1973

Modifying pivot elements in Gaussian elimination

G. W. (Gilbert W.) Stewart
Carnegie Mellon University

Follow this and additional works at: http://repository.cmu.edu/compsci
NOTICE WARNING CONCERNING COPYRIGHT RESTRICTIONS:
The copyright law of the United States (title 17, U.S. Code) governs the making of photocopies or other reproductions of copyrighted material. Any copying of this document without permission of its author may be prohibited by law.
MODIFYING PIVOT ELEMENTS IN GAUSSIAN ELIMINATION

G. W. Stewart

Departments of Computer Science and Mathematics
Carnegie-Mellon University
Pittsburgh, Pa.

March 1973

This work was supported in part by the Office of Naval Research under Contract No. N00014-67-A-0314-0018.
The rounding-error analysis of Gaussian elimination shows that the method is stable only when the elements of the matrix do not grow excessively in the course of the reduction. Usually such growth is prevented by interchanging rows and columns of the matrix so that the pivot element is acceptably large. In this paper the alternative of simply altering the pivot element is examined. The alteration, which amounts to a rank one modification of the matrix, is undone at a later stage by means of the well-known formula for the inverse of a modified matrix. The technique should prove useful in applications in which the pivoting strategy has been fixed, say to preserve sparseness in the reduction.
1. INTRODUCTION

Let $A$ be a real matrix of order $n$. The method of Gaussian elimination may be regarded as a technique for computing the LU decomposition of $A$ into the product of a unit lower triangular matrix $L$ and an upper triangular matrix $U$. Specifically, at the $k$-th step of the reduction, we have

$$A = \begin{pmatrix}
L_{11}^{(k)} & 0 \\
L_{21}^{(k)} & 1
\end{pmatrix}\begin{pmatrix}
U_{11}^{(k)} & U_{12}^{(k)} \\
0 & U_{22}^{(k)}
\end{pmatrix}$$

where $L_{11}^{(k)}$ and $U_{11}^{(k)}$ are of order $k$. The $(k+1)$-th row of $U$ is then given by

$$u_{k+1,j}^{(k)} = a_{k+1,j}^{(k)} (j = k+1, k+2, \ldots, n),$$

the $(k+1)$-th column of $L$ by,

$$\ell_{i,k+1} = \frac{a_{i,k+1}^{(k)}}{a_{k+1,k+1}^{(k)}} (i = k+1, k+2, \ldots, n),$$

and the matrix $A_{22}^{(k+1)}$ by

$$a_{i,j}^{(k+1)} = a_{i,j}^{(k)} - \ell_{i,k+1} u_{k+1,j}^{(k)} (i, j = k+2, \ldots, n).$$

The element $a_{k+1,k+1}^{(k)}$ is called a pivot element for the algorithm. If it is zero the algorithm fails, and if it is too small the algorithm becomes unstable in the presence of rounding errors. Usually this problem is avoided by interchanging two rows and perhaps two columns of $A_{22}^{(k)}$ to bring an acceptably large element into the pivot position. However, in applications involving large sparse matrices an unhappy pivot selection may destroy
the sparsity of the subsequent matrices. Indeed in some applications
the choice of pivots is determined entirely from the sparsity structure
of A, leaving no freedom to pivot for stability (e.g. see [1]).

In this paper we shall examine the technique of modifying the pivot
element so that it is acceptably large and then undoing the modification
later after the LU decomposition of the modified matrix has been computed.
Since the emergence of a small pivot element in Gaussian elimination be-
tokens a numerical ill-determination of the LU decomposition, we shall
not try to obtain the LU decomposition of A itself; rather we shall show
how the LU decomposition of the modified matrix may be used to solve linear
systems involving A.

The next section will be devoted to describing the mechanics of the
technique. The effects of rounding error will be discussed in Section 3.

2. MODIFYING PIVOT ELEMENTS IN THE SOLUTION OF LINEAR EQUATIONS

In this section we shall show how the solution of the equation

\[(2.1) \quad Ax = b\]

can be obtained from the solution of

\[(2.2) \quad By = b,\]

where A and B differ only in their \((1,1)\)-elements. We shall then indicate
the applications of this technique in Gaussian elimination.

