

COURSE 8

4. Numerical methods for solving linear systems

Practical solving of many problems eventually leads to solving linear systems.

Classification of the methods:

- *direct methods* - with low number of unknowns (up to several tens of thousands); they provide the exact solution of the system in a finite number of steps.
- *iterative methods* - with medium number of unknowns; it is obtained an approximation of the solution as the limit of a sequence.

- *semiiterative methods* - with large number of unknowns; it is obtained an approximation of the solution.

4.1. Perturbation of linear systems.

Consider the linear system

$$Ax = b.$$

Definition 1 *The number $\text{cond}(A) = \|A\| \|A^{-1}\|$ is called **conditioning number** of the matrix A . It measures the sensibility of the solution x of the system $Ax = b$ to the perturbation of A and b .*

*The system is **good conditioned** if $\text{cond}(A)$ is small (< 1000) or it is **ill conditioned** if $\text{cond}(A)$ is great.*

Remark 2 1. $\text{cond}(A) \geq 1$.

2. $\text{cond}(A)$ depends on the norm used.

Consider an example

$$\begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix},$$

with the solution $\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$.

We perturbate the right hand side:

$$\begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix} \begin{pmatrix} x_1 + \delta x_1 \\ x_2 + \delta x_2 \\ x_3 + \delta x_3 \\ x_4 + \delta x_4 \end{pmatrix} = \begin{pmatrix} 32.1 \\ 22.9 \\ 33.1 \\ 30.9 \end{pmatrix},$$

and obtain the exact solution $\begin{pmatrix} 9.2 \\ -12.6 \\ 4.5 \\ -1.1 \end{pmatrix}$.

Consider, for example,

$$\left| \frac{b_2 - (b_2 + \delta b_2)}{b_2} \right| = \left| \frac{\delta b_2}{b_2} \right| = \frac{1}{229} \approx \frac{1}{200},$$

where δb_i , $i = \overline{1,3}$ denote the perturbations of b , and

$$\left| \frac{x_2 - (x_2 + \delta x_2)}{x_2} \right| = \left| \frac{\delta x_2}{x_2} \right| = 13.6 \approx 10.$$

Thus, a relative error of order $\frac{1}{200}$ on the right hand side (precision of $\frac{1}{200}$ for the data in a linear system) attracts a relative error of order 10 on the solution, 2000 times larger.

Consider the same system, and perturb the matrix A :

$$\begin{pmatrix} 10 & 7 & 8.1 & 7.2 \\ 7.08 & 5.04 & 6 & 5 \\ 8 & 5.98 & 9.89 & 9 \\ 6.99 & 4.99 & 9 & 9.98 \end{pmatrix} \begin{pmatrix} x_1 + \delta x_1 \\ x_2 + \delta x_2 \\ x_3 + \delta x_3 \\ x_4 + \delta x_4 \end{pmatrix} = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix},$$

with exact solution $\begin{pmatrix} -81 \\ 137 \\ -34 \\ 22 \end{pmatrix}$.

The matrix A seems to have good properties (symmetric, with determinant 1), and the inverse $A^{-1} = \begin{pmatrix} 25 & -41 & 10 & -6 \\ -41 & 68 & -17 & 10 \\ 10 & -17 & 5 & -3 \\ -6 & 10 & -3 & 2 \end{pmatrix}$ is also with integer numbers.

This example is very concerning as such orders of the errors in many experimental sciences are considered as satisfactory.

Remark 3 *For this example we have $\text{cond}(A) = 2984$ (in euclidian norm).*

Analyze the phenomenon:

◆ In the first case, when b is perturbed, we compare the exact solutions x and $x + \delta x$ of the systems

$$Ax = b$$

and

$$A(x + \delta x) = b + \delta b.$$

Let $\|\cdot\|$ be a norm on \mathbb{R}^n and $\|\cdot\|$ the induced matrix norm.

