JOURNAL OF COMPUTATIONAL PHYSICS 38, 114–123 (1980)

# Note

## On the Problem of Unstable Pivots in the Incomplete *LU*-Conjugate Gradient Method*

In Ref. [1] the point was raised that in incomplete Cholesky decompositions pivots ($L_{ii}$ in Eq.(7b) of Ref. [1]) may arise which are $\leqslant 0$. Similarly, in incomplete $LU$ decompositions pivots ($U_{ii}$ in Appendix A of Ref. [1]) may arise which are $= 0$. In either case this leads to a breakdown of the algorithms unless something is done to "fix" these "bad" pivots. In Ref. [1] it was suggested that if in the course of an incomplete Cholesky decomposition one comes up with $L_{ii} \leqslant 0$, one should set

$$L_{ii} = \sum_{j=1}^{(i-1)} |L_{ij}| + \sum_{j=(i+1)}^{N} |L_{ji}|$$

thus assuring diagonal dominance of the $i$th row and column of $L$. This has worked quite well in practice. In [1] it was also suggested that if in the course of an incomplete $LU$ decomposition one comes up with $U_{ii} = 0$, "simply set $U_{ii}$ to a nonzero value and go on with the algorithm." This is too vague to be very satisfactory and in what follows we shall derive a quantitative theory to determine:

(1) Exactly how small must the pivot be before we alter it? Obviously on a computer of given accuracy there is some pivot value which is small enough to make the algorithm go unstable but is not yet a hard zero.

(2) If a pivot needs to be "fixed," what value should we set it to so as to get the best possible approximate inverse?

We show that for the special case of complete $LU$ decompositions of tridiagonal matrices our pivot fixing prescription will always work, and we derive an error bound on $|(LU)_{ij} - A_{ij}|$. In the case of complete $LU$ decompositions of dense matrices we present a variety of numerical examples which indicate the method works quite well.

It shall be assumed in what follows that we are dealing with a floating-point computer where the mantissa is stored with $t$ binary digits so the accuracy with which any number can be stored is 1 part in $2^t$.

* Work performed under the auspices of the U.S. Department of Energy by the Lawrence Livermore Laboratory under Contract W-7405-ENG-48.



0021-9991/80/130114–10$02.00/0

## 1. Basic Theory

Consider first complete $LU$ decomposition.

We first note the theorem of Wilkinson [2] which states that if one performs an $LU$ decomposition of a matrix $A$ with exact arithmetic except for an error which one makes in computing one element $L_{ij}(i>j)$, then all elements of the error matrix $E=A-LU$ will be zero except $E_{ij}$. To see this we note that even if an error was made in computing some element $L_{ij}$, any later elements $L_{kl}$ or $U_{kl}$ satisfy

$$U_{ll}L_{kl}=A_{kl}-\sum_{n<l}L_{kn}U_{nl}$$

or

$$U_{kl}=A_{kl}-\sum_{n<k}L_{kn}U_{nl},$$

so even if $L_{ij}$ enters into the summation on the right-hand side of these equations and has the wrong value, $L_{kl}$ or $U_{kl}$ will be altered in just such a way as to make $(LU)_{kl}$ exactly equal to $A_{kl}$. Similarly, if we make an error only in $U_{ij}$ $(i\leqslant j)$, then only $E_{ij}$ will be nonzero. Thus *in exact arithmetic* an error in computing one $L_{ij}$ or $U_{ij}$ may cause later elements of $L$ or $U$ to change but later elements of $(LU)$ will be unaffected.

