiagonal reduction matrix inversion method Algorithm BR for computing matrix inverses.

All numerical experiments were done on a Dell PC with a Pentium-IV processor, in MATLAB version 7.2 with machine epsilon $\varepsilon \approx 2.22 \times 10^{-16}$.

In the numerical experiments we use stopping criterion $\delta = n^{\frac{1}{4}}\sqrt{\varepsilon}$, where n is the size of matrices. For Newton's method with the $(1,\infty)$-scaling the scale factor is shifted to 1 when $\|X_{k+1} - X_k\|_F/\|X_{k+1}\|_F < 10^{-2}$. Based on the results in Corollary 4.6, if $\widehat{Q}$ and $\widehat{H}$ are the computed unitary and Hermitian polar factors produced by Algorithm 3.1, then we expect to observe $\|A - \widehat{Q}\widehat{H}\|_2/\|A\|_2$, $\|H - \widehat{H}\|_2/\|H\|_2$, and $\|Q - \widehat{Q}\|_2/(\theta\|A\|_2)$ not much larger than ε.

In the tables we will use the following notations

$$e_Q = \frac{\|Q - \widehat{Q}\|_2}{\theta\|A\|_2}, \quad e_H = \frac{\|H - \widehat{H}\|_2}{\|H\|_2}, \quad res = \frac{\|A - \widehat{Q}\widehat{H}\|_2}{\|A\|_2}, \quad ror = \|\widehat{Q}^*\widehat{Q} - I\|_2,$$

where the "exact" factors Q and H for the matrices in the first example are obtained from the SVD of A using MATLAB's variable precision arithmetic `vpa` with 24 significant decimal digits. p is the number of iterations, and n is the dimension of matrices. The symbol "HSF" refers to Newton's method (1.1) with Higham's $(1,\infty)$-scaling. The symbol "SUB" refers to (1.1) with scaling (1.6) using the Frobenius norms for the initial interval, i.e., $a = \|A^{-1}\|_F^{-1}$ and $b = \|A\|_F$.

EXAMPLE 1. Three groups of twenty real matrices were constructed with dimension 20 by using MATLAB' `gallery('randsvd',20,kappa,5)` with `kappa` equal to $10^2, 10^8, 10^{15}$, respectively. The singular values of the generated matrices are random values with uniformly distributed logarithm. For each group the ranges of the condition numbers $\kappa_2(A)$ and $\theta\|A\|_2$ are listed below.

`kappa` $= 10^2 : \kappa_2(A) \in [37.7,\ 90.6]$, $\theta\|A\|_2 \in [33.5,\ 83.7]$

`kappa` $= 10^8 : \kappa_2(A) \in [1.13,\ 6.75] \times 10^7$, $\theta\|A\|_2 \in [0.2,\ 5.44] \times 10^7$

`kappa` $= 10^{15} : \kappa_2(A) \in [6.35 \times 10^{10},\ 7.53 \times 10^{14}]$, $\theta\|A\|_2 \in [1.78 \times 10^{10},\ 5.22 \times 10^{14}]$

The test results are summarized in Table 5.1, where for each group the minimum and maximum values of the errors, residuals, and numbers of iterations are listed.

TABLE 5.1
Extreme values of errors, residuals, and iteration counts from Example 1.

`kappa`		10^2		10^8		10^{15}	
		Min	Max	Min	Max	Min	Max
p	SUB	6	6	8	8	8	9
	HSF	6	7	8	8	8	9
e_Q	SUB	2.6e−17	7.2e−17	7.4e−19	1.7e−17	6.9e−19	2.3e−17
	HSF	2.2e−17	7.7e−17	7.4e−19	1.7e−17	6.9e−19	2.3e−17
e_H	SUB	2.2e−16	4.1e−16	2.5e−16	4.1e−16	1.9e−16	4.4e−16
	HSF	1.7e−16	3.9e−16	1.9e−16	3.9e−16	2.2e−16	4.0e−16
res	SUB	4.1e−16	8.9e−16	4.0e−16	7.5e−16	3.5e−16	6.3e−16
	HSF	3.5e−16	8.2e−16	2.7e−16	5.9e−16	3.9e−16	7.2e−16
ror	SUB	7.9e−16	1.1e−15	8.0e−16	1.1e−15	7.8e−16	1.3e−15
	HSF	7.0e−16	1.2e−15	7.9e−16	1.2e−15	8.2e−16	1.1e−15

