MA 214: Introduction to numerical analysis
Lecture 38

Shripad M. Garge.
IIT Bombay

(shripad@math.iitb.ac.in)

2021-2022

System of linear equations

We are interested in solving the system of linear equations. We have the system

$$\begin{aligned}
a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\
a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\
&\vdots \\
a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= b_n
\end{aligned}$$

which is written as $A \cdot x = b$. The matrices A and b have real entries and we want to solve for the matrix x.

As we noted before, this system has a unique solution if the matrix A is invertible, this is so if and only if the determinant of A is non-zero.

Gaußian elimination method

We noted that the case where the matrix A is upper triangular is easier to solve. In that case the system is

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{nn}x_n &= b_n \end{aligned}$$

We can then solve for the variable x_n using the last equation.

Once the value of x_n is found, we use it in the second-last equation

$$a_{n-1,n-1}x_{n-1} + a_{nn}x_n = b_{n-1}$$

to solve for x_{n-1} and so on.

Gaußian elimination method

The procedure will fail if the element a_{ii} is zero in the step of eliminating x_i because then the operation

$$\left(E_j - \frac{a_{ji}}{a_{ii}} E_i\right) \to (E_j)$$

can not be performed.

The system may still have a solution, but the technique for finding the solution must be altered. Let us understand this through an example:

$$\begin{pmatrix} 1 & -1 & 2 \\ 2 & -2 & 3 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} -6 \\ -14 \\ 0 \end{pmatrix}.$$

We perform $(E_2 - 2E_1) \to (E_2)$ and $(E_3 - E_1) \to (E_3)$.

Gaußian elimination method

The resulting system is

$$\begin{pmatrix} 1 & -1 & 2 \\ 0 & 0 & -1 \\ 0 & 2 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} -6 \\ -2 \\ 6 \end{pmatrix}.$$

The diagonal entry a_{22} is now zero. So the procedure can not continue in its present form.

We observe that $a_{32} = 2 \neq 0$, hence we perform the operation $(E_2) \leftrightarrow (E_3)$ to obtain the system

$$\begin{pmatrix} 1 & -1 & 2 \\ 0 & 2 & -1 \\ 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} -6 \\ 6 \\ -2 \end{pmatrix}.$$

We then have $x_3 = 2$, $x_2 = 4$ and $x_1 = -6$.

Gaußian elimination method

The above example illustrates what is to be done if $a_{kk} = 0$ for some k.

We search for the first nonzero entry below a_{kk} in the k-th column.

If $a_{pk} \neq 0$ for some $k < p \leqslant n$, then the operation $(E_k) \leftrightarrow (E_p)$ is performed to obtain the new matrix A.

The procedure can then be continued.

If $a_{pk} = 0$ for each p, then the matrix A is not invertible and hence the system does not have a unique solution. The procedure then stops.

Finally, if $a_{nn} = 0$, the linear system does not have a unique solution, and again the procedure stops.

Operation count

After learning the method, we are interested in finding the total number of arithmetic operations performed in the method.

In general, the amount of time required to perform a multiplication or division on a computer is approximately the same and is considerably greater than that required to perform an addition or subtraction.

The actual differences in execution time, however, depend on the particular computing system.

To demonstrate the counting operations for a given method, we will count the operations required to solve a typical linear system of n equations in n unknowns using GEM.

We will keep the count of the additions/subtractions separate from the count of the multiplications/divisions because of the time differential.

Operation count

Assume that the variables $x_1, \ldots, x_{i-1}$ have been eliminated from the appropriate equations.

For eliminating x_i from the equations E_{i+1} onwards, we define $m_{ji} = a_{ji}/a_{ii}$ and perform $(E_j - m_{ji}E_i) \rightarrow (E_j)$.

Here $(n-i)$ divisions are performed, then m_{ji} is multiplied to each term of E_i, thus resulting in $(n-i)(n-i+1)$ multiplications, and then each term of the resulting equation is subtracted from the equation E_j, requiring $(n-i)(n-i+1)$ subtractions.

Mult/Div: $(n-i) + (n-i)(n-i+1) = (n-i)(n-i+2)$.

Add/Sub: $(n-i)(n-i+1)$.

The total number of operations required is obtained by summing the operation counts for each i.

Operation count

Mult/Div:

$$
\begin{aligned}
\sum_{i=1}^{n-1}(n-i)(n-i+2) &= \sum_{i=1}^{n-1}(n-i)^2 + 2\sum_{i=1}^{n-1}(n-i) \\
&= \sum_{i=1}^{n-1} i^2 + 2\sum_{i=1}^{n-1} i \\
&= \frac{(n-1)n(2n-1)}{6} + 2\frac{(n-1)n}{2} \\
&= \frac{2n^3+3n^2-5n}{6}.
\end{aligned}
$$

This is the count of multiplication or division operations.

Operation count

Add/Sub:

$$\begin{aligned}
\sum_{i=1}^{n-1}(n-i)(n-i+1) &= \sum_{i=1}^{n-1}(n-i)^2 + \sum_{i=1}^{n-1}(n-i) \\
&= \sum_{i=1}^{n-1} i^2 + \sum_{i=1}^{n-1} i \\
&= \frac{(n-1)n(2n-1)}{6} + \frac{(n-1)n}{2} \\
&= \frac{n^3-n}{3}.
\end{aligned}$$

This is the count of addition or subtraction operations.

These are the operations required to make the matrix upper-triangular.

Operation count

We now count the number of operations required in the backward substitutions.

Each such step requires $(n-i)$ multiplications, $(n-i-1)$ additions for each summation term followed by one subtraction and finally one division.

