

Gaussian Elimination and LU-Decomposition

Gary D. Knott
Civilized Software Inc.
12109 Heritage Park Circle
Silver Spring MD 20906
phone:301-962-3711
email:knott@civilized.com
URL:www.civilized.com

March 9, 2010

1 Gaussian Elimination

Solving a set of linear equations arises in many contexts in applied mathematics. Indeed, a claim can be made that solving sets of linear equations (generally as a component of dealing with larger problems like partial-differential-equation solving, or optimization, consumes more computer time than any other computational procedure. (Competitors would be the Gram-Schmidt process and the fast Fourier transform computation.)

Often the subject of linear algebra is approached by starting with the topic of solving linear equations, and Gaussian elimination methodology is elaborated to introduce matrix inverses, rank, nullspaces, etc.

We have seen above that computing a preimage vector $x \in \mathcal{R}^n$ of a vector $v \in \mathcal{R}^k$ with respect to the $n \times k$ matrix A consists of finding a solution (x_1, \dots, x_n) to the k linear equations:

$$\begin{aligned}A_{11}x_1 + A_{21}x_2 + \cdots + A_{n1}x_n &= v_1 \\A_{12}x_1 + A_{22}x_2 + \cdots + A_{n2}x_n &= v_2 \\&\vdots \\A_{1k}x_1 + A_{2k}x_2 + \cdots + A_{nk}x_n &= v_k.\end{aligned}$$

This corresponds to $xA = v$.

If $v \in \mathcal{R}^k - \text{rowspace}(A)$ then there are no solutions x ; the equations $xA = v$ are inconsistent. For example, $[1x_1 + 2x_2 = 0, 2x_1 + 4x_2 = 1]$.

If $\text{nullspace}(A) = \{0\}$ and $v \in \text{rowspace}(A)$, then there is a unique solution $x = vA^+$, where the $k \times n$ matrix A^+ is the Moore-Penrose pseudo-inverse matrix of A . Necessarily $k \geq n$; in case $n = k$, A is non-singular and $A^+ = A^{-1}$ so $x = vA^{-1}$. The vector vA^+ always belongs to $\text{colspace}(A)$ regardless of the choice of v or the dimension of $\text{nullspace}(A)$.

If $\dim(\text{nullspace}(A)) > 0$ and $v \in \text{rowspace}(A)$, there is a $\dim(\text{nullspace}(A))$ -dimensional flat of solutions x . The vectors in $\text{nullspace}(A) + vA^+ \subseteq \mathcal{R}^n$ comprise all the solution vectors, x , that satisfy $xA = v$.

In general, the matrix A^+A is the $k \times k$ projection matrix onto $\text{rowspace}(A) \subseteq \mathcal{R}^k$, and for any vector $v \in \mathcal{R}^k$, the vector vA^+ is the unique vector in $\text{colspace}(A)$ such that $|v - vA^+A|$ is minimal; moreover vA^+ is the shortest minimizing vector in \mathcal{R}^n .

We often wish to determine which of these cases hold for a given $n \times k$ matrix A and a given right-hand-side vector $v \in \mathcal{R}^k$, and when $v \in \text{rowspace}(A)$, we wish to compute the solution vector $x = vA^+$ without the expense of computing the Moore-Penrose pseudo-inverse matrix A^+ . The classic step-wise approach to computing x is to add a multiple of one equation to another at each step until the system of equations is reduced to a form which is easy to either solve or to see that no solution or no unique solution exists. This process is called *Gaussian elimination*, since we generally aim to eliminate successive variables from successive equations by simple algebraic modifications as was proceduralized by C. F. Gauss.

The form that is most commonly sought is a *triangular* system of equations. We will usually only need to deal with such a triangular system in the case where $n = k$ and we have a unique solution vector x , *i.e.* where we have a consistent system of equations with $n = k$, and the matrix of coefficients is non-singular. In this case, we can obtain:

$$\begin{array}{rcccccc} L_{11}x_1 & + & L_{21}x_2 & + & \cdots & + & L_{n-1,1}x_{n-1} & + & L_{n1}x_n & = & y_1 \\ & & L_{22}x_2 & + & \cdots & + & L_{n-1,2}x_{n-1} & + & L_{n2}x_n & = & y_2 \\ & & & & & & \vdots & & & & \\ & & & & & & L_{n-1,n-1}x_{n-1} & + & L_{n,n-1}x_n & = & y_{n-1} \\ & & & & & & & & L_{nn}x_n & = & y_n \end{array}$$

This corresponds to $xL = y$ where L is an $n \times n$ lower-triangular matrix. Let us assume there is a unique solution, so L must be non-singular, and thus L_{ii} is necessarily non-zero for $i = 1, \dots, n$. If we have such a triangular form, then we have one equation involving just x_n , one involving just x_n and x_{n-1} , and so on, and it is easy to compute the solution vector x . The process of computing the solution is called “back-substitution.” An algorithm for back-substitution is given below.

[for $i = n, n-1, \dots, 1$: ($x_i \leftarrow y_i$; for $j = i+1, \dots, n$: ($x_i \leftarrow x_i - L_{ji}x_j$); $x_i \leftarrow x_i/L_{ii}$)].

