ANALYSIS OF SPARSE ELIMINATION*

James R. Bunch

TR 73 - 183

August 1973

Supersedes TR73-158

Department of Computer Science
Cornell University
Ithaca, New York 14850

*This research was supported by ONR grant N00014-67-A-0077-0021.

ANALYSIS OF SPARSE ELIMINATION */

James R. Bunch
Department of Computer Science
Cornell University

Abstract

An error analysis is presented for Gaussian elimination when the matrix is arbitrarily sparse. Error analyses for elimination on band matrices and full matrices follow as special cases.

*/ This research was supported by ONR grant N00014-67-A-0077-0021

1. Introduction

Since the direct solution of systems of linear equations by elimination is now well understood when the matrix is full or band, attention has turned to the problem of elimination of systems when the matrix is sparse. When the sparse matrix arises from the discretization of an elliptic partial differential equation, it has a special structure and special properties. But all too frequently the matrix has a sparsity which makes useful structural properties hard to identify.

Here we present an error analysis of (point) Gaussian elimination when the matrix is arbitrarily sparse. As a model for our analysis, we use the excellent analysis of the full matrix case given in Forsythe and Moler [3]. Our only assumption is that the LU decomposition of the matrix exists, i.e., that the $(1,1)$ entry in each reduced matrix is nonzero ([3], pp. 27-36).

In Section 2 we relate matrices and graphs and relate Gaussian elimination and an elimination process on graphs. We express operation counts and storage in terms of the degree sequences of graphs and apply these to some examples, in particular, to full and band matrices.

In Section 3 we give an error analysis for the inner product of two sparse vectors. This analysis is the fundamental step toward the analysis of Gaussian elimination for sparse systems.

In Section 4 we discuss the numerical solution of linear equations. In Section 5 a detailed error analysis is given for solving sparse triangular systems of linear equations. In Section 5 a detailed error analysis of triangular factorization (LU decomposition) is given. In Section 7 these results are combined to give an error analysis of Gaussian elimination.

In Section 8 we apply the analysis to the important special cases of full and band matrices. For band matrices we note that the error is dependent on the bandwidth but not on the order of the matrix.

2. Matrices and Graphs

Let M be an $n \times n$ matrix with non-zero leading principal minors; hence M is non-singular and the LU decomposition of M exists and is unique. If M is also symmetric, then $M = LDL^t$.

Rose [4] associated an undirected graph with a symmetric matrix and interpreted the LDL^t factorization graph theoretically by undirected elimination graphs. Rose and Bunch [5] made the extension of the association of directed graphs with general square matrices and of the interpretation of the LU factorization by directed elimination graphs.

The directed graph of M , $G(M) = (X, \mathcal{A})$, with vertex set X and arc set $\mathcal{A}$, is defined as follows: a vertex $x_i \in X$ is associated with row i of M , and $(x_i, x_j) \in \mathcal{A}$ (an arc from x_i to x_j is in G) if and only if $m_{ij} \neq 0$ and $i \neq j$. The vertices X are regarded as ordered; i.e., $X = \{x_i\}_{i=1}^n$.

Let the matrix M be written as $M = \begin{bmatrix} a & r^t \\ c & B \end{bmatrix}$, where a is 1×1 , r and c are $(n-1) \times 1$ and B is $(n-1) \times (n-1)$. Then the first step of the LU factorization of M can be written as

$$M = \begin{bmatrix} 1 & 0 \\ c/a & I \end{bmatrix} \begin{bmatrix} a & r^t \\ 0 & B - cr^t/a \end{bmatrix}.$$

If $G(M)$ is the directed graph of M , the elimination graph G_y is obtained from G by deleting y together with its incident arcs and adding an arc (x, z) whenever there exists a (directed) x, z path of length 2 containing y . G_y is the graph of the matrix obtained by "eliminating" the variable corresponding to y in Gaussian elimination; e.g., G_{x_1} is the graph of $B - cr^t/a$. The accidental creation of zeros

during the elimination process is ignored. For a more detailed discussion, see Bunch and Rose [2], Section 2.

Let $G_1, \dots, G_{n-1}$ be the sequence of elimination graphs defined recursively by $G_0 = G(M)$ and $G_i = (G_{i-1})_{x_i}$. Let $|S|$ be the number of elements in the set S . Let $r_i = |\{y \in X_{i-1} : (x_i, y) \in A_{i-1}, G_{i-1} = (X_{i-1}, A_{i-1})\}|$ and $c_i = |\{y \in X_{i-1} : (y, x_i) \in A_{i-1}, G_{i-1} = (X_{i-1}, A_{i-1})\}|$ be the out-degree and in-degree, respectively, of vertex x_i in the elimination graph G_{i-1} .

Note that r_i+1, c_i+1 is the number of non-zero elements in the first row, column of the reduced matrix of order $n-i+1$, i.e., the reduced matrix whose graph is G_{i-1} .

Define $e_{ik} = 1$ if there is an arc from x_i to x_k and $e_{ki} = 1$ if there is an arc from x_k to x_i in G_{i-1} . Otherwise, let $e_{ik} = 0$ and $e_{ki} = 0$ for $k \neq i$; and let $e_{ii} = 1$ for all i . We count a division as a multiplication and a subtraction as an addition. For convenience, we shall count an addition whenever fill-in occurs, cf. [2], Section 2.

When M is symmetric, then $r_i = c_i \equiv d_i$, and Rose [4] has shown:

Theorem 1. The factorization $M = LDL^t$ requires

$$\sum_{i=1}^{n-1} d_i(d_i+3)/2 \quad \text{multiplications and}$$

$$\sum_{i=1}^{n-1} d_i(d_i+1)/2 \quad \text{additions,}$$

while the backsolving of $LDL^t x = b$ requires

$$n+2 \sum_{i=1}^{n-1} d_i \quad \text{multiplications and}$$

$$2 \sum_{i=1}^{n-1} d_i \quad \text{additions;}$$

the storage required is

$$2n + \sum_{i=1}^{n-1} d_i.$$

When M is general, Bunch and Rose [2] have shown

Theorem 2. The $M = LU$ factorization requires

$$\sum_{i=1}^{n-1} (r_i + 1)c_i \quad \text{multiplications and}$$

$$\sum_{i=1}^{n-1} r_i c_i \quad \text{additions,}$$

while the backsolving $LUx = b$ requires

$$n + \sum_{i=1}^{n-1} (r_i + c_i) \quad \text{multiplications and}$$

$$\sum_{i=1}^{n-1} (r_i + c_i) \quad \text{additions;}$$

the storage required is

$$2n + \sum_{i=1}^{n-1} (r_i + c_i).$$

Examples. Let M be an $n \times n$ matrix.

(1) If M is symmetric and full, then $d_i = n-i$ and solving $Mx = LDL^t x = b$ requires

$$n + \sum_{i=1}^{n-1} d_i(d_i+7)/2 = \frac{1}{6} n^3 + \frac{3}{2} n^2 - \frac{2}{3} n \quad \text{multiplications,}$$

$$\sum_{i=1}^{n-1} d_i(d_i+5)/2 = \frac{1}{6} n^3 + n^2 - \frac{7}{6} n \quad \text{additions, and}$$

$$2n + \sum_{i=1}^{n-1} d_i = \frac{1}{2} n^2 + \frac{3}{2} n \quad \text{storage.}$$

(2) If M is symmetric and band, with bandwidth $2m+1$, then

$$d_i = \begin{cases} m & \text{for } 1 \leq i \leq n-m \\ n-i & \text{for } n-m+1 \leq i \leq n-1 \end{cases}, \text{ since no fill-in occurs outside}$$

the band (see Figure 1). Solving $Mx = LDL^t x = b$ requires

$$\left(\frac{1}{2} m^2 + \frac{7}{2} m + 1\right)n - \frac{1}{3} m^3 - 2m^2 - \frac{5}{3} m \quad \text{multiplications,}$$

$$\left(\frac{1}{2} m^2 + \frac{5}{2} m\right)n - \frac{1}{3} m^3 - \frac{3}{2} m^2 - \frac{7}{6} m \quad \text{additions, and}$$

$$(m+2)n - \frac{1}{2} m(m+1) \quad \text{storage.}$$

Figure 1. $n \times n$ matrix with bandwidth $2m+1$.

(3) If M is general and full, then $r_i = n-i = c_i$ and solving $Mx = LUx = b$ requires

$$n + \sum_{i=1}^{n-1} (r_i c_i + 2c_i + r_i) = \frac{1}{3} n^3 + n^2 - \frac{1}{3} n \quad \text{multiplications,}$$

$$\sum_{i=1}^{n-1} (r_i c_i + c_i + r_i) = \frac{1}{3} n^3 + \frac{1}{2} n^2 - \frac{5}{6} n \quad \text{additions, and}$$

$$n^2 + n \quad \text{storage.}$$

(4) If M is general and band, with bandwidth $2m+1$, then

$$r_i = c_i = \begin{cases} m & \text{for } 1 \leq i \leq n-m \\ n-i & \text{for } n-m+1 \leq i \leq n-1 \end{cases}$$

when no pivoting is needed, (see Figure 1). Then solving $Mx = LUx = b$ requires

$$(m^2 + 3m + 1)n - \frac{2}{3} m^3 - 2m^2 - \frac{4}{3} m \quad \text{multiplications,}$$

$$(m^2 + 2m)n - \frac{2}{3} m^3 - \frac{3}{2} m^2 - \frac{5}{6} m \quad \text{additions, and}$$

$$2(m+1)n - m^2 - m \quad \text{storage.}$$

When partial pivoting is used, then U may be filled an extra width m (see Figure 2). Then

$$c_i = \begin{cases} m & \text{for } 1 \leq i \leq n-m \\ n-i & \text{for } n-m+1 \leq i \leq n-1 \end{cases} \quad \text{and}$$

$$r_i \leq \begin{cases} 2m & \text{for } 1 \leq i \leq n-2m \\ n-i & \text{for } n-2m+1 \leq i \leq n-1 \end{cases}.$$

Then solving $PMx = LUx = Pb$ requires

$$\leq (2m^2 + 4m + 1)n - \frac{13}{6} m^3 - 4m^2 - \frac{11}{6} m \quad \text{multiplications,}$$

$$\leq (2m^2 + 3m)n - \frac{13}{6} m^3 - \frac{7}{2} m^2 - \frac{4}{3} m \quad \text{additions, and}$$

$$\leq (3m+2)n - \frac{3}{2} m^2 - \frac{3}{2} m \quad \text{storage;}$$

(and an extra n -vector is needed to record the interchanges).

Figure 2. Structure of $L+U$ when partial pivoting is used on $n \times n$ matrix with bandwidth $2m+1$.

