COURSE 8

3.6. Quadrature formulas of Gauss type (continuation)

Definition 1 *A formula of the following form*

$$\int_a^b w(x)f(x)dx = \sum_{k=1}^{m} A_k f(x_k) + R_m(f) \tag{1}$$

is called **a quadrature formula of Gauss type** *or* **with maximum degree of exactness** *if the coefficients A_k and the nodes x_k, $k = 1, ..., m$ are determined such that the formula has the maximum degree of exactness.*

Remark 2 *The coefficients and the nodes are determined such that to minimize the error, to produce exact results for the largest class of polynomials.*

Example 3 *For $m = 2$, the Gauss type quadrature formula is*

$$\int_a^b f(x)dx = A_1 f(x_1) + A_2 f(x_2) + R_2(f).$$

The corresponding Legendre polynomial is

$$u(x) = \frac{2}{4!}\left[(x-a)^2(x-b)^2\right]'' \\ = x^2 - (a+b)x + \frac{1}{6}(a^2+b^2+4ab),$$

with the roots

$$x_1 = \frac{a+b}{2} - \frac{(b-a)\sqrt{3}}{6}, \\ x_2 = \frac{a+b}{2} + \frac{(b-a)\sqrt{3}}{6}.$$

For finding A_1 and A_2 we use the first two equations:

$$\begin{cases} A_1 + A_2 = b - a \\ A_1x_1 + A_2x_2 = (b^2-a^2)/2. \end{cases}$$

We get

$$A_1 = A_2 = (b-a)/2,$$

so the Gauss type quadrature formula with two nodes is

$$\int_a^b f(x)dx = \frac{b-a}{2}\left[f\left(\frac{a+b}{2} - \frac{b-a}{6}\sqrt{3}\right) + f\left(\frac{a+b}{2} + \frac{b-a}{6}\sqrt{3}\right)\right] + R_2(f).$$

Remark 4 *a) The coefficients A_k, $k = 1, ..., m$ of a Gauss type formula are positive.*

b) The coefficients A_k, $k = 1, ..., m$ and the roots of the Legendre polynomials can be found in tables, for $a = -1$, $b = 1$. For example, for $m = 2$ and $m = 3$:

m	*nodes*	*coefficients*
2	0.577	1
	-0.577	1
3	0.774	0.555
	0	0.888
	-0.774	0.555

c) For different weight functions, tables are available for both the nodes and the coefficients.

Example 5 *Approximate* $\int_{-1}^{1} e^x \cos x dx$ *using a Gauss type quadrature formula for* $m = 3$.

Sol.

$$\int_{-1}^{1} e^x \cos x dx \simeq 0.55 e^{0.77} \cos 0.77 + 0.88 \cos 0 + 0.55 e^{-0.77} \cos(-0.77)$$
$$= 1.9333904$$

(Exact value is 1.9334214.) Absolute error is $< 3.2 \cdot 10^{-5}$. ■

The integral $\int_a^b f(x)dx$ for an arbitrary interval $[a, b]$ could be transfomed in an integral on $[-1, 1]$ using the change of variable

$$t = \frac{2x - a - b}{b - a} \Leftrightarrow x = \frac{1}{2}\left[(b - a)t + a + b\right].$$

The Gauss type quadrature formulas may be applied on the following way:

$$\int_a^b f(x)dx = \int_{-1}^{1} f\left(\frac{(b-a)t + (b+a)}{2}\right) \frac{(b-a)}{2} dt. \qquad (2)$$

Example 6 *Consider the integral* $\int_1^3 (x^6 - x^2 \sin(2x))dx = 317.3442466.$

a) Compare the result obtained for Newton-Cotes type formula with $m = 1$ *and the gaussian formula for* $m = 2$;

b) Compare the result obtained for Newton-Cotes type formula with $m = 2$ *and the gaussian formula for* $m = 3$.

Sol. a) Each formula needs 2 evaluations of the function $f(x) = x^6 - x^2 \sin(2x)$. We have

$$\text{Trapezium formula } (m = 1) : \frac{2}{2}[f(1) + f(3)] = 731.6054420;$$

and a Gauss type formula for $m = 2$, using (2):

$$\int_1^3 (x^6 - x^2 \sin(2x))dx = \int_{-1}^1 ((t+2)^6 - (t+2)^2 \sin(2(t+2)))dt$$
$$\simeq f(-0.577 + 2) + f(0.577 + 2) = 306.8199344$$

b) Each formula needs 3 evaluations of the function. We have

$$\text{Simpson's F. } (m=2) : \frac{1}{3}[f(1)+4f(2)+f(3)] = 333.23;$$

and a Gauss type formula for $m = 3$, using (2):

$$\begin{aligned}\int_1^3 (x^6 - x^2 \sin(2x))dx &= \int_{-1}^1 ((t+2)^6 - (t+2)^2 \sin(2(t+2)))dt \\ &\simeq 0.55f(-0.77+2) + 0.88f(2) + 0.55f(0.77+2) \\ &= 317.2641516\end{aligned}$$

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 7 *The number* 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* cond (A) *is small (<1000) or it is* **ill conditioned** *if* cond (A) *is great.*

Remark 8 *1.* cond $(A) \geq 1$*.*

2. 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 9 *For this example we have* $\operatorname{cond}(A) = 2984$ *(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

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

♦ 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

$$\frac{\|\delta x\|}{\|x + \delta x\|} \leq \left\|A^{-1}\right\| \|\delta A\| = \left(\|A\| \left\|A^{-1}\right\|\right) \frac{\|\delta A\|}{\|A\|} = \text{cond}\,(A) \frac{\|\delta A\|}{\|A\|}. \tag{4}$$

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:

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

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 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.,

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

The method consists of two stages:

- reducing the system (5) 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

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

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

Example 11 *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 $p \in \{k, ..., n\}$ 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 12 *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

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

Example 13 *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 14 *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.,

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}. \tag{6}$$

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:

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

The solution is obtained by

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

Definition 15 *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 16 *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.*