ON PARLETT'S MATRIX NORM INEQUALITY FOR THE CHOLESKY DECOMPOSITION

ALAN EDELMAN
Department of Mathematics
and Lawrence Berkeley Laboratory
University of California
Berkeley, California 94720
edelman@math.berkeley.edu

WALTER F. MASCARENHAS
Dept. de Matematica
Universidade Estadual de Campinas
Caixa Postal 6065
Barao Geraldo
Campinas SP 13081, Brazil
walterm@dcc.unicamp.br

Dedicated to our friends Beresford and Velvel
on the occasion of their sixtieth birthdays.

ABSTRACT

We show that a certain matrix norm ratio studied by Parlett has a supremum that is $O(\sqrt{n})$ when the chosen norm is the Frobenius norm, while it is $O(\log n)$ for the 2-norm. This ratio arises in Parlett's analysis of the Cholesky decomposition of an n by n matrix.

Keywords: Cholesky, norm inequality, perturbation.

1 Introduction

Let U be a non-zero upper triangular matrix with diagonal entries $u_{ii} \geq -1$. Define

$$\tau(U) \;\equiv\; \frac{||U||}{||U^T + U + U^T U||}.$$

Parlett [4] asked for bounds for

$$\chi(n) \equiv \sup_{u_{ii} \geq -1} \tau(U),$$

where the supremum is taken over all such upper triangular U of dimension n.

The quantity $\chi(n)$ arises in Parlett's [4] perturbation theory of the Cholesky decomposition. The term $U^T U$ in the denominator would be neglected by first order perturbation theory but, according to Parlett, it actually helps in the analysis.

Consider perturbations δA to the identity matrix. (The analysis for perturbing any positive definite matrix can be reduced to this case through an appropriate change of coordinates [4].) The Cholesky factorization is

$$I + \delta A = (I + \delta U)^T (I + \delta U),$$

where δU is upper triangular, so that

$$\delta A = \delta U^T + \delta U + \delta U^T \delta U.$$

It follows that

$$\frac{||\delta U||}{||\delta A||} = \tau(\delta U).$$

Given $||\delta A||$, it is natural to ask for the maximum value of $||\delta U||$ and hence we study $\chi(n)$.

We bound the quantity $\chi(n)$ for both the Frobenius norm $||A||_F \equiv \sqrt{\sum a_{ij}^2} \equiv \sqrt{\sum \sigma_i^2}$ and the 2-norm $||A||_2 = \sigma_{\max}$, where the σ_i denote the singular values of A. We will denote our supremum as $\chi_F(n)$ and $\chi_2(n)$ respectively for the Frobenius norm and the 2-norm. Section 2 discusses the bounds for the Frobenius norm while Section 3 discusses the bounds for the 2-norm.

Other approaches to this problem may be found in [1] and [5]. Our bounds are tighter and have shorter proofs. The results indicate quite a difference in asymptotic behaviors as $n \to \infty$:

Frobenius norm bound:

$$\sqrt{2n-1} \le \chi_F(n) < \sqrt{n}\left(1 + \sqrt{1 + n^{-1/2}}\right). \tag{1}$$

2-norm bound:

$$0.22 \log_2(n-1) - 0.362 < \chi_2(n) < 2\log_2 n + 4. \tag{2}$$

2 The Frobenius norm's $\chi_F(n)$

Lower Bound: Let U be the $n \times n$ matrix defined by

$$u_{ij} = \begin{cases} -1 & j = i \\ 1 & j = i+1 \\ 0 & \text{otherwise.} \end{cases}$$

U has Frobenius norm $\sqrt{2n-1}$ and $U^T + U + U^T U$ has norm 1. Therefore $\chi(n) \ge \sqrt{2n-1}$.

Upper Bound: Though not logically necessary, we will find it convenient at times to make the change of variables $R = I + U$. We may then define

$$\rho(R) \equiv \tau(R - I) = \frac{||I - R||_F}{||I - R^T R||_F},$$

and ask for

$$\chi(n) \equiv \sup_{r_{ii} \geq 0} \rho(R)$$

where the supremum is over the set of n by n upper triangular matrices R with non-negative diagonals r_{ii}, excluding the identity matrix.