Now consider the effect of altering a pivot in $LU$ decomposition. We take

$$U_{ii}=A_{ii}-\sum_{j<i}L_{ij}U_{ji}-\delta_i=\tilde{U}_{ii}-\delta_i,$$

where $\delta_i$ is our "fix" that alters the $i$th pivot away from its unstable value, and $\tilde{U}_{ii}=A_{ii}-\sum_{j<i}L_{ij}U_{ji}$ is the unfixed pivot. By Wilkinson's theorem the sole effect of this on the error matrix in exact arithmetic is to introduce one nonzero element,

$$E_{ii}=+\delta_i,$$

into $E=A-LU$. What is the effect on round-off error in the rest of the decomposition? First, all the elements of $L$ in the $i$th column become

$$\begin{aligned}L_{ji}&=\left(A_{ji}-\sum_{k<i}L_{jk}U_{ki}\right)\Big/U_{ii}\\&=S_{ji}/(\tilde{U}_{ii}-\delta_i),\qquad j>i\end{aligned}$$

where

$$S_{ji}=A_{ji}-\sum_{k<i}L_{jk}U_{ki}.$$

(Note that $E_{ji}$ is independent of $\delta_i$ since $L_{ji}(\tilde{U}_{ii} - \delta_i)$ is a constant independent of the choice of $\delta_i$.) Now consider the round-off error in the calculation of any subsequent $L$ or $U$ element which depends on $L_{ji}$. We have

$$L_{jn}U_{nn} = A_{jn} - L_{ji}U_{in} - \sum_k L_{jk}U_{kn}, \qquad n < j, \quad k < n, \quad k \neq i, \quad j, n > i$$

and

$$U_{jn} = A_{jn} - L_{ji}U_{in} - \sum_k L_{jk}U_{kn}, \qquad n \geqslant j, \quad k < j, \quad k \neq i, \quad j, n > i.$$

Therefore the contribution from round-off error to $E_{jn}$ due to the $L_{ji}$ term is given by

$$\begin{aligned} E_{jn} &= \pm L_{ji}U_{in}\,2^{-t} \\ &= \pm S_{ji}U_{in}\,2^{-t}/(\tilde{U}_{ii} - \delta_i), \qquad j, n > i. \end{aligned}$$

To proceed further we need a definition of what constitutes the best $LU$ decomposition. We define the best pivot fix, $\delta_i$, to be the one that minimizes the maximum of all the $\delta_i$-dependent contributions to the error matrix $E$. These $\delta_i$-dependent error contributions are to $E_{ii}$, which has an error contribution $\delta_i$ and to all $E_{jn}$ (with $j, n > i$) which have error contributions $S_{ji}U_{in}\,2^{-t}/(\tilde{U}_{ii} - \delta_i)$. If we set

$$\sigma_i = \max_{j>1} |S_{ji}|$$

and

$$\mu_i = \max_{n>i} |U_{in}|,$$

then the maximum error contribution will be

$$\max(|\delta_i|, 2^{-t}\sigma_i\mu_i/|\tilde{U}_{ii} - \delta_i|).$$

The value of $\delta_i$ which minimizes this is

$$\delta_i = -\operatorname{sign}(\tilde{U}_{ii})((\tilde{U}_{ii}/2)^2 + 2^{-t}\sigma_i\mu_i)^{1/2} - (\tilde{U}_{ii}/2))$$

or

$$U_{ii} = \tilde{U}_{ij} - \delta_i = (\tilde{U}_{ii}/2) + \operatorname{sign}(\tilde{U}_{ii})((\tilde{U}_{ii}/2)^2 + 2^{-t}\sigma_i\mu_i)^{1/2}.$$

For ease of computation we approximate this by

$$U_{ii} = \begin{cases} \tilde{U}_{ii}, & (\tilde{U}_{ii})^2 > 2^{-t}\sigma_i\mu_i \\ \operatorname{sign}(\tilde{U}_{ii})(2^{-t}\sigma_i\mu_i)^{1/2}, & (\tilde{U}_{ii})^2 < 2^{-t}\sigma_i\mu_i. \end{cases}$$

Thus our prescription for $LU$ decomposition is as follows:

(1) As we compute the $i$th row of $U$ we find

$$\mu_i = \max_{n>i} |U_{in}|.$$

(2) As we compute the $i$th column of $L$ (but before dividing by the $i$th pivot) we find

$$\sigma_i = \max_{j>i} |S_{ji}|.$$

(3) If $(U_{ii})^2 > 2^{-t}\sigma_i\mu_i$, we leave the $i$th pivot $(U_{ii})$ unchanged. If $(U_{ii})^2 < 2^{-t}\sigma_i\mu_i$, we set the $i$th pivot (the new $U_{ii}$) $= \text{sign}(U_{ii})(2^{-t}\sigma_i\mu_i)^{1/2}$.