(5) Let M be the 14×14 matrix having the nonzero structure given by the X's in Figure 3. Suppose $M = LU$ exists; ignoring accidental creation of zeros, $L+U$ has the structure in Figure 3, with the O's indicating the fill-in which occurred during the elimination. Then

$$r_i = \begin{cases} 8 & \text{for } i = 1, 2, 4 \\ 7 & \text{for } i = 3, 5 \\ 6 & \text{for } i = 6 \\ 5 & \text{for } i = 7 \\ 4 & \text{for } i = 8 \\ 3 & \text{for } i = 9, 10, 11 \\ 2 & \text{for } i = 12 \\ 1 & \text{for } i = 13 \end{cases}$$

and

$$c_i = \begin{cases} 5 & \text{for } i = 2, 5 \\ 4 & \text{for } i = 6 \\ 3 & \text{for } i = 1, 3, 7, 8, 9, 10, 11 \\ 2 & \text{for } i = 12 \\ 1 & \text{for } i = 4, 13 \end{cases}$$

Then solving $Mx = b$ requires

$$n + \sum_{i=1}^{n-1} (r_i c_i + r_i + 2c_i) = 368 \quad \text{multiplications,}$$

$$\sum_{i=1}^{n-1} (r_i c_i + r_i + c_i) = 315 \quad \text{additions, and}$$

$$2n + \sum_{i=1}^{n-1} (r_i + c_i) = 104 \quad \text{storage.}$$

3. Inner Product of Sparse Vectors

Let x, y be sparse n -vectors, and
$$e_i = \begin{cases} 1 & \text{if } x_i \neq 0 \text{ and } y_i \neq 0 \\ 0 & \text{if } x_i = 0 \text{ or } y_i = 0 \end{cases}$$

Following Forsythe and Moler ([3], p. 91), let u be the unit round-off

error, i.e.,
$$u = \begin{cases} \frac{1}{2} \beta^{1-t} & \text{(rounded operations)} \\ \beta^{1-t} & \text{(chopped operations)} \end{cases}, \text{ when using } t \text{ digit,}$$

base β arithmetic.

Lemma ([3], p. 92). If $|\delta_i| \leq u$ for $1 \leq i \leq n$ and if $nu \leq 0.01$, then

$$1 - nu \leq \prod_{i=1}^n (1 + \delta_i) \leq 1 + 1.01 nu.$$

Note that we can also write $\prod_{i=1}^n (1 + \delta_i) = 1 + 1.01 n \theta u$, where $|\theta| \leq 1$.

Now, we can prove a theorem on the error for the floating point inner product of two sparse vectors. This generalizes Theorem (20.18) in [3], p. 93, on the error for the floating point inner product of two full vectors.

Theorem 3. If $nu \leq 0.01$, then
$$fl\left(\sum_{i=1}^n x_i \times y_i\right) = \sum_{i=1}^n x_i y_i [1 + 1.01 u \theta_i (1 + \beta_i + \sum_{k=i+1}^n e_k)] e_i,$$
 where $|\theta_i| \leq 1$ and
$$\beta_i = \min(1, \sum_{j=1}^{i-1} e_j) \text{ for } 1 \leq i \leq n.$$

Proof. For convenience, let us consider $n = 4$. Let $|\delta_i| \leq u$ and $|\theta_i| \leq 1$. If $x_1 \neq 0$ and $y_1 \neq 0$, then when x_1 and y_1 are multiplied, an error is made and $x_1 y_1 (1 + \delta_1)$ is obtained. If $x_1 = 0$ or $y_1 = 0$, then the result of multiplying x_1 and y_1 is 0 and no error is made.

Thus we may write the result as $x_1 y_1 (1 + \delta_1) e_1$. Similarly, multiplying x_2 and y_2 yields $x_2 y_2 (1 + \delta_2) e_2$. If $e_2 \neq 0$, we add $x_2 y_2 (1 + \delta_2) e_2$ to $x_1 y_1 (1 + \delta_1) e_1$; an error occurs in the addition if and only if $e_1 \neq 0$, i.e., if and only if an addition of non-zero numbers actually occurs.

The result of the addition is thus

$$(3.1) \quad [x_1 y_1 (1 + \delta_1 e_1) + x_2 y_2 (1 + \delta_2 e_2)] (1 + \delta_3 e_1 e_2) \quad .$$

Multiplying x_3 and y_3 yields

$$(3.2) \quad x_3 y_3 (1 + \delta_4) e_3 \quad .$$

If (3.2) is nonzero, we add (3.2) to (3.1), and an error will occur if (3.1) is nonzero, i.e., if $e_1 \neq 0$ and $e_2 \neq 0$, i.e., if

$$\beta_3 \equiv \min(1, e_1 + e_2) \neq 0 \quad .$$

Let $\beta_i \equiv \min(1, \sum_{j=1}^{i-1} e_j)$. Then $\beta_1 = 0$ and $\beta_2 = \min(1, e_1) = e_1$.

Thus adding (3.2) to (3.1) yields

$$(3.3) \quad \{[x_1 y_1 (1 + \delta_1) e_1 + x_2 y_2 (1 + \delta_2) e_2] (1 + \delta_3 \beta_2 e_2) + x_3 y_3 (1 + \delta_4 e_3)\} (1 + \delta_5 \beta_3 e_3) \quad .$$

Similarly, multiplying x_4 and y_4 yields

$$(3.4) \quad x_4 y_4 (1 + \delta_6) e_4 \quad .$$

If $e_4 \neq 0$, we add (3.4) to (3.3), and make an error in the addition

if and only if (3.3) is nonzero, i.e., if and only if $e_1 + e_2 + e_3 \neq 0$

if and only if $\beta_4 = \min(1, e_1 + e_2 + e_3) \neq 0$. Thus, the final result is

$$\begin{aligned}
(3.5) \quad & fl \left(\sum_{i=1}^4 x_i \times y_i \right) \\
&= \left([x_1 y_1 (1 + \delta_1) e_1 + x_2 y_2 (1 + \delta_2) e_2] (1 + \delta_3 \beta_2 e_2) + x_3 y_3 (1 + \delta_4) e_3 \right) (1 + \delta_5 \beta_3 e_3) \\
&\quad + x_4 y_4 (1 + \delta_6) e_4 \Big) (1 + \delta_7 \beta_4 e_4) \\
&= x_1 y_1 (1 + \delta_1) (1 + \delta_3 \beta_2 e_2) (1 + \delta_5 \beta_3 e_3) (1 + \delta_7 \beta_4 e_4) e_1 + \\
&\quad x_2 y_2 (1 + \delta_2) (1 + \delta_3 \beta_2 e_2) (1 + \delta_5 \beta_3 e_3) (1 + \delta_7 \beta_4 e_4) e_2 + \\
&\quad x_3 y_3 (1 + \delta_4) (1 + \delta_5 \beta_3 e_3) (1 + \delta_7 \beta_4 e_4) e_3 + \\
&\quad x_4 y_4 (1 + \delta_6) (1 + \delta_7 \beta_4 e_4) e_4 \\
&= x_1 y_1 [1 + 1.01 u \theta_1 (1 + \beta_2 e_2 + \beta_3 e_3 + \beta_4 e_4)] e_1 + \\
&\quad x_2 y_2 [1 + 1.01 u \theta_2 (1 + \beta_2 e_2 + \beta_3 e_3 + \beta_4 e_4)] e_2 + \\
&\quad x_3 y_3 [1 + 1.01 u \theta_3 (1 + \beta_3 e_3 + \beta_4 e_4)] e_3 + \\
&\quad x_4 y_4 [1 + 1.01 u \theta_4 (1 + \beta_4 e_4)] e_4 \\
&= x_1 y_1 [1 + 1.01 u \theta_1 (1 + e_2 + e_3 + e_4)] e_1 + \\
&\quad x_2 y_2 [1 + 1.01 u \theta_2 (1 + \beta_2 + e_3 + e_4)] e_2 + \\
&\quad x_3 y_3 [1 + 1.01 u \theta_3 (1 + \beta_3 + e_4)] e_3 + \\
&\quad x_4 y_4 [1 + 1.01 u \theta_4 (1 + \beta_4)] e_4 \\
&= \sum_{i=1}^4 x_i y_i [1 + 1.01 u \theta_i (1 + \beta_i + \sum_{j=i+1}^4 e_j)] e_i,
\end{aligned}$$

since $e_i^2 = e_i$, and for $j > i$, $\beta_j e_j e_i = [\min(1, e_1 + e_2 + \dots + e_i + \dots + e_{j-i})] e_j e_i = e_j e_i$.

A proof for arbitrary n follows similarly. q.e.d.

Example.

Let x^t be the j -th row of an $n \times n$ tridiagonal matrix, i.e.,
 $x_i = 0$ for $i < j-1$ and $i > j+1$; assume $x_{j-1}, x_j, x_{j+1} \neq 0$; let
 y be full. Then

$$\begin{aligned}
 fl(x^t y) &= \sum_{i=j-1}^{j+1} x_i \times y_i = \{[x_{j-1}y_{j-1}(1+\delta_1) + x_j y_j(1+\delta_2)](1+\delta_3) \\
 &\quad + x_{j+1}y_{j+1}(1+\delta_4)\}(1+\delta_5) \\
 &= x_{j-1}y_{j-1}(1+\delta_1)(1+\delta_3)(1+\delta_5) + x_j y_j(1+\delta_2)(1+\delta_3)(1+\delta_5) \\
 &\quad + x_{j+1}y_{j+1}(1+\delta_4)(1+\delta_5) \\
 &= x_{j-1}y_{j-1}(1+1.01 \theta_{j-1} 3u) + x_j y_j(1+1.01 \theta_j 3u) \\
 &\quad + x_{j+1}y_{j+1}(1+1.01 \theta_{j+1} 2u) \\
 &= x_{j-1}y_{j-1}[1+1.01 \theta_{j-1} u(1+e_j+e_{j+1})] + \\
 &\quad x_j y_j[1+1.01 \theta_j (1+\beta_j+e_{j+1})] + \\
 &\quad x_{j+1}y_{j+1}[1+1.01 \theta_{j+1} (1+\beta_{j+1})] \\
 &= \sum_{i=1}^n x_i y_i [1+1.01 \theta_i (1+\beta_i + \sum_{k=i+1}^n e_k)] .
 \end{aligned}$$

Let us now consider the error occurring when accumulated inner products are performed, i.e., x and y are t digit floating point numbers but the intermediate steps in the calculation are carried to $2t$ digits, with the final result, $fl_2(x^t y)$, being rounded to t digits.

Then, as in [3], p. 94, we have

Theorem 4. If $nu \leq 0.01$, then

$$fl_2\left(\sum_{i=1}^n x_i \times y_i\right) =$$

$$(1 + \theta_0 \frac{1}{2} \beta^{1-t}) \sum_{i=1}^n x_i y_i [1 + 1.01 u \theta_i (\beta_i + \sum_{k=i+1}^n e_k)^F \frac{1}{2} \beta^{1-2t}] e_k ,$$

where $F = 1 + 1/\beta$, $|\theta_i| \leq 1$ for $0 \leq i \leq n$, and $\beta_i = \min(1, \sum_{j=1}^{i-1} e_j)$
for $1 \leq i \leq n$.

Note that $\beta_i = 1$ in Theorems 3 and 4 denotes that an addition actually occurs.

In the following, we shall always assume $nu \leq 0.01$.

4. Numerical Solution of Linear Equations

We wish to solve $Mz = b$, where M is $n \times n$ and non-singular. If the LU decomposition of M exists, i.e., if the $n-1$ leading principal minors of M are nonzero ([3], p. 27), then $M = \tilde{L}\tilde{U}$, where $\tilde{L}$ is unit lower triangular and $\tilde{U}$ is upper triangular.

Since M is non-singular, we will now show that the LU decomposition of M is unique. If $M = L_1U_1 = L_2U_2$, where L_1, L_2 are unit lower triangular and U_1, U_2 are upper triangular, then $\det L_1 = 1 = \det L_2$ and $\det U_1 = \det U_2 = \det M \neq 0$, so $L_1^{-1}L_2 = U_1U_2^{-1}$. But $L_1^{-1}L_2$ is unit lower triangular and $U_1U_2^{-1}$ is upper triangular. Thus, $L_1^{-1}L_2 = I = U_1U_2^{-1}$, and, hence, $L_1 = L_2$ and $U_1 = U_2$.