Our upper bound (1) is

$$\rho(R) < \sqrt{n}\left(1 + \sqrt{1 + n^{-1/2}}\right). \tag{3}$$

for any upper triangular matrix R with $r_{ii} \geq 0$ and $R \neq I$. For n large, this is roughly $2\sqrt{n}$.

Given our expression for $\rho(R)$, it is natural to study three cases: a large numerator, a small denominator, or neither a large numerator nor small denominator. Each possibility yields a bound for $\rho(R)$. The three cases are indicated schematically in Figure 1.

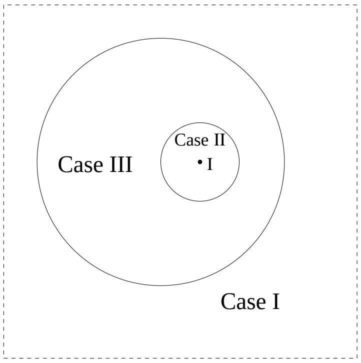

Figure 1 Proof outline: I: R large, II: $R \approx I$, III: R in between

Notice that if $||R||_F$ is very large, the quadratic term in the denominator of $\rho(R)$ is roughly the square of the numerator. This is the beauty of Parlett's suggestion of keeping the quadratic term – it will allow us to bound $\rho(R)$ for $||R||_F$ large. As R tends to the identity I, a simple argument shows that $\rho(R) \leq 2^{-1/2}$ suggesting the existence of a small bound for $\rho(R)$ when R is near the identity. If neither of these two hypotheses is true, then we again obtain a bound because the numerator in our expression for $\rho(R)$ is not too large, and also the denominator is not too small. We proceed to quantify these ideas.

- **Case I:** If $||R||_F$ is so large that $||R||_F > \sqrt{n}$, then

$$\rho(R) \leq \frac{\sqrt{n}}{||R||_F - \sqrt{n}}$$

Proof: Jensen's inequality states that the square of an average is no bigger than the average of squares:

$$(\frac{1}{n}\sum \sigma_i^2)^2 \leq \frac{1}{n}\sum(\sigma_i^2)^2.$$

Thanks to the singular value definition of the Frobenius norm

$$\frac{1}{\sqrt{n}}||R||_F^2 \leq ||R^TR||_F.$$

Therefore, $||I - R^TR||_F \geq ||R^TR||_F - ||I||_F \geq \frac{1}{\sqrt{n}}(||R||_F^2 - n)$ giving a bound for the denominator.

The triangle inequality bounds the numerator: $||I - R||_F \leq ||R||_F + ||I||_F = ||R||_F + \sqrt{n}$, and the result follows upon division.

• **Case II:** If $||I - R^TR||_F = \kappa$ for some $0 < \kappa < 1$, then

$$\rho(R) < \sqrt{2n-1}.$$

Proof: If $||I - R^TR||_F = \kappa$, then the upper left entry of $I - R^TR$ tells us that $(1 - r_{11}^2)^2 \leq \kappa^2$, so

$$(1 - r_{11})^2 \leq \kappa^2/(1 + r_{11})^2 < \kappa^2$$

since $r_{11} \geq 0$ and equality would violate $\kappa < 1$.

Let w_1 be the row vector obtained by deleting r_{11} from the first row of R. The first row (column) of $I - R^TR$ with its first component deleted is $r_{11}w_1$ (transposed). Since $||I - R^TR||_F = \kappa$,

$$(1 - r_{11})^2 + 2r_{11}^2||w_1||_2^2 \leq \kappa^2.$$

We claim that $||w_1||_2^2 < \kappa^2$ for otherwise $(1 - r_{11}^2)^2 \leq (1 - 2r_{11}^2)\kappa^2$, which implies that $\kappa^2 \geq \frac{(1-r_{11}^2)^2}{1-2r_{11}^2} \geq 1$, which would contradict the hypothesis $\kappa < 1$.

