LAPACK Working Note 14

On Floating Point Errors in Cholesky

James Demmel
Courant Institute
New York, NY 10012

October 1989

Abstract

Let H be a symmetric positive definite matrix. Consider solving the linear system $Hx = b$ using Cholesky, forward and back substitution in the standard way, yielding a computed solution $\hat{x}$. The usual floating point error analysis says that the relative error $\|x - \hat{x}\|_2 / \|\hat{x}\|_2 = O(\varepsilon)\kappa(H)$, where $\kappa(H)$ is the condition number of H. Now write $H = DAD$, where D is diagonal and A has unit diagonal; then $\kappa(A) \le n \cdot \min_{\tilde{D}} \kappa(\tilde{D} H \tilde{D})$ and it may be that $\kappa(A) \ll \kappa(H)$. We show that the scaled error may be bounded by $\|D(x - \hat{x})\|_2 / \|D\hat{x}\|_2 = O(\varepsilon)\kappa(A)$. This often provides better error bounds than the standard formula. We show that $\kappa(A)$ is the "right" condition number in several senses. First, its reciprocal is approximately the smallest componentwise relative perturbation that makes H singular. Second, it provides a nearly sharp criterion for the successful termination of Cholesky in floating point. Third, the bound on $\|D(x - \hat{x})\|_2$ is nearly attainable.

1 Introduction

We consider the floating point error analysis of using Cholesky to solve the n by n symmetric positive definite linear system $Hx = b$. Let $\hat{x}$ be the solution computed by forward and back substitution in the usual way. The standard error analysis bounds the error by

$$\frac{\|x - \hat{x}\|_2}{\|\hat{x}\|_2} \le O(\varepsilon)\kappa(H) \tag{1}$$

where $\kappa(H) = \|H^{-1}\|_2 \|H\|_2$ is the condition number of H and ε is the machine precision.

Now write $H = DAD$, where $D = \text{diag}\,(H_{11}^{1/2}, \ldots, H_{nn}^{1/2})$ is diagonal and A has unit diagonal. It is well known [6] that the condition number $\kappa(A)$ of A is at most $n \cdot \min_{\tilde{D}} \kappa(\tilde{D} H \tilde{D})$, i.e. it is nearly the best possible diagonal scaling of H. This leads to the following algorithm for solving $Hx = b$:

1. scale H to get $A = D^{-1} H D^{-1}$,
2. solve the linear system $Ax' = D^{-1}b$, and
3. unscale to get $x = D^{-1}x'$.

This algorithm has the error bound

$$\frac{\|D(\hat{x} - x)\|_2}{\|D\hat{x}\|_2} = O(\varepsilon)\kappa(A) \tag{2}$$

which may be much better than (1), since $\kappa(H)$ has been replaced by $\kappa(A)$.

In this paper we show that this scaling of H is unnecessary because the standard unscaled Cholesky algorithm satisfies bound (2). Therefore, nothing is gained by scaling. Furthermore, we show the error bound (2) can often be interpreted as meaning that small components of the solution $\hat{x}$ are computed with as high relative accuracy as the large components, independent of D.

This observation was originally made in [8], but stated informally and without proof.

We give further evidence for considering $\kappa(A)$ to be the "natural" condition number for solving $Hx = b$. Assume without loss of generality that $\|H\|_2 = 1$. The condition number $\kappa(H)$ has the property that its reciprocal is the 2-norm of the smallest perturbation δH such that $H + \delta H$ is singular. Here we show that $\kappa(A)$ is approximately the smallest componentwise relative perturbation of H that makes it singular. Componentwise relative error is often a better model of uncertainty in H than normwise error, especially when D causes H to have entries of widely varying magnitudes.

We also give a nearly sharp condition based on $\kappa(A)$ for deciding whether Cholesky applied to H will succeed in floating point arithmetic. Finally, we show that the error bound (2) is nearly attainable for the appropriate right hand side b.

The rest of this paper is organized as follows. Section 2 discusses our error analysis of Cholesky. We also show the bounds are as good as bounds based on the Skeel condition number [1], and that estimating $\kappa(A)$ is as easy as estimating $\kappa(H)$. Section 3 discusses the componentwise relative distance of H to singularity, uses it to give a nearly sharp criterion for the successful termination of Cholesky applied to H, and shows that the error bound is attainable.