We solve for z by:

$$(4.1) \quad \begin{cases} \text{(i)} & \text{decomposing } M = \tilde{L}\tilde{U}, \\ \text{(ii)} & \text{solving } \tilde{L}f = b \text{ for } f, \\ \text{(iii)} & \text{solving } \tilde{U}z = f \text{ for } z. \end{cases}$$

If the LU decomposition of M does not exist, then, since M is non-singular, there exist permutation matrices P_1, P_2, Q_1, Q_2 (not necessarily unique) such that each of the following matrices has a unique LU decomposition:

- (a) $P_1M = L_1U_1$, e.g., interchanging rows only during Gaussian elimination (or Gaussian elimination with partial pivoting by rows),
- (b) $MQ_1 = L_2U_2$, e.g., interchanging columns only during Gaussian elimination (or Gaussian elimination with partial pivoting by columns),

- (c) $P_2 M Q_2 = L_3 U_3$, e.g., interchanging rows and columns during Gaussian elimination (or Gaussian elimination with complete pivoting).

Using (a), we solve for z by:

$$(4.2) \quad \begin{cases} \text{(i)} & \text{decomposing } P_1 M = L_1 U_1 , \\ \text{(ii)} & \text{solving } L_1 f_1 = P_1 b \text{ for } f_1 , \\ \text{(iii)} & \text{solving } U_1 z = f_1 \text{ for } z . \end{cases}$$

Using (b), we solve for z by:

$$(4.3) \quad \begin{cases} \text{(i)} & \text{decomposing } M Q_1 = L_2 U_2 , \\ \text{(ii)} & \text{solving } L_2 f_2 = b \text{ for } f_2 , \\ \text{(iii)} & \text{solving } U_2 y = f_2 \text{ for } y , \\ \text{(iv)} & \text{recovering } z = Q_1 y . \end{cases}$$

Using (c), we solve for z by:

$$(4.4) \quad \begin{cases} \text{(i)} & \text{decomposing } P_2 M Q_2 = L_3 U_3 , \\ \text{(ii)} & \text{solving } L_3 f_3 = P_2 b \text{ for } f_3 , \\ \text{(iii)} & \text{solving } U_3 w = f_3 \text{ for } w , \\ \text{(iv)} & \text{recovering } z = Q_2 w . \end{cases}$$

Without loss of generality, let us assume that the LU decomposition of M exists, i.e., that we need not do any interchanges and that we solve for z by (4.1). However, if (4.1) is performed using finite precision arithmetic, then we obtain the exact decomposition of a perturbation of M , i.e., $LU = M + F$. Similarly, when attempting to solve (ii) in (4.1), we obtain the exact solution y of $(L + \delta L)y = b$

where δL is lower triangular, and when attempting to solve (iii) in (4.1), we obtain the exact solution x of $(U + \delta U)x = y$ where δU is upper triangular.

Hence, when attempting to solve $Mz = b$ by (4.1), we obtain the exact solution x of $(M + \delta M)x = b$, where $\delta M = F + (\delta L)U + L(\delta U) + (\delta L)(\delta U)$, by the procedure of:

$$(4.1)' \quad \begin{cases} \text{(i)} & \text{decomposing } M + F = LU, \\ \text{(ii)} & \text{solving } (L + \delta L)y = b \text{ for } y, \\ \text{(iii)} & \text{solving } (U + \delta U)x = y \text{ for } x. \end{cases}$$

When M is symmetric, if the LU decomposition of M exists, $M = \tilde{L}\tilde{U}$, then $\tilde{U} = \tilde{D}\tilde{L}^t$, where $\tilde{D}$ is diagonal. Then we obtain z by:

$$(4.4) \quad \begin{cases} \text{(i)} & \text{decomposing } M = \tilde{L}\tilde{D}\tilde{L}^t, \\ \text{(ii)} & \text{solving } \tilde{L}f = b \text{ for } f, \\ \text{(iii)} & \text{solving } \tilde{D}w = f \text{ for } w, \\ \text{(iv)} & \text{solving } \tilde{L}^t z = w \text{ for } z. \end{cases}$$

Similarly, if (4.4) is performed using finite precision arithmetic, then we obtain the exact decomposition of a perturbation of M , i.e., $LDL^t = M + F$. Then we obtain the exact solution x of $(M + \delta M)x = b$ by:

$$(4.4)' \quad \begin{cases} \text{(i)} & \text{decomposing } M + F = LDL^t, \\ \text{(ii)} & \text{solving } (L + \delta L)y = b \text{ for } y, \text{ where } \delta L \text{ is lower triangular,} \\ \text{(iii)} & \text{solving } (D + \delta D)v = y \text{ for } v, \text{ where } \delta D \text{ is diagonal,} \\ \text{(iv)} & \text{solving } (L^t + \delta L^t)x = v \text{ for } x, \text{ where } \delta L^t \text{ is upper triangular.} \end{cases}$$

Now $\delta M = F + LD(\delta L^t) + [L(\delta D) + (\delta L)D + (\delta L)(\delta D)][L^t + \delta L^t]$.

5. Error Analysis of Triangular Systems

In attempting to solve $Lf = b$, we obtain the exact solution y to
 (4.1)'(ii) $(L + \delta L)y = b$, where L is unit lower triangular and δL is lower triangular.

In exact arithmetic, for $1 \leq i \leq n$:

$$(5.1) \quad f_i = -l_{i1}f_1 - \dots - l_{i,i-1}f_{i-1} + b_i,$$

which may be rewritten as

$$(5.2) \quad f_i = -L_i^t g_i + b_i,$$

$$\text{where } L_i = \begin{bmatrix} l_{i1} \\ \vdots \\ l_{i,i-1} \end{bmatrix} \text{ and } g_i = \begin{bmatrix} f_1 \\ \vdots \\ f_{i-1} \end{bmatrix}.$$

Let us assume we compute

$$(5.3) \quad fl(-L_i^t g_i + b_i) = \frac{-fl(L_i^t g_i) + b_i}{(1 + \eta_{ii})},$$

where $|\eta_{ii}| \leq u$ for $2 \leq i \leq n$ and $\eta_{11} = 0$ (since $L_1^t f_1 = 0$ and no addition occurs for $i = 1$).

Recall that $e_{ij} = \begin{cases} 0 & \text{if } l_{ij} = 0 \\ 1 & \text{if } l_{ij} \neq 0 \end{cases}$, for $1 \leq j \leq i \leq n$.

Let $\beta_{ij} = \min(1, \sum_{k=1}^{j-1} e_{ik})$. So $\beta_{i1} = 0$, and $\beta_{ij} = 1$ will indicate that an addition actually occurs during the computation of the inner product $L_i^t g_i$.

$$(5.4) \quad f_l(L_i^t g_i) = \sum_{j=1}^{i-1} l_{ij} f_j [1 + l_{01} u_{\theta_i} (1 + \beta_{ij} + \sum_{k=j+1}^{i-1} e_{ik})] e_{ij} \quad ,$$

Now we combine (5.1) - (5.4) to obtain

$$(5.5) \quad y_i = \frac{-\ell_{i1}(1 + \eta_{i1})y_1 - \dots - \ell_{i,i-1}(1 + \eta_{i,i-1})y_{i-1} + b_i}{1 + \eta_{ii}}$$
$$|\eta_{ij}| \leq 1.01 \, u(1 + \beta_{ij} + \sum_{k=j+1}^{i-1} e_{ik})e_{ij} \quad \text{for } 1 \leq j < i \leq n \quad .$$
$$\delta_L = \begin{bmatrix} 0 & & & & & & & & \\ \eta_{21}^{\ell_{21}} & \eta_{22} & & & & & & & \\ \eta_{31}^{\ell_{21}} & \eta_{32}^{\ell_{32}} & \eta_{33} & & & & & & \\ . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . \\ \eta_{n1}^{\ell_{n1}} & . & . & . & . & . & . & . & \eta_{nn} \end{bmatrix}$$
$$\|x\|_1 \equiv \sum_{i=1}^n |x_i|, \quad \|M\|_1 \equiv \max_{1 \leq j \leq n} \sum_{i=1}^n |m_{ij}| = \max \text{ column sum}. \quad \text{If } M = M^t,$$

20

5. Error Analysis of Triangular Systems

In attempting to solve $Lf = b$, we obtain the exact solution y to
 (4.1)'(ii) $(L + \delta L)y = b$, where L is unit lower triangular and δL is lower triangular.

In exact arithmetic, for $1 \leq i \leq n$:

$$(5.1) \quad f_i = -l_{i1}f_1 - \dots - l_{i,i-1}f_{i-1} + b_i,$$

which may be rewritten as

$$(5.2) \quad f_i = -L_i^t g_i + b_i,$$

$$\text{where } L_i = \begin{bmatrix} l_{i1} \\ \vdots \\ l_{i,i-1} \end{bmatrix} \text{ and } g_i = \begin{bmatrix} f_1 \\ \vdots \\ f_{i-1} \end{bmatrix}.$$

Let us assume we compute

$$(5.3) \quad fl(-L_i^t g_i + b_i) = \frac{-fl(L_i^t g_i) + b_i}{(1 + \eta_{ii})},$$

where $|\eta_{ii}| \leq u$ for $2 \leq i \leq n$ and $\eta_{11} = 0$ (since $L_1^t f_1 = 0$ and no addition occurs for $i = 1$).

Recall that $e_{ij} = \begin{cases} 0 & \text{if } l_{ij} = 0 \\ 1 & \text{if } l_{ij} \neq 0 \end{cases}$, for $1 \leq j \leq i \leq n$.

Let $\beta_{ij} = \min(1, \sum_{k=1}^{j-1} e_{ik})$. So $\beta_{i1} = 0$, and $\beta_{ij} = 1$ will indicate that an addition actually occurs during the computation of the inner product $L_i^t g_i$.

;

$$(5.4) \quad f l(L_i^t g_i) = \sum_{j=1}^{i-1} l_{ij} f_j [1 + 1.01 u_{\theta_i} (1 + \beta_{ij} + \sum_{k=j+1}^{i-1} e_{ik})] e_{ij} \quad ,$$

where $|\theta_j| \leq 1$ for $1 \leq j \leq i-1$.

Now we combine (5.1) - (5.4) to obtain

Theorem 5. The exact solution y of $(L + \delta L)y = b$ is given by:

$$(5.5) \quad y_i = \frac{f_l(-l_{i1}y_1 - l_{i2}y_2 - \dots - l_{i,i-1}y_{i-1} + b_i) - l_{i1}(1 + \eta_{i1})y_1 - \dots - l_{i,i-1}(1 + \eta_{i,i-1})y_{i-1} + b_i}{1 + \eta_{ii}},$$

where $\eta_{11} = 0$, $|\eta_{ii}| \leq u$ for $2 \leq i \leq n$,

$$|\eta_{ij}| \leq 1.01 u(1 + \beta_{ij} + \sum_{k=j+1}^{i-1} e_{ik})e_{ij} \quad \text{for } 1 \leq j < i \leq n \quad .$$

From (5.5),

$$\delta_L = \begin{bmatrix} 0 & & & & & & & & \\ \eta_{21} l_{21} & \eta_{22} & & & & & & & \\ \eta_{31} l_{21} & \eta_{32} l_{32} & \eta_{33} & & & & & & \\ . & & & & . & & & & \\ . & & & & & . & & & \\ . & & & & & & . & & \\ \eta_{n1} l_{n1} & . & . & . & . & . & . & . & \eta_{nn} \end{bmatrix}$$

$$\text{Let } \|x\|_{\infty} \equiv \max_{1 \leq i \leq n} |x_i|, \quad \|M\|_{\infty} \equiv \max_{1 \leq i \leq n} \sum_{j=1}^n |m_{ij}| = \max \text{ row sum}.$$

$$\|x\|_1 \equiv \sum_{i=1}^n |x_i|, \quad \|M\|_1 \equiv \max_{1 \leq j \leq n} \sum_{i=1}^n |m_{ij}| = \max \text{ column sum}. \quad \text{If } M = M^t,$$

then $\|M\|_1 = \|M\|_\infty$.

