

## 0. Erori

### §0.1. Tipuri și surse de erori

De regulă, în matematică, prin *eroare* se înțelege diferența dintre valoarea exactă a unui număr și valoarea sa aproximativă. Se disting trei tipuri de erori.

1. **Erorile inerente** sunt cele care provin din simplificarea modelului fizic, pentru a putea fi descris printr-un model matematic. În această categorie intră și erorile de date (aparatele de măsură lucrează inevitabil cu anumite abateri) .

2. **Erorile de metodă (trunchiere)** apar datorită faptului că formulele și ecuațiile exacte se înlocuiesc cu formule și ecuații aproximative, pentru a permite calculul printr-un număr finit de operații aritmetice. De exemplu, numărul  $e$  este suma seriei

$1 + \frac{1}{1!} + \frac{1}{2!} + \dots + \frac{1}{n!} + \dots$ , care nu poate fi calculată exact. Însușind numai un număr finit de termeni apare în mod inevitabil o eroare de trunchiere.

3. **Erorile de rotunjire** se datorează faptului că în calcule, numerele cu un număr infinit de zecimale se aproximează prin numere cu un număr finit de zecimale. Pe de altă parte, orice calculator operează doar cu un număr finit de numere reale (evident cu un număr finit de zecimale), în timp ce mulțimea numerelor reale este infinită. Apare astfel o aproximare inevitabilă a numerelor reale care intervin în calcule cu numerele reprezentate în calculator.

În acest capitol sunt analizate erorile inerente și erorile de rotunjire, urmând ca erorile de trunchiere să fie studiate pe parcursul cărții, odată cu prezentarea metodelor numerice respective.

## §0.2. Reprezentarea numerelor în calculator

La baza construcției majorității tipurilor de calculatoare numerice stau elementele bistabile și de aceea se folosește ca bază de reprezentare a numerelor în calculator baza 2, ale cărei cifre sunt 0 și 1.

Este cunoscut că orice număr real  $x$  se poate reprezenta în baza 2 sub forma:

$$x = \pm(\alpha_n 2^n + \alpha_{n-1} 2^{n-1} + \dots + \alpha_0 2^0 + \alpha_{-1} 2^{-1} + \alpha_{-2} 2^{-2} + \dots),$$

cu  $\alpha_i = 0$  sau 1.

De exemplu numărul 13.5 în baza 10 se poate scrie în baza 2 sub forma:

$$13.5 = 1 \cdot 2^3 + 1 \cdot 2^2 + 0 \cdot 2 + 1 \cdot 2^0 + 1 \cdot 2^{-1} = 1101.1$$

Reprezentarea numerelor în calculator (*reprezentarea internă*) se face pe un număr finit de poziții,  $n$ , numit *lungimea cuvântului*, care este fixată la construcția calculatorului. Pentru precizii mai bune, unele calculatoare au posibilitatea reprezentării numărului pe  $2n$ ,  $3n$ , ... poziții (reprezentare pe dublu cuvânt, triplu cuvânt, ...).

Pe cele  $n$  poziții ale unui cuvânt, un număr poate fi reprezentat în *virgulă fixă* (proprie numerelor întregi sau numerelor subunitare) sau în *virgulă mobilă*.

Majoritatea tipurilor de calculatoare numerice folosesc pentru calcule științifice reprezentarea numerelor în virgulă mobilă. În această reprezentare, poziția virgulei zecimale nu este fixă. Orice număr real  $x$  se poate scrie sub forma  $x = a \cdot 10^b$  sau  $x = a \cdot 2^b$  cu  $|a| < 1$  și  $b \in \mathbb{Z}$ ;  $a$  se numește *mantisa* numărului real  $x$ , iar  $b$  *exponentul*.

Reprezentarea în virgulă mobilă este *normalizată* dacă prima cifră a mantisei este nenulă, deci dacă  $|a| \geq 10^{-1}$ , respectiv  $|a| \geq 2^{-1}$ . În acest caz reprezentarea este unică. *Cifrele semnificative* ale unui număr sunt cifrele mantisei, neluând în seamă zerourile care le preced.

**Exemplul** Reprezentarea normalizată, în virgulă mobilă a numărului în baza 10  $x = 13.5$  este:  $0.135 \cdot 10^2 = 0.135_{10} 2$ , indicele 10 arătând că reprezentarea este în baza 10. În acest caz mantisa este  $a = 0.135$ , iar exponentul  $b = 2$ .

În baza 2, același număr are reprezentarea normalizată în virgulă mobilă  $0.11011_2 100 = 0.11011 \cdot 2^4$ , având mantisa  $a = 0.11011$  și exponentul  $b = 100$ .

Pentru orice calculator numeric există numerele fixe  $t$  și  $e$  care reprezintă numărul de cifre ale mantisei, respectiv ale exponentului unui număr real ce poate fi reprezentat în calculatorul respectiv ( $n = t + e$ ). Numerele  $t$  și  $e$  determină

împreună cu baza de numerație (10 sau 2) o mulțime finită de numere reale  $F \subset \mathbb{R}$  care pot fi reprezentate exact în calculator.

### §0.3. Erorile de rotunjire și calculele în virgulă mobilă

Deoarece mulțimea  $F$  a numerelor reprezentabile într-un calculator este finită, se pune problema aproximării unui număr real  $x \notin F$  printr-un număr  $g \in F$ . Această problemă apare nu numai în datele de intrare în calculator, ci și pentru rezultate intermediare sau finale în urma calculelor efectuate în calculator.

Sunt frecvente cazurile când  $x, y \in F$  și  $x \pm y$  sau  $x \cdot y$  sau  $\frac{x}{y}$  nu fac parte din mulțimea  $F$ .

În mod natural, orice număr  $x \notin F$  se aproximează printr-un număr din  $F$ , notat cu  $rd(x)$ , care este cel mai apropiat număr de  $x$  ce aparține lui  $F$ . Numărul  $rd(x)$  va satisface

$$|x - rd(x)| \leq |x - g|, \quad \forall g \in F,$$

obținându-se în multe cazuri prin *rotunjire*.

**Exemplul 1.** Fie un calculator cu  $t = 4$  și  $e = 1$ . Atunci

$$rd(0.14285_{100}) = 0.1429_{100}$$

$$rd(3.14159_{100}) = 3.1416_{100}$$

$$rd(0.142842_{102}) = 0.1428_{102}.$$

În general, pentru  $t$  fixat, dacă  $x \notin F$ ,  $rd(x)$  se poate determina astfel. Se aduce  $x$  la forma normalizată  $x = a \cdot 10^b$ , cu  $|a| \geq 10^{-1}$ , adică

$$|a| = 0.\alpha_1\alpha_2\dots\alpha_i\alpha_{i+1}\dots, \quad 0 \leq \alpha_i \leq 9, \quad \alpha_1 \neq 0.$$

Se determină

$$a' = \begin{cases} 0.\alpha_1\alpha_2\dots\alpha_t, & \text{dacă } 0 \leq \alpha_{t+1} \leq 4 \\ 0.\alpha_1\alpha_2\dots\alpha_t + 10^{-t}, & \text{dacă } \alpha_{t+1} \geq 5 \end{cases},$$

adică  $\alpha_t$  crește cu 1 dacă  $\alpha_{t+1} \geq 5$  și se renunță la celelalte zecimale începând cu  $\alpha_{t+1}$ .

În final, definim  $rd(x) = \text{sign}(x) \cdot a' \cdot 10^b$ .

Pentru orice număr real  $x \neq 0$ , eroarea relativă este raportul  $\left| \frac{rd(x) - x}{x} \right|$ .

Deoarece  $|a| \geq 10^{-1}$ , eroarea relativă admite următoarea margine

$$\left| \frac{r\tilde{d}(x) - x}{x} \right| \leq \frac{5 \cdot 10^{-(t+1)}}{|a|} \leq 5 \cdot 10^{-t}.$$

Notăm  $\text{eps} = 5 \cdot 10^{-t}$ . Dacă  $r\tilde{d}(x) = x(1 + \varepsilon)$ , atunci, din inegalitatea de mai sus rezultă  $|\varepsilon| \leq \text{eps}$ . Numărul real  $\text{eps}$  se numește *precizia calculatorului*.

Pentru rotunjirea în sistemul binar se procedează analog. Se aduce  $x \notin F$  la forma normală  $x = a \cdot 2^b$  cu  $2^{-1} \leq |a| < 1$  și  $|a| = 0.\alpha_1\alpha_2\dots\alpha_i\alpha_{i+1}\dots$  cu  $\alpha_i = 0$  sau 1 și  $\alpha_1 = 1$ .

Se determină  $a' = \begin{cases} 0.\alpha_1\alpha_2\dots\alpha_t, & \text{dacă } \alpha_{t+1} = 0 \\ 0.\alpha_1\alpha_2\dots\alpha_t + 2^{-t}, & \text{dacă } \alpha_{t+1} = 1 \end{cases}$ , iar  $r\tilde{d}(x) = \text{sgn}(x) \cdot a' \cdot 2^b$ . În

acest caz  $\text{eps} = 2^{-t}$ .

În cazul în care  $r\tilde{d}(x) \in F$ , atunci  $rd(x)$  este chiar  $r\tilde{d}(x)$ . Deoarece numărul pozițiilor pentru exponentul  $e$  este finit, există, din păcate, numere  $x \notin F$  pentru care  $r\tilde{d}(x) \notin F$ .

**Exemplul 2.** Considerăm  $t = 4$  și  $e = 2$ . Atunci

- a)  $r\tilde{d}(0.31794_{10}110) = 0.3179_{10}110 \notin F$
- b)  $r\tilde{d}(0.99997_{10}99) = 0.1000_{10}100 \notin F$
- c)  $r\tilde{d}(0.012345_{10} - 99) = 0.1235_{10} - 100 \notin F$
- d)  $r\tilde{d}(0.54321_{10} - 110) = 0.5432_{10} - 110 \notin F$

În cazurile a) și b) exponentul pozitiv este prea mare ca să poată fi reprezentat pe spațiul alocat ( $e = 2$ ). În situațiile acestea se spune că avem *depășire superioară* a exponentului. În cazul b) depășirea superioară a exponentului apare abia după rotunjire. În exemplele c) și d) are loc *depășire de exponent inferioară*, adică exponentul negativ este prea mic pentru a putea fi reprezentat pe spațiul alocat. În aceste două situații, depășirea inferioară a exponentului poate fi prevenită definind  $rd(0.012345_{10} - 99) = 0.0123_{10} - 99 \in F$  (reprezentarea nu mai este normalizată) și

$$rd(0.54321_{10} - 110) = 0 \in F.$$

Atunci  $rd$  nu satisface egalitatea  $rd(x) = x(1 + \varepsilon)$ , deci eroarea relativă poate fi mai mare ca  $\text{eps}$ .

Situațiile de depășire de exponent superioară sau inferioară sunt tratate de calculatoarele numerice ca fiind excepții.

În mod obișnuit,  $rd(x)$  se definește prin egalitatea  $rd(x) = r\tilde{d}(x)$ .

În continuare, ținând seama că depășirile de exponent superioare și/sau inferioare nu sunt frecvente, vom considera cazul ideal  $e = \infty$  și

$rd : \mathbb{R} \rightarrow F$  prin  $rd(x) = x(1 + \varepsilon)$  cu  $|\varepsilon| \leq eps$ ,  $(\forall) x \in \mathbb{R}$ .

Se poate întâmpla ca rezultatul operațiilor aritmetice  $x \pm y$ ,  $x \cdot y$ ,  $\frac{x}{y}$  să nu fie elemente ale mulțimii  $F$ , chiar dacă operandii  $x$  și  $y \in F$ .

Vom nota cu  $+^*$ ,  $-^*$ ,  $\cdot^*$ ,  $/^*$  operațiile în virgulă mobilă corespunzătoare operațiilor aritmetice care sunt definite astfel:

$$\begin{aligned} x +^* y &\stackrel{def}{=} rd(x + y), \\ x -^* y &\stackrel{def}{=} rd(x - y), \\ x \cdot^* y &\stackrel{def}{=} rd(x \cdot y), \\ x /^* y &\stackrel{def}{=} rd(x / y), \end{aligned} \quad \text{pentru orice } x, y \in F$$

deci

$$\begin{aligned} x +^* y &= (x + y)(1 + \varepsilon_1), \\ x -^* y &= (x - y)(1 + \varepsilon_2), \\ x \cdot^* y &= (x \cdot y)(1 + \varepsilon_3), \\ x /^* y &= (x / y)(1 + \varepsilon_4). \end{aligned} \quad \text{cu } |\varepsilon_i| \leq eps, i = \overline{1,4}.$$

Aceste operații în virgulă mobilă nu au proprietățile binecunoscute ale operațiilor aritmetice. De exemplu:

1)  $x +^* y = x$ , dacă  $|y| < \frac{eps}{B}|x|$ , unde  $x, y \in F$ , iar  $B$  este baza de numerație.

Precizia mașinii  $eps$  ar putea fi definită ca fiind cel mai mic număr  $g \in F$  pentru care  $1 +^* g > 1$ , adică  $eps = \min\{g \in F \mid 1 +^* g > 1, g > 0\}$ .

2) Asociativitatea nu se mai păstrează, așa cum va rezulta din următorul exemplu.

**Exemplul 3.**

Fie  $a = 0.23371258_{10} - 4$ ,  $b = 0.33678429_{10} 2$ ,  $c = -0.33677811_{10} 2$ .

Într-un calculator cu  $t = 8$  și  $e = \infty$ , operația  $+^*$  conduce la:

$$\begin{aligned} a +^* (b +^* c) &= 0.23371258_{10} - 4 +^* 0.61800000_{10} - 3 = \\ &= 0.02337126_{10} - 3 +^* 0.61800000_{10} - 3 = 0.64137126_{10} - 3 \\ (a +^* b) +^* c &= (0.23371258_{10} - 4 +^* 0.33678429_{10} 2) -^* 0.33677811_{10} 2 = \\ &= (0.00000023_{10} 2 +^* 0.33678429_{10} 2) -^* 0.33677811_{10} 2 = \\ &= 0.33678452_{10} 2 -^* 0.33677811_{10} 2 = 0.64100000_{10} - 3 \end{aligned}$$

Rezultatul exact al adunării este:

$$a + b + c = 0.00000023371258_{10}2 + 0.33678429_{10}2 - 0.33677811_{10}2 = \\ = 0.641371258_{10} - 3 .$$

Dacă  $E$  este o expresie aritmetică, va rezulta din context cum se evaluează  $E$ . Dacă este nevoie se pot folosi paranteze care să precizeze ordinea operațiilor. Vom nota cu  $f(E)$  valoarea expresiei  $E$  obținută din calculul în virgulă mobilă. De exemplu:

$$f(x + y) \stackrel{def}{=} x + * y \\ f(x + (y + z)) \stackrel{def}{=} x + * (y + * z) \\ f((x + y) + z) \stackrel{def}{=} (x + * y) + * z$$

### §0.4. Propagarea erorilor

Așa cum am văzut în Exemplul 3 din paragraful precedent, în funcție de schema aleasă pentru evaluarea unei expresii am obținut rezultate diferite în calcule în aritmetica virgulei mobile. De aceea este necesar să distingem între diferitele scheme de calcul, chiar dacă din punct de vedere matematic ele sunt echivalente.

Desemnăm cu termenul *algoritm* o secvență finită de operații elementare care descriu cum se calculează soluția unei probleme.

În cele ce urmează vom formaliza noțiunea de algoritm, pentru a putea descrie propagarea erorilor.

Presupunem că numerele  $y_1, y_2, \dots, y_m$  constituie soluția unei probleme ale cărei date de intrare sunt  $x_1, \dots, x_n$ . Dacă introducem vectorii coloană

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \quad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}, \quad \text{atunci algoritmul pentru rezolvarea problemei de mai sus}$$

revine la determinarea funcției vectoriale  $y = \varphi(x)$ ,  $\varphi: D \rightarrow \mathbb{R}^m$ ,  $D \subset \mathbb{R}^n$ ,  $\varphi$  fiind dată de  $m$  funcții  $\varphi_i: D \rightarrow \mathbb{R}$ ,  $y_i = \varphi_i(x_1, \dots, x_n)$ ,  $i = \overline{1, m}$ .

La fiecare etapă de calcul există o *mulțime operand* de numere, care sunt fie numerele de intrare  $x_i$  sau au rezultat din operații anterioare. O operație elementară calculează un nou număr din unul sau mai multe elemente ale mulțimii operand. Acest nou număr este, fie un rezultat intermediar, fie unul final și se adaugă mulțimii operand, care este curățată de datele care nu mai sunt necesare

pentru restul calculelor. Mulțimea operand finală va consta din rezultatele dorite  $y_1, y_2, \dots, y_m$ . În concluzie, o operație corespunde unei transformări a mulțimii operand. Scriind mulțimile operand consecutive ca vectori coloană

$$x^{(i)} = \begin{pmatrix} x_1^{(i)} \\ x_2^{(i)} \\ \vdots \\ x_{n_i}^{(i)} \end{pmatrix} \in \mathbb{R}^{n_i}, \text{ putem asocia fiecărei operații elementare o funcție vectorială}$$

elementară astfel:  $\varphi^{(i)} : D_i \rightarrow \mathbb{R}^{n_{i+1}}, D_i \subset \mathbb{R}^{n_i}$  astfel încât  $\varphi^{(i)}(x^{(i)}) = x^{(i+1)}$  cu  $x^{(0)} = x$ , unde  $x^{(i+1)}$  este o reprezentare vectorială a mulțimii operand transformate.

Fiind dat un algoritm, șirul său de operații elementare dă naștere unei descompuneri a lui  $\varphi$  într-un șir de funcții elementare  $\varphi^{(i)} : D_i \rightarrow D_{i+1}, i = \overline{0, r}, D_j \subseteq \mathbb{R}^{n_j}, \varphi = \varphi^{(r)} \circ \varphi^{(r-1)} \circ \dots \circ \varphi^{(0)}, D_0 = D, D_{r+1} \subseteq \mathbb{R}^{n_{r+1}} = \mathbb{R}^m$ .

**Exemplul 1.** Pentru  $\varphi(a, b, c) = a + b + c$  să evidențiem doi algoritmi:

a) Fie  $\eta = a + b$  și  $y = c + \eta$ . Atunci, descompunerea de mai sus este:

$$\varphi^{(0)}(a, b, c) = \begin{pmatrix} a + b \\ c \end{pmatrix} \in \mathbb{R}^2, \varphi^{(1)}(u, v) = u + v \in \mathbb{R}, \varphi(a, b, c) = \varphi^{(1)}(\varphi^{(0)}(a, b, c)).$$

b) Fie  $\eta = b + c$  și  $y = a + \eta$ . În acest caz:

$$\varphi^{(0)}(a, b, c) = \begin{pmatrix} a \\ b + c \end{pmatrix} \in \mathbb{R}^2, \varphi^{(1)}(u, v) = u + v \in \mathbb{R}, \varphi(a, b, c) = \varphi^{(1)}(\varphi^{(0)}(a, b, c)).$$

Un argument pentru alegerea unui algoritm îl constituie propagarea erorilor în aritmetica virgulei mobile.

Să analizăm algoritmi a) și b) din exemplul de mai sus pentru intrările numerice din Exemplul 3 al secțiunii precedente. În cazul a)

$$\begin{aligned} \eta &= fl(a + b) = (a + b)(1 + \varepsilon_1), \\ \tilde{y} &= fl(\eta + c) = (\eta + c)(1 + \varepsilon_2) = [(a + b)(1 + \varepsilon_1) + c](1 + \varepsilon_2) = \\ &= (a + b + c) \left[ 1 + \frac{a + b}{a + b + c} \varepsilon_1 (1 + \varepsilon_2) + \varepsilon_2 \right]. \end{aligned}$$

Eroarea relativă a lui  $\tilde{y}$ ,  $\varepsilon_y = \frac{\tilde{y} - y}{y} = \frac{a + b}{a + b + c} \varepsilon_1 (1 + \varepsilon_2) + \varepsilon_2$ , sau, după

renunțarea la termenii neliniari în  $\varepsilon$  avem  $\varepsilon_y \approx \frac{a + b}{a + b + c} \varepsilon_1 + 1 \cdot \varepsilon_2$ .

Coefficienții lui  $\varepsilon_1$  și  $\varepsilon_2$  măsoară efectul erorilor de rotunjire  $\varepsilon_1$  și  $\varepsilon_2$  asupra erorii  $\varepsilon_y$  a rezultatului.

Factorul  $\frac{a+b}{a+b+c}$  este critic; în funcție de care dintre numerele  $|a+b|$  sau  $|b+c|$  este mai mic, este mai bine să se procedeze via  $(a+b)+c$  decât  $a+(b+c)$  pentru a calcula  $a+b+c$ .

Metoda de mai sus, de studiu a propagării erorilor, neglijând termenii de ordin superior, se poate extinde, conducând la *analiza diferențială a erorilor* unui algoritm, pentru a calcula  $\varphi(x)$ , dacă  $\varphi = \varphi^{(r)} \circ \varphi^{(r-1)} \circ \dots \circ \varphi^{(0)}$ .

Pentru aceasta vom investiga cum erorile  $\Delta x$  asupra datelor de intrare, ca și erorile de rotunjire acumulate de-a lungul algoritmului afectează rezultatul final  $y = \varphi(x)$ .

Fie  $\varphi: D \rightarrow \mathbb{R}^m$ ,  $\varphi(x) = \begin{pmatrix} \varphi_1(x_1, \dots, x_n) \\ \vdots \\ \varphi_m(x_1, \dots, x_n) \end{pmatrix}$ ,  $D$  fiind o submulțime deschisă

a lui  $\mathbb{R}^n$ . Presupunem că funcțiile  $\varphi_i$ ,  $i = \overline{1, m}$ , au derivate continue pe  $D$ . Fie  $\tilde{x}$  o valoare aproximativă pentru  $x$ .

Notăm  $\Delta x = \tilde{x} - x$ ,  $\Delta x_i = \tilde{x}_i - x_i$  erorile absolute ale lui  $\tilde{x}$  și respectiv  $\tilde{x}_i$ .

Fie  $\varepsilon_{\tilde{x}_i} = \frac{\tilde{x}_i - x_i}{x_i}$ , dacă  $x_i \neq 0$  erorile relative.

Înlocuind datele de intrare  $x$  cu  $\tilde{x}$ , obținem  $\tilde{y} = \varphi(\tilde{x})$  în loc de  $y = \varphi(x)$ .

Dezvoltând în serie Taylor și renunțând la termenii neliniari în  $\Delta x_i$  obținem:

$$\Delta y_i = \tilde{y}_i - y_i = \varphi_i(\tilde{x}) - \varphi_i(x) = \sum_{j=1}^n (\tilde{x}_j - x_j) \frac{\partial \varphi_i}{\partial x_j}(x) = \sum_{j=1}^n \frac{\partial \varphi_i}{\partial x_j}(x) \Delta x_j, \quad i = \overline{1, m}.$$

Matriceal putem scrie:

$$\Delta y = D\varphi(x)\Delta x,$$

unde  $D\varphi(x)$  este matricea jacobiană a funcției  $\varphi$ ,

$$D\varphi(x) = \begin{pmatrix} \frac{\partial \varphi_1}{\partial x_1}(x) & \dots & \frac{\partial \varphi_1}{\partial x_n}(x) \\ \dots & \dots & \dots \\ \frac{\partial \varphi_m}{\partial x_1}(x) & \dots & \frac{\partial \varphi_m}{\partial x_n}(x) \end{pmatrix}.$$

Cantitățile  $\frac{\partial \varphi_i}{\partial x_j}(x)$  semnifică modul în care  $y_i$  sunt afectați de erorile absolute  $\Delta x_j$  ale lui  $x_j$ .

Dacă  $y_i \neq 0, i = \overline{1, m}$  și  $x_j \neq 0, j = \overline{1, n}$  atunci formula pentru propagarea erorilor relative devine:

$$\varepsilon_{y_i} = \sum_{j=1}^n \frac{x_j}{\varphi_i(x)} \cdot \frac{\partial \varphi_i(x)}{\partial x_j} \cdot \varepsilon_{x_j} \quad (1)$$

Cantitățile  $\frac{x_j}{\varphi_i} \cdot \frac{\partial \varphi_i}{\partial x_j}$  se numesc *numere de condiționare*.

Dacă numerele de condiționare au valori (absolute) mari se spune că respectiva problemă este *rău condiționată*, altfel problema este *bine condiționată*.

Pentru problemele rău condiționate, erori relative mici în datele de intrare  $x$ , pot cauza erori relative mari în rezultatele  $y, y = \varphi(x)$ .

**Exempul 3.** Pentru  $y = \varphi(a, b, c) = a + b + c$  avem:

$$\varepsilon_y = \frac{a}{a+b+c} \varepsilon_a + \frac{b}{a+b+c} \varepsilon_b + \frac{c}{a+b+c} \varepsilon_c.$$

Această problemă este bine condiționată dacă termenii  $a, b, c$  sunt mici în comparație cu suma lor  $a+b+c$ .

Eroarea relativă (1) pentru câteva operații elementare, în care operanzii  $x$  și  $y$  sunt nenuli, este dată în tabelul:

$\varphi$	$\varepsilon_\varphi$
$x+y$	$\frac{x}{x+y} \varepsilon_x + \frac{y}{x+y} \varepsilon_y$ , dacă $x+y \neq 0$
$x-y$	$\frac{x}{x-y} \varepsilon_x - \frac{y}{x-y} \varepsilon_y$ , dacă $x-y \neq 0$
$x \cdot y$	$\varepsilon_x + \varepsilon_y$
$x/y$	$\varepsilon_x - \varepsilon_y$
$\sqrt{x}$	$\frac{1}{2} \varepsilon_x$

Se observă că la înmulțire, împărțire și extragere de rădăcină pătrată erorile relative ale operanzilor nu se propagă puternic în rezultate. Același lucru se întâmplă și în cazul adunării dacă cei doi operanzi au același semn. Atunci

numerele de condiționare  $\left| \frac{x}{x+y} \right|$  și  $\left| \frac{y}{x+y} \right|$  sunt în  $(0,1)$  și o margine a erorii relative este  $|\varepsilon_{x+y}| \leq \max\{|\varepsilon_x|, |\varepsilon_y|\}$ .

Dacă operanzii care se adună au semne diferite, atunci cel puțin unul dintre numerele de condiționare  $\left| \frac{x}{x+y} \right|, \left| \frac{y}{x+y} \right|$  este mai mare ca 1 și erorile relative  $\varepsilon_x, \varepsilon_y$  sunt amplificate. Dacă  $x \approx -y$  atunci se abandonează calculele (apare *cancelarea rezultatului*).

Pentru descrierea propagării erorilor de rotunjire într-un algoritm dat  $\varphi$ , vom apela tot la formula  $\Delta y = D\varphi(x) \cdot \Delta x$ .

Presupunem că  $\varphi$  admite reprezentarea  $\varphi = \varphi^{(r)} \circ \varphi^{(r-1)} \circ \dots \circ \varphi^{(0)}$  și că rezultatele calculelor pornind de la vectorul datelor de intrare,  $x = x^{(0)}$  sunt date de relațiile:

$$x = x^{(0)}, x^{(1)} = \varphi^{(0)}(x^{(0)}), \dots, y = x^{(r+1)} = \varphi^{(r)}(x^{(r)}), \varphi^{(i)} \in C^1(D_i), i = \overline{0, r}.$$

$$\text{Notăm } \psi^{(i)} = \varphi^{(r)} \circ \dots \circ \varphi^{(i)} : D_i \rightarrow R^m, i = \overline{0, r} \text{ și } \psi^{(0)} = \varphi.$$

Erorile din datele de intrare și erorile de rotunjire vor perturba rezultatele intermediare exacte  $x^{(i)}$ , obținându-se în aritmetica virgulei mobile aproximările  $\tilde{x}^{(i)}$  date de relațiile:  $\tilde{x}^{(i+1)} = fl(\varphi^{(i)}(\tilde{x}^{(i)}))$ . Atunci erorile absolute se pot evalua astfel:

$$\Delta x^{(i+1)} = [fl(\varphi^{(i)}(\tilde{x}^{(i)})) - \varphi^{(i)}(\tilde{x}^{(i)})] + [\varphi^{(i)}(\tilde{x}^{(i)}) - \varphi^{(i)}(x^{(i)})]. \quad (2)$$

Dar

$$\varphi^{(i)}(\tilde{x}^{(i)}) - \varphi^{(i)}(x^{(i)}) \approx D\varphi^{(i)}(x^{(i)})\Delta x^{(i)} \quad (3).$$

Cum  $\varphi^{(i)}$  este o funcție elementară, evaluarea sa în virgulă mobilă va fi

$$fl(\varphi^{(i)}(u)) = rd(\varphi^{(i)}(u)),$$

care pe componente devine:

$$fl(\varphi_j^{(i)}(u)) = rd(\varphi_j^{(i)}(u)) = (1 + \varepsilon_j)\varphi_j^{(i)}(u) \text{ cu } |\varepsilon_j| \leq eps, j = \overline{1, n_{i+1}}.$$

Aici  $\varepsilon_j$  este eroarea de rotunjire generată în timpul calculului în virgulă mobilă a componentei a  $j$ -a, a lui  $\varphi^{(i)}$ .

Relațiile de mai sus se scriu matriceal sub forma

$$fl(\varphi^{(i)}(u)) = (I + E_{i+1})\varphi^{(i)}(u),$$

unde  $I$  este matricea unitate de ordinul  $n_{i+1}$ , iar  $E_{i+1}$  este matricea diagonală a erorilor:

$$E_{i+1} = \begin{pmatrix} \varepsilon_1 & 0 & \dots & \dots & 0 \\ 0 & \varepsilon_2 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \dots & \varepsilon_{n_{i+1}} \end{pmatrix} \text{ cu } |\varepsilon_j| \leq eps.$$

Prima paranteză dreaptă din  $\Delta x^{(i+1)}$  devine:

$$fl(\varphi^{(i)}(\tilde{x}^{(i)})) - \varphi^{(i)}(\tilde{x}^{(i)}) = E_{i+1} \cdot \varphi^{(i)}(\tilde{x}^{(i)}) \approx E_{i+1} \cdot \varphi^{(i)}(x^{(i)}) = E_{i+1} x^{(i+1)} = \alpha_{i+1}.$$

Vectorul coloană  $\alpha_{i+1}$  poate fi interpretat ca eroarea absolută de rotunjire apărută atunci când se evaluează în aritmetica virgulei mobile  $\varphi^{(i)}$ , iar elementele diagonale ale matricei  $E_{i+1}$  pot fi interpretate ca erori relative de rotunjire corespunzătoare. Ținând seamă de (2) și (3) rezultă:

$$\Delta x^{(i+1)} = \alpha_{i+1} + D\varphi^{(i)}(x^{(i)}) \cdot \Delta x^{(i)} = E_{i+1} \cdot x^{(i+1)} + D\varphi^{(i)}(x^{(i)}) \cdot \Delta x^{(i)}$$

$$i \geq 0, \Delta x^{(0)} = \Delta x.$$

Prin urmare:

$$\Delta x^{(1)} = D\varphi^{(0)}(x) \cdot \Delta x + \alpha_1$$

$$\Delta x^{(2)} = D\varphi^{(1)}(x^1) [D\varphi^{(0)}(x) \cdot \Delta x + \alpha_1] + \alpha_2$$

$$\dots \dots \dots \Delta y = \Delta x^{(r+1)} = D\varphi^{(r)} \dots D\varphi^{(0)} \Delta x + D\varphi^{(r)} \dots D\varphi^{(1)} \alpha_1 + \dots + \alpha_{r+1}.$$

Sau dacă ținem seama de ceea ce am notat cu  $\psi$  avem:

$$\begin{aligned} \Delta y &= D\varphi(x)\Delta x + D\psi^{(1)}(x^{(1)}) \cdot \alpha_1 + \dots + D\psi^{(r)}(x^{(r)}) \cdot \alpha_r + \alpha_{r+1} = \\ &= D\varphi(x)\Delta x + D\psi^{(1)}(x^{(1)}) \cdot E_1 x^{(1)} + \dots + D\psi^{(r)}(x^{(r)}) \cdot E_r x^{(r)} + E_{r+1} y \end{aligned}$$

Aceasta arată că matricea jacobiană  $D\psi^{(i)}$  este importantă pentru efectul erorilor de rotunjire intermediare  $\alpha_i$  sau  $E_i$  asupra rezultatului final.

**Exemplul 5.** Fie de calculat expresia  $a^2 - b^2 = (a + b)(a - b)$ .

Vom prezenta 2 algoritmi:

*Algoritmul 1:*

$$\eta_1 = a \times a, \eta_2 = b \times b, y = \eta_1 - \eta_2. \text{ Atunci } \varphi^{(0)}(a, b) = \begin{pmatrix} a^2 \\ b \end{pmatrix},$$

$$\varphi^{(1)}(u, v) = \begin{pmatrix} u \\ v^2 \end{pmatrix}, \varphi^{(2)}(s, t) = s - t, \varphi = \varphi^{(2)} \circ \varphi^{(1)} \circ \varphi^{(0)}.$$

*Algoritmul 2:*

$$\eta_1 = a + b, \eta_2 = a - b, y = \eta_1 \cdot \eta_2. \text{ Atunci } \varphi^{(0)}(a, b) = \begin{pmatrix} a \\ b \\ a + b \end{pmatrix}$$

$$\varphi^{(1)}(a, b, u) = \begin{pmatrix} u \\ a - b \end{pmatrix}; \varphi^{(2)}(s, t) = s - t, \varphi = \varphi^{(2)} \circ \varphi^{(1)} \circ \varphi^{(0)}.$$

În primul caz avem:

$$x = x^{(0)} = \begin{pmatrix} a \\ b \end{pmatrix}, x^{(1)} = \begin{pmatrix} a^2 \\ b \end{pmatrix}, x^{(2)} = \begin{pmatrix} a^2 \\ b^2 \end{pmatrix}, x^{(3)} = y = a^2 - b^2,$$

$$\psi^{(1)}(u, v) = u - v^2, \psi^{(2)}(u, v) = u - v, D\varphi(x) = (2a, -2b),$$

$$D\psi^{(1)}(x^{(1)}) = (1, -2b), \quad D\psi^{(2)}(x^{(2)}) = (1, -1), \quad \alpha_1 = \begin{pmatrix} \varepsilon_1 a^2 \\ 0 \end{pmatrix}, E_1 = \begin{pmatrix} \varepsilon_1 & 0 \\ 0 & 0 \end{pmatrix},$$

deoarece  $f(\varphi^{(0)}(x^{(0)})) - \varphi^{(0)}(x^{(0)}) = \begin{pmatrix} a \cdot^* a \\ b \end{pmatrix} - \begin{pmatrix} a^2 \\ b \end{pmatrix}$  (eroarea de rotunjire în virgulă mobilă apare doar pe prima poziție). Similar

$$\alpha_2 = \begin{pmatrix} 0 \\ \varepsilon_2 b^2 \end{pmatrix}, E_2 = \begin{pmatrix} 0 & 0 \\ 0 & \varepsilon_2 \end{pmatrix}, \alpha_3 = \varepsilon_3 (a^2 - b^2), |\varepsilon_i| \leq eps, \quad i = \overline{1,3}.$$

$$\text{Dacă } \Delta x = \begin{pmatrix} \Delta a \\ \Delta b \end{pmatrix}, \Delta y = 2a\Delta a - 2b\Delta b + a^2\varepsilon_1 - b^2\varepsilon_2 + (a^2 - b^2)\varepsilon_3.$$

În cazul algoritmului 2, avem:

$$x = x^{(0)} = \begin{pmatrix} a \\ b \end{pmatrix}, x^{(1)} = \begin{pmatrix} a+b \\ a-b \end{pmatrix}, x^{(2)} = y = a^2 - b^2; \psi^{(1)}(u, v) = u \cdot v,$$

$$D\varphi(x) = (2a, -2b), D\psi^{(1)}(x^{(1)}) = (a-b, a+b); \alpha_1 = \begin{pmatrix} \varepsilon_1(a+b) \\ \varepsilon_2(a-b) \end{pmatrix},$$

$$\alpha_2 = \varepsilon_3(a^2 - b^2) \quad \text{și } E_1 = \begin{pmatrix} \varepsilon_1 & 0 \\ 0 & \varepsilon_2 \end{pmatrix}; |\varepsilon_i| \leq eps.$$

$$\text{Atunci } \Delta y = 2a\Delta a - 2b\Delta b + (a^2 - b^2)(\varepsilon_1 + \varepsilon_2 + \varepsilon_3).$$

Se poate da o margine a efectului erorilor de rotunjire astfel:

- în algoritmul 1:

$$\left| a^2\varepsilon_1 - b^2\varepsilon_2 + (a^2 - b^2)\varepsilon_3 \right| \leq (a^2 + b^2 + |a^2 - b^2|)eps$$

- în algoritmul 2:

$$\left| (a^2 - b^2)(\varepsilon_1 + \varepsilon_2 + \varepsilon_3) \right| \leq 3|a^2 - b^2|eps.$$

Când selectăm un anumit algoritm de calcul pentru un  $\varphi(x)$  (cu alte cuvinte, o anumită descompunere a lui  $\varphi$  în funcții elementare),  $D\varphi$  rămâne neschimbat; matricele jacobiene  $D\psi^{(i)}$  care măsoară propagarea rotunjirilor vor fi totuși diferite, efectul total al erorilor de rotunjire va fi:

$$D\psi^{(1)}\alpha_1 + \dots + D\psi^{(2)}\alpha_2 + \alpha_{r+1}.$$

Un algoritm este numeric mai bun decât alt algoritm pentru calculul lui  $\varphi(x)$ , dacă pentru o mulțime de date  $x$ , efectul total al erorilor de rotunjire este mai mic în cazul primului algoritm.

## 1. Sisteme de ecuații liniare

Reamintim că un sistem de  $n$  ecuații algebrice liniare cu  $n$  necunoscute este de forma:

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

Dacă notăm cu  $A$  matricea coeficienților, cu  $x$  vectorul coloană format cu necunoscutele sistemului și cu  $b$  coloana termenilor liberi, sistemul (1) se scrie sub formă matriceală :

$$Ax=b, \quad (2)$$

unde:

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$

Metodele numerice de rezolvare a sistemelor algebrice de ecuații liniare sunt de două tipuri: *metode directe* și *metode indirecte* (sau *iterative*).

*Metodele directe* constau în transformarea sistemului (1) într-un sistem triunghiular echivalent, care se rezolvă ușor. Cele mai cunoscute metode directe sunt: *metoda Gauss*, *metoda Cholesky* (utilizată pentru sistemele în care matricea  $A$  este simetrică și pozitiv definită) și *metoda Householder*.

Metodele directe permit determinarea soluției exacte a sistemului în cazul ideal, când nu avem erori de rotunjire. Numărul operațiilor aritmetice efectuate este de ordinul  $n^3$ . Pentru sisteme cu un număr de ecuații mai mare de 100, metodele directe devin inutilizabile datorită acumulării erorilor de rotunjire care alterează soluția.

*Metodele indirecte* (sau *iterative*) constau în construcția unui șir  $\{x^{(k)}\}$  de vectori  $n$ -dimensionali, care converge la soluția exactă a sistemului. Se alege ca

soluție aproximativă a sistemului un termen  $x^{(s)}$  al șirului, al cărui ordin depinde de precizia impusă.

O iterație presupune efectuarea unui număr de operații aritmetice de ordinul  $n^2$ . Metodele iterative sunt utilizate la rezolvarea sistemelor mari de ecuații. Cele mai cunoscute metode iterative sunt: *Jacobi*, *Gauss–Seidel*, *metodele de relaxare*.

### §1.1. Metoda Gauss. Factorizarea LU

Fie

$$m_r = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ m_{r+1,r} \\ \vdots \\ m_{n,r} \end{pmatrix} \quad \text{și} \quad e_r = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

(elementul 1 din  $e_r$  se află pe linia  $r$ ).

O matrice de forma  $M_r = I_n - m_r \cdot e_r^T$ , unde  $e_r^T = (0, \dots, 1, \dots, 0)$ , se numește *matrice Frobenius*. O astfel de matrice are următoarea structură:

$$M_r = \begin{pmatrix} 1 & & 0 & 0 & \dots & 0 \\ & \ddots & & & & \\ 0 & & 1 & 0 & \dots & \\ 0 & \dots & -m_{r+1,r} & 1 & \dots & 0 \\ & & \vdots & & & \\ 0 & \dots & -m_{nr} & 0 & \dots & 1 \end{pmatrix}$$

De exemplu, dacă  $n=4$  și  $r=2$ , avem:

$$\begin{aligned} M_2 &= \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ m_{32} \\ m_{42} \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & m_{32} & 0 & 0 \\ 0 & m_{42} & 0 & 0 \end{pmatrix} = \\ &= \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -m_{32} & 1 & 0 \\ 0 & -m_{42} & 0 & 1 \end{pmatrix} \end{aligned}$$

**Propoziția 1.** Orice matrice Frobenius  $M_r$  este inversabilă și inversa sa este:

$$M_r^{-1} = I_n + m_r \cdot e_r^T.$$

**Demonstrație.**

$$(I_n - m_r \cdot e_r^T)(I_n + m_r \cdot e_r^T) = I_n - m_r \cdot e_r^T + m_r \cdot e_r^T - m_r(e_r^T m_r)e_r^T.$$

Deoarece  $e_r^T \cdot m_r = 0$ , rezultă:

$$M_r(I_n + m_r \cdot e_r^T) = I_n, \text{ și deci } M_r^{-1} = I_n + m_r \cdot e_r^T. \quad \square$$

**Teorema 1.** Fie  $A$  o matrice pătrată de ordinul  $n$  care satisface condiția:

$$(*) \quad \det \begin{pmatrix} a_{11} & \dots & a_{1r} \\ \vdots & & \vdots \\ a_{r1} & \dots & a_{rr} \end{pmatrix} \neq 0 \text{ pentru orice } r = \overline{1, n-1}.$$

Atunci există o matrice inferior triunghiulară  $M \in M_n(\mathbb{R})$  astfel încât matricea  $U = MA$  este superior triunghiulară.

**Demonstrație.** Deoarece  $a_{11} \neq 0$ , putem considera matricea Frobenius

$$M_1 = \begin{pmatrix} 1 & 0 & \dots & 0 \\ -\frac{a_{21}}{a_{11}} & 1 & 0 & \dots & 0 \\ a_{11} & & & & \\ \vdots & & & & \vdots \\ -\frac{a_{n1}}{a_{11}} & 0 & \dots & & 1 \\ a_{11} & & & & \end{pmatrix}.$$

Dacă notăm  $A_1 = A$  și  $A_2 = M_1 A_1$ , atunci avem

$$A_2 = \begin{pmatrix} a_{11}^{(2)} & a_{12}^{(2)} & \dots & a_{1n}^{(2)} \\ 0 & a_{22}^{(2)} & \dots & a_{2n}^{(2)} \\ \vdots & \vdots & & \vdots \\ 0 & a_{n2}^{(2)} & \dots & a_{nn}^{(2)} \end{pmatrix},$$

unde, notând cu  $a_{ij}^{(1)} = a_{ij}$ , pentru  $i, j = \overline{1, n}$ , avem:  $a_{1j}^{(2)} = a_{1j}^{(1)}$  pentru

$$j = \overline{1, n}; \quad a_{ij}^{(2)} = a_{ij}^{(1)} - \frac{a_{i1}^{(1)} a_{1j}^{(1)}}{a_{11}^{(1)}}, \quad \text{pentru orice } i, j = \overline{2, n}.$$

Observăm că

$$a_{22}^{(2)} = a_{22} - \frac{a_{21} a_{12}}{a_{11}} = \frac{1}{a_{11}} \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \neq 0.$$

Dacă notăm

$$M_2 = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & -\frac{a_{32}^{(2)}}{a_{22}^{(2)}} & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & -\frac{a_{n2}^{(2)}}{a_{22}^{(2)}} & 0 & \dots & 1 \end{pmatrix},$$

atunci

$$A_3 = M_2 A_2 = \begin{pmatrix} a_{11}^{(3)} & a_{12}^{(3)} & a_{13}^{(3)} & \dots & a_{1n}^{(3)} \\ 0 & a_{22}^{(3)} & a_{23}^{(3)} & \dots & a_{2n}^{(3)} \\ 0 & 0 & a_{33}^{(3)} & \dots & a_{3n}^{(3)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & a_{nn}^{(3)} \end{pmatrix},$$

unde  $a_{ij}^{(3)} = a_{ij}^{(2)}$  pentru  $i=1, 2, j=\overline{1, n}$  și  $a_{ij}^{(3)} = a_{ij}^{(2)} - \frac{a_{i2}^{(2)} a_{2j}^{(2)}}{a_{22}^{(2)}}$ ,  $i, j = \overline{3, n}$ .

Un calcul simplu ne arată că

$$a_{33}^{(3)} = \frac{1}{a_{11}^{(1)} a_{22}^{(2)}} \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} \neq 0.$$

În general,  $a_{rr}^{(r)} \neq 0$  și se poate considera matricea Frobenius:

$$M_r = \begin{pmatrix} 1 & & 0 & 0 & \dots & 0 \\ & \ddots & & & & \\ 0 & & 1 & 0 & \dots & \\ 0 & \dots & -\frac{a_{r+1,r}^{(r)}}{a_{rr}^{(r)}} & 1 & \dots & 0 \\ & & \vdots & & \ddots & \\ 0 & \dots & -\frac{a_{nr}^{(r)}}{a_{rr}^{(r)}} & 0 & \dots & 1 \end{pmatrix}.$$

Dacă notăm cu  $A_{r+1} = M_r A_r$ , atunci

$$A_{r+1} = \begin{pmatrix} a_{11}^{(r+1)} & a_{12}^{(r+1)} & \dots & a_{1r}^{(r+1)} & \dots & a_{1n}^{(r+1)} \\ 0 & a_{22}^{(r+1)} & \dots & a_{2r}^{(r+1)} & \dots & a_{2n}^{(r+1)} \\ \vdots & \vdots & & \vdots & & \vdots \\ 0 & 0 & \dots & a_{rr}^{(r+1)} & \dots & a_{rn}^{(r+1)} \\ 0 & 0 & \dots & 0 & a_{r+1,r+1}^{(r+1)} & \dots & a_{r+1,n}^{(r+1)} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 0 & a_{n,r+1}^{(r+1)} & \dots & a_{nn}^{(r+1)} \end{pmatrix},$$

unde  $a_{ij}^{(r+1)} = a_{ij}^{(r)}$ , pentru  $i = \overline{1, r}$ ,  $j = \overline{1, n}$ ,  $a_{ij}^{(r+1)} = a_{ij}^{(r)} - \frac{a_{ir}^{(r)} a_{rj}^{(r)}}{a_{rr}^{(r)}}$ ,

$i, j = \overline{r+1, n}$ .

