Stiffness matrix from basics: beam elements

$$EI\frac{d^4w}{dx^4} = p(x) = 0$$

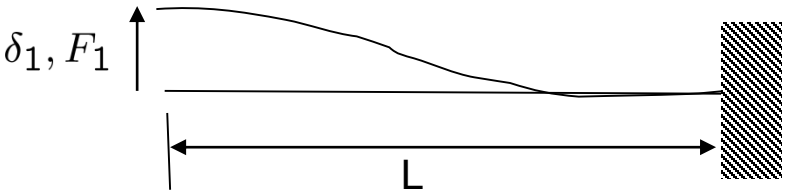
$$w = \frac{1}{6}C_1x^3 + \frac{1}{2}C_2x^2 + C_3x + C_4$$

At $x = 0$ $\quad \frac{dw}{dx} = 0 \Rightarrow C_3 = 0$

$\theta_1 = \theta_2 = \delta_2 = 0$

At $x = L$ $\quad \frac{dw}{dx} = 0, w = 0 \Rightarrow$

$$C_1L = -2C_2, C_4 = -\frac{1}{6}C_2L^2$$



$$\begin{aligned} \frac{d^2w}{dx^2} &= \frac{M(x)}{EI} \\ \frac{dM}{dx} &= V \\ \Rightarrow \qquad & EI\frac{d^3w}{dx^3} = V. \end{aligned} \tag{1}$$

At $x = 0$

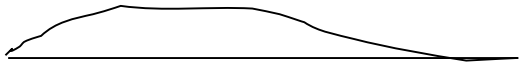
$$\frac{d^3w}{dx^3} = \frac{F_1}{EI} \Rightarrow C_1 = \frac{F_1}{EI}$$

Thus,

$$K_{11} = \frac{F_1}{\delta_1} = \frac{F_1}{w(x=0)} = \frac{12EI}{L^3}$$

Intuitive but not easy

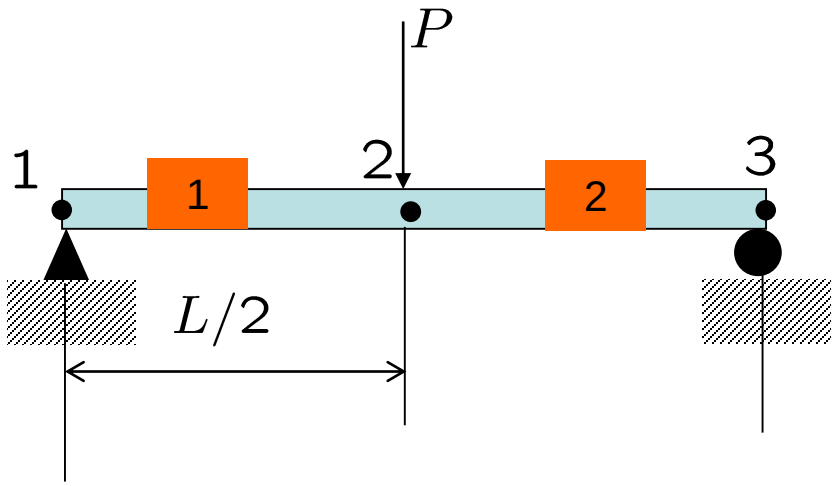
$$\boldsymbol{K}_L = \frac{EI}{L}\begin{pmatrix} 12/L^2 & -6/L & -12/L^2 & -6/L \\ & 4 & 6/L & 2 \\ & & 12/L^2 & 6/L \\ & & & 4 \end{pmatrix}$$



22

12

13



For each element

$$\boldsymbol{K}_L^{1/2} = \frac{8EI}{L^3}\begin{pmatrix} 12 & 6(L/2) & -12 & 6(L/2) \\ 6L/2 & 4(L/2)^2 & -6(L/2) & 2(L/2)^2 \\ -12 & -6(L/2) & 12 & -6(L/2) \\ 6(L/2) & 2(L/2)^2 & -6(L/2) & 4(L/2)^2 \end{pmatrix}$$

The assembled global stiffness

$$\boldsymbol{K}_G = \frac{8EI}{L^3}\begin{pmatrix} 12 & 6(L/2) & -12 & 6(L/2) & 0 & 0 \\ 6L/2 & 4(L/2)^2 & -6(L/2) & 2(L/2)^2 & 0 & 0 \\ -12 & -6(L/2) & 12+12 & -6(L/2)-6(L/2) & -12 & 6(L/2) \\ 6(L/2) & 2(L/2)^2 & -6(L/2)+6(L/2) & 4(L/2)^2+4(L/2)^2 & -6(L/2) & 2(L/2)^2 \\ 0 & 0 & -12 & -6(L/2) & 12 & -6(L/2) \\ 0 & 0 & 6(L/2) & 2(L/2)^2 & -6(L/2) & 4(L/2)^2 \end{pmatrix}$$

The global force vector

$$\boldsymbol{F} = \begin{Bmatrix} 0 \\ 0 \\ -P \\ 0 \\ 0 \\ 0 \end{Bmatrix}$$

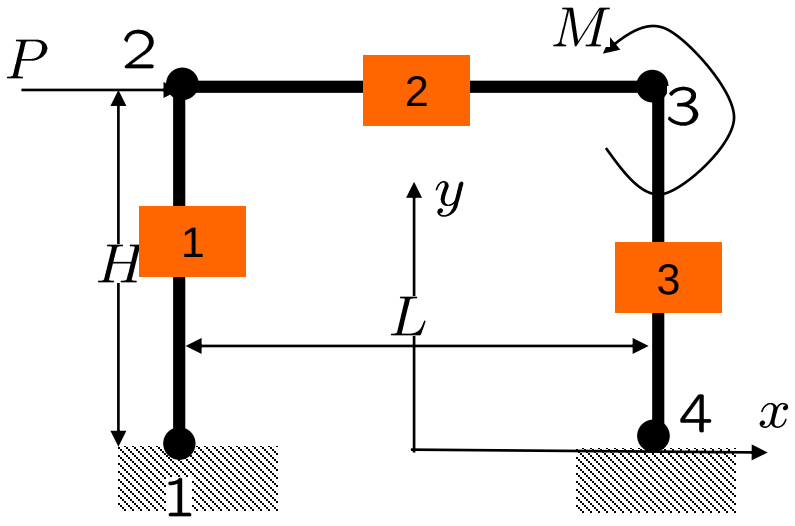
The global displacement vector

$$\boldsymbol{F} = \begin{Bmatrix} w_1 = 0 \\ \theta_1 \\ w_2 \\ \theta_2 \\ w_3 = 0 \\ \theta_3 \end{Bmatrix}$$

