MATHEMATICS OF COMPUTATION, VOLUME 35, NUMBER 152
OCTOBER 1980, PAGES 1235–1249

Numerical Stability of Nested Dissection Orderings

By Indu Mati Anand

Abstract. Rigorous bounds on rounding errors for sparse positive definite matrices are obtained. When used for nested dissection orderings of finite element matrices, the analysis furnishes bounds which are stronger than those for band orderings.

1. Introduction. It is the purpose of this paper to obtain rigorous bounds on rounding errors associated with sparse positive definite matrices, particularly those associated with nested dissection orderings of finite element matrices, and to compare these bounds with those obtainable for band elimination. Nested dissection was introduced by George [6] for regular $n \times n$ grids and later generalized by Birkhoff and George [2], and by George [8] to any grid. These methods provide very efficient orderings for the Cholesky factorization of linear systems

$$Ax = b, \tag{1.1}$$

arising from a finite element discretization of two-dimensional boundary value problems. It has been shown that nested dissection is an optimal strategy, in the sense that we cannot reduce the arithmetic and storage requirements, by an order of magnitude, by any other ordering of such a system; see, Hoffman et al. [11]. This has led to the conjecture, see Birkhoff and George [2], that nested dissection ordering may be accompanied by a smaller accumulated round-off than band elimination. Our analysis shows that indeed to be the case.

Backward error analysis, as applied to Gaussian elimination, seeks to establish bounds for $\|E\|$, for some norm $\|\cdot\|$, where E is the perturbation matrix such that the computed triangular factors L and U of A satisfy

$$LU = A + E.$$

For a specific ordering of the rows and columns of A, L and U are obtained by the well-known recursive procedure:

Let $A^{(1)} = A$, and denote by $a_{ij}^{(k)}$ the elements of $A^{(k)}$. Define $A^{(k+1)}$ for $k = 1, 2, \ldots, n$, by

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - l_{ik}a_{kj}^{(k)}, \qquad i, j > k, \tag{1.2a}$$

Received May 11, 1979; revised December 10, 1979.
1980 *Mathematics Subject Classification.* Primary 65F05, 65N20.

0025-5718/80/0000-0165/$04.75

where

$$l_{ik} = a_{ik}^{(k)}/a_{kk}^{(k)}, \tag{1.2b}$$

$$a_{ik}^{(k+1)} = 0, \quad i > k. \tag{1.2c}$$

For all other values of i and j

$$a_{ij}^{(k+1)} = a_{ij}^{(k)}. \tag{1.2d}$$

After the completion of the kth major step, which results in the generation of the first k columns of L and the matrix $A^{(k+1)}$, we obtain the incomplete factorization:

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{12}^T & A_{22} \end{bmatrix} = \begin{bmatrix} L_{11} & 0 \\ A_{12}^T U_{11}^{-1} & I_{n-k} \end{bmatrix} \begin{bmatrix} U_{11} & L_{11}^{-1} A_{12} \\ 0 & A_{22}^{(k+1)} \end{bmatrix},$$

where the matrix on the left-hand side is a block form of A with A_{11} of order k, and the second matrix on the right that of $A^{(k+1)}$. Further,

$$A_{11} = L_{11} U_{11},$$

and the submatrix $A_{22}^{(k+1)}$, known as the Schur complement, is defined by

$$A_{22}^{(k+1)} = A_{22} - A_{12}^T U_{11}^{-1} L_{11}^{-1} A_{12} = A_{22} - A_{12}^T A_{11}^{-1} A_{12}.$$

If A is positive definite, then we can take advantage of symmetry by using the Cholesky algorithm, which produces the factorization $A = \tilde{L}\tilde{L}^T$ in about half as many operations as Gaussian elimination.

The accumulation of round-off errors is related to the growth of the elements of the reduced matrices, as the following lemma given by Widlund [14] shows.

LEMMA 1. *Let the matrix A be factored, according to the formula* (1.2) *on a t-digit, base β computer using floating-point operations. Then the computed triangular matrices $\bar{L}$ and $\bar{U}$ satisfy*

$$\bar{L}\bar{U} = A + E,$$

where the element e_{ij} of the error matrix E satisfies

$$|e_{ij}| \leqslant \epsilon \sum_{k=2}^{\min(i,j)} |a_{ij}^{(k)} - a_{ij}^{(k-1)}| + \epsilon(1+\epsilon) \sum_{k=2}^{\min(i,j)*} |a_{jk}^{(k)}|. \tag{1.3}$$

Here ϵ, $\beta^{1-t}/2 \leqslant \epsilon \leqslant \beta^{1-t}$, is a machine-dependent constant and Σ^* denotes the sum over the values of k, for which the (i, j)th element of the intermediary matrix $A^{(k-1)}$ undergoes a change at step $k-1$.

Lemma 1 is an improved version of a result proved by Reid [13], which itself extends a result of Wilkinson [15]. We note here for future reference that Lemma 1 holds also for Cholesky decomposition and is well suited to applications where A is sparse.

If the norm of E is acceptably small compared to that of A, then the decomposition of A is said to be stable. The method defined by (1.2) is known to be stable for a positive definite symmetric matrix.

Thus, $\|E\|$ provides a satisfactory basis for a comparison of the stability properties of the various methods of factorization.

Although nested dissection has been the subject of much analysis and many experiments to compare it with the other methods, we know of only one study of its stability properties. Birkhoff and George [2] compared the stability properties of nested dissection and band elimination, and presented some arguments which suggest that for a subclass $\mathcal{S}$ of weakly diagonally dominant matrices, called S-matrices, [$A \in \mathcal{S} \iff a_{i,j} \leqslant 0$ for $i \neq j$ and $\Sigma_j a_{i,j} \geqslant 0$ for all i], nested dissection is at least as stable as band-elimination.