EXAMPLE 2. In this example the test matrices are the Hilbert matrices which are $n \times n$ matrices with entries $a_{ij} = 1/(i + j - 1)$. The example uses dimensions $n = 6$, $n = 8$, $n = 10$, $n = 12$ and $n = 14$. For every Hilbert matrix, the polar decomposition is $A_n = I_n A_n$.

The condition number $\kappa_2(A_n)$ ranges from 1.5×10^7 to 5.1×10^{17}, and $\theta\|A_n\|_2$ ranges from 2.6×10^5 to 1.0×10^{17}. The test results are reported in Table 5.2.

TABLE 5.2
Errors, residuals and iteration counts for Hilbert matrices from Example 2.

n		6	8	10	12	14
p	SUB	8	8	9	9	9
	HSF	7	8	8	9	9
e_Q	SUB	$1.2e{-}18$	$1.6e{-}19$	$2.7e{-}19$	$4.1e{-}19$	$2.0e{-}17$
	HSF	$1.2e{-}18$	$1.6e{-}19$	$2.7e{-}19$	$4.1e{-}19$	$2.0e{-}17$
e_H	SUB	$2.3e{-}16$	$1.9e{-}16$	$8.7e{-}17$	$1.4e{-}16$	$2.3e{-}16$
	HSF	$1.9e{-}16$	$1.4e{-}16$	$8.7e{-}17$	$1.6e{-}16$	$2.7e{-}16$
res	SUB	$2.6e{-}16$	$2.4e{-}16$	$1.8e{-}16$	$3.0e{-}16$	$3.8e{-}16$
	HSF	$2.5e{-}16$	$2.5e{-}16$	$1.8e{-}16$	$3.3e{-}16$	$6.3e{-}16$
ror	SUB	$2.6e{-}16$	$3.9e{-}16$	$6.2e{-}16$	$6.3e{-}16$	$6.5e{-}16$
	HSF	$2.8e{-}16$	$5.3e{-}16$	$6.8e{-}16$	$8.5e{-}16$	$1.0e{-}15$

Newton's method with the sub-optimal scaling performed well in both examples calculating the polar factors to nearly full precision. As predicted it takes at most 9 iterations. Newton's method with the $(1,\infty)$-scaling performs equally well.

6. Conclusion. The sub-optimal scaling scheme (1.6) is essentially costless and simplifies the algorithm of Newton's iteration (1.1) for computing polar factors. In a typical floating point system, with this scaling scheme, for matrices with $\kappa_2(A) < 10^{16}$ no more than 9 iterations are needed for convergence to the unitary polar factor with a convergence tolerance roughly equal to the machine epsilon. By employing the bidiagonal factorization for matrix inversion, (1.1) with (1.6) forms a provably backward stable algorithm. Newton's method with $(1,\infty)$-scaling and scaling (1.4) is also proved to be backward stable, provided the number of iterations is not too large.

Acknowledgment. We thank Nick Higham for detailed suggestions and sending part of his unpublished book [11]. We thank Andrzej Kiełbasiński for pointing out a mistake in an earlier version and the referees for their comments and suggestions.

REFERENCES

[1] Z. Bai and J. Demmel. Design of a parallel nonsymmetric eigenroutine toolbox. Part I. In R. F. Sincovec et. al., editor, *Proceedings of the Sixth SIAM Conference on Parallel Processing for Scientific Computing*. SIAM, Philadelphia, PA, 1993. Also available as Computer Science Report CSD-92-718, University of California, Berkeley, CA 1992.

[2] Z. Bai and J. Demmel. Design of a parallel nonsymmetric eigenroutine toolbox. Part II. Technical Report Department of Mathematics Research Report 95-11, University of Kentucky, Lexington, KY, 1995.