Finally we solve

$$\frac{8EI}{L^3}\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 4(L/2)^2 & -6(L/2) & 2(L/2)^2 & 0 & 0 \\ 0 & -6(L/2) & 12+12 & -6(L/2)-6(L/2) & 0 & 6(L/2) \\ 0 & 2(L/2)^2 & -6(L/2)+6(L/2) & 4(L/2)^2+4(L/2)^2 & 0 & 2(L/2)^2 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 6(L/2) & 2(L/2)^2 & 0 & 4(L/2)^2 \end{pmatrix} \begin{Bmatrix} w_1 \\ \theta_1 \\ w_2 \\ \theta_2 \\ w_3 \\ \theta_3 \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \\ -P \\ 0 \\ 0 \\ 0 \end{Bmatrix}$$

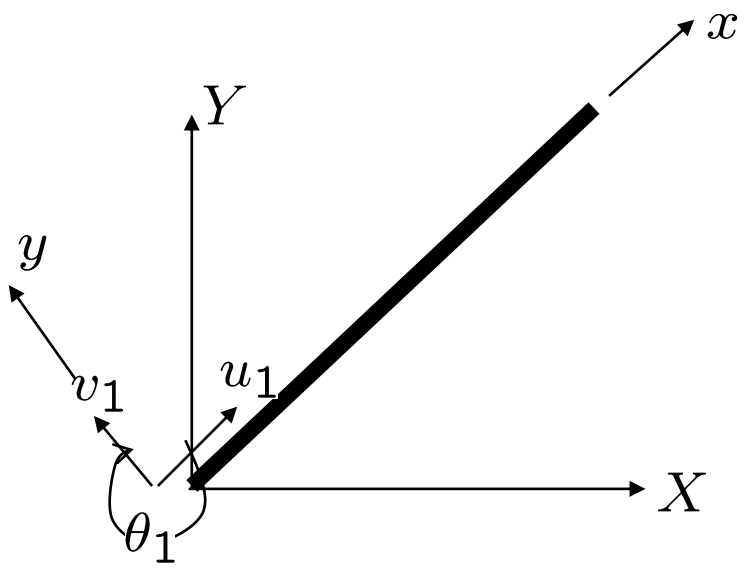
Beams in 2-d



The local stiffness for any element (say 2)

$$\boldsymbol{K}_L^2 = \begin{pmatrix} AE/L & 0 & 0 & -AE/L & 0 & 0 \\ 0 & 12EI/L^3 & 6EI/L^2 & 0 & -12EI/L^3 & 6EI/L^2 \\ 0 & 6EI/L^2 & 4EI/L & 0 & -6EI/L^2 & 2EI/L \\ -AE/L & 0 & 0 & AE/L & 0 & 0 \\ 0 & -12EI/L^3 & -6EI/L^2 & 0 & 12EI/L^3 & -6EI/L^2 \\ 0 & -6EI/L^2 & -4EI/L & 0 & 6EI/L^2 & -2EI/L \end{pmatrix}$$

Transformation Matrix for a 2-d beam element



$$\boldsymbol{u}_1 = \left\{ \begin{array}{c} u_1 \\ v_1 \\ \theta_1 \end{array} \right\} = \left(\begin{array}{ccc} \cos\phi & \sin\phi & 0 \\ -\sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{array} \right) \left\{ \begin{array}{c} U_1 \\ V_1 \\ \Theta_1 \end{array} \right\} = \boldsymbol{t}\boldsymbol{U}$$

$$\boldsymbol{u} = \left(\begin{array}{cc} \boldsymbol{t} & 0 \\ 0 & \boldsymbol{t} \end{array} \right) \boldsymbol{U} = \boldsymbol{T}\boldsymbol{U}$$

$$\boldsymbol{K}_G = \boldsymbol{T}^T \boldsymbol{K}_L \boldsymbol{T}$$

Solving the equations

The stiffness matrix in the worst case is $N \times N$. We can take advantage of its sparsity, diagonal dominance and symmetry to design storage and solution methodologies. The basic problem is to start from

$$\boldsymbol{K}\boldsymbol{U} = \boldsymbol{F}$$

and determine the unknown displacements $\boldsymbol{U}$ by

$$\boldsymbol{U} = \boldsymbol{K}^{-1}\boldsymbol{F}$$

We will look at a few numerical techniques for tackling this problem. In particular, we will look at direct methods. There are other iterative methods like conjugate gradient schemes that are suitable for FE calculations in large parallel computing environments.

Basic Gauss elimination:an example

$$\begin{pmatrix} 18 & -6 & -6 & 0 \\ -6 & 12 & 0 & -6 \\ -6 & 0 & 12 & -6 \\ 0 & -6 & -6 & 12 \end{pmatrix} \begin{Bmatrix} U_1 \\ U_2 \\ U_3 \\ U_4 \end{Bmatrix} = \begin{Bmatrix} 60 \\ 0 \\ 20 \\ 0 \end{Bmatrix}$$

Original problem

$$\begin{pmatrix} 18 & -6 & -6 & 0 \\ 0 & 10 & -2 & -6 \\ 0 & -2 & 10 & -6 \\ 0 & -6 & -6 & 12 \end{pmatrix} \begin{Bmatrix} U_1 \\ U_2 \\ U_3 \\ U_4 \end{Bmatrix} = \begin{Bmatrix} 60 \\ 20 \\ 40 \\ 0 \end{Bmatrix}$$

1st elimination:

Row2- Row 1*(-6/18)

Row3-Row1*(-6/18)

$$\begin{pmatrix} 18 & -6 & -6 & 0 \\ 0 & 10 & -2 & -6 \\ 0 & 0 & 9.6 & -7.2 \\ 0 & 0 & -7.2 & 8.4 \end{pmatrix} \begin{Bmatrix} U_1 \\ U_2 \\ U_3 \\ U_4 \end{Bmatrix} = \begin{Bmatrix} 60 \\ 20 \\ 44 \\ 12 \end{Bmatrix}$$

2nd elimination

Row3-Row2*(-2/10)

Row4-Row2*(-6/10)

$$\begin{pmatrix} 18 & -6 & -6 & 0 \\ 0 & 10 & -2 & -6 \\ 0 & 0 & 9.6 & -7.2 \\ 0 & 0 & 0 & 3 \end{pmatrix} \begin{Bmatrix} U_1 \\ U_2 \\ U_3 \\ U_4 \end{Bmatrix} = \begin{Bmatrix} 60 \\ 20 \\ 44 \\ 45 \end{Bmatrix}$$

3rd elimination:

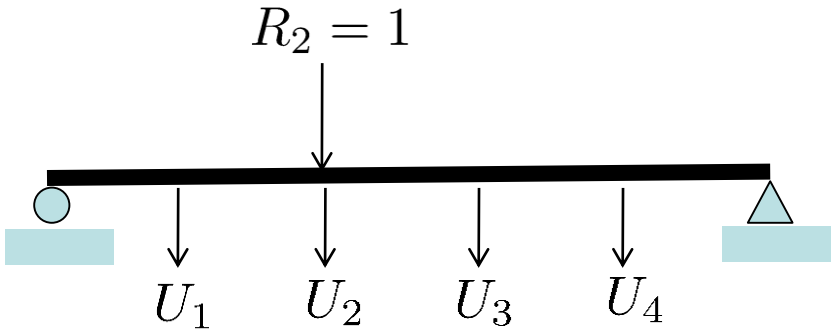
Row4-Row3*(-7.2/9.6)

To solve for $u_1, \ldots u_4$, we now *back substitute*:

$$\begin{aligned} u_4 &= 45/3 \\ u_3 &= (44 + 7.2U_4)/9.6 = 15.83 \\ u_2 &= (20 + 2U_3 + 6U_4)/10 = 14.17 \\ u_1 &= (60 + 6U_2 + 6U_3)/18 = 13.33 \end{aligned}$$

Note that, at any stage if we encounter a zero on the diagonal, the procedure will fail. It is advisable to resort to *pivoting* to take care of this issue.

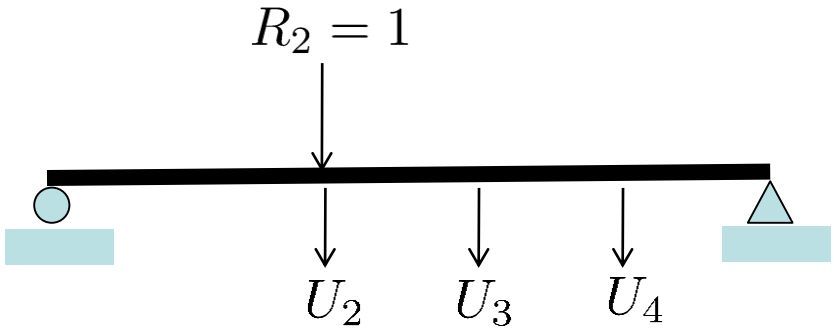
A physical picture of the Gauss elimination procedure



To illustrate the Gauss elimination procedure, consider a simply supported beam as shown in the figure. We wish to solve for dof's $U_1 \ldots U_4$. Assume that the final equation of equilibrium for this problem is

$$\begin{pmatrix} 5 & -4 & 1 & 0 \\ -4 & 6 & -4 & 1 \\ 1 & -4 & 6 & -4 \\ 0 & 1 & -4 & 5 \end{pmatrix} \begin{Bmatrix} U_1 \\ U_2 \\ U_3 \\ U_4 \end{Bmatrix} = \begin{Bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{Bmatrix}$$

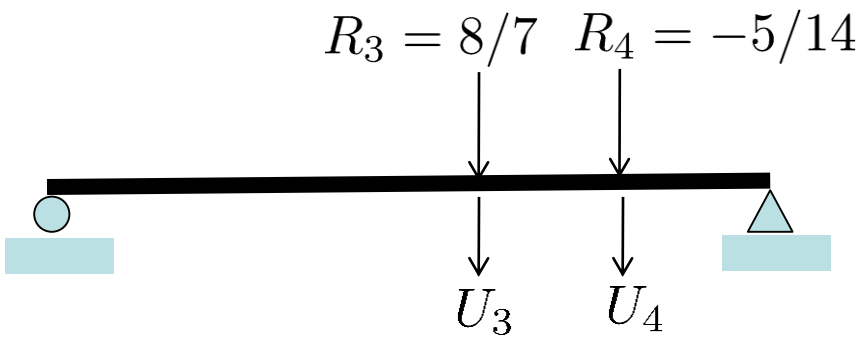
The first elimination will be equivalent to getting rid of the first dof U_1 by modifying the load accordingly.



Thus the physical problem being solved after the first elimination is

$$\begin{pmatrix} \frac{14}{5} & -\frac{16}{5} & 1 \\ -\frac{16}{5} & \frac{29}{5} & -4 \\ 1 & -4 & 5 \end{pmatrix} \begin{Bmatrix} U_2 \\ U_3 \\ U_4 \end{Bmatrix} = \begin{Bmatrix} 1 \\ 0 \\ 0 \end{Bmatrix}$$

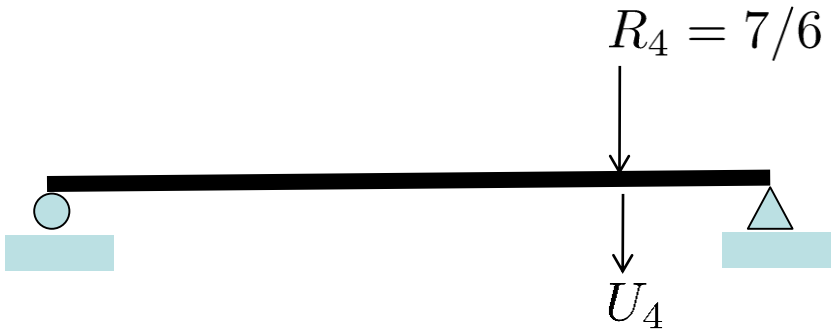
The physical problem now has 3 dof's.



