Prefaţă

În cadrul acestei cărţi sunt prezentate metodele de bază ale analizei numerice. Regăsim cele mai importante metode numerice prin care obţinem soluţii aproximative pentru diferite probleme matematice studiate în anii precedenţi la analiză matematică, ecuaţii diferenţiale şi algebră liniară. După fiecare metodă dăm şi algoritmizarea corespunzătoare într-un limbaj pseudocod. Activitatea studenţilor pe parcursul laboratoarelor constă în a proba valabilitatea acestor algoritmi prin alegerea unui limbaj evoluat de programare, cum ar fi de exemplu Pascal sau C. Menţionăm că în pachetele de programe Mathcad, Matlab, Mathematica, etc. sunt înglobate metodele numerice prezentate de noi.

Parcurgerea şi înţelegerea acestei cărţi necesită cunoştinţe din analiza matematică, algebra liniară şi ecuaţii diferenţiale, precum şi noţiuni de programare. Recomandăm călduros şi cu generozitate această carte pentru studenţii Universităţii "Petru Maior".

Tg. Mureş, 01.09.2004

conf.dr. Finta Béla

Cuprins

1 Introducere **9**

2 Erori **13**

2.1 Surse de erori 13

2.2 Studiul erorilor sau propagarea erorilor 14

2.3 Propagarea erorilor la cele patru operaţii aritmetice 14

2.4 Formula generală a propagării erorilor sau eroarea introdusă de o funcţie . 15

3 Folosirea sumelor şi a seriilor numerice în aproximări **17**

3.1 Serii alternate 17

3.2 Serii cu termeni pozitivi 19

3.3 Transformarea seriilor numerice slab convergente în serii mai rapid convergente 20

3.3.1 Transformarea lui Euler 20

3.3.2 Transformarea lui Kummer 21

4 Evaluarea funcţiilor **23**

4.1 Evaluarea polinoamelor 23

4.2 Evaluarea funcţiilor analitice 24

4.3 Evaluarea funcţiilor date prin relaţii de recurenţă 25

4.4 Evaluarea funcţiilor cu ajutorul fracţiilor continue 26

5 Ecuaţii neliniare **29**

5.1 Metoda lui Bairstow 29

5.2 Separarea rădăcinilor reale în cazul ecuaţiilor neliniare reale 31

5.3 Metoda bisecţiei sau metoda înjumătăţirii intervalului 32

5.4 Metode iterative pentru rezolvarea ecuaţiilor neliniare 33
5.4.1 Metoda tangentei sau metoda lui Newton 34
5.4.2 Metoda paralelelor 40
5.4.3 Metoda coardei 41
5.4.4 Metoda secantei 42
5.4.5 Metoda lui Steffensen 43
5.4.6 Teoria generală a metodelor iterative în cazul ecuaţiilor neliniare . . 46

6 Sisteme de ecuaţii liniare **51**
6.1 Metode directe de rezolvare numerică a sistemelor liniare 52
6.1.1 Rezolvarea unor sisteme liniare particulare 52
6.1.1.1 Rezolvarea sistemelor liniare diagonale 52
6.1.1.2 Rezolvarea sistemelor liniare superior triunghiulare 52
6.1.1.3 Rezolvarea sistemelor liniare inferior triunghiulare 53
6.1.2 Metoda lui Gauss 54
6.1.3 Aplicaţii ale metodei lui Gauss 59
6.1.3.1 Calculul determinantului folosind metoda lui Gauss 59
6.1.3.2 Calculul matricii inverse folosind metoda lui Gauss 60
6.1.3.3 Calculul rangului unei matrici folosind metoda lui Gauss . 61
6.1.4 Metode de factorizare 62
6.1.4.1 Metoda descompunerii LU 62
6.1.4.2 Metoda descompunerii LL^T (Cholesky) 67
6.1.4.3 Metoda descompunerii QR (Householder) 71
6.1.4.4 Sisteme de ecuaţii liniare cu matrice tridiagonală 75
6.2 Norme vectoriale şi matriciale 80
6.3 Perturbaţii 83
6.4 Metode iterative pentru rezolvarea sistemelor liniare 84
6.4.1 Metoda lui Jacobi 84
6.4.2 Metoda lui Gauss-Seidel 88
6.4.3 Teoria generală a metodelor iterative pentru sistemele liniare 91
6.4.4 Metoda SOR 94
6.5 Rezolvarea sistemelor liniare supradeterminate cu metoda celor mai mici pătrate 97

7 Sisteme de ecuaţii neliniare pe spaţii finit dimensionale 101

7.1 Metoda lui Jacobi pe spaţiul finit dimensional $\mathbb{R}^n$ 101

7.2 Metoda lui Newton-Raphson-Kantorovici pe $\mathbb{R}^n$ 105

7.3 Metoda gradientului pe $\mathbb{R}^n$ 116

8 Aproximarea funcţiilor 119

8.1 Aproximarea uniformă a funcţiilor continue cu ajutorul polinoamelor . . . 119

8.1.1 Teorema lui Weierstrass şi teorema lui Korovkin 119

8.1.2 Teorema lui Stone 128

8.2 Aproximarea funcţiilor prin interpolare 133

8.2.1 Interpolare liniară 135

8.2.2 Interpolarea polinomială a lui Lagrange 136

8.2.3 Polinoamele lui Cebâşev de speţa întâia 139

8.2.4 Evaluarea restului pentru polinomul de interpolare Lagrange 143

8.2.5 Teorema lui Faber asupra divergenţei procedeului de interpolare . . 144

8.2.6 Diferenţe finite şi divizate. Polinomul de interpolare al lui Newton . 148

8.3 Aproximarea funcţiilor cu ajutorul funcţiilor spline 153

8.4 Cea mai bună aproximare a funcţiilor în spaţii normate 157

9 Formulele de derivare şi integrare numerică 161

9.1 Formulele de derivare numerică 161

9.2 Formulele de integrare numerică 168

9.2.1 Formula dreptunghiului 168

9.2.2 Formula trapezului 173

9.2.3 Formula lui Simpson 176

10 Metode numerice pentru calculul valorilor şi vectorilor proprii ale unei matrici 183

10.1 Metoda lui Krylov 183

10.2 Metoda puterii 189

11 Metode numerice pentru rezolvarea ecuaţiilor diferenţiale şi ale sistemelor de ecuaţii diferenţiale 191

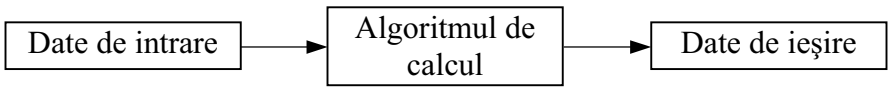
11.1 Metoda lui Euler pentru rezolvarea numerică a ecuaţiei diferenţiale ordinare de ordinul unu ca problemă Cauchy 191
11.2 Metoda lui Runge-Kutta de ordinul patru 195
11.3 Rezolvarea numerică a problemei lui Dirichlet pe un pătrat 197

Bibliografie **199**

Capitolul 1

Introducere

Metoda numerică (algoritmul numeric) este o metodă de rezolvare a unei probleme practice utilizând un număr finit de operaţii aritmetice şi logice (operaţiile uzuale pe care le poate executa un procesor sau coprocesor matematic). În practică apar probleme concrete cu date de intrare cunoscute. De obicei se asociază un model matematic acestei probleme, mai fin sau mai puţin fin. Rezolvarea problemei matematice în general nu se poate face cu mâna printr-un număr finit de paşi (operaţii), deci se caută să se rezolve problema printr-o metodă numerică. Algoritmul găsit se poate programa pe un calculator, iar rezultatele obţinute prin calculator se verifică practic. Aceste date de ieşire ar trebui să fie o aproximare reală pentru problema practică iniţială.



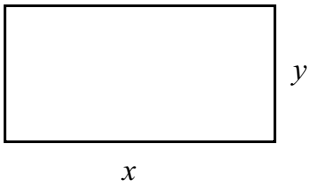
Un algoritm trebuie să aibă următoarele proprietăţi:

1. **Generalitate**, prin care se înţelege că algoritmul nu trebuie să rezolve numai o problemă ci toate problemele din clasa respectivă.

2. **Finitudine**, adică numărul de transformări intermediare aplicate datelor de intrare pentru a obţine datele de ieşire este finit.

3. **Unicitatea algoritmului**, adică transformările intermediare trebuie să fie unic determinate.

Exemplul 1. Se consideră următoarea problemă izoperimetrică din geometria elementară: dintre toate dreptunghiurile cu perimetrul constant să se determine cel cu arie maximă. (Rezolvare: pătrat)



Modelul matematic:

$$\begin{aligned} 2x + 2y &= \text{constant, deci} \\ x + y &= c \text{ (o constantă)} \\ x \cdot y &\to \text{maxim} \end{aligned}$$

Rezolvarea modelului matematic se poate face în acest caz fără calculator:

a) prin aplicarea materiei de gimnaziu se obţine: din inegalitatea mediilor avem: $\sqrt{xy} \le \dfrac{x+y}{2} = \dfrac{c}{2}$, deci $xy \le \frac{c^2}{4}$ şi egalitatea se realizează pentru $x = y = \dfrac{c}{2}$.

b) prin aplicarea materiei de liceu se obţine:

$$y = c - x, \quad x \cdot y = x(c-x) = cx - x^2,$$

se caută punctul maxim pentru funcţia $f : \mathbb{R} \to \mathbb{R}$, $f(x) = -x^2 + cx$.

Atunci $x_{\max} = \dfrac{c}{2}$ şi $y = \dfrac{c}{2}$ sunt datele de ieşire.

c) prin aplicarea materiei de universitate se obţine: să se calculeze maximul lui $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x,y) = x \cdot y$ cu restricţia $x+y-c = 0$. Se aplică metoda multiplicatorilor lui Lagrange: $L(x, y; \lambda) = xy + \lambda(x + y - c)$. Se calculează derivatele parţiale:

$$\begin{cases} \dfrac{\partial L}{\partial x} = 0 \\ \dfrac{\partial L}{\partial y} = 0 \\ \dfrac{\partial L}{\partial \lambda} = 0 \end{cases} \quad \begin{cases} y + \lambda = 0 \\ x + \lambda = 0 \\ x + y - c = 0 \end{cases} \quad \begin{cases} y = -\lambda \\ x = -\lambda \\ -2\lambda - c = 0 \end{cases} \quad \begin{cases} \lambda = \dfrac{-c}{2} \\ x = \dfrac{c}{2} \\ y = \dfrac{c}{2} \end{cases}$$

În final se face interpretarea rezultatelor teoretice obţinute. Rezultatul este un pătrat de latura $\dfrac{c}{2}$.

Exemplul 2. Să se rezolve următorul sistem liniar, care se poate obţine în urma unei probleme practice:

$$\begin{cases} a_{11}x + a_{12}y = b_1 \\ a_{21}x + a_{22}y = b_2 \end{cases}$$

Datele de intrare sunt parametrii $a_{11}, a_{12}, a_{21}, a_{22}, b_1, b_2$, datele de ieşire sunt x, y. Presupunem că: $\Delta = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \neq 0$, deci avem un sistem Cramer. Atunci: $x = \dfrac{\begin{vmatrix} b_1 & a_{12} \\ b_2 & a_{22} \end{vmatrix}}{\Delta}$ şi $y = \dfrac{\begin{vmatrix} a_{11} & b_1 \\ a_{21} & b_2 \end{vmatrix}}{\Delta}$.

Program sistem

Date de intrare: a_{ij}, b_i pentru $i, j = \overline{1,2}$.

Subrutina calcul (a_1, a_2, a_3, a_4)

calcul: $= a_1 * a_4 - a_2 * a_3$

$\Delta =$ calcul $(a_{11}, a_{12}, a_{21}, a_{22})$

Dacă $\Delta \neq 0$ atunci

$x :=$ calcul $(b_1, a_{12}, b_2, a_{22})/\Delta$

$y :=$ calcul $(a_{11}, b_1, a_{21}, b_2)/\Delta$

altfel nu este sistem de tip Cramer.

Tipăreşte: x, y.

La acest exemplu se observă că dacă se consideră un sistem de tip Cramer cu zece ecuaţii şi cu zece necunoscute şi vrem să-l rezolvăm folosind teoria determinanţilor din liceu, atunci în dezvoltarea unui determinant de ordinul zece, după definiţie am avea 10! de termeni care şi în cazul calculatorului reprezintă un volum imens de calcule. De aceea este necesar de a găsi alte metode de rezolvare a problemei, numite metode numerice.

Exemplul 3. Fie sistemul:

$$\begin{cases} 4,0000x + 0,8889y = 4,0000 \\ 1,0000x + 0,2222y = 1,0000. \end{cases}$$

Cum $\Delta \neq 0$, este un sistem Cramer şi are soluţia unică $x = 1$ şi $y = 0$.

Se face o mică perturbare a coeficienţilor sistemului. Se consideră sistemul cu trei zecimale exacte:

$$\begin{cases} 4,000x + 0,888y = 4,000 \\ 1,000x + 0,222y = 1,000. \end{cases}$$

Acest sistem este un sistem nedeterminat. Se păstrează continuitatea datelor de intrare, însă la datele de ieşire se produce nu un salt cantitativ, ci calitativ. Acest fenomen se numeşte instabilitatea numerică a sistemului liniar. Din punct de vedere geometric unghiul dintre cele două drepte este foarte mic, iar perturbaţia face ca dreptele să coincidă.

Capitolul 2

Erori

2.1 Surse de erori

Soluţiile obţinute prin metodele numerice sunt aproximative datorită erorilor. Surse de erori pot fi:

- gradul de adecvare al modelului matematic. Dacă modelul matematic este mai fin erorile se pot diminua.

- erori iniţiale sau erori în datele de intrare. Erorile iniţiale se formează din erori de măsurare datorite impreciziei instrumentului de măsurat. Erorile de observaţie sunt neregulate sau întâmplătoare.

- erorile de metodă. Apar prin folosirea unei metode numerice.

- erorile de calcul, care sunt de două tipuri: de trunchiere şi de rotunjire. Eroarea de trunchiere se obţine, de exemplu, prin calculul sumei unei serii înlocuind-o cu o sumă parţială. Eroarea de rotunjire se obţine în felul următor: dacă pe ultima zecimală a unui număr real avem cifrele 0, 1, 2, 3, 4 atunci denumim prin lipsă, când ultima cifră se lasă la o parte şi penultima cifră zecimală rămâne neschimbată, altfel denumim prin adaos, când ultima cifră zecimală lăsată la o parte este 5, 6, 7, 8, 9 şi penultima cifră se măreşte cu o unitate.

2.2 Studiul erorilor sau propagarea erorilor

Fie $a \in \mathbb{R}$ valoarea exactă sau ideală a unei mărimi. În practică în locul valorii exacte "a" se lucrează cu valoarea aproximativă $\tilde{a} \in \mathbb{R}$. Considerând în locul lui "a" pe "$\tilde{a}$" se comite o eroare care trebuie măsurată. Vom nota $\Delta(\tilde{a}) = |a - \tilde{a}|$ şi se foloseşte denumirea de **eroare absolută**.

Exemplu. Să se determine a cu două zecimale exacte. Acest lucru este posibil dacă se cunoaşte valoarea practică $\tilde{a}$ şi eroarea absolută $\Delta(\tilde{a}) \leq 10^{-2}$.

Valoarea

$$\delta(\tilde{a}) = \frac{|a - \tilde{a}|}{|\tilde{a}|} = \frac{\Delta(\tilde{a})}{|\tilde{a}|}$$

se numeşte **eroare relativă** dacă $\tilde{a} \neq 0$. Se poate întâmpla ca în alte lucrări eroarea relativă să fie definită prin formula $\delta(\tilde{a}) = \dfrac{\Delta(\tilde{a})}{|a|}$ pentru $a \neq 0$.

Numărul $\bar{\Delta}(\tilde{a})$ se numeşte o **limită pentru eroarea absolută** dacă $\bar{\Delta}(\tilde{a}) \geq \Delta(\tilde{a})$.

Numărul $\bar{\delta}(\tilde{a})$ se numeşte o **limită pentru eroarea relativă** dacă $\bar{\delta}(\tilde{a}) \geq \delta(\tilde{a})$.

Dacă $\bar{\Delta}(\tilde{a})$ este o limită pentru eroarea absolută, atunci $\bar{\delta}(\tilde{a}) := \dfrac{\bar{\Delta}(\tilde{a})}{|\tilde{a}|}$ este o delimitare pentru eroarea relativă.

2.3 Propagarea erorilor la cele patru operaţii aritmetice

Fie $a, b \in \mathbb{R}$ două valori exacte, $\tilde{a}, \tilde{b} \in \mathbb{R}$ valorile aproximative, iar $\Delta(\tilde{a})$, $\Delta(\tilde{b})$ erorile absolute respective.

Ne interesează $\Delta(\widetilde{a+b})$, adică o evaluare pentru eroarea absolută care se comite la adunare. Avem:

$$\Delta(\widetilde{a+b}) = |(a+b) - (\widetilde{a+b})| = |(a+b) - (\tilde{a}+\tilde{b})| \leq |a-\tilde{a}| + |b-\tilde{b}| = \Delta(\tilde{a}) + \Delta(\tilde{b}).$$

Analog şi la scădere:

$$\Delta(\widetilde{a-b}) = |(a-b) - (\widetilde{a-b})| = |(a-b) - (\tilde{a}-\tilde{b})| \leq |a-\tilde{a}| + |b-\tilde{b}| = \Delta(\tilde{a}) + \Delta(\tilde{b}).$$

La înmulţire se obţine pe rând:

$$\begin{aligned} \Delta(\widetilde{a \cdot b}) &= |ab - \widetilde{ab}| = |ab - \tilde{a}\cdot\tilde{b}| = |(ab - \tilde{a}b) + (\tilde{a}b - \tilde{a}\tilde{b})| \leq \\ &\leq \Delta(\tilde{a})\cdot|b| + \Delta(\tilde{b})\cdot|\tilde{a}| \leq \Delta(\tilde{a})\cdot\Delta(\tilde{b}) + \Delta(\tilde{a})\cdot|\tilde{b}| + \Delta(\tilde{b})\cdot|\tilde{a}|, \end{aligned}$$

căci $|b| = |b - \tilde{b} + \tilde{b}| \leq \Delta(\tilde{b}) + |\tilde{b}|$.

Cum valoarea $\Delta(\tilde{a}) \cdot \Delta(\tilde{b})$ este mică în comparaţie cu partea $\Delta(\tilde{a}) \cdot |\tilde{b}| + \Delta(\tilde{b}) \cdot |\tilde{a}|$, adică $\Delta(\tilde{a}) \cdot \Delta(\tilde{b}) \ll \Delta(\tilde{a}) \cdot |\tilde{b}| + \Delta(\tilde{b}) \cdot |\tilde{a}|$ (mult mai mică), de aceea la înmulţire se reţine numai partea: $\Delta(\widetilde{a \cdot b}) \leq \bar{\Delta}(\widetilde{a \cdot b}) \approx \Delta(\tilde{a}) \cdot |\tilde{b}| + \Delta(\tilde{b}) \cdot |\tilde{a}|$.

La împărţire se obţine pe rând:

$$\begin{aligned} \Delta(\widetilde{a/b}) &= \left| \frac{a}{b} - \widetilde{\left(\frac{a}{b}\right)} \right| = \left| \frac{a}{b} - \frac{\tilde{a}}{\tilde{b}} \right| = \frac{|a \cdot \tilde{b} - \tilde{a} \cdot b|}{|b \cdot \tilde{b}|} = \frac{|a \cdot \tilde{b} - \tilde{a}\tilde{b} + \tilde{a}\tilde{b} - \tilde{a}b|}{|b| \cdot |\tilde{b}|} \leq \\ &\leq \frac{\Delta(\tilde{a}) \cdot |\tilde{b}| + \Delta(\tilde{b}) \cdot |\tilde{a}|}{|b| \cdot |\tilde{b}|} \end{aligned}$$

şi cum $|\tilde{b}| = |b - \tilde{b} - b| \leq \Delta(\tilde{b}) + |b|$ rezultă că $|b| \geq |\tilde{b}| - \Delta(\tilde{b}) > 0$, adică

$$\Delta(\widetilde{a/b}) \leq \frac{\Delta(\tilde{a}) \cdot |\tilde{b}| + \Delta(\tilde{b}) \cdot |\tilde{a}|}{|\tilde{b}| \cdot (|\tilde{b}| - \Delta(\tilde{b}))}.$$

În cazul erorilor relative se pot deduce formule asemănătoare.

2.4 Formula generală a propagării erorilor sau eroarea introdusă de o funcţie

Dacă operaţiile aritmetice le concepem sub formă de funcţii (ca operaţii interne), avem de exemplu: $f = f(x) = f((x_1, x_2)) = x_1 + x_2$. O generalizare naturală este următoarea: se consideră funcţia $f = f(x) = f(x_1, x_2, \ldots, x_n)$ şi fie punctul $x^0 = (x_1^0, x_2^0, \ldots, x_n^0)$. Să se calculeze $f(x^0)$ unde x^0 este o valoare ideală şi $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n)$ este o aproximare pentru punctul x^0. Se poate calcula $f(\tilde{x})$ şi valoarea $\Delta(\widetilde{f(x^0)}) = |f(\tilde{x}) - f(x^0)|$ va fi eroarea comisă.

Dar $f(\tilde{x}) - f(x^0) = df(\xi) \cdot (\tilde{x} - x^0)$ din formula de medie a lui Lagrange, cu ξ o valoare intermediară pe segmentul determinat de capetele x^0 şi $\tilde{x}$. Din reprezentarea diferenţialei lui Fréchet cu ajutorul derivatelor parţiale se obţine că:

$$f(\tilde{x}) - f(x^0) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\xi) \cdot (\tilde{x}_i - x_i^0) \text{ adică}$$

$$|f(\tilde{x}) - f(x^0)| \leq \sum_{i=1}^{n} \left| \frac{\partial f}{\partial x_i}(\xi) \right| \cdot |\tilde{x}_i - x_i^0|.$$

Deoarece aproximarea se face în jurul valorii x^0, se poate presupune că pentru un domeniu D_0 în jurul valorii x^0, derivatele parţiale $\dfrac{\partial f}{\partial x_i}$ sunt mărginite, adică pentru $i = \overline{1,n}$ avem $\sup\left\{\left|\dfrac{\partial f}{\partial x_i}(x)\right| / x \in D_0\right\} = M_i$. Astfel $\Delta(\widetilde{f(x^0)}) \leq \sum_{i=1}^{n} M_i \Delta(\widetilde{x_i^0})$. Dacă se precizează de la bun început eroarea permisă ε, atunci se urmăreşte ca să obţinem $|f(\tilde{x}) - f(x^0)| \leq \varepsilon$. Prin alegerea: $\Delta(\widetilde{x_i^0}) = |\tilde{x}_i - x_i^0| \leq \dfrac{\varepsilon}{n \cdot M_i}$ se realizează abaterea cerută, adică $\tilde{x}_i$ trebuie luat destul de aproape de x_i^0.

Pentru interpretare avem următoarea figură geometrică:

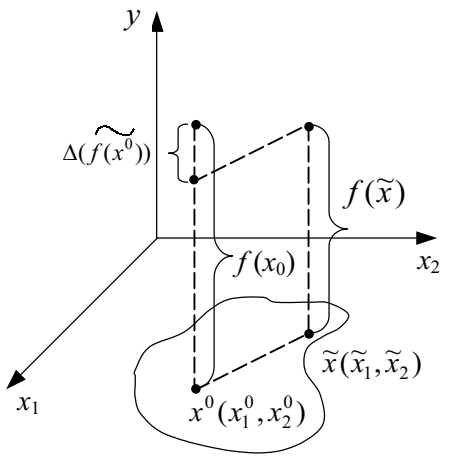


Figura 2.1:

De exemplu: dacă se dă funcţia $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x,y) = x^2 + 2y^2$, se cere ca să se evalueze eroarea care se comite dacă punctul $(0,0)$ se înlocuieşte cu punctul $(0,05; -0,025)$. Se alege dreptunghiul $D_0 = \{(x,y) \in \mathbb{R}^2 / \ |x| \leq 0,05, \ |y| \leq 0,025\}$. Se calculează:

$$M_1 = \sup\left\{\left|\frac{\partial f}{\partial x}(x,y)\right| / \ (x,y) \in D_0\right\} = \sup\{|2x| \ / \ |x| \leq 0,05, \ |y| \leq 0,025\} = 0,1 \quad \text{şi}$$

$$M_2 = \sup\left\{\left|\frac{\partial f}{\partial x}(x,y)\right| / \ (x,y) \in D_0\right\} = \sup\{|4y| \ / \ |x| \leq 0,5, \ |y| \leq 0,025\} = 0,1$$

Astfel $\Delta(\widetilde{f((0,0))}) \leq M_1 \cdot 0,05 + M_2 \cdot 0,025 = 0,1 \cdot 0,05 + 0,1 \cdot 0,025 = 0,0075$.

Capitolul 3

Folosirea sumelor şi a seriilor numerice în aproximări

Fie $\sum_{n=1}^{\infty} a_n$ o serie numerică convergentă, având suma seriei $S = a_1 + a_2 + \ldots + a_n + \ldots$. Se notează cu $S_n = \sum_{i=1}^{n} a_i$ şirul sumelor parţiale şi $R_n = \sum_{i=n+1}^{\infty} a_i$ şirul resturilor seriei pentru fiecare $n \neq 0$ număr natural. Cum seria este convergentă avem: $\lim_{n\to\infty} S_n = S$. În locul sumei S se consideră suma parţială $S_n (S \approx S_n)$ şi astfel se comite o eroare. Valoarea erorii absolute este $|S - S_n| = |R_n|$. Cum seria este covergentă avem $\lim_{n\to\infty} R_n = 0$, deci pentru orice $\varepsilon > 0$ precizie dinainte dată, rezultă faptul că $|R_n| < \varepsilon$, pentru $n \geq n(\varepsilon)$, $n(\varepsilon)$ fiind pragul corespunzător valorii lui ε.

3.1 Serii alternate

Seria $S := a_1 - a_2 + a_3 - a_4 + \ldots (-1)^{n+1} a_n + \ldots$, unde $a_i \geq 0$ pentru orice $i \in \mathbb{N}^*$ se numeşte alternată.

Avem: $|R_n| = |(-1)^{n+2} a_{n+1} + (-1)^{n+3} a_{n+2} + \ldots|$ şi se urmăreşte să obţinem o delimitare pentru rest. Dacă $n = 2m + 1$, atunci

$$\begin{aligned} |R_n| &= |-a_{2m+2} + a_{2m+3} - a_{2m+4} + a_{2m+5} + \ldots| = \\ &= |a_{2m+2} - a_{2m+3} + a_{2m+4} - a_{2m+5} + \ldots| \end{aligned}$$

Dacă $n = 2m$, atunci

$$|R_n| = |a_{2m+1} - a_{2m+2} + a_{2m+3} - a_{2m+4} + \ldots|.$$

Dacă se impune condiţia suplimentară ca şirul $\{a_i\}_{i\in\mathbb{N}^*}$ este monoton descrescător după un anumit rang, atunci în primul caz se obţine că $|R_n| = a_{2m+2} - a_{2m+3} + a_{2m+4} - a_{2m+5} + \ldots$, căci $a_{2m+2} \geq a_{2m+3}$, $a_{2m+4} \geq a_{2m+5}, \ldots$, deci

$$|R_n| = a_{2m+2} - a_{2m+3} + a_{2m+4} - a_{2m+5} + \ldots \leq a_{2m+2}$$

deoarece $a_{2m+3} \geq a_{2m+4}$, $a_{2m+5} \geq a_{2m+6}, \ldots$.

Procedând analog, în al doilea caz se obţine că

$$|R_n| = a_{2m+1} - a_{2m+2} + a_{2m+3} - a_{2m+4} + \ldots \leq a_{2m+1}.$$

Deci $|R_n| \leq \begin{cases} a_{2m+2} & \text{dacă } n = 2m+1 \\ a_{2m+1} & \text{dacă } n = 2m \end{cases}$ echivalent cu $|R_n| \leq a_{n+1}$. Deoarece $\lim\limits_{n\to\infty} a_{n+1} = 0$, pentru condiţia de oprire se poate impune cerinţa ca pentru orice $\varepsilon > 0$ precizie dinainte dată să avem $|R_n| \leq a_{n+1} \leq \varepsilon$.

Exemplu. Să se calculeze $\ln 2$ cu două zecimale exacte, folosind seria alternată $\ln 2 = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \ldots$. Din $|R_n| \leq a_{n+1} = \frac{1}{n+1} \leq 10^{-2}$ rezultă că $n \geq 99$. Prin urmare se calculează suma $S_{99} = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \ldots + \frac{1}{99}$.

Observaţie. Practica arată că pentru a obţine valoarea lui $\ln 2$ cu două zecimale exacte nu este necesar să adunăm 99 de termeni ai seriei.

Program serie alternată 1

Fie $S := 0$; $n := 99$; semn: $= -1$;

Pentru $i = \overline{1, 99}$ execută semn: $=$ semn $*(-1)$; $S := S +$ semn$/i$;

Tipăreşte S.

Un alt program pentru calculul lui $\ln 2$ cu o precizie ε se poate da în felul următor:

Program serie alternată 2

Fie $S' := 0$; $i := 0$; semn:$= -1$; $\varepsilon := 10^{-2}$;

Execută $S := S'$; $i := i + 1$; semn:$=$ semn $*(-1)$ $S' := S +$ semn$/i$;

până când $|S' - S| \geq \varepsilon$.

Tipăreşte S.

3.2 Serii cu termeni pozitivi

Fie $S = \sum_{n\geq 1} a_n$ o serie convergentă, S fiind suma seriei. Dacă $a_n \geq 0$, pentru orice $n \geq 1$, spunem că se dă o serie cu termeni pozitivi. Presupunem că termenii seriei verifică criteriul raportului (Criteriul lui d'Alembert) după un anumit rang încolo, adică $\frac{a_{n+1}}{a_n} \leq q$, unde $q < 1$ pentru $n \geq M$, $M \in \mathbb{N}$ fiind rangul corespunzător. Prin urmare $a_{M+1} \leq q \cdot a_M$, $a_{M+2} \leq q \cdot a_{M+1} \leq q^2 \cdot a_M$ şi aşa mai departe, se obţine:

$$\begin{aligned}\sum_{i\geq 1} a_i &\leq \sum_{i=1}^{M-1} a_i + a_M + q \cdot a_M + q^2 a_M + \ldots = \sum_{i=1}^{M-1} a_i + a_M(1 + q + q^2 + \ldots) = \\ &= \sum_{i=1}^{M-1} a_i + a_M \cdot \lim_{n\to\infty} \frac{1-q^n}{1-q} = \sum_{i=1}^{M-1} a_i + a_M \cdot \frac{1}{1-q}.\end{aligned}$$

Atunci se impune condiţia ca $R_M = a_M \cdot \frac{1}{1-q} \leq \varepsilon$ unde ε este precizia dinainte fixată.

Exemplu. Să se calculeze suma seriei $\sum_{n\geq 1} \frac{2^n}{n \cdot 3^n}$ cu două zecimale exacte.

În acest caz $\frac{a_{n+1}}{a_n} = \frac{\frac{2^{n+1}}{(n+1)3^{n+1}}}{\frac{2^n}{n\cdot 3^n}} = \frac{2}{3} \cdot \frac{n}{n+1} < \frac{2}{3} < 1$, deci $q = \frac{2}{3}$.

Se obţine că

$$R_M \leq \frac{2^M}{M \cdot 3^M} \cdot \frac{1}{1-\frac{2}{3}} = \frac{2^M}{M \cdot 3^{M-1}} < \varepsilon.$$

De aici se poate găsi o condiţie teoretică pentru rangul M, care ne asigură teoretic faptul că, dacă se face însumarea până la indicele M se obţine precizia dorită.

Program serie cu termeni pozitivi

Fie $S := \frac{2}{3}$, $x := \frac{2}{3}$, $n := 0$, $\varepsilon := 10^{-2}$;

Execută $n := n + 1$; $x := x * \frac{n}{n+1} * \frac{2}{3}$; $S := S + x$;

până când $3x > \varepsilon$.

Tipăreşte S.

3.3 Transformarea seriilor numerice slab convergente în serii mai rapid convergente

Se consideră două serii numerice $\sum_{n\geq 1} a_n$ şi $\sum_{n\geq 1} b_n$.

Definiţia 3.3.1. *Seria numerică* $\sum_{n\geq 1} a_n$ *este mai rapid convergentă decât seria* $\sum_{n\geq 1} b_n$ *dacă* $\lim_{n\to\infty} \frac{a_n}{b_n} = 0$.

Exemplu. Seria $\sum_{n\geq 1} \frac{1}{n^\alpha}$, $(\alpha > 1)$ este mai rapid convergentă, decât seria $\sum_{n\geq 1} \frac{1}{n^\beta}$, $(\beta > 1)$ dacă

$$\lim_{n\to\infty} \frac{a_n}{b_n} = \lim_{n\to\infty} \frac{n^\beta}{n^\alpha} = \lim_{n\to\infty} n^{\beta-\alpha} = 0$$

deci $\beta - \alpha < 0$, adică $\beta < \alpha$.

Observaţie. Se pot impune şi alte condiţii suficiente pentru a compara rapiditatea de convergenţă a două serii.

3.3.1 Transformarea lui Euler

Se consideră seria numerică convergentă $S = a_1 - a_2 + a_3 - a_4 + \ldots + (-1)^{n+1}a_n + \ldots$, $a_n \in \mathbb{R}$, pentru orice $n \in \mathbb{N}^*$. Menţionăm că orice serie numerică convergentă se poate considera sub forma anterioară. Se construieşte seria $S' = \frac{a_1}{2} - \frac{a_2 - a_1}{2} + \frac{a_3 - a_2}{2} - \frac{a_4 - a_3}{2} + \ldots + (-1)^{n+1}\frac{a_n - a_{n-1}}{2} + \ldots$, având termenul al $n-$lea $(-1)^{n+1} \cdot \frac{a_n - a_{n-1}}{2}$.

Se arată că noua serie are aceeaşi sumă ca şi seria iniţială. Într-adevăr $S_n - S'_n = (-1)^{n+1} \cdot \frac{a_n}{2}$, unde S_n şi S'_n sunt sumele parţiale de ordinul n ale seriilor cu sumele S şi S'. Din faptul că prima serie este convergentă rezultă că $a_n \to 0$, deci $\lim_{n\to\infty}(S_n - S'_n) = 0$. Prin urmare $S = S'$.

Mai trebuie arătat că a doua serie este mai rapid convergentă. Într-adevăr

$$\lim_{n\to\infty} \frac{(-1)^{n+1}\frac{a_n - a_{n-1}}{2}}{(-1)^{n+1}a_n} = \frac{1}{2}\lim_{n\to\infty}\left(1 - \frac{a_{n-1}}{a_n}\right) = 0,$$

dacă are loc condiţia $\lim_{n\to\infty} \frac{a_{n-1}}{a_n} = 1$. Deci, dacă este satisfăcută condiţia $\lim_{n\to\infty} \frac{a_{n-1}}{a_n} = 1$ pentru prima serie numerică, atunci prin transformarea lui Euler se obţine o serie mai rapid convergentă.

Exemplu. Se ştie că $\ln 2 = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \ldots + (-1)^{n+1}\frac{1}{n} + \ldots$. Dacă se aplică transformarea lui Euler se obţine

$$\begin{aligned}\ln 2 &= \frac{1}{2} - \frac{\frac{1}{2}-1}{2} + \frac{\frac{1}{3}-\frac{1}{2}}{2} - \frac{\frac{1}{4}-\frac{1}{3}}{2} + \ldots + (-1)^{n+1}\frac{\frac{1}{n}-\frac{1}{n-1}}{2} + \ldots = \\ &= \frac{1}{2} + \frac{1}{4} - \frac{1}{12} + \frac{1}{24} + \ldots + (-1)^n \frac{1}{2n(n-1)} + \ldots\end{aligned}$$

Program transformarea Euler

Fie $S := \frac{1}{2}$; $n := 1$; semn:$= -1$; $\varepsilon := 10^{-2}$;

Execută $n := n + 1$; semn: $=$ semn$*(-1)$; $S := S + \text{semn} * \frac{1}{2n(n-1)}$

până când $\frac{1}{2n(n-1)} \geq \varepsilon$

Tipăreşte S.

3.3.2 Transformarea lui Kummer

Se dă seria convergentă $\sum_{n\geq 1} a_n$ şi se consideră o serie ajutătoare: $u = \sum_{n\geq 1} u_n$ astfel încât $\lim_{n\to\infty} \frac{a_n}{u_n} = \lambda \in \mathbb{R}^*$. Transformarea lui Kummer defineşte o nouă serie $\lambda \cdot u + \sum_{n\geq 1}(a_n - \lambda \cdot u_n)$, care are aceeaşi sumă ca şi seria iniţială $\sum_{n\geq 1} a_n$. Mai trebuie să arătăm că

$$lim_{n\to\infty} \frac{a_n - \lambda u_n}{a_n} = 1 - \lambda \lim_{n\to\infty} \frac{u_n}{a_n} = 0.$$

Exemplu. Din analiză se cunoaşte că seria $\sum_{n\geq 1} \frac{1}{n^2}$ este convergentă şi suma seriei este egală cu $\frac{\pi^2}{6}$. Să se mărească rapiditatea convergenţei acestei serii, folosind transformarea lui Kummer. Se consideră seria ajutătoare $\sum_{n\geq 1} \frac{4}{4n^2-1} = 2$. În acest caz avem:

$$\lim_{n\to\infty} \frac{a_n}{u_n} = \lim_{n\to\infty} \frac{4n^2-1}{4n^2} = 1 = \lambda.$$

Deci

$$2 + \sum_{n\geq 1}\left(\frac{1}{n^2} - \frac{4}{4n^2-1}\right) = 2 + \sum_{n\geq 1} \frac{-1}{n^2(4n^2-1)} = 2 - \sum_{n\geq 1} \frac{1}{n^2(4n^2-1)}.$$

Program transformarea Kummer

Fie $S := 0$; $n := 0$; $\varepsilon := 10^{-2}$;
Execută $n := n + 1$; $S := S + \dfrac{1}{n^2(4n^2-1)}$;
Până când $\dfrac{1}{n^2(4n^2-1)} \geq \varepsilon$;
$S := 2 - S$;
Tipăreşte S.

Capitolul 4

Evaluarea funcţiilor

Evaluarea numerică a funcţiilor este una dintre problemele fundamentale ale calculului numeric. Este important să se găsească formule şi algoritmi corespunzători care să nu conducă la rezultate inacceptabile prin acumularea aberantă a erorilor.

4.1 Evaluarea polinoamelor

Fie polinomul $P(x) = a_n x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0$, $a_n \neq 0$ şi $a_i \in \mathbb{R}$ (sau $\mathbb{C}$) pentru $i = \overline{0, n}$. Dacă se dă $\xi \in \mathbb{R}$ (sau $\mathbb{C}$) să se calculeze $P(\xi)$. Pentru $P(\xi)$ avem următoarea scriere:

$$P(\xi) = (\ldots((a_n\xi + a_{n-1})\xi + a_{n-2})\xi + \ldots + a_1)\xi + a_0.$$

Conform schemei lui Horner:

	a_n	a_{n-1}	$\ldots$	a_0
ξ	a_n	$a_n \cdot \xi + a_{n-1}$	$\ldots$	$P(\xi)$

Program evaluarea polinoamelor

Date de intrare: a_i pentru $i = \overline{0, n}$; ξ;

Fie valpol:= a_n;

Pentru $i = \overline{1, n}$ execută: valpol: = valpol $* \xi + a_{n-i}$;

Tipăreşte valpol.

4.2 Evaluarea funcţiilor analitice

Se consideră funcţia $f : I \to \mathbb{R}$, unde I este un interval al axei reale şi $x_0 \in I$ un punct interior. Presupunem că f este derivabilă de o infinitate de ori în x_0. Spunem că f este **analitică** în x_0 dacă

$$f(x) = f(x_0) + \frac{f'(x_0)}{1!}(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \ldots + \frac{f^n(x_0)}{n!}(x - x_0)^n + \ldots$$

pentru orice $x \in (x_0 - \alpha, x_0 + \alpha) \subset I$ unde $\alpha > 0$ este dat.

Observaţie. În cazul funcţiilor reale noţiunea de analiticitate şi noţiunea de o infinitate de ori derivabilă nu sunt echivalente. Putem considera următorul contraexemplu:

$$f : \mathbb{R} \to \mathbb{R},\ f(x) = \begin{cases} e^{-1/x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

Această funcţie este derivabilă de o infinitate de ori pe $\mathbb{R}$, dar nu este analitică în $x_0 = 0$.

Introducem următoarele notaţii: C^1 clasa funcţiilor derivabile o dată şi cu prima derivată continuă, C^2 clasa funcţiilor de două ori derivabile şi cu a doua derivată continuă, şi în general C^n pentru $n \in \mathbb{N}^*$ clasa funcţiilor derivabile de n ori şi cu a n-a derivată continuă, C^∞ clasa funcţiilor de o infinitate de ori derivabile şi A mulţimea funcţiilor analitice. În cazul real avem următorul şir de incluziuni stricte: $C^1 \supsetneqq C^2 \supsetneqq \ldots \supsetneqq C^n \supsetneqq C^{n+1} \supsetneqq \ldots \supsetneqq C^\infty \supsetneqq A$. Însă în cazul funcţiilor complexe, are loc următorul şir de egalităţi:

$$C^1 = C^2 = \ldots = C^n = C^{n+1} = \ldots = C^\infty = A.$$

În cazul funcţiilor analitice dezvoltarea Taylorianã se descompune în polinomul lui Taylor de ordinul n :

$$T_n(x) = f(x_0) + \frac{f'(x_0)}{1!}(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \ldots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n,$$

şi în restul seriei Taylor de ordinul n :

$$R_n(x) = \frac{f^{(n+1)}(x_0)}{(n+1)!}(x - x_0)^{n+1} + \frac{f^{(n+2)}(x_0)}{(n+2)!}(x - x_0)^{n+2} + \ldots$$

Prin urmare avem reprezentarea funcţiei $f(x) = T_n(x) + R_n(x)$. Restul seriei Taylor se poate scrie în mai multe moduri, cel mai des fiind folosită forma Lagrange: $R_n(x) = \frac{f^{(n+1)}(\theta)}{(n+1)!}(x - x_0)^{n+1}$, unde $\theta \in (x_0, x)$ sau $\theta \in (x, x_0)$.

Funcţiile elementare sunt funcţii analitice fiindcă în cazul lor restul seriei lui Taylor tinde la zero.

Din analiză sunt cunoscute următoarele dezvoltări de tip Mac-Laurin pentru $x_0 = 0$:

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \ldots + \frac{x^n}{n!} + \ldots$$
$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} + \ldots + (-1)^n \frac{x^{2n+1}}{(2n+1)!} + \ldots$$
$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} + \ldots + (-1)^n \frac{x^{2n}}{(2n)!} + \ldots \text{ etc.}$$

Exemplu. Să se calculeze valoarea lui ε^ξ pentru $\xi \in \mathbb{R}$ fixat cu o precizie ε.

În cazul funcţiei exponenţiale restul lui Lagrange este de forma $\frac{e^\theta \xi^{n+1}}{(n+1)!}$. Pentru a obţine condiţiile de oprire procedăm în felul următor:

- dacă $\xi > 0$, atunci $\theta \in (0, \xi)$, deci vom impune ca $\frac{e^\xi \cdot \xi^{n+1}}{(n+1)!} < \varepsilon$, adică să fie satisfăcută condiţia $\frac{3^{[\xi]+1} \cdot \xi^{n+1}}{(n+1)!} < \varepsilon$;
- dacă $\xi < 0$, atunci $\theta \in (\xi, 0)$, deci vom impune condiţia teoretică $\frac{|\xi|^{n+1}}{(n+1)!} < \varepsilon$.

Program funcţii analitice

Fie $S := 1$; $x = \xi$; $n = 1$; ε; $y := \xi$;

Execută

dacă $\xi < 0$, atunci $S := S + x$; $n := n + 1$; $x := \frac{\xi}{n} x$;

altfel $S := S + x$; $n := n + 1$; $x := \frac{\xi}{n} x$; $y := \frac{\xi}{n} y$;

dacă $n \leq [\xi] + 1$, atunci $y := 3\frac{\xi}{n} y$;

altfel $y := \frac{\xi}{n} y$;

Până când $|y| \geq \varepsilon$.

Tipăreşte S.

4.3 Evaluarea funcţiilor date prin relaţii de recurenţă

Există o serie de funcţii speciale, care sunt definite prin relaţii de recurenţă, cum ar fi de exemplu polinoamele ortogonale, care satisfac următoarea relaţie de recurenţă:

$$a_i P_{i+1}(x) + (b_i + c_i x) P_i(x) + d_i P_{i-1}(x) = 0;\ i \geq 1,$$

unde numerele $a_i, b_i, c_i, d_i \in \mathbb{R}$ şi polinoamele $P_0(x)$ şi $P_1(x)$ sunt date.

Mai târziu (vezi paragraful 8.2.3) sunt tratate polinoamele lui Cebâşev date prin formula: $T_n(x) = \cos(n \arccos x)$, $x \in [-1, 1]$. Aceste polinoame verifică relaţia de recurenţă

$$T_{i+1}(x) - 2xT_i(x) + T_{i-1}(x) = 0,\ i \geq 1,\ T_0(x) = 1,\ T_1(x) = x.$$

Pentru un $\xi \in [-1, 1]$ şi $n \in \mathbb{N}^*$ dat să se calculeze $T_n(\xi)$.

Program relaţii de recurenţă

Fie $k := 1$; n; ξ; $x := 1$; $y := \xi$;

Execută $z := 2 * \xi * y - x$; $x := y$; $y := z$; $k := k + 1$;

Până când $k \neq n$.

Tipăreşte z.

4.4 Evaluarea funcţiilor cu ajutorul fracţiilor continue

O expresie de forma $a_0 + \dfrac{b_1}{a_1 + \frac{b_2}{a_2 + \dots}}$ se numeşte fracţie continuă şi se foloseşte notaţia $[a_0, b_1, a_1, b_2, a_2, \dots]$. În general elementele a_i, b_i ale fracţiei continue pot fi numere reale sau complexe sau funcţii de una sau mai multe variabile. Orice număr real se poate reprezenta sub forma unei fracţii continue, elementele fracţiei continue fiind numere întregi. De exemplu: $\sqrt{2} = 1 + \dfrac{1}{x}$, de unde $x = \dfrac{1}{\sqrt{2} - 1} = \sqrt{2} + 1$, deci $\sqrt{2} = 1 + \dfrac{1}{\sqrt{2} + 1}$. Însă $\sqrt{2} + 1 = 2 + \dfrac{1}{y}$, de unde $y = \dfrac{1}{\sqrt{2} - 1} = \sqrt{2} + 1$, deci $\sqrt{2} = 1 + \dfrac{1}{2 + \frac{1}{\sqrt{2}+1}}$, care se continuă în mod analog, deci $\sqrt{2} = [1; 1; 2; 1; 2; \dots]$.

Dacă fracţia continuă are un număr finit de termeni, ea se identifică cu fracţia obţinută prin reduceri succesive la numitor comun. De exemplu

$$1 + \frac{2}{3 + \frac{1}{4}} = 1 + \frac{2}{\frac{13}{4}} = 1 + \frac{8}{13} = \frac{21}{13} = [1; 2; 3; 1; 4].$$

Expresiile $R_0 = a_0$; $R_1 = a_0 + \dfrac{b_1}{a_1}$; $R_2 = a_0 + \dfrac{b_1}{a_1 + \frac{b_2}{a_2}}$; $\dots$ se numesc convergenţii fracţiei continue.

Dacă şirul convergenţilor este convergent şi $\lim\limits_{n \to \infty} R_n = A$, $A \in \mathbb{R}$, atunci prin definiţie numărul real A este valoarea atribuită la fracţia continuă. Dacă şirul convergenţilor este divergent, atunci fracţia continuă este divergentă.

Teorema 4.4.1. *(Legea formării convergenţilor) Fie* $P_0 = a_0$, $P_{-1} = 1$, $Q_0 = 1$, $Q_{-1} = 0$. *Se formează şirurile de numere* P_i *şi* Q_i *pentru* $i \geq 1$ *date prin relaţiile de recurenţă* $P_i = a_iP_{i-1} + b_iP_{i-2}$ *şi* $Q_i = a_iQ_{i-1} + b_iQ_{i-2}$. *Atunci* P_i *şi* Q_i *pentru* $i \geq 0$ *vor fi numărătorii respectiv numitorii convergenţilor* R_i.

DEMONSTRAŢIE. Se face prin inducţie matematică. Se verifică direct că

$$R_0 = \frac{P_0}{Q_0} = \frac{a_0}{1} = a_0 \quad \text{şi}$$
$$R_1 = \frac{P_1}{Q_1} = \frac{a_1P_0 + b_1P_1}{a_1Q_0 + b_1Q_1} = \frac{a_1 \cdot a_0 + b_1 \cdot 1}{a_1 \cdot 1 + b_1 \cdot 0} = a_0 + \frac{b_1}{a_1}.$$

Presupunem că $R_i = \dfrac{P_i}{Q_i}$ şi va trebui să demonstrăm că $R_{i+1} = \dfrac{P_{i+1}}{Q_{i+1}}$. Avem

$$R_i = \frac{P_i}{Q_i} = \frac{a_i \cdot P_{i-1} + b_i \cdot P_{i-2}}{a_i \cdot Q_{i-1} + b_i \cdot Q_{i-2}}.$$

Deoarece R_{i+1} se obţine din R_i, dacă expresia lui a_i se înlocuieşte prin $a_i + \dfrac{b_{i+1}}{a_{i+1}}$, rezultă că

$$\begin{aligned} R_{i+1} &= \frac{\left(a_i + \frac{b_{i+1}}{a_{i+1}}\right) P_{i-1} + b_iP_{i-2}}{\left(a_i + \frac{b_{i+1}}{a_{i+1}}\right) Q_{i-1} + b_iQ_{i-2}} = \frac{a_{i+1}(a_iP_{i-1} + b_iP_{i-2}) + b_{i+1}P_{i-1}}{a_{i+1}(a_iQ_{i-1} + b_iQ_{i-2}) + b_{i+1}Q_{i-1}} = \\ &= \frac{a_{i+1}P_i + b_{i+1}P_{i-1}}{a_{i+1}Q_i + b_{i+1}Q_{i-1}} = \frac{P_{i+1}}{Q_{i+1}}. \end{aligned}$$

Exemplu. Se consideră dezvoltarea funcţiei tangentă în fracţie continuă:

$$\text{tg}x = [0; x; 1; -x^2; 3; -x^2; 5; \ldots; -x^2; 2n-1; \ldots]$$

Să se determine valoarea tgξ pentru $\xi \in \left(-\dfrac{\pi}{2}, \dfrac{\pi}{2}\right)$ dat.

Program fracţii continue

Fie $k := 1$; ξ; ε; $R := \xi$; $A := \xi$; $B := 0$; $D := 1$; $E := 1$;

Execută $R' := R$; $C := B$; $B := A$; $F := E$; $E := D$; $k = k + 1$;

$A := (2k-1)B + (-\xi^2)C$; $D := (2k-1)E + (-\xi^2)F$;

$R := \dfrac{A}{D}$;

Până când $|R - R'| \geq \varepsilon$.

Tipăreşte R.

Capitolul 5

Ecuaţii neliniare

În acest capitol se tratează ecuaţiile neliniare algebrice şi transcendente cu o singură necunoscută.

Exemple.

1. $\sin 2x - \text{arctg}x + 0,3 = 0,\ x \in \mathbb{R}$ sau
2. $x^5 - 4x^2 + 0,3x^2 + \pi x - e = 0,\ x \in \mathbb{R}$.

Se observă că aceste ecuaţii nu se pot rezolva cu mâna folosind formulele de trigonometrie sau de algebră. Prima dată se discută rezolvarea ecuaţiilor polinomiale care au o structură aparte.

5.1 Metoda lui Bairstow

Fie ecuaţia polinomială $P_n(x) = a_n x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0 = 0$, $a_i \in \mathbb{R}$; $i = \overline{0,n}$ cu $a_n \neq 0$. Cu această metodă se determină cu aproximaţie divizorii polinomiali de gradul doi ai polinomului dat. Dacă pe polinoamele de gradul doi astfel obţinuţi egalăm cu zero atunci se obţin rădăcinile polinomului iniţial. Dacă $n = 2m$ atunci avem m bucăţi de polinoame de gradul doi, care au fiecare sau două rădăcini reale sau două rădăcini complex conjugate, respectiv dacă $n = 2m + 1$ pe lângă m bucăţi de polinoame de gradul doi mai obţinem şi un polinom de gradul unu.

Fie $m_1(x) = x^2 + p_1 x + q_1$, unde se poate alege $p_1 = \dfrac{a_{n-1}}{a_n}$ şi $q_1 = \dfrac{a_{n-2}}{a_n}$. Atunci $P_n(x) = m_1(x)Q_1(x) + A_1 x + B_1$. Mai departe scriem pe $Q_1(x) = m_1(x)Q_2(x) + a_1^* x + b_1^*$.

Se rezolvă sistemul

$$\begin{cases} (b_1^* - a_1^* p_1)p_1^* + a_1^* q_1^* = A_1 \\ -a_1^* q_1 p_1^* + b_1^* q_1^* = B_1 \end{cases}$$

pentru necunoscutele p_1^* şi q_1^*.

Se face corecţia $p_2 = p_1 + p_1^*$ şi $q_2 = q_1 + q_1^*$.

În locul polinomului de gradul doi $m_1(x)$ considerăm polinomul de gradul doi $m_2(x) = x^2 + p_2 x + q_2$, care are coeficienţii mai precişi. Aceşti paşi se repetă până când coeficienţii polinomului de gradul doi nu ating precizia dorită, după care factorul de gradul doi $m_k(x)$ astfel obţinut se anulează şi se rezolvă.

Vom obţine un nou polinom $P_{n-2}(x) = \dfrac{P_n(x)}{m_k(x)}$, pentru care iarăşi aplicăm procedeul de mai sus, considerând în locul lui $P_n(x)$ pe $P_{n-2}(x)$.

Program metoda Bairstow

Datele de intrare ε; n (gradul polinomului); $Q[n]$-vector

$$Q[0] = a_n,\ Q[1] = a_{n-1}, \ldots, Q[n] = a_0.$$

(se verifică $Q[0] \neq 0$).

Repetă $P := Q$;

(Calculează polinomul $M[2]$)

$$\begin{aligned} M[0] &:= 1; \\ M[1] &:= \frac{P[1]}{P[0]}; \\ M[2] &:= \frac{P[2]}{P[0]}; \end{aligned} \tag{5.1}$$

Repetă

Aplică subrutina împărţire (P, M, Q_1, R_1)

Aplică subrutina împărţire (Q_1, M, Q_2, R_2)

Calculează pe p_1^* şi q_1^* (Direct cu subrutina de calcul al sistemului liniar de două ecuaţii şi cu două necunoscute)

$M[1] := M[1] + p_1^*$; $M[2] := M[2] + q_1^*$.

Până când $|p_1^*| + |q_1^*| \geq \varepsilon$

Rezolvă ecuaţia $M = 0$ (de gradul doi) $x[n] :=$ şi $x[n-1] :=$.

(subrutină de rezolvare a ecuaţiei de gradul doi)

Aplică subrutina de împărţire (P, M, Q, R);

$n := n - 2$;

Până când $n \geq 3$

Rezolvă $Q = 0$ (Dacă $n = 2$ $x[2] :=$ $x[1] :=$)

(Dacă $n = 1$ $x[1] :=$)

Tipăreşte x ($x[i]$ pentru $i = \overline{1, n}$)

5.2 Separarea rădăcinilor reale în cazul ecuaţiilor neliniare reale

Se consideră ecuaţia sub forma generală $f(x) = 0$ unde $f : D \to \mathbb{R}$, $D \subseteq \mathbb{R}$, $D \neq \emptyset$ şi trebuie găsită soluţia $x^* \in D$.

Prima dată se efectuează separarea rădăcinilor, adică a găsi intervale în domeniul de definiţie D, astfel încât fiecare interval să conţină o singură rădăcină a ecuaţiei.

Prima metodă de acest gen este şirul lui Rolle. Presupunem că $f \in C^1(D)$, adică este o funcţie derivabilă pe D şi cu derivata întâia continuă. În acest caz se rezolvă ecuaţia $f'(x) = 0$, $x \in D$. Dacă $x_1, x_2, \ldots, x_n$ sunt rădăcinile ecuaţiei derivate, atunci se formează şirul lui Rolle cu valorile $f(x_1), f(x_2), \ldots, f(x_n)$. Orice schimbare de semn la doi termeni consecutivi $f(x_i), f(x_{i+1})$ garantează existenţa unei singure rădăcini în intervalul (x_i, x_{i+1}), conform teoremei lui Rolle.

Exemplu. Să se găsească intervalele de separare ale rădăcinilor ecuaţiei: $2x^3 - 9x^2 + 12x - 4,5 = 0$. Se consideră funcţia polinomială $f : \mathbb{R} \to \mathbb{R}$, $f(x) = 2x^3 - 9x^2 + 12x - 4,5$ şi ecuaţia neliniară, algebrică corespunzătoare $f(x) = 0$, $x \in \mathbb{R}$. Se ataşează ecuaţia $f'(x) = 0$, adică $x^2 - 3x + 2 = 0$ cu rădăcinile $x_1 = 1$, $x_2 = 2$. Se întocmeşte următorul tabel folosind şirul lui Rolle:

x	$-\infty$	1	2	$+\infty$
$f(x)$	$f(-\infty) = -\infty$	$f(1) = 0,5$	$f(2) = -0,5$	$f(+\infty) = +\infty$

Se observă că ecuaţia $f(x) = 0$ admite trei rădăcini reale în intervalele $(-\infty, 1)$; $(1, 2)$; $(2, +\infty)$. Deoarece $f(0) = -4,5 < 0$ şi $f(3) = 4,5 > 0$ se poate afirma că cele trei rădăcini reale ale ecuaţiei polinomiale se află în intervalele $(0, 1); (1, 2); (2, 3)$.

Ca a două metodă se poate folosi polinomul de interpolare al lui Lagrange sau al lui Newton. Dacă în domeniul de definiţie al funcţiei $f : D \to \mathbb{R}$, $D \subseteq \mathbb{R}$ se aleg punctele $x_1 < x_2 < \ldots < x_n$ atunci se construieşte polinomul de interpolare care trece prin punctele $(x_1, f(x_1)), (x_2, f(x_2)), \ldots, (x_n, f(x_n))$. Rădăcinile acestui polinom (vezi metoda lui Bairstow, paragraful 5.1) pot fi considerate ca aproximări pentru rădăcinile ecuaţiei $f(x) = 0$. În capitolul interpolări se dă efectiv construcţia polinomului de interpolare (vezi paragrafele 8.2.2 şi 8.2.6).

5.3 Metoda bisecţiei sau metoda înjumătăţirii intervalului

Fie funcţia $f : [a, b] \to \mathbb{R}$ de clasă $C([a, b])$ (continuă pe intervalul $[a, b]$) care admite o singură rădăcină x^* în intervalul $[a, b]$ (presupunem că, deja rădăcinile ecuaţiei $f(x) = 0$, $x \in D$ sunt separate). Acest lucru se întâmplă dacă $f(a) \cdot f(b) \leq 0$. Metoda înjumătăţirii intervalului constă în următoarea procedură: intervalul $[a, b]$ se divide în două părţi egale folosind mijlocul intervalului, punctul $\dfrac{a+b}{2}$. Rădăcina ecuaţiei, punctul x^* se află într-unul din intervalele $\left[a, \dfrac{a+b}{2}\right]$ sau $\left[\dfrac{a+b}{2}, b\right]$, după care procedeul se repetă pentru unul din aceste subintervale în care se află rădăcina x^*.

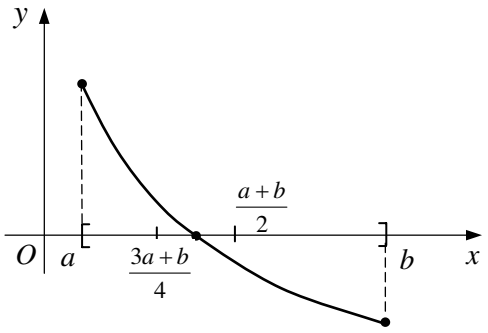


Figura 5.1:

Se construiesc două şiruri $\{a_n\}_{n\in\mathbb{N}}$, $\{b_n\}_{n\in\mathbb{N}}$ în felul următor: se alege $a_0 = a$, $b_0 = b$. Avem una dintre următoarele situaţii: $f(a) \cdot f\left(\dfrac{a+b}{2}\right) \leq 0$ sau $f\left(\dfrac{a+b}{2}\right) \cdot f(b) \leq 0$, corespunzător cărora se aleg $a_1 = a$ şi $b_1 = \dfrac{a+b}{2}$ respectiv $a_1 = \dfrac{a+b}{2}$ şi $b_1 = b$ etc. La pasul n găsim intervalul $[a_n, b_n]$ în care se află rădăcina şi iarăşi aplicăm bisecţia acestui

interval. Dacă $f(a_n) \cdot f\left(\frac{a_n+b_n}{2}\right) \leq 0$, atunci $a_{n+1} = a_n$ şi $b_{n+1} = \frac{a_n+b_n}{2}$, iar dacă $f\left(\frac{a_n+b_n}{2}\right) \cdot f(b_n) \leq 0$, atunci $a_{n+1} = \frac{a_n+b_n}{2}$ şi $b_{n+1} = b_n$.

Teorema 5.3.1. *Şirul $\{a_n\}_{n\in\mathbb{N}}$ este monoton crescător şi şirul $\{b_n\}_{n\in\mathbb{N}}$ este monoton descrescător, având aceeaşi limită, soluţia x^* a ecuaţiei $f(x) = 0$.*

DEMONSTRAŢIE. Monotonia celor două două şiruri rezultă din construcţia lor. Deoarece x^* este soluţia ecuaţiei $f(x) = 0$ avem $a_n \leq x^* \leq b_n$ pentru orice $n \in \mathbb{N}$, deci rezultă şi mărginirea celor două şiruri. Prin urmare şirurile sunt convergente şi trecând la limită în relaţia $b_n - a_n = \frac{b-a}{2^n}$ se obţine că $\lim\limits_{n\to\infty} a_n = \lim\limits_{n\to\infty} b_n = x^*$. Se observă că de fapt cele două şiruri reprezintă un cleşte pentru localizarea soluţiei x^*.

Program metoda bisecţiei

Fie a; b; f; ε (precizia dorită)
Execută $x := \frac{a+b}{2}$;
dacă $f(a) * f(x) \leq 0$, atunci $b := x$;
altfel $a := x$;
Până când $b - a \geq \varepsilon$.
Tipăreşte $x = \frac{a+b}{2}$.

5.4 Metode iterative pentru rezolvarea ecuaţiilor neliniare

La aceste metode prima dată ecuaţia $f(x) = 0$, $x \in D$ se transformă sub o formă iterativă $\varphi(x) = x$, $x \in D'$, adică din funcţia $f : D \to \mathbb{R}$ se construieşte o nouă funcţie iterativă $\varphi : D' \to \mathbb{R}$ astfel încât soluţia x^* a ecuaţiei $f(x) = 0$ să fie soluţie pentru ecuaţia $\varphi(x) = x$ şi invers: $f(x^*) = 0$ dacă şi numai dacă $\varphi(x^*) = x^*$. Menţionăm că un punct $x^* \in D'$ pentru care $\varphi(x^*) = x^*$ se numeşte **punct fix** pentru funcţia iterativă φ.

Exemplu. Fie $f : [1, 2] \to \mathbb{R}$, $f(x) = x^2 - 2$, care generează ecuaţia $x^2 - 2 = 0$. Asupra acestei ecuaţii efectuăm următoarele transformări echivalente: $x^2 - 2 = 0$; $x^2 = 2$; $x = \frac{2}{x}$; $2x = x + \frac{2}{x}$; $x = \frac{1}{2}\left(x + \frac{2}{x}\right)$. Se alege funcţia iterativă $\varphi : [1, 2] \to \mathbb{R}$, $\varphi(x) = \frac{1}{2}\left(x + \frac{2}{x}\right)$.

Soluţia ecuaţiei $x^2 - 2 = 0$ este $x^* = \sqrt{2}$ care se poate obţine cu ajutorul şirului iterativ $\{x_k\}_{k\in\mathbb{N}}$ generat de $x_{k+1} = \varphi(x_k)$ şi $x_0 = 1$.

5.4.1 Metoda tangentei sau metoda lui Newton

Se consideră funcţia $f : [a, b] \to \mathbb{R}$ de clasă $C^1([a, b])$ şi presupunem că în intervalul $[a, b]$ ecuaţia $f(x) = 0$ admite o singură rădăcină $x^* \in [a, b]$.

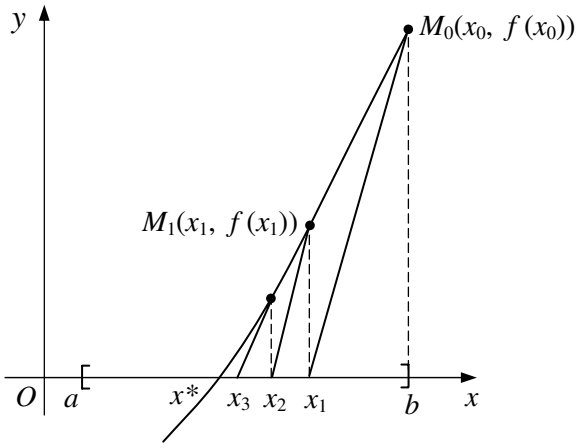


Figura 5.2:

După figură se poate urmări construcţia şirului iterativ $\{x_k\}_{k\in\mathbb{N}}$. Se alege ca punct de plecare $x_0 = b$ şi în punctul $M_0(x_0, f(x_0))$ se duce o tangentă la graficul funcţiei f care taie axa Ox în punctul x_1, după care tangenta în punctul $M_1(x_1, f(x_1))$ intersectează din nou axa Ox în punctul x_2 etc. Şirul $\{x_k\}_{k\in\mathbb{N}}$ astfel construit converge către x^*. Folosind formalismul matematic ecuaţia tangentei la graficul funcţiei f în punctul $M_n(x_n, f(x_n))$ este: $y - f(x_n) = f'(x_n)(x - x_n)$. Intersecţia tangentei cu axa Ox ne dă punctul de coordonate $(x_{n+1}, 0)$ adică $-f(x_n) = f'(x_n)(x_{n+1} - x_n)$ de unde $x_{n+1} = x_n - \dfrac{f(x_n)}{f'(x_n)}$. Astfel se obţine funcţia iterativă $\varphi : [a, b] \to \mathbb{R}$, $\varphi(x) = x - \dfrac{f(x)}{f'(x)}$ presupunând că $f'(x) \neq 0$ pentru orice $x \in [a, b]$. Această funcţie iterativă generează şirul iterativ $\{x_n\}_{n\in\mathbb{N}}$ dat de $x_{n+1} = \varphi(x_n)$.

Pentru a justifica existenţa şirului iterativ precum şi convergenţa lui către x^* dăm mai jos un rezultat teoretic al lui A.M. Ostrowski.

Teorema 5.4.1. *Fie $f \in C^2([a,b])$ şi $x_0 \in [a,b]$. Presupunem că următoarele condiţii sunt îndeplinite:*

1. $f'(x_0) \neq 0$,
2. *intervalul* $I_0 = [x_0, x_0 + 2h_0]$, *unde* $h_0 = -\dfrac{f(x_0)}{f'(x_0)}$, *este inclus în* $[a,b]$,
3. $2|h_0|M \leq |f'(x_0)|$, *unde* $M = \max\limits_{x \in I_0} |f''(x)|$.

