Backward Error Analysis for Gaussian Elimination

A. K. Cline
Department of Computer Sciences
University of Texas at Austin
Austin, Texas 78712

The purpose of this note is to show that the Gaussian elimination algorithm when performed on a matrix A with floating point arithmetic can be shown to result in the form

$$LU = A + E$$

where L is unit, lower triangular, U is upper triangular, and E is a matrix of perturbations. Formally written, the algorithm is

$$a^1_{i,j} = a_{i,j}, \text{for } i,\ j = 1,...n \tag{1}$$

and for $k = 1,...,n-1$

$$u_{k,j} = a^k_{k,j}, \text{for } j = k,...,n. \tag{2}$$

$$l_{i,k} = a^k_{i,k} / u_{k,k}, \tag{3}$$

and

$$a^{k+1}_{i,j} = a^k_{i,j} - l_{i,k}u_{k,j}. \tag{4}$$

Since floating point arithmetic is being employed, equations (3) and (4) are actually

$$l_{i,k} = fl(a^k_{i,k} / u_{k,k}), \tag{5}$$

and

$$a^{k+1}_{i,j} = fl(a^k_{i,j} - l_{i,k}u_{k,j}). \tag{6}$$

(Notice that the floating point arithmetic has no affect on equation (2) .) Using the model of errors introduced by this arithmetic, these equations could be recast as

$$l_{i,k} = a^k_{i,k} / u_{k,k}(1+\delta_1), \tag{7}$$

and

$$a^{k+1}_{i,j} = (a^k_{i,j} - l_{i,k}u_{k,j}(1+\delta_2))(1+\delta_3). \tag{8}$$

where $|\delta_1|, |\delta_2|, \text{and } |\delta_3| \le \varepsilon_0$, the machine unit floating point precision. (The dependency of the δ s upon i, j, and k has been omitted from the notation for clarity.)

Consider the error in "backward fashion" (i.e., the operations were performed correctly on perturbed data), equations (7) and (8) could be stated as

$$l_{i,k} = (a^k_{i,k} + e^k_{i,k}) / u_{k,k}, \tag{9}$$

and

$$a^{k+1}_{i,j} = a^k_{i,j} + e^k_{i,j} - l_{i,k}u_{k,j}. \tag{10}$$

These equations actually define the perturbations $e^k_{i,k}$ and $e^k_{i,j}$, thus

$$e^k_{i,k} = l_{i,k}u_{k,k} - a^k_{i,k}, \tag{11}$$

and

$$e^k_{i,j} = l_{i,k}u_{k,j} - a^k_{i,j} + a^{k+1}_{i,j}. \tag{12}$$

Equating (7) and (9), we obtain

$$e_{i,k}^{k} = \delta_1 a_{i,k}^{k}, \tag{13}$$

and we may conclude that

$$|e_{i,k}^{k}| \le \varepsilon_0 |a_{i,k}^{k}|. \tag{14}$$

Notice from (8) that

$$l_{i,k}u_{k,j} = (a_{i,j}^{k} - a_{i,j}^{k+1}/(1+\delta_3))/(1+\delta_2), \tag{15}$$

so (12) could be rewritten as

$$e_{i,j}^{k} = a_{i,j}^{k}(1/(1+\delta_2)-1) - a_{i,j}^{k+1}(1/((1+\delta_3)(1+\delta_2))-1). \tag{16}$$

Ignoring terms of size ε_0^2, this is

$$e_{i,j}^{k} = a_{i,j}^{k}(-\delta_2) - a_{i,j}^{k+1}(-\delta_2-\delta_3), \tag{17}$$

so

$$|e_{i,j}^{k}| \le \varepsilon_0 |a_{i,j}^{k}| + 2\varepsilon_0 |a_{i,j}^{k+1}| \le 3\varepsilon_0 \max\{|a_{i,j}^{k}|, |a_{i,j}^{k+1}|\}. \tag{18}$$

If we sum equation (10) over values of k from 1 through $i-1$ and assume $i \le j$, we obtain

$$a_{i,j}^{2} + \ldots + a_{i,j}^{i} = a_{i,j}^{1} + \ldots + a_{i,j}^{i-1} + (e_{i,j}^{1} + \ldots + e_{i,j}^{i-1}) - (l_{i,1}u_{1,j} + \ldots + l_{i,i-1}u_{i-1,j}), \tag{19}$$

and after canceling identical terms

$$a_{i,j}^{i} = a_{i,j}^{1} + (e_{i,j}^{1} + \ldots + e_{i,j}^{i-1}) - (l_{i,1}u_{1,j} + \ldots + l_{i,i-1}u_{i-1,j}), \tag{20}$$