(4) We divide all the $S_{ji}$ $(j > i)$ by the new $U_{ii}$ to get the $i$th column of $L$.

For incomplete $LU$ decompositions the arguments and the resulting prescriptions are exactly the same except only elements within the sparsity pattern chosen are calculated, used, and stored.

For complete $LU$ decompositions one may ask if this prescription will always work. We have shown that with our pivot prescription the error in any element of $(LU)$ is less than

$$\max_i (2^{-t}\sigma_i\mu_i)^{1/2}.$$

Therefore the prescription will work as long as $\sigma_i$ and $\mu_i$ are not very large.

For the case of tridiagonal matrices which is of great practical importance to the computational physicist, our method will always work because for this case $\sigma_i = A_{i+1,i}$ and $\mu_i = A_{i,i+1}$. The errors in $(LU)_{i+1,i}$ and $(LU)_{i,i+1}$ are zero since

$$(LU)_{i+1,i} = L_{i+1,i}U_{ii} = (A_{i+1,i}/U_{ii})\,U_{ii},$$

and

$$(LU)_{i,i+1} = L_{ii}U_{i,i+1} = A_{i,i+1}.$$

The error in $(LU)_{ii}$ is

$$E_{ii} < (2^{-t}\sigma_i\mu_i)^{1/2} = 2^{-t/2}(|A_{i+1,i}A_{i,i+1}|)^{1/2}.$$

Thus the error in the diagonal elements of $(LU)$ will always be less than $2^{-t/2}$ times the gemetric mean of the off-diagonal elements.

In the general case, earlier pivot shifts can affect later $\sigma_i$ and $\mu_i$ and cause them to grow. For example, consider the $N$-dimensional nonsingular matrix (for which I thank my reviewer).

$$A_{ij} = 1, \qquad i > j \quad \text{and} \quad j = 1, 2, 3, \ldots, (N-1)$$

$$A_{iN} = 1, \qquad i = 1, 2, \ldots, (N-1)$$

$$A_{ij} = 0 \qquad \text{otherwise.}$$

The first $N-1$ pivots of this matrix are all bad. Our pivot prescription leads to the following series of pivots if we take sign(0) = +1,

$$U_{11} = 2^{-t/2},$$
$$U_{ii} = U_{(i-1)(i-1)}((1/U_{(i-1)(i-1)}) - 1)^{1/2}$$

and

$$(\sigma_i \mu_i)^{1/2} = 2^{t/2} U_{ii}$$

so $U_{ii} \to 1/2$ from below. If we take sign(0) = −1 we get

$$U_{11} = 2^{-t/2},$$
$$U_{ii} = U_{(i-1)(i-1)}(1 - (1/U_{(i-1)(i-1)})^{1/2}.$$

So $U_{ii} \to -\infty$. Thus in the general case if the matrix is artfully enough arranged cumulative growth of the $\sigma_i \mu_i$ is *possible*. Typically though, cumulative growth does not seem to happen as is shown in Example 4 of Section II.

One can protect against the possibility of cumulative growth of pivot error by keeping a count of how many pivots had to be changed while decomposing a given matrix and printing a warning message if more than a few pivots were changed.

We have chosen our pivot shifts so as to minimize the elements of

$$E = A - LU.$$

Given that $E$ is small, what error bounds can be put on the error in the solution $X$ of $AX = Y$, due to $A \neq LU$.

Let $X_a = (LU)^{-1} Y = (A - E)^{-1} Y$. Then the error in $X$,

$$\begin{aligned} X - X_a &= X - (A - E)^{-1} Y \\ &= (I - (A - E)^{-1} A) X = (A - E)^{-1} (A - E - A) X \\ &= -(A - E)^{-1} EX. \end{aligned}$$

now let $\|X\| = (\sum X_i^2)^{1/2}$, and let $\|A\|$ be the subordinate norm

$$\|A\| = \sup(\|AX\|/\|X\|)$$

Then

$$\begin{aligned} \|X - X_a\| &= \|(A - E)^{-1} EX\| \\ &\leqslant \|(A - E)^{-1} E\| \, \|X\| \\ &\leqslant \|(A - E)^{-1}\| \, \|E\| \, \|X\| \end{aligned}$$