Atunci ecuaţia $f(x) = 0$ are o soluţie unică $x^* \in I_0$ iar şirul $\{x_k\}_{k\in\mathbb{N}}$ dat de metoda lui Newton este bine definit şi converge la x^*. Eroarea aproximaţiei este dată de

$$|x^* - x_k| \leq \frac{2M}{|f'(x_{k-1})|}(x_k - x_{k-1})^2.$$

DEMONSTRAŢIE. Aplicăm formula creşterilor finite funcţiei f' în punctele x_0, x_1 : $|f'(x_1) - f'(x_0)| = |x_1 - x_0| \cdot |f''(\xi)|$ unde $\xi \in (x_0, x_1)$. Ţinând seama de condiţia 3)

$$|f'(x_1) - f'(x_0)| \leq |x_1 - x_0| \cdot M = \left| x_0 - \frac{f(x_0)}{f'(x_0)} - x_0 \right| \cdot M = \left| -\frac{f(x_0)}{f'(x_0)} \right| \cdot M = |h_0| \cdot M \leq \frac{|f'(x_0)|}{2}$$

deci folosind inegalitatea triunghiului avem:

$$|f'(x_1)| \geq |f'(x_0)| - |f'(x_1) - f'(x_0)| \geq |f'(x_0)| - \frac{|f'(x_0)|}{2} = \frac{|f'(x_0)|}{2}.$$

Se demonstrează că, condiţiile teoremei lui Ostrowski vor fi adevărate şi dacă în locul punctului x_0 considerăm punctul x_1.

Din condiţia 1) şi din $|f'(x_1)| \geq \dfrac{|f'(x_0)|}{2}$ rezultă că şi $f'(x_1) \neq 0$, deci x_2 este bine definit. Vom calcula în două moduri diferite integrala $I = \displaystyle\int_{x_0}^{x_1} (x_1 - x) f''(x)dx$. Mai întâi integrând prin părţi obţinem:

$$\begin{aligned} I &= (x_1 - x)f'(x)\Big|_{x_0}^{x_1} - \int_{x_0}^{x_1} (-1)f'(x)dx = -(x_1 - x_0)f'(x_0) + \int_{x_0}^{x_1} f'(x)dx = \\ &= -\left(x_0 - \frac{f(x_0)}{f'(x_0)} - x_0 \right) f'(x_0) + f(x)\Big|_{x_0}^{x_1} = \frac{f(x_0)}{f'(x_0)} f'(x_0) + f(x_1) - f(x_0) = f(x_1). \end{aligned}$$

Apoi efectuând schimbarea de variabilă $x = x_0 + h_0 t$, avem

$$\begin{aligned} I &= \int_{x_0}^{x_1} (x_1 - x) f''(x)dx = \int_0^1 (x_1 - x_0 - h_0 t) f''(x_0 + h_0 t) h_0 dt = \\ &= \int_0^1 (x_0 + h_0 - x_0 - h_0 t) f''(x_0 + h_0 t) h_0 dt = h_0^2 \int_0^1 (1 - t) f''(x_0 + h_0 t) dt. \end{aligned}$$

De aici rezultă

$$\begin{aligned} |I| &= |f(x_1)| = h_0^2 \left| \int_0^1 (1-t) f''(x_0 + h_0 t) dt \right| \leq h_0^2 \int_0^1 |1-t||f''(x_0+h_0t)|dt \leq \\ &\leq h_0^2 \int_0^1 |1-t| M dt = h_0^2 M \int_0^1 (1-t)dt = h_0^2 M \frac{1}{2}, \end{aligned}$$

adică $|f(x_1)| \leq \dfrac{h_0^2 M}{2}$. Deoarece $f'(x_1) \neq 0$ există $h_1 = \dfrac{-f(x_1)}{f'(x_1)}$ şi avem

$$|h_1| = \left| \frac{-f(x_1)}{f'(x_1)} \right| = \frac{|f(x_1)|}{|f'(x_1)|} \leq \frac{\frac{h_0^2 M}{2}}{\frac{|f'(x_0)|}{2}} = \frac{h_0^2 M}{|f'(x_0)|}.$$

Prin urmare

$$\frac{2|h_1|M}{|f'(x_1)|} = 2M|h_1| \frac{1}{|f'(x_1)|} \leq 2M \cdot \frac{h_0^2 M}{|f'(x_0)|} \cdot \frac{1}{\frac{|f'(x_0)|}{2}} = \left(\frac{2|h_0|M}{|f'(x_0)|} \right)^2.$$

De aici şi din condiţia 3) rezultă că $\dfrac{2|h_1|M}{|f'(x_1)|} \leq 1^2$, adică $2|h_1|M \leq |f'(x_1)|$ ceea ce este condiţia 3) pentru punctul x_1. Pentru a arăta condiţia 2) şi în cazul punctului x_1 folosim inegalitatea demonstrată: $|h_1| \leq \frac{h_0^2 M}{|f'(x_0)|} = \frac{|h_0|}{2} \cdot \frac{2|h_0|M}{|f'(x_0)|} \leq \frac{|h_0|}{2}$ deoarece conform condiţiei 3) $\dfrac{2|h_0|M}{|f'(x_0)|} \leq 1$.

Intervalul $I_1 = [x_1, x_1 + 2h_1]$, analog prin construcţie cu I_0, este inclus în I_0.

Într-adevăr $x_1 = x_0 + h_0$ este tocmai mijlocul intervalului $I_0 = [x_0, x_0 + 2h_0]$ şi condiţia $|h_1| \leq \dfrac{|h_0|}{2}$ ne asigură că $x_1 + 2h_1 \in I_0$.

Prin recurenţă, se obţine pentru orice $k \in \mathbb{N} : f'(x_k) \neq 0$, $|h_k| \leq \dfrac{1}{2}|h_{k-1}|$, $2|h_k|M \leq |f'(x_k)|$, $I_k \subset I_{k-1}$, unde $h_k = -\dfrac{f(x_k)}{f'(x_k)}$ şi $I_k = [x_k, x_k + 2h_k]$. De aici rezultă, în primul rând, că şirul $\{x_k\}_{k \in \mathbb{N}}$ este definit.

Vom demonstra că acest şir iterativ este şir Cauchy sau fundamental. Pentru orice $\varepsilon > 0$ există $k_{(\varepsilon)} \in \mathbb{N}$ astfel încât $2|h_k| \leq \varepsilon$ deoarece $|h_k| \leq \dfrac{1}{2}|h_{k-1}| \leq \dfrac{1}{2^2}|h_{k-2}| \leq \ldots \leq \dfrac{1}{2^k}|h_0| \to 0$. Pentru orice $p \in \mathbb{N}$, $x_{k+p} \in I_{k+p} \subset I_k$ deci $|x_{k+p} - x_k| \leq 2|h_k| \leq \varepsilon$, ceea ce trebuia arătat.

Fie $\lim\limits_{k \to \infty} x_k = x^*$, vom demonstra că x^* este tocmai soluţia ecuaţiei $f(x) = 0$. Notăm $M_1 = \sup\limits_{x \in I_0} |f'(x)|$ şi din $h_k = \dfrac{-f(x_k)}{f'(x_k)}$ rezultă $f(x_k) = -h_k f'(x_k)$, deci $|f(x_k)| = |h_k f'(x_k)| \leq |h_k| M_1 \to 0$. Prin urmare $\lim\limits_{k \to \infty} f(x_k) = 0$, adică $f\left(\lim\limits_{k \to \infty} x_k \right) = f(x^*) = 0$. Să

arătăm că x^* este singura soluţie a ecuaţiei $f(x) = 0$ în I_0. Pentru orice $x \in (x_0, x_0 + 2h_0)$ avem $|x - x_0| < 2|h_0|$. Din teorema creşterilor finite şi din condiţia 3) obţinem

$$|f'(x) - f'(x_0)| = |x - x_0| \cdot |f''(\xi)| \leq |x - x_0|M < 2|h_0|M \leq |f'(x_0)|,$$

adică $f'(x) \neq 0$ pentru orice $x \in (x_0, x_0 + 2h_0)$. Prin urmare $f'(x) > 0$ sau $f'(x) < 0$ pentru orice $x \in (x_0, x_0 + 2h_0)$, adică f este strict monotonă pe intervalul $(x_0, x_0 + 2h_0)$, deci x^* este soluţie unică.

Eroarea aproximaţiei se obţine din inegalitatea $|h_k| \leq \dfrac{Mh_{k-1}^2}{|f'(x_{k-1})|}$ analoagă cu inegalitatea demonstrată $|h_1| \leq \dfrac{Mh_0^2}{|f'(x_0)|}$.

Ţinând seama că $|x_{k+p} - x_k| \leq 2|h_k|$ precum şi de faptul că $h_{k-1} = x_k - x_{k-1}$, avem

$$|x_{k+p} - x_k| \leq \frac{2M}{|f'(x_{k-1})|}(x_k - x_{k-1})^2.$$

Trecând la limită pentru $p \to \infty$, se obţine

$$|x^* - x_k| \leq \frac{2M}{|f'(x_{k-1})|}(x_k - x_{k-1})^2,$$

adică relaţia din enunţul teoremei q.e.d.

În continuare, dacă se presupune în plus că există $M_2 = \inf\limits_{x \in I_0} |f'(x)| > 0$, atunci se obţine următoarea condiţie de oprire la iterarea şirului $\{x_k\}_{k \in \mathbb{N}}$: $|x^* - x_k| \leq \dfrac{2M}{M_2}(x_k - x_{k-1})^2 \leq \varepsilon$ unde ε este precizia dinainte dată. Prin urmare $|x_k - x_{k-1}| \leq \sqrt{\dfrac{M_2}{2M}}\varepsilon$.

Avantajul metodei tangentei este că e o metodă rapid convergentă. Pentru a studia acest aspect matematic avem nevoie de

Definiţia 5.4.1. *Spunem că şirul iterativ $\{x_k\}_{k \in \mathbb{N}}$ convergent către x^*,* ***are ordinul de convergenţă*** *$p \in \mathbb{R}$, $p \geq 1$ dacă*

$$\lim_{k \to \infty} \frac{\Delta(x_{k+1})}{\Delta^p(x_k)} = \lim_{k \to \infty} \frac{|x_{k+1} - x^*|}{|x_k - x^*|^p} = c,$$

$c \in \mathbb{R}$ fiind o constantă.

Menţionăm că nu orice şir convergent admite ordin de convergenţă. De exemplu şirul $\{x_k\}_{k \in \mathbb{N}}$, $x_k = \dfrac{(-1)^k + 1}{k} \to x^* = 0$, dar nu are sens expresia $\dfrac{|x_{k+1} - x^*|}{|x_k - x^*|^p}$ pentru k număr natural impar. Totodată, dacă există ordinul de convergenţă, atunci acesta este unic.

Dacă $p = 1$ atunci spunem că avem o convergenţă liniară, iar dacă $p = 2$ avem o convergenţă pătratică. Cu cât p este mai mare, cu atât şirul iterativ va fi mai rapid convergent.

Teorema 5.4.2. *Şirul iterativ obţinut prin metoda tangentei are ordinul de convergenţă pătratică.*

DEMONSTRAŢIE. Conform formulei lui Taylor în punctul x_n avem $0 = f(x^*) = f(x_n) + \frac{f'(x_n)}{1!}(x^* - x_n) + \frac{f''(\xi_n)}{2!}(x^* - x_n)^2$ unde x^* este soluţia ecuaţiei $f(x) = 0$ şi $\xi_n \in (x_n, x^*)$ sau $\xi_n \in (x^*, x_n)$. Împărţind egalitatea la $f'(x_n)$ se obţine

$$\frac{f(x_n)}{f'(x_n)} + (x^* - x_n) = -\frac{f''(\xi_n)}{f'(x_n) \cdot 2}(x^* - x_n)^2,$$

deci

$$x^* - x_{n+1} = -\frac{1}{2}\frac{f''(\xi_n)}{f'(x_n)}(x^* - x_n)^2.$$

În această egalitate trecem la modul

$$|x^* - x_{n+1}| = \frac{1}{2}\frac{|f''(\xi_n)|}{|f'(x_n)|}(x^* - x_n)^2,$$

adică

$$\Delta(x_{n+1}) = \frac{1}{2}\frac{|f''(\xi_n)|}{|f'(x_n)|} \cdot \Delta^2(x_n).$$

Prin urmare

$$\frac{\Delta(x_{n+1})}{\Delta^2(x_n)} = \frac{1}{2}\frac{|f''(\xi_n)|}{|f'(x_n)|}.$$

Trecând la limită pentru $n \to \infty$ se obţine:

$$\lim_{n\to\infty} \frac{\Delta(x_{n+1})}{\Delta^2(x_n)} = \frac{1}{2}\frac{|f''(x^*)|}{|f'(x^*)|} = c,$$

deoarece $x_n \to x^*$ şi $\xi_n \in (x_n, x^*)$ sau $\xi_n \in (x^*, x_n)$ rezultă că şi $\lim\limits_{n\to\infty} \xi_n = x^*$ q.e.d.

Folosind rezultatele din teorema precedentă se obţine şi un alt criteriu de oprire. Dacă $M = \sup\limits_{x\in[a,b]} |f''(x)|$ şi $M_2 = \inf\limits_{x\in[a,b]} |f'(x)|$, $M_2 \neq 0$, atunci $\dfrac{\Delta(x_{n+1})}{\Delta^2(x_n)} \leq \dfrac{1}{2} \cdot \dfrac{M}{M_2} = C$ şi $C\Delta(x_{n+1}) \leq [C\Delta(x_n)]^2$. Prin intermediul acestei inegalităţi se obţine că $C\Delta(x_{n+1}) \leq [C\Delta(x_0)]^{2^{n+1}}$. Pentru a asigura ca $x_{n+1} \to x^*$ trebuie ca $\Delta(x_{n+1}) \to 0$ ceea ce putem asigura prin condiţia suficientă $|C\Delta(x_0)| < 1$. De aici se observă că $\Delta(x_0)$ trebuie să fie suficient de mic, adică punctul de plecare x_0 să fie suficient de aproape de soluţia ideală x^*. Acest fapt constituie un dezavantaj al acestei metode, fiind o metodă locală.

Un alt dezavantaj al metodei tangentei este că necesită calculul derivatei funcţiei f la fiecare pas de iteraţie. Pentru calculul aproximativ al derivatei se pot folosi formulele de derivare numerică (vezi paragraful 9.1). În unele cazuri ca să nu calculăm derivata funcţiei

f la fiecare pas al iteraţiei se foloseşte metoda tangentei simple, unde se calculează numai $f'(x_0)$ şi avem iteraţia $x_{k+1} = x_k - \dfrac{f(x_k)}{f'(x_0)}$ pentru orice $k \in \mathbb{N}$. Însă astfel se pierde convergenţa pătratică.

Înainte de algoritmizarea metodei avem nevoie de alegerea punctului de pornire $x_0 \in \{a, b\}$. Presupunem că f este de clasă $C^2([a, b])$ şi că f' şi f'' păstrează semnul pe tot intervalul $[a, b]$, adică f este monotonă pe $[a, b]$ şi este fie convexă, fie concavă pe tot $[a, b]$. În figura 5.2 am considerat cazul $f' > 0$ şi $f'' > 0$ când se alege $x_0 = b$.

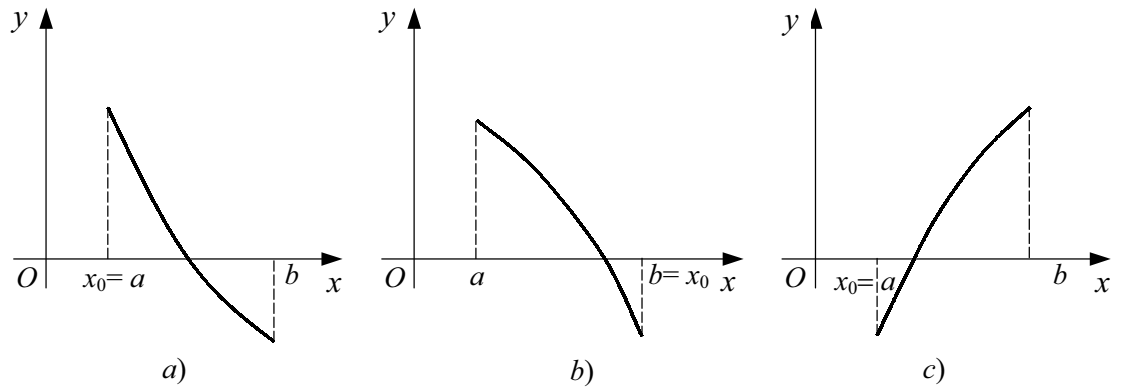


Figura 5.3:

În fig. 5.3 a) $f' < 0$ şi $f'' > 0$ se alege $x_0 = a$, în fig. 5.3 b) $f' < 0$, $f'' < 0$ se alege $x_0 = b$ iar în fig. 5.3 c) $f' > 0$, $f'' < 0$ se alege $x_0 = a$.

În final se observă că punctul de plecare x_0 verifică următoarea condiţie $f(x_0) \cdot f''(x_0) > 0$. Derivata a doua a funcţiei f în x_0 se poate calcula aproximativ folosind formulele de derivare numerică (vezi paragraful 9.1).

Program metoda tangentei

Datele de intrare: $a; b; f; \varepsilon$;

Fie $y = a$; dacă $f(y) * f''(y) < 0$ atunci $y := b$.

Repetă $x := y$; $y := \varphi(x)$ $\left(\varphi(x) = x - \dfrac{f(x)}{f'(x)}\right)$

Până când $|y - x| \geq \varepsilon$.

Tipăreşte y.

5.4.2 Metoda paralelelor

Ideea metodei constă în a construi o familie de drepte paralele ale căror intersecţii cu axa Ox ne dau şirul iterativ $\{x_k\}_{k\in\mathbb{N}}$ care converge către rădăcina x^* a ecuaţiei $f(x) = 0$.

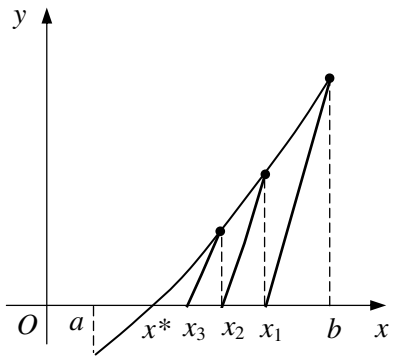


Figura 5.4:

Punctul de plecare $x_0 \in \{a, b\}$ se alege identic ca la metoda tangentei, adică se alege acel capăt al intervalului pentru care $f(x_0) \cdot f''(x_0) > 0$. Presupunem că am construit punctul $M_k(x_k, f(x_k))$. Prin acest punct ducem o dreaptă paralelă care are panta $\lambda > f'(x_0)$ în cazul figurii 5.4, şi taie axa Ox în punctul x_{k+1}. Ecuaţia dreptei este $\dfrac{y - f(x_k)}{x - x_k} = \lambda = \text{tg}\alpha$ (înclinaţia dreptei faţă de axa Ox).

Punând $y = 0$ şi $x = x_{k+1}$ obţinem $\dfrac{-f(x_k)}{x_{k+1} - x_k} = \lambda$, adică $x_{k+1} = x_k - \dfrac{1}{\lambda} f(x_k)$. Funcţia iterativă $\varphi : [a, b] \to \mathbb{R}$ este dată de $\varphi(x) = x - \dfrac{1}{\lambda} f(x)$. Metoda paralelelor are avantajul faţă de metoda tangentei că nu necesită calculul derivatei, însă este o metodă mult mai lent convergentă, având ordinea de convergenţă cel puţin liniară. Algoritmul acestei metode este analog cu algoritmul metodei tangentei, numai funcţia iterativă este schimbată.

Program metoda paralelelor

Datele de intrare $a; b; f; \varepsilon; \lambda$;
Fie $y := a$; dacă $f(y) * f''(y) < 0$, atunci $y := b$;
Repetă $x := y$; $y := \varphi(x)$ $\left(\varphi(x) = x - \dfrac{1}{\lambda} f(x)\right)$;
Până când $|y - x| \geq \varepsilon$;
Tipăreşte y.

5.4.3 Metoda coardei

La metoda coardei unul dintre capetele intervalului $[a, b]$ se fixează conform condiţiei $f(x) \cdot f''(x) > 0$.

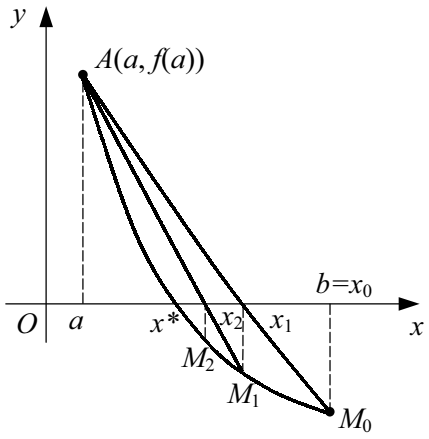


Figura 5.5:

În cazul figurii 5.5 avem $x = a$ şi ca punct de plecare pentru construcţia şirului iterativ se alege celălalt capăt $x_0 = b$. Punctul $A(a, f(a))$ se uneşte printr-o coardă cu punctul $M_0(x_0, f(x_0)) = M_0(b, f(b))$ care taie axa Ox în punctul x_1. Mai departe punctul $M_1(x_1, f(x_1))$ se uneşte printr-o coardă tot cu punctul $A(a, f(a))$ care intersectează axa Ox în punctul x_2 ş.a.m.d. Presupunem că s-a construit termenul x_k şi se scrie ecuaţia coardei care leagă punctele $A(a, f(a))$ şi $M_k(x_k, f(x_k))$:

$$\frac{y - f(a)}{x - a} = \frac{f(x_k) - f(a)}{x_k - a}.$$

Pentru a obţine intersecţia cu axa Ox punem $y = 0$ şi $x = x_{k+1}$: $\dfrac{-f(a)}{x_{k+1} - a} = \dfrac{f(x_k) - f(a)}{x_k - a}$, de unde $x_{k+1} = a - f(a)\dfrac{x_k - a}{f(x_k) - f(a)}$ sau $x_{k+1} = \dfrac{af(x_k) - x_k f(a)}{f(x_k) - f(a)}$ sau $x_{k+1} = x_k - f(x_k)\frac{x_k - a}{f(x_k) - f(a)}$. Prin urmare funcţia iterativă $\varphi : [a, b] \to \mathbb{R}$ este $\varphi(x) = \dfrac{af(x) - xf(a)}{f(x) - f(a)}$. Avantajul metodei faţă de metoda tangentei este că nu necesită calculul derivatei funcţiei f, însă este mai lent convergentă având ordinul de convergenţă $p = \dfrac{1 + \sqrt{5}}{2} < 2$.

Algoritmul acestei metode este analog cu cel al metodei tangentei, numai se schimbă funcţia iterativă φ.

Program metoda coardei

Datele de intrare $a; b; f; \varepsilon$;

Fie $c := a$; $y := b$;

Dacă $f(c) * f''(c) < 0$ atunci $c := b$; $y := a$

Repetă $x := y$; $y := \varphi(x)$ $\left(\varphi(x) = \frac{af(x)-xf(a)}{f(x)-f(a)}\right)$;

Până când $|y - x| \geq \varepsilon$;

Tipăreşte y.

5.4.4 Metoda secantei

Ideea metodei constă în următoarele: se alege punctul de plecare $x_0 \in \{a, b\}$ conform condiţiei $f(x_0)f''(x_0) > 0$. În cazul fig. 5.6 avem $x_0 = b$.

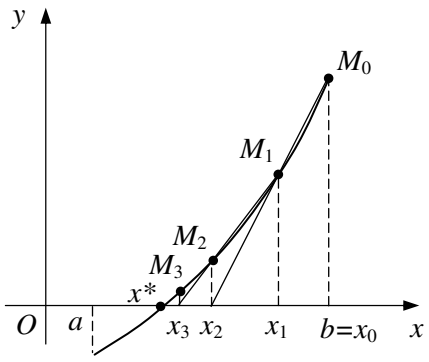


Figura 5.6:

Punctul x_1 se alege între x^* şi x_0. Prin punctele $M_0(x_0, f(x_0)) = M_0(b, f(b))$ şi $M_1(x_1, f(x_1))$ se duce o secantă care intersectează axa Ox în punctul x_2. Prin punctele M_1 şi $M_2(x_2, f(x_2))$ se duce o nouă secantă care taie axa Ox în punctul x_3, ş.a.m.d. Presupunem că s-a obţinut punctul x_k şi prin punctele $M_{k-1}(x_{k-1}, f(x_{k-1}))$ şi $M_k(x_k, f(x_k))$ se duce o secantă:

$$\frac{y - f(x_k)}{x - x_k} = \frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}}$$

care determină pe axa Ox $(y = 0)$ punctul $x = x_{k+1}$:

$$\frac{-f(x_k)}{x_{k+1} - x_k} = \frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}},$$

de unde

$$x_{k+1} = x_k - f(x_k)\frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})} \text{ sau } x_{k+1} = \frac{f(x_k)x_{k-1} - f(x_{k-1})x_k}{f(x_k) - f(x_{k-1})}.$$

Această relaţie de recurenţă se poate obţine şi din metoda tangentei $x_{k+1} = x_k - \dfrac{f(x_k)}{f'(x_k)}$ prin aproximarea $f'(x_k) \approx \dfrac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}}$. Funcţia iterativă $\varphi : [a,b]^2 \to \mathbb{R}$ este de forma $\varphi(x,y) = \dfrac{f(x)y - f(y)x}{f(x) - f(y)}$ care generează şirul iterativ $\{x_k\}_{k\in\mathbb{N}}$ prin relaţia iterativă $x_{k+1} = \varphi(x_k, x_{k-1})$.

Metoda secantei are avantajul faţă de metoda tangentei, că nu necesită calculul derivatei funcţiei f, însă are dezavantajul, că are ordinul de convergenţă $p = \dfrac{1+\sqrt{5}}{2} < 2$, deci este mai lent convergentă.

Program metoda secantei

Datele de intrare: $a; b; f; \varepsilon$;
Fie $x := x_0$; $z := x_1$;
Repetă $y := x$; $x := z$; $z := \varphi(x,y)$ $\left(\varphi(x,y) = \dfrac{f(x)y - f(y)x}{f(x) - f(y)}\right)$
Până când $|z - x| \geq \varepsilon$;
Tipăreşte z.

5.4.5 Metoda lui Steffensen

La această metodă plecăm de la relaţia de recurenţă a lui Newton $x_{k+1} = x_k - \dfrac{f(x_k)}{f'(x_k)}$. Ideea lui Steffensen este de a înlocui $f'(x_k)$ prin formula aproximativă $\dfrac{f(x_k + f(x_k)) - f(x_k)}{f(x_k)}$, căci se poate presupune că $|f(x_k)|$ are valoare mică. Dacă $\beta_k = f(x_k) \approx 0$ atunci $f'(x_k) \approx \dfrac{f(x_k + \beta_k) - f(x_k)}{\beta_k}$. Metoda lui Steffensen generează şirul iterativ $\{x_k\}_{k\in\mathbb{N}}$, dat prin relaţia

$$x_{k+1} = x_k - \frac{f(x_k)}{\frac{f(x_k + f(x_k)) - f(x_k)}{f(x_k)}} = x_k - \frac{f^2(x_k)}{f(x_k + f(x_k)) - f(x_k)}.$$

Funcţia recursivă $\varphi : [a,b] \to \mathbb{R}$ este $\varphi(x) = x - \dfrac{f^2(x)}{f(x + f(x)) - f(x)}$. Avantajul acestei metode faţă de cea a tangentei este că nu necesită calculul derivatei funcţiei f şi păstrează ordinul de convergenţă pătratică de la metoda lui Newton.

Teorema 5.4.3. *Metoda lui Steffensen are ordinul de convergenţă pătratică.*

DEMONSTRAŢIE. Fie $\beta_k = f(x_k)$. Se dezvoltă f în serie Taylor în punctul x_k :

$$f(x_k + \beta_k) = f(x_k) + \beta_k f'(x_k) + \frac{\beta_k^2}{2} f''(x_k) + o(\beta_k^2)$$

unde $o(\beta_k^2)$ este restul seriei, care are ordinul strict mai mic decât β_k^2, mai precis înseamnă că $\lim\limits_{\beta_k \to 0} \dfrac{o(\beta_k^2)}{\beta_k^2} = 0$. Din dezvoltarea taylorianã se obţine

$$\frac{f(x_k + \beta_k) - f(x_k)}{\beta_k} = f'(x_k) \left[1 + \frac{1}{2}\beta_k \frac{f''(x_k)}{f'(x_k)}\right] + o(\beta_k).$$

Notând cu $h_k = \dfrac{-f(x_k)}{f'(x_k)}$ vom avea

$$\frac{f(x_k + \beta_k) - f(x_k)}{\beta_k} = f'(x_k) \left[1 - \frac{1}{2} h_k f''(x_k)\right] + o(\beta_k).$$

Conform recurenţei lui Steffensen $x_{k+1} = x_k - \dfrac{f(x_k)}{\frac{f(x_k+\beta_k)-f(x_k)}{\beta_k}}$ şi înlocuind egalitatea obţinutã anterior în această recurenţă se obţine

$$\begin{aligned} x_{k+1} &= x_k - \frac{f(x_k)}{f'(x_k)\left[1 - \frac{1}{2}h_k f''(x_k)\right] + o(\beta_k)} \approx x_k - \frac{f(x_k)}{f'(x_k)} \cdot \frac{1}{1 - \frac{1}{2}h_k f''(x_k)} = \\ &= x_k + h_k \left[1 - \frac{1}{2} h_k f''(x_k)\right]^{-1}. \end{aligned}$$

În continuare considerăm dezvoltarea Mac-Laurin a funcţiei $\frac{1}{1+x}$ pentru $|x| < 1$:

$$(1+x)^{-1} = 1 - \frac{1}{1!}x + \frac{1}{2!}2x^2 + \ldots = 1 - x + x^2 - x^3 + \ldots$$

Înlocuind aici x cu $-x$ se obţine dezvoltarea funcţiei $(1-x)^{-1}$ în origine: $(1-x)^{-1} = 1 + x + x^2 + x^3 + \ldots$ Înlocuind acum $x = \dfrac{1}{2} h_k f''(x_k)$ se poate continua dezvoltarea

$$\begin{aligned} x_{k+1} &= x_k + h_k \left[1 - \frac{1}{2} h_k f''(x_k)\right]^{-1} = x_k + h_k \left[1 + \frac{1}{2} h_k f''(x_k) + o(h_k^2)\right] \approx \\ &\approx x_k + h_k + \frac{1}{2} h_k^2 f''(x_k). \end{aligned}$$

Scădem valoarea x^* din ambele părţi ale egalităţii:

$$x_{k+1} - x^* = x_k - x^* + h_k + \frac{1}{2} h_k^2 f''(x_k).$$

Introducem notaţia $\varepsilon_k = x_k - x^*$ şi astfel egalitatea anterioară ia forma $\varepsilon_{k+1} = \varepsilon_k + h_k + \frac{1}{2}h_k^2 f''(x_k)$. În continuare funcţia f se dezvoltă în serie Taylor în punctul x_k apropiat de x^*, soluţia ecuaţiei $f(x) = 0$:

$$0 = f(x^*) = f(x_k) + \frac{f'(x_k)}{1!}(x^* - x_k) + \frac{f''(\xi_k)}{2!}(x^* - x_k)^2$$

unde $\xi_k \in (x^*, x_k)$ sau $\xi_k \in (x_k, x^*)$. Această dezvoltare se împarte la $f'(x_k)$:

$$\frac{f(x_k)}{f'(x_k)} + x^* - x_k = -\frac{1}{2}\frac{(x^* - x_k)^2}{f'(x_k)} f''(\xi_k),$$

adică $-h_k - \varepsilon_k = -\frac{1}{2}\frac{\varepsilon_k^2}{f'(x_k)} f''(\xi_k)$, de unde $h_k = -\varepsilon_k + \frac{1}{2}\frac{\varepsilon_k^2}{f'(x_k)} f''(\xi_k)$. Ultima relaţie obţinută se înlocuieşte în egalitatea

$$\begin{aligned}
\varepsilon_{k+1} &= \varepsilon_k + h_k + \frac{1}{2}h_k^2 f''(x_k) = \\
&= \varepsilon_k - \varepsilon_k + \frac{1}{2}\frac{\varepsilon_k^2}{f'(x_k)} f''(\xi_k) + \frac{1}{2}\left(-\varepsilon_k + \frac{1}{2}\frac{\varepsilon_k^2}{f'(x_k)} f''(\xi_k)\right)^2 f''(x_k) = \\
&= \frac{1}{2}\frac{f''(\xi_k)}{f'(x_k)}\varepsilon_k^2 + \frac{1}{2}f''(x_k)\varepsilon_k^2 + o(\varepsilon_k^2).
\end{aligned}$$

Trecând la module

$$|\varepsilon_{k+1}| \approx \left|\frac{1}{2}\frac{f''(\xi_k)}{f'(x_k)}\varepsilon_k^2 + \frac{1}{2}f''(x_k)\varepsilon_k^2\right|$$

şi cum $\Delta(x_k) = |\varepsilon_k|$ avem

$$\Delta(x_{k+1}) = \frac{1}{2}\left|\frac{f''(\xi_k)}{f'(x_k)} + f''(x_k)\right| \Delta^2(x_k).$$

Trecând la limită pentru $k \to \infty$

$$\lim_{k\to\infty} \frac{\Delta(x_{k+1})}{\Delta^2(x_k)} = \lim_{k\to\infty} \frac{1}{2}\left|\frac{f''(\xi_k)}{f'(x_k)} + f''(x_k)\right| = \frac{1}{2}\left|\frac{f''(x^*)}{f'(x^*)} + f''(x^*)\right| = c \quad \text{q.e.d.}$$

Dezavantajul metodei lui Steffensen este că e o metodă locală ca şi metoda lui Newton.

Program metoda Steffensen

Datele de intrare: $a; b; f; \varepsilon$;
Fie $y := a$; dacă $f(y) * f''(y) < 0$ atunci $y := b$;
Repetă $x := y$; $y := \varphi(x)$ $\left(\varphi(x) = x - \dfrac{[f(x)]^2}{f(x + f(x)) - f(x)}\right)$
Până când $|y - x| \geq \varepsilon$;
Tipăreşte y.

5.4.6 Teoria generală a metodelor iterative în cazul ecuaţiilor neliniare

Fie funcţia $f : [a, b] \to \mathbb{R}$ şi ecuaţia $f(x) = 0$, $x \in [a, b]$. Se presupune că funcţia f admite o unică soluţie în $[a, b]$ şi ecuaţia $f(x) = 0$ prin tranformări echivalente se poate aduce la forma iterativă $\varphi(x) = x$, $x \in [a, b]$ unde $\varphi : [a, b] \to \mathbb{R}$ este funcţia iterativă. Orice soluţie x^* a ecuaţiei $f(x) = 0$ va fi punct fix pentru φ şi invers. Prin urmare pentru a asigura existenţa şi unicitatea soluţiei ecuaţiei $f(x) = 0$ este de ajuns să asigurăm existenţa şi unicitatea punctului fix al lui φ.

În continuare avem nevoie de teorema de punct fix a lui Banach.

Teorema 5.4.4. *Fie (X, ρ) un spaţiu metric complet şi $\varphi : X \to X$ o contracţie, adică pentru care există constanta $\alpha \in [0, 1)$ astfel încât $\rho(\varphi(x), \varphi(y)) \leq \alpha \cdot \rho(x, y)$ pentru orice $x, y \in X$. Atunci funcţia φ va admite un unic punct fix care se poate obţine ca limita şirului iterativ $\{x_k\}_{k\in\mathbb{N}}$ dat de $x_{k+1} = \varphi(x_k)$ pentru $x_0 \in X$ arbitrar.*

Avem următoarele evaluări apriori: $\rho(x^*, x_k) \leq \dfrac{\alpha^k}{1-\alpha}\rho(x_1, x_0)$ respectiv aposteriori $\rho(x^*, x_k) \leq \dfrac{\alpha}{1-\alpha}\rho(x_k, x_{k-1})$.

În continuare enunţăm o teoremă cu condiţii suficiente pentru a asigura soluţia ecuaţiei iterative în cazul ecuaţiilor neliniare cu o singură necunoscută.

Teorema 5.4.5. *(teoremă generală) Dacă φ este derivabilă pe intervalul $J = [x_0 - \delta, x_0 + \delta]$, $\delta > 0$ şi derivata φ' satisface inegalitatea $0 \leq |\varphi'(x)| \leq m < 1$ pe J şi punctul $x_1 = \varphi(x_0)$ verifică inegalitatea $|x_1 - x_0| \leq (1 - m)\delta$, atunci:*

- *putem forma şirul $\{x_k\}_{k\in\mathbb{N}}$ cu regula iterativă $x_{k+1} = \varphi(x_k)$ astfel încât pentru orice $k \in \mathbb{N}$ avem $x_k \in J$;*

- *$\lim\limits_{k\to\infty} x_k = x^*$ cu $x^* \in J$;*

- *x^* este singura rădăcină a ecuaţiei $\varphi(x) = x$ în intervalul J.*

Demonstraţie. Se aplică teorema de punct fix a lui Banach în acest caz. Se alege $X = J$ şi cum J este un interval închis al axei reale se poate considera ca un spaţiu complet în raport cu distanţa obişnuită de pe axa reală. Mai trebuie să arătăm că pentru

orice $x \in J$ avem $\varphi(x) \in J$. Într-adevăr:

$$\begin{aligned} |\varphi(x) - x_0| &= |\varphi(x) - \varphi(x_0) + x_1 - x_0| \leq |\varphi(x) - \varphi(x_0)| + |x_1 - x_0| = \\ &= |\varphi'(\xi)| \cdot |x - x_0| + |x_1 - x_0| \leq m \cdot |x - x_0| + (1 - m) \cdot \delta \leq \\ &\leq m \cdot \delta + (1 - m)\delta = \delta. \end{aligned}$$

Funcţia φ este o contracţie căci $|\varphi(x) - \varphi(y)| = |\varphi'(\xi)| \cdot |x - y| \leq m \cdot |x - y|$ cu $m < 1$. Prin urmare sunt satisfăcute toate cerinţele teoremei lui Banach, de unde rezultă consecinţele acestei teoreme. q.e.d.

În continuare aplicăm această teoremă pentru metodele iterative anterioare alegând în mod convenabil funcţia iterativă φ.

Aplicaţia 5.4.1. *Plecând de la ecuaţia $f(x) = 0$ prin translaţie se obţine $x + f(x) = x$ şi prin alegerea funcţiei iterative $\varphi : [a, b] \to \mathbb{R}$, $\varphi(x) = x + f(x)$ se obţine următorul rezultat: dacă f este derivabilă pe $J = [x_0 - \delta, x_0 + \delta]$, $\delta > 0$ şi $|1 + f'(x)| \leq m < 1$ pe J şi $|f(x_0)| \leq (1 - m) \cdot \delta$ atunci putem forma şirul $\{x_k\}_{k \in \mathbb{N}}$ cu regula iterativă $x_{k+1} = \varphi(x_k)$ astfel încât pentru orice $k \in \mathbb{N}$ avem $x_k \in J$ şi $\lim_{k \to \infty} x_k = x^*$ cu $x^* \in J$, x^* este singura rădăcină în J a ecuaţiei $f(x) = 0$.*

Aplicaţia 5.4.2. *Plecând de la ecuaţia $f(x) = 0$ prin omotopie simplă se obţine $x + \omega f(x) = x$ cu $\omega \neq 0$ şi prin alegerea funcţiei iterative $\varphi : [a, b] \to \mathbb{R}$, $\varphi(x) = x + \omega f(x)$ se obţine următorul rezultat: dacă f este derivabilă pe $J = [x_0 - \delta, x_0 + \delta]$, $\delta > 0$ şi $|1 + \omega f'(x)| \leq m < 1$ pe J şi $|\omega f(x_0)| \leq (1 - m) \cdot \delta$ atunci putem forma şirul $\{x_k\}_{k \in \mathbb{N}}$ cu regula iterativă $x_{k+1} = \varphi(x_k)$ astfel încât pentru orice $k \in \mathbb{N}$ avem $x_k \in J$ şi $\lim_{k \to \infty} x_k = x^*$ cu $x^* \in J$, x^* este singura rădăcină în J a ecuaţiei $f(x) = 0$.*

Aplicaţia 5.4.3. *Plecând de la ecuaţia $f(x) = 0$ prin transformarea lui Newton se obţine $x - \frac{f(x)}{f'(x)} = x$ şi prin alegerea funcţiei iterative $\varphi : [a, b] \to \mathbb{R}$, $\varphi(x) = x - \frac{f(x)}{f'(x)}$ se obţine următorul rezultat pentru metoda tangentei: dacă f este derivabilă de două ori pe $J = [x_0 - \delta, x_0 + \delta]$, $\delta > 0$ şi $f'(x) \neq 0$ pentru orice $x \in J$ şi $\frac{|f(x) \cdot f''(x)|}{[f'(x)]^2} \leq m < 1$ pe J şi $\left|\frac{f(x_0)}{f'(x_0)}\right| \leq (1 - m)\delta$ atunci putem forma şirul $\{x_k\}_{k \in \mathbb{N}}$ cu regula iterativă $x_{k+1} = \varphi(x_k)$ astfel încât pentru orice $k \in \mathbb{N}$ avem $x_k \in J$ şi $\lim_{k \to \infty} x_k = x^*$ cu $x^* \in J$, x^* este singura rădăcină în J a ecuaţiei $f(x) = 0$.*

Aplicaţia 5.4.4. *Plecând de la ecuaţia $f(x) = 0$ prin transformarea metodei paralelelor se obţine $x - \frac{1}{\lambda} \cdot f(x) = x$, $\lambda \neq 0$. Prin alegerea funcţiei iterative $\varphi : [a, b] \to \mathbb{R}$, $\varphi(x) =$*

$x - \frac{1}{\lambda} f(x)$ *se obţine următorul rezultat pentru metoda paralelelor: dacă* f *este derivabilă pe* $J = [x_0 - \delta, x_0 + \delta]$, $\delta > 0$ *şi* $\left|1 - \frac{1}{\lambda} \cdot f'(x)\right| \leq m < 1$ *pe* J *şi* $|f(x_0)| \leq (1-m) \cdot \delta \cdot |\lambda|$ *atunci putem forma şirul* $\{x_k\}_{k \in \mathbb{N}}$ *cu regula iterativă* $x_{k+1} = \varphi(x_k)$ *astfel încât pentru orice* $k \in \mathbb{N}$ *avem* $x_k \in J$ *şi* $\lim_{k \to \infty} x_k = x^*$ *cu* $x^* \in J$, x^* *este singura rădăcină în* J *a ecuaţiei* $f(x) = 0$.

Aplicaţia 5.4.5. *Plecând de la ecuaţia* $f(x) = 0$ *prin transformarea metodei coardei se obţine* $x - f(x)\frac{x-a}{f(x)-f(a)} = x$. *Prin alegerea funcţiei iterative* $\varphi : (a, b] \to \mathbb{R}$, $\varphi(x) = x - f(x) \cdot \frac{x-a}{f(x)-f(a)}$ *se obţine următorul rezultat pentru metoda coardei: dacă* f *este derivabilă pe* $J = [x_0 - \delta, x_0 + \delta]$, $\delta > 0$ *şi*

$$\frac{|f(a)| \ |f(a) - f(x) + f'(x) \cdot (x-a)|}{[f(x) - f(a)]^2} \leq m < 1$$

pe J *şi*

$$\left|\frac{(x_0 - a) f(x_0)}{f(x_0) - f(a)}\right| \leq (1-m) \cdot \delta$$

atunci putem forma şirul $\{x_k\}_{k \in \mathbb{N}}$ *cu regula iterativă* $x_{k+1} = \varphi(x_k)$ *astfel încât pentru orice* $k \in \mathbb{N}$ *avem* $x_k \in J$ *şi* $\lim_{k \to \infty} x_k = x^*$ *cu* $x^* \in J$, x^* *este singura rădăcină în* J *a ecuaţiei* $f(x) = 0$.

Aplicaţia 5.4.6. *Plecând de la ecuaţia* $f(x) = 0$ *prin transformarea lui Steffensen se obţine* $x - \frac{f^2(x)}{f(x+f(x))-f(x)}$. *Prin alegerea funcţiei iterative* $\varphi : [a, b] \to \mathbb{R}$, $\varphi(x) = x - \frac{f^2(x)}{f(x+f(x))-f(x)}$ *se obţine următorul rezultat: dacă* f *este derivabilă pe* $J = [x_0 - \delta, x_0 + \delta]$, $\delta > 0$ *şi*

$$\begin{aligned} & \left|1 - \frac{2f(x) \cdot f'(x)[f(x + f(x)) - f(x)] - f^2(x) \cdot [f'(x + f(x)) \cdot (1 + f'(x)) - f'(x)]}{[f(x + f(x)) - f(x)]^2}\right| = \\ = \ & \left|1 - \frac{f^2(x) \cdot f'(x) + 2f(x) \cdot f'(x) \cdot f(x + f(x)) - f^2(x) \cdot f'(x + f(x)) \cdot (1 + f'(x))}{[f(x + f(x)) - f(x)]^2}\right| \leq \\ \leq \ & m < 1 \end{aligned}$$

pe J *şi* $\left|\frac{f^2(x_0)}{f(x_0+f(x_0))-f(x_0)}\right| \leq (1-m) \cdot \delta$, *atunci putem forma şirul* $\{x_k\}_{k \in \mathbb{N}}$ *cu regula iterativă* $x_{k+1} = \varphi(x_k)$ *astfel încât pentru orice* $k \in \mathbb{N}$ *avem* $x_k \in J$ *şi* $\lim_{k \to \infty} x_k = x^*$ *cu* $x^* \in J$, x^* *este singura rădăcină în* J *a ecuaţiei* $f(x) = 0$.

Aplicaţia 5.4.7. *Plecând de la ecuaţia* $f(x) = x^2 - p = 0$ *prin transformări echivalente se obţine* $\frac{1}{2}\left(x + \frac{p}{x}\right) = x$ *unde* $p \in [0, +\infty)$. *Prin alegerea funcţiei iterative* $\varphi : [a, b] \to \mathbb{R}$, $\varphi(x) = \frac{1}{2}\left(x + \frac{p}{x}\right)$ *se obţine următorul rezultat pentru calcul aproximativ al lui* $\sqrt{p}$: *dacă*

$\left|1-\frac{p}{x^2}\right| \leq 2m < 2$ *pe* $J = [x_0 - \delta, x_0 + \delta]$, $\delta > 0$ *şi* $\left|x_0 - \frac{p}{x_0}\right| \leq 2(1-m)\cdot\delta$ *atunci putem forma şirul* $\{x_k\}_{k\in\mathbb{N}}$ *cu regula iterativă* $x_{k+1} = \varphi(x_k)$ *astfel încât pentru orice* $k \in \mathbb{N}$ *avem* $x_k \in J$ *şi* $\lim_{k\to\infty} x_k = x^* = \sqrt{p}$ *cu* $x^* \in J$, x^* *este singura rădăcină în* J *a ecuaţiei* $f(x) = 0$. *Prin urmare numărul* $\sqrt{p}$ *se poate aproxima oricât de bine cu termenii şirului iterativ* $\{x_k\}_{k\in\mathbb{N}}$, *generate prin relaţia de recurenţă* $x_{k+1} = \dfrac{1}{2}\left(x_k + \dfrac{p}{x_k}\right)$.

Capitolul 6

Sisteme de ecuaţii liniare

În funcţie de numărul necunoscutelor se disting mai multe metode:

1. dacă numărul necunoscutelor are ordinul mai mic decât 10^3 atunci se folosesc metodele directe de rezolvare numerică a sistemelor liniare (de exemplu: metoda lui Gauss, metoda descompunerii LU, metoda descompunerii LL^T, metoda descompunerii QR, etc.), fiindcă acumularea erorilor încă nu influenţează rezultatul final.

2. dacă numărul necunoscutelor are ordinul mai mare decât 10^3 şi mai mic decât 10^6 atunci se folosesc metodele iterative de rezolvare numerică a sistemelor liniare (de exemplu metoda lui Jacobi, metoda lui Seidel, metoda SOR), fiindcă aceste metode sunt autocorectoare.

3. dacă numărul necunoscutelor are ordinul mai mare decât 10^6 se folosesc metodele probabilistice de rezolvare numerică a sistemelor liniare (de exemplu: metode de tip Monte Carlo).

6.1 Metode directe de rezolvare numerică a sistemelor liniare

6.1.1 Rezolvarea unor sisteme liniare particulare

6.1.1.1 Rezolvarea sistemelor liniare diagonale

Un sistem liniar de forma $\begin{cases} a_{11}x_1 = b_1 \\ a_{22}x_2 = b_2 \\ \vdots \\ a_{nn}x_n = b_n \end{cases}$ se numeşte sistem liniar diagonal, căci în matricea sistemului $A = \begin{pmatrix} a_{11} & & 0 \\ & a_{22} & \\ 0 & & \ddots \\ & & a_{nn} \end{pmatrix}$ toate elementele în afara elementelor de pe diagonala principală sunt egale cu zero. Condiţia necesară şi suficientă ca un sistem diagonal să admită o unică soluţie este ca $a_{ii} \neq 0$ pentru orice $i = \overline{1,n}$. În acest caz soluţia sistemului este $x_1 = \dfrac{b_1}{a_{11}}, x_2 = \dfrac{b_2}{a_{22}}, \ldots, x_n = \dfrac{b_n}{a_{nn}}$.

6.1.1.2 Rezolvarea sistemelor liniare superior triunghiulare

Un sistem liniar de forma

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

se numeşte sistem liniar superior triunghiular, căci în matricea sistemului

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ & a_{22} & \ldots & a_{2n} \\ & 0 & & \vdots \\ & & & a_{nn} \end{pmatrix}$$

toate elementele de sub diagonala principală sunt egale cu zero. Condiţia necesară şi suficientă ca un sistem liniar superior triunghiular să admită o singură soluţie este ca

$a_{ii} \neq 0$ pentru orice $i = \overline{1,n}$. În acest caz sistemul se rezolvă începând de la coadă, adică

$$x_n = \frac{b_n}{a_{nn}}, x_{n-1} = \frac{b_{n-1} - a_{n-1,n} \cdot x_n}{a_{n-1,n-1}}, \ldots, x_k = \frac{b_k - \sum_{i=k+1}^{n} a_{ki}x_i}{a_{kk}}, \ldots, x_1 = \frac{b_1 - \sum_{i=2}^{n} a_{1i}x_i}{a_{11}}.$$

Program sistem superior triunghiular

Datele de intrare: a_{ij} pentru $i, j = \overline{1,n}$ cu $i \leq j$ şi b_i pentru $i = \overline{1,n}$.
Pentru $k = \overline{n,1}$ execută
$S := 0$;
Pentru $i = \overline{k+1,n}$ execută $S := S + a_{ki}x_i$;
$x_k = (b_k - S)/a_{kk}$.
Tipăreşte x_k pentru $k = \overline{1,n}$.

6.1.1.3 Rezolvarea sistemelor liniare inferior triunghiulare

Un sistem liniar de forma $\begin{cases} a_{11}x_1 & = b_1 \\ a_{21}x_1 + a_{22}x_2 & = b_2 \\ \vdots & \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n & = b_n \end{cases}$ se numeşte sistem liniar inferior triunghiular, căci în matricea sistemului $A = \begin{pmatrix} a_{11} & & & \\ a_{21} & a_{22} & & 0 \\ \vdots & & & \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{pmatrix}$ toate elementele deasupra diagonalei principale sunt egale cu zero. Condiţia necesară şi suficientă ca un sistem liniar inferior triunghiular să admită o singură soluţie este ca $a_{ii} \neq 0$ pentru orice $i = \overline{1,n}$. În acest caz sistemul se rezolvă începând cu necunoscuta x_1, adică

$$x_1 = \frac{b_1}{a_{11}}, x_2 = \frac{b_2 - a_{21}x_1}{a_{22}}, \ldots, x_k = \frac{b_k - \sum_{i=1}^{k-1} a_{ki}x_i}{a_{kk}}, \ldots, x_n = \frac{b_n - \sum_{i=1}^{n-1} a_{ni}x_i}{a_{nn}}.$$

Program sistem inferior triunghiular

Datele de intrare: a_{ij} pentru $i, j = \overline{1,n}$ cu $i \geq j$ şi b_i pentru $i = \overline{1,n}$.
Pentru $k = \overline{1,n}$ execută

$S := 0$;

Pentru $i = \overline{1, k-1}$ execută $S := S + a_{ki}x_i$;

$x_k = (b_k - S)/a_{kk}$.

Tipăreşte x_k pentru $k = \overline{1, n}$.

6.1.2 Metoda lui Gauss

Se consideră sistemul

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

Ideea lui Gauss este de a reduce acest sistem la un sistem diagonal, inferior triunghiular, sau superior triunghiular, care se rezolvă conform paragrafului anterior. Noi vom reduce acest sistem la un sistem superior triunghiular. Se notează cu $a_{ij}^{(1)} = a_{ij}$ şi $b_i^{(1)} = b_i$ pentru $i, j = \overline{1, n}$. Prin urmare avem sistemul: $\begin{cases} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + \cdots + a_{1n}^{(1)}x_n = b_1^{(1)} \\ a_{21}^{(1)}x_1 + a_{22}^{(1)}x_2 + \cdots + a_{2n}^{(1)}x_n = b_2^{(1)} \\ \vdots \\ a_{n1}^{(1)}x_1 + a_{n2}^{(1)}x_2 + \cdots + a_{nn}^{(1)}x_n = b_n^{(1)} \end{cases}$. Se presupune că $a_{11}^{(1)} \neq 0$, care se numeşte elementul pivot. Prima ecuaţie se normează, adică se împarte la elementul pivot $a_{11}^{(1)}$, după care prima ecuaţie se înmulţeşte pe rând cu elementele $-a_{21}^{(1)}, -a_{31}^{(1)}, \ldots, -a_{n1}^{(1)}$ şi se adună la a doua, la a treia, etc. respectiv la ultima ecuaţie. În acest fel necunoscuta x_1 se elimină din a doua, a treia, etc., şi ultima ecuaţie. Se obţine următorul sistem

$$\begin{cases} x_1 + \dfrac{a_{12}^{(1)}}{a_{11}^{(1)}}x_2 + \cdots + \dfrac{a_{1n}^{(1)}}{a_{11}^{(1)}}x_n = \dfrac{b_1^{(1)}}{a_{11}^{(1)}} & /(-a_{21}^{(1)}), \ldots, /(-a_{n1}^{(1)}) \\ 0 + \left(a_{22}^{(1)} - \dfrac{a_{12}^{(1)}}{a_{11}^{(1)}}a_{21}^{(1)}\right)x_2 + \cdots + \left(a_{2n}^{(1)} - \dfrac{a_{1n}^{(1)}}{a_{11}^{(1)}}a_{21}^{(1)}\right)x_n = b_2^{(1)} - \dfrac{b_1^{(1)}}{a_{11}^{(1)}}a_{21}^{(1)} \\ \vdots \\ 0 + \left(a_{n2}^{(1)} - \dfrac{a_{12}^{(1)}}{a_{11}^{(1)}}a_{n1}^{(1)}\right)x_2 + \cdots + \left(a_{nn}^{(1)} - \dfrac{a_{1n}^{(1)}}{a_{11}^{(1)}}a_{n1}^{(1)}\right)x_n = b_n^{(1)} - \dfrac{b_1^{(1)}}{a_{11}^{(1)}}a_{n1}^{(1)} \end{cases}$$

Se notează $a_{11}^{(2)} = 1$; pentru $j = \overline{2,n}$ $a_{1j}^{(2)} = \dfrac{a_{1j}^{(1)}}{a_{11}^{(1)}}$; $b_1^{(2)} = \dfrac{b_1^{(1)}}{a_{11}^{(1)}}$; pentru $i = \overline{2,n}$ $a_{i1}^{(2)} = 0$; pentru $i,j = \overline{2,n}$ $a_{ij}^{(2)} = a_{ij}^{(1)} - \dfrac{a_{1j}^{(1)}}{a_{11}^{(1)}} \cdot a_{i1}^{(1)}$; pentru $i = \overline{2,n}$ $b_i^{(2)} = b_i^{(1)} - \dfrac{b_1^{(1)}}{a_{11}^{(1)}} a_{i1}^{(1)}$. În acest fel se obţine sistemul:

$$\begin{cases} a_{11}^{(2)}x_1 + a_{12}^{(2)}x_2 + \cdots + a_{1n}^{(2)}x_n = b_1^{(2)} \\ a_{21}^{(2)}x_1 + a_{22}^{(2)}x_2 + \cdots + a_{2n}^{(2)}x_n = b_2^{(2)} \\ \vdots \\ a_{n1}^{(2)}x_1 + a_{n2}^{(2)}x_2 + \cdots + a_{nn}^{(2)}x_n = b_n^{(2)} \end{cases} \quad \text{adică}$$

$$\begin{cases} a_{11}^{(2)}x_1 + a_{12}^{(2)}x_2 + \cdots + a_{1n}^{(2)}x_n = b_1^{(2)} \\ \qquad\qquad a_{22}^{(2)}x_2 + \cdots + a_{2n}^{(2)}x_n = b_2^{(2)} \\ \qquad\qquad\qquad\qquad \vdots \\ \qquad\qquad a_{n2}^{(2)}x_2 + \cdots + a_{nn}^{(2)}x_n = b_n^{(2)} \end{cases}$$

Se alege ca pivot elementul $a_{22}^{(2)} \neq 0$ şi a doua ecuaţie se normează, adică se împarte la elementul pivot, după care se elimină variabila x_2 din a treia, a patra, etc., ultima ecuaţie prin înmulţirea ecuaţiei a doua cu elementele $-a_{32}^{(2)}, \ldots, -a_{n2}^{(2)}$. Astfel se obţine următorul sistem:

$$\begin{cases} a_{11}^{(2)}x_1 + a_{12}^{(2)}x_2 + a_{13}^{(2)}x_3 + \cdots + a_{1n}^{(2)}x_n = b_1^{(2)} \\ \\ \qquad x_2 + \dfrac{a_{23}^{(2)}}{a_{22}^{(2)}}x_3 + \cdots + \dfrac{a_{2n}^{(2)}}{a_{22}^{(2)}}x_n = \dfrac{b_2^{(2)}}{a_{22}^{(2)}} \qquad /(-a_{32}^{(2)}) \ldots /(-a_{n2}^{(2)}) \\ \\ \qquad 0 + \left(a_{33}^{(2)} - \dfrac{a_{23}^{(2)}}{a_{22}^{(2)}}a_{32}^{(2)}\right)x_3 + \cdots + \left(a_{3n}^{(2)} - \dfrac{a_{2n}^{(2)}}{a_{22}^{(2)}}a_{32}^{(2)}\right)x_n = b_2^{(2)} - \dfrac{b_2^{(2)}}{a_{22}^{(2)}}a_{32}^{(2)} \\ \qquad \vdots \\ \qquad 0 + \left(a_{n3}^{(2)} - \dfrac{a_{23}^{(2)}}{a_{22}^{(2)}}a_{n2}^{(2)}\right)x_3 + \cdots + \left(a_{nn}^{(2)} - \dfrac{a_{2n}^{(2)}}{a_{22}^{(2)}}a_{n2}^{(2)}\right)x_n = b_n^{(2)} - \dfrac{b_2^{(2)}}{a_{22}^{(2)}}a_{n2}^{(2)}. \end{cases}$$

Se introduc notaţiile $a_{ij}^{(3)}$ şi $b_i^{(3)}$ pentru $i, j = \overline{1,n}$. Se presupune că la pasul k s-a obţinut următorul sistem:

$$\begin{cases} a_{11}^{(k)} + a_{12}^{(k)} x_2 + \cdots + a_{1k}^{(k)} x_k + \cdots + a_{1n}^{(k)} x_n = b_1^{(k)} \\ a_{22}^{(k)} x_2 + \cdots + a_{2k}^{(k)} x_k + \cdots + a_{2n}^{(k)} x_n = b_2^{(k)} \\ \vdots \\ a_{kk}^{(k)} x_k + \cdots + a_{kn}^{(k)} x_n = b_k^{(k)} \\ a_{k+1,k}^{(k)} x_k + \cdots + a_{k+1,n}^{(k)} x_n = b_{k+1}^{(k)} \\ \vdots \\ a_{n,k}^{(k)} x_k + \cdots + a_{n,n}^{(k)} x_n = b_n^{(k)} \end{cases}$$

Se presupune că pivotul $a_{kk}^{(k)}$ este diferit de zero şi ecuaţia k se normează, adică se împarte la pivot, după care se elimină necunoscuta x_k din ecuaţiile $k+1, \ldots, n$ prin înmulţirea succesivă a ecuaţiei normate cu elementele $-a_{k+1,k}^{(k)}, \ldots, -a_{n,k}^{(k)}$ şi prin adunarea acesteia la ecuaţiile $k+1, \ldots, n$. Astfel se obţine următorul sistem:

$$\begin{cases} a_{11}^{(k)} x_1 + a_{12}^{(k)} x_2 + \cdots + a_{1k}^{(k)} x_k + \cdots + a_{1n}^{(k)} x_n = b_1^{(k)} \\ a_{22}^{(k)} x_2 + \cdots + a_{2k}^{(k)} x_k + \cdots + a_{2n}^{(k)} x_n = b_2^{(k)} \\ \vdots \\ x_k + \cdots + \dfrac{a_{kn}^{(k)}}{a_{kk}^{(k)}} x_n = \dfrac{b_k^{(k)}}{a_{kk}^{(k)}} \\ 0 + \cdots + \left(a_{k+1,n}^{(k)} - \dfrac{a_{kn}^{(k)}}{a_{kk}^{(k)}} \cdot a_{k+1,k}^{(k)} \right) x_n = b_{k+1}^{(k)} - \dfrac{b_k^{(k)}}{a_{kk}^{(k)}} \cdot a_{k+1,k}^{(k)} \\ \vdots \\ 0 + \cdots + \left(a_{n,n}^{(k)} - \dfrac{a_{kn}^{(k)}}{a_{kk}^{(k)}} \cdot a_{n,k}^{(k)} \right) x_n = b_n^{(k)} - \dfrac{b_k^{(k)}}{a_{kk}^{(k)}} \cdot a_{n,k}^{(k)} \end{cases}$$

Folosind următoarele notaţii: pentru $i = \overline{1,k-1}$ şi $j = \overline{1,n}$ $a_{ij}^{(k+1)} = a_{ij}^{(k)}$ şi $b_i^{(k+1)} = b_i^{(k)}$; $a_{k,k}^{(k+1)} = 1$; pentru $j = \overline{k+1,n}$ $a_{kj}^{(k+1)} = \dfrac{a_{kj}^{(k)}}{a_{kk}^{(k)}}$; $b_k^{(k+1)} = \dfrac{b_k^{(k)}}{a_{kk}^{(k)}}$; pentru $i = \overline{k+1,n}$ $a_{i,k}^{(k+1)} = 0$;

pentru $i, j = \overline{k+1, n}$ $a_{ij}^{(k+1)} = a_{ij}^{(k)} - \dfrac{a_{kj}^{(k)}}{a_{kk}^{(k)}} \cdot a_{ik}^{(k)}$ şi $b_i^{(k+1)} = b_i^{(k)} - \dfrac{b_k^{(k)}}{a_{kk}^{(k)}} \cdot a_{ik}^{(k)}$; se obţine sistemul:

$$\begin{cases} a_{11}^{(k+1)}x_1 + a_{12}^{(k+1)}x_2 + \cdots + a_{1k}^{(k+1)}x_k + a_{1,k+1}^{(k+1)}x_{k+1} + \cdots + a_{1,n}^{(k+1)}x_n = b_1^{(k+1)} \\ a_{22}^{(k+1)}x_2 + \cdots + a_{2k}^{(k+1)}x_k + a_{2,k+1}^{(k+1)}x_{k+1} + \cdots + a_{2,n}^{(k+1)}x_n = b_2^{(k+1)} \\ \vdots \\ a_{k,k}^{(k+1)}x_k + a_{k,k+1}^{(k+1)}x_{k+1} + \cdots + a_{k,n}^{(k+1)}x_n = b_k^{(k+1)} \\ a_{k+1,k+1}^{(k+1)}x_{k+1} + \cdots + a_{k+1,n}^{(k+1)}x_n = b_{k+1}^{(k+1)} \\ \vdots \\ a_{n,k+1}^{(k+1)}x_{k+1} + \cdots + a_{n,n}^{(k+1)}x_n = b_n^{(k+1)} \end{cases}$$

Procedeul de triunghiularizare arătat mai sus se continuă până când $k \leq n-1$. Astfel se obţine un sistem superior triunghiular care se rezolvă conform § 6.1.1.2. Din formulele de recurenţă din faţa sistemului obţinut la pasul $k+1$ se deduce:

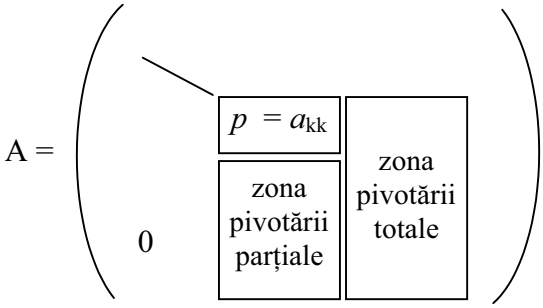
Algoritmul metoda Gauss

Datele de intrare: a_{ij}, b_i pentru $i, j = \overline{1, n}$.
Pentru $k = \overline{1, n-1}$ execută
    dacă $a_{kk} \neq 0$ atunci $p := a_{kk}$ altfel stop;
    pentru $j = \overline{k, n}$ execută $a_{kj} = a_{kj}/p$;
    $b_k := b_k/p$;
    pentru $i = \overline{k+1, n}$ execută
        pentru $j = \overline{k+1, n}$ execută
            $a_{ij} := a_{ij} - a_{kj} \cdot a_{ik}$;
        $b_i := b_i - b_k \cdot a_{ik}$;
Aplică algoritmul sistem superior triunghiular;
Tipăreşte x_k pentru $k = \overline{1, n}$.

Ca exerciţiu vă propunem să se facă reducerea sistemului iniţial la un sistem inferior triunghiular respectiv diagonal.

În algoritmul metoda Gauss se poate întâmpla ca la pasul k elementul pivot $p := a_{kk} = 0$ sau $p := a_{kk} \approx 0$, adică este o valoare în modul foarte mică. În acest caz pentru a continua algoritmul lui Gauss este necesar un procedeu de pivotare. Vom folosi două metode de pivotare: pivotare parţială respectiv pivotare totală. În cazul **pivotării**

parţiale presupunem că există cel puţin un element nenul printre elementele a_{ik}, unde $i = \overline{k+1, n}$, adică printre elementele care sunt situate sub elementul pivot în coloana pivotului. Dacă $a_{i_0 k} \neq 0$ unde $i_0 \in \{k+1, \ldots, n\}$, atunci se schimbă între ele valorile a_{kk} şi $a_{i_0 k}$, care practic înseamnă schimbarea liniilor k şi i_0 între ele. Schimbarea a două linii nu afectează cu nimic rezolvarea sistemului iniţial. **Pivotarea totală** se aplică în cazul când la pivotarea parţială nu se găseşte un element nenul. În acest caz se caută un element nenul a_{i_0, j_0}, unde $i_0 \in \{k, \ldots, n\}$ şi $j_0 \in \{k+1, \ldots, n\}$, după care elementul a_{i_0, j_0} se duce pe poziţia elementului pivot. Practic asta înseamnă că se schimbă între ele liniile i_0 şi k, după care se schimbă între ele coloanele j_0 şi k. Schimbarea a două linii nu afectează cu nimic rezolvarea sistemului iniţial, însă schimbarea a două coloane induce schimbarea a două necunoscute între ele.