În final se obține matricea superior triunghiulară

$$U = A_n = M_{n-1} \dots M_2 M_1 A = \begin{pmatrix} a_{11}^{(n)} & a_{12}^{(n)} & \dots & a_{1n}^{(n)} \\ 0 & a_{22}^{(n)} & \dots & a_{2n}^{(n)} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & a_{nn}^{(n)} \end{pmatrix}.$$

Notăm cu  $M = M_{n-1} M_{n-2} \dots M_2 M_1$  și demonstrația teoremei este completă.  $\square$

**Exemplu.**

$$A_1 = A = \begin{pmatrix} 5 & 2 & 1 \\ 5 & -6 & 2 \\ -4 & 2 & 1 \end{pmatrix}, \quad M_1 = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ \frac{4}{5} & 0 & 1 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 5 & 2 & 1 \\ 0 & -8 & 1 \\ 0 & \frac{18}{5} & \frac{9}{5} \end{pmatrix},$$

$$M_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \frac{9}{20} & 1 \end{pmatrix}, \quad U = A_3 = \begin{pmatrix} 5 & 2 & 1 \\ 0 & -8 & 1 \\ 0 & 0 & \frac{9}{4} \end{pmatrix}, \quad M = M_2 M_1 = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ \frac{7}{20} & \frac{9}{20} & 1 \end{pmatrix}.$$

Considerăm sistemul

$$\begin{cases} 5x_1 + 2x_2 + x_3 = 12 \\ 5x_1 - 6x_2 + 2x_3 = -1 \\ -4x_1 + 2x_2 + x_3 = 3 \end{cases},$$

a cărui soluție este  $x_1=1, x_2=2, x_3=3$ . Sub formă matriceală sistemul se scrie:

$Ax=b$ , unde  $b = \begin{pmatrix} 12 \\ -1 \\ 3 \end{pmatrix}$ . Acest sistem este echivalent cu următorul sistem:

$(M_2M_1A)x=(M_2M_1)b$ . Efectuând calculele obținem

$$\begin{cases} 5x_1 + 2x_2 + x_3 = 12 \\ -8x_2 + x_3 = -13 \\ \frac{9}{4}x_3 = \frac{27}{4} \end{cases}.$$

*Numărul operațiilor pentru determinarea matricei  $U$  și a vectorului  $Mb$*

Pentru o linie fixată  $i$  se calculează  $-\frac{a_{ir}^{(r)}}{a_{rr}^{(r)}}$ , apoi se fac înmulțirile cu

$a_{rj}^{(r)}, r+1 \leq j \leq n$ , și se adună  $a_{ij}^{(r)}, r+1 \leq j \leq n$ . La fel și cu  $b_i^{(r+1)}$ . Sunt

$2(n-r)+3$  operații elementare pentru fiecare linie  $i, r+1 \leq i \leq n$ , și pentru fiecare etapă  $r$  vor fi  $(n-r)[2(n-r)+3]$  operații. În total vor fi

$\sum_{r=1}^n [2(n-r)^2 + 3(n-r)] = \frac{2}{3}n^3 + \frac{1}{2}n^2 - \frac{7}{6}n$  operații elementare. Dacă adăugăm și

cele  $n^2$  operații pentru rezolvarea sistemului triunghiular, rezultă că numărul de

operații pentru rezolvarea sistemului  $Ax=b$  este  $\frac{2}{3}n^3 + \frac{3}{2}n^2 - \frac{7}{6}n$ .

În continuare notăm cu  $L_r = M_r^{-1}$ . Din Propoziția 1 rezultă că  $L_r$  este de forma:

$$L_r = \begin{pmatrix} 1 & \dots & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & & \vdots \\ 0 & & 1 & \dots & 0 \\ 0 & & \frac{a_{r+1,r}^{(r)}}{a_{rr}^{(r)}} & \dots & 0 \\ \vdots & & \vdots & & \vdots \\ 0 & \dots & \frac{a_{nr}^{(r)}}{a_{rr}^{(r)}} & 0 & \dots & 1 \end{pmatrix}.$$

Dacă notăm cu  $L=L_1L_2\dots L_{n-1}$ , atunci  $L$  este o matrice inferior triunghiulară de tipul următor

$$L = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ \ell_{21} & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ \ell_{n1} & \ell_{n2} & \ell_{n3} & \dots & 1 \end{pmatrix}.$$

Deoarece  $A = M_1^{-1}M_2^{-1}\dots M_{n-1}^{-1}U$ , rezultă că:

$$A = LU \quad (3)$$

Așadar, orice matrice pătratică ce îndeplinește condiția (\*) din Teorema 1 admite o descompunere unică de forma (3), unde  $L$  este inferior triunghiulară având elementele de pe diagonala principală egale cu 1 și  $U$  este superior triunghiulară. Descompunerea (3) este cunoscută sub numele de *factorizarea LU*.

*Algoritmul pentru factorizarea LU*

{ Determinarea matricelor  $U$  și  $L$  cu păstrarea matricii  $A$  }

Pentru  $i:=1, n$  execută

    Pentru  $j:=1, n$  execută

$u_{ij} := a_{ij}$ ;

        dacă  $i=j$  atunci  $l_{ii}:=1$  altfel  $l_{ij}:=0$ ;

    sfârșit pentru  $j$ ;

sfârșit pentru  $i$ ;

Pentru  $r:=1, n-1$  execută

    Pentru  $i:=r+1, n$  execută

        Pentru  $j:=r+1, n$  execută

$$u_{ij} := u_{ij} - \frac{u_{ir}u_{rj}}{u_{rr}};$$

        sfârșit pentru  $j$ ;

$$l_{ir} := \frac{u_{ir}}{u_{rr}};$$

    sfârșit pentru  $i$ ;

sfârșit pentru  $r$ ;

Pentru  $i:=2, n$  execută

    Pentru  $j:=1, i-1$  execută

$u_{ij}:=0$ ;

    sfârșit pentru  $j$ ;

sfârșit pentru  $i$ .

Algoritmul se află programat în MATLAB și poate fi apelat cu secvența:

$[L, U] = lu(A)$  { se afișează cele două matrice }

În exemplul precedent avem:

$$L_1 = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ -\frac{4}{5} & 0 & 1 \end{pmatrix}, \quad L_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -\frac{9}{20} & 1 \end{pmatrix},$$

$$A=LU = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ -\frac{4}{5} & -\frac{9}{20} & 1 \end{pmatrix} \begin{pmatrix} 5 & 2 & 1 \\ 0 & -8 & 1 \\ 0 & 0 & \frac{9}{4} \end{pmatrix} = \begin{pmatrix} 5 & 2 & 1 \\ 5 & -6 & 2 \\ -4 & 2 & 1 \end{pmatrix}$$

**Observația 1.** Dacă pivotul este “foarte mic”, adică  $|a_{rr}^{(r)}| \ll 1$ , atunci împărțirile la acest pivot produc erori de rotunjire foarte mari, care alterează soluția. În acest caz se recomandă schimbarea pivotului. Se poate alege un nou pivot

$$\pi_r = |a_{ir}^{(r)}| = \max \left\{ |a_{ij}^{(r)}| ; r \leq j \leq n \right\}$$

$$\text{sau } \pi_r = |a_{jr}^{(r)}| = \max \left\{ |a_{k\ell}^{(r)}| ; r \leq k, \ell \leq n \right\}$$

Aceasta presupune schimbarea între ele a două linii și eventual și a două coloane.

*Algoritmul Gauss pentru rezolvarea sistemelor de ecuații liniare*

Pentru  $r:=1, n-1$  execută

Pentru  $i:=r+1, n$  execută

Pentru  $j:=r+1, n$  execută

găsește pivotul conform cu (β);

schimbă linia  $i$  cu linia pivotului și coloana  $j$  cu coloana pivotului dacă este cazul;

sfârșit pentru  $j$

sfârșit pentru  $i$

Pentru  $i:=r+1, n$  execută

$$b_i = b_i - \frac{a_{ir} b_r}{a_{rr}}$$

Pentru  $j:=r+1, n$  execută

$$a_{ij} = a_{ij} - \frac{a_{ir} a_{rj}}{a_{rr}} ;$$

sfârșit pentru  $j$ ;

sfârșit pentru  $i$ ;

sfârșit pentru  $r$ ;

$$x_n := \frac{b_n}{a_{nn}} ;$$

Pentru  $i:=n-1,1,-1$  execută  
 $s:=0$  ;  
 Pentru  $j:=i+1,n$  execută  
 $s:=s+a_{ij}x_j$  ;  
 sfârșit pentru  $j$  ;  
 $x_i := \frac{(b_i - s)}{a_{ii}}$  ;  
 sfârșit pentru  $i$  .

## §1.2. Matrice simetrice pozitiv definite

Reamintim că o matrice simetrică se numește pozitiv definită, dacă forma pătratică asociată este pozitiv definită. Mai precis, dacă  $A$  este o matrice simetrică, atunci  $A$  se numește *pozitiv definită* dacă

$$\varphi(x)=x^T Ax > 0 ,$$

pentru orice  $x \neq 0$ , unde  $x = (x_1, x_2, \dots, x_n)^T$ .

Din Algebra Liniară, se știe că o matrice simetrică  $A$ , este pozitiv definită dacă și numai dacă  $\Delta_r > 0$  pentru orice  $r = \overline{1, n}$ , unde

$$\Delta_r = \det \begin{pmatrix} a_{11} & \dots & a_{1r} \\ \vdots & & \vdots \\ a_{r1} & \dots & a_{rr} \end{pmatrix}.$$

În practică aceste condiții sunt greu de verificat pentru matrice de dimensiuni mari. De aceea, în continuare vom prezenta unele condiții necesare, respectiv și suficiente, pentru ca o matrice simetrică să fie pozitiv definită.

**Propoziția 1.** Dacă  $A$  este o matrice simetrică pozitiv definită, atunci:

- (a)  $a_{ii} > 0$  pentru orice  $i = \overline{1, n}$ ,
- (b)  $a_{ii}a_{jj} > a_{ij}^2$  pentru orice  $i, j = \overline{1, n}$ .

**Demonstrație.**

$$\begin{aligned} \varphi(x) = x^T Ax &= (x_1, \dots, x_n) \cdot \begin{pmatrix} a_{11}x_1 + \dots + a_{1n}x_n \\ \vdots \\ a_{n1}x_1 + \dots + a_{nn}x_n \end{pmatrix} = (a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n)x_1 + \\ &+ (a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n)x_2 + \dots + (a_{n1}x_1 + a_{n2}x_2 + \dots + a_{nn}x_n)x_n \end{aligned}$$

Ținând seama că  $a_{ij} = a_{ji}$ , în continuare avem

$$\begin{aligned} \varphi(x) = \sum_{i=1}^n \sum_{j=1}^n a_{ij} x_i x_j = & a_{11} x_1^2 + 2a_{12} x_1 x_2 + \dots + 2a_{1n} x_1 x_n + \\ & + a_{22} x_2^2 + \dots + 2a_{2n} x_2 x_n + \\ & \vdots \\ & + a_{nn} x_n^2 \end{aligned}$$

În particular, pentru  $x = e_i = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix}$  avem  $\varphi(e_i) = a_{ii}$ . Cum  $\varphi$  este pozitiv definită și

$e_i \neq 0$ , rezultă că  $a_{ii} = \varphi(e_i) > 0$ , adică (a).

Pentru un număr real oarecare  $\lambda$  avem

$$\varphi(\lambda e_i + e_j) = a_{ii} \lambda^2 + 2a_{ij} \lambda + a_{jj} > 0. \quad (1)$$

Pentru ca inegalitatea (1) să fie adevărată pentru orice  $\lambda \in \mathbb{R}$ , trebuie ca

$$\Delta = 4(a_{ij}^2 - a_{ii} a_{jj}) < 0.$$

Așadar am demonstrat că  $a_{ij}^2 < a_{ii} a_{jj}$  pentru orice  $i, j = \overline{1, n}$ , adică (b).  $\square$

**Observația 2.** Condițiile care apar în Propoziția 1 sunt doar necesare nu și suficiente.

**Exemplu.**

Matricea  $A = \begin{pmatrix} 3 & 2 & -2 \\ 2 & 3 & 2 \\ -2 & 2 & 3 \end{pmatrix}$  satisface condițiile din Propoziția 1, dar nu este

pozitiv definită.

Într-adevăr,

$$\varphi(x) = 3(x_1^2 + x_2^2 + x_3^2) + 4(x_1 x_2 - x_1 x_3 + x_2 x_3).$$

Dacă  $x = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}$ , atunci  $\varphi(x) = 9 - 12 = -3 < 0$ , deci  $\varphi$  nu este pozitiv definită.

**Definiția 1.** Spunem că matricea  $A$  este tare diagonal dominantă dacă elementele sale satisfac inegalitățile:

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^n |a_{ij}|, \quad i = \overline{1, n}. \quad (d)$$

Dacă inegalitățile (d) devin egalități pentru anumiți indici, dar nu pentru toți, matricea se numește slab diagonal dominantă.

**Teorema 1.** Fie  $A$  o matrice simetrică cu următoarele proprietăți:

- (i)  $A$  este tare diagonal dominantă,
- (ii)  $a_{ii} > 0$  pentru  $i = \overline{1, n}$ .

Atunci  $A$  este pozitiv definită.

**Demonstrație.**

Din condiția (i) rezultă că dacă  $x \neq 0$ , atunci:

$$\begin{aligned} \varphi(x) &= \sum_{i=1}^n a_{ii} x_i^2 + \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n a_{ij} x_i x_j > \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n |a_{ij}| x_i^2 - \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n |a_{ij}| \cdot |x_i| \cdot |x_j| = \\ &= \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n |a_{ij}| \cdot |x_i| \cdot (|x_i| - |x_j|) \end{aligned}$$

Deoarece  $a_{ij} = a_{ji}$  avem și inegalitatea:

$$\varphi(x) > \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n |a_{ij}| \cdot |x_j| (|x_j| - |x_i|).$$

Adunând cele două inegalități rezultă

$$2\varphi(x) > \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n |a_{ij}| \cdot (|x_i| - |x_j|)^2 \geq 0.$$

Așadar,  $\varphi(x) > 0$  pentru orice  $x \neq 0$ , deci  $\varphi$  este pozitiv definită.  $\square$

**Definiția 2.** Fie  $M = \{1, 2, \dots, n\}$ . O matrice  $A$  se numește reductibilă dacă există două submulțimi  $S, T \subset M$  cu proprietățile:

- (i)  $S \neq \emptyset, T \neq \emptyset$
- (ii)  $S \cap T = \emptyset$
- (iii)  $S \cup T = M$
- (iv)  $a_{ij} = 0$  pentru orice  $i \in S$  și  $j \in T$ .

Matricea  $A$  se numește ireductibilă dacă oricare ar fi submulțimile  $S$  și  $T$  ale lui  $M$  cu proprietățile (i)–(iii), există  $i_0 \in S$  și  $j_0 \in T$  astfel încât  $a_{i_0 j_0} \neq 0$ .

Cel mai simplu exemplu de matrice reductibilă este matricea diagonală.

**Teorema 2.** Fie  $A$  o matrice simetrică având următoarele proprietăți:

- (i)  $A$  este slab diagonal dominantă,
- (ii)  $A$  este ireductibilă,
- (iii)  $a_{ii} > 0$  pentru orice  $i = \overline{1, n}$ .

Atunci  $A$  este pozitiv definită.

**Demonstrație.** Procedând ca în demonstrația Teoremei 1, rezultă:

$$\varphi(x) \geq \frac{1}{2} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n |a_{ij}| \cdot (|x_i| - |x_j|)^2 \geq 0.$$

Vom arăta că situația  $\varphi(x)=0$  pentru  $x \neq 0$  nu poate avea loc. Într-adevăr,  $\varphi(x)$  se anulează în următoarele cazuri:

1)  $a_{ij} = 0$  pentru orice  $i \neq j$ . Atunci matricea  $A$  are forma diagonală și este reductibilă.

2)  $|x_i| = |x_j| = \alpha \neq 0$  pentru orice  $i$  și  $j$ .

$$\varphi(x) = \sum_{i=1}^n a_{ii} \alpha^2 + \sum_{\substack{i=1 \\ j \neq i}}^n \sum_{j=1}^n a_{ij} x_i x_j \geq \sum_{i=1}^n (a_{ii} - \sum_{\substack{j=1 \\ j \neq i}}^n |a_{ij}|) \cdot \alpha^2 \geq 0.$$

Cum există cel puțin un indice  $i_0$  astfel încât  $\sum_{\substack{j=1 \\ j \neq i_0}}^n |a_{i_0 j}| < a_{i_0 i_0}$ , rezultă

$\varphi(x) > 0$  pentru  $x \neq 0$ .

3)  $a_{ij} = 0$  pentru orice pereche de indici  $(i, j)$  pentru care  $|x_i| \neq |x_j|$  și  $a_{ij} \neq 0$  dacă  $|x_i| = |x_j| \neq 0$ .

Fie  $M = \{1, 2, \dots, n\}$  și  $S = \{i, j \in M; |x_i| = |x_j| \neq 0\}$ .

Dacă  $S = M$ , atunci suntem în cazul 2).

Dacă  $S = \emptyset$ , atunci  $|x_i| \neq |x_j|$  pentru orice  $i$  și  $j$  și evident  $\varphi(x) > 0$  pentru  $x \neq 0$ .

Așadar, putem presupune că  $\emptyset \neq S \subset M$  (incluziune strictă). Dacă notăm cu  $T = M \setminus S$  atunci  $S$  și  $T$  satisfac condițiile (i)–(iv) din Definiția 2, deci  $A$  este reductibilă.  $\square$

### Exemplu.

Fie

$$A = \begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{pmatrix}.$$

Matricea  $A$  este simetrică, slab diagonal dominantă, ireductibilă și are elementele de pe diagonala principală strict pozitive. Din Teorema 2 rezultă că  $A$  este pozitiv definită.

**Oservația 2.** Teorema 2 este utilă la stabilirea faptului că anumite matrice care apar în rezolvarea numerică a ecuațiilor cu derivate parțiale de tip eliptic sunt pozitiv definite.

### §1.3. Metoda Cholesky

Fie  $A$  o matrice simetrică, pozitiv definită și

$$\varphi(x) = x^T Ax = \sum_{i=1}^n \sum_{j=1}^n a_{ij} x_i x_j$$

forma pătratică asociată. Deoarece  $a_{11} > 0$  avem:

$$\begin{aligned} \varphi(x) = a_{11}x_1^2 + 2a_{12}x_1x_2 + \dots + 2a_{1n}x_1x_n + \sum_{i=2}^n \sum_{j=2}^n a_{ij}x_ix_j = & \left( \sqrt{a_{11}}x_1 + \sum_{j=2}^n \frac{a_{1j}}{\sqrt{a_{11}}}x_j \right)^2 + \\ & + \sum_{i=2}^n \sum_{j=2}^n a_{ij}^{(1)}x_ix_j, \text{ unde } a_{ij}^{(1)} = a_{ij} - \frac{a_{1i}a_{1j}}{a_{11}}, \quad i, j = \overline{2, n} \end{aligned}$$

Dacă notăm cu

$$\varphi_1(x) = \sum_{i=2}^n \sum_{j=2}^n a_{ij}^{(1)}x_ix_j,$$

atunci  $\varphi_1$  este la rândul său o formă pătratică pozitiv definită.

Într-adevăr, să presupunem prin absurd că există  $z = \begin{pmatrix} z_2 \\ \vdots \\ z_n \end{pmatrix} \neq 0$  astfel încât

$$\varphi_1(z) \leq 0.$$

$$\text{Fie } z_1 = -\sum_{j=2}^n \frac{a_{1j}}{a_{11}}z_j \text{ și } \bar{z} = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix}.$$

În continuare avem

$$\varphi(\bar{z}) = \left( \sqrt{a_{11}}z_1 + \sum_{j=2}^n \frac{a_{1j}}{\sqrt{a_{11}}}z_j \right)^2 + \varphi_1(z) = 0 + \varphi_1(z) \leq 0,$$

ceea ce contrazice faptul că  $\varphi$  este pozitiv definită.

Așadar, am demonstrat că  $\varphi_1$  este pozitiv definită. În particular, rezultă că  $a_{22}^{(1)} > 0$ . Mai departe procedăm cu  $\varphi_1$  așa cum am procedat cu  $\varphi$  și obținem

$$\varphi_1(x) = \left( \sqrt{a_{22}^{(1)}} x_2 + \sum_{j=3}^n \frac{a_{2j}^{(1)}}{\sqrt{a_{22}^{(1)}}} x_j \right)^2 + \varphi_2(x) \quad ,$$

unde

$$\varphi_2(x) = \sum_{i=3}^n \sum_{j=3}^n a_{ij}^{(2)} x_i x_j$$

este pozitiv definită. În final  $\varphi(x)$  se reprezintă ca o sumă de pătrate. Mai precis  $\varphi(x)$  admite următoarea scriere:

$$\varphi(x) = \sum_{i=1}^n \left( \sqrt{a_{ii}^{(i-1)}} x_i + \sum_{j=i+1}^n \frac{a_{ij}^{(i-1)}}{\sqrt{a_{ii}^{(i-1)}}} x_j \right)^2 ,$$

unde

$$a_{ij}^{(0)} = a_{ij} \quad \text{și} \quad a_{ij}^{(p)} = a_{ij}^{(p-1)} - \frac{a_{pi}^{(p-1)} a_{pj}^{(p-1)}}{a_{pp}^{(p-1)}} , \quad p = \overline{1, n-1} .$$

Introducem notațiile:

$$\begin{aligned} r_{ii} &= \sqrt{a_{ii}^{(i-1)}} , \quad i = \overline{1, n} \\ r_{ij} &= \frac{a_{ij}^{(i-1)}}{r_{ii}} , \quad i < j \\ r_{ij} &= 0, \quad j < i \\ a_{ij}^{(p)} &= a_{ij}^{(p-1)} - r_{pi} r_{pj} , \quad p = \overline{1, n-1} , \quad i, j = \overline{p+1, n} . \end{aligned} \quad (1)$$

Cu aceste notații avem

$$\begin{aligned} \varphi(x) &= \sum_{i=1}^n \left( \sum_{j=i}^n r_{ij} x_j \right)^2 = (r_{11} x_1 + r_{12} x_2 + \dots + r_{1n} x_n)^2 + (r_{22} x_2 + \dots + r_{2n} x_n)^2 + \\ &\quad + \dots + (r_{nn} x_n)^2 \end{aligned}$$

Dacă notăm cu  $R$  următoarea matrice superior triunghiulară

$$R = \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ 0 & r_{22} & \dots & r_{2n} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & r_{nn} \end{pmatrix} ,$$

atunci  $\varphi(x) = (x^T R^T)(R x) = x^T (R^T R) x$ . Pe de altă parte,  $\varphi(x) = x^T A x$ . Se obține astfel următoarea descompunere a matricei  $A$

$$A = R^T R \quad (2)$$

unde  $R$  este o matrice superior triunghiulară.

Descompunerea (2) poartă numele de *factorizarea Cholesky* a matricei  $A$  și are loc pentru matrice simetrice pozitiv definite.

*Numărul de operații pentru determinarea matricei  $R$*

Pentru a calcula elementele liniei a  $i$ -a a matricei  $R$  sunt necesare  $(n-i)(2i-1)+2i-2$  operații elementare și o extragere de rădăcină pătrată. Pentru toate liniile sunt necesare

$$\sum_{i=1}^n [(n-i)(2i-1)+2i-2] = \frac{n^3}{3} + \frac{n^2}{2} - \frac{5n}{6}$$

operații elementare plus  $n$  extrageri de rădăcină pătrată.

**Exemplu.** Să se determine descompunerea Cholesky a matricei

$$A = \begin{pmatrix} 3 & 2 & 2 \\ 2 & 3 & 2 \\ 2 & 2 & 3 \end{pmatrix}$$

$$r_{11} = \sqrt{3}, \quad r_{12} = \frac{2}{\sqrt{3}}, \quad r_{13} = \frac{2}{\sqrt{3}}, \quad a_{22}^{(1)} = a_{22} - r_{12}^2 = \frac{5}{3}, \quad r_{22} = \sqrt{\frac{5}{3}},$$

$$a_{23}^{(1)} = a_{23} - r_{12}r_{13} = \frac{2}{3}, \quad r_{23} = \frac{a_{23}^{(1)}}{r_{22}} = \frac{2}{\sqrt{15}}, \quad a_{33}^{(1)} = a_{33} - r_{13}^2 = \frac{5}{3},$$

$$a_{33}^{(2)} = a_{33}^{(1)} - r_{23}^2 = \frac{7}{5}, \quad r_{33} = \sqrt{\frac{7}{5}}$$

$$R = \begin{pmatrix} \sqrt{3} & \frac{2}{\sqrt{3}} & \frac{2}{\sqrt{3}} \\ 0 & \sqrt{\frac{5}{3}} & \frac{2}{\sqrt{15}} \\ 0 & 0 & \sqrt{\frac{7}{5}} \end{pmatrix}.$$

Se verifică imediat că  $A = R^T R$ .

Rezolvarea sistemului  $Ax=b$  cu metoda Cholesky, în cazul când matricea  $A$  este simetrică și pozitiv definită, revine la rezolvarea a două sisteme triunghiulare și anume

$$\begin{cases} R^T y = b \\ Rx = y \end{cases}$$

*Algoritmul Cholesky pentru rezolvarea sistemelor de ecuații liniare*

Pentru  $p:=1, n-1$  execută

$r_{pp} := \sqrt{a_{pp}} ;$   
 Pentru  $k:=p+1, n-1$  execută  

$$r_{pk} := \frac{a_{pk}}{r_{pp}} ;$$
 sfârșit pentru  $k ;$   
 Pentru  $i:=p+1, n$  execută  
     Pentru  $j:=i, n$  execută  
          $a_{ij} := a_{ij} - r_{pi} r_{pj} ;$   
         sfârșit pentru  $j ;$   
     sfârșit pentru  $i ;$   
 sfârșit pentru  $p ;$   
 { Rezolvarea sistemului  $R^T y = b$  }  

$$y_1 = \frac{b_1}{r_{11}} ;$$
 Pentru  $i:=2, n$  execută  
      $s := 0 ;$   
     Pentru  $j:=1, i$  execută  
          $s := s + r_{ij} y_j ;$   
     sfârșit pentru  $j ;$   
     
$$y_i := \frac{b_i - s}{r_{ii}} ;$$
 sfârșit pentru  $i ;$   
 { Rezolvarea sistemului  $Rx = y$  }  

$$x_n = \frac{y_n}{r_{nn}} ;$$
 Pentru  $i:=n-1, 1$  execută  
      $s := 0 ;$   
     Pentru  $j:=i+1, n$  execută  
          $s := s + r_{ij} x_j ;$   
     sfârșit pentru  $j ;$   
     
$$x_i := \frac{y_i - s}{r_{ii}} ;$$
 sfârșit pentru  $i .$

Algoritmul se află programat și în MATLAB și se apelează cu secvența:

$R = \text{chol}(A);$

$x = R \setminus R \setminus b$  { pentru afișarea soluției }

### §1.4. Metoda Householder. Factorizarea QR

O matrice Householder este o matrice de forma  $H = I_n - 2hh^T$ , unde  $h^T = (0, \dots, 0, h_i, \dots, h_n)$  și  $\|h\|_2 = \sqrt{h_i^2 + \dots + h_n^2} = 1$ . Se observă imediat că o matrice Householder este simetrică și are următoarea structură:

$$H = \begin{pmatrix} 1 & & & & & & & & 0 \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ 0 & & & 1 - 2h_i^2 & -2h_i h_{i+1} & \dots & -2h_i h_n \\ & & & \vdots & \vdots & & \vdots \\ & & & -2h_n h_i & -2h_n h_{i+1} & \dots & 1 - 2h_n^2 \end{pmatrix}$$

Mai mult, constatăm că  $H$  este ortogonală. Într-adevăr,

$$H^2 = (I_n - 2hh^T)(I_n - 2hh^T) = I_n - 2hh^T - 2hh^T + 4h(h^T h)h^T.$$

Cum  $h^T h = 1$ , rezultă  $H^2 = I_n$ . Așadar, avem  $H^{-1} = H = H^T$ .

Un calcul simplu ne arată că  $(hh^T)x = (h^T x)h$ , pentru orice  $x = (x_1, x_2, \dots, x_n)^T$ .

În continuare ne punem următoarea problemă: *dat fiind un vector coloană  $x \neq 0$ , se poate determina o matrice Householder  $H$ , astfel încât  $Hx$  să fie colinear cu  $e_1$  ? ( unde  $e_1^T = (1, 0, \dots, 0)$  ).*

Cu alte cuvinte, se poate determina un vector coloană  $h$ , cu  $\|h\|_2 = 1$  și un număr real  $\sigma$  astfel încât  $Hx = x - 2hh^T x = \sigma e_1$  ?

Ținând seama de observația de mai sus, aceasta revine la  $x - 2(h^T x)h = \sigma e_1$ , de unde rezultă  $x - \sigma e_1 = 2(h^T x)h$ . Așadar,  $h$  trebuie să fie colinear cu  $x - \sigma e_1$ . Cum  $\|h\|_2 = 1$  rezultă

$$h = \frac{x - \sigma e_1}{\|x - \sigma e_1\|_2}. \quad (1)$$

Pe de altă parte,  $H$  fiind ortogonală avem

$$\|x\|_2 = \|Hx\|_2 = |\sigma| \cdot \|e_1\| = |\sigma|.$$

Alegem  $\sigma = -\operatorname{sgn}(x_1) \|x\|_2$  și facem convenția  $\operatorname{sgn}(x_1) = 1$  dacă  $x_1 = 0$ .

În continuare avem

$$x - \sigma \cdot e_1 = \begin{pmatrix} x_1 + \operatorname{sgn}(x_1) \cdot \|x\|_2 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} (|x_1| + \|x\|_2) \cdot \operatorname{sgn}(x_1) \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \text{ și}$$

$$\|x - \sigma e_1\|_2^2 = 2\|x\|_2^2 + 2|x_1| \cdot \|x\|_2 = 2\|x\|_2 ( \|x\|_2 + |x_1| ) .$$

Înlocuind în (1) obținem

$$h = \frac{1}{\sqrt{2\|x\|_2 ( \|x\|_2 + |x_1| )}} \begin{pmatrix} (|x_1| + \|x\|_2) \operatorname{sgn} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} . \quad (2)$$

Se obține astfel următorul algoritm pentru determinarea lui  $h$  și deci a matricei  $H$ :

$$\begin{aligned} H &= I_n - \beta u u^T \\ \beta &= \left( \|x\|_2 ( \|x\|_2 + |x_1| ) \right)^{-1} \\ u &= \left( (|x_1| + \|x\|_2) \operatorname{sgn}(x_1), x_2, \dots, x_n \right)^T \\ \operatorname{sgn}(x_1) &= 1 \text{ dacă } x_1 = 0. \end{aligned} \quad (3)$$

**Teorema 1.** Pentru orice matrice  $A \in M_n(\mathbb{R})$  nesingulară există o matrice ortogonală  $H$  astfel încât matricea  $R = HA$  este superior triunghiulară.

**Demonstrație.**

Fie  $a_1 = \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{pmatrix}$ , prima coloană a matricei  $A$ . Din cele arătate mai înainte rezultă

că există o matrice Householder  $H_1$  astfel încât  $H_1 a_1 = \sigma e_1$ . Matricea  $H_1$  se determină astfel:

$$s = \left( \sum_{j=1}^n a_{j1}^2 \right)^{1/2}, \quad \beta = (s(s + |a_{11}|))^{-1}, \quad u = \left( (|a_{11}| + s) \operatorname{sgn}(a_{11}), a_{21}, \dots, a_{n1} \right)^T,$$

$$\operatorname{sgn}(a_{11}) = 1 \text{ dacă } a_{11} = 0, \quad H_1 = I_n - \beta u u^T. \quad (4)$$

Dacă notăm cu  $A_1 = H_1 A$ , atunci  $A_1$  are următoarea formă:

$$A_1 = \begin{pmatrix} -\operatorname{sgn}(a_{11})s & a_{12}^{(1)} & \dots & a_{1n}^{(1)} \\ 0 & a_{22}^{(1)} & \dots & a_{2n}^{(1)} \\ \vdots & \vdots & \dots & \vdots \\ 0 & a_{n2}^{(1)} & \dots & a_{nn}^{(1)} \end{pmatrix}$$

În continuare considerăm vectorul  $a_2^{(1)} = \begin{pmatrix} a_{22}^{(1)} \\ \vdots \\ a_{n2}^{(1)} \end{pmatrix}$  și determinăm o matrice

ortogonală  $\tilde{H}_2 \in M_{n-1}(\mathbb{R})$  astfel încât

$$\tilde{H}_2 a_2^{(1)} = \sigma \cdot \tilde{e}_1,$$

unde  $\tilde{e}_1^T = (1, 0, \dots, 0) \in \mathbb{R}^{n-1}$ .

Notăm cu  $H_2 = \begin{pmatrix} 1 & 0 \\ 0 & \tilde{H}_2 \end{pmatrix} \in M_n(\mathbb{R})$  și cu  $A_2 = H_2 A_1$ . Matricea  $A_2$  va arăta astfel

$$A_2 = \begin{pmatrix} a_{11}^{(2)} & a_{12}^{(2)} & a_{13}^{(2)} & \dots & a_{1n}^{(2)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \dots & a_{2n}^{(2)} \\ 0 & 0 & a_{33}^{(2)} & \dots & a_{3n}^{(2)} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & a_{n3}^{(2)} & \dots & a_{nn}^{(2)} \end{pmatrix}, \text{ unde } a_{1j}^{(2)} = a_{1j}^{(1)}, j = \overline{1, n}.$$

În continuare se determină o matrice Householder  $\tilde{H}_3 \in M_{n-2}(\mathbb{R})$  cu proprietatea că  $\tilde{H}_3 a_3^{(2)} = \sigma \tilde{e}_1$ , unde  $\tilde{e}_1^T = (1, 0, \dots, 0) \in M_{n-2}(\mathbb{R})$ . Vom nota cu

$H_3 = \begin{pmatrix} I_2 & 0 \\ 0 & \tilde{H}_3 \end{pmatrix} \in M_n(\mathbb{R})$  și cu  $A_3 = H_3 A_2$ . Matricea  $A_3$  va avea toate elementele de

sub diagonala principală, din primele trei coloane, zero. Procedul continuă într-un mod evident. În final, obținem o matrice superior triunghiulară  $A_{n-1} = H_{n-1} A_{n-2} = \dots = H_{n-1} \dots H_2 H_1 A$ . Dacă notăm  $H = H_{n-1} \dots H_2 H_1$  și cu  $R = HA$ , atunci  $H$  este ortogonală și  $R$  superior triunghiulară.  $\square$

**Corolar.** Pentru orice matrice nesingulară  $A \in M_n(\mathbb{R})$  există o matrice ortogonală  $Q$  și o matrice superior triunghiulară  $R$  astfel încât  $A = QR$ .

*Algoritmul Householder pentru rezolvarea sistemelor de ecuații liniare*

Fie sistemul  $Ax = b$  cu  $A \in M_n(\mathbb{R})$ . Notăm cu  $C = (A|b) = (c_{ij}) \in M_{n, n+1}(\mathbb{R})$  matricea extinsă.

Pentru  $i = 1, n-1$  execută

$$s := \sqrt{\sum_{j=i}^n c_{ji}^2};$$

dacă  $s = 0$  atunci  $A$  este singulară. Stop!

altfel  $\beta := (s(|c_{ii}| + s))^{-1}$ ; dacă  $c_{ii} = 0$  atunci  $\text{sgn}(c_{ii}) := 1$ ;

$$u := (0, \dots, 0, (c_{ii} + s) \cdot \text{sgn}(c_{ii}), c_{i+1,i}, \dots, c_{ni})^T;$$

$$H_i = I_n - \beta uu^T; \quad C := H_i C;$$

sfârșit pentru  $i$ ;

**Exemplu.** Fie sistemul

$$\begin{cases} 5x_1 + 2x_2 + x_3 = 12 \\ 5x_1 - 6x_2 + 2x_3 = -1 \\ -4x_1 + 2x_2 + x_3 = 3 \end{cases}$$

Soluția exactă este  $x_1 = 1, x_2 = 2, x_3 = 3$ .

Aplicăm metoda Householder.

$$A = \begin{pmatrix} 5 & 2 & 1 \\ 5 & -6 & 2 \\ -4 & 2 & 1 \end{pmatrix}; \quad b = \begin{pmatrix} 12 \\ -1 \\ 3 \end{pmatrix}, \quad C = \begin{pmatrix} 5 & 2 & 1 & 12 \\ 5 & -6 & 2 & -1 \\ -4 & 2 & 1 & 3 \end{pmatrix};$$

Iterația I

$$a_1 = \begin{pmatrix} 5 \\ 5 \\ -4 \end{pmatrix}; \quad c_{11} = 5; \quad s = \sqrt{66} = 8.124038405;$$

$$\beta = \frac{1}{8.124038405 \cdot 13.124038405} = 9.379086466 \cdot 10^{-3}; \quad u = \begin{pmatrix} 13.124038405 \\ 5 \\ -4 \end{pmatrix};$$

$$H_1 = \begin{pmatrix} -0.615457455 & -0.615457455 & 0.492365964 \\ -0.615457455 & 0.765522838 & 0.187581729 \\ 0.492365964 & 0.187581729 & 0.849934617 \end{pmatrix}$$

$$A_1 = H_1 A = \begin{pmatrix} -8.12403840 & 3.44656174 & -1.35400640 \\ 0 & -5.44888848 & 1.10316995 \\ 0 & 1.55911078 & 1.71746403 \end{pmatrix};$$

$$b_1 = H_1 b = \begin{pmatrix} -5.292934112 \\ -7.588267109 \\ 8.270613687 \end{pmatrix}; \quad C = H_1 \cdot C = [A_1 | b_1]$$

Iterația a II-a

$$a_2^{(1)} = \begin{pmatrix} -5.448888481 \\ 1.559110785 \end{pmatrix}; \quad s = \sqrt{c_{22}^2 + c_{32}^2} = 5.667557862;$$

$$c_{22} = -5.448888481; \quad \beta = 0.015872234; \quad u = \begin{pmatrix} 0 \\ -11.116446343 \\ 1.559110785 \end{pmatrix};$$

$$H_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -0.961417354 & 0.275093933 \\ 0 & 0.275093933 & 0.961417354 \end{pmatrix};$$

$$A_2 = H_2 A_1 = \begin{pmatrix} -8.124038405 & 3.446561747 & -1.354006401 \\ 0 & 5.667557862 & -0.588142797 \\ 0 & 0 & 1.954675092 \end{pmatrix};$$

$$b_2 = H_2 b_1 = \begin{pmatrix} -5.292934112 \\ 9.570687333 \\ 5.864025277 \end{pmatrix}; \quad C = H_2 \cdot C = [A_2 | b_2];$$

$$H = H_2 \cdot H_1 = \begin{pmatrix} -0.615457455 & -0.615457455 & 0.492365964 \\ 0.727158367 & -0.684384346 & 0.053467527 \\ 0.30406057 & 0.390935018 & 0.868744486 \end{pmatrix};$$

$$R = H \cdot A = \begin{pmatrix} -8.124038405 & 3.446561747 & -1.354006401 \\ 0 & 5.667557862 & -0.588142797 \\ 0 & 0 & 1.954675092 \end{pmatrix};$$

Soluția sistemului inițial este  $x = R^{-1} H b$ , unde:

$$R^{-1} = \begin{pmatrix} -0.123091491 & 0.074854538 & -0.062742657 \\ 0 & 0.176442839 & 0.053089941 \\ 0 & 0 & 0.511593975 \end{pmatrix}.$$

Se obține soluția  $x_1=1$  ;  $x_2=2$  ;  $x_3=3.000000001$ .

## §1.5. Norme de matrice

Cele mai utilizate norme vectoriale pe  $\mathbb{R}^n$  sunt:

$$1) \quad \|x\|_\infty = \max \{ |x_1|, |x_2|, \dots, |x_n| \}$$

$$2) \quad \|x\|_p = \left( \sum_{i=1}^n |x_i|^p \right)^{1/p}, \quad 1 \leq p < \infty$$

unde  $x^T = (x_1, x_2, \dots, x_n) \in \mathbb{R}^n$ .

**Definiția 1.** Se numește normă de matrice orice aplicație

$$A \rightarrow \|A\|_M : M_n(\mathbb{R}) \rightarrow \mathbb{R}_+$$

cu proprietățile:

- (i)  $\|A\|_M = 0$  dacă și numai dacă  $A = 0$ ,
- (ii)  $\|\lambda A\|_M = |\lambda| \|A\|_M$ , ;  $\lambda \in \mathbb{R}$ ,  $A \in M_n(\mathbb{R})$ ,
- (iii)  $\|A + B\|_M \leq \|A\|_M + \|B\|_M$ ,
- (iv)  $\|AB\|_M \leq \|A\|_M \cdot \|B\|_M$ ,  $A, B \in M_n(\mathbb{R})$ .

Un exemplu de normă de matrice este *norma euclidiană* de matrice, care se definește astfel

$$\|A\|_E = \left( \sum_{i=1}^n \sum_{j=1}^n a_{ij}^2 \right)^{1/2}. \quad (1)$$

Proprietățile (i) și (ii) sunt evidente. Pentru a demonstra proprietățile (iii) și (iv) se folosește inegalitatea Cauchy–Buniakovski–Schwarz pe  $\mathbb{R}^n$ . Pentru exemplificare demonstrăm (iv). Fie  $C = AB$ . Atunci

$$c_{ij}^2 = \left( \sum_{k=1}^n a_{ik} b_{kj} \right)^2 \leq \left( \sum_{k=1}^n a_{ik}^2 \right) \left( \sum_{k=1}^n b_{kj}^2 \right)$$

În continuare avem

$$\|AB\|_E^2 = \sum_{i=1}^n \sum_{j=1}^n c_{ij}^2 \leq \sum_{i=1}^n \left( \sum_{k=1}^n a_{ik}^2 \right) \cdot \sum_{j=1}^n \left( \sum_{k=1}^n b_{kj}^2 \right) = \|A\|_E^2 \cdot \|B\|_E^2,$$

de unde rezultă  $\|AB\|_E \leq \|A\|_E \cdot \|B\|_E$ .

**Definiția 2.** O normă de matrice  $\|\cdot\|_M$  se numește *compatibilă cu norma vectorială*  $\|\cdot\|_p$  dacă  $\|Ax\|_p \leq \|A\|_M \|x\|_p$  pentru orice  $x$ .

**Observația 1.**  $\|Ax\|_2 \leq \|A\|_E \|x\|_2$ ,  $(\forall) x$ .

Într-adevăr,  $\|Ax\|_2^2 = \sum_{i=1}^n (a_{i1}x_1 + \dots + a_{in}x_n)^2 \leq \sum_{i=1}^n \left( \sum_{j=1}^n a_{ij}^2 \right) \left( \sum_{j=1}^n x_j^2 \right) = \|A\|_E^2 \cdot \|x\|_2^2$ .

**Observația 2.** Dacă  $\lambda$  este o valoare proprie a matricei  $A$ , atunci  $|\lambda| \leq \|A\|_M$  pentru orice normă de matrice compatibilă cu o normă vectorială.

Într-adevăr, fie  $v$  un vector propriu al matricei  $A$  care corespunde valorii proprii  $\lambda$ . Atunci avem

$$|\lambda| \cdot \|v\| = \|\lambda v\| = \|Av\| \leq \|A\|_M \|v\|,$$

deci  $|\lambda| \leq \|A\|_M$ .

După cum se știe, între mulțimea  $M_n(\mathbb{R})$  a matricelor pătratice cu elemente din  $\mathbb{R}$  și mulțimea  $L(\mathbb{R}^n)$  a aplicațiilor liniare și continue,  $U: \mathbb{R}^n \rightarrow \mathbb{R}^n$ , există o corespondență bijectivă. Mai precis, dacă  $A$  este matricea asociată transformării liniare  $U$ , atunci  $U(e_i^T) = (a_{1i}, a_{2i}, \dots, a_{ni})$ , unde  $e_i^T = (0, \dots, 1, \dots, 0) \in \mathbb{R}^n$  și  $U(x^T) = (Ax)^T$ . Pe de altă parte, spațiul  $L(\mathbb{R}^n)$  este un spațiu normat în raport cu norma operatorială:

$$\|U\|_0 = \sup \left\{ \|U(x^T)\|; \|x^T\| = 1 \right\} \quad (2)$$

unde cu  $\|\cdot\|$  am notat o normă oarecare pe  $\mathbb{R}^n$ .

Se știe de asemenea că:

$$\|U\|_0 = \inf \left\{ c > 0; \|U(x^T)\| \leq c \|x^T\|, (\forall) x^T \in \mathbb{R}^n \right\} \quad (3)$$

**Definiția 3.** Se numește norma matricei  $A$  subordonată normei vectoriale  $\|\cdot\|$  următorul număr:

$$\|A\| = \sup \left\{ \|Ax\|; \|x\| = 1 \right\} = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} \quad (4)$$

Ca și în cazul normei operatoriale, avem

$$\|A\| = \inf \left\{ c > 0; \|Ax\| \leq c \|x\|, (\forall) x \right\}. \quad (5)$$

Din relația (5) rezultă în particular că  $\|Ax\| \leq \|A\| \cdot \|x\|$ ,  $\forall x \in \mathbb{R}^n$ , deci norma matriceală definită de (4) este compatibilă cu norma vectorială căreia îi este subordonată.

Este evident că aplicația  $A \rightarrow \|A\|$  definită de (4) satisface proprietățile (i)–(iii) din definiția 1. De asemenea avem

$$\|ABx\| \leq \|A\| \cdot \|Bx\| \leq \|A\| \cdot \|B\| \cdot \|x\|,$$

de unde rezultă  $\|A \cdot B\| \leq \|A\| \cdot \|B\|$ .

Așadar, formula (4) definește într-adevăr o normă de matrice.

**Definiția 4.** Dacă  $\lambda_1, \dots, \lambda_n$  sunt valorile proprii ale matricei  $A$ , atunci se notează cu  $\rho(A) = \max_{1 \leq i \leq n} |\lambda_i|$  și  $\rho(A)$  se numește raza spectrală a matricei  $A$  (în această definiție  $\lambda_i$  pot fi reale sau complexe)

**Teorema 1.** Pentru  $A \in M_n(\mathbb{R})$  avem:

$$(1) \quad \|A\|_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^n |a_{ij}|,$$

$$(2) \quad \|A\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^n |a_{ij}| ,$$

$$(3) \quad \|A\|_2 = \left( \rho(A^T \cdot A) \right)^{\frac{1}{2}} ,$$

unde cu  $\|A\|_p$  am notat norma matricei  $A$  subordonată normei vectoriale  $\|x\|_p$ .

**Demonstrație.**

$$\|Ax\|_\infty = \max_{1 \leq i \leq n} \left| \sum_{j=1}^n a_{ij} x_j \right| \leq \max_{1 \leq i \leq n} \sum_{j=1}^n |a_{ij}| \cdot |x_j| \leq \|x\|_\infty \max_{1 \leq i \leq n} \sum_{j=1}^n |a_{ij}|$$

Rezultă  $\|A\|_\infty \leq \max_{1 \leq i \leq n} \sum_{j=1}^n |a_{ij}|$ . Rămâne să arătăm că există  $\tilde{x}$  cu  $\|\tilde{x}\|_\infty = 1$

astfel încât  $\|A\tilde{x}\|_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^n |a_{ij}|$ . Pentru aceasta, fie  $k$  astfel încât să avem

$$\sum_{j=1}^n |a_{kj}| = \max_{1 \leq i \leq n} \sum_{j=1}^n |a_{ij}| \quad (6)$$

$$\text{și fie } \tilde{x}_j = \begin{cases} 0 & \text{dacă } a_{kj} = 0 \\ \frac{a_{kj}}{|a_{kj}|} & \text{dacă } a_{kj} \neq 0 \end{cases} .$$

Evident că  $\|\tilde{x}\|_\infty = 1$  și  $\|A\tilde{x}\|_\infty = \sum_{j=1}^n |a_{kj}| = \max_{1 \leq i \leq n} \sum_{j=1}^n |a_{ij}|$ . Așadar, am demonstrat

(1). În continuare avem

$$\begin{aligned} \|Ax\|_1 &= \sum_{i=1}^n |a_{i1}x_1 + \dots + a_{in}x_n| \leq \sum_{i=1}^n (|a_{i1}|x_1 + \dots + |a_{in}|x_n) \leq \\ &\leq \left( \max_{1 \leq j \leq n} \sum_{i=1}^n |a_{ij}| \right) \cdot (|x_1| + \dots + |x_n|) = \max_{1 \leq j \leq n} \left( \sum_{i=1}^n |a_{ij}| \right) \cdot \|x\|_1 \end{aligned}$$

de unde rezultă  $\|A\|_1 \leq \max_{1 \leq j \leq n} \sum_{i=1}^n |a_{ij}|$ .

