

GAUSSIAN ELIMINATION IN FLOATING-POINT ARITHMETIC

Zvonimir Bohte and Marko Petkovšek

Department of Mathematics, University of Ljubljana,
p.o.b. 64, 61111 Ljubljana, Yugoslavia

ABSTRACT:

Slight improvements in the a priori error analysis of the Gaussian elimination with partial pivoting are obtained. The failure or non-failure of the method is expressed in terms of the bound on the condition number.

GAUSOVA ELIMINACIJA U ARITMETICI S POMIČNIM ZAREZOM. U radu je dobijeno poboljšanje u apriornoj analizi grešaka kod Gausove metode s parcijalnim pivotiranjem. Uspešnost ili neuspešnost metode izražena je s ograničenjem na broj uslovljenosti.

1. INTRODUCTION

Let A be a real square matrix of order n . It is well-known (see [3]) that the Gaussian elimination with partial pivoting for the solution of the system of linear equations

$$Ax = b$$

where A is non-singular yields a permutation matrix P , a unit lower triangular matrix L , and a non-singular upper triangular matrix U such that

$$PA = LU$$

The solution x is obtained by solving two triangular systems $Ly = Pb$ and $Ux = y$. Furthermore, the elements of L are bounded by 1 in modulus, and the growth of the elements during the elimination is bounded by 2^{n-1} .

This is true for the exact arithmetic only. The rounding errors may cause the failure of the method.

We shall analyse the rounding errors using the equation

$$\text{fl}(x \circ y) = (x \circ y)(1 + \epsilon), \quad |\epsilon| \leq u \quad (1)$$

where x and y are any standard floating-point numbers and $\text{fl}(x \circ y)$ denotes the computed result of any of the four arithmetic operations. We shall suppose that the relative error of an arithmetic operation is bounded by the unit rounding error u which is normally

$$\begin{aligned} u &= b^{1-t}/2 && \text{(for rounding)} \\ &= b^{1-t} && \text{(for chopping)} \end{aligned}$$

where t is the length of the mantissa in the base b (usually 2 or 10). It is of course assumed also that during the computation no overflow or underflow occurs.

In the following we shall be interested only in the triangular decomposition.

2. THE ALGORITHM

We denote the current calculated matrix at the r -th step by $B^{(r)}$. This matrix is already upper triangular in the first $r-1$ columns. It is assumed that the matrix $A = B^{(1)}$ is the matrix stored in the computer. The final calculated upper triangular matrix is denoted by V .

The algorithm for triangular decomposition is as follows:

for $r = 1, \dots, n-1$ do:

find the smallest p such that $r \leq p \leq n$ and

$$|b_{pr}^{(r)}| = \max |b_{ir}^{(r)}|, \quad i = r, \dots, n.$$

if $p > r$ interchange the rows p and r and call the new matrix $B^{(r)}$ again.

for $i = r+1, \dots, n$ set:

$$\begin{aligned} m_{ir} &= \text{fl}(b_{ir}^{(r)}/b_{rr}^{(r)}), \quad \text{if } b_{rr}^{(r)} \neq 0 \\ &= 0, && \text{otherwise} \end{aligned}$$

$$b_{ir}^{(r+1)} = 0.$$

for $k = r+1, \dots, n$ set:

$$b_{ik}^{(r+1)} = \text{fl}(b_{ik}^{(r)} - \text{fl}(m_{ir} b_{rk}^{(r)})).$$

If we assume $b_{ik}^{(r+1)} = b_{ik}^{(r)}$ whenever $i \leq r$ or $k < r$, then $B^{(n)} = V$.

Note that this algorithm never breaks down and yields an upper triangular matrix V even if some of the current matrices $B^{(r)}$ are singular. In such a case V is also singular.

3. THE ROUNDING ERROR ANALYSIS

The following rounding error analysis is in all essential ideas due to Wilkinson [2]. There are slight improvements only in some details.