In Section 2 we carry out backward error analysis for general sparse, positive definite, symmetric matrices. In Sections 3 and 4 we establish error bounds for row-by-row and nested dissection orderings of two-dimensional finite element problems and show the superiority of nested dissection.

We find that when the matrix in (1.1) is generated by the conventional row-by-row ordering of the unknowns in an $n \times n$ grid, then the l_2 norm of the perturbation matrix E satisfies

$$\|E\|_2 \leqslant cn^2 \in \|A\|_2, \quad \text{where } LL^T = A + E.$$

Here c is a constant. We also show that we cannot generally expect to reduce the exponent in the factor n^2 in this estimate.

The analysis of Section 4 shows that for nested dissection orderings we have

$$LL^T = A + E, \quad \text{where } \|E\|_2 \leqslant c_1 n^{8/5} \in \|A\|_2.$$

These results are proved for simple $n \times n$ grid problems in this paper. However, the proofs extend with only slight modifications to the nested dissection partitioning of graphs relating to the class of two-dimensional finite element problems described by George [8].

It has become increasingly evident that the usefulness of the dissection ideas is not limited to complete nested dissection. The one-way dissection given by George [5], and the incomplete nested dissection by George et al. [10], are useful variants of the method. The use of these methods for certain indefinite problems is examined in Anand [1], where dissection again appears to improve the numerical stability of computed solutions. We may also note here that the method of dissection is available for arbitrary grids. George and Liu [9] have developed an automatic algorithm for an economic generation of partitions based on heuristic considerations. They aim to find, at each step, a small separator set which disconnects a graph into two or more components of approximately equal size. More recently, Lipton, Rose and Tarjan [12] have designed an algorithm, which partitions any n-vertex planar graph G into three sets A, B, C, where C separates A and B, $|C| \leqslant 2\sqrt{2}\sqrt{n}$ and $|A|$ and $|B|$ do not exceed $2n/3$.

2. **Error Analysis of Sparse Matrices.** We now extend to sparse matrices the rigorous backward error analysis carried out by Wilkinson [16] for dense positive definite symmetric matrices. He proved the following theorem:

THEOREM (WILKINSON). *If A is a positive definite floating-point matrix of order $N \geqslant 10$, then, provided that*

$$\lambda_{\min} = \frac{1}{\|A^{-1}\|_2} \geqslant 20N^{3/2}2^{-t_1}\|A\|_2,$$

the Cholesky factor L can be computed without breakdown and the computed L satisfies

$$LL^T = A + E, \quad \text{where } \|E\|_2 \leqslant 2.5\,N^{3/2}2^{-t_1}\|A\|_2.$$

Here $t_1 = t - \log_2(1.06)$ for t-digit binary arithmetic.

He showed, furthermore, that since a_{NN} is involved in N independent operations, the factor $N^{3/2}$ cannot be improved upon in a significant way.

If A is a highly sparse matrix, then it is possible to improve the above bound, since a typical element in the matrix is only operated on a few times. Lemma 4 will show that the bound on the norm of the error depends mainly on σ, the maximum number of nonzero elements in a row of $L + L^T$.

We shall need the following result on Cholesky factorization in exact arithmetic, which we quote without proof; see Wilkinson [ibid.].

LEMMA 2. *Let $A^{(1)}$ be a positive definite matrix of order N and $L^{(1)}$ be its Cholesky factor so that $A^{(1)} = L^{(1)}(L^{(1)})^T$, and let*

$$A^{(1)} = \begin{bmatrix} a_{11}^{(1)} & a_1^T \\ a_1 & A^{(2)} \end{bmatrix}, \qquad L^{(1)} = \begin{bmatrix} l_{11} & 0 \\ l_1 & L^{(2)} \end{bmatrix},$$

where $B^{(2)} = L^{(2)}(L^{(2)})^T = A^{(2)} - l_1 l_1^T$.

Then

(i) *$B^{(2)}$ is positive definite,*

(ii) $\|B^{(2)}\|_2 \leqslant \|A^{(2)}\|_2 \leqslant \|A^{(1)}\|_2$, *and*

(iii) $\|l_1 l_1^T\|_2 = \|l_1\|_2^2 = \|a_1\|_2^2/a_{11}^{(1)} \leqslant \|A^{(2)}\|_2 \leqslant \|A^{(1)}\|_2$.

We assume now that for the actual computations floating-point arithmetic of relative accuracy β^{1-t_1} is used on a base β computer. When chopping arithmetic is used, $t_1 = t$, the number of significant digits in the mantissa, while t_1 is a nonintegral number greater than t, when rounding arithmetic is used. Thus, we assume that the floating-point basic operations satisfy

$$\text{fl}(x \oplus y) = (x \oplus y)(1 + \text{error}),$$

where $|\text{error}| < \beta^{1-t_1}$ and $\oplus$ represents addition, subtraction, multiplication or division. We further write $\epsilon = \alpha\beta^{1-t_1}$, where $\alpha > 0$ is a parameter close to unity. It is con-

venient to replace quantities such as $(1 \pm \beta^{1-t_1})^k$, which appear in extended computations, by $1 \pm k\beta^{1-t_1}$, and the parameter α is intended to allow for the second-order effects, which are ignored in such a replacement. Wilkinson [16], for example, chose the value $\alpha = 1.06$ under the assumption

$$k\beta^{1-t_1} < 0.1.$$

We shall also use quantities like ϵ_{ij}, η_{ij} to denote actual errors introduced in a computation. These quantities satisfy

$$|\epsilon_{ij}| < \epsilon, \tag{2.1a}$$

and

$$|\eta_{ij}| < 2\epsilon. \tag{2.1b}$$

We also assume, following Wilkinson [16], that the magnitude of relative error introduced in a square-root operation is at most 2ϵ.