Exercise 1.1: State the back-substitution algorithm that applies when we have a non-singular $n \times n$ upper-triangular matrix U with $xU = y$, so that we have one equation involving just x_1 , one involving just x_1 and x_2 , and so on.

In matrix form, adding a multiple of one equation to another consists of adding a multiple of one column of A to another, and at the same time adding that multiple of the same (single element)

column of v to the other corresponding (single element) column of v . (We manipulate columns rather than rows because we take vectors to be rows rather than columns and we apply matrices to vectors by multiplying on the right.) This can be nicely organized, when desired, by appending v to the matrix A as an additional row.

Adding a multiple of one column of a matrix to another column can be effected by multiplying on the right by a suitable *elementary* matrix E . Define $E_k[i, j, \alpha]$ to be the $k \times k$ matrix $I + \alpha e_j^T e_i$ where e_j is the k -vector $(0, \dots, 0, 1, 0, \dots, 0)$ with each component equal to 0 except component j which is 1. Now, for any k -column matrix A , $AE_k[i, j, \alpha]$ is the same matrix as A except that column i is replaced by $(A \text{ col } i) + \alpha(A \text{ col } j)$.

Exercise 1.2: Show that $e_j^T e_i$ is the $k \times k$ matrix each of whose elements is 0 except component $[j, i]$ which is 1.

Exercise 1.3: Show that $E_k[i, j, \alpha]^T = E_k[j, i, \alpha]$. Thus the transpose of an elementary matrix is an elementary matrix.

Exercise 1.4: A suitable conformable elementary matrix can also be used to add a multiple of a row of an $n \times m$ matrix A to another row of A . Show that $E_n[i, j, \alpha]A$ is the same matrix as A except that row j is replaced by $(A \text{ row } j) + \alpha(A \text{ row } i)$.

Exercise 1.5: Show that $E_k[i, i, \alpha - 1] = I$ except the $[i, i]$ element is α . Show that $B = AE_k[i, i, \alpha - 1]$ has the same columns as A except $B \text{ col } i = \alpha(A \text{ col } i)$.

Exercise 1.6: Show that, if $\alpha = 0$ or $i \neq j$ then $E_k[i, j, \alpha]^{-1} = E_k[i, j, -\alpha]$, if $\alpha \neq -1$ and $i = j$ then $E_k[i, j, \alpha]^{-1} = E_k[i, i, -\alpha/(1 + \alpha)]$, and if $\alpha = -1$ or $i = j$ then $E_k[i, j, \alpha]$ is singular.

Exercise 1.7: Let the $k \times k$ matrix $S = E_k[i, j, -1]E_k[j, i, 1]E_k[i, j, -1]E_k[i, i, -2]$, where $1 \leq i \leq k$ and $1 \leq j \leq k$ with $i \neq j$. Show that $B = AS$ has the same columns as A except $B \text{ col } i = A \text{ col } j$ and $B \text{ col } j = A \text{ col } i$, where $i \neq j$. What does right-multiplication by S do if $i = j$?

Recall that $transpose_k(i, j)$ denotes the permutation $\langle 1, \dots, j, \dots, i, \dots, k \rangle$ where component $t = t$, except component $i = j$ and component $j = i$. The $k \times k$ column permutation matrix corresponding to $transpose_k(i, j)$ is the matrix $I \text{ col } transpose_k(i, j)$; from the exercise above, this is:

$E_k[i, j, -1]E_k[j, i, 1]E_k[i, j, -1]E_k[i, i, -2]$ when $i \neq j$. The $n \times n$ row permutation matrix corresponding to $transpose_n(i, j)$ is the matrix $I \text{ row } transpose_n(i, j)$; when $i \neq j$, this is the matrix $(E_n[i, j, -1]E_n[j, i, 1]E_n[i, j, -1]E_n[i, i, -2])^T$. When $i = j$, $I \text{ col } transpose_k(i, i) = E_k[i, i, 0]$ and $I \text{ row } transpose_n(i, i) = E_n[i, i, 0]$.

Note that since transposition permutation matrices can be expressed as a product of elementary matrices and every permutation can be expressed as a composition of transpose permutations, any permutation matrix can be expressed as a product of elementary matrices.

It is convenient to define the $k \times k$ matrix $G_k[i, w] = I + e_i^T w$, where $w \in \mathcal{R}^k$. The matrix $G_k[i, w]$ is called a *Gauss* matrix. Note $(AG_k[i, w]) \text{ col } j = (A \text{ col } j) + w_j(A \text{ col } i)$ for $1 \leq j \leq k$ where $colsize(A) = k$.

Exercise 1.8: Show that $G_k[i, w] = E_k[1, i, w_1]E_k[2, i, w_2] \cdots E_k[k, i, w_k]$.

Exercise 1.9: Show that the matrices $E_k[1, i, w_1], \dots, E_k[i-1, i, w_{i-1}], E_k[i, i, w_i], E_k[i+1, i, w_{i+1}], \dots, E_k[k, i, w_k]$ commute with one-another.

Exercise 1.10: Show that if $w_i = 0$, then $G_k[i, w]^{-1} = G_k[i, -w] = I - e_i^T w$.