or

$$\|X - X_a\|/\|X\| \leqslant \|E\| \, \|(A-E)^{-1}\|$$
$$= (\|E\|/\|A-E\|)\kappa,$$

where $\kappa$ is the condition number $= \|A-E\| \, \|(A-E)^{-1}\|$. Typically in real problems we have tried (see Section II) we find

$$\|E\| \sim 2^{-t/2} \|A\|,$$

so

$$\|X_a - X\|/\|X\| < 2^{-t/2}\kappa.$$

Of course this is quite a pessimistic estimate since in most cases $X$ will not happen to be the largest eigenvalue of $E$ and the smallest eigenvalue of $A-E$.

Note that the total error in $X$ will be given by [3] $[A - E + (\delta L)U + L(\delta U) + (\delta L)(\delta U)]X_a = Y$, whose $\delta L$ and $\delta U$ come from round-off error during the solve, and $\delta L \sim 2^{-t}L$, $\delta U \sim 2^{-t}U$. Since our pivot shifts tend to prevent large growth in $L$ and $U$ and thus in $\delta L$ and $\delta U$, round-off errors in the solve also tend to be decreased.

## II. Numerical Results

We tested our new prescription on four matrices whose complete $LU$ decompositions (without pivoting) are totally unstable. For each matrix, complete $LU$ decomposition without pivoting but with our new pivot prescription was tried.

Example 1. $A_{i,i+1} = -A_{i+1,i} = 1$ and all other elements $= 0$. If the dimension $N$ is even, the matrix is nonsingular, while if $N$ is odd, it is singular. We took $N = 1000$. Since the determinant of every odd-dimensional principal minor is zero, every other pivot is bad and $LU$ decomposition (without pivoting) blows up immediately. On the CDC 7600 ($t = 48$ binary digit mantissa) we performed the complete $LU$ decomposition using our modified pivot prescription. We then calculated

$$A - LU = E_1$$

and

$$I - (LU)^{-1}A = E_2.$$

The largest elements of $E_1$ and $E_2$ had absolute values of $\approx 2^{-24}$.

Example 2.

$$A_{ij} = 1, \qquad i + j \leqslant N + 1,$$

and

$$A_{ij} = 0, \qquad i + j > N + 1,$$

where $N$ is the dimension and we used $N = 20$. $A$ is nonsingular but all its principal minors except for $A$ itself and the $1 \times 1$ principal minor are singular. Therefore $LU$ decomposition is totally unstable but with our modified prescription we again obtain an $LU$ decomposition which gives an $E_1$ and $E_2$ the largest elements of which have absolute values $\approx 2^{-24}$.

EXAMPLE 3.

$$A = B - \alpha I,$$

where $B_{ii} = 2$, $B_{i,i+1} = B_{i+1,i} = -1$, and all other elements of $B$ are zero. The eigenvalues of $B$ are

$$\lambda_j = 4 \sin^2(\pi j/(2N + 2)), \qquad j = 1, 2, \ldots, N,$$

where $N$ is the dimension. We tried three different matrices of this type with $N = 1000$ and $\alpha = 4 \sin^2(\pi/M)$, where $M = 5, 20$, and 100. Then the $i$th principal minor is singular for $i = 5n - 1$; $n = 1, 2, 3, \ldots, 200$ if $M = 5$, $i = 10n - 1$; $n = 1, 2, 3, \ldots, 100$ if $M = 20$, $i = 50n - 1$; $n = 1, 2, 3, \ldots, 20$ if $M = 100$. Thus every fifth ($M = 5$), 10th ($M = 20$) or 50th ($M = 100$) pivot is bad and regular $LU$ decomposition blows up. Our prescription produced $LU$ decompositions such that the largest elements of $E_1$ and $E_2$ had absolute values $\approx 2^{-24}$.