Let us denote the largest elements in modulus at each step by h_r :

$$h_r = \max |b_{ik}^{(r)}|, \quad i, k = r, \dots, n$$

and

$$h = \max h_r, \quad r = 1, \dots, n$$

Using (1) we have

$$m_{ir} = q_{ir}(1 + x_{ir}), \quad i = r+1, \dots, n \quad (2)$$

and

$$b_{ik}^{(r+1)} = (b_{ik}^{(r)} - m_{ir}b_{rk}^{(r)}(1 + y_{ik}^{(r)}))(1 + z_{ik}^{(r)}), \quad i, k = r+1, \dots, n \quad (3)$$

where

$$q_{ir} = b_{ir}^{(r)} / b_{rr}^{(r)}, \quad \text{if } b_{rr}^{(r)} \neq 0 \\ = 0 \quad \text{otherwise}$$

and $|x_{ir}|, |y_{ik}^{(r)}|, |z_{ik}^{(r)}| \leq u$. Since we are pivoting, $|q_{ir}| \leq 1$. It can be shown that $|fl(x/y)| \leq 1$ whenever $|x| \leq |y|$. Therefore $|m_{ir}| \leq 1$, too.

We can write

$$b_{ik}^{(r+1)} = b_{ik}^{(r)} - m_{ir}b_{rk}^{(r)} + d_{ik}^{(r)}, \quad i = r+1, \dots, n; \quad k = r, \dots, n$$

where

$$d_{ik}^{(r)} = -m_{ir}b_{rk}^{(r)}y_{ik}^{(r)}(1 + z_{ik}^{(r)}) + (b_{ik}^{(r)} - m_{ir}b_{rk}^{(r)})z_{ik}^{(r)}, \quad i, k = r+1, \dots, n$$

and

$$d_{ir}^{(r)} = b_{ir}^{(r)}x_{ir}, \quad i = r+1, \dots, n$$

Then

$$|d_{ik}^{(r)}| \leq duh_r, \quad i, k = r+1, \dots, n \quad (4)$$

and

$$|d_{ir}^{(r)}| \leq uh_r < duh_r, \quad i = r+1, \dots, n \quad (5)$$

where

$$d = 3 + u$$

Wilkinson expressed $d_{ik}^{(r)}$ in another way (see [2]):

$$d_{ik}^{(r)} = b_{ik}^{(r+1)} z_{ik}^{(r)} / (1 + z_{ik}^{(r)}) - m_{ir} b_{rk}^{(r)} y_{ik}^{(r)}$$

and hence

$$|d_{ik}^{(r)}| \leq uh_{r+1} / (1 - u) + uh_r, \quad i, k = r+1, \dots, n \quad (6)$$

It is possible to show (see [3]) that we have obtained the exact decomposition of the permuted and perturbed initial matrix

$$PA + E = MV \quad (7)$$

where M is a unit lower triangular matrix with the elements (2) below the diagonal, P a permutation matrix obtained by the same interchanges defined by the algorithm from the identity matrix I and E the matrix with the elements in the upper triangle

$$e_{ik} = \sum_{r=1}^{i-1} d_{ik}^{(r)}, \quad i \leq k \quad (8)$$

and in the lower triangle

$$e_{ik} = \sum_{r=1}^k d_{ik}^{(r)}, \quad i > k \quad (9)$$

From these equations and the bounds (5) and (6) it is easy to obtain the bounds of the form (see [1])

$$\|E\|_p \leq c_p n^2 uh, \quad p = 1, \infty, E \quad (10)$$

where the constant c_p is independent of A . For instance,

$$c_1 = (2 - u) / (2 - 2u)$$

4. A PRIORI BOUNDS

To find a priori bounds we need a bound on the growth of the computed elements. From (3) it follows

$$h_{r+1} \leq ch_r, \quad r = 1, \dots, n-1$$

where

$$c = 2 + 3u + u^2$$

Hence

$$h_r \leq c^{r-1} h_1 \quad (11)$$

and

$$h \leq c^{n-1} h_1 \quad (12)$$

Inserting (12) into (10) we obtain a priori bounds

$$\|E\|_p \leq c_p n^2 c^{n-1} u h_1 \quad (13)$$

As a slight improvement it is possible to avoid the factor n^2 in these bounds if we use (11) instead of (12) in the proof of (10).