Mult/Div:

$$\begin{aligned}
1+\sum_{i=1}^{n-1}\big((n-i)+1\big) &= 1+\sum_{i=1}^{n-1}(n-i)+(n-1)\\
&= \sum_{i=1}^{n-1}(n-i)+n=\sum_{i=1}^{n-1} i+n\\
&= \frac{n(n-1)}{2}+n=\frac{n^2+n}{2}.
\end{aligned}$$

Operation count

Finally, we count the number of additions/subtractions:

$$\sum_{i=1}^{n-1}\big((n-i-1)+1\big) = \sum_{i=1}^{n-1}(n-i) = \sum_{i=1}^{n-1} i = \frac{n^2-n}{2}.$$

The total number of operations in the GEM is

Mult/Div:

$$\frac{2n^3+3n^2-5n}{6} + \frac{n^2+n}{2} = \frac{n^3}{3} + n^2 - \frac{n}{3}$$

Add/Sub:

$$\frac{n^3-n}{3} + \frac{n^2-n}{2} = \frac{n^3}{3} + \frac{n^2}{2} - \frac{5n}{6}.$$

Operation count

For large n, the total number of multiplications and divisions is approximately $n^3/3$, as is the total number of additions and subtractions.

Thus the amount of computation and the time required increases with n in proportion to n^3, as shown below:

n	Mult/Div	Add/Sub
3	17	11
10	430	375
50	44, 150	42, 875
100	3, 43, 400	3, 38, 250

MA 214: Introduction to numerical analysis
Lecture 39

Shripad M. Garge.
IIT Bombay

(shripad@math.iitb.ac.in)

2021-2022

Pivoting

In GEM, we found that a row interchange was needed when one of the pivot elements a_{kk} is 0.

This row interchange has the form $(E_k) \leftrightarrow (E_p)$, where p is the smallest integer greater than k with $a_{pk} \neq 0$.

To reduce the round-off error, it is often necessary to perform row interchanges even when the pivot elements are not zero.

If a_{kk} is small in magnitude compared to a_{jk}, then the magnitude of the multiplier $m_{jk} = a_{jk}/a_{kk}$ will be much larger than 1.

Round-off error introduced in the computation of one of the terms a_{kl} is multiplied by m_{jk} when computing a_{jl}, which compounds the original error.

Pivoting

Also, when performing the backward substitution for

$$x_k = \frac{a_{k,n+1}x_{n+1} - \sum_{j=k+1}^{n} a_{kj}x_j}{a_{kk}}$$

with a small value of a_{kk}, any error in the numerator can be dramatically increased because of the division by a_{kk}.

In our next example, we will see that even for small systems, round-off error can dominate the calculations.

Consider the system:

$$\begin{aligned} E_1: &\quad 0.003x_1 + 59.14x_2 = 59.17 \\ E_2: &\quad 5.291x_1 - \ \ 6.13x_2 = 46.78 \end{aligned}$$

We use four-digit rounding arithmetic and compare the results to the exact solution $x_1 = 10$ and $x_2 = 1$.

Pivoting

The first pivot element, $a_{11} = 0.003$, is small and then

$$m_{21} = \frac{5.291}{0.003} = 1763.\overline{6}$$

rounds to 1764.

Performing $(E_2 - m_{21}E_1) \to (E_2)$ and the appropriate rounding gives the system

$$\begin{aligned} 0.003x_1 + 59.14x_2 &= \quad 59.17 \\ -104300x_2 &\approx -104400 \end{aligned}$$

instead of the exact system.

The disparity in the magnitudes of $m_{21}a_{13}$ and a_{23} has introduced a round-off error, but the round-off error has not yet been propagated.

Pivoting

Backward substitution yields

$$x_2 \approx 1.001$$

which is a close approximation to the actual value, $x_2 = 1$.

However, because of the small pivot $a_{11} = 0.003$

$$x_1 \approx \frac{59.17 - (59.14)(1.001)}{0.003} = -10$$

contains the small error of 0.001 multiplied by $\dfrac{59.14}{0.003} \approx 20000$.

This ruins the approximation to the actual value $x_1 = 10$.

This is clearly a contrived example but it does give an idea on how the round-off error can alter the results.

Partial pivoting

For larger systems it is much more difficult to predict in advance when devastating round-off error might occur.

To avoid this problem, pivoting is performed by selecting an element a_{pq} with a larger magnitude as the pivot, and interchanging the k-th and p-th rows. This can be followed by the interchange of the k-th and q-th columns, if necessary.

The simplest strategy is to select an element in the same column that is below the diagonal and has the largest absolute value; specifically, we determine the smallest $p \geqslant k$ such that

$$|a_{pk}| = \max_{k \leqslant i \leqslant n} |a_{ik}|$$

and perform $(E_k) \leftrightarrow (E_p)$.

In this case no interchange of columns is used.

Partial pivoting

Consider the system:

$$\begin{aligned} E_1: &\quad 0.003x_1 + 59.14x_2 = 59.17 \\ E_2: &\quad 5.291x_1 - \ \ 6.13x_2 = 46.78 \end{aligned}$$

We use *partial pivoting* and four-digit rounding arithmetic and compare the results to the exact solution $x_1 = 10$ and $x_2 = 1$.