Thus, we can bound the one-norm of δL by:

$$(5.6) \quad \|\delta L\|_1 = \max_{1 \leq j \leq n} \sum_{i=j}^n |\eta_{ij} \ell_{ij}| \leq \tau \max_{1 \leq j \leq n} \sum_{i=j}^n |\eta_{ij}| ,$$

where $\tau \equiv \max_{i,j} |\ell_{ij}|$.

But $\eta_{11} = 0$ and $\beta_{11} = 0$, so from Theorem 5,

$$(5.7) \quad \begin{aligned} \sum_{i=1}^n |\eta_{i1}| &\leq 1.01 u \sum_{i=2}^n (1 + \sum_{k=2}^{i-1} e_{ik}) e_{i1} \\ &= 1.01 u (c_1 + \sum_{i=2}^n \sum_{k=2}^{i-1} e_{ik} e_{i1}) , \end{aligned}$$

since $c_1 = \sum_{i=2}^n e_{i1}$ is the number of non-zeros in the first column of L off the diagonal.

For $2 \leq j \leq n$, from Theorem 5,

$$(5.8) \quad \begin{aligned} \sum_{i=j}^n |\eta_{ij}| &\leq u + 1.01 u \sum_{i=j+1}^n (1 + \beta_{ij} + \sum_{k=j+1}^{i-1} e_{ik}) e_{ij} \\ &\leq 1.01 u \{1 + c_j + \sum_{i=j+1}^n (\beta_{ij} + \sum_{k=j+1}^{i-1} e_{ik}) e_{ij}\} , \end{aligned}$$

since $c_j = \sum_{i=j+1}^n e_{ij}$ is the number of non-zeros in the j -th column of L off the diagonal.

From (5.6) - (5.8), we have

Theorem 6. Let $\tau = \max_{i,j} |\ell_{ij}|$. Then

$$\begin{aligned} \|\delta L\|_1 &\leq 1.01 u \tau \max \left\{ c_1 + \sum_{i=2}^n \sum_{k=2}^{i-1} e_{ik} e_{i1}, \max_{2 \leq j \leq n-1} \left[1 + c_j + \sum_{i=j+1}^n (\beta_{ij} + \sum_{k=j+1}^{i-1} e_{ik}) e_{ij} \right] \right\} \\ &\leq 1.01 u \tau \max_{1 \leq j \leq n-1} \left\{ 1 + c_j + \sum_{i=j+1}^n (\beta_{ij} + \sum_{k=j+1}^{i-1} e_{ik}) e_{ij} \right\} . \end{aligned}$$

For the ∞ -norm, we cannot express the bound in terms of the c_j , since the sum is across the rows and the c_j measure down the columns. Instead, we obtain

Theorem 7.
$$\|\delta L\|_{\infty} \leq 1.01 u \tau \max_{1 \leq i \leq n} \left\{ 1 + \sum_{j=1}^{i-1} (1 + \beta_{ij} + \sum_{k=j+1}^{i-1} e_{ik}) e_{ij} \right\} .$$

For a full matrix $e_{ij} = 1$ for all i, j . Thus, $\beta_{ij} = 1$ for $j \geq 2$,

$$c_j = n-j, \quad \sum_{i=j+1}^n \left(\sum_{k=j+1}^{i-1} e_{ik} \right) e_{ij} = \sum_{i=j+1}^n (i-1-j) = \frac{1}{2} (n-j)(n-j-1)$$

for $1 \leq j \leq n$. From Theorems 6 and 7,

$$(5.9) \quad \|\delta L\|_1 \leq 1.01 u \tau \max \left\{ n-1 + \frac{1}{2} (n-1)(n-2), \max_{2 \leq j \leq n-1} \left[1 + 2(n-j) + \frac{1}{2} (n-j)(n-j-1) \right] \right\}$$

$$= 1.01 u \tau n(n-1)/2 ,$$

and

$$(5.10) \quad \|\delta L\|_{\infty} \leq 1.01 u \tau \max_{1 \leq i \leq n} \left\{ 1 + i-2 + \sum_{j=2}^{i-1} (2 + i-1-j) + 1 \right\}$$

$$= 1.01 u \tau \left[\frac{n(n+1)}{2} - 1 \right] .$$

We summarize this as

Corollary 1. When M is a full matrix,

$$\|\delta L\|_1 \leq 1.01 u \tau \frac{n(n-1)}{2} \quad \text{and}$$

$$\|\delta L\|_{\infty} \leq 1.01 u \tau \left[\frac{n(n+1)}{2} - 1 \right] .$$

Now, let M be a band matrix with bandwidth $2m+1$, i.e.,
 $m_{ij} = 0$ for $|i-j| > m$. Assume M has nonzero elements within the
band. Then L is unit lower triangular with $l_{ij} = 0$ for $i > j+m$,

and so $e_{ij} = 0$ for $i > j+m$. Then, $\beta_{ij} = \begin{cases} 0 & \text{for } j \leq i-m \\ 1 & \text{for } i-m+1 \leq j \leq i-1 \end{cases}$,

$$\sum_{i=j+1}^n \beta_{ij} = \begin{cases} m-1 & \text{for } 1 \leq j \leq n-m \\ n-j-1 & \text{for } n-m+1 \leq j \leq n \end{cases}, \quad c_j = \begin{cases} m & \text{for } 1 \leq j \leq n-m \\ n-j & \text{for } n-m+1 \leq j \leq n \end{cases},$$

$$\text{and } \sum_{i=j+1}^n \left(\sum_{k=j+1}^{i-1} e_{ik} \right) e_{ij} = \sum_{i=j+1}^{\min(n, j+m)} \sum_{k=j+1}^{i-1} e_{ik} = \sum_{i=j+1}^{\min(n, j+m)} (i-1-j) =$$

$$\begin{cases} \frac{1}{2} m(m-1) & \text{for } 1 \leq j \leq n-m \\ \frac{1}{2} (n-j)(n-j-1) & \text{for } n-m+1 \leq j \leq n \end{cases}. \quad \text{Thus } \|\delta L\|_1 \leq 1.01 u \tau \left[\frac{1}{2} m^2 + \frac{3}{2} m \right].$$

Similarly, $\|\delta L\|_\infty \leq 1.01 u \tau \left[\frac{1}{2} m^2 + \frac{3}{2} m \right]$.

We summarize this in

Corollary 2. For an $n \times n$ band matrix with bandwidth $2m+1$,

$$\|\delta L\|_1, \|\delta L\|_\infty \leq 1.01 u \tau \left[\frac{1}{2} m^2 + \frac{3}{2} m \right].$$

Note that the bound is dependent on the bandwidth of the matrix
and is independent of the order n of the matrix.

In analogy with [3], pp. 104-106, let us consider the solution of
 $Rx = b$, where R is lower triangular and non-singular. In finite
precision arithmetic we obtain the exact solution y to $(R + \delta R)y = b$.

In exact arithmetic, for $1 \leq i \leq n$:

$$(5.11) \quad x_i = \frac{-r_{i1}x_1 - \dots - r_{i,i-1}x_{i-1} + b_i}{r_{ii}},$$

which may be rewritten as

$$(5.12) \quad x_i = \frac{-R_i^t z_i + b_i}{r_{ii}},$$

$$\text{where } R_i = \begin{bmatrix} r_{i1} \\ \vdots \\ r_{i,i-1} \end{bmatrix} \text{ and } z_i = \begin{bmatrix} x_1 \\ \vdots \\ x_{i-1} \end{bmatrix}.$$

Let us assume that we compute $R_i^t z_i$, subtract it from b_i , and then divide by r_{ii} . Then

$$(5.13) \quad y_i = fl\left(\frac{-R_i^t z_i + b_i}{r_{ii}}\right) = fl\left\{\frac{fl[-fl(R_i^t z_i) + b_i]}{r_{ii}}\right\} \\ = \frac{fl(-R_i^t z_i) + b_i}{r_{ii}(1 + \eta_{ii})(1 + \eta'_{ii})},$$

where $|\eta_{ii}| \leq u$ for $1 \leq i \leq n$, $\eta'_{11} = 0$, $|\eta'_{ii}| \leq u$ for $2 \leq i \leq n$.

Again, let $\beta_{ij} = \min(1, \sum_{k=1}^{j-1} e_{ik})$. From (5.4) we obtain

Theorem 8. The exact solution y of $(R + \delta R)y = b$, where R is lower triangular, is given by:

$$(5.14) \quad y_i = fl\left(\frac{-r_{i1}y_1 - r_{i2}y_2 - \dots - r_{i,i-1}y_{i-1} + b_i}{r_{ii}}\right) \\ = \frac{-r_{i1}(1 + \eta_{i1})y_1 - \dots - r_{i,i-1}(1 + \eta_{i,i-1})y_{i-1} + b_i}{r_{ii}(1 + \eta_{ii})(1 + \eta'_{ii})}$$

where $|\eta_{ii}| \leq u$ for $1 \leq i \leq n$, $\eta_{11} = 0$, $|\eta'_{ii}| \leq u$ for $2 \leq i \leq n$, and $|\eta_{ij}| \leq 1.01 u (1 + \beta_{ij} + \sum_{k=j+1}^{i-1} e_{ik}) e_{ij}$ for $1 \leq j < i \leq n$.

Corresponding to Theorems 6 and 7, we have

Theorem 9. Let $\rho = \max_{i,j} |r_{ij}|$. Then $\|\delta R\|_1 \leq 1.01 u \rho \max\{1 + c_1 +$

$$\sum_{i=2}^n \sum_{k=2}^{i-1} e_{ik} e_{il}, \max_{2 \leq j \leq n-1} [1 + c_j + \sum_{i=j+1}^n (\beta_{ij} + \sum_{k=j+1}^{i-1} e_{ik}) e_{ij}]\}, \text{ and}$$

$$\|\delta R\|_{\infty} \leq 1.01 u \rho \max_{1 \leq i \leq n} \{2 + \sum_{j=1}^{i-1} (1 + \beta_{ij} + \sum_{k=j+1}^{i-1} e_{ik}) e_{ij}\}.$$

Let R be a full lower triangular matrix, i.e., $R_{ij} \neq 0$ for $i \geq j$. Let $|\delta R|$ be the matrix whose (i,j) element is $|(\delta R)_{ij}|$.

Then

$$|\delta R| \leq 1.01 u \rho \begin{bmatrix} 1 & & & & & & & \\ 1 & 2 & & & & & & \\ 2 & 2 & 2 & & & & & \\ 3 & 3 & 2 & 2 & & & & \\ \vdots & & & & \ddots & & & \\ \vdots & & & & & \ddots & & \\ \vdots & & & & & & \ddots & \\ n-1 & n-1 & n-2 & n-3 & \cdot & \cdot & \cdot & 2 \end{bmatrix},$$

so

$$\|\delta R\|_1 \leq 1.01 u \rho \left[\frac{1}{2} n(n-1) + 1 \right] \quad \text{and}$$

$$\|\delta R\|_{\infty} \leq 1.01 u \rho \left[\frac{1}{2} n(n+1) \right].$$

Next, in attempting to solve $Uz = y$, we obtain the exact solution x to (4.1)'(iii) $(U + \delta U)x = y$, where U and δU are upper

$$\sigma \equiv \max_{i,j} |u_{ij}|.$$

Since $\|A^t\|_1 = \|A\|_\infty$ for all A , when U is a full upper triangular matrix, we have from the above that

$$\|\delta U\|_1 \leq 1.01 u \sigma \left[\frac{1}{2} n(n+1) \right] \quad \text{and}$$

$$\|\delta U\|_\infty \leq 1.01 u \sigma \left[\frac{1}{2} n(n-1) + 1 \right] .$$

However, when U is arbitrarily sparse, we cannot, in general, use the analysis for $R = U^t$.

