New Perturbation Bounds for the Unitary Polar Factor *

Ren-Cang Li
Department of Mathematics
University of California at Berkeley
Berkeley, California 94720

June, 1993

Computer Science Division Technical Report UCB//CSD-94-852, University of California, Berkeley, CA 94720, December, 1994. To appear in *SIAM Journal on Matrix Analysis and Applications.*

Abstract

Let A be an $m \times n$ $(m \geq n)$ complex matrix. It is known that there is a unique *polar decomposition* $A = QH$, where $Q^*Q = I$, the $n \times n$ identity matrix, and H is positive definite, provided A has full column rank. This note addresses the following question: how much may Q change if A is perturbed? For the square case $m = n$ our bound, which is valid for any unitarily invariant norm, is sharper and simpler than Mathias's (*SIAM J. Matrix Anal. Appl.,* **14***(1993), 588–597.*). For the non-square case, we also establish a bound for unitarily invariant norm, which has not been done in literature.

Let A be an $m \times n$ $(m \geq n)$ complex matrix. It is known that there are Q with orthonormal column vectors, i.e., $Q^*Q = I$, and a unique positive

*This material is based in part upon work supported by Argonne National Laboratory under grant No. 20552402 and the University of Tennessee through the Advanced Research Projects Agency under contract No. DAAL03-91-C-0047, by the National Science Foundation under grant No. ASC-9005933, and by the National Science Infrastructure grants No. CDA-8722788 and CDA-9401156.

semidefinite H such that

$$A = QH. \tag{1}$$

Hereafter I denotes an identity matrix with appropriate dimensions which should be clear from the context or specified. The decomposition (1) is called the *polar decomposition* of A. If, in addition, A has full column rank then Q is uniquely determined also. In fact,

$$H = (A^*A)^{1/2}, \quad Q = A(A^*A)^{-1/2}, \tag{2}$$

where superscript "$*$" denotes conjugate transpose. The decomposition (1) can also be computed from the *singular value decomposition* (SVD) $A = U\Sigma V^*$ by

$$H = V\Sigma_1 V^*, \quad Q = U_1 V^*, \tag{3}$$

where $U = (U_1, U_2)$ and V are unitary, U_1 is $m \times n$, $\Sigma = \begin{pmatrix} \Sigma_1 \\ 0 \end{pmatrix}$ and $\Sigma_1 = \mathrm{diag}\,(\sigma_1, \ldots, \sigma_n)$ is nonnegative.

There are published bounds upon how much the two factor matrices Q and H may change if entries of A are perturbed. Among papers [1, 2, 4, 6, 7, 8, 9, 10] written on this subject, the perturbation bounds for Q when $m = n$ proved by Mathias [9] covers every unitarily invariant norm, while others are for the Frobenius norm only. Chen and Sun [2, 10] and Li [7] also deal with the case $m \geq n$ as we do here. A surprise is how heavily the sensitivity of the Q factor depends upon whether the working number field is real or complex [1, 6, 9].

In this paper, we obtain some bounds for the perturbations of Q, assuming A is complex. Our bound for the case $m = n$ is achievable and improves Mathias' slightly for small perturbations and significantly for big ones.

For the sake of convenience in our presentation, we use A and $\widetilde{A}$ for two matrices having full column rank, one of which is a perturbation of the other. Let

$$A = QH, \quad \widetilde{A} = \widetilde{Q}\widetilde{H} \tag{4}$$

be the polar decompositions of A and $\widetilde{A}$ respectively, and let

$$A = U\Sigma V^*, \quad \widetilde{A} = \widetilde{U}\widetilde{\Sigma}\widetilde{V}^* \tag{5}$$

be the SVDs of A and $\widetilde{A}$, respectively, where $\widetilde{U} = (\widetilde{U}_1, \widetilde{U}_2)$, $\widetilde{U}_1$ is $m \times n$, and $\widetilde{\Sigma} = \begin{pmatrix} \widetilde{\Sigma}_1 \\ 0 \end{pmatrix}$ and $\widetilde{\Sigma}_1 = \mathrm{diag}\,(\widetilde{\sigma}_1, \ldots, \widetilde{\sigma}_n)$. Assume as usual that

$$\sigma_1 \geq \cdots \geq \sigma_n > 0, \quad \text{and} \quad \widetilde{\sigma}_1 \geq \cdots \geq \widetilde{\sigma}_n > 0. \tag{6}$$