The partial pivoting procedure first requires finding

$$\max\left\{|a_{11}|, |a_{21}|\right\} = \max\left\{|0.003|, |5.291|\right\} = |5.291| = |a_{21}|.$$

Then we perform $(E_2) \leftrightarrow (E_1)$ to get the system

$$\begin{aligned} E_1: &\quad 5.291x_1 - \ \ 6.13x_2 = 46.78 \\ E_2: &\quad 0.003x_1 + 59.14x_2 = 59.17 \end{aligned}$$

Partial pivoting

The multiplier for this system is

$$m_{21} = \frac{a_{21}}{a_{11}} = \frac{0.003}{5.291} = 0.0005670$$

and the operation $(E_2 - m_{21}E_1) \rightarrow (E_2)$ reduces the system to

$$\begin{aligned} E_1:&\quad 5.291x_1 - 6.13x_2 = 46.78 \\ E_2:&\quad 59.14x_2 \approx 59.14 \end{aligned}$$

The four-digit answers resulting from the backward substitution are the correct values $x_1 = 10$ and $x_2 = 1$.

Each multiplier m_{ji} in the partial pivoting algorithm has magnitude less than or equal to 1.

Although this strategy is sufficient for many linear systems, situations do arise when it is inadequate.

Scaled partial pivoting

The linear system:

$$\begin{aligned} E_1: \quad 30x_1 + 591400x_2 &= 591700 \\ E_2: \quad 5.291x_1 - \quad 6.13x_2 &= 46.78 \end{aligned}$$

is the same as that in the above examples, except that all the entries in the first equation, E_1, have been multiplied by 10^4.

The partial pivoting procedure with four-digit rounding arithmetic leads to the same results as obtained in the first example. The maximal value in the first column is 30, and the multiplier $m_{21} = \dfrac{5.291}{30} = 0.1764$ leads to the system

$$\begin{aligned} E_1: \quad 30x_1 + 591400x_2 &= \quad 591700 \\ E_2: \quad -104300x_2 &= -104400 \end{aligned}$$

which has the same inaccurate solutions $x_2 \approx 1.001$ and $x_1 \approx -10$.

Scaled partial pivoting

Scaled partial pivoting (or scaled-column pivoting) is needed for the system illustrated above.

It places the element in the pivot position that is largest relative to the entries in its row.

The first step in this procedure is to define a scale factor s_i for each row as

$$s_i = \max_{1 \leqslant j \leqslant n} |a_{ij}|.$$

Assuming that $s_i > 0$, the appropriate row interchange to place zeros in the first column is determined by choosing the least integer p with

$$\frac{|a_{p1}|}{s_p} = \max_{1 \leqslant k \leqslant n} \frac{|a_{k1}|}{s_1}$$

and performing $(E_1) \leftrightarrow (E_p)$.

Scaled partial pivoting

The effect of scaling is to ensure that the largest element in each row has a relative magnitude of 1 before the comparison for row interchange is performed.

In a similar manner, before eliminating the variable x_i using the operations $E_k - m_{ki}E_i$, for $k = i + 1, \ldots, n$, we select the smallest integer $p \geqslant i$ with

$$\frac{|a_{pi}|}{s_p} = \max_{i \leqslant k \leqslant n} \frac{|a_{ki}|}{s_k}$$

and perform the row interchange $(E_i) \leftrightarrow (E_p)$ if $i \neq p$.

The scale factors $s_1, \ldots, s_n$ are computed only once, at the start of the procedure.

They are row dependent, so they must also be interchanged when row interchanges are performed.

Scaled partial pivoting

We now apply scaled partial pivoting to the previous example. It gives

$$s_1 = \max\{|30.00|, |591400|\} = 591400, \quad s_2 = 6.130.$$

Consequently,

$$\frac{|a_{11}|}{s_1} = \frac{30}{591400} = 0.5073 \times 10^{-4}, \quad \frac{|a_{21}|}{s_2} = \frac{5.291}{6.130} = 0.8631$$

and the interchange $(E_1) \leftrightarrow (E_2)$ is made.

Applying GEM to the new system

$$\begin{aligned} E_1 &: \quad 5.291x_1 - \quad 6.13x_2 = 46.78 \\ E_2 &: \quad\quad 30x_1 + 591400x_2 = 591700 \end{aligned}$$

produces the correct results: $x_1 = 10$ and $x_2 = 1$.

Operation count

The first additional computations required for scaled partial pivoting result from the determination of the scale factors; there are $(n-1)$ comparisons for each of the n rows, for a total of $n(n-1)$ comparisons.

To determine the correct first interchange, n divisions are performed, followed by $n-1$ comparisons. So the first interchange determination adds n divisions and $n-1$ comparisons.

The scaling factors are computed only once, so the second step requires $(n-1)$ divisions and $(n-2)$ comparisons.

We proceed in a similar manner until there are zeros below the main diagonal in all but the n-th row.

The final step requires that we perform 2 divisions and 1 comparison.

Operation count

As a consequence, scaled partial pivoting adds a total of

$$n(n-1) + \sum_{k=1}^{n-1} k = n(n-1) + \frac{n(n-1)}{2} = \frac{3}{2}n(n-1)$$

comparisons and

$$\sum_{k=2}^{n} k = \frac{n(n+1)}{2} - 1 = \frac{1}{2}(n-1)(n+2)$$

divisions to the GEM.

The time required to perform a comparison is about the same as an addition/subtraction.

Operation count

The total time to perform the basic GEM is $O(n^3/3)$ multiplications/divisions and $O(n^3/3)$ additions/subtractions.

Hence, scaled partial pivoting does not add significantly to the computational time required to solve a system for large values of n.

To emphasize the importance of choosing the scale factors only once, imagine if the procedure were modified so that new scale factors were determined each time a row interchange was made.

In this case, the term $n(n-1)$ would be replaced by

$$\sum_{k=2}^{n} k(k-1) = \frac{1}{3}n(n^2-1).$$

As a consequence, this pivoting technique would add $O(n^3/3)$ comparisons, in addition to the $[n(n+1)/2] - 1$ divisions.