Pe de altă parte dacă  $e_j^T = (0, \dots, 1, \dots, 0)$ , atunci  $\|e_j\|_1 = 1$  și  $\|Ae_j\|_1 = \sum_{i=1}^n |a_{ij}|$ , de

unde rezultă  $\|A\|_1 \geq \sum_{i=1}^n |a_{ij}|$ , pentru orice  $j = \overline{1, n}$ . Așadar,  $\|A\|_1 \geq \max_{1 \leq j \leq n} \sum_{i=1}^n |a_{ij}|$

și cu aceasta afirmația (2) este dovedită.

Fie  $B = A^T A$  și fie

$$\mu_1 = \sup \left\{ x^T B x ; \|x\|_2 = 1 \right\} \quad (7)$$

Evident  $\mu_1 = \sup \left\{ (Ax)^T Ax ; \|x\|_2 = 1 \right\} = \|A\|_2^2$ , deci  $\|A\|_2 = \sqrt{\mu_1}$ .

Deoarece mulțimea  $S = \{x ; \|x\|_2 = 1\}$  este compactă, rezultă că există  $v$  cu proprietățile:  $\mu_1 = v^T B v$  și  $\|v\|_2 = 1$ .

Vom arăta în continuare că  $Bv = \mu_1 v$ , deci că  $v$  este un vector propriu pentru  $B$  și corespunde valorii proprii  $\mu_1$ .

Într-adevăr, pentru orice  $z \neq 0$  avem:  $\left( \frac{z}{\|z\|_2} \right)^T B \left( \frac{z}{\|z\|_2} \right) \leq \mu_1$  și deci

$$z^T B z \leq \mu_1 \|z\|_2^2 = \mu_1 z^T z \quad (8)$$

Pe de altă parte este evident că relația (8) este verificată și pentru  $z=0$ . Deci relația (8) are loc pentru orice  $z$ . De asemenea avem:

$$v^T B v = \mu_1 v^T v \quad (9)$$

Dacă notăm cu  $C = B - \mu_1 I_n$ , atunci avem:

$$z^T C z \leq 0, \quad (\forall) z \text{ și} \quad (8')$$

$$v^T C v = 0 \quad (9')$$

Fie  $z = v + ty$ , unde  $t \in \mathbb{R}$  este oarecare și  $y$  este un vector oarecare. Din (8') și din faptul că  $C$  este simetrică rezultă

$$v^T C v + 2ty^T(Cv) + t^2 y^T C y \leq 0.$$

Ținând seama de (9') avem

$$t^2 y^T C y + 2ty^T(Cv) \leq 0. \quad (10)$$

Pentru ca (10) să fie adevărată pentru orice  $t \in \mathbb{R}$  trebuie ca  $y^T C v = 0$ . Cum  $y$  a fost arbitrar rezultă  $0 = C v = (B - \mu_1 I_n) v = B v - \mu_1 v$ .

Așadar, avem  $Bv = \mu_1 v$ , deci  $\mu_1$  este valoare proprie pentru  $B$  și în plus  $\mu_1 = \|A\|_2^2$ .

Pe de altă parte, fie  $\mu$  o altă valoare proprie a matricei  $B$  și fie  $u \neq 0$ ,  $\|u\|_2 = 1$ , astfel încât  $Bu = \mu u$ . În continuare avem

$$\mu_1 = \|A\|_2^2 \geq \|Au\|_2^2 = u^T B u = u^T \mu u = \mu.$$

Așadar,  $\mu_1$  este cea mai mare valoare proprie a matricei  $B$ , deci am demonstrat și afirmația (3).  $\square$

În particular dacă presupunem că matricea  $A$  este simetrică, rezultă că  $B=A^2$ . Fie  $\lambda_1, \lambda_2, \dots, \lambda_n$  valorile proprii ale matricei  $A$ , care în acest caz sunt reale. Se știe că  $\lambda_1^2, \lambda_2^2, \dots, \lambda_n^2$  sunt valorile proprii ale matricei  $A^2$ . Să presupunem că  $\lambda_1^2 = \max_{1 \leq j \leq n} \lambda_j^2$ .

Din Teorema 1 rezultă că  $\|A\|_2 = |\lambda_1|$ . Dacă, în plus,  $A$  este pozitiv definită, atunci  $\lambda_i > 0$  pentru orice  $i$ . Să presupunem că:  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$ . Din cele de mai sus rezultă  $\|A\|_2 = \lambda_1$ , unde  $\lambda_1$  este cea mai mare valoare proprie a matricei simetrice și pozitiv definite  $A$ .

### **§1.6. Perturbarea sistemelor liniare. Numărul de condiționare al unei matrice**

Considerăm următorul sistem de ecuații liniare

$$\begin{cases} 10x_1 + 7x_2 + 8x_3 + 7x_4 = 32 \\ 7x_1 + 5x_2 + 6x_3 + 5x_4 = 23 \\ 8x_1 + 6x_2 + 10x_3 + 9x_4 = 33 \\ 7x_1 + 5x_2 + 9x_3 + 10x_4 = 31 \end{cases} \quad (1)$$

a cărui soluție exactă este  $x_1=x_2=x_3=x_4=1$ .

Să considerăm acum sistemul (1') în care am modificat "puțin" termenii liberi

$$\begin{cases} 10x_1 + 7x_2 + 8x_3 + 7x_4 = 32.1 \\ 7x_1 + 5x_2 + 6x_3 + 5x_4 = 22.9 \\ 8x_1 + 6x_2 + 10x_3 + 9x_4 = 33.1 \\ 7x_1 + 5x_2 + 9x_3 + 10x_4 = 30.9 \end{cases} \quad (1')$$

Soluția sistemului (1') este  $x_1 = 9.2$ ;  $x_2 = -12.6$ ;  $x_3 = 4.5$ ;  $x_4 = -1.1$ . Așadar, o eroare mică, de ordinul 0.1, a termenilor liberi, produce o eroare mare, de ordinul 10, a soluției sistemului.

Fie acum sistemul (1'') în care modificăm puțin coeficienții sistemului

$$\begin{cases} 10x_1 + 7x_2 + 8.1x_3 + 7.2x_4 = 32 \\ 7.08x_1 + 5.04x_2 + 6x_3 + 5x_4 = 23 \\ 8x_1 + 5.98x_2 + 9.89x_3 + 9x_4 = 33 \\ 6.99x_1 + 4.99x_2 + 9x_3 + 9.98x_4 = 31 \end{cases} \quad (1'')$$

Soluția sistemului (1'') este:  $x_1 = -81$ ;  $x_2 = 137$ ;  $x_3 = -34$ ;  $x_4 = 22$ .

Să analizăm acum efectul perturbării membrului drept asupra soluției unui sistem liniar  $Ax=b$ , în care matricea  $A$  este nesingulară.

Notăm cu  $\delta b$  perturbarea membrului drept și cu  $\delta x$  perturbarea care rezultă pentru soluție. Avem:  $A(x + \delta x) = b + \delta b$ , de unde rezultă  $A \delta x = \delta b$  și deci  $\delta x = A^{-1} \delta b$ . Pentru orice normă de matrice compatibilă avem:

$$\|\delta x\| \leq \|A^{-1}\| \cdot \|\delta b\| \quad (2)$$

Pe de altă parte  $\|b\| = \|Ax\| \leq \|A\| \cdot \|x\|$ , de unde rezultă:

$$\frac{1}{\|x\|} \leq \frac{\|A\|}{\|b\|} \quad (3)$$

Din relațiile (2) și (3) obținem  $\frac{\|\delta x\|}{\|x\|} \leq \|A\| \cdot \|A^{-1}\| \cdot \frac{\|\delta b\|}{\|b\|}$ .

Numărul de condiționare al unei matrice se definește astfel

$$\text{cond}(A) = \|A\| \cdot \|A^{-1}\|. \quad (4)$$

Așadar, între eroarea relativă a membrului drept și eroarea relativă a soluției sistemului avem următoarea inegalitate

$$\frac{\|\delta x\|}{\|x\|} \leq \text{cond}(A) \cdot \frac{\|\delta b\|}{\|b\|}. \quad (5)$$

Observăm că dacă numărul de condiționare al matricei coeficienților sistemului este mare, atunci la erori relativ mici ale termenilor liberi, pot apare erori relativ mari pentru soluția sistemului. În cazul exemplului (1) avem

$$A = \begin{pmatrix} 10 & 7 & 8 & 8 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix}, \quad A^{-1} = \begin{pmatrix} 25 & -41 & 10 & -6 \\ -41 & 68 & -17 & 10 \\ 10 & -17 & 5 & -3 \\ -6 & 10 & -3 & 2 \end{pmatrix}$$

și  $\text{cond}_2(A) \cong 2984$ . (S-a folosit norma de matrice  $\|\cdot\|_2$ ). După cum se vede, numărul de condiționare este destul de mare, ceea ce explică instabilitatea soluției sistemului.

Numărul de condiționare are următoarele proprietăți:

- (i)  $\text{cond}(I_n) \geq 1$
- (ii)  $\text{cond}(A) = \text{cond}(A^{-1})$
- (iii)  $\text{cond}(\alpha A) = \text{cond}(A)$  pentru orice  $\alpha \neq 0$
- (iv)  $\text{cond}_2(A) = \frac{\sqrt{\mu_1}}{\sqrt{\mu_n}}$ , unde  $\mu_1 \geq \mu_2 \geq \dots \geq \mu_n > 0$  sunt valorile proprii ale matricei

$$B = A^T A$$

- (v) Dacă  $A$  este simetrică, atunci  $\text{cond}_2(A) = \frac{\max|\lambda_i|}{\min|\lambda_i|}$ , unde  $\lambda_1, \dots, \lambda_n$  sunt valorile proprii ale matricei  $A$
- (vi) Dacă  $A$  este ortogonală, atunci  $\text{cond}(A)=1$ .
- Pentru a evalua eroarea soluției sistemului la o perturbare a coeficienților sistemului, avem nevoie de următoarele două leme.

**Lema 1.** Dacă  $A \in M_n(\mathbb{R})$  și  $\|A\| < 1$ , atunci:

- (i)  $A+I_n$  și  $A-I_n$  sunt nesingulare, și
- (ii)  $\frac{1}{\|A\|+1} \leq \|(A \pm I_n)^{-1}\| \leq \frac{1}{1-\|A\|}$ .

**Demonstrație.**

Prezentăm demonstrația pentru  $A+I_n$ .

Presupunem prin absurd că  $A+I_n$  este singulară. Atunci există  $x \neq 0$ ,  $\|x\|=1$  astfel încât  $(A+I_n) \cdot x = 0$ . În continuare avem  $x = -Ax$ , deci  $\|x\| \leq \|A\| \cdot \|x\|$ .

Rezultă  $\|A\| \geq 1$  ceea ce contrazice ipoteza  $\|A\| < 1$ .

Pentru a demonstra (ii) observăm că

$$1 = \|I_n\| = \|(I_n + A)(I_n + A)^{-1}\| \leq \|I_n + A\| \cdot \|(I_n + A)^{-1}\| \leq (1 + \|A\|) \cdot \|(I_n + A)^{-1}\|$$

de unde rezultă

$$\frac{1}{\|A\|+1} \leq \|(A + I_n)^{-1}\|.$$

Pe de altă parte avem

$$I_n = (I_n + A)^{-1} + A(I_n + A)^{-1},$$

de unde rezultă

$$\|(I_n + A)^{-1}\| = \|I_n - A(I_n + A)^{-1}\| \leq 1 + \|A\| \cdot \|(I_n + A)^{-1}\|,$$

și mai departe

$$\|(I_n + A)^{-1}\| \leq \frac{1}{1-\|A\|}. \quad \square$$

**Lema 2 (a perturbării).** Fie  $A, B \in M_n(\mathbb{R})$  cu proprietățile:

- (i)  $\|A^{-1}\| \leq \alpha$
- (ii)  $\|A^{-1}(B - A)\| \leq k < 1$ .

Atunci  $B$  este nesingulară și  $\|B^{-1}\| \leq \frac{\alpha}{1-k}$ .

**Demonstrație.**

Din Lema 1 rezultă că  $I_n + A^{-1}(B - A) = A^{-1}B$  este nesingulară.

Cum  $\det(A^{-1}B) = \det(A^{-1}) \cdot \det B$ , va rezulta  $\det B \neq 0$ , deci  $B$  este nesingulară. Tot din Lema 1 rezultă

$$\|B^{-1}A\| = \left\| [I_n + A^{-1}(B-A)]^{-1} \right\| \leq \frac{1}{1 - \|A^{-1}(B-A)\|} \leq \frac{1}{1-k}.$$

Mai departe avem

$$\|B^{-1}\| = \|(B^{-1}A)A^{-1}\| \leq \|B^{-1}A\| \cdot \|A^{-1}\| \leq \frac{\alpha}{1-k}. \quad \square$$

**Teorema 1.** Dacă perturbăm matricea coeficienților sistemului  $Ax = b$  cu  $\delta A$  și  $\|A^{-1}\delta A\| < 1$ , atunci între eroarea relativă a soluției și eroarea relativă a matricei coeficienților are loc inegalitatea

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\text{cond}(A)}{1 - \text{cond}(A)} \frac{\|\delta A\|}{\|A\|}.$$

**Demonstrație.**

Din egalitățile  $Ax=b$  și  $(A+\delta A)(x+\delta x)=b$  rezultă  $A\delta x + \delta Ax + \delta A\delta x=0$ . În continuare avem  $\delta x = -(A+\delta A)^{-1}\delta Ax$  și mai departe

$$\frac{\|\delta x\|}{\|x\|} \leq \|(A+\delta A)^{-1}\| \cdot \|\delta A\| \quad (6)$$

Dacă alegem în Lema 2  $\alpha = \|A^{-1}\|$  și  $B=A+\delta A$ , atunci

$$\|A^{-1}(B-A)\| = \|A^{-1}\delta A\| < 1$$

și va rezulta

$$\|B^{-1}\| = \|(A+\delta A)^{-1}\| \leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\delta A\|} \leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\| \cdot \|\delta A\|}. \quad (7)$$

Din (6) și (7) obținem

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

Ținând seama de definiția numărului de condiționare, ultima inegalitate devine

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\text{cond}(A)}{1 - \text{cond}(A)} \frac{\|\delta A\|}{\|A\|}. \quad \square$$

**Observația 1.** Dacă presupunem în plus că perturbăm și membrul drept al sistemului cu  $\delta b$  atunci rezultă

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

**Observația 2.** Rezolvarea sistemului  $Ax=b$ , cu metoda Gauss revine la rezolvarea a două sisteme triunghiulare  $Uy=b$  și  $Lx=y$ . Rezolvarea fiecărui sistem necesită  $n^2$  operații. Dacă unul din aceste sisteme este rău condiționat (ceea ce se poate întâmpla chiar dacă sistemul inițial  $Ax=b$  este bine condiționat) metoda Gauss conduce la erori mari.

Cu totul altfel stau lucrurile în cazul metodei Householder. Deoarece  $\text{cond}(Q) = 1$  și  $\text{cond}(QR) = \text{cond}(R)$ , rezultă că sistemul  $Qy = b$  este bine condiționat și deci că sistemul  $Ax = b$  are aceeași condiționare ca sistemul  $Rx = y$ . Așadar, algoritmul Householder are proprietăți de stabilitate mai bune decât algoritmul Gauss.

**Observația 3.** Pentru evaluarea numărului de condiționare  $\text{cond}(A)$ , este suficient să cunoaștem un majorant pentru  $\|A^{-1}\|$ .

Calculul lui  $\|(LU)^{-1}\|$  este mai ușor decât calculul lui  $\|A^{-1}\|$ , deoarece inversarea matricelor triunghiulare este ușoară. Să presupunem că

$$\|(LU)^{-1}\| = \alpha \quad \text{și} \quad \|A - LU\| < \frac{k}{\alpha}$$

unde  $0 < k < 1$ . Atunci din Lema 2 rezultă  $\|A^{-1}\| \leq \frac{\alpha}{1-k}$ .

Într-adevăr, să alegem în Lema 2 matricea  $LU$  în loc de  $A$  și matricea  $A$  în loc de  $B$ . Avem

$$\|A^{-1}(A - LU)\| \leq \|A^{-1}\| \cdot \|A - LU\| \leq \alpha \cdot \frac{k}{\alpha} = k < 1.$$

Atunci rezultă  $\|A^{-1}\| \leq \frac{\alpha}{1-k}$  și deci  $\text{cond}(A) \leq \frac{\alpha \|A\|}{1-k}$ . □

### §1.7. Metode iterative de rezolvare a sistemelor de ecuații liniare

Metodele directe de rezolvare numerică a sistemelor de ecuații liniare se utilizează pentru sisteme care au matricea coeficienților densă (aproape toți coeficienții sunt nenuli) și cu un număr de ecuații moderat (până la 100 de ecuații). Pentru sisteme mari de ecuații de ordinul  $10^3 \rightarrow 10^5$  și care au matricea coeficienților rară (cu multe elemente nule), se utilizează metode iterative de rezolvare numerică.

Să presupunem că sistemul

$$Ax = b \quad (1)$$

se poate pune sub forma echivalentă

$$x = Bx + c \quad (2)$$

Forma echivalentă (2) ne sugerează următorul proces iterativ:

$$x^{(m+1)} = Bx^{(m)} + c, \quad m \geq 0, \quad (3)$$

unde  $x^{(0)}$  este un vector arbitrar.

Dacă notăm cu  $x^*$  soluția exactă a sistemului, atunci avem

$$x^* = Bx^* + c \quad (4)$$

Fie  $e^{(m)} = x^* - x^{(m)}$  vectorul eroare.

Din (3) și (4) rezultă  $e^{(m+1)} = Be^{(m)}$ ,  $m \in \mathbb{N}^*$  și mai departe

$$e^{(m)} = B^m e^{(0)} \quad (5)$$

**Teorema 1.** Dacă  $\|B\| < 1$ , atunci șirul  $(x^{(m)})$  este convergent și  $\lim_{m \rightarrow \infty} x^{(m)} = x^*$ .

**Demonstrație.**

Este suficient să arătăm că  $\lim_{m \rightarrow \infty} e^{(m)} = 0$ .

Din (5) avem

$$\|e^{(m)}\| \leq \|B^m\| \|e^{(0)}\| \leq \|B\|^m \cdot \|e^{(0)}\|.$$

Deoarece  $\lim_{m \rightarrow \infty} \|B\|^m = 0$ , rezultă  $\lim_{m \rightarrow \infty} e^{(m)} = 0$ .  $\square$

**Teorema 2.** Condiția necesară și suficientă ca șirul  $(x^{(m)})$  definit de (3) să fie convergent este ca  $\rho(B) < 1$ , unde cu  $\rho(B)$  s-a notat raza spectrală a matricei  $B$ .

**Demonstrație.**

Este suficient să arătăm că  $\lim_{m \rightarrow \infty} B^m = 0$  dacă și numai dacă  $\rho(B) < 1$ . Din

Algebra liniară se știe că matricea  $B$  se poate aduce la *forma canonică Jordan*, deci că există o matrice nesingulară  $C$  astfel încât

$$C^{-1} \cdot B \cdot C = J = \begin{pmatrix} J_{p_1}(\lambda_1) & \dots & 0 \\ \vdots & J_{p_2}(\lambda_2) & \vdots \\ 0 & & J_{p_r}(\lambda_r) \end{pmatrix},$$

unde

$$J_p(\lambda) = \begin{pmatrix} \lambda & 1 & 0 & \dots & 0 \\ & \lambda & 1 & & \vdots \\ \vdots & & & \ddots & 1 \\ 0 & & & & \lambda \end{pmatrix}$$

este o *celulă Jordan*,  $\lambda_1, \dots, \lambda_r$  sunt valorile proprii ale matricei  $B$  și  $p_1, \dots, p_r$  sunt ordinele de multiplicitate ale acestor valori proprii. Deoarece  $C^{-1} B^m C = J^m$ , rezultă că  $\lim_{m \rightarrow \infty} B^m = 0$  dacă și numai dacă  $\lim_{m \rightarrow \infty} J^m = 0$ . Pe

de altă parte,  $J = D + N$ , unde

$$D = \begin{pmatrix} \lambda_1 & & & & 0 \\ & \ddots & & & \\ 0 & & \lambda_1 & & \\ & & & \ddots & \\ & & & & \lambda_r & & 0 \\ & & & & & \ddots & \\ 0 & & & & 0 & & \lambda_r \end{pmatrix}$$

este o matrice diagonală de ordinul  $n$ , iar  $N$  este o *matrice nilpotentă de ordinul  $n$* , adică  $N^n = 0$ .

În continuare avem  $J^m = \sum_{k=0}^m C_m^k D^{m-k} N^k$ . Deoarece  $N^k = 0$  pentru  $k \geq n$ , vom

avea

$$J^m = \sum_{k=0}^n C_m^k D^{m-k} N^k. \quad (6)$$

Observăm că  $\|D\|_\infty = \max_{1 \leq i \leq r} |\lambda_i| = \rho(B) < 1$ . Din (6) rezultă:

$$\|J^m\|_\infty \leq \sum_{k=0}^n \frac{m(m-1)\dots(m-k+1)}{k!} \|D\|_\infty^{m-k} \|N\|_\infty^k < \sum_{k=0}^n \frac{m^k}{k!} \|N\|_\infty^k \cdot (\rho(B))^{m-k}$$

Cum  $\lim_{m \rightarrow \infty} m^k (\rho(B))^{m-k} = 0$ , rezultă că  $\lim_{m \rightarrow \infty} \|J^m\|_{\infty} = 0$ , deci că  $\lim_{m \rightarrow \infty} J^m = 0$ .

Reciproc, să presupunem că  $\lim_{m \rightarrow \infty} B^m = 0$  și că  $\rho(B) \geq 1$ . Atunci există un vector propriu  $x \neq 0$  și o valoare proprie  $\lambda$ , cu  $|\lambda| \geq 1$ , astfel încât  $Bx = \lambda x$  și deci  $B^m x = \lambda^m x$ .

Cum  $(\lambda^m x)$  nu converge la 0, rezultă că  $B^m$  nu converge la 0, ceea ce contrazice ipoteza făcută.  $\square$

Una din cele mai cunoscute metode iterative este *metoda Jacobi*.

Să presupunem că matricea sistemului  $Ax = b$  are proprietatea  $a_{ii} \neq 0, i = \overline{1, n}$ . Dacă notăm cu  $D = \text{diag}(a_{11}, \dots, a_{nn})$  și cu  $E = D - A$ , atunci obținem sistemul echivalent  $(D - E)x = b$  și mai departe

$$x = D^{-1}Ex + D^{-1}b \quad (7)$$

Cum  $D^{-1} = \text{diag}\left(\frac{1}{a_{11}}, \dots, \frac{1}{a_{nn}}\right)$ , rezultă că  $\|D^{-1}E\|_{\infty} = \max_{1 \leq i \leq n} \sum_{\substack{j=1 \\ j \neq i}}^n \frac{|a_{ij}|}{|a_{ii}|}$ .

Observăm că dacă matricea  $A$  este tare diagonal dominantă, atunci  $\|D^{-1}E\|_{\infty} < 1$  și

în virtutea Teoremei 1, șirul

$$x^{(m+1)} = (D^{-1}E)x^{(m)} + D^{-1}b, \quad m \geq 0 \quad (8)$$

este convergent pentru orice aproximație inițială  $x^{(0)}$ . Așadar, metoda Jacobi constă în următoarele:

Sistemul  $Ax = b$  se pune sub forma echivalentă (7).

Scris pe componente, sistemul (7) arată astfel

$$x_i = \frac{1}{a_{ii}} \left( b_i - \sum_{\substack{j=1 \\ j \neq i}}^n a_{ij} x_j \right), \quad i = \overline{1, n} \quad (7')$$

Se obține șirul recurent  $\{x^{(m)}\}$  unde

$$x_i^{(m+1)} = \frac{1}{a_{ii}} \left( b_i - \sum_{\substack{j=1 \\ j \neq i}}^n a_{ij} x_j^{(m)} \right), \quad i = \overline{1, n} \quad (8')$$

Dacă matricea  $A$  este tare diagonal dominantă, șirul  $(x^{(m)})$  converge la soluția exactă a sistemului.

**Exemplu.** Fie sistemul

$$\begin{pmatrix} 5 & -1 & -1 & -1 \\ -1 & 10 & -1 & -1 \\ -1 & -1 & 5 & -1 \\ -1 & -1 & -1 & 10 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} -4 \\ 12 \\ 8 \\ 34 \end{pmatrix}$$

Soluția exactă este  $x_1 = 1, x_2 = 2, x_3 = 3, x_4 = 4$ .

$$D = \begin{pmatrix} 5 & 0 & 0 & 0 \\ 0 & 10 & 0 & 0 \\ 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 10 \end{pmatrix}, E = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}, D^{-1}E = \begin{pmatrix} 0 & 0,2 & 0,2 & 0,2 \\ 0,1 & 0 & 0,1 & 0,1 \\ 0,2 & 0,2 & 0 & 0,2 \\ 0,1 & 0,1 & 0,1 & 0 \end{pmatrix},$$

$$D^{-1}b = \begin{pmatrix} -0,8 \\ 1,2 \\ 1,6 \\ 3,4 \end{pmatrix}, \|D^{-1}E\|_{\infty} = 0,6 < 1.$$

Obținem următorul proces iterativ:

$$\begin{cases} x_1^{(m+1)} = 0,2x_2^{(m)} + 0,2x_3^{(m)} + 0,2x_4^{(m)} - 0,8 \\ x_2^{(m+1)} = 0,1x_1^{(m)} + 0,1x_3^{(m)} + 0,1x_4^{(m)} - 1,2 \\ x_3^{(m+1)} = 0,2x_1^{(m)} + 0,2x_2^{(m)} + 0,2x_4^{(m)} + 1,6 \\ x_4^{(m+1)} = 0,1x_1^{(m)} + 0,1x_2^{(m)} + 0,1x_3^{(m)} + 3,4 \end{cases}$$

Dacă alegem aproximația inițială  $x_1^0 = x_2^0 = x_3^0 = x_4^0 = 0$  atunci după 5 iterații obținem

$$x_1^{(5)} = 0,948; \quad x_2^{(5)} = 1,969; \quad x_3^{(5)} = 2,948; \quad x_4^{(5)} = 3,969$$

O altă metodă iterativă cunoscută este *metoda Gauss-Seidel* și care corespunde următoarei *spargerii* a matricei coeficienților:

$$A = (D+L)+U \text{ unde } D = \text{diag}(a_{11}, \dots, a_{nn}),$$

$$L = \begin{pmatrix} 0 & 0 & \dots & 0 \\ a_{21} & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & 0 \end{pmatrix}, \text{ iar } U = \begin{pmatrix} 0 & a_{12} & \dots & a_{1n} \\ 0 & 0 & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 \end{pmatrix}.$$

Sistemul (1) devine  $(D+L)x = -Ux + b$  și mai departe obținem următorul proces iterativ:

$$(D+L)x^{(m+1)} = -Ux^{(m)} + b. \quad (9)$$

Pe componente obținem

$$x_i^{(m+1)} = \frac{1}{a_{ii}} \left( b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(m+1)} - \sum_{j=i+1}^n a_{ij} x_j^{(m)} \right), \quad i = \overline{1, n}. \quad (10)$$

Din algoritmul (10) se observă că fiecare nouă componentă,  $x_j^{(m+1)}$ , este imediat utilizată la calculul următoarei componente.

Se poate arăta că procesul iterativ Gauss-Seidel este convergent dacă matricea  $A$  este tare diagonal dominantă.

În cazul exemplului precedent obținem

$$\begin{pmatrix} 5 & 0 & 0 & 0 \\ -1 & 10 & 0 & 0 \\ -1 & -1 & 5 & 0 \\ -1 & -1 & -1 & 10 \end{pmatrix} \begin{pmatrix} x_1^{(m+1)} \\ x_2^{(m+1)} \\ x_3^{(m+1)} \\ x_4^{(m+1)} \end{pmatrix} = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1^{(m)} \\ x_2^{(m)} \\ x_3^{(m)} \\ x_4^{(m)} \end{pmatrix} + \begin{pmatrix} -4 \\ 12 \\ 8 \\ 34 \end{pmatrix}$$

sau

$$\begin{cases} x_1^{(m+1)} = \frac{1}{5} (x_2^{(m)} + x_3^{(m)} + x_4^{(m)} - 4) \\ x_2^{(m+1)} = \frac{1}{10} (x_1^{(m+1)} + x_3^{(m)} + x_4^{(m)} + 12) \\ x_3^{(m+1)} = \frac{1}{5} (x_1^{(m+1)} + x_2^{(m+1)} + x_4^{(m)} + 8) \\ x_4^{(m+1)} = \frac{1}{10} (x_1^{(m+1)} + x_2^{(m+1)} + x_3^{(m+1)} + 34) \end{cases}$$

Pentru  $x_1^{(0)} = x_2^{(0)} = x_3^{(0)} = x_4^{(0)} = 0$ , după 5 iterații obținem

$$x_1^{(5)} = 0.995 ; \quad x_2^{(5)} = 1.998 ; \quad x_3^{(5)} = 2.998 ; \quad x_4^{(5)} = 3.999 .$$

### §1.8. Metode de relaxare. Principiile de bază

Metodele de relaxare sunt metode iterative și sunt utilizate pentru rezolvarea numerică a sistemelor liniare care au matricea coeficienților simetrică și pozitiv definită.

Fie sistemul liniar

$$Ax - b = 0 \quad (1)$$

unde matricea  $A$  este simetrică și pozitiv definită. Dacă  $v = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}$  este un vector de

probă oarecare, atunci notăm cu

$$r = Av - b. \quad (2)$$

Vectorul  $r$  se numește *vectorul rezidual*.

Scopul oricărei metode de relaxare este ca prin schimbarea sistematică a vectorului  $v$ , vectorul rezidual corespunzător  $r$  să se micșoreze, eventual să se anuleze.

În cele ce urmează, pentru orice doi vectori

$$u = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} \quad \text{și} \quad v = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}$$

vom nota produsul lor scalar cu

$$\langle u, v \rangle = v^T u = u_1 v_1 + u_2 v_2 + \dots + u_n v_n \quad (3)$$

Asociem sistemului (1) funcția pătratică

$$F(v) = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n a_{ij} v_i v_j - \sum_{i=1}^n b_i v_i = \frac{1}{2} \langle Av, v \rangle - \langle b, v \rangle \quad (4)$$

Deoarece  $A$  este pozitiv definită, rezultă  $Q(v) > 0$  pentru orice  $v \neq 0$ , unde  $Q(v) = \langle Av, v \rangle$ . Observăm de asemenea că pentru orice  $i = \overline{1, n}$  avem

$$\frac{\partial F}{\partial v_i} = \sum_{j=1}^n a_{ij} v_j - b_i,$$

deci vectorul rezidual

$$r = \text{grad} F. \quad (5)$$

**Teorema 1.** Problema determinării soluției sistemului (1) este echivalentă cu problema determinării punctului de minim al funcției pătratice (4).

**Demonstrație.**

Fie  $v_0$  soluția sistemului (1). Atunci  $r_0 = Av_0 - b = 0$ . Cum

$r_0 = \text{grad} F(v_0)$ , rezultă  $\frac{\partial F}{\partial v_i}(v_0) = 0$ . Așadar,  $v = v_0$  este punct critic pentru  $F$ . Pe

de altă parte,

$$d^2 F(v_0) = \sum_{i=1}^n \sum_{j=1}^n a_{ij} dv_i dv_j > 0.$$

Rezultă că  $v = v_0$  este un punct de minim global pentru  $F$ .

Reciproc, dacă  $v=v_0$  este punct de minim pentru  $F$  atunci

$$\frac{\partial F}{\partial v_i}(v_0) = 0, \quad i = \overline{1, n}.$$

Rezultă  $\sum_{j=1}^n a_{ij}v_j^0 - b_i = 0, \quad i = \overline{1, n}$ , deci  $v=v_0$  este soluție pentru (1).  $\square$

În continuare prezentăm principiul de bază al metodei relaxării. Fie  $v$  un vector de probă oarecare,  $p$  o direcție dată și  $D = \{v' = v + tp; t \in \mathbb{R}\}$ , dreapta care trece prin  $v$  și este paralelă cu  $p$ . Ne propunem să determinăm  $v'_0 \in D$  astfel încât  $F(v'_0) = \min\{F(v'); v' \in D\}$ . Ținând seama de (4), rezultă

$$\begin{aligned} F(v') &= \frac{1}{2} \langle A(v + tp), v + tp \rangle - \langle b, v + tp \rangle = \\ &= \frac{1}{2} \langle Av, v \rangle - \langle b, v \rangle + \frac{t}{2} \langle Av, p \rangle + \frac{t^2}{2} \langle Ap, p \rangle + \frac{t}{2} \langle Ap, v \rangle - t \langle b, p \rangle = \\ &= F(v) + \frac{t^2}{2} \langle Ap, p \rangle + t \langle Av, p \rangle - t \langle b, p \rangle = F(v) + \frac{t^2}{2} \langle Ap, p \rangle + t \langle Av - b, p \rangle = \\ &= F(v) + \frac{t^2}{2} \langle Ap, p \rangle + t \langle r, p \rangle. \end{aligned}$$

Folosim notația

$$f(t) = F(v') = F(v + tp) = F(v) + \frac{t^2}{2} \langle Ap, p \rangle + t \langle r, p \rangle \quad (6)$$

Determinăm pe  $t$  astfel încât  $f(t) = F(v')$  să fie minimă. Pentru aceasta trebuie să avem  $f'(t) = 0$ , de unde rezultă  $t \langle Ap, p \rangle + \langle r, p \rangle = 0$ . Așadar, obținem:

$$t_{\min} = -\frac{\langle r, p \rangle}{\langle Ap, p \rangle} \quad (7)$$

Cum  $f''(t) = \langle Ap, p \rangle > 0$ , rezultă că vectorul  $v'_0 = v + t_{\min} p$  este un punct de minim pentru  $F(v')$ .

În continuare avem

$$f(t_{\min}) = F(v'_0) = F(v) - \frac{1}{2} \frac{\langle r, p \rangle^2}{\langle Ap, p \rangle}$$

de unde rezultă

$$\Delta F = F(v'_0) - F(v) = -\frac{1}{2} \frac{\langle r, p \rangle^2}{\langle Ap, p \rangle} \leq 0.$$

Pentru ca  $\Delta F < 0$ , trebuie ca  $\langle r, p \rangle \neq 0$ . Rezultă că direcția  $p$  se alege astfel încât  $p$  să nu fie perpendiculară pe  $r$ .

**Observația 1.** Dacă  $r'_0 = Av'_0 - b$  este vectorul rezidual corespunzător vectorului  $v'_0 = v + t_{\min} p$ , atunci  $\langle r'_0, p \rangle = 0$ .

Într-adevăr,

$$\langle r'_0, p \rangle = \langle Av - b, p \rangle + t_{\min} \langle Ap, p \rangle = \langle r, p \rangle - \frac{\langle r, p \rangle}{\langle Ap, p \rangle} \langle Ap, p \rangle = 0 .$$

Pentru interpretarea geometrică a principiului relaxării să considerăm cazul particular  $n = 2$ .

Ecuțiile  $F(v) = \text{constant}$ , reprezintă ecuațiile unor elipse concentrice, al căror centru comun, coincide cu punctul de minim al funcției  $F$ . Într-adevăr, ecuația  $F(v) = c$  revine la

$$a_{11}v_1^2 + 2a_{12}v_1v_2 + a_{22}v_2^2 - 2b_1v_1 - 2b_2v_2 = 2c . \quad (8)$$

Deoarece  $A$  este pozitiv definită, rezultă că

$$\delta = \begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix} > 0 ,$$

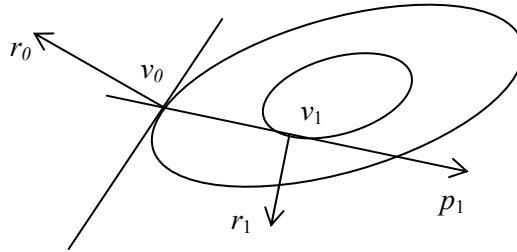
deci (8) reprezintă o elipsă.

Fie  $v_0$  un vector de probă oarecare și  $c_0 = F(v_0)$ . Ecuația  $F(v) = c_0$  reprezintă o elipsă și  $v = v_0$  aparține acestei elipse. Deoarece  $r_0 = \text{grad}F(v_0)$ , rezultă că  $r_0$  este perpendicular pe tangenta în  $v = v_0$  la elipsă. Direcția  $p_1$  o alegem astfel încât să nu fie perpendiculară pe  $r_0$ . Fie  $v_1 = v_0 + t_{\min} p_1$  și fie  $c_1 = F(v_1)$ .

Punctul  $v = v_1$  aparține elipsei  $F(v) = c_1$  și de asemenea aparține dreptei ce trece prin  $v_0$  și are direcția  $p_1$ .

Fie  $r_1 = Av_1 - b$ . Din Observația 1, rezultă că  $r_1$  este perpendicular pe direcția  $p_1$ . Pe de altă parte  $r_1 = \text{grad}F(v_1)$  este perpendicular pe tangenta în  $v = v_1$  la elipsa  $F(v) = c_1$ . Rezultă că  $v = v_1$ , este punctul de tangență la elipsa

$F(v) = c_1$  al dreptei care trece prin  $v_0$  și are direcția  $p_1$ .



## §1.9. Metoda relaxării simple

Este o metodă specifică calculului de mână, având mai ales o semnificație istorică.

Fie  $v$  un vector de probă oarecare și fie  $r = Av - b$  vectorul rezidual corespunzător.

Dacă  $\max_{1 \leq i \leq n} |r_i| = |r_j|$ , atunci alegem  $p = e_j$  unde  $e_j^T = (0, \dots, 1, \dots, 0)$ . Rezultă

$$t_{\min} = -\frac{\langle r, p \rangle}{\langle Ap, p \rangle} = -\frac{r_j}{a_{jj}} \quad \text{și}$$

$$v' = v + t_{\min} p = v - \frac{r_j}{a_{jj}} e_j \quad (1)$$

Pe componente avem:

$$v'_i = \begin{cases} v_i & \text{daca } i \neq j \\ v_j - \frac{r_j}{a_{jj}} & \text{daca } i = j \end{cases} \quad (2)$$

De asemenea vom avea  $r' = Av' - b = r - \frac{r_j}{a_{jj}} Ae_j$  și mai departe

$$\begin{cases} r'_1 = r_1 - \frac{r_j}{a_{jj}} a_{1j} \\ \dots\dots\dots \\ r'_j = 0 \\ \dots\dots\dots \\ r'_n = r_n - \frac{r_j}{a_{jj}} a_{nj} \end{cases} \quad (3)$$

$$\Delta F = F(v) - F(v') = -\frac{1}{2} \frac{r_j^2}{a_{jj}} < 0,$$

ceea ce asigură convergența metodei. Deși convergența este asigurată, experiențele numerice arată că aceasta este foarte lentă. Convergența este îmbunătățită dacă matricea  $A$  este tare diagonal dominantă.

**Exemplu.**

$$\begin{cases} -x_1 + 0.2x_2 + 0.2x_3 + 0.6 = 0 \\ 0.2x_1 - x_2 + 0.2x_3 + 0.5 = 0 \\ 0.2x_1 + 0.2x_2 - x_3 + 0.4 = 0 \end{cases}$$

Dacă alegem  $v^{(1)} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$ , atunci

$$r^{(1)} = \begin{pmatrix} 0.6 \\ 0.5 \\ 0.4 \end{pmatrix} \text{ și } \max(r_1^{(1)}, r_2^{(1)}, r_3^{(1)}) = r_1^{(1)} = 0.6 .$$

Așadar

$$p_1 = e_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad v^{(2)} = v^{(1)} - \frac{r_1^{(1)}}{a_{11}} e_1 \quad \text{și} \quad r^{(2)} = r^{(1)} - \frac{r_1^{(1)}}{a_{11}} A e_1 .$$

Pe componente avem

$$\begin{cases} v_1^{(2)} = 0.6 \\ v_2^{(2)} = 0 \\ v_3^{(2)} = 0 \end{cases}, \quad \begin{cases} r_1^{(2)} = 0 \\ r_2^{(2)} = 0.62 \\ r_3^{(2)} = 0.52 \end{cases},$$

$$r_2^{(2)} = \max(r_1^{(2)}, r_2^{(2)}, r_3^{(2)}) = 0.62 .$$

Rezultă

$$p_2 = e_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad v^{(3)} = v^{(2)} - \frac{r_2^{(2)}}{a_{22}} e_2 \quad \text{și} \quad r^{(3)} = r^{(2)} - \frac{r_2^{(2)}}{a_{22}} A e_2 ;$$

$$\begin{cases} v_1^{(3)} = 0.6 \\ v_2^{(3)} = 0.62 \\ v_3^{(3)} = 0 \end{cases}, \quad \begin{cases} r_1^{(3)} = 0.124 \\ r_2^{(3)} = 0 \\ r_3^{(3)} = 0.644 \end{cases}$$

$$r_3^{(3)} = \max(r_1^{(3)}, r_2^{(3)}, r_3^{(3)}) = 0.644 .$$

În continuare

$$p_3 = e_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \quad v^{(4)} = v^{(3)} - \frac{r_3^{(3)}}{a_{33}} e_3 \quad \text{și} \quad r^{(4)} = r^{(3)} - \frac{r_3^{(3)}}{a_{33}} A e_3 .$$

$$\begin{cases} v_1^{(4)} = 0.6 \\ v_2^{(4)} = 0.62 \\ v_3^{(4)} = 0.644 \end{cases}, \quad \begin{cases} r_1^{(4)} = 0.2528 \\ r_2^{(4)} = 0.1288 \\ r_3^{(4)} = 0 \end{cases}, \text{ etc.}$$

### §1.10. Metoda deplasărilor succesive (Gauss - Seidel)

În metoda deplasărilor succesive, direcția de relaxare urmează ciclic direcțiile  $e_1, e_2, \dots, e_n$ , indiferent de reziduurile respective, după care ciclul se reia. Pentru simplificare să presupunem că avem sistemul

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 - b_1 = 0 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 - b_2 = 0 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 - b_3 = 0 \end{cases}$$

Fie  $v^{(0)}$  vectorul de probă inițial și fie  $p' = e_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$ . Conform formulei (1) din §9

rezultă  $v' = v^{(0)} - \frac{r_1^{(0)}}{a_{11}} e_1$ , iar pe componente

$$\begin{cases} v'_1 = v_1^{(0)} - \frac{1}{a_{11}}(a_{11}v_1^{(0)} + a_{12}v_2^{(0)} + a_{13}v_3^{(0)} - b_1) \\ v'_2 = v_2^{(0)} \\ v'_3 = v_3^{(0)} \end{cases}$$

În continuare alegem direcția de relaxare

$$p'' = e_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$$

și obținem vectorul  $v''$  de componente

$$\begin{cases} v''_1 = v'_1 \\ v''_2 = v_2^{(0)} - \frac{1}{a_{22}}(a_{21}v'_1 + a_{22}v'_2 + a_{23}v'_3 - b_2) \\ v''_3 = v_3^{(0)} \end{cases}$$

În sfârșit, pentru direcția de relaxare

$$p''' = e_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix},$$

obținem vectorul  $v'''$  de componente

$$\begin{cases} v_1''' = v_1'' \\ v_2''' = v_2'' \\ v_3''' = v_3^{(0)} - \frac{1}{a_{33}}(a_{31}v_1'' + a_{32}v_2'' + a_{33}v_3'' - b_3) \end{cases}$$

După încheierea acestui ciclu, vectorul găsit va fi notat cu  $v^{(1)}$  și va avea componentele:

$$\begin{cases} v_1^{(1)} = v_1''' = -\frac{a_{12}}{a_{11}}v_2^{(0)} - \frac{a_{13}}{a_{11}}v_3^{(0)} + \frac{b_1}{a_{11}} \\ v_2^{(1)} = v_2''' = -\frac{a_{21}}{a_{22}}v_1^{(1)} - \frac{a_{23}}{a_{22}}v_3^{(0)} + \frac{b_2}{a_{22}} \\ v_3^{(1)} = v_3''' = -\frac{a_{31}}{a_{33}}v_1^{(1)} - \frac{a_{32}}{a_{33}}v_2^{(1)} + \frac{b_3}{a_{33}} \end{cases} \quad (1)$$

Efectuând calculele obținem:

$$\begin{cases} a_{11}v_1^{(1)} + a_{12}v_2^{(0)} + a_{13}v_3^{(0)} = b_1 \\ a_{21}v_1^{(1)} + a_{22}v_2^{(1)} + a_{23}v_3^{(0)} = b_2 \\ a_{31}v_1^{(1)} + a_{32}v_2^{(1)} + a_{33}v_3^{(1)} = b_3 \end{cases}$$

În general, pentru un sistem de  $n$  ecuații, după  $(m+1)$  cicluri se obține vectorul  $v^{(m+1)}$  care verifică ecuațiile:

$$\begin{cases} a_{11}v_1^{(m+1)} + a_{12}v_2^{(m)} + a_{13}v_3^{(m)} + \dots + a_{1n}v_n^{(m)} = b_1 \\ a_{21}v_1^{(m+1)} + a_{22}v_2^{(m+1)} + a_{23}v_3^{(m)} + \dots + a_{2n}v_n^{(m)} = b_2 \\ \dots \\ a_{n1}v_1^{(m+1)} + a_{n2}v_2^{(m+1)} + a_{n3}v_3^{(m+1)} + \dots + a_{nn}v_n^{(m+1)} = b_n \end{cases} \quad (2)$$

**Observația 1.** Formulele (1) coincid cu formulele (10) din §7.

Dacă notăm cu

$$E = \begin{pmatrix} 0 & 0 & 0 \dots & 0 \\ a_{21} & 0 & 0 \dots & 0 \\ a_{31} & a_{32} & 0 \dots & 0 \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & 0 \end{pmatrix}, \quad F = E^T \quad \text{și} \quad D = \begin{pmatrix} a_{11} & 0 & \dots & 0 \\ 0 & a_{22} & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & a_{nn} \end{pmatrix},$$

atunci matricea  $A$  admite descompunerea  $A=E+D+F$  și șirul de vectori  $v^{(m)}$  verifică relația matricială

$$(D+E)v^{(m+1)} + Fv^{(m)} = b, \quad (3)$$

de unde rezultă

$$v^{(m+1)} = -(D+E)^{-1}Fv^{(m)} + (D+E)^{-1}b. \quad (4)$$

În sfârșit, notând

$$M = -(D+E)^{-1}F \quad \text{și} \quad C = (D+E)^{-1}b \quad (5)$$

obținem procesul iterativ

$$v^{(m+1)} = Mv^{(m)} + C. \quad (6)$$

**Exemplu.** Să se găsească soluția aproximativă obținută după 5 iterații cu metoda deplasărilor succesive, luând vectorul inițial  $(0, 0, 0)$ , pentru sistemul  $Ax = b$ , unde:

$$A = \begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 3 & -1 & 0 \\ 0 & -1 & 3 & -1 \\ 0 & 0 & -1 & 2 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}.$$

*Rezolvare*

$$(D+E)^{-1} = \frac{1}{36} \begin{pmatrix} 18 & 0 & 0 & 0 \\ 6 & 12 & 0 & 0 \\ 2 & 4 & 12 & 0 \\ 1 & 2 & 6 & 18 \end{pmatrix} \quad F = \begin{pmatrix} 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

$$M = -(D+E)^{-1}F = \frac{1}{36} \begin{pmatrix} 0 & 18 & 0 & 0 \\ 0 & 6 & 12 & 0 \\ 0 & 2 & 4 & 12 \\ 0 & 1 & 2 & 6 \end{pmatrix}$$

$$\det(M - \lambda I) = \lambda^2 \left( \lambda^2 - \frac{4}{9}\lambda + \frac{1}{36} \right).$$

Valorile proprii ale matricei  $M$  sunt  $\lambda_1 = 0.369$ ;  $\lambda_2 = 0.077$ ;  $\lambda_3 = 0$ ;  $\lambda_4 = 0$ .  
 $\rho(M) = 0.369 < 1$ , deci procesul este convergent.