It follows from (2) and (5) that

$$Q = U_1 V^*, \quad \widetilde{Q} = \widetilde{U}_1 \widetilde{V}^*.$$

In what follows, $\|X\|_2$ denotes the spectral norm which is the biggest singular value of X, and $\|X\|_{\mathrm{F}}$ the Frobenius norm which is the square root of the trace of X^*X. We shall use $|||\cdot|||$ to denote a general unitarily invariant norm [5, 11]. Two particular ones are $\|\cdot\|_2$ and $\|\cdot\|_{\mathrm{F}}$. Consider

$$\begin{aligned} \left|\left|\left| A - \widetilde{A} \right|\right|\right| &= \left|\left|\left| U^*(A - \widetilde{A})\widetilde{V} \right|\right|\right| = \left|\left|\left| \Sigma V^* \widetilde{V} - U^* \widetilde{U} \widetilde{\Sigma} \right|\right|\right| & (7) \\ &= \left|\left|\left| \widetilde{U}^*(\widetilde{A} - A)V \right|\right|\right| = \left|\left|\left| \widetilde{\Sigma} \widetilde{V}^* V - \widetilde{U}^* U \Sigma \right|\right|\right|. & (8) \end{aligned}$$

Define

$$\begin{aligned} E &\stackrel{\mathrm{def}}{=} \Sigma V^* \widetilde{V} - U^* \widetilde{U} \widetilde{\Sigma}, \qquad \text{and} & (9) \\ \widetilde{E} &\stackrel{\mathrm{def}}{=} \widetilde{\Sigma} \widetilde{V}^* V - \widetilde{U}^* U \Sigma & (10) \end{aligned}$$

to infer from (7) and (8) that

$$|||E||| = \left|\left|\left| \widetilde{E} \right|\right|\right| = \left|\left|\left| A - \widetilde{A} \right|\right|\right|. \tag{11}$$

Notice that by (9) and (10)

$$\begin{aligned} (I, 0)E &= \Sigma_1 V^* \widetilde{V} - U_1^* \widetilde{U}_1 \widetilde{\Sigma}_1, \qquad \text{and} \\ (I, 0)\widetilde{E} &= \widetilde{\Sigma}_1 \widetilde{V}^* V - \widetilde{U}_1^* U_1 \Sigma_1, \end{aligned}$$

where I is $n \times n$. Adding the conjugate transpose of the second to the first yields

$$\Sigma_1 (V^* \widetilde{V} - U_1^* \widetilde{U}_1) + (V^* \widetilde{V} - U_1^* \widetilde{U}_1) \widetilde{\Sigma}_1 = (I, 0)E + \widetilde{E}^* \begin{pmatrix} I \\ 0 \end{pmatrix}. \tag{12}$$

This is our perturbation equation to derive our perturbation bounds for Q because for any unitarily invariant norm $|||\cdot|||$,

$$\left|\left|\left| V^* \widetilde{V} - U_1^* \widetilde{U}_1 \right|\right|\right| = \left|\left|\left| I - V U_1^* \widetilde{U}_1 \widetilde{V}^* \right|\right|\right| = \left|\left|\left| I - Q^* \widetilde{Q} \right|\right|\right|. \tag{13}$$

We shall use following lemma, a special case of Davis and Kahan [3, Theorem 5.2].

Lemma 1 *Let M and N be two Hermitian matrices, and let S be a complex matrix with suitable dimensions. Suppose there are two disjoint intervals separated by a gap of width at least η, one of which contains the spectrum of M and the other contains that of N. If $\eta > 0$, then there is a unique solution X to the matrix equation $MX - XN = S$, and moreover $|||X||| \leq \frac{1}{\eta} |||S|||$ for every unitarily invariant norm $|||\cdot|||$.*

Applying this lemma to (11), (12) and (13) with $M = \Sigma_1$, $N = -\widetilde{\Sigma}_1$ and $X = V^*\widetilde{V} - U_1^*\widetilde{U}_1$ yields

Lemma 2

$$\left|\left|\left| I - Q^*\widetilde{Q} \right|\right|\right| \leq \frac{2}{\sigma_n + \widetilde{\sigma}_n} \left|\left|\left| A - \widetilde{A} \right|\right|\right|. \tag{14}$$

When $m = n$, both Q and $\widetilde{Q}$ are unitary. Thus $\left|\left|\left| I - Q^*\widetilde{Q} \right|\right|\right| = \left|\left|\left| Q - \widetilde{Q} \right|\right|\right|$, and Lemma 2 yields