MA 214: Introduction to numerical analysis
Lecture 40

Shripad M. Garge.
IIT Bombay

(shripad@math.iitb.ac.in)

2021-2022

LU decomposition

The Gaußian elimination method is the principal tool in solving linear systems of equations, so it should be no surprise that it appears in other guises.

We now see that the steps used to solve a system of the form $Ax = b$ can be used to factor the matrix A.

The factorisation is particularly useful when it has the form $A = LU$, where L is lower triangular and U is upper triangular.

Although not all matrices have this type of representation, many do that occur frequently in the application of numerical techniques.

To see which matrices have an LU factorisation and to find how it is determined, first suppose that Gaußian elimination method can be performed on the system $Ax = b$ without row interchanges.

This is equivalent to having nonzero pivot elements a_{ii}, for each i.

LU decomposition

The first step in the Gaussian elimination method consists of performing, for each $j = 2, 3, ..., n$, the operations

$$(E_j - m_{j1}E_1) \to (E_j), \text{ where } m_{j1} = \frac{a_{j1}}{a_{11}}.$$

These operations transform the system into one in which all the entries in the first column below the diagonal are zero.

The above system of operations can be viewed in another way.

It is simultaneously accomplished by multiplying the original matrix A on the left by the matrix

$$M^{(1)} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -m_{21} & 1 & & 0 \\ \vdots & & \ddots & 0 \\ -m_{n1} & 0 & \cdots & 1 \end{pmatrix}.$$

LU decomposition

It is a matrix which differs from the identity matrix only in the first column, below the diagonal element, where the entry 0 is replaced by the negatives of multipliers, $-m_{j1}$.

We then have

$$M^{(1)}Ax = M^{(1)}b \text{ which gives } A^{(2)}x = b^{(2)}.$$

In a similar manner, we construct $M^{(2)}$, which is obtained by replacing the entries below the diagonal in the second column of the identity matrix by the negatives of the multipliers, $-m_{j2}$.

This changes the system by

$$M^{(2)}A^{(2)}x = M^{(2)}b^{(2)} \text{ giving } A^{(3)}x = b^{(3)}.$$

LU decomposition

In general, with

$$A^{(k)}x = b^{(k)}$$

already formed, we multiply by the matrix $M^{(k)}$.

This is the matrix which differs from the identity matrix only in the k-th column, below the diagonal entry, where 0 is replaced by $-m_{jk}$ to obtain the next system

$$A^{(k+1)}x = M^{(k)}A^{(k)}x = M^{(k)}b^{(k)} = b^{(k+1)}.$$

The process ends with the formation of $A^{(n)}x = b^{(n)}$, where $A^{(n)}$ is the upper triangular matrix, the result of the Gaußian elimination method.

We have

$$A^{(n)} = M^{(n-1)}M^{(n-2)}\dots M^{(1)}A.$$

LU decomposition

The matrix $A^{(n)}$ is an upper triangular matrix.

This is the U matrix in our desired decomposition $A = LU$.

The matrix L then has to be $L = AU^{-1}$ but we will obtain it using the above matrices $M^{(k)}$.

Note that each of the $M^{(k)}$ is lower triangular and so is the inverse of each $M^{(k)}$.

The matrix $M^{(k)}$ effects the row operation $(E_j - m_{jk}E_k) \rightarrow (E_j)$ and its inverse $L^{(k)}$ should reverse this operation, hence should perform the operation $(E_j + m_{jk}E_k) \rightarrow (E_j)$.

Thus, the matrix $L^{(k)} = (M^{(k)})^{-1}$ differs from the identity matrix only in the k-th column, below the diagonal entry, where 0 is replaced by the multiplier, m_{jk}, for $j = k + 1, \dots, n$.

LU decomposition

Therefore $L = L^{(1)}L^{(2)}\cdots L^{(n-1)}$ is a lower triangular matrix and one has

$$A = LU = \left(L^{(1)}L^{(2)}\cdots L^{(n-1)}\right)\left(M^{(n-1)}M^{(n-2)}\cdots M^{(1)}A\right).$$

Thus, if the GEM can be performed on a system $Ax = b$ without any row changes, then the matrix A has a factorisation $A = LU$ as a product of a lower triangular matrix and an upper triangular matrix.

Consider the system

$$\begin{pmatrix} 1 & 1 & 0 & 3 \\ 2 & 1 & -1 & 1 \\ 3 & -1 & -1 & 2 \\ -1 & 2 & 3 & -1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ -3 \\ 4 \end{pmatrix}.$$

An example

We obtain the LU factorisation. The first operations are: $(E_2 - 2E_1) \to (E_2)$, $(E_3 - 3E_1) \to (E_3)$, $(E_4 - (-1)E_1) \to (E_4)$. This gives

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 & 3 \\ 0 & -1 & -1 & -5 \\ 0 & -4 & -1 & -7 \\ 0 & 3 & 3 & 2 \end{pmatrix} = \left(L^{(1)}\right)\left(M^{(1)}A\right).$$

The next operations are: $(E_3 - 4E_2) \to (E_3)$, $(E_4 - (-3)E_2) \to (E_4)$ which gives

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 4 & 1 & 0 \\ 0 & -3 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 & 3 \\ 0 & -1 & -1 & -5 \\ 0 & 0 & 3 & 13 \\ 0 & 0 & 0 & -13 \end{pmatrix}$$

$$A = \left(L^{(1)}L^{(2)}\right)\left(M^{(2)}M^{(1)}A\right).$$

The example, continued

The resulting matrix $M^{(2)}M^{(1)}A$ is upper triangular now. Hence $U = M^{(2)}M^{(1)}A$ and $L = L^{(1)}L^{(2)}$. We thus get

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 4 & 1 & 0 \\ -1 & -3 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 & 3 \\ 0 & -1 & -1 & -5 \\ 0 & 0 & 3 & 13 \\ 0 & 0 & 0 & -13 \end{pmatrix}.$$

To solve the system $Ax = LUx = b$, we introduce the variable y, defined by $y = Ux$. We solve for y first from the system $Ly = b$:

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 4 & 1 & 0 \\ -1 & -3 & 0 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 8 \\ 7 \\ 14 \\ -7 \end{pmatrix}.$$

This gives $y_1 = 8$, $y_2 = -9$, $y_3 = 26$ and $y_4 = -26$.

