COURSE 9

3.5. General quadrature formulas

Example 1 *If we know only the values of $f'(a)$ and $f(b)$ which is the corresponding quadrature formula?*

Sol. We have $f(x) = (B_1 f)(x) + (R_1 f)(x)$ with

$$(B_1 f)(x) = b_{01}(x) f'(a) + b_{10}(x) f(b). \tag{1}$$

Formula (1) generates the quadrature formula

$$\int_a^b f(x)dx = \sum_{k=0}^{1} \sum_{j=I_k} A_{kj} f^{(j)}(x_k) + R_1(f),$$

with

$$A_{01} = \int_a^b b_{01}(x)dx = -\frac{(b-a)^2}{2}$$
$$A_{10} = \int_a^b b_{10}(x)dx = (b-a).$$

Example 2 *Find the coefficients A, B and C of the following quadrature formula:*

$$\int_0^1 f(x)dx = Af(0) + Bf'(0) + Cf(1) + R(f).$$

4. Numerical methods for solving linear systems

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

Real-World Application:

1. Price of Fruits: Peter buys two apples and three bananas for $4. Nadia buys four apples and six bananas for $8 from the same store. How much does one banana and one apple costs?

2. A baker sells plain cakes for $7 and decorated cakes for $11. On a busy Saturday the baker started with 120 cakes, and sold all but three. His takings for the day were $991. How many plain cakes did he sell that day, and how many were decorated before they were sold?

3. Twice John's age plus five times Claire's age is 204. Nine times John's age minus three times Claire's age is also 204. How old are John and Claire?

4. Assume an electric network consisting of two voltage sources and three resistors. Applying Kirchhoff's laws it is determined the current going through each resistor.

5. Network analysis: networks composed of branches and junctions are used as models in such fields as economics, traffic analysis, and electrical engineering.

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 3 *The number* $\text{cond}\,(A) = \|A\| \left\|A^{-1}\right\|$ *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 4 *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 5 *For this example we have* $\mathrm{cond}\,(A) = 2984$ *(in euclidian norm).*

Analyze the phenomenon:

♦ In the first case, when b is perturbed, we compare de 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

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

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

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

$$\frac{\|\delta x\|}{\|x\|} \leq \left(\|A\| \left\|A^{-1}\right\|\right) \frac{\|\delta b\|}{\|b\|} \overset{denoted}{=} \text{cond}\,(A) \frac{\|\delta b\|}{\|b\|}. \tag{2}$$

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

$$Ax = b$$

and

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

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

$$\frac{\|\delta x\|}{\|x + \delta x\|} \leq \left\|A^{-1}\right\| \|\delta A\| = \left(\|A\| \left\|A^{-1}\right\|\right) \frac{\|\delta A\|}{\|A\|} = \text{cond}\,(A) \frac{\|\delta A\|}{\|A\|}. \tag{3}$$

4.2. Direct methods for solving linear systems

People not dealing with numerical analysis may naturally ask: "which is the role of such a study, since the Cramer formulas and the Gauss method are known for such a long time, and may be applied to any nonsingular linear system?". The answer to this question is dictated by practical considerations, specific to the numerical analysis. More precisely, both of the mentioned methods present major impediments when we try to use them with computers, for large number of unknowns.

Drawbacks of the Cramer's method. 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:

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

Consider, hypothetically, a volume $V = 1$ km^3 of cubic processors of each having the side $l = 10^{-8}$ 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 c=300.000 km/s.) Let us admit that the time required by an element for performing an elementary arithmetic operation is $t = l/c$. In such a case, the amount of elementary arithmetic operations performed in a second is

$$\text{No. of op.} = \frac{\text{No. of elem.}}{\text{Time required by an elem.}} = \frac{V}{l^3} / \frac{l}{c} = \frac{cV}{l^4} = \frac{3{\cdot}10^8{\cdot}10^9}{10^{-40}}.$$

Let us admit that the Cramer formulas for the considered system require the performing of only 100! elementary arithmetic operations. Since $100! = 10^{157,9\ldots}$, this means that this computer would need approximately 10^{94} years to compute the solution of the system.

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

The Cramer formulas are not used for practical problems even in the case of small numbers of unknowns, because the number of elementary operations (if admitting that are just of order $\mathcal{O}(n!)$) is much higher than of other direct methods, which require only $\mathcal{O}\left(n^3\right)$ operations. As an example, Ciarlet mentions that for a linear system with $n = 10$ unknowns, the Gauss method requires 700 operations, while the Cramer rule requires 400.000.000 operations. On the other hand, the size of the cumulated errors in the Cramer rule may lead to meaningless solutions.

Drawbacks of the Gauss method when used at the computer. In the case of the partial pivoting variant for solving the system $Ax = b$, the relative error of the obtained "solution" $\tilde{x}$ is bounded in the following way

$$\frac{\|x^* - \tilde{x}\|_\infty}{\|x^*\|_\infty} \leq 4n^2 \rho \epsilon \cdot cond(A),$$

where

- n is the dimension of the system;

- ϵ is the machine epsilon (the exact upper bound of the relative errors appearing in the representation of the real numbers and in performing the elementary arithmetic operations in floating point arithmetic);

- $cond(A)$ is the condition number of the matrix A in the Chebyshev norm;

- ρ is a parameter specific to any matrix, its maximum value being 2^{n-1}; Wilkinson has shown in 1965 that this bound is exact and has provided theoretical examples of matrices when this bound is attained, but for the majority of the practical problems the value of ρ is small. Wilkinson has further stated that in all his activity he has not found practical applications for which the amplification factor to be larger than 16.

The above relation shows that the solution computed by the Gauss method may be far away from the exact one when the number of unknowns is large or when the matrix A is ill conditioned.

On the other hand, the number of elementary arithmetic operations performed by the algorithm makes that the time required to be exaggerated long as n increases.

For $n \geq 100.000$ no linear system with "full" matrix has been solved by direct methods; the "record" seems to be attained by a system having dimension $n = 76.800$. Most often, the large linear systems have sparse matrices, coming from different discretizations. The Gauss method does not take into account such structures.

4.2.1. Gauss method for solving linear systems

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

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

The method consists of two stages:

- reducing the system (4) 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

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

After $n-1$ steps we obtain the system

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

Remark 6 *The total number of elementary operations is of order* $\frac{2}{3}n^3$.

Example 7 *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: $m_{21} = \frac{a_{21}}{a_{11}} = \frac{1}{0.0001}$

$$\begin{pmatrix} 0.0001 & 1 \\ 1 - 0.0001 * m_{21} = 0 & 1 - 1 * m_{21} = -9999 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$
$$= \begin{pmatrix} 1 \\ 2 - 1 * m_{21} = -9998 \end{pmatrix}$$

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

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 $p \in \{k, ..., n\}$ such that:

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

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

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

Example 8 *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

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

Subtracting multiplies of the first equation from the three others gives

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

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

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

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

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

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 9 *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 $L_2 - \frac{2}{3}L_1 \to L_2$ and obtain

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

so the system becames

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

Solution is

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

Example 10 *Solve the following system using Gauss elimination method:*

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

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

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

$$\begin{pmatrix} a_{11} & a_{12} & ... & a_{1n} \\ a_{21} & a_{22} & ... & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & ... & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}. \tag{5}$$

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

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

The solution is obtained by

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

Definition 11 *A $n \times n$ matrix A is* **strictly diagonally dominant** *if*

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

Theorem 12 *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 rows or columns interchanges, and the computations are stable with respect to the growth of rounding errors.*