In exact arithmetic, we backsolve, for $i = n, n-1, \dots, 1$:

$$(5.15) \quad z_i = \frac{-u_{i,i+1}z_{i+1} - \dots - u_{i,n}z_n + y_i}{u_{ii}} ,$$

which may be rewritten as

$$(5.16) \quad z_i = (-U_{ii}^t h_i + y_i) / u_{ii} ,$$

$$\text{where } U_i = \begin{bmatrix} u_{i,i+1} \\ \vdots \\ u_{in} \end{bmatrix} \quad \text{and} \quad h_i = \begin{bmatrix} z_{i+1} \\ \vdots \\ z_n \end{bmatrix} , \quad \text{for } i = n, n-1, \dots, 1 .$$

The usual order for calculating (5.16) is first computing the inner product $U_{ii}^t h_i$ (i.e., $u_{i,i+1}z_{i+1} + \dots + u_{in}z_n$ is computed from left to right), subtract this from y_i , and then divide by u_{ii} . Thus, in finite precision, we have, for $i = n, n-1, \dots, 1$:

$$(5.17) \quad x_i = fl\{fl[-fl(U_{ii}^t h_i) + y_i] / u_{ii}\} \\ = \frac{-fl(U_{ii}^t h_i) + y_i}{u_{ii}(1+\mu_{ii})(1+\mu'_{ii})} ,$$

where $\mu'_{nn} = 0$, $|\mu'_{ii}| \leq u$ for $1 \leq i \leq n-1$, and $|\mu_{ii}| \leq u$ for $1 \leq i \leq n$.

Using the analysis for $R = U^t$ would imply that one calculates $U_{ii}^t h_i$ from right to left: $u_{in} z_n + u_{i,n-1} z_{n-1} + \dots + u_{i,i+1} z_{i+1}$.

Thus, the structure of $(\delta R)^t$ in the sparse case is quite different from that of δU , where we have computed $U_{ii}^t h_i$ in the usual order from left to right: $u_{i,i+1} z_{i+1} + \dots + u_{in} z_n$.

$$\text{Recall that } e_{ij} = \begin{cases} 0 & \text{if } u_{ij} = 0 \\ 1 & \text{if } u_{ij} \neq 0 \end{cases}, \text{ for } 1 \leq i \leq j \leq n.$$

Let $\gamma_{ij} = \min(1, \sum_{k=i+1}^{j-1} e_{ik})$; so $\gamma_{i,i+1} = 0$, and $\gamma_{ij} = 1$ will

indicate that an addition actually occurs during the computation of the inner product $U_{ii}^t h_i$.

From Theorem 3 once again, we obtain, for $i = n, n-1, \dots, 1$:

$$(5.18) \quad fl(U_{ii}^t h_i) = \sum_{j=i+1}^n u_{ij} x_j [1 + 1.01 u \varphi_i (1 + \gamma_{ij} + \sum_{k=j+1}^n e_{ik})] e_{ij},$$

where $|\varphi_i| \leq 1$ for $i+1 \leq j \leq n$.

Combining (5.15) - (5.18), we have

Theorem 10. The exact solution x of $(U + \delta U)x = y$ is given by:

$$(5.19) \quad x_i = fl \left(\frac{-u_{i,i+1} x_{i+1} - \dots - u_{in} x_n + y_i}{u_{ii}} \right) \\ = \frac{-u_{i,i+1} (1 + \mu_{i,i+1}) x_{i+1} - \dots - u_{in} (1 + \mu_{in}) x_n + y_i}{u_{ii} (1 + \mu_{ii}) (1 + \mu'_{ii})},$$

where $\mu'_{nn} = 0$, $|\mu'_{ii}| \leq u$ for $1 \leq i \leq n-1$, $|\mu_{ii}| \leq u$ for $1 \leq i \leq n$, and

$$|\mu_{ij}| \leq 1.01 u (1 + \gamma_{ij} + \sum_{k=j+1}^n e_{ik}) e_{ij} \quad \text{for } 1 \leq i < j \leq n.$$

$$\text{From (5.19), } (\delta U)_{ij} = \begin{cases} \mu_{ij} u_{ij} & \text{for } j > i \\ [\mu_{ii} + \mu'_{ii}(1 + \mu_{ii})] u_{ii} & \text{for } j = i \\ 0 & \text{for } j < i \end{cases}.$$

The one-norm of δU is bounded by:

$$(5.20) \quad \|\delta U\|_1 = \max_{1 \leq j \leq n} \{ |\mu_{jj} + \mu'_{jj}(1 + \mu_{jj})| |u_{jj}| + \sum_{i=1}^{j-1} |\mu_{ij} u_{ij}| \} \\ \leq \sigma \max_{1 \leq j \leq n} \{ |\mu_{jj}| + |\mu'_{jj}|(1 + |\mu_{jj}|) + \sum_{i=1}^{j-1} |\mu_{ij}| \},$$

where $\sigma \equiv \max_{i,j} |u_{ij}|$.

For $1 \leq j \leq n-1$, $|\mu_{jj}| + |\mu'_{jj}|(1 + |\mu_{jj}|) \leq u + u(1+u) \leq 2(1.01)u$
since $u \leq 0.02$, and $|\mu_{nn}| + |\mu'_{nn}|(1 + |\mu_{nn}|) = |\mu_{nn}| \leq u$.

Thus, from (5.19) and (5.20), we have

Theorem 11. Let $\sigma = \max_{i,j} |u_{ij}|$. Then

$$\|\delta U\|_1 \leq 1.01 u \sigma \max \left[\max_{1 \leq j \leq n-1} \left\{ 2 + \sum_{i=1}^{j-1} (1 + \gamma_{ij} + \sum_{k=j+1}^n e_{ik}) e_{ij} \right\}, 1 + \sum_{i=1}^{n-1} (1 + \gamma_{in}) e_{in} \right].$$

The ∞ -norm is bounded by:

$$(5.21) \quad \|\delta U\|_{\infty} = \max_{1 \leq i \leq n} \{ |\mu_{ii} + \mu'_{ii}(1 + \mu_{ii})| |u_{ii}| + \sum_{j=i+1}^n |\mu_{ij}| |u_{ij}| \} \\ \leq \sigma \max_{1 \leq i \leq n} \{ |\mu_{ii}| + |\mu'_{ii}|(1 + |\mu_{ii}|) + \sum_{j=i+1}^n |\mu_{ij}| \}.$$

$$\begin{aligned}
(5.22) \quad \sum_{j=i+1}^n |\mu_{ij}| &\leq 1.01 u \sum_{j=i+1}^n (1 + \gamma_{ij} + \sum_{k=j+1}^n e_{ik}) e_{ij} \\
&= 1.01 u [r_i + \sum_{j=i+1}^n (\gamma_{ij} + \sum_{k=j+1}^n e_{ik}) e_{ij}] ,
\end{aligned}$$

since $r_i = \sum_{j=i+1}^n e_{ij}$ = number of nonzeros in the i -th row of U off the diagonal.

From (5.21) and (5.22), we obtain

Theorem 12. $\|\delta U\|_{\infty} \leq 1.01 u_{\sigma} \max_{1 \leq i \leq n} \{2 + r_i + \sum_{j=i+1}^n (\gamma_{ij} + \sum_{k=j+1}^n e_{ik}) e_{ij}\} .$

For a full matrix, $e_{ij} = 1$ for all i, j ; $\gamma_{ij} = 1$ for $j > i$;

$$r_i = n-i ; \quad \sum_{i=1}^{j-1} \left(\sum_{k=j+1}^n e_{ik} \right) e_{ij} = \sum_{i=1}^{j-1} (n-j) = (n-j)(j-1) ; \text{ and}$$

$$\sum_{j=i+1}^n \left(\sum_{k=j+1}^n e_{ik} \right) e_{ij} = \sum_{j=i+1}^n (n-j) = (n-i)(n-i-1)/2 . \text{ From Theorems 11}$$

and 12,

$$\begin{aligned}
\|\delta U\|_1 &\leq 1.01 u_{\sigma} \max\{2 + (n-j+2)(j-2) + n-j+1\} \\
&= 1.01 u_{\sigma} \left\{ 1 + \left(\frac{n+1}{2} \right)^2 \right\} \\
&= 1.01 u_{\sigma} \left\{ \frac{1}{4} n^2 + \frac{1}{2} n + \frac{5}{4} \right\} ,
\end{aligned}$$

$$\begin{aligned}
\|\delta U\|_{\infty} &\leq 1.01 u_{\sigma} \max_{1 \leq i \leq n} \left\{ 2 + n - i + n - i - 1 + \frac{1}{2} (n - i)(n - i - 1) \right\} \\
&= 1.01 u_{\sigma} \left\{ 2 + 2n - 3 + \frac{1}{2} (n - 1)(n - 2) \right\} \\
&= 1.01 u_{\sigma} \left[\frac{n(n+1)}{2} \right] , \quad \text{and}
\end{aligned}$$

$$|\delta U| \leq 1.01 u_{\sigma} \begin{bmatrix} 2 & n-1 & n-1 & n-2 & n-3 & . & . & . & 3 & 2 \\ & 2 & n-2 & n-2 & n-3 & . & . & . & 3 & 2 \\ & & 2 & n-3 & n-3 & . & . & . & 3 & 2 \\ & & & 2 & n-4 & . & . & . & 3 & 2 \\ & & & & \ddots & & & & \vdots & \vdots \\ & & & & & & 2 & 3 & 3 & 2 \\ & & & & & & & 2 & 2 & 2 \\ & & & & & & & & 2 & 1 \\ & & & & & & & & & 2 \end{bmatrix}$$

We summarize this as

Corollary 3. When M is a full matrix,

$$\begin{aligned}
\|\delta U\|_1 &\leq 1.01 u_{\sigma} \left[\frac{1}{4} n^2 + \frac{1}{2} n + \frac{5}{4} \right] \quad \text{and} \\
\|\delta U\|_{\infty} &\leq 1.01 u_{\sigma} \left[\frac{n(n+1)}{2} \right] .
\end{aligned}$$

Let M be a band matrix with bandwidth $2m+1$. When no pivoting is needed, $u_{ij} = 0 = e_{ij}$ for $j > i+m$. Then, for $m \leq j \leq n$,

$$\sum_{i=1}^{j-1} (1 + \gamma_{ij} + \sum_{k=j+1}^n e_{ik}) e_{ij} = \sum_{i=j-m}^{j-1} (1 + \gamma_{ij} + \sum_{k=j+1}^{i+m} 1) = m + \sum_{i=j-m}^{j-2} (2+i+m-j) =$$

$$\frac{1}{2} m^2 + \frac{3}{2} m - 1. \quad \text{For } 1 \leq i \leq n-m, \quad r_i = m \quad \text{and}$$

$$\sum_{j=i+1}^n (\gamma_{ij} + \sum_{k=j+1}^n e_{ik}) e_{ij} = \sum_{j=i+1}^{i+m} (\gamma_{ij} + \sum_{k=j+1}^{i+m} 1) = m - 1 + \sum_{j=i+2}^{i+m} (1+i+m-j) =$$

$$\frac{1}{2} m^2 + \frac{1}{2} m - 1. \quad \text{Thus, } \|\delta U\|_1 \leq 1.01 u \sigma \left[\frac{1}{2} m^2 + \frac{3}{2} m + 1 \right] \quad \text{and}$$

$$\|\delta U\|_\infty \leq 1.01 u \sigma \left[\frac{1}{2} m^2 + \frac{3}{2} m + 1 \right].$$