The example, continued

We then solve for $Ux = y$

$$\begin{pmatrix} 1 & 1 & 0 & 3 \\ 0 & -1 & -1 & -5 \\ 0 & 0 & 3 & 13 \\ 0 & 0 & 0 & -13 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 8 \\ -9 \\ 26 \\ -26 \end{pmatrix}.$$

Using backward substitution, we obtain $x_4 = 2$, $x_3 = 0$, $x_2 = -1$ and $x_1 = 3$.

Observe that we had the diagonal entry in the matrix L, $l_{ii} = 1$ for all i in this method.

The methods differ by putting conditions on the diagonal elements of L. The one we studied just now had $l_{ii} = 1$. Some other method demands that $u_{ii} = 1$ and yet another demands that $l_{ii} = u_{ii}$.

Permutation matrices

Now, we worry ourselves with the question about row interchange.

If the row interchange is not needed then A can be factored as $A = LU$.

Performing a row interchange $(E_i) \leftrightarrow (E_j)$ of the system $Ax = b$ is achieved by multiplying to the system on the left by a certain matrix P.

This matrix P differs from the identity matrix only in 4 places, $p_{ii} = p_{jj} = 0$ and $p_{ji} = p_{ji} = 1$.

Such a matrix is called a permutation matrix.

In general, a permutation matrix is a matrix whose every row and every column has only one non-zero entry which is equal to 1.

Permutation matrices

We write the permutation matrices of size 2.

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

There will be 6 permutation matrices of size 3, and in general $n!$ of size n.

Note that the product of two permutation matrices is a permutation matrix, the determinant of a permutation matrix is ± 1 and the inverse of a permutation matrix is again a permutation matrix.

If we knew the row interchanges that would be needed to solve a particular system by GEM, then we could arrange the original equations in an order that would ensure that no row operations are required.

PLU decomposition

This means that for an invertible matrix A, a permutation matrix P exists such that the system $PAx = Pb$ does not require row interchanges.

Then the Gaußian elimination method can be applied to it.

As a consequence, PA can be factored as LU where L is a lower triangular matrix and U is an upper triangular matrix.

Then we get

$$A = P^{-1}LU = P'LU$$

which is a decomposition of A as a product of a permutation matrix P', a lower triangular matrix L and an upper triangular matrix M, in that order.

This is called a PLU decomposition.

MA 214: Introduction to numerical analysis
Lecture 41

Shripad M. Garge.
IIT Bombay

(shripad@math.iitb.ac.in)

2021-2022

PLU decomposition

Given a system of linear equations $Ax = b$, if we knew the row interchanges that would be needed to solve it by GEM, then we could arrange the original equations in an order that would ensure that no row operations are required.

This means that for an invertible matrix A, a permutation matrix P exists such that the system $PAx = Pb$ does not require row interchanges. Then the matrix PA can be factored as LU where L is a lower triangular matrix and U is an upper triangular matrix.

Then we get $A = P^{-1}LU = P'LU$ which is a decomposition of A as a product of a permutation matrix P', a lower triangular matrix L and an upper triangular matrix M, in that order.

This is called a PLU decomposition.

PLU decomposition

Consider the matrix

$$A = \begin{pmatrix} 0 & 0 & -1 & 1 \\ 1 & 1 & -1 & 2 \\ -1 & -1 & 2 & 0 \\ 1 & 2 & 0 & 2 \end{pmatrix}.$$

This matrix can not have factorisation as $A = LU$ because $a_{11} = 0$.

We therefore perform $(E_1) \leftrightarrow (E_2)$ and then $(E_3 + E_1) \rightarrow (E_3)$, $(E_4 - E_1) \rightarrow (E_4)$ to get the matrix

$$\begin{pmatrix} 1 & 1 & -1 & 2 \\ 0 & 0 & -1 & 1 \\ 0 & 0 & 1 & 2 \\ 0 & 1 & 1 & 0 \end{pmatrix}.$$

This necessitates a row interchange $(E_2) \leftrightarrow (E_4)$.

PLU decomposition

After that we perform $(E_4 + E_3) \to (E_4)$ to get the matrix

$$\begin{pmatrix} 1 & 1 & -1 & 2 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 3 \end{pmatrix}$$

which is upper triangular.

The permutation matrix associated to the row interchanges $(E_1) \leftrightarrow (E_2)$ followed by $(E_2) \leftrightarrow (E_4)$ is

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}.$$

PLU decomposition

We have

$$PA = \begin{pmatrix} 1 & 1 & -1 & 2 \\ 1 & 2 & 0 & 2 \\ -1 & -1 & 2 & 0 \\ 0 & 0 & -1 & 1 \end{pmatrix}.$$

This is the matrix on which we can perform the Gaußian elimination, the operations being the same as the ones above:

$$(E_2 - E_1) \to (E_2), (E_3 + E_1) \to (E_3) \text{ and } (E_4 + E_3) \to (E_4).$$

The nonzero multipliers for PA are consequently, $m_{21} = 1$, $m_{31} = -1$ and $m_{43} = -1$.