The next elimination results in:

$$\begin{pmatrix} \frac{15}{7} & -\frac{20}{7} \\ -\frac{20}{7} & \frac{65}{14} \end{pmatrix} \begin{Bmatrix} U_3 \\ U_4 \end{Bmatrix} = \begin{Bmatrix} \frac{8}{7} \\ -\frac{5}{14} \end{Bmatrix}$$

Note now that a different loading has to be applied to maintain the equivalence with the original problem.



Finlly we get

$$\frac{5}{6}U_4 = \frac{7}{6}.$$

Thus, from a physical standpoint, we can arge that we should not encounter a zero diagonal element while performing a Gauss elimination on a FE stiffness matrix. This is because the i th diagonal at any stage of elimination represents the stiffness of the structure when $(i-1)$ dof's have been released. This stiffness should be positive if the structure is stable.

The numerics of Gauss elimination

In this chapter we will look at methods to solve the linear system of simultaneous equations

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b},$$

where, the bold capitals denote matrices and bold small symbols denote vectors.

It is particularly easy to solve a $n \times n$ *upper triangular system* like

$$\begin{aligned}
a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + a_{14}x_4 \cdots + a_{1n}x_n &= b_1 \\
a_{22}x_2 + a_{23}x_3 + a_{24}x_4 \cdots + a_{2n}x_n &= b_2 \\
a_{33}x_3 + a_{34}x_4 \cdots + a_{3n}x_n &= b_3 \\
&\vdots \\
a_{nn}x_n &= b_n
\end{aligned}$$

Clearly,

$$x_n = \frac{b_n}{a_{nn}}.$$

Given that $a_{jj} \neq 0$, for $j = n-1, \ldots 1$, we can perform *backward substitution* to obtain:

$$x_j = \frac{1}{a_{jj}} \left(b_j - \sum_{k=j+1}^{n} a_{jk} x_k \right).$$

To count the number of operations, we assume that multiplication/division requires the longest time (though, division takes somewhat longer than multiplication in a computer) and addition/subtraction takes short times. In each step, we have $n - j$ multiplications and 1 division. Thus, for the entire back substitution process, we have

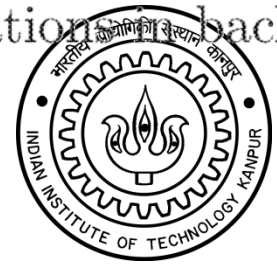
$$n^2 - \frac{n(n-1)}{2} + n = \frac{n^2}{2} + \frac{3n}{2},$$

multiplications/divisions. We also have $n - j - 1$ additions and 1 subtraction per step, which leads to

$$\frac{n^2 + n}{2}$$

addition/subtraction operations. The number of arithmetic operations in back substitution is:

$$\frac{n^2}{2} + O(n)$$



In the *Gauss elimination method*, a series of "elimination" steps transforms the system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ to an upper traingular system $\boldsymbol{R}\boldsymbol{x} = \boldsymbol{c}$, which is mathematically equivalent (may not be exactly equivalent due to round off errors) to the original system. Two operations *permutation of rows* and *scaling of rows* are performed in the elimination process. The process goes through the following steps:

- Form the augmented matrix
$$[\boldsymbol{A}|\boldsymbol{b}]\,.$$
- Set
$$\boldsymbol{A}^0 = \boldsymbol{A}, \boldsymbol{b}^0 = \boldsymbol{b}$$
- For $r \in 1, \ldots n$, find $a_{r1} \neq 0$. Such an element must exist as otherwise $\boldsymbol{A}$ will be singular. Permute 1st and r th rows. Let the new augmented matrix be
$$\left[\tilde{\boldsymbol{A}^0}|\tilde{\boldsymbol{b}^0}\right].$$
- For $j = 2, \ldots n$, multiply row 1 by q_{j1} and subtract the result from j th row, where,
$$q_{j1} = \frac{\tilde{a}^0_{j1}}{\tilde{a}^0_{11}} \quad a^1_{jk} = \tilde{a}^0_{jk} - q_{j1}\tilde{a}^0_{1k}, b^1_j = \tilde{b}^0_j - q_{j1}\tilde{b}^0_1.$$

After the last step, the result is

$$[\boldsymbol{A}^1|\boldsymbol{b}^1] = \left(\begin{array}{cccc|c} \tilde{a}^0_{11} & \tilde{a}^0_{12} & \dots & \tilde{a}^0_{1n} & b^0_1 \\ 0 & a^1_{21} & \dots & a^1_{2n} & b^1_2 \\ & & & & \\ 0 & \vdots & \vdots & \vdots & \vdots \\ 0 & a^1_{n1} & \dots & a^1_{nn} & b^1_n \end{array}\right)$$

We can express the transformation $[\boldsymbol{A}|\boldsymbol{b}] \rightarrow [\tilde{\boldsymbol{A}}^0|\tilde{\boldsymbol{b}}^0]$ as $[\tilde{\boldsymbol{A}}^0|\tilde{\boldsymbol{b}}^0] = \boldsymbol{P}_1[\boldsymbol{A}|\boldsymbol{b}]$, where $\boldsymbol{P}$ is a *permutation matrix*. This matrix has the form

$$\begin{array}{c} \\ \mathbf{1} \\ \\ \\ \\ \\ \mathbf{r} \\ \\ \\ \end{array}\!\!\begin{array}{c} \begin{array}{ccccccc} \mathbf{1} & & & & & \mathbf{r} & \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & \dots & 0 & 1 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 & \dots & 0 \\ \vdots & & & \ddots & \vdots & \dots & \dots & 0 \\ & & & & \ddots & \vdots & \dots & 0 \\ 1 & & \dots & & & 0 & \dots & 0 \\ & & & \vdots & & & \ddots & \vdots \\ 0 & & & & & & & 1 \end{pmatrix} \end{array}$$