[3] L. A. Balzer. Accelerated convergence of the matrix sign function method of solving Lyapunov, Riccati and other matrix equations. *Internat. J. Control*, 32(6):1057–1078, 1980.

[4] R. Byers. Solving the algebraic Riccati equation with the matrix sign function. *Linear Algebra Appl.*, 85:267–279, 1987.

[5] A. A. Dubrulle. An optimum iteration for the matrix polar decomposition. *Electron. Trans. Numer. Anal.*, 8:21–25 (electronic), 1999.

[6] W. Gander. Algorithms for the polar decomposition. *SIAM J. Sci. Statist. Comput.*, 11(6):1102–1115, 1990.

[7] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, third edition, 1996.

[8] N. J. Higham. Computing the polar decomposition—with applications. *SIAM J. Sci. Statist. Comput.*, 7(4):1160–1174, 1986.

[9] N. J. Higham. Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra Appl.*, 103:103–118, 1988.
[10] N. J. Higham. *Accuracy and Stability of Numerical Algorithms*. SIAM Publications, Philadelphia, PA,USA, second edition, 2002.
[11] N. J. Higham. *Functions of Matrices: Theory and Computation*. SIAM Publications, Philadelphia, PA,USA, 2008.
[12] N. J. Higham and P. Papadimitriou. Parallel singular value decomposition via the polar decomposition. Technical Report Numerical Analysis Report No. 239, Department of Mathematics, University of Manchester, Manchester M13 9PL, England, 1993.
[13] N. J. Higham and P. Papadimitriou. A parallel algorithm for computing the polar decomposition. *Parallel Computing*, 20(8):1161–1173, 1994.
[14] N. J. Higham and R. S. Schreiber. Fast polar decomposition of an arbitrary matrix. *SIAM J. Sci. Statist. Comput.*, 11:648–655, 1990.
[15] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 1990. Corrected reprint of the 1985 original.
[16] C. S. Kenney and A. J. Laub. On scaling Newton's method for polar decomposition and the matrix sign function. *SIAM J. Matrix Anal. Appl.*, 13:688–706, 1992.
[17] C. S. Kenney and A. J. Laub. The matrix sign function. *IEEE Trans. Automat. Control*, 40(8):1330–1348, 1995.
[18] A. Kiełbasiński and K. Ziętak. Numerical behavior of Higham's scaled method for polar decomposition. *Numerical Algorithms*, 32:105–140, 2003.
[19] A. Kiełbasiński, P. Zieliński, and K. Ziętak. Higham's scaled method for polar decomposition and numerical matrix-inversion. Technical Report Institute of Mathematics and Computer Science Report I18/2007/P-045, Wroclaw University of Technology, Wroclaw, Poland, 2007.
[20] R.-C. Li. New perturbation bounds for the unitary polar factor. *SIAM J. Matrix Anal. Appl.*, 16:327–332, 1995.
[21] R. Mathias. Perturbation bounds for the polar decomposition. *SIAM J. Matrix Anal. Appl.*, 14:588–597, 1993.
[22] J. D. Roberts. Linear model reduction and solution of the algebraic Riccati equation by use of the sign function. *Internat. J. Control*, 32:677–687, 1980. (Reprint of Technical Report No. TR-13, CUED/B-Control, Cambridge University, Engineering Department, 1971).
[23] P. Zieliński and K. Ziętak. Polar decomposition - properties, applications and algorithms. *Applied Mathematics, Annals of the Polish Mathematical Society*, 38:23–49, 1995.

Appendix A.

A.1. Bidiagonal reduction-based matrix inversion algorithm.

Algorithm BR
Input: Nonsingular matrix $A \in \mathbb{C}^{n\times n}$
Output: $G = A^{-1}$

Step 1: Compute $A = UBV^*$ with U, V unitary and B upper bidiagonal
Step 2: Solve $BY = U^*$ for Y by back substitution
Step 3: Compute $G = VY$

In Step 1 one may use the Householder reflectors to perform the reduction. The reduction needs $\frac{8}{3}n^3$ flops and computing U needs $\frac{4}{3}n^3$ flops. The matrix V is stored in factorized form. The cost for solving the matrix equation is $O(n^2)$ flops. With the factorized form of V it needs $2n^3$ flops to compute G. So the total cost is $6n^3$ flops.