This gives the LU factorisation for the matrix PA.

PLU decomposition

$$PA = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & -1 & 2 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 3 \end{pmatrix} = LU.$$

The final decomposition for A as a product of a permutation matrix, a lower triangular matrix and an upper triangular matrix can be obtained as $A = P^{-1}LU$:

$$\begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & -1 & 2 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 3 \end{pmatrix}.$$

While it would be of interest to some to write $A = PLU$, it is desirable to know the matrices which admit GEM without row interchanges.

Diagonally dominant matrices

We now turn attention to two classes of matrices for which Gaussian elimination can be performed effectively without row interchanges.

An $n \times n$ matrix A is said to be diagonally dominant when

$$|a_{ii}| \geqslant \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|$$

holds for each $i = 1, 2, \ldots, n$.

A diagonally dominant matrix is said to be strictly diagonally dominant when the above inequality is strict for each n.

Diagonally dominant matrices

A strictly diagonally dominant matrix A is invertible.

Moreover, in this case, the Gaußian elimination method can be performed on any linear system of the form $Ax = b$ to obtain its unique solution without row interchanges.

Furthermore, the computations will be stable with respect to the growth of round-off errors.

The above statement is a serious theorem, which we will not prove in this course, but we need to understand that this is the reason the strictly dominant matrices are useful.

Positive definite matrices

A matrix A is called positive definite if A is symmetric and if $x^t Ax > 0$ for every $0 \neq x$.

If $A = [a_{ij}]$ and the column vector x is given by $x = [x_1, \ldots, x_n]^t$ then

$$x^t Ax = \left(\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j \right).$$

The matrix

$$A = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}$$

is a positive definite matrix.

To begin with, A is a symmetric matrix.

Positive definite matrices

Further,

$$\begin{aligned} x^t A x &= 2x_1^2 - 2x_1x_2 + 2x_2^2 - 2x_2x_3 + 2x_3^2 \\ &= x_1^2 + (x_1 - x_2)^2 + (x_2 - x_3)^2 + x_3^2. \end{aligned}$$

If $(x_1, x_2, x_3) \neq (0, 0, 0)$ then it follows that

$$x^t A x > 0$$

and hence A is a positive definite matrix.

It is clear from this example that using the definition to determine if a matrix is positive definite is difficult.

Fortunately, there are more easily verified criteria, which will be presented later in our course, for identifying members of this important class.

Positive definite matrices

The next result provides some necessary conditions that can be used to eliminate certain matrices from consideration.

If $A = [a_{ij}]$ is a positive definite matrix then

- A is invertible,
- $a_{ii} > 0$ for each i,
- $(a_{ij})^2 < a_{ii}a_{jj}$ for each $i \neq j$,
- $\max\limits_{1\leqslant k,j\leqslant n} |a_{kj}| \leqslant \max\limits_{1\leqslant i\leqslant n} |a_{ii}|$.

We should note, however, that a matrix A may satisfy all these conditions and may still not be positive definite!

We now give a necessary and sufficient criterion for a matrix to be positive definite.

Positive definite matrices

A leading principal submatrix of a matrix A is a matrix of the form

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kk} \end{pmatrix}$$

for some $1 \leqslant k \leqslant n$.

A matrix A is positive definite if and only if each leading principal submatrix of A has positive determinant.

Consider the matrix discussed in the above example:

$$A = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}.$$

Positive definite matrices

We verify that it is positive definite using the criteria.

$$A_1 = (2), \;\; A_2 = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}.$$

We then have $\det(A_1) = 2$ and $\det(A_2) = 3$.

Since $\det(A) = 4$, we have that A is positive definite.

We now record the results that make positive definite matrices interesting.

A symmetric matrix A is positive definite if and only if Gaußian elimination method can be applied to any system $Ax = b$ without interchanging rows with all pivot elements positive.

Moreover, in this case, the computations are stable with respect to the growth of round-off errors.

Positive definite matrices

The positive definite matrices admit interesting factorisations.

A matrix A is positive definite if and only if $A = LDL^t$ where L is lower triangular with 1's on the diagonal and D is a diagonal matrix with positive diagonal entries.

A matrix A is positive definite if and only if $A = LL^t$ where L is a lower triangular matrix.

A factorisation of a matrix A, $A = LU$ with $l_{ii} = u_{ii}$ is called a Cholesky factorisation.

The factorisation $A = L^t L$ is one such.

Note that being positive definite is not necessary to admit a Cholesky factorisation. The zero matrix admits one, for instance.

Cholesky factorisation

Determine the Cholesky factorisation of the positive definite matrix

$$A = \begin{pmatrix} 4 & -1 & 1 \\ -1 & 4.25 & 2.75 \\ 1 & 2.75 & 3.5 \end{pmatrix}.$$

If

$$A = \begin{pmatrix} 4 & -1 & 1 \\ -1 & 4.25 & 2.75 \\ 1 & 2.75 & 3.5 \end{pmatrix} = \begin{pmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{pmatrix} \begin{pmatrix} l_{11} & l_{21} & l_{31} \\ 0 & l_{22} & l_{32} \\ 0 & 0 & l_{33} \end{pmatrix}$$

then we get

$$\begin{pmatrix} 4 & -1 & 1 \\ -1 & 4.25 & 2.75 \\ 1 & 2.75 & 3.5 \end{pmatrix} = \begin{pmatrix} l_{11}^2 & l_{11}l_{21} & l_{11}l_{31} \\ l_{11}l_{21} & l_{21}^2 + l_{22}^2 & l_{21}l_{31} + l_{22}l_{32} \\ l_{11}l_{31} & l_{21}l_{31} + l_{22}l_{32} & l_{31}^2 + l_{32}^2 + l_{33}^2 \end{pmatrix}$$