EXAMPLE 4. $A$ was a $25 \times 25$ matrix whose elements were independent random numbers evenly distributed between $-1$ and $+1$. Then $A_{6,6}$, $A_{12,12}$, $A_{15,15}$, $A_{18,18}$, $A_{21,21}$, and $A_{24,24}$ were shifted so as to make the sixth, 12th, 15th, 18th, 21st, and 24th principal minors singular. One hundred random matrices each with six bad pivots were generated in this way. $LU$ decomposition of each of these 100 matrices is totally unstable but $LU$ decomposition with our modified prescription worked very well for all 100 matrices. For each matrix let $\varepsilon_1$ and $\varepsilon_2$ be the largest element of $E_1$ and $E_2$ respectively, i.e.:

$$\varepsilon_1 = \max_{i,j} |E_{1ij}|$$

and

$$\varepsilon_2 = \max_{i,j} |E_{2ij}|.$$

Then the average over all 100 matrices of $\varepsilon_1$ was

$$\bar{\varepsilon}_1 = 5.5 \times 10^{-7},$$

and the same average for $\varepsilon_2$ was

$$\bar{\varepsilon}_2 = 4 \times 10^{-6}.$$

The maximum over all 100 matrices of $\varepsilon_1$ was

$$\varepsilon_1^M = 3.7 \times 10^{-6}$$

and the same maximum for $\varepsilon_2$ was

$$\varepsilon_2^M = 1.8 \times 10^{-4}.$$

Thus, in general, with our modified pivot prescription the elements of $L$ and $U$ do not seem to grow cumulatively and only in very special cases such as that presented in Section I are we likely to encounter this problem.

Thus on all four test problems our modified prescription was able to take matrices whose regular $LU$ decomposition (without pivoting) was completely unstable and produce an approximate $LU$ decomposition such that on the average $(LU)^{-1}A = I$ and $LU = A$ to 24 binary digit accuracy on a 48 binary digit machine, and in the worst case (Example 4—worst of 100 matrices) $(LU)^{-1}A = I$ to 4 decimal place accuracy and $LU = A$ to 6 place accuracy.

## III. Additional Applications

Our experience with the four test problems suggests that, in addition to its intended application in incomplete factorization schemes, our pivot prescription might be a very viable alternative to pivoting in complete $LU$ decomposition algorithms. Complete $LU$ decomposition with our modified pivot prescription (and no permutations of rows or columns) followed by conjugate gradient if needed (i.e., if any pivots required modification) to further improve the answer has many advantages over the usual pivoting schemes such as:

(1) In the sparse matrix case pivoting usually causes many *more* elements which were zero in the original matrix to fill in and become nonzero in $L$ and $U$ than would have been the case if there had been no permutations of rows or columns. In our method we suffer from none of this additional fill in.

Furthermore, various very effective minimal storage schemes (such as nested dissection) for Gauss elimination with sparse matrices have so far only been applicable to positive definite symmetric matrices because for more general matrices pivoting was required. Our modified pivot prescription would make these methods applicable to any matrices with the appropriate sparsity patterns.

(2) $LU$ decomposition without pivoting vectorizes very well and Fong and Jordan [4] have shown that it can go at super vector speeds on the Cray 1 computer. However, permuting rows and columns does not vectorize well and so our modified pivot method, which already entails much less work than pivoting in scalar computing, will have an even greater advantage on vector machines.

(3) Computational experience shows that in many production code applications of $LU$ decomposition most of the matrices produced do not have "bad"

pivots and so, with conventional methods, either most of the time a lot of unnecessary permuting of rows and columns is being done when just as good results would have been obtained from $LU$ decomposition without pivoting, or else nothing at all is done about bad pivots in which case the code occasionally blows up. It is only the occasional matrix with "bad" pivots that needs some special attention and our method does very little extra work and when a "bad" pivot shows up, it keeps the code from crashing.

It will now be shown that if we do an exact $LU$ decomposition of some matrix $M$ except that we shift ("fix") $p$ of the pivots, and if we follow our $\tilde{L}U$ decomposition with conjugate gradient iterations, then we will get the exact answer in $(2p + 1)$ iterations.