Since A and B differ only in their \((1,1)\)-elements, B can be written
in the form
where \( e_1 \) is the first column of the identity matrix. Then it follows from the well known modification formula (see [2, p. 123]) that

\[
A^{-1} = B^{-1} - \tau B^{-1} e_1 e_1^T B^{-1},
\]

where

\[
\tau = \frac{1}{e_1^T B^{-1} e_1 - \sigma^{-1}}.
\]

Since \( x = A^{-1} b \), we have from (2.2) and (2.3)

\[
x = y - \tau B^{-1} e_1 e_1^T y
\]

\[
= y - \tau y_1 c_1,
\]

where \( y_1 \) is the first component of \( y \) and \( c_1 \) is the first column of \( B^{-1} \).

Thus the solution of (2.1) can be obtained from the solution of (2.2) by subtracting a suitable correction vector.

The economics of this technique are favorable. The system (2.2) costs no more to solve than (2.1). The vector \( c_1 \) can be obtained at the same time and at very little additional cost by solving the system

\[
Bc_1 = e_1.
\]

The computation of \( \tau \) (n. b., \( e_1^T B^{-1} e_1 \) is the first component of \( c_1 \) ) and \( x \) entails a negligible amount of additional calculation. Note that once \( c_1 \) has been calculated it can be saved and used to solve other systems of the form (2.1) with differing right hand sides.

Concerning Gaussian elimination, suppose that at the k-th step an unacceptably small pivot element emerges. Then a solution of (2.1) may
be obtained in the form \( x^T = (x_1^{(k)T}, x_2^{(k)T}) \) as follows.

1) Solve the system

\[
\begin{pmatrix}
L^{(k)}_{11} & 0 \\
L^{(k)}_{21} & I
\end{pmatrix}
\begin{pmatrix}
b^{(k)}_1 \\
b^{(k)}_2
\end{pmatrix}
= \begin{pmatrix}
b_1 \\
b_2
\end{pmatrix}
\]

2) Set \( B_{22}^{(k)} = A_{22}^{(k)} + \sigma^k e_1 e_1^T \), where \( \sigma^k \) is chosen to make the pivot element acceptably large.

3) Solve the systems

\[
B_{22}^{(k)} y_2^{(k)} = b_2^{(k)}, \quad B_{22}^{(k)} c_1^{(k)} = e_1
\]

4) Correct \( y_2^{(k)} \) to yield a solution of the system

\[
A_{22}^{(k)} x_2^{(k)} = b_2^{(k)}
\]

5) Solve the system

\[
U_{11}^{(k)} x_1^{(k)} = b_1^{(k)} - U_{12}^{(k)} x_2^{(k)}
\]

This process can be repeated should an unacceptably small pivot be encountered in step 3 of the above algorithm; however, here the economics are not as favorable. The time considerations are roughly the same; each application of the technique requires the solution of an additional set of equations involving the matrix \( B_{22}^{(k)} \), a negligible increase over the Gaussian reduction of \( B_{22}^{(k)} \) itself. However, each \( c_1^{(k)} \) must be stored, and, since they are columns of inverse matrices, they need not be sparse, even when
the original matrices are. Thus in applications involving large sparse matrices, the technique cannot be used too many times.

There remains the problem of choosing \( \sigma_k \). It is clear that \( \sigma_k \) must not be too large; for as \( \sigma_k \) increases, \( \sum_{k} e_{T}^{T} \) becomes a slight perturbation of the singular matrix \( \sigma_k e_{T} \). A natural choice is to take \( \sigma_k \) to be just large enough to dominate the elements in the first column of \( A_{22}^{(k)} \), which corresponds to partial pivoting in the elimination process. Since the value of the next pivot element can easily be computed, it may be desirable to alter \( \sigma_k \) slightly, say multiply it by a factor of two, whenever cancellation would occur in the calculation of the next pivot element.

It is hardly necessary to add that our results hold also for the Crout and Doolittle variants of Gaussian elimination, for which the discussion above remains valid with some slight and obvious modifications.*

3. ERROR ANALYSIS

The algorithm described in the last section must be implemented in finite precision arithmetic, and it is important to assess the effects of the resulting rounding errors on the solution. For simplicity we shall first assume that a modification is made at the first stage of the elimination and drop the superscripts \( (k) \). We shall determine conditions under which the computed solution \( x \) has a residual

\[
r = b - Ax
\]

*This baroque variation on the famous theme of Laplace (il est aisé à voir) is due to Ostrowski [Arch. Rational Mech. Anal. 1 (1958), p.241].
that is small. Note that, whatever the value of $r$, $x$ is the solution of the system

$$(A+E)x = b,$$

where $E = rx^T/\|x\|^2$ satisfies

$$\frac{|r|}{\|A\|} = \frac{|x|}{\|A\|}$$

in the Frobenius norm defined by $\|A\|^2 = \text{trace } A^TA$. Thus a small residual implies that $x$, however inaccurate, is the solution of a slightly perturbed problem.