Exercise 1.11: Let A be a $k \times m$ matrix. Show that $(G_k[i, w]^T A)$ row $j = (A$ row $j) + w_i(A$ row $i)$.

Exercise 1.12: Let A be a $k \times k$ matrix with $A_{ri} \neq 0$.

Let $v = \left[\frac{-A_{r1}}{A_{ri}}, \dots, \frac{-A_{r,i-1}}{A_{ri}}, \frac{1 - A_{ri}}{A_{ri}}, \frac{-A_{r,i+1}}{A_{ri}}, \dots, \frac{-A_{rk}}{A_{ri}} \right]$. What is $(AG_k[i, v])$ row r ?

The matrix $G_k[h, w]$ can be used to convert the last $h-1$ components of a k -vector to 0, when component h of the vector is non-zero.

Let $a = (a_1, a_2, \dots, a_k)$ with $a_h \neq 0$. Take $w = (0, \dots, 0, 0, -a_{h+1}/a_h, -a_{h+2}/a_h, \dots, -a_k/a_h)$. Then $aG_k[h, w] = (a_1, a_2, \dots, a_h, 0, \dots, 0)$. We use only this form of Gauss matrix below; thus we shall henceforth consider only restricted Gauss matrices $G_k[h, w]$ where w col $(1 : h) = 0$.

Exercise 1.13: Take $s \in \{1, \dots, k\}$ and let v_1, \dots, v_s be vectors in \mathcal{R}^k such that (v_h) col $(1 : h) = 0$ for $h = 1, \dots, s$. Let $M_h = G_k[h, v_h]$ be the indicated restricted Gauss matrix. Show that $M_1 M_2 \cdots M_s = I + \sum_{1 \leq i \leq s} e_i^T v_i$ and also show that $(M_1 M_2 \cdots M_s)^{-1} = I - \sum_{1 \leq i \leq s} e_i^T v_i$. Hint: show that $(e_i^T v_i)(e_j^T v_j) = 0$ when $i > j$ (and (v_i) col $(1 : i) = 0$).

We can transform an $n \times k$ matrix A into a lower-triangular or diagonal form by multiplying by suitable permutation matrices and restricted Gauss matrices appropriately on the left and on the right of A . Gaussian elimination can be systematized and cast in a more general form by considering an associated matrix factorization called *LU-decomposition* [GV89]. Again, by multiplying by suitable permutation matrices and restricted Gauss matrices appropriately on the left and on the right of A , we can obtain the complete *LU-decomposition* of the $n \times k$ matrix A .

Let $r = \text{rank}(A)$. We will give an algorithm below based on the complete pivoting algorithm in [GV89] that determines $n \times n$ transposition permutation matrices R_1, \dots, R_r and $k \times k$ transposition permutation matrices C_1, \dots, C_r and $k \times k$ Gauss matrices M_1, \dots, M_r , and an $n \times k$ lower-triangular matrix L and a non-singular $k \times k$ upper-triangular matrix U such that

$$R_r \cdots R_1 A C_1 M_1 \cdots C_r M_r = L \quad \text{and} \quad R_r \cdots R_1 A C_1 \cdots C_r = LU.$$

The matrix L is an $n \times k$ lower-triangular matrix of the form $\begin{bmatrix} J & 0 \\ K & 0 \end{bmatrix}$, where J is an $r \times r$ non-singular lower-triangular matrix and K is an $(n-r) \times r$ matrix.

LU-Decomposition by Gaussian Elimination with Complete Pivoting:

input: A , $n \geq 1$, $k \geq 1$.

output: $L, U, b, c, r, R_1, \dots, R_r, C_1, \dots, C_r, M_1, \dots, M_r$

1. $L \leftarrow A; h \leftarrow 1; U \leftarrow I_{k \times k}; b \leftarrow \langle 1, 2, \dots, n \rangle; c \leftarrow \langle 1, 2, \dots, k \rangle$.
2. Determine indices $p \in \{h, \dots, n\}$ and $q \in \{h, \dots, k\}$ such that $|L_{pq}| = \max_{\substack{h \leq i \leq n \\ h \leq j \leq k}} |L_{ij}|$.
3. $a \leftarrow L_{pq}$; if $a = 0$ then ($r \leftarrow h - 1$; exit).
4. $b_h \leftarrow p; c_h \leftarrow q$.

$$\left[\begin{array}{l} \text{Let } u = \text{transpose}_k(h, q). \text{ Define } C_h = I \text{ col } u. \\ \text{Let } u = \text{transpose}_n(h, p). \text{ Define } R_h = I \text{ row } u. \end{array} \right]$$
5. Swap L col h and L col q in L ;
 Swap L row h and L row p in L .
 { Now $L_{hh} = a$. }
6. $\left\{ \begin{array}{l} \text{Subtract multiples of } L \text{ col } h \text{ from } L \text{ col } (h+1), L \text{ col } (h+2), \dots, L \text{ col } k \\ \text{to make } L \text{ row } h \text{ col } [(h+1) : k] = 0. \text{ Also compute } U \text{ row } h \text{ col } [(h+1) : k]. \end{array} \right\}$
 for $i = h+1, \dots, k$:
 $(z \leftarrow L_{hi}/a; L_{hi} \leftarrow 0; U_{hi} \leftarrow z; \text{ for } j = h+1, \dots, n : (L_{ji} \leftarrow L_{ji} - zL_{jh}))$.

$$\left[\begin{array}{l} \text{Let } w \text{ col } (1 : h) = 0 \text{ and } w \text{ col } ((h+1) : k) = -[L \text{ row } h \text{ col } ((h+1) : k)]/L_{hh}. \\ \text{Define } M_h = G_k[h, w]. \end{array} \right]$$
7. if $h = n$ or $h = k$ then ($r \leftarrow h$; exit);
 $h \leftarrow h + 1$; go to step 2.