In order to show that a matrix inverse computed by Algorithm BR follows Assumption 4.1, we need the following lemma.

LEMMA A.1. *Consider the system $Bx = z$, where $z \in \mathbb{C}^n$ and B is nonsingular and upper bidiagonal denoted by*

$$B = \begin{bmatrix} \alpha_1 & \beta_1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & \beta_{n-1} \\ 0 & & & \alpha_n \end{bmatrix}.$$

Let $\hat{x}$ be the numerical solution with back substitution. Then $\hat{x}$ satisfies

$$\hat{x} = B^{-1}(z + \delta z) + \delta x,$$

where

$$|\delta z| \le 3n\varepsilon|z| + O(\varepsilon^2), \qquad |\delta x| \le 3n\varepsilon|\hat{x}| + O(\varepsilon^2).$$

Proof. The components of the computed vector $\hat{x}$ can be formulated as

$$\hat{x}_n = \frac{z_n}{\alpha_n(1+\epsilon_n)}, \qquad \hat{x}_k = \frac{z_k(1+\delta_k) - \beta_k\hat{x}_{k+1}}{\alpha_k(1+\epsilon_k)}, \quad 1 \le k \le n-1,$$

where $|\epsilon_n|, |\delta_k| < \varepsilon$, $|\epsilon_k| < 3\varepsilon$, $k \le n-1$. So we have

$$\begin{aligned} &\alpha_n\hat{x}_n(1+\epsilon_n) = z_n \\ &\alpha_{n-1}\hat{x}_{n-1}(1+\epsilon_{n-1}) + \beta_{n-1}\hat{x}_n = z_{n-1}(1+\delta_{n-1}) \\ &\qquad\vdots \\ &\alpha_1\hat{x}_1(1+\epsilon_1) + \beta_1\hat{x}_2 = z_1(1+\delta_1). \end{aligned} \tag{A.1}$$

Define $\tilde{x} = [\tilde{x}_1, \ldots, \tilde{x}_n]^T$, $\tilde{z} = [\tilde{z}_1, \ldots, \tilde{z}_n]^T$ with

$$\tilde{x}_n = \hat{x}_n(1+\epsilon_n), \quad \tilde{x}_{n-1} = \hat{x}_{n-1}(1+\epsilon_{n-1})(1+\epsilon_n), \quad \ldots, \quad \tilde{x}_1 = \hat{x}_1\prod_{k=1}^{n}(1+\epsilon_k),$$

$$\tilde{z}_n = z_n, \quad \tilde{z}_{n-1} = z_{n-1}(1+\delta_{n-1})(1+\epsilon_n), \quad \ldots, \quad \tilde{z}_1 = z_1(1+\delta_1)\prod_{k=2}^{n}(1+\epsilon_k);$$

By multiplying $(1+\epsilon_n)$ to the 2nd equation, $(1+\epsilon_n)(1+\epsilon_{n-1})$ to the 3rd equation, and so on, in (A.1), we obtain that $\tilde{x}$ and $\tilde{z}$ satisfy

$$B\tilde{x} = \tilde{z}.$$

Let $\delta z = \tilde{z} - z$ and $\delta x = \hat{x} - \tilde{x}$. Then from $\tilde{x} = B^{-1}\tilde{z}$ we have

$$\hat{x} = B^{-1}(z + \delta z) + \delta x.$$

The error bounds for $|\delta x|$ and $|\delta z|$ follows simply from the definitions. ∎

THEOREM A.2. *Let X be the inverse of A computed by Algorithm BR then X satisfies Assumption 4.1.*

Proof. We only consider the first order errors.