The following notation will be used in the error analysis. The symbol $\langle k \rangle$, called a relative counter, will stand generically for a quotient of the form

$$(1+p_1)(1+p_2)\ldots(1+p_k),$$

(3.1) \( \langle k \rangle = \frac{(1+p_1)(1+p_2)\ldots(1+p_k)}{(1+p_{k+1})(1+p_{k+2})\ldots(1+p_k)} \),

where the numbers $|p_i|$ are uniformly bounded by some small quantity. We shall also use the notation $\#k\#$ for the deviation of $\langle k \rangle$ from unity:

$$\#k\# = 1 - \langle k \rangle.$$

(3.2) \(|\#k\#| \leq k \cdot \varepsilon \leq \cdot 1$$

for some number $\varepsilon$ of approximately the same size as the bound on the $|p_i|$. 

The symbol $\#k\#$ will be called an absolute counter. We shall assume that the bounds on the $p_i$ in (3.1) and on the integer $k$ are so restricted that
If \( x \) is a vector, then \( x^{\langle k \rangle} \) will denote the vector \((x_1^{\langle k \rangle}, x_2^{\langle k \rangle}, \ldots, x_n^{\langle k \rangle})^T\), where each appearance of the counter \( k \) may stand for a different value.

The relative and absolute counters have the following easily verified properties:

\[
\langle k \rangle \langle l \rangle = \langle k+l \rangle,
\]

\[
1/\langle k \rangle = k,
\]

and

\[
(3.3) \quad \langle k \rangle - \langle l \rangle = \#k + \#l.
\]

The usual backward error bounds for \( t \)-digit, base \( \beta \), floating-point arithmetic (see, e.g., [3]) can be expressed in the form

\[
fl(a \circ b) = (a \circ b)^{\langle 1 \rangle} \quad (\circ = x, \div)
\]

and

\[
fl(a + b) = a^{\langle 1 \rangle} + b^{\langle 1 \rangle},
\]

where \( e \) in (3.2) is of the order \( \beta^{-t} \).

We turn now to the analysis of the effect of modifying the 1-1 element of \( A \). All quantities will denote the computed values, with the exception of \( B = A + e e_1^T \). The first step is to solve the systems \( B y = b \) and \( B c_1 = e_1 \). We assume this is done stably so that the computed solutions satisfy

\[
B_1 y = (B + E_1) y = b
\]

and

\[
B_2 c_1 = (B + E_2) c_1 = e_1,
\]

where

\[
\|E_i\| \leq \epsilon_i \|B\| \quad (i=1,2)
\]

for some small \( \epsilon_i \). Note that the single rounding error made in forming \( B \) from \( A \) may be absorbed in the error matrices \( E_i \).
The next step is to compute $\tau$. There is no rounding error in the computation of $c_{11} = e_1^T B_2 e_1$. Hence

$$\tau = \frac{1}{c_{11} \langle 1 | \sigma^{-1} | 2 \rangle}$$

or

$$(3.4) \quad \tau + \sigma \langle 3 \rangle - \tau \sigma c_{11} \langle 3 \rangle = 0.$$  

Finally one computes

$$x = \frac{1}{c_{11}} y_{1} \langle 4 \rangle - x \langle 3 \rangle.$$  

From this it follows that

$$\tau y_{1} c_{1} = y \langle 4 \rangle - x \langle 3 \rangle,$$

whence from (3.2)

$$(3.5) \quad |\tau| |x_{1}| \|c_{1}\| \leq 1.1 \left(\|x\| + \|y\|\right)$$

Now

$$r = b - Ax$$

$$= b - (B - \sigma e_{1} e_{1}^T) x$$

$$= b - (B_{2} - \sigma e_{1} e_{1}^T) x + E_{2} x$$

$$(3.6)$$

$$= b - (B_{2} - \sigma e_{1} e_{1}^T) \left(y \langle 1 \rangle - \tau y_{1} c_{1} \langle 3 \rangle\right) - E_{2} x$$