We shall now consider the effect of round-off on Cholesky decomposition, and prove first a result on the growth of elements.

LEMMA 3. *If A is a symmetric, positive definite, sparse floating-point matrix of order $N \geqslant 10$, then, provided that*

$$\lambda_{\min} = \frac{1}{\|A^{-1}\|_2} \geqslant 20N\sigma^{1/2}\epsilon\|A\|_2, \tag{2.2}$$

the Cholesky factor L of A can be computed in single-precision arithmetic without breakdown. If $B^{(k)}$ denotes the computed Schur-complement of order $N - k + 1$, then

$$\|B^{(k)}\|_2 \leqslant [1 + \epsilon(\sigma^{1/2} + 2.2)]^{k-1}\|A\|_2, \tag{2.3}$$

where σ is the maximum number of nonzero elements in a row of $L + L^T$.

Proof. As we see below, the condition (2.2) will ensure that the matrices $B^{(k)}$ are positive definite. Since $\|A^{-1}\|_2\|A\|_2 \geqslant 1$ for any A, we must have

$$2N\sigma^{1/2}\epsilon \leqslant 0.1. \tag{2.4}$$

In order to evaluate the role of sparsity, we shall first obtain bounds ignoring sparsity in the manner of Wilkinson [16], and make adjustments for it subsequently.

Now consider the computation of the first column of $L^{(1)}$ and the Schur-complement $B^{(2)}$ for the matrix $A = A^{(1)}$. Let $F^{(1)}$ be the error matrix associated with the computation of the first column of $L^{(1)}$ and $E^{(1)}$ be that associated with the computation of $L^{(1)}$ and $B^{(2)}$. Using the notations of Lemma 2,

$$l_{11} = \text{fl}[\text{sqrt } a_{11}^{(1)}],$$

which gives

$$l_{11}^2 = a_{11}^{(1)}(1 + 2\epsilon_{11}^{(1)}),$$

and

$$l_{i1} = \mathrm{fl}\left[\frac{a_{i1}^{(1)}}{a_{11}^{(1)}}\right] = \frac{a_{i1}^{(1)}}{a_{11}^{(1)}}(1 + \epsilon_{i1}^{(1)}).$$

Thus, l_{11} and l_1 are exact for the matrix $A^{(1)} + F^{(1)}$, where $f_{11}^{(1)} = 2a_{11}^{(1)}\epsilon_{11}^{(1)}$, $f_{1i}^{(1)} = f_{i1}^{(1)} = a_{i1}^{(1)}\epsilon_{i1}^{(1)}$, $f_{ij}^{(1)} = 0$ otherwise. This gives $\|F^{(1)}\|_2 \leqslant \| |F^{(1)}| \|_2 \leqslant 2\epsilon \| |A| \|_2 \leqslant 2N^{1/2}\epsilon \|A^{(1)}\|_2$. The condition (2.2) then ensures that $A^{(1)} + F^{(1)}$ is positive definite, so that it follows from Lemma 2 that

$$\|l_1\|_2^2 \leqslant \|A^{(1)} + F^{(1)}\|_2 \leqslant (1 + 2N^{1/2}\epsilon)\|A^{(1)}\|_2, \tag{2.5}$$

where now l_1 denotes the computed vector l_1.

The elements of $B^{(2)}$ are given by

$$b_{ij}^{(2)} = \mathrm{fl}(a_{ij}^{(1)} - l_{i1}l_{1j}) = a_{ij}^{(1)}(1 + \epsilon_{ij}^{(1)}) - l_{i1}l_{1j}(1 + \eta_{ij}^{(1)})$$

$$= a_{ij}^{(1)}(1 + \epsilon_{ij}^{(1)}) - l_{i1}l_{1j}\eta_{ij}^{(1)} - l_{il}l_{1j}.$$

Thus, the computed $L^{(1)}$ and $B^{(2)}$ are exact for the matrix $A^{(1)} + E^{(1)}$, where the elements of $E^{(1)}$ are given by

$$e_{11}^{(1)} = 2a_{11}^{(1)}\epsilon_{11}^{(1)}, \qquad e_{1i}^{(1)} = e_{i1}^{(1)} = a_{i1}^{(1)}\epsilon_{i1},$$

$$e_{ij}^{(1)} = a_{ij}^{(1)}\epsilon_{ij}^{(1)} - l_{i1}l_{1j}\eta_{ij}^{(1)}, \qquad i, j \geqslant 2,$$

and satisfy

$$|e_{11}^{(1)}| \leqslant 2\epsilon|a_{11}^{(1)}|, \qquad |e_{i1}^{(1)}| = |e_{1i}^{(1)}| \leqslant \epsilon|a_{i1}^{(1)}|, \qquad i > 1,$$

$$|e_{ij}^{(1)}| \leqslant \epsilon|a_{ij}^{(1)}| + 2\epsilon|l_{i1}l_{1j}|, \qquad i, j \neq 1.$$