We have the systems

$$Ax = b$$

and

$$Ax + A\delta x = b + \delta b \iff A\delta x = \delta b.$$

From $\delta x = A^{-1}\delta b$ we get $\|\delta x\| \leq \|A^{-1}\| \|\delta b\|$

and from $b = Ax$ we get $\|b\| \leq \|A\| \|x\| \Leftrightarrow \frac{1}{\|x\|} \leq \frac{\|A\|}{\|b\|}$,

so the relative error of the result is bounded by

$$\frac{\|\delta x\|}{\|x\|} \leq \left(\|A\| \|A^{-1}\| \right) \frac{\|\delta b\|}{\|b\|} \stackrel{\text{denoted}}{=} \text{cond}(A) \frac{\|\delta b\|}{\|b\|}. \quad (1)$$

◆ In the second case, when the matrix A is perturbed, we compare the exact solutions of the linear systems

$$Ax = b$$

and

$$\begin{aligned} (A + \delta A)(x + \delta x) = b &\iff Ax + A\delta x + \delta Ax + \delta A\delta x = b \\ &\iff A\delta x = -\delta A(x + \delta x). \end{aligned}$$

From $\delta x = -A^{-1}\delta A(x + \delta x)$, we get $\|\delta x\| \leq \|A^{-1}\| \|\delta A\| \|x + \delta x\|$, or

$$\frac{\|\delta x\|}{\|x + \delta x\|} \leq \|A^{-1}\| \|\delta A\| = \left(\|A\| \|A^{-1}\| \right) \frac{\|\delta A\|}{\|A\|} = \text{cond}(A) \frac{\|\delta A\|}{\|A\|}. \quad (2)$$

4.2. Direct methods for solving linear systems

Why Cramer's method is not suitable for solving linear systems for $n \geq 100$ and it will not be in near future?

For applying Cramer's method for a $n \times n$ system we need in a rough evaluation the following number of operations:

$$\begin{cases} (n+1)! & \text{additions} \\ (n+2)! & \text{multiplications} \\ n & \text{divisions} \end{cases}$$

Consider, hypothetically, a volume $V = 1 \text{ km}^3$ of cubic processors of each having the side $l = 10^{-8} \text{ cm}$ (radius of an atom), the time for execution of an operation is supposed to be equal to the time needed for the light to pass through an atom. (Light speed is 300.000 km/s.)

In this hypothetically case, the time necessary for solving the $n \times n$ system, $n \geq 100$, will be more than 10^{94} years!

4.2.1. Gauss method for solving linear systems

Consider the linear system $Ax = b$, i.e.,

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}. \quad (3)$$

The method consists of two stages:

- reducing the system (3) to an equivalent one, $Ux = d$, with U an upper triangular matrix.
- solving of the upper triangular linear system $Ux = d$ by backward substitution.

At least one of the elements on the first column is nonzero, otherwise A is singular. We choose one of these nonzero elements (using some criterion) and this will be called the first elimination **pivot**.

If the case, we change the line of the pivot with the first line, both in A and in b , and next we successively make zeros under the first pivot:

$$\begin{pmatrix} a_{11}^1 & a_{12}^1 & \dots & a_{1n}^1 \\ 0 & a_{22}^1 & \dots & a_{2n}^1 \\ \vdots & \vdots & & \vdots \\ 0 & a_{n2}^1 & \dots & a_{nn}^1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1^1 \\ b_2^1 \\ \vdots \\ b_n^1 \end{pmatrix}.$$

Analogously, after k steps we obtain the system

$$\begin{pmatrix} a_{11}^1 & a_{12}^1 & \dots & a_{1k}^1 & a_{1,k+1}^1 & \dots & a_{1n}^1 \\ 0 & a_{22}^2 & \dots & a_{2k}^2 & a_{2,k+1}^2 & \dots & a_{2n}^2 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & a_{kk}^k & a_{k,k+1}^k & \dots & a_{kn}^k \\ 0 & 0 & \dots & 0 & a_{k+1,k+1}^k & \dots & a_{k+1,n}^k \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 0 & a_{n,k+1}^k & \dots & a_{nn}^k \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \\ x_{k+1} \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1^1 \\ b_2^2 \\ \vdots \\ b_k^k \\ b_{k+1}^k \\ \vdots \\ b_n^k \end{pmatrix}.$$