If pivots $i(j), j = 1, 2, \ldots, p$ have been shifted by an amount $\delta_{i(j)}$, then

$$M = LU + \sum_{j=1}^{p} \delta_{i(j)} e_{i(j)} e_{i(j)}^T,$$

where $e_i$ is the unit column vector whose $i$th component is 1 while all other components are 0. We then use conjugate gradient to solve

$$Mx = y,$$

but in the form

$$(L^{-1}MU^{-1})(Ux) = (L^{-1}y) = Nz = w,$$

where

$$\begin{aligned} N &= I + \sum_{j=1}^{p} \delta_{i(j)} f_j g_j^T, \\ w &= L^{-1}y, \\ z &= Ux, \\ f_j &= L^{-1}e_{i(j)}, \\ g_j^T &= e_{i(j)}^T U^{-1}. \end{aligned}$$

The conjugate gradient algorithm for nonsymmetric matrices converges in $r$ iterations if $N^TN$ has only $r$ distinct eigenvalues.

Now

$$\begin{aligned} N^TN = I &+ \sum_{j=1}^{p} \delta_{i(j)}(f_j g_j^T + g_j f_j^T) \\ &+ \sum_{k,j=1}^{p} \delta_{i(j)} \delta_{i(k)} g_j (f_j^T f_k) g_k^T. \end{aligned}$$

Consider the linear subspace $V$ spanned by $f_1, f_2, \ldots, f_p, g_1, g_2, \ldots, g_p$, and $\bar{V}$ the linear subspace of all vectors orthogonal to $V$. Then if $x \in \bar{V}$, $N^T N x = x$, so $N^T N$ is nontrivial only on the $2p$-dimensional linear subspace $V$ and so $N^T N$ can have at most $(2p+1)$ distinct eigenvalues. If we take as our initial guess vector $x_0 = (LU)^{-1} y$, then our initial error vector $(\delta z)_0 = U(x_0 - M^{-1} y)$ lies entirely within $V$ and so conjugate gradient sets the exact answer in $(2p)$ iterations.

This suggests that conjugate gradient is a good choice for an iterative scheme to follow complete $LU$ decomposition with pivot shifting to "fix" unstable pivots. I wish to thank Gene Golub for bringing this point to my attention.

## IV. A more Efficient Form for the Incomplete $LU$-Conjugate Gradient Algorithm

The algorithm given in Ref. [1, Appendix A, Eqs. (9′a)–(9′e)], may be made computationally more efficient by the substitution $p_i(\text{old}) = (U^T U)^{-1} p_i(\text{new})$. This gives the more efficient form of the algorihtm:

$$r_0 = y - AX_0 \qquad \text{and} \qquad p_0 = A^T (LL^T)^{-1} r_0,$$

$$a_i = \frac{(r_i, (LL^T)^{-1} r_i)}{(p_i, (U^T U)^{-1} p_i)}, \tag{9′a}$$

$$x_{i+1} = x_i + a_i (U^T U)^{-1} p_i, \tag{9′b}$$

$$r_{i+1} = r_i - a_i A (U^T U)^{-1} p_i, \tag{9′c}$$

$$b_i = \frac{(r_{i+1}, (LL^T)^{-1} r_{i+1})}{(r_i, (LL^T)^{-1} r_i)}, \tag{9′d}$$

$$p_{i+1} = A^T (LL^T)^{-1} r_{i+1} + b_i p_i, \qquad i = 0, 1, 2, \ldots. \tag{9′e}$$

## References

1. D. S. Kershaw, *J. Comput. Phys.* **26** (1978), 43.
2. J. H. Wilkinson, "The Algebraic Eigenvalue Problem," Chap. 4, Sect. 40, Oxford Univ. Press (Clarendon), London, 1965.
3. J. H. Wilkinson, "The Algebraic Eigenvalue Problem," Chap. 4 p. 63, Oxford Univ. Press (Clarendon), London, 1965.
4. K. Fong and T. L. Jordan, Los Alamos Scientific Laboratory Report, LA-6774, June 1977.

Received: July 12, 1978; Revised: September 14, 1979

David S. Kershaw

*University of California*
*Lawrence Livermore Laboratory*
*Livermore, California 94550*