When partial pivoting is used, then $u_{ij} = 0 = e_{ij}$ for $j > i+2m$.

$$\text{For } 2m \leq j \leq n, \quad \sum_{i=1}^{j-1} (1 + \gamma_{ij} + \sum_{k=j+1}^n e_{ik}) e_{ij} \leq \sum_{i=j-2m}^{j-1} (1 + \gamma_{ij} + \sum_{k=j+1}^{i+2m} 1) =$$

$$2m + \sum_{i=j-2m}^{j-2} (2+i+2m-j) = 2m^2 + 3m - 1. \quad \text{For } 1 \leq i \leq n-2m, \quad r_i \leq 2m,$$

$$\sum_{j=i+1}^n (\gamma_{ij} + \sum_{k=j+1}^n e_{ik}) e_{ij} \leq \sum_{j=i+1}^{i+2m} (\gamma_{ij} + \sum_{k=j+1}^{i+2m} 1) = 2m - 1 + \sum_{j=i+2}^{i+2m} (1+i+2m-j) =$$

$$2m^2 + m - 1. \quad \text{Thus, } \|\delta U\|_1 \leq 1.01 u \sigma [2m^2 + 3m + 1] \quad \text{and}$$

$$\|\delta U\|_\infty \leq 1.01 u \sigma [2m^2 + 3m + 1].$$

We summarize this as

Corollary 4. When M is a band matrix with bandwidth $2m+1$,

$$\|\delta U\|_1, \|\delta U\|_\infty \leq 1.01 u \sigma \left[\frac{1}{2} m^2 + \frac{3}{2} m + 1 \right]$$

when no pivoting is needed, while

$$\|\delta U\|_1, \|\delta U\|_\infty \leq 1.01 u \sigma [2m^2 + 3m + 1]$$

when partial pivoting (by rows) is used.

Note once again that for band matrices the bounds depend on the bandwidth but not on the order of M .

When M is symmetric and if $M = LU$ exists, then $U = DL^t$, where D is diagonal. Here $r_i = c_i \equiv d_i$, and we obtain $(D + \delta D)v = y$ and $(L^t + \Delta L^t)x = v$ in finite precision arithmetic.

We then have immediately

Theorem 13. When M is symmetric, the exact solution v of $(D + \delta D)v = y$ is given by:

$$(5.23) \quad v_i = fl(y_i/D_{ii}) = \frac{y_i}{D_{ii}(1 + \delta_{ii})},$$

where $|\delta_{ii}| \leq u$ for $1 \leq i \leq n$. Thus,

$$(5.24) \quad \|\delta D\|_1 = \|\delta D\|_\infty \leq u \max_i |D_{ii}|.$$

Solving $(L^t + \Delta L^t)x = v$, we have for $i = n, n-1, \dots, 1$:

$$(5.25) \quad x_i = fl[-fl(\ell_{i,i+1}x_{i+1} + \dots + \ell_{in}x_n) + v_i] \\ = \frac{-fl(\ell_{i,i+1}x_{i+1} + \dots + \ell_{in}x_n) + v_i}{(1 + \mu_{ii})},$$

where $|\mu_{ii}| \leq u$.

An analysis similar to (5.15)-(5.18) gives

Theorem 14. When M is symmetric, the exact solution x of $(L^t + \Delta L^t)x = v$ is given, for $i = n, n-1, \dots, 1$, by:

$$(5.26) \quad x_i = fl(-l_{i,i+1}x_{i+1} - \dots - l_{in}x_n + v_i) \\ = \frac{-l_{i,i+1}(1+\mu_{i,i+1})x_{i+1} - \dots - l_{in}(1+\mu_{in})x_n + v_i}{(1+\mu_{ii})},$$

where $\mu_{nn} = 0$, $|\mu_{ii}| \leq u$ for $1 \leq i \leq n-1$, and

$$|\mu_{ij}| \leq 1.01 u (1 + \gamma_{ij} + \sum_{k=j+1}^n e_{ik}) e_{ij} \quad \text{for } 1 \leq i < j \leq n, \text{ and}$$

$$\gamma_{ij} = \min(1, \sum_{k=i+1}^{j-1} e_{ik}). \quad \text{Then}$$

$$(5.27) \quad \|\Delta L^t\|_1 \leq 1.01 u \tau \max \left[\max_{1 \leq j \leq n-1} \left\{ 1 + \sum_{i=1}^{j-1} (1 + \gamma_{ij} + \sum_{k=j+1}^n e_{ik}) e_{ij} \right\}, \sum_{i=1}^{n-1} (1 + \gamma_{in}) e_{in} \right]$$

and

$$(5.28) \quad \|\Delta L^t\|_\infty \leq 1.01 u \tau \max_{1 \leq i \leq n} \left\{ 1 + d_i + \sum_{j=i+1}^n (\gamma_{ij} + \sum_{k=j+1}^n e_{ik}) e_{ij} \right\}.$$

From Theorem 14 we have immediately

Corollary 5. When M is a symmetric full matrix,

$$\|\Delta L^t\|_1 \leq 1.01 u \tau \left[\frac{1}{4} n^2 + \frac{1}{2} n + \frac{1}{4} \right] \quad \text{and}$$

$$\|\Delta L^t\|_\infty \leq 1.01 u \tau \left[\frac{n(n+1)}{2} - 1 \right]$$

and

Corollary 6. When M is a symmetric band matrix with bandwidth $2m+1$,

$$\|\Delta L^t\|_1, \|\Delta L^t\|_\infty \leq 1.01 u \tau \left[\frac{1}{2} m^2 + \frac{3}{2} m \right].$$

6. Error Analysis of Triangular Factorization

Now we shall determine F in (4.1)'(i), i.e., a matrix F such that due to the finite precision arithmetic we now have the exact LU decomposition of $M+F$ rather than M .

Let $M^{(1)} \equiv M$. Then the elimination is defined sequentially for $1 \leq k \leq n-1$ by the following:

given $M^{(k)}$, let $L^{(k)}$ be defined by

$$\ell_{ij}^{(k)} \equiv \begin{cases} 1 & \text{for } 1 \leq i \leq n, \quad j = i \\ f\ell(m_{ik}^{(k)}/m_{kk}^{(k)}) & \text{for } k+1 \leq i \leq n \\ 0 & \text{otherwise;} \end{cases}$$

$$L^{(k)} = \begin{bmatrix} 1 & & & & & \\ & 1 & & & & \\ & & \ddots & & & \\ & & & 1 & & \\ & & & \ell_{k+1,k}^{(k)} & & \\ & & & \vdots & & \\ & & & \ell_{nk}^{(k)} & & \\ & & & & 1 & \\ & & & & & \ddots \\ & & & & & & 1 \end{bmatrix}$$

Then $M^{(k+1)} \equiv fl[(L^{(k)})^{-1} M^{(k)}]$, i.e.,

$$m_{ij}^{(k+1)} \equiv \begin{cases} 0 & \text{for } k+1 \leq i \leq n, \quad j = k \\ f \ell(m_{ij}^{(k)} - \ell_{ik}^{(k)} \times m_{kj}^{(k)}) & \text{for } k+1 \leq i, j \leq n \\ m_{ij}^{(k)} & \text{otherwise;} \end{cases}$$

$$M^{(k+1)} =$$

Then $U = M^{(n)}$ and $L = L^{(1)} L^{(2)} \dots L^{(n-1)}$, i.e.,

$$L_{ij} = \begin{cases} 0 & \text{for } i < j \\ 1 & \text{for } i = j \\ L_{ij}^{(j)} & \text{for } i > j \end{cases}.$$

Thus, if $m_{ik}^{(k)} \neq 0$, i.e., if $e_{ik} = 1$, then $l_{ik}^{(k)} \neq 0$. So, for $k+1 \leq i \leq n$,

$$(6.1) \quad \ell_{ik}^{(k)} = (m_{ik}^{(k)} / m_{kk}^{(k)}) (1 + \delta_{ik} e_{ik}) \quad ,$$

where $|\delta_{ik}| \leq u$, or we may express (6.1) as

$$(6.2) \quad 0 = m_{ik}^{(k)} - \ell_{ik}^{(k)} m_{kk}^{(k)} + \varepsilon_{ik}^{(k)},$$

where $\varepsilon_{ik}^{(k)} \equiv m_{ik}^{(k)} \delta_{ik} e_{ik}$.

If $l_{ik}^{(k)} \neq 0$ and $m_{kj}^{(k)} \neq 0$, i.e., if $e_{ik} = 1 = e_{kj}$, then we multiply $l_{ik}^{(k)}$ and $m_{kj}^{(k)}$ and subtract from $m_{ij}^{(k)}$. Thus, for $k+1 \leq i, j \leq n$:

$$(6.3) \quad m_{ij}^{(k+1)} = \frac{m_{ij}^{(k)} - \ell_{ik}^{(k)} m_{kj}^{(k)} (1 + \delta_{ij} e_{ik} e_{kj})}{1 + \delta'_{ij} e_{ik} e_{kj}},$$

where $|\delta'_{ij}|, |\delta_{ij}| \leq u$.

Note that in (6.3) we are counting an addition if $\ell_{ik}^{(k)} m_{kj}^{(k)} \neq 0$, i.e., whether or not $m_{ij}^{(k)} \neq 0$. In other words, we are counting an addition whenever fill-in occurs, cf. Section 2 and Bunch and Rose [2], Section 2.

We may express (6.3), for $k+1 \leq i, j \leq n$ as

$$(6.4) \quad m_{ij}^{(k+1)} = m_{ij}^{(k)} - \ell_{ik}^{(k)} m_{kj}^{(k)} + \varepsilon_{ij}^{(k)},$$

where $\varepsilon_{ij}^{(k)} \equiv -\ell_{ik}^{(k)} m_{kj}^{(k)} \delta_{ij} e_{ik} e_{kj} - m_{ij}^{(k+1)} \delta'_{ij} e_{ik} e_{kj}$.