If $a_{kk}^k \neq 0$, denote $m_{ik} = \frac{a_{ik}^k}{a_{kk}^k}$ and we get

$$\begin{aligned} a_{ij}^{k+1} &= a_{ij}^k - m_{ik} a_{kj}^k, \quad j = k, \dots, n \\ b_i^{k+1} &= b_i^k - m_{ik} b_k^k, \quad i = k + 1, \dots, n. \end{aligned}$$

After $n - 1$ steps we obtain the system

$$\begin{pmatrix} a_{11}^1 & a_{12}^1 & \dots & a_{1n}^1 \\ 0 & a_{22}^2 & \dots & a_{2n}^2 \\ 0 & 0 & \dots & a_{3n}^3 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & a_{nn}^{n-1} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1^1 \\ b_2^2 \\ b_3^3 \\ \vdots \\ b_n^{n-1} \end{pmatrix}.$$

Remark 4 *The total number of elementary operations is of order $\frac{2}{3}n^3$.*

Example 5 *Consider the system*

$$\begin{pmatrix} 0.0001 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$$

Gauss algorithm yields: $m_{11} = \frac{a_{21}}{a_{11}} = \frac{1}{0.0001}$

$$\begin{pmatrix} 0.0001 & 1 \\ 1 - 0.0001 * m_{11} = 0 & 1 - 1 * m_{11} = -9999 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \\ = \begin{pmatrix} 1 \\ 2 - 1 * m_{11} = -9998 \end{pmatrix}$$

$$\Rightarrow y = \frac{9998}{9999} = 0.(9998) \approx 1.$$

Replacing in the first equation we get

$$x = 1.000(1000) \approx 1.$$

By division with a pivot of small absolute value there could be induced errors. For avoiding this there are two ways:

A) Partial pivoting: finding an index $p \in \{k, \dots, n\}$ such that:

$$|a_{p,k}^k| = \max_{i=\overline{k,n}} |a_{i,k}^k|.$$

B) Total pivoting: finding $p, q \in \{k, \dots, n\}$ such that:

$$|a_{p,q}^k| = \max_{i,j=\overline{k,n}} |a_{ij}^k|,$$

Example 6 *Solve the following system of equations using partial pivoting:*

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 3 & 1 & 5 \\ -1 & 1 & -5 & 3 \\ 3 & 1 & 7 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 10 \\ 31 \\ -2 \\ 18 \end{bmatrix}.$$

The pivot is a_{41} . We interchange the 1–st line and the 4–th line. We have

$$\begin{bmatrix} 3 & 1 & 7 & -2 \\ 2 & 3 & 1 & 5 \\ -1 & 1 & -5 & 3 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 18 \\ 31 \\ -2 \\ 10 \end{bmatrix},$$

then

$$\begin{aligned} \text{pivot element} &\rightarrow \\ m_{21} &= \frac{2}{3} \\ m_{31} &= -\frac{1}{3} \\ m_{41} &= \frac{1}{3} \end{aligned} \left[\begin{array}{cccc|c} 3 & 1 & 7 & -2 & 18 \\ 0 & 2.33 & -3.66 & 6.33 & 19 \\ 0 & 1.33 & -2.66 & 2.33 & 4 \\ 0 & 0.66 & -1.33 & 1.66 & 4 \end{array} \right].$$

Subtracting multiplies of the first equation from the three others gives

$$\begin{aligned} \text{pivot element} &\rightarrow \\ m_{32} &= \frac{1.33}{2.33} \\ m_{42} &= \frac{0.66}{2.33} \end{aligned} \left[\begin{array}{cccc|c} 3 & 1 & 7 & -2 & 18 \\ 0 & \mathbf{2.33} & -3.66 & 6.33 & 19 \\ 0 & 1.33 & -2.66 & 2.33 & 4 \\ 0 & 0.66 & -1.33 & 1.66 & 4 \end{array} \right].$$

Subtracting multiplies, of the second equation from the last two equations, gives

$$\begin{aligned} \text{pivot element} &\rightarrow \\ m_{43} &= \frac{0.28}{0.57} \end{aligned} \left[\begin{array}{cccc|c} 3 & 1 & 7 & -2 & 18 \\ 0 & 2.33 & -3.66 & 6.33 & 19 \\ 0 & 0 & -\mathbf{0.57} & -1.28 & -6.85 \\ 0 & 0 & -0.28 & -0.14 & -1.42 \end{array} \right].$$