Pe componente algoritmul conduce la:

Iterația I

$$\begin{cases} x_1^{(1)} = \frac{1}{a_{11}}(b - a_{12}x_2^{(0)} - a_{13}x_3^{(0)} - a_{14}x_4^{(0)}) \\ x_2^{(1)} = \frac{1}{a_{22}}(b - a_{21}x_1^{(1)} - a_{23}x_3^{(0)} - a_{24}x_4^{(0)}) \\ x_3^{(1)} = \frac{1}{a_{33}}(b - a_{31}x_1^{(1)} - a_{32}x_2^{(1)} - a_{34}x_4^{(0)}) \\ x_4^{(1)} = \frac{1}{a_{44}}(b - a_{41}x_1^{(1)} - a_{42}x_2^{(1)} - a_{43}x_3^{(1)}) \end{cases} \text{rezultă } x^{(1)} = \begin{pmatrix} 0.5 \\ 0.5 \\ 0.5 \\ 0.75 \end{pmatrix}$$

Iterația a II –a

$$\begin{cases} x_1^{(2)} = \frac{1}{a_{11}}(b - a_{12}x_2^{(1)} - a_{13}x_3^{(1)} - a_{14}x_4^{(1)}) \\ x_2^{(2)} = \frac{1}{a_{22}}(b - a_{21}x_1^{(2)} - a_{23}x_3^{(1)} - a_{24}x_4^{(1)}) \\ x_3^{(2)} = \frac{1}{a_{33}}(b - a_{31}x_1^{(2)} - a_{32}x_2^{(2)} - a_{34}x_4^{(1)}) \\ x_4^{(2)} = \frac{1}{a_{44}}(b - a_{41}x_1^{(2)} - a_{42}x_2^{(2)} - a_{43}x_3^{(2)}) \end{cases} \text{rezultă } x^{(2)} = \begin{pmatrix} 0.75 \\ 0.75 \\ 0.83333 \\ 0.91667 \end{pmatrix}$$

Iterația a III –a

$$\begin{cases} x_1^{(3)} = \frac{1}{a_{11}}(b - a_{12}x_2^{(2)} - a_{13}x_3^{(2)} - a_{14}x_4^{(2)}) \\ x_2^{(3)} = \frac{1}{a_{22}}(b - a_{21}x_1^{(3)} - a_{23}x_3^{(2)} - a_{24}x_4^{(2)}) \\ x_3^{(3)} = \frac{1}{a_{33}}(b - a_{31}x_1^{(3)} - a_{32}x_2^{(3)} - a_{34}x_4^{(2)}) \\ x_4^{(3)} = \frac{1}{a_{44}}(b - a_{41}x_1^{(3)} - a_{42}x_2^{(3)} - a_{43}x_3^{(3)}) \end{cases} \text{rezultă } x^{(3)} = \begin{pmatrix} 0.875 \\ 0.90278 \\ 0.93981 \\ 0.96991 \end{pmatrix}$$

Iterația a IV –a

$$\begin{cases} x_1^{(4)} = \frac{1}{a_{11}}(b - a_{12}x_2^{(3)} - a_{13}x_3^{(3)} - a_{14}x_4^{(3)}) \\ x_2^{(4)} = \frac{1}{a_{22}}(b - a_{21}x_1^{(4)} - a_{23}x_3^{(3)} - a_{24}x_4^{(3)}) \\ x_3^{(4)} = \frac{1}{a_{33}}(b - a_{31}x_1^{(4)} - a_{32}x_2^{(4)} - a_{34}x_4^{(3)}) \\ x_4^{(4)} = \frac{1}{a_{44}}(b - a_{41}x_1^{(4)} - a_{42}x_2^{(4)} - a_{43}x_3^{(4)}) \end{cases} \text{rezultă } x^{(4)} = \begin{pmatrix} 0.95139 \\ 0.96373 \\ 0.97788 \\ 0.98894 \end{pmatrix}$$

Iterația a V –a

$$\begin{cases} x_1^{(5)} = \frac{1}{a_{11}}(b - a_{12}x_2^{(4)} - a_{13}x_3^{(4)} - a_{14}x_4^{(4)}) \\ x_2^{(5)} = \frac{1}{a_{22}}(b - a_{21}x_1^{(5)} - a_{23}x_3^{(4)} - a_{24}x_4^{(4)}) \\ x_3^{(5)} = \frac{1}{a_{33}}(b - a_{31}x_1^{(5)} - a_{32}x_2^{(5)} - a_{34}x_4^{(4)}) \\ x_4^{(5)} = \frac{1}{a_{44}}(b - a_{41}x_1^{(5)} - a_{42}x_2^{(5)} - a_{43}x_3^{(5)}) \end{cases} \text{ rezultă } x^{(5)} = \begin{pmatrix} 0.98187 \\ 0.98658 \\ 0.99184 \\ 0.99592 \end{pmatrix}.$$

**Teorema 1.** Dacă matricea  $A$  este simetrică și pozitiv definită, procesul iterativ Gauss–Seidel este convergent.

Demonstrația rezultă din analiza descreșterii funcției pătratice  $F$  prin trecerea de la o iterație la alta. O altă demonstrație se bazează pe faptul că se poate arăta că dacă  $A$  este simetrică și pozitiv definită, atunci  $\rho(M) < 1$  și conform Teoremei 2 din §7, procesul iterativ este convergent.

### §1.11. Metoda suprarelaxării

Pentru sisteme mari de ecuații, procesul iterativ Gauss–Seidel converge lent, deoarece raza spectrală  $\rho(M)$  este în vecinătatea lui 1.

Metoda suprarelaxării este o generalizare a metodei Gauss–Seidel, care constă în introducerea unui parametru  $\omega$  în vederea accelerării convergenței. Ca și în metoda Gauss–Seidel, direcția de relaxare urmează ciclic direcțiile  $e_1, e_2, \dots, e_n$  dar se înlocuiește  $t_{\min}$  cu  $t = \omega t_{\min}$ . Exemplificăm metoda pe cazul particular al unui sistem de trei ecuații. Fie  $v^{(0)}$  vectorul de probă inițial. După un ciclu în care direcția de relaxare urmează succesiv direcțiile  $e_1, e_2, e_3$  obținem vectorul  $v^{(1)}$  de componente:

$$\begin{cases} v_1^{(1)} = v_1^{(0)} - \frac{\omega}{a_{11}}(a_{11}v_1^{(0)} + a_{12}v_2^{(0)} + a_{13}v_3^{(0)} - b_1) \\ v_2^{(1)} = v_2^{(0)} - \frac{\omega}{a_{22}}(a_{21}v_1^{(1)} + a_{22}v_2^{(0)} + a_{23}v_3^{(0)} - b_2) \\ v_3^{(1)} = v_3^{(0)} - \frac{\omega}{a_{33}}(a_{31}v_1^{(1)} + a_{32}v_2^{(1)} + a_{33}v_3^{(0)} - b_3) \end{cases} \quad (1)$$

Dacă  $\omega = 1$ , obținem din nou formulele (1) din §10. După efectuarea calculelor rezultă:

$$\begin{cases} \omega^{-1}a_{11}v_1^{(1)} + (1-\omega^{-1})a_{11}v_1^{(0)} + a_{12}v_2^{(0)} + a_{13}v_3^{(0)} = b_1 \\ a_{21}v_1^{(1)} + a_{22}\omega^{-1}v_2^{(1)} + (1-\omega^{-1})a_{22}v_2^{(0)} + a_{23}v_3^{(0)} = b_2 \\ a_{31}v_1^{(1)} + a_{32}v_2^{(1)} + a_{33}\omega^{-1}v_3^{(1)} + (1-\omega^{-1})a_{33}v_3^{(0)} = b_3 \end{cases} \quad (2)$$

Dacă introducem notațiile

$$E = \begin{pmatrix} 0 & 0 & 0 \\ a_{21} & 0 & 0 \\ a_{31} & a_{32} & 0 \end{pmatrix}, \quad D = \begin{pmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & 0 \\ 0 & 0 & a_{33} \end{pmatrix} \quad \text{și } F = E^T,$$

relațiile (2) capătă forma matricială

$$(E + \omega^{-1}D)v^{(1)} + [F + (1 - \omega^{-1})D]v^{(0)} = b. \quad (3)$$

În general, pentru un sistem de  $n$  ecuații, șirul de vectori  $v^{(m)}$  satisface relația:

$$(E + \omega^{-1}D)v^{(m+1)} + [F + (1 - \omega^{-1})D]v^{(m)} = b. \quad (4)$$

În continuare avem

$$v^{(m+1)} = -(E + \omega^{-1}D)^{-1} [F + (1 - \omega^{-1})D] v^{(m)} + (E + \omega^{-1}D)^{-1} b.$$

Notăm cu

$$\begin{cases} M(\omega) = -(E + \omega^{-1}D)^{-1} [F + (1 - \omega^{-1})D] \\ C(\omega) = [E + \omega^{-1}D]^{-1} b \end{cases}. \quad (5)$$

Obținem astfel procesul iterativ

$$v^{(m+1)} = M(\omega)v^{(m)} + C(\omega). \quad (6)$$

Pentru  $\omega = 1$  obținem algoritmul Gauss –Seidel (vezi (4), (5), (6) din §10).

Parametrul optim,  $\omega_{\text{opt}}$ , va fi acela pentru care raza spectrală a matricei  $M(\omega)$  va fi minimă. Evident, pentru acest parametru se obține cea mai rapidă convergență.

Se poate demonstra următoarea teoremă.

**Teorema 1.** *Dacă matricea  $A$  este simetrică și pozitiv definită, metoda suprarelaxării este convergentă pentru orice  $0 < \omega < 2$ .*

În particular, rezultă că metoda Gauss –Seidel este convergentă dacă  $A$  este simetrică și pozitiv definită, deoarece corespunde cazului particular  $\omega = 1$ .

Determinarea parametrului optim,  $\omega_{\text{opt}}$ , este posibilă în cazul *matricelor bloc tridiagonale*.

**Definiția 1.** *O matrice  $A$  se numește bloc tridiagonală, dacă are următoarea structură:*

$$A = \begin{pmatrix} D_1 & F_1 & 0 & 0 & \cdots & & 0 \\ E_1 & D_2 & F_2 & 0 & \cdots & & 0 \\ 0 & E_2 & D_3 & F_3 & \cdots & & 0 \\ \vdots & & & & \ddots & & \vdots \\ 0 & & & & \cdots & E_{m-2} & D_{m-1} & F_{m-1} \\ 0 & & & & \cdots & 0 & E_{m-1} & D_m \end{pmatrix},$$

unde  $D_i$  sunt matrice pătratice de diferite ordine,  $E_k$  și  $F_k$  sunt, în general, matrice dreptunghiulare.  $F_k$  are același număr de linii ca matricea  $D_k$  și același număr de coloane ca matricea  $D_{k+1}$ .  $E_k$  are același număr de linii cu  $D_{k+1}$  și același număr de coloane cu  $D_k$ . În afara matricelor care intră în bandă, toate elementele sunt nule. Dacă, în plus, matricele  $D_i$  sunt diagonale,  $A$  se numește diagonal bloc tridiagonală.

Prezentăm de asemenea fără demonstrație următoarea teoremă.

**Teorema 2.** Fie  $A$  o matrice simetrică, pozitiv definită și diagonal bloc tridiagonală.

Atunci parametrul optim de relaxare este dat de relația  $\omega_{\text{opt}} = \frac{2}{1 + \sqrt{1 - \lambda_1^2}}$ , unde

$\lambda_1$  este cea mai mare valoare proprie a matricei  $-D^{-1}(E+F)$ .

**Exemplu.** Fie sistemul  $Ax = b$  din exemplul din paragraful precedent, care are matricea diagonal bloc tridiagonală

$$A = \begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 3 & -1 & 0 \\ 0 & -1 & 3 & -1 \\ 0 & 0 & -1 & 2 \end{pmatrix}, \quad E + F = \begin{pmatrix} 0 & -1 & 0 & 0 \\ -1 & 0 & -1 & 0 \\ 0 & -1 & 0 & -1 \\ 0 & 0 & -1 & 0 \end{pmatrix},$$

$$D = \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix}, \quad -D^{-1}(E+F) = \begin{pmatrix} 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & 0 \\ 0 & \frac{1}{3} & 0 & \frac{1}{3} \\ 0 & 0 & \frac{1}{2} & 0 \end{pmatrix}.$$

Ecuația caracteristică este

$$\begin{vmatrix} -\lambda & \frac{1}{2} & 0 & 0 \\ \frac{1}{3} & -\lambda & \frac{1}{3} & 0 \\ 0 & \frac{1}{3} & -\lambda & \frac{1}{3} \\ 0 & 0 & \frac{1}{2} & -\lambda \end{vmatrix} = 0,$$

care este echivalentă cu

$$\begin{vmatrix} 2\lambda & -1 & 0 & 0 \\ -1 & 3\lambda & -1 & 0 \\ 0 & -1 & 3\lambda & -1 \\ 0 & 0 & -1 & 2\lambda \end{vmatrix} = 36\lambda^4 + 16\lambda^2 + 1 = 0.$$

Rezultă  $\lambda_i = \pm \sqrt{\frac{4 \pm \sqrt{7}}{18}}$  și  $\lambda_1 = 0.60763$ .

Parametrul optim de relaxare  $\omega_{opt} = \frac{2}{1 + \sqrt{1 - \lambda_1^2}} = 1.11469$ .

Procedeul iterativ  $x^{(m+1)} = M(\omega)x^{(m)} + C(\omega)$ , unde

$$M(\omega) = \begin{pmatrix} -0.11469 & 0.55734 & 0 & 0 \\ -0.04261 & 0.0924 & 0.37156 & 0 \\ -0.01583 & 0.03433 & 0.02337 & 0.37156 \\ -0.00882482 & 0.01913 & 0.01303 & 0.0924 \end{pmatrix}, \quad C(\omega) = \begin{pmatrix} 0.55734 \\ 0.57865 \\ 0.58657 \\ 0.88426 \end{pmatrix},$$

conduce la următoarele valori ale vectorului soluțiilor pentru primele cinci iterații:

$$x^{(1)} = \begin{pmatrix} 0.55734 \\ 0.57865 \\ 0.58657 \\ 0.88426 \end{pmatrix}, \quad x^{(2)} = \begin{pmatrix} 0.81593 \\ 0.82631 \\ 0.93988 \\ 0.97976 \end{pmatrix}, \quad x^{(3)} = \begin{pmatrix} 0.92431 \\ 0.96946 \\ 0.98803 \\ 0.99565 \end{pmatrix}, \quad x^{(4)} = \begin{pmatrix} 0.99166 \\ 0.99595 \\ 0.99825 \\ 0.99953 \end{pmatrix},$$

$$x^{(5)} = \begin{pmatrix} 0.9987 \\ 0.99933 \\ 0.99978 \\ 0.99993 \end{pmatrix}.$$

### §1.12. Metoda gradientilor conjugați

Fie sistemul  $Ax=b$ , unde  $A$  este simetrică și pozitiv definită.

**Definiția 1.** Spunem că direcțiile  $p$  și  $q$  sunt direcții conjugate în raport cu matricea  $A$  dacă  $\langle Ap, q \rangle = \langle p, Aq \rangle = 0$ .

Fie  $v^{(0)}$  un vector de probă și  $r^{(0)} = Av^{(0)} - b$  vectorul rezidual corespunzător. În metoda gradientilor conjugați pentru rezolvarea sistemului  $Ax=b$ , prima direcție de relaxare se alege  $p^{(1)} = -r^{(0)}$ . În continuare avem:

$$t_{\min} = -\frac{\langle r^{(0)}, p^{(1)} \rangle}{\langle Ap^{(1)}, p^{(1)} \rangle} = \frac{\langle r^{(0)}, r^{(0)} \rangle}{\langle Ap^{(1)}, p^{(1)} \rangle}, \quad (1)$$

$$v^{(1)} = v^{(0)} + q_1 p^{(1)}, \quad (2)$$

unde

$$q_1 = t_{\min} = -\frac{\langle r^{(0)}, p^{(1)} \rangle}{\langle Ap^{(1)}, p^{(1)} \rangle} \quad (3)$$

Valoarea minimă a funcției  $F=F(v)$ , când  $v$  parcurge dreapta ce trece prin  $v^{(0)}$  și are direcția  $p^{(1)} = -r^{(0)}$  este  $v^{(1)}$ .

Fie

$$r^{(1)} = \text{grad}F(v^{(1)}) = Av^{(1)} - b.$$

Din Observația 1 din §8 rezultă că

$$\langle r^{(1)}, p^{(1)} \rangle = -\langle r^{(1)}, r^{(0)} \rangle = 0. \quad (4)$$

Următoarea direcție de relaxare  $p^{(2)}$  se alege de forma  $p^{(2)} = -r^{(1)} + c_1 p^{(1)}$  și în plus să fie conjugată direcției  $p^{(1)}$  în raport cu  $A$ .

Așadar, avem:

$$0 = \langle Ap^{(2)}, p^{(1)} \rangle = \langle p^{(2)}, Ap^{(1)} \rangle = -\langle r^{(1)}, Ap^{(1)} \rangle + c_1 \langle Ap^{(1)}, p^{(1)} \rangle$$

de unde rezultă

$$c_1 = \frac{\langle r^{(1)}, Ap^{(1)} \rangle}{\langle Ap^{(1)}, p^{(1)} \rangle}. \quad (5)$$

Avem de asemenea

$$v^{(2)} = v^{(1)} + t_{\min} p^{(2)} = v^{(1)} - \frac{\langle r^{(1)}, p^{(2)} \rangle}{\langle Ap^{(2)}, p^{(2)} \rangle} p^{(2)} . \quad (6)$$

În general, pentru orice  $k \geq 2$  obținem

$$c_{k-1} = \frac{\langle r^{(k-1)}, Ap^{(k-1)} \rangle}{\langle Ap^{(k-1)}, p^{(k-1)} \rangle} , \quad (7)$$

$$p^{(k)} = -r^{(k-1)} + c_{k-1} p^{(k-1)} , \quad (8)$$

$$q_k = -\frac{\langle r^{(k-1)}, p^{(k)} \rangle}{\langle Ap^{(k)}, p^{(k)} \rangle} , \quad (9)$$

$$v_k = v^{(k-1)} + q_k p^{(k)} . \quad (10)$$

Metoda gradientilor conjugați este definită de formulele (7) – (10).

În continuare prezentăm unele simplificări și proprietăți suplimentare.

Deoarece  $r^{(k-1)}$  este ortogonal pe direcția  $p^{(k-1)}$  rezultă

$$\langle r^{(k-1)}, p^{(k)} \rangle = \langle r^{(k-1)}, -r^{(k-1)} + c_{k-1} p^{(k-1)} \rangle = -\langle r^{(k-1)}, r^{(k-1)} \rangle$$

și deci

$$q_k = \frac{\langle r^{(k-1)}, r^{(k-1)} \rangle}{\langle Ap^{(k)}, p^{(k)} \rangle} > 0 . \quad (9')$$

Pe de altă parte, din (10) avem

$$r^{(k)} = Av^{(k)} - b = Av^{(k-1)} + q_k Ap^{(k)} - b = r^{(k-1)} + q_k Ap^{(k)} .$$

Obținem deci următoarea relația de recurență

$$r^{(k)} = r^{(k-1)} + q_k Ap^{(k)} . \quad (11)$$

Observăm că

$$\langle r^{(k)}, r^{(k-1)} \rangle = 0 . \quad (12)$$

Într-adevăr, din (11) rezultă

$$\langle r^{(k)}, r^{(k-1)} \rangle = \langle r^{(k-1)}, r^{(k-1)} \rangle + q_k \langle r^{(k-1)}, Ap^{(k)} \rangle . \quad (13)$$

Pe de altă parte, ținând seama de (8), de (9') și de faptul că  $p^{(k)}$  și  $p^{(k-1)}$  sunt  $A$ -conjugate, rezultă

$$q_k \langle r^{(k-1)}, Ap^{(k)} \rangle = \langle r^{(k-1)}, r^{(k-1)} \rangle \frac{\langle r^{(k-1)}, Ap^{(k)} \rangle}{\langle Ap^{(k)}, -r^{(k-1)} + c_{k-1} p^{(k-1)} \rangle} = -\langle r^{(k-1)}, r^{(k-1)} \rangle$$

Din (13) și (14) rezultă acum (12) .

Deoarece  $Ap^{(k)}$  oricum trebuie calculat, rezultă că vectorul rezidual  $r^{(k)}$  se va calcula din relația de recurență (11) și nu prin înlocuirea directă a lui  $v^{(k)}$  în expresia  $Av=b$ .

În continuare vom stabili o altă formulă pentru coeficientul  $c_{k-1}$ .

Din (11) și (12) rezultă

$$\langle r^{(k-1)}, Ap^{(k-1)} \rangle = \left\langle r^{(k-1)}, \frac{1}{q_{k-1}} (r^{(k-1)} - r^{(k-2)}) \right\rangle = \frac{1}{q_{k-1}} \langle r^{(k-1)}, r^{(k-1)} \rangle.$$

Ținând seama de (7) și de (9') obținem

$$c_{k-1} = \frac{\langle r^{(k-1)}, Ap^{(k-1)} \rangle}{\langle Ap^{(k-1)}, p^{(k-1)} \rangle} = \frac{\langle r^{(k-1)}, r^{(k-1)} \rangle}{\langle r^{(k-2)}, r^{(k-2)} \rangle}.$$

*Algoritm pentru rezolvarea sistemelor de ecuații liniare cu metoda gradientilor conjugați*

Calculează

$$r^{(0)} = Av^{(0)} - b; \quad p^{(1)} = -r^{(0)};$$

$$q_1 = \frac{\langle r^{(0)}, r^{(0)} \rangle}{\langle Ap^{(1)}, p^{(1)} \rangle}; \quad v^{(1)} = v^{(0)} + q_1 p^{(1)}; \quad r^{(1)} = r^{(0)} + q_1 Ap^{(1)};$$

Pentru  $k:=2, n$  calculează

$$c_{k-1} = \frac{\langle r^{(k-1)}, r^{(k-1)} \rangle}{\langle r^{(k-2)}, r^{(k-2)} \rangle}; \quad p^{(k)} = -r^{(k-1)} + c_{k-1} p^{(k-1)};$$

$$q_k = \frac{\langle r^{(k-1)}, r^{(k-1)} \rangle}{\langle Ap^{(k)}, p^{(k)} \rangle}; \quad v^{(k)} = v^{(k-1)} + q_k p^{(k)}; \quad r^{(k)} = r^{(k-1)} + q_k Ap^{(k)};$$

sfârșit pentru  $k$ .

În Mathcad algoritmul de mai sus este aplicat unui exemplu.

**Metoda gradientilor conjugati**

Folosind metoda gradientilor conjugati sa se gaseasca solutia sistemului de ecuatii liniare  $Ax=b$

$$A := \begin{bmatrix} 5 & 1 & -1 & 1 \\ 1 & 6 & 2 & -2 \\ -1 & 2 & 7 & 1 \\ 1 & -2 & 1 & 8 \end{bmatrix} \quad b := \begin{bmatrix} 6 \\ 7 \\ 9 \\ 8 \end{bmatrix} \quad \text{Vectorul de proba} \quad x := \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad n := 4$$

Algoritmul metodei gradientilor conjugati

$$\text{GrCon}(A, b, x, n) := \left. \begin{array}{l} r^{<0>} \leftarrow A \cdot x - b \\ p \leftarrow -r^{<0>} \\ q \leftarrow \frac{r^{<0>} \cdot r^{<0>}}{A \cdot p \cdot p} \\ x \leftarrow x + q \cdot p \\ r^{<1>} \leftarrow r^{<0>} + q \cdot A \cdot p \\ \text{for } k \in 2..n \\ \quad \left. \begin{array}{l} c \leftarrow \frac{r^{<k-1>} \cdot r^{<k-1>}}{r^{<k-2>} \cdot r^{<k-2>}} \\ p \leftarrow -r^{<k-1>} + c \cdot p \\ q \leftarrow \frac{r^{<k-1>} \cdot r^{<k-1>}}{A \cdot p \cdot p} \\ x \leftarrow x + q \cdot p \\ r^{<k>} \leftarrow r^{<k-1>} + q \cdot A \cdot p \end{array} \right\} \\ x \end{array} \right|$$

Apelarea programului si afisarea rezultatelor

$$\text{GrCon}(A, b, x, n) = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

**Teorema 1.** În metoda gradientilor conjugați direcțiile de relaxare  $p^{(k)}$ , ( $k=1,2,\dots$ ) sunt conjugate două câte două în raport cu matricea  $A$ , iar vectorii reziduali  $r^{(k)}$ , ( $k=0,1,\dots$ ) sunt ortogonali doi câte doi.

**Demonstrație.**

Demonstrația se face prin inducție relativ la  $k$ . Pentru  $k=1$  avem  $\langle r^{(1)}, r^{(0)} \rangle = 0$

din (4), iar pentru prima afirmație nu avem ce arăta. Ipoteza de inducție este:

$$\langle r^{(i)}, r^{(j)} \rangle = 0 \quad \text{pentru } i \neq j, \quad 0 \leq i, j \leq k, \quad (15)$$

$$\langle p^{(i)}, Ap^{(j)} \rangle = 0 \quad \text{pentru } i \neq j, \quad 1 \leq i \leq j \leq k. \quad (16)$$

Va trebui să arătăm că

$$\langle r^{(k+1)}, r^{(j)} \rangle = 0, \quad \text{pentru } j = \overline{0, k} \quad (17)$$

$$\langle p^{(k+1)}, Ap^{(j)} \rangle = 0, \quad \text{pentru } j = \overline{1, k}. \quad (18)$$

Fie  $j=k$ , atunci

$$\langle p^{(k+1)}, Ap^{(k)} \rangle = \langle Ap^{(k+1)}, p^{(k)} \rangle = 0$$

deoarece  $p^{(k)}$  și  $p^{(k+1)}$  sunt  $A$ -conjugate.

Fie  $1 \leq j < k$ , atunci

$$\langle p^{(k+1)}, Ap^{(j)} \rangle = \langle -r^{(k)} + c_k p^{(k)}, Ap^{(j)} \rangle = -\langle r^{(k)}, Ap^{(j)} \rangle,$$

deoarece  $\langle p^{(k)}, Ap^{(j)} \rangle = 0$  conform ipotezei (16).

Pe de altă parte,

$$Ap^{(j)} = \frac{1}{q_j} (r^{(j)} - r^{(j-1)})$$

și din (15) rezultă  $\langle r^{(k)}, Ap^{(j)} \rangle = 0$ .

Așadar am demonstrat (18).

Pentru  $j=k$ , (17) este adevărată din (12).

Fie  $0 \leq j < k$ . Din (11) și (15) rezultă:

$$\begin{aligned} \langle r^{(k+1)}, r^{(j)} \rangle &= \langle r^{(k)} + q_{k+1} Ap^{(k+1)}, r^{(j)} \rangle = q_{k+1} \langle Ap^{(k+1)}, r^{(j)} \rangle = \\ &= q_{k+1} \langle Ap^{(k+1)}, -p^{(j)} + c_{j-1} p^{(j-1)} \rangle \end{aligned}$$

Ținând seama de (18) și de faptul că  $A$  este simetrică, rezultă  $\langle r^{(k+1)}, r^{(j)} \rangle = 0$  și cu aceasta teorema este demonstrată.  $\square$

Din Teorema 1 rezultă că vectorii reziduali  $r^{(k)}$  sunt ortogonali doi câte doi și deci sunt liniar independenți (dacă sunt nenuli). Așadar, nu pot exista  $(n+1)$  vectori reziduali nenuli. Rezultă că în metoda gradientilor conjugați soluția exactă



În general, sistemul (1) nu este compatibil și  $\min_{x \in \mathbb{R}^n} f(x) = f(x^*) > 0$ , iar  $x = x^*$  este un substitut pentru soluția sistemului și anume soluția în sensul celor mai mici pătrate.

Funcția  $f$  se poate pune sub forma

$$f(x) = \langle r, r \rangle = \langle Ax - b, Ax - b \rangle = \langle Ax, Ax \rangle - 2\langle Ax, b \rangle + \langle b, b \rangle$$

și mai departe

$$f(x) = \langle A^T Ax, x \rangle - 2\langle A^T b, x \rangle + \langle b, b \rangle \quad (4)$$

**Teorema 1.** Dacă  $\text{rang} A = n$ , atunci sistemul (1) admite o singură soluție în sensul celor mai mici pătrate și aceasta este soluția (unică) a sistemului.

$$A^T Ax = A^T b \quad (5)$$

(Sistemul (5) se numește sistemul normal al lui Gauss).

**Demonstrație.**

Punctele de extrem ale funcției pătratice  $f$  dată de (4), se caută printre punctele sale critice, iar acestea, se află rezolvând sistemul:

$$\text{grad} f = 0$$

Cum  $\text{grad} f = A^T Ax - A^T b$ , obținem sistemul  $A^T Ax = A^T b$ . Pe de altă parte se știe că:

$$\text{rang} A = \text{rang} A^T = \text{rang} \begin{pmatrix} A^T & A \end{pmatrix} = \text{rang} \begin{pmatrix} A & A^T \end{pmatrix}.$$

Matricea  $B = A^T A$  este o matrice pătratică de ordinul  $n$  și  $\text{rang} B = n$ , conform celor de mai sus. Rezultă că sistemul (5) admite o soluție unică,  $x = x^*$ , care este punct critic pentru  $f$ .

Matricea  $B$  este evident simetrică și semipozitiv definită. Mai mult, în ipoteza noastră, matricea  $B$  este pozitiv definită. Într-adevăr, dacă presupunem că  $\langle Bx, x \rangle = 0$ , atunci rezultă  $\langle Ax, Ax \rangle = 0$  și deci  $Ax = 0$ . Cum  $\text{rang} A = n < m$  rezultă  $x = 0$ .

Pe de altă parte avem

$$d^2 f(x) = \sum_{i=1}^n \sum_{j=1}^n b_{ij} dx_i dx_j > 0,$$

de unde rezultă că  $x = x^*$  este punct de minim pentru  $f$  și cu aceasta teorema este demonstrată.  $\square$

Așadar, în ipoteza  $\text{rang} A = n$ , soluția sistemului (1), în sensul celor mai mici pătrate, este unică și se află rezolvând sistemul (5). Acest sistem este simetric pozitiv definit. Rezolvarea sa se poate face prin metoda Cholesky sau una din metodele de relaxare.

**Observația 1.** Teoretic, soluția sistemului (5) este  $x^* = (A^T A)^{-1} A^T b$ . Matricea  $P = (A^T A)^{-1} A^T$  se numește pseudoinversa matricei (dreptunghiulară)  $A$ .

Se observă că dacă  $A$  este pătratică, atunci  $P=A^{-1}(A^T)^{-1}A^T=A^{-1}$ , deci noțiunea de matrice pseudoinversă generalizează noțiunea de matrice inversă (pentru matrice dreptunghiulare).

Rezolvarea practică a sistemului (5) ridică probleme din cauza faptului că numărul de condiționare al matricei  $B=A^T A$  este mare. Fie  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n > 0$  valorile proprii ale matricei  $B$ . Atunci:

$$\text{cond}(B) = \frac{\lambda_1}{\lambda_n} . \quad (6)$$

Cum

$$\lambda_1 = \sup_{x \neq 0} \frac{\langle Bx, x \rangle}{\langle x, x \rangle} \geq \max_i \langle B e_i, e_i \rangle = \max_i b_{ii} \quad \text{și} \quad \lambda_n = \inf_{x \neq 0} \frac{\langle Bx, x \rangle}{\langle x, x \rangle} \leq \min_i \langle B e_i, e_i \rangle = \min_i b_{ii} ,$$

rezultă

$$\text{cond}(B) \geq \frac{\max b_{ii}}{\min b_{ii}} . \quad (7)$$

### Exemplul 1. (Dreapta de regresie)

Să presupunem că vrem să găsim o dreaptă  $y=mx+n$  care să treacă prin punctele:

$$M_1(0,0) ; M_2(2,1) ; M_3(5,3) ; M_4(8,5) ; M_6(10,6)$$

Se obține astfel sistemul

$$\begin{cases} 0 \cdot m + n = 0 \\ 2 \cdot m + n = 1 \\ 5 \cdot m + n = 3 \\ 8 \cdot m + n = 5 \\ 10 \cdot m + n = 6 \end{cases} . \quad (8)$$

Evident, sistemul (8) este supradimensionat și incompatibil. Avem:

$$A = \begin{pmatrix} 0 & 1 \\ 2 & 1 \\ 5 & 1 \\ 8 & 1 \\ 10 & 1 \end{pmatrix} ; b = \begin{pmatrix} 0 \\ 1 \\ 3 \\ 5 \\ 6 \end{pmatrix} ; B = A^T A = \begin{pmatrix} 193 & 25 \\ 25 & 5 \end{pmatrix} ; A^T b = \begin{pmatrix} 117 \\ 15 \end{pmatrix} .$$

Ecuțiile normale ale lui Gauss sunt

$$\begin{cases} 193m + 25n = 117 \\ 25m + 5n = 15 \end{cases} .$$

Soluția exactă este  $m = \frac{21}{34}, n = -\frac{3}{34}$ , iar valorile proprii sunt  $\lambda_1 = 196.268$  și

$\lambda_2 = 1.732$ . Rezultă  $\text{cond}(B) \cong 113$ . Dacă folosim estimarea (7) obținem





Folosind metoda Gauss să se rezolve următoarele sisteme de ecuații liniare:

$$1. \quad \begin{cases} 6x_1 + x_2 - x_3 + x_4 = 2 \\ x_1 + 7x_2 + 2x_3 - x_4 = -1 \\ -x_1 + 2x_2 + 8x_3 + x_4 = 12 \\ x_1 - x_2 + x_3 + 9x_4 = -5 \end{cases}$$

$$R. \text{ Matricea triunghiulară } \tilde{A} = \begin{pmatrix} 1.0 & 0.1667 & -0.1667 & 0.1667 \\ 0 & 1.0 & 0.3171 & -0.1707 \\ 0 & 0 & 1.0 & 0.2150 \\ 0 & 0 & 0 & 1.0 \end{pmatrix},$$

$$\text{vectorul termenilor liberi transformat } \tilde{b} = \begin{pmatrix} 0.333 \\ -0.1951 \\ 1.7850 \\ -1.0 \end{pmatrix}.$$

$$\text{Soluția sistemului } x = \begin{pmatrix} 1 \\ -1 \\ 2 \\ -1 \end{pmatrix}.$$

$$2. \quad \begin{cases} 4x_1 + x_2 - x_3 + x_4 = 1 \\ x_1 + 8x_2 + x_3 + x_4 = 7 \\ x_1 + x_2 + 6x_3 + x_4 = -10 \\ x_1 + x_2 + x_3 + 7x_4 = 12 \end{cases}$$

$$R. \text{ Matricea triunghiulară } \tilde{A} = \begin{pmatrix} 1.0 & 0.2500 & -0.2500 & 0.2500 \\ 0 & 1.0 & 0.1613 & 0.0968 \\ 0 & 0 & 1.0 & 0.1105 \\ 0 & 0 & 0 & 1.0 \end{pmatrix},$$

$$\text{vectorul termenilor liberi transformat } \tilde{b} = \begin{pmatrix} 0.2500 \\ 0.8710 \\ -1.7789 \\ 2.0 \end{pmatrix}.$$

$$\text{Soluția sistemului } x = (-1 \ 1 \ -2 \ 2)'$$

$$3. \quad \begin{cases} 5x_1 + x_2 + x_3 + x_4 = 14 \\ x_1 + 8x_2 + 2x_3 + x_4 = 127 \\ -x_1 + x_2 + 6x_3 - x_4 = 15 \\ x_1 + x_2 + x_3 + x_4 = 10 \end{cases}$$

$$R. \text{ Matricea triunghiulară } \tilde{A} = \begin{pmatrix} 1.0 & 0.2000 & -0.2000 & 0.2000 \\ 0 & 1.0 & 0.2308 & 0.1026 \\ 0 & 0 & 1.0 & -0.1558 \\ 0 & 0 & 0 & 1.0 \end{pmatrix},$$

$$\text{vectorul termenilor liberi transformat } \tilde{b} = \begin{pmatrix} 2.8000 \\ 3.1026 \\ 2.3766 \\ 4.0 \end{pmatrix}.$$

Soluția sistemului  $x = (1 \ 2 \ 3 \ 4)'$ .

Să se rezolve următoarele sisteme de ecuații liniare folosind metoda Cholesky:

$$4. \quad \begin{cases} 10x_1 + x_2 - x_3 - x_4 = 9 \\ x_1 + 9x_2 + x_3 - x_4 = -6 \\ -x_1 + x_2 + 11x_3 + x_4 = 8 \\ -x_1 - x_2 + x_3 + 8x_4 = -7 \end{cases}$$

$$R. \quad R^T = \begin{pmatrix} 3.1623 & 0 & 0 & 0 \\ 0.3162 & 2.9833 & 0 & 0 \\ -0.3162 & 0.3687 & 0.32809 & 0 \\ -0.3162 & -0.3017 & 0.3082 & 2.7774 \end{pmatrix}. \quad \text{Soluția sistemului}$$

$$R^T y = b \quad \text{este} \quad y = \begin{pmatrix} 2.846 \\ -2.3129 \\ 2.9726 \\ -2.7774 \end{pmatrix}, \quad \text{iar cea a sistemului } Rx=y, \quad \text{și deci soluția}$$

$$\text{sistemului inițial este } x = \begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix}.$$

$$5. \quad \begin{cases} 8x_1 - x_2 - x_3 + x_4 & = 7 \\ -x_1 + 6x_2 + x_3 - 2x_4 & = 6 \\ -x_1 + x_2 + 7x_3 + x_4 & = -8 \\ x_1 - 2x_2 + x_3 + 9x_4 & = -11 \end{cases}$$

$$R. \quad R^T = \begin{pmatrix} 2.82843 & 0 & 0 & 0 \\ -0.35355 & 2.42384 & 0 & 0 \\ -0.35355 & 0.361 & 2.59705 & 0 \\ 0.35355 & -0.77357 & 0.54071 & 2.82564 \end{pmatrix}.$$

Soluția sistemului  $R^T y = b$  este  $y = \begin{pmatrix} 2.47487 \\ 2.83641 \\ -3.13776 \\ -2.82564 \end{pmatrix}$  iar cea a sistemului  $Rx = y$ , și

deci soluția sistemului inițial este  $x = \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix}$ .

$$6. \quad \begin{cases} 6x_1 + x_2 - 2x_3 - x_4 & = 19 \\ x_1 + 5x_2 + x_3 - x_4 & = 15 \\ -2x_1 + x_2 + 11x_3 - x_4 & = 8 \\ -x_1 - x_2 - x_3 + 4x_4 & = -10 \end{cases}$$

$$R. \quad R^T = \begin{pmatrix} 2.44949 & 0 & 0 & 0 \\ 0.40825 & 2.19848 & 0 & 0 \\ -0.8165 & 0.60648 & 3.15682 & 0 \\ -0.40825 & -0.37905 & -0.34954 & 1.88878 \end{pmatrix}.$$

Soluția sistemului  $R^T y = b$  este  $y = \begin{pmatrix} 7.75672 \\ 5.3825 \\ 3.50636 \\ -1.88878 \end{pmatrix}$ , iar cea a sistemului  $Rx = y$ , și

deci soluția sistemului inițial este  $x = (3 \ 2 \ 1 \ -1)'$ .

Folosind metoda Householder, să se rezolve sistemele:

$$7. \begin{cases} 4x_1 + 3x_2 - 2x_3 + x_4 = 2 \\ 5x_1 + 6x_2 + 7x_3 - 8x_4 = -14 \\ 9x_1 - 8x_2 + 7x_3 - 6x_4 = -30 \\ x_1 + x_2 - x_3 - x_4 = 0 \end{cases}$$

R.  $\det(A) = -1112$ ; descompunerea  $A = QR$  este dată de

$$Q = \begin{pmatrix} 0.3607 & -0.3882 & 0.7877 & -0.3143 \\ 0.4508 & -0.7068 & -0.5445 & 0.0257 \\ 0.8115 & 0.5787 & -0.0784 & 0.02 \\ 0.0902 & -0.1217 & 0.2774 & 0.9487 \end{pmatrix},$$

$$R = \begin{pmatrix} 110905 & -26148 & 80249 & -80249 \\ 0 & -101569 & 00016 & 1.9155 \\ 0 & 0 & -6.2130 & 5.83368 \\ 0 & 0 & 0 & -1.58889 \end{pmatrix}.$$

Soluția sistemului  $Qy = b$  este  $y = \begin{pmatrix} 299354 \\ -8.2430 \\ 11.5498 \\ -1.5888 \end{pmatrix},$

iar cea sistemului  $Rx = y$  este  $x = (-1 \ 1 \ -1 \ 1)'$ .

$$8. \begin{cases} x_1 + 2x_2 - 3x_3 - 4x_4 = 16 \\ 5x_1 - 6x_2 + 7x_3 - 8x_4 = 2 \\ 9x_1 + 8x_2 + 7x_3 + 6x_4 = 6 \\ 4x_1 - 3x_2 + 2x_3 - x_4 = -2 \end{cases}$$

R.  $\det(A) = -2808$ ; descompunerea  $A = QR$  este dată

$$Q = \begin{pmatrix} 0.09017 & 0.17005 & 0.78843 & -0.58424 \\ 0.45083 & -0.71359 & -0.2494 & -0.47469 \\ 0.8115 & 0.55307 & -0.18504 & 0.03651 \\ 0.36067 & -0.39494 & 0.53098 & 0.65727 \end{pmatrix},$$

$$R = \begin{pmatrix} 11.09054 & 2.88534 & 9.2872 & 0.541 \\ 0 & 10.23107 & -2.42367 & 8.7419 \\ 0 & 0 & -4.3444 & -2.79972 \\ 0 & 0 & 0 & 5.69631 \end{pmatrix}.$$

Soluția sistemului  $Qy=b$  este  $y = \begin{pmatrix} 6.49202 \\ 5.40201 \\ 9.94384 \\ -11.39263 \end{pmatrix}$ ,

iar cea sistemului  $Rx=y$  este  $x = \begin{pmatrix} 1 \\ 2 \\ -1 \\ -2 \end{pmatrix}$ .

9. 
$$\begin{cases} 4x_1 + 3x_2 + 2x_3 + x_4 = 2 \\ x_1 + 2x_2 + 3x_3 + 4x_4 = -2 \\ 9x_1 - 8x_2 + 6x_3 - 7x_4 = 30 \\ x_1 - x_2 + x_3 - x_4 = 4 \end{cases}$$

R.  $\det(A) = -80$ ; descompunerea  $A=QR$  este dată de

$$Q = \begin{pmatrix} 0.40202 & -0.82257 & 0.3961 & 0.0697 \\ 0.1005 & -0.39663 & -0.90448 & -0.12039 \\ 0.90453 & 0.4028 & -0.05925 & -0.12673 \\ 0.1005 & 0.06173 & -0.14663 & 0.98213 \end{pmatrix},$$

$$R = \begin{pmatrix} 9.94987 & -5.92972 & 6.63325 & -5.62821 \\ 0 & -6.5451 & -0.3565 & -5.29041 \\ 0 & 0 & -2.42341 & -2.66043 \\ 0 & 0 & 0 & -0.50691 \end{pmatrix}.$$

Soluția sistemului  $Qy=b$  este  $y = \begin{pmatrix} 28.14106 \\ 11.47901 \\ 0.23702 \\ 0.50691 \end{pmatrix}$ ,

iar cea sistemului  $Rx=y$  este  $x = \begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix}$ .

10. Pentru matricea  $A = \begin{pmatrix} 1 & 2 & 1 & -1 \\ 2 & 1 & -1 & 3 \\ 2 & 1 & -1 & 1 \\ 1 & -1 & -1 & 1 \end{pmatrix}$ , să se calculeze:

- a)  $\det(A)$ ,  $A^{-1}$ ,  $\det(A^{-1})$ ;  
 b)  $\|A\|_1$ ,  $\|A\|_2$ ,  $\|A\|_\infty$ ;  $\|A^{-1}\|_1$ ,  $\|A^{-1}\|_2$ ,  $\|A^{-1}\|_\infty$ ;  
 c)  $\text{cond}_1(A)$ ,  $\text{cond}_2(A)$ ,  $\text{cond}_\infty(A)$ .

R. a)  $\det(A)=6$ ,  $A^{-1} = \begin{pmatrix} 0.6667 & 0 & -0.3333 & 1 \\ -0.3333 & 0 & 0.6667 & -1 \\ 1 & 0.5 & -1.5 & 1 \\ 0 & 0.5 & -0.5 & 0 \end{pmatrix}$ ,

$\det(A^{-1})=-0.1667$ ;

- b)  $\begin{cases} \|A\|_1 = 6, \|A\|_2 = 5.7446, \|A\|_\infty = 7; \\ \|A^{-1}\|_1 = 3, \|A^{-1}\|_2 = 2.848, \|A^{-1}\|_\infty = 4. \end{cases}$   
 c)  $\text{cond}_1(A)=18$ ,  $\text{cond}_2(A)=12.7511$ ,  $\text{cond}_\infty(A)=28$ .

11. Să se calculeze:

- a)  $\det(A)$ ,  $A^{-1}$ ,  $\det(A^{-1})$ ;  
 b)  $\|A\|_1$ ,  $\|A\|_2$ ,  $\|A\|_\infty$ ;  $\|A^{-1}\|_1$ ,  $\|A^{-1}\|_2$ ,  $\|A^{-1}\|_\infty$ ;  
 c)  $\text{cond}_1(A)$ ,  $\text{cond}_2(A)$ ,  $\text{cond}_\infty(A)$

pentru matricea  $A = \begin{pmatrix} 4 & -1 & 5 \\ 3 & 2 & -2 \\ 6 & 1 & 3 \end{pmatrix}$ .

R. a)  $\det(A)=8$ ,  $A^{-1} = \begin{pmatrix} 1 & 1 & -1 \\ -2.625 & -2.25 & 2.875 \\ -1.125 & -1.25 & 1.375 \end{pmatrix}$ ,  $\det(A^{-1})=0.125$ ;

- b)  $\begin{cases} \|A\|_1 = 6, \|A\|_2 = 5.7446, \|A\|_\infty = 7; \\ \|A^{-1}\|_1 = 3, \|A^{-1}\|_2 = 2.848, \|A^{-1}\|_\infty = 4. \end{cases}$

- c)  $\text{cond}_1(A)=68.25$ ,  $\text{cond}_2(A)=48.28525$ ,  $\text{cond}_\infty(A)=77.5$ .