Let $\hat{U}\hat{B}\hat{V}^*$ be the computed bidiagonal factorization of A. Then $\hat{U} = U + \Delta U_1$, $\hat{V} = V + \Delta V_1$, where U, V are unitary and $\|\Delta U_1\|_2 \le d_1\varepsilon$, $\|\Delta V_1\|_2 \le d_2\varepsilon$, and

$$U\hat{B}V^* = A + E,$$

where $\|E\|_2 \le d_3\varepsilon\|A\|_2$ for some modest constants d_1, d_2, d_3. Let $\hat{Y}$ be the numerical solution of $\hat{B}Y = \hat{U}^*$ computed by back substitution. By Lemma A.1,

$$\hat{Y} = \hat{B}^{-1}(\hat{U}^* + \Delta U_2) + \Delta Y,$$

where

$$\|\Delta U_2\|_2 \le 3n^{\frac{3}{2}}\varepsilon, \qquad \|\Delta Y\|_2 \le 3n^{\frac{3}{2}}\varepsilon\|\hat{Y}\|_2.$$

Let X be the computed matrix product $\hat{V}\hat{Y}$. We have

$$X = \hat{V}\hat{Y} + \Delta X,$$

where $\|\Delta X\|_2 \le d_4\varepsilon\|\hat{Y}\|_2$ for some modest constant d_4. Now

$$\begin{aligned} X &= \hat{V}\hat{B}^{-1}(\hat{U}^* + \Delta U_2) + \hat{V}\Delta Y + \Delta X = \hat{V}\hat{B}^{-1}\hat{U}^* + \hat{V}\hat{B}^{-1}\Delta U_2 + \hat{V}\Delta Y + \Delta X \\ &=: V\hat{B}^{-1}U^* + F = (A+E)^{-1} + F, \end{aligned}$$

where

$$F = \Delta V_1\hat{B}^{-1}U^* + V\hat{B}^{-1}(\Delta U_1)^* + \hat{V}\hat{B}^{-1}\Delta U_2 + \hat{V}\Delta Y + \Delta X.$$

It is easily verified $\|F\|_F \le (6n^{\frac{3}{2}} + d_1 + d_2 + d_4)\varepsilon\|A^{-1}\|_2$. □

A.2. Proof of Lemma 4.3. Equations (4.5) and (4.7) are established in the proof of Theorem 2.5 in [8]. Here we slightly modify that proof to establish (4.6).

Let $A(t) = A + tF$ have polar decompositions $Q(t)H(t)$. Note that $H(t) = (A(t)^*A(t))^{1/2}$ (positive definite square root) and $Q(t) = A(t)H(t)^{-1}$ are sums, differences, products, quotients and compositions with C^∞ functions of the entries of $A(t)$ which is trivially a C^∞ function. Hence, $Q(t)$ and $H(t)$ are also C^∞. (Here we use the fact that A is nonsingular to observe that $H(t) = (A(t)^*A(t))^{1/2}$ avoids the singularity of the square root at zero.) Using $\dot{Q}$ and $\dot{H}$ to denote differentiation by t, Taylor's theorem implies that

$$\begin{aligned} Q(t) &= Q(0) + t\dot{Q}(0) + O(t^2) = Q(I + tQ^*\dot{Q}(0)) + O(t^2) \\ H(t) &= H(0) + t\dot{H}(0) + O(t^2). \end{aligned}$$

The proof of Theorem 2.5 in [8] shows $\dot{H}(0) = G = G^*$ with G given by (4.5) and $Q^*(0)\dot{Q}(0) = E = -E^*$ with E given by (4.7).

Differentiate $A + tF = Q(t)H(t)$ to get $F = \dot{Q}H + Q\dot{H}$. Evaluating at $t = 0$, letting $E = Q^*(0)\dot{Q}(0)$, $G = \dot{H}(0)$ gives $Q^*F = EH + G$. Using the facts that E is skew-Hermitian and G is Hermitian while subtracting this equation to its Hermitian transpose gives (4.6). The Lyapunov operator on the right is nonsingular, because A nonsingular implies that the eigenvalues of the Hermitian polar factor H are real and positive. Hence, the solution E is unique.