Prin urmare în cazul aplicării pivotării parţiale în algoritmul metoda Gauss în locul instrucţiunii "stop" se pune următorul bloc de instrucţiuni:

pentru $i_0 = \overline{k+1, n}$ caută $a_{i_0, k} \neq 0$;
schimbă între ele liniile i_0 şi k;

iar în cazul aplicării pivotării totale se pune următorul bloc de instrucţiuni:

pentru $i_0 = \overline{k, n}$ execută
    pentru $j_0 = \overline{k+1, n}$ execută
        caută $a_{i_0, j_0} \neq 0$
schimbă între ele liniile i_0 şi k;
schimbă între ele coloanele j_0 şi k.

În acest ultim caz la sfârşitul algoritmului trebuie pusă instrucţiunea schimbă între ele valorile x_k şi x_{j_0}.

Din liceu se cunoaşte următorul rezultat:

Teorema 6.1.1. *Sistemul iniţial cu n ecuaţii şi n necunoscute este un sistem de tip Cramer, adică admite o unică soluţie, dacă şi numai dacă* $\det A \neq 0$, *unde* $A = (a_{ij})_{i,j=\overline{1,n}}$

este matricea sistemului.

În acest caz, dacă elementul pivot devine nul, cu ajutorul pivotării parţiale sau a pivotării totale putem continua mai departe algoritmul metodei Gauss.

În cazul când $\det A = 0$ găsim un $k \in \{1, \ldots, n\}$ astfel încât $a_{ij} = 0$ pentru orice $i, j = \overline{k, n}$. Dacă există un $i \in \{k, \ldots, n\}$ astfel încât $b_i \neq 0$ atunci sistemul este incompatibil, fiindcă ecuaţia a $i - a$ nu are loc, iar dacă pentru orice $i = \overline{k, n}$ $b_i = 0$ sistemul este compatibil nedeterminat. Practic se poate întâmpla ca valorile $b_i \approx 0$ pentru $i = \overline{k, n}$, sunt mici în valoare absolută şi se pot considera ca fiind egale cu zero. Menţionăm că în acest caz deşi din punct de vedere teoretic sistemul liniar este incompatibil, totuşi calculatorul ne indică un sistem compatibil şi nedeterminat.

6.1.3 Aplicaţii ale metodei lui Gauss

6.1.3.1 Calculul determinantului folosind metoda lui Gauss

Se consideră matricea $A = (a_{ij})_{i,j=\overline{1,n}}$. Se pune problema calculului determinantului matricii A, notata cu $\det A$. Se observă că, calculul determinantului nu se poate face după definiţia clasică a determinantului învăţată în liceu, fiindcă ne conduce la un volum imens de calcule având $n!$ de termeni. De aceea e nevoie de metode numerice adecvate. Aici vom prezenta calculul determinantului folosind metoda lui Gauss. Fie " det " variabila corespunzătoare, pe care o iniţializăm la începutul algoritmului cu det $:= 1$. Dacă în algoritmul metodei Gauss elementul $a_{kk} \neq 0$ pentru $k = \overline{1, n-1}$, atunci $p := a_{kk}$ şi punem instrucţiunea det $:=$ det $* p$. După ce se termină în metoda Gauss ciclul în raport cu contorul k punem instrucţiunea det $:=$ det $* a_{nn}$. În acest fel în variabila det vom avea tocmai valoarea determinantului matricii A. Într-adevăr, algoritmul lui Gauss transformă matricea A într-o matrice superior triunghiulară, a cărei determinant este tocmai produsul elementelor de pe diagonala principală. Pe primele $n - 1$ poziţii de pe diagonala principală a matricii superior triunghiulare avem valoarea unu, dar pivotul fiind scos ca un factor forţat de pe linii, intră ca un factor de multiplicare la calculul determinantului matricii A. Menţionăm că, dacă pivotul la un anumit pas este zero se aplică pivotarea parţială sau totală, care înseamnă schimbarea a două linii sau a două coloane între ele. Din punct de vedere al calculului determinantului, aceste operaţii induc schimbarea semnului determinantului. De aceea în algoritmul Gauss în cazul pivotării parţiale respectiv

totale de fiecare dată se pune câte o instrucţiune de forma: det := − det . La sfârşitul algoritmului Gauss punem instrucţiunea de afişare a rezultatului variabilei det.

6.1.3.2 Calculul matricii inverse folosind metoda lui Gauss

Fie A o matrice pătratică de ordinul n, nesingulară (cu determinant nenul). Se ştie că în acest caz există matricea inversă A^{-1} şi are loc egalitatea: $A \cdot A^{-1} = I_n$, unde I_n este matricea unitate de ordinul n. Notăm $A = (a_{ij})_{i,j=\overline{1,n}}$, $A^{-1} = (x_{ij})_{i,j=\overline{1,n}}$. Astfel avem egalitatea

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} \cdot \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nn} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{pmatrix}.$$

Dacă liniile matricii A se înmulţesc numai cu prima coloană a matricii inverse, cu vectorul $\begin{pmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{pmatrix}$, se obţine un sistem liniar având termenul liber prima coloană a matricii I_n :

$$\begin{cases} a_{11}x_{11} + a_{12}x_{21} + \dots + a_{1n}x_{n1} = 1 \\ a_{21}x_{11} + a_{22}x_{21} + \dots + a_{2n}x_{n1} = 0 \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ a_{n1}x_{11} + a_{n2}x_{21} + \dots + a_{nn}x_{n1} = 0 \end{cases}.$$

Acest sistem liniar are matricea tocmai matricea A cu determinant diferit de zero, deci este un sistem liniar de tip Cramer. Unica soluţie a sistemului se obţine cu algoritmul metodei Gauss având termenul liber $e_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$, primul vector din baza canonică a spaţiului $\mathbb{R}^n$. Soluţia obţinută este tocmai prima coloană a lui A^{-1}. La al doilea pas se

fixează a doua coloană a matricii inverse A^{-1}, vectorul $\begin{pmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{n2} \end{pmatrix}$ şi se obţine în mod similar un sistem liniar de forma:

$$A \cdot \begin{pmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{n2} \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix} = e_2,$$

care iarăşi se rezolvă cu metoda lui Gauss, ş.a.m.d. Unica soluţie a sistemului liniar se aşează pe a doua coloană a lui A^{-1}. Prin urmare, calculul matricii inverse pentru o matrice A de ordinul n, nesingulară, se reduce la rezolvarea consecutivă a n sisteme liniare cu metoda lui Gauss, alegând pentru termenii liberi vectorii $e_1, e_2, \ldots, e_n$ din baza canonică a lui $\mathbb{R}^n$.

Soluţiile obţinute se aranjează pe rând pe coloanele matricii inverse A^{-1}.

Program matrice inversă

Datele de intrare: $n; A$;
Pentru $i = \overline{1, n}$ execută
    $b := e_i$;
    Rezolvă sistemul $Ax = b$ cu metoda lui Gauss;
    Pentru $j = \overline{1, n}$ execută
        $A^{-1}[j, i] := x[j]$;
Tipăreşte A^{-1}.

6.1.3.3 Calculul rangului unei matrici folosind metoda lui Gauss

Fie $A = (a_{ij})_{i,j=\overline{1,n}}$ o matrice pătratică de ordinul n. Ne interesează rangul matricii A. Aplicând algoritmul lui Gauss pentru matricea A, se obţine o nouă matrice, care are toate elementele de sub diagonala principală egale cu zero, iar pe diagonala principală apar elementele nenule sau nule. Numărul elementelor nenule de pe diagonala principală este tocmai rangul matricii A. Prin urmare, în algoritmul Gauss, la sfârşitul programului

numărăm toate elementele nenule de pe diagonala principală, începând cu elementul care se află pe poziţia primei linii şi a primei coloane. Menţionăm că elementele matricii A trebuie să fie în aşa fel, încât în algoritmul Gauss la un anumit pas să nu apară numere foarte mari în valoare absolută.

Program rangul matricii

Datele de intrare: $n; A$;

Aplică algoritmul metoda Gauss; (vezi paragraful 6.1.2)

Fie $k := 0$;

Pentru $k = \overline{1,n}$ execută

dacă $a_{kk} \neq 0$ atunci $k := k + 1$;

Tipăreşte k.

6.1.4 Metode de factorizare

La aceste metode numerice ideea de bază este să factorizăm, să descompunem matricea sistemului liniar într-un produs de două matrici cu "structură" mai simplă.

6.1.4.1 Metoda descompunerii LU

Dacă se dă o matrice pătratică de ordinul n, $A = (a_{ij})_{i,j=\overline{1,n}}$ cu elemente numere reale se pune problema de a descompune matricea A într-un produs de două matrici triunghiulare: $A = L \cdot U$, unde L este o matrice inferior triunghiulară (vezi paragraful 6.1.1.3) iar U este o matrice superior triunghiulară (vezi paragraful 6.1.1.2). Rostul acestei descompuneri se poate argumenta în felul următor: dacă se consideră sistemul $A \cdot x = b$, unde $x = (x_1, x_2, \ldots, x_n)^T$ şi $b = (b_1, b_2, \ldots, b_n)^T$ şi urmărim să-l rezolvăm, adică să determinăm necunoscuta x în funcţie de datele problemei A şi b, o posibilitate ar fi efectuarea descompunerii matricii A în produsul $L \cdot U$, fiindcă dacă am reuşit această descompunere soluţia sistemului $A \cdot x = b$ se reduce la rezolvarea sistemelor $L \cdot y = b$ şi $U \cdot x = y$ ($A \cdot x = (L \cdot U) \cdot x = L \cdot (U \cdot x) = L \cdot y = b$). Rezolvarea acestor sisteme se face uşor conform paragrafelor 6.1.1.2 şi 6.1.1.3. Prin urmare să ne întoarcem la posibilitatea

efectuării descompunerii LU pentru o matrice A dată. În matricea

$$L = \begin{pmatrix} l_{11} & & 0 & \\ l_{21} & l_{22} & & \\ \vdots & \vdots & \ddots & \\ l_{n1} & l_{n2} & \dots & l_{nn} \end{pmatrix}$$

avem $1 + 2 + \cdots + n = \dfrac{n^2 + n}{2}$ necunoscute, iar în matricea

$$U = \begin{pmatrix} u_{11} & u_{12} & \dots & u_{1n} \\ & u_{22} & \dots & u_{2n} \\ & & \ddots & \vdots \\ & 0 & & u_{nn} \end{pmatrix}$$

avem iarăşi $n + (n-1) + \cdots + 1 = \dfrac{n^2 + n}{2}$ necunoscute. În total avem de determinat $n^2 + n$ necunoscute. Prin înmulţirea directă a lui L cu U şi prin egalarea rezultatelor cu elementele corespunzătoare ale matricii A obţinem numai n^2 de egalităţi. Prin urmare, în general din n^2 egalităţi nu se pot determina $n^2 + n$ necunoscute, adică pentru o matrice A dată, descompunerea $L \cdot U$ nu este unică, deci problema descompunerii este nedeterminată. Într-adevăr, dacă $A = L \cdot U$ este o descompunere şi considerăm pe D o matrice diagonală astfel încât fiecare element de pe diagonala principală a lui D nu este zero, atunci există D^{-1} şi putem obţine o nouă descompunere gen LU pentru matricea A :

$$A = L \cdot U = L(D \cdot D^{-1}) \cdot U = (L \cdot D) \cdot (D^{-1} \cdot U) = L' \cdot U',$$

unde $L' = L \cdot D$ este noua matrice inferior triunghiulară iar $U' = D^{-1} \cdot U$ este noua matrice superior triunghiulară. Ca să asigurăm unicitatea descompunerii LU, luăm sau diagonala principală a lui L sau diagonala principală a lui U să fie egală cu un şir de unu. Noi fixăm diagonala lui L cu numerele reale unu. Deci avem:

$$L = \begin{pmatrix} 1 & & & & \\ l_{21} & 1 & & 0 & \\ l_{31} & l_{32} & 1 & & \\ \dots & \dots & \dots & & \\ l_{n1} & l_{n2} & l_{n3} & \dots & 1 \end{pmatrix}$$

şi

$$U = \begin{pmatrix} u_{11} & u_{12} & \dots & u_{1n} \\ & u_{22} & \dots & u_{2n} \\ & & \ddots & \vdots \\ & 0 & & u_{nn} \end{pmatrix}$$

În continuare prezentăm un algoritm de determinare a elementelor l_{ij} din L şi u_{ij} din U, pas cu pas:

pas 1. se determină u_{11} : $\quad 1 \cdot u_{11} = a_{11}$;

pas 2. se determină $l_{21}, l_{31}, \dots, l_{n1}$:
$$\begin{aligned} & l_{21} \cdot u_{11} = a_{21}; \\ & l_{31} \cdot u_{11} = a_{31}; \\ & \vdots \\ & l_{n1} \cdot u_{11} = a_{n1}. \end{aligned}$$

Observăm că pentru determinarea elementelor $l_{21}, l_{31}, \dots, l_{n1}$ trebuie să impunem condiţia ca $u_{11} \neq 0$.

pas 3. se determină $u_{12}, u_{13}, \dots, u_{1n}$:
$$\begin{aligned} & 1 \cdot u_{12} = a_{12}; \\ & 1 \cdot u_{13} = a_{13}; \\ & \vdots \\ & 1 \cdot u_{1n} = a_{1n}. \end{aligned}$$

După ce prima coloană a lui L şi prima linie a lui U sunt determinate urmează a doua coloană a lui L şi a doua linie a lui U :

pas 4. se determină u_{22} : $\quad l_{21} \cdot u_{12} + 1 \cdot u_{22} = a_{22}$ cu necunoscuta u_{22}.

pas 5. se determină $l_{32}, \dots, l_{n2}$:
$$\begin{aligned} & l_{31} \cdot u_{12} + l_{32} \cdot u_{22} = a_{22}; \\ & \vdots \\ & l_{n1} \cdot u_{12} + l_{n2} \cdot u_{22} = a_{n2}. \end{aligned}$$

Observăm că pentru determinarea elementelor $l_{32}, \dots, l_{n2}$ avem nevoie ca $u_{22} \neq 0$.

pas 6. se determină $u_{22}, \dots, u_{2n}$:
$$\begin{aligned} & l_{21} \cdot u_{13} + 1 \cdot u_{23} = a_{23}; \\ & \vdots \\ & l_{21} \cdot u_{1n} + 1 \cdot u_{2n} = a_{2n}. \end{aligned}$$

Mai departe algoritmul se continuă în mod similar pentru celelalte coloane ale lui L şi celelalte linii ale lui U.

Teorema 6.1.4.1. *Dacă într-o matrice A minorii principali sunt nenuli, atunci are loc descompunerea LU şi în mod unic.*

DEMONSTRAŢIE. Prin minorii principali ai matricii A vom înţelege determinanţii care se formează începând cu elementul, care se află pe prima linie şi pe prima coloană a matricii A:

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \dots & a_{3n} \\ \dots & \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} \end{pmatrix}, \text{ adică } D_1 = |a_{11}|,$$

$$D_2 = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}, \; D_3 = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}, \dots, \; D_n = \det(A).$$

Teorema afirmă faptul că, dacă $D_i \neq 0$ pentru $i = \overline{1,n}$, atunci există descompunerea LU şi în mod unic. Presupunem că avem descompunerea

$$A = LU = \begin{pmatrix} 1 & & & & \\ l_{21} & 1 & & & \\ l_{31} & l_{32} & 1 & & \\ \dots & \dots & \dots & \dots & \\ l_{n1} & l_{n2} & l_{n3} & \dots & 1 \end{pmatrix} \cdot \begin{pmatrix} u_{11} & u_{12} & \dots & u_{1n} \\ & u_{22} & \dots & u_{2n} \\ & & \ddots & \vdots \\ & 0 & & u_{nn} \end{pmatrix},$$

de unde $\det(A) = \det(L \cdot U) = \det L \cdot \det U = 1 \cdot u_{11}u_{22} \dots u_{nn} = u_{11}u_{22} \dots u_{nn}$. În mod analog dacă alegem matricile din matricea A, care corespund minorilor principali din A, atunci trebuie să alegem tot matricile corespunzătoare minorilor principali din L şi U. Adică: $D_1 = 1 \cdot u_{11}$, $D_2 = 1 \cdot u_{11} \cdot u_{22}$, $D_3 = 1 \cdot u_{11} \cdot u_{22} \cdot u_{33}, \dots$, $D_n = \det(A) = 1 \cdot u_{11} \cdot u_{22} \dots u_{nn}$. Înaintea teoremei, la algoritmul de determinare a elementelor lui L şi U singura restricţie care se impune este ca pe rând u_{11}, $u_{22}, \dots, u_{n-1,n-1}$ să fie diferite de zero. Observăm că pentru $u_{n,n}$ nu e nevoie de condiţia ca $u_{n,n} \neq 0$. Prin urmare $u_{11} \neq 0$ este echivalent cu $D_1 \neq 0$, $u_{22} \neq 0$ este echivalent cu $D_2 \neq 0$, $u_{33} \neq 0$ este

echivalent cu $D_3 \neq 0, \ldots, u_{n-1,n-1} \neq 0$ este echivalent cu $D_{n-1,n-1} \neq 0$. Sigur că, condiţia $D_n = \det(A) \neq 0$ este echivalentă cu $u_{n,n} \neq 0$. □

Observaţia 6.1.4.1. *Din demonstraţie reiese un calcul foarte simplu pentru* $\det(A)$. *Într-adevăr dacă am făcut descompunerea LU pentru matricea A, atunci se cunosc elementele matricii U de pe diagonala principală şi avem* $\det(A) = u_{11} \cdot u_{22} \cdot \ldots \cdot u_{nn}$.

Menţionăm că descompunerea LU are loc şi într-un caz mai general, când se cere numai ca matricea A să fie inversabilă. Într-adevăr, $\det(A) = D_n = u_{11}u_{22} \ldots u_{nn} \neq 0$ implică $u_{ii} \neq 0$ pentru orice $i = \overline{1,n}$. În general când efectuăm descompunerea LU se poate întâmpla, că vom fi nevoiţi să schimbăm între ele două linii sau două coloane ale matricii A pentru a avea $u_{ii} \neq 0$.

Program factorizare LU

Datele de intrare: n; A;

Pentru $k = \overline{1,n}$ execută: $l_{kk} = 1$;

Pentru $k = \overline{1,n}$ execută:

$$u_{kk} := a_{kk} - \sum_{j=1}^{k-1} l_{kj} * u_{jk};$$

(se obţine prin înmulţirea liniei k din L cu coloana k din U)

pentru $i = \overline{k+1,n}$ execută:

$$l_{ik} := \left(a_{ik} - \sum_{j=1}^{k-1} l_{ij} * u_{jk} \right) / u_{kk};$$

(se obţine prin înmulţirea liniei i din L cu coloana k din U);

$$u_{ki} := a_{ki} - \sum_{j=1}^{k-1} l_{kj} * u_{ji};$$

(se obţine prin înmulţirea liniei k din L cu coloana i din U);

Tipăreşte L şi U.

Dacă se cere să se rezolve sistemul liniar $A \cdot x = b$ cu metoda LU atunci avem următorul program:

Program 2 LU

Datele de intrare: n; A; b;

Aplică subrutina program factorizare LU;

Aplică subrutina matrice inferior triunghiulară

(conform §6.1.1.3 pentru sistemul $Ly = b$);

Aplică subrutina matrice superior triunghiulară

(conform §6.1.1.2 pentru sistemul $Ux = y$);

Tipăreşte x.

6.1.4.2 Metoda descompunerii LL^T (Cholesky)

Prin descompunerea LL^T a unei matrici $A = (a_{ij})_{i,j=\overline{1,n}}$ vom înţelege aceea descompunere, unde L este o matrice inferior triunghiulară, iar L^T înseamnă transpunerea matricii L, adică L^T va fi o matrice superior triunghiulară. Pentru a enunţa şi a demonstra o teoremă referitoare la existenţa şi unicitatea descompunerii LL^T avem nevoie în prealabil de câteva noţiuni asupra matricii A :

Definiţia 6.1.4.1. *O matrice A se numeşte simetrică, dacă $a_{ij} = a_{ji}$ pentru orice $i, j = \overline{1,n}$, adică elementele situate simetric faţă de diagonala principală a matricii A sunt egale.*

Definiţia 6.1.4.2. *O matrice A se numeşte pozitiv definită dacă $x^T Ax > 0$ pentru orice*

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n \setminus \{\theta_{\mathbb{R}^n}\},$$

unde x^T înseamnă transpusa vectorului coloană x. Prin urmare, dacă se efectuează înmulţirile în expresia $x^T Ax$ se obţine că matricea A este pozitiv definită, dacă forma pătratică corespunzătoare verifică $\sum_{i,j=1}^{n} a_{ij}x_i x_j > 0$ pentru orice $x_i \in \mathbb{R}$ cu $i = \overline{1,n}$, şi nu toţi x_i fiind deodată egali cu zero.

Aici menţionăm teorema lui Sylvester: o matrice A este pozitiv definită dacă şi numai dacă toţi minorii principali sunt pozitivi: $D_1 = |a_{11}| > 0$, $D_2 = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} > 0$, $D_3 =$

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} > 0, \ldots, D_n = \begin{vmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{vmatrix} = \det(A) > 0.$$

Teorema 6.1.4.2. *(Cholesky) Dacă matricea A este simetrică şi pozitiv definită atunci există factorizarea LL^T pentru A şi această descompunere este unică.*

DEMONSTRAŢIE. Demonstraţia vom face folosind metoda inducţiei matematice. Pentru $n = 1$ avem $A = L \cdot L^T$, adică $a_{11} = l_{11} \cdot l_{11} = l_{11}^2$. Dar $D_1 = |a_{11}| = a_{11} > 0$, deci există un singur număr real şi pozitiv l_{11} pentru care $l_{11} = \sqrt{a_{11}}$. Observăm că pentru a asigura unicitatea descompunerii luăm radicalul aritmetic, care la orice număr real şi pozitiv atribuie tot un număr real şi pozitiv. Presupunem că pentru matricea $A = A_{n-1}$ de ordinul $n-1$, simetrică şi pozitiv definită are loc descompunerea LL^T, adică există o matrice inferior triunghiulară L_{n-1} de ordinul $n-1$ astfel ca $A_{n-1} = L_{n-1} \cdot L_{n-1}^T$. Fie acum $A = A_n$ o matrice de ordinul n simetrică şi pozitiv definită:

$$A_n = \begin{pmatrix} a_{11} & \ldots & a_{1,n-1} & a_{1,n} \\ a_{21} & \ldots & a_{2,n-1} & a_{2,n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n-1,1} & \ldots & a_{n-1,n-1} & a_{n-1,n} \\ a_{n,1} & \ldots & a_{n,n-1} & a_{n,n} \end{pmatrix}.$$

Matricea A_n o tăiem în blocuri de matrici separând ultima linie şi ultima coloană a matricii A_n : $A_n = \left(\begin{array}{c|c} A_{n-1} & a \\ \hline a^T & a_{nn} \end{array}\right)$, unde $A_{n-1} = \begin{pmatrix} a_{11} & \ldots & a_{1,n-1} \\ a_{21} & \ldots & a_{2,n-1} \\ \ldots & \ldots & \ldots \\ a_{n-1,1} & \ldots & a_{n-1,n-1} \end{pmatrix}$, iar $a = \begin{pmatrix} a_{1,n} \\ a_{2,n} \\ \vdots \\ a_{n-1,n} \end{pmatrix}$. Aici menţionăm că, deoarece conform presupunerii A_n este o matrice simetrică, rezultă că A_{n-1} va fi tot o matrice simetrică, iar pe ultima linie a lui A_n putem pune a^T, care înseamnă transpunerea matricii coloană $a \in \mathbb{R}^{n-1}$ într-o ma-

trice linie. Fie $x = \left(\begin{array}{c|c} x' \\ \hline 0 \end{array}\right) \in \mathbb{R}^n$, unde $x' = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \end{pmatrix} \in \mathbb{R}^{n-1}$ este un vector arbitrar. Conform presupunerii A_n este o matrice pozitiv definită, deci $x^T Ax > 0$ pentru orice $x \in \mathbb{R}^n \setminus \{\theta_{\mathbb{R}^n}\}$ adică $(x'^T \mid 0) \left(\begin{array}{c|c} A_{n-1} & a \\ \hline a^T & a_{nn} \end{array}\right) \cdot \left(\begin{array}{c|c} x' \\ \hline 0 \end{array}\right) > 0$ pentru orice $x' \in \mathbb{R}^{n-1} \setminus \{\theta_{\mathbb{R}^{n-1}}\}$. Făcând înmulţirea matricială, care se reduce la înmulţirea matricială cu blocuri de matrici compatibile din punct de vedere al dimensiunilor, se obţine că $x'^T \cdot A_{n-1} \cdot x' > 0$ pentru orice $x' \in \mathbb{R}^{n-1} \setminus \{\theta_{\mathbb{R}^{n-1}}\}$. Prin urmare A_{n-1} este o matrice pozitiv definită. Astfel ajungem la următoarea concluzie: dacă matricea A_n este simetrică şi pozitiv definită rezultă că A_{n-1} este tot o matrice simetrică şi pozitiv definită. Prin urmare conform presupunerii inductive pentru matricea A_{n-1} putem aplica descompunerea LL^T, adică există o matrice L_{n-1}, de ordinul $n-1$, inferior triunghiulară, cu elemente numere reale astfel ca $A_{n-1} = L_{n-1} \cdot L_{n-1}^T$. Pornind de la matricea L_{n-1} astfel obţinută construim o matrice L_n, de ordinul n, inferior triunghiulară, în felul următor: $L_n = \left(\begin{array}{c|c} L_{n-1} & \theta_{\mathbb{R}^{n-1}} \\ \hline l^T & l_{nn} \end{array}\right)$, unde $\theta_{\mathbb{R}^{n-1}} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbb{R}^{n-1}$, $l = \begin{pmatrix} l_{1n} \\ l_{2n} \\ \vdots \\ l_{n-1,n} \end{pmatrix} \in \mathbb{R}^{n-1}$, deci $l^T = (l_{1n}\ l_{2n} \ldots l_{n-1,n})$, iar $l_{nn} \in \mathbb{R}$. Elementele $l_{1n}, l_{2n}, \ldots, l_{n-1,n}, l_{nn}$ deocamdată sunt necunoscute şi urmează să le determinăm din condiţia $L_n \cdot L_n^T = A_n$:

$$\left(\begin{array}{c|c} L_{n-1} & \theta_{\mathbb{R}^{n-1}} \\ \hline l^T & l_{nn} \end{array}\right) \cdot \left(\begin{array}{c|c} L_{n-1}^T & l \\ \hline \theta_{\mathbb{R}^{n-1}}^T & l_{nn} \end{array}\right) = A_n = \left(\begin{array}{c|c} A_{n-1} & a \\ \hline a^T & a_{nn} \end{array}\right).$$

Efectuând înmulţirea cu blocurile de matrici compatibile rezultă:

$$\left(\begin{array}{c|c} L_{n-1} \cdot L_{n-1}^T & L_{n-1} \cdot l \\ \hline l^T \cdot L_{n-1}^T & l^T \cdot l + l_{nn}^2 \end{array}\right) = \left(\begin{array}{c|c} A_{n-1} & a \\ \hline a^T & a_{nn} \end{array}\right).$$

Prin identificarea blocurilor rezultă că: $A_{n-1} = L_{n-1} \cdot L_{n-1}^T$, egalitate care are loc conform presupunerii pasului inductiv, $L_{n-1} \cdot l = a$, care este un sistem inferior triunghiular şi se rezolvă conform §6.1.1.3 ($\det(A_{n-1}) > 0$ implică $\det(L_{n-1}) > 0$). Deci determinăm elementele lui l. Însă $l^T \cdot L_{n-1}^T = a^T$ nu este altceva decât transpunerea sistemului $L_{n-1} \cdot l = a$. Ne rămâne determinarea lui $l_{nn} \in \mathbb{R}$. Din egalitatea $A_n = L_n L_n^T$ rezultă $\det(A_n) = \det(L_n L_n^T) = \det(L_n) \cdot \det(L_n^T) = \det(L_n) \cdot \det(L_n) = [\det(L_n)]^2$. Dacă determinantul

matricii L_n dezvoltăm după ultima coloană avem: $\det(L_n) = \det(L_{n-1}) \cdot l_{nn}$. Prin urmare $\det(A_n) = [\det(L_{n-1})]^2 \cdot l_{nn}^2$. Însă matricea A_n este pozitiv definită, deci teorema lui Sylvester ne asigură că $\det(A_n) > 0$, deci există $l_{nn} \in \mathbb{R}$ număr real pozitiv astfel încât

$$l_{nn} = \frac{\sqrt{\det(A_n)}}{|\det(L_{n-1})|} = \frac{\sqrt{\det(A_n)}}{\det(L_{n-1})} > 0,$$

căci $\det(L_{n-1}) > 0$. Observăm că dacă elementele l_{ii}, pentru $i = \overline{1, n-1}$ se fixează prin $l_{ii} > 0$ atunci şi elementul $l_{nn} > 0$.

Observaţia 6.1.4.2. *Dacă pentru o matrice A simetrică şi pozitiv definită am efectuat descompunerea LL^T atunci determinantul matricii A se calculează foarte simplu:* $\det(A) = \det(L \cdot L^T) = \det(L) \cdot \det(L^T) = \det(L) \cdot \det(L) = [\det(L)]^2 = (l_{11}\ l_{22} \dots l_{nn})^2$, *unde $l_{ii} > 0$ sunt elementele strict pozitive de pe diagonala principală a lui L.*

Program factorizare LL^T

Datele de intrare: n; A; (conform presupunerii A este o matrice simetrică şi pozitiv definită)

Pentru $i = \overline{1, n}$ execută:

$$L[i,i] := \left(A[i,i] - \sum_{j=1}^{i-1} (L[i,j])^2 \right)^{\frac{1}{2}}$$

(unde linia i din matricea L se înmulţeşte cu coloana i din matricea L^T, care este identică cu linia i din matricea L)

Pentru $k = \overline{i+1, n}$ execută

$$L[k,i] := \left(A[k,i] - \textstyle\sum_{j=1}^{i-1} L[k,j] * L[i,j] \right) / L[i,i];$$

(unde linia k din matricea L se înmulţeşte cu coloana i din matricea L^T, care este identică cu linia i din matricea L);

Tipăreşte L.

Dacă avem de rezolvat numeric un sistem liniar de forma $A \cdot x = b$, unde A este o matrice simetrică şi pozitiv definită atunci avem următorul program:

Program 2 LL^T

Datele de intrare: n; A; b;

Aplică subrutina program factorizare LL^T;

Aplică subrutina matrice inferior triunghiulară (conform §6.1.1.3) pentru sistemul $Ly = b$;

Aplică subrutina matrice superior triunghiulară (conform §6.1.1.2) pentru sistemul $L^T x = y$;

Tipăreşte x.

6.1.4.3 Metoda descompunerii QR (Householder)

Pentru a formula problema matematică a descompunerii QR, în prealabil trebuie să fixăm nişte noţiuni. Reamintim că o matrice R se numeşte superior triunghiulară, dacă toate elementele matricii R aflate sub diagonala principală sunt egale cu zero (vezi §6.1.1.3).

Definiţia 6.1.4.3. *O matrice Q se numeşte matrice ortogonală, dacă $Q^T \cdot Q = D$, unde Q^T înseamnă matricea transpusă obţinută din matricea Q, iar D este o matrice diagonală inversabilă.*

Prin urmare, dacă $Q = (q_{ij})_{i,j=\overline{1,n}}$, atunci condiţia $Q^T Q = D$ înseamnă că liniile lui Q^T se înmulţesc cu coloanele lui Q, adică coloanele lui Q se înmulţesc cu coloanele lui Q. Dacă alegem două coloane diferite ale lui Q produsul lor scalar trebuie să fie zero, iar dacă alegem aceeaşi coloană produsul scalar al coloanei respective cu ea însăşi trebuie să fie diferit de zero:

$$\sum_{i=1}^{n} q_{ij} \cdot q_{ik} = \begin{cases} 0, & \text{dacă } j \neq k, \\ d_{jj} \neq 0, & \text{dacă } k = j. \end{cases}$$

Un exemplu simplu de matrice ortogonală pentru $n = 2$ este matricea

$$Q = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}$$

cu $\alpha \in \mathbb{R}$, iar în cazul $n \geq 3$ putem alege de exemplu

$$Q = \begin{pmatrix} \cos\alpha & -\sin\alpha & & & 0 \\ \sin\alpha & \cos\alpha & & & \\ & & 1 & & \\ & & & \ddots & \\ 0 & & & & 1 \end{pmatrix}$$

unde apar $n-2$ bucăţi de 1 pe diagonala principală.

Prin descompunerea QR a matricii A vom înţelege determinarea unei matrici ortogonale Q şi determinarea unei matrici superior triunghiulare R, având un şir de numere reale unu pe diagonala principală, astfel ca să aibă loc egalitatea: $A = Q \cdot R$.

Teorema 6.1.4.3. *Dacă matricea A este nesingulară* ($\det(A) \neq 0$) *atunci există în mod unic descompunerea QR a matricii A.*

Demonstraţie. Fie $A = (a_{ij})_{i,j=\overline{1,n}} = (a_1\ a_2\ \dots\ a_n)$, unde cu a_i am notat coloana i a matricii A pentru $i = \overline{1,n}$, $Q = (q_{ij})_{i,j=\overline{1,n}} = (q_1\ q_2 \dots\ q_n)$, unde cu q_i am notat coloana i a matricii Q pentru $i = \overline{1,n}$, iar $R = \begin{pmatrix} 1 & r_{12} & r_{13} & \dots & r_{1n} \\ 0 & 1 & r_{23} & \dots & r_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & & 1 \end{pmatrix}$, unde $r_{ij} \in \mathbb{R}$ pentru $n \geq j > i \geq 1$. Am fixat $r_{ii} = 1$ pentru orice $i = \overline{1,n}$. Are loc egalitatea: $A = QR$, adică

$$(a_1\ a_2\ \dots\ a_n) = (q_1\ q_2\ \dots\ q_n) \cdot \begin{pmatrix} 1 & r_{12} & r_{13} & \dots & r_{1n} \\ 0 & 1 & r_{23} & \dots & r_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & & 1 \end{pmatrix}.$$

La primul pas toate liniile lui Q se înmulţesc cu prima coloană a lui R şi se obţine prima coloană a lui A : $q_1 \cdot 1 = a_1$, adică $q_1 = a_1$. Deoarece $\det(A) \neq 0$ rezultă că $a_1 \neq \theta_{\mathbb{R}^n}$, deci $q_1 \neq \theta_{\mathbb{R}^n}$. La pasul doi toate liniile lui Q se înmulţesc cu a doua coloană a lui R şi se obţine a doua coloană din A : $q_1 \cdot r_{12} + q_2 \cdot 1 = a_2$, adică $r_{12}q_1 + q_2 = a_2$. În această egalitate avem două necunoscute: numărul real r_{12} şi coloana q_2. Considerând pe $\mathbb{R}^n$ produsul scalar euclidian obişnuit vom impune condiţia $(q_2, q_1) = 0$ ca să formăm matricea Q ca pe o matrice ortogonală. Observăm că pentru determinarea pe rând a coloanelor matricii

Q vom folosi o metodă generală Gram-Schmidt. Cum $q_2 = a_2 - r_{12}q_1$ din condiţia de ortogonalitate $(q_2, q_1) = 0$ rezultă $(a_2 - r_{12}q_1, q_1) = 0$, adică $(a_2, q_1) - r_{12}(q_1, q_1) = 0$, deci $r_{12} = \dfrac{(a_2, q_1)}{\|q_1\|^2}$. Deoarece $q_1 \neq \theta_{\mathbb{R}^n}$ avem asigurată existenţa elementului r_{12}. Ştiind pe r_{12}, îl înlocuim în egalitatea $q_2 = a_2 - r_{12}q_1$ şi obţinem pe q_2. Menţionăm că $q_2 \neq \theta_{\mathbb{R}^n}$, căci în caz contrar avem $q_2 = \theta_{\mathbb{R}^n} = a_2 - r_{12}q_1$, adică $a_2 = r_{12}q_1 = r_{12}a_1$, deci a doua coloană a matricii A se obţine din prima coloană a matricii A printr-o înmulţire cu un factor real, deci $\det(A) = 0$, ceea ce înseamnă o contradicţie.

La pasul trei liniile lui Q se înmulţesc cu a treia coloană a lui R, deci avem: $q_1 \cdot r_{13} + q_2 \cdot r_{23} + q_3 \cdot 1 = a_3$, deci $q_3 = a_3 - r_{13}q_1 - r_{23}q_2$. Impunem condiţia de ortogonalitate a lui q_3 cu q_1 şi cu q_2 : $(q_3, q_1) = 0$ şi $(q_3, q_2) = 0$, deci avem: $(a_3 - r_{13}q_1 - r_{23}q_2, q_1) = 0$, adică $(a_3, q_1) - r_{13}(q_1, q_1) - r_{23}(q_2, q_1) = 0$, deci $(a_3, q_1) - r_{13}(q_1, q_1) = 0$, prin urmare:

$$r_{13} = \frac{(a_3, q_1)}{(q_1, q_1)} = \frac{(a_3, q_1)}{\|q_1\|^2}.$$

Analog avem şirul de relaţii echivalente:

$$\begin{aligned}
&(q_3, q_2) = 0 \Leftrightarrow (a_3 - r_{13}q_1 - r_{23}q_2, q_2) = 0 \Leftrightarrow \\
&\Leftrightarrow (a_3, q_2) - r_{13}(q_1, q_2) - r_{23}(q_2, q_2) = 0 \Leftrightarrow \\
&\Leftrightarrow (a_3, q_2) - r_{23}(q_2, q_2) = 0 \Leftrightarrow r_{23} = \frac{(a_3, q_2)}{(q_2, q_2)} = \frac{(a_3, q_2)}{\|q_2\|^2}.
\end{aligned}$$

Cum $q_1 \neq \theta_{\mathbb{R}^n}$ şi $q_2 \neq \theta_{\mathbb{R}^n}$ avem asigurată existenţa elementelor r_{13} şi r_{23}. Ştiind pe r_{13} şi r_{23} avem

$$q_3 = a_3 - \frac{(a_3, q_1)}{\|q_1\|^2} \cdot q_1 - \frac{(a_3, q_2)}{\|q_2\|^2} \cdot q_2.$$

În mod analog arătăm că $q_3 \neq \theta_{\mathbb{R}^n}$, fiindcă în caz contrar, dacă $q_3 = \theta_{\mathbb{R}^n}$ avem $a_3 = r_{13} \cdot q_1 + r_{23}q_2$, de unde deducem că a_3 depinde liniar de a_1 şi a_2, adică $\det(A) = 0$, ceea ce înseamnă o contradicţie.

În general liniile lui Q se înmulţesc cu coloana k din matricea R: $a_k = q_1 \cdot r_{1k} + q_2 \cdot r_{2k} + \cdots + q_{k-1} \cdot r_{k-1,k} + q_k \cdot 1$, adică $a_k = r_{1k}q_1 + r_{2k}q_2 + \cdots + r_{k-1,k} \cdot q_{k-1} + q_k$. De aici: $q_k = a_k - r_{1k}q_1 - r_{2k}q_2 - \cdots - r_{k-1,k} \cdot q_{k-1}$ şi impunem condiţiile de ortogonalitate asupra lui Q : $(q_k, q_j) = 0$ pentru orice $j = \overline{1, k-1}$. Prin urmare:

$$\begin{aligned}
&(a_k - r_{1k}q_1 - r_{2k}q_2 - \ldots - r_{k-1,k} \cdot q_{k-1}, q_j) = 0 \Leftrightarrow \\
&\Leftrightarrow (a_k, q_j) - r_{1k}(q_1, q_j) - r_{2k}(q_2, q_j) - \ldots - r_{jk}(q_j, q_j) - \ldots - r_{k-1,k}(q_{k-1}, q_j) = 0 \Leftrightarrow \\
&\Leftrightarrow (a_k, q_j) - r_{jk} \cdot \|q_j\|^2 = 0 \Leftrightarrow r_{jk} = \frac{(a_k, q_j)}{\|q_j\|^2}
\end{aligned}$$

pentru fiecare $j = \overline{1, k-1}$. În final

$$q_k = a_k - \frac{(a_k, q_1)}{\|q_1\|^2} \cdot q_1 - \frac{(a_k, q_2)}{\|q_2\|^2} \cdot q_2 - \ldots - \frac{(a_k, q_{k-1})}{\|q_{k-1}\|^2} \cdot q_{k-1}.$$

Program factorizare QR

Datele de intrare: n; A;

$q_1 := a_1$;

Pentru $k := \overline{2, n}$ execută

Pentru $j := \overline{1, k-1}$ execută

$r_{jk} := \dfrac{(a_k, q_j)}{\|q_j\|^2}$;

$q_k := a_k - \sum_{j=1}^{k-1} r_{jk} q_j$;

Tipăreşte R şi Q;

Dacă se pune problema rezolvării numerice a sistemului de ecuaţii liniare $A \cdot x = b$ folosind metoda descompunerii QR, unde $A = (a_{ij})_{i,j=\overline{1,n}}$ şi $b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} \in \mathbb{R}^n$ sunt date, $x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n$ fiind necunoscuta, atunci procedăm în felul următor: $Ax = b \Leftrightarrow QRx = b \Leftrightarrow Q^TQRx = Q^Tb \Leftrightarrow DRx = Q^Tb \Leftrightarrow D^{-1}DRx = D^{-1}Q^Tb \Leftrightarrow Rx = D^{-1}Q^Tb$. Prin urmare rezolvarea sistemului $Ax = b$ este echivalent cu rezolvarea sistemului superior triunghiular $Rx = D^{-1}Q^Tb$. (vezi paragraful §6.1.1.2).

Program 2 QR

Datele de intrare: n; A; b

Aplică subrutina program factorizare QR;

Aplică subrutina matrice superior triunghiulară (conform §6.1.1.2) pentru sistemul:

$Rx = D^{-1} \cdot Q^T \cdot b$;

Tipăreşte x.

6.1.4.4 Sisteme de ecuaţii liniare cu matrice tridiagonală

Definiţia 6.1.4.4. *Printr-o matrice tridiagonală înţelegem o matrice care are elemente nenule pe diagonala principală şi pe cele două codiagonale imediat vecine acesteia. Prin urmare avem o matrice tridiagonală, dacă are următoarea formă:*

$$A = \begin{pmatrix} b_1 & c_1 & & & & & \\ a_2 & b_2 & c_2 & & 0 & & \\ & \ddots & \ddots & \ddots & & & \\ & & a_i & b_i & c_i & & \\ & 0 & & \ddots & \ddots & \ddots & \\ & & & & a_{n-1} & b_{n-1} & c_{n-1} \\ & & & & & a_n & b_n \end{pmatrix}$$

Vrem să aplicăm pentru o matrice tridiagonală A metoda descompunerii LU (vezi paragraful §6.1.4.1) numai că în acest caz pe diagonala matricii superior triunghiulare U vom fixa un şir de unu. Alegem:

$$L = \begin{pmatrix} t_1 & & & & & \\ p_2 & t_2 & & & 0 & \\ & \ddots & \ddots & \ddots & & \\ & & p_i & t_i & & \\ 0 & & & \ddots & \ddots & \\ & & & & p_n & t_n \end{pmatrix} \quad \text{şi} \quad U = \begin{pmatrix} 1 & s_1 & & & & & \\ & 1 & s_2 & & & & \\ & & \ddots & \ddots & & 0 & \\ & & & 1 & s_i & & \\ & & & & \ddots & \ddots & \\ & 0 & & & & 1 & s_{n-1} \\ & & & & & & 1 \end{pmatrix}$$

şi din egalitatea $A = L \cdot U$ prin identificarea elementelor corespunzătoare vom obţine:
Pentru prima linie din L avem:

1. dacă se înmulţeşte prima linie din L cu prima coloană din U atunci: $t_1 \cdot 1 = b_1$;
2. dacă se înmulţeşte prima linie din L cu a doua coloană din U atunci: $t_1 \cdot s_1 + 0 \cdot 1 = c_1$;

   Pentru liniile $i = \overline{2, n-1}$ din L avem:

$$\begin{matrix} & & & \text{coloanele } i-1 & & i & & & \\ L = & & & & \vdots & \vdots & & & \\ & 0 & \dots & 0 & p_i & t_i & 0 & \dots & 0 \\ & & & & \vdots & \vdots & & & \end{matrix} \quad \text{linia } i$$

$$U = \begin{array}{c} \text{coloanele } i-1 \quad i \quad i+1 \\ \left(\begin{array}{ccc} \vdots & \vdots & \vdots \\ 1 & s_{i-1} & \vdots \\ & 1 & s_i \\ & & 1 \quad s_{i+1} \\ & & \vdots \quad \vdots \end{array}\right) \end{array} \begin{array}{l} \text{liniile} \\ i-1 \\ i \\ i+1 \end{array}$$

3. dacă se înmulţeşte linia i din L cu coloana $i-1$ din U atunci: $p_i \cdot 1 + t_i \cdot 0 = a_i$;

4. dacă se înmulţeşte linia i din L cu coloana i din U atunci: $p_i \cdot s_{i-1} + t_i \cdot 1 = b_i$;

5. dacă se înmulţeşte linia i din L cu coloana $i+1$ din U atunci: $p_i \cdot 0 + t_i \cdot s_i = c_i$;

   Pentru ultima linie din L avem:

6. dacă se înmulţeşte ultima linie a lui L cu penultima coloană a lui U atunci: $p_n \cdot 1 + t_n \cdot 0 = a_n$;

7. dacă se înmulţeşte ultima linie a lui L cu ultima coloană a lui U atunci: $p_n \cdot s_{n-1} + t_n \cdot 1 = b_n$.

Prin urmare rezultă egalităţile:

$t_1 = b_1$; $t_1 \cdot s_1 = c_1$;

pentru $i = \overline{2, n-1}$ avem $p_i = a_i$;

$p_i s_{i-1} + t_i = b_i$;

$t_i s_i = c_i$;

$p_n = a_n$;

$p_n \cdot s_{n-1} + t_n = b_n$.

Astfel deducem următoarele formule pentru calculul elementelor necunoscute ale matricilor L şi U : Pentru $i = \overline{2, n}$ avem $p_i := a_i$. Mai departe $t_1 := b_1$ şi $s_1 := \frac{c_1}{t_1}$, dacă $t_1 \neq 0$. Pentru $i = \overline{2, n-1}$ avem $t_i := b_i - p_i s_{i-1}$ şi $s_i := \frac{c_i}{t_i}$, dacă $t_i \neq 0$. În final: $t_n := b_n - p_n \cdot s_{n-1}$.

Teorema 6.1.4.4. *Dacă matricea tridiagonală A este nesingulară* ($\det(A) \neq 0$), *atunci există descompunerea particulară LU de mai sus.*

DEMONSTRAŢIE. Din egalitatea: $\det(A) = \det(L \cdot U) = \det(L) \cdot \det(U) = t_1 \cdot t_2 \cdot \ldots \cdot t_n \cdot 1 \cdot 1 \cdot \ldots \cdot 1 = t_1 \cdot t_2 \cdot \ldots \cdot t_n$ avem $\det(A) \neq 0$ dacă şi numai dacă $t_i \neq 0$ pentru orice $i = \overline{1,n}$. Dar aceste condiţii ne asigură existenţa formulelor pentru calculul lui s_i, $\left(s_i = \dfrac{c_i}{t_i}, i = \overline{1,n-1}\right)$.

Program matrice tridiagonală

Datele de intrare: A, adică a_i pentru $i = \overline{2,n}$;

b_i pentru $i = \overline{1,n}$;

c_i pentru $i = \overline{1,n-1}$;

Pentru $i = \overline{2,n}$ execută: $p_i = a_i$;

$t_1 := b_1$; $s_1 := \dfrac{c_1}{t_1}$;

Pentru $i = \overline{2,n-1}$ execută

$t_i := b_i - p_i s_{i-1}$;

$s_i := \dfrac{c_i}{t_i}$;

$t_n := b_n - p_n \cdot s_{n-1}$;

Tipăreşte: L şi U (adică p_i pentru $i = \overline{2,n}$;

t_i pentru $i = \overline{1,n}$;

s_i pentru $i = \overline{1,n-1}$).

În continuare vrem să rezolvăm sistemul liniar $Ax = d$, unde A este o matrice tridiagonală de ordinul n cu $\det(A) \neq 0$, $d = \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_n \end{pmatrix} \in \mathbb{R}^n$ este vectorul dat al termenului liber iar $x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n$ este coloana necunoscutelor. Rezolvarea numerică a sistemului $A \cdot x = d$ vom face prin metoda descompunerii particulare LU a matricii tridiagonale A.

Prin urmare: $A \cdot x = d \Leftrightarrow LUx = d \Leftrightarrow \begin{cases} Ly = d \\ Ux = y \end{cases}$. Avem de rezolvat $Ly = d$, adică

$$\begin{pmatrix} t_1 & & & & & \\ p_2 & t_2 & & 0 & & \\ & \ddots & \ddots & & & \\ & & p_i & t_i & & \\ 0 & & & \ddots & \ddots & \\ & & & & p_n & t_n \end{pmatrix} \cdot \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_i \\ \vdots \\ d_n \end{pmatrix}$$

deci

$$\begin{cases} t_1 y_1 = d_1 \\ p_2 y_1 + t_2 y_2 = d_2 \\ \quad \vdots \\ p_i y_{i-1} + t_i y_i = d_i \\ \quad \vdots \\ p_n y_{n-1} + t_n y_n = d_n \end{cases}$$

de unde: $y_1 = \dfrac{d_1}{t_1}$; $y_2 = \dfrac{d_2 - p_2 y_1}{t_2}$; $\ldots$; $y_i = \dfrac{d_i - p_i y_{i-1}}{t_i}$; $\ldots$; $y_n = \dfrac{d_n - p_n y_{n-1}}{t_n}$, adică: $y_1 = \dfrac{d_1}{t_1}$ şi pentru $i = \overline{2, n}$ avem:

$$y_i := \frac{d_i - p_i y_{i-1}}{t_i}. \tag{6.1}$$

Considerăm sistemul: $Ux = y$, adică

$$\begin{pmatrix} 1 & s_1 & & & & & \\ & 1 & s_2 & & & & \\ & & \ddots & \ddots & & 0 & \\ & & & 1 & s_i & & \\ 0 & & & & \ddots & \ddots & \\ & & & & & 1 & s_{n-1} \\ & & & & & & 1 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_i \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix}$$

deci:

$$\begin{cases} 1 \cdot x_1 + s_1 x_2 = y_1 \\ 1 \cdot x_2 + s_2 x_3 = y_2 \\ \quad \vdots \\ 1 \cdot x_i + s_i x_{i+1} = y_i \\ \quad \vdots \\ 1 \cdot x_{n-1} + s_{n-1} \cdot x_n = y_{n-1} \\ 1 \cdot x_n = y_n, \end{cases}$$

adică:

$$\begin{cases} x_n = y_n \\ x_{n-1} + s_{n-1} \cdot x_n = y_{n-1} \\ \quad \vdots \\ x_i + s_i x_{i+1} = y_i \\ \quad \vdots \\ x_2 + s_2 x_3 = y_2 \\ x_1 + s_1 x_2 = y_1, \end{cases}$$

care are soluţia, conform unui sistem inferior triunghiular:

$$\begin{cases} x_n = y_n \\ x_{n-1} = y_{n-1} - s_{n-1} \cdot x_n \\ \quad \vdots \\ x_i = y_i - s_i \cdot x_{i+1} \\ \quad \vdots \\ x_2 = y_2 - s_2 x_3 \\ x_1 = y_1 - s_1 x_2, \end{cases}$$

deci: $x_n = y_n$ şi pentru $i = \overline{n-1, 1}$ avem formulele

$$x_i := y_i - s_i \cdot x_{i+1}. \tag{6.2}$$

Combinând metoda descrisă la Program matrice tridiagonală cu formulele (6.1) şi (6.2) obţinem:

Program 2 matrice tridiagonală

Datele de intrare: A, adică a_i pentru $i = \overline{2,n}$;

b_i pentru $i = \overline{1,n}$;

c_i pentru $i = \overline{1,n-1}$;

şi d, adică d_i pentru $i = \overline{1,n}$.

$t_1 := b_1$; $s_1 := \dfrac{c_1}{t_1}$; $y_1 := \dfrac{d_1}{t_1}$;

Pentru $i = \overline{2,n-1}$ execută:

$t_i := b_i - a_i s_{i-1}$;

$s_i := \dfrac{c_i}{t_i}$;

$y_i := \dfrac{d_i - a_i y_{i-1}}{t_i}$;

$t_n := b_n - a_n \cdot s_{n-1}$; $x_n := \dfrac{d_n - a_n \cdot y_{n-1}}{t_n}$;

Pentru $i = \overline{n-1,1}$ execută:

$x_i := y_i - s_i \cdot x_{i+1}$.

Tipăreşte: x, adică x_i pentru $i = \overline{1,n}$.

6.2 Norme vectoriale şi matriciale

Considerăm spaţiul $\mathbb{R}^n$ şi orice normă definită pe acest spaţiu o vom denumi normă vectorială. Să reamintim axiomele normei: norma $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ este o funcţie care verifică axiomele:

1. $\|x+y\| \leq \|x\| + \|y\|$ pentru orice $x, y \in \mathbb{R}^n$;
2. $\|\alpha x\| = |\alpha| \cdot \|x\|$ pentru orice $\alpha \in \mathbb{R}^n$ şi orice $x \in \mathbb{R}^n$;
3. $\|x\| \geq 0$ pentru orice $x \in \mathbb{R}^n$ şi $\|x\| = 0$ dacă şi numai dacă $x = \theta_{\mathbb{R}^n}$.

În continuare dăm exemple de norme vectoriale:

1. norma euclidiană: $\|x\|_2 = \sqrt{\sum_{i=1}^n x_1^2}$, unde $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$;
2. norma maximum: $\|x\|_\infty = \max\{|x_i| \mid i = \overline{1,n}\}$, unde $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$;
3. norma octaedrică: $\|x\|_1 = \sum_{i=1}^n |x_i|$, unde $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$.

Lăsăm pe seama cititorului să verifice axiomele normei pentru aceste norme particulare. O generalizare naturală a acestor norme ar fi norma dată de formula: $\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}$, unde $p \in [1, +\infty]$ este un număr real fixat sau simbolul $+\infty$, iar $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$. În cazul $p = +\infty$ dăm următoarea interpretare:

$$\|x\|_\infty = \lim_{p \to +\infty} \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p} = \max\{|x_i| \ / \ i = 1, n\}.$$

Menţionăm faptul că pe spaţiul $\mathbb{R}^n$ oricare două norme vectoriale sunt compatibile, adică oricare ar fi o normă vectorială $\|\cdot\|$ pe $\mathbb{R}^n$ există constantele $c_1, c_2 > 0$ astfel încât $c_1 \cdot \|x\|_2 \leq \|x\| \leq c_2 \cdot \|x\|_2$ pentru orice $x \in \mathbb{R}^n$.

Considerăm spaţiul linear al matricelor cu elemente numere reale având m linii şi n coloane, notat cu $\mathcal{M}_{m\times n}(\mathbb{R})$.

O normă, $\|\cdot\| : \mathcal{M}_{m\times n}(\mathbb{R}) \to \mathbb{R}$ definită pe spaţiul linear $\mathcal{M}_{m\times n}(\mathbb{R})$ o vom numi normă matricială, şi se defineşte cu aceleaşi axiome:

1. $\|A + B\| \leq \|A\| + \|B\|$ pentru orice $A, B \in \mathcal{M}_{m\times n}(\mathbb{R})$;
2. $\|\alpha A\| = |\alpha| \cdot \|A\|$ pentru orice $\alpha \in \mathbb{R}$ şi orice $A \in \mathcal{M}_{m\times n}(\mathbb{R})$;
3. $\|A\| \geq 0$ pentru orice $A \in \mathcal{M}_{m\times n}(\mathbb{R})$ şi $\|A\| = 0$ dacă şi numai dacă $A = O_{m\times n}$, unde cu $O_{m\times n}$ am notat matricea nulă cu m linii şi n coloane.

   Un exemplu de normă matricială ar fi:

   $$\|A\| = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2}, \text{ unde } A = (a_{ij})_{\substack{i=\overline{1,m}\\ j=\overline{1,n}}} \in \mathcal{M}_{m\times n}(\mathbb{R}).$$

   Lăsăm pe seama cititorului să verifice axiomele normei pentru acest exemplu.

   Dacă $m = n$ atunci avem matricile pătratice de ordinul n cu elemente reale, notate simplu cu $\mathcal{M}_n(\mathbb{R})$ în loc de $\mathcal{M}_{n\times n}(\mathbb{R})$. În acest caz mai adăugăm şi următoarea axiomă la axiomele normei matriciale:

4. $\|A \cdot B\| \leq \|A\| \cdot \|B\|$ pentru orice $A, B \in \mathcal{M}_n(\mathbb{R})$.

   Lăsăm pe seama cititorului să verifice valabilitatea acestei axiome pentru norma matricială

   $$\|A\| = \sqrt{\sum_{i,j=1}^{n} a_{ij}^2}, \text{ unde } A = (a_{ij})_{i,j=\overline{1,n}}.$$

Definiţia 6.2.1. *O normă matricială pe spaţiul linear $\mathcal{M}_n(\mathbb{R})$ se numeşte compatibilă cu o normă vectorială pe spaţiul $\mathbb{R}^n$ dacă are loc inegalitatea $\|A \cdot x\| \leq \|A\| \cdot \|x\|$ pentru orice matrice $A \in \mathcal{M}_n(\mathbb{R})$ şi orice vector $x \in \mathbb{R}^n$.*

Lăsăm pe seama cititorului să verifice că norma matricială $\|A\| = \sqrt{\sum_{i,j=1}^{n} a_{ij}^2}$, unde $A = (a_{ij})_{i,j=\overline{1,n}} \in \mathcal{M}_n(\mathbb{R})$, este compatibilă cu norma vectorială euclidiană $\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$, unde $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$.

Menţionăm faptul că inegalitatea $\|A \cdot x\| \leq \|A\| \cdot \|x\|$ exprimă continuitatea aplicaţiilor lineare definite pe spaţiul $\mathbb{R}^n$ reprezentate prin matricile A.

Dintre toate normele matriciale compatibile cu o normă vectorială să alegem pe cea mai mică, dată de formula

$$\sup_{\substack{x \in \mathbb{R}^n \\ x \neq \theta_{\mathbb{R}^n}}} \frac{\|Ax\|}{\|x\|} =: \|A\|.$$

Se poate arăta că, astfel într-adevăr se defineşte o normă, ceea ce se verifică la un curs de analiză matematică. Tot acolo se demonstrează că:

$$\|A\| := \sup \left\{ \frac{\|Ax\|}{\|x\|} \ / \ x \neq \theta_{\mathbb{R}^n} \right\} = \sup\{\|Ax\| \ / \ \|x\| \leq 1\}.$$

Această normă matricială vom denumi norma matricială subordonată normei vectoriale. În continuare prezentăm normele matriciale subordonate diferitelor norme vectoriale:

1. pentru norma vectorială euclidiană avem norma matricială subordonată dată de formula: $\|A\|_2 = \sqrt{\lambda_1}$, unde λ_1 este cea mai mare valoare proprie a matricii A^*A, unde A^* este matricea adjunctă a matricii A. Menţionăm că toate valorile proprii ale lui A^*A sunt pozitive, iar matricea adjunctă A^* se obţine din matricea A printr-o transpunere în cazul real.

2. pentru norma vectorială maximum avem norma matricială linie dată de formula:

$$\|A\|_\infty = \max_i \sum_{j=1}^{n} |a_{ij}|.$$

3. pentru norma vectorială octaedrică avem norma matricială coloană dată de formula:

$$\|A\|_1 = \max_j \sum_{i=1}^{n} |a_{ij}|.$$

6.3 Perturbaţii

Fie dat sistemul liniar sub forma matricială $A \cdot x = b$, unde $A = (a_{ij})_{i,j=\overline{1,n}} \in \mathcal{M}_n(\mathbb{R})$, $x = (x_1, x_2, \ldots, x_n)^T \in \mathbb{R}^n$, $b = (b_1, b_2, \ldots, b_n)^T \in \mathbb{R}^n$. Presupunem că avem un sistem de tip Cramer cu $\det A \neq 0$. Totodată presupunem că datele de intrare: elementele matricii A şi elementele termenului liber b suferă mici perturbaţii, şi am aştepta ca şi la datele de ieşire x să obţinem perturbaţii de acelaşi ordin. Însă acest lucru nu se adevereşte, şi ne vom exprima sub forma: sistemul liniar este instabil din punct de vedere numeric, dacă numărul de condiţionare $\mathcal{H}(A) = \|A\| \cdot \|A^{-1}\|$ este "mare", de ordinul $10^5, 10^6, \ldots$.

1. la acest subpunct presupunem că termenul liber b suferă mici perturbaţii, iar matricea A rămâne neschimbată. Variaţia termenului liber b, notată cu δb, induce o variaţie a soluţiei x, notată cu δx: $A \cdot (x + \delta x) = b + \delta b$, de unde rezultă $Ax + A\delta x = b + \delta b$, deci $A \cdot \delta x = \delta b$, adică, $\delta x = A^{-1} \cdot \delta b$. Prin urmare: $\|\delta x\| = \|A^{-1} \cdot \delta b\| \leq \|A^{-1}\| \cdot \|\delta b\|$. Aşadar

$$\frac{\|\delta x\|}{\|x\|} = \frac{\|A\| \cdot \|\delta x\|}{\|A\| \cdot \|x\|} \leq \frac{\|A\| \cdot \|A^{-1}\|}{\|A\| \cdot \|x\|} \cdot \|\delta b\|.$$

   Însă $\|A\| \cdot \|x\| \geq \|A \cdot x\| = \|b\|$, deci

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\|A\| \cdot \|A^{-1}\|}{\|b\|} \cdot \|\delta b\| = \|A\| \cdot \|A^{-1}\| \cdot \frac{\|\delta b\|}{\|b\|} = \mathcal{H}(A) \cdot \frac{\|\delta b\|}{\|b\|}.$$

   Prin urmare dacă numărul de condiţionare este "mic" de ordinul $10^{-1}, 10^0, 10^1$, atunci o perturbaţie mică a termenului liber δb, va implica conform inegalităţii anterioare o perturbaţie mică a soluţiei δx.

   Însă, dacă numărul de condiţionare $\mathcal{H}(A) = \|A\| \cdot \|A^{-1}\|$ este mare, de ordinul $10^5, 10^6, \ldots$, atunci se poate întâmpla ca la o variaţie mică a termenului liber δb, de exemplu de ordinul 10^{-2}, să obţinem o variaţie a soluţiei δx de ordinul $10^{-2} \cdot 10^5 = 10^3$. Prin urmare, dacă numărul de condiţionare este "mic" atunci sistemul liniar este stabil din punct de vedere numeric, iar dacă numărul de condiţionare este "mare", atunci sistemul liniar poate să fie instabil din punct de vedere numeric.

2. la acest subpunct presupunem că elementele matricii A suferă mici perturbaţii, notate cu δA, iar termenul liber b rămâne neschimbat. Obţinem o perturbaţie a

soluţiei x notată cu δx : $(A+\delta A)(x+\delta x) = b$, adică $A\cdot x+\delta A\cdot x+A\cdot\delta x+\delta A\cdot\delta x = b$, deci $A \cdot \delta x = -\delta A(x + \delta x)$. Prin urmare: $\delta x = -A^{-1} \cdot \delta A \cdot (x + \delta x)$ de unde:

$$\begin{aligned} \|\delta x\| &= \| - A^{-1} \cdot \delta A \cdot (x + \delta x)\| = \|A^{-1} \cdot \delta A \cdot (x + \delta x)\| \leq \\ &\leq \|A^{-1}\| \cdot \|\delta A\| \cdot \|x + \delta x\|, \end{aligned}$$

adică

$$\frac{\|\delta x\|}{\|x + \delta x\|} \leq \|A^{-1}\| \cdot \|\delta A\| = \|A^{-1}\| \cdot \|A\| \cdot \frac{\|\delta A\|}{\|A\|} = \mathcal{H}(A) \cdot \frac{\|\delta A\|}{\|A\|}.$$

Se observă că şi la acest subpunct putem trage aceleaşi concluzii ca la subpunctul anterior, şi anume: dacă numărul de condiţionare $\mathcal{H}(A)$ este "mic", atunci sistemul este stabil din punct de vedere numeric, iar dacă numărul de condiţionare este "mare", atunci sistemul liniar devine instabil din punct de vedere numeric.

6.4 Metode iterative pentru rezolvarea sistemelor liniare

6.4.1 Metoda lui Jacobi

Se consideră sistemul linear sub forma matricială $A \cdot x = b$, unde $A = (a_{ij})_{i,j=\overline{1,n}}$, $x = (x_1, x_2, \ldots, x_n)^T$ şi $b = (b_1, b_2, \ldots, b_n)^T$. Forma algebrică a sistemului linear este:

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

Dacă presupunem că $a_{ii} \neq 0$ pentru orice $i = \overline{1, n}$, atunci sistemul linear se poate pune sub forma echivalentă, denumită forma iterativă, în felul următor: din prima ecuaţie se scoate necunoscuta x_1, din a doua necunoscuta $x_2, \ldots$, din ultima ecuaţie necunoscuta

x_n, şi se obţine:

$$\begin{cases} x_1 = \dfrac{-a_{12}x_2 - a_{13}x_3 - \cdots - a_{1n}x_n}{a_{11}} + \dfrac{b_1}{a_{11}} \\ \\ x_2 = \dfrac{-a_{21}x_1 - a_{23}x_3 - \cdots - a_{2n}x_n}{a_{22}} + \dfrac{b_2}{a_{22}} \\ \vdots \\ x_n = \dfrac{-a_{n1}x_1 - a_{n2}x_2 - \cdots - a_{n,n-1}x_{n-1}}{a_{nn}} + \dfrac{b_n}{a_{nn}} \end{cases}$$

Transcriem sistemul iterativ sub forma matricială:

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} 0 & -\dfrac{a_{12}}{a_{11}} & \dots & -\dfrac{a_{1n}}{a_{11}} \\ -\dfrac{a_{21}}{a_{22}} & 0 & \dots & -\dfrac{a_{2n}}{a_{22}} \\ \vdots & \vdots & & \vdots \\ -\dfrac{a_{n1}}{a_{nn}} & -\dfrac{a_{n2}}{a_{nn}} & \dots & 0 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} \dfrac{b_1}{a_{11}} \\ \dfrac{b_2}{a_{22}} \\ \vdots \\ \dfrac{b_n}{a_{nn}} \end{pmatrix}.$$

Introducem notaţiile:

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, B_J = \begin{pmatrix} 0 & -\dfrac{a_{12}}{a_{11}} & \dots & -\dfrac{a_{1n}}{a_{11}} \\ -\dfrac{a_{21}}{a_{22}} & 0 & \dots & -\dfrac{a_{2n}}{a_{22}} \\ \vdots & \vdots & & \vdots \\ -\dfrac{a_{n1}}{a_{nn}} & -\dfrac{a_{n2}}{a_{nn}} & \dots & 0 \end{pmatrix} \quad \text{şi} \quad c_J = \begin{pmatrix} \dfrac{b_1}{a_{11}} \\ \dfrac{b_2}{a_{22}} \\ \vdots \\ \dfrac{b_n}{a_{nn}} \end{pmatrix}.$$

Astfel sistemul iniţial apare sub forma matricială iterativă echivalentă: $x = B_J \cdot x + c_J$. Alegem un punct iniţial de pornire $x^0 = (x_1^0, x_2^0, \dots, x_n^0) \in \mathbb{R}^n$ şi începem să iterăm şirul $(x^k)_{k\in\mathbb{N}} \subset \mathbb{R}^n$ în felul următor:

$$x^1 = B_J \cdot x^0 + c_J,\ x^2 = B_J \cdot x^1 + c_J, \dots, x^{k+1} = B_J \cdot x^k + c_J, \dots$$

Iteraţia $x^{k+1} = B_J x^k + c_J$ scrisă pe larg înseamnă:

$$x_i^{k+1} = \frac{-\sum\limits_{\substack{j=1 \\ j\neq i}}^{n} a_{ij}x_j^k}{a_{ii}} + \frac{b_i}{a_{ii}}$$

ceea ce este echivalent cu

$$\sum_{j=1}^{i-1} a_{ij}x_j^k + a_{ii}x_i^{k+1} + \sum_{j=i+1}^{n} a_{ij}x_j^k = b_i$$

pentru orice $i = \overline{1,n}$. Astfel putem determina matricile B_J şi c_J în funcţie de datele iniţiale: matricile A şi b. Într-adevăr, fie $A = L+D+U$ descompunerea matricii A în matricea strict inferior triunghiulară L, matricea diagonală D şi matricea strict superior triunghiulară U. Menţionăm că matricile L, D, U sunt matrici pătratice de ordinul n, unde elementele nealese din matricea A sunt egale cu zero. Prin urmare: $L \cdot x^k + D \cdot x^{k+1} + U \cdot x^k = b$, adică: $x^{k+1} = -D^{-1}(L+U) \cdot x^k + D^{-1}b$, ceea ce înseamnă că $B_J = -D^{-1}(L+U)$ şi $c_J = D^{-1}b$. Deoarece conform presupunerii $a_{ii} \neq 0$ pentru orice $i = \overline{1,n}$ rezultă că există matricea inversă D^{-1}, căci $\det(D) = \prod_{i=1}^{n} a_{ii} \neq 0$.