Folosind metoda Iacobi să se găsească soluția aproximativă pentru următoarele sisteme de ecuații liniare :

$$12. \begin{cases} 5x_1 - x_2 - x_3 = 5 \\ x_1 + 6x_2 + x_3 = -4 \\ x_1 - x_2 + 7x_3 = 9 \end{cases}$$

R. Sistemul se poate pune sub forma  $x = Bx + c$ , unde  $B = D^{-1} \cdot (D - A)$ , iar  $D =$

$$\begin{pmatrix} 5 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 7 \end{pmatrix}, \quad E = D - A = \begin{pmatrix} 0 & 1 & 1 \\ -1 & 0 & -1 \\ -1 & 1 & 0 \end{pmatrix},$$

$$B = \begin{pmatrix} 0 & 0.2 & 0.2 \\ -0.16667 & 0 & -0.16667 \\ -0.14286 & 0.14286 & 0 \end{pmatrix}, \quad c = D^{-1} \cdot b = \begin{pmatrix} 1 \\ -0.66667 \\ 1.28571 \end{pmatrix}, \quad \|B\|_2 = 0.41997,$$

$\rho(B) = 0.29277$ . Soluția la fiecare iterație, pornind cu  $x^{(0)} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$ , este:

$$x^{(1)} = \begin{pmatrix} 1 \\ -0.66667 \\ 1.28571 \end{pmatrix}, \quad x^{(2)} = \begin{pmatrix} 1.12381 \\ -1.04762 \\ 1.04762 \end{pmatrix},$$

$$x^{(3)} = \begin{pmatrix} 1 \\ -1.02857 \\ 0.97551 \end{pmatrix}, \quad x^{(4)} = \begin{pmatrix} 0.98939 \\ -0.99592 \\ 0.99592 \end{pmatrix}.$$

Soluția exactă fiind  $x^* = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}$ , eroarea care se face dacă se reține ca soluție

aproximativă  $x^{(4)}$ , este  $\|x^* - x^{(4)}\| = 0.01208$ .

$$13. \begin{cases} 6x_1 + x_2 - x_3 + x_4 = 2 \\ x_1 + 7x_2 + x_3 - x_4 = -4 \\ x_1 + 2x_2 + 8x_3 + x_4 = 6 \\ x_1 + x_2 + x_3 + 9x_4 = -8 \end{cases} \quad (\text{să se scrie soluția obținută după patru}$$

iterații).

R. Se poate pune sistemul sub forma  $x = Bx + c$ , unde  $B = D^{-1} \cdot E$ , iar

$$D = \begin{pmatrix} 6 & 0 & 0 & 0 \\ 0 & 7 & 0 & 0 \\ 0 & 0 & 8 & 0 \\ 0 & 0 & 0 & 9 \end{pmatrix}, \quad E = D - A = \begin{pmatrix} 0 & -1 & 1 & 2 \\ -1 & 0 & -1 & 1 \\ -1 & -2 & 0 & -1 \\ 1 & 1 & 1 & 0 \end{pmatrix},$$

$$B = \begin{pmatrix} 0 & -0.16667 & 0.16667 & -0.33333 \\ -0.14286 & 0 & -0.14286 & 0.14286 \\ -0.125 & -0.25 & 0 & -0.125 \\ -0.11111 & -0.11111 & -0.11111 & 0 \end{pmatrix},$$

$$c = D^{-1} \cdot b = \begin{pmatrix} 0.33333 \\ -0.57143 \\ 0.75 \\ -0.88889 \end{pmatrix}, \quad \|B\|_2 = 0.5989, \quad \rho(B) = 0.27926. \quad \text{Soluția pentru primele}$$

$$4 \text{ iterații, pornind cu } x^{(0)} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \text{ este: } x^{(1)} = \begin{pmatrix} 0.33333 \\ -0.57143 \\ 0.75 \\ -0.88889 \end{pmatrix},$$

$$x^{(2)} = \begin{pmatrix} 0.84987 \\ -0.85317 \\ 0.9623 \\ -0.94577 \end{pmatrix}, \quad x^{(3)} = \begin{pmatrix} 0.95117 \\ -0.96542 \\ 0.97528 \\ -0.99544 \end{pmatrix}, \quad x^{(4)} = \begin{pmatrix} 0.9886 \\ -0.98884 \\ 0.99689 \\ -0.99567 \end{pmatrix}.$$

$$\text{Soluția exactă fiind } x^* = \begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix}, \text{ eroarea care se face dacă se reține ca soluție}$$

$$\text{aproximativă } x^{(4)}, \text{ este } \|x^* - x^{(4)}\| = 0.01682.$$

$$14. \quad \begin{cases} 4x_1 - x_2 - x_3 + x_4 = 3 \\ x_1 + 7x_2 + x_3 - x_4 = 14 \\ x_1 - x_2 + 6x_3 + x_4 = 21 \\ x_1 - x_2 - x_3 + 9x_4 = 38 \end{cases}$$

R. Se poate pune sistemul sub forma  $x = Bx + c$ , unde  $B = D^{-1} \cdot E$ , iar

$$D = \begin{pmatrix} 4 & 0 & 0 & 0 \\ 0 & 7 & 0 & 0 \\ 0 & 0 & 6 & 0 \\ 0 & 0 & 0 & 9 \end{pmatrix}, \quad E = D - A = \begin{pmatrix} 0 & 1 & 1 & -1 \\ -1 & 0 & -1 & 1 \\ -1 & 1 & 0 & -1 \\ -1 & 1 & -1 & 0 \end{pmatrix},$$

$$B = \begin{pmatrix} 0 & 0.25 & 0.25 & -0.25 \\ -0.14286 & 0 & -0.14286 & 0.14286 \\ -0.16667 & 0.16667 & 0 & -0.16667 \\ -0.11111 & 0.11111 & -0.11111 & 0 \end{pmatrix},$$

$$c = D^{-1} \cdot b = \begin{pmatrix} 0.75 \\ 2 \\ 3.5 \\ 4.22222 \end{pmatrix}, \quad \|B\|_2 = 0.60753, \quad \rho(B) = 0.22758. \text{ Pentru primele 4 iterații,}$$

$$\text{pornind cu } x^{(0)} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \text{ soluția este: } x^{(1)} = \begin{pmatrix} 0.75 \\ 2 \\ 3.5 \\ 4.22222 \end{pmatrix},$$

$$x^{(2)} = \begin{pmatrix} 1.06944 \\ 1.99603 \\ 3.00463 \\ 3.97222 \end{pmatrix}, \quad x^{(3)} = \begin{pmatrix} 1.00711 \\ 1.98545 \\ 2.99239 \\ 3.99133 \end{pmatrix}, \quad x^{(4)} = \begin{pmatrix} 0.99663 \\ 1.99883 \\ 2.99784 \\ 3.99844 \end{pmatrix}.$$

$$\text{Soluția exactă fiind } x^* = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}, \text{ eroarea care se face dacă se reține ca soluție}$$

aproximativă  $x^{(4)}$ , este  $\|x^* - x^{(4)}\| = 0.00417$ .

Să se rezolve cu metoda Gauss–Seidel următoarele sisteme:

$$15. \quad \begin{cases} 5x_1 - 2x_2 - 2x_3 = -3 \\ -x_1 + 6x_2 - x_3 = 2 \\ -x_1 - 3x_2 + 4x_3 = 8 \end{cases}$$

R. Se observă că matricea coeficienților sistemului  $A = \begin{pmatrix} 5 & -2 & -2 \\ -1 & 6 & -1 \\ -1 & -3 & 4 \end{pmatrix}$  este tare

diagonal dominantă și atunci algoritmul Gauss–Seidel este convergent.

Sistemul se poate pune sub forma  $x^{(m+1)} = (D+L)^{-1}(-Ux^{(m)}+b)$ ,  $m \geq 0$ , unde:

$$L = \begin{pmatrix} 0 & 0 & 0 \\ -1 & 0 & 0 \\ -1 & -3 & 0 \end{pmatrix}, \quad U = \begin{pmatrix} 0 & -2 & -2 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{pmatrix}, \quad D = \begin{pmatrix} 5 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 4 \end{pmatrix}.$$

Atunci

$$(D+L)^{-1} = \begin{pmatrix} 0.2 & 0 & 0 \\ 0.03333 & 0.16667 & 0 \\ 0.075 & 0.125 & 0.25 \end{pmatrix},$$

$$-(D+L)^{-1}U = \begin{pmatrix} 0 & 0.4 & 0.4 \\ 0 & 0.06667 & 0.23333 \\ 0 & 0.15 & 0.275 \end{pmatrix},$$

$$(D+L)^{-1}b = \begin{pmatrix} -0.6 \\ 0.23333 \\ 2.025 \end{pmatrix}. \text{ Se obțin pornind cu } x^{(0)} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \text{ vectorii:}$$

$$x^{(1)} = \begin{pmatrix} -0.6 \\ 0.23333 \\ 2.025 \end{pmatrix}, x^{(2)} = \begin{pmatrix} 0.30333 \\ 0.72139 \\ 2.61687 \end{pmatrix}, x^{(3)} = \begin{pmatrix} 0.73531 \\ 0.89203 \\ 2.85285 \end{pmatrix},$$

$$x^{(4)} = \begin{pmatrix} 0.89795 \\ 0.95847 \\ 2.94334 \end{pmatrix}.$$

$$16. \begin{cases} 6x_1 - x_2 - 2x_3 + x_4 & = 2 \\ 2x_1 + 6x_2 - x_3 - x_4 & = 8 \\ -x_1 - 3x_2 + 8x_3 + 2x_4 & = 2 \\ x_1 + x_2 + x_3 + 4x_4 & = -1 \end{cases}$$

R. Se observă că matricea coeficienților sistemului

$$A = \begin{pmatrix} 6 & -1 & -2 & 1 \\ 2 & 6 & -1 & -1 \\ -1 & -3 & 8 & 2 \\ 1 & 1 & 1 & 4 \end{pmatrix}$$

este tare diagonal dominantă și algoritmul Gauss–Seidel este convergent.

Sistemul se poate pune sub forma  $x^{(m+1)} = (D+L)^{-1}(-Ux^{(m)}+b)$ ,  $m \geq 0$ , unde:

$$L = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 \\ -1 & -3 & 0 & 0 \\ 1 & 1 & 1 & 0 \end{pmatrix}, U = \begin{pmatrix} 0 & -1 & -2 & 1 \\ 0 & 0 & -1 & -1 \\ 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

$$D = \begin{pmatrix} 6 & 0 & 0 & 0 \\ 0 & 6 & 0 & 0 \\ 0 & 0 & 8 & 0 \\ 0 & 0 & 0 & 4 \end{pmatrix} .$$

$$\text{Atunci } (D+L)^{-1} = \begin{pmatrix} 0.16667 & 0 & 0 & 0 \\ -0.05556 & 0.16667 & 0 & 0 \\ 0 & 0.0625 & 0.125 & 0 \\ -0.02778 & -0.05729 & -0.03125 & 0.25 \end{pmatrix} ,$$

$$-(D+L)^{-1}U = \begin{pmatrix} 0 & 0.16667 & 0.33333 & -0.16667 \\ 0 & -0.05556 & 0.05556 & 0.22222 \\ 0 & 0 & 0.0625 & -0.1875 \\ 0 & -0.02778 & -0.11285 & 0.03299 \end{pmatrix} ,$$

$$(D+L)^{-1}b = \begin{pmatrix} 0.33333 \\ 1.22222 \\ 0.75 \\ -0.82639 \end{pmatrix} .$$

$$\text{Alegând } x^{(0)} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} , \text{ obținem: } x^{(1)} = \begin{pmatrix} 0.5 \\ 1 \\ 0.9375 \\ -0.85938 \end{pmatrix} , x^{(2)} = \begin{pmatrix} 0.95573 \\ 1.02778 \\ 0.96973 \\ -0.98831 \end{pmatrix} ,$$

$$x^{(3)} = \begin{pmatrix} 0.99259 \\ 0.99937 \\ 0.99592 \\ -0.99697 \end{pmatrix} , x^{(4)} = \begin{pmatrix} 0.99803 \\ 1.00048 \\ 0.99918 \\ -0.99942 \end{pmatrix} , x^{(5)} = \begin{pmatrix} 0.99971 \\ 1.00006 \\ 0.99984 \\ -0.9999 \end{pmatrix} .$$

$$17. \begin{cases} 10x_1 - x_2 - x_3 + x_4 & = 7 \\ x_1 + 11x_2 - x_3 - x_4 & = -10 \\ -x_1 - x_2 + 12x_3 + x_4 & = 22 \\ x_1 + x_2 + x_3 + 13x_4 & = -24 \end{cases}$$

R. Se observă că matricea coeficienților sistemului

$$A = \begin{pmatrix} 10 & -1 & -1 & 1 \\ 1 & 11 & -1 & -1 \\ -1 & -1 & 12 & 1 \\ 1 & 1 & 1 & 13 \end{pmatrix}$$

este tare diagonal dominantă și algoritmul Gauss–Seidel este convergent.

Sistemul se poate pune sub forma  $x^{(m+1)} = (D+L)^{-1}(-Ux^{(m)} + b)$ ,  $m \geq 0$ , unde:

$$L = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ -1 & -1 & 0 & 0 \\ 1 & 1 & 1 & 0 \end{pmatrix}, \quad U = \begin{pmatrix} 0 & -1 & -1 & 1 \\ 0 & 0 & -1 & -1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

$$D = \begin{pmatrix} 10 & 0 & 0 & 0 \\ 0 & 11 & 0 & 0 \\ 0 & 0 & 12 & 0 \\ 0 & 0 & 0 & 13 \end{pmatrix}$$

Atunci

$$(D+L)^{-1} = \begin{pmatrix} 0.1 & 0 & 0 & 0 \\ -0.00909 & 0.09091 & 0 & 0 \\ 0.00758 & 0.00758 & 0.08333 & 0 \\ -0.00758 & -0.00758 & -0.00641 & 0.07692 \end{pmatrix},$$

$$-(D+L)^{-1}U = \begin{pmatrix} 0 & 0.1 & 0.1 & -0.1 \\ 0 & -0.00909 & 0.08182 & 0.1 \\ 0 & 0.00758 & 0.01515 & -0.08333 \\ 0 & -0.00758 & -0.01515 & 0.00641 \end{pmatrix},$$

$$(D+L)^{-1}b = \begin{pmatrix} 0.7 \\ -0.97273 \\ 1.81061 \\ -1.96445 \end{pmatrix}.$$

Alegând  $x^{(0)} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$ , obținem:  $x^{(1)} = \begin{pmatrix} 0.7 \\ -0.97273 \\ 1.81061 \\ -1.96445 \end{pmatrix},$

$$x^{(2)} = \begin{pmatrix} 0.98023 \\ -1.01219 \\ 1.99437 \\ -1.99711 \end{pmatrix}, \quad x^{(3)} = \begin{pmatrix} 0.99793 \\ -1.00001 \\ 1.99958 \\ -1.99999 \end{pmatrix}, \quad x^{(4)} = \begin{pmatrix} 0.99993 \\ -1.00001 \\ 1.99998 \\ -1.99999 \end{pmatrix},$$

$$x^{(5)} = \begin{pmatrix} 1 \\ -1 \\ 2 \\ -2 \end{pmatrix}.$$

Folosind metoda relaxării simple să se scrie soluția aproximativă pentru următoarele sisteme:

$$18. \begin{cases} 8x_1 - x_2 + x_3 = 3 \\ -x_1 + 6x_2 - x_3 = 14 \\ x_1 - x_2 + 9x_3 = -28 \end{cases}$$

R. Matricea sistemului este  $A = \begin{pmatrix} 8 & -1 & 1 \\ -1 & 6 & -1 \\ 1 & -1 & 9 \end{pmatrix}$ .

Luând vectorul de probă  $x^{(1)} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$  se obține  $r^{(1)} = \begin{pmatrix} -3 \\ -14 \\ 28 \end{pmatrix}$ , deci prima direcție

de relaxare este  $p^{(1)} = e_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$ . Rezultă  $x^{(2)} = \begin{pmatrix} 0 \\ 0 \\ -3.11111 \end{pmatrix}$ .

Analog :

$$r^{(2)} = \begin{pmatrix} -6.11111 \\ -10.88889 \\ 0 \end{pmatrix}, p^{(2)} = e_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, x^{(3)} = \begin{pmatrix} 0 \\ 1.81481 \\ -3.11111 \end{pmatrix};$$

$$r^{(3)} = \begin{pmatrix} -7.92593 \\ 0 \\ -1.81481 \end{pmatrix}, p^{(3)} = e_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, x^{(3)} = \begin{pmatrix} 0.99074 \\ 1.81481 \\ -3.11111 \end{pmatrix};$$

$$r^{(4)} = \begin{pmatrix} 0 \\ -0.99074 \\ -0.824070 \end{pmatrix}, p^{(4)} = e_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, x^{(4)} = \begin{pmatrix} 0.99074 \\ 1.91994 \\ -3.11111 \end{pmatrix}.$$

ș. a. m. d.

$$19. \begin{cases} 10x_1 + x_2 - x_3 - x_4 = 9 \\ x_1 + 9x_2 - x_3 + x_4 = -24 \\ -x_1 - x_2 + 11x_3 + x_4 = 30 \\ -x_1 + x_2 + x_3 + 8x_4 = -32 \end{cases}$$

R. Matricea sistemului este  $A = \begin{pmatrix} 10 & 1 & -1 & -1 \\ 1 & 9 & -1 & 1 \\ -1 & -1 & 11 & 1 \\ -1 & 1 & 1 & 8 \end{pmatrix}$ .

Alegând  $x^{(1)} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$  se obține  $r^{(1)} = \begin{pmatrix} -9 \\ 24 \\ -30 \\ 32 \end{pmatrix}$ .

Deci  $p^{(1)} = e_4 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$ ,  $x^{(2)} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ -4 \end{pmatrix}$ ,  $r^{(2)} = \begin{pmatrix} -5 \\ 20 \\ -34 \\ 0 \end{pmatrix}$ .

Mai departe obținem :

$$p^{(2)} = e_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, x^{(3)} = \begin{pmatrix} 0 \\ 0 \\ 3.090901 \\ -4 \end{pmatrix}; r^{(3)} = \begin{pmatrix} -8.09091 \\ 16.90909 \\ 0 \\ 3.09091 \end{pmatrix}, p^{(3)} = e_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix},$$

$$x^{(4)} = \begin{pmatrix} 0 \\ -1.87879 \\ 3.09091 \\ -4 \end{pmatrix}; p^{(4)} = e_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, x^{(5)} = \begin{pmatrix} 0.99697 \\ -1.87879 \\ 3.09091 \\ -4 \end{pmatrix}; r^{(5)} = \begin{pmatrix} 0 \\ 0.99697 \\ 0.88182 \\ 0.21515 \end{pmatrix}$$

$$, p^{(5)} = e_1 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, x^{(6)} = \begin{pmatrix} 0.99697 \\ -1.98956 \\ 3.09091 \\ -4 \end{pmatrix}.$$

$$20. \begin{cases} 4x_1 - 2x_2 + x_3 = 9 \\ -2x_1 + 6x_2 - x_3 = 5 \\ x_1 - x_2 + 5x_3 = 6 \end{cases}$$

R. Matricea  $A = \begin{pmatrix} 4 & -2 & 1 \\ -2 & 6 & -1 \\ 1 & -1 & 5 \end{pmatrix}$  este simetrică, tare diagonal dominantă:

$|4| > |-2| + |1|$ ,  $|6| > |-2| + |-1|$ ,  $|5| > |1| + |-1|$ ,  $\Delta_1=4$ ,  $\Delta_2=20$ ,  $\Delta_3=94$ , deci pozitiv definită și se poate aplica metoda. Luând vectorul de probă

$$\begin{aligned}
 x^{(1)} &= \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \text{ se obțin: } r^{(1)} = \begin{pmatrix} -9 \\ -5 \\ -6 \end{pmatrix}, p^{(1)} = e_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, x^{(2)} = \begin{pmatrix} 2.25 \\ 0 \\ 0 \end{pmatrix}; \\
 r^{(2)} &= \begin{pmatrix} 0 \\ -9.5 \\ -3.75 \end{pmatrix}, \text{ este } p^{(2)} = e_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, x^{(3)} = \begin{pmatrix} 2.25 \\ 1.58333 \\ 0 \end{pmatrix}; r^{(3)} = \begin{pmatrix} -3.16667 \\ 0 \\ -5.33333 \end{pmatrix} \\
 , p^{(3)} &= e_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, x^{(3)} = \begin{pmatrix} 2.25 \\ 1.5833 \\ 1.06667 \end{pmatrix}; r^{(4)} = \begin{pmatrix} -2.1 \\ -1.06667 \\ 0 \end{pmatrix}, p^{(4)} = e_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \\
 x^{(4)} &= \begin{pmatrix} 2.775 \\ 1.58333 \\ 1.06667 \end{pmatrix}.
 \end{aligned}$$

Să se scrie soluția aproximativă pentru următoarele sisteme de ecuații liniare obținută cu metoda suprarelaxării:

$$21. \begin{cases} 2x_1 - x_2 &= 3 \\ -x_1 + 3x_2 - x_3 &= -5 \\ -x_2 + 4x_3 &= 5 \end{cases}$$

R. Matricea coeficienților sistemului este bloc tridiagonală, simetrică și pozitiv definită, deci metoda suprarelaxării este convergentă. Sistemul se poate pune sub forma:

$$\begin{aligned}
 x^{(m+1)} &= Mx^{(m)} + C, \quad m \geq 0, \text{ unde: } M = -\left(E + \frac{1}{\omega}D\right)^{-1} \cdot \left[F + \left(1 - \frac{1}{\omega}\right)D\right], \\
 C &= \left(E + \frac{1}{\omega}\right)^{-1} \cdot b, \quad E = \begin{pmatrix} 0 & 0 & 0 \\ -1 & 0 & 0 \\ 0 & -1 & 0 \end{pmatrix}, \quad D = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 4 \end{pmatrix}, \quad F = E^T,
 \end{aligned}$$

$$\omega = \frac{2}{1 + \sqrt{1 - \lambda_1^2}}, \quad \lambda_1 \text{ fiind cea mai mare valoare proprie a matricei } -D^{-1}(E+F),$$

$\lambda_1 = 0.5$ ,  $\omega = 1.0718$  și considerând vectorul inițial  $x^{(0)} = b$ , rezultă

$$x^{(1)} = \begin{pmatrix} -1.28719 \\ -0.10088 \\ 0.95373 \end{pmatrix}, \quad x^{(2)} = \begin{pmatrix} 1.64605 \\ -0.85027 \\ 1.04344 \end{pmatrix}, \quad x^{(3)} = \begin{pmatrix} 1.03385 \\ -0.98314 \\ 1.0014 \end{pmatrix},$$

$$x^{(4)} = \begin{pmatrix} 1.00661 \\ -0.99835 \\ 1.00034 \end{pmatrix}.$$

$$22. \quad \begin{cases} 14x_1 + x_2 & = 13 \\ x_1 + 5x_2 - x_3 & = -3 \\ -x_2 + 14x_3 - 2x_4 & = -15 \\ -2x_3 + 5x_4 & = 7 \end{cases}$$

R. Matricea coeficienților sistemului este bloc tridiagonală, simetrică și pozitiv definită, deci metoda suprarelaxării este convergentă. Sistemul se poate pune sub forma

$$x^{(m+1)} = Mx^{(m)} + C, \quad m \geq 0, \quad \text{unde} \quad M = -\left(E + \frac{1}{\omega}D\right)^{-1} \cdot \left[F + \left(1 - \frac{1}{\omega}\right)D\right],$$

$$C = \left(E + \frac{1}{\omega}\right)^{-1} \cdot b, \quad E = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -2 & 0 \end{pmatrix}, \quad D = \begin{pmatrix} 14 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 \\ 0 & 0 & 14 & 0 \\ 0 & 0 & 0 & 5 \end{pmatrix}, \quad F = E^T,$$

$$\omega = \frac{2}{1 + \sqrt{1 - \lambda_1^2}}, \quad \lambda_1 \text{ fiind cea mai mare valoare proprie a matricei } -D^{-1}(E+F),$$

$\lambda_1 = 0.2735$ ,  $\omega = 1.01943$  și considerând vectorul inițial  $x^{(0)} = b$ , rezultă:

$$x^{(1)} = \begin{pmatrix} 0.91242 \\ -3.79769 \\ -0.05784 \\ 1.26758 \end{pmatrix}, \quad x^{(2)} = \begin{pmatrix} 1.20542 \\ -0.79542 \\ -0.96444 \\ 1.0093 \end{pmatrix}, \quad x^{(3)} = \begin{pmatrix} 0.98111 \\ -0.99288 \\ -0.99882 \\ 1.0003 \end{pmatrix},$$

$$x^{(4)} = \begin{pmatrix} 0.99985 \\ -0.99987 \\ -0.99997 \\ 1.00001 \end{pmatrix}, \quad x^{(5)} = \begin{pmatrix} 0.99999 \\ -0.99999 \\ -1 \\ 1 \end{pmatrix}.$$

Să se găsească soluția aproximativă obținută cu metoda Gauss-Seidel pentru sistemele de ecuații liniare următoare:

$$23. \begin{cases} 14x_1 + x_2 & = 13 \\ x_1 + 5x_2 - x_3 & = -3 \\ -x_2 + 14x_3 - 2x_4 & = -15 \\ -2x_3 + 5x_4 & = 7 \end{cases}$$

Să se precizeze numărul de iterații necesare pentru ca eroarea să se micșoreze de 10 ori.

R. Matricea  $A$  fiind simetrică și pozitiv definită procedeul iterativ Gauss-Seidel este convergent. Sistemul se poate pune sub forma

$$x^{(m+1)} = Mx^{(m)} + C, \quad m \geq 0,$$

unde

$$M = -(D+E)^{-1}F, \quad C = (D+E)^{-1}b, \quad E = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -2 & 0 \end{pmatrix}, \quad F = E^T,$$

$$D = \begin{pmatrix} 14 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 \\ 0 & 0 & 14 & 0 \\ 0 & 0 & 0 & 5 \end{pmatrix}.$$

$$\text{Deci } M = \begin{pmatrix} 0 & -0.07143 & 0 & 0 \\ 0 & 0.01429 & 0.2 & 0 \\ 0 & 0.00102 & 0.01429 & 0.14286 \\ 0 & 0.0004 & 0.000571 & 0.05714 \end{pmatrix}, \quad C = \begin{pmatrix} 0.92857 \\ -0.78571 \\ -1.12755 \\ 0.94898 \end{pmatrix}.$$

Dacă se pornește cu  $v^{(0)} = b$  obțin vectorii:

$$v^{(1)} = \begin{pmatrix} 1.14286 \\ -3.82857 \\ -0.3449 \\ 1.26204 \end{pmatrix}, \quad v^{(2)} = \begin{pmatrix} 1.20204 \\ -0.90939 \\ -0.95609 \\ 1.01756 \end{pmatrix}, \quad v^{(3)} = \begin{pmatrix} 0.99353 \\ -0.98992 \\ -0.99677 \\ 1.00129 \end{pmatrix},$$

$$v^{(4)} = \begin{pmatrix} 0.99928 \\ -0.99921 \\ -0.99976 \\ 1.0001 \end{pmatrix}, \quad v^{(5)} = \begin{pmatrix} 0.99994 \\ -0.99994 \\ -0.99998 \\ 1.00001 \end{pmatrix}.$$

Raza spectrală este  $\rho(M) = 0.0748$ , rata de convergență este  $R(M) = -\lg(\rho(M)) = 1.12609$  și numărul de iterații după care eroarea scade de 10 ori

este  $K \approx \frac{1}{R(M)} = 0.88803$ , adică la fiecare iterație eroarea scade de 10 ori.

$$24. \quad \begin{cases} 5x_1 + x_2 - x_3 & = 5 \\ x_1 + 8x_2 + x_3 & = 10 \\ -x_1 + x_2 + 15x_3 & = 15 \end{cases}$$

Să se precizeze numărul de iterații necesare pentru ca eroarea să se micșoreze de 10 ori.

R. Matricea  $A$  fiind simetrică și pozitiv definită procedeul iterativ

Gauss–Seidel este convergent. Sistemul se poate pune sub forma

$$x^{(m+1)} = Mx^{(m)} + C, \quad m \geq 0,$$

unde

$$M = -(D+E)^{-1}F, \quad C = (D+E)^{-1}b, \quad E = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ -1 & 1 & 0 \end{pmatrix}, \quad F = E^T, \quad D = \begin{pmatrix} 5 & 0 & 0 \\ 0 & 8 & 0 \\ 0 & 0 & 15 \end{pmatrix}$$

$$\text{Deci} \quad M = \begin{pmatrix} 0 & -0.2 & 0.2 \\ 0 & 0.02 & -0.12 \\ 0 & -0.01467 & 0.02133 \end{pmatrix}, \quad C = \begin{pmatrix} 1 \\ 0.9 \\ 1.0067 \end{pmatrix}.$$

Dacă se pornește cu  $v^{(0)} = b$  obțin vectorii:

$$v^{(1)} = \begin{pmatrix} 2 \\ -0.7 \\ 1.18 \end{pmatrix}, \quad v^{(2)} = \begin{pmatrix} 1.376 \\ 0.7444 \\ 1.04211 \end{pmatrix}, \quad v^{(3)} = \begin{pmatrix} 1.05954 \\ 0.78984 \\ 1.01798 \end{pmatrix}, \quad v^{(4)} = \begin{pmatrix} 1.04563 \\ 0.79364 \\ 1.0168 \end{pmatrix},$$

$$v^{(5)} = \begin{pmatrix} 1.04463 \\ 0.79386 \\ 1.01672 \end{pmatrix}.$$

Raza spectrală este  $\rho(M) = 0.06262$ , rata de convergență este  $R(M) = -\lg(\rho(M)) = 1.20326$  și numărul de iterații după care eroarea scade de 10 ori

este  $K \approx \frac{1}{R(M)} = 0.83108$ , adică la fiecare iterație eroarea scade de 10 ori.

Să se scrie soluția aproximativă obținută cu metoda gradientilor conjugați aplicată sistemelor de ecuații liniare:

$$25. \quad \begin{cases} 6x_1 + x_2 - x_3 & = 6 \\ x_1 + 8x_2 + x_3 & = -8 \\ -x_1 + x_2 + 10x_3 & = -12 \end{cases}.$$

R. Matricea coeficienților sistemului este simetrică și pozitiv definită și deci se

poate aplica metoda. Luând  $x^{(0)} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$  ca vector de probă, se obțin rezultatele.

Iterația I.

$$r^{(0)} = Ax_0 - b = \begin{pmatrix} -6 \\ 8 \\ 12 \end{pmatrix}, \text{ prima direcție de relaxare, } p^{(1)} = -r^{(0)},$$

$$q_1 = -\frac{\langle r^{(0)}, p^{(1)} \rangle}{\langle Ap^{(1)}, p^{(1)} \rangle} = 0.10133, \quad x^{(1)} = x^{(0)} + q_1 p^{(1)} = \begin{pmatrix} 0.60797 \\ -0.81063 \\ -1.21595 \end{pmatrix}.$$

Iterația a II-a .

$$r^{(1)} = r^{(0)} + q_1 Ap^{(1)} = \begin{pmatrix} -1.94684 \\ 0.90698 \\ -1.57807 \end{pmatrix}, \quad c_1 = \frac{\langle r^{(1)}, r^{(1)} \rangle}{\langle r^{(0)}, r^{(0)} \rangle} = 0.02911$$

$$p^{(2)} = -r^{(1)} + c_1 p^{(1)} = \begin{pmatrix} 2.12151 \\ -1.13987 \\ 1.22874 \end{pmatrix}, \quad q_2 = -\frac{\langle r^{(1)}, p^{(2)} \rangle}{\langle Ap^{(2)}, p^{(2)} \rangle} = 0.17916,$$

$$x^{(2)} = x^{(1)} + q_2 p^{(2)} = \begin{pmatrix} 0.98807 \\ -1.01485 \\ -0.9958 \end{pmatrix}.$$

Iterația a III-a.

$$r^{(2)} = r^{(1)} + q_2 Ap^{(2)} = \begin{pmatrix} -0.09062 \\ -0.12656 \\ 0.03906 \end{pmatrix}, \quad c_2 = \frac{\langle r^{(2)}, r^{(2)} \rangle}{\langle r^{(1)}, r^{(1)} \rangle} = 3.62608 \cdot 10^{-3}$$

$$p^{(3)} = -r^{(2)} + c_2 p^{(2)} = \begin{pmatrix} 0.09832 \\ 0.12243 \\ -0.03461 \end{pmatrix}, \quad q_3 = -\frac{\langle r^{(2)}, p^{(3)} \rangle}{\langle Ap^{(3)}, p^{(3)} \rangle} = 0.12133,$$

$$x^{(3)} = x^{(2)} + q_3 p^{(3)} = \begin{pmatrix} 1 \\ -1 \\ -1 \end{pmatrix}.$$

26. Să se determine traseul optim pentru o conductă de gaze naturale care să treacă prin "apropierea" localităților  $L_i$ ,  $i=1, \dots, 10$ , care raportate la un sistem cartezian de referință au coordonatele următoare:

$$L_1(1,2), L_2(2,2), L_3(5,3), L_4(7,4), L_5(10,2), L_6(11,3), L_7(15,4), L_8(16,5), L_9(18,1), \\ L_{10}(20,4).$$

R. Luând traseul după o dreaptă, se obține sistemul:

$$\begin{cases} a + b = 2 \\ 2a + b = 2 \\ 5a + b = 3 \\ 7a + b = 4 \\ 10a + b = 2 \\ 11a + b = 3 \\ 15a + b = 4 \\ 16a + b = 5 \\ 18a + b = 1 \\ 20a + b = 4 \end{cases}$$

care este supradimensionat. Se formează sistemul normal al lui Gauss  $Bu=c$ , unde

$$B=A^T A, \quad c=A^T b, \quad \text{adică: } B = \begin{pmatrix} 1505 & 105 \\ 105 & 10 \end{pmatrix}, \quad c = \begin{pmatrix} 340 \\ 30 \end{pmatrix}, \quad \text{a cărui soluție este}$$

$u_1=0.06211$ ,  $u_2=2.34783$ . Raportat la acel sistem de coordonate traseul conductei trebuie să urmeze dreapta  $y = 0.06211x + 2.34783$ .

## 2. Sisteme de ecuații neliniare

În acest capitol abordăm problema rezolvării numerice a sistemelor de ecuații algebrice neliniare.

Considerăm următorul sistem de ecuații

$$\begin{cases} f_1(x_1, x_2, \dots, x_n) = 0 \\ \dots \\ f_n(x_1, x_2, \dots, x_n) = 0, \end{cases} \quad (1)$$

în care cel puțin una din funcțiile  $f_i$ ,  $i = \overline{1, n}$  nu este liniară. Sub formă vectorială sistemul se scrie

$$F(x) = 0, \quad (2)$$

unde

$$x = (x_1, x_2, \dots, x_n)^T \quad \text{și} \quad F(x) = [f_1(x_1, \dots, x_n), \dots, f_n(x_1, \dots, x_n)]^T$$

Dacă adunăm  $x$  în ambii membri și notăm cu  $G(x) = x + F(x)$ , sistemul (2) se poate pune sub forma echivalentă

$$x = G(x) \quad (3)$$

Evident, există și alte metode de a pune sistemul (2) sub forma (3).

### Exemplul 1.

$$\begin{cases} f_1(x_1, x_2) \equiv 2x_1^2 + x_2^2 - 5 = 0 \\ f_2(x_1, x_2) \equiv x_1 + 2x_2 - 3 = 0 \end{cases} \quad (4)$$

Se observă că prima ecuație nu este liniară. Acest sistem se poate pune sub forma echivalentă

$$\begin{cases} x_1 = 2x_1^2 + x_2^2 - 5 + x_1 \equiv g_1(x_1, x_2) \\ x_2 = x_1 + 2x_2 - 3 \equiv g_2(x_1, x_2) \end{cases} \quad (4')$$

Sistemul fiind foarte simplu se poate rezolva cu metoda substituției. Înlocuind  $x_1 = 3 - 2x_2$  în prima ecuație obținem  $9x_2^2 - 24x_2 + 13 = 0$ , ecuație care admite rădăcinile  $x_{1,2} = \frac{4 \pm \sqrt{3}}{3}$ .

Așadar soluțiile exacte ale sistemului sunt

$$M_1\left(\frac{1-2\sqrt{3}}{3}, \frac{4+\sqrt{3}}{3}\right) \text{ și } M_2\left(\frac{1+2\sqrt{3}}{3}, \frac{4-\sqrt{3}}{3}\right).$$

Fie  $D = [1, 2] \times \left[\frac{1}{2}, \frac{3}{2}\right]$  o vecinătate a punctului  $M_2$ . În această vecinătate

sistemul (4) se poate pune sub forma echivalentă

$$\begin{cases} x_1 = \sqrt{\frac{5-x_2^2}{2}} \equiv \tilde{g}_1(x_1, x_2) \\ x_2 = \frac{1}{2}(3-x_1) \equiv \tilde{g}_2(x_1, x_2) \end{cases} \quad (4'')$$

(4') și (4'') sunt variante echivalente (de tipul (3)) ale sistemului (4), în vecinătatea punctului  $M_2$ .

În continuare prezentăm două metode numerice de rezolvare aproximativă a sistemelor neliniare.

## §2.1. Metoda aproximațiilor succesive

Fie  $D \subset \mathbb{R}^n$  o mulțime convexă, mărginită și închisă și fie sistemul

$$x = G(x), \quad x \in D.$$

Presupunem de asemenea că  $G \in C^1(D)$ . În aceste condiții există

$$m_{ij} = \sup_{x \in D} \left| \frac{\partial \tilde{g}_i}{\partial x_j}(x) \right|, \quad i = \overline{1, n} \quad j = \overline{1, n}.$$

Notăm cu

$$M = \begin{pmatrix} m_{11} & m_{12} & \dots & m_{1n} \\ \cdot & \cdot & \cdot & \cdot \\ m_{n1} & m_{n2} & \dots & m_{nn} \end{pmatrix}.$$

**Teorema 1.** Dacă  $G \in C^1(D)$ ,  $G(D) \subset D$  și  $\|M\|_\infty < 1$ , atunci sistemul  $x = G(x)$  admite o singură soluție în domeniul  $D$ , care se află cu metoda aproximațiilor succesive.

**Demonstrație.** Fie  $x \in D$  și  $y \in D$  oarecare. Din Teorema lui Lagrange pentru funcții de mai multe variabile, rezultă că pentru orice  $i = \overline{1, n}$ , există

$$\xi_i = x + \theta_i(y - x), \quad 0 < \theta_i < 1,$$

astfel încât

$$g_i(x) - g_i(y) = \frac{\partial g_i}{\partial x_1}(\xi_i)(x_1 - y_1) + \dots + \frac{\partial g_i}{\partial x_n}(\xi_i)(x_n - y_n) .$$

Ținând seama că

$$\left| \frac{\partial g_i}{\partial x_j}(x) \right| \leq m_{ij}, (\forall) x \in D,$$

rezultă

$$|g_i(x) - g_i(y)| \leq m_{i1}|x_1 - y_1| + \dots + m_{in}|x_n - y_n| \leq \|x - y\|_\infty \sum_{j=1}^n m_{ij}$$

și mai departe

$$\|G(x) - G(y)\|_\infty = \max_{1 \leq i \leq n} |g_i(x) - g_i(y)| \leq \|x - y\|_\infty \max_{1 \leq i \leq n} \sum_{j=1}^n m_{ij} = \|x - y\|_\infty \|M\|_\infty .$$

Cum  $\|M\|_\infty < 1$ , rezultă că aplicația  $G : D \rightarrow D$  este o contracție.

Conform teoremei de punct fix a lui Banach rezultă că există  $x^* \in D$ , unic, astfel încât  $x^* = G(x^*)$ . Așadar  $x^*$  este soluția unică a sistemului (3) din domeniul  $D$ . Această soluție se află cu metoda aproximațiilor succesive. Fie  $x^{(0)} \in D$  oarecare și fie șirul aproximațiilor succesive

$$x^{(k+1)} = G(x^{(k)}), k \geq 0 .$$

Acest șir este convergent în  $\mathbb{R}^n$  și limita sa  $x^* = \lim_{k \rightarrow \infty} x^{(k)}$  este soluția sistemului

(3) și deci a sistemului echivalent (1), respectiv (2). Teorema lui Banach ne dă și evaluarea erorii și anume

$$\|x^{(k)} - x^*\|_\infty \leq \frac{\|M\|_\infty^k}{1 - \|M\|_\infty} \|x^{(1)} - x^{(0)}\|_\infty . \quad \square \quad (5)$$

**Observația 1.** Teorema rămâne valabilă și dacă norma  $\|M\|_\infty$  se înlocuiește cu altă normă de matrice, de exemplu  $\|M\|_1$  sau  $\|M\|_2$ .

Considerăm din nou sistemul (4) din exemplul 1. În domeniul  $D = [1, 2] \times \left[ \frac{1}{2}, \frac{3}{2} \right]$ ,

acest sistem este echivalent cu sistemul

$$\begin{cases} x_1 = \sqrt{\frac{5 - x_2^2}{2}} \equiv g_1(x_1, x_2) \\ x_2 = \frac{1}{2}(3 - x_1) \equiv g_2(x_1, x_2) \end{cases} .$$

În acest domeniu, sistemul admite o singură soluție și anume

$$\begin{cases} x_1 = \frac{1+2\sqrt{3}}{3} \approx 1.4880338 \\ x_2 = \frac{4-\sqrt{3}}{3} \approx 0.7559831 \end{cases}$$

Deoarece  $x_1 \in [1,2]$  și  $x_2 \in \left[\frac{1}{2}, \frac{3}{2}\right]$  rezultă

$$\frac{1}{2} \leq \frac{3-x_1}{2} = g_2(x_1, x_2) \leq \frac{3}{2} \quad \text{și} \quad \sqrt{\frac{11}{8}} \leq \sqrt{\frac{5-x_2^2}{2}} = g_1(x_1, x_2) \leq \sqrt{\frac{19}{8}},$$

deci  $1 \leq g_1(x_1, x_2) \leq 2$ .

Așadar, dacă  $(x_1, x_2) \in D$  atunci  $(g_1(x_1, x_2), g_2(x_1, x_2)) \in D$ .

În continuare avem

$$\frac{\partial g_1}{\partial x_1} = 0, \quad \frac{\partial g_1}{\partial x_2} = -\frac{x_2}{\sqrt{10-2x_2^2}}; \quad \frac{\partial g_2}{\partial x_1} = -\frac{1}{2}; \quad \frac{\partial g_2}{\partial x_2} = 0$$

$$m_{11} = m_{22} = 0; \quad m_{12} = \frac{3}{\sqrt{22}}; \quad m_{21} = \frac{1}{2}$$

$$\|M\|_{\infty} = \|M\|_1 = \frac{3}{\sqrt{22}} < 1 \quad \text{și} \quad \|M\|_2 = \sqrt{\frac{29}{44}} < 1.$$

Alegem  $x_1^{(0)} = 1.5$  și  $x_2^{(0)} = 1$  (centrul dreptunghiului).

Se obțin următoarele valori pentru șirul aproximațiilor succesive

Nr. de iterații	0	1	2	3	4	5	6
$x_1$	1.5	1.414	1.490	1.478	1.488	1.487	1.488
$x_2$	1.0	0.750	0.793	0.755	0.761	0.756	0.756

În continuare prezentăm metoda aproximațiilor succesive pentru o singură ecuație neliniară.

Fie deci ecuația

$$f(x) = 0, \quad x \in [a, b].$$

Această ecuație se pune sub forma echivalentă

$$x = g(x), \quad x \in [a, b].$$

Din Teorema 1 rezultă că dacă

$$g \in C^1[a, b], \quad g: [a, b] \rightarrow [a, b] \quad \text{și} \quad \|g'\| = \sup\{|g'(x)|; x \in [a, b]\} < 1$$

atunci ecuația admite o singură rădăcină în intervalul  $[a, b]$  și aceasta este  $x^* = \lim_{k \rightarrow \infty} x_k$ , unde  $x_{k+1} = g(x_k)$ ,  $k \geq 0$ , iar  $x_0 \in [a, b]$  este arbitrar.

**Exemplu 2.** Fie ecuația

$$x^5 - x - 0.2 = 0; \quad x \in [-0.3; -0.2].$$

Forma echivalentă este

$$x = x^5 - 0,2 = 0, \quad x \in [-0,3; -0,2].$$

Avem  $g' = 5x^4$  și  $\|g'\| = 0.0405 < 1$ .

Se poate alege  $x_0 = -0.3$ . Șirul aproximațiilor succesive este

$$\begin{cases} x_{k+1} = x_k^5 - 0.2 \\ x_0 = -0.3 \end{cases}.$$

Se obțin următoarele valori pentru șirul aproximațiilor succesive

Numărul iterației	0	1	2	3	4	5
$x$	-0.3	-0.20243	-0.20034	-0.20032	-0.200322	-0.20032

## §2.2. Metoda Newton - Raphson

Fie  $D \subset \mathbb{R}^n$  o mulțime convexă, mărginită și închisă și fie sistemul neliniar

$$\begin{cases} f_1(x_1, x_2, \dots, x_n) = 0 \\ \dots \\ f_n(x_1, x_2, \dots, x_n) = 0, \end{cases} \quad (1)$$

Presupunem că  $\alpha = (\alpha_1, \dots, \alpha_n)^T \in D$  este o soluție izolată a sistemului

(1) și că  $x^{(0)} = (x_1^{(0)}, \dots, x_n^{(0)})^T \in D$  este un punct "apropiat" de  $\alpha$  (adică  $\|\alpha - x^{(0)}\| \ll 1$ ). Presupunem, de asemenea, că funcțiile  $f_i, i = \overline{1, n}$

sunt de clasă  $C^1$  pe  $D$ . În aceste condiții, dacă  $x \in D$  se află într-o vecinătate suficient de mică a punctului  $x^{(0)}$  avem

$$f_i(x) \approx f_i(x^{(0)}) + df_i(x^{(0)})(x - x^{(0)}), \quad i = \overline{1, n}.$$

Rezultă că sistemul (1) se poate înlocui cu sistemul liniar apropiat

$$\begin{cases} f_1(x^{(0)}) + df_1(x^{(0)})(x - x^{(0)}) = 0 \\ \dots \\ f_n(x^{(0)}) + df_n(x^{(0)})(x - x^{(0)}) = 0 \end{cases} \quad (2)$$

Sub forma vectorială sistemul (2) se scrie

$$F(x^{(0)}) + dF(x^{(0)})(x - x^{(0)}) = 0 \quad (3)$$

unde

$$F = (f_1, f_2, \dots, f_n)^T \quad \text{și} \quad dF = (df_1, \dots, df_n)^T.$$

Deoarece sistemul (3) este "apropiat" de sistemul (1), ne așteptăm ca soluția sa,  $x^{(1)}$ , să fie "apropiată" de soluția  $\alpha$  a sistemului (1).

Așadar  $x^{(1)}$  verifică relația

$$F(x^{(0)}) + dF(x^{(0)})(x^{(1)} - x^{(0)}) = 0.$$

În continuare considerăm sistemul liniar

$$F(x^{(1)}) + dF(x^{(1)})(x - x^{(1)}) = 0$$

și ne așteptăm ca soluția sa,  $x^{(2)}$ , să se "apropie" mai mult de  $\alpha$ . Așadar  $x^{(2)}$  verifică relația

$$F(x^{(1)}) + dF(x^{(1)})(x^{(2)} - x^{(1)}) = 0.$$

În general, considerăm șirul de vectori  $\{x^{(p)}\}$  cu proprietatea:

$$F(x^{(p)}) + dF(x^{(p)})(x^{(p+1)} - x^{(p)}) = 0 \quad (4)$$

și ne așteptăm că  $\{x^{(p)}\}$  să convergă la  $\alpha$ .