2 Error Analysis

In order to derive error bound (2), we state Cholesky's algorithm to establish notation:

Algorithm 2.1 *Cholesky decomposition $H = LL^T$ for an n by n symmetric positive definite matrix H.*

for $i = 1$ *to* n
  $L_{ii} = (H_{ii} - \sum_{k=1}^{i-1} L_{ik}^2)^{1/2}$
  for $j = i+1$ *to* n
    $L_{ji} = (H_{ji} - \sum_{k=1}^{i-1} L_{jk}L_{ik})/L_{ii}$
  endfor
endfor

Lemma 2.1 *Let L be the Cholesky factor of H computed using Algorithm 2.1 in finite precision arithmetic with precision ε. Then $LL^T = H + E$ where*

$$|E_{ij}| \le \frac{(n+1)\varepsilon}{1-(n+1)\varepsilon} \cdot (H_{ii}H_{jj})^{1/2}$$

PROOF. Let subscripted εs denote independent quantities bounded in magnitude by ε. Applying the usual rules for floating point arithmetic yields

$$L_{ii} = (1+\varepsilon_1)((1+\varepsilon_2)H_{ii} - \sum_{k=1}^{i-1} L_{ik}^2(1+i\varepsilon_{k+2}))^{1/2} \tag{3}$$

whence $\sum_{k=1}^{i} L_{ik}^2 = H_{ii} + E_{ii}$ where $|E_{ii}| \le (i+1)\varepsilon \sum_{k=1}^{i} L_{ik}^2$. Rearranging, we see $\sum_{k=1}^{i} L_{ik}^2 \le H_{ii}(1-(i+1)\varepsilon)^{-1}$ and so $|E_{ii}| \le (i+1)\varepsilon(1-(i+1)\varepsilon)^{-1}H_{ii}$ as claimed.

Next we have

$$L_{ji} = (1+\varepsilon_1)((1+\varepsilon_2)H_{ji} - \sum_{k=1}^{i-1} L_{jk}L_{ik}(1+i\varepsilon_{k+2}))/L_{ii} \tag{4}$$

whence $\sum_{k=1}^{i} L_{jk}L_{ik} = H_{ji} + E_{ji}$ where $|E_{ji}| \le (i+1)\varepsilon \sum_{k=1}^{i} |L_{jk}L_{ik}|$. By Cauchy-Schwartz

$$\sum_{k=1}^{i} |L_{jk}L_{ik}| \le (\sum_k L_{jk}^2 \cdot \sum_k L_{ik}^2)^{1/2} \le \frac{(H_{ii}H_{jj})^{1/2}}{1-(j+1)\varepsilon}$$

yielding the desired result. ∎

To simplify notation, define $\tilde{H}$ by $\tilde{H}_{ij} = (H_{ii}H_{jj})^{1/2}$. Note that $D^{-1}\tilde{H}D^{-1}$ is the matrix of all ones. Let $|E|$ denote the matrix of absolute values of entries of E: $|E|_{ij} = |E_{ij}|$. Let inequalities like $X \le Y$ between matrices X and Y be interpreted componentwise. Then Lemma 1 may be restated as $LL^T = H + E$ where $|E| \le (n+1)\varepsilon(1-(n+1)\varepsilon)^{-1}\tilde{H}$. Also $|L| \cdot |L^T| \le (1-(n+1)\varepsilon)^{-1}\tilde{H}$.

Theorem 2.1 *Let $H = DAD$ be symmetric positive definite, D diagonal, and A have unit diagonal. Let $x_T = H^{-1}b$. Let δH be a perturbation satisfying $|\delta H_{ij}| \le \eta(H_{ii}H_{jj})^{1/2}$ (δH represents initial errors in H). Consider solving the system $(H+\delta H)x = b$ by Cholesky followed by forward and back substitution. Let $\hat{x}$ be the computed solution. Then the scaled error $D(\hat{x}-x_T)$ satisfies*

$$\frac{\|D(\hat{x}-x_T)\|_2}{\|D\hat{x}\|_2} \le \left(n\eta + \frac{(3n^2+n+n^3\varepsilon)\varepsilon}{1-(n+1)\varepsilon} \right) \kappa(A)$$