Theorem 1 *Let A and $\widetilde{A}$ be two $n \times n$ nonsingular complex matrices whose polar decompositions are given by (4), and let σ_n and $\widetilde{\sigma}_n$ be the smallest singular values of A and $\widetilde{A}$ respectively. Then*

$$\left|\left|\left| Q - \widetilde{Q} \right|\right|\right| \leq \frac{2}{\sigma_n + \widetilde{\sigma}_n} \left|\left|\left| A - \widetilde{A} \right|\right|\right|. \tag{15}$$

If, however, $m > n$, then it follows from (9) and (10) that

$$\begin{aligned} (0, I)E &= -U_2^*\widetilde{U}_1\widetilde{\Sigma}_1, \qquad \text{and} \\ (0, I)\widetilde{E} &= -\widetilde{U}_2^*U_1\Sigma_1, \end{aligned}$$

where I is $(m-n) \times (m-n)$. Therefore

$$\left|\left|\left| U_2^*\widetilde{U}_1 \right|\right|\right| \leq \left|\left|\left| -U_2^*\widetilde{U}_1\widetilde{\Sigma}_1 \right|\right|\right| \, \|\widetilde{\Sigma}_1^{-1}\|_2 \leq \frac{|||(0, I)E|||}{\widetilde{\sigma}_n} \leq \frac{\left|\left|\left| A - \widetilde{A} \right|\right|\right|}{\widetilde{\sigma}_n}.$$

Similarly,

$$\left|\left|\left| \widetilde{U}_2^*U_1 \right|\right|\right| \leq \frac{\left|\left|\left| (0, I)\widetilde{E} \right|\right|\right|}{\sigma_n} \leq \frac{\left|\left|\left| A - \widetilde{A} \right|\right|\right|}{\sigma_n}.$$

Notice that $(U_1V^*, U_2) = (Q, U_2)$ and $(\widetilde{U}_1\widetilde{V}^*, \widetilde{U}_2) = (\widetilde{Q}, \widetilde{U}_2)$ are unitary. Hence $U_2^*Q = 0$ and

$$\begin{aligned}
\left|\left|\left| Q - \widetilde{Q} \right|\right|\right| = \left|\left|\left| (Q, U_2)^*(Q - \widetilde{Q}) \right|\right|\right| &= \left|\left|\left| \begin{pmatrix} I - Q^*\widetilde{Q} \\ -U_2^*\widetilde{Q} \end{pmatrix} \right|\right|\right| \\
&\le \left|\left|\left| I - Q^*\widetilde{Q} \right|\right|\right| + \left|\left|\left| -U_2^*\widetilde{U}_1\widetilde{V}^* \right|\right|\right| \\
&= \left|\left|\left| I - Q^*\widetilde{Q} \right|\right|\right| + \left|\left|\left| U_2^*\widetilde{U}_1 \right|\right|\right| \\
&\le \left(\frac{2}{\sigma_n + \widetilde{\sigma}_n} + \frac{1}{\widetilde{\sigma}_n} \right) \left|\left|\left| A - \widetilde{A} \right|\right|\right|. \quad (16)
\end{aligned}$$

Similarly, we can prove

$$\left|\left|\left| Q - \widetilde{Q} \right|\right|\right| \le \left(\frac{2}{\sigma_n + \widetilde{\sigma}_n} + \frac{1}{\sigma_n} \right) \left|\left|\left| A - \widetilde{A} \right|\right|\right|. \tag{17}$$

Therefore, generally, we have

Theorem 2 *Let A and $\widetilde{A}$ be two $m \times n$ $(m > n)$ complex matrices having full column rank and with the polar decompositions (4), and let σ_n and $\widetilde{\sigma}_n$ be the smallest singular values of A and $\widetilde{A}$ respectively. Then*

$$\left|\left|\left| Q - \widetilde{Q} \right|\right|\right| \le \left(\frac{2}{\sigma_n + \widetilde{\sigma}_n} + \frac{1}{\max\{\sigma_n, \widetilde{\sigma}_n\}} \right) \left|\left|\left| A - \widetilde{A} \right|\right|\right|. \tag{18}$$