Since R^TR is similar to RR^T, we deduce that $||I - RR^T||_F = ||I - R^TR||_F = \kappa$. Let R_k be the submatrix of R obtained by taking rows and columns k through n. The matrix in the corresponding position of $I - RR^T$ is $I - R_kR_k^T$ so that $||I - R_k^TR_k||_F = ||I - R_kR_k^T||_F \leq \kappa$. Now the argument that we applied to the first row of R may be applied to the first row of each R_k so that

$$(1 - r_{kk})^2 < \kappa^2 \quad \text{and} \quad ||w_k||_2^2 < \kappa^2,$$

for every row k. Here w_k denotes the row vector past the diagonal of R in row k for $k = 1, \ldots, n-1$. Add up the contributions from the rows of $I - R$ to conclude

$$||I - R||_F^2 < (2n-1)\kappa^2,$$

yielding the upper bound for $\rho(R)$ claimed in this case.

We remark that, since this upper bound matches the lower bound obtained earlier from an example U on the boundary of this case, the worst case cannot fall in Case II. We suspect that $\rho(R)$ achieves its maximum $\chi(n)$ for some $R = I + U$ similar to the example, but slightly outside Case II.

• **Case III:** If $||I - R^T R||_F > 1$ then

$$\rho(R) < ||R||_F + \sqrt{n}.$$

Proof: The triangle inequality.

•**Final Assembly** If $||R||_F > \sqrt{n}\sqrt{1 + n^{-1/2}}$ apply Case I. If $||I - R^T R||_F \leq 1$ apply Case II, possibly taking the limiting case of $\kappa = 1$ by continuity. Otherwise $||R||_F \leq \sqrt{n}\sqrt{1 + n^{-1/2}}$ and Case III applies completing the proof.

3 The 2-norm's $\chi_2(n)$

Lower Bound: We are indebted to Roy Mathias [3] for the realization that an example of Kahan [2] serves as a lower bound for the 2-norm. The discussion that follows is a reformulation and enhancement of Mathias' observation.

We begin with the observation that if U is any non-zero upper triangular matrix, then

$$\lim_{\epsilon \to 0} \tau(\epsilon U) = \frac{||U||_2}{||U^T + U||_2}.$$

Therefore

$$\chi_2(n) \geq \sup \frac{||U||_2}{||U^T + U||_2},$$

the supremum taken over all non-zero upper triangular matrices.

Kahan [2] shows that the upper triangular matrix $W \in \mathbb{R}^{m \times m}$ with

$$W_{ij} = \begin{cases} \frac{1}{j-i} & j > i \\ 0 & j \leq i \end{cases}$$

satisfies $||W + W^T||_2 > 2\log m + \frac{1}{2} - 2\log 2 + \frac{1}{m} > 2\log m - 0.8863$. Therefore $||W||_2 > \log m - 0.44315$ by the triangle inequality.

Kahan also shows $||W - W^T||_2 < \pi$ for all m. Define a $2m$ by $2m$ matrix

$$Y = \begin{pmatrix} 0 & W \\ -W & 0 \end{pmatrix}.$$

It follows that $||Y||_2 = ||W||_2 > \log m - 0.44315$ and that $||Y + Y^T||_2 = ||W - W^T||_2 < \pi$.

Performing a perfect shuffle[1] n the rows and columns of Y produces a $2m$ by $2m$ strictly upper triangular matrix U for which

$$\frac{||U||_2}{||U + U^T||_2} > \frac{1}{\pi}\log(m) - \frac{0.44315}{\pi}.$$

[1] A perfect shuffle rearranges the numbers $1, \ldots, 2n$ into $1, n+1, 2, n+2, 3, n+3, \ldots, n, 2n$.

Replacing $2m$ with n or $n-1$ depending on whether n is even or odd, numerically computing $(0.44315+\log 2)/\pi$, and switching to base 2 logarithms, we may conclude that there exists an $n \times n$ strictly upper triangular matrix U' such that

$$\tau(U') > \frac{1}{\pi}\log(n-1) - 0.362 > 0.22\log_2(n-1) - 0.362.$$

Upper Bound: In the following analysis we always assume that U is not the zero matrix.