Subtracting multiplies, of the third equation form the last one, gives the upper triangular system

$$\left[\begin{array}{cccc|c} 3 & 1 & 7 & -2 & 18 \\ 0 & 2.33 & -3.66 & 6.33 & 19 \\ 0 & 0 & -0.57 & -1.28 & -6.85 \\ 0 & 0 & 0 & 0.5 & 2 \end{array} \right].$$

The process of the back substitution algorithm applied to the triangular system produces the solution

$$\begin{aligned} x_4 &= \frac{2}{0.5} = 4 \\ x_3 &= \frac{-6.85 + 1.28x_4}{-0.57} = 3 \\ x_2 &= \frac{19 + 3.66x_3 - 6.33x_4}{2.33} = 2 \\ x_1 &= \frac{18 - x_2 - 7x_3 + 2x_4}{3} = 1. \end{aligned}$$

Example 7 *Solve the system:*

$$\begin{cases} 2x + y = 3 \\ 3x - 2y = 1 \end{cases}$$

Sol.

$$\begin{cases} 2x + y = 3 \\ 3x - 2y = 1 \end{cases}$$

The extended matrix is

$$\left[\begin{array}{cc|c} 2 & 1 & 3 \\ 3 & -2 & 1 \end{array} \right]$$

and the pivot is 3. We interchange the lines:

$$\left[\begin{array}{cc|c} 3 & -2 & 1 \\ 2 & 1 & 3 \end{array} \right]$$

We have $L_2 - \frac{2}{3}L_1 \rightarrow L_2$ and obtain

$$\left[\begin{array}{cc|c} 3 & -2 & 1 \\ 0 & \frac{7}{3} & \frac{7}{3} \end{array} \right]$$

so the system becomes

$$\begin{cases} 3x - 2y = 1 \\ \frac{7}{3}y = \frac{7}{3} \end{cases} .$$

Solution is

$$\begin{cases} x = 1 \\ y = 1 \end{cases}.$$

Example 8 *Solve the system:*

$$\begin{cases} x_1 + x_2 + x_3 = 4 \\ 2x_1 - 2x_2 + 3x_3 = 5 \\ x_1 - x_2 + 4x_3 = 5. \end{cases}$$

4.2.2. Gauss-Jordan method ("total elimination" method)

Consider the linear system $Ax = b$, i.e.,

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}. \quad (4)$$

We make transformations, like in Gauss elimination method, to make zeroes in the lines $i+1, i+2, \dots, n$ and then, also in the lines $1, 2, \dots, i-1$ such that the system to be reducing to:

$$\begin{pmatrix} a_{11}^1 & 0 & \dots & 0 \\ 0 & a_{22}^2 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & a_{nn}^n \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1^1 \\ b_2^2 \\ b_3^3 \\ \vdots \\ b_n^n \end{pmatrix}.$$

The solution is obtained by

$$x_i = \frac{b_i^i}{a_{ii}^i}, \quad i = 1, \dots, n.$$

Definition 9 A $n \times n$ matrix A is **strictly diagonally dominant** if

$$|a_{ii}| > \sum_{j=1, j \neq i}^n |a_{ij}|, \text{ for } i = 1, 2, \dots, n.$$

Theorem 10 If A is a strictly diagonally dominant matrix, then A is nonsingular and moreover, Gaussian elimination can be performed on any linear system $Ax = b$ without row or column interchanges, and the computations are stable with respect to the growth of rounding errors.

4.2.3. Factorization methods - LU methods

The matrix A can be factored into the product of a lower triangular matrix L and an upper triangular matrix U , namely $A = LU$.

$$Ax = b \iff L U x = b,$$

where

$$L = \begin{pmatrix} l_{11} & 0 & \dots & 0 \\ l_{21} & l_{22} & \dots & 0 \\ \vdots & & & \\ l_{n1} & l_{n2} & \dots & l_{nn} \end{pmatrix} \quad U = \begin{pmatrix} u_{11} & u_{12} & \dots & u_{1n} \\ 0 & u_{22} & \dots & u_{2n} \\ \vdots & & & \\ 0 & 0 & \dots & u_{nn} \end{pmatrix}.$$

We solve the systems in two stages:

First stage: Solve $Lz = b$,

Second stage: Solve $Ux = z$.