PROOF. Abbreviate $(1-(n+1)\varepsilon)^{-1}$ by χ. In computing the Cholesky decomposition of $H+\delta H$ we get $H = LL^T - E - \delta H$, where E is bounded by Lemma 2.1. In solving $Ly = b$ with forward substitution, we actually get $(L+\delta L_1)\hat{y} = b$, where $|\delta L_{1,ij}| \le n\varepsilon|L_{ij}|$ [7]. In solving $L^T x = \hat{y}$ we actually get $(L+\delta L_2)^T\hat{x} = \hat{y}$ where $|\delta L_{2,ij}| \le n\varepsilon|L_{ij}|$. Altogether

$$(H + \delta H + E + \delta L_1 L^T + L\delta L_2^T + \delta L_1 \delta L_2^T)\hat{x} \equiv (H+F)\hat{x} = b$$

Now write this as

$$D^{-1}(H+F)D^{-1}D\hat{x} = (A + D^{-1}FD^{-1})D\hat{x} = D^{-1}b$$

The usual techniques show that

$$\frac{\|D(\hat{x}-x_T)\|_2}{\|D\hat{x}\|_2} \le \kappa(A) \cdot \frac{\|D^{-1}FD^{-1}\|_2}{\|A\|_2}$$

so it suffices to estimate

$$\begin{aligned}
\|D^{-1}FD^{-1}\|_2 &\leq \|D^{-1}\delta HD^{-1}\|_2 + \|D^{-1}ED^{-1}\|_2 + \|D^{-1}|\delta L_1| \cdot |L^T|D^{-1}\|_2 \\
&\quad + \|D^{-1}|L| \cdot |\delta L_2^T|D^{-1}\|_2 + \|D^{-1}|\delta L_1| \cdot |\delta L_2^T|D^{-1}\|_2 \\
&\leq \eta\|D^{-1}\tilde{H}D^{-1}\|_2 + \chi(n+1)\varepsilon\|D^{-1}\tilde{H}D^{-1}\|_2 + \chi n\varepsilon\|D^{-1}\tilde{H}D^{-1}\|_2 \\
&\quad + \chi n\varepsilon\|D^{-1}\tilde{H}D^{-1}\|_2 + \chi n^2\varepsilon^2\|D^{-1}\tilde{H}D^{-1}\|_2 \\
&\leq n\eta + \chi(n^2+n)\varepsilon + \chi n^2\varepsilon + \chi n^2\varepsilon + \chi n^3\varepsilon^2
\end{aligned}$$

This proves the result. ∎

Now we interpret this error bound. Suppose $\kappa(A)$ is moderate and D has diagonal entries of widely varying magnitudes. Then we claim the solution components $x = H^{-1}b = D^{-1}A^{-1}D^{-1}b$ will often be scaled like D^{-1}; this is because A's moderate conditioning will usually mean the components of $A^{-1}D^{-1}b$ will be comparable in size. In other words, the components of Dx and $D\hat{x}$ will be moderate in size. Thus, Theorem 2.1 says we will get the entries of Dx to moderate absolute accuracy, and hence the tiny as well as the large components of x to moderate relative accuracy.

We illustrate this reasoning with the following somewhat extreme example. Let $H = DAD$ where

$$A = \begin{bmatrix} 1 & -.11 & .24 & -.34 \\ -.11 & 1 & .07 & .30 \\ .24 & .07 & 1 & .65 \\ -.34 & .30 & .65 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 42 \\ -26 \\ 24 \\ 34 \end{bmatrix}$$

and $D = \text{diag}\ (1, 10^5, 10^{-10}, 10^{15})$. Here $\kappa(H) \approx 10^{50}$ but $\kappa(A) \approx 2.0$. Then the computed solution from Cholesky (with $\varepsilon \approx 2.2 \cdot 10^{-16}$) is

$$\hat{x} = \begin{bmatrix} -3.886390563036245 \cdot 10^{11} \\ 9.923477775911641 \cdot 10^{5} \\ 7.473020816898015 \cdot 10^{21} \\ -6.476540655693383 \cdot 10^{-4} \end{bmatrix}$$