Hence, using (2.5),

$$\|E^{(1)}\|_2 \leqslant \| |E^{(1)}| \|_2 \leqslant 2\epsilon \| |A^{(1)}| \|_2 + 2\epsilon\|l_1\|_2^2$$

$$\leqslant 2\epsilon N^{1/2}\|A^{(1)}\|_2 + 2\epsilon(1 + 2N^{1/2}\epsilon)\|A^{(1)}\|_2$$

$$= 2\epsilon(N^{1/2} + 1 + 2N^{1/2}\epsilon)\|A^{(1)}\|_2,$$

so that using (2.4), we certainly get

$$\|E^{(1)}\|_2 \leqslant 2\epsilon(N^{1/2} + 1.1)\|A^{(1)}\|_2. \tag{2.6}$$

If $G^{(1)}$ is the matrix $E^{(1)}$ with its first row and column deleted, then

$$\|G^{(2)}\|_2 \leqslant \| |G^{(2)}| \|_2 \leqslant \epsilon \| |A^{(2)}| \|_2 + 2\epsilon\|l_1\|_2^2$$

$$\leqslant N^{1/2}\epsilon\|A^{(2)}\|_2 + 2\epsilon(1 + 2N^{1/2}\epsilon)\|A^{(1)}\|_2.$$

Now from Lemma 2,

$$\|B^{(2)}\|_2 \leqslant \|A^{(2)} + G^{(2)}\|_2,$$

so that by the above estimate for $\|G^{(2)}\|_2$,

$$\|B^{(2)}\|_2 \leqslant \|A^{(2)}\|_2 + \|G^{(2)}\|_2$$

$$\leqslant \|A^{(2)}\|_2 + N^{1/2}\epsilon\|A^{(2)}\|_2 + 2\epsilon(1 + 2N^{1/2}\epsilon)\|A^{(1)}\|_2$$

or, using (2.4) again,

$$\|B^{(2)}\|_2 \leqslant \{1 + \epsilon(N^{1/2} + 2.2)\}\|A^{(1)}\|_2. \tag{2.7}$$

Proceeding in the same way to obtain the kth column of $L^{(k)}$ and the Schur-complement $B^{(k+1)}$ from $B^{(k)}$, we shall find that they are exact for a matrix $B^{(k)} + E^{(k)}$, where

$$\|E^{(k)}\|_2 \leqslant 2\epsilon\{(N - k + 1)^{1/2} + 1.1\}\|B^{(k)}\|_2, \tag{2.8a}$$

and

$$\|B^{(k+1)}\|_2 \leqslant [1 + \epsilon\{(N - k + 1)^{1/2} + 2.2\}]\|B^{(k)}\|_2. \tag{2.8b}$$

The formula (2.8b) used recursively for $k - 1, k - 2, \ldots, 1$, will, in conjunction with (2.7), give

$$\|B^{(k)}\|_2 \leqslant [1 + \epsilon(N^{1/2} + 2.2)][1 + \epsilon((N - 1)^{1/2} + 2.2)] \cdots [1 + \epsilon\{(N - k + 2)^{1/2} + 2.2\}]\|A^{(1)}\|_2.$$

This certainly gives

$$\|B^{(k)}\|_2 \leqslant [1 + \epsilon(N^{1/2} + 2.2)]^{k-1}\|A^{(1)}\|_2. \tag{2.9}$$

In the case when it is known that L has no more than σ nonzero elements in a column

$$\|F^{(1)}\|_2 \leqslant \| \, |F^{(1)}| \, \|_2 \leqslant 2\epsilon\| \, |S^{(1)}| \, \|_2,$$

where, as before, $F^{(1)}$ is such that the first column of $L^{(1)}$ is exact for $A^{(1)} + F^{(1)}$, and $S^{(1)}$ is the submatrix of A, obtained by deleting from A all rows i and columns j such that $a_{i1}, a_{1j} = 0$. Since the order of $S^{(1)}$ is no more than σ, and $S^{(1)}$ is a principal submatrix of $A^{(1)}$,

$$\| \, |S^{(1)}| \, \|_2 \leqslant \sigma^{1/2}\|S^{(1)}\|_2 \leqslant \sigma^{1/2}\|A^{(1)}\|_2.$$

Therefore,

$$\|F^{(1)}\|_2 \leqslant 2\epsilon\sigma^{1/2}\|A^{(1)}\|_2,$$

and instead of (2.5), we shall have

$$\|l_1\|_2^2 \leqslant (1 + 2\sigma^{1/2}\epsilon)\|A^{(1)}\|_2.$$

Further, in the computation of $B^{(2)}$, we observe that if either a_{i1} or a_{1j} equals zero, then $l_{i1}l_{1j} = 0$, so that in that case $b_{ij}^{(2)} = a_{ij}^{(1)}$ and $e_{ij}^{(1)} = 0$. It follows that $E^{(1)}$ has no more than σ nonnull rows and columns, and

$$\|E^{(1)}\|_2 \leqslant \| \, |E^{(1)}| \, \|_2 \leqslant 2\epsilon\| \, |S^{(1)}| \, \|_2 + 2\epsilon\|l_1\|_2^2,$$

or

$$\|E^{(1)}\|_2 \leqslant 2\epsilon(\sigma^{1/2} + 1 + 2\sigma^{1/2}\epsilon)\|A^{(1)}\|_2 \leqslant 2\epsilon(\sigma^{1/2} + 1.1)\|A^{(1)}\|_2.$$

Similarly (2.7) can be replaced by

$$\|B^{(2)}\|_2 \leqslant \{1 + \epsilon(\sigma^{1/2} + 2.2)\}\|A^{(1)}\|_2,$$

and (2.8a) and (2.8b) can be replaced, respectively, by

$$\|E^{(k)}\|_2 \leqslant 2\epsilon\{\sigma^{1/2} + 1.1\}\|B^{(k)}\|_2, \tag{2.10a}$$

and

$$\|B^{(k+1)}\|_2 \leqslant [1 + \epsilon\{\sigma^{1/2} + 2.2\}]\|B^{(k)}\|_2. \tag{2.10b}$$

Finally, in place of (2.9), we shall obtain

$$\|B^{(k)}\|_2 \leqslant [1 + \epsilon(\sigma^{1/2} + 2.2)]^{k-1}\|A^{(1)}\|_2,$$

which completes the proof of the lemma.