At exit, this algorithm has determined the value r , the permutation matrices C_1, \dots, C_r and R_1, \dots, R_r , the Gauss matrices M_1, \dots, M_r , the lower-triangular matrix L , the upper-triangular matrix U , and the permutations b and c in transposition vector form.

Exercise 1.14: Show that the matrix products $R_r \cdots R_1$ and $M_1 \cdots M_r$ and $C_1 \cdots C_r$ are all products of elementary matrices.

Let $P = R_r \cdots R_1$ and let $B = C_1 M_1 \cdots C_r M_r$. Let $Q = C_1 \cdots C_r$. If $r = 0$, take $P = I$, $B = I$, and $Q = I$. Finally, let $U = B^{-1}Q$; this is the matrix U computed in the algorithm above.

The matrix P is an $n \times n$ permutation matrix, B is a $k \times k$ non-singular matrix, Q is a $k \times k$ permutation matrix, and the matrix U is a $k \times k$ non-singular upper-triangular matrix with $U_{ii} = 1$ for $i = 1, \dots, k$. Also, b represents the same n -permutation as the row-permutation matrix P , and c represents the same k -permutation as the column-permutation matrix Q .

Then $R_r \cdots R_1 A C_1 M_1 \cdots C_r M_r = L$, so $PAB = L$, so $PA = LB^{-1}$, so $PAQ = LB^{-1}Q$, so $PAQ = LU$.

$PAQ = LU$ is the so-called *LU-decomposition* for A . Given $PAQ = LU$, we may determine if the set of linear equations $xA = v$ is consistent, and if so, compute x such that $xA = v$ as follows.

Note $A = P^{-1}LUQ^{-1} = P^T LUQ^T$. Then $xA = v$ implies $xP^T LUQ^T = v$ implies $xP^T LU = vQ$. Let $y = xP^T L$. Then $yU = vQ$. Since U is non-singular and upper-triangular, we can compute y with back-substitution. Now recall $y = xP^T L$. Let $z = xP^T$. Then $y = zL$.

If $n = k$ and L is non-singular, we can compute z with back-substitution. Otherwise, recall L is an $n \times k$ lower-triangular matrix with $L = \begin{bmatrix} J & 0 \\ K & 0 \end{bmatrix}$ where J is an $r \times r$ non-singular lower-triangular matrix and K is an $(n-r) \times r$ matrix. If $y \text{ col } [(r+1) : k] \neq 0$, our equations are inconsistent and x does not exist. Otherwise we may take $z_{r+1} = z_{r+2} = \dots = z_n = 0$ and solve for z_1, \dots, z_r in $[z \text{ col } 1 : r]J = [y \text{ col } 1 : r]$ via back-substitution.

Finally, we must appropriately permute the components of z according to the the permutation matrix P to obtain $x = zP$. (In order to compute zP within z , we may use the following algorithm. [for $i = r, r-1, \dots, 1$: swap z_i with z_{b_i}]. Also, note that in order to compute vQ within v , we may use the algorithm: [for $i = 1, 2, \dots, r$: swap v_i with v_{c_i}].)

Exercise 1.15: What is computed in the Gaussian elimination LU-decomposition algorithm when the $n \times k$ matrix $A = 0$? What is computed when $k = 1$ and $A = e_n^T$? What is the result of the Gaussian elimination LU-decomposition algorithm when $n \geq k$? What is the result when $n < k$?

Exercise 1.16: What is computed in the Gaussian elimination LU-decomposition algorithm when the $n \times k$ matrix A is $\text{diag}(v_1, v_2, \dots, v_{\min(n,k)})$?

Exercise 1.17: Show that if $n = k$ and $r = k$, the linear equations $xA = v$ are necessarily consistent and have a unique solution.

Exercise 1.18: Why do we seek the largest magnitude element of L row $(h : n)$ col $(h : k)$ in step 2? Would determining *any* non-zero element of L row $(h : n)$ col $(h : k)$ suffice? Hint: think about the round-off error in computing a quotient b/a .

Exercise 1.19: Show that L is non-singular if and only if $n = k = r$ and $L = J$.

Exercise 1.20: Show that if $n = k$ then $\det(A) = L_{11} \cdot L_{22} \cdots L_{nn}$.

Exercise 1.21: Show that $R_t = I \text{ row } \text{transpose}_n(t, b_t)$ and $C_t = I \text{ col } \text{transpose}_k(t, c_t)$ where b and c are computed in the Gaussian elimination LU-decomposition algorithm. Also show that $C_t = C_t^T = C_t^{-1}$ and $R_t = R_t^T = R_t^{-1}$.