भारतीय प्रौद्योगिकी संस्थान कानपुर INDIAN INSTITUTE OF TECHNOLOGY KANPUR

Similarly, we can represent $[\boldsymbol{A}^1|\boldsymbol{b}^1] = \boldsymbol{G}_1[\tilde{\boldsymbol{A}}^0|\tilde{\boldsymbol{b}}^0]$, where $\boldsymbol{G}_1$ is called a *Frobenius matrix* and is given by:

$$\begin{pmatrix} 1 & & & \\ -q_{21} & 1 & & \\ \vdots & & \ddots & \\ -q_{n1} & \dots & & 1 \end{pmatrix}$$

It can be shown that $\det \boldsymbol{P}_1 = \det \boldsymbol{G}_1 = 1$ and $\boldsymbol{P}_1^{-1} = \boldsymbol{P}_1$. The inverse of $\boldsymbol{G}_1$ is

$$\boldsymbol{G}_1^{-1} = \begin{pmatrix} 1 & & & \\ q_{21} & 1 & & \\ \vdots & & \ddots & \\ q_{n1} & \dots & & 1 \end{pmatrix}$$

Clearly, $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ has the same solution as $\boldsymbol{A}^1\boldsymbol{x} = \boldsymbol{b}^1$ as

$$\boldsymbol{G}_1\boldsymbol{P}_1\boldsymbol{A}\boldsymbol{x} = \boldsymbol{G}_1\boldsymbol{P}_1\boldsymbol{b} = \boldsymbol{b}^1.$$

The series of steps

$$[\boldsymbol{A}|\boldsymbol{b}] \to [\boldsymbol{A}^1|\boldsymbol{b}^1] \to \cdots \to [\boldsymbol{A}^{n-1}|\boldsymbol{b}^{n-1}] \to [\boldsymbol{R}|\boldsymbol{c}],$$

leads to an upper triangular matrix $\boldsymbol{R}$, where

$$[\boldsymbol{R}|\boldsymbol{c}] = \boldsymbol{G}_{n-1}\boldsymbol{P}_{n-1}\ldots\boldsymbol{G}_1\boldsymbol{P}_1[\boldsymbol{A}|\boldsymbol{b}],$$

is equivalent to the original system. Here,

$$\boldsymbol{P}_i = \begin{pmatrix} 1 & & & & & & & \\ & \ddots & & & & & & \\ & & 1 & & & & & \\ & & & 0 & \ldots & 1 & \ldots & \\ & & & \vdots & 1 & & & \\ & & & \vdots & & & & \\ & & & 1 & \ldots & 0 & & \\ & & & & & & 1 & \\ & & & & & & & \ddots \\ & & & & & & & & 1 \end{pmatrix} \begin{matrix} \\ \\ \\ \text{i} \\ \\ \\ \text{r} \\ \\ \\ \end{matrix} \quad \begin{matrix} \\ \\ \\ \text{i} \\ \\ \\ \text{r} \\ \\ \\ \end{matrix}$$

$$\boldsymbol{G}_i = \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & -q_{i+1,i} & 1 & \\ & & \vdots & & \ddots \\ & & -q_{ni} & & & 1 \end{pmatrix} \begin{matrix} \\ \\ \text{i} \\ \\ \\ \\ \end{matrix}$$

The element $a_{r1} = \tilde{a}_{11}^0$ is called the *pivot element* and is chosen so that

$$|a_{r1}| = \max_{1 \le j \le n} |a_{j1}|.$$

Permutation of rows is called *column pivoting*. Though *total pivoting* where columns are also permuted is possible, it is expensive and rarely used.

We can perform an operation count for Gauss elimination. For simplicity, we will not count the steps required for pivoting. Thus,

- The kth elimination step involves

$$a_{ij}^k = a_{ij}^{k-1} - \frac{a_{ik}^{k-1}}{a_{kk}^{k-1}} a_{kj}^{k-1}, b_i^k = b_i^{k-1} - \frac{a_{ik}^{k-1}}{a_{kk}^{k-1}} b_k^{k-1}$$

  for $i, j = k \ldots n$.

  - This involves $n - k$ divisions to form $a_{ik}^{k-1}/a_{kk}^{k-1}$
  - followed by $(n-k)(n-k+1)$ multiplications.
  - Total number of multiplications/divisions over $k = 1, \ldots n$

$$\sum_{k=1}^{n-1} (n-k)(n-k+2) = \frac{2n^3 + 3n^2 - 5n}{6} = O(n^3)$$

Example

As an example, we solve the system

$$\begin{pmatrix} 3 & 1 & 6 \\ 2 & 1 & 3 \\ 1 & 1 & 1 \end{pmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ x_3 \end{Bmatrix} = \begin{Bmatrix} 2 \\ 7 \\ 4 \end{Bmatrix}$$

Step 1: pivoting

$$\left(\begin{array}{ccc|c} \boxed{3} & 1 & 6 & 2 \\ 2 & 1 & 3 & 7 \\ 1 & 1 & 1 & 4 \end{array}\right)$$

Step 2: elimination

Instead of 0 a_{j1} can be stored.

$$\left(\begin{array}{ccc|c} 3 & 1 & 6 & 2 \\ 2/3 & 1/3 & -1 & 17/3 \\ 1/3 & \boxed{2/3} & -1 & 10/3 \end{array}\right)$$

भारतीय प्रौद्योगिकी संस्थान कानपुर INDIAN INSTITUTE OF TECHNOLOGY KANPUR

Step 3: pivoting

$$\left(\begin{array}{ccc|c} 3 & 1 & 6 & 2 \\ 1/3 & 2/3 & -1 & 10/3 \\ 2/3 & 1/3 & -1 & 17/3 \end{array}\right)$$