A.3. Proof for (4.15). Let $\widehat{Q}_p = U_p\Sigma_pV_p^*$ be the SVD. Recall that the singular values satisfy $\sigma_i^{(p)} \ge 1$. Then $\|\widehat{Q}_p - \widehat{Q}_p^{-*}\|_2 < \delta$ implies

$$\sigma_i^{(p)} - \frac{1}{\sigma_i^{(p)}} < \delta, \quad i = 1, 2, \ldots, n.$$

Because

$$\sigma_i^{(p)} - \frac{1}{\sigma_i^{(p)}} = \frac{(\sigma_i^{(p)} + 1)(\sigma_i^{(p)} - 1)}{\sigma_i^{(p)}},$$

we have

$$\sigma_i^{(p)} - 1 < \frac{\delta\sigma_i^{(p)}}{\sigma_i^{(p)}+1}.$$

Then

$$\frac{1}{2}\left(\sigma_i^{(p)} + \frac{1}{\sigma_i^{(p)}}\right) - 1 = \frac{(\sigma_i^{(p)}-1)^2}{2\sigma_i^{(p)}} < \frac{\sigma_i^{(p)}}{2(\sigma_i^{(p)}+1)^2}\delta^2.$$

Since the function $x/(x+1)^2$ is decreasing when $x \geq 1$, we have $\sigma_i^{(p)}/(\sigma_i^{(p)}+1)^2 \leq 1/4$. Hence

$$\frac{1}{2}\left(\sigma_i^{(p)} + \frac{1}{\sigma_i^{(p)}}\right) - 1 < \delta^2/8.$$

and using $W_p = U_p V_p^*$,

$$\begin{aligned}\|\Delta\widehat{Q}_p\|_2 &= \|(\widehat{Q}_p + \widehat{Q}_p^{-*})/2 - W_p\|_2 = \|U_p((\Sigma_p + \Sigma_p^{-1})/2 - I)V_p^*\|_2 \\ &= \max_i \left(\frac{1}{2}\left(\sigma_i^{(p)} + \frac{1}{\sigma_i^{(p)}}\right) - 1\right) < \delta^2/8.\end{aligned}$$

A.4. Proof of Lemma 4.4. By (2.4), $\sigma_{\max}^{(k)} \leq b_k$ for $k \geq 0$. So we only need to show

$$\frac{b_k}{\sigma_i^{(k)} + \sigma_j^{(k)}} \leq \frac{b}{\sigma_i + \sigma_j}, \quad k \geq 0.$$

An easy calculation shows for all i and j,

$$\sigma_i^{(k)} + \sigma_j^{(k)} = \frac{\zeta_{k-1}}{2}\left(\sigma_i^{(k-1)} + \sigma_j^{(k-1)}\right)\left(1 + \frac{1}{\zeta_{k-1}^2 \sigma_i^{(k-1)}\sigma_j^{(k-1)}}\right).$$

Recall b_k satisfies

$$b_k = \frac{1}{2}\left(\zeta_{k-1}b_{k-1} + \frac{1}{\zeta_{k-1}b_{k-1}}\right) = \frac{\zeta_{k-1}}{2}\, b_{k-1}\left(1 + \frac{1}{\zeta_{k-1}^2 b_{k-1}^2}\right)$$

Because $\sigma_i^{(k-1)}\sigma_j^{(k-1)} \leq b_{k-1}^2$,

$$\begin{aligned}\frac{b_k}{\sigma_i^{(k)} + \sigma_j^{(k)}} &= \frac{\frac{\zeta_{k-1}}{2}\, b_{k-1}\left(1 + \frac{1}{\zeta_{k-1}^2 b_{k-1}^2}\right)}{\frac{\zeta_{k-1}}{2}\left(\sigma_i^{(k-1)} + \sigma_j^{(k-1)}\right)\left(1 + \frac{1}{\zeta_{k-1}^2 \sigma_i^{(k-1)}\sigma_j^{(k-1)}}\right)} \\ &\leq \frac{b_{k-1}}{\sigma_i^{(k-1)} + \sigma_j^{(k-1)}}.\end{aligned}$$

An easy induction on k now implies

$$\frac{b_k}{\sigma_i^{(k)} + \sigma_j^{(k)}} \leq \frac{b_0}{\sigma_i^{(0)} + \sigma_j^{(0)}} = \frac{b}{\sigma_i + \sigma_j}, \quad k \geq 0.$$

Remark 2.