Menţionăm următorul rezultat referitor la convergenţa şirului $(x^k)_{k\in\mathbb{N}}$ ca o condiţie suficientă:

Propoziţia 6.4.1. *Dacă matricea A este dominant diagonală, adică $|a_{ii}| > \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|$ pentru orice $i = \overline{1,n}$, atunci şirul iterativ $(x^k)_{k\in\mathbb{N}} \subset \mathbb{R}^n$ generat mai sus va fi convergent pentru orice punct iniţial de pornire $x^0 \in \mathbb{R}^n$.*

Demonstraţie. Să alegem pe $\mathbb{R}^n$ norma vectorială maximum: $\|x\|_\infty = \max\{|x_i| \ / \ i = \overline{1,n}\}$ şi norma matricială subordonată, numită norma linie, dată de formula $\|B_J\|_\infty = \max\limits_{1\leq i\leq n} \sum\limits_{j=1}^{n} |b_{ij}|$, unde $B = (b_{ij})_{i,j=\overline{1,n}}$. Pentru orice $i = \overline{1,n}$ condiţia $|a_{ii}| > \sum\limits_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|$ este echivalentă cu $\sum\limits_{\substack{j=1 \\ j\neq i}}^{n} \left|-\dfrac{a_{ij}}{a_{ii}}\right| < 1$, adică suma elementelor în modul în linia i din matricea B_J este strict mai mică decât unu. Prin urmare

$$\|B_J\|_\infty = \max_{1\leq i\leq n} \sum_{\substack{j=1 \\ j\neq i}}^{n} |b_{ij}| = \max_{1\leq i\leq n} \sum_{\substack{j=1 \\ j\neq i}}^{n} \left|-\frac{a_{ij}}{a_{ii}}\right| < 1.$$

Astfel, dacă notăm cu $\Phi : \mathbb{R}^n \to \mathbb{R}^n$, $\Phi(x) = B_J x + c_J$, funcţia iterativă prin care generăm şirul iterativ $(x^k)_{k\in\mathbb{N}}$ ($x^{k+1} = \Phi(x^k) = B_J x^k + c_J$), atunci pentru orice $x, y \in \mathbb{R}^n$ avem:

$$\|\Phi(x) - \Phi(y)\|_\infty = \|(B_J x + c_J) - (B_J y + c_J)\|_\infty = \|B_J(x-y)\|_\infty \leq \|B_J\|_\infty \cdot \|x-y\|_\infty.$$

Această condiţie exprimă faptul că Φ este o contracţie pe $\mathbb{R}^n$ având constanta de contracţie $\|B_J\|_\infty < 1$. Suntem gata cu demonstraţia teoremei, fiindcă este de ajuns să

aplicăm teorema de punct fix a lui Banach pentru contracţia Φ pe spaţiul metric complet $\mathbb{R}^n$, unde metrica este generată de către norma vectorială maximum. q.e.d.

Presupunem că şirul iterativ $(x^k)_{k\in\mathbb{N}} \subset \mathbb{R}^n$ este convergent, şi fie $\lim_{k\to\infty} x^k = x^*$. Trecând la limită în recurenţa $x^{k+1} = B_J \cdot x^k + c_J$ pentru $k \to \infty$ obţinem că $x^* = B_J x^* + c_J$, ceea ce este echivalent cu $Ax^* = b$. Prin urmare în acest fel generăm un şir iterativ $(x^k)_{k\in\mathbb{N}}$ care converge către soluţia sistemului iniţial $A \cdot x = b$.

În continuare prezentăm un algoritm pentru metoda lui Jacobi folosind o condiţie practică de oprire:

Program Jacobi

Datele de intrare: $n; A; b; \varepsilon;\ x^0$;

Fie $y := x^0$;

Repetă $\quad x := y$;

$\qquad\qquad y := B_J \cdot x + c_J$;

Până când $\|y - x\| \geq \varepsilon$;

Tipăreşte y.

În acest program ε înseamnă precizia dată dinainte, ($\varepsilon = 10^{-2}, 10^{-3}, \dots$), $x^0 = (x_1^0, x_2^0, \dots, x_n^0) \in \mathbb{R}^n$ este punctul iniţial de pornire, atribuirile $y := x^0$ şi $x := y$ înseamnă atribuiri de matrici coloane termen cu termen. Instrucţiunea $y := B_J \cdot x + c_J$ se referă la iteraţia $x^{k+1} = B_J \cdot x^k + c_J$, care se poate scrie pe larg în felul următor:

pentru $i = \overline{1,n}$ execută: $x_i^{k+1} := \dfrac{b_i - \sum\limits_{\substack{j=1 \\ j\neq i}}^{n} a_{ij} x_j^k}{a_{ii}}$. Pentru vectorul "vechi" x^k folosim vectorul x şi pentru vectorul "nou" x^{k+1} folosim vectorul y :

pentru $i = \overline{1,n}$ execută: $y[i] := \dfrac{b[i] - \sum\limits_{\substack{j=1 \\ j\neq i}}^{n} a[i,j] * x[j]}{a[i,i]}$. Din punct de vedere al programării ultima instrucţiune se poate realiza în felul următor:

pentru $i = \overline{1,n}$ execută:

$\qquad y[i] := 0$;

$\qquad$ pentru $j := \overline{1,n}$ execută:

$\qquad\qquad$ dacă $j \neq i$ atunci $y[i] := y[i] + a[i,j] * x[j]$

$y[i] := (b[i] - y[i])/a[i,i];$

La condiţia $\|y - x\| \geq \varepsilon$ prima dată se calculează într-o subrutină o anumită normă vectorială a vectorului $y - x$:

Fie $s := 0;$

Pentru $i = \overline{1,n}$ execută: $s := s + |y[i] - x[i]|;$

şi în final în variabila s avem valoarea normei $\|y - x\|$.

6.4.2 Metoda lui Gauss-Seidel

Se consideră sistemul linear sub forma matricială $A \cdot x = b$, unde $A = (a_{ij})_{i,j=\overline{1,n}}$, $x = (x_1, x_2, \ldots, x_n)^T$ şi $b = (b_1, b_2, \ldots, b_n)^T$. La fel ca la metoda lui Jacobi şi aici vom transforma sistemul iniţial $A \cdot x = b$ sub forma iterativă echivalentă: $x = B_J \cdot x + c_J$, presupunând că $a_{ii} \neq 0$ pentru orice $i = \overline{1,n}$. În mod analog alegem la întâmplare un punct de pornire $x^0 \in \mathbb{R}^n$, şi iterăm şirul $(x^k)_{k\in\mathbb{N}}$ într-un mod diferit de iteraţia metodei lui Jacobi. Ideea lui Seidel constă în următoarele: la calculul componentelor vectorului $x^{k+1} = (x_1^{k+1}, x_2^{k+1}, \ldots, x_n^{k+1})^T$ pe lângă componentele lui $x^k = (x_1^k, x_2^k, \ldots, x_n^k)^T$ vom folosi componentele noi calculate ale lui x^{k+1} :

$$\begin{cases} x_1^{k+1} = \dfrac{1}{a_{11}} \left(b_1 - \displaystyle\sum_{j=2}^{n} a_{1j} x_j^k \right) \\ \\ x_2^{k+1} = \dfrac{1}{a_{22}} \left(b_2 - a_{21} \cdot x_1^{k+1} - \displaystyle\sum_{j=3}^{n} a_{2j} x_j^k \right) \\ \vdots \\ x_i^{k+1} = \dfrac{1}{a_{ii}} \left(b_i - \displaystyle\sum_{j=1}^{i-1} a_{ij} \cdot x_j^{k+1} - \sum_{j=i+1}^{n} a_{ij} x_j^k \right) \\ \vdots \\ x_n^{k+1} = \dfrac{1}{a_{nn}} \left(b_n - \displaystyle\sum_{j=1}^{n-1} a_{nj} x_j^{k+1} \right). \end{cases} \tag{6.3}$$

Observăm că în prima egalitate din (6.3) x_1^{k+1} depinde de $x_2^k, x_3^k, \ldots, x_n^k$. În a doua egalitate din (6.3) x_2^{k+1} depinde de $x_1^{k+1}, x_3^k, \ldots, x_n^k$, şi dacă înlocuim pe x_1^{k+1} folosind prima egalitate din (6.3) obţinem că în esenţă x_2^{k+1} depinde numai de $x_1^k, x_2^k, \ldots, x_n^k$. În mod analog ne dăm seama că x_i^{k+1} depinde de $x_1^k, x_2^k, \ldots, x_n^k$ pentru fiecare $i = \overline{2,n}$. Să notăm această

dependenţă sub forma matricială: $x^{k+1} = B_{GS}x^k + c_{GS}$, unde B_{GS} este o matrice de ordinul n iar c_{GS} este o matrice coloană cu n elemente. Menţionăm că matricea B_{GS} şi c_{GS} se alege tocmai în aşa fel, încât iteraţia $x^{k+1} = B_{GS}x^k + c_{GS}$ este echivalentă cu relaţiile din (6.3). În continuare determinăm la concret matricile B_{GS} şi c_{GS} folosind datele iniţiale: pe A şi pe b. Dacă în relaţiile (6.3) înmulţim pe rând ecuaţiile cu $a_{ii} \neq 0$ pentru $i = \overline{1,n}$ şi ducem toţi termenii în membrul stâng obţinem următoarea formă echivalentă cu (6.3):

$$\begin{cases} a_{11}x_1^{k+1} + \sum_{j=2}^{n} a_{1j}x_j^k = b_1 \\ a_{21}x_1^{k+1} + a_{22}x_2^{k+1} + \sum_{j=3}^{n} a_{2j}x_j^k = b_2 \\ \vdots \\ \sum_{j=1}^{i} a_{ij}x_j^{k+1} + \sum_{j=i+1}^{n} a_{ij}x_j^k = b_i \\ \vdots \\ \sum_{j=1}^{n} a_{nj}x_j^{k+1} = b_n \end{cases} \tag{6.4}$$

Scriem matricea A sub forma $A = L + D + U$, unde L este parte strict superior inferioară a matricii A, D este diagonala matricii A iar U este partea strict superior triunghiulară a matricii A. Menţionăm că matricile L, D şi U sunt matrici pătratice de ordinul n, ale căror elemente sunt alese dintre elementele corespunzătoare ale matricii A, iar restul elementelor fiind completate cu zerouri. Prin urmare relaţiile (6.4) apar sub forma echivalentă matricială în felul următor: $(L+D)\cdot x^{k+1} + Ux^k = b$. Cum $\det(L+D) = \det(D) = \prod_{i=1}^{n} a_{ii} \neq 0$ obţinem că există matricea inversă $(L+D)^{-1}$, deci $x^{k+1} = -(L+D)^{-1}\cdot U\cdot x^k + (L+D)^{-1}\cdot b$. Prin urmare avem: $B_{GS} = -(L+D)^{-1}\cdot U$ şi $c_{GS} = (L+D)^{-1}\cdot b$. Alegem orice punct de pornire $x^0 = (x_1^0, x_2^0, \ldots, x_n^0) \in \mathbb{R}^n$ şi iterăm şirul $(x^k)_{k\in\mathbb{N}}$ conform relaţiilor (6.3) care este echivalent cu iteraţia:

$$x^1 = B_{GS}x^0 + c_{GS},\ x^2 = B_{GS}x^1 + c_{GS}, \ldots, x^{k+1} = B_{GS}x^k + c_{GS}, \ldots$$

Menţionăm fără demonstraţie următorul rezultat referitor la convergenţa şirului $(x^k)_{k\in\mathbb{N}}$ cu o condiţie suficientă:

Propoziţia 6.4.2. *Dacă matricea A este dominant diagonală, adică $|a_{ii}| > \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|$ pentru orice $i = \overline{1,n}$, atunci şirul iterativ $(x^k)_{k\in\mathbb{N}}$ generat mai sus va fi convergent pentru*

orice punct iniţial de pornire $x^0 \in \mathbb{R}^n$.

Presupunem că şirul iterativ $(x^k)_{k\in\mathbb{N}} \subset \mathbb{R}^n$ este convergent şi fie $\lim_{k\to\infty} x^k = x^*$. Trecând la limită în recurenţa $x^{k+1} = B_{GS}x^k + c_{GS}$ pentru $k \to \infty$ obţinem că $x^* = B_{GS}x^* + c_{GS}$, ceea ce este echivalent cu $Ax^* = b$. Prin urmare în acest fel generăm un şir iterativ $(x^k)_{k\in\mathbb{N}}$ care converge către soluţia sistemului iniţial $Ax = b$.

În continuare prezentăm un algoritm pentru metoda lui Gauss-Seidel folosind o condiţie practică de oprire:

Program Gauss-Seidel

Datele de intrare: $n; A; b; \varepsilon; x^0$;

Fie $y := x^0$;

Repetă $x := y$;

$y := B_{GS}x + c_{GS}$;

Până când $\|y - x\| \geq \varepsilon$;

Tipăreşte y.

În acest program ε înseamnă precizia dinainte dată, $(\varepsilon = 10^{-2}, 10^{-3}, \dots)$, $x^0 = (x_1^0, x_2^0, \dots, x_n^0) \in \mathbb{R}^n$ este punctul iniţial de pornire, atribuirile $y := x^0$ şi $x := y$ înseamnă atribuiri de matrici coloane termen cu termen. Instrucţiunea $y := B_{GS}x + c_{GS}$ se referă la iteraţia $x^{k+1} = B_{GS}x^k + c_{GS}$, care se poate scrie pe larg în felul următor: pentru $i = \overline{1,n}$ execută:

$$x_i^{k+1} := \frac{b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{k+1} - \sum_{j=i+1}^{n} a_{ij}x_j^k}{a_{ii}}.$$

Pentru vectorul "vechi" x^k folosim vectorul x şi pentru vectorul "nou" x^{k+1} folosim vectorul y :

Pentru $i = \overline{1,n}$ execută:

$$y[i] := \frac{b[i] - \sum_{j=1}^{i-1} a[i,j] * y[j] - \sum_{j=i+1}^{n} a[i,j] * x[j]}{a[i,i]}.$$

Din punct de vedere al programării ultima instrucţiune se poate realiza în felul următor:

Fie $s_1 := 0$; $s_2 := 0$;

Pentru $j = \overline{1, i-1}$ execută:

$$s_1 := s_1 + a[i,j] * y[j];$$

Pentru $j = \overline{i+1, n}$ execută:

$$s_2 := s_2 + a[i,j] * x[j];$$

$y[i] := (b[i] - s_1 - s_2)/a[i,i];$

La condiţia $\|y - x\| \geq \varepsilon$ prima dată se calculează într-o subrutină o anumită normă vectorială a vectorului $y - x$:

Fie $s := 0;$

Pentru $i := \overline{1, n}$ execută:

$$s := s + |y[i] - x[i]|;$$

şi în final în variabila s avem valoarea normei $\|y - x\|$.

6.4.3 Teoria generală a metodelor iterative pentru sistemele liniare

Am văzut că la metoda iterativă a lui Jacobi (vezi §6.4.1) şirul iterativ $(x^k)_{k\in\mathbb{N}}$, care converge către soluţia sistemului iniţial $Ax = b$, este generat de formula de recurenţă: $x^{k+1} = B_J x^k + c_J$. La fel şi la metoda iterativă a lui Gauss-Seidel (vezi §6.4.2) şirul iterativ $(x^k)_{k\in\mathbb{N}}$ convergent către soluţia sistemului iniţial $Ax = b$ este generat de formula de recurenţă $x^{k+1} = B_{GS} x^k + c_{GS}$. Observăm că atât metoda lui Jacobi cât şi metoda lui Gauss-Seidel generează şirul iterativ $(x^k)_{k\in\mathbb{N}}$ prin aceeaşi formulă de recurenţă:

$$x^{k+1} = B \cdot x^k + c, \tag{6.5}$$

unde B este o matrice pătrată de ordinul n iar c este o matrice coloană cu n elemente.

Definiţia 6.4.1. *O metodă iterativă, care duce la formarea unui şir de aproximaţii succesive de forma (6.5), se numeşte staţionară (matricile B şi c rămân constante pe parcursul iteraţiilor).*

Astfel atât metoda lui Jacobi cât şi metoda lui Gauss-Seidel sunt metode staţionare.

Teorema 6.4.1. *Dacă norma matricii B, în raport cu o anumită normă matricială, este strict mai mică decât unu, atunci metoda iterativă staţionară (6.5) este convergentă pentru orice punct iniţial de pornire $x^0 \in \mathbb{R}^n$.*

DEMONSTRAŢIE. Notăm cu $\Phi : \mathbb{R}^n \to \mathbb{R}^n$, $\Phi(x) = Bx + c$ funcţia iterativă care generează şirul iterativ $(x^k)_{k\in\mathbb{N}}$ conform (6.5) : $x^{k+1} = \Phi(x^k) = Bx^k + c$. Alegem o normă vectorială pe $\mathbb{R}^n$ şi astfel spaţiul normat $\mathbb{R}^n$ devine un spaţiu metric complet. Fie pe $\mathcal{M}_n(\mathbb{R})$ o normă matricială subordonată sau compatibilă cu norma vectorială dată pe $\mathbb{R}^n$. Arătăm că Φ este o contracţie pe $\mathbb{R}^n$:

$$\|\Phi(x) - \Phi(y)\| = \|(Bx + c) - (By + c)\| = \|B(x - y)\| \leq \|B\| \cdot \|x - y\|$$

pentru orice $x, y \in \mathbb{R}^n$. Însă conform presupunerii $\|B\| < 1$. Teorema de punct fix a lui Banach aplicată pentru contracţia Φ ne dă că pentru orice punct de pornire $x^0 \in \mathbb{R}^n$ şirul iterativ $(x^k)_{k\in\mathbb{N}}$ este convergent şi tinde către singurul punct fix $x^* \in \mathbb{R}^n$ al lui $\Phi : x^* = \Phi(x^*) = Bx^* + c$. q.e.d.

Observaţia 6.4.1. *Deoarece sistemul iniţial $Ax = b$ se pune sub forma iterativă $x = Bx + c$ prin transformări echivalente, rezultă că punctul fix al lui Φ este tocmai soluţia ecuaţiei iniţiale, adică şirul iterativ $(x^k)_{k\in\mathbb{N}}$ dat prin formula de recurenţă $x^{k+1} = \Phi(x^k)$ tinde către $x^* \in \mathbb{R}^n$, punctul fix al lui Φ, care este în acelaşi timp şi soluţia ecuaţiei iniţiale $Ax^* = b$.*

În continuare vrem să prezentăm un rezultat mai puternic, prin care dăm o condiţie necesară şi suficientă ca o metodă staţionară să fie convergentă. Însă în prealabil avem nevoie de nişte cunoştinţe de algebră liniară. Reamintim că pentru o matrice pătratică B, de ordinul n, ecuaţia caracteristică are forma $\det(B - \lambda I_n) = 0$, unde I_n este matricea unitate de ordinul n. Ecuaţia caracteristică este o ecuaţie polinomială de grad n în variabila $\lambda \in \mathbb{R}$. Soluţiile ecuaţiei caracteristice, adică cele n rădăcini reale şi complexe ale ecuaţiei polinomiale de grad n se numesc valorile proprii ale matricii B. Să le notăm cu $\lambda_1, \lambda_2, \ldots, \lambda_n$. Prin raza spectrală a matricii B, notată cu $r(B)$ se înţelege: $r(B) = \max\{|\lambda_i| \ / \ i = \overline{1, n}\}$, adică raza celui mai mic disc din planul complex, centrat în originea planului complex şi care conţine, acoperă toate valorile proprii ale matricii B.

Teorema 6.4.2. *Condiţia necesară şi suficientă ca o metodă staţionară să fie convergentă este ca raza spectrală a matricii B să fie strict mai mic decât unu.*

DEMONSTRAŢIE. Matricea B putem să reducem la forma canonică Jordan, adică va exista o matrice de schimbare a bazei C astfel încât $C^{-1}BC = A$, unde A este forma canonică Jordan. Noi în continuare vom da demonstraţia într-un caz particular şi anume când forma canonică Jordan A este o matrice diagonală. Se ştie din algebra liniară că acest caz are loc, când multiplicitatea algebrică a valorii proprii coincide cu multiplicitatea geometrică a vectorului propriu corespunzător valorii proprii. Avem un caz particular important, când ecuaţia caracteristică are toate rădăcinile simple. Menţionăm că demonstraţia teoremei în cazul general, pentru blocurile Jordan se face în mod asemănător. Fie deci forma lui A egală cu:

$$A = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & 0 \\ & & \ddots & \\ & 0 & & \lambda_n \end{pmatrix}$$

unde pe diagonală apar toate valorile proprii ale matricii B cu multiplicităţile corespunzătoare. Din $C^{-1}BC = A$ deducem că $B = CAC^{-1}$, deci $B^2 = B \cdot B = (CAC^{-1})(CAC^{-1}) = CA^2C^{-1}$ căci $C^{-1}C = I_n$. Mai departe putem arăta simplu prin metoda inducţiei matematice că $B^k = C \cdot A^k \cdot C^{-1}$ pentru orice $k \geq 1$ număr natural. Fie $x^0 \in \mathbb{R}^n$ un punct de pornire. Atunci $x^1 = Bx^0 + c$ şi $x^2 = Bx^1 + c = B(Bx^0 + c) + c = B^2x^0 + Bc + c = B^2x^0 + (B + I_n) \cdot c$. Cum $x^3 = Bx^2 + c$ rezultă că $x^3 = B \cdot [B^2x^0 + (B + I_n) \cdot c] + c = B^3x^0 + B(B + I_n) \cdot c + c = B^3x^0 + (B^2 + B + I_n) \cdot c$. Presupunem că $x^k = B^kx^0 + (B^{k-1} + B^{k-2} + \cdots + B + I_n) \cdot c$ şi deducem că $x^{k+1} = Bx^k + c = B[B^kx^0 + (B^{k-1} + B^{k-2} + \cdots + B + I_n) \cdot c] + c = B^{k+1} \cdot x^0 + (B^k + B^{k-1} + \cdots + B^2 + B + I_n) \cdot c$. Acum în egalitatea $x^k = B^kx^0 + (B^{k-1} + B^{k-2} + \cdots + B + I_n) \cdot c$ înlocuim puterile lui B prin formula $B^k = C \cdot A^k \cdot C^{-1}$ pentru orice $k \geq 1$. Deci $x^k = C \cdot A^k \cdot C^{-1} \cdot x^0 + C(A^{k-1} + A^{k-2} + \cdots + A + I_n) \cdot C^{-1} \cdot c$. Observăm că x^k este convergent pentru orice punct $x^0 \in \mathbb{R}^n$, dacă $A^k \to O_n$ şi $A^{k-1} + A^{k-2} + \cdots + A + I_n$ tinde către o matrice dată. Din $A = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & 0 \\ & & \ddots & \\ & 0 & & \lambda_n \end{pmatrix}$ avem $A^2 = A \cdot A = \begin{pmatrix} \lambda_1^2 & & & \\ & \lambda_2^2 & & 0 \\ & & \ddots & \\ & 0 & & \lambda_n^2 \end{pmatrix}$ şi prin inducţie se arată imediat că $A^k = \begin{pmatrix} \lambda_1^k & & & \\ & \lambda_2^k & & 0 \\ & & \ddots & \\ & 0 & & \lambda_n^k \end{pmatrix}$. Din analiză se ştie că $\lim_{k\to\infty} q^k = 0$ dacă

şi numai dacă $|q| < 1$, adică dacă $|\lambda_1|, |\lambda_2|, \ldots, |\lambda_n| < 1$ de unde rezultă că $A^k \to O_n$. În acelaşi timp

$$A^{k-1} + A^{k-2} + \cdots + A + I_n =$$
$$= \begin{pmatrix} \lambda_1^{k-1} + \lambda_1^{k-2} + \cdots + \lambda_1 + 1 & & & \\ & \lambda_2^{k-1} + \lambda_2^{k-2} + \cdots + \lambda_2 + 1 & & 0 \\ & & \ddots & \\ & 0 & & \lambda_n^{k-1} + \lambda_n^{k-2} + \cdots + \lambda_n + 1 \end{pmatrix}$$

este o serie de matrici convergente, dacă seria de forma $q^{k-1} + q^{k-2} + \cdots + q + 1$ converge, adică există

$$lim_{k\to\infty}(q^{k-1} + q^{k-2} + \cdots + q + 1) = \lim_{k\to\infty} \frac{1 - q^k}{1 - q} = \frac{1}{1 - q}$$

dacă şi numai dacă $|q| < 1$. Şi această condiţie implică faptul că $|\lambda_1|, |\lambda_2|, \ldots, |\lambda_n| < 1$. Prin urmare $r(B) = \max\{|\lambda_i|,\ i = \overline{1, n}\} < 1$. Totodată din demonstraţie rezultă că

$$A^{k-1} + A^{k-2} + \cdots + A + I_n \to \begin{pmatrix} \frac{1}{1-\lambda_1} & & & \\ & \frac{1}{1-\lambda_2} & & 0 \\ & & \ddots & \\ & 0 & & \frac{1}{1-\lambda_n} \end{pmatrix} = (I_n - A)^{-1}.$$

Prin urmare $x^k \to C \cdot O_n \cdot C^{-1} \cdot x^0 + C \cdot (I_n - A)^{-1} \cdot C^{-1} \cdot c = C \cdot (I_n - A)^{-1} \cdot C^{-1} \cdot c$ pentru $k \to \infty$, independent de punctul de pornire x^0. q.e.d.

Observaţia 6.4.2. *Deoarece din teoria spectrală a operatorilor liniari pe spaţii finit dimensionale se ştie că $r(B) \leq \|B\|$ pentru orice normă matricială deducem că teorema 6.4.2 este un rezultat mai puternic decât teorema 6.4.1. Totuşi din punct de vedere numeric recomandăm folosirea teoremei 6.4.1.*

6.4.4 Metoda SOR

Denumirea metodei provine din literatura de specialitate în limba engleză fiind prescurtarea metodei "successiv over relaxation", adică metoda suprarelaxărilor succesive. Este o metodă iterativă de rezolvare a sistemelor liniare de forma $A \cdot x = b$. Ideea metodei

constă în introducerea unui factor de relaxare $\omega \in \mathbb{R}$ în iteraţia de la metoda lui Gauss-Seidel:

$$x_i^{k+1} = x_i^k + \omega \cdot \frac{b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{k+1} - \sum_{j=i}^{n} a_{ij} x_j^k}{a_{ii}},$$

pentru $i = \overline{1,n}$.

Observăm că pentru $\omega = 1$ din iteraţia de tip SOR se reobţine iteraţia de tip Gauss-Seidel. Să arătăm că metoda SOR este tot o metodă iterativă staţionară. Într-adevăr, relaţia iterativă de tip SOR este echivalentă cu:

$$\omega \cdot \sum_{j=1}^{i-1} a_{ij} x_j^{k+1} + a_{ii} x_i^{k+1} - (1-\omega) \cdot a_{ii} x_i^k + \omega \cdot \sum_{j=i+1}^{n} a_{ij} x_j^k = \omega \cdot b_i$$

pentru $i = \overline{1,n}$. Se consideră descompunerea lui A în forma $A = L + D + U$, unde L este partea strict inferior triunghiulară a lui A, D este partea diagonală a lui A iar U este partea strict superior triunghiulară a lui A, matricile L, D, U fiind pătratice de ordinul n cu elemente nule în afara elementelor luate din matricea A. Aşadar avem: $\omega \cdot L \cdot x^{k+1} + D \cdot x^{k+1} - (1-\omega) \cdot Dx^k + \omega \cdot U \cdot x^k = \omega \cdot b$, adică $(D+\omega L) \cdot x^{k+1} = [(1-\omega) \cdot D - \omega \cdot U] \cdot x^k + \omega \cdot b$, ceea ce este echivalent cu $x^{k+1} = (D+\omega L)^{-1} \cdot [(1-\omega) \cdot D - \omega \cdot U] \cdot x^k + (D+\omega L)^{-1} \cdot \omega \cdot b$, fiindcă $\det(D + \omega L) = \det(D) = \prod_{i=1}^{n} a_{ii} \neq 0$, deci există $(D + \omega L)^{-1}$. Notăm $B_\omega = (D+\omega L)^{-1} \cdot [(1-\omega) \cdot D - \omega U]$ şi $c_\omega = (D+\omega L)^{-1} \cdot \omega b$, deci avem iteraţia SOR sub forma matricială $x^{k+1} = B_\omega x^k + c_\omega$, adică avem tocmai o metodă staţionară.

Teorema 6.4.3. *Are loc inegalitatea* $r(B_\omega) \geq |1 - \omega|$.

Demonstraţie. Prima dată să calculăm $\det(B_\omega)$. Vom folosi formulele $\det(A_1 \cdot A_2) = \det(A_1) \cdot \det(A_2)$, respectiv $\det(A^{-1}) = \frac{1}{\det(A)}$. Avem pe rând:

$$\begin{aligned}
\det(B_\omega) &= \det\{(D+\omega L)^{-1} \cdot [(1-\omega) \cdot D - \omega U]\} = \\
&= \det[(D+\omega L)^{-1}] \cdot \det[(1-\omega) \cdot D - \omega U] = \\
&= \frac{1}{\det(D+\omega L)} \cdot \det[(1-\omega) \cdot D - \omega U] = \\
&= \frac{1}{\det D} \cdot \det[(1-\omega) \cdot D] = \frac{1}{\prod_{i=1}^{n} a_{ii}} \cdot \prod_{i=1}^{n} (1-\omega) \cdot a_{ii} = (1-\omega)^n.
\end{aligned}$$

Pe de altă parte $B_\omega = C^{-1} \cdot B^* \cdot C$, unde B^* este forma canonică Jordan a matricii B_ω iar C este matricea de trecere ca să obţinem această formă canonică. Avem pe rând:

$$\begin{aligned} \det(B_\omega) &= \det(C^{-1} \cdot B^* \cdot C) = \det(C^{-1}) \cdot \det(B^*) \cdot \det(C) = \\ &= \frac{1}{\det(C)} \cdot \left(\prod_{i=1}^{n} \lambda_i \right) \cdot \det(C) = \prod_{i=1}^{n} \lambda_i, \end{aligned}$$

unde $\lambda_1, \lambda_2, \ldots, \lambda_n$ sunt valorile proprii ale matricii B_ω, care sunt aşezate pe diagonala principală a matricii B^*, sub care avem numai zerouri. Prin urmare $(1-\omega)^n = \lambda_1 \lambda_2 \ldots \lambda_n$, adică $|(1-\omega)^n| = |\lambda_1 \lambda_2 \ldots \lambda_n|$, deci $|1-\omega|^n = |\lambda_1| \cdot |\lambda_2| \cdot \ldots \cdot |\lambda_n| \leq [r(B_\omega)]^n$, adică $|1-\omega| \leq r(B_\omega)$. q.e.d.

Consecinţa 6.4.1. *Metoda SOR este convergentă dacă $\omega \in (0,2)$. Într-adevăr, din teorema 6.4.2 de la metodele staţionare rezultă că metoda SOR este convergentă, dacă $r(B_\omega) < 1$. Însă conform teoremei 6.4.3 $|1-\omega| \leq r(B_\omega)$, deci $|1-\omega| < 1$, adică $\omega \in (0,2)$.*

Program SOR

Datele de intrare: $n; A; b; \varepsilon; x^0$;

Fie $y := x^0$;

Repetă $x := y$;

$y := B_\omega x + C_\omega$;

Până când $\|y - x\| \geq \varepsilon$;

Tipăreşte y.

În acest program ε înseamnă precizia dinainte dată, $(\varepsilon = 10^{-2}, 10^{-3}, \ldots)$, $x^0 = (x_1^0, x_2^0, \ldots, x_n^0) \in \mathbb{R}^n$ este punctul iniţial de pornire, atribuirile $y := x^0$ şi $x := y$ înseamnă atribuiri de matrici coloane termen cu termen. Instrucţiunea $y := B_\omega x + c_\omega$ se referă la iteraţia $x^{k+1} = B_\omega x^k + c_\omega$, care se poate scrie pe larg în felul următor: pentru $i = \overline{1,n}$ execută:

$$x_i^{k+1} := x_i^k + \omega \cdot \frac{b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{k+1} - \sum_{j=i}^{n} a_{ij} x_j^k}{a_{ii}}.$$

Pentru vectorul "vechi" x^k folosim vectorul x şi pentru vectorul "nou" x^{k+1} folosim vectorul y :

Pentru $i = \overline{1,n}$ execută:

$$y[i] := x[i] + \omega \cdot \frac{b[i] - \sum_{j=1}^{i-1} a[i,j] * y[j] - \sum_{j=i}^{n} a[i,j] * x[j]}{a[i,i]}.$$

Din punct de vedere al programării ultima instrucţiune se poate realiza în felul următor:

Fie $s_1 := 0$; $s_2 := 0$;

Pentru $j = \overline{1,i-1}$ execută:

$$s_1 := s_1 + a[i,j] * y[j];$$

Pentru $j = \overline{i,n}$ execută:

$$s_2 := s_2 + a[i,j] * x[j];$$

$y[i] := x[i] + \omega \cdot (b[i] - s_1 - s_2)/a[i,i]$;

La condiţia $\|y - x\| \geq \varepsilon$ prima dată se calculează într-o subrutină o anumită normă vectorială a vectorului $y - x$:

Fie $s := 0$;

Pentru $i := \overline{1,n}$ execută:

$$s := s + |y[i] - x[i]|;$$

şi în final în variabila s avem valoarea normei $\|y - x\|$.

6.5 Rezolvarea sistemelor liniare supradeterminate cu metoda celor mai mici pătrate

Definiţia 6.5.1. *Un sistem liniar se numeşte supradeterminat dacă numărul ecuaţiilor este mai mare (sau mult mai mare) decât numărul neunoscutelor.*

Fie deci sistemul liniar sub forma matricială $A \cdot x = b$, unde $A = (a_{ij})_{\substack{i=\overline{1,n} \\ j=\overline{1,m}}}$, $x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix} \in \mathbb{R}^m$, $b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} \in \mathbb{R}^n$, şi $n > m$. În general un sistem liniar de acest gen

este incompatibil în sensul clasic al soluţiilor, adică nu există un vector $x \in \mathbb{R}^m$ care să satisfacă simultan fiecare ecuaţie a sistemului liniar. De aceea se formează vectorul rezidual $r = A \cdot x - b \in \mathbb{R}^n$ şi urmărim ca $\|r\|_2^2$ să fie minimă, unde am considerat norma euclideană în spaţiul $\mathbb{R}^n$. Dacă există un vector $x \in \mathbb{R}^n$ pentru care se minimizează expresia $\|r\|_2^2 = \|Ax - b\|_2^2$, atunci vectorul x se acceptă ca soluţie a sistemului iniţial $Ax = b$, dar nu în sens clasic, ci în sensul celor mai mici pătrate.

Fie $f : \mathbb{R}^m \to \mathbb{R}$,

$$f(x) = f(x_1, x_2, \ldots, x_m) = \sum_{i=1}^{n} \left(\sum_{j=1}^{m} a_{ij} x_j - b_i \right)^2 = \|Ax - b\|_2^2 = \|r\|_2^2.$$

Observăm că $f(x) \geq 0$ pentru orice $x \in \mathbb{R}^m$, şi se caută o valoare $x \in \mathbb{R}^m$ pentru care f ia valoarea minimă. Se determină punctele staţionare ale sistemului $\dfrac{\partial f}{\partial x_k}(x) = 0$ pentru $k = \overline{1, m}$. Avem următorul calcul

$$\frac{\partial f}{\partial x_k}(x) = \sum_{i=1}^{n} 2 \cdot \left(\sum_{j=1}^{m} a_{ij} x_j - b_i \right) \cdot a_{ik} = 0,$$

adică

$$\sum_{i=1}^{n} a_{ik} \cdot \left(\sum_{j=1}^{m} a_{ij} x_j - b_i \right) = 0$$

pentru orice $k = \overline{1, m}$. Acest sistem are forma matricială $A^T(Ax - b) = \theta_{\mathbb{R}^m}$. Prin urmare soluţia clasică a sistemului linear $(A^T \cdot A) \cdot x = A^T b$ se acceptă ca soluţie a sistemului linear iniţial supradeterminat în sensul celor mai mici pătrate.

În continuare arătăm că punctul staţionar $x \in \mathbb{R}^m$ care verifică relaţia $(A^T A)x = A^T b$ va fi punct de minim pentru funcţia f.

Teorema 6.5.1. *Dacă A este o matrice de tip $m \times n$, b este o matrice coloană de tip $n \times 1$, iar $x \in \mathbb{R}^m$ este soluţia clasică a sistemului linear $A^T(Ax - b) = \theta_{\mathbb{R}^m}$ atunci pentru orice $y \in \mathbb{R}^m$ se obţine $\|b - Ax\|_2 \leq \|b - Ay\|_2$.*

Demonstraţie. Notăm vectorii reziduali cu $r_x = b - Ax$ şi $r_y = b - Ay$. Avem $r_y = b - Ay = b - Ax + Ax - Ay = r_x + A(x - y)$. Folosind formulele matricilor transpuse $(A + B)^T = A^T + B^T$ şi $(A \cdot B)^T = B^T \cdot A^T$ se obţine că $r_y^T = (r_x + A(x - y))^T = r_x^T + (x - y)^T A^T$. Atunci

$$\begin{aligned} r_y^T \cdot r_y &= (r_x^T + (x - y)^T A^T) \cdot (r_x + A(x - y)) = \\ &= r_x^T \cdot r_x + (x - y)^T \cdot A^T \cdot r_x + r_x^T \cdot A(x - y) + (x - y)^T A^T \cdot A(x - y). \end{aligned}$$

Însă $A^T \cdot r_x = \theta_{\mathbb{R}^m}$ şi $r_x^T A = (A^T r_x)^T = \theta_{\mathbb{R}^m}$ fiindcă $(A^T)^T = A$. Prin urmare $r_y^T \cdot r_y = r_x^T \cdot r_x + (x-y)^T A^T \cdot A(x-y)$. Se poate verifica uşor că $\|r_y\|_2^2 = r_y^T \cdot r_y$ conform definiţiei normei euclidiene. Prin urmare $\|r_y\|_2^2 = \|r_x\|_2^2 + \|A(x-y)\|_2^2 \geq \|r_x\|_2^2$, ceea ce trebuia demonstrat. □

Exemplu. Să se rezolve următorul sistem linear supradeterminat în sensul celor mai mici pătrate: $\begin{cases} x+y=2 \\ x+2y=3 \\ 2x+y=4 \end{cases}$. Dacă se rezolvă sistemul linear $\begin{cases} x+y=2 \\ x+2y=3 \end{cases}$ avem soluţia unică $x=y=1$. Se observă imediat că $x=y=1$ nu verifică ecuaţia $2x+y=4$. Ajungem la acelaş rezultat dacă se calculează determinantul caracteristic

$$\begin{vmatrix} 1 & 1 & 2 \\ 1 & 2 & 3 \\ 2 & 1 & 4 \end{vmatrix} = 8+2+6-8-3-4 = 1 \neq 0.$$

Deoarece determinantul caracteristic este diferit de zero sistemul este incompatibil, deci nu admite soluţie în sens clasic. Se introduce funcţia $f : \mathbb{R}^2 \to \mathbb{R}$,

$$f(x,y) = (x+y-2)^2 + (x+2y-3)^2 + (2x+y-4)^2.$$

Se rezolvă sistemul

$$\begin{cases} \dfrac{\partial f}{\partial x}(x,y) = 0 \\ \\ \dfrac{\partial f}{\partial y}(x,y) = 0 \end{cases}$$

şi se obţine:

$$\begin{cases} 2(x+y-2) + 2(x+2y-3) + 2(2x+y-4) \cdot 2 = 0 \\ 2(x+y-2) + 2(x+2y-3) \cdot 2 + 2(2x+y-4) = 0. \end{cases}$$

Prin urmare $\begin{cases} 6x+5y-13=0 \\ 5x+6y-12=0 \end{cases}$ care admite soluţia clasică $x = \dfrac{18}{11}$ şi $y = \dfrac{7}{11}$. Această soluţie se acceptă ca soluţia sistemului iniţial în sensul celor mai mici pătrate. Soluţia $x = \dfrac{18}{11}$ şi $y = \dfrac{7}{11}$ încearcă să verifice toate cele trei ecuaţii liniare ale sistemului iniţial minimizând vectorul rezidual corespunzător.

Menţionăm că $A^TAx = A^Tb$ ne dă acelaşi sistem liniar cu alegerea: $A = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 2 & 1 \end{bmatrix}$ şi $b = \begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix}$.

Capitolul 7

Sisteme de ecuaţii neliniare pe spaţii finit dimensionale

7.1 Metoda lui Jacobi pe spaţiul finit dimensional $\mathbb{R}^n$

Fie $F : D \to \mathbb{R}^n$, $D \neq \emptyset$, $D \subseteq \mathbb{R}^n$ o funcţie dată, la care se ataşează ecuaţia corespunzătoare $F(x) = \theta_{\mathbb{R}^n}$, unde $x = (x_1, x_2, \ldots, x_n) \in D$ iar cu $\theta_{\mathbb{R}^n} = (0, 0, \ldots, 0)$ am notat originea spaţiului $\mathbb{R}^n$. Ecuaţia se poate scrie sub forma unui sistem de ecuaţii neliniare:

$$\begin{cases} F_1(x_1, x_2, \ldots, x_n) = 0 \\ F_2(x_1, x_2, \ldots, x_n) = 0 \\ \vdots \\ F_n(x_1, x_2, \ldots, x_n) = 0, \end{cases}$$

unde $F = (F_1, F_2, \ldots, F_n)$, $F_i : D \subset \mathbb{R}^n \to \mathbb{R}$ fiind componentele funcţiei F şi unde presupunem că cel puţin o componentă F_i, $i \in \{1, 2, \ldots, n\}$ nu este o aplicaţie liniară. Pentru a rezolva acest sistem de ecuaţii, prima dată prin transformări echivalente sistemul neliniar se pune sub o formă iterativă $\Phi(x) = x$, unde $\Phi : D \subset \mathbb{R}^n \to \mathbb{R}^n$ se numeşte funcţia iterativă. Dacă $\Phi = (\Phi_1, \Phi_2, \ldots, \Phi_n)$, unde $\Phi_i : D \subset \mathbb{R}^n \to \mathbb{R}$ $i = \overline{1, n}$ sunt componentele

lui Φ, atunci se obţine sistemul

$$\begin{cases} \Phi_1(x_1, x_2, \ldots, x_n) = x_1 \\ \Phi_2(x_1, x_2, \ldots, x_n) = x_2 \\ \vdots \\ \Phi_n(x_1, x_2, \ldots, x_n) = x_n. \end{cases}$$

Exemplul 7.1.1. *Se consideră sistemul* $\begin{cases} x_1 - \sin x_2 + x_2^2 = 0 \\ x_1^3 - 3x_2 + 1 = 0 \end{cases}$ *. În cazul nostru* $n = 2$, $F_1(x_1, x_2) = x_1 - sinx_2 + x_2^2$ *şi* $F_2(x_1, x_2) = x_1^3 - 3x_2 + 1$. *Sistemul dat este echivalent cu sistemul* $\begin{cases} x_1 = \sin x_2 - x_2^2 \\ x_2 = \dfrac{1}{3}(x_1^3 + 1) \end{cases}$ *. Prin urmare cu această scriere se obţine funcţia iterativă* $\Phi = (\Phi_1, \Phi_2) : D \subset \mathbb{R}^2 \to \mathbb{R}^2$, *unde* $\Phi_1(x_1, x_2) = \sin x_2 - x_2^2$ *şi* $\Phi_2(x_1, x_2) = \dfrac{1}{3}(x_1^3 + 1)$.

Exemplul 7.1.2. *Dacă* $\omega \in \mathbb{R}^* = \mathbb{R} \setminus \{0\}$, *atunci sistemul de ecuaţii* $F(x) = \theta_{\mathbb{R}^n}$ *este echivalent cu* $x + \omega \cdot F(x) = x$, *deci putem alege ca funcţie iterativă* $\Phi(x) = x + \omega F(x)$.

Rostul transformării ecuaţiei $F(x) = \theta_{\mathbb{R}^n}$ *sub forma iterativă* $\Phi(x) = x$ *este de a aplica teoreme şi tehnici de punct fix. Vom spune că un vector* $x^* \in \mathbb{R}^n$ *este punct fix pentru funcţia iterativă* Φ, *dacă* $\Phi(x^*) = x^*$. *Utilizând de exemplu teorema de punct fix a lui Banach putem asigura existenţa şi unicitatea punctului fix* x^*. *Deoarece ecuaţiile* $\Phi(x) = x$ *şi* $F(x) = \theta_{\mathbb{R}^n}$ *sunt echivalente, rezultă că* $F(x^*) = \theta_{\mathbb{R}^n}$, *adică asigurăm existenţa şi unicitatea soluţiei pentru ecuaţia* $F(x) = \theta_{\mathbb{R}^n}$.

Teorema 7.1.1. *(Jacobi) Fie* $D \subset \mathbb{R}^n$, $D \neq \emptyset$ *un domeniu (*D *este deschis şi conex), iar* $\Phi : D \to \mathbb{R}^n$, $\Phi = (\Phi_1, \Phi_2, \ldots, \Phi_n)$ *o funcţie diferenţiabilă. Dacă* $D^* \neq \emptyset$, $D^* \subset D$ *este o submulţime închisă şi convexă (pentru orice două puncte* $x, y \in D^*$ *segmentul determinat de capetele* x *şi* y *cade în întregime în mulţimea* D^**), care este invariantă faţă de* Φ, *adică* $\Phi(D^*) \subset D^*$, *şi există* $\lambda \in [0, 1)$ *astfel încât*

$$\sum_{j=1}^{n} \left| \frac{\partial \Phi_i}{\partial x_j}(x) \right| \leq \lambda$$

pentru orice $i = \overline{1, n}$ *şi orice* $x \in D^*$, *atunci sistemul de ecuaţii* $\Phi(x) = x$ *admite o unică soluţie* x^* *în mulţimea* D^*.

Pentru a demonstra teorema lui Jacobi avem nevoie de teorema de punct fix a lui Banach:

Teorema 7.1.2. *(Banach) Dacă $D^* \subset \mathbb{R}^n$, $D^* \neq \emptyset$ este o submulţime închisă în spaţiul $\mathbb{R}^n$, iar $\Phi : D^* \to D^*$ este o contracţie, adică există $\alpha \in [0,1)$ astfel încât $\|\Phi(x) - \Phi(y)\| \leq \alpha \cdot \|x - y\|$ pentru orice $x, y \in D^*$, atunci pentru orice punct de pornire $x^0 \in D^*$ şirul $(x^k)_{k\in\mathbb{N}}$ dat prin relaţia de recurenţă $x^{k+1} = \Phi(x^k)$ este convergent către un punct $x^* \in D^*$, care va fi unicul punct fix al lui Φ în D^*. Pentru evaluarea erorilor avem estimarea apriori*

$$\|x^* - x^k\| \leq \frac{\alpha^k}{1-\alpha} \cdot \|x^1 - x^0\|$$

respectiv estimarea aposteriori

$$\|x^* - x^k\| \leq \frac{\alpha}{1-\alpha} \cdot \|x^k - x^{k-1}\|.$$

Menţionăm că în general teorema lui Banach se enunţă pe tot spaţiul $\mathbb{R}^n$, însă putem considera în locul lui $\mathbb{R}^n$ o submulţime închisă $D^* \subset \mathbb{R}^n$ care să fie invariantă faţă de $\Phi : \Phi(D^*) \subset D^*$.

Pentru a demonstra teorema a lui Jacobi mai avem nevoie şi de o teoremă de medie:

Teorema 7.1.3. *Dacă $D \subset \mathbb{R}^n$, $D \neq \emptyset$ este o submulţime deschisă şi convexă, iar funcţia $\Phi_1 : D \to \mathbb{R}$ este diferenţiabilă, atunci pentru orice două puncte $x, y \in D$ există o valoare intermediară $\theta \in D$ aflat pe segmentul cu capetele x şi y pentru care $\Phi_1(x) - \Phi_1(y) = D\Phi_1(\theta)(x - y)$, unde $D\Phi_1(\theta)$ înseamnă diferenţiala funcţiei Φ_1 în punctul θ.*

Demonstraţie. (teorema 7.1.1 a lui Jacobi) Deoarece $D^* \neq \emptyset$, $D^* \subset \mathbb{R}^n$ este o submulţime închisă, convexă şi invariantă faţă de Φ ne rămâne de verificat că $\Phi : D^* \to D^*$, $\Phi = (\Phi_1, \Phi_2, \ldots, \Phi_n)$ este o contracţie: conform teoremei 7.1.3 pentru orice două puncte $x, y \in D^*$ există $\theta = \lambda^* x + (1 - \lambda^*)y$ cu $\lambda^* \in (0,1)$ astfel încât:

$$\Phi_i(x) - \Phi_i(y) = D\Phi_i(\theta)(x - y).$$

Folosind reprezentarea diferenţialei cu ajutorul derivatelor parţiale obţinem:

$$\begin{aligned}
\Phi_i(x) - \Phi_i(y) &= \sum_{j=1}^{n} \frac{\partial \Phi_i}{\partial x_j}(\theta)(x_j - y_j), \text{ adică} \\
|\Phi_i(x) - \Phi_i(y)| &= \left| \sum_{j=1}^{n} \frac{\partial \Phi_i}{\partial x_j}(\theta)(x_j - y_j) \right| \leq \sum_{j=1}^{n} \left| \frac{\partial \Phi_i}{\partial x_j}(\theta) \right| \cdot |x_j - y_j| \leq \\
&\leq \sum_{j=1}^{n} \left| \frac{\partial \Phi_i}{\partial x_j}(\theta) \right| \cdot \|x - y\|_\infty \leq \left(\sum_{j=1}^{n} \left| \frac{\partial \Phi_i}{\partial x_j}(\theta) \right| \right) \cdot \|x - y\|_\infty \leq \lambda \cdot \|x - y\|_\infty
\end{aligned}$$

unde am ales norma maximum dată de formula: $\|x\|_\infty = \max\{|x_i| \ \mid i = \overline{1,n}\}$. Prin urmare:

$$\|\Phi(x) - \Phi(y)\|_\infty = \max\{|\Phi_i(x) - \Phi_i(y)| \ \mid i = \overline{1,n}\} \leq \lambda \cdot \|x - y\|_\infty$$

adică Φ este o contracţie cu $\alpha = \lambda \in [0, 1)$. q.e.d.

Exemplul 7.1.3. *Să se rezolve următorul sistem folosind metoda lui Jacobi:*

$$\begin{cases} x_1^2 + 2x_2^2 - 7x_1 = 0 \\ 2x_1^2 - x_2^2 - 8x_2 = 0 \end{cases}$$

Avem $n = 2$, *alegem* $D = \mathbb{R}^2$, $D^* = [-1, 1]^2 = [-1, 1] \times [-1, 1]$, $F = (F_1, F_2) : \mathbb{R}^2 \to \mathbb{R}^2$, $F_1(x_1, x_2) = x_1^2 + 2x_2^2 - 7x_1$ *şi* $F_2(x_1, x_2) = 2x_1^2 - x_2^2 - 8x_2$. *Transcriem sistemul sub forma iterativă*

$$\begin{cases} x_1 = \dfrac{1}{7}(x_1^2 + 2x_2^2) \\ \\ x_2 = \dfrac{1}{8}(2x_1^2 - x_2^2), \end{cases}$$

deci $\Phi : D^* \to D^*$ *este dată de* $\Phi_1(x_1, x_2) = \dfrac{1}{7}(x_1^2 + 2x_2^2)$ *şi* $\Phi_2(x_1, x_2) = \dfrac{1}{8}(2x_1^2 - x_2^2)$. *Observăm că din* $(x_1, x_2) \in D^* = [-1, 1]^2$ *rezultă că* $x_1, x_2 \in [-1, 1]$, *deci* $|x_1| \leq 1$ *şi* $|x_2| \leq 1$. *Prin urmare*

$$\begin{aligned} |\Phi_1(x_1, x_2)| &= \frac{1}{7}(x_1^2 + 2x_2^2) \leq \frac{1}{7}(1 + 2 \cdot 1) = \frac{3}{7} \leq 1 \quad \text{şi} \\ |\Phi_2(x_1, x_2)| &= \left|\frac{1}{8}(2x_1^2 - x_2^2)\right| \leq \frac{1}{8}(2 \cdot 1 + 1) = \frac{3}{8} \leq 1. \end{aligned}$$

Mai departe avem pentru orice $x = (x_1, x_2) \in D^*$:

$$\begin{aligned} \left|\frac{\partial \Phi_1}{\partial x_1}(x)\right| + \left|\frac{\partial \Phi_1}{\partial x_2}(x)\right| &= \left|\frac{2}{7}x_1\right| + \left|\frac{4}{7}x_2\right| \leq \frac{2}{7} + \frac{4}{7} = \frac{6}{7} \quad \text{şi} \\ \left|\frac{\partial \Phi_2}{\partial x_1}(x)\right| + \left|\frac{\partial \Phi_2}{\partial x_2}(x)\right| &= \left|\frac{4}{8}x_1\right| + \left|\frac{2}{8}x_2\right| \leq \frac{4}{8} + \frac{2}{8} = \frac{3}{4}. \end{aligned}$$

Prin urmare putem alege $\lambda = \max\left\{\dfrac{6}{7}, \dfrac{3}{4}\right\} = \dfrac{6}{7} < 1$.

Deoarece

$$\frac{\lambda}{1 - \lambda} = \frac{\frac{6}{7}}{1 - \frac{6}{7}} = \frac{\frac{6}{7}}{\frac{1}{7}} = 6$$

pentru condiţia de oprire putem alege $6 \cdot \|x^k - x^{k-1}\| \leq \varepsilon$, *adică* $\|x^k - x^{k-1}\| \leq \dfrac{\varepsilon}{6}$. *Noi totuşi în locul inegalităţii exacte* $\|x^k - x^{k-1}\| \leq \dfrac{\varepsilon}{6}$ *vom considera inegalitatea practică* $\|x^k - x^{k-1}\| \leq \varepsilon$.

Algoritmul metoda lui Jacobi

Datele de intrare: n; $\Phi = (\Phi_1, \Phi_2, \ldots, \Phi_n)$; $x^0 = (x_1^0, x_2^0, \ldots, x_n^0)$; ε;
Fie $y := x^0$;
Repetă $x := y$;
$y := \Phi(x)$;
Până când $\|y - x\| \geq \varepsilon$;
Tipăreşte y.

Menţionăm că ciclul se repetă până când condiţia $\|y - x\| < \varepsilon$ devine adevărată şi când $\|y - x\| < \varepsilon$ atunci ieşim din ciclu.

7.2 Metoda lui Newton-Raphson-Kantorovici pe $\mathbb{R}^n$

Dacă se alege $n = 1$ atunci metoda lui Newton-Raphson-Kantorovici este tocmai metoda tangentei pe axa reală (vezi paragraful 5.4.1), dată de recurenţa $x^{k+1} = x^k - [f'(x^k)]^{-1} f(x^k)$. Considerăm cazul $n = 2$. Fie $F : D \subset \mathbb{R}^2 \to \mathbb{R}$, $D \neq \emptyset$ o funcţie dată, unde $F = (F_1, F_2)$, $F_1, F_2 : D \subset \mathbb{R}^2 \to \mathbb{R}$ şi $F(x) = \theta_{\mathbb{R}^2}$, $x = (x_1, x_2) \in D$ sistemul de ecuaţii dat sub forma

$$\begin{cases} F_1(x_1, x_2) = 0 \\ F_2(x_1, x_2) = 0 \end{cases} \quad \text{sau} \quad \begin{cases} F_1(x) = 0 \\ F_2(x) = 0. \end{cases}$$

Considerăm suprafeţele $z = F_1(x_1, x_2)$ şi $z = F_2(x_1, x_2)$ cu $x = (x_1, x_2) \in D$. Atunci pentru a obţine soluţia sistemului $F(x) = \theta_{\mathbb{R}^2}$ considerăm curba spaţială dată de intersecţia celor două suprafeţe:

$$\begin{cases} x_3 = F_1(x_1, x_2) \\ x_3 = F_2(x_1, x_2) \end{cases}$$

şi unde această curbă intersectează planul x_1Ox_2 ($x_3 = 0$) vom avea soluţia sistemului $F(x) = \theta_{\mathbb{R}^2}$, notată cu $x^* = (x_1^*, x_2^*) \in D$. Fie $x^0 = (x_1^0, x_2^0) \in D$ un punct iniţial de pornire aflat într-o vecinătate suficient de mică a lui x^*. Construcţia geometrică pentru aproximarea soluţiei x^* plecând de la punctul x^0 se face în felul următor: în punctele $(x^0, F_1(x^0)) = (x_1^0, x_2^0, F_1(x_1^0, x_2^0))$ şi $(x^0, F_2(x^0)) = (x_1^0, x_2^0, F_2(x_1^0, x_2^0))$ se duc două plane

tangente la suprafeţele $F_1(x_1,x_2)-x_3=0$ şi $F_2(x_1,x_2)-x_3=0$ scrise sub forma implicită, având ecuaţiile:

$$\frac{\partial F_1}{\partial x_1}(x^0)\cdot(x_1-x_1^0)+\frac{\partial F_1}{\partial x_2}(x^0)(x_2-x_2^0)-1\cdot(x_3-F_1(x^0))=0 \quad \text{şi}$$
$$\frac{\partial F_2}{\partial x_1}(x^0)\cdot(x_1-x_1^0)+\frac{\partial F_2}{\partial x_2}(x^0)(x_2-x_2^0)-1\cdot(x_3-F_2(x^0))=0,$$

unde $(x,x_3)=(x_1,x_2,x_3)\in\mathbb{R}^3$ este un punct curent din spaţiul $\mathbb{R}^3$. Cele două plane tangente se intersectează după o dreaptă având ecuaţia:

$$\begin{cases} x_3=F_1(x^0)+\dfrac{\partial F_1}{\partial x_1}(x^0)(x_1-x_1^0)+\dfrac{\partial F_1}{\partial x_2}(x^0)(x_2-x_2^0) \\ x_3=F_2(x^0)+\dfrac{\partial F_2}{\partial x_1}(x^0)(x_1-x_1^0)+\dfrac{\partial F_2}{\partial x_2}(x^0)(x_2-x_2^0). \end{cases}$$

Deoarece cele două plane tangente sunt aproximaţii liniare pentru suprafeţele $x_3=F_1(x)$ şi $x_3=F_2(x)$, rezultă că dreapta obţinută prin intersecţia celor două plane tangente va aproxima curba spaţială dată de $\begin{cases} x_3=F_1(x) \\ x_3=F_2(x) \end{cases}$. Astfel intersecţia dreptei cu planul x_1Ox_2 $(x_3=0)$ va fi un punct $x^1=(x_1^1,x_2^1)\in D$ care aproximează soluţia $x^*\in D$. Prin urmare pentru $x_3=0$ obţinem $x_1=x_1^1$ şi $x_2=x_2^1$:

$$\begin{cases} 0=F_1(x^0)+\dfrac{\partial F_1}{\partial x_1}(x^0)(x_1^1-x_1^0)+\dfrac{\partial F_1}{\partial x_2}(x^0)\cdot(x_2^1-x_2^0) \\ 0=F_2(x^0)+\dfrac{\partial F_2}{\partial x_1}(x^0)(x_1^1-x_1^0)+\dfrac{\partial F_2}{\partial x_2}(x^0)\cdot(x_2^1-x_2^0), \text{ deci} \end{cases}$$
$$\begin{cases} \dfrac{\partial F_1}{\partial x_1}(x^0)\cdot(x_1^1-x_1^0)+\dfrac{\partial F_1}{\partial x_2}(x^0)\cdot(x_2^1-x_2^0)=-F_1(x^0) \\ \dfrac{\partial F_2}{\partial x_1}(x^0)\cdot(x_1^1-x_1^0)+\dfrac{\partial F_2}{\partial x_2}(x^0)\cdot(x_2^1-x_2^0)=-F_2(x^0) \end{cases}$$

Forma matricială a acestui sistem este:

$$\begin{bmatrix} \dfrac{\partial F_1}{\partial x_1}(x^0) & \dfrac{\partial F_1}{\partial x_2}(x^0) \\ \dfrac{\partial F_2}{\partial x_1}(x^0) & \dfrac{\partial F_2}{\partial x_2}(x^0) \end{bmatrix}\cdot\begin{bmatrix} x_1^1-x_1^0 \\ x_2^1-x_2^0 \end{bmatrix}=-\begin{pmatrix} F_1(x^0) \\ F_2(x^0) \end{pmatrix}.$$

Notăm cu

$$F'(x)=\frac{D(F)}{D(x)}=\frac{D(F_1,F_2)}{D(x_1,x_2)}=J=\begin{bmatrix} \dfrac{\partial F_1}{\partial x_1}(x^0) & \dfrac{\partial F_1}{\partial x_2}(x^0) \\ \dfrac{\partial F_2}{\partial x_1}(x^0) & \dfrac{\partial F_2}{\partial x_2}(x^0) \end{bmatrix}$$

matricea jacobiană, despre care presupunem că este inversabilă în punctul x^0, notând cu J^{-1} matricea inversă:

$$\begin{pmatrix} x_1^1 - x_1^0 \\ x_2^1 - x_2^0 \end{pmatrix} = -J^{-1} \cdot \begin{pmatrix} F_1(x^0) \\ F_2(x^0) \end{pmatrix}, \text{ adică } \begin{pmatrix} x_1^1 \\ x_2^1 \end{pmatrix} = \begin{pmatrix} x_1^0 \\ x_2^0 \end{pmatrix} - J^{-1} \cdot \begin{pmatrix} F_1(x^0) \\ F_2(x^0) \end{pmatrix}.$$

Prin urmare în acest mod am obţinut formula de recurenţă pentru trecerea de la punctul de start $x^0 = (x_1^0, x_2^0)$ la următorul punct $x^1 = (x_1^1, x_2^1)$. În general presupunând că am găsit punctul $x^k = (x_1^k, x_2^k)$ construcţia următorului punct $x^{k+1} = (x_1^{k+1}, x_2^{k+1})$ se va face în acelaşi mod cum am trecut de la punctul x^0 la punctul x^1, adică

$$\begin{bmatrix} x_1^{k+1} \\ x_2^{k+1} \end{bmatrix} = \begin{bmatrix} x_1^k \\ x_2^k \end{bmatrix} - \begin{bmatrix} \frac{\partial F_1}{\partial x_1}(x^k) & \frac{\partial F_1}{\partial x_2}(x^k) \\ \frac{\partial F_2}{\partial x_1}(x^k) & \frac{\partial F_2}{\partial x_2}(x^k) \end{bmatrix}^{-1} \cdot \begin{pmatrix} F_1(x^k) \\ F_2(x^k) \end{pmatrix}.$$

De aici deja se vede forma iteraţiei lui Newton-Raphson în cazul general. Fie deci sistemul: $F(x) = \theta_{\mathbb{R}^n}$, unde $F : D \subset \mathbb{R}^n \to \mathbb{R}^n$, $F = (F_1, F_2, \ldots, F_n)$ cu $F_i : D \subset \mathbb{R}^n \to \mathbb{R}$. Dacă $x^0 = (x_1^0, x_2^0, \ldots, x_n^0) \in \mathbb{R}^n$ este un punct de pornire atunci presupunând că am obţinut la pasul k punctul $x^k = (x_1^k, x_2^k, \ldots, x_n^k)$ următorul punct $x^{k+1} = (x_1^{k+1}, x_2^{k+1}, \ldots, x_n^{k+1})$ la pasul $k+1$ se va obţine conform formulei:

$$x^{k+1} = x^k - J^{-1} \cdot F(x^k) = x^k - (F'(x^k))^{-1} \cdot F(x^k),$$

pe larg:

$$\begin{pmatrix} x_1^{k+1} \\ x_2^{k+1} \\ \vdots \\ x_n^{k+1} \end{pmatrix} = \begin{pmatrix} x_1^k \\ x_2^k \\ \vdots \\ x_n^k \end{pmatrix} - \begin{pmatrix} \frac{\partial F_1}{\partial x_1}(x^k) & \frac{\partial F_1}{\partial x_2}(x^k) & \ldots & \frac{\partial F_1}{\partial x_n}(x^k) \\ \frac{\partial F_2}{\partial x_1}(x^k) & \frac{\partial F_2}{\partial x_2}(x^k) & \ldots & \frac{\partial F_2}{\partial x_n}(x^k) \\ \vdots & \vdots & & \vdots \\ \frac{\partial F_n}{\partial x_1}(x^k) & \frac{\partial F_n}{\partial x_2}(x^k) & \ldots & \frac{\partial F_n}{\partial x_n}(x^k) \end{pmatrix}^{-1} \begin{pmatrix} F_1(x^k) \\ F_2(x^k) \\ \vdots \\ F_n(x^k) \end{pmatrix}$$

Menţionăm că formal se poate obţine această recurenţă, dacă considerăm dezvoltarea taylorianăa funcţiei F în punctul x^k :

$$F(x) = F(x^k) + F'(x^k)(x - x^k) + \frac{F''(x^k)}{2!}(x - x^k)^2 + \ldots$$

Trunchiem această dezvoltare reţinând primii doi termeni, practic facem o "aproximare liniară":

$$F(x) = F(x^k) + F'(x^k)(x - x^k).$$

Punând $x = x^{k+1}$ avem

$$F(x^{k+1}) = F(x^k) + F'(x^k)(x^{k+1} - x^k).$$

Presupunând că $F(x^{k+1}) \approx \theta_{\mathbb{R}^n}$ obţinem următorul şir de relaţii echivalente între ele (presupunând că există $[F'(x^k)]^{-1}$):

$$\begin{aligned} & \theta_{\mathbb{R}^n} = F(x^k) + F'(x^k)(x^{k+1} - x^k) \Leftrightarrow \\ \Leftrightarrow \quad & F'(x^k)(x^{k+1} - x^k) = -F(x^k) \Leftrightarrow \\ \Leftrightarrow \quad & x^{k+1} - x^k = -[F'(x^k)]^{-1} \cdot F(x^k) \Leftrightarrow \\ \Leftrightarrow \quad & x^{k+1} = x^k - [F'(x^k)]^{-1} \cdot F(x^k). \end{aligned}$$

Pentru a asigura existenţa şi convergenţa şirului iterativ $(x^k)_{k\in\mathbb{N}}$ dat de formula de recurenţă de mai sus, avem un rezultat binecunoscut sub numele de teorema lui Kantorovici. Menţionăm că în acelaşi timp acest rezultat a fost obţinut şi de către Ostrowski şi în literatura de specialitate anglo-americană apare sub denumirea de metoda lui Newton-Raphson.