Let $\varepsilon_{ij}^{(k)} = 0$ otherwise. From Section 5, $\tau \equiv \max_{i,j} |\ell_{ij}|$ and $\sigma \equiv \max_{i,j} |u_{ij}|$. Hence, $|\ell_{ik}^{(k)}| \leq \tau$ for all i, k and $|m_{ij}^{(k)}| \leq \sigma$ for all i, j, k . Then

$$(6.5) \quad \begin{cases} |\varepsilon_{ik}^{(k)}| \leq \sigma u e_{ik} & \text{for } k+1 \leq i \leq n \\ |\varepsilon_{ij}^{(k)}| \leq (\tau+1)\sigma u e_{ik} e_{kj} & \text{for } k+1 \leq i, j \leq n \end{cases}$$

For $1 \leq k \leq n-1$, let

$$(6.6) \quad \begin{aligned} F^{(k)} &\equiv M^{(k+1)} - 2M^{(k)} + L^{(k)} M^{(k)} \\ &= M^{(k+1)} - (L^{(k)})^{-1} M^{(k)} \end{aligned}$$

Then $F = \sum_{k=1}^{n-1} F^{(k)}$, and we conclude with

Theorem 15. In finite precision arithmetic $M+F = LU$, where F is given by:

$$(6.7) \quad |F_{ij}| \leq \begin{cases} u(\tau+1)_\sigma \sum_{k=1}^{i-1} e_{ik} e_{kj} & \text{for } i \leq j \\ u(\tau+1)_\sigma \sum_{k=1}^{j-1} e_{ik} e_{kj} + u\sigma e_{ij} & \text{for } i > j \end{cases},$$

and hence

$$(6.8) \quad \|F\|_1 \leq u\sigma \max_{1 \leq j \leq n} \left\{ \sum_{i=j+1}^n e_{ij} + (\tau+1) \left[\sum_{i=2}^j \sum_{k=1}^{i-1} e_{ik} e_{kj} + \sum_{i=j+1}^n \sum_{k=1}^{j-1} e_{ik} e_{kj} \right] \right\}$$

and

$$(6.9) \quad \|F\|_\infty \leq u\sigma \max_{1 \leq i \leq n} \left\{ \sum_{j=1}^{i-1} e_{ij} + (\tau+1) \left[\sum_{j=1}^{i-1} \sum_{k=1}^{j-1} e_{ik} e_{kj} + \sum_{j=i}^n \sum_{k=1}^{i-1} e_{ik} e_{kj} \right] \right\}.$$

When M is full, $\sum_{i=j+1}^n e_{ij} = n-j$, $\sum_{i=2}^j \sum_{k=1}^{i-1} e_{ik} e_{kj} = \sum_{i=2}^j (i-1) =$

$\frac{1}{2} j(j-1)$, and $\sum_{i=j+1}^n \left(\sum_{k=1}^{j-1} e_{ik} \right) e_{kj} = \sum_{i=j+1}^n (j-1) = (n-j)(j-1)$;

so $\|F\|_1 \leq u\sigma \max_{1 \leq j \leq n} \{n-j + (\tau+1)[(n-\frac{1}{2}j)(j-1)]\} \leq u\sigma\{n-1 + \frac{1}{2}(\tau+1)n(n-1)\}$.

Similarly, $\sum_{j=1}^{i-1} e_{ij} = i-1$, $\sum_{j=1}^{i-1} \sum_{k=1}^{j-1} e_{ik} e_{kj} = \sum_{j=1}^{i-1} (j-1) = \frac{1}{2}(i-1)(i-2)$, and

$\sum_{j=i}^n \sum_{k=1}^{i-1} e_{ik} e_{kj} = \sum_{j=i}^n (i-1) = (n-i+1)(i-1)$; so

$$\|F\|_\infty \leq u\sigma \max_{1 \leq i \leq n} \left\{ i-1 + (\tau+1) \left[\frac{1}{2}(i-1)(i-2) + (n-i+1)(i-1) \right] \right\} \leq$$

$u\sigma\{n-1 + \frac{1}{2}(\tau+1)n(n-1)\}$.

Thus, we have

Corollary 7. When M is a full $n \times n$ matrix,

$$\|F\|_1, \|F\|_\infty \leq u\sigma[n-1 + \frac{1}{2}(\tau+1)n(n-1)] .$$

When M is a band matrix with bandwidth $2m+1$, $e_{ij} = 0$ for $i > j+m$.

If no pivoting is needed, then $e_{ij} = 0$ for $j > i+m$. Then,

$$\text{for } m \leq j \leq n-m, \quad \sum_{i=j+1}^n e_{ij} = \sum_{i=j+1}^{j+m} 1 = m, \quad \sum_{i=2}^j \sum_{k=1}^{i-1} e_{ik} e_{kj} =$$

$$\sum_{i=j-m}^j \sum_{k=j-m}^{i-1} e_{ik} e_{kj} = \sum_{i=j-m}^j (i-j+m) = \sum_{r=0}^m r = \frac{1}{2} m(m+1), \text{ and}$$

$$\sum_{i=j+1}^n \sum_{k=1}^{j-1} e_{ik} e_{kj} = \sum_{i=j+1}^{j+m} \sum_{k=j-m}^{j-1} e_{ik} e_{kj} = \sum_{i=j+1}^{j+m} (m-i+j) = \sum_{r=0}^{m-1} r = \frac{1}{2} m(m-1) .$$

So $\|F\|_1 \leq u\sigma[m + (\tau+1)m^2]$. Similarly, $\|F\|_\infty \leq u\sigma[m + (\tau+1)m^2]$.

If partial pivoting (by rows) is used, then $\tau = 1$ and $e_{ij} = 0$ for $j > i+2m$. Then, for $2m \leq j \leq n-2m$, $\sum_{i=j+1}^n e_{ij} = \sum_{i=j+1}^{j+m} 1 = m$,

$$\sum_{i=2}^j \sum_{k=1}^{i-1} e_{ik} e_{kj} = \sum_{i=j-2m}^{j-m-1} (i-j+2m) + \sum_{i=j-m}^j m = \left(\sum_{r=0}^{m-1} r \right) + m(m+1) = \frac{1}{2} m(m-1) + m(m+1) ,$$

$$\text{and } \sum_{i=j+1}^n \sum_{k=1}^{j-1} e_{ik} e_{kj} = \sum_{i=j+1}^{j+m} \sum_{k=j-m}^{j-1} e_{ik} e_{kj} = \sum_{i=j+1}^{j+m} (m-i+j) = \frac{1}{2} m(m-1) .$$

Thus, $\|F\|_1 \leq u\sigma[4m^2 + m]$. Similarly, $\|F\|_\infty \leq u\sigma[4m^2 + m]$.

We summarize this as

Then $F = \sum_{k=1}^{n-1} F^{(k)}$, and we conclude with

Theorem 15. In finite precision arithmetic $M+F = LU$, where F is given by:

$$(6.7) \quad |F_{ij}| \leq \begin{cases} u(\tau+1)\sigma \sum_{k=1}^{i-1} e_{ik}e_{kj} & \text{for } i \leq j \\ u(\tau+1)\sigma \sum_{k=1}^{j-1} e_{ik}e_{kj} + u\sigma e_{ij} & \text{for } i > j \end{cases},$$

and hence

$$(6.8) \quad \|F\|_1 \leq u\sigma \max_{1 \leq j \leq n} \left\{ \sum_{i=j+1}^n e_{ij} + (\tau+1) \left[\sum_{i=2}^j \sum_{k=1}^{i-1} e_{ik}e_{kj} + \sum_{i=j+1}^n \sum_{k=1}^{j-1} e_{ik}e_{kj} \right] \right\}$$

and

$$(6.9) \quad \|F\|_{\infty} \leq u\sigma \max_{1 \leq i \leq n} \left\{ \sum_{j=1}^{i-1} e_{ij} + (\tau+1) \left[\sum_{j=1}^{i-1} \sum_{k=1}^{j-1} e_{ik}e_{kj} + \sum_{j=i}^n \sum_{k=1}^{i-1} e_{ik}e_{kj} \right] \right\}.$$

When M is full, $\sum_{i=j+1}^n e_{ij} = n-j$, $\sum_{i=2}^j \sum_{k=1}^{i-1} e_{ik}e_{kj} = \sum_{i=2}^j (i-1) =$

$\frac{1}{2} j(j-1)$, and $\sum_{i=j+1}^n \left(\sum_{k=1}^{j-1} e_{ik} \right) e_{kj} = \sum_{i=j+1}^n (j-1) = (n-j)(j-1)$;

so $\|F\|_1 \leq u\sigma \max_{1 \leq j \leq n} \{n-j + (\tau+1)[(n - \frac{1}{2} j)(j-1)]\} \leq u\sigma\{n-1 + \frac{1}{2} (\tau+1)n(n-1)\}$.

Similarly, $\sum_{j=1}^{i-1} e_{ij} = i-1$, $\sum_{j=1}^{i-1} \sum_{k=1}^{j-1} e_{ik}e_{kj} = \sum_{j=1}^{i-1} (j-1) = \frac{1}{2} (i-1)(i-2)$, and

$\sum_{j=i}^n \sum_{k=1}^{i-1} e_{ik}e_{kj} = \sum_{j=i}^n (i-1) = (n-i+1)(i-1)$; so

$$\|F\|_{\infty} \leq u\sigma \max_{1 \leq i \leq n} \left\{ i-1 + (\tau+1) \left[\frac{1}{2} (i-1)(i-2) + (n-i+1)(i-1) \right] \right\} \leq$$

$u\sigma\{n-1 + \frac{1}{2} (\tau+1)n(n-1)\}$.

Thus, we have

Corollary 7. When M is a full $n \times n$ matrix,

$$\|F\|_1, \|F\|_\infty \leq u_\sigma[n-1 + \frac{1}{2}(\tau+1)n(n-1)] .$$

When M is a band matrix with bandwidth $2m+1$, $e_{ij} = 0$ for $i > j+m$.

If no pivoting is needed, then $e_{ij} = 0$ for $j > i+m$. Then,

$$\text{for } m \leq j \leq n-m, \quad \sum_{i=j+1}^n e_{ij} = \sum_{i=j+1}^{j+m} 1 = m, \quad \sum_{i=2}^j \sum_{k=1}^{i-1} e_{ik} e_{kj} =$$

$$\sum_{i=j-m}^j \sum_{k=j-m}^{i-1} e_{ik} e_{kj} = \sum_{i=j-m}^j (i-j+m) = \sum_{r=0}^m r = \frac{1}{2} m(m+1), \text{ and}$$

$$\sum_{i=j+1}^n \sum_{k=1}^{j-1} e_{ik} e_{kj} = \sum_{i=j+1}^{j+m} \sum_{k=j-m}^{j-1} e_{ik} e_{kj} = \sum_{i=j+1}^{j+m} (m-i+j) = \sum_{r=0}^{m-1} r = \frac{1}{2} m(m-1) .$$

So $\|F\|_1 \leq u_\sigma[m + (\tau+1)m^2]$. Similarly, $\|F\|_\infty \leq u_\sigma[m + (\tau+1)m^2]$.

If partial pivoting (by rows) is used, then $\tau = 1$ and $e_{ij} = 0$

$$\text{for } j > i+2m. \text{ Then, for } 2m \leq j \leq n-2m, \quad \sum_{i=j+1}^n e_{ij} = \sum_{i=j+1}^{j+m} 1 = m,$$

$$\sum_{i=2}^j \sum_{k=1}^{i-1} e_{ik} e_{kj} = \sum_{i=j-2m}^{j-m-1} (i-j+2m) + \sum_{i=j-m}^j m = \left(\sum_{r=0}^{m-1} r \right) + m(m+1) = \frac{1}{2} m(m-1) + m(m+1),$$

$$\text{and } \sum_{i=j+1}^n \sum_{k=1}^{j-1} e_{ik} e_{kj} = \sum_{i=j+1}^{j+m} \sum_{k=j-m}^{j-1} e_{ik} e_{kj} = \sum_{i=j+1}^{j+m} (m-i+j) = \frac{1}{2} m(m-1) .$$

Thus, $\|F\|_1 \leq u_\sigma[4m^2 + m]$. Similarly, $\|F\|_\infty \leq u_\sigma[4m^2 + m]$.

We summarize this as

Corollary 8. When M is an $n \times n$ band matrix with bandwidth $2m+1$, then

$$\|F\|_1, \|F\|_\infty \leq u\sigma[m + (\tau+1)m^2]$$

when no pivoting is needed, and

$$\|F\|_1, \|F\|_\infty \leq u\sigma[4m^2 + m]$$

when partial pivoting (by rows) is used.

Note that the bounds once again are independent of n and depend on the bandwidth.

