Numerical Methods and
Approximation Theory
Niš, September 26-28, 1984

GAUSSIAN ELIMINATION FOR DIAGONALLY DOMINANT MATRICES

Zvonimir Bohte, Marko Petkovšek

ABSTRACT:
Wilkinson [1] proved that the property of columnwise diagonal dominancy is preserved during the Gaussian elimination. This is true only for exact arithmetic. In this paper a corresponding theorem for floating point arithmetic is proved.

GAUSOVA ELIMINACIJA ZA DIJAGONALNO DOMINANTNE MATRICE.
Wilkinson [1] je dokazao da se osobina dijagonalne dominantnosti po kolonama u toku Gausove eliminacije ne narušava. To je tačno samo za egzaktnu aritmetiku. U ovom radu je dokazana odgovarajuća teorema za aritmetiku u pomičnom zarezu.

1. INTRODUCTION

Let A be a real square matrix of order n. The Gaussian elimination for the solution of the system of linear equations

$$Ax = b$$

yields a set of equivalent systems

$$A^{(r)}x = b^{(r)} \quad , \quad r = 1,\dots,n$$

where $A^{(1)} = A$, $b^{(1)} = b$ and $A^{(n)}$ is an upper triangular matrix. The matrix $A^{(r)}$ has the following block structure

$$A^{(r)} = \begin{bmatrix} U_r & X_r \\ \emptyset & A_r \end{bmatrix} \tag{1}$$

where U_r is an upper triangular matrix of order r-1 and A_r a square matrix of order n-r+1.

Wilkinson [1] proved: If the original matrix A is columnwise diagonally dominant, i.e. if

$$|a_{kk}| > \sum_{\substack{i=1 \\ i\neq k}}^{n} |a_{ik}| \quad , \quad k = 1,\dots,n$$

then the matrix A_r is also columnwise diagonally dominant, i.e.

$$|a_{kk}^{(r)}| > \sum_{\substack{i=r \\ i\neq k}}^{n} |a_{ik}^{(r)}| \quad , \quad k = r,\dots,n$$

for all $r = 2,\dots,n-1$. He also proved that

$$\max_{i,k,r} |a_{ik}^{(r)}| \leqq 2.\max_{i,k}|a_{ik}| .$$

Unfortunately, the presence of rounding errors may destroy the original diagonal dominancy. Therefore, to ensure the nonfailure of the method it is necessary to require more than just a mere diagonal dominancy.

In the analysis of rounding errors we shall use the equation

$$(2) \qquad fl(x \circ y) = (x \circ y)(1 + e) \ , \quad |e| \leqq u$$

where x and y are any standard floating point numbers and $fl(x \circ y)$ 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 unit rounding error which is normally

$$u = b^{1-t}/2 \quad \text{(for rounding)}$$
$$= b^{1-t} \quad \text{(for chopping)}$$

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 leave out all work with the right-hand sides.

2. THE ALGORITHM AND ERROR ANALYSIS

We denote the current calculated matrix at the r-th step by $B^{(r)}$. It has the same block structure as the matrix (1)

$$B^{(r)} = \begin{bmatrix} V_r & Y_r \\ \emptyset & B_r \end{bmatrix}$$

It is assumed that the matrix $A = B^{(1)}$ is the matrix stored in the computer.

The algorithm for the calculation of the upper triangular matrix $B^{(n)}$ is as follows:

$$r = 1,\dots,n-1:$$
$$i = r+1,\dots,n:$$
(3) $$m_{ir} = fl(b_{ir}^{(r)}/b_{rr}^{(r)})$$
$$k = r+1,\dots,n:$$
(4) $$b_{ik}^{(r+1)} = fl(b_{ik}^{(r)} - fl(m_{ir}b_{rk}^{(r)}))$$