Under the condition (2.2), (2.3) gives the following bound

$$\|B^{(N)}\|_2 \leqslant c\|A\|_2,$$

where c is a constant no more than 1.18.

The above analysis also furnishes a bound for the norm of the error matrix which is of the order of $N\sigma^{1/2}$. A more satisfactory bound is obtained in Lemma 4 by using Lemma 1.

When A is sparse, L will typically contain more nonzero elements than the lower triangular part of A. This phenomenon is known as the "fill" suffered by A. To understand how this fill takes place we return to the formulas (1.2). If at the first step of the algorithm a_{i1} and a_{1j} are nonzero while a_{ij} is zero, then $a_{ij}^{(2)}$ will be nonzero. In the following steps this can lead to a propagation of fill. We may note here that $a_{ij}^{(k)}$, $i, j > k$, might be zero, even if $a_{ij}^{(k-1)}$ differs from zero. However, exact cancellation rarely occurs and we may disregard such accidental creations of zeros. We shall always assume that if $a_{ij}^{(k-1)} \neq 0$, then so is a $a_{ij}^{(p)}$, $p \geqslant k$.

Furthermore, the (i, j)th element will not change at the kth step, unless $a_{kj}^{(k)}$ is different from zero. Therefore, if accidental creation of zeros is ruled out, then the number of times the (i, j)th element changes will be no more than the number of nonzeros in the sequence $a_{1j}^{(1)}, a_{2j}^{(2)}, a_{3j}^{(3)}, \ldots, a_{mj}^{(m)}$, where $m = \min(i - 1, j - 1)$.

LEMMA 4. *If A is a symmetric, positive definite, sparse floating-point matrix of order $N \geqslant 10$, and*

$$\lambda_{\min} \geqslant 20N\sigma^{1/2}\|A\|_2,$$

then the computed Cholesky factor L of A is exact for a matrix $A + E$, where

$$\|E\|_2 \leqslant 3\epsilon[1 + \epsilon(\sigma^{1/2} + 2.2)]^{N-1}\sigma^2\|A\|_2. \tag{2.11}$$

Proof. Since the number of times the (i, j)th element changes during the elimination process is no more than the number of nonzeros in the sequence $a_{1j}^{(1)}, a_{2j}^{(2)}$,

$a_{3j}^{(3)}, \ldots, a_{mj}^{(m)}$, where $m = \min(i-1, j-1)$, and σ is the maximum number of nonzeros in a column of $L + L^T$, no element undergoes more than $\sigma - 1$ changes. Let σ_{ij} denote the exact number of changes suffered by the (i, j)th element.

From Lemma 1, we find that the elements e_{ij} of $E = \Sigma E^{(k)}$ satisfy

$$|e_{ij}| \leqslant \epsilon \sum_{k=2}^{\min(i,j)} |a_{ij}^{(k)} - a_{ij}^{(k-1)}| + \epsilon(1+\epsilon) \sum_{k=1}^{\min(i,j)*} |a_{ij}^{(k)}|.$$

Clearly, if the (i, j)th element undergoes change σ_{ij} times, then from the above estimate, we get

$$|e_{ij}| \leqslant 3\epsilon\sigma_{ij}\rho, \quad \text{where } \rho = \max_{i,j,k} |a_{ij}^{(k)}|.$$

Consequently,

$$\|E\|_2 \leqslant 3\epsilon\rho \max_i \sum_j \sigma_{ij}, \tag{2.12}$$

and therefore,

$$\|E\|_2 \leqslant 3\epsilon\sigma^2\rho, \tag{2.13}$$

since the 2-norm of E is bounded by the sum of its elements in a row and there are no more than σ nonzero elements in any row of E, assuming that no zeros are created during elimination by cancellation.

But

$$\rho \leqslant \max_k \|B^{(k)}\|_2 \leqslant [1 + \epsilon(\sigma^{1/2} + 2.2)]^{N-1} \|A\|_2$$

from Lemma 3. Hence from (2.13), we get

$$\|E\|_2 \leqslant 3\epsilon[1 + \epsilon(\sigma^{1/2} + 2.2)]^{N-1}\sigma^2\|A\|_2,$$

which is the result of (2.11).

As an immediate corollary of Lemmas 3 and 4, we can prove the following

LEMMA 5. *If A is a symmetric, positive definite bandmatrix of order $N \geqslant 10$ and bandwidth $m = \max_{a_{ij} \neq 0} |i - j|$, and if*

$$\lambda_{\min} = \frac{1}{\|A^{-1}\|_2} \geqslant 20Nm^{1/2}\epsilon\|A\|_2,$$

then the computed Cholesky factor L of A is exact for a matrix $A + E$, where

$$\|E\|_2 \leqslant 3\epsilon[1 + \epsilon(m^{1/2} + 2.2)]^{N-1}m^2\|A\|_2. \tag{2.14}$$

Proof. It can easily be shown that if the Cholesky factorization of a bandmatrix is obtained, then, in the absence of pivoting, fill takes place only within the band. Therefore, there will be no more than $2m$ nonzero elements in a row of E. Also, the (i, j)th element will undergo changes between 1 and m times depending on its position within the band. The result (2.14) then follows from (2.12).

3. **Error Analysis of Finite Element Matrices: Row-by-Row Ordering.** We assume now that the system

$$Ax = b$$

is associated with a finite element mesh M_0, formed by subdividing the unit square $(0, 1) \times (0, 1)$ into n^2 small square elements of side length $1/n$. The mesh M_0 has a node at each of the $N = (n + 1)^2$ vertices, and these nodes are ordered in the standard row-by-row manner. We assume further that each x_i is associated with a node of M_0, and $a_{ij} \neq 0$ if and only if x_i and x_j are associated with nodes of the same element.