and the bound of Theorem (2.1) is

$$\left\| \begin{bmatrix} -.3641849059026662 - c_1x_1 \\ .09299067506030982 - c_2x_2 \\ .7002799484168281 - c_3x_3 \\ -.6069020369959377 - c_4x_4 \end{bmatrix} \right\|_2 \leq 68\varepsilon \approx 1.5 \cdot 10^{-14}$$

where $c_i = 9.370774758627102 \cdot 10^{-13} D_{ii}$. This implies each solution component is correct to almost 14 decimal digits, even though the traditional condition number $\kappa(H) \approx 10^{50}$.

We note that the Skeel condition number $\|\ |A^{-1}| \cdot |A|\ \|_\infty$ of A cannot differ from $\kappa(A)$ by more than a factor of n. Thus, even though Cholesky does not guarantee small componentwise backward error without single precision iterative refinement [1], the error bounds are essentially the same with or without single precision iterative refinement.

Finally, given the Cholesky factors of H, we show that it is just as easy to estimate $\kappa(A)$ as $\kappa(H)$. In order to use the Hager estimator [3, 5, 4] to estimate $\|A^{-1}\|_1$ (which overestimates $\|A^{-1}\|_2$ by at most a factor of $\sqrt{n}$), one needs to be able to multiply an arbitrary vector by A^{-1}. But since $A^{-1} = D^{-1}L^{-T}L^{-1}D^{-1}$, this can be done by multiplying by D^{-1}, doing forward and back substitution with L (multiplying by H^{-1}), and again multiplying by D^{-1}.

3 Optimality of the Bound

It is common in numerical analysis that the reciprocal of the condition number estimates the smallest perturbation of the problem that make it singular [2]. For linear equation solving the relationship is exact: Assume without loss of generality that $\|H\|_2 = 1$. Then $\kappa(H) = \|H^{-1}\|_2$ is exactly the reciprocal of the 2-norm of the smallest δH such that $H + \delta H$ is singular.

Here we show that $\kappa(A)$ has a similar property if we measure the distance to singularity in terms of the largest componentwise relative perturbation of H. Define the componentwise relative distance between H and H' as

$$reldist(H, H') \equiv \max_{ij} \frac{|H_{ij} - H'_{ij}|}{|H_{ij}|}$$

We seek a singular H' which (nearly) minimizes $reldist(H, H')$.

Theorem 3.1 *Let H be symmetric positive definite, and write $H = DAD$ where D is diagonal and A has unit diagonal. Let H' be singular. Then*

$$reldist(H, H') \geq \frac{1}{n \cdot \kappa(A)} \tag{5}$$

Now let $H' = H - \lambda_{\min}(A) \cdot D^2$. Then H' is singular and

$$reldist(H, H') \leq \frac{n}{\kappa(A)} \tag{6}$$

Thus, to within a factor of n, $\kappa(A)$ is the reciprocal of the smallest componentwise relative change in H that makes it singular.

Proof. Since *reldist* is independent of D, we may assume $D = I$ and $H = A$. Then A positive definite with unit diagonal means $|A_{ij}| \leq 1$ and $\|A\|_2 \leq n$. Suppose $\delta H = \delta A$ has the property that $A + \delta A$ is singular. Then

$$\max_{ij} \frac{|\delta A_{ij}|}{|A_{ij}|} \geq \max_{ij} |\delta A_{ij}| \geq \frac{\|\delta A\|_2}{n} \geq \frac{\lambda_{\min}(A)}{n} = \frac{\lambda_{\min}(A)}{\lambda_{\max}(A)} \frac{\lambda_{\max}(A)}{n} \geq \frac{1}{n \cdot \kappa(A)}$$

proving (5). The choice $H' = H - \lambda_{\min}(A) \cdot D^2$ means $\delta A = -\lambda_{\min} I$, and so

$$\max_{ij} \frac{|\delta A_{ij}|}{|A_{ij}|} = \lambda_{\min}(A) = \lambda_{\max}(A) \cdot \frac{\lambda_{\min}(A)}{\lambda_{\max}(A)} \leq \frac{n}{\kappa(A)}$$

proving (6). ∎

We interpret this theorem as follows. In matrices like the example in section 2, tiny components may be known as accurately as large components; this may be because of the unit in which they are measured, for example. Thus, it is inappropriate to measure the uncertainty in the matrix with a norm like the 2-norm, since this permits equal size perturbations in each entry; a measure like *reldist* is more appropriate. And, from Theorem 2.1, we see that that the componentwise relative distance to singularity bears the proper (reciprocal) relationship to the condition number.