1. The condition $a \le \|A^{-1}\|_2^{-1}$ is essential for proving Lemma 4.4. Without it one may not have the inequality $\sigma_j^{(k)} \le b_k$ and the result can't be proved.
2. In the case that the optimal scaling is employed, $b_k = \sigma_{\max}^{(k)}$ and one has the same result.

A.5. Relation (4.23) for the scalings (1.3) and (1.4).

LEMMA A.3. *Suppose that the matrix sequence Q_k is generated by Newton's iteration with scaling ζ_k that is either $\zeta_k^{(1,\infty)}$ in (1.3) or $\zeta_k^{(F)}$ in (1.4). The singular values of Q_k satisfy*

$$\frac{\sigma_{\max}^{(k)}}{\sigma_i^{(k)} + \sigma_j^{(k)}} \le \left(\prod_{\ell=0}^{k-1} \psi_\ell\right) \frac{\|A\|_2}{\sigma_i + \sigma_j} \le n^{\frac{k}{2}} \frac{\|A\|_2}{\sigma_i + \sigma_j}, \quad k \ge 0, \quad 1 \le i, j \le n, \tag{A.2}$$

where $\psi_\ell = \max\left\{1, \frac{1}{\zeta_\ell^2 \sigma_{\max}^{(\ell)} \sigma_{\min}^{(\ell)}}\right\} \le \sqrt{n}$.

Proof. Let

$$w_k = \frac{1}{\zeta_k^2 \sigma_{\min}^{(k)} \sigma_{\max}^{(k)}}, \quad k \ge 0.$$

If

$$\zeta_k = \zeta_k^{(1,\infty)} = \sqrt[4]{\frac{\|Q_k^{-1}\|_1 \|Q_k^{-1}\|_\infty}{\|Q_k\|_1 \|Q_k\|_\infty}},$$

using the inequalities $\|A\|_2 \le \sqrt{\|A\|_1 \|A\|_\infty} \le \sqrt{n}\|A\|_2$, we have ([11, Page 208])

$$\frac{1}{\sqrt[4]{n}}\, \zeta_k \le \frac{1}{\sqrt{\sigma_{\min}^{(k)} \sigma_{\max}^{(k)}}} \le \sqrt[4]{n}\, \zeta_k. \tag{A.3}$$

If

$$\zeta_k = \zeta_k^{(F)} = \sqrt{\frac{\|Q_k^{-1}\|_F}{\|Q_k\|_F}},$$

using the inequalities $\|A\|_2 \le \|A\|_F \le \sqrt{n}\, \|A\|_2$, we also have (A.3).

So in both cases we have

$$\frac{1}{\sqrt{n}} \le w_k \le \sqrt{n}.$$

Construct an interval $[a_k, b_k]$ according to the rule

$$\begin{array}{lll} a_k = \sigma_{\min}^{(k)} w_k, & b_k = \sigma_{\max}^{(k)} & w_k \le 1 \\ a_k = \sigma_{\min}^{(k)}, & b_k = \sigma_{\max}^{(k)} w_k, & w_k \ge 1 \end{array}.$$

Because $a_k \le \sigma_{\min}^{(k)}$ and $b_k \ge \sigma_{\max}^{(k)}$, we have $\sigma_1^{(k)}, \ldots, \sigma_n^{(k)} \in [a_k, b_k]$. Also, in both cases we have $(\zeta_k a_k)^{-1} = \zeta_k b_k$. So $\rho(\zeta_k a_k) = \rho(\zeta_k b_k)$ and

$$\sigma_j^{(k+1)} = \rho(\zeta_k \sigma_j^{(k)}) \in \rho(\zeta_k [a_k, b_k]) = [1, \rho(\zeta_k b_k)], \quad j = 1, 2, \ldots, n.$$