While we have chosen to exhibit our results for a simple mesh, the methods of proof, in this and the next section, are not limited to it but may be extended to other planar finite element problems. The other assumptions are also not as restrictive, as they might seem. If there are no nodes (unknowns) on the boundary, as in the Dirichlet problem, then the system will simply have an appropriate number of identities, which will not affect our bounds seriously. Similarly, if some $a_{ij} = 0$, even when x_i and x_j are associated with the same element, the bounds obtained will not be tight, although they will still provide a useful overall estimate of accumulated round-off errors.

George [6] has proved the following lemma and corollary, which we quote here without proof.

LEMMA 6. *Let M_0 be the regular finite element mesh described above; and let M_k, $k = 1, 2, \ldots, N = (n + 1)^2$, be the mesh sequence generated through the elimination process by an arbitrary ordering of M_0. Then at least one element having $n + 1$ or more unknowns associated with it, appears in the mesh sequence $\{M_k\}$.*

The next result is a corollary of Lemma 6.

LEMMA 7. *Every ordering of M_0 results in a bandwidth satisfying $m \geqslant n$.*

We can now obtain a bound on the error for the Cholesky factorization of a matrix which corresponds to a row-by-row ordering of M_0.

LEMMA 8. *If A is a matrix of minimum bandwidth, obtained by a row-by-row ordering of M_0, then the computed Cholesky factor L satisfies*

$$LL^T = A + E,$$

where

$$\|E\|_2 \leqslant 3\epsilon[1 + \epsilon(n^{1/2} + 2.2)]^N n^2 \|A\|_2. \tag{3.1}$$

Proof. The lemma follows from Lemmas 5 and 7, since we may safely assume $N \geqslant 10$.

In order to interpret the estimate (2.14), we observe that under the assumption (2.2) of Lemma 3 the factor $[1 + (n^{1/2} + 2.2)\epsilon]^N$ reduces to a constant < 1.18 so that the bound on the error is proportional to n^2. It can be shown that for band-matrices arising from finite element problems, fill propagates to all the elements within the band outside the first block. Therefore, we should expect most elements outside

the first block to change n times, and to have $2n$ nonzero elements in a row of $L + L^T$. Therefore, we cannot expect to improve the factor n^2 in (3.1) by an order of magnitude for row-by-row ordering by using a general a priori analysis for single-precision arithmetic ignoring statistical distribution of the errors.

4. Error Analysis of Finite Element Matrices: Nested Dissection Orderings. We now consider the stability of factorization corresponding to nested dissection orderings of mesh M_0 described in Section 3.

THEOREM. *If A is the matrix associated with a nested dissection ordering of the mesh M_0, then the computed Cholesky factor L satisfies*

$$LL^T = A + E,$$

where

$$\|E\|_2 \leqslant C_1 \epsilon n^{8/5}[1 + 4\epsilon\{n^{1/2} + 1.1\}]^N \|A\|_2, \tag{4.1}$$

and C_1 is a constant.

Proof. For the purposes of the proof, we first assume that $n = 2^l$, where l is a multiple of 5. It is also convenient to follow the description of the dissection strategy as given by George [6].

The nested dissection ordering of the mesh M_0 is defined as follows: Suppose x_{ij} is the unknown associated with the node (ih, jh).

Let

$$\pi(i) = p + 1 \quad \text{if } i = 2^p(2q + 1),$$
$$\pi(0) = 1,$$
$$\pi(n) = 1.$$

Now define sets of nodes P_k by

$$P_k = \{x_{ij}(\max(\pi(i), \pi(j)) = k)\}.$$

For $k > 1$, the sets P_k are unions of +-shaped sets. They are displayed in the paper by George [6]. The strategy is to number the unknowns corresponding to the nodes in P_1, followed by those in P_2 and so on, finally numbering those in P_l. As observed by George, each P_k consists of $n^2/2^{2k}$ independent +-shaped sets of nodes, which remain independent during the elimination. Furthermore, each independent set has no more than 2^{k+1} unknowns in it, and each unknown in P_k is connected to no more than $6 \cdot 2^k - 3$ unknowns, at the time of its elimination.

Let n_k = the number of nodes in P_k. Then by the above observation,

$$n_k \leqslant \frac{n^2}{2^{2k}} \cdot 2^{k+1} = 2^{2l-k+1}. \tag{4.2}$$

Let m_k = the maximum number of unknowns to which any unknown in P_k is connected. Then

$$m_k \leqslant 6 \cdot 2^k - 3 < 6 \cdot 2^k. \tag{4.3}$$

Let further σ_k = the maximum number of unknowns in P_k to which an unknown in P_i, $i \geqslant k$, may be connected during the elimination of P_k.

An unknown in P_i cannot be connected to more than four independent subsets of P_k, $i > k$. Therefore,

$$\sigma_k \leqslant 4 \cdot 2^{k+1} = 8 \cdot 2^k. \tag{4.4}$$

Now, turning to the elimination process, we know that the computed L satisfies

$$LL^T = A + E,$$

where

$$E = \sum_{i=1}^{N-1} E^{(i)}, \tag{4.5}$$

$E^{(i)}$ being the error matrix corresponding to the elimination of the ith unknown.

We may rewrite (4.5) as

$$E = \sum_{k=1}^{l} \left(\sum_{i=\nu_{k-1}+1}^{\nu_k} E^{(i)} \right), \tag{4.6}$$

where

$$\nu_k = \sum_{i=1}^{k} n_i.$$

The sum in the parentheses in (4.6) arises from the elimination of the n_k unknowns in P_k.