In addition, we can make the following rather sharp statement about when Cholesky succeeds or fails when applied to H:

Theorem 3.2 *Let $H = DAD$ as before, let ε be machine precision, and let $\chi = (1-(n-1)\varepsilon)^{-1} \approx 1$. Then if $\lambda_{\min}(A) > \varepsilon \cdot \chi(n^2+n)$, Cholesky applied to H is guaranteed to succeed (compute a real nonsingular L); we assume no underflow occurs. If $\lambda_{\min}(A) \le \varepsilon$, then there exist rounding errors which will cause Cholesky to fail. Finally, If $\lambda_{\min}(A) < -\varepsilon \cdot \chi(n^2+5n)$, then Cholesky is guaranteed to fail (H and A are not positive definite).*

PROOF. If Cholesky fails, then by Lemma 2.1 we must have $H+E$ indefinite or $A+D^{-1}ED^{-1}$ indefinite, or $\lambda_{\min}(A) \le \|D^{-1}ED^{-1}\|_2 \le \chi(n^2+n)\varepsilon$ so the converse inequality means Cholesky must succeed. Similarly, if Cholesky succeeds, $A+D^{-1}ED^{-1}$ must be definite so $0 < \lambda_{\min}(A) + \|D^{-1}ED^{-1}\|_2 \le \lambda_{\min}(A) + \chi(n^2+n)\varepsilon$. Now suppose $\lambda_{\min}(A) \le \varepsilon$. Choose the rounding errors ε_i in equations (3) so that $\varepsilon_2 = -\varepsilon$ and the other $\varepsilon_i = 0$. Choose all the rounding errors in equation (4) to be zero. Then L would be the exact Cholesky factor of $H - \varepsilon D^2$, which by Theorem 3.1 is not positive definite. ∎

Finally, we note that the bound of Theorem 2.1 is optimal in the following sense. For any H there is a b and a diagonal perturbation δH for which the bound on $\|D(\hat{x}-x_T)\|_2$ is nearly attainable. We simply choose $\delta H_{ii} = \eta H_{ii}$ and b so that Dx_T is parallel to the eigenvector of the smallest eigenvalue of A. In particular, this δH makes only small relative changes in the entries of H.

Acknowledgements. The author acknowledges NSF grants DCR-8552474 and ASC-8715728. He is a Presidential Young Investigator. He also thanks Nick Higham for several valuable comments.

References

[1] M. Arioli, J. Demmel, and I. S. Duff. Solving sparse linear systems witwh sparse backward error. *SIAM J. Matrix Anal. Appl.*, 10(2):165–190, April 1989.

[2] James Demmel. On condition numbers and the distance to the nearest ill-posed problem. *Numerische Mathematik*, 51(3):251–289, July 1987.

[3] W. W. Hager. Condition estimators. *SIAM Journal on Scientific and Statistical Computing*, 5:311–316, 1984.

[4] N. Higham. Algorithm 674: fortran codes for estimating the one-norm of a real or complex matrix, with applications to condition estimation. *ACM Trans. Math. Software*, 15(2):168, 1989.

[5] N. Higham. A survey of condition number estimation for triangular matrices. *SIAM Review*, 29:575–596, 1987.

[6] A. Van Der Sluis. Condition numbers and equilibration of matrices. *Numerische Mathematik*, 14:14–23, 1969.

[7] J. H. Wilkinson. *The Algebraic Eigenvalue Problem.* Oxford University Press, 1965.

[8] J. H. Wilkinson. A priori error analysis of algebraic processes. In I. G. Petrovsky, editor, *Proceedings of the International Congress of Mathematicians, Moscow 1966*, pages 629–640, Mir Publishers, 1968.