$$= b - B_{2} y \langle 1 \rangle - E_{2} x + y_{1} (\tau + \sigma \langle 1 \rangle - \tau \sigma c_{11} \langle 3 \rangle) e_{1}$$

$$+ y_{1} \tau B_{2} c_{1} \ #3\#$$

the last equality following from the fact that

$$B_{2} c_{1} \langle 3 \rangle = B_{2} c_{1} + B_{2} c_{1} \ #3\# = e_{1} + B_{2} c_{1} \ #3\#.$$
On subtracting zero in the guise of (3.4) from the term in parentheses in the last member of (3.6), we obtain from (3.3)

\[ r = b - B_2 y + E_2 x + y_1 (\sigma + \tau c_{11} \#6\#) e_1 + y_1TB_2 c_1 \#3\#. \]

Hence

\[ r = b - B_1 y + (E_1 - E_2) y + B_2 y \#1\# - E_2 x \]

\[ + y_1 (\sigma + \tau c_{11} \#6\#) e_1 + y_1TB_2 c_1 \#3\#. \]

Since \( b - B_1 y = 0 \), if we let

\[ \mu = \max \left( \frac{|B_2|}{|A|}, \frac{|\sigma|}{|A|} \right) \]

and

\[ \lambda = \frac{|b|}{|x|}, \]

then from (3.2)

\[
\frac{|x|}{|A||x|} \leq (\varepsilon_1 + \varepsilon_2) \lambda + \varepsilon_2 \lambda + \varepsilon_2 + 4 \varepsilon \lambda \mu \\
+ 5.6 \varepsilon \mu (1+\lambda) + 3.3 \varepsilon \mu (1+\lambda) \\
\leq \varepsilon_2 + 10 \varepsilon \mu + \lambda (\varepsilon_1 + \varepsilon_2 + 15 \varepsilon \mu) \]

This result is quite satisfactory. For reasonable modifications of the pivot element, the number \( \mu \) will be of order unity. Thus \( \lambda \), the ratio of the sizes of the computed solutions of the equations \( Ax = b \) and \( By = b \), is the controlling factor. If \( \lambda \) is large, that is if severe numerical cancellation occurs in the passage from \( y \) to \( x \), the result cannot be guaranteed to have a small residual. Note that this cannot happen if \( B \) is well conditioned, whatever the condition of \( A \). In any event, the condition
is one that can be easily check.

There remains one point to clear up. The modification step is only one part of the algorithm described in the last section, and we must show that this algorithm as a whole is stable. The usual rounding error analysis for triangular systems shows that the computed vectors $b_1^{(k)}$ and $b_2^{(k)}$ satisfy

\[
\begin{pmatrix}
L_{11}^{(k)} + F_{11} & 0 \\
L_{21}^{(k)} + F_{21} & I
\end{pmatrix}
\begin{pmatrix}
b_1^{(k)} \\
b_2^{(k)}
\end{pmatrix}
= 
\begin{pmatrix}
b_1 \\
b_2
\end{pmatrix},
\]

where $F_{11}$ and $F_{21}$ are small compared to $L_{11}^{(k)}$ and $L_{21}^{(k)}$. The results of this section imply that, if all has gone well, the computed vector $x_2^{(k)}$ will satisfy

\[
(A_2^{(k)} + G_{22})x_2^{(k)} = b_2^{(k)},
\]

where $G_{22}$ is small compared with $A_2^{(k)}$. Since the solution for $x_1^{(k)}$ amounts to no more than the completion of the solution of a triangular system, the computed vector $x$ satisfies

\[
\begin{pmatrix}
U_{11}^{(k)} + G_{11} & U_{12}^{(k)} + G_{12} \\
0 & A_2^{(k)} + G_{22}
\end{pmatrix}
\begin{pmatrix}
x_1^{(k)} \\
x_2^{(k)}
\end{pmatrix}
= 
\begin{pmatrix}
b_1^{(k)} \\
b_2^{(k)}
\end{pmatrix},
\]

where $G_{11}$ and $G_{22}$ are small compared with $U_{11}^{(k)}$ and $U_{12}^{(k)}$. Equations (3.7) and (3.8) can be combined in the usual way to show that the computed solution satisfies

\[
(A + H)x = b
\]

where $H$ is small compared with $A$ (see, for example, [3, p. 108], in which the final bound must be supplemented by a factor of $\|L\|$ since no assumptions about pivoting strategy have been made).
REFERENCES