Since

$$\rho(\zeta_k b_k) = \frac{1}{2}\left(\zeta_k b_k + \frac{1}{\zeta_k b_k}\right) = \frac{\zeta_k b_k}{2}\left(1 + \frac{1}{(\zeta_k b_k)^2}\right),$$

we have

$$\begin{aligned}\frac{\sigma_{\max}^{(k+1)}}{\sigma_i^{(k+1)} + \sigma_j^{(k+1)}} \le \frac{\rho(\zeta_k b_k)}{\sigma_i^{(k+1)} + \sigma_j^{(k+1)}} &= \frac{\frac{\zeta_k b_k}{2}\left(1 + \frac{1}{(\zeta_k b_k)^2}\right)}{\frac{\zeta_k}{2}(\sigma_i^{(k)} + \sigma_j^{(k)})\left(1 + \frac{1}{\zeta_k^2 \sigma_i^{(k)} \sigma_j^{(k)}}\right)} \\ &= \frac{b_k\left(1 + \frac{1}{(\zeta_k b_k)^2}\right)}{(\sigma_i^{(k)} + \sigma_j^{(k)})\left(1 + \frac{1}{\zeta_k^2 \sigma_i^{(k)} \sigma_j^{(k)}}\right)}.\end{aligned}$$

If $w_k \le 1$, then $b_k = \sigma_{\max}^{(k)}$. Because $\sigma_i^{(k)}\sigma_j^{(k)} \le \left(\sigma_{\max}^{(k)}\right)^2 = b_k^2$, we have

$$\frac{\sigma_{\max}^{(k+1)}}{\sigma_i^{(k+1)} + \sigma_j^{(k+1)}} \le \frac{b_k}{\sigma_i^{(k)} + \sigma_j^{(k)}} = \frac{\sigma_{\max}^{(k)}}{\sigma_i^{(k)} + \sigma_j^{(k)}}.$$

If $w_k \ge 1$ then $b_k = \sigma_{\max}^{(k)} w_k$. Because $\sigma_i^{(k)}\sigma_j^{(k)} \le \left(\sigma_{\max}^{(k)}\right)^2 \le b_k^2$, we have

$$\frac{\sigma_{\max}^{(k+1)}}{\sigma_i^{(k+1)} + \sigma_j^{(k+1)}} \le \frac{b_k}{\sigma_i^{(k)} + \sigma_j^{(k)}} = w_k \frac{\sigma_{\max}^{(k)}}{\sigma_i^{(k)} + \sigma_j^{(k)}}.$$

Hence

$$\frac{\sigma_{\max}^{(k+1)}}{\sigma_i^{(k+1)} + \sigma_j^{(k+1)}} \le \psi_k \frac{\sigma_{\max}^{(k)}}{\sigma_i^{(k)} + \sigma_j^{(k)}}, \quad \psi_k = \max\{1, w_k\} \le \sqrt{n}.$$

Then the inequalities in (A.2) can be easily derived. ∎

REMARK 3. Suppose that Newton's method with scaling $\zeta_k^{(1,\infty)}$ or $\zeta_k^{(F)}$ terminates after p iterations. We have the same error bounds as in Theorem 4.5 but with a factor $n^{\frac{p}{2}}$ in the first term of the bounds for $\|\widehat{Q}\widehat{H} - A\|_2$, $\|\widehat{H} - H\|_2$, and $\|\widehat{Q} - Q\|_2$. When n and p are both large, this factor is notably large and one may not be able to use the bounds to claim backward stability. Unfortunately, we are unable to provide an upper bound for p, although it is observed that p is usually moderate in practice ($p \le 9$ for the $(1,\infty)$-scaling for all examples in Section 5).

However, we argue that the factor $n^{\frac{p}{2}}$ is an overestimate. The point is that when Q_k is getting close to a unitary matrix, $\zeta_k^{(1,\infty)}$ and $\zeta_k^{(F)}$ are getting close to 1. Then w_k as well as ψ_k will be close to 1. So in practice w_k will be around 1 after couple iterations.