When M is symmetric and $LDL^t = M+F$, then $F = F^t$ and $e_{ij} = e_{ji}$ for all i, j . Then $\sum_{k=1}^{i-1} e_{ik} e_{ki} = \sum_{k=1}^{i-1} e_{ik}$, and (6.7) becomes:

$$(6.10) \quad |F_{ij}| \leq \begin{cases} u(\tau+1)\sigma \sum_{k=1}^{i-1} e_{ik} e_{jk} + u\sigma e_{ij} & \text{for } i < j \\ u(\tau+1)\sigma \sum_{k=1}^{i-1} e_{ik} & \text{for } i = j \\ u(\tau+1)\sigma \sum_{k=1}^{j-1} e_{ik} e_{jk} + u\sigma e_{ij} & \text{for } i > j \end{cases}$$

Theorem 16. When $M = M^t$, then $LDL^t = M+F$ and $\|F\|_1 = \|F\|_\infty \leq$

$$u\sigma \max_{1 \leq i \leq n} \left\{ \sum_{\substack{j=1 \\ j \neq i}}^n e_{ij} + (\tau+1) \left[\sum_{j=1}^{i-1} \sum_{k=1}^{j-1} e_{ik} e_{jk} + \sum_{k=1}^{i-1} e_{ik} + \sum_{j=i+1}^n \sum_{k=1}^{i-1} e_{ik} e_{jk} \right] \right\}.$$

We then have immediately

Corollary 9. When M is symmetric and full,

$$\|F\|_1 = \|F\|_\infty \leq u_\sigma \left\{ n - 1 + \frac{1}{2} (\tau+1)n(n-1) \right\}$$

and

Corollary 10. When M is an $n \times n$ symmetric band matrix with bandwidth $2m+1$,

$$\|F\|_1 = \|F\|_\infty \leq u_\sigma \{ m + (\tau+1)m^2 \} .$$

7. Error Analysis of Elimination

Now we can combine the analyses in Sections 4-6. From Section 4, when attempting to solve $Mz = b$ we obtain the exact solution x of $(M + \delta M)x = b$, where

$$(7.1) \quad \delta M = F + (\delta L)U + L(\delta U) + (\delta L)(\delta U) .$$

Taking norms, we have

$$(7.2) \quad \|\delta M\| \leq \|F\| + \|\delta L\| \|U\| + \|L\| \|\delta U\| + \|\delta L\| \|\delta U\| .$$

Since L is unit lower triangular,

$$(7.3) \quad \|L\|_1 = \max_{1 \leq j \leq n} \sum_{i=j}^n |l_{ij}| \leq 1 + \tau \max_{1 \leq j \leq n-1} c_j ,$$

since $\tau = \max_{i,j} |l_{ij}|$ and c_j is the number of nonzero elements in the j -th column of L off the diagonal.

For the ∞ -norm we have

$$(7.4) \quad \|L\|_{\infty} \leq 1 + \tau \max_{2 \leq i \leq n} \sum_{j=1}^{i-1} e_{ij} \leq 1 + \tau(n-1) .$$

Since U is nonsingular and upper triangular,

$$(7.5) \quad \|U\|_1 \leq \sigma(1 + \max_{2 \leq j \leq n} \sum_{i=1}^{j-1} e_{ij}) \leq \sigma n ,$$

where $\sigma = \max_{i,j} |u_{ij}|$.

Since r_j is the number of nonzeros in the j -th row of U off the diagonal,

$$(7.6) \quad \|U\|_{\infty} \leq \sigma(1 + \max_{1 \leq j \leq n-1} r_j) .$$

We now have bounds on all the norms in (7.2) for the one-norm and the ∞ -norm, and thus can bound $\|\delta M\|_1$ and $\|\delta M\|_\infty$. However, we shall consider in detail in the next section only the special cases of full and band matrices.

When M is symmetric, from Section 4 we have

$$(7.7) \quad \delta M = F + LD(\Delta L^t) + [L(\delta D) + (\delta L)D + (\delta L)(\delta D)][L^t + \Delta L^t] .$$

Thus, taking norms, we have

$$(7.8) \quad \|\delta M\| \leq \|F\| + \|L\| \|D\| \|\Delta L^t\| + [\|L\| \|\delta D\| + \|\delta L\| \|D\| + \|\delta L\| \|\delta D\|] [\|L^t\| + \|\Delta L^t\|] .$$

Since L is unit lower triangular,

$$(7.9) \quad \|L^t\|_1 = \|L\|_\infty \leq 1 + \tau \max_{2 \leq i \leq n} \sum_{j=1}^{i-1} e_{ij} \leq 1 + \tau(n-1)$$

and

$$(7.10) \quad \|L^t\|_\infty = \|L\|_1 \leq 1 + \tau \max_{1 \leq j \leq n-1} d_j .$$

Thus, we could now bound (7.8) when $M = M^t$.

8. The Full and Band Matrix Cases

When M is an $n \times n$ full matrix,

$$(8.1) \quad \|L\|_1, \|L\|_\infty \leq 1 + \tau(n-1)$$

and

$$(8.2) \quad \|U\|_1, \|U\|_\infty \leq \sigma n.$$

From Corollaries 1, 3, 7 and (7.2), (8.1), and (8.2) we obtain

Theorem 17. When M is an $n \times n$ full matrix,

$$(8.3) \quad \|\delta M\|_1 \leq 1.01 u \sigma \frac{1}{4} [3\tau n^3 + (\tau+3)n^2 + (\tau+4)n + 1 + 1.01 u \tau \frac{1}{2} (n^4 + n^3 + 3n^2 - 5n)]$$

and

$$(8.4) \quad \|\delta M\|_\infty \leq 1.01 u \sigma [\tau n^3 + (\tau+1)n^2 + (1-2\tau)n - 1 + 1.01 u \tau \frac{1}{4} (n^4 + 2n^3 - n^2 - 2n)] .$$

With partial pivoting, $\tau = 1$ and $\sigma \leq 2^{n-1} \max_{i,j} |M_{ij}|$. With

complete pivoting, $\tau = 1$ and $\sigma \leq \sqrt{n} f(n) \max_{i,j} |M_{ij}|$, where

$$f(n) = \left(\prod_{k=2}^n k^{\frac{1}{k-1}} \right)^{1/2} \leq 1.8 n^{\frac{1}{4}} \log n \quad ([3], \text{ p. 108}).$$

If M is (column) diagonally dominant, i.e., $|M_{jj}| \geq \sum_{\substack{i=1 \\ i \neq j}}^n |M_{ij}|$

for $1 \leq j \leq n$, then $\tau = 1$ and $\sigma \leq 2$ (Wilkinson [7]).

We state the above in

Corollary 11. Let M be an $n \times n$ full matrix. If M is (column) diagonally dominant or if partial or complete pivoting is used, then

$\tau = 1$ and

$$\|\delta M\|_1 \leq 1.01 u \sigma \left[\frac{3}{4} n^3 + n^2 + \frac{5}{4} n + \frac{1}{4} + 1.01 u \frac{1}{8} (n^4 + n^3 + 3n^2 - 5n) \right]$$

and

$$\|\delta M\|_\infty \leq 1.01 u \sigma [n^3 + 2n^2 - 3n - 1 + 1.01 u \frac{1}{4} (n^4 + 2n^3 - n^2 - 2n)] \dots$$

When M is an $n \times n$ band matrix with bandwidth $2m+1$,

$$(8.5) \quad \|L\|_1, \|L\|_\infty \leq 1 + \tau m$$

and

$$(8.6) \quad \|U\|_1, \|U\|_\infty \leq \begin{cases} \sigma(m+1) & \text{when no pivoting is needed,} \\ \sigma(2m+1) & \text{when partial pivoting is used.} \end{cases}$$

From Corollaries 2, 4, 8 and (7.2), (8.5), and (8.6) we have

Theorem 18. When M is a band matrix with bandwidth $2m+1$, and if no pivoting is needed, then

$$(8.7) \quad \|\delta M\|_1, \|\delta M\|_\infty \leq 1.01 u \sigma \left\{ \tau m^3 + \frac{1}{2} (9\tau+3)m^2 + \frac{5}{2} (\tau+1)m + \tau+1 + 1.01 u \tau \left(\frac{1}{4} m^4 + \frac{3}{2} m^3 + \frac{11}{4} m^2 + \frac{3}{2} m \right) \right\}.$$

If M is also (column) diagonally dominant, then no pivoting is needed and $\tau = 1$, $\sigma \leq 2$; giving

Corollary 12. If M is a (column) diagonally dominant band matrix with bandwidth $2m+1$, then

$$(8.8) \quad \|\delta M\|_1, \|\delta M\|_\infty \leq 1.01 u \left\{ 2m^3 + 12m^2 + 10m + 4 + 1.01 u \left(\frac{1}{2} m^4 + 3m^3 + \frac{11}{2} m^2 + 3m \right) \right\}.$$

If partial pivoting is used for a band matrix, then $\tau = 1$, and $\sigma \leq 2^{2m-1} - (m-1)2^{m-2}$ (Wilkinson [6]). Note that the bound on σ depends

only on the bandwidth and is independent of the order n of the matrix.

From Corollaries 2, 4, 8 and (7.2), (8.5) and (8.6) we have

Theorem 19. When partial pivoting is used on a band matrix M with bandwidth $2m+1$,

$$(8.9) \quad \|\delta M\|_1, \|\delta M\|_\infty \leq 1.01 u_\sigma \left[3m^3 + \frac{25}{2}m^2 + \frac{13}{2}m + 1 + 1.01 u \left(m^4 + \frac{5}{2}m^3 + 2m^2 + \frac{1}{2}m \right) \right] .$$

Note that in Corollaries 12-14 the bounds depend on the bandwidth and not on the order of the matrix.

Using (7.8), we can obtain similar bounds for $\|\delta M\|_1$ and $\|\delta M\|_\infty$ when M is a full or band symmetric matrix.

9. Remarks

From Sections 5-8 we see that the error matrix δM arising from performing the elimination process in finite precision depends on the fill-in occurring during the elimination. We could seek an ordering of the equations so that the bound on $\|\delta M\|$ is minimized. This would not be equivalent to the seeking of an ordering to minimize the fill-in. However, we see that minimizing fill-in helps to keep the bound on $\|\delta M\|$ from becoming too large. The problem is even more difficult if we need to pivot for stability. For a further discussion of these ordering problems, see Bunch [1]. However, given a fixed ordering, our analysis here bounds the errors occurring in the computation.

Acknowledgment

The author wishes to thank Mr. David M. Gay for his careful reading of the manuscript and the referee for helpful suggestions.

References

- [1] Bunch, J. R., "Complexity of sparse elimination", Complexity of Parallel and Sequential Numerical Algorithms, J. F. Traub, editor, Academic Press, 1973.
- [2] Bunch, J. R., and D. J. Rose, "Partitioning, tearing, and modification of sparse linear systems", Linear Algebra and its Applications, submitted.
- [3] Forsythe, G. E., and C. B. Moler, Computer Solution of Linear Algebraic Systems, Prentice-Hall, 1967.
- [4] Rose, D. J., "A graph-theoretic study of the numerical solution of sparse positive definite systems of linear equations", Graph Theory and Computing, R. Read, editor, Academic Press, New York, 1972.
- [5] Rose, D. J., and J. R. Bunch, "The role of partitioning in the numerical solution of sparse systems", Sparse Matrices and their Applications, D. J. Rose and R. A. Willoughby, editors, Plenum Press, 1972.
- [6] Wilkinson, J. H., private communication.
- [7] Wilkinson, J. H., The Algebraic Eigenvalue Problem, Clarendon Press, 1965.