Methods for computing matrices L and U : **Doolittle method** where all diagonal elements of L have to be 1; **Crout method** where all diagonal elements of U have to be 1 and **Choleski method** where $l_{ii} = u_{ii}$ for $i = 1, \dots, n$.

Remark 11 *LU factorizations are modified forms of Gauss elimination method.*

Doolittle method

We consider that the hypothesis of Theorem 10 is fulfilled, so $a_{kk} \neq 0$, $k = \overline{1, n-1}$. Denote

$$l_{i,k} := \frac{a_{i,k}^{(k-1)}}{a_{k,k}^{(k-1)}}, \quad i = \overline{k+1, n}$$

$$t^{(k)} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ l_{k+1,k} \\ \dots \\ l_{n,k} \end{bmatrix},$$

having zeros for the first k -th lines, and

$$M_k = I_n - t^{(k)} e_k \in \mathcal{M}_{n \times n}(\mathbb{R}) \quad (5)$$

where $e_k = \begin{pmatrix} 0 & \dots & 1 & \dots & 0 \end{pmatrix}$ is the k -unit vector of dimension n , (has

1 on the k -th position) and $I_n = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & & & \\ 0 & 0 & \dots & 1 \end{pmatrix}$ is the identity matrix of order n .

$a_{i,k}^{(0)}$ are elements of A ; $a_{i,k}^{(1)}$ are elements of $M_1 \cdot A$; ...; $a_{i,k}^{(k-1)}$ are elements of $M_{k-1} \dots \cdot M_1 \cdot A$.

Definition 12 The matrix M_k is called the **Gauss matrix**, the components $l_{i,k}$ are called the **Gauss multiplies** and the vector $t^{(k)}$ is the **Gauss vector**.

Remark 13 If $A \in \mathcal{M}_{n \times n}(\mathbb{R})$, then the Gauss matrices M_1, \dots, M_{n-1} can be determined such that

$$U = M_{n-1} \cdot M_{n-2} \dots M_2 \cdot M_1 \cdot A$$

is an upper triangular matrix. Moreover, if we choose

$$L = M_1^{-1} \cdot M_2^{-1} \dots M_{n-1}^{-1}$$

then

$$A = L \cdot U.$$

Example 14 *Find LU factorization for the matrix*

$$A = \begin{pmatrix} 2 & 1 \\ 6 & 8 \end{pmatrix}.$$

Solve the system $\begin{cases} 2x_1 + x_2 = 3 \\ 6x_1 + 8x_2 = 9 \end{cases}.$

Sol.

$$\begin{aligned} M_1 &= I_2 - t^{(1)}e_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} 0 \\ 6 \\ 2 \end{pmatrix} \begin{pmatrix} 1 & 0 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} 0 & 0 \\ 3 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -3 & 1 \end{pmatrix}. \end{aligned}$$

We have

$$U = M_1 A = \begin{pmatrix} 1 & 0 \\ -3 & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 \\ 6 & 8 \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 0 & 5 \end{pmatrix}$$

$$L = M_1^{-1} = \begin{pmatrix} 1 & 0 \\ 3 & 1 \end{pmatrix}.$$

So

$$A = \begin{pmatrix} 2 & 1 \\ 6 & 8 \end{pmatrix} = L \cdot U = \begin{pmatrix} 1 & 0 \\ 3 & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 \\ 0 & 5 \end{pmatrix}.$$

We have

$$L \cdot U \cdot x = \begin{pmatrix} 3 \\ 9 \end{pmatrix}$$
$$Ux = z$$

and

$$\begin{pmatrix} 1 & 0 \\ 3 & 1 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} 3 \\ 9 \end{pmatrix} \Rightarrow z = \begin{pmatrix} 3 \\ 0 \end{pmatrix}$$
$$\begin{pmatrix} 2 & 1 \\ 0 & 5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 3 \\ 0 \end{pmatrix} \Rightarrow x = \begin{pmatrix} 1.5 \\ 0 \end{pmatrix}.$$