Estimates (16) and (17) can be sharpened a little bit when $|||\cdot||| = \|\cdot\|_{\mathrm{F}}$. As a matter of fact, we shall have

$$\begin{aligned}
\left\| Q - \widetilde{Q} \right\|_{\mathrm{F}} &= \sqrt{\left\| I - Q^*\widetilde{Q} \right\|_{\mathrm{F}}^2 + \left\| U_2^*\widetilde{U}_1 \right\|_{\mathrm{F}}^2} \\
&\le \sqrt{\left(\frac{2}{\sigma_n + \widetilde{\sigma}_n} \right)^2 + \frac{1}{\widetilde{\sigma}_n^2}} \left\| A - \widetilde{A} \right\|_{\mathrm{F}}, \quad \text{and} \\
\left\| Q - \widetilde{Q} \right\|_{\mathrm{F}} &\le \sqrt{\left(\frac{2}{\sigma_n + \widetilde{\sigma}_n} \right)^2 + \frac{1}{\sigma_n^2}} \left\| A - \widetilde{A} \right\|_{\mathrm{F}}.
\end{aligned}$$

A consequence of these two inequalities is

Theorem 3 *Under the conditions of Theorem 2,*

$$\left\| Q - \widetilde{Q} \right\|_{\mathrm{F}} \le \sqrt{\left(\frac{2}{\sigma_n + \widetilde{\sigma}_n} \right)^2 + \left(\frac{1}{\max\{\sigma_n, \widetilde{\sigma}_n\}} \right)^2} \left\| A - \widetilde{A} \right\|_{\mathrm{F}}. \tag{19}$$

We conclude this paper with a few remarks.

1. The bound in (15) is best possible, in the sense that the equality can be achieved. Take the following case for an example: both A and $\widetilde{A}$ are $n \times n$ unitary matrices. Thus $\sigma_n = \widetilde{\sigma}_n = 1$, $Q = A$, $\widetilde{Q} = \widetilde{A}$, and

   $$\left\| Q - \widetilde{Q} \right\| = \frac{2}{\sigma_n + \widetilde{\sigma}_n} \left\| A - \widetilde{A} \right\|.$$

   It is even achievable in the real number field by taking A and $\widetilde{A}$ to be two $n \times n$ orthogonal matrices though, as we know Q behaves quite differently in the real number field (ref. 5 below).

   All previously published bounds do not achieve this!

2. Bounds (15), (18) and (19) involve both σ_n and $\widetilde{\sigma}_n$. To obtain bounds involving σ_n alone, one can weaken them by utilizing the following fact

   $$\left\| A - \widetilde{A} \right\|_2 \geq |\sigma_n - \widetilde{\sigma}_n|.$$

   For example, (15) yields

   $$\left\| Q - \widetilde{Q} \right\| \leq \frac{2}{2\sigma_n - \left\| A - \widetilde{A} \right\|_2} \left\| A - \widetilde{A} \right\|, \tag{20}$$

   provided $\left\| A - \widetilde{A} \right\|_2 < 2\sigma_n$.

3. Mathias [9] proved that for $m = n$ if $\left\| A - \widetilde{A} \right\|_2 < \sigma_n$ then

   $$\left\| Q - \widetilde{Q} \right\| \leq - \frac{\left\| A - \widetilde{A} \right\|}{\left\| A - \widetilde{A} \right\|_2} \times \ln \left(1 - \frac{\left\| A - \widetilde{A} \right\|_2}{\sigma_n} \right). \tag{21}$$

   Although his bound uses slightly different information than ours, it is always a bigger bound than (15), and sometimes much bigger (since the left hand-side of (21) could blow up). To see why, we claim that even (20), the weakened form of (15), is still no weaker than Mathias' because their ratio (his/ours) is

   $$-\frac{\ln(1-x)}{x} \cdot \left(1 - \frac{x}{2} \right) = 1 + \sum_{j=2}^{\infty} \left(\frac{1}{j+1} - \frac{1}{2j} \right) x^j > 1$$

   for $0 < x = \left\| A - \widetilde{A} \right\|_2 \Big/ \sigma_n < 1$.