Fie deci ecuaţia operatorială $F(x) = \theta_{\mathbb{R}^n}$, unde $F : D \subset \mathbb{R}^n \to \mathbb{R}^n$, $D \neq \emptyset$ este o funcţie dată.

Teorema 7.2.1. *(Kantorovici-Ostrowski) Fie $F : D \subset \mathbb{R}^n \to \mathbb{R}^n$, $D \neq \emptyset$ o funcţie diferenţiabilă Fréchet pe mulţimea deschisă D, iar $D_0 \subset D$ o submulţime convexă. Alegem $x^0 \in D_0$. Presupunem că următoarele condiţii sunt satisfăcute:*

1. *există $(F'(x^0))^{-1}$ şi $\|(F'(x^0))^{-1}\| \leq \beta_0$;*
2. *există o constantă $\gamma > 0$ astfel ca $\|F'(x) - F'(y)\| \leq \gamma \cdot \|x - y\|$ pentru orice $x, y \in D_0$;*
3. *$\|(F'(x^0))^{-1} \cdot F(x^0)\| \leq \eta_0$;*
4. *$\alpha_0 = \beta_0 \cdot \gamma \cdot \eta_0 \leq \frac{1}{2}$;*
5. *sfera $B_0 = B(x^0, r_0) = \{x \in \mathbb{R}^n \ / \ \|x - x^0\| \leq r_0\}$ este inclusă în D_0, unde*

$$r_0 = \frac{1 - \sqrt{1 - 2\alpha_0}}{\alpha_0} \cdot \eta_0.$$

Atunci ecuaţia operatorială $F(x) = \theta_{\mathbb{R}^n}$ admite o unică soluţie $x^* \in D_0$, iar şirul $(x^k)_{k\in\mathbb{N}}$ dat de formula de recurenţă a lui Newton $x^{k+1} = x^k - (F'(x^k))^{-1} \cdot F(x^k)$ este bine definit şi converge la x^*. Eroarea la iteraţia k este dată de $\|x^k - x^*\| \leq 2^{1-k} \cdot (2\alpha_0)^{2^k - 1} \cdot \eta_0$.

Observaţia 7.2.1.

1. *În unele cazuri în locul condiţiei 2, de lipschitzianitate a lui F se foloseşte o condiţie mai puternică, şi anume ca F să fie diferenţiabilă Fréchet de două ori şi $\|F''(x)\| \leq \gamma$ dacă $\|x - x^0\| \leq 2\eta_0$. Din inegalitatea mediilor în spaţiul $\mathbb{R}^n$ obţinem:*

$$\|F'(x) - F'(y)\| \leq \|F''(\xi)\| \cdot \|x - y\| \leq \gamma \cdot \|x - y\|,$$

unde ξ este o valoare intermediară pe segmentul determinat de punctele $x, y \in \mathbb{R}^n$.

2. *Dacă pe $\mathbb{R}^n$ alegem norma vectorială infinit, adică $\|x\|_\infty = \max\{|x_i| \ / \ i = \overline{1,n}\}$ cu $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$, atunci se alege norma matricială subordonată cunoscută sub numele de norma matricială linie:*

$$\|A\| = \|(a_{ij})_{ij=\overline{1,n}}\| = \max_{1 \leq i \leq n} \sum_{j=1}^{n} |a_{ij}|.$$

În acest fel

$$\|[F'(x^0)]^{-1}\| = \max_{1 \leq i \leq n} \sum_{j=1}^{n} \frac{|\det(A_{ij})|}{|\det([F'(x^0)]^{-1})|} = \max_{1 \leq i \leq n} \frac{1}{|\det([F'(x^0)]^{-1})|} \sum_{j=1}^{n} |\det(A_{ij})| \leq \beta_0,$$

unde A_{ij} este complementul algebric corespunzător elementului a_{ij} obţinut în matricea transpusă a matricii inverse a lui Jacobi $[F'(x^0)]^{-1}$. Pentru calculul normei diferenţialei de ordinul doi avem estimarea:

$$\|F''(x)\| \leq \max_{1 \leq i \leq n} \sum_{j=1}^{n} \sum_{k=1}^{n} \left| \frac{\partial^2 F_i(x)}{\partial x^j \partial x^k} \right| \leq n^2 \cdot L \leq \gamma,$$

unde L este o margine superioară pentru derivatele parţiale de ordinul doi ale lui F_i, $i = \overline{1,n}$, pe D.

3. *Dacă pe $\mathbb{R}^n$ alegem norma matricială euclidiană $\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$ cu $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$, atunci se alege o normă matricială compatibilă cu norma vectorială euclidiană (şi nu cea subordonată), dată de formula*

$$\|A\| = \|(a_{ij})_{i,j=\overline{1,n}}\| = \sqrt{\sum_{i,j=1}^{n} a_{ij}^2}.$$

Prin urmare în cazul nostru:

$$\|[F'(x^0)]^{-1}\| = \frac{1}{|\det([F'(x^0)]^{-1})|} \cdot \left(\sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij}^2 \right)^{1/2} \leq \beta_0,$$

unde A_{ij} este complementul algebric corespunzător elementului a_{ij} în matricea transpusă obţinută din matricea inversă a lui Jacobi $[F'(x^0)]^{-1}$. Pentru calculul normei diferenţialei de ordinul doi se foloseşte evaluarea

$$\|F''(x)\| \leq \left[\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\left(\frac{\partial^2 F_i(x)}{\partial x^j \partial x^k}\right)^2\right]^{1/2} \leq \gamma.$$

Pentru a demonstra teorema lui Kantorovici avem nevoie în prealabil de un şir de leme.

Fixăm spaţiul $\mathbb{R}^n$ cu $n \geq 1$ număr natural dat, cu o anumită normă dată, şi notăm cu $L(\mathbb{R}^n)$ algebra aplicaţiilor liniare (şi în mod automat continue) care aplică spaţiul $\mathbb{R}^n$ tot în $\mathbb{R}^n$. Pentru un element $T \in L(\mathbb{R}^n)$, prin norma acestui element T vom înţelege

$$\|T\| = \sup_{x \neq \theta_{\mathbb{R}^n}} \frac{\|T(x)\|}{\|x\|}.$$

Astfel $L(\mathbb{R}^n)$ este o algebră normată, chiar o algebră Banach.

Lema 7.2.1. *Dacă $T \in L(\mathbb{R}^n)$ verifică $\|I - T\| < 1$, unde I este aplicaţia identică a lui $\mathbb{R}^n$, atunci T este un element inversabil şi*

$$\|T^{-1}\| \leq \frac{1}{1 - \|I - T\|}.$$

Demonstraţie. Şirul

$$\left(\sum_{n=0}^{N}(I-T)^n\right)_{N\in\mathbb{N}}$$

este un şir Cauchy în $L(\mathbb{R}^n)$: pentru $N > M$ avem

$$\left\|\sum_{n=0}^{N}(I-T)^n - \sum_{n=0}^{M}(I-T)^n\right\| = \left\|\sum_{n=M+1}^{N}(I-T)^n\right\| \leq \sum_{n=M+1}^{N}\|(I-T)^n\| \leq$$

$$\leq \sum_{n=M+1}^{N}\|I-T\|^n = \|I-T\|^{M+1}\cdot(1+\|I-T\|+\cdots+\|I-T\|^{N-M-1}) \leq$$

$$\leq \|I-T\|^{M+1}\cdot\frac{1}{1-\|I-T\|} \to 0$$

când $M \to \infty$. Dar $L(\mathbb{R}^n)$ este un spaţiu complet, deci există

$$S = \sum_{n=0}^{\infty}\cdot(I-T)^n \in L(\mathbb{R}^n).$$

Avem

$$\begin{aligned} T \cdot S &= [I-(I-T)] \cdot \left[\sum_{n=0}^{\infty}(I-T)^n\right] = \lim_{N\to\infty}[I-(I-T)] \cdot \left[\sum_{n=0}^{N}(I-T)^n\right] = \\ &= \lim_{N\to\infty}[I-(I-T)^{N+1}] = I, \end{aligned}$$

căci $\|(I-T)^{N+1}\| \to 0$. Analog $S \cdot T = I$. Deci aplicaţia T este inversabilă în $L(\mathbb{R}^n)$ şi

$$\begin{aligned} \|T^{-1}\| = \|S\| &= \left\|\lim_{N\to\infty}\sum_{n=0}^{N}(I-T)^n\right\| = \lim_{N\to\infty}\left\|\sum_{n=0}^{N}(I-T)^n\right\| \leq \\ &\leq \lim_{N\to\infty}\sum_{n=0}^{N}\|I-T\|^n = \frac{1}{1-\|I-T\|}\ . \qquad \text{q.e.d.} \end{aligned}$$

Lema 7.2.2. *Dacă $T, S \in L(\mathbb{R}^n)$ sunt astfel încât există T^{-1} şi $\|T^{-1}\| \leq \alpha$, $\|T-S\| \leq \beta$, $\alpha \cdot \beta < 1$, atunci există S^{-1} şi $\|S^{-1}\| \leq \frac{\alpha}{1-\alpha\cdot\beta}$.*

Demonstraţie. Fie $U := T^{-1}(T-S) = I - T^{-1} \cdot S$. Atunci

$$\|I-(I-U)\| = \|U\| = \|T^{-1}(T-S)\| \leq \|T^{-1}\| \cdot \|T-S\| \leq \alpha \cdot \beta < 1.$$

Conform lemei 7.2.1 există inversa aplicaţiei $I-U$ şi

$$\|(I-U)^{-1}\| \leq \frac{1}{1-\|U\|} \leq \frac{1}{1-\alpha\beta}.$$

Prin urmare există inversa lui $S = T(I-U)$ şi $S^{-1} = (I-U)^{-1}T^{-1}$ şi

$$\|S^{-1}\| = \|(I-U)^{-1}T^{-1}\| \leq \|(I-U)^{-1}\| \cdot \|T^{-1}\| \leq \frac{\alpha}{1-\alpha\beta}. \quad \text{q.e.d.}$$

Lema 7.2.3. *Fie $G : D \subset \mathbb{R}^n \to \mathbb{R}^n$ un operator diferenţiabil Fréchet pe mulţimea deschisă $D \neq \emptyset$, iar $D_0 \subset D$ o submulţime convexă. Presupunem că derivata Fréchet a lui G are proprietatea lui Lipschitz pe D_0 : există o constantă $L > 0$ astfel ca $\|G'(x) - G'(y)\| \leq L \cdot \|x-y\|$ pentru orice $x, y \in D_0$. Atunci pentru orice $x, y \in D_0$ avem*

$$\|G(x) - G(y) - G'(x)(x-y)\| \leq \frac{L}{2} \cdot \|x-y\|^2.$$

DEMONSTRAŢIE.

$$\begin{aligned}
\|G(x)-G(y)-G'(x)(x-y)\| &= \|G(y+t(x-y))\,\big|_0^1 - G'(x)(x-y)\| = \\
&= \left\| \int_0^1 G'(y+t(x-y))(x-y)dt - \int_0^1 G'(x)(x-y)dt \right\| = \\
&= \| \int_0^1 (G'(y+t(x-y))-G'(x))(x-y)dt\| \leq \\
&\leq \int_0^1 \|G'(y+t(x-y))-G'(x)\| \cdot \|x-y\|dt \leq \\
&\leq \int_0^1 L \cdot \|y+t(x-y)-x\| \cdot \|x-y\|dt = \\
&= \int_0^1 L \cdot \|(1-t)\cdot(y-x)\| \cdot \|x-y\|dt = \\
&= L \cdot \int_0^1 |1-t| \cdot \|y-x\| \cdot \|x-y\|dt = \\
&= L \cdot \|x-y\|^2 \cdot \int_0^1 (1-t)dt = \frac{L}{2} \cdot \|x-y\|^2. \quad \text{q.e.d.}
\end{aligned}$$

DEMONSTRAŢIE. (Teorema lui Kantorovici-Ostrowski)

Să observăm în primul rând că $\|x^1 - x^0\| = \|(F'(x^0))^{-1}F(x^0)\| \leq \eta_0$ şi întrucât $\eta_0 < r_0$ rezultă că $x^1 \in B_0$, (avem următorul şir de echivalenţe: $\eta_0 < r_0 \Leftrightarrow \eta_0 < \dfrac{1-\sqrt{1-2\alpha_0}}{\alpha_0} \cdot \eta_0 \Leftrightarrow \alpha_0 < 1-\sqrt{1-2\alpha_0} \Leftrightarrow \sqrt{1-2\alpha_0} < 1-\alpha_0 \Leftrightarrow 1-2\alpha_0 < (1-\alpha_0)^2 \Leftrightarrow \alpha_0^2 > 0$). Aşadar, putem folosi condiţiile 2 şi 4 şi avem:

$$\|F'(x^1)-F'(x^0)\| \leq \gamma \cdot \|x^1-x^0\| \leq \gamma \cdot \eta_0 \leq \frac{1}{2\beta_0} < \frac{1}{\|(F'(x^0))^{-1}\|}.$$

Conform lemei 7.2.2 (prin alegerea $T = F'(x^0)$, $S = F'(x^1)$, $\alpha = \|(F'(x^0))^{-1}\|$, $\beta = \|F'(x^1)-F'(x^0)\|$) există $(F'(x^1))^{-1}$ şi

$$\|(F'(x^1))^{-1}\| \leq \frac{\|(F'(x^0))^{-1}\|}{1-\|(F'(x^0))^{-1}\| \cdot \|F'(x^1)-F'(x^0)\|} \leq \frac{\beta_0}{1-\beta_0\cdot\gamma\cdot\eta_0} = \frac{\beta_0}{1-\alpha_0}.$$

Definim $\beta_1 = \dfrac{\beta_0}{1-\alpha_0}$ şi atunci vom avea $\|(F'(x^1))^{-1}\| \leq \beta_1$. Aceasta înseamnă că iteraţia următoare x^1 satisface o relaţie de forma 1. Vom căuta să arătăm că sunt satisfăcute relaţii analoage şi cu 3, 4 şi 5. Să observăm că $(F'(x^0))^{-1}F'(x^1)$ verifică

$$\begin{aligned}
\|I-(F'(x^0))^{-1}F'(x^1)\| &= \|(F'(x^0))^{-1}(F'(x^0)-F'(x^1))\| \leq \\
&\leq \|(F'(x^0))^{-1}\| \cdot \|F'(x^0)-F'(x^1)\| \leq \\
&\leq \beta_0\cdot\gamma\cdot\eta_0 = \alpha_0 \leq \frac{1}{2} < 1,
\end{aligned}$$

deci conform lemei 7.2.1 (cu alegerea $T = (F'(x^0))^{-1}F'(x^1)$ există $[(F'(x^0))^{-1}F'(x^1)]^{-1} = (F'(x^1))^{-1} \cdot F'(x^0)$, deci $(F'(x^1))^{-1} = [(F'(x^0))^{-1}F'(x^1)]^{-1} \cdot (F'(x^0))^{-1}$. Prin urmare $(F'(x^1))^{-1}F(x^1) = [(F'(x^0))^{-1} \cdot F'(x^1)]^{-1} \cdot (F'(x^0))^{-1} \cdot F(x^1)$, deci

$$\begin{aligned} \|(F'(x^1))^{-1} \cdot F(x^1)\| &\leq \|((F'(x^0))^{-1} \cdot F'(x^1))^{-1}\| \cdot \|(F'(x^0))^{-1}F(x^1)\| \leq \\ &\leq \frac{1}{1-\alpha_0} \cdot \|(F'(x^0))^{-1}F(x^1)\| \end{aligned}$$

conform lemei 7.2.1 (alegând $T = (F'(x^0))^{-1}F'(x^1)$ şi $\alpha = \alpha_0$). Pentru a obţine o majorare pentru $\|(F'(x^0))^{-1}F(x^1)\|$ să introducem operatorul $G(x) = x - (F'(x^0))^{-1}F(x)$ şi să observăm că G este diferenţiabil Fréchet pe D_0 şi $G'(x) = I - (F'(x^0))^{-1}F'(x)$, unde cu I am notat matricea unitate de ordinul n sau aplicaţia identică pe spaţiul $\mathbb{R}^n$. Aşadar

$$\begin{aligned} \|G'(x) - G'(y)\| &= \|(F'(x^0))^{-1}(F'(x) - F'(y))\| \leq \|(F'(x^0))^{-1}\| \cdot \|F'(x) - F'(y)\| \leq \\ &\leq \beta_0 \cdot \gamma \cdot \|x - y\|. \end{aligned}$$

Întrucât $G'(x^0) = O_n$, matricea nulă de ordinul n, aplicăm lema 7.2.3 şi avem:

$$\begin{aligned} \|(F'(x^0))^{-1}F(x^1)\| &= \|G(x^1) - G(x^0) - G'(x^0)(x^1 - x^0)\| \leq \frac{1}{2}\beta_0 \cdot \gamma \cdot \|x^1 - x^0\|^2 \leq \\ &\leq \frac{1}{2}\,\beta_0 \cdot \gamma \cdot \eta_0^2 = \frac{1}{2}\,\alpha_0 \cdot \eta_0. \end{aligned}$$

Dacă notăm cu $\eta_1 = \dfrac{1}{2} \cdot \dfrac{\alpha_0}{1-\alpha_0} \cdot \eta_0$, atunci $\|(F'(x^1))^{-1}F(x^1)\| \leq \dfrac{1}{1-\alpha_0} \cdot \dfrac{1}{2} \cdot \alpha_0 \cdot \eta_0 = \eta_1$. Notând cu

$$\alpha_1 = \beta_1 \cdot \gamma \cdot \eta_1 = \frac{\beta_0}{1-\alpha_0} \cdot \gamma \cdot \frac{1}{2} \cdot \frac{\alpha_0}{1-\alpha_0} \cdot \eta_0 = \frac{1}{2} \cdot \left(\frac{\alpha_0}{1-\alpha_0}\right)^2 \leq \frac{1}{2}$$

(avem următorul şir de echivalenţe: $\left(\dfrac{\alpha_0}{1-\alpha_0}\right)^2 \leq 1 \Leftrightarrow \dfrac{\alpha_0}{1-\alpha_0} \leq 1 \Leftrightarrow \alpha_0 \leq 1 - \alpha_0 \Leftrightarrow \alpha_0 \leq \dfrac{1}{2}$). Să considerăm sfera $B_1 = B(x^1, r_1)$ unde $r_1 = \dfrac{1 - \sqrt{1-2\alpha_1}}{\alpha_1} \cdot \eta_1$. Înlocuind pe α_1, η_1 cu expresiile lor în funcţie de α_0, η_0, obţinem că

$$\begin{aligned} r_1 &= \frac{1 - \sqrt{1 - 2 \cdot \dfrac{1}{2} \cdot \left(\dfrac{\alpha_0}{1-\alpha_0}\right)^2}}{\dfrac{1}{2} \cdot \left(\dfrac{\alpha_0}{1-\alpha_0}\right)^2} \cdot \frac{1}{2} \cdot \frac{\alpha_0}{1-\alpha_0} \cdot \eta_0 = \frac{1 - \dfrac{\sqrt{(1-\alpha_0)^2 - \alpha_0^2}}{1-\alpha_0}}{\dfrac{1}{2} \cdot \dfrac{\alpha_0}{1-\alpha_0}} \cdot \eta_0 = \\ &= \frac{1 - \alpha_0 - \sqrt{1-2\alpha_0}}{\alpha_0} \cdot \eta_0 = \frac{1 - \sqrt{1-2\alpha_0}}{\alpha_0} \cdot \eta_0 - \eta_0 = r_0 - \eta_0. \end{aligned}$$

Să arătăm că $B_1 \subset B_0$. Fie $x \in B_1$, deci $\|x - x^1\| \leq r_1$. Prin urmare

$$\|x - x^0\| \leq \|x - x^1\| + \|x^1 - x^0\| \leq r_1 + \eta_0 = r_0 - \eta_0 + \eta_0 = r_0,$$

adică $x \in B_0$ şi $B_1 \subset B_0$.

Prin inducţie matematică (trecerea de la k la $k+1$ este absolut similară cu trecerea de la 0 la 1) se poate arăta că pentru orice k, $(F'(x^k))^{-1}$ există şi $\|(F'(x^k))^{-1}\| \leq \beta_k$, $\|(F'(x^k))^{-1}F(x^k)\| \leq \eta_k$, $\alpha_k = \beta_k\gamma\eta_k \leq \frac{1}{2}$, $B_k = B(x^k, r^k) \subset B_{k-1}$, unde $\beta_k = \frac{\beta_{k-1}}{1-\alpha_{k-1}}$, $\eta_k = \frac{1}{2} \cdot \frac{\alpha_{k-1}}{1-\alpha_{k-1}} \cdot \eta_{k-1}$, $r_k = \frac{1-\sqrt{1-2\alpha_k}}{\alpha_k} \cdot \eta_k$. Din aceste relaţii rezultă că şirul $(x^k)_{k\in\mathbb{N}}$ dat de metoda lui Newton este bine definit şi aparţine lui B_0. Pentru convergenţa lui, să observăm mai întâi că

$$\alpha_k = \frac{1}{2} \cdot \frac{\alpha_{k-1}^2}{(1-\alpha_{k-1})^2} \leq \frac{1}{2} \cdot (2 \cdot \alpha_{k-1})^2,$$

deci $2\alpha_k \leq (2\alpha_{k-1})^2 \leq (2\alpha_{k-2})^{2^2} \leq \cdots \leq (2\alpha_0)^{2^k}$. Prin urmare $\alpha_k \leq 2^{-1} \cdot (2\alpha_0)^{2^k}$. Pe de altă parte

$$\begin{aligned} \eta_k &= \frac{1}{2} \cdot \frac{\alpha_{k-1}}{1-\alpha_{k-1}} \cdot \eta_{k-1} \leq \alpha_{k-1} \cdot \eta_{k-1} \leq \alpha_{k-1} \cdot \alpha_{k-2} \cdot \eta_{k-2} \leq \cdots \leq \\ &\leq \alpha_{k-1}\alpha_{k-2}\ldots\alpha_0 \cdot \eta_0 \leq 2^{-k} \cdot (2\alpha_0)^{1+2+2^2+\cdots+2^{k-1}} \cdot \eta_0 = 2^{-k} \cdot (2\alpha_0)^{2^k-1} \cdot \eta_0, \end{aligned}$$

de unde rezultă că $\eta_k \to 0$ pentru $k \to \infty$. Prin calcul direct obţinem $r_k \leq 2\eta_k$ ($r_k = \frac{1-\sqrt{1-2\alpha_k}}{\alpha_k} \cdot \eta_k \leq 2\eta_k \Leftrightarrow 1 - \sqrt{1-2\alpha_k} \leq 2\alpha_k \Leftrightarrow 1 - 2\alpha_k \leq \sqrt{1-2\alpha_k} \Leftrightarrow \sqrt{1-2\alpha_k} \leq 1 \Leftrightarrow 1 - 2\alpha_k \leq 1 \Leftrightarrow \alpha_k \geq 0$). Prin urmare $\{B_k\}_{k\in\mathbb{N}}$ este un şir descendent de mulţimi cu diametrul tinzând la zero. Conform teoremei lui Cantor există un singur punct x^* care aparţine la toate mulţimile B_k iar şirul $(x^k)_{k\in\mathbb{N}}$ (centrele sferelor) converge la x^*.

Pentru a arăta că x^* este soluţia ecuaţiei $F(x) = \theta_{\mathbb{R}^n}$, să observăm că deoarece funcţionala $\|F'(x)\|$ este continuă pe B_0 (din condiţia 2 de lipschitzianitate a lui F' rezultă imediat continuitatea lui F'), rezultă că este mărginită pe B_0, $\|F'(x)\| \leq M$ pentru orice $x \in B_0$ şi

$$\|F(x^k)\| = \|F'(x^k)(F'(x^k))^{-1}F(x^k)\| \leq \|F'(x^k)\| \cdot \|(F'(x^k))^{-1}F(x^k)\| \leq M \cdot \eta_k,$$

de unde $F(x^k) \to \theta_{\mathbb{R}^n}$ pentru $k \to \infty$ şi $F(x^*) = \theta_{\mathbb{R}^n}$, căci F este continuă. În sfârşit eroarea la iteraţia k se obţine din $\eta_k \leq 2^{-k} \cdot (2\alpha_0)^{2^k-1} \cdot \eta_0$, întrucât $x^* \in B_k$ şi $\|x^k - x^*\| \leq r_k \leq 2\eta_k \leq 2^{1-k} \cdot (2\alpha_0)^{2^k-1} \cdot \eta_0$. $\square$

Menţionăm că metoda lui Newton-Raphson are dezavantajul că este o metodă locală, fiindcă contează foarte mult poziţia punctului de pornire x^0, care trebuie să fie "suficient de aproape" de soluţia căutată x^*. Altfel se poate întâmpla ca şirul iterativ obţinut cu metoda lui Newton să diveargă sau dacă chiar este convergent să conveargă la o altă soluţie a sistemului iniţial, diferită de soluţia căutată x^*. Avantajul acestei metode este că este o metodă rapid convergentă având ordinul de convergenţă doi.

Dacă vrem să determinăm soluţia x^* cu o precizie $\varepsilon > 0$ dinainte dată, atunci este de ajuns să impunem condiţia $2^{1-k} \cdot (2\alpha_0)^{2^k-1} \cdot \eta_0 \leq \varepsilon$ de unde se determină prima dată indicele $k_{(\varepsilon)}$. Pentru acest indice avem $\|x_{k_{(\varepsilon)}} - x^*\| \leq 2^{1-k_{(\varepsilon)}} \cdot (2\alpha_0)^{2^{k_{(\varepsilon)}}-1} \cdot \eta_0 \leq \varepsilon$, deci termenul $x_{k_{(\varepsilon)}}$, din şirul iterativ $(x_k)_{k\in\mathbb{N}}$ generat de noi se acceptă ca soluţia sistemului iniţial cu o precizie ε, fiindcă eroarea absolută este $\|x_{k_{(\varepsilon)}} - x^*\| \leq \varepsilon$.

În final prezentăm un algoritm pentru metoda lui Newton-Raphson cu o condiţie practică de oprire:

Algoritm Newton

Datele de intrare: n; $F = (F_1, F_2, \ldots, F_n)$; x^0; ε;
Fie $y := x^0$;
Repetă $x := y$;
$y := \Phi(x)$;
Până când $\|y - x\| \geq \varepsilon$;
Tipăreşte y.

Menţionăm că atribuirile din acest program se referă la vectori, deci se fac pe componente. De exemplu: $y := x^0$ înseamnă: pentru $i = \overline{1,n}$ execută: $y[i] := x^0[i]$. Funcţia Φ se numeşte funcţia iterativă a metodei lui Newton în cazul spaţiului finit dimensional $\mathbb{R}^n$ şi este dată de formula: $\Phi(x) = x - (F'(x))^{-1}F(x)$. Pentru calculul lui Φ avem de efectuat următorii paşi: se determină elementele matricii $F'(x)$, adică numerele $\frac{\partial F_i}{\partial x_j}(x)$, pentru $i, j = \overline{1,n}$. Cu aceste numere se formează matricea jacobiană $F'(x)$, la care se ia matricea inversă $(F'(x))^{-1}$, care se înmulţeşte cu vectorul coloană $F(x)$ şi acest rezultat scădem din vectorul coloană x. Din punct de vedere numeric pentru calculul lui $\frac{\partial F_i}{\partial x_j}(x)$ se foloseşte următoarea formulă de derivare numerică:

$$\frac{\partial F_i}{\partial x_j}(x) \approx \frac{F_i(x + he_j) - F_i(x)}{h},$$

unde cu $(e_i)_{i=\overline{1,n}}$ notăm baza canonică a spaţiului $\mathbb{R}^n$, iar $h = 10^{-2}, 10^{-3}, \ldots$ (vezi capitolul de formule de derivare numerică pentru funcţii reale de variabilă reală paragraful 9.1). Pentru calculul matricii inverse a matricii jacobiene $F'(x)$, putem aplica algoritmul de calcul al matricii inverse folosind ca subrutină algoritmul lui Gauss (vezi paragraful 6.1.3.2).

7.3 Metoda gradientului pe $\mathbb{R}^n$

Problema matematică propusă este rezolvarea aproximativă a sistemului de ecuaţii neliniare:

$$\begin{cases} F_1(x_1, x_2, \ldots, x_n) = 0 \\ F_2(x_1, x_2, \ldots, x_n) = 0 \\ \vdots \\ F_n(x_1, x_2, \ldots, x_n) = 0, \end{cases}$$

unde $F_1, F_2, \ldots, F_n : D \subset \mathbb{R}^n \to \mathbb{R}$ sunt funcţii date, $D \neq \emptyset$, iar $F = (F_1, F_2, \ldots, F_n) : D \subset \mathbb{R}^n \to \mathbb{R}^n$. Introducem funcţia $G : D \to \mathbb{R}$, $G(x) = \|F(x)\|$, unde $\|\cdot\|$ este o normă arbitrară pe spaţiul $\mathbb{R}^n$. Dacă $x^* \in D$ este soluţia ecuaţiei $F(x) = \theta_{\mathbb{R}^n}$, adică $F(x^*) = \theta_{\mathbb{R}^n}$, atunci $G(x^*) = \|F(x^*)\| = \|\theta_{\mathbb{R}^n}\| = 0$ şi invers, dacă $G(x^*) = 0$ rezultă că $\|F(x^*)\| = 0$ de unde obţinem că $F(x^*) = \theta_{\mathbb{R}^n}$. Deoarece funcţia G ia numai valori pozitive din cauza normei, ajungem la concluzia că soluţia $x^* \in D$ a sistemului este punct de minim pentru funcţia G şi invers, orice punct $x^* \in D$ pentru care G se anulează va fi soluţie pentru ecuaţia operatorială $F(x) = \theta_{\mathbb{R}^n}$. Într-un caz particular putem să alegem norma euclidiană pe spaţiul $\mathbb{R}^n$ şi atunci $G(x) = \sqrt{\sum_{i=1}^{n} F_i^2(x)}$. A obţine minimul lui G este echivalent de a obţine minimul lui $G^2(x) = \sum_{i=1}^{n} F_i^2(x)$. Prin urmare $x^* \in D$ este soluţia ecuaţiei $F(x) = \theta_{\mathbb{R}^n}$ dacă şi numai dacă x^* este acel punct de minim al lui G^2 pentru care G^2 se anulează.

Fie deci $\mathcal{H} : D \subset \mathbb{R}^n \to \mathbb{R}$, $D \neq \emptyset$ o funcţie pentru care vrem să determinăm punctul minim. Un caz fericit este când $\mathcal{H}$ admite un singur punct de minim global pe D. Acest caz are loc de exemplu când graficul lui $\mathcal{H}$ este convex. Pentru a determina punctul

minim al lui $\mathcal{H}$ alegem un punct de pornire $x^0 \in D$ şi o direcţie $h_0 \in \mathbb{R}^n$. În cazul metodei gradientului, (sau metoda celei mai rapide descreşteri) se alege $h_0 = -\operatorname{grad}\mathcal{H}(x^0)$. Se cunoaşte din analiza matematică că cea mai rapidă descreştere a lui $\mathcal{H}$ se obţine în punctul x_0 tocmai în direcţia - grad $\mathcal{H}(x_0)$. Deplasarea în direcţia aleasă se face conform condiţiei de a obţine minimul:

$$\min\{\mathcal{H}(x^0 + t \cdot h_0) \mid t \geq 0\} = \min\{\mathcal{H}(x^0 - t \cdot \operatorname{grad}\mathcal{H}(x^0)) \mid t \geq 0\}$$

Fie t_0 pentru care se obţine minimul dorit. Atunci în punctul $x^1 = x^0 + t_0 \cdot h_0 = x^0 - t_0 \cdot \operatorname{grad}\mathcal{H}(x^0)$ avem cu siguranţă $\mathcal{H}(x^1) \leq \mathcal{H}(x^0)$. În continuare se alege în punctul x^1 direcţia $h_1 = -\operatorname{grad}\mathcal{H}(x^1)$ şi se determină

$$t_1 = \min\{\mathcal{H}(x^1 + t \cdot h_1) \mid t \geq 0\} = \min\{\mathcal{H}(x^1 - t \cdot \operatorname{grad}\mathcal{H}(x^1)) \mid t \geq 0\}.$$

Aşadar

$$\mathcal{H}(x^2) = \mathcal{H}(x^1 + t_1 \cdot h_1) = \mathcal{H}(x^1 - t_1 \cdot \operatorname{grad}\mathcal{H}(x^1)) \leq \mathcal{H}(x^1).$$

În general, dacă am determinat punctul $x^n \in D$, atunci următorul punct

$$x^{n+1} = x^n + t_n \cdot h_n = x^n - t_n \cdot \operatorname{grad}\mathcal{H}(x^n),$$

unde se alege direcţia $h_n = -\operatorname{grad}\mathcal{H}(x^n)$ iar

$$t_n = \min\{\mathcal{H}(x^n + t \cdot h_n) \mid t \geq 0\} = \min\{\mathcal{H}(x^n - t \cdot \operatorname{grad}\mathcal{H}(x^n)) \mid t \geq 0\}.$$

Menţionăm faptul că metoda gradientului în final se reduce la minimizarea unui şir de funcţii reale cu variabilă reală $G_n : [0, +\infty) \to \mathbb{R}$,

$$G_n(t) = \mathcal{H}(x^n + t \cdot h_n) = \mathcal{H}(x^n - t \cdot \operatorname{grad}\mathcal{H}(x^n)).$$

Totodată presupunem că vectorul $x^n + t \cdot h_n \in D$, când se caută minimul lui G_n în funcţie de $t \geq 0$.

Algoritmul se opreşte când pentru un $n \in \mathbb{N}$ suficient de mare $\|x^{n+1} - x^n\| \leq \varepsilon$ sau când $|\mathcal{H}(x^{n+1}) - \mathcal{H}(x^n)| \leq \varepsilon$, unde $\varepsilon > 0$ este o precizie dinainte dată.

Capitolul 8

Aproximarea funcţiilor

Fie (A, ρ) un spaţiu metric, iar $B \subset A$ o parte a lui A. Vom spune că elementele lui A se aproximează cu elementele lui B, dacă pentru orice $\varepsilon > 0$ şi orice $f \in A$ există $g \in B$ astfel încât $\rho(f, g) \leq \varepsilon$.

În analiza matematică această proprietate se exprimă în felul următor: submulţimea B este densă în mulţimea A. De exemplu, dacă alegem $A = \mathbb{R}$, $B = \mathbb{Q}$, iar distanţa $\rho : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ este dată de formula $\rho(x, y) = |x - y|$ pentru orice $x, y \in \mathbb{R}$, atunci obţinem proprietatea binecunoscută, că numerele raţionale sunt dense pe axa reală.

În general, dacă A este un spaţiu de funcţii, iar B o parte a acestui spaţiu formată din funcţii de o "structură mai simplă", atunci vom spune că funcţia $f \in A$, de o "structură mai complexă" se aproximează cu funcţia $g \in B$ de o "structură mai simplă" în raport cu metrica ρ.

8.1 Aproximarea uniformă a funcţiilor continue cu ajutorul polinoamelor

8.1.1 Teorema lui Weierstrass şi teorema lui Korovkin

Vom nota cu $C([0, 1], \mathbb{R}) = \{f : [0, 1] \to \mathbb{R} \mid f \text{ este continuă}\}$ spaţiul funcţiilor continue cu valori reale definite pe intervalul $[0, 1]$. Pe acest spaţiu se defineşte adunarea punctuală a două funcţii: $\forall\ f, g \in C([0, 1], \mathbb{R})$ avem $(f + g) : [0, 1] \to \mathbb{R}$, $(f + g)(x) = f(x) + g(x)$ $\forall\ x \in [0, 1]$, şi înmulţirea unei funcţii cu scalari: $\forall\ f \in C([0, 1], \mathbb{R})$ şi $\forall\ \alpha \in \mathbb{R}$ avem

$(\alpha \cdot f) : [0,1] \to \mathbb{R}$, $(\alpha \cdot f)(x) = \alpha \cdot f(x)\ \forall\ x \in [0,1]$. Se ştie că suma a două funcţii continue va fi tot o funcţie continuă, şi dacă o funcţie continuă se înmulţeşte cu un scalar se obţine tot o funcţie continuă.

Lema 8.1.1. *Spaţiul $C([0,1],\mathbb{R})$ este un spaţiu liniar (vectorial) real.*

DEMONSTRAŢIE. Se verifică cu uşurinţă axiomele spaţiului liniar. Menţionăm că elementul neutru faţă de adunare este funcţia $\theta : [0,1] \to \mathbb{R}$, $\theta(x) = 0$ pentru $\forall\ x \in [0,1]$, iar elementul simetric faţă de adunare este funcţia $(-f) : [0,1] \to \mathbb{R}$, $(-f)(x) = -f(x)$ pentru $\forall\ x \in [0,1]$. q.e.d.

Pe spaţiul $C([0,1],\mathbb{R})$ se defineşte o normă $\|\cdot\|_\infty : C([0,1],\mathbb{R}) \to \mathbb{R}$ în felul următor: $\forall\ f \in C([0,1],\mathbb{R})$ fie $\|f\|_\infty = \sup\{|f(x)|\ /\ x \in [0,1]\}$. Deoarece funcţia f este continuă pe intervalul compact [0, 1], conform teoremei lui Weierstrass este mărginită şi îşi atinge marginile. Prin urmare

$$\|f\|_\infty = \sup\{|f(x)|\ /\ x \in [0,1]\} = \max\{|f(x)|\ /\ x \in [0,1]\}.$$

Lema 8.1.2. *Spaţiul $(C([0,1],\mathbb{R}), \|\cdot\|_\infty)$ este un spaţiu liniar normat.*

DEMONSTRAŢIE. Trebuie să verificăm că funcţia $\|\cdot\|_\infty : C([0,1],\mathbb{R}) \to \mathbb{R}$ satisface axiomele normei. Într-adevăr, pentru $f, g \in C([0,1],\mathbb{R})$ şi orice $x \in [0,1]$ avem

$$\begin{aligned} |(f+g)(x)| &= |f(x)+g(x)| \le |f(x)| + |g(x)| \le \\ &\le \max\{|f(x)|\ /\ x \in [0,1]\} + \max\{|g(x)|\ /\ x \in [0,1]\} = \|f\|_\infty + \|g\|_\infty. \end{aligned}$$

Prin urmare:

$$\begin{aligned} \|f+g\|_\infty &= \max\{|(f+g)(x)|\ /\ x \in [0,1]\} \le \|f\|_\infty + \|g\|_\infty. \\ \|\alpha \cdot f\| &= \max\{|(\alpha \cdot f)(x)|\ /\ x \in [0,1]\} = \max\{|\alpha \cdot f(x)|\ /\ x \in [0,1]\} = \\ &= \max\{|\alpha| \cdot |f(x)|\ /\ x \in [0,1]\} = |\alpha| \cdot \max\{|f(x)|\ /\ x \in [0,1]\} = |\alpha| \cdot \|f\|_\infty. \end{aligned}$$

În final avem $\|f\|_\infty = \max\{|f(x)|\ /\ x \in [0,1]\} \ge 0$ şi $\|f\|_\infty = 0$ implică că $|f(x)| = 0$ pentru orice $x \in [0,1]$, deci $f = \theta$. q.e.d.

Norma $\|\cdot\|_\infty$ definită pe $(C[0,1],\mathbb{R})$ se numeşte norma Cebâşev, sau norma supremum sau norma maximum. Se ştie că orice normă induce o metrică. Vom defini metrica $\rho_\infty : C([0,1],\mathbb{R}) \times C([0,1],\mathbb{R}) \to \mathbb{R}$ prin formula $\rho_\infty(f,g) = \|f-g\|_\infty$ pentru orice $f, g \in C([0,1],\mathbb{R})$.

Lema 8.1.3. *Spaţiul* $(C([0,1],\mathbb{R}),\rho_\infty)$ *este un spaţiu metric.*

DEMONSTRAŢIE. Trebuie să verificăm că funcţia ρ_∞ satisface axiomele metricii. Într-adevăr:

$$\begin{aligned}\rho_\infty(f,g) &= \|f-g\|_\infty = \|(f-h)+(h-g)\|_\infty \le \|f-h\|_\infty + \|h-g\|_\infty = \\ &= \rho_\infty(f,h)+\rho_\infty(h,g) \quad \text{pentru orice} \quad f,g,h \in C([0,1],\mathbb{R}),\end{aligned}$$

iar $\rho_\infty(f,g)=\rho_\infty(g,f)$ şi $\rho_\infty(f,g)\ge 0$ şi $\rho_\infty(f,g)=0 \Leftrightarrow f=g$ sunt uşor de arătat. q.e.d.

Metrica ρ_∞ definită pe $C([0,1],\mathbb{R})$ se numeşte metrica Cebâşev sau metrica supremum sau metrica maximum.

Observaţia 8.1.1. *În analiza matematică se arată mai mult, şi anume că spaţiul* $(C([0,1],\mathbb{R}),\rho_\infty)$ *este un spaţiu Banach, adică orice şir Cauchy sau fundamental este convergent în acest spaţiu. Într-adevăr, fie* $(f_n)_{n\in\mathbb{N}} \subset (C[0,1],\mathbb{R})$ *un şir fundamental de funcţii continue, adică pentru orice* $\varepsilon>0$ *există un indice* $n=n_{(\varepsilon)}\in\mathbb{N}$ *astfel ca pentru orice* $n,m\in\mathbb{N}$ *cu* $n,m\ge n_{(\varepsilon)}$ *să avem* $\rho(f_n,f_m)\le\varepsilon$. *În acest caz va exista o funcţie continuă* $f:[0,1]\to\mathbb{R}$ *astfel încât pentru orice* $\varepsilon>0$ *există un indice* $n_{(\varepsilon)}\in\mathbb{N}$ *pentru care dacă* $n\ge n_{(\varepsilon)}$, *avem* $\rho(f_n,f)\le\varepsilon$.

Observaţia 8.1.2. *Convergenţa şirului de funcţii continue* $(f_n)_{n\in\mathbb{N}}$ *către* $f\in(C[0,1],\mathbb{R})$ *se mai numeşte şi convergenţa uniformă pe intervalul [0, 1]. Din punct de vedere geometric distanţa între două funcţii* $f,g\in(C[0,1],\mathbb{R})$ *are următoarea semnificaţie:*

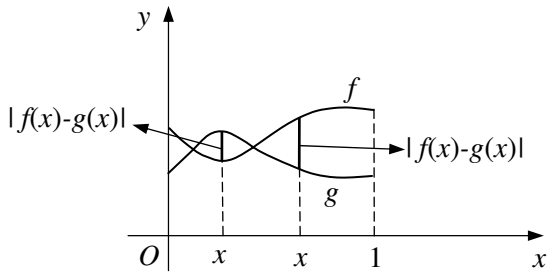


Figura 8.1:

Vom avea că $\rho_\infty(f,g)\le\varepsilon$ *cu* $\varepsilon>0$, *dacă pentru orice* $x\in[0,1]$ *avem* $|f(x)-g(x)|\le\varepsilon$. *Aşadar faptul că şirul de funcţii continue* $(f_n)_{n\in\mathbb{N}}$ *converge la funcţia* $f\in(C[0,1],\mathbb{R})$ *uniform în raport cu metrica* ρ_∞ *are următoarea interpretare geometrică:*

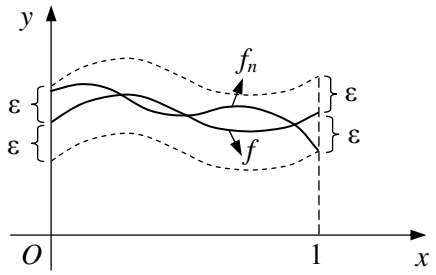


Figura 8.2:

Pentru orice $\varepsilon > 0$ există un indice $n_{(\varepsilon)} \in \mathbb{N}$ astfel ca orice funcţie continuă f_n cu $n \geq n_{(\varepsilon)}$ se află în banda de rază ε desenată în jurul funcţiei f.

Lema 8.1.4. *Orice funcţie continuă $f : [0,1] \to \mathbb{R}$ este uniform continuă, adică pentru orice $\varepsilon > 0$ există $\delta_{(\varepsilon)} > 0$ astfel încât pentru orice $x', x'' \in [0,1]$ cu $|x' - x''| < \delta_{(\varepsilon)}$ rezultă că $|f(x') - f(x'')| < \varepsilon$.*

Demonstraţie. Demonstraţia vom da prin metoda reducerii la absurd. Presupunem contrariul, adică există $\varepsilon > 0$ astfel încât pentru orice $\delta > 0$ există $x'_\delta, x''_\delta \in [0,1]$ cu $|x'_\delta - x''_\delta| < \delta$ şi $|f(x'_\delta) - f(x''_\delta)| > \varepsilon$. Alegem $\delta = \dfrac{1}{n}$, $n \in \mathbb{N}^*$. Prin urmare există două şiruri $\{x'_n\}_{n\in\mathbb{N}^*}$ şi $\{x''_n\}_{n\in\mathbb{N}^*}$ astfel ca $|x'_n - x''_n| < \dfrac{1}{n}$ şi $|f(x'_n) - f(x''_n)| > \varepsilon$.

Deoarece şirul $\{x'_n\}_{n\in\mathbb{N}^*} \subset [0,1]$ este mărginit, conform teoremei lui Bolzano admite un subşir $\{x'_{n_k}\}_{k\in\mathbb{N}^*}$ convergent, adică $x'_{n_k} \to x^* \in [0,1]$ pentru $k \to \infty$. Trecând la limită în inegalitatea

$$|x''_{n_k} - x^*| \leq |x''_{n_k} - x'_{n_k}| + |x'_{n_k} - x^*| \leq \frac{1}{n_k} + |x'_{n_k} - x^*|$$

pentru $k \to \infty$, obţinem că $x''_{n_k} \to x^*$ când $k \to \infty$. Acum trecem la limită în inegalitatea $|f(x'_{n_k}) - f(x''_{n_k})| > \varepsilon > 0$ pentru $k \to \infty$ şi folosind faptul că f este continuă rezultă $|f(x^*) - f(x^*)| = 0 > \varepsilon$, ceea ce reprezintă o contradicţie. q.e.d.

Teorema 8.1.1. *(Weierstrass) Orice funcţie continuă $f \in C([0,1], \mathbb{R})$ se poate reprezenta ca limita unui şir de polinoame $P_n : [0,1] \to \mathbb{R}$ în raport cu metrica supremum, adică pentru orice $\varepsilon > 0$ există un polinom $P : [0,1] \to \mathbb{R}$ astfel ca $\rho_\infty(f, P) \leq \varepsilon$.*

Demonstraţie. Pentru teorema lui Weierstrass vom da o demonstraţie constructivă

folosind polinoamele lui Bernstein: $B_n(f,\cdot):[0,1]\to\mathbb{R}$,

$$B_n(f,x)=\sum_{k=0}^{n}C_n^k\cdot x^k\cdot(1-x)^{n-k}\cdot f\left(\frac{k}{n}\right).$$

Se observă imediat că făcând calculele se obţine un polinom de grad n în x, şi vom arăta că şirul de polinoame $B_n(f,\cdot)$ tinde uniform către f.

Pasul 1. La început vom demonstra acest fapt pentru trei funcţii particulare apoi pentru o funcţie continuă arbitrară f.

Fie $f_0:[0,1]\to\mathbb{R}$, $f_0(x)=1$. Atunci

$$B_n(f_0,x)=\sum_{k=0}^{n}C_n^kx^k(1-x)^{n-k}\cdot f_0\left(\frac{k}{n}\right)=\sum_{k=0}^{n}C_n^kx^k\cdot(1-x)^{n-k}=[x+(1-x)]^n=1,$$

unde am folosit formula binomului lui Newton. Deoarece $B_n(f_0,x)=1=f_0(x)$ avem că şirul de polinoame $B_n(f_0,\cdot)$ tinde uniform către f_0.

Fie $f_1:[0,1]\to\mathbb{R}$, $f_1(x)=x$. Atunci

$$\begin{aligned}
B_n(f_1,x) &= \sum_{k=0}^{n}C_n^k\cdot x^k\cdot(1-x)^{n-k}\cdot f_1\left(\frac{k}{n}\right)=\sum_{k=0}^{n}C_n^k\cdot x^k\cdot(1-x)^{n-k}\cdot\frac{k}{n}=\\
&= \sum_{k=1}^{n}C_n^k\cdot\frac{k}{n}\cdot x^k\cdot(1-x)^{n-k}=\sum_{k=1}^{n}\frac{n!}{k!(n-k)!}\cdot\frac{k}{n}\cdot x^k\cdot(1-x)^{n-k}=\\
&= \sum_{k=1}^{n}\frac{(n-1)!}{(k-1)!(n-k)!}\cdot x^k\cdot(1-x)^{n-k}=\sum_{k=1}^{n}C_{n-1}^{k-1}\cdot x^k\cdot(1-x)^{n-k}=\\
&= x\cdot\sum_{k=1}^{n}C_{n-1}^{k-1}\cdot x^{k-1}\cdot(1-x)^{(n-1)-(k-1)}=x\cdot[x+(1-x)]^{n-1}=x=f_1(x).
\end{aligned}$$

Prin urmare $B_n(f_1,\cdot)$ tinde uniform către f_1.

Fie $f_2:[0,1]\to\mathbb{R}$, $f_2(x)=x^2$. Atunci

$$\begin{aligned}
B_n(f_2,x) &= \sum_{k=0}^{n}C_n^k\cdot x^k\cdot(1-x)^{n-k}\cdot f_2\left(\frac{k}{n}\right)=\sum_{k=0}^{n}C_n^k\cdot x^k\cdot(1-x)^{n-k}\cdot\left(\frac{k}{n}\right)^2=\\
&= \sum_{k=1}^{n}C_n^k\cdot\frac{k^2}{n^2}\cdot x^k\cdot(1-x)^{n-k}=\sum_{k=1}^{n}\frac{n!}{k!(n-k)!}\cdot\frac{k^2}{n^2}\cdot x^k\cdot(1-x)^{n-k}=
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!(n-k)!} \cdot \frac{k}{n} \cdot x^k \cdot (1-x)^{n-k} = \\
&= \frac{n-1}{n} \cdot \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)! \cdot (n-k)!} \cdot \left(\frac{k-1}{n-1} + \frac{1}{n-1} \right) \cdot x^k \cdot (1-x)^{n-k} = \\
&= \frac{n-1}{n} \cdot \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!(n-k)!} \cdot \frac{k-1}{n-1} \cdot x^k \cdot (1-x)^{n-k} + \\
&\quad + \frac{n-1}{n} \cdot \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!(n-k)!} \cdot \frac{1}{n-1} \cdot x^k \cdot (1-x)^{n-k} = \\
&= \frac{n-1}{n} \cdot \sum_{k=2}^{n} \frac{(n-1)!}{(k-1)!(n-k)!} \cdot \frac{k-1}{n-1} \cdot x^k \cdot (1-x)^{n-k} + \\
&\quad + \frac{1}{n} \cdot \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!(n-k)!} \cdot x^k \cdot (1-x)^{n-k} = \\
&= \frac{n-1}{n} \cdot \sum_{k=2}^{n} \frac{(n-2)!}{(k-2)!(n-k)!} \cdot x^k \cdot (1-x)^{n-k} + \\
&\quad + \frac{1}{n} \cdot \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!(n-k)!} \cdot x^k \cdot (1-x)^{n-k} = \\
&= \frac{n-1}{n} \cdot x^2 \cdot \sum_{k=2}^{n} C_{n-2}^{k-2} \cdot x^{k-2} \cdot (1-x)^{(n-2)-(k-2)} + \\
&\quad + \frac{1}{n} \cdot x \cdot \sum_{k=1}^{n} C_{n-1}^{k-1} \cdot x^{k-1} \cdot (1-x)^{(n-1)-(k-1)} = \\
&= \frac{n-1}{n} \cdot x^2 \cdot [x + (1-x)]^{n-2} + \frac{1}{n} \cdot x \cdot [x + (1-x)]^{n-1} = \\
&= \frac{n-1}{n} \cdot x^2 + \frac{1}{n} \cdot x = x^2 + \frac{x - x^2}{n} = f_2(x) + \frac{x - x^2}{n}.
\end{aligned}
$$

Deoarece

$$\left| \frac{x - x^2}{n} \right| \leq \frac{x + x^2}{n} \leq \frac{2}{n}$$

pentru orice $x \in [0,1]$ rezultă că $B_n(f_2, \cdot)$ converge uniform către f_2.

Pasul 2. Din polinoamele lui Bernstein putem construi operatorii Bernstein $B_n : C([0,1], \mathbb{R}) \to C([0,1], \mathbb{R})$, $B_n(f) = B_n(f, \cdot)$, adică o funcţie care asociază la orice funcţie continuă f polinomul Bernstein $B_n(f, \cdot)$. Operatorii Bernstein B_n posedă următoarele proprietăţi:

1. $B_n(f+g) = B_n(f) + B_n(g)$ (B_n este aditivă);

2. $B_n(\alpha \cdot f) = \alpha \cdot B_n(f)$ (B_n este omogenă);

3. pentru $f \geq \theta$ ($f(x) \geq 0$ pentru orice $x \in [0,1]$) rezultă că $B_n(f) \geq \theta$ ($B_n(f)(x) \geq 0$ pentru orice $x \in [0,1]$) (B_n este pozitiv).

1 şi 2 împreună înseamnă că B_n este un operator liniar. Alegând $f = g = \theta$ în 1, se obţine că $B_n(\theta) = \theta$, şi alegând $\alpha = -1$ deducem că

$$B_n(f-g) = B_n(f + (-1)\cdot g) = B_n(f) + B_n((-1)\cdot g) = B_n(f) - B_n(g).$$

Luând în considerare şi punctul 3 rezultă proprietatea de monotonie a operatorului lui Bernstein: $f \leq g$ (adică $f(x) \leq g(x)$ pentru orice $x \in [0,1]$) implică $B_n(f) \leq B_n(g)$ (adică $B_n(f)(x) \leq B_n(g)(x)$ pentru orice $x \in [0,1]$). Într-adevăr, $g - f \geq \theta$, deci $B_n(g-f) \geq B_n(\theta) = \theta$. Prin urmare $B_n(g) - B_n(f) \geq \theta$, adică $B_n(g) \geq B_n(f)$.

În continuare vom demonstra proprietăţile 1, 2 şi 3:

1. Avem de arătat că: $B_n(f+g)(x) = (B_n(f) + B_n(g))(x)$, adică $B_n(f+g)(x) = B_n(f)(x) + B_n(g)(x)$ pentru orice $x \in [0,1]$, deci $B_n(f+g,x) = B_n(f,x) + B_n(g,x)$. Dar

$$\begin{aligned} B_n(f+g,x) &= \sum_{k=0}^{n} C_n^k \cdot x^k \cdot (1-x)^{n-k} \cdot (f+g)\left(\frac{k}{n}\right) = \\ &= \sum_{k=0}^{n} C_n^k \cdot x^k \cdot (1-x)^{n-k} \cdot \left[f\left(\frac{k}{n}\right) + g\left(\frac{k}{n}\right)\right] = \\ &= \sum_{k=0}^{n} C_n^k \cdot x^k \cdot (1-x)^{n-k} \cdot f\left(\frac{k}{n}\right) + \sum_{k=0}^{n} C_n^k \cdot x^k \cdot (1-x)^{n-k} \cdot g\left(\frac{k}{n}\right) = \\ &= B_n(f,x) + B_n(g,x). \end{aligned}$$

2. Avem de arătat că $B_n(\alpha\cdot f)(x) = \alpha \cdot B_n(f)(x)$, care este echivalent cu: $B_n(\alpha\cdot f; x) = \alpha \cdot B_n(f,x)$. Dar

$$\begin{aligned} B_n(\alpha\cdot f, x) &= \sum_{k=0}^{n} C_n^k \cdot x^k \cdot (1-x)^{n-k} \cdot (\alpha\cdot f)\left(\frac{k}{n}\right) = \\ &= \sum_{k=0}^{n} C_n^k \cdot x^k \cdot (1-x)^{n-k} \cdot \alpha \cdot f\left(\frac{k}{n}\right) = \\ &= \alpha \cdot \sum_{k=0}^{n} C_n^k \cdot x^k \cdot (1-x)^{n-k} \cdot f\left(\frac{k}{n}\right) = \alpha \cdot B_n(f,x). \end{aligned}$$

3. Avem de arătat că, dacă $f(x) \geq 0$ pentru orice $x \in [0,1]$ atunci $B_n(f)(x) = B_n(f,x) \geq 0$ pentru orice $x \in [0,1]$. Într-adevăr

$$B_n(f,x) = \sum_{k=0}^{n} C_n^k \cdot x^k \cdot (1-x)^{n-k} \cdot f\left(\frac{k}{n}\right) \geq 0, \forall\, x \in [0,1.]$$

Pasul 3. Deoarece $f \in C([0,1],\mathbb{R})$ este continuă pe $[0,\ 1]$ conform teoremei lui Weierstrass este mărginită, adică există o constantă $M > 0$ astfel încât $|f(x)| \leq M$ pentru orice $x \in [0,1]$. Prin urmarea pentru orice $x,t \in [0,1]$ avem $|f(x) - f(t)| \leq |f(x)| + |f(t)| \leq 2M$.

Pe de altă parte conform lemei 8.1.4 $f \in C([0,1],\mathbb{R})$ este uniform continuă, deci pentru orice $\varepsilon > 0$ există $\delta_{(\varepsilon)} > 0$ astfel încât pentru orice $x,t \in [0,1]$ cu $|x-t| \leq \delta_{(\varepsilon)}$ rezultă $|f(x) - f(t)| \leq \varepsilon$. Dacă pentru $t \in [0,1]$ număr real fixat considerăm funcţia $\Psi_t : [0,1] \to \mathbb{R}$, $\Psi_t(x) = (x-t)^2$, atunci cele două inegalităţi anterioare se pot scrie într-o singură inegalitate:

$$|f(x) - f(t)| \leq \varepsilon + \frac{2M}{\delta_{(\varepsilon)}^2} \cdot \psi_t(x).$$

Într-adevăr, dacă $|x-t| \leq \delta_{(\varepsilon)}$, atunci

$$|f(x) - f(t)| \leq \varepsilon \leq \varepsilon + \frac{2M}{\delta_{(\varepsilon)}^2} \cdot \Psi_t(x),$$

iar dacă $|x-t| > \delta_{(\varepsilon)}$, atunci $\dfrac{|x-t|^2}{\delta_{(\varepsilon)}^2} > 1$, deci

$$|f(x) - f(t)| \leq 2M < \varepsilon + 2M \cdot \frac{|x-t|^2}{\delta_{(\varepsilon)}^2} = \varepsilon + \frac{2M}{\delta_{(\varepsilon)}^2} \cdot \Psi_t(x).$$

Însă

$$B_n(\Psi_t,t) = B_n(f_2 - 2tf_1 + t^2 f_0, t) = t^2 + \frac{t-t^2}{n} - 2t \cdot t + t^2 \cdot 1 = \frac{t-t^2}{n}$$

folosind de la pasul 1 calculul lui $B_n(f_0,x)$, $B_n(f_1,x)$ şi $B_n(f_2,x)$.

Aplicăm operatorul lui Bernstein asupra inegalităţii: $|f(x) - f(t)| \leq \varepsilon + \frac{2M}{\delta_{(\varepsilon)}^2} \cdot \psi_t(x)$, adică la inegalitatea

$$-\varepsilon - \frac{2M}{\delta_{(\varepsilon)}^2}\psi_t(x) \leq f(x) - f(t) \leq \varepsilon + \frac{2M}{\delta_{(\varepsilon)}^2} \cdot \psi_t(x)$$

şi se obţine conform proprietăţilor 1, 2, 3:

$$-\varepsilon - \frac{2M}{\delta_{(\varepsilon)}^2} \cdot \frac{t-t^2}{n} \leq B_n(f,t) - f(t) \leq \varepsilon + \frac{2M}{\delta_{(\varepsilon)}^2} \cdot \frac{t-t^2}{n},$$

adică

$$|B_n(f,t) - f(t)| \leq \varepsilon + \frac{2M}{\delta^2_{(\varepsilon)}} \cdot \frac{t - t^2}{n} \leq \varepsilon + \frac{2M}{\delta^2_{(\varepsilon)} \cdot n},$$

pentru orice $t \in [0,1]$. Se observă că pentru orice $\varepsilon > 0$ există un $n_{(\varepsilon)} \in \mathbb{N}$ astfel încât pentru orice $n \geq n_{(\varepsilon)}$ avem $\dfrac{2M}{\delta^2_{(\varepsilon)} \cdot n} \leq \varepsilon$. Prin urmare pentru $n \geq n_{(\varepsilon)}$ avem $|B_n(f,t) - f(t)| \leq 2\varepsilon$ pentru orice $t \in [0,1]$. Această condiţie exprimă faptul că polinoamele lui Bernstein converg uniform către funcţia dată f:

$$\|B_n(f) - f\|_\infty = \max\{|B_n(f,x) - f(x) \ / \ x \in [0,1]\} \leq 2\varepsilon. \qquad \square$$

Observaţia 8.1.3. *Dacă în locul intervalului [0, 1] se consideră un interval arbitrar* $[a,b]$, *atunci prin transformarea* $x = \dfrac{z-a}{b-a}$ *cu* $z \in [a,b]$ *se face o corespondenţă biunivocă între punctele* z *din intervalul* $[a,b]$ *respectiv punctele* x *din intervalul [0, 1]. Menţionăm că prin această transformare polinoamele Bernstein definite pe intervalul* $[0,1]$ *se transformă tot în polinoame definite pe intervalul* $[a,b]$.

Ca aplicaţie vom pune problema de a calcula valoarea polinomului lui Bernstein $B_n(f,x)$ de ordinul n într-un punct $x \in [0,1]$ dat.

Program Bernstein

Datele de intrare: f, n, x;

Dacă $x = 1$ atunci $s := f(1)$;

Dacă $x \neq 1$ atunci

Fie $s := 0$; $z := 1$;

Pentru $i = \overline{1,n}$ execută $z := z * (1 - x)$;

$y := z * f(0)$;

$s := s + y$;

Pentru $k = \overline{0,n-1}$ execută

$$z := z * \frac{n-k}{k+1} * \frac{x}{1-x};$$

$$y := z * f\left(\frac{k+1}{n}\right);$$

$$s := s + y;$$

Tipăreşte s.

Plecând de la demonstraţia teoremei lui Weierstrass cu ajutorul polinoamelor lui Bernstein se obţine următoarea generalizare naturală cunoscută sub numele de teorema lui Korovkin:

Teorema 8.1.2. *(Korovkin) Dacă şirul de operatori liniari şi pozitivi* $L_n : C([0,1],\mathbb{R}) \to C([0,1],\mathbb{R})$, $n \in \mathbb{N}$ *verifică condiţiile ca* $L_n(f_i)$ *tinde uniform către* f_i *pentru orice* $i = \overline{0,2}$ *şi orice* $n \in \mathbb{N}$, *atunci avem că* $L_n(f)$ *tinde uniform către* f, *pentru orice* $f \in C([0,1],\mathbb{R})$, *când* $n \to \infty$.

Pentru demonstraţie trebuie să luăm numai partea 2 şi partea 3 din demonstraţia teoremei lui Weierstrass punând în locul operatorilor B_n operatorii L_n, $n \in \mathbb{N}$.

8.1.2 Teorema lui Stone

Teorema lui Stone este o generalizare esenţială a teoremei lui Weierstrass. Fixăm spaţiul de bază $C([0,1],\mathbb{R})$ cu norma supremum sau norma maximum sau norma Cebâşev:

$$\|f\| = \sup\{|f(x)| \ / \ x \in [0,1]\} = \max\{|f(x)| \ / \ x \in [0,1]\}.$$

Ştim că convergenţa unui şir de funcţii continue $(f_n)_{n\in\mathbb{N}} \subset C([0,1],\mathbb{R})$ în acest spaţiu înseamnă convergenţa uniformă a acestui şir de funcţii continue. Totodată se ştie că $(C([0,1],\mathbb{R}), \|\cdot\|)$ este chiar un spaţiu Banach. Mai mult pe spaţiul $C([0,1],\mathbb{R})$ se poate introduce o nouă operaţie de înmulţire punctuală a două funcţii continue: $f \cdot g : [0,1] \to \mathbb{R}$, $(f \cdot g)(x) = f(x) \cdot g(x)$ pentru orice $x \in [0,1]$. Această operaţie este o operaţie internă pe $C([0,1],\mathbb{R})$, fiindcă se ştie că produsul punctual a două funcţii continue este tot o funcţie continuă. Astfel $C([0,1],\mathbb{R})$ devine chiar o algebră Banach: $C([0,1],\mathbb{R})$ cu adunarea şi înmulţirea cu scalari este un spaţiu vectorial real şi $C([0,1],\mathbb{R})$ cu adunarea şi înmulţirea punctuală este un inel unitar, mai are loc şi relaţia: $\alpha(fg) = (\alpha f)g = f(\alpha g)$ pentru orice $\alpha \in \mathbb{R}$ şi orice $f, g \in C([0,1],\mathbb{R})$, iar la axiomele normei adăugăm axioma $\|fg\| \leq \|f\| \cdot \|g\|$, care se poate verifica cu destulă uşurinţă pentru $C([0,1],\mathbb{R})$ cu norma maximum. O submulţime $A \subset C([0,1],\mathbb{R})$, $A \neq \emptyset$ care la rândul său verifică axiomele algebrei se numeşte subalgebră a algebrei $C([0,1],\mathbb{R})$. Pentru o subalgebră A putem considera închiderea topologică a lui A în $C([0,1],\mathbb{R})$, în raport cu norma supremum (care generează o metrică supremum şi care la rândul său generează o topologie pe $C([0,1],\mathbb{R})$) şi să o notăm cu $\bar{A}$.

Lema 8.1.5. *Dacă $A \subset C([0,1],\mathbb{R})$, $A \neq \emptyset$ este o subalgebră atunci $\bar{A} \subset C([0,1],\mathbb{R})$ va fi tot o subalgebră.*

DEMONSTRAŢIE. Fie $f, g \in \bar{A}$. Atunci f şi g sunt două puncte aderente relativ la subalgebra A, adică există două şiruri $(f_n)_{n\in\mathbb{N}}$ şi $(g_n)_{n\in\mathbb{N}} \subset A$ astfel ca f_n tinde uniform la f şi g_n tinde uniform la g. Astfel $(f_n + g_n)_{n\in\mathbb{N}}$ este un şir din subalgebra A şi tinde uniform către $f+g$. Dar acesta implică că $(f+g) \in \bar{A}$. În mod analog rezultă că $f \cdot g \in \bar{A}$ şi $\alpha \cdot f \in \bar{A}$. q.e.d.

Este interesant că numărul de operaţii se poate lărgi în cazul algebrei Banach standard $C([0,1],\mathbb{R})$ prin a organiza mulţimea $C([0,1],\mathbb{R})$ ca o latice. Într-adevăr vom defini pentru orice $f, g \in C([0,1],\mathbb{R})$ funcţia $\min(f,g) : [0,1] \to \mathbb{R}$, $[\min(f,g)](x) = \min\{f(x), g(x)\}$ pentru orice $x \in [0,1]$, respectiv funcţia $\max(f,g) : [0,1], \to \mathbb{R}$, $[\max(f,g)](x) = \max\{f(x), g(x)\}$ pentru orice $x \in [0,1]$. Se ştie că dacă f şi g sunt două funcţii continue atunci funcţiile $\min(f,g)$ şi $\max(f,g)$ sunt tot funcţii continue. Luăm ca infimum al lui f, g în $C([0,1],\mathbb{R})$ notat cu $f \wedge g = \min(f,g)$ respectiv supremum al lui f, g în $C([0,1],\mathbb{R})$ notat cu $f \vee g = \max(f,g)$. Astfel $C([0,1],\mathbb{R})$ devinde o latice, fiindcă axiomele laticii se verifică imediat. Vom spune că o submulţime $A \subset C([0,1],\mathbb{R})$, $A \neq \emptyset$ este o sublatice, dacă la rândul său este o latice, adică se verifică pe A axiomele laticii.