Let us denote

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

and

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

Using (2) in (3) and (4) we have

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

and

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

where

(8) $$q_{ir} = b_{ir}^{(r)}/b_{rr}^{(r)}$$

and

(9) $$|x_{ir}|, \ |y_{ik}^{(r)}|, \ |z_{ik}^{(r)}| \leqq u$$

Let us suppose that

(10) $$|q_{ir}| \leqq 1$$

We can write equation (7) in the form

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

where

$$d_{ik}^{(r)} = -q_{ir}b_{rk}^{(r)}(x_{ir} + y_{ik}^{(r)} + x_{ir}y_{ik}^{(r)})(1 + z_{ik}^{(r)}) + \\ +(b_{ik}^{(r)} - q_{ir}b_{rk}^{(r)})z_{ik}^{(r)}$$

Then we can obtain the bound for $d_{ik}^{(r)}$ using (5), (9) and (10)

(12) $$|d_{ik}^{(r)}| \leqq h_r(2u+u^2)(1+u) + 2uh_r = (4 + 3u + u^2)uh_r$$

Now, we can formulate the theorem.

3. THE THEOREM

Let A be a columnwise diagonally dominant matrix of order n and furthermore, let

$$(13) \qquad |a_{kk}| > \sum_{\substack{i=1\\ i\neq k}}^{n} |a_{ik}| + cun(n-1)|a_{kk}| \ , \quad k = 1,\dots,n$$

where $c = 4 + 3u + u^2$, and u is the unit rounding error.

Then the following is true for $r = 1,\dots,n$:

(i) the matrix B_r is columnwise diagonally dominant and furthermore,

$$|b_{kk}^{(r)}| > \sum_{\substack{i=r\\ i\neq k}}^{n} |b_{ik}^{(r)}| + cu(n-r+1)(n-r)|a_{kk}| \ , \quad k = r,\dots,n$$

$$(ii) \quad \sum_{i=r}^{n} |b_{ik}^{(r)}| \leq \sum_{i=1}^{n} |a_{ik}| + cu(2n-r)(r-1)|a_{kk}| \ , \quad k = r,\dots,n$$

$$(iii) \quad |b_{ik}^{(r)}| \leq (2 - cu(n-r+1)(n-r))|a_{kk}| \ , \quad i,k = r,\dots,n$$

PROOF. We shall prove the theorem by the mathematical induction with respect to r. Let $r = 1$. Then, since $B^{(1)} = B_1 = A$, proposition (i) coincides with (13). Obviously, (13) implies that $cun(n-1) < 1$. Therefore, (ii) and (iii) hold trivially for $r = 1$.

Let propositions (i) - (iii) hold for some r, $1 \leq r \leq$ $\leq n-1$, and let $r+1 \leq k \leq n$. From (11) and (8) we obtain

$$(14) \qquad \sum_{\substack{i=r+1\\ i\neq k}}^{n} |b_{ik}^{(r+1)}| \leq |b_{rk}^{(r)}|/|b_{rr}^{(r)}| \sum_{\substack{i=r+1\\ i\neq k}}^{n} |b_{ir}^{(r)}| + \sum_{\substack{i=r+1\\ i\neq k}}^{n} |b_{ik}^{(r)}| + \sum_{\substack{i=r+1\\ i\neq k}}^{n} |d_{ik}^{(r)}|$$

From (i) and (8) it follows that the inequality (10) holds. Therefore, we can use the bound (12) in (14). From (i) it follows

$$\sum_{\substack{i=r+1\\ i\neq k}}^{n} |b_{ir}^{(r)}| < |b_{rr}^{(r)}| - |b_{kr}^{(r)}|$$

So, from (14) we have

$$\sum_{\substack{i=r+1\\ i\neq k}}^{n} |b_{ik}^{(r+1)}| \leq |b_{rk}^{(r)}|(|b_{rr}^{(r)}| - |b_{kr}^{(r)}|)/|b_{rr}^{(r)}| + \sum_{\substack{i=r+1\\ i\neq k}}^{n} |b_{ik}^{(r)}| + cuh_r(n-r-1) =$$

$$= \sum_{\substack{i=r\\ i\neq k}}^{n} |b_{ik}^{(r)}| - |q_{kr}||b_{rk}^{(r)}| + cuh_r(n-r-1)$$