4. Chen and Sun [2] studied the case $m > n$, also. But only the Frobenius norm was considered. They proved

$$\left\|Q - \widetilde{Q}\right\|_{\mathrm{F}} \leq \frac{2}{\sigma_n} \left\|A - \widetilde{A}\right\|_{\mathrm{F}}. \tag{22}$$

Without loss of generality, assume $\widetilde{\sigma}_n \leq \sigma_n$. Then it is easy to see our bound (19) is sharper than (22) when

$$\widetilde{\sigma}_n \leq \sigma_n \leq \frac{\sqrt{3}}{2-\sqrt{3}}\widetilde{\sigma}_n \approx 6.5\widetilde{\sigma}_n;$$

Otherwise (22) is sharper by a little because

$$\sqrt{\left(\frac{2}{\sigma_n + \widetilde{\sigma}_n}\right)^2 + \left(\frac{1}{\sigma_n}\right)^2} \leq \frac{\sqrt{5}}{\sigma_n} \approx \frac{2.2}{\sigma_n}$$

always. More generally, Sun and Chen [10] and Li [7] treated the cases when A and $\widetilde{A}$ do not necessarily have full column rank. Applied to our full column rank case here, the perturbation bound for the polar factor in [10] reads exactly the same as (22), and that in [7] reads

$$\left\|Q - \widetilde{Q}\right\|_{\mathrm{F}} \leq \frac{1}{\min\{\sigma_n, \widetilde{\sigma}_n\}} \left\|A - \widetilde{A}\right\|_{\mathrm{F}},$$

which is clearly sharper than (19) and (22) when $\sigma_n \approx \widetilde{\sigma}_n$. However, it may be very bad if one of σ_n and $\widetilde{\sigma}_n$ is much smaller than the other.

5. Perturbation bounds for the Q factor in polar decomposition illustrate that the change in Q is proportional to the reciprocal of the smallest singular value of A when $m = n$ and the working number field is complex. However, it has been discovered by Barrlund [1], Kenney and Laub [6] and Mathias [9] that for the real case the change in Q is proportional to the reciprocal of the sum of the two smallest singular values of A if $m = n$, which means Q is (much) less sensitive to perturbations in A in the real case than in the complex case. Our derivation above of the perturbation bound (15) for the complex case is very elementary while giving the best among those that have been published. But the author was unable to extend our derivation to do a better job for the real case. It is worth saying (as pointed out by one of the anonymous referees) that even in the real number field when

$m > n$ the change in Q is not proportional to $\frac{1}{\sigma_{n-1}+\sigma_n}$ instead of $\frac{1}{\sigma_n}$. The following example offered by the referee makes this point very clear:

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 0.8 \times 10^{-6} \\ 0 & 0 \end{pmatrix}, \quad Q = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix};$$

$$\widetilde{A} = \begin{pmatrix} 1 & 0 \\ 0 & 0.8 \times 10^{-6} \\ 0 & 0.6 \times 10^{-6} \end{pmatrix}, \quad \widetilde{Q} = \begin{pmatrix} 1 & 0 \\ 0 & 0.8 \\ 0 & 0.6 \end{pmatrix}.$$

Acknowledgements. *The author is grateful for Professor W. Kahan's and Professor J. Demmel's encouragement. He thanks Dr. N. J. Higham for his valuable comments concerning the manuscript. He is indebted to the referees for their many tips, especially the last part of* Remark 5 *to which he did not previously pay attention.*

References

[1] A. Barrlund, Perturbation bounds on the polar decomposition, *BIT*, **30**(1990), 101–113.

[2] Chun-hui Chen and Ji-guang Sun, Perturbation bounds for the polar factors, *J. Comp. Math.*, **7**(1989), 397–401.

[3] Ch. Davis and W. M. Kahan, The rotation of eigenvectors by a perturbation. III, *SIAM J. Numer. Anal.*, **7**:1 (1970), 1–46.

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

[5] R. A. Horn and Ch. R. Johnson, *Matrix analysis*, Cambridge University Press, New York, 1985.

[6] C. Kenney and A. J. Laub, Polar decompostion and matrix sign function condition estimates, *SIAM J. Sci. Statist. Comput.*, **12**(1991), 488–504.

[7] R.-C. Li, A perturbation bound for the generalized polar decomposition, *BIT*, **33**(1993), 304–308.

[8] Jian-qin Mao, The perturbation analysis of the product of singular vector matrices UV^H, *J. Comp. Math.*, **4**(1986), 245–248.

[9] R. Mathias, Perturbation bounds for the polar decomposition, *SIAM J. Matrix Anal. Appl.*, **14**(1993), 588–597.

[10] Ji-guang Sun and Chun-hui Chen, Generalized polar decomposition, *Math. Numer. Sinica*, **11**(1989), 262–273 (Chinese).

[11] G. W. Stewart and Ji-guang Sun, *Matrix Perturbation Theory*, Academic Press, New York, 1990.