Cholesky factorisation

We can solve for l_{ij}:

$$\begin{aligned}
a_{11} &: \quad 4 = l_{11}^2 && \Longrightarrow \quad l_{11} = 2, \\
a_{21} &: \, -1 = l_{11} l_{21} && \Longrightarrow \quad l_{21} = -0.5, \\
a_{31} &: \quad 1 = l_{11} l_{31} && \Longrightarrow \quad l_{31} = 0.5, \\
a_{22} &: 4.25 = l_{21}^2 + l_{22}^2 && \Longrightarrow \quad l_{22} = 2 \\
a_{32} &: 2.75 = l_{21} l_{31} + l_{22} l_{32} && \Longrightarrow \quad l_{32} = 1.5 \\
a_{33} &: \; 3.5 = l_{31}^2 + l_{32}^2 + l_{33}^2 && \Longrightarrow \quad l_{33} = 1
\end{aligned}$$

and hence

$$A = \begin{pmatrix} 4 & -1 & 1 \\ -1 & 4.25 & 2.75 \\ 1 & 2.75 & 3.5 \end{pmatrix} = \begin{pmatrix} 2 & 0 & 0 \\ -0.5 & 2 & 0 \\ 0.5 & 1.5 & 1 \end{pmatrix} \begin{pmatrix} 2 & -0.5 & 0.5 \\ 0 & 2 & 1.5 \\ 0 & 0 & 1 \end{pmatrix} = LL^t.$$

MA 214: Introduction to numerical analysis
Lecture 42

Shripad M. Garge.
IIT Bombay

(shripad@math.iitb.ac.in)

2021-2022

Iterative methods

The methods we used to solve systems $Ax = b$ until the last lecture are called the direct methods.

Now we want to try the analogues of iterative techniques which were so useful in finding solutions of equations in one variable.

An initial approximation (or approximations) was found, and new approximations are then determined based on how well the previous approximations satisfied the equation.

The objective is to find a way to minimise the difference between the approximations and the exact solution.

To discuss iterative methods for solving linear systems, we first need to determine a way to measure the distance between n-dimensional column vectors.

Distance in $\mathbb{R}^n$

Our underlying set is $\mathbb{R}^n$, the vector space of all column vectors of size $n \times 1$.

We define two types of distances on pairs of vectors in $\mathbb{R}^n$. Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ with

$$\mathbf{x} = (x_1, \ldots, x_n)^t \text{ and } \mathbf{y} = (y_1, \ldots, y_n)^t.$$

The l_2-distance:

$$\|\mathbf{x} - \mathbf{y}\|_2 = \left(\sum_{i=1}^{n} (x_i - y_i)^2\right)^{1/2}.$$

The l_∞-distance:

$$\|\mathbf{x} - \mathbf{y}\|_\infty = \max_{1 \leqslant i \leqslant n} |x_i - y_i|.$$

Distance in $\mathbb{R}^n$

Consider the system of linear equations:

$$\begin{aligned} 3.3330x_1 + 15920x_2 - 10.333x_3 &= 15913, \\ 2.2220x_1 + 16.710x_2 + 9.6120x_3 &= 28.544, \\ 1.5611x_1 + 5.1791x_2 + 1.6852x_3 &= 8.4254 \end{aligned}$$

which has the exact solution $\mathbf{x} = (1, 1, 1)^t$ but the Gaußian elimination method performed using five-digit rounding arithmetic and partial pivoting produces the approximate solution

$$\tilde{\mathbf{x}} = (1.2001, 0.99991, 0.92538)^t.$$

We now compute the distance of $\tilde{\mathbf{x}}$ in both the distances, l_2 and l_∞.

Distance in $\mathbb{R}^n$

We have

$$\begin{aligned}\|\mathbf{x}-\tilde{\mathbf{x}}\|_2 &= \left((0.2001)^2+(0.00009)^2+(0.07462)^2\right)^{1/2}\\ &= 0.21356\end{aligned}$$

and

$$\begin{aligned}\|\mathbf{x}-\tilde{\mathbf{x}}\|_\infty &= \max\{0.2001, 0.00009, 0.07462\}\\ &= 0.2001.\end{aligned}$$

Here, the second and the third components are good approximations but the first component is a poor approximation and that error dominates the computations of both the above distances.

The above notions of distance can be used to define convergence of sequences in $\mathbb{R}^n$.

Note that $\|\mathbf{x}-\mathbf{y}\|_\infty \leqslant \|\mathbf{x}-\mathbf{y}\|_2 \leqslant \sqrt{n}\|\mathbf{x}-\mathbf{y}\|_\infty$.

Norms on matrices

We want to generalise the notion of absolute value in $\mathbb{R}$ to the space of $n \times n$ matrices.

Note that we also have the analogues of absolute value on $\mathbb{R}^n$, $\|\mathbf{x}\|_2 = \|\mathbf{x} - 0\|_2$ and $\|\mathbf{x}\|_\infty = \|\mathbf{x} - 0\|_\infty$.

They are called the l_2 and l_∞ norms on $\mathbb{R}^n$.

