Gauss Elimination Method with Partial Pivoting

The reduction of a matrix A to its row echelon form may necessitate row interchanges as the example shows: $A := \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$ shows. In this case, after interchanging the two rows, we obtain $x_1 = 1 = x_2$ as the (unique) solution of the linear system

$$\begin{aligned} 0x_1 + x_2 &= 1 \\ x_1 + x_2 &= 2. \end{aligned}$$

Let us modify the matrix A in the above example by replacing the 0 in the top left corner by a small number .0001 and consider the following linear system:

$$\begin{aligned} .0001x_1 + x_2 &= 1 \\ x_1 + x_2 &= 2. \end{aligned}$$

We observe that in order to satisfy the above two equations, x_1 and x_2 will be close to 1. In fact,

$$x_1 = \frac{1}{1 - .0001} \quad \text{and} \quad x_2 = 2 - \frac{1}{1 - .0001}.$$

In real life problems, large systems of linear equations need to be solved on computers. Any computer can use only a finite number of real numbers in the calculations. Most computers use a **floating point arithmetic** rather than a fixed point arithmetic. The floating point arithmetic on a particular computer is characterized by its base β, its precision t and its exponent range $[\alpha_1, \alpha_2]$. The floating point numbers consist of the number 0 and the numbers of the form

$$\pm\ .d_1 d_2 \dots d_t \times \beta^e,$$

where $d_1, \dots, d_t$ are nonnegative integers less than β, $d_1 \neq 0$ and e is an integer between α_1 and α_2. Typically, we may have $\beta = 2$, $t = 8$, $\alpha_1 = -64$ and $\alpha_2 = 64$; or $\beta = 10$, $t = 3$, $\alpha_1 = -16$ and $\alpha_2 = 16$. If x is a given real number, then its floating point representative $fl(x)$ is given by the floating point number which is nearest to it. This is known as the **rounded arithmetic**.

Let us use the Gaussian Elimination Method (GEM) to solve the linear system

$$\begin{aligned} .0001x_1 + x_2 &= 1 \\ x_1 + x_2 &= 2 \end{aligned}$$

by using the floating point (rounded) arithmetic with base 10, precision 3 and exponent range $[-16, 16]$. This means that the nonzero numbers that can be used in the calculations are $\pm\ .d_1 d_2 d_3 \times 10^e$, where d_1, d_2, d_3 are nonnegative integers between 0 and 9, $d_1 \neq 0$ and e is an integer between -16 and 16. Using 0.0001 as the pivot, we may eliminate x_1 from the second equation as follows:

$$\left[\begin{array}{cc|c} .0001 & 1 & 1 \\ 1 & 1 & 2 \end{array}\right] \overset{R_2 \longrightarrow R_2 - 10^4 R_1}{\sim} \left[\begin{array}{cc|c} .0001 & 1 & 1 \\ 0 & -9999 & -9998 \end{array}\right].$$

Since we are using base 10 and precision 3 in the rounded arithmetic, the equations $.0001x_1 + x_2 = 1$ and $-9999x_2 = -9998$ will be written as $(.1) \times 10^{-3} x_1 + x_2 = 1$ and $-(.1) \times 10^5 x_2 = -(.1) \times 10^5$. The second equation gives $x_2 = 1$. Back substitution in the first equation gives $x_1 = 0$. However, $x_1 = 0, x_2 = 1$ is a completely wrong solution of the given linear system. We can trace the reason for this anomaly to the pivot .0001 being too small. So although $.0001 \neq 0$ and we can, in **theory**, use it as a pivot in the GEM, it is not advisable to do so in **practice**. Using small pivots means dividing rows by small numbers for elimination. This may introduce errors of underflow, overflow and round-off. Therefore, in each step, it is advisable to use a pivot that is largest in absolute value. Let us carry out the GEM with this precaution, that is, let us interchange the two rows, and solve the above system once again as follows:

$$\left[\begin{array}{rr|r} 1 & 1 & 2 \\ .0001 & 1 & 1 \end{array}\right] \overset{R_2 \longrightarrow R_2 - 10^{-4} R_1}{\sim} \left[\begin{array}{cc|c} 1 & 1 & 2 \\ 0 & .9999 & .9998 \end{array}\right].$$

As before, the equation $.9999x_2 = .9998$ will be written as $.1 \times 10^1 x_2 = .1 \times 10^1$. This gives $x_2 = 1$, and back substitution into the equation $x_1 + x_2 = 2$ gives $x_1 = 1$. Hence we obtain a reasonably correct solution $x_1 = 1 = x_2$ of the given linear system in our arithmetic.

Thus in order to avoid underflow, overflow and roundoff errors, it is necessary to ensure that the pivots are not too small. This means that when we multiply an equation by a number and subtract it from another equation, the multiplier should not be too large. For this purpose, we modify the GEM for an $m \times n$ matrix as far as the selection of pivots is concerned by adopting the following strategy known as **partial pivoting**. For $j = 1, \ldots, m-1$, in the jth step of the GEM, compare the entries in the jth, $(j+1)$st, $\ldots$, mth positions in the relevant column and instead of selecting one which is merely nonzero, select the entry that has the largest absolute value, and make it as the pivot at that step by interchanging two rows. (If there are more than one such entries having the same largest absolute value, then select the one which is uppermost.) Then all the multipliers m_{ij} for the row operations $R_i \longrightarrow R_i - m_{ij} R_j$ for $i = j+1, \ldots, m$ will have absolute values at most 1. In order to guard against any distortion of solution due to the pivots being small, the strategy of partial pivoting is built into the GEM irrespective of the actual sizes of the coefficients of the system. Gaussian elimination with partial pivoting is considered to be one of the most fundamental algorithms in numerical linear algebra.

Example. Let us use the GEM with partial pivoting to solve the following system:

$$\begin{aligned} 2x_1 + x_2 + x_3 &= 5 \\ 4x_1 - 6x_2 &= -2 \\ -2x_1 + 7x_2 + 2x_3 &= 9. \end{aligned}$$

$$\left[\begin{array}{rrr|r} 2 & 1 & 1 & 5 \\ 4 & -6 & 0 & -2 \\ -2 & 7 & 2 & 9 \end{array}\right] \overset{R_1 \longleftrightarrow R_2}{\sim} \left[\begin{array}{rrr|r} 4 & -6 & 0 & -2 \\ 2 & 1 & 1 & 5 \\ -2 & 7 & 2 & 9 \end{array}\right] \overset{R_2 \longrightarrow R_2 - (1/2)R_1,\ R_3 \longrightarrow R_3 + (1/2)R_1}{\sim} \left[\begin{array}{rrr|r} 4 & -6 & 0 & -2 \\ 0 & 4 & 1 & 6 \\ 0 & 4 & 2 & 8 \end{array}\right]$$

$$\overset{R_3 \longrightarrow R_3 - R_2}{\sim} \left[\begin{array}{rrr|r} 4 & -6 & 0 & -2 \\ 0 & 4 & 1 & 6 \\ 0 & 0 & 1 & 2 \end{array}\right].$$

Thus $x_3 = 2$. Then $4x_2 + x_3 = 6$ gives $x_2 = 1$, and $4x_1 - 6x_2 = -2$ gives $x_1 = 1$. Hence $x_1 = 1, x_2 = 1, x_3 = 2$ is the (unique) solution of the given linear system.