Noticing that $a_{i,j}^{i} = u_{i,j} = l_{i,i}u_{i,j}$ (since the diagonal elements of L are one) and that $a_{i,j}^{1} = a_{i,j}$, we have

$$\sum_{k=1}^{i} l_{i,k}u_{k,j} = a_{i,j} + \varepsilon_{i,j}, \tag{21}$$

where $\varepsilon_{i,j} = \sum_{k=1}^{i-1} e_{i,j}^{k}$. Similarly for $j < i$, if we sum (10) for values of k from 1 through $j-1$, we obtain

$$a_{i,j}^{2} + \ldots + a_{i,j}^{j} = a_{i,j}^{1} + \ldots + a_{i,j}^{j-1} + (e_{i,j}^{1} + \ldots + e_{i,j}^{j-1}) - (l_{i,1}u_{1,j} + \ldots + l_{i,j-1}u_{j-1,j}). \tag{22}$$

Again, after canceling identical terms, we have

$$a_{i,j}^{j} = a_{i,j}^{1} + (e_{i,j}^{1} + \ldots + e_{i,j}^{j-1}) - (l_{i,1}u_{1,j} + \ldots + l_{i,j-1}u_{j-1,j}). \tag{23}$$

From (9) we have $l_{i,j}u_{j,j} = a_{i,j}^{j} + e_{i,j}^{j}$, and thus

$$\sum_{k=1}^{j} l_{i,k}u_{k,j} = a_{i,j} + \varepsilon_{i,j}, \tag{24}$$

where $\varepsilon_{i,j} = \sum_{k=1}^{j} e_{i,j}^{k}$.

Notice that for $i \le j$, from (18)

$$|\varepsilon_{i,j}| \le \sum_{k=1}^{i-1} |e_{i,j}^{k}| \le 3(i-1)\varepsilon_0 \max\{|a_{i,j}^{1}|, \ldots. |a_{i,j}^{i}|\}, \tag{25}$$

and for $j < i$, from (18) and (14)

$$|e_{i,j}| \le \sum_{k=1}^{j} |e_{i,j}^{k}| \le 3(j-1)e_0 \max\{|a_{i,j}^{1}|,....|a_{i,j}^{j}|\} + e_0 |a_{i,j}^{j}|. \tag{26}$$

Thus, in general

$$|e_{i,j}| \le 3e_0 \min\{i-1, j\} \max\{|a_{i,j}^{1}|,....|a_{i,j}^{\min\{i,j\}}|\}, \tag{27}$$

for $i, j = 1,...,n$.

A slightly different approach results in the production of the factorization

$$L^{n-1} \cdots L^2 L^1 (A+E) = U$$

where $A, E,$ and U are exactly as before and $L^{n-1} \cdots L^2 L^1$ is inverse of L. This variant of the algorithm is more efficient when row interchanges are introduced since only portions of rows are interchanged here whereas entire rows must be interchanged with the LU factorization. (As before however the actual permutations will be ignored to avoid the great complexity they add to the notation.)

Equation (9) could be also be stated as

$$l_{i,k} a_{k,k}^{k} = a_{i,k}^{k} + e_{i,k}^{k} \tag{28}$$

which is

$$-l_{i,k} a_{k,k}^{k} + 1(a_{i,k}^{k} + e_{i,k}^{k}) = 0 \tag{29}$$

and equation (10) could be stated as

$$-l_{i,k} a_{k,j}^{k} + 1(a_{i,j}^{k} + e_{i,j}^{k}) = a_{i,j}^{k+1}. \tag{30}$$

By defining

$$a_{i,j}^{k+1} = a_{i,j}^{k} \tag{31}$$

for $i = 1,...,k$ and $j = 1,...,n$ and for $i = 1,...,n$ and $j = 1,...,k$, and

$$e_{i,j}^{k} = 0 \tag{32}$$

for $i = 1,...,k-1$ and $j = 1,...,n$ and for $i = 1,...,n$ and $j = 1,...,k$, and constructing the lower triangular matrix L^k with unit diagonal, $-l_{i,k}$ in the i,k position, and zeroes elsewhere, we have

$$L^k (A^k + E^k) = A^{k+1}, \tag{33}$$

for $k = 1,...,n-1$. Using the fact that

$$L^{k-1} \cdots L^2 L^1 E^k = E^k \tag{34}$$

(which follows from the structure of the L^i matrices) a simple induction argument shows that

$$L^k \cdots L^2 L^1 (A^1 + E^1 + \cdots + E^k) = A^{k+1} \tag{35}$$

for $k = 1,...,n-1$. By defining $E = E^1 + \cdots + E^{n-1}$ (which is the same E as in the previous analysis), we have

$$L^{n-1} \cdots L^2 L^1 (A+E) = A^n, \tag{36}$$

but from (2)and (30) we see that $U = A^n$ and this is the factorization we sought.