Exercise 1.22: Show that the transposition vectors b and c computed in the algorithm above determine the same permutations as the row permutation matrix $P = R_r \cdots R_1$ and the column permutation matrix $Q = C_1 \cdots C_r$ respectively. Thus the matrices P and Q need not be explicitly computed.

Solution 1.22: The matrix P is the $n \times n$ row permutation matrix I row p where the n -permutation p is determined by the transposition vector b via the algorithm: [$p \leftarrow \langle 1, 2, \dots, n \rangle$: for $i = 1, 2, \dots, r$: swap p_i with p_{b_i}], i.e. $p = (1 b_1)(2 b_2) \cdots (r b_r)$.

The matrix Q is the $k \times k$ column permutation matrix $I \text{ col } m$ where the k -permutation m is determined by the transposition vector c via the algorithm: $[m \leftarrow \langle 1, 2, \dots, k \rangle$: for $i = 1, 2, \dots, r$: swap m_i with m_{c_i}], i.e. $m = (1 \ c_1)(2 \ c_2) \cdots (r \ c_r)$.

Note in practical application, none of the matrices $C_1, \dots, C_r, R_1, \dots, R_r$, or M_1, \dots, M_r need to be computed. They are effectively replaced by b, c , and U . Thus none of the bracketed operations in the LU-decomposition algorithm need to be done.

Exercise 1.23: Show that $|U_{ij}| \leq 1$ for $1 \leq i < j \leq k$.

It remains to demonstrate that the matrix U computed in the Gaussian-elimination LU-decomposition algorithm is the same as the matrix $B^{-1}Q$. We have

$$B^{-1}Q = (C_1 M_1 C_2 M_2 \cdots C_r M_r)^{-1} C_1 C_2 \cdots C_r = M_r^{-1} C_r M_{r-1}^{-1} C_{r-1} \cdots C_2 M_1^{-1} C_1 C_1 C_2 \cdots C_r.$$

Recall that $C_i = C_i^{-1} = C_i^T$ is a $k \times k$ permutation matrix corresponding to a transposition $\text{transpose}_k(i, j)$ where $i \leq j \leq r$. Thus

$$B^{-1}Q = M_r^{-1}(C_r(M_{r-1}^{-1}(C_{r-1}(\cdots(C_3(M_2^{-1}(C_2 M_1^{-1} C_2))C_3))\cdots))C_r).$$

Now, let w_h be the k -vector such that $M_h = I + e_h^T w_h$. Recall that M_1, M_2, \dots, M_r are restricted Gauss matrices. Thus, for $h = 1, \dots, r$, $(w_h) \text{ col } 1 : h = 0$ and $(w_h) \text{ col } ((h+1) : k) = -[L \text{ row } h \text{ col } ((h+1) : k)]/L_{hh}$ as computed in step 6. Then $M_h^{-1} = I - e_h^T w_h$.

Now $C_2 M_1^{-1} C_2 = C_2(I - e_1^T w_1)C_2 = C_2 I C_2 - (C_2 e_1^T)(w_1 C_2) = I - e_1^T(w_1 C_2)$. This follows because the row permutation matrix C_2 exchanges row 2 with row j where $j \geq 2$, so that $C_2 e_1^T = e_1^T$. Finally, $C_2 = C_2^{-1}$, so $C_2 I C_2 = I$.

Also note the column permutation matrix C_2 exchanges column 2 with column j where $j \geq 2$, so that $w_1 C_2 \text{ col } 1 = (w_1) \text{ col } 1$; thus $(w_1) \text{ col } 1$ remains 0 and $I - e_1^T(w_1 C_2)$ remains a restricted Gauss matrix.

Next, $M_2^{-1}(C_2 M_1^{-1} C_2) = (I - e_2^T w_2)(I - e_1^T(w_1 C_2)) = I - e_1^T(w_1 C_2) - e_2^T w_2$, since $e_2^T w_2 e_1^T(w_1 C_2) = O_{k \times k}$. And thus, $C_3(M_2^{-1}(C_2 M_1^{-1} C_2))C_3 = I - e_1^T(w_1 C_2 C_3) - e_2^T(w_2 C_3)$.

Continuing in this manner, we finally obtain

$$B^{-1}Q = I - \sum_{1 \leq i \leq r} e_i^T(w_i C_{i+1} \cdots C_r).$$

But, $[w_i C_{i+1} \cdots C_r] \text{ col } (1 : i) = 0$ and $[w_i C_{i+1} \cdots C_r] \text{ col } ((i+1) : k)$ is the final value of $U \text{ row } i \text{ col } ((i+1) : k)$ computed in the Gaussian elimination LU-decomposition algorithm above. Thus $B^{-1}Q = U$, where $U \text{ row } i = e_i + w_i C_{i+1} \cdots C_r$.

When we have obtained L and U and P (or equivalently b) and Q (or equivalently c), we can use the process described above to solve $xA = v$ for any given right-side vector v that admits a solution with two back-substitution steps, and two vector permutations.