Reamintim că pentru orice  $a \in D$  și orice

$$h = (h_1, \dots, h_n)^T \in \mathbb{R}^n, \quad dF(a)(h) = J_F(a)h,$$

unde

$$J_F(a) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(a) & \dots & \frac{\partial f_1}{\partial x_n}(a) \\ \dots & \dots & \dots \\ \frac{\partial f_n}{\partial x_1}(a) & \dots & \frac{\partial f_n}{\partial x_n}(a) \end{pmatrix}$$

Dacă presupunem că  $J_F(\alpha)$  este nesingulară, atunci, din continuitate, rezultă că există o vecinătate  $V$  a punctului  $x = \alpha$ , astfel încât  $J_F(x)$  este nesingulară pentru orice  $x \in V$ .

În această condiție, din (4) rezultă

$$x^{(p+1)} = x^{(p)} - J_F^{-1}(x^{(p)})F(x^{(p)}), \quad p \geq 0 \quad (5)$$

**Teorema 1.** Fie  $D \subset \mathbb{R}^n$  o mulțime convexă, mărginită și închisă,  $\alpha \in D$  o soluție izolată a sistemului  $F(x) = 0$  și fie  $r > 0$  astfel încât bila

$$B_r(\alpha) = \{x \in \mathbb{R}^n; \|x - \alpha\|_\infty < r\} \subset D.$$

Presupunem că

- (i)  $F \in C^2(D)$
- (ii) Există  $M_1 > 0$  astfel încât  $\|J_F^{-1}(x)\|_\infty \leq M_1$ ,  $x \in B_r(\alpha)$
- (iii) Există  $M_2 > 0$  astfel încât  $\left| \frac{\partial^2 f_k}{\partial x_i \partial x_j}(x) \right| \leq M_2$ ,  $x \in B_r(\alpha)$ ,

oricare ar fi  $i, j, k = \overline{1, n}$ .

Atunci șirul  $\{x^{(p)}\}$  definit de (5) are proprietățile:

- a)  $\|x^{(p+1)} - \alpha\|_\infty \leq \frac{1}{2} n^2 M_1 M_2 \|x^{(p)} - \alpha\|_\infty^2$
- b) Dacă  $\frac{1}{2} n^2 M_1 M_2 \|x^{(0)} - \alpha\|_\infty^2 < 1$ , atunci șirul  $\{x^{(p)}\}$  este convergent și  $\lim_{p \rightarrow \infty} x^{(p)} = \alpha$ .

**Demonstrație.** Din formula Taylor rezultă că pentru orice  $k = \overline{1, n}$  și orice  $p \in \mathbb{N}$ , există un punct  $\xi_k^{(p)}$  pe segmentul de dreaptă deschis de capete  $\alpha$  și  $x^{(p)}$  astfel încât

$$f_k(\alpha) - f_k(x^{(p)}) = df_k(x^{(p)})(\alpha - x^{(p)}) + \frac{1}{2!} d^2 f_k(\xi_k^{(p)})(\alpha - x^{(p)})^2.$$

Ținând seama că

$$d^2 f_k(\xi_k^{(p)})(\alpha - x^{(p)}) = \sum_{i=1}^n \sum_{j=1}^n \frac{\partial^2 f_k}{\partial x_i \partial x_j}(\xi_k^{(p)})(\alpha_i - x_i^{(p)})(\alpha_j - x_j^{(p)})$$

și de ipoteza (iii), rezultă

$$\begin{aligned} \left| f_k(\alpha) - f_k(x^{(p)}) - df_k(x^{(p)})(\alpha - x^{(p)}) \right| &\leq \frac{1}{2} M_2 \sum_{i=1}^n \sum_{j=1}^n |\alpha_i - x_i^{(p)}| |\alpha_j - x_j^{(p)}| \\ &\leq \frac{1}{2} M_2 n^2 \|\alpha - x^{(p)}\|_\infty^2, \quad (\forall) k = \overline{1, n}. \end{aligned}$$

În continuare, avem

$$\|F(\alpha) - F(x^{(p)}) - dF(x^{(p)})(\alpha - x^{(p)})\|_\infty = \frac{1}{2} M_2 n^2 \|\alpha - x^{(p)}\|_\infty^2 \quad (6)$$

Pe de altă parte din (4) rezultă

$$-F(x^{(p)}) + dF(x^{(p)})(x^{(p)}) = dF(x^{(p)})(x^{(p+1)}) \quad (7)$$

Cum  $F(\alpha) = 0$ , din (6) și (7) obținem

$$\begin{aligned} \|dF(x^{(p)})(x^{(p+1)} - \alpha)\|_\infty &\leq \frac{1}{2} n^2 M_2 \|\alpha - x^{(p)}\|_\infty^2 \quad \text{sau} \\ \|J_F(x^{(p)})(x^{(p+1)} - \alpha)\|_\infty &\leq \frac{1}{2} n^2 M_2 \|\alpha - x^{(p)}\|_\infty^2 \quad (8) \end{aligned}$$

În sfârșit, ținând seama și de ipoteza (ii) avem

$$\begin{aligned} \|(x^{(p+1)} - \alpha)\|_\infty &= \|J_F^{-1}(x^{(p)}) J_F(x^{(p)})(x^{(p+1)} - \alpha)\|_\infty \leq \\ &\leq \|J_F^{-1}(x^{(p)})\| \cdot \|J_F(x^{(p)})(x^{(p+1)} - \alpha)\|_\infty \leq M_1 \frac{1}{2} M_2 n^2 \|\alpha - x^{(p)}\|_\infty^2 \end{aligned}$$

Așadar, am demonstrat afirmația a).

Dacă notăm cu  $c = \frac{1}{2}n^2M_1M_2$ , din a) rezultă

$$\|x^{(p+1)} - \alpha\|_{\infty} \leq c \|x^{(p)} - \alpha\|_{\infty}^2 \quad (9)$$

Particularizând indicele  $p$ , obținem succesiv

$$\|x^{(1)} - \alpha\|_{\infty} \leq c \|x^{(0)} - \alpha\|_{\infty}^2$$

$$\|x^{(2)} - \alpha\|_{\infty} \leq c \|x^{(1)} - \alpha\|_{\infty}^2 \leq c^3 \|x^{(0)} - \alpha\|_{\infty}^4$$

$$\dots\dots\dots$$

$$\|x^{(p+1)} - \alpha\|_{\infty} \leq \frac{1}{c} \left( c \|x^{(0)} - \alpha\|_{\infty} \right)^{2^{p+1}}.$$

Dacă  $c \|x^{(0)} - \alpha\|_{\infty} < 1$ , atunci  $\lim_{p \rightarrow \infty} \|x^{(p+1)} - \alpha\|_{\infty} = 0$ , deci  $\{x^{(p)}\}$  este convergent și  $\lim_{p \rightarrow \infty} x^{(p)} = \alpha$ . Cu aceasta teorema este demonstrată.  $\square$

**Exemplu.** Reluăm sistemul (4) din §1

$$\begin{cases} f_1(x_1, x_2) \equiv 2x_1^2 + x_2^2 - 5 = 0 \\ f_2(x_1, x_2) \equiv x_1 + 2x_2 - 3 = 0 \end{cases}$$

În domeniul  $D = [1, 2] \times \left[\frac{1}{2}, \frac{3}{2}\right]$  sistemul admite o singură soluție și anume

$$\begin{cases} \alpha_1 = \frac{1+2\sqrt{3}}{3} \approx 1.4880338 \\ \alpha_2 = \frac{4-\sqrt{3}}{3} \approx 0.7559831 \end{cases}$$

$$x_1^{(0)} = \frac{3}{2} \quad \text{și} \quad x_2^{(0)} = 1; \quad J_F(x_1, x_2) = \begin{pmatrix} 4x_1 & 2x_2 \\ 1 & 2 \end{pmatrix};$$

$$J_F\left(\frac{3}{2}, 1\right) = \begin{pmatrix} 6 & 2 \\ 1 & 2 \end{pmatrix}; \quad J_F^{-1}\left(\frac{3}{2}, 1\right) = \begin{pmatrix} \frac{1}{5} & -\frac{1}{5} \\ -\frac{1}{10} & \frac{3}{5} \end{pmatrix}$$

$$f_1\left(\frac{3}{2}, 1\right) = \frac{1}{2}; \quad f_2\left(\frac{3}{2}, 1\right) = \frac{1}{2}.$$

Conform (5) avem

$$\begin{pmatrix} x_1^{(1)} \\ x_2^{(1)} \end{pmatrix} = \begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix} - \begin{pmatrix} \frac{1}{5} & -\frac{1}{5} \\ -\frac{1}{10} & \frac{3}{5} \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix} - \begin{pmatrix} 0 \\ 1 \\ 4 \end{pmatrix} = \begin{pmatrix} 3 \\ 2 \\ 3 \\ 4 \end{pmatrix}.$$

Primele 3 iterații sunt prezentate în tabelul următor

Numărul iterației	0	1	2	3
$x_1$	1.5	1.5	1.488095	1.488034
$x_2$	1.0	0.75	0.755952	0.755983

$$J_F^{-1}(x) = \begin{pmatrix} \frac{1}{4x_1 - x_2} & \frac{-x_2}{4x_1 - x_2} \\ -\frac{1}{2(4x_1 - x_2)} & \frac{4x_1}{2(4x_1 - x_2)} \end{pmatrix}$$

$$x_1 \in [1, 2] ; x_2 \in \left[ \frac{1}{2}, \frac{3}{2} \right]$$

$$\|J_F^{-1}(x)\|_{\infty} = \max \left\{ \frac{1+x_2}{4x_1 - x_2}, \frac{1+4x_1}{2(4x_1 - x_2)} \right\} \leq \frac{9}{5}.$$

Așadar, putem lua  $M_1 = \frac{9}{5}$ .

$$\frac{\partial^2 f_1}{\partial x_1^2} = 4 ; \frac{\partial^2 f_1}{\partial x_1 \partial x_2} = 0 ; \frac{\partial^2 f_1}{\partial x_2^2} = 2 ; \frac{\partial^2 f_2}{\partial x_1^2} = \frac{\partial^2 f_2}{\partial x_1 \partial x_2} = \frac{\partial^2 f_2}{\partial x_2^2} = 0$$

și deci  $M_2 = 4$ .

$$x^{(0)} = \begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix}, \alpha \cong \begin{pmatrix} 1,488 \\ 0,756 \end{pmatrix}, \|x^{(0)} - \alpha\|_{\infty}^2 \approx 0.06.$$

$$\frac{1}{2} M_1 M_2 n^2 \|x^{(0)} - \alpha\|_{\infty}^2 \approx \frac{1}{2} \cdot \frac{9}{5} \cdot 4 \cdot 4 \cdot 0.06 = 0.864 < 1$$

Rezultă că algoritmul Newton–Raphson este convergent în acest caz.

**Observația 1.** Metoda Newton expusă aici are un inconvenient major și anume faptul că la fiecare pas trebuie calculată inversa  $J_F^{-1}(x^{(p)})$ . Din motive de continuitate, putem presupune că într-o vecinătate suficient de mică a punctului  $x^{(0)}$  avem  $J_F^{-1}(x^{(p)}) \cong J_F^{-1}(x^{(0)})$ . Se obține astfel metoda Newton modificată

$$\begin{cases} v^{(p+1)} &= v^{(p)} - J_F^{-1}(x^{(0)}) F(v^{(p)}) \\ v^{(0)} &= x^{(0)} \end{cases}, \quad p \geq 0 \quad (10)$$

Observăm că  $v^{(1)} = x^{(1)}$  dar, în general  $v^{(p)} \neq x^{(p)}$  pentru  $p > 1$ .

L. Kantorovici a studiat metoda Newton modificată și a dat condiții suficiente care asigură convergența algoritmului (10).

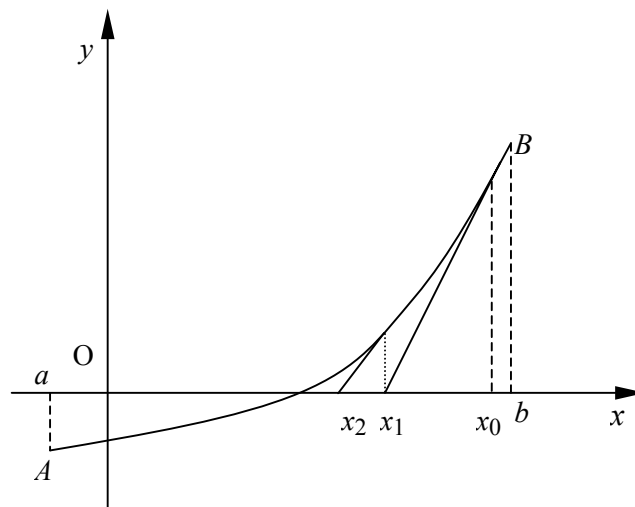
În continuare să analizăm metoda Newton–Raphson pentru o singură ecuație neliniară.

$$F(x)=0, \quad x \in [a, b]. \quad (11)$$

Presupunem că ecuația (11) admite o singură rădăcină  $\alpha \in [a, b]$ .

Algoritmul (5) revine la

$$\begin{cases} x_{p+1} = x_p - \frac{f(x_p)}{f'(x_p)}, & p \geq 0 \\ x_0 \in [a, b] \end{cases} \quad (12)$$



Din punct de vedere geometric,  $x_{p+1}$  reprezintă abscisa punctului în care tangenta la graficul funcției  $f$  în punctul  $M_p[x_p, f(x_p)]$  întâlnește axa  $Ox$ .

Într-adevăr, ecuația tangentei la grafic în punctul  $M_0[x_0, f(x_0)]$  este

$$y - f(x_0) = f'(x_0)(x - x_0).$$

Fie  $x_1$  abscisa punctului în care această tangentă întâlnește axa  $Ox$ .

Avem

$$0 - f(x_0) = f'(x_0)(x_1 - x_0)$$

și mai departe

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

adică prima iterație din (12).

Fie  $m_1 = \inf\{|f'(x)|; x \in [a, b]\}$ . Atunci putem lua  $M_1 = \frac{1}{m_1}$ . Evident

$m_2 = \sup\{|f''(x)|; x \in [a, b]\}$ . Algoritmul este convergent dacă

$$\frac{1}{2}M_1M_2|x_0 - \alpha|^2 < 1.$$

**Exemplul 2.** Fie ecuația  $F(x) \equiv x^3 - 2x - 5 = 0$ ;  $x \in [2,3]$ . Ecuația admite o singură rădăcină reală  $\alpha \in (2,3)$ .

$$F(2) = -1 < 0; F(3) = 16 > 0.$$

Algoritmul este

$$\begin{cases} x_{p+1} = x_p - \frac{x_p^3 - 2x_p - 5}{3x_p^2 - 2}, & p \geq 0 \\ x_0 \in (2,3) \text{ arbitrar} \end{cases} \quad (13)$$

$$f'(x) = 3x^2 - 2; \quad M_1 = \frac{1}{10}; \quad f''(x) = 6x; \quad M_2 = 18$$

$$\frac{1}{2} M_1 M_2 |x_0 - \alpha|^2 < \frac{9}{10} < 1,$$

de unde rezultă convergența șirului  $\{x_p\}$  definit de (13). Valorile obținute după primele 5 iterații sunt trecute în tabelul de mai jos.

Numărul iterației	0	1	2	3	4	5
$x$	2.5	2.16418	2.09714	2.09456	2.09455	2.09455

### Exerciții

1. Să se găsească soluția aproximativă a sistemului

$$\begin{cases} y^3 - 20x - 1 = 0 \\ x^3 + xy - 10y + 10 = 0 \end{cases}$$

situată în dreptunghiul  $D = [-1, 1] \times [0, 2]$ , folosind metoda aproximațiilor succesive.

$$R. \text{ Considerăm } G : D \rightarrow D \quad \text{unde} \quad G(x, y) = \begin{pmatrix} g_1(x, y) \\ g_2(x, y) \end{pmatrix} = \begin{pmatrix} \frac{y^3 - 1}{20} \\ \frac{x^3 + xy + 10}{10} \end{pmatrix}$$

$$M = (m_{ij})_{1 \leq i, j \leq 2} = \left( \sup_{x \in D} \left| \frac{\partial g_i}{\partial x_j} \right| \right)_{1 \leq i, j \leq 2} = \begin{pmatrix} 0 & \frac{3}{5} \\ \frac{1}{2} & \frac{1}{10} \end{pmatrix}, \quad \text{iar} \quad \|M\|_{\infty} = \frac{3}{5}, \quad \text{deci} \quad G$$

este o contracție și șirul aproximațiilor succesive  $x^{(p+1)} = G(x^{(p)})$  converge la soluția sistemului.

Valorile obținute după primele 3 iterații sunt trecute în tabelul de mai jos.

Numărul iterației	0	1	2	3
$x$	0.5	-0.0437	0.00583	-0.000679
$y$	0.5	1.0375	0.99545	0.99993

2. Să se găsească soluția aproximativă a sistemului

$$\begin{cases} x^3 + y^3 - 6x + 3 = 0 \\ x^3 - y^3 - 6y + 2 = 0 \end{cases}$$

situată în dreptunghiul  $D = \left[ \frac{1}{2}, \frac{5}{6} \right] \times \left[ \frac{1}{6}, \frac{1}{2} \right]$ , folosind metoda aproximațiilor succesive.

R. Punem sistemul sub forma 
$$\begin{cases} x = \frac{x^3 + y^3}{6} + \frac{1}{2} \\ y = \frac{x^3 - y^3}{6} + \frac{1}{3} \end{cases}$$
 și atunci

$G(x, y) = \begin{pmatrix} g_1(x, y) = \frac{x^3 + y^3}{6} + \frac{1}{2} \\ g_2(x, y) = \frac{x^3 - y^3}{6} + \frac{1}{3} \end{pmatrix}$  este o contracție a lui  $D$ . Într-adevăr,

$$M = (m_{ij})_{1 \leq i, j \leq 2} = \left( \sup_{x \in D} \left| \frac{\partial g_i}{\partial x_j} \right| \right)_{1 \leq i, j \leq 2} = \begin{pmatrix} \left( \frac{5}{6} \right)^2 \cdot \frac{1}{2} & \left( \frac{1}{2} \right)^2 \cdot \frac{1}{2} \\ \left( \frac{5}{6} \right)^2 \cdot \frac{1}{2} & \left( \frac{1}{2} \right)^2 \cdot \frac{1}{2} \end{pmatrix}, \quad \text{iar}$$

$\|M\|_\infty = 0.47222$ , deci  $G$  este o contracție și șirul aproximațiilor succesive  $x^{(p+1)} = G(x^{(p)})$  converge la soluția sistemului. Considerând  $x_0=0.5$  și  $y_0=0.5$  avem:

Numărul iterației	0	1	2	3
$x$	0.5	0.54167	0.53266	0.53256
$y$	0.5	0.33333	0.35365	0.35115

3. Să se găsească soluția aproximativă a ecuației  $e^{-x} + 10x - 5 = 0$  situată în intervalul  $[0, 1]$ , folosind metoda aproximațiilor succesive.

R. Ecuația se poate pune sub forma  $x = \frac{5 - e^{-x}}{10} = \varphi(x)$ , unde  $\varphi(x)$  este o contracție și șirul aproximațiilor succesive  $x = \varphi(x)$  converge la soluția ecuației. Valorile obținute după primele 5 iterații sunt trecute în tabelul de mai jos.

Nr. de iterații	0	1	2	3	4	5
$x$	0	0.4	0.43297	0.43514	0.43528	0.43529

$$|x_5 - x^*| \leq \frac{\left(\frac{1}{3}\right)^5}{1 - \frac{1}{3}} |x_1 - x_0| = 0.00525.$$

4. Să se găsească soluția aproximativă din cadranul întâi pentru sistemul

$$\begin{cases} x_1 + 3 \lg x_1 - x_2^2 = 0 \\ 2x_1^2 - x_1 x_2 - 5x_1 + 1 = 0 \end{cases}$$

folosind metoda Newton.

R. Șirul aproximațiilor succesive  $x^{(p+1)} = x^{(p)} - J_F^{-1}(x^{(p)})F(x^{(p)})$ ,  $p \geq 0$

unde:

$$F(x, y) = \begin{pmatrix} x_1 + 3 \lg x_1 - x_2^2 \\ 2x_1^2 - x_1 x_2 - 5x_1 + 1 \end{pmatrix}, \quad J_F(x^{(p)}) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} \end{pmatrix}.$$

Se obțin:

$$J^{-1}(x^{(0)}) = \begin{pmatrix} -0.19112 & 0.25482 \\ -0.31853 & 0.09137 \end{pmatrix}, \quad x^{(1)} = \begin{pmatrix} 3.59209 \\ 2.32015 \end{pmatrix};$$

$$J^{-1}(x^{(1)}) = \begin{pmatrix} -0.12916 & 0.16685 \\ -0.25343 & 0.049 \end{pmatrix}, \quad x^{(2)} = \begin{pmatrix} 3.49059 \\ 2.26341 \end{pmatrix};$$

$$J^{-1}(x^{(2)}) = \begin{pmatrix} -0.13672 & 0.1773 \\ -0.26238 & 0.05379 \end{pmatrix}, \quad x^{(3)} = \begin{pmatrix} 3.48745 \\ 2.26163 \end{pmatrix};$$

$$J^{-1}(x^{(3)}) = \begin{pmatrix} -0.13697 & 0.17765 \\ -0.26267 & 0.05395 \end{pmatrix}, \quad x^{(4)} = \begin{pmatrix} 3.48744 \\ 2.26163 \end{pmatrix}$$

$$J^{-1}(x^{(4)}) = \begin{pmatrix} -0.13697 & 0.17765 \\ -0.26267 & 0.05395 \end{pmatrix}, \quad x^{(5)} = \begin{pmatrix} 3.48744 \\ 2.26163 \end{pmatrix} \text{ ș. a. m. d.}$$

5. Să se găsească soluția aproximativă ( $x > 0$ ,  $y > 0$ ) pentru sistemul

$$\begin{cases} x^2 - y = 0 \\ x^2 + y^2 - 3x = 0 \end{cases}$$

folosind metoda Newton.

R. Șirul aproximațiilor  $x^{(p+1)} = x^{(p)} - J_F^{-1}(x^{(p)})F(x^{(p)})$ ,  $p \geq 0$  unde:

$$F(x, y) = \begin{pmatrix} x^2 - y \\ x^2 + y^2 - 3x \end{pmatrix}, \quad J_F(x^{(p)}) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} \end{pmatrix}.$$

Se obțin următoarele rezultate dacă se pornește cu  $x^{(0)} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ :

$$J^{-1}(x^{(0)}) = \begin{pmatrix} 0.66667 & 0.33333 \\ 0.33333 & 0.66667 \end{pmatrix}, \quad x^{(1)} = \begin{pmatrix} 1.33333 \\ 1.66667 \end{pmatrix};$$

$$J^{-1}(x^{(1)}) = \begin{pmatrix} 0.38961 & 0.11688 \\ 0.03896 & 0.31169 \end{pmatrix}, \quad x^{(2)} = \begin{pmatrix} 1.22511 \\ 1.48918 \end{pmatrix};$$

$$J^{-1}(x^{(2)}) = \begin{pmatrix} 0.44138 & 0.1482 \\ 0.08148 & 0.36311 \end{pmatrix}, \quad x^{(3)} = \begin{pmatrix} 1.21353 \\ 1.47253 \end{pmatrix};$$

$$J^{-1}(x^{(3)}) = \begin{pmatrix} 0.44792 & 0.15209 \\ 0.08714 & 0.36914 \end{pmatrix}, \quad x^{(4)} = \begin{pmatrix} 1.21341 \\ 1.47237 \end{pmatrix}$$

$$J^{-1}(x^{(4)}) = \begin{pmatrix} 0.44799 & 0.15213 \\ 0.0872 & 0.3692 \end{pmatrix}, \quad x^{(5)} = \begin{pmatrix} 1.21341 \\ 1.47237 \end{pmatrix}.$$

6. Folosind metoda Newton să se aproximeze soluția pozitivă a sistemului de ecuații neliniare

$$\begin{cases} x^2 + y^2 + z^2 = 1 \\ 2x^2 + y^2 - 4z = 0 \\ 3x^2 - 4y + z^2 = 0 \end{cases}, \quad \text{considerând } X^{(0)} = \begin{pmatrix} 0.5 \\ 0.5 \\ 0.5 \end{pmatrix}.$$

R. Șirul aproximațiilor  $x^{(p+1)} = x^{(p)} - J_F^{-1}(x^{(p)})F(x^{(p)})$ ,  $p \geq 0$ , unde:

$$F(X) = \begin{pmatrix} x^2 + y^2 + z^2 - 1 \\ 2x^2 + y^2 - 4z \\ 3x^2 - 4y + z^2 \end{pmatrix}; \quad X = \begin{pmatrix} x \\ y \\ z \end{pmatrix};$$

$$J_F(X) = \begin{pmatrix} \frac{\partial f_1}{\partial x} & \frac{\partial f_1}{\partial y} & \frac{\partial f_1}{\partial z} \\ \frac{\partial f_2}{\partial x} & \frac{\partial f_2}{\partial y} & \frac{\partial f_2}{\partial z} \\ \frac{\partial f_3}{\partial x} & \frac{\partial f_3}{\partial y} & \frac{\partial f_3}{\partial z} \end{pmatrix} = \begin{pmatrix} 2x & 2y & 2z \\ 4x & 2y & -4 \\ 6x & -4 & 2z \end{pmatrix} .$$

Se obțin următoarele rezultate dacă se pornește cu  $X^{(0)} = \begin{pmatrix} 0.5 \\ 0.5 \\ 0.5 \end{pmatrix}$  :

$$J_F^{-1}(X^{(0)}) = \begin{pmatrix} 0.375 & 0.125 & 0.125 \\ 0.35 & 0.05 & -0.15 \\ 0.275 & -0.175 & 0.025 \end{pmatrix} ; X^{(1)} = \begin{pmatrix} 0.875 \\ 0.5 \\ 0.375 \end{pmatrix}$$

$$J_F^{-1}(X^{(1)}) = \begin{pmatrix} 0.23552 & 0.05792 & 0.07336 \\ 0.36486 & 0.04054 & -0.14865 \\ 0.2973 & -0.18919 & 0.02703 \end{pmatrix} ; X^{(2)} = \begin{pmatrix} 0.78982 \\ 0.49662 \\ 0.36993 \end{pmatrix}$$

$$J_F^{-1}(X^{(2)}) = \begin{pmatrix} 0.26276 & 0.06359 & 0.08104 \\ 0.36652 & 0.04023 & -0.149 \\ 0.29855 & -0.18978 & 0.02701 \end{pmatrix} ; X^{(3)} = \begin{pmatrix} 0.78521 \\ 0.49661 \\ 0.36992 \end{pmatrix}$$

7. Să se găsească soluția aproximativă ( $x > 0$ ,  $y > 0$ ) pentru sistemul

$$\begin{cases} x^2 - y = 0 \\ x^2 + y^2 - 3x = 0 \end{cases}$$

folosind metoda Newton modificată.

R. Șirul aproximațiilor succesive  $x^{(p+1)} = -J_F^{-1}(x^{(0)})F(x^{(p)})$ ,  $p \geq 0$

unde:

$$F(x, y) = \begin{pmatrix} x^2 - y \\ x^2 + y^2 - 3x \end{pmatrix}, \quad J_F(x^{(p)}) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} \end{pmatrix} .$$

Se obțin următoarele rezultate dacă se pornește cu  $x^{(0)} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$  :

$$\begin{aligned} J^{-1}(x^{(0)}) &= \begin{pmatrix} 0.66667 & 0.33333 \\ 0.33333 & 0.66667 \end{pmatrix}, \quad x^{(1)} = \begin{pmatrix} 1.23077 \\ 1.46154 \end{pmatrix}, \\ x^{(2)} &= \begin{pmatrix} 1.20791 \\ 1.46908 \end{pmatrix}, \quad x^{(3)} = \begin{pmatrix} 1.21586 \\ 1.47494 \end{pmatrix}, \quad x^{(4)} = \begin{pmatrix} 1.21217 \\ 1.47093 \end{pmatrix}, \\ x^{(47)} &= \begin{pmatrix} 1.21341 \\ 1.47237 \end{pmatrix}. \end{aligned}$$

### 3. Vectori și valori proprii

Reamintim că dacă  $A$  este o matrice pătratică, atunci un vector  $x \in \mathbb{R}^n$  se numește *vector propriu* în raport cu  $A$ , dacă  $x \neq 0$  și există un număr  $\lambda$  (real sau complex) astfel încât  $Ax = \lambda x$ . Numărul  $\lambda$  se mai numește și *valoarea proprie*. Valorile proprii ale matricei  $A$  sunt rădăcinile *polinomului caracteristic*  $P(\lambda) = \det(A - \lambda I)$  și sunt invariante la transformările de similitudine ale lui  $A$ ; acest lucru înseamnă că valorile proprii ale matricei  $A$  coincid cu valorile proprii ale matricei  $C^{-1}AC$ , oricare ar fi matricea nesingulară  $C$ .

Dacă matricea  $A$  este simetrică, atunci valorile sale proprii sunt reale și există o bază ortonormală formată din vectori proprii, deci cu proprietatea  $Av_i = \lambda_i v_i$ ,  $i = \overline{1, n}$ , în raport cu care matricea  $A$  se reduce la forma diagonală

$$D = \begin{pmatrix} \lambda_1 & 0 & 0 & \dots & 0 \\ 0 & \lambda_2 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \lambda_n \end{pmatrix} \quad (1)$$

Baza  $v_1, \dots, v_n$  se poate alege astfel încât  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$ . Dacă, în plus,  $A$  este și pozitiv definită, atunci  $\lambda_1 \geq \dots \geq \lambda_n > 0$  și

$$\lambda_1 = \|A\|_2 = \sup_{x \neq 0} \frac{\langle Ax, x \rangle}{\langle x, x \rangle}.$$

Fie  $V$  matricea de trecere de la baza canonică a spațiului  $\mathbb{R}^n$  la baza  $v_1, v_2, \dots, v_n$ . Se verifică imediat că  $V^T V = I$ , deci  $V$  este ortogonală. Rezultă că  $V^{-1} = V^T$  și că  $D = V^T \cdot AV$ .

În practică, valorile proprii ale matricei  $A$  nu se determină rezolvând numeric ecuația caracteristică  $\det(A - \lambda I) = 0$ , deoarece, așa cum vom arăta în continuare, rădăcinile unui polinom sunt foarte "sensibile" la orice modificare a coeficienților polinomului.

Într-adevăr, fie polinomul

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x + a_0 ,$$

și fie

$$h(x) = f(x) + \varepsilon g(x)$$

polinomul modificat, în care  $\varepsilon > 0$  este arbitrar, iar

$$g(x) = b_n x^n + b_{n-1} x^{n-1} + \dots + b_1 x + b_0$$

este un polinom oarecare. Cum  $g$  este arbitrar, putem considera că  $b_i = a_i$ ,  $i = \overline{1, n}$  sau  $b_i = 0$  pentru  $i \neq j$  și  $b_j = a_j$  etc. Așadar, cazul considerat este practic cazul cel mai general. Fie  $x_1, x_2, \dots, x_n$  rădăcinile polinomului  $f$ . Pentru simplificare, vom presupune că aceste rădăcini sunt simple, deci că  $f(x_k) = 0$  și  $f'(x_k) \neq 0$ ,  $k = \overline{1, n}$ .

Să presupunem că vrem să determinăm rădăcinile ecuației  $h(x) = 0$  cu una din metodele numerice cunoscute, de exemplu metoda Newton - Raphson. Ne așteptăm ca pentru  $\varepsilon > 0$  foarte mic, rădăcinile ecuației  $h(x) = 0$  să fie apropiate de rădăcinile ecuației inițiale  $f(x) = 0$ . Notăm cu  $z_k$  o rădăcină oarecare a ecuației  $h(x) = 0$ . Conform algoritmului Newton avem

$$z_k = x_k - \frac{h(x_k)}{h'(x_k)} = x_k - \frac{\varepsilon g(x_k)}{f'(x_k) + \varepsilon g'(x_k)} \quad (2)$$

Dacă notăm cu

$$q(\varepsilon) = \frac{g(x_k)}{f'(x_k) + \varepsilon g'(x_k)} ,$$

atunci

$$q'(\varepsilon) = - \frac{g(x_k)g'(x_k)}{[f'(x_k) + \varepsilon g'(x_k)]^2} . \quad (3)$$

Cum  $q(\varepsilon) \approx q(0) + \varepsilon q'(0)$  pentru  $\varepsilon > 0$  suficient de mic, din (2) și (3) rezultă

$$z_k = x_k - \varepsilon \left( \frac{g(x_k)}{f'(x_k)} - \varepsilon \frac{g(x_k)g'(x_k)}{[f'(x_k)]^2} \right) \approx x_k - \varepsilon \frac{g(x_k)}{f'(x_k)} . \quad (4)$$

Să presupunem că  $b_i = 0$  pentru  $i \neq j$  și  $b_j = a_j$ . Așadar, modificarea polinomului  $f$  constă în faptul că se înlocuiește coeficientul  $a_j$  cu coeficientul  $\tilde{a}_j = (1 + \varepsilon)a_j$ , iar ceilalți coeficienți rămân neschimbați. Din (4) rezultă

$$z_k - x_k \approx \varepsilon \frac{a_j x_k^j}{f'(x_k)} . \quad (5)$$

**Exemplul 1.** Fie

$$f(x) = \prod_{k=1}^{12} (x - k) = (x - 1)(x - 2) \dots (x - 12) .$$

Evident

$$x_k = k, \quad k = \overline{1,12} \quad \text{și} \quad f'(x_k) = (-1)^k (12-k)!(k-1)!$$

Conform (5) avem

$$|z_k - x_k| = \varepsilon \frac{|a_j| k^j}{(12-k)!(k-1)!} .$$

Se poate arăta că  $a_7 = -6.926.634$  .

Să presupunem că  $\varepsilon = 10^{-11}$ , ceea ce înseamnă că modificarea coeficientului  $a_7$  se face cu cantitatea  $\varepsilon \cdot a_7 = -0.00006926634 \approx -0.00007$  .

Acest lucru este oricând posibil datorită erorilor inerente la introducerea datelor. Să analizăm efectul acestei modificări asupra rădăcinii  $x_9 = 9$  a ecuației  $f(x) = 0$ . Un calcul direct ne arată că

$$|z_9 - x_9| = 0.00137 \approx 0.0014 .$$

Așadar, modificând un singur coeficient și anume  $a_7$  cu  $0.00007$ , rădăcina  $x_9$  se modifică cu  $0.0014$ . Raportul dintre modificarea rădăcinii  $x_9$  și modificarea coeficientului  $a_7$  este  $20$ , ceea ce arată *sensibilitatea* rădăcinilor unui polinom la modificarea coeficienților.

Din cele de mai sus rezultă că nu se recomandă determinarea valorilor proprii ale unei matrice pe calea rezolvării numerice a ecuației caracteristice.

Metoda recomandată este să se aducă, printr-un procedeu oarecare, matricea la forma diagonală și atunci valorile proprii se determină global (toate odată), ele fiind, de fapt, elementele de pe diagonala principală. Se urmărește deci, ca prin transformări de similitudine, care nu modifică valorile proprii, să micșorăm, eventual până la dispariție, elementele nediagonale ale matricei, astfel încât, în final, să obținem practic matricea diagonală.

### §3.1. Metoda rotațiilor a lui Jacobi

Fie  $A$  o matrice simetrică. Metoda Jacobi constă în efectuarea unei suite de transformări de similitudine ale matricei  $A$  utilizând cele mai simple matrice ortogonale netriviabile (matricele de rotație) de forma

$$U = \begin{pmatrix} 1 & 0 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & 1 & 0 & \cdot & \cdot & \cdot & 0 \\ & & \ddots & \cdot & \cdot & \cdot & \\ 0 & \cdot & \cdot & \cos \varphi & \cdot & \cdot & \sin \varphi & \cdot & \cdot & 0 \\ \vdots & & & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \vdots \\ & & & \cdot & \cdot & \ddots & \cdot & \cdot & \cdot & \\ 0 & \cdot & \cdot & -\sin \varphi & \cdot & \cdot & \cos \varphi & \cdot & \cdot & 0 \\ & & & \cdot & \cdot & \cdot & \cdot & \ddots & \cdot & \\ 0 & \cdot & \cdot & \cdot & \cdot & \cdot & 0 & 1 & 0 \\ & & & & & & 0 & 0 & 1 \end{pmatrix} \begin{matrix} \leftarrow p \\ \\ \\ \leftarrow q \\ \\ \end{matrix} \quad (1)$$

Așadar, elementele matricei  $U$  sunt:

$$\begin{cases} u_{ii} = 1 & \text{dacă } i \neq p \text{ și } i \neq q \\ u_{pp} = \cos \varphi, & u_{pq} = \sin \varphi \\ u_{qp} = -\sin \varphi, & u_{qq} = \cos \varphi \\ u_{ij} = 0 & \text{în rest} \end{cases} \quad (2)$$

O asemenea matrice este ortogonală ( $U^T U = I$  și deci  $U^{-1} = U^T$ ) și reprezintă din punct de vedere geometric o rotație de unghi  $\varphi$  în planul determinat de direcțiile  $e_p$  și  $e_q$ . Notăm cu  $A' = U^T A$  și cu  $A'' = A' U = U^T A U$ . În cazul particular  $n = 5$ ,  $p = 2$  și  $q = 4$ , matricea  $A''$  arată astfel

$a_{11}$	$a_{12} \cos \varphi - a_{14} \sin \varphi$	$a_{13}$	$a_{12} \sin \varphi + a_{14} \cos \varphi$	$a_{15}$
$a_{21} \cos \varphi -$ $a_{41} \sin \varphi$	$a_{22} \cos^2 \varphi - 2a_{24} \sin \varphi$ $\cos \varphi + a_{44} \sin^2 \varphi$	$a_{23} \cos \varphi -$ $a_{43} \sin \varphi$	$(a_{22} - a_{44}) \sin \varphi \cos \varphi +$ $a_{24} \cos 2\varphi$	$a_{25} \cos \varphi -$ $a_{45} \sin \varphi$
$a_{31}$	$a_{32} \cos \varphi - a_{34} \sin \varphi$	$a_{33}$	$a_{32} \sin \varphi + a_{34} \cos \varphi$	$a_{35}$
$a_{21} \sin \varphi +$ $a_{41} \cos \varphi$	$(a_{22} - a_{44}) \sin \varphi$ $\cos \varphi + a_{24} \cos 2\varphi$	$a_{23} \sin \varphi +$ $a_{43} \cos \varphi$	$a_{22} \sin^2 \varphi + 2a_{24} \sin \varphi$ $\cos \varphi + a_{44} \cos^2 \varphi$	$a_{25} \sin \varphi +$ $a_{45} \cos \varphi$
$a_{51}$	$a_{52} \cos \varphi - a_{54} \sin \varphi$	$a_{53}$	$a_{52} \sin \varphi + a_{54} \cos \varphi$	$a_{55}$

În general, elementele matricei  $A'$  sunt

$$\begin{cases} a'_{ij} = a_{ij} & \text{dacă } i \neq p \text{ și } i \neq q \\ a'_{pj} = a_{pj} \cos \varphi - a_{qj} \sin \varphi \\ a'_{qj} = a_{pj} \sin \varphi + a_{qj} \cos \varphi \end{cases}, \quad j = \overline{1, n} \quad (3)$$

iar cele ale matricei  $A''$  sunt

$$\begin{cases} a''_{ij} = a'_{ij} & \text{dacă } j \neq p \text{ și } j \neq q \\ a''_{ip} = a'_{ip} \cos \varphi - a'_{iq} \sin \varphi \\ a''_{iq} = a'_{ip} \sin \varphi + a'_{iq} \cos \varphi \end{cases}, i = \overline{1, n} . \quad (4)$$

Din (3) și (4) rezultă

$$\begin{cases} a''_{pp} = a_{pp} \cos^2 \varphi - 2a_{pq} \cos \varphi \sin \varphi + a_{qq} \sin^2 \varphi \\ a''_{qq} = a_{pp} \sin^2 \varphi + 2a_{pq} \cos \varphi \sin \varphi + a_{qq} \cos^2 \varphi \\ a''_{pq} = (a_{pp} - a_{qq}) \sin \varphi \cos \varphi + a_{pq} \cos 2\varphi \\ a''_{qp} = a''_{pq} \end{cases} . \quad (5)$$

Cum intenția noastră este ca elementul nedijagonal cel mai mare (în valoare absolută) să se anuleze în urma rotației, vom alege liniile  $p$  și  $q$  astfel încât  $a_{pq}$  să fie cel mai mare element (în valoare absolută) de deasupra diagonalei principale și vom pune condiția ca  $a''_{pq} = 0$ . Ținând seama de (5) rezultă

$$\frac{1}{2}(a_{pp} - a_{qq}) \sin 2\varphi + a_{pq} \cos 2\varphi = 0$$

și mai departe

$$\operatorname{ctg} 2\varphi = \frac{a_{qq} - a_{pp}}{2a_{pq}} . \quad (6)$$

Așadar, unghiul de rotație se află din relația (6). Introducem notațiile

$$\theta = \frac{a_{qq} - a_{pp}}{2a_{pq}} \text{ și } \operatorname{tg} \varphi = t . \quad (7)$$

Cum  $\operatorname{ctg} 2\varphi = \frac{1 - \operatorname{tg}^2 \varphi}{2\operatorname{tg} \varphi}$ , din (6) și (7) rezultă  $t^2 + 2\theta \cdot t - 1 = 0$ . Rezolvând

această ecuație obținem

$$t_{1,2} = -\theta \pm \sqrt{\theta^2 + 1} = \frac{1}{\theta \pm \sqrt{\theta^2 + 1}} .$$

Pentru a evita ca numitorul să fie mic, luăm

$$t = \begin{cases} \frac{1}{\theta + \operatorname{sgn}(\theta)\sqrt{\theta^2 + 1}} & \text{dacă } \theta \neq 0 \\ 1 & \text{dacă } \theta = 0 \end{cases} . \quad (8)$$

Conform unor formule elementare de trigonometrie avem

$$\begin{cases} c = \cos \varphi = \frac{1}{\sqrt{1+t^2}} \\ s = \sin \varphi = \frac{t}{\sqrt{1+t^2}} \end{cases} . \quad (9)$$

Din (8) și (9) rezultă că  $|t| \leq 1$ ,  $c \geq \frac{1}{\sqrt{2}}$ ,  $|s| \leq \frac{1}{\sqrt{2}}$  și deci că

$$\varphi \in \left[ -\frac{\pi}{4}, \frac{\pi}{4} \right] .$$

Dacă notăm cu  $S(B)$  suma pătratelor elementelor nediagonale ale unei matrice  $B$  oarecare, atunci din (3) și (4), un calcul direct ne conduce la

$$S(A'') = [S(A) - 2a_{pq}^2] + 2a_{pq}''^2 .$$

Așadar, dacă alegem unghiul rotație  $\varphi$  conform (8) și (9) rezultă

$$a_{pq}'' = 0 \quad \text{și deci} \quad S(A'') = S(A) - 2a_{pq}^2 . \quad (10)$$

Deoarece  $a_{ij}^2 \leq a_{pq}^2$  pentru  $i \neq j$ , vom avea

$$S(A) \leq n(n-1)a_{pq}^2 \quad \text{sau} \quad -\frac{2}{n(n-1)}S(A) \geq -2a_{pq}^2 . \quad (11)$$

Din (10) și (11) rezultă

$$S(A'') \leq S(A) \left( 1 - \frac{2}{n(n-1)} \right) < S(A) \quad \text{pentru} \quad n \geq 2 . \quad (12)$$

Să considerăm acum un șir de rotații în urma cărora se obțin matricele  $A_0, A_1, A_2, \dots, A_k, \dots$  unde  $A_0 = A, A_1 = A'', A_2 = A_1''$ , etc.

Din (12) rezultă

$$S(A_k) \leq \left( 1 - \frac{2}{n(n-1)} \right)^k S(A) . \quad (13)$$

Cum  $1 - \frac{2}{n(n-1)} \in (0,1)$  pentru  $n > 2$ , din (13) rezultă  $\lim_{k \rightarrow \infty} S(A_k) = 0$ .

Așadar, la limită șirul  $\{A_k\}$  tinde la matricea diagonală.

Se poate demonstra următoarea teoremă

**Teorema 1.** Fie  $\lambda_i$  valorile proprii ale matricei  $A$  și fie  $a_{jj}^{(k)}$  elementele diagonale ale matricei  $A_k$ . Atunci

$$\left| a_{jj}^{(k)} - \lambda_j \right| \leq \sqrt{S(A_k)} .$$

Deoarece  $a_{pq}^{(k)}$  este cel mai mare (în valoare absolută) element nediagonal al matricei  $A_k$ , rezultă

$$S(A_k) \leq (n^2 - n)(a_{pq}^{(k)})^2 < n^2 (a_{pq}^{(k)})^2 .$$

Din Teorema 1 obținem

$$\left| a_{jj}^{(k)} - \lambda_j \right| < n \left| a_{pq}^{(k)} \right| . \quad (14)$$

Inegalitatea (14) poate fi luată drept *criteriu de oprire*. Din inegalitatea  $n \left| a_{pq}^{(k)} \right| < \varepsilon$ , va rezulta numărul  $k$  al rotațiilor necesare pentru a aproxima valorile

proprii  $\lambda_j$  ale matricei  $A$  cu elementele diagonale  $a_{jj}^{(k)}$  ale matricei  $A_k$ . Șirul de matrice  $A_k$  se calculează recursiv

$$\begin{cases} A_k = U_k^T A_{k-1} U_k, & k \geq 1 \\ A_0 = A \end{cases} . \quad (15)$$

*Algoritm pentru determinarea valorilor proprii prin metoda rotațiilor a lui Jacobi*

Intrare  $A$ ,  $\varepsilon$ ;

Repetă

Determină :  $max$  := elementul maxim în valoare absolută de  
deasupra diagonalei principale a matricei  $A$  ;

Fie  $(p, q)$  poziția acestui element ;

Calculează cu formulele (7), (8), (9) respectiv  $\theta$ ,  $t$ ,  $c$ ,  $s$  ;

Determină  $U$  prin înlocuirea în  $I_n$  a elementelor  $i_{pp}$  și  $i_{qq}$  cu  
 $c$  și  $i_{pq}$  cu  $s$ , iar  $i_{qp}$  cu  $-s$  ;

Calculează  $A := U^T \cdot A \cdot U$ ; calculează  $S := \sqrt{\sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n a_{ij}^2}$

până când  $S < \varepsilon$  .

**Exemplul 1.** Pentru matricea

$$A = \begin{pmatrix} 3 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & 1 & 3 \end{pmatrix}, \quad a_{12} = 1; \quad p = 1; \quad q = 2; \quad \theta = 0; \quad t = 1; \quad c = s = \frac{1}{\sqrt{2}}$$

$$U_1 = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 1 \end{pmatrix}; \quad A_1 = U_1^T A U_1 = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 4 & \sqrt{2} \\ 0 & \sqrt{2} & 3 \end{pmatrix}; \quad a_{23} = \sqrt{2}; \quad p = 2; \quad q = 3$$

$$\theta = \frac{3-4}{2\sqrt{2}} = -\frac{1}{2\sqrt{2}}; \quad 1 + \theta^2 = \frac{9}{8}; \quad t = -\frac{\sqrt{2}}{2}; \quad 1 + t^2 = \frac{3}{2}; \quad c = \frac{\sqrt{2}}{\sqrt{3}}; \quad s = -\frac{1}{\sqrt{3}}$$

$$U_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{\sqrt{2}}{\sqrt{3}} & -\frac{1}{\sqrt{3}} \\ 0 & \frac{1}{\sqrt{3}} & \frac{\sqrt{2}}{\sqrt{3}} \end{pmatrix}; \quad A_2 = U_2^T A_1 U_2 = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

Rezultă:  $\lambda_1 = 2$  ;  $\lambda_2 = 5$  ;  $\lambda_3 = 2$  .