Observaţia 8.1.4. *Nu orice subalgebră a lui $C([0,1],\mathbb{R})$ este o sublatice. Într-adevăr, alegem*

$$A = \{a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 \ / \ n \in \mathbb{N},\ a_i \in \mathbb{R},\ i = \overline{0,n},\ x \in [0,1]\}.$$

Rezultă imediat că A este o subalgebră a lui $C([0,1],\mathbb{R})$, însă nu este o sublatice a lui $C([0,1],\mathbb{R})$ fiindcă dacă alegem $f : [0,1] \to \mathbb{R}$, $f(x) = x$ pentru orice $x \in [0,1]$ şi $g : [0,1] \to \mathbb{R}$, $g(x) = 1 - x$, pentru orice $x \in [0,1]$ se obţine că $f \wedge g \notin A$. Într-adevăr

$$(f \wedge g)(x) = \min\{x, 1-x\} = \begin{cases} x, & \text{dacă} \quad x \in \left[0, \dfrac{1}{2}\right] \\ 1 - x, & \text{dacă} \quad x \in \left(\dfrac{1}{2}, 1\right]. \end{cases}$$

Dar $f \wedge g$ nu este derivabilă în punctul $x = \dfrac{1}{2}$ pe când orice element al lui A este derivabilă în orice punct $x \in [0,1]$. Deci această subalgebră A nu este o sublatice a lui $C([0,1],\mathbb{R})$.

În $C([0,1],\mathbb{R})$ se defineşte elementul notat cu $|f| = f \vee (-f)$ unde $-f : [0,1] \to \mathbb{R}$, $(-f)(x) = -f(x)$ pentru orice $x \in [0,1]$.

Lema 8.1.6. *O subalgebră $A \subset C([0,1],\mathbb{R})$ este o sublatice dacă şi numai dacă din $f \in A$ rezultă că $|f| \in A$ pentru orice $f \in A$.*

DEMONSTRAŢIE. Presupunem că subalgebra $A \subset C([0,1],\mathbb{R})$ este o sublatice. Atunci din $f \in A$ rezultă că $-f = (-1)\cdot f \in A$, deci $|f| = f \vee (-f) \in A$.

Invers, să arătăm că dacă pentru o subalgebră $A \subset C([0,1],\mathbb{R})$ din $f \in A$ rezultă că $|f| \in A$ pentru orice $f \in A$, atunci A este o sublatice. Din f şi $|f| \in A$ rezultă că f^+ şi $f^- \in A$ unde

$$f^+ : [0,1] \to \mathbb{R},\ f^+(x) = \frac{f(x)+|f|(x)}{2} = \frac{f(x)+|f(x)|}{2} \quad \text{şi}$$
$$f^- : [0,1] \to \mathbb{R},\ f^-(x) = \frac{f(x)-|f|(x)}{2} = \frac{f(x)-|f(x)|}{2}.$$

Menţionăm proprietăţile imediate ca $f = f^+ + f^-$, $f^+ = f \vee \theta$, $f^- = f \wedge \theta$, unde θ este funcţia nulă, adică $\theta : [0,1] \to \mathbb{R}$, $\theta(x) = 0$ pentru orice $x \in [0,1]$. Prin urmare din $f, \theta \in A$ rezultă că $f \vee \theta$ şi $f \wedge \theta \in A$. Suntem gata cu demonstraţia fiindcă pentru $f, g \in A$ avem $f - g \in A$ şi folosind cele anterioare avem $f \vee g = (f-g) \vee \theta + g \in A$ şi $f \wedge g = (f-g) \wedge \theta + g \in A$. q.e.d.

Definiţia 8.1.1. *Spunem că o subalgebră $A \subset C([0,1],\mathbb{R})$ separă punctele mulţimii [0, 1], dacă pentru oricare două puncte distincte $x_1, x_2 \in [0,1]$ cu $x_1 \neq x_2$ există o funcţie $f \in A$ astfel ca $f(x_1) \neq f(x_2)$.*

Teorema 8.1.3. *(Stone, cazul real) Fie $A \subset C([0,1],\mathbb{R})$ o subalgebră care conţine funcţiile constante pe [0, 1] şi separă punctele lui [0, 1]. Atunci A este densă în $C([0,1],\mathbb{R})$, adică $\bar{A} = C([0,1],\mathbb{R})$.*

DEMONSTRAŢIE.

Etapa 1. Se arată că $\bar{A}$ este o sublatice a lui $C([0,1],\mathbb{R})$. Conform lemei 8.1.5 rezultă că $\bar{A}$ este o subalgebră. Pentru ca $\bar{A}$ să fie o sublatice este de ajuns să arătăm că odată cu $f \in \bar{A}$ avem că şi $|f| \in \bar{A}$. Fie $f \in \bar{A}$. Putem presupune că $f \in \theta$, întrucât dacă $f = \theta$ atunci şi $|f| = \theta$ şi $|f| \in \bar{A}$, fiindcă orice algebră conţine elementul nul. Deci fie $f \in \bar{A}$ cu $f \neq \theta$. Cum pentru orice două numere $\alpha, \beta \in \mathbb{R}$ aşa ca $|\alpha| \leq \beta$ şi $\beta > 0$, are loc

$$|\alpha| = \beta \cdot \sqrt{1 - \left(1 - \frac{\alpha^2}{\beta^2}\right)}$$

rezultă că dacă notăm prin $\gamma = 1 - \dfrac{\alpha^2}{\beta^2}$, atunci $0 \leq \gamma \leq 1$ şi avem $|\alpha| = \beta \cdot \sqrt{1-\gamma}$. Dar conform dezvoltării binomiale, avem:

$$\sqrt{1-u} = 1 - \frac{u}{2} - \sum_{n=1}^{\infty} \frac{(2n-1)!!}{(2n)!!} \cdot \frac{u^{n+1}}{2n+2} \text{ cu } |u| < 1.$$

Menţionăm că avem notaţiile: $(2n-1)!! := 1 \cdot 3 \ldots (2n-1)$ şi $(2n)!! = 2 \cdot 4 \ldots (2n)$ şi dezvoltarea anterioară are sens şi pentru capetele $u = \pm 1$ (pentru $u = -1$ avem o serie alternată iar pentru $u = 1$ avem seria convergentă

$$\sum_{n=1}^{\infty} \frac{(2n-1)!!}{(2n)!!} \cdot \frac{1}{2n+2},$$

căci

$$\frac{1 \cdot 3 \ldots (2n-1)}{2 \cdot 4 \ldots (2n)} \leq \frac{1}{\sqrt{2n+1}}$$

ceea ce se arată imediat prin inducţie matematică). Să alegem $\alpha = f(x)$ şi $\beta = \|f\|$, de unde $\gamma = 1 - \left(\frac{f(x)}{\|f\|}\right)^2$. Utilizând dezvoltarea în serie obţinem:

$$|f(x)| = \|f\| \left\{ 1 - \frac{1}{2}\left[1 - \left(\frac{f(x)}{\|f\|}\right)^2\right] - \sum_{n=1}^{\infty} \frac{(2n-1)!!}{(2n)!!} \cdot \frac{1}{2n+2} \cdot \left[1 - \left(\frac{f(x)}{\|f\|}\right)^2\right]^{n+1} \right\}.$$

Se observă că seria din membrul drept este uniform convergentă (prin majorare se obţine seria numerică convergentă de mai sus), şi cum termenii seriei sunt în $\bar{A}$ rezultă că şi suma seriei va fi tot în $\bar{A}$.

Etapa 2. Să arătăm că dacă $f \in C([0,1], \mathbb{R})$ şi x_1, x_2 sunt două puncte oarecare din intervalul [0, 1], atunci există o funcţie $g \in A$ astfel ca $g(x_1) = f(x_1)$ şi $g(x_2) = f(x_2)$. Dacă $x_1 = x_2$, vom lua ca funcţie g funcţia constantă $g(x) = f(x_1)$ pentru orice $x \in [0,1]$. Dacă $x_1 \neq x_2$, atunci din condiţia de separare a punctelor rezultă că există o funcţie $h \in A$ aşa ca $h(x_1) \neq h(x_2)$. De aici se vede că α, β sunt unic determinaţi prin sistemul liniar

$$\begin{cases} \alpha h(x_1) + \beta = f(x_1) \\ \alpha h(x_2) + \beta = f(x_2), \end{cases}$$

cu determinantul principal

$$\begin{vmatrix} h(x_1) & 1 \\ h(x_2) & 1 \end{vmatrix} = h(x_1) - h(x_2) \neq 0.$$

Fie funcţia $g : [0,1] \to \mathbb{R}$ dată prin $g(x) = \alpha \cdot h(x) + \beta$ pentru orice $x \in [0,1]$, care evident satisface condiţiile cerute. Întrucât α, β depind de alegerea punctelor $x_1, x_2 \in [0,1]$ vom mai folosi şi notaţia $g(x) = g_{x_1,x_2}(x)$ pentru orice $x \in [0,1]$.

Etapa 3. Fie $f \in C([0,1],\mathbb{R})$ şi $\varepsilon > 0$. Dacă $s,t \in [0,1]$ atunci conform etapei a doua există o funcţie $g_{s,t} \in A$ aşa ca $g_{s,t}(s) = f(s)$ şi $g_{s,t}(t) = f(t)$. Cum f şi $g_{s,t}$ sunt funcţii continue pe [0, 1] la fel este şi funcţia φ, dată de $\varphi(x) = g_{s,t}(x) - f(x)$ pentru orice $x \in [0,1]$. Să exprimăm continuitatea lui φ în punctul "s". Pentru orice $\varepsilon > 0$ există o vecinătate deschisă $U(s)$ a lui "s" aşa ca $|\varphi(x) - \varphi(s)| = |\varphi(x)| < \varepsilon$ pentru orice $x \in U(s)$, de unde rezultă că $g_{s,t}(x) > f(x) - \varepsilon$ pentru orice $x \in U(s)$. Dar pentru t fixat mulţimile $U(s)$ $(s \in [0,1])$ formează o acoperire deschisă pentru [0, 1] şi cum [0, 1] este un interval compact rezultă că există $s_1, \ldots, s_m \in [0,1]$ aşa ca $\bigcup_{i=1}^{m} U(s_i) = [0,1]$. Fie $\Psi_t = \sup_{1 \le i \le m} \{g_{s_i,t}\}$. Deoarece $g_{s_i,t} \in A \subset \bar{A}$ conform etapei 1 avem că $\Psi_t \in \bar{A}$ pentru orice $t \in [0,1]$ şi $\Psi_t(x) > f(x) - \varepsilon$ pentru orice $x \in [0,1]$, adică $\Psi_t(x) - f(x) > -\varepsilon$ pentru orice $x \in [0,1]$. Pe de altă parte $\Psi_t - f$ este o funcţie continuă pe [0, 1], deci şi în "t", şi atunci pentru orice $\varepsilon > 0$ există o vecinătate deschisă a punctului "t", notată cu $V(t)$, aşa ca $|\Psi_t(x) - f(x)| < \varepsilon$ pentru orice $x \in V(t)$, de unde rezultă că $\Psi_t(x) < f(x) + \varepsilon$ pentru orice $x \in V(t)$. Dar intervalul [0, 1] fiind compact, există $t_1, \ldots, t_n \in [0,1]$ aşa încât $[0,1] = \bigcup_{i=1}^{n} V(t_i)$. Fie $F = \inf_{1 \le i \le n} \{\Psi_{t_i}\}$. Atunci $F(x) < f(x) + \varepsilon$ şi din $\Psi_t(x) > f(x) - \varepsilon$ pentru orice $x \in [0,1]$ rezultă că $F(x) > f(x) - \varepsilon$ pentru orice $x \in [0,1]$, adică $|F(x) - f(x)| < \varepsilon$ pentru orice $x \in [0,1]$. Cum prin construcţie $F \in \bar{A}$ avem că $\bar{A} = C([0,1],\mathbb{R})$ q.e.d.

Consecinţa 8.1.1. *Dacă se alege $A = \mathbb{R}[X]$, adică mulţimea tuturor polinoamelor definite pe intervalul compact [0, 1], atunci A este o subalgebră a lui $C([0,1],\mathbb{R})$, care conţine funcţiile constante, separă punctele lui [0, 1] (fiindcă de exemplu funcţia $f : [0,1] \to \mathbb{R}$, $f(x) = x$ are această proprietate), prin urmare reobţinem teorema lui Weierstrass, de densitate a polinoamelor în $C([0,1],\mathbb{R})$.*

Consecinţa 8.1.2. *Dacă se alege A mulţimea tuturor polinoamelor trigonometrice definite pe intervalul compact $[0, 2\pi]$, atunci $\bar{A} = C([0,2\pi],\mathbb{R})$. Într-adevăr, fie*

$$A = \left\{ a_0 + \sum_{i=1}^{n} a_i \sin ix + \sum_{j=1}^{m} b_j \cos jx \mid n,m \in \mathbb{N}^*,\ a_i, b_j \in \mathbb{R},\ x \in [0,2\pi] \right\}$$

mulţimea tuturor polinoamelor trigonometrice. Imediat putem constata că A este o subalgebră a lui $C([0,2\pi],\mathbb{R})$ (la înmulţire se aplică formulele trigonometrice de transformare

a produsurilor de sinuşi şi cosinuşi în sume şi diferenţe). Totodată menţionăm că nu contează dacă în locul intervalului compact [0, 1] am ales intervalul compact $[0, 2\pi]$, *fiindcă putem să facem o corespondenţă biunivocă între cele două intervale, şi această transformare nu-şi schimbă esenţa problemei. Deoarece* A *conţine funcţiile constante şi separă punctele intervalului* $[0, 2\pi]$ *avem* $\bar{A} = C([0, 2\pi], \mathbb{R})$, *deci orice funcţie continuă pe intervalul* $[0, 2\pi]$ *se poate aproxima cu un şir de polinoame trigonometrice în raport cu norma maximum sau supremum.*

În încheierea acestui paragraf enunţăm fără demonstraţie şi versiunea complexă a teoremei lui Stone. Pentru asta considerăm $C([0, 1], \mathbb{C}) = \{f : [0, 1] \to \mathbb{C} \mid f$ *continuă}, care este tot o algebră Banach, dar peste corpul numerelor complexe. Definim funcţia* $\bar{f} \in C([0, 1], \mathbb{C})$ *în felul următor:* $\bar{f} : [0, 1] \to \mathbb{C}$, $\bar{f}(x) = \overline{f(x)}$ *pentru orice* $x \in [0, 1]$, *unde* $\overline{f(x)}$ *înseamnă conjugata numărului complex* $f(x)$.

Teorema 8.1.4. *(Stone, cazul complex) Dacă* A *este o subalgebră a lui* $C([0, 1], \mathbb{C})$ *care conţine funcţiile constante, separă punctele intervalului [0, 1] şi din* $f \in A$ *rezultă că* $\bar{f} \in A$ *pentru orice* $f \in A$ *atunci* A *este densă în* $C([0, 1], \mathbb{C})$.

Ca consecinţe importante se pot obţine şi în acest caz variantele complexe ale consecinţelor 8.1.1 şi 8.1.2, considerând polinoamele cu coeficienţi complecşi respectiv polinoamele trigonometrice cu coeficienţi complecşi.

8.2 Aproximarea funcţiilor prin interpolare

Introducem mulţimea funcţiilor definite pe un interval $[a, b]$ dat cu valori reale, notată cu $\mathcal{F}([a, b], \mathbb{R})$. Fie $\Delta : a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b$ o diviziune dată a intervalului $[a, b]$. Pe mulţimea $\mathcal{F}([a, b], \mathbb{R})$ introducem o seminormă dată de formula: $\|f\|_\Delta = \sum_{i=1}^{n} |f(x_i)|$. Într-adevăr, în acest fel definim o seminormă:

1.

$$\begin{aligned} \|f + g\|_\Delta &= \sum_{i=1}^{n} |(f + g)(x_i)| = \sum_{i=1}^{n} |f(x_i) + g(x_i)| \leq \sum_{i=1}^{n} (|f(x_i)| + |g(x_i)|) = \\ &= \sum_{i=1}^{n} |f(x_i)| + \sum_{i=1}^{n} |g(x_i)| = \|f\|_\Delta + \|g\|_\Delta. \end{aligned}$$

2.

$$\begin{aligned}\|\alpha \cdot f\|_\Delta &= \sum_{i=1}^{n} |(\alpha \cdot f)(x_i)| = \sum_{i=1}^{n} |\alpha \cdot f(x_i)| = \sum_{i=1}^{n} |\alpha| \cdot |f(x_i)| = \\ &= |\alpha| \cdot \sum_{i=1}^{n} |f(x_i)| = |\alpha| \cdot \|f\|_\Delta.\end{aligned}$$

3. $\|f\|_\Delta = \sum_{i=1}^{n} |f(x_i)| \geq 0$ însă din condiţia $\|f\|_\Delta = 0$ nu rezultă că $f = \theta$, adică $f(x) = 0$ pentru orice $x \in [a, b]$, deci funcţia $\|\cdot\|_\Delta$ nu va fi o normă.

Cu ajutorul acestei seminorme putem defini o semimetrică $\rho_\Delta : \mathcal{F}([a, b], \mathbb{R}) \times \mathcal{F}([a, b], \mathbb{R}) \to \mathbb{R}$, dată de formula $\rho_\Delta(f, g) = \|f - g\|_\Delta$ pentru orice $f, g \in \mathcal{F}([a, b], \mathbb{R})$. Într-adevăr ρ_Δ este o semimetrică:

1.

$$\rho_\Delta(f, g) = \|f-g\|_\Delta = \|(f-h)+(h-g)\|_\Delta \leq \|f-h\|_\Delta+\|h-g\|_\Delta = \rho_\Delta(f, h)+\rho_\Delta(h, g).$$

2.

$$\rho_\Delta(f, g) = \|f - g\|_\Delta = \| - (g - f)\|_\Delta = | - 1| \cdot \|g - f\|_\Delta = \|g - f\|_\Delta = \rho_\Delta(g, f).$$

3. $\rho_\Delta(f, g) = \|f - g\|_\Delta \geq 0$, însă din $\rho_\Delta(f, g) = 0$ nu rezultă că $f = g$, fiindcă $\rho_\Delta(f, g) = \|f - g\|_\Delta = 0$ şi de aici nu rezultă că $f - g = \theta$, adică $f = g$, deci ρ_Δ va fi numai o semimetrică.

Fie $(\mathcal{F}([a, b], \mathbb{R}), \rho_\Delta)$ spaţiul semimetric al funcţilor definite pe intervalul $[a, b]$ cu valori reale, iar $B \subset \mathcal{F}([a, b], \mathbb{R})$. Vom spune că elementele mulţimii B aproximează în sens interpolator elementele mulţimii $\mathcal{F}([a, b], \mathbb{R})$, dacă pentru orice $f \in \mathcal{F}([a, b], \mathbb{R})$ există $g \in B$ astfel încât $\rho_\Delta(f, g) = 0$. Practic această condiţie înseamnă următoarele: pentru orice $f \in \mathcal{F}([a, b], \mathbb{R})$ putem găsi un $g \in B$ astfel ca $g(x_i) = f(x_i)$ pentru $i = \overline{0, n}$. Vom spune că funcţia g aproximează funcţia f prin interpolare.

8.2.1 Interpolare liniară

Dacă se alege submulţimea B mulţimea funcţiilor definite pe intervalul $[a, b]$ cu valori reale, segmentar liniare (care admit graficul o linie frântă) vom spune că avem o interpolare liniară. Prin urmare problema interpolării liniare se formulează în felul următor: fie $f : [a, b] \to \mathbb{R}$ o funcţie dată, $\Delta : a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b$ o diviziune dată, şi se cere să se determine aceea funcţie g segmentar liniară pentru care $g(x_i) = f(x_i)$ pentru orice $i = \overline{0, n}$. Pentru $i = \overline{0, n-1}$ fixăm intervalul $[x_i, x_{i+1}]$ şi fie $g/_{[x_i, x_{i+1}]}(x) = m_i x + n_i$, unde $m_i, n_i \in \mathbb{R}$. Se impun condiţiile ca $g(x_i) = m_i x_i + n_i = f(x_i)$ şi $g(x_{i+1}) = m_i x_{i+1} + n_i = f(x_{i+1})$. Astfel se obţine un sistem cu două ecuaţii şi cu două necunoscute m_i, n_i :

$$\begin{cases} x_i m_i + n_i = f(x_i) \\ x_{i+1} m_i + n_i = f(x_{i+1}) \end{cases}$$

cu soluţiile:

$$m_i = \frac{\begin{vmatrix} f(x_i) & 1 \\ f(x_{i+1}) & 1 \end{vmatrix}}{\begin{vmatrix} x_i & 1 \\ x_{i+1} & 1 \end{vmatrix}} = \frac{f(x_i) - f(x_{i+1})}{x_i - x_{i+1}} = \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i} \quad \text{şi}$$

$$n_i = \frac{\begin{vmatrix} x_i & f(x_i) \\ x_{i+1} & f(x_{i+1}) \end{vmatrix}}{\begin{vmatrix} x_i & 1 \\ x_{i+1} & 1 \end{vmatrix}} = \frac{x_i \cdot f(x_{i+1}) - x_{i+1} \cdot f(x_i)}{x_i - x_{i+1}} = \frac{x_{i+1} \cdot f(x_i) - x_i \cdot f(x_{i+1})}{x_{i+1} - x_i}$$

Prin urmare avem

$$g(x) = \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i} \cdot x + \frac{x_{i+1} \cdot f(x_i) - x_i \cdot f(x_{i+1})}{x_{i+1} - x_i}$$

pentru orice $x \in [x_i, x_{i+1}]$. Evident graficul lui g este un segment liniar pe intervalul $[x_i, x_{i+1}]$, iar pe intervalul $[a, b]$ se obţine ca grafic o linie frântă.

8.2.2 Interpolarea polinomială a lui Lagrange

Prima dată vom formula problema interpolării polinomiale: fie dată o funcţie $f : [a,b] \to \mathbb{R}$, fie $\Delta : a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b$ o diviziune dată şi se caută acel polinom de gradul cel mult n, $P : [a,b] \to \mathbb{R}$, pentru care $P(x_i) = f(x_i)$ pentru orice $i = \overline{0,n}$. Se observă că în acest caz submulţimea B se alege mulţimea polinoamelor de gradul cel mult n, care au o "structură mai simplă". Problema pusă admite o singură soluţie:

Teorema 8.2.1. *Există un unic polinom de gradul cel mult n, $P : [a,b] \to \mathbb{R}$, pentru care $P(x_i) = f(x_i)$ oricare ar fi $i = \overline{0,n}$.*

Demonstraţie. Vom demonstra existenţa şi unicitatea polinomului de interpolare P. Fie

$$P(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0,$$

unde $a_i \in \mathbb{R}$, $i = \overline{1,n}$. Se impun condiţiile de interpolare:

$$\begin{aligned}
P(x_0) &= a_n x_0^n + a_{n-1}x_0^{n-1} + \cdots + a_1 x_0 + a_0 = f(x_0)\\
P(x_1) &= a_n x_1^n + a_{n-1}x_1^{n-1} + \cdots + a_1 x_1 + a_0 = f(x_1)\\
&\vdots\\
P(x_n) &= a_n x_n^n + a_{n-1}x_n^{n-1} + \cdots + a_1 x_n + a_0 = f(x_n).
\end{aligned}$$

Astfel se obţine un sistem liniar cu $n+1$ ecuaţii şi $n+1$ necunoscute: $a_n, a_{n-1}, \ldots, a_1, a_0$:

$$\begin{cases}
x_0^n \cdot a_n + x_0^{n-1} \cdot a_{n-1} + \cdots + x_0 \cdot a_1 + a_0 = f(x_0)\\
x_1^n \cdot a_n + x_1^{n-1} \cdot a_{n-1} + \cdots + x_1 \cdot a_1 + a_0 = f(x_0)\\
\vdots\\
x_n^n \cdot a_n + x_n^{n-1} \cdot a_{n-1} + \cdots + x_n \cdot a_1 + a_0 = f(x_n)
\end{cases}$$

Deoarece determinantul sistemului liniar este un determinant de tip Vandermonde, avem:

$$\det = \begin{vmatrix} x_0^n & x_0^{n-1} & \dots & x_0 & 1 \\ x_1^n & x_1^{n-1} & \dots & x_1 & 1 \\ \dots & \dots & \dots & \dots & \dots \\ x_n^n & x_n^{n-1} & \dots & x_n & 1 \end{vmatrix} = \begin{vmatrix} x_0^n & x_1^n & \dots & x_n^n \\ x_0^{n-1} & x_1^{n-1} & \dots & x_n^{n-1} \\ \dots & \dots & \dots & \dots \\ x_0 & x_1 & \dots & x_n \\ 1 & 1 & \dots & 1 \end{vmatrix} =$$

$$= (-1)^{n+(n-1)+\dots+1} \cdot \begin{vmatrix} 1 & 1 & \dots & 1 \\ x_0 & x_1 & \dots & x_n \\ \dots & \dots & \dots & \dots \\ x_0^{n-1} & x_1^{n-1} & \dots & x_n^{n-1} \\ x_0^n & x_1^n & \dots & x_n^n \end{vmatrix} = (-1)^{\frac{n(n+1)}{2}} \cdot \prod_{n \geq i > j \geq 1} (x_i - x_j) \neq 0$$

fiindcă conform presupunerii nodurile x_i verifică condiţia: $a = x_0 < x_1 < \dots < x_{n-1} < x_n = b$. Astfel se obţine un sistem de tip Cramer cu $n+1$ ecuaţii şi $n+1$ necunoscute, sistem care admite soluţie şi această soluţie este unică. □

Sigur, prin rezolvarea concretă a sistemului liniar cu ajutorul determinanţilor putem obţine la concret coeficienţii a_i, însă în continuare vom prezenta o altă construcţie pentru acest polinom de interpolare, cunoscută sub numele de procedeul lui Lagrange şi de aceea polinomul de interpolare poartă denumirea de polinomul de interpolare al lui Lagrange. Construcţia constă în următoarele: se caută polinomul

$$P(x) = L_n(x) = \sum_{i=0}^{n} L_{ni}(x) \cdot f(x_i),$$

unde pe P am notat cu L_n, numit polinomul de interpolare al lui Lagrange de ordinul n, care se formează cu ajutorul polinoamelor de bază de grad n ale lui Lagrange L_{ni}, în felul următor: condiţia $L_n(x_0) = f(x_0)$ se poate asigura dacă impunem următoarele condiţii: $L_{n0}(x_0) = 1$, $L_{n1}(x_0) = 0, \dots, L_{nn}(x_0) = 0$, fiindcă $L_n(x_0) = L_{n0}(x_0)f(x_0) + L_{n1}(x_0)f(x_1) + \dots + L_{nn}(x_0) \cdot f(x_n) = 1 \cdot f(x_0) + 0 \cdot f(x_1) + \dots + 0 \cdot f(x_n) = f(x_0)$. Pornind de la condiţia $L_n(x_1) = f(x_1)$ în mod similar impunem condiţiile $L_{n0}(x_1) = 0$, $L_{n1}(x_1) = 1$, $L_{n2}(x_1) = 0, \dots, L_{nn}(x_1) = 0$, care asigură valabilitatea egalităţii $L_n(x_1) = f(x_1)$. Tot aşa, pentru a îndeplini ultima condiţie $L_n(x_n) = f(x_n)$ vom impune pe rând condiţiile: $L_{n0}(x_n) = 0, \dots, L_{n,n-1}(x_n) = 0$, $L_{nn}(x_n) = 1$. În acest fel pentru polinomul de bază de gradul n L_{n0} avem următoarele condiţii impuse: $L_{n0}(x_0) = 1$, $L_{n0}(x_1) = 0, \dots, L_{n0}(x_n) =$

0. Prin urmare nodurile $x_1, x_2, \ldots, x_n$ sunt rădăcinile polinomului L_{n0}, deci numai din aceste condiţii rezultă forma $L_{n0}(x) = (x - x_1)(x - x_2)\ldots(x - x_n)$. Însă îndeplinirea egalităţii $L_{n0}(x_0) = 1$ înseamnă normarea polinomului L_{n0}, şi astfel se obţine forma finală:

$$L_{n0}(x) = \frac{(x - x_1)(x - x_2)\ldots(x - x_n)}{(x_0 - x_1)(x_0 - x_2)\ldots(x_0 - x_n)}.$$

În mod analog rezultă pe rând:

$$\begin{aligned} L_{n1}(x) &= \frac{(x - x_0)(x - x_2)\ldots(x - x_n)}{(x_1 - x_0)(x_1 - x_2)\ldots(x_1 - x_n)}, \ldots, \\ L_{nn}(x) &= \frac{(x - x_0)(x - x_1)\ldots(x - x_{n-1})}{(x_n - x_0)(x_n - x_1)\ldots(x_n - x_{n-1})}. \end{aligned}$$

Dacă introducem polinomul ω_n de gradul $n + 1$ prin formula

$$\omega_n(x) = (x - x_0)(x - x_1)\ldots(x - x_n)$$

atunci se obţin următoarele forme pentru polinoamele de bază ale lui Lagrange:

$$L_{n0}(x) = \frac{\omega_n(x)}{(x - x_0) \cdot \omega_n'(x_0)},\ L_{n1}(x) = \frac{\omega_n(x)}{(x - x_1) \cdot \omega_n'(x_1)}, \ldots, L_{nn}(x) = \frac{\omega_n(x)}{(x - x_n) \cdot \omega_n'(x_n)}.$$

Menţionăm că derivarea polinomului ω_n se face ca un produs de polinoame de gradul unu, derivând pe rând fiecare componentă de gradul unu, şi făcând însumarea acestor rezultate:

$$\omega_n'(x) = (x - x_1)\ldots(x - x_n) + (x - x_0)(x - x_2)\ldots(x - x_n) + \cdots + (x - x_0)\ldots(x - x_{n-1}).$$

În final avem polinomul de interpolare al lui Lagrange de ordinul n, dată de formula:

$$L_n(x) = \sum_{i=0}^{n} \frac{\omega_n(x)}{(x - x_i) \cdot \omega'(x_i)} \cdot f(x_i) = \omega_n(x) \cdot \sum_{i=0}^{n} \frac{f(x_i)}{(x - x_i) \cdot \omega_n'(x_i)}.$$

Ca un exemplu concret propunem să se calculeze polinomul de interpolare Lagrange care se asociază funcţiei $f : [4, 16] \to \mathbb{R}$, $f(x) = \sqrt{x}$ luând nodurile $x_0 = 4$, $x_1 = 9$, $x_2 = 16$.

Avem $\omega_3(x) = (x - 4)(x - 9)(x - 16)$, $\omega_3'(x) = (x - 9)(x - 16) + (x - 4)(x - 16) + (x - 4)(x - 9)$,

$$\begin{aligned} L_{20}(x) &= \frac{\omega_3(x)}{(x - 4)\omega_3'(4)} = \frac{(x - 9)(x - 16)}{(4 - 9)(4 - 16)} = \frac{1}{60}(x - 9)(x - 16), \\ L_{21}(x) &= \frac{\omega_3(x)}{(x - 9)\omega_3'(9)} = \frac{(x - 4)(x - 16)}{(9 - 4)(9 - 16)} = -\frac{1}{35}(x - 4)(x - 16), \\ L_{22}(x) &= \frac{\omega_3(x)}{(x - 16)\omega_3'(16)} = \frac{(x - 4)(x - 9)}{(16 - 4)(16 - 9)} = \frac{1}{84}(x - 4)(x - 9). \end{aligned}$$

Astfel

$$\begin{aligned} L_3(x) &= L_{20}(x)\cdot f(4) + L_{21}(x)\cdot f(9) + L_{22}(x)\cdot f(16) = \\ &= \frac{1}{30}(x-9)(x-16) - \frac{3}{35}(x-4)(x-16) + \frac{1}{21}(x-4)(x-9) = \\ &= \frac{1}{210}[7(x-9)(x-16) - 18(x-4)(x-16) + 10(x-4)(x-9)] = \\ &= \frac{1}{210}(-x^2 + 55x + 216). \end{aligned}$$

Practic se obţine parabola (funcţia de gradul al doilea) care trece prin punctele $(4,2)$, $(9,3)$, $(16,4)$.

Ca aplicaţie să se determine valoarea polinomului de interpolare Lagrange $L_n(t)$ într-un punct $t \in [a,b]$ dat.

Program Lagrange

Datele de intrare: f; $x = (x[0], x[1], \ldots, x[n])$; t;
Pentru $i := \overline{0,n}$ execută $y[i] := 1$;
$s := 0$
Pentru $i := \overline{0,n}$ execută
    Pentru $j := \overline{0,n}$ execută
        Dacă $j \neq i$ atunci execută
            $y[i] := \dfrac{t - x[j]}{x[i] - x[j]} * y[i]$;
    $y[i] := y[i] * f(x[i])$;
    $s := s + y[i]$;
Tipăreşte s.

8.2.3 Polinoamele lui Cebâşev de speţa întâia

Pentru a evalua eroarea care se comite atunci când funcţia $f : [a,b] \to \mathbb{R}$ se înlocuieşte prin polinomul său de interpolare Lagrange de ordinul n corespunzător unei diviziuni Δ fixate avem nevoie tehnic de polinoamele lui Cebâşev de speţa întâia.

Definiţia 8.2.1. *Polinoamele lui Cebâşev de speţa întâia se definesc în felul următor:* $T_n : [-1,1] \to \mathbb{R}$, $T_n(x) = \cos(n \arccos x)$ *pentru orice* $x \in [-1,1]$.

Observăm că funcţia T_n este bine definită, fiindcă funcţia arccos are sens pentru orice $x \in [-1,1]$, iar $\arccos x$ este un unghi din intervalul $[0,\pi]$, care se înmulţeşte cu n şi se ia cosinusul unghiului $n \cdot \arccos x$, rezultatul fiind tot un număr real din intervalul [-1, 1]. Astfel putem scrie $T_n : [-1,1] \to [-1,1]$, $T_n(x) = \cos(n \cdot \arccos x)$ pentru orice $x \in [-1,1]$.

Deşi nu se observă din prima vedere totuşi funcţiile T_n sunt legi polinomiale. Pentru $n = 0$ avem $T_0(x) = \cos(0 \cdot \arccos x) = \cos 0 = 1$. Pentru $n = 1$ avem $T_1(x) = \cos(\arccos x) = x$. Pentru $n = 2$ avem $T_2(x) = \cos(2 \cdot \arccos x) = 2 \cdot \cos^2(\arccos x) - 1 = 2x^2 - 1$. Pentru a arăta că pentru un $n \in \mathbb{N}$ arbitrar T_n este tot un polinom vom folosi metoda inducţiei matematice. Plecăm de la formula trigonometrică:

$$\cos(n+1) \cdot \varphi = 2 \cdot \cos\varphi \cdot \cos n\varphi - \cos(n-1)\varphi.$$

Punând $\varphi = \arccos x$ se obţine formula de recurenţă $T_{n+1}(x) = 2x \cdot T_n(x) - T_{n-1}(x)$. Presupunând că T_{n-1} şi T_n sunt polinoame, imediat deducem că şi T_{n+1} este tot un polinom. Astfel avem:

$$\begin{aligned}
T_3(x) &= 2xT_2(x) - T_1(x) = 2x(2x^2-1) - x = 4x^3 - 3x,\\
T_4(x) &= 2xT_3(x) - T_2(x) = 2x(4x^3-3x) - (2x^2-1) = 8x^4 - 8x^2 + 1,\\
T_5(x) &= 2xT_4(x) - T_3(x) = 2x(8x^4-8x^2+1) - (4x^3-3x) = 16x^5 - 20x^3 + 5x, \text{ etc.}
\end{aligned}$$

În continuare enumerăm câteva proprietăţi de bază ale polinoamelor lui Cebâşev de speţa întâia:

1. Dacă n este un număr natural par, atunci polinomul T_n este o funcţie pară, iar dacă n este un număr natural impar, atunci polinomul T_n este o funcţie impară.

   Verificarea acestei proprietăţi putem face printr-o simplă inducţie matematică: $T_0(x) = 1$ este funcţia pară, iar $T_1(x) = x$ este o funcţie impară. Dacă n este impar atunci conform presupunerii T_n este o funcţie impară, iar deoarece $n-1$ va fi un număr par rezultă că T_{n-1} este o funcţie pară. Trebuie să arătăm că T_{n+1} va fi o funcţie pară, $n+1$ fiind un număr par. Într-adevăr:

   $$\begin{aligned}
   T_{n+1}(-x) &= 2 \cdot (-x) \cdot T_n(-x) - T_{n-1}(-x) = 2 \cdot (-x) \cdot (-T_n(x)) - T_{n-1}(x) =\\
   &= 2x \cdot T_n(x) - T_{n-1}(x) = T_{n+1}(x).
   \end{aligned}$$

   Dacă n este par atunci conform presupunerii T_n este o funcţie pară, iar deoarece $n-1$ va fi un număr impar rezultă că T_{n-1} este o funcţie impară. Trebuie să arătăm

că T_{n+1} va fi o funcţie impară, $n+1$ fiind un număr impar. Într-adevăr:

$$\begin{aligned} T_{n+1}(-x) &= 2\cdot(-x)\cdot T_n(-x) - T_{n-1}(-x) = 2\cdot(-x)\cdot T_n(x) - (-T_{n-1}(x)) = \\ &= -2xT_n(x) + T_{n-1}(x) = -(2xT_n(x) - T_{n-1}(x)) = -T_{n+1}(x) \end{aligned}$$

2. Pentru orice $n \geq 1$ număr natural coeficientul termenului maxim din polinomul T_n este egal cu 2^{n-1}. Pentru $n = 1$ avem $T_1(x) = x$, deci coeficientul este $2^0 = 1$. Se poate arăta uşor prin inducţie că T_n este un polinom de grad n având coeficientul lui x^n egal cu 2^{n-1}. Într-adevăr, folosind formula de recurenţă $T_{n+1}(x) = 2x \cdot T_n(x) - T_{n-1}(x)$ deducem că în polinomul T_{n-1} necunoscuta x apare la puterea maximă $n-1$, în polinomul T_n x apare la puterea maximă n şi înmulţit cu x, rezultă în final că în polinomul T_{n+1} x apare la puterea maximă $n+1$. Coeficientul lui x^{n+1} în T_{n+1} se obţine din coeficientul lui x^n din T_n care conform presupunerii este 2^{n-1}, şi prin înmulţirea cu $2x$ în T_{n+1} apare $2^n x^{n+1}$.

3. Polinomul T_n are toate rădăcinile reale şi distincte situate în intervalul $(-1, 1)$ date de formula $x_i = \cos\dfrac{(2i+1)\cdot\pi}{2n}$, cu $i = \overline{0, n-1}$. Într-adevăr, T_n fiind de grad n are cel mult n rădăcini reale. Ne rămâne să verificăm că numerele $x_i = \cos\dfrac{(2i+1)\pi}{2n}$ cu $i = \overline{0, n-1}$ sunt rădăcini pentru T_n :
$T_n(x_i) = \cos\left[n \cdot \arccos\left(\cos\dfrac{(2i+1)\pi}{2n}\right)\right] = \cos\left(n \cdot \dfrac{(2i+1)\cdot\pi}{2n}\right) = \cos\dfrac{(2i+1)\pi}{2} = 0.$

4. Polinomul lui Cebâşev de speţa întâia $T_n : [-1, 1] \to [-1, 1]$ ia valorile extreme -1 şi 1 în punctele $x_m = \cos\dfrac{m\pi}{n}$, cu $m = \overline{0, n}$. Într-adevăr, vrem să verificăm când are loc egalitatea $T_n(x) = \pm 1$. Aceasta înseamnă că $\cos(n \cdot \arccos x) = \pm 1$, adică $n \cdot \arccos x = m \cdot \pi \in [0, n\pi]$ cu $m = \overline{0, n}$. Prin urmare $\arccos x = \dfrac{m\pi}{n}$, deci $x_m = \cos\dfrac{m\pi}{n}$, cu $m = \overline{0, n}$. Aşadar $T_n(x_m) = \cos m\pi = (-1)^m$ pentru $m = \overline{0, n}$.

În continuare enunţăm o proprietate remarcabilă a polinoamelor lui Cebâşev de speţa întâia, fapt pentru care le-am considerat noi în cadrul acestui curs:

Teorema 8.2.2. *Printre polinoamele de grad n cu coeficientul termenului de grad maxim egal cu unu, polinoamele* $\bar{T}_n : [-1, 1] \to \mathbb{R}$, $\bar{T}_n(x) = 2^{1-n} \cdot T_n(x)$, $n \geq 1$, *au cea mai mică abatere de la zero pe intervalul [-1, 1], adică pentru orice alt polinom* $\bar{P}_n : [-1, 1] \to \mathbb{R}$, *de grad n cu coeficientul termenului de grad maxim egal cu unu avem*

$$\max_{x\in[-1,1]} |\bar{P}_n(x)| \geq \max_{x\in[-1,1]} |\bar{T}_n(x)| = 2^{1-n}.$$

DEMONSTRAŢIE. Demonstraţia teoremei începem cu o mică observaţie: deoarece $T_n : [-1,1] \to [-1,1]$ şi are gradul n, coeficientul lui x^n fiind 2^{n-1} rezultă că $\bar{T}_n(x) = 2^{1-n} \cdot T_n(x)$ duce intervalul [-1, 1] în intervalul $[-2^{1-n}, 2^{1-n}]$ fiind tot un polinom de grad n având coeficientul termenului lui x^n egal cu $2^{n-1} \cdot 2^{1-n} = 2^0 = 1$.

Demonstraţia teoremei facem cu metoda reducerii la absurd. Presupunem că există un polinom

$$\bar{P}_n(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$$

pentru care

$$\max_{x\in[-1,1]} |\bar{P}_n(x)| < \max_{x\in[-1,1]} |\bar{T}_n(x)| = 2^{1-n}.$$

Considerăm polinomul $\bar{T}_n - \bar{P}_n$, care are gradul maxim mai mic sau egal cu $n-1$. Deoarece

$$\max_{x\in[-1,1]} |\bar{P}_n(x)| < \max_{x\in[-1,1]} |\bar{T}_n(x)|$$

rezultă că $\bar{P}_n \neq \bar{T}_n$, deci $\bar{T}_n - \bar{P}_n$ este diferit de polinomul nul. Luând valorile $x_m = \cos\dfrac{m\pi}{n}$ pentru $m = \overline{0,n}$ calculăm $\bar{T}_n(x_m) - \bar{P}_n(x_m)$. Deoarece $\bar{T}_n(x_m) = 2^{1-n} \cdot (-1)^m$ şi cum

$$\max_{x\in[-1,1]} |\bar{P}_n(x)| < \max_{x\in[-1,1]} |\bar{T}_n(x)| = 2^{1-n}$$

rezultă că în punctele x_m, $m = \overline{0,n}$ semnul valorii $\bar{T}_n(x_m) - \bar{P}_n(x_m)$ este determinat de semnul numărului $\bar{T}_n(x_m)$. Prin urmare polinomul $\bar{T}_n - \bar{P}_n$ îşi schimbă semnul de n ori, ceea ce implică că admite n rădăcini reale şi distincte. Însă polinomul $\bar{T}_n - \bar{P}_n$ are gradul maxim $n-1$, deci va fi polinomul nul, adică $\bar{T}_n - \bar{P}_n \equiv 0$, ceea ce înseamnă o contradicţie. □

Ca problemă propunem să se calculeze valoarea polinomului Cebâşev de speţa întâia $T_n(x)$ în punctul $x \in [-1,1]$ dat, pentru o valoare $n \geq 2$ dată, folosind relaţia de recurenţă: $T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$.

Program polinoame Cebâşev (vezi la capitolul de evaluare a funcţiilor date prin relaţii de recurenţă, paragraful 4.3).

8.2.4 Evaluarea restului pentru polinomul de interpolare Lagrange

Fie $f : [a,b] \to \mathbb{R}$ o funcţie dată, $\Delta : a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b$ o diviziune dată a intervalului $[a,b]$ şi fie $L_n : [a,b] \to \mathbb{R}$ polinomul de interpolare Lagrange care se asociază la funcţia f şi diviziunea Δ. Evident L_n aproximează funcţia iniţială f. Putem scrie $f(x) = L_n(x) + R_n(x)$, unde $R_n : [a,b] \to \mathbb{R}$ se numeşte restul dintre funcţia f şi polinomul său de interpolare L_n. Vrem să evaluăm restul într-un punct x dat din intervalul $[a,b]$, $x \neq x_i$, $i = \overline{0,n}$. Punând $x = x_i$ în egalitatea $f(x) = L_n(x) + R_n(x)$ se obţine $f(x_i) = L_n(x_i) + R_n(x_i)$ adică $R_n(x_i) = 0$ pentru $i = \overline{0,n}$, adică funcţia R_n admite punctele $x = x_i$ ca rădăcini. Dacă $\omega_n(x) = \prod_{i=0}^{n}(x - x_i)$, atunci putem scrie $R_n(x) = \omega_n(x) \cdot r_n(x)$. Fie acum $x \in [a,b]$ fixat şi $x \neq x_i$ pentru $i = \overline{0,n}$. Introducem funcţia $\varphi : [a,b] \to \mathbb{R}$ dată de formula $\varphi(t) = L_n(t) + \omega_n(t) \cdot r_n(x) - f(t)$. Observăm că

$$\begin{aligned} \varphi(x_i) &= L_n(x_i) + \omega_n(x_i) \cdot r_n(x) - f(x_i) = (L_n(x_i) - f(x_i)) + \omega_n(x_i) \cdot r_n(x) = \\ &= 0 + 0 \cdot r_n(x) = 0 \end{aligned}$$

pentru orice $i = \overline{0,n}$ şi încă $\varphi(x) = L_n(x) + \omega_n(x) \cdot r_n(x) - f(x) = 0$, adică φ se anulează în $n+2$ puncte distincte. Impunem cerinţa ca $f \in C^{n+1}([a,b],\mathbb{R})$, adică este o funcţie derivabilă de $n+1$ ori şi cu derivata de ordinul $n+1$ continuă. Prin urmare şi funcţia $\varphi \in C^{n+1}([a,b],\mathbb{R})$. Deoarece φ se anulează în $n+2$ puncte din teorema lui Rolle rezultă că φ' se anulează în $n+1$ puncte, φ'' se anulează în n puncte, ş.a.m.d. şi în final $\varphi^{(n+1)}$ se anulează cel puţin o dată în intervalul $[a,b]$, adică există un $\xi \in [a,b]$ astfel încât $\varphi^{(n+1)}(\xi) = 0$. Derivăm formula

$$\varphi(t) = L_n(t) + \omega_n(t) \cdot r_n(x) - f(t)$$

în raport cu t de $n+1$ ori. Deoarece L_n este un polinom de grad n, derivata lui de ordinul $n+1$ este zero, iar ω_n fiind un polinom de grad $n+1$ având coeficientul lui x^{n+1} egal cu unu, derivata lui de ordinul $n+1$ va fi egală cu $(n+1)!$. Prin urmare $\varphi^{(n+1)}(t) = 0 + (n+1)! \cdot r_n(x) - f^{(n+1)}(t)$, şi din condiţia $\varphi^{(n+1)}(\xi) = 0$ se obţine $(n+1)! \cdot r_n(x) - f^{(n+1)}(\xi) = 0$, de unde $r_n(x) = \dfrac{f^{(n+1)}(\xi)}{(n+1)!}$. Prin urmare pentru restul R_n avem reprezentarea

$$R_n(x) = \omega_n(x) \cdot \frac{f^{(n+1)}(\xi)}{(n+1)!}.$$

Conform presupunerii $f^{(n+1)}$ este continuă pe $[a,b]$, rezultă în conformitate cu teorema lui Weierstrass că este mărginită pe $[a,b]$, adică există o constantă $M > 0$ astfel ca $|f^{(n+1)}(x)| \leq M$ pentru orice $x \in [a,b]$. În acest fel pentru evaluarea restului obţinem

$$\begin{aligned} |f(x) - L_n(x)| &= |R_n(x)| \leq \sup_{x\in[a,b]} |R_n(x)| = \sup_{x\in[a,b]} \left|\omega_n(x) \cdot \frac{f^{(n+1)}(\xi)}{(n+1)!}\right| \leq \\ &\leq \sup_{x\in[a,b]} |\omega_n(x)| \cdot \frac{M}{(n+1)!}. \end{aligned}$$

Imediat se pune problema de a alege punctele de diviziune x_i, $i = \overline{0,n}$ încât $\sup|\omega_n(x)|$ să fie minim, adică abaterea polinomului ω_n de la zero să fie minimă. Dacă facem transformarea intervalului $[a,b]$ în intervalul [-1, 1] prin aplicaţia $z = -1 + \dfrac{x-a}{b-a} \cdot 2$, unde $x \in [a,b]$ iar $z \in [-1,1]$, atunci se pune problema de a determina nodurile $z_i \in [-1,1]$ pentru care ω_n are abaterea minimă de la zero pe intervalul [-1, 1]. Evident răspunsul la această întrebare îl constituie polinomul lui Cebâşev de speţa întâia de ordinul n, notat de noi cu T_n. În general problema este o problema deschisă.

8.2.5 Teorema lui Faber asupra divergenţei procedeului de interpolare

Fixăm ca spaţiu de lucru spaţiul Banach $(C([0,1],\mathbb{R}), \|\cdot\|)$ unde $C([0,1],\mathbb{R}) = \{f : [0,1] \to \mathbb{R} \mid f \text{ este continuă}\}$ iar norma $\|\cdot\|$ este norma supremum sau norma maximum sau norma Cebâşev:

$$\|f\| = \sup\{|f(x)| \ / \ x \in [0,1]\} = \max\{|f(x)| \ / \ x \in [0,1]\}.$$

Fie $X = \{x_n = (x_n^1, x_n^2, \ldots, x_n^n) \mid n \in \mathbb{N}\}$ sistemul punctelor de diviziune: $x_n^i \in [0,1]$, $x_n^i \neq x_n^j$ pentru $i \neq j$, $x_n^1 < x_n^2 < \cdots < x_n^n$. Putem forma tabelul cu şirul nodurilor:

$$X = \begin{pmatrix} x_1^1 & & & & & \\ x_2^1 & x_2^2 & & & & \\ x_3^1 & x_3^2 & x_3^3 & & & \\ \vdots & \vdots & \vdots & \ddots & & \\ x_n^1 & x_n^2 & x_n^3 & \ldots & x_n^n & \\ \vdots & \vdots & \vdots & & \vdots & \ddots \end{pmatrix}$$

În continuare cu L_n vom nota operatorul lui Lagrange de ordinul n în locul polinomului de interpolare Lagrange de ordinul n, iar pentru polinomul de interpolare Lagrange de ordinul n asociat la o funcţie f vom folosi notaţia $L_n(f)$. Fie deci $L_n : C([0,1],\mathbb{R}) \to C([0,1],\mathbb{R})$ operatorul lui Lagrange care asociază la orice funcţie continuă $f \in C([0,1],\mathbb{R})$ polinomul de interpolare al lui Lagrange de ordinul n, $L_n(f)$ dat de formula

$$L_n(f)(x) = \sum_{i=1}^{n} L_{ni}(x) f(x_n^i),\ x \in [0,1],$$

unde $L_{ni} : [0,1] \to \mathbb{R}$ sunt polinoamele de bază ale lui Lagrange, date de formulele:

$$L_{ni}(x) = \frac{(x-x_n^1)\dots(x-x_n^{i-1})(x-x_n^{i+1})\dots(x-x_n^n)}{(x_n^i-x_n^1)\dots(x_n^i-x_n^{i-1})(x_n^i-x_n^{i+1})\dots(x_n^i-x_n^n)},$$

pentru orice $i = \overline{1,n}$.

Propoziţia 8.2.1. *Operatorul de interpolare Lagrange L_n este linear şi continuu şi norma sa $\|L_n\|$ este dată de $\|L_n\| = \lambda_n$, unde λ_n se numesc constantele lui Lebesque şi sunt date de formulele $\lambda_n = \max\{l_n(x) \mid x \in [0,1]\}$, unde $l_n(x) = \sum_{i=1}^{n} |L_{ni}(x)|$.*

Demonstraţie. Operatorul Lagrange L_n este aditiv: $L_n(f+g) = L_n(f) + L_n(g)$, pentru orice $f, g \in C[0,1]$. Într-adevăr, avem:

$$\begin{aligned} L_n(f+g)(x) &= \sum_{i=1}^{n} L_{ni}(x)\cdot(f+g)(x_n^i) = \sum_{i=1}^{n} L_{ni}(x)\cdot f(x_n^i) + \sum_{i=1}^{n} L_{ni}(x)\cdot g(x_n^i) = \\ &= L_n(f)(x) + L_n(g)(x) = (L_n(f) + L_n(g))(x) \end{aligned}$$

pentru orice $x \in [0,1]$.

Operatorul Lagrange L_n este omogen: $L_n(\alpha f) = \alpha \cdot L_n(f)$ pentru orice $\alpha \in \mathbb{R}$ şi orice $f \in C[0,1]$. Într-adevăr

$$\begin{aligned} L_n(a\ f)(x) &= \sum_{i=1}^{n} L_{ni}(x)\cdot(\alpha f)(x_n^i) = \sum_{i=1}^{n} L_{ni}(x)\cdot\alpha\cdot f(x_n^i) = \\ &= \alpha\cdot\sum_{i=1}^{n} L_{ni}(x)\cdot f(x_n^i) = \alpha\cdot L_n(f)(x) = (\alpha\cdot L_n(f))(x) \end{aligned}$$

pentru orice $x \in [0,1]$. Prin urmare operatorul L_n este linear, fiind aditiv şi omogen.

Operatorul Lagrange L_n este continuu. Trebuie să demonstrăm că există o constantă $M > 0$ astfel încât $\|L_n(f)\| \leq M \cdot \|f\|$ pentru orice $f \in C[0,1]$. Într-adevăr:

$$\begin{aligned}
\|L_n(f)\| &= \max\{|L_n(f)(x)| \ / \ x \in [0,1]\} = \\
&= \max\left\{\left|\sum_{i=1}^{n} L_{ni}(x) f(x_n^i)\right| \ / \ x \in [0,1]\right\} \leq \\
&\leq \max\left\{\sum_{i=1}^{n} |L_{ni}(x)| \cdot |f(x_n^i)| \ / \ x \in [0,1]\right\} \leq \\
&\leq \max\left\{\sum_{i=1}^{n} |L_{ni}(x)| \cdot \|f\| \ / \ x \in [0,1]\right\} = \\
&= \max\left\{\sum_{i=1}^{n} |L_{ni}(x)| \ / \ x \in [0,1]\right\} \cdot \|f\| = \\
&= \max\{l_n(x) \ / \ x \in [0,1]\} \cdot \|f\| = \lambda_n \cdot \|f\|,
\end{aligned}$$

deci putem alege constanta $M = \lambda_n$. Deoarece norma lui L_n este cea mai mică constantă M care verifică inegalitatea $\|L_n(f)\| \leq M\|f\|$ pentru orice $f \in C([0,1], \mathbb{R})$, rezultă că $\|L_n\| \leq \lambda_n$. Pe de altă parte: $\|L_n\| = \sup\{\|L_n(f)\| \ / \ \|f\| \leq 1\} \geq \|L_n(f_0)\|$, unde vom alege pe f_0 astfel încât $\|f_0\| \leq 1$ şi $\|L_n(f_0)\| = \lambda_n$. Alegem pe $\tau_n \in [0,1]$ astfel încât $\lambda_n = l_n(\tau_n)$, căci l_n este continuă şi conform teoremei lui Weierstrass îşi atinge maximul. Se construieşte funcţia $f_0 : [0,1] \to \mathbb{R}$ astfel încât $f_0(x_n^i) = \operatorname{sign} L_{ni}(\tau_n) \in \{-1, 0, +1\}$ şi lineară între oricare două puncte din plan de forma $(x_n^i, f_0(x_n^i))$ şi $(x_n^{i+1}, f_0(x_n^{i+1}))$ unde $i = \overline{1, n-1}$. Prin urmare graficul lui f_0 este o linie frântă continuă aflat în banda delimitată de dreptele $y = -1$ şi $y = 1$. Prin urmare $f_0 \in C[0,1]$ şi mai mult $\|f_0\| \leq 1$, din cauza construcţiei. Deci

$$\begin{aligned}
\|L_n\| &\geq \|L_n(f_0)\| \geq |L_n(f_0)(\tau_n)| = \left|\sum_{i=1}^{n} L_{ni}(\tau_n) \cdot f_0(x_n^i)\right| = \\
&= \left|\sum_{i=1}^{n} L_{ni}(\tau_n) \cdot \operatorname{sign} L_{ni}(\tau_n)\right| = \left|\sum_{i=1}^{n} \operatorname{sign} L_{ni}(\tau_n) \cdot L_{ni}(\tau_n)\right| = \\
&= \left|\sum_{i=1}^{n} |L_{ni}(\tau_n)|\right| = \sum_{i=1}^{n} |L_{ni}(\tau_n)| = l_n(\tau_n) = \lambda_n.
\end{aligned}$$

Prin urmare $\|L_n\| \geq \lambda_n$. În consecinţă $\|L_n\| = \lambda_n$.□

În continuarea vom prezenta câteva rezultate auxiliare din analiză şi analiză funcţională de care vom avea nevoie:

Teorema 8.2.3. *(Bernstein, vezi de exemplu: Natanson, Teoria constructivă a funcţiilor) Pentru constanta λ_n avem următoarea evaluare:*

$$\lambda_n > \frac{\ln n}{8 \cdot \sqrt{\pi}}.$$

Definiţia 8.2.2. *Fie S un spaţiu tolopogic, iar $S_0 \subset S$. Vom spune că S_0 este superdensă în S, dacă S_0 este o submulţime de tip G_δ, (adică se poate scrie ca o intersecţie a unei familii numărabile de părţi deschise ale lui S; $S = \bigcap_{n\in\mathbb{N}} G_n$ cu $G_n \subset S$ deschise), $\operatorname{card} S_0 \geq c$ (unde c este puterea continuului; $\operatorname{card}\mathbb{R} = c$), şi $\bar{S}_0 = S$ (adică S_0 este densă în S).*

Teorema 8.2.4. *(Principiul condensării singularităţilor)*

Fie X un spaţiu Banach, Y un spaţiu normat iar $\mathcal{A} \subset L(X,Y) = \{A : X \to Y \mid A$ e lineară şi continuă$\}$ o parte nemărginită uniform, adică astfel încât: $\sup\{\|A\| \ / \ A \in \mathcal{A}\} = +\infty$. Atunci mulţimea $S_{\mathcal{A}} = \{x \in X \ / \ \sup\{\|A(x)\| \ / \ A \in \mathcal{A}\} = +\infty\}$ este superdensă în X.

Teorema 8.2.5. *(Faber) Pentru orice matrice*

$$X = \begin{pmatrix} x_1^1 & & & & & \\ x_2^1 & x_2^2 & & & & \\ x_3^1 & x_3^2 & x_3^3 & & & \\ \vdots & \vdots & \vdots & \ddots & & \\ x_n^1 & x_n^2 & x_n^3 & \dots & x_n^n & \\ \vdots & \vdots & \vdots & & \vdots & \ddots \end{pmatrix}$$

avem că $S = \{f \in C[0,1] \ / \ \sup\{\|L_n(f)\| \ / \ n \in \mathbb{N}\} = +\infty\}$ este superdensă în $C[0,1]$.

Demonstraţie. Se foloseşte principiul condensării singularităţilor. Fie $X = C([0,1],\mathbb{R})$ (care este spaţiu Banach), $Y = C([0,1],\mathbb{R})$ (care este un spaţiu normat cu norma supremum), $\mathcal{A} = \{L_n \mid n \in \mathbb{N}\} \subset L(C([0,1],\mathbb{R}), C([0,1],\mathbb{R}))$. Deoarece

$$\sup\{\|L_n\| \ / \ n \in \mathbb{N}\} = \sup\{\lambda_n \mid n \in \mathbb{N}\} > \sup\left\{\frac{\ln n}{8\sqrt{\pi}} \ / \ n \in \mathbb{N}\right\} = +\infty,$$

rezultă că S este superdensă în $C([0,1],\mathbb{R})$.

Consecinţa 8.2.1. *Există funcţia continuă $f \in C([0,1],\mathbb{R})$ pentru care şirul de polinoame de interpolare Lagrange este divergent. Prin urmare în acest caz dacă mărim numărul de noduri nu rezultă că graficele polinoamelor de interpolare Lagrange corespunzătoare "se vor apropia" de graficul lui f. Practic apare fenomenul de "tijă", adică într-un punct din intervalul [0, 1] graficul polinoamelor de interpolare tinde la $+\infty$ sau la $-\infty$.*

8.2.6 Diferenţe finite şi divizate. Polinomul de interpolare al lui Newton

Scopul introducerii diferenţelor finite şi divizate este de a da o nouă construcţie pentru polinomul de interpolare, cunoscută sub forma polinomului de interpolare al lui Newton.

Fie $f : [a, b] \to \mathbb{R}$ o funcţie dată, iar $h > 0$ un număr real fixat astfel încât $a + h < b$.

Definiţia 8.2.3. *Prin diferenţa finită la dreapta de ordinul unu pentru funcţia f în punctul x vom înţelege o nouă funcţie notată cu $\Delta f : [a, b - h] \to \mathbb{R}$ dată de formula $\Delta f(x) = f(x + h) - f(x)$.*

Se observă că diferenţa finită la dreapta de ordinul unu dată de noi în definiţie este bine definită, căci pentru $x \in [a, b-h]$ avem $x+h \in [a+h, b] \subset [a, b]$. Menţionăm că prin diferenţa de ordinul zero notată cu $\Delta^0 f(x)$ vom înţelege $\Delta^0 f(x) = f(x)$.

Diferenţele finite la dreapta de ordin superior (de ordin n) se definesc în mod recursiv $\Delta^n f = \Delta(\Delta^{n-1} f)$. Pentru $n = 2$ se obţine $\Delta^2 f = \Delta(\Delta f)$. Aici menţionăm că pentru a putea defini diferenţa finită la dreapta de ordinul doi $\Delta^2 f$ pe intervalul $[a, b - 2h]$ avem nevoie ca $h > 0$ să îndeplinească condiţia $a + 2h < b$, şi în mod analog pentru diferenţa finită la dreapta de ordinul n trebuie ca $a + nh < b$ cu $h > 0$, fiindcă $\Delta^n f$ se defineşte pe intervalul $[a, b - nh]$.

Este uşor de arătat că $\Delta^n(\Delta^m f) = \Delta^{n+m} f$, făcând o simplă inducţie matematică fie după n, fie după m.

În continuare deducem şi alte proprietăţi pentru diferenţele finite la dreapta:

1. Diferenţa finită la dreapta de ordinul unu se poate interpreta ca un operator Δ care asociază la orice funcţie f funcţia Δf. Într-adevăr, dacă notăm cu $\mathcal{F}([a, b], \mathbb{R})$ mulţimea funcţiilor definite pe intervalul $[a, b]$ cu valori reale, atunci $\Delta : \mathcal{F}([a, b], \mathbb{R}) \to \mathcal{F}([a, b - h], \mathbb{R})$ este dată de formula $\Delta(f) = \Delta f$ pentru orice $f \in \mathcal{F}([a, b], \mathbb{R})$.

2. Operatorul de diferenţă finită la dreapta de ordinul unu este linear. Într-adevăr, avem $\Delta(f + g) = \Delta(f) + \Delta(g)$, căci

$$\begin{aligned} \Delta(f + g)(x) &= (f + g)(x + h) - (f + g)(x) = \\ &= (f(x + h) + g(x + h)) - (f(x) + g(x)) = \end{aligned}$$

$$= (f(x+h)-f(x))+(g(x+h)-g(x)) = (\Delta f)(x)+(\Delta g)(x) =$$
$$= \Delta(f)(x)+\Delta(g)(x) = (\Delta(f)+\Delta(g))(x)$$

pentru orice $x \in [a, b-h]$, precum şi $\Delta(\alpha f) = \alpha \Delta f$ pentru orice $\alpha \in \mathbb{R}$ şi $f \in \mathcal{F}([a,b],\mathbb{R})$, căci

$$\begin{aligned}\Delta(\alpha f)(x) &= (\alpha f)(x+h)-(\alpha f)(x) = \alpha \cdot f(x+h)-\alpha f(x) = \\ &= \alpha(f(x+h)-f(x)) = \alpha \cdot \Delta f(x)\end{aligned}$$

pentru orice $x \in [a, b-h]$.

3. Cum $\Delta f(x) = f(x+h)-f(x)$ avem

$$\begin{aligned}\Delta^2 f(x) &= \Delta(\Delta f)(x) = \Delta(f(x+h)-f(x)) = \\ &= \Delta f(x+h)-\Delta f(x) = \\ &= (f(x+2h)-f(x+h))-(f(x+h)-f(x)) = \\ &= f(x+2h)-2\cdot f(x+h)+f(x),\end{aligned}$$

iar pentru $n = 3$ putem deduce uşor că:

$$\begin{aligned}\Delta^3 f(x) &= \Delta(\Delta^2 f)(x) = \Delta(f(x+2h)-2\cdot f(x+h)+f(x)) = \\ &= \Delta f(x+2h)-2\Delta f(x+h)+\Delta f(x) = \\ &= (f(x+3h)-f(x+2h))-2(f(x+2h)-f(x+h))+(f(x+h)- \\ &\quad -f(x)) = f(x+3h)-3f(x+2h)+3f(x+h)-f(x).\end{aligned}$$

Generalizând aceste egalităţi pentru diferenţa finită la dreapta de ordinul n vom obţine:

$$\Delta^n f(x) = \sum_{k=0}^{n}(-1)^k \cdot C_n^k \cdot f(x+(n-k)\cdot h).$$

Această egalitate vom demonstra prin inducţie matematică. Este de ajuns să o verificăm pentru $n+1$:

$$\begin{aligned}\Delta^{n+1} f(x) &= \Delta(\Delta^n f)(x) = \Delta\left(\sum_{k=0}^{n}(-1)^k \cdot C_n^k \cdot f(x+(n-k)\cdot h)\right) = \\ &= \sum_{k=0}^{n}(-1)^k \cdot C_n^k \Delta f(x+(n-k)\cdot h) =\end{aligned}$$

$$= \sum_{k=0}^{n}(-1)^k \cdot C_n^k \cdot (f(x+(n-k+1)\cdot h) - f(x+(n-k)\cdot h)) =$$
$$= \sum_{k=0}^{n}(-1)^k \cdot C_n^k \cdot f(x+(n-k+1)\cdot h) - \sum_{k=0}^{n}(-1)^k \cdot C_n^k \cdot f(x+(n-k)\cdot h) =$$

în a doua sumă facem schimbarea indicilor dată de formula $k = l-1$, adică $l = k+1$:

$$= \sum_{k=0}^{n}(-1)^k \cdot C_n^k \cdot f(x+(n-k+1)\cdot h) - \sum_{l=1}^{n+1}(-1)^{l-1} \cdot C_n^{l-1} \cdot f(x+(n-l+1)\cdot h) =$$
$$= C_n^0 \cdot f(x+(n+1)\cdot h) + \sum_{k=1}^{n}(-1)^k \cdot C_n^k \cdot f(x+(n-k+1)\cdot h) +$$
$$+ \sum_{l=1}^{n}(-1)^l \cdot C_n^{l-1} \cdot f(x+(n-l+1)\cdot h) + (-1)^{n+1} \cdot C_n^n \cdot f(x) =$$
$$= C_{n+1}^0 f(x+(n+1)\cdot h) + \sum_{k=1}^{n}(-1)^k \cdot (C_n^k + C_n^{k-1}) \cdot f(x+(n-k+1)\cdot h) +$$
$$+ (-1)^{n+1} \cdot C_{n+1}^{n+1} \cdot f(x) =$$
$$= C_{n+1}^0 \cdot f(x+(n+1)\cdot h) + \sum_{k=1}^{n}(-1)^k \cdot C_{n+1}^k \cdot f(x+(n-k+1)\cdot h) +$$
$$+ (-1)^{n+1} \cdot C_{n+1}^{n+1} \cdot f(x) = \sum_{k=0}^{n+1}(-1)^k \cdot C_{n+1}^k \cdot f(x+(n-k+1)\cdot h).$$

În mod analog se pot defini şi diferenţele finite la stânga de orice ordin. De exemplu pentru $f : [a,b] \to \mathbb{R}$ funcţie dată, $h > 0$ astfel ca $a + h < b$ vom defini diferenţa finită la stânga $\Delta_s f(x) = f(x) - f(x-h)$ pentru orice $x \in [a+h, b]$. Vor avea loc proprietăţi similare cu 1, 2, 3 şi pentru diferenţele finite la stânga.