Let us rewrite (4.6) as

$$E = \sum_{k=1}^{\delta} \left(\sum_{i=\nu_{k-1}+1}^{\nu_k} E^{(i)} \right) + \sum_{k=\delta+1}^{l} \left(\sum_{i=\nu_{k-1}+1}^{\nu_k} E^{(i)} \right), \tag{4.7}$$

where we shall choose δ later.

The idea here is the following:

For the first δP_k's, n_k is large while m_k is small and we shall get a better bound using the method of Lemma 4. For the last $l - \delta$ P_k's, n_k is small while m_k is relatively larger and better bounds are obtained by using Wilkinson's approach in [16] for a dense matrix. We shall determine δ so as to balance the two contributions.

Consider the elimination of the unknowns in P_k, $k \leqslant \delta$. Since any unknown in P_i is connected to no more than σ_k unknowns in P_k, a row in $\Sigma_{i=\nu_{k-1}+1}^{\nu_k} E^{(i)}$ contains $\leqslant \sigma_k$ nonzeros in the columns $\nu_{k-1} + 1$ to ν_k, each changing no more than $\sigma_k/4$ times; the columns $\nu_k + 1$ to N contain no more than $m_{k+1} = 2m_k$ nonzeros, all except nine of which change at most $\sigma_k/2$ times. Adding all the changes in a row, we get as in Lemma 4, for $k \leqslant \delta$,

$$\left\| \sum_{i=\nu_{k-1}+1}^{\nu_k} E^{(i)} \right\|_2 \leqslant 4\epsilon m_k \sigma_k \rho_k, \tag{4.8}$$

where $\rho_k = \max_{p \leqslant \nu_k, i, j} |a_{ij}^{(p)}|$.

Clearly from a recursive use of (2.10b),

$$\rho_k < \rho_\delta \leqslant [1 + 2\epsilon(\sigma_\delta^{1/2} + 1.1)]^{\nu_\delta}\|A\|_2.$$

Therefore, (4.8) gives

$$\left\| \sum_{i=\nu_{k-1}+1}^{\nu_k} E^{(i)} \right\|_2 \leqslant 4\epsilon[1 + 2\epsilon(\sigma^{1/2} + 2.2)]^{\nu_\delta} m_k \sigma_k \|A\|_2. \tag{4.9}$$

It follows that

$$\begin{aligned}
\left\| \sum_{k=1}^{\delta} \left(\sum_{i=\nu_{k-1}+1}^{\nu_k} E^{(i)} \right) \right\|_2 &\leqslant \sum_{k=1}^{\delta} \left\| \sum_{i=\nu_{k-1}+1}^{\nu_k} E^{(i)} \right\|_2 \\
&\leqslant 4\epsilon\{1 + 2\epsilon(\sigma_\delta^{1/2} + 1.1)\}^{\nu_\delta} \left(\sum_{k=1}^{\delta} m_k \sigma_k \right) \|A\|_2 \\
&\leqslant 4\epsilon\{1 + 2\epsilon(\sigma_\delta^{1/2} + 1.1)\}^{\nu_\delta} \|A\|_2 \cdot 48 \sum_{k=1}^{\delta} 2^{2k} \\
&\qquad \text{and from (4.3) and (4.4)} \\
&\leqslant 4\epsilon \cdot 64\{1 + 2\epsilon(\sigma_\delta^{1/2} + 1.1)\}^{\nu_\delta} \|A\|_2 \cdot 2^{2\delta}.
\end{aligned} \tag{4.10}$$

For values of $i > \nu_\delta$, we can use the estimates obtained from (2.10a) and (2.10b) of Lemma 3, so that for $\nu_{k-1} + 1 \leqslant i \leqslant \nu_k$ we get

$$\|E^{(i)}\|_2 \leqslant 2\epsilon\{(m_k^{1/2} + 1.1)\}\|B^{(i)}\|_2, \tag{4.11a}$$

and

$$\|B^{(i+1)}\|_2 \leqslant [1 + \epsilon\{m_k^{1/2} + 2.2\}]\|B^{(i)}\|_2. \tag{4.11b}$$

Therefore,

$$\begin{aligned}
\left\| \sum_{k=\delta+1}^{l} \sum_{i=\nu_{k-1}+1}^{\nu_k} E^{(i)} \right\|_2 &\leqslant \sum_{k=\delta+1}^{l} \left[\sum_{i=\nu_{k-1}+1}^{\nu_k} 2\epsilon(m_k^{1/2} + 1.1) \right] \\
&\quad \cdot [1 + 2\epsilon(m_l^{1/2} + 1.1)]^{N-\nu_\delta - 1} \\
&\quad \cdot [1 + 2\epsilon(m_\delta^{1/2} + 1.1)]^{\nu_\delta} \|A\|_2,
\end{aligned}$$

from (4.11) and from the estimate of $\|B^{(\nu_\delta)}\|_2$ obtained from a recursive use of (2.10b). Now, since $\nu_k - \nu_{k-1} = n_k$ and $m_\delta < m_l$, it follows that

$$\begin{aligned}
\left\| \sum_{k=\delta+1}^{l} \sum_{i=\nu_{k-1}+1}^{\nu_k} E^{(i)} \right\|_2 &\leqslant \sum_{k=\delta+1}^{l} 2\epsilon[n_k m_k^{1/2} + 1.1 n_k] \\
&\quad \cdot [1 + 2\epsilon(m_l^{1/2} + 1.1)]^{N-1} \|A\|_2.
\end{aligned}$$

Now

$$\sum_{k=\delta+1}^{l} n_k \leqslant 2^{2l-\delta+1},$$

and

$$\sum_{k=\delta+1}^{l} m_k^{1/2} n_k \leqslant \frac{2 \cdot 6^{1/2}}{2^{1/2}-1} \cdot 2^{2l-\delta/2}.$$

Moreover, $n_l = 2n$.