Step 4: elimination

$$\left(\begin{array}{ccc|c} 3 & 1 & 6 & 2 \\ 1/3 & 2/3 & -1 & 10/3 \\ 2/3 & 1/2 & -1/2 & 4 \end{array}\right)$$

Step 5: back substitute

$$\begin{aligned} x_3 &= -8 \\ x_2 &= \frac{3}{2}\left(\frac{10}{3} - x_3\right) = -7 \\ x_1 &= \frac{1}{3}\left(2 - x_2 - 6x_3\right) = 19 \end{aligned}$$

Solution also using the demo code `gauss_elim.m`

After the elimination is complete, the transformed matrix can be written as

$$\begin{pmatrix} 1 & 0 & 0 \\ 1/3 & 1 & 0 \\ 2/3 & 1/2 & 1 \end{pmatrix} \begin{pmatrix} 3 & 1 & 6 \\ 0 & 2/3 & -1 \\ 0 & 0 & -1/2 \end{pmatrix} = \begin{pmatrix} 3 & 1 & 6 \\ 1 & 1 & 1 \\ 2 & 1 & 3 \end{pmatrix},$$

which is same as the matrix we started with (with two rows permuted). This is called a LU decomposition.

Solution also using the demo code `LU.m`

Formally,

$$\boldsymbol{L} = \begin{pmatrix} l_{11} & & & 0 \\ l_{21} & 1 & & \\ \vdots & \ddots & \ddots & \\ l_{n1} & \cdots & l_{n.n-1} & \end{pmatrix}, \boldsymbol{U} = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ & r_{22} & \cdots & r_{2n} \\ & & \ddots & \\ 0 & & & r_{nn} \end{pmatrix}$$

are the factor in a LU decomposition such that

$$\boldsymbol{LU} = \boldsymbol{PA},$$

where,

$$\boldsymbol{P} = \boldsymbol{P}_{n-1} \ldots \boldsymbol{P}_1.$$

In the i th elimination step, i.e. $[\boldsymbol{A}^{i-1}|\boldsymbol{b}^{i-1}] \rightarrow [\boldsymbol{A}^{i}|\boldsymbol{b}^{i}]$, we store the factors $q_{i+1,i} \ldots q_{n,i}$ of the Frobenius matrix $\boldsymbol{G}_i^{-1}$ in place of the zeros created. Thus, we work with the augmented matrix:

$$\left(\begin{array}{ccccccc|c}
r_{11} & r_{12} & \cdots & r_{1i} & r_{1,i+1} & \cdots & r_{1n} & c_1 \\
\lambda_{21} & r_{22} & \cdots & r_{2i} & r_{2,i+1} & \cdots & r_{2n} & c_2 \\
\lambda_{31} & \lambda_{32} & \cdots & r_{3i} & r_{3,i+1} & \cdots & r_{3n} & c_3 \\
\vdots & \vdots & \ddots & & \vdots & & \vdots & \vdots \\
\lambda_{i1} & \lambda_{i2} & & r_{ii} & r_{i,i+1} & & r_{in} & c_i \\
\lambda_{i+1,1} & \lambda_{i+1,2} & & \lambda_{i+1,i} & a^i_{i+1,i+1} & \cdots & a^i_{i+1,n} & b^i_{i+1} \\
\vdots & & & \vdots & \vdots & & \vdots & \vdots \\
\lambda_{n1} & \lambda_{n2} & \cdots & \lambda_{ni} & a^i_{n,i+1} & \cdots & a^i_{nn} & b^i_n
\end{array}\right)$$

Here, the λ_{ki} are pertutations of q_{ki} and the matrix above finally leads to $\boldsymbol{L}$ and $\boldsymbol{U}$. Once the decomposition is done, two sets of equations need to be solved to obtain the solution $\boldsymbol{x}$:

$$\boldsymbol{Ly} = \boldsymbol{Pb}, \text{ and } \boldsymbol{Ux} = \boldsymbol{y}.$$

Exercise

The importance of the pivoting process can be understood from the following example:

$$\begin{pmatrix} 10^{-4} & 1 \\ 1 & 1 \end{pmatrix} \begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix} = \begin{Bmatrix} 1 \\ 2 \end{Bmatrix}$$

The exact solution to this is $x_1 = 1.00010001$ and $x_2 = 0.99989999$. Let us first do this without pivoting. The elimination step leads to

$$\left(\begin{array}{cc|c} 0.1 \cdot 10^{-3} & 0.1 \cdot 10^{1} & 0.1 \cdot 10^{1} \\ 0 & -0.1 \cdot 10^{5} & -0.1 \cdot 10^{5} \end{array} \right).$$

After backward substitution, the solution is $x_1 = 0, x_2 = 1$. With pivoting, the two rows are first swapped leading to:

$$\left(\begin{array}{cc|c} 0.1 \cdot 10^{1} & 0.1 \cdot 10^{1} & 0.2 \cdot 10^{1} \\ 0 & 0.1 \cdot 10^{1} & 0.1 \cdot 10^{1} \end{array} \right).$$

which gives the correct result $x_1 = x_2 = 1$.

भारतीय प्रौद्योगिकी संस्थान कानपुर INDIAN INSTITUTE OF TECHNOLOGY KANPUR

Sometimes, even pivoting cannot lead us to the correct solution and a *scaling* is neccessary. For example, multiplying the first row of the matrix in this example by a large number (say 20000), we get the system

$$\begin{pmatrix} 2 & 20000 \\ 1 & 1 \end{pmatrix} \begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix} \begin{Bmatrix} 20000 \\ 2 \end{Bmatrix}$$

With or without pivoting, this system gives the incorrect solution $x_1 = 0, x_2 = 1$. However, we can apply a *equilibriation step* to this system before solving it, by multiplying it with a diagonal matrix, i.e.

$$\boldsymbol{DAx} = \boldsymbol{Db},$$

where

$$D_{ii} = \sum_{j=1}^{n} \left(|a_{ij}|\right)^{-1}.$$

An even better method for stabilising matrices that have elements of very different size is to apply a *total pivoting* where both rows and columns are shuffled in a equilibriation step.