### **§3.2. Metoda Householder pentru tridiagonalizarea matricelor simetrice**

Pentru matricele simetrice tridiagonale există o metodă specială de determinare a valorilor proprii, bazată pe conceptul algebric de *șir Sturm*; această metodă va fi prezentată în paragraful următor.

Prezintă deci interes cunoașterea unor metode de tridiagonalizare a matricelor simetrice. Cele mai cunoscute metode din această categorie sunt *metoda Givens* și *metoda Householder*.

Așa cum am văzut în Capitolul I, §4, pentru orice vector  $x \in \mathbb{R}^n$ ,  $x \neq 0$ , există o matrice Householder  $H$ , astfel încât  $Hx = \sigma e_1$ , unde  $\sigma$  este un număr real. Algoritmul pentru determinarea matricei  $H$  este prezentat în (3), respectiv (4). Fie  $A \in M_n(\mathbb{R})$  o matrice simetrică și fie  $a_i = (a_{1i}, a_{2i}, \dots, a_{ni})^T$ ,  $i = \overline{1, n}$ , vectorii săi coloană.

Căutăm o matrice Householder  $H_1$  astfel încât

$$H_1 a_1 = \begin{pmatrix} a_{11} \\ a_{21}^{(1)} \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Pentru aceasta, alegem  $H_1$  de forma  $H_1 = \begin{pmatrix} 1 & 0 \\ 0 & \tilde{H}_1 \end{pmatrix}$ , unde  $\tilde{H}_1$  este matricea Householder de ordinul  $(n-1)$  cu proprietatea că

$$\tilde{H}_1 \cdot \begin{pmatrix} a_{21} \\ \cdot \\ \cdot \\ a_{n1} \end{pmatrix} = \begin{pmatrix} a_{21}^{(1)} \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Conform algoritmului (4) descris în Capitolul I §4 avem:

$$s = (a_{21}^2 + \dots + a_{n1}^2)^{1/2}, \quad \beta = (s(|a_{21}| + s))^{-1}, \quad u = (a_{21} + s \cdot \text{sgn}(a_{21}), a_{31}, \dots, a_{n1})^T,$$

$$\text{sgn}(a_{21}) = 1 \text{ dacă } a_{21} > 0, \quad \tilde{H}_1 = I_{n-1} - \beta uu^T.$$

Dacă notăm cu  $\tilde{a}_1 = (a_{21}, \dots, a_{n1})^T$ , atunci

$$\tilde{H}_1 \tilde{a}_1 = \begin{pmatrix} -s \cdot \text{sgn}(a_{21}) \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$\text{Mai departe avem } H_1 a_1 = \begin{pmatrix} 1 & 0 \\ 0 & \tilde{H}_1 \end{pmatrix} \begin{pmatrix} a_{11} \\ \tilde{a}_1 \end{pmatrix} = \begin{pmatrix} a_{11} \\ \tilde{H}_1 \cdot \tilde{a}_1 \end{pmatrix} = \begin{pmatrix} a_{11} \\ -s \cdot \text{sgn}(a_{21}) \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Fie  $A_1 = H_1 A H_1^T$ . Atunci

$$A_1 = \begin{pmatrix} a_{11} & -s \cdot \text{sgn}(a_{21}) & 0 & \dots & 0 \\ -s \cdot \text{sgn}(a_{21}) & a_{22}^{(1)} & a_{23}^{(1)} & \dots & a_{2n}^{(1)} \\ 0 & a_{32}^{(1)} & a_{33}^{(1)} & \dots & a_{3n}^{(1)} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \dots & \dots & \dots & \dots & \dots \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & a_{n2}^{(1)} & a_{n3}^{(1)} & \dots & a_{nn}^{(1)} \end{pmatrix}.$$

În continuare se caută o matrice Householder  $H_2$  cu proprietatea că elementele  $a_{i2}^{(2)}$  ( $i = \overline{4, n}$ ) din matricea  $A_2 = H_2 A_1 H_2^T$  sunt nule, etc.

*Algoritm pentru tridiagonalizarea matricei A*

$A_0 := A$ ;

Pentru  $i := 1, n-1$  calculează

$$s := \left( \sum_{j=i+1}^n a_{ij}^2 \right)^{1/2}; \quad \beta := (s(|a_{i+1,i}| + s))^{-1};$$

$$u = (a_{i+1,i} + s \cdot \operatorname{sgn}(a_{i+1,i}), a_{i+2,i}, \dots, a_{ni})^T ;$$

dacă  $a_{i,i+1} \geq 0$  atunci  $\operatorname{sgn}(a_{i,i+1}) := 1$  altfel  $\operatorname{sgn}(a_{i,i+1}) := -1$  ;

$$\tilde{H}_i := I_{n-1} - \beta \cdot u \cdot u^T ;$$

$$H_i := \begin{pmatrix} I_i & 0 \\ 0 & \tilde{H}_i \end{pmatrix} ;$$

$$A_i := H_i A_{i-1} H_i^T ;$$

sfârșit pentru  $i$ .

### §3.3. Determinarea valorilor proprii ale matricelor simetrice tridiagonale

Următoarea teoremă precizează mulțimea din planul complex (respectiv intervalul din  $\mathbb{R}$ ) în care se află valorile proprii ale unei matrice.

**Teorema 1. (Gerschgorin).** Fie  $A$  o matrice pătratică de ordinul  $n$ ,

$$r_i = \sum_{\substack{j=1 \\ j \neq i}}^n |a_{ij}| \quad \text{și} \quad D_i = \{z \in \mathbb{C}; |z - a_{ii}| < r_i\} \quad ; \quad i = \overline{1, n}.$$

Dacă  $\lambda$  este valoarea proprie a matricei  $A$ , atunci  $\lambda \in \bigcup_{i=1}^n D_i$ .

**Demonstrație.** Fie  $\lambda$  o valoare proprie a matricei  $A$  și fie  $x = (x_1, \dots, x_n)^T$  un vector propriu corespunzător lui  $\lambda$ . Atunci  $x \neq 0$  și  $Ax = \lambda x$ . Rezultă

$$a_{i1}x_1 + \dots + a_{ii}x_i + \dots + a_{in}x_n = \lambda x_i$$

sau

$$\sum_{\substack{j=1 \\ j \neq i}}^n a_{ij}x_j = (\lambda - a_{ii})x_i, \quad i = \overline{1, n} \quad (1)$$

Fie  $p \in \{1, 2, \dots, n\}$  astfel încât  $|x_p| = \|x\|_\infty > 0$ .

Din (1) rezultă

$$|\lambda - a_{pp}| \leq \sum_{\substack{j=1 \\ j \neq p}}^n |a_{pj}| \left| \frac{x_j}{x_p} \right| \leq \sum_{\substack{j=1 \\ j \neq p}}^n |a_{pj}| = r_p .$$

Așadar,  $\lambda \in D_p \subset \bigcup_{i=1}^n D_i$ .

În cazul particular când  $A \in M_n(\mathbb{R})$  și are toate valorile proprii reale, rezultă că

$$\lambda \in \bigcup_{i=1}^n [a_{ii} - r_i, a_{ii} + r_i] \subset \mathbb{R}. \quad \square$$

**Exemplul 1.** Fie  $A = \begin{pmatrix} 1 & 2 & 1 \\ 2 & -3 & -2 \\ 1 & -2 & 2 \end{pmatrix}$ ,  $r_1=3$ ;  $r_2=4$ ;  $r_3=3$ .

$$\lambda \in [-2, 4] \cup [-7, 1] \cup [-1, 5] = [-7, 5].$$

O matrice simetrică tridiagonală este de forma

$$J = \begin{pmatrix} a_1 & b_1 & 0 & 0 & \dots & \dots & \dots & 0 \\ b_1 & a_2 & b_2 & 0 & \dots & \dots & \dots & 0 \\ 0 & b_2 & a_3 & b_3 & \dots & \dots & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & b_{n-2} & a_{n-1} & b_{n-1} \\ 0 & 0 & 0 & 0 & \dots & 0 & b_{n-1} & a_n \end{pmatrix}. \quad (2)$$

Pentru o astfel de matrice avem

$$a_{ii} = a_i; \quad r_1 = |b_1|; \quad r_n = |b_{n-1}|; \quad r_i = |b_{i-1}| + |b_i|, \quad i = \overline{2, n-1}.$$

Fie  $a = \min_{1 \leq i \leq n} (a_i - r_i)$  și  $b = \max_{1 \leq i \leq n} (a_i + r_i)$ .

Valorile proprii ale matricei  $A$  vor aparține intervalului  $[a, b]$ .

**Definiția 1.** Un șir ordonat și finit de polinoame reale  $f_n, f_{n-1}, \dots, f_1, f_0$  se numește șir Sturm, dacă:

1. Polinoamele vecine nu au rădăcini comune;
2.  $f_0$  nu are rădăcini reale;
3. Dacă  $x = \alpha$  este o rădăcină a unuia din polinoamele intermediare

$$f_i, \quad i = \overline{1, n-1}, \quad \text{atunci} \quad f_{i-1}(\alpha)f_{i+1}(\alpha) < 0;$$

4. Dacă  $f_n(\alpha) = 0$ , atunci pentru  $h > 0$ , suficient de mic, avem

$$\operatorname{sgn} \frac{f_n(\alpha - h)}{f_{n-1}(\alpha - h)} = -1 \quad \text{și} \quad \operatorname{sgn} \frac{f_n(\alpha + h)}{f_{n-1}(\alpha + h)} = +1.$$

În continuare, pentru orice  $x \in \mathbb{R}$ , notăm cu  $S(x)$  numărul schimbărilor de semn din șirul

$$f_n(x), f_{n-1}(x), \dots, f_1(x), f_0(x),$$

după ce am eliminat elementele nule.

**Teorema 2. (Sturm)** Fie  $f_n, f_{n-1}, \dots, f_1, f_0$  un șir Sturm de polinoame. Dacă numerele reale  $a$  și  $b$ ,  $a < b$ , nu sunt rădăcini ale polinomului  $f_n$  și

dacă polinomul  $f_n$  nu are rădăcini multiple, atunci  $S(a) \geq S(b)$  și diferența  $S(a) - S(b)$  este egală cu numărul rădăcinilor reale ale polinomului  $f_n$  din intervalul  $(a, b)$ .

**Demonstrație.** Deoarece polinoamele sunt funcții continue, atât timp cât  $x$  crescând, nu întâlnește nici o rădăcină a vreunui din polinoamele din șir, semnele polinoamelor din șir nu se schimbă și deci  $S(x)$  rămâne neschimbat. Rămâne să analizăm următoarele cazuri posibile:

a)  $x = \alpha$  este rădăcină pentru unul din polinoamele intermediare. Fie  $i \in \{1, 2, \dots, (n-1)\}$  astfel încât  $f_i(\alpha) = 0$ . Din Definiția 1 rezultă  $f_{i-1}(\alpha)f_{i+1}(\alpha) < 0$ . Să presupunem că  $f_{i-1}(\alpha) < 0$  și  $f_{i+1}(\alpha) > 0$ . Din continuitate rezultă că există  $h > 0$  astfel încât  $f_{i-1}(x) < 0$  și  $f_{i+1}(x) > 0$  pentru orice  $x \in [\alpha - h, \alpha + h]$ . Avem următoarea situație

$x$	$f_{i-1}(x)$	$f_i(x)$	$f_{i+1}(x)$
$\alpha - h$	-	$\pm$	+
$\alpha$	-	0	+
$\alpha + h$	-	$\mp$	+

Rezultă  $S(\alpha + h) = S(\alpha - h)$ .

În mod analog, dacă  $f_{i-1}(\alpha) > 0$  și  $f_{i+1}(\alpha) < 0$  avem următorul tabel ale semnelor

$x$	$f_{i-1}(x)$	$f_i(x)$	$f_{i+1}(x)$
$\alpha - h$	+	$\pm$	-
$\alpha$	+	0	-
$\alpha + h$	+	$\mp$	-

Rezultă, de asemenea  $S(\alpha + h) = S(\alpha - h)$ .

b)  $x = \alpha$  este o rădăcină a polinomului  $f_n$ . Evident, în acest caz  $f_{n-1}(\alpha) \neq 0$  (Definiția 1, proprietatea 1).

Din continuitate și din proprietatea 4 a Definiției 1, rezultă că nu putem avea decât următoarele situații

$x$	$f_{n-1}(x)$	$f_n(x)$	$x$	$f_{n-1}(x)$	$f_n(x)$
$\alpha - h$	-	+	$\alpha - h$	+	-
$\alpha$	-	0	$\alpha$	+	0
$\alpha + h$	-	-	$\alpha + h$	+	+

Așadar, la trecerea printr-o rădăcină a polinomului  $f_n$ ,  $S$  scade cu o unitate ( $S(\alpha + h) = S(\alpha - h) - 1$ ).

În definitiv, am demonstrat că numărul rădăcinilor reale ale polinoamelor  $f_n$ , cuprinse în intervalul  $(a, b)$  este egal cu

$S(a) - S(b)$ . □

**Exemplul 2.**

Fie polinoamele:

$$f_3(x) = x^3 - 3x^2 + 1;$$

$$f_2(x) = x^2 - 3x + 1;$$

$$f_1(x) = x - 1;$$

$$f_0(x) = 1$$

Din reprezentarea grafică a polinoamelor se observă că ele formează un șir Sturm. Alegem  $a = -1$  și  $b = 3$ .

$$f_3(-1) = -3; \quad f_2(-1) = 5; \quad f_1(-1) = -2;$$

$$f_0(-1) = 1$$

$$f_3(3) = 1; \quad f_2(3) = 1; \quad f_1(3) = 2;$$

$$f_0(3) = 1; \quad S(-1) = 3; \quad S(3) = 0;$$

Numărul rădăcinilor reale ale polinomului  $f_3$  cuprinse în intervalul  $(-1, 3)$  este 3.

Fie  $J$  matricea simetrică tridiagonală dată de (2) și fie

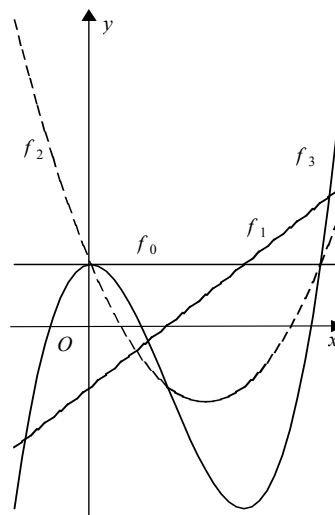
$$P(\lambda) = \det(\lambda I - J) = \begin{vmatrix} \lambda - a_1 & -b_1 & 0 & 0 & \dots & 0 & 0 \\ -b_1 & \lambda - a_2 & -b_2 & 0 & \dots & 0 & 0 \\ 0 & -b_2 & \lambda - a_3 & -b_3 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & -b_{n-1} & \lambda - a_n \end{vmatrix}$$

Introducem următoarele notații

$$\begin{cases} f_0(\lambda) = 1 \\ f_1(\lambda) = \lambda - a_1 \\ f_2(\lambda) = (\lambda - a_2)f_1(\lambda) - b_1^2 f_0(\lambda) \\ f_3(\lambda) = (\lambda - a_3)f_2(\lambda) - b_2^2 f_1(\lambda) \\ \dots \\ f_n(\lambda) = (\lambda - a_n)f_{n-1}(\lambda) - b_{n-1}^2 f_{n-2}(\lambda) \end{cases} \quad (3)$$

Se observă imediat că  $f_n(\lambda) \equiv P(\lambda)$  este polinomul caracteristic atașat matricei  $A$ .

**Teorema 3.** Dacă  $b_i \neq 0$ ,  $i = \overline{1, n-1}$ , atunci fiecare polinom  $f_k$ ,  $k = \overline{0, n}$ , are exact  $k$  rădăcini reale simple. Mai mult, pentru orice  $1 \leq k \leq n-1$ , rădăcinile polinomului  $f_k$  separă rădăcinile polinomului  $f_{k+1}$ .



**Demonstrație.** Polinomul  $f_1$  admite rădăcina  $\lambda_1^{(1)} = a_1$ . Din (3) și din ipoteza  $b_1 \neq 0$  rezultă  $f_2(\lambda_1^{(1)}) = -b_1^2 < 0$ .

Pe de altă parte, deoarece  $f_2(\lambda) = \lambda^2 + \dots$  și deci  $\lim_{\lambda \rightarrow \pm\infty} f_2(\lambda) = +\infty$ , rezultă că există  $\lambda_1^{(2)}, \lambda_2^{(2)}$  rădăcini reale ale lui  $f_2$ , astfel încât

$$\lambda_1^{(2)} < \lambda_1^{(1)} < \lambda_2^{(2)}.$$

Ținând din nou seama de (3) și de ipoteza  $b_2 \neq 0$ , rezultă

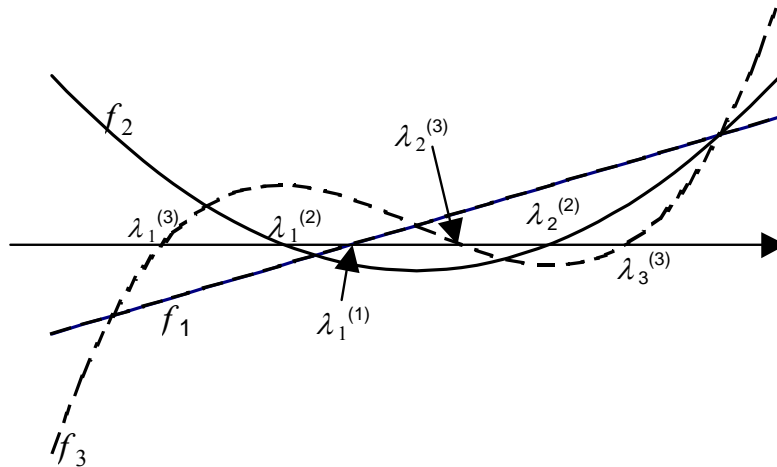
$$f_3(\lambda_1^{(2)}) = -b_2^2 f_1(\lambda_1^{(2)}) > 0 \quad \text{și} \quad f_3(\lambda_2^{(2)}) = -b_2^2 f_1(\lambda_2^{(2)}) < 0$$

Cum  $f_3(\lambda) = \lambda^3 + \dots$  avem  $\lim_{\lambda \rightarrow -\infty} f_3(\lambda) = -\infty$  și  $\lim_{\lambda \rightarrow \infty} f_3(\lambda) = +\infty$ .

Așadar, polinomul  $f_3$  admite 3 rădăcini reale simple

$$\lambda_1^{(3)} \in (-\infty, \lambda_1^{(2)}), \quad \lambda_2^{(3)} \in (\lambda_1^{(2)}, \lambda_2^{(2)}) \quad \text{și} \quad \lambda_3^{(3)} \in (\lambda_2^{(2)}, \infty).$$

Prin inducție matematică se poate arăta că  $f_k$  are  $k$  rădăcini reale simple și separă rădăcinile polinomului  $f_{k+1}$ .  $\square$



**Corolarul 1.** Orice matrice simetrică tridiagonală ireductibilă are  $n$  valori proprii reale distincte.

Într-adevăr, conform Definiției 2, Capitolul I, §2, dacă matricea  $J$  este ireductibilă, atunci  $b_i \neq 0$ ,  $i = \overline{1, n-1}$ . Afirmția rezultă acum din Teorema 3 și din observația că  $f_n(\lambda) \equiv P(\lambda)$  este polinomul caracteristic al matricei  $J$ .

**Teorema 4.** Dacă  $J$  este o matrice simetrică tridiagonală ireductibilă și

$$f_0(\lambda) = 1; f_1(\lambda) = \lambda - a_1; f_k(\lambda) = (\lambda - a_k)f_{k-1}(\lambda) - b_{k-1}^2 f_{k-2}(\lambda), \quad k = \overline{2, n},$$

atunci  $f_n, f_{n-1}, \dots, f_1, f_0$  este un șir Sturm.

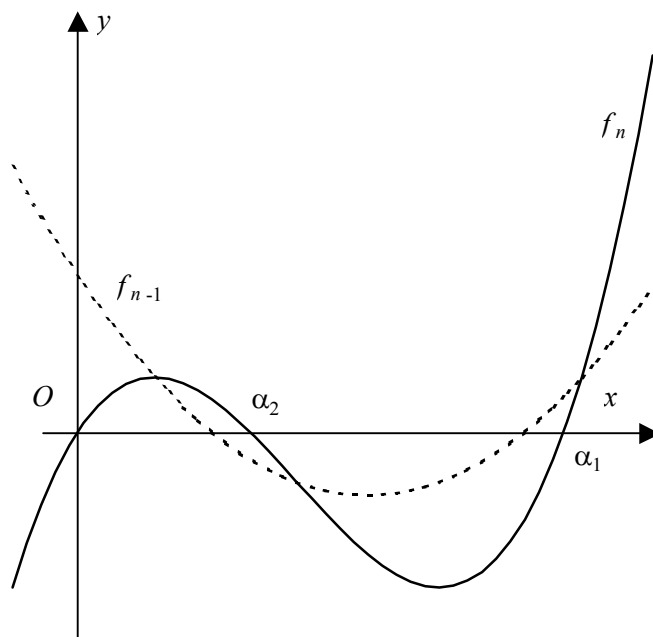
**Demonstrație.** Evident  $f_0(\lambda) \neq 0$ , pentru orice  $\lambda \in \mathbb{R}$ . Fie  $k \in \{1, 2, \dots, n-1\}$  și  $\alpha \in \mathbb{R}$  astfel încât  $f_k(\alpha) = 0$ . Atunci  $f_{k+1}(\alpha) = -b_k^2 f_{k-1}(\alpha)$ . Din Teorema 3 rezultă  $f_{k+1}(\alpha) \neq 0$  și  $f_{k-1}(\alpha) \neq 0$ , iar din egalitatea precedentă rezultă  $f_{k+1}(\alpha)f_{k-1}(\alpha) < 0$ .

Fie  $x = \alpha_1$  cea mai mare rădăcină a polinomului  $f_n$ . Din Teorema 3 și din faptul că  $\lim_{x \rightarrow \infty} f_n(x) = \lim_{x \rightarrow \infty} f_{n-1}(x) = +\infty$  rezultă  $f_{n-1}(\alpha_1) > 0$  și mai departe că

$$\operatorname{sgn} \frac{f_n(\alpha_1 + h)}{f_{n-1}(\alpha_1 + h)} = 1 \quad \text{și} \quad \operatorname{sgn} \frac{f_n(\alpha_1 - h)}{f_{n-1}(\alpha_1 - h)} = -1$$

pentru  $h > 0$  suficient de mic. Dacă  $x = \alpha_2$  este următoarea rădăcină a polinomului  $f_n$ , vom avea  $f_{n-1}(\alpha_2) < 0$  și deci, pentru  $h > 0$  suficient de mic

$$\operatorname{sgn} \frac{f_n(\alpha_2 + h)}{f_{n-1}(\alpha_2 + h)} = 1 \quad \text{și} \quad \operatorname{sgn} \frac{f_n(\alpha_2 - h)}{f_{n-1}(\alpha_2 - h)} = -1 \quad \text{ș. a. m. d.} \quad \square$$



**Teorema 5.** Fie  $J$  o matrice simetrică tridiagonală ireductibilă și fie  $a \in \mathbb{R}$  oarecare. Atunci numărul valorilor proprii ale matricei  $J$ , mai mari ca  $a$ , este egal cu  $S(a)$ .

Afirmația rezultă din Teorema 4, Teorema 2 și din observația că dacă  $b > \alpha_1$ , unde  $\alpha_1$  este cea mai mare rădăcină a polinomului  $f_n$ , atunci  $S(b) = 0$ , deoarece  $f_i(b) > 0$ ,  $i = \overline{0, n}$ .

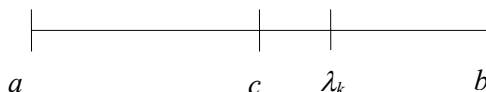
Teorema 5 ne permite să determinăm valorile proprii ale unei matrice simetrice tridiagonale ireductibilă cu metoda înjumătățirii.

Fie  $a, b \in \mathbb{R}$  astfel încât

$$a < \lambda_n < \dots < \lambda_2 < \lambda_1 < b.$$

Evident  $S(a) = n$  și  $S(b) = 0$ . Fie  $c$  mijlocul intervalului  $[a, b]$ .

Dorim să localizăm valoarea proprie  $\lambda_k$



Dacă  $S(c) \geq k$ , atunci la dreapta lui  $c$  se află  $k$  valori proprii, deci inclusiv  $\lambda_k$ . În acest caz notăm  $a_1 = c$ ,  $b_1 = b$ . Dacă dimpotrivă  $S(c) < k$ , atunci  $\lambda_k \in (a, c)$  și notăm  $a_1 = a$ ,  $b_1 = c$ .

Să presupunem că  $\lambda_k \in (c, b)$ . Fie  $c_1 = \frac{a_1 + b_1}{2}$ . Dacă  $S(c_1) \geq k$ , atunci notăm  $a_2 = c_1$ ,  $b_2 = b_1$ , iar dacă  $S(c_1) < k$ , atunci  $a_2 = a_1$ ,  $b_2 = c_1$  etc.

Rezultă că  $\lambda_k \in (a_p, b_p)$ , unde  $b_p - a_p = \frac{b - a}{2^p}$ . Putem alege  $\lambda_k \approx \frac{a_p + b_p}{2}$  și eroarea care se face va fi mai mică decât  $\frac{b - a}{2^p}$ .

**Exemplul 3.** Fie  $A = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix}$ . Atunci  $r_1 = 1, r_2 = 2, r_3 = 1$ ,

$$a = \min(2-1, 2-2, 2-1) = 0; \quad b = \max(2+1, 2+2, 2+1) = 4.$$

Din Teorema Gershgorin rezultă că valorile proprii se află în intervalul  $[0, 4]$ .

Fie  $\lambda_3 < \lambda_2 < \lambda_1$  aceste valori proprii. Să presupunem că vrem să determinăm valoarea proprie  $\lambda_1$ . Notăm cu

$$c = \frac{a + b}{2} = 2.$$

$$f_0(\lambda) = 1; \quad f_1(\lambda) = \lambda - 2; \quad f_2(\lambda) = (\lambda - 2)^2 - 1; \quad f_3(\lambda) = (\lambda - 2)^3 - 2(\lambda - 2)$$

$$f_0(2) = 1; \quad f_1(2) = 0; \quad f_2(2) = -1; \quad f_3(2) = 0; \quad S(2) = 1.$$

Rezultă că în intervalul  $[2,4]$  se află o singură valoare proprie, deci

$$\lambda_1 \in [2,4]. \text{ Fie } c_1 = \frac{2+4}{2} = 3 ,$$

$$f_0(3) = 1; f_1(3) = 1; f_2(3) = 0; f_3(3) = -1; S(3) = 1.$$

$$\text{Rezultă } \lambda_1 \in [3,4] \text{ Fie } c_2 = \frac{3+4}{2} = \frac{7}{2} .$$

$$f_0\left(\frac{7}{2}\right) = 1; f_1\left(\frac{7}{2}\right) = \frac{3}{2}; f_2\left(\frac{7}{2}\right) = \frac{5}{4}; f_3\left(\frac{7}{2}\right) = \frac{3}{8}; S\left(\frac{7}{2}\right) = 0 .$$

Așadar, la dreapta lui  $\frac{7}{2}$  nu se află nici o valoare proprie. Rezultă

$$\lambda_1 \in \left(3, \frac{7}{2}\right) \text{ etc.}$$

### Exerciții

1. Folosind metoda rotațiilor a lui Jacobi să se calculeze valorile și vectorii proprii

$$\text{pentru matricea } A = \begin{pmatrix} 1 & 1 & 3 \\ 1 & 3 & 1 \\ 3 & 1 & 1 \end{pmatrix} .$$

R.  $\max_{i < j} |a_{ij}| = 3 = a_{13}$  ,  $p = 1$  ,  $q = 3$  . Rezultă că

$$\theta = \frac{a_{qq} - a_{pp}}{2a_{pq}} = \frac{a_{33} - a_{11}}{2a_{13}} = \frac{1-1}{6} = 0 \text{ și } t = 1 ,$$

$$\cos \varphi = \frac{1}{\sqrt{1+t^2}} = \frac{1}{\sqrt{2}} ; \sin \varphi = \frac{t}{\sqrt{1+t^2}} = \frac{1}{\sqrt{2}} , \text{ iar}$$

$$U_1 = \begin{pmatrix} \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \\ -\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{pmatrix}.$$

$$\begin{aligned} A^{(1)} = U_1^T \cdot A \cdot U_1 &= \begin{pmatrix} \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 & 3 \\ 1 & 3 & 1 \\ 3 & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \\ -\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{pmatrix} = \\ &= \begin{pmatrix} -2 & 0 & 0 \\ 0 & 5 & \sqrt{2} \\ 0 & \sqrt{2} & 4 \end{pmatrix} \end{aligned}$$

$\max_{i < j} |a_{ij}^{(1)}| = \sqrt{2} = a_{23}^{(1)}$ ,  $p = 2$ ,  $q = 3$ . Rezultă că

$$\theta = \frac{a_{qq} - a_{pp}}{2a_{pq}} = \frac{a_{33}^{(1)} - a_{22}^{(1)}}{2a_{23}^{(1)}} = \frac{4 - 5}{2\sqrt{2}} = -\frac{1}{2\sqrt{2}} \text{ și } t = \frac{1}{\theta + \operatorname{sgn} \theta \sqrt{1 + \theta^2}} = -\frac{1}{2\sqrt{2}}$$

$$\cos \varphi = \frac{1}{\sqrt{1 + t^2}} = \sqrt{\frac{2}{3}}; \quad \sin \varphi = \frac{t}{\sqrt{1 + t^2}} = -\frac{1}{\sqrt{3}}, \text{ iar}$$

$$U_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{\sqrt{2}}{\sqrt{3}} & -\frac{1}{\sqrt{3}} \\ 0 & \frac{1}{\sqrt{3}} & \frac{\sqrt{2}}{\sqrt{3}} \end{pmatrix}.$$

$$A^{(2)} = U_2^T \cdot A^{(1)} \cdot U_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{\sqrt{2}}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ 0 & -\frac{1}{\sqrt{3}} & \frac{\sqrt{2}}{\sqrt{3}} \end{pmatrix} \cdot \begin{pmatrix} -2 & 0 & 0 \\ 0 & 5 & \sqrt{2} \\ 0 & \sqrt{2} & 4 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{\sqrt{2}}{\sqrt{3}} & -\frac{1}{\sqrt{3}} \\ 0 & \frac{1}{\sqrt{3}} & \frac{\sqrt{2}}{\sqrt{3}} \end{pmatrix} =$$

$$\begin{pmatrix} -2 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 3 \end{pmatrix}$$

Deoarece  $S(A^{(2)}) = 0$  am obținut chiar valorile proprii exacte pentru matricea  $A$ .

$$V = U_1 \cdot U_2 = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \\ 0 & \frac{\sqrt{2}}{\sqrt{3}} & -\frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \end{pmatrix} \text{ reprezintă matricea de trecere de la baza în}$$

care matricea  $A$  este dată (canonică) la baza în care  $A$  are forma diagonală. Se știe de la cursul de Algebră liniară că această bază este dată de coloanele matricei de trecere. Deci vectorii proprii se obțin ca fiind coloanele matricei de trecere, astfel:

$$\lambda_1 = -2, v_1 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \sqrt{2} \\ 0 \\ 1 \\ -\sqrt{2} \end{pmatrix}, \lambda_2 = 6, v_2 = \begin{pmatrix} \frac{1}{\sqrt{6}} \\ \sqrt{6} \\ \frac{\sqrt{2}}{\sqrt{3}} \\ 1 \\ \frac{1}{\sqrt{6}} \end{pmatrix}, \lambda_3 = 3, v_3 = \begin{pmatrix} \frac{1}{\sqrt{3}} \\ \sqrt{3} \\ 1 \\ \sqrt{3} \\ -\frac{1}{\sqrt{3}} \end{pmatrix}.$$

2. Folosind metoda Jacobi să se determine valorile proprii aproximative ale

$$\text{matricei } A = \begin{pmatrix} 1 & 2 & 4 & 3 \\ 2 & 1 & 3 & 5 \\ 4 & 3 & 1 & 4 \\ 3 & 5 & 4 & 1 \end{pmatrix}.$$

R. Procedând ca în exercițiul de mai sus se obțin succesiv:

$$U_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0.70711 & 0 & 0.70711 \\ 0 & 0 & 1 & 0 \\ 0 & -0.70711 & 0 & 0.70711 \end{pmatrix},$$

$$A_1 = \begin{pmatrix} 1 & -0.70711 & 4 & 3.53553 \\ -0.70711 & -4 & -0.70711 & 0 \\ 4 & -0.70711 & 1 & 4.94975 \\ 3.53553 & 0 & 4.94975 & 6 \end{pmatrix},$$

$$U_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0.85171 & 0.52401 \\ 0 & 0 & -0.52401 & 0.85171 \end{pmatrix},$$

$$A_2 = \begin{pmatrix} 1 & -0.70711 & 1.55422 & 5.10729 \\ -0.70711 & -4 & -0.60225 & -0.37053 \\ 1.55422 & -0.60225 & -2.04527 & 0 \\ 5.10729 & -0.37053 & 0 & 9.04527 \end{pmatrix},$$

$$U_3 = \begin{pmatrix} 0.89965 & 0 & 0 & 0.43661 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0.52401 \\ -0.43661 & 0 & 0 & 0.89965 \end{pmatrix},$$

$$A_3 = \begin{pmatrix} -1.4786 & -0.47438 & 1.39825 & 0 \\ -0.47438 & -4 & -0.60225 & -0.64207 \\ 1.39825 & -0.60225 & -2.04527 & 0.67858 \\ 0 & -0.64207 & 0.67858 & 11.52386 \end{pmatrix},$$

$$U_4 = \begin{pmatrix} 0.63301 & 0 & 0.77414 & 0 \\ 0 & 1 & 0 & 0 \\ -0.77414 & 0 & 0.63301 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

$$A_4 = \begin{pmatrix} -3.1886 & 0.16595 & 0 & -0.52532 \\ 0.16595 & -4 & -0.74847 & -0.64207 \\ 0 & -0.74847 & -3.3526 & 0.42955 \\ -0.52532 & -0.64207 & 0.42955 & 11.52386 \end{pmatrix},$$

și așa mai departe se obține la iterația a noua

$$U_9 = \begin{pmatrix} 0.1458 & 0 & 0 & 0.98931 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -0.98931 & 0 & 0 & 0.1458 \end{pmatrix},$$

$$A_9 = \begin{pmatrix} -4.18733 & 0.01986 & -0.00427 & 0 \\ 0.01986 & -0.21355 & 0.00026 & 0.00452 \\ -0.00452 & 0.00026 & 11.58733 & 0.00062 \\ 0 & 0.00452 & 0.00062 & -3.18645 \end{pmatrix}, \quad S(A_9)=0.02944,$$

valorile proprii exacte fiind:

$$\lambda_1=-3.18646, \lambda_2=-4.18743, \lambda_3=-0.21344, \lambda_4=11.58733.$$

3. Să se determine valorile proprii aproximative ale matricei

$$A = \begin{pmatrix} 1 & 1 & 1 & 4 \\ 1 & 1 & 0 & 5 \\ 1 & 0 & 1 & 4 \\ 4 & 5 & 4 & 1 \end{pmatrix}$$

folosind metoda Jacobi .

R. Procedând ca în exercițiul de mai sus se obține succesiv:

$$U_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0.70711 & 0 & 0.70711 \\ 0 & 0 & 1 & 0 \\ 0 & -0.70711 & 0 & 0.70711 \end{pmatrix},$$

$$A_1 = \begin{pmatrix} 1 & -2.12132 & 1 & 3.53553 \\ -2.12132 & -4 & -2.82843 & 0 \\ 1 & -2.82843 & 1 & 2.82843 \\ 3.53553 & 0 & 2.82843 & 6 \end{pmatrix},$$

$$U_2 = \begin{pmatrix} 0.88807 & 0 & 0 & 0.4597 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -0.4597 & 0 & 0 & 0.88807 \end{pmatrix},$$

$$A_2 = \begin{pmatrix} -0.83013 & -1.88389 & -0.41216 & 0 \\ -1.88389 & -4 & -2.82843 & -0.97517 \\ -0.41216 & -2.82843 & 1 & 2.97155 \\ 0 & -0.97517 & 2.97155 & 7.83013 \end{pmatrix},$$

$$U_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0.93659 & 0.35043 \\ 0 & 0 & -0.35043 & 0.93659 \end{pmatrix},$$

$$A_3 = \begin{pmatrix} -0.83013 & -1.88389 & -0.38602 & -0.14443 \\ -1.88389 & -4 & -2.30734 & -1.90451 \\ -0.38602 & -2.30734 & -0.11183 & 0 \\ -0.14443 & -1.90451 & 0 & 8.94196 \end{pmatrix},$$

$$U_4 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0.4217 & 0.90674 & 0 \\ 0 & -0.90674 & 0.4217 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

$$A_4 = \begin{pmatrix} -0.83013 & -0.44441 & -1.87097 & -0.14443 \\ -0.44441 & 0.96123 & 0 & -0.83313 \\ -1.87097 & 0 & -5.07308 & -1.72689 \\ -0.14443 & -0.83313 & -1.72689 & 8.94196 \end{pmatrix},$$

și așa mai departe se obține la iterația a zecea

$$U_{10} = \begin{pmatrix} 0.99999 & 0.005 & 0 & \\ -0.005 & 0.99999 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

$$A_{10} = \begin{pmatrix} 9.22995 & -0.02005 & -0.00611 & 0.02598 \\ -0.02005 & 1.01318 & 0.00089 & 0 \\ -0.00611 & 0.00089 & -0.26839 & -0.0252 \\ 0.02598 & 0 & -0.0252 & -5.97535 \end{pmatrix}, \quad S(A_{10})=0.05855,$$

valorile proprii exacte fiind:

$$\lambda_1=1.01373, \lambda_2=-0.26828, \lambda_3=-5.9755, \lambda_4=9.23005.$$

Să se aducă la forma tridiagonală matricele următoare folosind metoda Householder.

$$4. \quad A = \begin{pmatrix} 4 & 1 & -1 & 1 \\ 1 & 5 & 0 & 1 \\ -1 & 0 & 6 & 0 \\ 1 & 1 & 0 & 7 \end{pmatrix}$$

$$R. \quad s = \sqrt{\sum_{j=2}^4 a_{1,j}^2} = 1.73205, \quad \beta = \frac{1}{s(|a_{1,2}| + s)} = 0.21132,$$

$$u = \begin{pmatrix} a_{2,1} + s \\ a_{3,2} \\ a_{4,2} \end{pmatrix} = \begin{pmatrix} 1 + 1.73205 \\ -1 \\ 1 \end{pmatrix},$$

$$\tilde{H}_1 = I_3 - \beta \cdot u \cdot u^T = \begin{pmatrix} -0.57735 & 0.57735 & -0.57735 \\ 0.57735 & 0.78867 & 0.21132 \\ -0.57735 & 0.21132 & 0.78867 \end{pmatrix},$$

$$H_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -0.57735 & 0.57735 & -0.57735 \\ 0 & 0.57735 & 0.78867 & 0.21132 \\ 0 & -0.57735 & 0.21132 & 0.78867 \end{pmatrix},$$

$$A_1 = H_1 \cdot A \cdot H_1^T = \begin{pmatrix} 4 & -0.57735 & 0 & 0 \\ -0.57735 & -5.33333 & 2.13076 & 4.79743 \\ 0 & 2.13076 & 5.95534 & 0.83333 \\ 0 & 4.79743 & 0.83333 & 5.37799 \end{pmatrix};$$

$$s = \sqrt{\sum_{j=3}^4 a_{2,j}^2} = 0.52493, \quad \beta = \frac{1}{s(|a_{1,2}| + s)} = 0.02581,$$

$$u = \begin{pmatrix} a_{3,2}^1 + s \\ a_{4,2}^1 \end{pmatrix} = \begin{pmatrix} 2.13076 + 0.52493 \\ 4.79743 \end{pmatrix},$$

$$\tilde{H}_2 = I_2 - \beta \cdot u \cdot u^T = \begin{pmatrix} -0.40591 & -0.91391 \\ -0.91391 & 0.40591 \end{pmatrix},$$

$$H_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -0.40591 & -0.91391 \\ 0 & 0 & -0.91391 & 0.40591 \end{pmatrix},$$

$$A_2 = H_2 \cdot A_1 \cdot H_2^T = \begin{pmatrix} 4 & -0.57735 & 0 & 0 \\ -0.57735 & -5.33333 & -5.24933 & 0 \\ 0 & -5.24933 & 6.09139 & 0.77290 \\ 0 & 0 & 0.77290 & 5.24193 \end{pmatrix}.$$

5. 
$$A = \begin{pmatrix} 5 & 1 & 1 & 2 \\ 1 & 7 & -1 & 1 \\ 1 & -1 & 6 & 1 \\ 2 & 1 & 1 & 8 \end{pmatrix}$$

R. 
$$s = \sqrt{\sum_{j=2}^4 a_{1,j}^2} = 2.44948, \quad \beta = \frac{1}{s(|a_{1,2}| + s)} = 0.11835,$$

$$u = \begin{pmatrix} a_{2,1} + s \\ a_{3,2} \\ a_{4,2} \end{pmatrix} = \begin{pmatrix} 1 + 1.73205 \\ 1 \\ 2 \end{pmatrix},$$

$$\tilde{H}_1 = I_3 - \beta \cdot u \cdot u^T = \begin{pmatrix} -0.40824 & -0.40824 & -0.81649 \\ -0.40824 & 0.88164 & -0.23670 \\ -0.81649 & -0.23670 & 0.52659 \end{pmatrix},$$

$$H_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -0.57735 & 0.57735 & -0.57735 \\ 0 & 0.57735 & 0.78867 & 0.21132 \\ 0 & -0.57735 & 0.21132 & 0.78867 \end{pmatrix},$$

$$A_1 = H_1 \cdot A \cdot H_1^T = \begin{pmatrix} 5 & 0.81649 & 0 & 0 \\ -0.57735 & 5.83333 & -1.76598 & 4.63299 \\ 0 & -1.76598 & 6.77447 & 1.20585 \\ 0 & 4.63299 & 1.20585 & 5.72552 \end{pmatrix};$$

$$s = \sqrt{\sum_{j=3}^4 a_{2,j}^2} = 4.95815, \quad \beta = \frac{1}{s(|a_{1,2}| + s)} = 0.02999,$$

$$u = \begin{pmatrix} a_{3,2}^1 + s \\ a_{4,2}^1 \end{pmatrix} = \begin{pmatrix} -1.76598 - 4.95815 \\ 4.63299 \end{pmatrix},$$

$$\tilde{H}_2 = I_2 - \beta \cdot u \cdot u^T = \begin{pmatrix} -0.35617 & 0.93441 \\ 0.93441 & 0.35617 \end{pmatrix},$$

$$H_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -0.35617 & 0.93441 \\ 0 & 0 & 0.93441 & 0.35617 \end{pmatrix},$$

$$A_2 = H_2 \cdot A_1 \cdot H_2^T = \begin{pmatrix} 5 & 0.81649 & 0 & 0 \\ 0.81649 & 5.83333 & 4.95815 & 0 \\ 0 & 4.95815 & 5.05593 & 0.55078 \\ 0 & 0 & 0.55078 & 7.44406 \end{pmatrix}.$$

6. Să se găsească cea mai mare valoare proprie în valoare absolută, pentru matricea

$$A = \begin{pmatrix} 4 & 1 & 0 & 0 \\ 1 & 5 & -1 & 0 \\ 0 & -1 & 6 & 1 \\ 0 & 0 & 1 & 7 \end{pmatrix}$$

folosind polinoamele Sturm.

R.  $r_1 = 2$ ,  $r_2 = 2$ ,  $r_3 = 2$ ,  $r_4 = 1$ , iar rădăcinile polinomului caracteristic se află în intervalul  $[a, b]$ , unde :

$$a = \min_{1 \leq i \leq 4} (a_{ii} - r_i) = \min \{ 4-2, 5-2, 6-2, 7-1 \} = 3 \text{ și}$$

$$b = \max_{1 \leq i \leq 4} (a_{ii} + r_i) = \max \{ 4+2, 5+2, 6+2, 7+1 \} = 8 .$$

Polinoamele Sturm pentru această matrice sunt:

$$f_0(\lambda) = 1, \quad f_1(\lambda) = \lambda - a_{11} = \lambda - 4,$$

$$f_2(\lambda) = (\lambda - a_{22})f_1(\lambda) - a_{12}^2 f_0(\lambda) = (\lambda - 5)(\lambda - 4) - 1,$$

$$f_3(\lambda) = (\lambda - a_{33})f_2(\lambda) - a_{23}^2 f_1(\lambda) = (\lambda - 6)f_2(\lambda) - f_1(\lambda),$$

$$f_4(\lambda) = (\lambda - a_{44})f_3(\lambda) - a_{34}^2 f_2(\lambda) = (\lambda - 7)f_3(\lambda) - f_2(\lambda).$$

Schimbările de semn în șirul Sturm de mai sus :

$$f_0(a) = 1, \quad f_1(a) = -1, \quad f_2(a) = 1, \quad f_3(a) = -2, \quad f_4(a) = 7,$$

arată că la dreapta lui  $a$  se află 4 rădăcini ale ecuației caracteristice,

$$f_0(b) = 1, \quad f_1(b) = 4, \quad f_2(b) = 11, \quad f_3(b) = 18, \quad f_4(b) = 7,$$

iar la dreapta lui  $b$  nu se află nici o rădăcină a ecuației caracteristice.

Luând  $c = \frac{a+b}{2}$ ,  $f_0(c) = 1$ ,  $f_1(c) = 1.5$ ,  $f_2(c) = -0.25$ ,  $f_3(c) = -1.375$ ,

$$f_4(c) = 2.3125,$$

la dreapta lui  $c$  se află 2 rădăcini ale ecuației caracteristice și atunci  $a := c$ ;

$$c = \frac{a+b}{2}, f_0(c) = 1, f_1(c) = 2.75, f_2(c) = 3.8125, f_3(c) = 0.10938,$$

$$f_4(c) = -3.83984,$$

la dreapta lui  $c$  se află o rădăcină a ecuației caracteristice,  $a := c$ ;

$$c = \frac{a+b}{2}, f_0(c) = 1, f_1(c) = 3.375, f_2(c) = 7.01563, f_3(c) = 6.27148,$$

$$f_4(c) = -4.66382,$$

la dreapta lui  $c$  se află o rădăcină a ecuației caracteristice,  $a := c$ ;

$$c = \frac{a+b}{2}, f_0(c) = 1, f_1(c) = 3.6875, f_2(c) = 8.91016, f_3(c) = 11.3439,$$

$$f_4(c) = -1.10814,$$

la dreapta lui  $c$  se află o rădăcină a ecuației caracteristice,  $a := c$ ;

$$c = \frac{a+b}{2}, f_0(c) = 1, f_1(c) = 3.84375, f_2(c) = 9.93066, f_3(c) = 14.46591,$$

$$f_4(c) = 2.27495,$$

la dreapta lui  $c$  nu se află nici o rădăcină a ecuației caracteristice,  $b := c$ ;

$$c = \frac{a+b}{2}, f_0(c) = 1, f_1(c) = 3.76563, f_2(c) = 9.41431, f_3(c) = 12.85651,$$

$$f_4(c) = 0.42896,$$

la dreapta lui  $c$  nu se află nici o rădăcină a ecuației caracteristice, ș. a. m. d.  
La iterația a 16-a se obține aproximația  $c = 7.74529$ , iar  $P(c) = 0.00023$ ,  $P(\lambda)$  fiind polinomul caracteristic.