Algorithmically, this process is:

1. Compute $v \leftarrow vQ$ by permuting v according to the transposition vector c .
2. Compute y such that $yU = v$ by back-substitution.
3. If $y \text{ col } [(r + 1) : k] \neq 0$ then ($'xA = v'$ is inconsistent; exit.)
4. $z \text{ col } [(r + 1) : k] \leftarrow 0$.
5. Compute $[z \text{ col } (1 : r)]$ such that $[z \text{ col } (1 : r)][L \text{ row } (1 : r) \text{ col } (1 : r)] = [y \text{ col } (1 : r)]$ via back-substitution.
6. Compute $x \leftarrow zP$ by permuting z according to the transposition vector b .
7. exit.

If we wish to solve just the single system of equations $xA = v$, we can save some time by appending v to A as A row $(n + 1)$, and then using Gaussian elimination to convert $\begin{bmatrix} A \\ v \end{bmatrix}$ to lower-triangular

form $\begin{bmatrix} J & 0 \\ K & 0 \\ w & u \end{bmatrix}$ where the row $[w \ u]$ is the result of processing v with Gaussian elimination. (The vector $[w \ u]$ is equal to the vector y computed in step 2, above.)

Exercise 1.24: What is the maximum number of times we *read* an element of the matrix L , given as a function of n and k , in the Gaussian elimination LU-decomposition algorithm above, not including the optional steps in brackets?

Solution 1.24: Let $m = \min(n, k)$. Then the maximum number of L -reads is:

$$\sum_{1 \leq h \leq m} [(n - h + 1)(k - h + 1) + 1 + 2n + 2k + (k - h)(1 + 2(n - h))] = m^3 + \frac{3}{2}(k - n + 2)m^2 + 3[(n + 2)(k + \frac{1}{2}) - \frac{1}{6}k]m.$$

There are several modifications to the Gaussian elimination LU-decomposition algorithm given above that are practically desirable in a computer program.

The first issue has to do with dividing by L_{kk} . Because of round-off error and the possibility of computing too-large or too-small values, we should use an overflow handler that either substitutes the largest representable correctly-signed value in place of the unrepresentable overflowing value (with a warning,) or declares our coefficient matrix to be undecomposable. Also, we should use a machine with “soft” (unnormalized) underflow. Moreover the test “ $a = 0$ ” in step 3 might usefully be replaced by “ $|a| < \epsilon$ ”, where ϵ is a small value near the smallest normalized positive value of the machine. Note that the errors that arise in the various values of a due to round-off error means that the computed value of the rank r may be incorrect.

In some cases we want to enforce a requirement that the $n \times k$ matrix A be of full-rank: $\text{rank}(A) = \min(n, k)$. We can modify A if necessary to do this. In this circumstance, we may replace the statement “if $a = 0$ then ($r \leftarrow h - 1$; exit)” in step 3 with “if $|a| < \epsilon$ then ($a \leftarrow a + 2 \cdot mv$)”, where mv

is a small positive value [PTVF92]. Suitable choices for mv are the value ϵ or $\epsilon + \min_{A_{ij} \neq 0} |A_{ij}|/100$. This device of forcing A to have full-rank is similar to the related device of adding suitable constants to each diagonal element of an $n \times n$ matrix to force it to be non-singular, and to improve its “condition” for processing by the Gaussian elimination LU-decomposition algorithm. (Another way to improve the “condition” of the matrix A is to replace A by AS and v by vS , where the $k \times k$ matrix S is a diagonal matrix that scales the j -th equation $x(A \text{ col } j) = v_j$ by the constant S_{jj} to obtain the equivalent equation $x(A \text{ col } j)S_{jj} = v_jS_{jj}$. Generally we want to choose the scaling factors $S_{11}, S_{22}, \dots, S_{kk}$ to make each equation similar in “size”. For example, we could use $S_{jj} = 1/|A \text{ col } j|$, or $S_{jj} = 1/\max_{1 \leq i \leq n} |A_{ij}|$. Note this scaling could be done within the Gaussian elimination LU-decomposition algorithm given above.)

The second issue has to do with efficiency. The Gaussian elimination LU-decomposition algorithm processes the columns of A one-by-one; The processing of a single column consists of scaling to introduce the value 1 as the diagonal component and subtracting multiples of the scaled column from other columns to “zero-out” the row or tail of the row of that diagonal component. The non-zero element that we scale to one is called the *pivot element*, and the process of “zeroing-out” the pivot-element row is called *pivoting*, so the Gaussian elimination LU-decomposition algorithm consists of $\min(n, k)$ pivoting operations. A single pivoting step applied to a matrix A with respect to the pivot element A_{ij} is generally taken to be just the pre- or post-multiplication of A by the appropriate Gauss matrix that converts A row i to e_j or A col j to e_i^T .

In step 2 of the Gaussian elimination LU-decomposition algorithm, the search for the element with the largest absolute value in the sub-array L row $[h : n]$ col $[h : k]$ is called *complete pivoting* search, and the element $L_{pq} = a$ that is found is the pivot element at iteration h . Using the element with the largest absolute value generally results in the best numerical “stability” - we obtain close to the least practicable error in the resulting solution vector or vectors computed based on the lower-triangular matrix L . However, complete pivoting is time-consuming. Nevertheless, if we want to guarantee the exact form of the matrix L specified above, we need to use complete-pivoting or, at least, search for a non-zero element.