We now define:

$$\|A\|_2 = \max_{\|\mathbf{x}\|_2=1} \|A\mathbf{x}\|_2, \qquad \|A\|_\infty = \max_{\|\mathbf{x}\|_\infty=1} \|A\mathbf{x}\|_\infty$$

The l_∞ norm on matrices can be computed quite easily. If $A = (a_{ij})$ then

$$\|A\|_\infty = \max_i \sum_j |a_{ij}|.$$

Norms on matrices

Compute the l_∞ norm of

$$A = \begin{pmatrix} 1 & 2 & -1 \\ 0 & 3 & -1 \\ 5 & -1 & 1 \end{pmatrix}.$$

We have

$$\sum_{j=1}^{3} |a_{1j}| = 1 + 2 + 1 = 4, \quad \sum_{j=1}^{3} |a_{2j}| = 0 + 3 + 1 = 4$$

and

$$\sum_{j=1}^{3} |a_{3j}| = 5 + 1 + 1 = 7.$$

Hence $\|A\|_\infty = 7$.

Eigenvalues, eigenvectors

An $n \times n$ matrix A acts as a function from $\mathbb{R}^n$ to $\mathbb{R}^n$.

The eigenvectors of A are the vectors that are sent to their multiples by the matrix A.

More precisely, a nonzero vector $\mathbf{v} \in \mathbb{R}^n$ is an eigenvector for A if there is a $\lambda \in \mathbb{R}$ such that $A\mathbf{v} = \lambda\mathbf{v}$.

A real number $\lambda \in \mathbb{R}$ is called an eigenvalue of A if there is a non-zero $\mathbf{v} \in \mathbb{R}^n$ with $A\mathbf{v} = \lambda\mathbf{v}$.

Now, if λ is an eigenvalue of A then $A - \lambda I$ is not invertible, because it sends a nonzero $\mathbf{v}$ to zero. Hence $\det(A - \lambda I) = 0$.

Thus, the eigenvalues of A are the roots of the characteristic polynomial of A which is the $\det(A - \lambda I)$.

Eigenvalues, eigenvectors

Determine the eigenvalues and eigenvectors of

$$A = \begin{pmatrix} 2 & 0 & 0 \\ 1 & 1 & 2 \\ 1 & -1 & 4 \end{pmatrix}.$$

Note that the characteristic polynomial of A is

$$-(\lambda^3 - 7\lambda^2 + 16\lambda - 12) = -(\lambda - 3)(\lambda - 2)^2.$$

Thus the eigenvalues of A are 2 and 3.

Now we compute the corresponding eigenvectors.

The eigenvectors for λ are the vectors $\mathbf{v}$ with $(A - \lambda I)(\mathbf{v}) = 0$.

Eigenvalues, eigenvectors

$\lambda = 3$: $(A - 3I)(\mathbf{v}) = 0$ gives

$$\begin{pmatrix} -1 & 0 & 0 \\ 1 & -2 & 2 \\ 1 & -1 & 1 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

This gives

$$v_1 = 0 \text{ and } v_2 = v_3$$

hence an eigenvector for eigenvalue 3 is of the form $c(0,1,1)^t$ for $0 \neq c \in \mathbb{R}$.

$\lambda = 2$: A similar analysis as above tells us that an eigenvector for eigenvalue 2 is a non-zero vector $\mathbf{v}$ satisfying $v_1 - v_2 + 2v_3 = 0$.

In particular, it is of the form $c_1(0,2,1)^t + c_2(-2,0,1)^t$ where $(c_1, c_2) \neq (0,0)$.

Spectral radius

Note that even though our A has real entries, some roots of the characteristic polynomial of A may be (non-real) complex numbers.

The spectral radius of A, $\rho(A)$, is defined by

$$\rho(A) = \max |\lambda|$$

where λ varies over all roots of the characteristic polynomial of A.

If A is an $n \times n$ matrix then

- $\|A\|_2 = \left[\rho(A^t A)\right]^{1/2}$
- $\rho(A) \leqslant \|A\|_2$ and
- $\rho(A) \leqslant \|A\|_\infty$.

Let us now compute l_2-norm of a matrix.

Spectral radius

Let

$$A = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ -1 & 1 & 2 \end{pmatrix}.$$

We apply the above result, and for that we need to compute $\rho(A^tA)$.

Here

$$A^tA = \begin{pmatrix} 1 & 1 & -1 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ -1 & 1 & 2 \end{pmatrix} = \begin{pmatrix} 3 & 2 & -1 \\ 2 & 6 & 4 \\ -1 & 4 & 5 \end{pmatrix}$$

and the eigenvalues of A^tA are $0, 7 \pm \sqrt{7}$. Hence

$$\|A\|_2 = \sqrt{\rho(A^tA)} = \sqrt{7 + \sqrt{7}} \approx 3.10576.$$

Convergent matrices

In studying iterative matrix techniques, it is of particular importance to know when powers of a matrix become small, that is, when all the entries approach zero.

Matrices of this type are called convergent.

An $n \times n$ matrix A is called convergent if for each $1 \leqslant i, j \leqslant n$

$$\lim_{k\to\infty} (A^k)_{ij} = 0.$$

For instance, if

$$A = \begin{pmatrix} 1/2 & 0 \\ 1/4 & 1/2 \end{pmatrix} \text{ then } A^k = \begin{pmatrix} 1/2^k & 0 \\ k/2^{k+1} & 1/2^k \end{pmatrix}.$$

It then follows that A is a convergent matrix.

Convergent matrices

The following statements are equivalent:

1. A is a convergent matrix.
2. $\lim_{n\to\infty} \|A^n\|_2 = 0$.
3. $\lim_{n\to\infty} \|A^n\|_\infty = 0$.
4. $\rho(A) < 1$.
5. $\lim_{n\to\infty} A^n\mathbf{v} = 0$ for every vector $\mathbf{v}$.