Computation of inverse by LU decomposition

In principle, the inverse of a square matrix $\boldsymbol{A}$ can be computed by first LU decomposing $\boldsymbol{PA}$. Solution of the systems

$$\boldsymbol{L}\boldsymbol{y}^i = \boldsymbol{P}\boldsymbol{e}^i, \boldsymbol{U}\boldsymbol{x}^i = \boldsymbol{y}^i,$$

where, $\boldsymbol{e}^i$ are the Cartesian basis vectors, yields the inverse $\boldsymbol{A}^{-1} = \left[\boldsymbol{x}^1\ \boldsymbol{x}^2\ \ldots\ \boldsymbol{x}^n\right]$. The practical implementation is illustrated through an example.

Exercise

Consider again the matrix

$$\begin{pmatrix} 3 & 1 & 6 \\ 2 & 1 & 3 \\ 1 & 1 & 1 \end{pmatrix}$$

Find its inverse. Step 1: pivoting

$$\left(\begin{array}{ccc|ccc} \boxed{3} & 1 & 6 & 1 & 0 & 0 \\ 2 & 1 & 3 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 \end{array}\right)$$

Step 2: elimination

$$\left(\begin{array}{ccc|ccc} 3 & 1 & 6 & 1 & 0 & 0 \\ 0 & 1/3 & -1 & -2/3 & 0 & 1 \\ 0 & \boxed{2/3} & -1 & -1/3 & 1 & 0 \end{array}\right)$$

भारतीय प्रौद्योगिकी संस्थान कानपुर INDIAN INSTITUTE OF TECHNOLOGY KANPUR

Step 3: Permute rows

$$\left(\begin{array}{ccc|ccc} 3 & 1 & 6 & 1 & 0 & 0 \\ 0 & 2/3 & -1 & -1/3 & 0 & 1 \\ 0 & 1/3 & -1 & -2/3 & 1 & 0 \end{array}\right)$$

Step 4: Next elimination

$$\left(\begin{array}{ccc|ccc} 3 & 1 & 6 & 1 & 0 & 0 \\ 0 & 2/3 & -1 & -1/3 & 0 & 1 \\ 0 & 0 & -1/2 & -1/2 & 1 & -1/2 \end{array}\right)$$

Step 4: Back substitution

$$\left(\begin{array}{ccc|ccc} 3 & 1 & 0 & -5 & 12 & -6 \\ 0 & 2/3 & 0 & 2/3 & -2 & 2 \\ 0 & 0 & -1/2 & -1/2 & 1 & -1/2 \end{array}\right)$$

भारतीय प्रौद्योगिकी संस्थान कानपुर INDIAN INSTITUTE OF TECHNOLOGY KANPUR

Back substitution leads to the inverse matrix:

$$\boldsymbol{A}^{-1} = \begin{pmatrix} -2 & 5 & -3 \\ 1 & -3 & 3 \\ 1 & -2 & 1 \end{pmatrix}$$

भारतीय प्रौद्योगिकी संस्थान कानपुर
INDIAN INSTITUTE OF TECHNOLOGY KANPUR

Matrix norms and conditioning

Assuming a basic exposure to linear algebra, we will quickly recapitulate some essential aspects. We have a vector space $\mathbb{K}^n$ of all n dimensional vectors. That is, we deal with vectors like $\boldsymbol{x} \in \mathbb{K}^n$ where, $\boldsymbol{x}^T = \langle x_1 \; x_2 \; \ldots \; x_n \rangle$, where x_i are components of the vector with respect to a fixed Cartesian basis $\boldsymbol{e}_i$,

$$\boldsymbol{x} = \sum_{i=1}^{n} x_i \boldsymbol{e}_i.$$

A mapping $||\cdot|| : \mathbb{K}^n \rightarrow \mathbb{R}$ is called a *vector norm* if it satisfies:

- *Definiteness:* $||\boldsymbol{x}|| \geq 0, ||\boldsymbol{x}|| = 0, \Rightarrow \boldsymbol{x} = 0, \forall \boldsymbol{x} \in \mathbb{K}^n$.
- *Homogenity:* $||\alpha \boldsymbol{x}|| = |\alpha| ||\boldsymbol{x}||, \alpha \in \mathbb{K}, \boldsymbol{x} \in \mathbb{K}^n$
- *Triangle inequality:* $||\boldsymbol{x} + \boldsymbol{y}|| \leq ||\boldsymbol{x}|| + ||\boldsymbol{y}||, \boldsymbol{x}, \boldsymbol{y} \in \mathbb{K}^n$

The norm can be defined over any vector space V over $\mathbb{K}$. The resulting pair $\{V, ||\cdot||\}$ is called a *normed vector space*

Examples of vector norms are the L_2 norm

$$||\boldsymbol{x}||_2 = \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2},$$

or the more general L_p norm

$$||\boldsymbol{x}||_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}.$$

More useful examples include the L_∞ norm and the L_1 norm defined as

$$||\boldsymbol{x}||_\infty = \max_{i=1,\ldots n} |x_i|, ||\boldsymbol{x}||_1 = \sum_{i=1}^{n} |x_i|.$$

A mapping $(\boldsymbol{x}, \boldsymbol{y}) : \mathbb{K}^n \times \mathbb{K}^n \to \mathbb{K}$ is called a *scalar product* if it satisfies

- *Symmetry:* $(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{y}, \boldsymbol{x}) \;\; \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{K}$
- *Linearity:* $(\alpha\boldsymbol{x} + \beta\boldsymbol{y}, \boldsymbol{z}) = \alpha(\boldsymbol{x}, \boldsymbol{z}) + \beta(\boldsymbol{y}, \boldsymbol{z}),\ \forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathbb{K}^n,\ \alpha, \beta \in \mathbb{K}$
- *Definiteness:* $(\boldsymbol{x}, \boldsymbol{x}) \in \mathbb{R}, (\boldsymbol{x}, \boldsymbol{x}) > 0\ \boldsymbol{x} \in \mathbb{K}^n / \{0\}$.