Exercise 1.25: Can we really guarantee we will get an (approximate) solution to $xA = v$ in the case where A is non-singular by using the LU-decomposition algorithm given above, followed by two back-substitution computations done in 64-bit floating-point arithmetic?

Solution 1.25: It depends on what the meaning of ‘approximation’ is.

When all we care about is computing the unique solution to $xA = v$ when it exists, then there is a practical compromise called *cross-row partial pivoting* search, where we replace the pivot-value search in step 2 with: “Determine the index $p \in \{h, \dots, n\}$ such that $|L_{ph}| = \max_{h \leq i \leq n} |L_{ih}|$,” and then take $a = L_{ph}$ in step 3. Much experience has shown that this is almost always stable and it is much less costly than complete-pivoting. When the matrix A is non-singular, partial pivoting will generally succeed in stably computing the LU-decomposition.

Exercise 1.26: Show that when cross-row partial pivoting is used, the permutation matrices C_1, \dots, C_r will all be the $k \times k$ identity matrix I .

We could also replace the pivot-value search in step 2 with: “Determine the index $q \in \{h, \dots, k\}$ such that $|L_{hq}| = \max_{h \leq i \leq k} |L_{hi}|$,” and then take $a = L_{hq}$ in step 3. In this case, the permutation

matrices R_1, \dots, R_r will all be the $n \times n$ identity matrix. We shall call this variant of partial pivoting, *cross-column partial pivoting*. When A is non-singular and cross-column partial pivoting is used, we have $AC_1M_1 \cdots C_rM_r = L$, and $L_{ii} \neq 0$ for $1 \leq i \leq n$.

Exercise 1.27: Why must A be non-singular to ensure that $AC_1M_1 \cdots C_rM_r = L$ is obtained with cross-column partial pivoting?

Exercise 1.28: When can we avoid searching for a pivot value entirely, *i. e.* when can we replace step 2 with “ $p \leftarrow h; q \leftarrow h$.”? (Does it suffice for A_{ii} to be non-zero for $1 \leq i \leq n$?)

Exercise 1.29: Can we save any time by searching for the *next* pivot value in L row $((h+1) : n)$ col $((h+1) : k)$ while we are subtracting suitable multiples of L row $((h+1) : n)$ col h from L row $((h+1) : n)$ col $((h+1) : k)$ in step 6, and not using step 2 after the first initial pivot value is determined?

Solution 1.29: Probably we can obtain a constant-factor speed-up. The exact improvement depends on how good a coder you or your compiler is. But the maximum running time of the LU-decomposition algorithm remains $O(\min(n, k)^3)$.

Finally, by replacing the commands “ $L_{hi} \leftarrow 0; U_{hi} \leftarrow z$ ” with “ $L_{hi} \leftarrow z$ ” in step 6, we can save space in the Gaussian-elimination LU-decomposition algorithm by storing the elements U_{ij} for $2 \leq i \leq \min(n, k)$ and $i < j \leq k$ in the strictly-upper-triangular part of the matrix L (which would otherwise be 0) as it is being formed; U_{ii} is known to be 1 for $i = 1, \dots, k$, U_{ij} is known to be 0 for $2 \leq i < j \leq k$, and U row $(n+1 : k) = [O_{k-n, n} \ I_{k-n, k-n}]$ when $n < k$. Thus U need not be explicitly created as a separate array.

Exercise 1.30: Given the $n \times k$ rank r matrix A , suppose we have the LU-decomposition $PAQ = LU$. Describe how we can easily obtain the decomposition $A = FG^T$ where F is an $n \times r$ rank r matrix with linearly-independent columns, and G is a $k \times r$ rank r matrix with linearly-independent columns. (Recall that this decomposition is the starting point for constructing the Moore-Penrose pseudo-inverse matrix A^+ .)

Exercise 1.31: Suppose A is an $n \times n$ non-singular matrix. Explain how to use the LU-decomposition $PAQ = LU$ to compute A^{-1} . Hint: look at $A^{-1}A = I$ as n sets of linear equations: $xA = e_1, \dots, xA = e_n$.

We stated above that when the $n \times n$ matrix A is non-singular, use of cross-column partial pivoting ensures that we can write $AC_1M_1 \cdots C_rM_r = L$, where L is an $n \times n$ non-singular lower-triangular matrix (so that $L_{ii} \neq 0$ for $1 \leq i \leq n$).

By multiplying at most $v := n(n+1)/2$ particular non-singular elementary matrices on the right of L , the non-singular matrix L can be reduced to the $n \times n$ identity matrix. These elementary matrices H_1, \dots, H_v are chosen to effect the same transformations as achieved by the following program: [for $i = n, n-1, \dots, 2$: (for $j = 1, 2, \dots, i-1$: ($L_{ij} \leftarrow L_{ij} - (L_{ij}/L_{ii})L_{ii}$)); for $i = 1, 2, \dots, n$: ($L_{ii} \leftarrow L_{ii}/L_{ii}$)].