Theorem 3.1 *If $U \in \mathbb{R}^{2^k \times 2^k}$ is upper triangular, with $u_{ii} \geq -1$ for $1 \leq i \leq 2^k$, then*

$$\tau(U) = \frac{\|U\|_2}{\|U + U^T + U^T U\|_2} \; < \; 2k+2.$$

Furthermore, if $U \in \mathbb{R}^{n\times n}$ with the same hypotheses,

$$\tau(U) < 2\log_2 n + 4.$$

Proof: The second statement follows from the first by adding rows and columns of zeros so that 2^k is the smallest power of 2 bigger than n. The proof of the first statement is a divide and conquer style argument by induction on k, the $\log_2$ of the dimension of U. The theorem is true for $k=0$ since for 1×1 matrices with $u_{11} \geq -1$,

$$\tau(U) \; = \; \frac{|\, u_{11} \,|}{|\, 2u_{11} + u_{11}^2 \,|} \; = \; \frac{1}{2+u_{ii}} \; \leq \; 1 \; < \; 2\times 0 \; + \; 2.$$

Now let us assume that the theorem holds for dimension $2^{(k-1)}$. Take a $2^k \times 2^k$ matrix U satisfying the hypotheses of Theorem 3.1. We once again take the same strategy as in Section 2. We first consider the case that the numerator in $\tau(U)$ is large, and then second that the denominator is not too small. The only non-trivial part of the argument is the case when the denominator may be small, but the numerator is not large. We chose two convenient numbers for large and small: 3 and 3/4 respectively. Slightly better bounds may be obtained with different choices.

If $\|U\|_2 \geq 3$ then

$$\|U^T + U + U^T U\|_2 \geq \|U^T U\|_2 - \|U^T\|_2 - \|U\|_2 = \|U\|_2\,(\|U\|_2 - 2) \geq \|U\|_2.$$

Therefore, if $\|U\|_2 \geq 3$ then $\tau(U) \leq \|U\|_2/\|U\|_2 = 1 < 2k+2$. On the other hand, if $\|U\|_2 < 3$, $k \geq 1$ and $\|U + U^T + U^T U\|_2 \geq 3/4$, then $\tau(U) \; < \; 4 \; \leq \; 2k+2$.

We have only left the case $\|U + U^T + U^T U\|_2 < 3/4$ and $\|U\|_2 < 3$. In this case we write

$$U = \begin{pmatrix} X & Y \\ 0 & Z \end{pmatrix} = \begin{pmatrix} X & 0 \\ 0 & Z \end{pmatrix} + \begin{pmatrix} 0 & Y \\ 0 & 0 \end{pmatrix} \qquad (4)$$

where each block X, Y, Z has dimension $2^{(k-1)}$. The X and Z may be thought of as the easy part of the divide and conquer, while the Y is more of a nuisance term that needs to be handled gingerly.

Our partition of U leads to

$$UU^T = \begin{pmatrix} X & Y \\ 0 & Z \end{pmatrix} \begin{pmatrix} X^T & 0 \\ Y^T & Z^T \end{pmatrix} = \begin{pmatrix} XX^T + YY^T & YZ^T \\ ZY^T & ZZ^T \end{pmatrix}, \tag{5}$$

$$B \equiv U + U^T + UU^T = \begin{pmatrix} ? & Y(I+Z^T) \\ ? & Z + Z^T + ZZ^T \end{pmatrix}, \tag{6}$$

and

$$B' \equiv U + U^T + U^T U = \begin{pmatrix} X + X^T + X^T X & ? \\ ? & ? \end{pmatrix}. \tag{7}$$

The question marks indicate matrix elements that are of no immediate interest to us. Since $(I+U)(I+U)^T$ is similar to $(I+U)^T(I+U)$, we learn that the B and B' defined in (6) and (7) are similar symmetric matrices. In particular, $\|B\| = \|B'\|$. The same holds if U is replaced with Z or X.

The proof follows from the three claims below:

- Claim 1: $\|X\|_2 \ < \ 2k\ \|B\|_2$.
- Claim 2: $\|Z\|_2 \ < \ 2k\ \|B\|_2$.
- Claim 3: $\|Y\|_2 \ < \ 2\ \|B\|_2$.