We will mostly use the Euclidean scalar product

$$(\boldsymbol{x}, \boldsymbol{y})_2 = \sum_{j=1}^{n} x_j y_j.$$

Two vectors are orthogonal if

$$(\boldsymbol{x}, \boldsymbol{y})_2 = 0.$$

A set of vectors in $\{\boldsymbol{a}^1, \boldsymbol{a}^2 \ldots \boldsymbol{a}^n\}$, with $\boldsymbol{a}^i \neq 0$ is called an *orthogonal basis* if $(\boldsymbol{a}^i, \boldsymbol{a}^j)_2 = 0$ for $i \neq j$. Further, they are called an *orthonormal basis* if, additionally, $(\boldsymbol{a}^k, \boldsymbol{a}^k) = 1$ for $k = 1, \ldots n$.

With an orthonormal basis in place, we can easily see that

$$(\boldsymbol{x}, \boldsymbol{a}^i)_2 = \left(\sum \alpha_j \boldsymbol{a}^j, \boldsymbol{a}^i \right)_2 = \alpha_i.$$

This leads to the representation

$$\boldsymbol{x} = \sum_{i=1}^{n} \left(\boldsymbol{x}, \boldsymbol{a}^i\right)_2 \boldsymbol{a}^i.$$

Further, it can easily be seen that

$$||\boldsymbol{x}||_2^2 = (\boldsymbol{x}, \boldsymbol{x})_2 = \sum_{i=1}^{n} \left|(\boldsymbol{x}, \boldsymbol{a}^i)_2\right|^2.$$

भारतीय प्रौद्योगिकी संस्थान कानपुर INDIAN INSTITUTE OF TECHNOLOGY KANPUR

Consider the vector space of all $n \times n$ matrices $\boldsymbol{A} \in \mathbb{K}^{n \times n}$. The norm induced by a vector norm $||\cdot||$ is

$$||\boldsymbol{A}|| = \max\left(\frac{||\boldsymbol{A}\boldsymbol{x}||}{||\boldsymbol{x}||} : ||\boldsymbol{x}|| \neq 0\right),$$

for $\boldsymbol{x} \in \mathbb{K}^n$. Through this norm, we are trying to determine the direction that is amplified most by $\boldsymbol{A}$.

A matrix norm satisfies

$$||\boldsymbol{A}\boldsymbol{x}|| \leq ||\boldsymbol{A}||||\boldsymbol{x}||, \boldsymbol{x} \in \mathbb{K}^n, \boldsymbol{A} \in \mathbb{K}^{n \times n}$$

and,

$$||\boldsymbol{A}\boldsymbol{B}|| \leq ||\boldsymbol{A}||||\boldsymbol{B}||, \boldsymbol{A}, \boldsymbol{B} \in \mathbb{K}^{n \times n}.$$

Further, a *natural matrix norm* satisfies

$$||\boldsymbol{I}|| = 1.$$

The natural matrix norm generated from the Euclidean norm is called the *spectral norm*. Note that it is not

$$||\boldsymbol{A}|| = \left(\sum_{j,k=1}^{n} |a_{jk}^2| \right)^{1/2},$$

as, for this norm, $||\boldsymbol{I}|| = \sqrt{n}$. If the eigenvalues of $\boldsymbol{A}^T\boldsymbol{A}$ are λ, then the spectral norm is

$$||\boldsymbol{A}||_2 = \max |\lambda| .$$

Now, we can do an error analysis of the linear system

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}.$$

This implies that if the matrix $\boldsymbol{A}$ and vector $\boldsymbol{b}$ are faulty by $\delta\boldsymbol{A}$ and $\delta\boldsymbol{b}$ respectively, we wish to find the error $\delta\boldsymbol{x}$ in the solution of the system

$$\tilde{\boldsymbol{A}}\tilde{\boldsymbol{x}} = \tilde{\boldsymbol{b}},$$

with $\tilde{\boldsymbol{A}} = \boldsymbol{A} + \delta\boldsymbol{A}, \tilde{\boldsymbol{b}} = \boldsymbol{b} + \delta\boldsymbol{b}$ and $\tilde{\boldsymbol{x}} = \boldsymbol{x} + \delta\boldsymbol{x}$.

भारतीय प्रौद्योगिकी संस्थान कानपुर INDIAN INSTITUTE OF TECHNOLOGY KANPUR

We state the so-called *perturbation theorem* without proof. Let $\boldsymbol{A} \in \mathbb{K}^{n \times n}$ be an invertible matrix such that $||\delta \boldsymbol{A}|| \leq ||\boldsymbol{A}^{-1}||^{-1}$, then, the perturbed matrix $\tilde{\boldsymbol{A}} = \boldsymbol{A} + \delta \boldsymbol{A}$ is also invertible and

$$\frac{||\delta \boldsymbol{x}||}{||\boldsymbol{x}||} \leq \frac{\text{cond}\boldsymbol{A}}{1 - \text{cond}\boldsymbol{A}\frac{||\delta \boldsymbol{A}||}{||\boldsymbol{A}||}} \left\{ \frac{||\delta \boldsymbol{b}||}{||\boldsymbol{b}||} + \frac{||\delta \boldsymbol{A}||}{||\boldsymbol{A}||} \right\},$$

where $\text{cond}\boldsymbol{A} = ||\boldsymbol{A}||||\boldsymbol{A}^{-1}||$.

Further, it can be shown that for the spectral norm

$$\text{cond}_2 \boldsymbol{A} = \frac{|\lambda_{max}|}{|\lambda_{min}|},$$

where λ_{max} and λ_{min} are the largest and smallest eigenvalues of $\boldsymbol{A}$.

For cases where $\text{cond}||\boldsymbol{A}||||\boldsymbol{A}^{-1}|| << 1$, for $\text{cond}\boldsymbol{A} \sim 10^s$, and relative errors in $\boldsymbol{A}$ and $\boldsymbol{b} \sim 10^{-k}$, the relative error in $\boldsymbol{x}$ will be 10^{s-k}.