Exercise 1.32: Let $1 \leq j < i \leq n$. Show that the elementary matrix that zeros L_{ij} is $E_n[i, j, -L_{ij}/L_{ii}]$. Then show that for $1 \leq k \leq n(n-1)/2$, the elementary matrix H_k is

$E_n[i, j, -L_{ij}/L_{ii}]$ where $i = p+1$ and $j = s - i(i-1)/2$ with $s = 1 + n(n-1)/2 - k$ and $p = \lfloor \sqrt{2s} \rfloor$. (Also, for $n(n-1)/2 + 1 \leq k \leq n(n+1)/2$, $H_k = E_n[r, r, L_{rr}^{-1} - 1]$ where $r = k - n(n-1)/2$.)

Thus $AC_1M_1 \cdots C_rM_rH_1 \cdots H_v = I$, so $A^{-1} = C_1M_1 \cdots C_rM_rH_1 \cdots H_v$ and $A = H_v^{-1} \cdots H_1^{-1}C_r^{-1}H_v^{-1}M_r^{-1}C_r^{-1} \cdots M_1^{-1}C_1^{-1}$. Note that $H_1^{-1}, \dots, H_v^{-1}, M_1^{-1}, \dots, M_r^{-1}$, and $C_1^{-1}, \dots, C_r^{-1}$ are all elementary matrices or products of elementary matrices; thus A is written as a product of elementary matrices. (What is the minimum number of elementary matrices that must be multiplied to guarantee that any non-singular matrix can be produced?)

Exercise 1.33: Let A be an $n \times n$ non-singular matrix. Show that application of the algorithm below to the matrix A exits at step 7 and that, at exit, B is indeed A^{-1} .

1. $h \leftarrow 1$; $B \leftarrow I_{n \times n}$.
2. $a \leftarrow 0$; for $i = h, \dots, n$: if $(A_{hi} \neq 0)$ then $\{a \leftarrow A_{hi}$; $q \leftarrow i$; goto step 3 $\}$.
3. if $a = 0$ then exit(“ A is singular.”).
4. if $(q \neq h)$ then $\{\text{swap } A \text{ col } h \text{ and } A \text{ col } q \text{ in } A$; swap B row h and B row q in B ; $\}$.
5. $A \text{ col } h \leftarrow (A \text{ col } h)/a$; $B \text{ col } h \leftarrow (B \text{ col } h)/a$.
6. for $i = 1, \dots, n$:
if $(i \neq h)$ then $\{A \text{ col } i \leftarrow (A \text{ col } i) - A_{hi}(A \text{ col } h)$; $B \text{ col } i \leftarrow (B \text{ col } i) - A_{hi}(B \text{ col } h)$ $\}$.
7. if $h = n$ then exit(“ $B = A^{-1}$.”).
8. $h \leftarrow h + 1$; go to step 2.

Exercise 1.34: Can any $n \times n$ matrix, not necessarily non-singular, be represented by a product of elementary matrices, as defined here?

Exercise 1.35: Let A be an $n \times n$ symmetric matrix. Show that there exists a non-singular matrix F such that FAF^T is a diagonal matrix of the form $\text{diag}(1, \dots, 1, -1, \dots, -1, 0, \dots, 0)$ where there are s_1 ones, followed by s_2 minus-ones, followed by s_3 zeros, with $s_1 \geq 0$, $s_2 \geq 0$, $s_3 \geq 0$, and $s_1 + s_2 + s_3 = n$. Hint: choose Y^T to be the matrix $C_1M_1 \cdots C_rM_r$ in the identity $AC_1M_1 \cdots C_rM_r = L$ expressing the reduction of A to lower-triangular form via cross-column partial pivoting. Now express Y^T as a product of elementary matrices $E_k \cdots E_2E_1$. Then define $F^T = Y^T S_1^T \cdots S_n^T$ where $S_i = E_n[i, i, |L_{ii}|^{1/2} - 1]$.

There is an often helpful device called *iterative improvement* [PTVF92] for improving our solution x for $xA = v$ obtained using the LU-decomposition of A . The idea is as follows. Suppose we have an approximate solution y with $y = x + d$, where d is the error in the vector y . Then $yA = (x + d)A = v + c$ where $c = dA = yA - x$. Thus we can compute the error d by solving for d in $dA = c$ (and we can do this using the already-obtained LU-decomposition of A .) Then $y - d$ is an improved (although generally still approximate when ordinary floating-point arithmetic is used) solution to $xA = v$.

Although computing one error vector is usually sufficient (and, indeed, if the vector d that we compute is very small, it need not be used at all,) we can repeat this process until the error vector is suitably small, e.g., until $\max_{1 \leq i \leq n} |d_i| \leq p \cdot \max_{1 \leq i \leq n} |y_i|$ where the “precision” p is the largest floating-point value such that $1 + p = 1$ in machine arithmetic. (On a 64-bit IEEE floating-point machine, $p = 2^{-53}$.)

Exercise 1.36: Is it possible for iterative improvement to diverge? That is, is it possible that $y - d$ is a worse solution to $xA = v$ than y itself is?

References

- [GV89] Gene H. Golub and Charles F. Van Loan. *Matrix Computations, second edition*. John Hopkins University Press, 1989.
- [PTVF92] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes in FORTRAN - The Art of Scientific Computing (2nd. ed.)*. Cambridge Univ. Press, N.Y., 1992.