From (8) and (9) using (11) in (4) and (5) we obtain

$$|e_{ik}| \leq du \sum_{r=1}^{i-1} h_r \leq du h_1 (c^{i-1} - 1)/(c - 1), \quad i \leq k$$

and

$$|e_{ik}| \leq du \sum_{r=1}^k h_r \leq du h_1 (c^k - 1)/(c - 1), \quad i > k$$

Then, for instance,

$$\|E\|_1 \leq du h_1 \|F\|_1 / (c - 1)$$

where

$$\begin{aligned} f_{ik} &= c^{i-1} - 1, \quad i \leq k \\ &= c^k - 1, \quad i > k \end{aligned}$$

Since $c > 1$ it is obvious that

$$\|F\|_1 = \sum_{i=1}^n (c^{i-1} - 1) = (c^n - 1)/(c - 1) - n$$

and

$$\|E\|_1 \leq d(c^n - 1 - n(c - 1)) u h_1 / (c - 1)^2 \quad (14)$$

which is about $n^2/6$ times smaller than the bound (13) for $p=1$.

Similarly we can get the bounds for the other two norms.

A priori bounds are therefore of the form

$$\|E\|_p \leq d_p c^n u h_1, \quad p = 1, \infty, E \quad (15)$$

where d_p is independent of A .

5. CONCLUSIONS

On the basis of this analysis it is possible to state sufficient conditions which will guarantee that the computed upper triangular matrix V is non-singular and therefore that the Gaussian elimination with partial pivoting does not break down.

THEOREM. If $\|E\|_p \leq e_p$ and $\text{cond}_p(A) < h_1/e_p$ then the matrix V is non-singular.

Proof. For the norms $1, \infty, E$ we have

$$h_1 \leq \|A\|_p \quad (16)$$

From equation (7) it follows that the matrix V is non-singular if the matrix $PA + E$ is non-singular or if

$$\|A^{-1}P^T E\|_p < 1$$

This inequality follows directly from the hypotheses, (16), and the fact that

$$\|P^T E\|_p = \|E\|_p, \quad p = 1, \infty, E \quad \text{Q.E.D.}$$

If the failure of the Gaussian elimination actually occurs then it follows from the theorem that

$$\text{cond}_p(A) \geq h_1/e_p$$

where e_p is any bound for $\|E\|_p$. Here, we could use in (10) the actual value of h . To find it we must determine h_r at each step. But then it would be more appropriate to perform complete pivoting since no additional arithmetic operations would be necessary.

Because of the exponential factor in the a priori bounds (10) and (15) the theorem is very weak. The following table shows the precision required by the theorem using (14) in order to guarantee non-singular V when rounding in base 10 for some typical condition numbers and a few n :

$\text{cond}_1(A)$	minimal length of mantissa		
	$n = 5$	$n = 10$	$n = 100$
1	3	5	32
10^2	5	7	34
10^4	7	9	36
10^6	9	11	38

For large n the results are very unrealistic. This shows the weakness of this a priori error analysis.

R E F E R E N C E S

1. Z. Bohte, Bounds for rounding errors in the Gaussian elimination for band systems, *J. Inst. Maths Applics*, 16 (1975), 133 - 142.
2. J.H. Wilkinson, Error analysis of direct methods of matrix inversion, *J. Assoc. Comput. Mach.*, 8 (1961), 281 - 330.
3. J.H. Wilkinson, *The algebraic eigenvalue problem*, Clarendon Press, Oxford, 1965.