În continuare dăm un exemplu pentru diferenţele finite la dreapta. Să considerăm funcţia $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^2 - 2x + 3$. În acest caz alegem $h = 1$ şi avem

$$\Delta f(x) = f(x+h) - f(x) = f(x+1) - f(x) = (x+1)^2 - 2(x+1) + 3 - (x^2 - 2x + 3) =$$
$$= x^2 + 2x + 1 - 2x - 2 + 3 - x^2 + 2x - 3 = 2x - 1,$$
$$\Delta^2 f(x) = \Delta(\Delta f)(x) = [2(x+1) - 1] - [2x - 1] = 2, \text{ iar}$$
$$\Delta^3 f(x) = \Delta(\Delta^2 f)(x) = 2 - 2 = 0.$$

Putem constata o proprietate interesantă, şi anume în cazul polinoamelor de grad n diferenţa finită la dreapta de ordinul $n+1$ este egală cu zero.

În continuare să presupunem că domeniul de definiţie al funcţiei f este un interval finit de forma $[a, +\infty)$. Fie $x_0 \in [a, +\infty)$, $h > 0$ un număr real pozitiv fixat şi considerăm nodurile $(x_k)_{k\in\mathbb{N}}$ date sub forma unei progresii aritmetice crescătoare: $x_k = x_0 + k \cdot h$. Dacă notăm cu $f_k = f(x_k)$, atunci avem următoarele relaţii definite şi demonstrate mai anterior scrise sub următoarea formă: $\Delta^0 f_k = f_k$, $\Delta^1 f_k = f_{k+1} - f_k$, $\Delta^2 f_k = f_{k+2} - 2f_{k+1} + f_k$, $\ldots$, $\Delta^{n+1} f_k = \Delta^n(\Delta f_k) = \Delta^n(f_{k+1} - f_k) = \Delta^n f_{k+1} - \Delta^n f_n$, etc. Folosind aceste notaţii putem enunţa următoarea lemă:

Lema 8.2.1. *Dacă $f \in C^{n+1}([x_k, x_{k+n}], \mathbb{R})$ atunci există un punct $\xi \in (x_k, x_{k+n})$ astfel încât $\Delta^n f_k = h^n \cdot f^{(n)}(\xi)$.*

Demonstraţie. Demonstraţia vom face prin inducţie matematică. Pentru $n = 1$ avem $\Delta^1 f_k = h \cdot f'(\xi)$, unde $\xi \in (x_k, x_{k+1})$, adică $f_{k+1} - f_k = h \cdot f'(\xi)$, deci $f(x_{k+1}) - f(x_k) = (x_{k+1} - x_k) \cdot f'(\xi)$, ceea ce este tocmai teorema de medie a lui Lagrange. Demonstraţia se poate termina cu ajutorul inducţiei matematice.

Fie $f : [a, b] \to \mathbb{R}$ o funcţie dată, iar $a = x_0 < \cdots < x_{n-1} < x_n = b$ o diviziune a intervalului $[a, b]$.

Definiţia 8.2.4. *Valorile funcţiei f în punctele diviziunii, adică valorile $f(x_0), f(x_1), \ldots, f(x_n)$ se numesc diferenţele divizate de ordinul zero ale funcţiei f. Numărul, notat cu $f(x_0; x_1) : \dfrac{f(x_1) - f(x_0)}{x_1 - x_0}$, se numeşte diferenţa divizată de ordinul unu a funcţiei f pe nodurile x_0 şi x_1. Diferenţa divizată de ordinul n a funcţiei f pe nodurile $x_0, x_1, \ldots, x_n$ se defineşte în mod recursiv cu ajutorul diferenţelor divizate de ordinul $n - 1$ ale funcţiei f conform formulei:*

$$f(x_0; x_1; \ldots; x_n) = \frac{f(x_1; \ldots; x_n) - f(x_0; \ldots; x_{n-1})}{x_n - x_0}.$$

În continuare enumerăm câteva proprietăţi ale diferenţelor divizate:

Propoziţia 8.2.2. *Diferenţa divizată este simetrică în raport cu argumentele sale şi avem următoarea formulă:*

$$f(x_0; x_1; \ldots; x_n) = \sum_{i=0}^{n} \frac{f(x_i)}{(x_i - x_0) \ldots (x_i - x_{i-1})(x_i - x_{i+1}) \ldots (x_i - x_n)}.$$

Demonstraţie. Demonstraţia se poate face prin inducţie matematică. Noi vom proba valabilitatea formulei pentru $n = 1$ şi $n = 2$ printr-un calcul direct. Fie $n = 1$:

$$f(x_0; x_1) = \frac{f(x_1) - f(x_0)}{x_1 - x_0} = \frac{f(x_0)}{x_0 - x_1} + \frac{f(x_1)}{x_1 - x_0}.$$

Fie $n = 2$:

$$f(x_0; x_1; x_2) = \frac{f(x_1, x_2) - f(x_0, x_1)}{x_2 - x_0} = \frac{1}{x_2 - x_0} \cdot \left[\frac{f(x_2) - f(x_1)}{x_2 - x_1} - \frac{f(x_1) - f(x_0)}{x_1 - x_0} \right] =$$
$$= \frac{f(x_0)}{(x_1 - x_0)(x_2 - x_0)} + \frac{f(x_1)}{x_2 - x_0} \cdot \left[-\frac{1}{x_2 - x_1} - \frac{1}{x_1 - x_0} \right] + \frac{f(x_2)}{(x_2 - x_0)(x_2 - x_1)} =$$
$$= \frac{f(x_0)}{(x_1 - x_0)(x_2 - x_0)} + \frac{f(x_1)}{x_2 - x_0} \cdot \frac{-(x_1 - x_0) - (x_2 - x_1)}{(x_2 - x_1)(x_1 - x_0)} + \frac{f(x_2)}{(x_2 - x_0)(x_2 - x_1)} =$$
$$= \frac{f(x_0)}{(x_1 - x_0)(x_2 - x_0)} + \frac{f(x_1)}{(x_1 - x_0)(x_1 - x_2)} + \frac{f(x_2)}{(x_2 - x_0)(x_2 - x_1)}.$$

Propoziţia 8.2.3. *Dacă nodurile* $(x_k)_{k=\overline{0,n}}$ *sunt echidistante cu pasul* $h = \dfrac{b-a}{n}$, *atunci diferenţa divizată se poate exprima cu ajutorul diferenţelor finite prin formula:* $f(x_0; x_1; \ldots; x_n) = \dfrac{\Delta^n f_0}{n! \cdot h^n}$.

Demonstraţie. Demonstraţia se poate face prin inducţie matematică. Verificăm printr-un calcul direct valabilitatea formulei pentru $n = 1$ şi $n = 2$. Fie $n = 1$:

$$f(x_0; x_1) = \frac{f(x_1) - f(x_0)}{x_1 - x_0} = \frac{f_1 - f_0}{h} = \frac{\Delta f_0}{h}.$$

Fie $n = 2$:

$$f(x_0; x_1; x_2) = \frac{f(x_1; x_2) - f(x_0; x_1)}{x_2 - x_0} = \frac{\dfrac{\Delta f_1}{h} - \dfrac{\Delta f_0}{h}}{2h} = \frac{\Delta f_1 - \Delta f_0}{2h^2} = \frac{\Delta^2 f_0}{2h^2}.$$

Presupunem că formula are loc pentru $n - 1$ şi vom demonstra pentru n :

$$f(x_0; x_1; \ldots; x_n) = \frac{f(x_1; \ldots; x_n) - f(x_0; \ldots; x_{n-1})}{x_n - x_0} = \frac{\dfrac{\Delta^{n-1} f_1}{(n-1)! h^{n-1}} - \dfrac{\Delta^{n-1} f_0}{(n-1)! h^{n-1}}}{n \cdot h} =$$
$$= \frac{\Delta^{n-1} f_1 - \Delta^{n-1} f_0}{n! h^n} = \frac{\Delta^n f_0}{n! h^n} \qquad \square$$

Propoziţia 8.2.4. *Dacă nodurile* $(x_k)_{k=\overline{0,n}}$ *sunt echidistante cu pasul* $h = \dfrac{b-a}{n}$ *şi* $f \in C^{n+1}([x_0, x_n], \mathbb{R}) = C^{n+1}([a, b], \mathbb{R})$, *atunci există un punct* $\xi \in (a, b)$ *astfel încât* $f(x_0; x_1; \ldots; x_n) = \dfrac{f^{(n)}(\xi)}{n!}$.

Demonstraţie. Valabilitatea formulei rezultă imediat din lema 8.2.1 şi propoziţia 8.2.3. $\square$

În continuare prezentăm construcţia lui Newton pentru polinomul de interpolare. Fie $\Delta : a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b$ o diviziune în care se cunosc valorile funcţiei

$f : [a, b] \to \mathbb{R}$.. Să se găsească acel polinom de gradul n, notat cu N_n, pentru care $N_n(x_i) = f(x_i)$ pentru orice $i = \overline{0, n}$. Ideea lui Newton este de a căuta polinomul de interpolare sub forma:

$$N_n(x) = \alpha_0 + \alpha_1(x - x_0) + \alpha_2(x - x_0)(x - x_1) + \cdots + \alpha_n(x - x_0)(x - x_1)\ldots(x - x_{n-1}),$$

unde necunoscutele $\alpha_0, \alpha_1, \ldots, \alpha_n$ se determină din condiţiile de interpolare. Într-adevăr: $N_n(x_0) = f(x_0)$, deci $\alpha_0 = f(x_0); N_n(x_1) = f(x_1)$, adică $\alpha_0 + \alpha_1(x_1 - x_0) = f(x_1)$, de unde $\alpha_1 = \dfrac{f(x_1) - f(x_0)}{x_1 - x_0} = f(x_0; x_1)$; $N_n(x_2) = f(x_2)$, adică $\alpha_0 + \alpha_1(x_2 - x_0) + \alpha_2(x_2 - x_0)(x_2 - x_1) = f(x_2)$, de unde $f(x_0) + f(x_0; x_1)(x_2 - x_0) + \alpha_2(x_2 - x_0)(x_2 - x_1) = f(x_2)$. Avem următorul şir de egalităţi echivalente:

$$\begin{aligned}
&f(x_0; x_1)(x_2 - x_0) + \alpha_2(x_2 - x_0)(x_2 - x_1) = f(x_2) - f(x_0) \Leftrightarrow \\
&\Leftrightarrow f(x_0; x_1) + \alpha_2(x_2 - x_1) = \frac{f(x_2) - f(x_0)}{x_2 - x_0} \Leftrightarrow f(x_0; x_1) + \alpha_2(x_2 - x_1) = f(x_0; x_2) \Leftrightarrow \\
&\Leftrightarrow \alpha_2 = \frac{f(x_0; x_2) - f(x_0; x_1)}{x_2 - x_1} \Leftrightarrow \alpha_2 = f(x_0; x_1; x_2).
\end{aligned}$$

În general se poate demonstra prin inducţie că $\alpha_k = f(x_0; x_1; \ldots; x_k)$ pentru orice $k = \overline{0, n}$. Prin urmare se obţine polinomul de interpolare al lui Newton:

$$\begin{aligned}
N_n(x) = f(x_0) + f(x_0; x_1) \cdot (x - x_0) + f(x_0; x_1; x_2) \cdot (x - x_0)(x - x_1) + \cdots + \\
+ f(x_0; x_1; \ldots; x_n) \cdot (x - x_0)(x - x_1)\ldots(x - x_{n-1}).
\end{aligned}$$

8.3 Aproximarea funcţiilor cu ajutorul funcţiilor spline

Se consideră funcţia $f : [a, b] \to \mathbb{R}$ şi intervalul $[a, b]$ împărţim în n părţi egale alegând pasul $h = \frac{b-a}{n}$ şi nodurile $x_0 = a < x_1 < x_2 < \cdots < x_{n-1} < x_n = b$, date de formulele $x_k = x_0 + k \cdot h = a + k \cdot h$, unde $k = \overline{0, n}$. Prin funcţia spline $S : [a, b] \to \mathbb{R}$ vom înţelege o nouă funcţie definită pe intervalul $[a, b]$ care aproximează funcţia f şi are următoarea construcţie: pe fiecare subinterval de forma $[x_i, x_{i+1}]$, cu $i = \overline{0, n-1}$, $S_i = S/_{[x_i, x_{i+1}]}$ este un polinom algebric de un anumit grad şi cerem ca S să fie continuă pe tot intervalul $[a, b]$ împreună cu derivatele până la un anumit ordin. Această cerinţă practic înseamnă că polinoamele date pe fiecare subinterval trebuie să fie "lipite" în nodurile $x_i, i = \overline{1, n-1}$, prin care asigurăm continuitatea lui S pe tot intervalul $[a, b]$ respectiv şi

derivatele la stânga ale polinaomelor să coincidă cu derivatele la dreapta ale polinoamelor în nodurile $x_i, i = \overline{1, n-1}$, astfel reuşind să impunem condiţia de derivabilitate a lui S pe tot intervalul $[a, b]$. Gradul polinoamelor (sau gradul maxim al polinoamelor) folosite pentru construcţia funcţiei spline S se numeşte gradul funcţiei spline S, iar prin defectul funcţiei spline S vom înţelege diferenţa dintre gradul funcţiei spline şi cel mai mare ordin al derivatelor polinoamelor pentru care funcţia spline S încă este derivabilă.

1. **Funcţia spline de gradul unu**. În acest caz funcţia spline $S : [a, b] \to \mathbb{R}$ pe fiecare subinterval este un polinom de grad maxim unu, deci $S_i(x) = S_{/[x_i, x_{i+1}]}(x) = m_i x + n_i$, unde $m_i, n_i \in \mathbb{R}$ pentru $i = \overline{0, n-1}$. Vrem să "lipim" aceste segmente în nodurile $x_i, i = \overline{1, n-1}$, deci impunem condiţia de continuitate a lui S pe tot intervalul $[a, b]$: pentru orice $i = \overline{0, n-2}$, $S_i(x_{i+1}) = S_{i+1}(x_{i+1}) = f(x_{i+1})$. Prin urmare putem să impunem pe fiecare subinterval următoarele cerinţe: $S_i(x_i) = f(x_i)$ şi $S_i(x_{i+1}) = f(x_{i+1})$ cu $i = \overline{0, n-1}$. Observăm că în acest fel obţinem următorul sistem liniar cu două ecuaţii şi două necunoscute:

$$\begin{cases} m_i x_i + n_i = f(x_i) \\ m_i x_{i+1} + n_i = f(x_{i+1}). \end{cases}$$

   Prin urmare suntem în cazul interpolării liniare (vezi paragraful 8.2.1). În acest caz se determină complet $m_i, n_i \in \mathbb{R}$ pentru $i = \overline{0, n-1}$ şi este de ajuns să alegem $f \in C([a, b], \mathbb{R})$, iar funcţia spline corespunzătoare de gradul unu va fi $S \in C([a, b], \mathbb{R})$ dată de

$$S_{/[x_i, x_{i+1}]}(x) = S_i(x) = \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i} \cdot x + \frac{x_{i+1} \cdot f(x_i) - x_i \cdot f(x_{i+1})}{x_{i+1} - x_i}$$

   cu $i = \overline{0, n-1}$. Observăm că în acest caz nu are rost să cerem f derivabilă, fiindcă nu avem posibilitatea să reglăm funcţia spline S ca să aproximeze şi derivatele lui f. Graficul funcţiei spline S de gradul unu este o linie frântă continuă, care trece prin punctele $(x_i, f(x_i))$ dar nu este derivabilă în punctele de lipire x_i, $i = \overline{1, n-1}$. Prin urmare defectul lui S este tot unu.

2. **Funcţia spline de gradul doi**. În acest caz funcţia spline $S : [a, b] \to \mathbb{R}$ pe fiecare subinterval este un polinom de grad maxim doi, deci $S_i(x) = S_{/[x_i, x_{i+1}]}(x) = a_i x^2 + b_i x + c_i$, unde $a_i, b_i, c_i \in \mathbb{R}$ pentru $i = \overline{0, n-1}$. Prima dată vrem să "lipim" aceste arce

de parabolă între ele în nodurile $x_i, i = \overline{1, n-1}$. Prin aceasta asigurăm continuitatea lui S pe tot intervalul $[a, b]$ şi impunem condiţiile: $S_i(x_{i+1}) = S_{i+1}(x_{i+1}) = f(x_{i+1})$ pentru orice $i = \overline{0, n-2}$. Prin urmare pe fiecare subinterval $[x_i, x_{i+1}]$ avem de impus cerinţele: $S_i(x_i) = f(x_i)$ şi $S_i(x_{i+1}) = f(x_{i+1})$. În acest fel obţinem un sistem liniar cu trei necunoscute şi numai două ecuaţii:

$$\begin{cases} a_i x_i^2 + b_i x_i + c_i = f(x_i) \\ a_i x_{i+1}^2 + b_i x_{i+1} + c_i = f(x_{i+1}). \end{cases} \tag{8.1}$$

Observăm că o necunoscută rămâne nedeterminată. Prin urmare pentru a determina şi a treia necunoscută putem să luăm în considerare şi derivata de ordinul întâi a lui f. Prin urmare în acest caz fie $f \in C^1([a, b], \mathbb{R})$. Dacă vrem să "lipim" şi derivatele de ordinul întâi atunci avem de impus condiţiile: $S_i'(x_{i+1}) = S_{i+1}'(x_{i+1}) = f'(x_{i+1})$ pentru orice $i = \overline{0, n-2}$. Prin urmare în acest fel pe fiecare subinterval $[x_i, x_{i+1}]$ avem următoarele condiţii: $S_i(x_i) = f(x_i)$, $S_i(x_{i+1}) = f(x_{i+1})$, $S_i'(x_i) = S'(x_i + 0) = f'(x_i)$ şi $S_i'(x_{i+1}) = S'(x_{i+1} - 0) = f'(x_{i+1})$ cu $i = \overline{0, n-1}$. Însă în acest fel obţinem un sistem liniar cu patru ecuaţii şi numai trei necuscute: a_i, b_i, c_i, sistem care în general este incompatibil. Deci cu funcţia spline de gradul doi în general pe lângă aproximarea în noduri a funcţiilor ($S(x_i) = f(x_i)$ pentru orice $i = \overline{0, n}$) nu putem asigura aproximarea în noduri şi a derivatelor de ordinul unu ale funcţiilor ($S'(x_i) = f'(x_i)$ pentru orice $i = \overline{0, n}$). Însă ceea ce putem realiza în acest caz este existenţa globală a primei derivate a funcţiei spline S fără a aproxima prima derivată a lui f în nodurile date. Într-adevăr, pentru a determina funcţia spline de gradul doi avem de determinat coeficienţii a_i, b_i, c_i pentru $i = \overline{0, n-1}$, adică avem de determinat $3n$ necunoscute. Sistemele de forma (8.1) ne dau $2n$ legături, deci mai avem de impus încă n legături. Cerem ca funcţia spline S să fie derivabilă în nodurile x_i, $i = \overline{1, n-1}$, adică pentru orice $i = \overline{1, n-1}$ $S_{i-1}'(x_i) = S'(x_i - 0) = S'(x_i + 0) = S_i'(x_i)$, adică: $2a_{i-1}x_i + b_{i-1} = 2a_i x_i + b_i$, $i = \overline{1, n-1}$. Astfel până acum avem în total $2n + n - 1 = 3n - 1$ legături. Pentru ultima legătură putem să cerem de exemplu ca una dintre următoarele două egalităţi să aibă loc

$$\begin{aligned} S'(x_0 + 0) &= S_0'(x_0) = 2a_0 x_0 + b_0 = \frac{f(x_1) - f(x_0)}{x_1 - x_0} = \frac{f(a+h) - f(a)}{h} \text{ sau} \\ S'(x_n - 0) &= S_{n-1}'(x_n) = 2a_{n-1} x_n + b_{n-1} = \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}} = \frac{f(b) - f(b-h)}{h} \end{aligned}$$

Astfel avem în final $3n$ legături cu $3n$ necunoscute. Practic asta înseamnă rezolvarea unui sistem liniar cu $3n$ ecuaţii şi $3n$ necunoscute. Astfel defectul acestei funcţii spline este unu.

3. **Funcţia spline de gradul trei**. În practică cea mai des folosită este funcţia spline de gradul trei. În acest caz funcţia spline $S : [a, b] \to \mathbb{R}$ pe fiecare subinterval este un polinom de gradul trei, deci

$$S_i(x) = S/_{[x_i, x_{i+1}]} = a_i x^3 + b_i x^2 + c_i x + d_i,$$

unde $a_i, b_i, c_i, d_i \in \mathbb{R}$ pentru $i = \overline{0, n-1}$. Prima dată "lipim" aceste arce de curbe de gradul trei, după care "lipim" şi derivatele acestor arce de curbe impunând condiţiile: $S_i(x_i) = f(x_i)$, $S_i(x_{i+1}) = f(x_{i+1})$, $S_i'(x_i) = f'(x_i)$ şi $S_i'(x_{i+1}) = f'(x_{i+1})$. Pentru a realiza aceste condiţii presupunem că $f \in C^1([a, b], \mathbb{R})$ iar curba spline astfel construită este egală pe noduri cu funcţia f şi în plus derivata curbei spline pe noduri coincide cu derivata funcţiei f pe noduri. Într-adevăr, sistemul de condiţii inpus mai sus are forma algebrică:

$$\begin{cases} a_i x_i^3 + b_i x_i^2 + c_i x_i + d_i = f(x_i) \\ a_i x_{i+1}^3 + b_i x_{i+1}^2 + c_i x_{i+1} + d_i = f(x_{i+1}) \\ 3a_i x_i^2 + 2b_i x_i + c_i = f'(x_i) \\ 3a_i x_{i+1}^2 + 2b_i x_{i+1} + c_i = f'(x_{i+1}) \end{cases}$$

cu $i = \overline{0, n-1}$.

Observăm că am obţinut un sistem liniar cu patru ecuaţii şi cu patru necunoscute. Prin rezolvarea sistemului liniar cu ajutorul determinanţilor putem să obţinem valorile lui a_i, b_i, c_i şi d_i, deci pe S_i pentru orice $i = \overline{0, n-1}$. În continuare arătăm că polinomul căutat S_i are următoarea formă:

$$\begin{aligned} S_i(x) \;=\; & \frac{(x - x_{i+1})^2[2(x - x_i) + h]}{h^3} \cdot f(x_i) + \\ & + \frac{(x - x_i)^2 \cdot [2(x_{i+1} - x) + h]}{h^3} \cdot f(x_{i+1}) + \\ & + \frac{(x - x_{i+1})^2 (x - x_i)}{h^2} \cdot f'(x_i) + \frac{(x - x_i)^2 (x - x_{i+1})}{h^2} \cdot f'(x_{i+1}) \end{aligned}$$

pentru orice $i = \overline{0, n-1}$. Într-adevăr, S_i este un polinom de gradul trei în x şi verifică condiţiile: $S_i(x_i) = f(x_i)$, $S_i(x_{i+1}) = f(x_{i+1})$,

$$\begin{aligned} S_i'(x) &= \frac{2(x-x_{i+1})\cdot[2(x-x_i)+h]}{h^3}\cdot f(x_i) + \frac{(x-x_{i+1})^2\cdot(+2)}{h^3}\cdot f(x_i) + \\ &+ \frac{2(x-x_i)\cdot[2(x_{i+1}-x)+h]}{h^3}\cdot f(x_{i+1}) + \frac{(x-x_i)^2\cdot(-2)}{h^3}\cdot f(x_{i+1}) + \\ &+ \frac{2(x-x_{i+1})\cdot(x-x_i)}{h^2}\cdot f'(x_i) + \frac{(x-x_{i+1})^2\cdot 1}{h^2} f'(x_i) + \\ &+ \frac{2(x-x_i)(x-x_{i+1})}{h^2}\cdot f'(x_{i+1}) + \frac{(x-x_i)^2\cdot 1}{h^2}\cdot f'(x_{i+1}), \end{aligned}$$

deci $S_i'(x_i) = f'(x_i)$ şi $S_i'(x_{i+1}) = f'(x_{i+1})$.

Observăm că funcţia spline cubică astfel construită are defectul doi. Se poate realiza ca defectul funcţiei spline de gradul trei să fie unu, dar în acest caz trebuie să renunţăm la faptul că derivata funcţiei spline pe noduri coincide cu derivata funcţiei f pe noduri. Pur şi simplu impunem condiţiile: pentru $i = \overline{0, n-1}$ să avem: $S_i(x_i) = f(x_i)$ şi $S_i(x_{i+1}) = f(x_{i+1})$ şi pentru $i = \overline{1, n-1}$ să avem: $S_{i-1}'(x_i) = S'(x_i - 0) = S'(x_i + 0) = S_i'(x_i)$ şi $S_{i-1}''(x_i) = S''(x_i - 0) = S''(x_i + 0) = S_i''(x_i)$. Astfel obţinem $2n + n - 1 + n - 1 = 4n - 2$ legături. Pentru cele două legături rămase putem să cerem de exemplu valabilitatea următoarelor două egalităţi:

$$\begin{aligned} S'(x_0 + 0) &= S_0'(x_0) = \frac{f(x_1) - f(x_0)}{x_1 - x_0} = \frac{f(a+h) - f(a)}{h} \quad \text{şi} \\ S'(x_n - 0) &= S_{n-1}'(x_n) = \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}} = \frac{f(b) - f(b-h)}{h}. \end{aligned}$$

Ca problemă de algoritmizat propunem să se determine valoarea funcţiei spline cubică $S : [a, b] \to \mathbb{R}$ într-un punct $x \in [a, b]$ dat.

8.4 Cea mai bună aproximare a funcţiilor în spaţii normate

Vom considera noţiunea de aproximare (în cazul special şi cel mai important aproximarea funcţiilor) în cadrul abstract al spaţiilor normate. Vom da definiţia riguroasă precum şi câteva rezultate teoretice în această direcţie.

Fie $(X, \|\cdot\|)$ un spaţiu normat, $Y \subset X$, $Y \neq \emptyset$.

Definiţia 8.4.1. *Dacă $f \in X$ şi $g \in Y$ atunci spunem că f se aproximează cu g sau că g este o aproximaţie pentru elementul f. Dacă $g_1, g_2 \in Y$ sunt două aproximaţii ale lui f şi dacă $\|f - g_1\| < \|f - g_2\|$ atunci spunem că g_1 este o aproximare mai bună decât g_2 pentru f. Un element $g^* \in Y$ pentru care $\|g^* - f\| \leq \|g - f\|$ pentru orice $g \in Y$ se numeşte elementul de cea mai bună aproximare a lui f.*

În continuare enunţăm câteva teoreme cu condiţii suficiente care asigură existenţa elementului de cea mai bună aproximare în spaţii normate. Însă iniţial vom avea nevoie de câteva rezultate din analiză şi analiză funcţională.

Fie deci $(X, \|\cdot\|)$ un spaţiu normat, iar $Y \subset X$, $Y \neq \emptyset$ o submulţime.

Definiţia 8.4.2. *Vom spune că submulţimea Y în spaţiul normat X este o submulţime compactă, dacă din orice acoperire deschisă a lui Y se poate extrage o subacoperire finită (definiţia compacităţii în sensul lui Lebesque).*

Dacă Y este o submulţime compactă atunci este mărginită şi închisă. Însă, invers în general nu are loc într-un spaţiu normat arbitrar. Într-adevăr, avem rezultatul lui Riesz în această direcţie, care afirmă că bila închisă $B(\theta, 1) = \{x \in X \ / \ \|x\| \leq 1\} \subset X$ centrată în originea spaţiului normat θ şi de rază unu (care evident este o submulţime închisă şi mărginită) este compactă dacă şi numai dacă spaţiul normat X este spaţiu linear finit dimensional. În cazul spaţiilor normate finit dimensionale submulţimile compacte se caracterizează prin a fi închise şi mărginite.

O caracterizare a compacităţii într-un spaţiu normat se poate da prin compacitatea secvenţială.

Propoziţia 8.4.1. *Într-un spaţiu normat $(X, \|\cdot\|)$ următoarele afirmaţii sunt echivalente pentru o submulţime $Y \subset X$:*

i) Y este compactă;

ii) orice şir din Y admite cel puţin un subşir convergent în Y.

Teorema 8.4.1. *(de existenţă a elementului de cea mai bună aproximare) Dacă $Y \subset X$, $Y \neq \emptyset$ este o submulţime compactă a spaţiului normat X, atunci orice $f \in X$ admite element de cea mai bună aproximare în Y.*

Demonstraţie. Fie $d = \inf_{g \in Y} \|f - g\| \geq 0$. Dacă există un $g^* \in Y$ astfel încât infimumul este atins pentru acest element g^*, adică $d = \|f - g^*\|$, atunci g^* va fi un element

de cea mai bună aproximare pentru f. În caz contrar există un şir $(g_n)_{n\in\mathbb{N}} \subset Y$ astfel încât $\lim\limits_{n\to\infty} \|f - g_n\| = d$. Dar conform presupunerii Y este o submulţime compactă, deci şi secvenţial compactă, prin urmare din şirul $(g_n)_{n\in\mathbb{N}}$ putem extrage un subşir $(g_{n_k})_{k\in\mathbb{N}} \subset Y$ astfel încât $\lim\limits_{k\to\infty} g_{n_k} = g^* \in Y$. De aici se obţine că $d = \lim\limits_{k\to\infty} \|f - g_{n_k}\| = \|f - g^*\|$, cu $g^* \in Y$. Astfel $g^* \in Y$ este elementul de cea mai bună aproximare pentru $f \in X$. Aici am folosit faptul că norma este o funcţie continuă. □

Teorema 8.4.2. *(de existenţă a elementului de cea mai bună aproximare) Fie $(X, \|\cdot\|)$ un spaţiu normat, $Y \subset X$, $Y \neq \emptyset$ un subspaţiu finit dimensional al lui X. Atunci orice $f \in X$ admite element de cea mai bună aproximare în Y.*

Demonstraţie. Fie $Y_0 = \{g \in Y \ / \ \|g\| \leq 2 \cdot \|f\|\}$. Avem $Y_0 \neq \emptyset$ fiindcă Y_0 conţine elementul neutru θ şi $\|\theta\| = 0 \leq 2 \cdot \|f\|$. Submulţimea Y_0 este o submulţime închisă şi mărginită dintr-un spaţiu finit dimensional, fiind o bilă închisă cu centrul în θ şi de rază $2 \cdot \|f\|$. Prin urmare Y_0 este o submulţime compactă şi conform teoremei 8.4.1 pentru f există un element de cea mai bună aproximare $g^* \in Y_0$. Vom arăta că acest element g^* va fi elementul de cea mai bună aproximare a lui f relativ şi la submulţimea Y. Într-adevăr fie $g \in Y - Y_0$. Atunci $\|g\| > 2 \cdot \|f\|$. Vom avea pe rând:

$$\|f - g\| \ \geq \ |\, \|f\| - \|g\| \,| = \|g\| - \|f\| > 2 \cdot \|f\| - \|f\| = \|f\| = \|f - \theta\| \geq \|f - g^*\|.$$

Deci g^* va fi elementul de cea mai bună aproximare pentru Y. □

Lema 8.4.1. *Într-un spaţiu Hilbert $(X, (\cdot,\cdot))$ are loc formula paralelogramului: $\|f+g\|^2 + \|f-g\|^2 = 2 \cdot (\|f\|^2 + \|g\|^2)$ pentru orice $f, g \in X$.*

Demonstraţie. Avem pe rând:

$$\begin{aligned}\|f+g\|^2 + \|f-g\|^2 &= (f+g, f+g) + (f-g, f-g) = \\ &= (f,f) + (g,f) + (f,g) + (g,g) + (f,f) - (g,f) - (f,g) + (g,g) = \\ &= 2 \cdot ((f,f) + (g,g)) = 2 \cdot (\|f\|^2 + \|g\|^2) \ \square\end{aligned}$$

Teorema 8.4.3. *(de existenţă şi unicitate a elementului de cea mai bună aproximare în spaţii Hilbert). Dacă $(X, (\cdot,\cdot))$ este un spaţiu Hilbert, $Y \subset X$, $Y \neq \emptyset$ este convexă şi închisă atunci pentru orice element $f \in X$ există în mod unic un element de cea mai bună aproximare $g^* \in Y$.*

DEMONSTRAŢIE. Dacă $f \in Y$ atunci să alegem $g^* = f$ şi avem elementul de cea mai bună aproximare. Dacă $f \notin Y$ atunci fie $d = \inf_{g \in Y} \|f - g\| \geq 0$. Din definiţia infimumului rezultă că există un şir $(g_n)_{n \in \mathbb{N}} \subset Y$ astfel încât $\lim_{n \to \infty} \|f - g_n\| = d$. Aplicăm formula paralelogramului:

$$2\left(\|f - g_m\|^2 + \|f - g_n\|^2\right) = \|2f - (g_m + g_n)\|^2 + \|g_m - g_n\|^2,$$

adică

$$\|g_m - g_n\|^2 = 2 \cdot \left(\|f - g_m\|^2 + \|f - g_n\|^2\right) - 4 \cdot \left\|f - \frac{g_m + g_n}{2}\right\|^2.$$

Submulţimea Y fiind convexă rezultă că $\frac{g_m + g_n}{2} \in Y$, deci $\left\|f - \frac{g_m + g_n}{2}\right\| \geq d$, adică

$$\|g_m - g_n\|^2 \leq 2\left(\|f - g_m\|^2 + \|f - g_n\|^2\right) - 4d^2.$$

Făcând ca n şi m să tindă la ∞ vom obţine

$$0 \leq \lim_{n,m \to \infty} \|g_m - g_n\|^2 \leq 2\left(\lim_{m \to \infty} \|f - g_m\|^2 + \lim_{n \to \infty} \|f - g_n\|^2\right) - 4d^2 = 2(d^2 + d^2) - 4d^2 = 0.$$

Prin urmare şirul $(g_n)_{n \in \mathbb{N}} \subset Y$ este un şir Cauchy sau fundamental. Dar X fiind un spaţiu Hilbert, este complet, deci există un element $g^* \in X$ astfel ca $\lim_{n \to \infty} g_n = g^*$. Însă Y este şi închisă, deci va conţine limita şirului $(g_n) \subset Y$, punctul g^*. Vom arăta că $g^* \in Y$ este elementul de cea mai bună aproximare pentru $f \in X$. Într-adevăr, din $\lim_{n \to \infty} \|f - g_n\| = d$ rezultă că $\|f - g^*\| = d$, folosind continuitatea normei. Să arătăm că elementul de cea mai bună aproximare a lui f este unic. Presupunem prin absurd că $g^{**} \in Y$ ar fi un alt element astfel ca $\|f - g^{**}\| = d$. Y fiind convexă, avem $\frac{g^* + g^{**}}{2} \in Y$, deci $\left\|f - \frac{g^* + g^{**}}{2}\right\| \geq d$. Folosind iarăşi formula paralelogramului:

$$2\left(\|f - g^*\|^2 + \|f - g^{**}\|^2\right) = \|2 \cdot f - (g^* + g^{**})\| + \|g^* - g^{**}\|^2,$$

obţinem

$$\begin{aligned} 0 \leq \|g^* - g^{**}\|^2 &= 2(d^2 + d^2) - \|2f - (g^* + g^{**})\|^2 = \\ &= 4d^2 - 4 \cdot \left\|f - \frac{g^* + g^{**}}{2}\right\|^2 \leq 4d^2 - 4d^2 = 0, \end{aligned}$$

deci $\|g^* - g^{**}\| = 0$, adică $g^* = g^{**}$. □

Capitolul 9

Formulele de derivare şi integrare numerică

9.1 Formulele de derivare numerică

Formulele de derivare numerică sunt nişte formule care dau valori oricât "de bune", oricât de precise într-un punct dat pentru derivatele de orice ordin ale unei funcţii derivabile. Se cunoaşte definiţia derivatei de ordinul întâi într-un punct dat. Fie $f : (a, b) \to \mathbb{R}$ o funcţie derivabilă în punctul $x_0 \in (a, b)$. Atunci există

$$\lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} = f'(x_0) \in \mathbb{R}.$$

Dacă notăm cu $h = x - x_0$ atunci $x = x_0 + h$ şi avem următoarea definiţie a derivabilităţii unei funcţii într-un punct dat: există

$$\lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h} = f'(x_0).$$

Astfel dacă f este derivabilă o dată în punctul x_0, atunci pentru $h > 0$ cu h "foarte aproape" de zero ($h = 10^{-2}, h = 10^{-3}$, etc.) formula $\frac{f(x_0+h)-f(x_0)}{h}$ ne dă o valoare "apropiată" de $f'(x_0)$. Astfel obţinem prima formulă de derivare numerică, şi anume dacă se dă funcţia $f \in C^1((a, b), \mathbb{R})$ şi punctul $x_0 \in (a, b)$ atunci pentru $h > 0$, $h \approx 0$ valoarea fracţiei

$$\frac{f(x_0 + h) - f(x_0)}{h} \approx f'(x_0).$$

În continuarea vrem să demonstrăm convergenţa acestei formule, adică cu cât h ia valori pozitive din ce în ce mai apropiate de zero, valorile formulei de derivare numerică $\frac{f(x_0+h)-f(x_0)}{h}$ tind către $f'(x_0)$, şi în acelaşi timp putem da evaluarea erorii:

Propoziţia 9.1.1. *Fie* $f \in C^2([a,b],\mathbb{R})$, $x_0 \in (a,b)$ *şi* $h > 0$ *cu* $x_0 + h \in [a,b]$. *Atunci*

$$\left|\frac{f(x_0+h)-f(x_0)}{h} - f'(x_0)\right| \leq \frac{M_2}{2}\cdot h, \quad \textit{unde}$$
$$M_2 = \sup\{|f''(x)| \ / \ x \in [a,b]\} = \max\{|f''(x)| \ / \ x \in [a,b]\}.$$

DEMONSTRAŢIE. Deoarece $f \in C^2([a,b],\mathbb{R})$ şi $x_0 \in (a,b)$ considerăm dezvoltarea în serie Taylor a lui f în punctul x_0 cu restul sub forma Lagrange de ordinul doi:

$$f(x_0+h) = f(x_0) + \frac{f'(x_0)}{1!}\cdot h + \frac{f''(\xi)}{2!}\cdot h^2,$$

unde $\xi \in (x_0, x_0+h) \subset [a,b]$. Deci

$$\left|\frac{f(x_0+h)-f(x_0)}{h} - f'(x_0)\right| = \left|\frac{f''(\xi)}{2}\cdot h\right| \leq \frac{M_2}{2}\cdot h,$$

unde M_2 este o constantă care depinde de f şi nu de h. Menţionăm că deoarece $f \in C^2([a,b],\mathbb{R})$, deci f'' este continuă, conform teoremei lui Weierstrass f'' este mărginită pe $[a,b]$ şi îşi atinge marginile, deci există constanta finită

$$M_2 = \sup\{|f''(x)| \ / \ x \in [a,b]\} = \max\{|f''(x)| \ / \ x \in [a,b]\}. \text{ q.e.d.}$$

Din evaluarea prezentată în propoziţia anterioară deducem că dacă ne interesează valoarea lui $f'(x_0)$ cu o precizie $\varepsilon > 0$ atunci alegem h astfel ca să aibă loc inegalitatea $\frac{M_2}{2}\cdot h < \varepsilon$, adică $h < \frac{2\varepsilon}{M_2}$, din care rezultă că

$$\left|\frac{f(x_0+h)-f(x_0)}{h} - f'(x_0)\right| < \varepsilon.$$

În continuare prezentăm un algoritm pentru calculul lui $f'(x_0)$ cu o precizie $\varepsilon > 0$ folosind o condiţie practică de oprire:

Algoritm 1 prima derivată

Datele de intrare: $f; x_0; \varepsilon$;
Fie $h = 10^{-1}$; $A := \dfrac{f(x_0+h)-f(x_0)}{h}$;
Repetă: $B := A$; $h := \dfrac{h}{2}$ $\left(\text{sau } h := \dfrac{h}{10}\right)$; $A := \frac{f(x_0+h)-f(x_0)}{h}$;
Până când: $|B - A| \geq \varepsilon$;
Tipăreşte A.

Deoarece

$$\lim_{\substack{h\to 0\\ h<0}} \frac{f(x_0+h)-f(x_0)}{h} = f'(x_0),$$

notând cu $t=-h>0$ avem

$$\lim_{\substack{t\to 0\\ t>0}} \frac{f(x_0-t)-f(x_0)}{-t} = \lim_{\substack{t\to 0\\ t>0}} \frac{f(x_0)-f(x_0-t)}{t} = f'(x_0).$$

Prin urmare expresia $\frac{f(x_0)-f(x_0-t)}{t}$ pentru o valoare $t>0$ şi $t\approx 0$ generează o nouă formulă de derivare numerică pentru $f'(x_0)$.

Propoziţia 9.1.2. *Fie $f\in C^2([a,b],\mathbb{R})$ şi $x_0\in(a,b)$ şi $t>0$ astfel ca $x_0-t\in[a,b]$. Atunci*

$$\left|\frac{f(x_0)-f(x_0-t)}{t}-f'(x_0)\right| \le \frac{M_2}{2}t, \quad \textit{unde}$$
$$M_2=\sup\{|f''(x)| \ / \ x\in[a,b]\}=\max\{|f''(x)| \ / \ x\in[a,b]\}.$$

Demonstraţie. Deoarece $f\in C^2([a,b],\mathbb{R})$, $x_0\in(a,b)$ şi $t>0$ cu $x_0-t\in[a,b]$ vom folosi dezvoltarea lui f în serie Taylor în punctul x_0 cu restul sub forma Lagrange de ordinul doi:

$$f(x_0-t)=f(x_0)-\frac{f'(x_0)}{1!}\cdot t+\frac{f''(\xi)}{2!}t^2,$$

unde $\xi\in(x_0-t,x_0)$. De aici deducem că

$$\frac{f(x_0)-f(x_0-t)}{t}-f'(x_0)=-\frac{f''(\xi)}{2}\cdot t,$$

adică

$$\left|\frac{f(x_0)-f(x_0-t)}{t}-f'(x_0)\right|=\left|\frac{f''(\xi)}{2}\cdot t\right|\le\frac{M_2}{2}\cdot t, \text{ q.e.d.}$$

Prin urmare, dacă vrem să determinăm valoarea lui $f'(x_0)$ cu o precizie $\varepsilon>0$ dată cu ajutorul formulei de derivare numerică $\dfrac{f(x_0)-f(x_0-t)}{t}$ se foloseşte condiţia teoretică ca $\dfrac{M_2}{2}\cdot t\le\varepsilon$ de unde $t\le\dfrac{2\varepsilon}{M_2}$.

În continuare dăm un algoritm cu o condiţie practică de oprire:

Algoritm 2 prima derivată

Datele de intrare: $f; x_0; \varepsilon$;

Fie $t:=10^{-1}$; $A:=\dfrac{f(x_0)-f(x_0-t)}{t}$;

Repetă $B := A$; $t := \dfrac{t}{2}$ $\left(\text{sau } t := \dfrac{t}{10}\right)$; $A := \dfrac{f(x_0) - f(x_0 - t)}{t}$;

Până când $|B - A| \geq \varepsilon$;

Tipăreşte A.

Mai prezentăm încă o ultimă formulă de derivare numerică pentru prima derivată:

$$\lim_{\substack{h \to 0 \\ h > 0}} \frac{f(x_0 + h) - f(x_0 - h)}{2h} = \frac{1}{2} \lim_{h \to 0} \left[\frac{f(x_0 + h) - f(x_0)}{h} + \frac{f(x_0) - f(x_0 - h)}{h} \right] =$$

$$= \frac{1}{2} \cdot (f'(x_0) + f'(x_0)) = f'(x_0),$$

de unde pentru $h > 0$ şi $h \approx 0$ valoarea expresiei $\dfrac{f(x_0 + h) - f(x_0 - h)}{2h}$ ne dă o valoare aproximativă pentru $f'(x_0)$.

În continuare dăm o interpretare geometrică pentru cele trei formule de derivare numerică prezentate mai anterior:

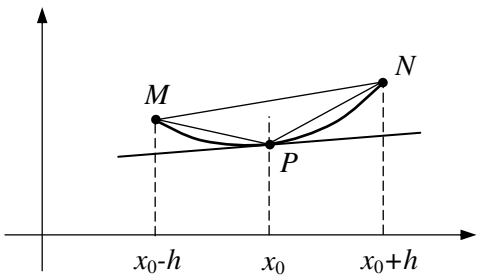


Figura 9.1:

Formula de derivare numerică $\dfrac{f(x_0 + h) - f(x_0)}{h}$ ne dă direcţia pantei PN, iar $\dfrac{f(x_0) - f(x_0 - h)}{h}$ ne dă direcţia pantei MP, şi formula $\dfrac{f(x_0 + h) - f(x_0 - h)}{2h}$ ne dă direcţia pantei MN. Intuitiv se observă că dintre cele trei pante, panta MN ar fi cea mai bună aproximare pentru direcţia tangentei în punctul P la graficul lui f.

În continuare vom demonstra şi teoretic acest fapt, arătând că evaluarea erorii la formula $\dfrac{f(x_0 + h) - f(x_0 - h)}{2h}$ depinde de h^2 şi nu de h (spunem că această formulă are ordinul de convergenţă pătratică, pe când cele două formule de derivare numerică anterioare au ordinul de convergenţă liniară, fiindcă acolo cum am văzut erorile depind de h).

Pentru asta avem nevoie de o lemă teoretică:

Lema 9.1.1. *Fie $f \in C([a,b],\mathbb{R})$ şi $\xi_1, \xi_2 \in [a,b]$ puncte arbitrare. Atunci există un $\xi \in [a,b]$ astfel ca* $\dfrac{f(\xi_1)+f(\xi_2)}{2} = f(\xi)$.

DEMONSTRAŢIE. Deoarece $f \in C([a,b],\mathbb{R})$, conform teoremei lui Weierstrass îşi atinge marginile pe $[a,b]$ deci

$$\min_{x\in[a,b]} f(x) \le \frac{f(\xi_1)+f(\xi_2)}{2} \le \max_{x\in[a,b]} f(x).$$

Pe de altă parte orice funcţie continuă are proprietatea lui Darboux, deci pentru valoarea

$$\frac{f(\xi_1)+f(\xi_2)}{2} \in \left[\min_{x\in[a,b]} f(x), \max_{x\in[a,b]} f(x)\right]$$

există un punct $\xi \in [a,b]$ astfel încât $\dfrac{f(\xi_1)+f(\xi_2)}{2} = f(\xi)$. q.e.d.

Evident această lemă are o generalizare naturală şi pentru n puncte.

Propoziţia 9.1.3. *Fie $f \in C^3([a,b],\mathbb{R})$ şi $x_0 \in (a,b)$ şi $h > 0$ astfel ca $x_0 - h$, $x_0 + h \in [a,b]$. Atunci*

$$\left|\frac{f(x_0+h)-f(x_0-h)}{2h} - f'(x_0)\right| \le \frac{M_3}{6}\cdot h^2, \quad \textit{unde}$$

$$M_3 = \sup\{|f'''(x)| \ / \ x \in [a,b]\} = \max\{|f'''(x)| \ / \ x \in [a,b]\}.$$

DEMONSTRAŢIE. Din dezvoltările tayloriene obţinem:

$$\begin{aligned} f(x_0+h) &= f(x_0) + \frac{f'(x_0)}{1!}\cdot h + \frac{f''(x_0)}{2!}\cdot h^2 + \frac{f'''(\xi_1)}{3!}\cdot h^3, \quad \text{unde} \quad \xi_1 \in (x_0, x_0+h) \\ f(x_0-h) &= f(x_0) - \frac{f'(x_0)}{1!}\cdot h + \frac{f''(x_0)}{2!}\cdot h^2 - \frac{f'''(\xi_2)}{3!}\cdot h^3, \quad \text{unde} \quad \xi_2 \in (x_0-h, x_0). \end{aligned}$$

Prin scădere se obţine:

$$f(x_0+h) - f(x_0-h) = 2f'(x_0)\cdot h + \frac{f'''(\xi_1)+f'''(\xi_2)}{6}\cdot h^3, \text{ deci}$$

$$\frac{f(x_0+h)-f(x_0-h)}{2h} - f'(x_0) = \frac{f'''(\xi_1)+f'''(\xi_2)}{12}\cdot h^2.$$

Conform lemei 9.1.1 $\dfrac{f'''(\xi_1)+f'''(\xi_2)}{2} = f'''(\xi)$, cu $\xi \in [x_0-h, x_0+h]$, deci:

$$\left|\frac{f(x_0+h)-f(x_0-h)}{2h} - f'(x_0)\right| = \left|\frac{f'''(\xi)}{6}\cdot h^2\right| \le \frac{M_3}{6}\cdot h^2.$$

Dacă vrem să determinăm teoretic valoarea lui h care trebuie aleasă astfel încât formula de derivare numerică $\dfrac{f(x_0+h)-f(x_0-h)}{2h}$ să ne dea valoarea lui $f'(x_0)$ cu o precizie $\varepsilon > 0$ este de ajuns să impunem condiţii ca $\dfrac{M_3}{6}\cdot h^2 \le \varepsilon$, adică $h \le \sqrt{\dfrac{6\varepsilon}{M_3}}$.

În continuare prezentăm un algoritm numeric cu o condiţie practică de oprire:

Algoritm 3 prima derivată

Datele de intrare: f; x_0; ε;

Fie $h := 10^{-1}$; $A := \dfrac{f(x_0+h)-f(x_0-h)}{2h}$;

Repetă $B := A$; $h := \dfrac{h}{2}$ $\left(\text{sau } h := \dfrac{h}{10}\right)$; $A := \frac{f(x_0+h)-f(x_0-h)}{2h}$;

Până când $|B - A| \geq \varepsilon$;

Tipăreşte A.

În continuare vom considera o formulă de derivare numerică pentru a doua derivată a unei funcţii într-un punct dat:

$$\frac{f(x_0-h)-2f(x_0)+f(x_0+h)}{h^2},$$

care aproximează pe $f''(x_0)$.

Într-adevăr, presupunând că $f \in C^2([a,b],\mathbb{R})$ cu $x_0 - h$, $x_0 + h \in [a,b]$ atunci din formula $\Delta^2 f(x_0-h) = h^2 \cdot f''(\xi)$, cu $\xi \in (x_0-h, x_0+h)$ obţinem că

$$\frac{f(x_0-h)-2f(x_0)+f(x_0+h)}{h^2} = f''(\xi).$$

(vezi lema 8.2.1 de la diferenţe finite pentru $n = 2$). Făcând $h \to 0$ avem că $\xi \to x_0$, deci

$$\lim_{h\to 0} \frac{f(x_0-h)-2f(x_0)+f(x_0+h)}{h^2} = \lim_{\xi\to x_0} f''(\xi) = f''(x_0).$$

În continuare vrem să studiem evaluarea restului pentru această formulă de derivare numerică arătând că restul este de ordinul al doilea, adică depinde de h^2.

Propoziţia 9.1.4. *Fie $f \in C^4([a,b],\mathbb{R})$, $x_0 \in (a,b)$ şi $h > 0$ astfel ca x_0-h, $x_0+h \in [a,b]$. Atunci*

$$\left|\frac{f(x_0-h)-2f(x_0)+f(x_0+h)}{h^2} - f''(x_0)\right| \leq \frac{M_4}{12}\cdot h^4, \quad \textit{unde}$$

$$M_4 = \sup\{|f^{(IV)}(x)| \ / \ x \in [a,b]\} = \max\{|f^{(IV)}(x) \ / \ x \in [a,b]\}.$$

Demonstraţie. Vom folosi dezvoltările tayloriene în punctul x_0 cu restul Lagrange de ordinul patru:

$$f(x_0+h) = f(x_0) + \frac{f'(x_0)}{1!}h + \frac{f''(x_0)}{2!}h^2 + \frac{f'''(x_0)}{3!}h^3 + \frac{f^{(IV)}(\xi_1)}{4!}h^4,$$

unde $\xi_1 \in (x_0, x_0 + h)$,

$$f(x_0 - h) = f(x_0) - \frac{f'(x_0)}{1!}h + \frac{f''(x_0)}{2!}h^2 - \frac{f'''(x_0)}{3!}h^3 + \frac{f^{(IV)}(\xi_2)}{4!}h^4,$$

unde $\xi_2 \in (x_0 - h, x_0)$. Prin adunarea celor două relaţii se obţine:

$$f(x_0 + h) + f(x_0 - h) = 2f(x_0) + f''(x_0) \cdot h^2 + \frac{f^{(IV)}(\xi_1) + f^{(IV)}(\xi_2)}{24}h^4,$$

deci

$$\left|\frac{f(x_0 - h) - 2f(x_0) + f(x_0 + h)}{h^2} - f''(x_0)\right| = \left|\frac{f^{(IV)}(\xi_1) + f^{(IV)}(\xi_2)}{2}\right| \cdot \frac{h^2}{12} =$$
$$= \frac{|f^{(IV)}(\xi)|}{12}h^2 \leq \frac{M_4}{12} \cdot h^2,$$

unde $\xi \in [x_0 - h, x_0 + h]$ se obţine din lema 9.1.1. q.e.d.

Dacă $\varepsilon > 0$ este precizia dată, atunci determinarea lui h se poate face din condiţia $\frac{M_4}{12} \cdot h^2 \leq \varepsilon$, adică $h \leq \sqrt{\frac{12 \cdot \varepsilon}{M_4}}$, care asigură că formula de derivare numerică

$$\frac{f(x_0 - h) - 2f(x_0) + f(x_0 + h)}{h^2}$$

aproximează pe $f''(x_0)$ cu o precizie ε.

În continuare prezentăm un algoritm cu o condiţie practică de oprire:

Algoritm a doua derivată

Datele de intrare: f; x_0; ε;

Fie $h := 10^{-1}$; $A := \dfrac{f(x_0 + h) - 2f(x_0) + f(x_0 - h)}{h^2}$;

Repetă: $B := A$; $h := \dfrac{h}{2}$ $\left(\text{sau } h := \dfrac{h}{10}\right)$; $A := \dfrac{f(x_0 + h) - 2f(x_0) + f(x_0 - h)}{h^2}$;

Până când: $|B - A| \geq \varepsilon$;

Tipăreşte A.

Menţionăm că există formule de derivare numerică şi pentru derivatele de ordin superior respectiv pentru derivatele de ordinul unu şi doi luând mai multe noduri în considerare. Totodată derivata de orice ordin a unei funcţii derivabile de o infinitate de ori se poate calcula aproximativ printr-un algoritm recursiv, folosind numai formulele de derivare din propoziţiile 9.1.1 şi 9.1.2.

9.2 Formulele de integrare numerică

9.2.1 Formula dreptunghiului

Prima dată vom demonstra o lemă tehnică:

Lema 9.2.1. *(prima teoremă a valorii medii generalizate): Fie $f, g \in C([a,b], \mathbb{R})$ două funcţii continue astfel ca $g(x) \geq 0$ pentru orice $x \in [a,b]$. Atunci există un punct $\xi \in [a,b]$ astfel încât*

$$\int_a^b f(x)g(x)dx = f(\xi) \cdot \int_a^b g(x)dx.$$

DEMONSTRAŢIE. Deoarece $f \in C([a,b], \mathbb{R})$ rezultă conform teoremei lui Weierstrass că este mărginită şi îşi atinge marginile. Fie $m = \min\limits_{x\in[a,b]} f(x)$ şi $M = \max\limits_{x\in[a,b]} f(x)$. Cum $g(x) \geq 0$ pentru orice $x \in [a,b]$ din inegalitatea $m \leq f(x) \leq M$ rezultă inegalitatea $m \cdot g(x) \leq f(x) \cdot g(x) \leq M \cdot g(x)$ pentru $x \in [a,b]$. Prin urmare

$$m \cdot \int_a^b g(x)dx \leq \int_a^b f(x)g(x)dx \leq M \cdot \int_a^b g(x)dx.$$

Dacă $g(x) = 0$ pentru orice $x \in [a,b]$ avem $\int_a^b f(x)g(x)dx = 0$ şi $\int_a^b g(x)dx = 0$, deci pentru orice punct $\xi \in [a,b]$ are loc egalitatea

$$\int_a^b f(x)g(x)dx = f(\xi) \cdot \int_a^b g(x)dx$$

în mod trivial. În caz contrar există un $x \in [a,b]$ astfel ca $g(x) > 0$ şi din cauza continuităţii lui g rezultă că $\int_a^b g(x)dx > 0$. Astfel

$$m \leq \frac{\int_a^b f(x)g(x)dx}{\int_a^b g(x)dx} \leq M.$$

Funcţia f fiind continuă are proprietatea lui Darboux, deci pentru valoarea

$$\frac{\int_a^b f(x)g(x)dx}{\int_a^b g(x)dx}$$

intermediară între m şi M va exista un $\xi \in [a,b]$ astfel încât să avem

$$f(\xi) = \frac{\int_a^b f(x)g(x)dx}{\int_a^b g(x)dx},$$

adică

$$\int_a^b f(x)g(x)dx = f(\xi) \cdot \int_a^b g(x)dx. \qquad \square$$

Caz particular: dacă $g(x) = 1$ pentru orice $x \in [a, b]$ din lema 9.2.1 obţinem

Lema 9.2.2. *(prima teoremă a valorii medii) Dacă $f \in C([a, b], \mathbb{R})$ este o funcţie continuă atunci există un punct $\xi \in [a, b]$ astfel încât*

$$\int_a^b f(x)dx = f(\xi)(b - a).$$

Prima dată vom deduce formula elementară a dreptunghiului: fie $f : [0, h] \to \mathbb{R}$ o funcţie de clasă $C^1([0, h], \mathbb{R})$, adică o dată derivabilă şi cu prima derivată continuă. Atunci avem următoarele formule elementare pentru calculul ariei folosind formule de aproximare cu ajutorul ariilor unor dreptunghiuri:

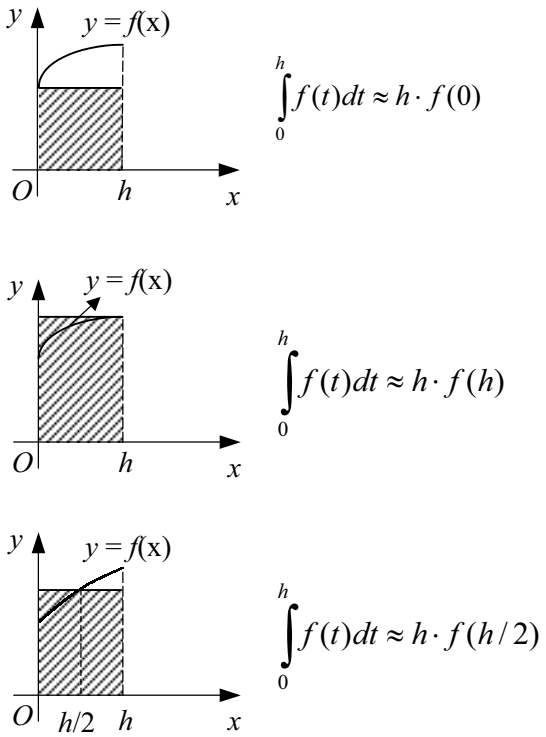


Figura 9.2:

Alegând $x = 0$ sau $x = \frac{h}{2}$ sau $x = h$ obţinem formulele de cvadratură elementare prezentate mai sus.

În continuare arătăm convergenţa acestor formule de cvadratură elementare evaluând eroarea comisă: dacă $x \in [0,h]$ atunci conform lemei 9.2.2 avem:

$$\left|\int_0^h f(t)dt - f(x)\cdot h\right| = |f(\xi)\cdot(h-0) - f(x)\cdot h| = h\cdot|f(\xi)-f(x)| = h\cdot|f'(\eta)\cdot(\xi-x)|$$

unde am aplicat teorema medie a lui Lagrange cu $\eta \in (\xi, x)$ dacă $\xi < x$ sau $\eta \in (x,\xi)$ dacă $x < \xi$. Dar $\xi, x \in [0,h]$, deci

$$\left|\int_0^h f(t)dt - f(x)h\right| = h\cdot|f'(\eta)|\cdot|\xi - x| \le h\cdot M_1\cdot h = M_1h^2,$$

unde $M_1 = \sup\{|f'(x)| \ / \ x\in[0,h]\} = \max\{|f'(x)| \ / \ x\in[0,h]\}$.

Fie acum $f \in C^1([a,b],\mathbb{R})$ şi dividem intervalul $[a,b]$ în n părţi egale cu nodurile echidistante date de formulele $x_k = x_0 + h\cdot k$, unde $x_0 = a$, $h = \frac{b-a}{n}$, iar $k = \overline{0,n}$. Atunci obţinem următoarele formule aproximative de cvadratură cunoscute sub numele de formula dreptunghiurilor:

$$\begin{aligned}
\int_a^b f(x)dx &\approx h\cdot\left(\sum_{i=0}^{n-1} f(x_i)\right) = h\cdot(f(x_0)+f(x_1)+\cdots+f(x_{n-1}))\\
\int_a^b f(x)dx &\approx h\cdot\left(\sum_{i=1}^{n} f(x_i)\right) = h\cdot(f(x_1)+f(x_2)+\cdots+f(x_n))\\
\int_a^b f(x)dx &\approx h\cdot\left(\sum_{i=0}^{n-1} f\left(\frac{x_i+x_{i+1}}{2}\right)\right) =\\
&= h\cdot\left(f\left(\frac{x_0+x_1}{2}\right)+f\left(\frac{x_1+x_2}{2}\right)+\cdots+f\left(\frac{x_{n-1}+x_n}{2}\right)\right).
\end{aligned}$$

În continuare deducem o formulă de evaluare a erorii în acest caz şi arătăm convergenţa acestor formule, în sensul că, dacă $n\to\infty$, adică numărul nodurilor echidistante mărim nelimitat, atunci formulele dreptunghiurilor corespunzătoare ne dau nişte valori din ce în ce mai apropiate de valoarea integralei definite. Într-adevăr:

$$\begin{aligned}
\left|\int_a^b f(x)dx - h\cdot\sum_{i=0}^{n-1} f(x_i)\right| &= \left|\sum_{i=0}^{n-1}\int_{x_i}^{x_{i+1}} f(x)dx - h\cdot\sum_{i=0}^{n-1} f(x_i)\right| =\\
&= \left|\sum_{i=0}^{n-1}\left(\int_{x_i}^{x_{i+1}} f(x)dx - h\cdot f(x_i)\right)\right| \le\\
&\le \sum_{i=0}^{n-1}\left|\int_{x_i}^{x_{i+1}} f(x)dx - h\cdot f(x_i)\right| \le \sum_{i=0}^{n-1} M_1\cdot h^2 =\\
&= n\cdot M_1\cdot h^2 = n\cdot M_1\cdot\left(\frac{b-a}{n}\right)^2 = \frac{1}{n}\cdot M_1\cdot(b-a)^2 =\\
&= M_1(b-a)\cdot h.
\end{aligned}$$

De aici deducem că pentru $n \to \infty$ valorile $h \cdot \sum_{i=0}^{n-1} f(x_i)$ aproximează din ce în ce mai bine $\int_a^b f(x)dx$. În mod analog:

$$\left| \int_a^b f(x)dx - h \cdot \sum_{i=1}^{n} f(x_i) \right| \leq \frac{1}{n} \cdot M_1 \cdot (b-a)^2 = M_1 \cdot (b-a) \cdot h \quad \text{şi}$$

$$\left| \int_a^b f(x)dx - h \cdot \sum_{i=0}^{n-1} f\left(\frac{x_i + x_{i+1}}{2} \right) \right| \leq \frac{1}{n} \cdot M_1 \cdot (b-a)^2 = M_1 \cdot (b-a) \cdot h.$$

Menţionăm că aici $M_1 = \sup\{|f'(x)| \ / \ x \in [a,b]\} = \max\{|f'(x)| \ / \ x \in [a,b]\}$. Dacă ne interesează să calculăm valoarea integralei definite $\int_a^b f(x)dx$ cu o precizie dată $\varepsilon > 0$, atunci este de ajuns să determinăm pe h astfel ca $M_1 \cdot (b-a) \cdot h \leq \varepsilon$, adică $h \leq \frac{\varepsilon}{M_1(b-a)}$ şi să determinăm numărul natural n care fixează numărul de noduri, adică $\frac{1}{n} \cdot M_1 \cdot (b-a)^2 \leq \varepsilon$, de unde rezultă că $n \geq \frac{M_1 \cdot (b-a)^2}{\varepsilon}$.

Totuşi noi în continuare prezentăm un alt algoritm folosind o condiţie practică de oprire:

Algoritm metoda dreptunghiului

Datele de intrare: $f; n; a; b; \varepsilon$;

Fie $h := \frac{b-a}{n}$; $A := 0$;

Pentru $i = \overline{0, n-1}$ execută

$A := A + h * f(a + i * h)$;

Repetă

$B := A$;

$n := 2 * n$;

$h := \frac{b-a}{n}$;

$A := 0$;

For $i = \overline{0, n-1}$ do

$A := A + h * f(a + i * h)$;

Până când $|B - A| \geq \varepsilon$;

Tipăreşte A.

Observăm că la formulele dreptunghiurilor cu tehnica prezentată am reuşit să arătăm că aceste formule de cvadratură au ordinul de convergenţă unu, fiindcă la evaluarea erorii h apare la prima putere. În continuare prin rafinarea unor raţionamente vom arăta că:

$$\left| \int_0^h f(t)dt - h \cdot f\left(\frac{h}{2}\right) \right| \leq \frac{M_2}{24} \cdot h^3,$$

unde presupunem că $f \in C^2([0,h];\mathbb{R})$, iar

$$M_2 = \sup\{|f''(x)| \ / \ x \in [0,h]\} = \max\{|f''(x)| \ / \ x \in [0,h]\}.$$

Într-adevăr, fie $F : [0,h] \to \mathbb{R}$, $F(x) = \int_0^x f(t)dt$. Atunci $F'(x) = f(x)$, $F''(x) = f'(x)$ şi $F'''(x) = f''(x)$. Din dezvoltările tayloriene obţinem:

$$F(h) = F\left(\frac{h}{2}\right) + \frac{F'\left(\frac{h}{2}\right)}{1!} \cdot \frac{h}{2} + \frac{F''\left(\frac{h}{2}\right)}{2!} \cdot \left(\frac{h}{2}\right)^2 + \frac{F'''(\xi_1)}{3!} \cdot \left(\frac{h}{2}\right)^3,$$

unde $\xi_1 \in \left(\frac{h}{2}, h\right)$.

$$F(0) = F\left(\frac{h}{2}\right) - \frac{F'\left(\frac{h}{2}\right)}{1!} \cdot \frac{h}{2} + \frac{F''\left(\frac{h}{2}\right)}{2!} \cdot \left(\frac{h}{2}\right)^2 - \frac{F'''(\xi_2)}{3!} \cdot \left(\frac{h}{2}\right)^3,$$

unde $\xi_2 \in \left(0, \frac{h}{2}\right)$. Prin scăderea celor două egalităţi se obţine:

$$F(h) - F(0) = F'\left(\frac{h}{2}\right) \cdot h + \frac{F'''(\xi_1) + F'''(\xi_2)}{3!} \cdot \left(\frac{h}{2}\right)^3.$$

Însă $F(h) = \int_0^h f(t)dt$, $F(0) = 0$, $F'\left(\frac{h}{2}\right) = f\left(\frac{h}{2}\right)$, $F'''(\xi_1) = f''(\xi_1)$ şi $F'''(\xi_2) = f''(\xi_2)$, deci obţinem relaţia:

$$\int_0^h f(t)dt = f\left(\frac{h}{2}\right) \cdot h + \frac{f''(\xi_1) + f''(\xi_2)}{48} \cdot h^3.$$

Folosind lema 9.1.1 deducem că există $\xi \in [0,h]$ astfel încât $\frac{f''(\xi_1) + f''(\xi_2)}{2} = f''(\xi)$. Prin urmare

$$\left| \int_0^h f(t)dt - f\left(\frac{h}{2}\right) \cdot h \right| \leq \frac{|f''(\xi)|}{24} \cdot h^3 \leq \frac{M_2}{24} \cdot h^3,$$

unde $M_2 = \sup\{|f''(x)| \ / \ x \in [0,h]\} = \max\{|f''(x)| \ / \ x \in [0,h]\}$.