Finally, from (i), (iii), (11) and (12) it follows

$$\sum_{\substack{i=r+1\\ i\neq k}}^{n} |b_{ik}^{(r+1)}| < |b_{kk}^{(r)}| - cu(n-r+1)(n-r)|a_{kk}| - |q_{kr}||b_{rk}^{(r)}| + 2cu(n-r-1)|a_{kk}| \leq$$

$$\leq |b_{kk}^{(r+1)} - d_{kk}^{(r)}| - cu((n-r)(n-r-1) + 2)|a_{kk}| \leq$$

$$\leq |b_{kk}^{(r+1)}| + 2cu|a_{kk}| - cu((n-r)(n-r-1) + 2)|a_{kk}| \leq$$

$$\leq |b_{kk}^{(r+1)}| - cu(n-r)(n-r-1)|a_{kk}|$$

which proves (i).

To prove (ii), note that

$$\sum_{i=r+1}^{n} |q_{ir}| < 1 \tag{15}$$

because B_r is columnwise strictly diagonally dominant. Therefore, (11), (12), (15) and (iii) imply that

$$\sum_{i=r+1}^{n} |b_{ik}^{(r+1)}| \leq \sum_{i=r+1}^{n} |b_{ik}^{(r)}| + |b_{rk}^{(r)}| \sum_{i=r+1}^{n} |q_{ir}| + \sum_{i=r+1}^{n} |d_{ik}^{(r)}| \leq$$

$$\leq \sum_{i=r}^{n} |b_{ik}^{(r)}| + 2cu(n-r)|a_{kk}|$$

Then, using (ii) it follows

$$\sum_{i=r+1}^{n} |b_{ik}^{(r+1)}| \leq \sum_{i=1}^{n} |a_{ik}| + cu(2n-r)(r-1)|a_{kk}| + 2cu(n-r)|a_{kk}| = \tag{16}$$

$$= \sum_{i=1}^{n} |a_{ik}| + cu(2n-r-1)r|a_{kk}|$$

and we have obtained the same inequality (ii) in which r is replaced by r+1.

If we proceed and use the inequality (13) in (16) we get

$$\sum_{i=r+1}^{n} |b_{ik}^{(r+1)}| \leq 2|a_{kk}| - cun(n-1)|a_{kk}| + cu(2n-r-1)r|a_{kk}| =$$
$$= 2|a_{kk}| - cu(n-r)(n-r-1)|a_{kk}|$$

Therefore, for each pair $i,k = r+1,\dots,n$

$$|b_{ik}^{(r+1)}| \leq (2 - cu(n-r)(n-r-1))|a_{kk}|$$

which proves (iii).

4. CONCLUSIONS

The assumptions of the Theorem are sufficient to ensure that the Gaussian elimination in floating point cannot break down. All the quotients m_{ir} are bounded in modulus by 1 and the pivotal growth of the computed elements is bounded by 2. Therefore, in view of Wilkinson's error analysis [1] the Gaussian elimination for matrices which satisfy (13) is numerically stable.

The Theorem also enables us to determine the minimal length of the mantissa which ensures that the breakdown of the Gaussian elimination cannot occur. Let the matrix A be such that

$$|a_{kk}| \geq d \sum_{\substack{i=1 \\ i\neq k}}^{n} |a_{ik}| , \quad k = 1,\dots,n$$

The following table shows the minimal length of the mantissa in dependance on d and n with rounding in base 10.

minimal length of the mantissa

d	n = 5	n = 10	n = 100
1·001	6	7	9
1·01	5	6	8
1·1	4	5	7
1·5	4	4	6
2	3	4	6

REFERENCES:

1. WILKINSON J.H.: *Error analysis of direct methods of matrix inversion.* J. ACM 8 (1961), 281 - 330.