In fact, if these claims are true then from (4),

$$\|U\|_2 \ \leq \ \max\{\ \|X\|_2,\ \|Z\|_2\ \} \ + \ \|Y\|_2 \ < \ (2k+2)\ \|B\|_2,$$

completing the induction.

Let us now prove the three claims above. Since $Z + Z^T + ZZ^T$ is the lower right corner of B in (6) we can use the induction hypothesis to conclude that

$$\|Z\|_2 \ < \ (2(k-1)+2)\ \|Z + Z^T + ZZ^T\|_2 \leq \ 2k\ \|B\|_2.$$

Analogously, Since $X + X^T + XX^T$ is the upper right corner of B' ,

$$\|X\|_2 \ < \ (2(k-1)+2)\ \|X + X^T + XX^T\|_2 \leq \ 2k\ \|B'\|_2 = 2k\ \|B\|_2,$$

and we have proved the first two claims.

Now the proof of the third claim. Using (6) and standard norm inequalities we obtain,

$$\|B\|_2 \ \geq \ \|Y(I+Z^T)\|_2 \ \geq \ \frac{\|Y\|_2}{\|(I+Z^T)^{-1}\|_2} \ = \ \frac{\|Y\|_2}{\|(I+Z)^{-1}\|_2} \tag{8}$$

So as to obtain information about $(I+Z)^{-1}$, we look at the partition

$$(I+U)^{-1} = \begin{pmatrix} ? & ? \\ 0 & (I+Z)^{-1} \end{pmatrix},$$

from which it is clear that

$$\|(I+Z)^{-1}\|_2 \;\le\; \|(I+U)^{-1}\|_2. \tag{9}$$

Finally, the assumption that $\|(I+U)^T(I+U)-I\|_2 < 3/4$ means that all the eigenvalues of the positive definite matrix $(I+U)^T(I+U)$ are greater than $1/4$, from which it readily follows that

$$\|(I+U)^{-1}\|_2 < 2. \tag{10}$$

Claim 3 is a direct consequence of (8), (9), and (10). This completes our proof.

4 Concluding Remarks

We suspect more precise bounds for $\chi(n)$ are obtainable. Indeed there is evidence that $\chi_F(n)$ may be $\sqrt{2n}$ asymptotically as it was in our Case II. We can also slightly improve the 2-norm bound. We satisfy ourselves here with the bounds (1) and (2) as they are tight enough to demonstrate Parlett's point that keeping the quadratic term helps produce better bounds.

Acknowledgements

We would like to thank Beresford Parlett for inviting us to work on this problem while one author was visiting the other author at Berkeley during May of 1992. We would further like to thank Velvel Kahan who served as a not very anonymous referee suggesting many valuable improvements including the observation that our original upper bound for $\chi_F(N)$ of $2.7\sqrt{n}$ could be reduced to a quantity that is roughly $2\sqrt{n}$. This work was supported by the Applied Mathematical Sciences subprogram of the Office of Energy Research, U.S. Department of Energy under Contract DE-AC03-76SF00098.

References

[1] Z.Drmač, M.Omladič, K.Veselić, On the perturbation of the Cholesky factorization, November 1992, preprint.

[2] W. Kahan, Every $n \times n$ matrix Z with real spectrum satisfies $\|Z - Z^*\| \le \|Z + Z^*\|(\log_2 n + 0.038)$, *Proc. Amer. Math. Soc. 39* (1973), 235–241.
(On page 238 the term $(\log n + \frac{1}{4} - \frac{1}{2}\log 2 + 1/2n)$ that appears on Line 3 of Section 1 should be $(\log n + \frac{1}{4} - \log 2 + 1/2n)$. On the next line, the term $0.92\log_2 n$ should be $0.44\log_2 n$, while the $\log 2$ factor on line -2 of page 238, should be $2\log 2$. The author acknowledges these corrections.)

[3] R.Mathias, Personal Communication to B. Parlett, 1991.

[4] B.N.Parlett, Perturbation theory for QR, lecture given at the Fourth SIAM Conference on Applied Linear Algebra, Minneapolis, September 1991. Document in preparation.

[5] J.-G. Sun, Perturbation bounds for the Cholesky and QR factorizations, *BIT 31* (1991), 341–352.