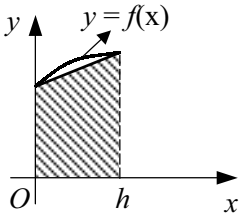
De aici pentru un interval oarecare $[a, b]$ se obţine că

$$\begin{aligned}
&\left|\int_a^b f(t)dt - h\cdot\left(\sum_{i=0}^{n-1} f\left(\frac{x_i+x_{i+1}}{2}\right)\right)\right| = \\
&= \left|\sum_{i=0}^{n-1}\int_{x_i}^{x_{i+1}} f(t)dt - h\cdot\left(\sum_{i=0}^{n-1} f\left(\frac{x_i+x_{i+1}}{2}\right)\right)\right| = \\
&= \left|\sum_{i=0}^{n-1}\left(\int_{x_i}^{x_{i+1}} f(t)dt - h\cdot f\left(\frac{x_i+x_{i+1}}{2}\right)\right)\right| \leq \\
&\leq \sum_{i=0}^{n-1}\left|\int_{x_i}^{x_i+1} f(t)dt - h\cdot f\left(\frac{x_i+x_{i+1}}{2}\right)\right| \leq \sum_{i=0}^{n-1}\frac{M_2}{24}\cdot h^3 = n\cdot\frac{M_2}{24}\cdot h^3 = \\
&= \frac{1}{n^2}\cdot\frac{M_2}{24}\cdot(b-a)^3 = \frac{M_2}{24}\cdot(b-a)\cdot h^2,
\end{aligned}$$

unde presupunem că $f \in C^2([a, b], \mathbb{R})$ iar $M_2 = \sup\{|f''(x)| \ / \ x \in [a, b]\} = \max\{|f''(x)| \ / \ x \in [a, b]\}$, deci avem o convergenţă de ordinul doi.

9.2.2 Formula trapezului

Prima dată vom deduce formula elementară a trapezului: fie $f \in C^2([0, h], \mathbb{R})$, adică o funcţie de două ori derivabilă şi cu derivata a doua continuă. Atunci avem următoarea formulă elementară pentru calculul ariei folosind formula de aproximare cu ajutorul ariei unui trapez:



$$\int_0^h f(t)dt \approx h\cdot\frac{f(0)+f(h)}{2}$$

Figura 9.3:

În continuare arătăm convergenţa acestei formule de cvadratură elementară evaluând eroarea comisă. Fie $F : [0, h] \to \mathbb{R}$, $F(x) = \int_0^x f(t)dt$, $F(0) = 0$, $F(h) = \int_0^h f(t)dt$, $F'(x) = f(x)$, $F''(x) = f'(x)$, $F'''(x) = f''(x)$. Considerăm dezvoltările tayloriene cu

restul sub forma integrală:

$$\begin{aligned} F(h) &= F(0) + \frac{F'(0)}{1!}h + \frac{F''(0)}{2!}h^2 + \frac{1}{2!}\cdot\int_0^h (h-t)^2\cdot F'''(t)dt \quad \text{şi} \\ f(h) &= f(0) + \frac{f'(0)}{1!}\cdot h + \int_0^h (h-t)\cdot f''(t)dt. \end{aligned}$$

În prima egalitate trecem de la F la f, iar a doua egalitate înmulţim cu $\dfrac{h}{2}$:

$$\int_0^h f(t)dt = 0 + \frac{f(0)}{1}\cdot h + \frac{f'(0)}{2}\cdot h^2 + \frac{1}{2}\cdot\int_0^h (h-t)^2\cdot f''(t)dt,$$
$$f(h)\cdot\frac{h}{2} = f(0)\cdot\frac{h}{2} + \frac{f'(0)}{1}\cdot\frac{h^2}{2} + \frac{h}{2}\cdot\int_0^h (h-t)\cdot f''(t)dt$$

şi din prima egalitate scădem pe cea de a doua:

$$\int_0^h f(t)dt - f(h)\cdot\frac{h}{2} = f(0)\cdot\frac{h}{2} + \frac{1}{2}\cdot\left[\int_0^h (h-t)^2 f''(t)dt - \int_0^h h(h-t)f''(t)dt\right],$$

adică

$$\int_0^h f(t)dt = \frac{f(0)+f(h)}{2}\cdot h + \frac{1}{2}\cdot\int_0^h [(h-t)^2 - (h^2-ht)]\cdot f''(t)dt, \text{ deci}$$
$$\int_0^h f(t)dt = \frac{f(0)+f(h)}{2}\cdot h + \frac{1}{2}\cdot\int_0^h (-ht+t^2)\cdot f''(t)dt.$$

Prin urmare:

$$\left|\int_0^h f(t)dt - \frac{f(0)+f(h)}{2}\cdot h\right| = \left|\frac{1}{2}\cdot\int_0^h (-ht+t^2)\cdot f''(t)dt\right| \le$$
$$\le \frac{1}{2}\cdot\int_0^h |-ht+t^2|\cdot|f''(t)|dt \le \frac{1}{2}\int_0^h (ht-t^2)\cdot M_2 dt =$$
$$= \frac{M_2}{2}\cdot\left(h\cdot\frac{t^2}{2}\Big|_0^h - \frac{t^3}{3}\Big|_0^h\right) = \frac{M_2}{2}\cdot\left(\frac{h^3}{2} - \frac{h^3}{3}\right) = \frac{M_2}{12}\cdot h^3,$$

unde $M_2 = \sup\{|f''(x)| \ / \ x\in[0,h]\} = \max\{|f''(x)| \ / \ x\in[0,h]\}$.

Fie acum $f\in C^2([a,b],\mathbb{R})$ şi împărţim intervalul $[a,b]$ în n părţi egale cu nodurile echidistante date de formulele $x_k = x_0 + h\cdot k$, unde $x_0 = a$, $h = \frac{b-a}{n}$ iar $k = \overline{0,n}$. Atunci din formula elementară a trapezului obţinem următoarea formulă de cvadratură cunoscută sub numele de formula trapezului:

$$\begin{aligned} \int_a^b f(x)dx &\approx \frac{h}{2}(f(x_0) + 2\cdot f(x_1) + \cdots + 2\cdot f(x_{n-1}) + f(x_n)) = \\ &= h\cdot\left(\frac{f(x_0)}{2} + f(x_1) + \cdots + f(x_{n-1}) + \frac{f(x_n)}{2}\right). \end{aligned}$$

În continuare deducem o formulă de evaluare a erorii în acest caz şi arătăm convergenţa formulei trapezului, în sensul că, dacă $n \to \infty$, adică numărul nodurilor echidistante mărim nelimitat, atunci formulele trapezelor corespunzătoare ne dau nişte valori din ce în ce mai apropiate de valoarea integralei definite. Într-adevăr:

$$\begin{aligned}
& \left| \int_a^b f(x)dx - \frac{h}{2} \cdot \left(f(x_0) + 2 \cdot \sum_{i=1}^{n-1} f(x_i) + f(x_n) \right) \right| = \\
= & \left| \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} f(x)dx - \frac{h}{2} \left(f(x_0) + 2 \cdot \sum_{i=1}^{n-1} f(x_i) + f(x_n) \right) \right| = \\
= & \left| \sum_{i=0}^{n-1} \left(\int_{x_i}^{x_{i+1}} f(x)dx - \frac{h}{2} \cdot (f(x_i) + f(x_{i+1})) \right) \right| \leq \\
\leq & \sum_{i=0}^{n-1} \left| \int_{x_i}^{x_{i+1}} f(x)dx - \frac{h}{2} \cdot (f(x_i) + f(x_{i+1})) \right| \leq \\
\leq & \sum_{i=0}^{n-1} \frac{M_2}{12} \cdot h^3 = n \cdot \frac{M_2}{12} \cdot h^3 = \frac{1}{n^2} \cdot \frac{M_2}{12} \cdot (b-a)^3 = \frac{M_2}{12} \cdot (b-a) \cdot h^2,
\end{aligned}$$

unde $M_2 = \sup\{|f''(x)| \ / \ x \in [a,b]\} = \max\{|f''(x)| \ / \ x \in [a,b]\}$. Prin urmare formula trapezului are ordinul de convergenţă pătratică.

Dacă vrem să determinăm valoarea integralei definite $\int_a^b f(x)dx$ cu o precizie $\varepsilon > 0$ atunci impunem condiţia ca $\frac{1}{n^2} \cdot \frac{M_2}{12}(b-a)^3 \leq \varepsilon$, adică $n \geq \sqrt{\frac{M_2}{12} \cdot \frac{(b-a)^3}{\varepsilon}}$. Prin urmare avem numărul de noduri necesare în formula trapezului, care este garanţia teoretică pentru precizia ε, conform teoriei prezentate mai sus.

Noi în continuare prezentăm un algoritm pentru metoda trapezului cu o condiţie practică de oprire:

Algoritm metoda trapezului

Datele de intrare: $f; n; a; b; \varepsilon$;

Fie $h := \dfrac{b-a}{n}$; $A := 0$; $A := A + \dfrac{h}{2} * f(a)$;

Pentru $i = \overline{1, n-1}$ execută $A := A + h * f(a + i * h)$;

$A := A + \dfrac{h}{2} * f(b)$;

Repetă $\quad B := A$; $n := 2 * n$; $h := \dfrac{b-a}{n}$;

$A := 0$; $A := A + \dfrac{h}{2} * f(a)$;

Pentru $i = \overline{1, n-1}$ execută

$A := A + h * f(a + i * h)$;

$A := A + \dfrac{h}{2} * f(b)$;

Până când $|B - A| \geq \varepsilon$;

Tipăreşte A.

9.2.3 Formula lui Simpson

Prima dată vom deduce formula lui Simpson pentru o arie elementară. Presupunem că $f \in C^4([-h, h], \mathbb{R})$ şi scopul nostru este de a calcula aproximativ integrala $\int_{-h}^{h} f(x)dx$. Ideea lui Simpson constă în următoarele: aria figurii determinată de graficele $x = -h$, $x = h$, $y = 0$ şi $y = f(x)$ se aproximează printr-o nouă arie, unde în locul curbei $y = f(x)$ se consideră un arc de parabolă, care să treacă prin punctele $A(-h, f(-h))$, $B(0, f(0))$ şi $C(h, f(h))$. Fie ecuaţia parabolei $y(x) = ax^2 + bx + c$ şi impunem condiţiile ca această funcţie de gradul al doilea să treacă prin punctele A, B şi C :

$$\begin{cases} a \cdot h^2 - b \cdot h + c = f(-h) \\ c = f(0) \\ a \cdot h^2 + b \cdot h + c = f(h) \end{cases}$$

Prin urmare $c = f(0)$ şi

$$\begin{cases} a \cdot h^2 - b \cdot h = f(-h) - f(0) \\ a \cdot h^2 + b \cdot h = f(h) - f(0) \end{cases}.$$

Prin adunarea şi scăderea celor două relaţii obţinem:

$$a = \frac{f(-h) - 2 \cdot f(0) + f(h)}{2h^2} \quad \text{şi} \quad b = \frac{f(h) - f(-h)}{2h}.$$

Aşadar

$$y(x) = \frac{f(-h) - 2 \cdot f(0) + f(h)}{2h^2} x^2 + \frac{f(h) - f(-h)}{2h} \cdot x + f(0).$$

Ne rămâne să calculăm:

$$\begin{aligned}\int_{-h}^{h} y(x)dx &= \int_{-h}^{h} \frac{f(-h)-2f(0)+f(h)}{2h^2}x^2dx + \int_{-h}^{h} \frac{f(h)-f(-h)}{2h}xdx + \int_{-h}^{h} f(0)dx = \\ &= \frac{f(-h)-2f(0)+f(h)}{2h^2}\cdot\frac{x^3}{3}\Big|_{-h}^{h} + \frac{f(h)-f(-h)}{2h}\cdot\frac{x^2}{2}\Big|_{-h}^{h} + f(0)\cdot x\Big|_{-h}^{h} = \\ &= \frac{f(-h)-2\cdot f(0)+f(h)}{2h^2}\cdot\left(\frac{h^3}{3}+\frac{h^3}{3}\right) + \frac{f(h)-f(-h)}{2h}\cdot\left(\frac{h^2}{2}-\frac{h^2}{2}\right) + \\ &\quad + f(0)\cdot(h+h) = (f(-h)-2f(0)+f(h))\cdot\frac{h}{3} + 2f(0)\cdot h = \\ &= \frac{h}{3}(f(-h)+4\cdot f(0)+f(h)).\end{aligned}$$

Prin urmare avem următoarea formulă aproximativă pentru calculul ariei elementare cunoscută sub numele de formula lui Simpson:

$$\int_{-h}^{h} f(x)dx \approx \frac{h}{3}(f(-h)+4\cdot f(0)+f(h)).$$

În continuare arătăm convergenţa acestei formule de cvadratură elementară evaluând eroarea comisă. Fie deci

$$\begin{aligned}&f\in C^4([-h,h],\mathbb{R}),\ F:[-h,h]\to\mathbb{R},\ F(x)=\int_0^x f(t)dt,\\ &F(0)=0,\ F(h)=\int_0^h f(t)dt,\ F(-h)=\int_0^{-h} f(t)dt=-\int_{-h}^0 f(t)dt,\\ &F'(x)=f(x),\ F''(x)=f'(x),\ F'''(x)=f''(x),\ F^{IV}(x)=f'''(x),\ F^V(x)=f^{IV}(x).\end{aligned}$$

Considerăm dezvoltările tayloriene cu restul sub forma integrală:

$$\begin{aligned}F(h) &= F(0)+\frac{F'(0)}{1!}h+\frac{F''(0)}{2!}h^2+\frac{F'''(0)}{3!}h^3+\frac{F^{IV}(0)}{4!}h^4+\\ &\quad +\frac{1}{4!}\cdot\int_0^h (h-t)^4\cdot F^{(V)}(t)dt;\\ F(-h) &= F(0)-\frac{F'(0)}{1!}h+\frac{F''(0)}{2!}h^2-\frac{F'''(0)}{3!}h^3+\frac{F^{IV}(0)}{4!}h^4-\\ &\quad -\frac{1}{4!}\cdot\int_0^h (h-t)^4\cdot F^{(V)}(-t)dt;\\ f(h) &= f(0)+\frac{f'(0)}{1!}h+\frac{f''(0)}{2!}h^2+\frac{f'''(0)}{3!}\cdot h^3+\frac{1}{3!}\cdot\int_0^h (h-t)^3\dot{f}^{IV}(t)dt; \quad (9.1)\\ f(-h) &= f(0)-\frac{f'(0)}{1!}h+\frac{f''(0)}{2!}h^2-\frac{f'''(0)}{3!}\cdot h^3+\frac{1}{3!}\cdot\int_0^h (h-t)^3\dot{f}^{IV}(-t)dt. \quad (9.2)\end{aligned}$$

Prin urmare primele două egalităţi au forma:

$$\begin{aligned}\int_0^h f(t)dt &= 0+\frac{f(0)}{1!}h+\frac{f'(0)}{2!}h^2+\frac{f''(0)}{3!}h^3+\frac{f'''(0)}{4!}h^4+\frac{1}{4!}\cdot\int_0^h (h-t)^4\cdot f^{IV}(t)dt;\\ \int_0^{-h} f(t)dt &= 0-\frac{f(0)}{1!}h+\frac{f'(0)}{2!}h^2-\frac{f''(0)}{3!}h^3+\frac{f'''(0)}{4!}h^4-\frac{1}{4!}\cdot\int_0^h (h-t)^4\cdot f^{IV}(-t)dt;\end{aligned}$$

Dacă le scădem se obţine:

$$\begin{aligned}\int_{-h}^h f(t)dt &= \int_0^h f(t)dt+\int_{-h}^0 f(t)dt=\int_0^h f(t)dt-\int_0^{-h} f(t)dt= \qquad (9.3)\\ &= 2\cdot f(0)\cdot h+\frac{1}{3}\cdot f''(0)\cdot h^3+\frac{1}{4!}\cdot\int_0^h (h-t)^4\cdot(f^{IV}(t)+f^{IV}(-t))dt\end{aligned}$$

În continuare adunăm egalităţile (9.1) şi (9.2):

$$f(h)+f(-h)=2\cdot f(0)+f''(0)\cdot h^2+\frac{1}{3!}\cdot\int_0^h (h-t)^3(f^{IV}(t)+f^{IV}(-t))dt.$$

Această ultimă egalitate înmulţim cu $\frac{h}{3}$:

$$\frac{h}{3}(f(h)+f(-h))=f(0)\cdot\frac{2}{3}h+f''(0)\cdot\frac{h^3}{3}+\frac{h}{18}\cdot\int_0^h (h-t)^3(f^{IV}(t)+f^{IV}(-t))dt$$

şi această ultimă egalitate scădem din egalitatea (9.3):

$$\begin{aligned}&\int_{-h}^h f(t)dt-\frac{h}{3}(f(h)+f(-h))=\\ &=4\cdot f(0)\cdot\frac{h}{3}+\frac{1}{4!}\cdot\int_0^h (h-t)^3\cdot\left(h-t-\frac{4}{3}h\right)\cdot\left(f^{IV}(t)+f^{IV}(-t)\right)dt,\end{aligned}$$

deci

$$\begin{aligned}&\int_{-h}^h f(t)dt-\frac{h}{3}(f(h)+4\cdot f(0)+f(-h))=\\ &=-\frac{1}{24}\cdot\int_0^h (h-t)^3\cdot\left(\frac{h}{3}+t\right)\cdot\left(f^{IV}(t)+f^{IV}(-t)\right)dt.\end{aligned}$$

În continuare aplicăm lema 9.2.1 pentru integrala din membrul drept, alegând

$$g(x)=(h-x)^3\cdot\left(\frac{h}{3}+x\right)\geq 0$$

pentru orice $x\in[0,h]$. Prin urmare există un $\xi\in[0,h]$ astfel încât:

$$\int_{-h}^h f(t)dt-\frac{h}{3}(f(h)+4\cdot f(0)+f(-h))=-\frac{1}{24}\cdot\left(f^{IV}(\xi)+f^{IV}(-\xi)\right)\cdot\int_0^h (h-t)^3\cdot\left(\frac{h}{3}+t\right)dt.$$

Dar conform lemei 9.1.1

$$\frac{f^{IV}(\xi)+f^{IV}(-\xi)}{2}=f^{IV}(\eta)$$

unde $\eta\in[-h,h]$. Prin urmare

$$\begin{aligned}
&\int_{-h}^{h}f(t)dt-\frac{h}{3}(f(h)+4\cdot f(0)+f(-h))=-\frac{1}{12}\cdot f^{IV}(\eta)\cdot\int_0^h(h-t)^3\cdot\left(\frac{h}{3}+t\right)dt=\\
=\;&-\frac{1}{12}\cdot f^{IV}(\eta)\cdot\left[\frac{h}{3}\cdot\int_0^h(h-t)^3dt+\int_0^h(h-t)^3\cdot tdt\right]=\\
=\;&-\frac{1}{12}\cdot f^{IV}(\eta)\cdot\left[\frac{h}{3}\cdot\int_0^h(h^3-3h^2t+3ht^2-t^3)dt+\int_0^h(h^3t-3h^2t^2+3ht^3-t^4)dt\right]=\\
=\;&-\frac{1}{12}\cdot f^{IV}(\eta)\cdot\left[\frac{h}{3}\cdot\left(h^3\cdot t\Big|_0^h-3h^2\cdot\frac{t^2}{2}\Big|_0^h+3h\cdot\frac{t^3}{3}\Big|_0^h-\frac{t^4}{4}\Big|_0^h\right)+\right.\\
&\left.+\;\left(h^3\cdot\frac{t^2}{2}\Big|_0^h-3h^2\cdot\frac{t^3}{3}\Big|_0^h+3h\cdot\frac{t^4}{4}\Big|_0^h-\frac{t^5}{5}\Big|_0^h\right)\right]=\\
=\;&-\frac{1}{12}\cdot f^{IV}(\eta)\cdot\left[\frac{h}{3}\left(h^4-\frac{3}{2}h^4+h^4-\frac{h^4}{4}\right)+\left(\frac{h^5}{2}-h^5+\frac{3}{4}h^5-\frac{h^5}{5}\right)\right]=\\
=\;&-\frac{1}{90}f^{IV}(\eta)\cdot h^5,\ \text{deci}
\end{aligned}$$

$$\left|\int_{-h}^{h}f(t)dt-\frac{h}{3}(f(h)+4\cdot f(0)+f(-h))\right|=\frac{1}{90}\cdot|f^{IV}(\eta)|\cdot h^5\le\frac{1}{90}\cdot M_4\cdot h^5,$$

unde $M_4=\sup\{|f^{IV}(x)|\ /\ x\in[-h,h]\}=\max\{|f^{IV}(x)|\ /\ x\in[-h,h]\}$.

În continuare scriem formula lui Simpson în cazul unei funcţii $f\in C^4([a,b],\mathbb{R})$ unde se face diviziunea intervalului $[a,b]$ în $2n$ subintervale considerând pasul $h=\dfrac{b-a}{2n}$ şi nodurile $x_k=x_0+h\cdot k$ cu $x_0=a$, unde $k=\overline{0,2n}$. Vom aplica formula lui Simpson pentru fiecare arie elementară limitată de: $x=x_{2k}$; $x=x_{2k+2}$; $y=0$; $y=f(x)$; unde $k=\overline{0,n-1}$. Astfel se obţine formula aproximativă de calcul al integralei definite cunoscută sub numele de formula lui Simpson:

$$\begin{aligned}
\int_a^b f(x)dx\;&\approx\;\frac{h}{3}(f(x_0)+4f(x_1)+2f(x_2)+4f(x_3)+\cdots+4\cdot f(x_{2n-1})+f(x_{2n}))=\\
&=\;\frac{h}{3}\left(f(x_0)+4\cdot\sum_{i=1}^{n}f(x_{2i-1})+2\cdot\sum_{i=1}^{n-1}f(x_{2i})+f(x_{2n})\right).
\end{aligned}$$

În continuare studiem restul care se obţine prin înlocuirea integralei definite cu formula

lui Simpson:

$$\begin{aligned}
&\left|\int_a^b f(x)dx - \frac{h}{3}\cdot(f(x_0)+4f(x_1)+2f(x_2)+4f(x_3)+\cdots+4f(x_{2n-1})+f(x_{2n}))\right| = \\
=\ &\left|\sum_{i=0}^{n-1}\int_{x_{2i}}^{x_{2i+2}} f(x)dx - \sum_{i=0}^{n-1}\frac{h}{3}\cdot(f(x_{2i})+4f(x_{2i+1})+f(x_{2i+2}))\right| = \\
=\ &\left|\sum_{i=0}^{n-1}\left(\int_{x_{2i}}^{x_{2i+2}} f(x)dx - \frac{h}{3}\cdot(f(x_{2i})+4f(x_{2i+1})+f(x_{2i+2}))\right)\right| \leq \\
\leq\ &\sum_{i=0}^{n-1}\left|\int_{x_{2i}}^{x_{2i+2}} f(x)dx - \frac{h}{3}\cdot(f(x_{2i})+4f(x_{2i+1})+f(x_{2i+2}))\right| \leq \\
\leq\ &\sum_{i=0}^{n-1}\frac{1}{90}\cdot M_4\cdot h^5 = n\cdot\frac{1}{90}\cdot M_4\cdot h^5 = \frac{1}{90}\cdot M_4\cdot(b-a)\cdot h^4 = \\
=\ &\frac{1}{n^4}\cdot\frac{1}{90}\cdot M_4\cdot(b-a)^5, \text{ unde} \\
&M_4 = \sup\{|f^{IV}(x)| \ / \ x\in[a,b]\} = \max\{|f^{IV}(x)| \ / \ x\in[a,b]\}.
\end{aligned}$$

Observăm că metoda lui Simpson are ordinul de convergenţă patru, fiindcă în evaluarea restului apare h^4. Totodată menţionăm că formula lui Simpson este formula cu cea mai rapidă convergenţă dintre cele trei formule de cvadratură prezentate şi în acelaşi timp cu cel mai mic număr de operaţii necesare pentru a obţine o precizie $\varepsilon > 0$ dată. Sigur că, dacă ne interesează valoarea integralei $\int_a^b f(x)dx$ cu o precizie $\varepsilon > 0$ dinainte dată, atunci folosind teoria anterioară numărul n se obţine din condiţia: $\frac{1}{n^4}\cdot\frac{1}{90}\cdot M_4\cdot(b-a)^5 \leq \varepsilon$, adică $n \geq \sqrt[4]{\frac{M_4}{90}\cdot\frac{(b-a)^5}{\varepsilon}}$.

În continuare dăm un algoritm de calcul al integralei definite folosind o condiţie practică de oprire.

Algoritmul metoda Simpson

Datele de intrare: $f; n; a; b; \varepsilon$;

Fie $h = \frac{b-a}{2n}$; $A := 0$; $A := A + \frac{h}{3} * f(a)$;

Pentru $i = \overline{1,n}$ execută $A := A + \frac{4}{3} * h * f(a + (2i-1) * h)$;

Pentru $i = \overline{1,n-1}$ execută $A := A + \frac{2}{3} * h * f(a + 2 * i * h)$;

$A := A + \frac{h}{3} * f(a + 2 * n * h)$;

Repetă $B := A$; $n := 2 * n$; $h = \dfrac{b-a}{2n}$;

$A := 0$; $A := A + \dfrac{h}{3} * f(a)$;

Pentru $i = \overline{1, n}$ execută $A := A + \dfrac{4}{3} * h * f(a + (2i-1) * h)$;

Pentru $i = \overline{1, n-1}$ execută $A := A + \dfrac{2}{3} * h * f(a + 2 * i * h)$;

$A := A + \dfrac{h}{3} * f(a + 2 * n * h)$;

Până când $|B - A| \geq \varepsilon$;

Tipăreşte A.

Capitolul 10

Metode numerice pentru calculul valorilor şi vectorilor proprii ale unei matrici

10.1 Metoda lui Krylov

Fie $A = (a_{ij})_{i,j=\overline{1,n}}$ o matrice pătratică de ordinul n, şi $x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \neq \theta_{\mathbb{R}^n}$ (diferit de vectorul nul). Fie $y = A \cdot x$. Dacă componentele vectorului y sunt proporţionale cu componentele vectorului x, adică există $\lambda \in \mathbb{R}$ ($\lambda \in \mathbb{C}$), astfel încât $y = \lambda x$, (adică pentru orice $i = \overline{1,n}$ avem $y[i] = \lambda \cdot x[i]$), atunci vom spune că x este un vector propriu pentru matricea A iar λ este o valoare proprie pentru matricea A. Prin urmare din $y = Ax$ şi $y = \lambda x$ obţinem $Ax = \lambda x$.

Exemplul 10.1.1. *Fie* $A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 6 & -2 \\ 3 & 4 & -1 \end{bmatrix}$, $x = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$. *Atunci*

$$y = A \cdot x = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 6 & -2 \\ 3 & 4 & -1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 6 \\ 6 \\ 6 \end{bmatrix} = 6 \cdot \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

Deci $\lambda = 6$ *este o valoare proprie pentru matricea* A, *iar* $x = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$ *este un vector propriu pentru matricea* A, *care corespune valorii proprii* $\lambda = 6$.

Din $Ax = \lambda x$ *rezultă* $(A - \lambda I_n) \cdot x = \theta_{\mathbb{R}^n}$, *unde* I_n *este matricea unitate de ordinul* n. *Prin urmare:*

$$\begin{pmatrix} a_{11} - \lambda & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} - \lambda & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn-\lambda} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Acest sistem liniar admite o soluţie x *nebanală, dacă şi numai dacă determinantul sistemului este egal cu zero:* $\det(A - \lambda \cdot I_n) = 0$, *adică*

$$\begin{vmatrix} a_{11} - \lambda & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} - \lambda & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn-\lambda} \end{vmatrix} = 0.$$

Dacă dezvoltăm acest determinant, atunci obţinem în variabila λ *un polinom de gradul* n :

$$\det(A - \lambda I_n) = (-1)^n \cdot [\lambda^n - p_1\lambda^{n-1} + p_2\lambda^{n-2} - \dots + (-1)^n p_n] = 0.$$

Rădăcinile λ_i, $i = \overline{1,n}$, *ale acestei ecuaţii polinomiale sunt valorile proprii ale matricii* A, *iar pentru fiecare* λ_i, $i = \overline{1,n}$ *valoare proprie fixată rezolvăm sistemul liniar şi omogen* $(A - \lambda_i I_n) \cdot x = \theta_{\mathbb{R}^n}$ *şi vom obţine vectorii proprii corespunzători valorii proprii* λ_i.

Metoda dezvoltării directe

Fie de exemplu $n = 2$ iar $A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$. Avem $\det(A - \lambda I_2) = 0 \Leftrightarrow \begin{vmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{vmatrix} = 0$ $\Leftrightarrow \lambda^2 - 4\lambda + 3 = 0 \Leftrightarrow \lambda_1 = 1,\ \lambda_2 = 3$.

Pentru $\lambda_1 = 1$ avem $(A - \lambda_1 I_2) \cdot x = \theta_{\mathbb{R}^2} \Leftrightarrow \begin{bmatrix} 2-1 & 1 \\ 1 & 2-1 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$.

Deci $x_1 = -x_2$, adică $\begin{cases} x_2 = c \\ x_1 = -c \end{cases}$ deci $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = +c \cdot \begin{pmatrix} -1 \\ 1 \end{pmatrix}$, cu vectorul propriu

$x = \begin{pmatrix} -1 \\ +1 \end{pmatrix}$. Pentru $\lambda_2 = 3$ avem $(A - \lambda_2 I_2) \cdot x = \theta_{\mathbb{R}^2}$, adică $x_1 = x_2$, adică $x_1 = x_2 = c$. Deci $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = c \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, cu vectorul propriu $x = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$.

Pentru $n = 3$ vom face în general: $A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$,

$$\begin{aligned} \det(A - \lambda I_3) &= \begin{vmatrix} a_{11} - \lambda & a_{12} & a_{13} \\ a_{21} & a_{22} - \lambda & a_{23} \\ a_{31} & a_{32} & a_{33-\lambda} \end{vmatrix} = \\ &= (a_{11} - \lambda)(a_{22} - \lambda)(a_{33} - \lambda) + a_{12} \cdot a_{23} \cdot a_{31} + a_{13} \cdot a_{32} \cdot a_{21} - \\ &\quad - a_{13} \cdot a_{31}(a_{22} - \lambda) - a_{23} \cdot a_{32} \cdot (a_{11} - \lambda) - a_{12} \cdot a_{21} \cdot (a_{33} - \lambda) = \\ &= (-1)^3 \cdot \left[\lambda^3 - \lambda^2(a_{11} + a_{22} + a_{33}) + \lambda \cdot \left(\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} + \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} \right] + \right. \\ &\quad \left. + \begin{vmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{vmatrix} \right) - \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} \Bigg], \end{aligned}$$

adică

$$\det(A - \lambda I_3) = (-1)^3 \cdot [\lambda^3 - p_1 \cdot \lambda^2 + p_2 \cdot \lambda - p_3] = 0,$$

unde $p_1 = Tr(A) = a_{11} + a_{22} + a_{33}$ (se numeşte urma matricii)

$$\begin{aligned} p_2 &= \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} + \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} + \begin{vmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{vmatrix} \\ p_3 &= \det(A) = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{22} & a_{23} \end{vmatrix}. \end{aligned} \tag{10.1}$$

Pentru n arbitrar se obţine un polinom de gradul n care se rezolvă numeric, de exemplu prin metoda lui Bairstow (vezi paragraful 5.1):

$$D(\lambda) = \det(A - \lambda I_n) = (-1)^n \cdot [\lambda^n - p_1 \lambda^{n-1} + p_2 \cdot \lambda^{n-2} - \cdots + (-1)^n \cdot p_n] = 0.$$

Reamintim o teoremă cunoscută din algebra lineară:

Teorema 10.1.1. *(Cayley-Hamilton)*

Matricea A verifică ecuaţia sa caracteristică, adică dacă

$$D(\lambda) = \det(A - \lambda I_n) = (-1)^n(\lambda^n + p_1\lambda^{n-1} + p_2\lambda^{n-2} + \cdots + p_n) = 0$$

este ecuaţia caracteristică a matricii A, atunci:

$$A^n + p_1A^{n-1} + p_2A^{n-2} + \cdots + p_nI_n = O_n,$$

unde O_n este matricea nulă de ordinul n.

Alegem un $y^{(0)} = \begin{bmatrix} y_1^{(0)} \\ y_2^{(0)} \\ \vdots \\ y_n^{(0)} \end{bmatrix} \neq \theta_{\mathbb{R}^n}$ şi înmulţim ecuaţia anterioară din dreapta cu $y^{(0)}$:

$$A^n y^{(0)} + p_1A^{n-1}y^{(0)} + p_2A^{n-2}y^{(0)} + \cdots + p_n y^{(0)} = \theta_{\mathbb{R}^n}.$$

Fie $y^{(k)} = Ay^{(k-1)}$ pentru $k = 1, 2, \ldots, n$, adică

$$y^{(1)} = Ay^{(0)}, y^{(2)} = Ay^{(1)} = A \cdot Ay^{(0)} = A^2y^{(0)}, \ldots, y^{(n)} = A \cdot y^{(n-1)} = A^n y^{(0)}.$$

Astfel obţinem relaţia:

$$y^{(n)} + p_1y^{(n-1)} + p_2y^{(n-2)} + \cdots + p_n y^{(0)} = \theta_{\mathbb{R}^n}, \text{adică}$$

$$p_1y^{(n-1)} + p_2y^{(n-2)} + \cdots + p_n y^{(0)} = -y^{(n)}, \text{adică}$$

$$p_1 \cdot \begin{pmatrix} y_1^{(n-1)} \\ y_2^{(n-1)} \\ \vdots \\ y_n^{(n-1)} \end{pmatrix} + p_2 \cdot \begin{pmatrix} y_1^{(n-2)} \\ y_2^{(n-2)} \\ \vdots \\ y_n^{(n-2)} \end{pmatrix} + \cdots + p_n \cdot \begin{pmatrix} y_1^{(0)} \\ y_2^{(0)} \\ \vdots \\ y_n^{(0)} \end{pmatrix} = - \begin{pmatrix} y_1^{(n)} \\ y_2^{(n)} \\ \vdots \\ y_n^{(n)} \end{pmatrix}, \text{adică}$$

$$\begin{cases} p_1y_1^{(n-1)} + p_2y_1^{(n-2)} + \cdots + p_n y_1^{(0)} = -y_1^{(n)} \\ p_1y_2^{(n-1)} + p_2y_2^{(n-2)} + \cdots + p_n y_2^{(0)} = -y_2^{(n)} \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ p_1y_n^{(n-1)} + p_2y_n^{(n-2)} + \cdots + p_n y_n^{(0)} = -y_n^{(n)} \end{cases}, \text{adică}$$

$$\begin{bmatrix} y_1^{(n-1)} & y_1^{(n-2)} & \ldots & y_1^{(0)} \\ y_2^{(n-1)} & y_2^{(n-2)} & \ldots & y_2^{(0)} \\ \ldots & \ldots & \ldots & \ldots \\ y_n^{(n-1)} & y_n^{(n-2)} & \ldots & y_n^{(0)} \end{bmatrix} \cdot \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{bmatrix} = - \begin{bmatrix} y_1^{(n)} \\ y_2^{(n)} \\ \vdots \\ y_n^{(n)} \end{bmatrix}.$$

Acest sistem liniar rezolvăm cu metoda lui Gauss, şi se obţin coeficienţii ecuaţiei caracteristice. Dacă cumva la metoda lui Gauss la un anumit pas elementul pivot devine zero, atunci alegem un alt vector nenul de pornire $y^{(0)} \in \mathbb{R}^n$.

Program 1 KRYLOV

Datele de intrare: n, A;
Pentru $i = \overline{1,n}$ execută
citeşte $Y[i][n]$;
Pentru $j = \overline{n,2}$ execută
Pentru $i = \overline{1,n}$ execută
$Y[i][j-1] := \sum_{k=1}^{n} A[i][k] * Y[k][j]$;
Pentru $i = \overline{1,n}$ execută
$b[i] := -\sum_{k=1}^{n} A[i][k] * Y[k][1]$;
Rezolvă sistemul liniar $Y \cdot p = b$ cu metoda lui Gauss.
Tipăreşte p ($p[i]$ pentru $i = \overline{1,n}$)

Exemplul 10.1.2. *Fie* $n = 4$; $A = \begin{bmatrix} -4 & -3 & 1 & 1 \\ 2 & 0 & 4 & -1 \\ 1 & 1 & 2 & -2 \\ 1 & 1 & -1 & -1 \end{bmatrix}$ *iar* $Y[4] := \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$ *vectorul de pornire, adică* $Y[1][4] := 1$ *şi pentru* $i = \overline{2,4}$ $Y[i][4] := 0$.

$Y \cdot p = b$ *are forma:*
$$\begin{cases} -39p_1 + 12p_2 - 4p_3 + p_4 = -120 \\ 20p_1 - 5p_2 + 2p_3 + 0p_4 = 47 \\ 11p_1 - 2p_2 + 1p_3 + 0p_4 = 23 \\ 13p_1 - 4p_2 + 1p_3 + 0p_4 = 43 \end{cases}$$

cu soluţia $p_1 := 3; \quad p_2 := -7; \quad p_3 := -24; \quad p_4 := -15$, *deci ecuaţia caracteristică este*

$$\lambda^4 + 3\lambda^3 - 7\lambda^2 - 24\lambda - 15 = 0.$$

În continuare prezentăm metoda lui Krylov pentru calculul vectorilor proprii.

Presupunem că coeficienţii polinomului caracteristic sunt cunoscute şi în acelaşi timp am reuşit să rezolvăm ecuaţia caracteristică şi se cunosc valorile proprii $\lambda_1, \lambda_2, \ldots, \lambda_n$. Vectorii proprii se află în felul următor, folosind metoda lui Krylov:

$$x^{(i)} = y^{(n-1)} + q_{1i}y^{(n-2)} + \cdots + q_{n-1,i} \cdot y^{(0)}$$

pentru $i = \overline{1,n}$, unde vectorii coloană $y^{(n-1)}, y^{(n-2)}, \ldots, y^{(0)}$ sunt cunoscute de la metoda lui Krylov pentru determinarea polinomului caracteristic, iar coeficienţii q_{ji} ($j = \overline{1,n-1}$ şi $i = \overline{1,n}$) se pot calcula cu metoda schemei lui Horner (vezi paragraful 4.1):

$$q_{0i} = 1; \quad q_{ji} = \lambda_i \cdot q_{j-1,i} + p_j.$$

Într-adevăr λ_i, $i = \overline{1,n}$ fiind o valoare proprie, este soluţia ecuaţiei caracteristice $\lambda^n + p_1\lambda^{n-1} + p_2\lambda^{n-2} + \cdots + p_{n-1}\lambda + p_n = 0$. Conform schemei lui Horner obţinem tabelul:

	1	p_1	p_2	$\ldots$	p_{n-1}	p_n
λ_i	q_{0i}	q_{1i}	q_{2i}	$\ldots$	$q_{n-1,i}$	$q_{n,i}$

unde $q_{0i} = 1$, $q_{1i} = \lambda_i \cdot q_{0i} + p_1$, $q_{2i} = \lambda_i \cdot q_{1i} + p_2, \ldots, q_{n,i} = \lambda_i q_{n-1,i} + p_n = 0$. Astfel

$$\lambda^n + p_1\lambda^{n-1} + p_2\lambda^{n-2} + \cdots + p_{n-1}\lambda + p_n = (\lambda - \lambda_i)(q_{0i}\lambda^{n-1} + q_{1i}\lambda^{n-2} + \cdots + q_{n-1,i}) = 0.$$

Utilizând teorema lui Cayley-Hamilton (vezi teorema 10.1.1) putem pune în locul lui λ matricea A :

$$A^n + p_1A^{n-1} + p_2A^{n-2} + \cdots + p_{n-1} \cdot A + p_nI_n = (A - \lambda_iI_n)(q_{0i}A^{n-1} + q_{1i}A^{n-2} + \cdots + q_{n-1,i}I_n) = O_n$$

unde O_n este matricea nulă de ordinul n. De aici prin înmulţirea ecuaţiei matriciale din dreapta cu un vector $y^{(0)}$ vom obţine:

$$(A - \lambda_iI_n)(q_{0i}A^{n-1}y^{(0)} + q_{1i}A^{n-2}y^{(0)} + \cdots + q_{n-1,i}I_ny^{(0)}) = O_n \cdot y^{(0)},$$

adică

$$(A - \lambda_iI_n)(y^{(n-1)} + q_{1i}y^{(n-2)} + \cdots + q_{n-1,i}y^{(0)}) = \theta_{\mathbb{R}^n}.$$

Notând cu $x^{(i)} = y^{(n-1)} + q_{1i}y^{(n-2)} + \cdots + q_{n-1,i}y^{(0)}$ avem $(A - \lambda_iI_n)x^{(i)} = \theta_{\mathbb{R}^n}$, adică $A \cdot x^{(i)} = \lambda_ix^{(i)}$, ceea ce înseamnă că $x^{(i)}$ este un vector propriu pentru valoarea proprie λ_i.

Program 2 Krylov

Datele de intrare: $n; Y; \lambda$ ($\lambda[i]$ pentru $i = \overline{1,n}$); (de la program 1 Krylov)
Pentru $i = \overline{1,n}$ execută
    $Q[0][i] := 1;$
    Pentru $j = \overline{1,n-1}$ execută

$$Q[j][i] := \lambda[i] * Q[j-1][i] + p[j];$$

Pentru $i = \overline{1,n}$ execută

Pentru $k = \overline{1,n}$ execută

$X[i][k] := Y[k][1]$

Pentru $j = \overline{1,n-1}$ execută

Pentru $k = \overline{1,n}$ execută

$X[i][k] := X[i][k] + Q[j][i] * Y[k][j+1];$

Tipăreşte X (pe linii vom avea vectorii proprii).

10.2 Metoda puterii

La această metodă presupunem că matricea A de ordinul n admite n valori proprii distincte $\lambda_1, \lambda_2, \ldots, \lambda_n$ astfel încât să avem $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|$. Din presupunerea făcută rezultă că matricea A va admite n vectori proprii linear independenţi, notaţi cu $x_1, x_2, \ldots, x_n$, pentru fiecare valoare proprie λ_i câte un vector propriu x_i : $Ax_i = \lambda_i x_i$, $i = \overline{1,n}$. Atunci orice vector $y \in \mathbb{R}^n$ se poate exprima în mod unic ca o combinaţie liniară a vectorilor proprii $x_1, x_2, \ldots, x_n$: $y = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n$. Avem:

$$\begin{aligned} Ay &= A(\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n) = \alpha_1 Ax_1 + \alpha_2 Ax_2 + \cdots + \alpha_n Ax_n = \\ &= \alpha_1\lambda_1 x_1 + \alpha_2\lambda_2 x_2 + \cdots + \alpha_n\lambda_n x_n. \end{aligned}$$

În mod analog se obţine:

$$\begin{aligned} A^2y &= A(Ay) = (\alpha_1\lambda_1 x_1 + \alpha_2\lambda_2 x_2 + \cdots + \alpha_n\lambda_n x_n) = \\ &= \alpha_1\lambda_1 Ax_1 + \alpha_2\lambda_2 Ax_2 + \cdots + \alpha_n\lambda_n Ax_n = \\ &= \alpha_1\lambda_1^2 x_1 + \alpha_2\lambda_2^2 x_2 + \cdots + \alpha_n\lambda_n^2 x_n. \end{aligned}$$

Prin urmare pentru un $k \in \mathbb{N}^*$ arbitrar se poate arăta cu ajutorul inducţiei matematice că:

$$A^k y = \alpha_1\lambda_1^k x_1 + \alpha_2\lambda_2^k x_2 + \cdots + \alpha_n\lambda_n^k x_n.$$

De aici rezultă că:

$$A^k y = \lambda_1^k \left[\alpha_1 x_1 + \alpha_2 \left(\frac{\lambda_2}{\lambda_1}\right)^k x_2 + \cdots + \alpha_n \left(\frac{\lambda_n}{\lambda_1}\right)^k x_n \right].$$

Trecând la limită pentru $k \to \infty$ în egalitatea anterioară, pentru un număr natural k suficient de mare, rezultă că: $A^k y \approx \lambda_1^k \alpha_1 x_1$ şi $A^{k+1} y \approx \lambda_1^{k+1} \alpha_1 x_1$, deoarece conform condiţiilor iniţiale:

$$\lim_{k\to\infty} \left(\frac{\lambda_2}{\lambda_1}\right)^k = 0, \ldots, \lim_{k\to\infty} \left(\frac{\lambda_n}{\lambda_1}\right)^k = 0.$$

Din egalităţile $A^k y = \lambda_1^k \alpha_1 x_1$ şi $A^{k+1} y = \lambda_1^{k+1} \alpha_1 x_1$ rezultă că vectorii coloană $A^{k+1} y$ şi $A^k y$ sunt proporţionale pe componentele corespunzătoare, raportul fiind λ_1. Astfel putem determina pe λ_1, valoarea proprie dominantă. Din egalitatea

$$A(A^k y) = A^{k+1} y = \lambda_1^{k+1} \alpha_1 x_1 = \lambda_1 (\lambda_1^k \alpha_1 x_1) = \lambda_1 (A^k y)$$

putem trage concluzia că vectorul $A^k y$ pentru un k suficient de mare este tocmai vectorul propriu corespunzător valorii proprii λ_1. După ce am determinat pe λ_1, pentru determinarea lui λ_2 se alege vectorul y de forma $y := \alpha_2 x_2 + \cdots + \alpha_n x_n$. Astfel utilizând raţionamentul anterior cu noul vector y se deduce că: $A^k y = \alpha_2 \lambda_2^k x_2 + \alpha_3 \lambda_3^k x_3 + \cdots + \alpha_n \lambda_n^n x_n$. De aici rezultă că:

$$A^k y = \lambda_2^k \cdot \left[\alpha_2 x_2 + \left(\frac{\lambda_3}{\lambda_2}\right)^k x_3 + \cdots + \left(\frac{\lambda_n}{\lambda_2}\right)^k x_n\right].$$

Prin urmare pentru k suficient de mare avem: $A^k y = \lambda_2^k \alpha_2 x_2$, căci

$$\lim_{k\to\infty} \left(\frac{\lambda_3}{\lambda_2}\right)^k = 0, \ldots, \lim_{k\to\infty} \left(\frac{\lambda_n}{\lambda_2}\right)^k = 0.$$

Din egalităţile $A^k y = \lambda_2^k \alpha_2 x_2$ şi $A^{k+1} y = \lambda_2^{k+1} \alpha_2 x_2$ deducem o valoare aproximativă a lui λ_2, folosind un raţionament analog cu determinarea lui λ_1. Vectorul $A^k y$ va fi vectorul propriu corespunzător valorii proprii λ_2. În final pentru determinarea valorilor proprii $\lambda_3, \ldots, \lambda_n$ şi a vectorilor proprii corespunzători utilizăm un raţionament analog.

Capitolul 11

Metode numerice pentru rezolvarea ecuaţiilor diferenţiale şi ale sistemelor de ecuaţii diferenţiale

11.1 Metoda lui Euler pentru rezolvarea numerică a ecuaţiei diferenţiale ordinare de ordinul unu ca problemă Cauchy

Se consideră ecuaţia diferenţială ordinară de ordinul unu: $y' = f(x, y)$, unde f este o funcţie dată şi se caută soluţia $y = y(x)$. Dacă în plus avem condiţia iniţială $y(x_0) = y_0$, unde $x_0, y_0 \in \mathbb{R}$ sunt date, atunci spunem că avem o problemă Cauchy. În analiză se demonstrează existenţa şi unicitatea soluţiei problemei Cauchy folosind anumite condiţii asupra lui f. Dacă se alege $f(x, y) = e^{x^2}$, atunci se obţine ecuaţia diferenţială de ordinul unu $y' = e^{x^2}$, care are soluţia: $y(x) = \int_{x_0}^{x} e^{t^2} dt + y_0$, unde $y(x_0) = y_0$. Se ştie că integrala obţinută nu se poate calcula cu mâna folosind tehnici de schimbare de variabilă şi integrare prin părţi. Prin urmare este nevoie de o metodă numerică.

În continuare prezentăm metoda lui Euler. Vrem să aflăm soluţia aproximativă a problemei Cauchy. Se consideră intervalul $[a, b]$, unde $a = x_0$ şi împărţim acest interval în n părţi egale cu nodurile $a = x_0 < x_1 < x_2 < \cdots < x_n = b$. Fie $h = \dfrac{b-a}{n}$. Atunci $x_1 = x_0 + h$, iar pentru y' vom folosi următoarea formulă de derivare numerică: $y' \approx$

$\frac{y_1-y_0}{x_1-x_0}=\frac{y_1-y_0}{h}$. Din egalitatea $y'(x_0)=f(x_0,y_0)$ obţinem $\frac{y_1-y_0}{h}=f(x_0,y_0)$, adică $y_1=y_0+h\cdot f(x_0,y_0)$. Aici valoarea y_1 va fi o valoare aproximativă pentru curba teoretică a ecuaţiei diferenţiale de ordinul unu în punctul x_1, adică $y_1\approx y(x_1)$. În general ştiind punctul (x_k,y_k) următorul punct se obţine prin formulele $x_{k+1}=x_k+h$ şi $y_{k+1}=y_k+h\cdot f(x_k,y_k)$. Prin urmare pentru a rezolva numeric problema Cauchy trebuie să întocmim tabelul:

x	x_0	x_1	$\ldots$	x_n
$y(x)$	y_0	y_1	$\ldots$	y_n

Se poate demonstra că dacă numărul nodurilor creşte, adică h tinde la zero atunci metoda lui Euler este o metodă convergentă. Într-adevăr, pentru orice $k=\overline{1,n}$ avem de evaluat expresia $|y(x_k)-y_k|$, unde $y(x_k)$ este valoarea teoretică a curbei integrale $y=y(x)$ în punctul $x=x_k$ (adică soluţia teoretică a problemei Cauchy $y'=f(x,y)$, $y_0=y(x_0)$ în punctul $x=x_k$), iar y_k este valoare numerică aproximativă în punctul $x=x_k$. Din problema Cauchy: $y'(x)=f(x,y(x))$ şi $y(x_0)=y_0$ obţinem

$$y(x)=\int_{x_0}^{x}f(t,y(t))dt+y_0.$$

Astfel avem:

$$\begin{aligned}
|y(x_k)-y_k| &= \left|\int_{x_0}^{x_k}f(t,y(t))dt+y_0-y_k|\right|= \\
&= \left|\sum_{i=0}^{k-1}\int_{x_i}^{x_{i+1}}f(t,y(t))dt+\sum_{i=0}^{k-1}(y_i-y_{i+1})\right|= \\
&= \left|\sum_{i=0}^{k-1}\left(\int_{x_i}^{x_{i+1}}f(t,y(t))dt+y_i-y_{i+1}\right)\right|\leq \\
&\leq \sum_{i=0}^{k-1}\left|\int_{x_i}^{x_{i+1}}f(t,y(t))dt+y_i-y_{i+1}\right|= \\
&= \sum_{i=0}^{k-1}\left|\int_{x_i}^{x_{i+1}}f(t,y(t))dt+y_i-(y_i+h\cdot f(x_i,y_i))\right|= \\
&= \sum_{i=0}^{k-1}\left|\int_{x_i}^{x_{i+1}}f(t,y(t))dt-h\cdot f(x_i,y_i)\right|= \\
&= \sum_{i=0}^{k-1}|(x_{i+1}-x_i)\cdot f(\xi_i,y(\xi_i))-h\cdot f(x_i,y_i)|=
\end{aligned}$$

$$\begin{aligned} &= \sum_{i=0}^{k-1} |h \cdot f(\xi_i, y(\xi_i)) - h \cdot f(x_i, y_i)| = \\ &= \sum_{i=0}^{k-1} h \cdot |f(\xi_i, y(\xi_i)) - f(x_i, y_i))| \end{aligned}$$

unde $\xi_i \in (x_i, x_{i+1})$ pentru orice $i = \overline{0, k-1}$, şi unde am aplicat teorema de medie pentru integrala definită

$$\int_{x_i}^{x_{i+1}} f(t, y(t))dt = (x_{i+1} - x_i) \cdot f(\xi_i, y(\xi_i))$$

(vezi lema 9.2.2). În continuare presupunem, că funcţia $f : [a, b] \times \mathbb{R} \to \mathbb{R}$ este lipschitziană, adică există constantele reale pozitive K_1 şi K_2, astfel încât

$$|f(u, v) - f(u', v')| \leq K_1 \cdot |u - u'| + K_2 \cdot |v - v'|$$

oricare ar fi $u, u' \in [a, b]$ şi $v, v' \in \mathbb{R}$. Prin urmare

$$\begin{aligned} &\sum_{i=0}^{k-1} h \cdot |f(\xi_i, y(\xi_i)) - f(x_i, y_i)| \leq \\ &\leq \sum_{i=0}^{k-1} h \cdot (K_1 \cdot |\xi_i - x_i| + K_2 \cdot |y(\xi_i) - y_i|) = \\ = &\sum_{i=0}^{k-1} h \cdot (K_1 \cdot |\xi_i - x_i| + K_2 \cdot |y(\xi_i) - y(x_i)|). \end{aligned}$$

Din condiţia de lipschitzianitate a lui f rezultă continuitatea lui f, deci $y \in C^1([a, b], \mathbb{R})$. Astfel din teorema de medie a lui Lagrange obţinem:

$$|y(\xi_i) - y(x_i)| = |y'(\eta_i) \cdot (\xi_i - x_i)|,$$

unde $\eta_i \in (\xi_i, x_i)$, dacă $\xi_i < x_i$ sau $\eta_i \in (x_i, \xi_i)$ dacă $x_i < \xi_i$. Fie $M_1 = \sup\{|y'(x)| \ / \ x \in [a, b]\} = \max\{|y'(x)| \ / \ x \in [a, b]\}$. Prin urmare

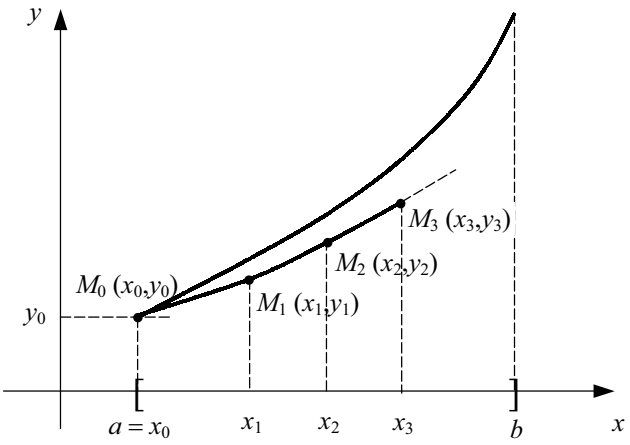
$$|y(\xi_i) - y(x_i)| = |y'(\eta_i) \cdot (\xi_i - x_i)| \leq M_1 \cdot |\xi_i - x_i|$$

Deci

$$\begin{aligned}
|y(x_k) - y_k| &\leq \sum_{i=0}^{k-1} h \cdot (K_1 \cdot |\xi_i - x_i| + K_2 \cdot |y(\xi_i) - y(x_i)|) \leq \\
&\leq \sum_{i=0}^{k-1} h \cdot (K_1 \cdot |\xi_i - x_i| + K_2 \cdot M_1 \cdot |\xi_i - x_i|) = \\
&= \sum_{i=0}^{k-1} h \cdot (K_1 + K_2 \cdot M_1) \cdot |\xi_i - x_i| \leq \\
&\leq (K_1 + K_2 \cdot M_1) \cdot h \cdot \sum_{i=0}^{k-1} |\xi_i - x_i| \leq \\
&\leq (K_1 + K_2 \cdot M_1) \cdot h \cdot \sum_{i=0}^{k-1} h = (K_1 + K_2 \cdot M_1) \cdot k \cdot h^2 \leq \\
&\leq (K_1 + K_2 \cdot M_1) \cdot n \cdot h^2 = (K_1 + K_2 \cdot M_1) \cdot (b-a) \cdot h.
\end{aligned}$$

Dacă se dă dinainte precizia $\varepsilon > 0$ se impune ca $(K_1 + K_2 \cdot M_1) \cdot (b-a) \cdot h < \varepsilon$, adică $h < \dfrac{\varepsilon}{(K_1 + K_2 \cdot M_1) \cdot (b-a)}$, sau $(K_1 + K_2 \cdot M_1) \cdot \dfrac{(b-a)^2}{n} < \varepsilon$, de unde se obţine $n > \dfrac{(K_1 + K_2 \cdot M_1)(b-a)^2}{\varepsilon}$.

Interpretare geometrică: curba teoretică se aproximează cu ajutorul unei linii frânte.



Algoritmul corespunzător metodei lui Euler este:

Algoritm metoda Euler

Datele de intrare: $a; b; n; f; y_0;$

Fie $h := \dfrac{b-a}{n}$; $x[0] := a$; $y[0] := y_0$;

Pentru $i = \overline{0, n-1}$ execută

$x[i+1] := x[i] + h;$

$y[i+1] := y[i] + h * f(x[i], y[i]);$

Tipăreşte $x; y;$ (adică $x[i]$ şi $y[i]$ pentru $i = 0, n$).

Metoda lui Euler se poate extinde şi pentru sisteme de ecuaţii diferenţiale ordinare. Se pune problema rezolvării numerice a următoarei probleme Cauchy: $\begin{cases} y' = f_1(x, y, z) \\ z' = f_2(x, y, z) \end{cases}$ cu $\begin{cases} y(x_0) = y_0 \\ z(x_0) = z_0 \end{cases}$, unde f_1 şi f_2 sunt funcţii date, iar x_0, y_0, z_0 sunt numere date. În acest caz metoda lui Euler are forma: $x_1 = x_0 + h$, $y_1 = y_0 + h \cdot f_1(x_0, y_0, z_0)$, $z_1 = z_0 + h \cdot f_2(x_0, y_0, z_0)$, unde $h = \dfrac{b-a}{n}$, luând nodurile: $a = x_0 < x_1 < x_2 < \cdots < x_n = b$. În general avem: $x_{k+1} = x_k + h$, $y_{k+1} = y_k + h \cdot f_1(x_k, y_k, z_k)$, $z_{k+1} = z_k + h \cdot f_2(x_k, y_k, z_k)$.

11.2 Metoda lui Runge-Kutta de ordinul patru

Pentru rezolvarea problemei Cauchy: $y' = f(x, y)$ şi $y(x_0) = y_0$ avem următoarea metodă: intervalul $[a, b]$ se divide în n părţi egale cu punctele $a = x_0 < x_1 < x_2 < \cdots < x_n = b$, cu pasul $h = \dfrac{b-a}{n}$. Dacă se ştie punctul (x_k, y_k) atunci punctul (x_{k+1}, y_{k+1}) se calculează în felul următor: $x_{k+1} = x_k + h$ iar $y_{k+1} = y_k + \dfrac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)$, unde:

$$\begin{aligned} k_1 &= h \cdot f(x_k, y_k) \\ k_2 &= h \cdot f\left(x_k + \frac{h}{2}, y_k + \frac{k_1}{2}\right) \\ k_3 &= h \cdot f\left(x_k + \frac{h}{2}, y_k + \frac{k_2}{2}\right) \\ k_4 &= h \cdot f\left(x_k + h, y_k + k_3\right) \end{aligned}$$

Această metodă are o acurateţe mare.

Algoritmul metoda Runge-Kutta

Datele de intrare: $a; b; n; f; y_0;$

Fie $h := \dfrac{b-a}{n}$; $x[0] := a$; $y[0] := y_0$;

Pentru $i = \overline{0, n-1}$ execută

$x[i+1] := x[i] + h;$

$K_1 := h * f(x[i], y[i]);$

$K_2 := h * f\left(x[i] + \dfrac{h}{2}, y[i] + \dfrac{K_1}{2}\right);$

$K_3 := h * f\left(x[i] + \dfrac{h}{2}, y[i] + \dfrac{K_2}{2}\right);$

$K_4 := h * f(x[i] + h, y[i] + K_3);$

$y[i+1] := y[i] + \frac{1}{6} * (K_1 + 2 * K_2 + 2 * K_3 + K_4);$

Tipăreşte $x; y;$

Menţionăm că şi această metodă se poate extinde pentru sisteme de ecuaţii diferenţiale cu problemă Cauchy. Fie $\begin{cases} y' = f_1(x, y, z) \\ z' = f_2(x, y, z) \end{cases}$ cu $\begin{cases} y(x_0) = y_0 \\ z(x_0) = z_0 \end{cases}$. În acest caz avem următoarele formule numerice: $x_{k+1} = x_k + h$, $y_{k+1} = y_k + \dfrac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)$, $z_{k+1} = z_k + \dfrac{1}{6}(l_1 + 2l_2 + 2l_3 + l_4)$, unde

$$\begin{aligned}
k_1 &= h \cdot f_1(x_k, y_k, z_k) \quad \text{şi} \quad l_1 = h \cdot f_2(x_k, y_k, z_k); \\
k_2 &= h \cdot f_1\left(x_k + \frac{h}{2}, y_k + \frac{k_1}{2}, z_k + \frac{l_1}{2}\right) \\
l_2 &= h \cdot f_2\left(x_k + \frac{h}{2}, y_k + \frac{k_1}{2}, z_k + \frac{l_1}{2}\right) \\
k_3 &= h \cdot f_1\left(x_k + \frac{h}{2}, y_k + \frac{k_2}{2}, z_k + \frac{l_2}{2}\right) \\
l_3 &= h \cdot f_2\left(x_k + \frac{h}{2}, y_k + \frac{k_2}{2}, z_k + \frac{l_2}{2}\right) \\
k_4 &= h \cdot f_1\left(x_k + h, y_k + k_3, z_k + l_3\right) \\
l_4 &= h \cdot f_2\left(x_k + h, y_k + k_3, z_k + l_3\right).
\end{aligned}$$

11.3 Rezolvarea numerică a problemei lui Dirichlet pe un pătrat

Se consideră pătratul: $D = \{(x, y) \in \mathbb{R}^2 \mid 0 \leq x, y \leq 1\}$, şi ecuaţia lui Laplace: $\dfrac{\partial^2 f}{\partial x^2} + \dfrac{\partial^2 f}{\partial y^2} = 0$ pe D. Se caută soluţia ecuaţiei lui Laplace, funcţia $f = f(x, y)$, cu $(x, y) \in D$ astfel încât se cunosc valorile lui f pe frontiera pătratului: $f(0, y) = \varphi_1(y)$, $f(1, y) = \varphi_2(y)$, $f(x, 0) = \varphi_3(x)$, $f(x, 1) = \varphi_4(x)$, unde $\varphi_1, \varphi_2, \varphi_3, \varphi_4$ sunt funcţii date. O astfel de problemă se numeşte problema lui Dirichlet. În analiză se demonstrează existenţa şi unicitatea problemei lui Dirichlet dacă se impun anumite restricţii asupra funcţiilor $\varphi_1, \varphi_2, \varphi_3$ şi φ_4. În continuare prezentăm o metodă numerică pentru rezolvarea aproximativă a problemei de mai sus. Împărţim laturile pătratului în N părţi egale. Fie $h = \dfrac{1}{N}$, $x_n = n \cdot h$, $y_m = m \cdot h$, unde $n, m = 0, 1, \ldots, N$. Astfel se formează o reţea a pătratului. Să notăm cu $u_{n,m}$ valoarea aproximativă a lui $f(x_n, y_m)$. Atunci folosind formulele de derivare numerică

$$f''(x_0) \simeq \frac{f(x_0 - h) - 2 \cdot f(x_0) + f(x_0 + h)}{h^2}$$

(vezi paragraful 9.1), obţinem:

$$\frac{\partial^2 f}{\partial x^2}(x_n, y_m) = \frac{1}{h^2} \cdot (u_{n-1,m} - 2u_{n,m} + u_{n+1,m})$$
$$\frac{\partial^2 f}{\partial y^2}(x_n, y_m) = \frac{1}{h^2} \cdot (u_{n,m-1} - 2u_{n,m} + u_{n,m+1})$$

Astfel se obţine următoarea formă discretă a ecuaţiei lui Laplace:

$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = \frac{1}{h^2}(u_{n-1,m} + u_{n+1,m} + u_{n,m-1} + u_{n,m+1} - 4u_{n,m}) = 0$$

unde $n, m = \overline{1, N-1}$, adică se obţine un sistem liniar cu $(N-1)^2$ ecuaţii şi $(N-1)^2$ necunoscute. Necunoscutele sistemului liniar sunt notate cu $(u_{n,m})_{n,m=\overline{1,N-1}}$, fiindcă din condiţiile la limită avem:

$$u_{0,m} = f(x_0, y_m) = f(0, y_m) = \varphi_1(y_m);$$
$$u_{N,m} = f(x_N, y_m) = f(1, y_m) = \varphi_2(y_m);$$
$$u_{n,0} = f(x_n, y_0) = f(x_n, 0) = \varphi_3(x_n);$$
$$u_{n,N} = f(x_n, y_N) = f(x_n, 1) = \varphi_4(x_n);$$

oricare ar fi $n, m = \overline{0, N}$.

Rezolvarea sistemului liniar în necunoscutele $(u_{n,m})_{n,m=\overline{1,N-1}}$ se poate face cu metodele numerice corespunzătoare (vezi capitolul 6.)

Bibliografie

[1] Gh. Coman, G. Pavel, I. Rus, A.I. Rus, *Introducere în teoria ecuaţiilor operatoriale*, Editura Dacia, Cluj-Napoca, 1976.

[2] Gh. Coman, *Analiză numerică*, Editura Libris, Cluj-Napoca, 1995.

[3] I. Cuculescu, *Analiză numerică*, Editura Tehnică, Bucureşti, 1967.

[4] B. Démidovitch, I. Maron, *Eléments de calcul numérique*, Editura MIR, Moscova, 1979.

[5] I. Ichim, G. Marinescu, *Metode de aproximare numerică*, Editura Academiei, Bucureşti, 1986.

[6] V. Iorga, B. Jora şi alţii, *Programare numerică*, Editura Teora, 1996.

[7] B. Jankó, *Rezolvarea ecuaţiilor operaţionale în spaţii Banach*, Editura Academiei, Bucureşti, 1969.

[8] Kis Ottó, Kovács Margit, *Metode numerice* (în lb. maghiară), Editura Tehnică, Budapesta, 1973.

[9] D. Larionescu, *Metode numerice*, Editura Tehnică, Bucureşti, 1989.

[10] G. Marinescu, *Analiză numerică*, Editura Academiei, Bucureşti, 1974.

[11] I. Păvăloiu, *Rezolvarea ecuaţiilor prin interpolare*, Editura Dacia, Cluj-Napoca, 1981.

[12] M. Postolache, *Metode numerice*, Editura Sirius, Bucureşti, 1994.

[13] V. Voiévodine, *Principes numériques d'algèbre linéaire*, Editura MIR, Moscova, 1980.

[14] E.A. Volkov, *Numerical Methods*, Editura MIR, Moscova, 1986.