Consequently,

$$\left\| \sum_{k=\delta+1}^{l} \sum_{i=\nu_{k-1}+1}^{\nu_k} E^{(i)} \right\|_2 \leqslant 2\epsilon \left[\frac{2 \cdot 6^{1/2}}{2^{1/2}-1} 2^{2l-\delta/2} + 1.1 \cdot 2^{2l-\delta+1} \right] \cdot [1 + 4\epsilon(n^{1/2} + 1.1)]^{N-1} \|A\|_2. \tag{4.12}$$

From (4.7), (4.10) and (4.12), we get

$$\|E\|_2 \leqslant \left[4\epsilon \cdot 64 \cdot 2^{2\delta} + 2\epsilon \left\{ \frac{2 \cdot 6^{1/2}}{2^{1/2}-1} \cdot 2^{2l-\delta/2} + 1.1 \cdot 2^{2l-\delta+1} \right\} \right] \cdot [1 + 4\epsilon(n^{1/2} + 1.1)]^{N-1} \|A\|_2. \tag{4.13}$$

We now find δ such that $2^{2\delta} = 2^{2l-\delta/2}$, which will be the case if $\delta = 4l/5$.

In case l is a multiple of 5, we get an integer value of δ, and recalling that $n = 2^l$, we get from (4.13) finally

$$\|E\|_2 \leqslant C_1 \epsilon n^{8/5} [1 + 4\epsilon(n^{1/2} + 1.1)]^N \|A\|_2,$$

where C_1 is a constant.

For a general value of n, we can always determine a value l which is a multiple of 5 such that $2^{l-5} < n \leqslant 2^l$. Since the lower and upper bounds are of the same order of magnitude and a larger problem clearly provides an upper bound for a smaller, the result follows for general values of n.

To put the exponent 8/5 in (4.1) in perspective, we observe that any ordering of the mesh M_0 would lead to a dense submatrix of order, at least, n during the elimination process and that Wilkinson [16] has shown that the appropriate factor for a dense matrix of order n is $n^{3/2}$. The exponent 8/5 is worse than 3/2 by only 1/10. Thus, it appears that for very large systems nested dissection orderings have very nearly the ideal stability properties.

Acknowledgement. The paper is based on part of the work done for the author's Ph.D. thesis at New York University. The author is greatly indebted to her advisor Professor Olof B. Widlund for many valuable discussions and critical comments.

15 Green Way
Chelmsford, Massachusetts 01824

1. I. M. ANAND, *Numerical Stability of Nested Dissection Orderings*, Ph.D. Thesis, New York University, New York, 1979.

2. G. BIRKHOFF & A. GEORGE, "Elimination by nested dissection," in *Complexity of Sequential and Parallel Algorithms* (J. Traub, Ed.), Academic Press, New York, 1973, pp. 221–269.

3. I. S. DUFF, A. M. ERISMAN & J. K. REID, "On George's nested dissection algorithm," *SIAM J. Numer. Anal.*, v. 13, 1976, pp. 686–695.

4. G. E. FORSYTHE & C. B. MOLER, *Computer Solution of Linear Algebraic Systems*, Prentice-Hall, Englewood Cliffs, N. J., 1967.

5. A. GEORGE, *An Efficient Band Oriented Scheme for Solving $n \times n$ Grid Problems*, Proc. Fall Joint Computer Conference, 1972.

6. A. GEORGE, "Nested dissection of a regular finite element mesh," *SIAM J. Numer. Anal.*, v. 10, 1973, pp. 345–363.

7. A. GEORGE, "Numerical experiments using dissection methods to solve $n \times n$ grid problems," *SIAM J. Numer. Anal.*, v. 14, 1977, pp. 161–179.

8. A. GEORGE, "Solution of linear systems of equations: Direct methods for finite element problems," *Sparse Matrix Techniques* (V. A. Barker, Ed.), Lecture Notes in Math., vol. 572, Springer-Verlag, Berlin and New York, 1977, pp. 52–101.

9. A. GEORGE & J. W. H. LIU, "An automatic nested dissection algorithm for irregular finite element problems," *SIAM J. Numer. Anal.*, v. 15, 1978, pp. 1053–1069.

10. A. GEORGE, W. G. POOLE, JR. & R. G. VOIGT, "Incomplete nested dissection for solving $n \times n$ grid problems," *SIAM J. Numer. Anal.*, v. 15, 1978, pp. 662–673.

11. A. HOFFMAN, M. S. MARTIN & D. J. ROSE, "Complexity bounds for regular finite difference and finite-element grids," *SIAM J. Numer. Anal.*, v. 10, 1973, pp. 364–369.

12. R. J. LIPTON, D. J. ROSE & R. E. TARJAN, "Generalized nested dissection," *SIAM J. Numer. Anal.*, v. 16, 1979, pp. 346–358.

13. J. K. REID, "A note on the stability of Gaussian elimination," *J. Inst. Math. Appl.*, v. 8, 1971, pp. 374–375.

14. O. WIDLUND, "On the use of sparsity of finite element systems of equations by Gaussian elimination-type methods," *Actas del Seminario Sobre Metodos Numericos Modernas*, Vol. 2, Universidad Central de Venezuela, 1974.

15. J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.

16. J. H. WILKINSON, "A priori error analysis of algebraic processes," *Proc. Internat. Congr. Math.* (*Moscow*, 1966), Izdat. "Mir", Moscow, 1968, pp. 629–640.