7. Să se găsească cea de-a doua valoare proprie pentru matricea

$$(|\lambda_1| > |\lambda_2| > |\lambda_3| > |\lambda_4|)$$

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

folosind polinoamele Sturm.

R.  $r_1 = 2$ ,  $r_2 = 3$ ,  $r_3 = 3$ ,  $r_4 = 2$ , iar rădăcinile polinomului caracteristic se află în intervalul  $[a, b]$ , unde :

$$a = \min_{1 \leq i \leq 4} (a_{ii} - r_i) = \min \{ 5-2, 6-3, 6-3, 8-2 \} = 3 \text{ și}$$

$$b = \max_{1 \leq i \leq 4} (a_{ii} + r_i) = \max \{ 5+2, 6+3, 6+3, 8+2 \} = 10 .$$

Polinoamele Sturm pentru această matrice sunt:

$$f_0(\lambda) = 1, \quad f_1(\lambda) = \lambda - a_{11} = \lambda - 5,$$

$$f_2(\lambda) = (\lambda - a_{22})f_1(\lambda) - a_{12}^2 f_0(\lambda) = (\lambda - 6)(\lambda - 5) - 4,$$

$$f_3(\lambda) = (\lambda - a_{33})f_2(\lambda) - a_{23}^2 f_1(\lambda) = (\lambda - 6)f_2(\lambda) - f_1(\lambda),$$

$$f_4(\lambda) = (\lambda - a_{44})f_3(\lambda) - a_{34}^2 f_2(\lambda) = (\lambda - 8)f_3(\lambda) - 4f_2(\lambda).$$

Schimbările de semn în șirul Sturm de mai sus :

$$f_0(a) = 1, \quad f_1(a) = -2, \quad f_2(a) = 2, \quad f_3(a) = -4, \quad f_4(a) = 12,$$

arată că la dreapta lui  $a$  se află 4 rădăcini ale ecuației caracteristice,

$$f_0(b) = 1, \quad f_1(b) = 5, \quad f_2(b) = 16, \quad f_3(b) = 59, \quad f_4(b) = 54,$$

iar la dreapta lui  $b$  nu se află nici o rădăcină a ecuației caracteristice.  
Luând

$$c = \frac{a+b}{2}, \quad f_0(c) = 1, \quad f_1(c) = 1.5, \quad f_2(c) = -3.25, \quad f_3(c) = 17.6875,$$

$$f_4(c) = 2.3125,$$

la dreapta lui  $c$  se află 2 rădăcini ale ecuației caracteristice și atunci  $a := c$ ;

$$c = \frac{a+b}{2}, f_0(c) = 1, f_1(c) = 3.25, f_2(c) = 3.3125, f_3(c) = 4.20312,$$

$$f_4(c) = -12.19921,$$

la dreapta lui  $c$  se află o rădăcină a ecuației caracteristice,  $b := c$ ;

$$c = \frac{a+b}{2}, f_0(c) = 1, f_1(c) = 2.375, f_2(c) = -0.73437, f_3(c) = -3.38476,$$

$$f_4(c) = 5.05297,$$

la dreapta lui  $c$  se află o rădăcină a ecuației caracteristice,  $a := c$ ;

$$c = \frac{a+b}{2}, f_0(c) = 1, f_1(c) = 2.8125, f_2(c) = 1.09765, f_3(c) = -0.82299,$$

$$f_4(c) = -4.23631,$$

la dreapta lui  $c$  se află o rădăcină a ecuației caracteristice,  $b := c$ ;

$$c = \frac{a+b}{2}, f_0(c) = 1, f_1(c) = 2.59375, f_2(c) = 0.13378, f_3(c) = -2.38052,$$

$$f_4(c) = 0.43193,$$

la dreapta lui  $c$  se află două rădăcini ale ecuației caracteristice,  $a := c$ ;

$$c = \frac{a+b}{2}, f_0(c) = 1, f_1(c) = 2.70312, f_2(c) = 0.60375, f_3(c) = -1.67484,$$

$$f_4(c) = -1.91781,$$

la dreapta lui  $c$  se află o rădăcină a ecuației caracteristice, ș. a. m. d.

La iterația a 20-a se obține aproximația  $c = 7.61384$ , iar  $P(c) = 0.00001$ ,  $P(\lambda)$  fiind polinomul caracteristic.

## 4. Interpolarea funcțiilor

Fie  $f : [a, b] \rightarrow \mathbb{R}$  și fie  $x_0, x_1, \dots, x_n$ ,  $(n+1)$  puncte distincte din intervalul  $[a, b]$ , numite *noduri*.

Problema interpolării funcției  $f$  în nodurile  $x_i$ ,  $i = \overline{0, n}$ , constă în determinarea unei funcții  $g : [a, b] \rightarrow \mathbb{R}$ , dintr-o clasă de funcții cunoscută, cu proprietatea  $g(x_i) = f(x_i)$ ,  $i = \overline{0, n}$ .

Pusă sub această formă generală problema poate să nu aibă soluție sau să aibă o infinitate de soluții.

Cea mai utilizată clasă de funcții de interpolare este clasa polinoamelor, datorită ușurinței cu care se integrează și se derivează.

Interpolarea funcțiilor prezintă o importanță deosebită pentru cazul când funcția nu este definită printr-o relație analitică, ci printr-un tablou de valori, ce reprezintă, de exemplu, rezultatele unei experiențe. Chiar și atunci când funcția este dată printr-o relație analitică, dar această relație este *complicată* se poate alege interpolarea în locul calculului direct.

### §4.1. Polinomul de interpolare al lui Lagrange

**Teorema 1.** Fie  $f : [a, b] \rightarrow \mathbb{R}$  și  $x_0, x_1, \dots, x_n$ ,  $(n+1)$  noduri din intervalul  $[a, b]$ . Atunci există un polinom unic  $P_n$ , de gradul  $n$ , care interpoalează funcția  $f$  în nodurile  $x_i$ ,  $i = \overline{0, n}$  ( $f(x_i) = P_n(x_i)$ ,  $i = \overline{0, n}$ ). Acest polinom se numește polinomul de interpolare al lui Lagrange.

**Demonstrație.**

Căutăm un polinom  $P_n$  sub forma următoare:

$$P_n(x) = L_0(x) \cdot f(x_0) + L_1(x) \cdot f(x_1) + \dots + L_n(x) \cdot f(x_n),$$

unde  $L_i$  sunt polinoame de gradul  $n$  ce urmează să fie determinate. Deoarece dorim ca  $P_n(x_i) = f(x_i)$ , vom pune condițiile:

$$L_i(x_j) = \delta_{ij} = \begin{cases} 1 & \text{dac\u0103 } j=i \\ 0 & \text{dac\u0103 } j \neq i \end{cases}.$$

Deoarece  $L_i(x_j)=0$  pentru  $i \neq j$ , rezult\u0103 c\u0103  $L_i$  admite r\u0103d\u0103cinile  $x_0, x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n$ .

A\u015fadar,

$$L_i(x) = a_i(x-x_0)\dots(x-x_{i-1})(x-x_{i+1})\dots(x-x_n).$$

Cum  $L_i(x_i)=1$ , rezult\u0103

$$a_i = \frac{1}{(x_i - x_0)\dots(x_i - x_{i-1})(x_i - x_{i+1})\dots(x_i - x_n)}.$$

\u00c0n concluzie avem

$$P_n(x) = \sum_{i=0}^n L_i(x) f(x_i) \quad (1)$$

unde

$$L_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^n \frac{(x - x_j)}{(x_i - x_j)}. \quad (2)$$

Evident polinomul (1) are gradul  $n$  \u015fi are proprietatea  $P_n(x_i)=f(x_i)$ ,  $i = \overline{0, n}$ .

Fie  $Q_n$  un alt polinom de gradul  $n$  cu proprietatea  $Q_n(x_i)=f(x_i)$ ,  $i = \overline{0, n}$  \u015fi fie  $R=P_n-Q_n$ . Deoarece  $\text{grad} R \leq n$  \u015fi  $R(x_i)=0$ ,  $i = \overline{0, n}$  rezult\u0103 c\u0103  $R$  este polinom identic nul, deci c\u0103  $P_n=Q_n$ .  $\square$

**Exemplu.** Fie nodurile  $x_0=-1, x_1=1$ , \u015fi  $x_2=2$  \u015fi  $f(x_0)=2, f(x_1)=f(x_2)=1$ . Atunci

$$P_2(x) = \frac{(x-1)(x-2)}{(-1-1)(-1-2)} 2 + \frac{(x+1)(x-2)}{(1+1)(1-2)} 1 + \frac{(x+1)(x-1)}{(2+1)(-2-1)} 1.$$

Efectu\u0103nd calculele ob\u015ftinem  $P_2(x) = \frac{1}{6}(x^2 - 3x + 8)$ .

\u00c0n continuare vom nota eroarea \u00een fiecare punct cu

$$E(f; x) = f(x) - P_n(x) \quad (3)$$

Evident  $E(f; x_i)=0$ ,  $i = \overline{0, n}$ . Introducem de asemenea nota\u015fia:

$$U_{n+1}(x) = \prod_{i=0}^n (x - x_i) \quad (4)$$

**Teorema 2.** *Dac\u0103  $f \in C^{(n+1)}[a, b]$ , atunci pentru orice  $x \in [a, b]$ , exist\u0103  $\xi_x \in (a, b)$  astfel \u00eenc\u0103t*

$$E(f; x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!} U_{n+1}(x) \quad (5)$$

**Demonstrație.**

Considerăm funcția auxiliară

$$g(t) = f(t) - P_n(t) - \frac{E(f;x)}{U_{n+1}(x)} U_{n+1}(t), \quad t \in [a, b], \quad x \neq x_i$$

Observăm că  $g$  se anulează în  $(n+2)$  puncte distincte  $x_0, x_1, \dots, x_n, x$ . Din teorema lui Rolle rezultă că există  $\xi_x \in (a, b)$  astfel încât  $g^{(n+1)}(\xi_x) = 0$ .

Cum

$$g^{(n+1)}(t) = f^{(n+1)}(t) - \frac{E(f;x)}{U_{n+1}(x)} (n+1)!$$

rezultă

$$E(f;x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!} U_{n+1}(x). \quad \square$$

**Corolar.** Dacă există  $M > 0$  astfel încât  $|f^{(n+1)}(x)| \leq M$  pentru orice  $x \in [a, b]$ , atunci:

$$|E(f;x)| \leq \frac{M}{(n+1)!} |U_{n+1}(x)|, \quad x \in [a, b].$$

**Exemplu.** Fie funcția  $f(x) = \ln x$  și nodurile 0.4; 0.5; 0.7; 0.8. Evaluăm eroarea în punctul  $x = 0.6$ .

$$U_4(0.6) = (0.2)(0.1)(-0.1)(-0.2) = 0.0004.$$

$$|f^{IV}(x)| = \left| -\frac{6}{x^4} \right| \leq \frac{6}{(0.4)^4} \cong 234.4, \quad x \in [0.4; 0.8].$$

Rezultă

$$|E(f; 0.6)| \leq \frac{1}{24} 234.4 \cdot 0.0004 \cong 0.0039,$$

acest număr fiind doar un majorant al erorii.

Dacă folosim următoarele valori în noduri

$X$	0.4	0.5	0.7	0.8
$f(x)$	-0.916291	-0.693147	-0.356675	-0.223144

și calculăm polinomul lui Lagrange obținem:  $P_3(0.6) = -0.509975$ .

Pe de altă parte  $\ln(0.6) = -0.510826$ . Rezultă că  $E(f; 0.6) = -0.000851$ , ceea ce confirmă afirmația de mai sus.

**Observația 1.** Dacă  $f = Q$  este un polinom de grad cel mult  $n$ , atunci  $E(f;x) = 0$ , oricare  $x \in [a, b]$ .

Afirmația rezultă din Teorema 2 deoarece, în acest caz  $f^{(n+1)}(x) = 0$ .

**Observația 2.**  $E(f+g;x)=E(f;x)+E(g;x)$

Într-adevăr, dacă  $P_n^f$  este polinomul de interpolare pentru  $f$  și  $P_n^g$  este polinomul de interpolare pentru  $g$ , atunci  $P_n^f + P_n^g$  este polinomul de interpolare pentru  $f+g$  și deci

$$E(f+g;x) = f(x) + g(x) - P_n^f(x) - P_n^g(x) = E(f;x) + E(g;x).$$

În continuare vom presupune că nodurile sunt echidistante, deci că  $x_i = x_0 + i \cdot h$ ,  $i = \overline{0, n}$ , unde

$$h = \frac{x_n - x_0}{n}. \quad (6)$$

Considerăm de asemenea schimbarea de variabilă

$$x = x_0 + th \quad (7)$$

Înlocuind (6) și (7) în (2) obținem:

$$\tilde{L}_i(t) = L_i(x_0 + th) = \prod_{\substack{j=0 \\ j \neq i}}^n \frac{(t-j)}{(i-j)}$$

Folosind notația:

$$\pi_{n+1}(t) = t(t-1)(t-2)\dots(t-n) = \prod_{j=0}^n (t-j) \quad (8)$$

obținem:

$$\tilde{L}_i(t) = \frac{(-1)^{n-i} \pi_{n+1}(t)}{i!(n-i)! (t-i)} = \frac{(-1)^{n-i}}{n!} C_n^i \frac{\pi_{n+1}(t)}{t-i}$$

Obținem astfel expresia polinomului lui Lagrange pentru noduri echidistante

$$\tilde{P}_n(t) = \frac{\pi_{n+1}(t)}{n!} \sum_{i=0}^n (-1)^{n-i} C_n^i \frac{f(x_i)}{t-i}. \quad (9)$$

Eroarea devine:

$$\tilde{E}(t) = \frac{\pi_{n+1}(t)}{(n+1)!} h^{n+1} f^{(n+1)}(\xi_t). \quad (10)$$

În continuare considerăm un șir de diviziuni  $\{\Delta_n\}$  ale intervalului  $[a, b]$  cu

$$\lim_{n \rightarrow \infty} \|\Delta_n\| = 0, \quad \Delta_n : a = x_0^{(n)} < x_1^{(n)} < \dots < x_n^{(n)} = b.$$

Notăm cu  $P_n$  - polinomul lui Lagrange care interpoalează funcția  $f$  în nodurile  $x_i^{(n)}$ ,  $i = \overline{0, n}$ . Dacă  $n$  este mare,  $P_n$  coincide cu  $f$  într-un număr mare de noduri, deci ne așteptăm ca eroarea

$$E_n(f;x) = f(x) - P_n(x)$$

să fie mică, eventual ca  $\lim_{n \rightarrow \infty} E_n(f;x) = 0$ .

Ajungem astfel la următoarea întrebare:

În ce condiții șirul de polinoame  $\{P_n\}$  converge punctual (eventual uniform) la funcția  $f$  pe intervalul  $[a, b]$ ?

În anul 1912, S. N. Bernstein a arătat că pentru funcția  $f(x) = |x|$ ,  $x \in [-1, 1]$ , dacă alegem nodurile echidistante  $x_i^{(n)} = -1 + \frac{2i}{n}$ ,  $i = \overline{0, n}$ , atunci

$$\lim_{n \rightarrow \infty} P_n(x) \neq f(x) \text{ dacă } x \notin \{-1, 0, 1\}.$$

S-ar putea crede că acest lucru se datorează faptului că funcția modul nu este derivabilă în origine. Următorul exemplu dat de C. Runge în 1901 arată că există funcții indefinit derivabile pentru care  $\{P_n\}$  nu converge la  $f$ .

$$\text{Fie } f(x) = \frac{1}{1+x^2}, \quad x \in [-5, 5].$$

Evident  $f \in C^\infty[-5, 5]$ . Fie nodurile echidistante

$$x_i = -5 + \frac{10}{n}i, \quad i = \overline{0, n}.$$

Se poate arăta că  $\lim_{n \rightarrow \infty} P_n(x) = f(x)$  dacă  $|x| \leq c$  și  $\lim_{n \rightarrow \infty} P_n(x) \neq f(x)$  dacă

$|x| > c$ , unde  $c \cong 3.6334$  este o rădăcină a ecuației:

$$(5+x)\ln(5+x) + (5-x)\ln(5-x) - 5\ln 26 - 2\arctg 5 = 0.$$

În anul 1914, S. N. Bernstein a arătat că pentru orice sistem de noduri  $\{x_i^{(n)}\}$ ,  $i = \overline{0, n}$  din intervalul  $[a, b]$ , dat dinainte, există o funcție continuă

$f: [a, b] \rightarrow \mathbb{R}$  astfel încât șirul polinoamelor lui Lagrange  $\{P_n\}$  care interpolează funcția  $f$  în aceste noduri nu converge uniform la  $f$  pe  $[a, b]$ .

Există totuși și situații când convergența are loc. Se poate demonstra următoarea teoremă:

**Teorema 3.** Dacă  $f \in C^\infty(\mathbb{R})$  și se dezvoltă în serie Taylor pe  $\mathbb{R}$ , atunci pentru orice sistem de noduri distincte și echidistante  $\{x_i^{(n)}\}$ ,  $i = \overline{0, n}$  din  $[a, b]$ , șirul polinoamelor  $\{P_n\}$  care interpolează funcția  $f$  în aceste noduri converge uniform la  $f$  pe  $[a, b]$ .

Se pune întrebarea dacă interpolarea cu polinoame Lagrange este utilă în practică, din moment ce așa cum am văzut, în general șirul polinoamelor de interpolare  $\{P_n\}$  nu converge la  $f$ .

Răspunsul este că interpolarea Lagrange este utilă. Se constată în practică faptul că pentru un punct  $\alpha \in [a, b]$ , eroarea  $|f(\alpha) - P_n(\alpha)|$  scade până la un punct, pe măsură ce  $n$  crește, și deci, pentru  $n$  relativ mic,  $P_n(\alpha)$  aproximează acceptabil valoarea  $f(\alpha)$ . Pentru valori mari ale lui  $n$ , interpolarea Lagrange nu este recomandată.

Din cele prezentate până acum, rezultă că șirul polinoamelor de interpolare asociate unei funcții continue nu converge uniform, în mod necesar, la această

funcție. Se pune întrebarea dacă o funcție continuă poate fi aproximată uniform cu polinoame. Răspunsul a fost dat de K. Weierstrass în anul 1885.

**Teorema 4.** Fie  $f: [a, b] \rightarrow \mathbb{R}$  continuă. Atunci, pentru orice  $\varepsilon > 0$ , există un polinom  $Q_\varepsilon$  astfel încât

$$\|f - Q_\varepsilon\| = \sup\{|f(x) - Q_\varepsilon(x)|; x \in [a, b]\} < \varepsilon.$$

Evident, dacă luăm  $\varepsilon = \frac{1}{n}$ , rezultă că există un șir de polinoame  $\{Q_n\}$  care converge uniform pe  $[a, b]$  la funcția  $f$ . Din teorema lui Weierstrass rezultă că polinoamele algebrice pe  $[a, b]$  sunt, în raport cu funcțiile continue pe  $[a, b]$ , în aceeași relație ca numerele raționale  $Q$  față de numerele reale  $\mathbb{R}$ .

Teorema lui Weierstrass este extrem de importantă în analiza matematică, în general, și în analiza numerică, în special. Dintre numeroasele demonstrații date acestei teoreme, cea mai cunoscută este demonstrația dată de S. N. Bernstein, în anul 1912. Bernstein a arătat cum se poate construi șirul de polinoame care aproximează funcția  $f$  și anume:

$$B_n(x) = \sum_{k=0}^n C_n^k (1-x)^{n-k} x^k f\left(\frac{k}{n}\right), \quad x \in [0, 1].$$

Acest șir de polinoame, care se numesc *polinoame Bernstein*, au proprietatea  $B_n \xrightarrow{u} f$  pe  $[0, 1]$ . Trecerea de la  $[0, 1]$  la  $[a, b]$  se face cu ușurință printr-o schimbare de variabilă. Evident, polinoamele Bernstein nu sunt polinoame de interpolare. Din păcate, convergența șirului  $\{B_n\}$  către  $f$  este destul de încetă, și din această cauză, în practică, polinoamele Bernstein nu se folosesc la aproximarea directă a funcțiilor. Teorema lui Weierstrass este importantă prin implicațiile sale teoretice, dar și practice, așa cum vom vedea, de exemplu, la integrarea numerică.

## §4.2. Interpolarea iterativă. Metoda Aitken

În acest paragraf vom nota polinomul lui Lagrange care interpoalează funcția  $f$  în nodurile  $x_i, i = \overline{0, n}$  cu  $P_n(x; x_0, x_1, \dots, x_n)$ . Evident,

$$P_0(x; x_0) = f(x_0).$$

**Teorema 1.** Are loc următoare relație de recurență:

$$P_n(x; x_0, x_1, \dots, x_n) = \frac{1}{x_n - x_{n-1}} \begin{vmatrix} P_{n-1}(x; x_0, x_1, \dots, x_{n-2}, x_{n-1}) & x_{n-1} - x \\ P_{n-1}(x; x_0, x_1, \dots, x_{n-2}, x_n) & x_n - x \end{vmatrix}.$$

**Demonstrație.**

Fie

$$Q(x) = \frac{1}{x_n - x_{n-1}} \begin{vmatrix} P_{n-1}(x; x_0, x_1, \dots, x_{n-2}, x_{n-1}) & x_{n-1} - x \\ P_{n-1}(x; x_0, x_1, \dots, x_{n-2}, x_n) & x_n - x \end{vmatrix}.$$

Observăm că pentru orice  $i = 0, n-2$  avem

$$Q(x_i) = \frac{1}{x_n - x_{n-1}} \begin{vmatrix} f(x_i) & x_{n-1} - x \\ f(x_i) & x_n - x \end{vmatrix} = f(x_i).$$

În continuare, avem:

$$Q(x_{n-1}) = \frac{1}{x_n - x_{n-1}} \begin{vmatrix} f(x_{n-1}) & 0 \\ P_{n-1}(x_{n-1}; x_0, \dots, x_{n-2}, x_n) & x_n - x_{n-1} \end{vmatrix} = f(x_{n-1}),$$

$$Q(x_n) = \frac{1}{x_n - x_{n-1}} \begin{vmatrix} P_{n-1}(x_n; x_0, \dots, x_{n-2}, x_{n-1}) & x_{n-1} - x_n \\ f(x_n) & 0 \end{vmatrix} = f(x_n).$$

Așadar,  $Q$  este un polinom de gradul  $n$  care interpolează funcția  $f$  în nodurile  $x_i, i = 0, n$ . Din unicitatea polinomului de interpolare al lui Lagrange, rezultă că  $Q = P_n$ .

Metoda Aitken este bine ilustrată de următorul tabel:

$x_0$	$x_0 - \alpha$	$f(x_0)$				
$x_1$	$x_1 - \alpha$	$f(x_1)$	$P_1(\alpha; x_0, x_1)$			
$x_2$	$x_2 - \alpha$	$f(x_2)$	$P_1(\alpha; x_0, x_2)$	$P_2(\alpha; x_0, x_1, x_2)$		
$x_3$	$x_3 - \alpha$	$f(x_3)$	$P_1(\alpha; x_0, x_3)$	$P_2(\alpha; x_0, x_1, x_3)$	$P_3(\alpha; x_0, x_1, x_2, x_3)$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$
$x_n$	$x_n - \alpha$	$f(x_n)$	$P_1(\alpha; x_0, x_n)$	$P_2(\alpha; x_0, x_1, x_n)$	$P_3(\alpha; x_0, x_1, x_2, x_n)$	$\dots P_n(\alpha; x_0, x_1, \dots, x_n)$

Algoritmul de interpolare iterativă (metoda Aitken)

Pentru  $i := 1, n$  execută

$$y_i := f(x_i), \quad d_i := x_i - \alpha;$$

sfârșit pentru  $i$ ;

Pentru  $i := 2, n$  execută

Pentru  $j := i, n$  execută

$$y_j := \frac{y_{i-1}d_j - y_jd_{i-1}}{x_j - x_{i-1}}$$

sfârșit pentru  $i$

sfârșit pentru  $j$ .

### §4.3. Polinoame Cebîșev

Polinoamele Cebîșev sunt definite pe intervalul  $[-1, 1]$  prin relația:

$$T_n(x) = \cos(n \cdot \arccos(x)) \quad (1)$$

Deoarece

$T_{n+1}(x) + T_{n-1}(x) = \cos[(n+1)\arccos x] + \cos[(n-1)\arccos x] = 2x \cos(n \cdot \arccos x)$ ,  
rezultă următoarea relație de recurență:

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x), \quad n \geq 1 \quad (2)$$

Cum

$$T_0(x) = 1 \quad \text{și} \quad T_1(x) = x,$$

din (2) rezultă

$$T_2(x) = 2x^2 - 1, \quad T_3(x) = 4x^3 - 3x, \quad T_4(x) = 8x^4 - 8x^2 + 1, \quad T_5(x) = 16x^5 - 20x^3 + 5x \text{ etc.}$$

$$\text{Observăm că} \quad T_n(x) = 2^{n-1}x^n + \dots$$

Dacă  $T_n(x) = 0$ , atunci

$$n \cdot \arccos(x) = (2k+1) \frac{\pi}{2},$$

de unde rezultă

$$x_k = \cos(2k+1) \frac{\pi}{2n}, \quad k = \overline{0, n-1} \quad (3)$$

Așadar, polinomul  $T_n$  are  $n$  rădăcini reale distincte, date de formula (3).

Pe de altă parte, avem

$$T'_n(x) = n \cdot \frac{\sin(n \cdot \arccos(x))}{\sqrt{1-x^2}}.$$

Dacă  $T'_n(x) = 0$ , atunci  $n \cdot \arccos(x) = k\pi$ , și deci

$$y_k = \cos\left(\frac{k\pi}{n}\right), \quad k = \overline{1, n-1} \quad (4)$$

sunt zerourile derivatei  $T'_n$ . Se observă că rădăcinile derivatei  $T'_n$  separă rădăcinile polinomului  $T_n$ . Într-adevăr,

$$(2k+1) \cdot \frac{\pi}{2n} < (k+1) \cdot \frac{\pi}{n} < (2k+3) \cdot \frac{\pi}{2n},$$

de unde rezultă

$$x_k = \cos(2k+1) \frac{\pi}{2n} > y_{k+1} = \cos(k+1) \frac{\pi}{n} > x_{k+1} = \cos(2k+3) \frac{\pi}{2n}.$$

Constatăm de asemenea că

$$T_n(y_k) = \cos\left[n \cdot \arccos\left(\cos\left(\frac{k\pi}{n}\right)\right)\right] = \cos(k\pi) = (-1)^k.$$

Cum  $|T_n(x)| \leq 1, x \in [-1, 1]$ , rezultă că  $y_k, k = \overline{1, n-1}$ , sunt puncte de extrem local pentru  $T_n$ . Pe de altă parte, avem  $T_n(-1) = (-1)^n$  și  $T_n(1) = 1$ .

Așadar,  $T_n$  are  $(n+1)$  puncte de extrem local și își schimbă semnul de  $n$  ori pe intervalul  $[-1, 1]$ .

Prezentăm în continuare tabelul de variație pentru polinoamele  $T_3$  și  $T_4$ .

$x$	-1		$-\frac{\sqrt{3}}{2}$		$-\frac{1}{2}$		0		$\frac{1}{2}$		$\frac{\sqrt{3}}{2}$		1
$T_3'$			+		0		-		0		+		
$T_3$	-1	↗	0	↗	1	↘	0	↘	-1	↗	0	↗	1

$x$	-1		$-\frac{1}{\sqrt{2}}$		0		$\frac{1}{\sqrt{2}}$		1
$T_4'$		-	0	+	0	-	0	+	
$T_4$	1	↘	-1	↗	1	↘	-1	↗	1

Următorul rezultat datorat lui Cebâșev pune în evidență o proprietate remarcabilă a zerourilor polinoamelor Cebîșev.

**Teorema 1.** Fie  $x_k = (2k+1)\frac{\pi}{2n}, k = \overline{0, n}$ , zerourile polinomului Cebîșev  $T_{n+1}$ .

Atunci, oricare ar fi  $(n+1)$  puncte distincte,  $z_i, i = \overline{0, n}$  din intervalul  $[-1, 1]$ , avem

$$\sup_{x \in [-1, 1]} |(x-x_0)(x-x_1)...(x-x_n)| \leq \sup_{x \in [-1, 1]} |(x-z_0)(x-z_1)...(x-z_n)| .$$

**Demonstrație.** Deoarece

$$T_{n+1}(x) = 2^n(x-x_0)...(x-x_n) ,$$

rezultă că trebuie să arătăm că

$$\sup_{x \in [-1, 1]} \frac{1}{2^n} |T_{n+1}(x)| \leq \sup_{x \in [-1, 1]} |(x-z_0)(x-z_1)...(x-z_n)|, (\forall) z_i \in [-1, 1] .$$

Presupunem prin absurd că există  $\bar{z}_0, \bar{z}_1, \dots, \bar{z}_n \in [-1, 1]$  astfel încât

$$\sup_{x \in [-1, 1]} |q_{n+1}(x)| < \sup_{x \in [-1, 1]} \frac{1}{2^n} |T_{n+1}(x)| = \frac{1}{2^n} \tag{5}$$

unde

$$q_{n+1}(x) = (x-\bar{z}_0)(x-\bar{z}_1)...(x-\bar{z}_n) . \tag{6}$$

Fie

$$r_n(x) = \frac{1}{2^n} T_{n+1}(x) - q_{n+1}(x), x \in [-1, 1] .$$

Evident,  $r_n$  este un polinom de grad cel mult  $n$ . Observăm ca  $r_n$  are același semn cu  $T_{n+1}$  în cele  $(n+2)$  puncte de extrem ale polinomului  $T_{n+1}$ . Într-adevăr, fie  $y_k$  un asemenea punct. Presupunem că  $T_{n+1}(y_k)=1$ . Dacă  $r_n(y_k) \leq 0$ , atunci

$$q_{n+1}(y_k) = \frac{1}{2^n} - r_n(y_k) \geq \frac{1}{2^n},$$

ceea ce contrazice relația (5). Dacă  $T_{n+1}(y_k)=-1$  și presupunem că  $r_n(y_k) > 0$ , atunci

$$-q_{n+1}(y_k) = r_n(y_k) + \frac{1}{2^n} > \frac{1}{2^n},$$

ceea ce contrazice relația (5). Așadar,  $r_n$  își schimbă semnul de  $(n+2)$  ori, deci  $r_n$  are  $(n+1)$  rădăcini. Acest lucru nu este posibil decât dacă  $r_n(x)=0$ ,  $(\forall)x \in [-1,1]$ .

Rezultă atunci că  $\frac{1}{2^n} T_{n+1} = q_{n+1}$ , ceea ce contrazice relația (5).  $\square$

Revenim acum la evaluarea erorii în interpolarea Lagrange.

Fie  $(n+1)$  noduri  $x_i$  în  $[-1,1]$  și  $f \in C^{(n+1)}[-1,1]$ . Dacă  $P_n$  este polinomul lui Lagrange care interpolează funcția  $f$  în nodurile  $x_i$ ,  $i = \overline{0, n}$ , atunci

$$|f(x) - P_n(x)| = \frac{f^{(n+1)}(\xi_x)}{(n+1)!} \cdot (x-x_0) \dots (x-x_n) \quad (7)$$

(vezi Capitolul 4, §1, Teorema 2).

Din (7) rezultă că

$$\|f - P_n\|_\infty = \sup_{-1 \leq x \leq 1} |f(x) - P_n(x)| \leq \frac{\|f^{(n+1)}\|_\infty}{(n+1)!} \sup_{-1 \leq x \leq 1} |(x-x_0) \dots (x-x_n)|.$$

Așadar, eroarea  $\|f - P_n\|_\infty$  va fi minimă dacă

$$\sup_{-1 \leq x \leq 1} |(x-x_0) \dots (x-x_n)|$$

va fi minimă. Pe de altă parte, din Teorema 1 rezultă că acest lucru se întâmplă dacă alegem nodurile

$$x_i = \cos(2i+1) \frac{\pi}{2(n+1)}, \quad i = \overline{0, n}$$

(adică  $x_i$  sunt zerourile polinomului Cebîșev  $T_{n+1}$ ). Din cele de mai sus rezultă că are loc următoarea teoremă:

**Teorema 2.** Fie  $P_n^*$  polinomul lui Lagrange care interpolează funcția  $f$  în nodurile

$$x_i = \cos(2i+1) \frac{\pi}{2(n+1)}, \quad i = \overline{0, n}.$$

Atunci  $\|f - P_n^*\|_\infty \leq \frac{1}{2^n (n+1)!} \|f^{(n+1)}\|_\infty$ . Pentru acele funcții care au proprietatea că  $\lim_{n \rightarrow \infty} \frac{1}{2^n (n+1)!} \|f^{(n+1)}\|_\infty = 0$  va rezulta că șirul  $P_n^* \xrightarrow{u} f$ .

#### §4.4. Funcții spline cubice

Fie

$$\Delta: a=x_0 < x_1 < \dots < x_{i-1} < x_i < \dots < x_n = b$$

o diviziune oarecare a intervalului  $[a, b]$ .

Se numește *funcție spline cubică* o funcție

$$s : [a, b] \rightarrow \mathbb{R}$$

cu următoarele proprietăți:

(i) Restricția lui  $s$  la fiecare subinterval  $[x_{i-1}, x_i]$  este un polinom de grad cel mult trei;

(ii)  $s, s', s''$  sunt continue pe  $[a, b]$ .

În continuare ne punem problema interpolării unei funcții  $f : [a, b] \rightarrow \mathbb{R}$  cu ajutorul unei funcții spline cubice. Cu alte cuvinte, ne punem problema să găsim o funcție spline cubică  $s$ , astfel încât

$$s(x_i) = f(x_i), \quad i = \overline{0, n}.$$

Deoarece restricția lui  $s$  la subintervalele  $[x_{i-1}, x_i]$  este un polinom de grad cel mult trei, rezultă că

$$s(x) = a_i + b_i x + c_i x^2 + d_i x^3$$

pentru orice  $x \in [x_{i-1}, x_i]$ . Determinarea funcției  $s$  presupune deci determinarea a  $4n$  coeficienți  $(a_i, b_i, c_i, d_i)$ .

Să evaluăm acum de câte condiții dispunem. Faptul că

$$s(x_i) = f(x_i), \quad i = \overline{0, n}$$

ne asigură  $(n+1)$  condiții. Pe de altă parte, din continuitatea lui  $s$  și a derivatelor  $s'$  și  $s''$ , rezultă:

$$s^{(k)}(x_{i-0}) = s^{(k)}(x_{i+0}), \quad i = \overline{1, n-1}, \quad k = \overline{0, 2},$$

care ne asigură  $3(n-1)$  condiții. În total, dispunem deci de  $(4n-2)$  condiții, cu două mai puțin decât numărul coeficienților ce urmează a fi determinați.

Dacă se cunosc derivatele  $f'(a)$  și  $f'(b)$ , atunci adăugăm condițiile

$$s'(a) = f'(a) \quad \text{și} \quad s'(b) = f'(b).$$

Dacă nu se cunosc aceste derivate, atunci se aproximează

$$f'(a) \cong y'_0 \quad \text{și} \quad f'(b) = y'_0$$

și se pun condițiile  $s'(a) = y'_0$  și  $s'(b) = y'_0$ . Dacă nu avem nici o informație despre  $f'(a)$  și  $f'(b)$  se pot pune condițiile:

$$s''(a) = s''(b) = 0.$$

În acest caz se obține așa numita *funcție spline cubică naturală*.

Înainte de a prezenta teorema fundamentală privind existența funcțiilor spline cubice, reamintim următorul rezultat de algebră liniară.

**Propoziția 1.** Orice matrice pătratică strict diagonal dominantă este nesingulară.

**Demonstrație.** Fie  $A \in M_n(\mathbb{R})$  cu proprietatea:

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^n |a_{ij}|. \quad (1)$$

Dacă vom arăta că sistemul  $Ax=0$  admite numai soluția banală, va rezulta că  $\det A \neq 0$ .

Presupunem prin absurd că există  $\alpha \neq 0$  astfel încât  $A\alpha = 0$ .

Fie

$$\alpha_j = \|\alpha\|_\infty = \max\{|\alpha_1|, |\alpha_2|, \dots, |\alpha_n|\}.$$

Cum  $\alpha$  este soluție pentru sistemul  $Ax = 0$  rezultă

$$a_{j1}\alpha_1 + \dots + a_{jj}\alpha_j + \dots + a_{jn}\alpha_n = 0 \quad \text{sau}$$

$$a_{jj} + \sum_{\substack{k=1 \\ k \neq j}}^n a_{jk} \frac{\alpha_k}{\alpha_j} = 0. \quad (2)$$

În continuare avem

$$|a_{jj}| \leq \sum_{\substack{k=1 \\ k \neq j}}^n |a_{jk}| \frac{|\alpha_k|}{|\alpha_j|} \leq \sum_{\substack{k=1 \\ k \neq j}}^n |a_{jk}|,$$

ceea ce contrazice (1).  $\square$

**Teorema 1.** Pentru orice  $(n+3)$  numere date  $y'_0, y_0, y_1, \dots, y_n, y'_n$ , există o funcție spline cubică  $s$ , unică cu proprietățile:

$$s(x_i) = y_i, \quad 0 \leq i \leq n, \quad s'(x_0) = y'_0, \quad s'(x_n) = y'_n.$$

**Demonstrație.**

Vom nota cu  $M_i = s''(x_i)$ ,  $i = \overline{0, n}$ . Deoarece  $s''$  este liniară pe intervalul  $[x_i, x_{i+1}]$ , rezultă că  $s''$  este de forma  $s''(x) = \alpha x + \beta$ . Din condițiile  $M_i = s''(x_i)$  și  $M_{i+1} = s''(x_{i+1})$  rezultă

$$a = \frac{M_{i+1} - M_i}{h_i} \quad \text{și} \quad \beta = \frac{M_i x_{i+1} - M_{i+1} x_i}{h_i},$$

unde  $h_i = x_{i+1} - x_i$ . Așadar pe intervalul  $[x_i, x_{i+1}]$ , avem:

$$s''(x) = \frac{(x_{i+1} - x)M_i + (x - x_i)M_{i+1}}{h_i}, \quad i = \overline{0, n-1}. \quad (3)$$

Integrând de două ori obținem

$$s(x) = \frac{(x_{i+1} - x)^3 M_i + (x - x_i)^3 M_{i+1}}{6h_i} + C(x_{i+1} - x) + D(x - x_i), \quad i = \overline{0, n-1} \quad (4)$$

unde  $C$  și  $D$  sunt constante arbitrare.

Punând condițiile de interpolare  $s(x_i) = y_i, 0 \leq i \leq n$ , rezultă

$$C = \frac{y_i}{h_i} - \frac{h_i M_i}{6} \quad \text{și} \quad D = \frac{y_{i+1}}{h_i} - \frac{h_i M_{i+1}}{6} \quad (5)$$

Înlocuind (5) în (4) obținem pentru  $x \in [x_i, x_{i+1}]$  și  $i = \overline{0, n-1}$ :

$$s(x) = \frac{(x_{i+1} - x)^3 M_i + (x - x_i)^3 M_{i+1}}{6h_i} + \frac{(x_{i+1} - x)y_i + (x - x_i)y_{i+1}}{h_i} - \frac{(x_{i+1} - x)M_i + (x - x_i)M_{i+1}}{6} h_i, \quad i = \overline{0, n-1} \quad (6)$$

Să observăm că funcția  $s$  definită în (6) este continuă pe  $[a, b]$ .

Într-adevăr

$$\lim_{\substack{x \rightarrow x_i \\ x < x_i}} s(x) = \lim_{\substack{x \rightarrow x_i \\ x < x_i}} \left[ \frac{(x_i - x)^3 M_{i-1} + (x - x_{i-1})^3 M_i}{6h_{i-1}} + \frac{(x_i - x)y_{i-1} + (x - x_{i-1})y_i}{h_{i-1}} - h_{i-1} \frac{(x_i - x)M_{i-1} + (x - x_{i-1})M_i}{6} \right] = \frac{h_{i-1}^3 M_i}{6h_{i-1}} + y_i - \frac{h_{i-1}^2 M_i}{6} = y_i$$

și analog  $\lim_{\substack{x \rightarrow x_i \\ x > x_i}} s(x) = y_i$ .

În continuare vom pune condiția ca derivata  $s'$  să fie continuă pe  $[a, b]$ .

Din (6) rezultă:

$$s'(x) = \frac{-(x_{i+1} - x)^2 M_i + (x - x_i)^2 M_{i+1}}{2h_i} + \frac{y_{i+1} - y_i}{h_i} - \frac{(M_{i+1} - M_i)h_i}{6}, \quad (7)$$

pentru  $x \in (x_i, x_{i+1}), i = \overline{0, n-1}$ .

Punem condiția ca  $\lim_{\substack{x \rightarrow x_i \\ x < x_i}} s'(x) = \lim_{\substack{x \rightarrow x_i \\ x > x_i}} s'(x)$  și obținem

$$\frac{h_{i-1}^2 M_i}{2h_{i-1}} + \frac{y_i - y_{i-1}}{h_{i-1}} - \frac{(M_i - M_{i-1})h_{i-1}}{6} = \frac{h_i^2 M_i}{2h_i} + \frac{y_{i+1} - y_i}{h_i} - \frac{(M_{i+1} - M_i)h_i}{6}$$

și mai departe

$$\frac{h_{i-1}}{6} M_{i-1} + \frac{h_i + h_{i-1}}{3} M_i + \frac{h_i}{6} M_{i+1} = \frac{y_{i+1} - y_i}{h_i} - \frac{y_i - y_{i-1}}{h_{i-1}} \quad (8)$$

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

La cele  $(n-1)$  ecuații date de (8) adăugăm două ecuații care corespund condițiilor:  $s'(x_0) = y'_0$  și  $s'(x_n) = y'_n$ .

Ținând seama de (7) aceste ecuații sunt:

$$\frac{h_0}{3} M_0 + \frac{h_0}{6} M_1 = \frac{y_1 - y_0}{h_0} - y'_0 \quad (9)$$

$$\frac{h_{n-1}}{6} M_{n-1} + \frac{h_{n-1}}{3} M_n = y'_n - \frac{y_n - y_{n-1}}{h_{n-1}} \quad (10)$$

Din (8), (9) și (10) rezultă următorul sistem  $AM=b$ , unde

$$A = \begin{pmatrix} \frac{h_0}{3} & \frac{h_0}{6} & 0 & 0 & \dots & 0 & 0 & 0 \\ \frac{h_0}{6} & \frac{h_0 + h_1}{3} & \frac{h_1}{6} & 0 & \dots & 0 & 0 & 0 \\ 0 & \frac{h_1}{3} & \frac{h_1 + h_2}{6} & \frac{h_2}{6} & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & \frac{h_{n-2}}{6} & \frac{h_{n-2} + h_{n-1}}{3} & \frac{h_{n-1}}{6} \\ 0 & 0 & 0 & 0 & \dots & 0 & \frac{h_{n-1}}{6} & \frac{h_{n-1}}{3} \end{pmatrix}$$

$$M = \begin{pmatrix} M_0 \\ M_1 \\ \vdots \\ M_n \end{pmatrix}, \text{ iar } b = \begin{pmatrix} \frac{y_1 - y_0}{h_0} - y'_0 \\ \frac{y_2 - y_1}{h_1} - \frac{y_1 - y_0}{h_0} \\ \vdots \\ \frac{y_n - y_{n-1}}{h_{n-1}} - \frac{y_{n-1} - y_{n-2}}{h_{n-2}} \\ y'_n - \frac{y_n - y_{n-1}}{h_{n-1}} \end{pmatrix}.$$

Observăm că matricea  $A$  este simetrică, tridiagonală și strict diagonal dominantă. Un asemenea sistem are soluție unică, care se obține ușor cu algoritmul Gauss. Înlocuind în (6) această soluție, găsim funcția spline cubică pe care o căutam. Evident această funcție este unică, deoarece soluția  $(M_0, M_1, \dots, M_n)$  este unică.  $\square$

**Exemplu:** Să se determine valorile funcției  $f$  în punctul  $\frac{5\pi}{24}$  folosind interpolarea spline cubică, știind că:

$x_i$	0	$\frac{\pi}{6}$	$\frac{\pi}{4}$	$\frac{\pi}{3}$	$\frac{\pi}{2}$
$y_i = f(x_i)$	0	0.5	0.70711	0.86603	1

și că valorile lui  $f'$  în punctele  $x_0$  și  $x_4$  sunt:  $y'_0 = 1, y'_4 = 0$ .

R. Aplicăm teorema 1 și vom găsi funcția spline cubică  $s$  ce interpolează funcția  $f$ , dacă determinăm coeficienții  $M_i, i = \overline{0, 3}$ .

Coeficienții  $M_i$  se determină prin rezolvarea sistemului liniar  $AM = b$ , unde

$$A := \begin{bmatrix} \frac{h_0}{3} & \frac{h_0}{6} & 0 & 0 & 0 \\ \frac{h_0}{6} & \frac{h_0+h_1}{3} & \frac{h_1}{6} & 0 & 0 \\ 0 & \frac{h_1}{6} & \frac{h_1+h_2}{3} & \frac{h_2}{6} & 0 \\ 0 & 0 & \frac{h_2}{6} & \frac{h_2+h_3}{3} & \frac{h_3}{6} \\ 0 & 0 & 0 & \frac{h_3}{6} & \frac{h_3}{3} \end{bmatrix} \quad b := \begin{bmatrix} \frac{y_1-y_0}{h_0} - y'_0 \\ \frac{y_2-y_1}{h_1} - \frac{y_1-y_0}{h_0} \\ \frac{y_3-y_2}{h_2} - \frac{y_2-y_1}{h_1} \\ \frac{y_4-y_3}{h_3} - \frac{y_3-y_2}{h_2} \\ y'_4 - \frac{y_4-y_3}{h_3} \end{bmatrix}, \quad M = \begin{pmatrix} M_0 \\ M_1 \\ M_2 \\ M_3 \\ M_4 \end{pmatrix}$$

și  $h_i = x_{i+1} - x_i, i = \overline{0, 3}$ . Obținem că:  $M = \begin{pmatrix} -5.15454 \cdot 10^{-3} \\ -0.50616 \\ -0.70767 \\ -0.88161 \\ -1.02524 \end{pmatrix}$ .

Punctul  $\frac{5\pi}{24} \in [x_1, x_2]$ . Scriem funcția de interpolare  $s$  pe acest interval:

$$s(x) = \frac{(x_2 - x)^3 M_1 + (x - x_1)^3 M_2}{6h_1} + \frac{(x_2 - x)y_1 + (x - x_1)y_2}{h_1} - \frac{(x_2 - x)M_1 + (x - x_1)M_2}{6} h_1$$

Calculăm valoarea funcției  $s$  în punctul dat:  $s\left(\frac{5\pi}{4}\right) = 0.60875$ .

Deci, valoarea aproximativă a lui  $f$  în punctul  $\frac{5\pi}{24}$  este 0.60875.

Funcțiile spline cubice au următoarea proprietate de optimizare.

**Teorema 2.** Fie  $G$  mulțimea funcțiilor  $g : [a, b] \rightarrow \mathbb{R}$ , de clasă  $C^2$  cu proprietățile:

- (i)  $g(x_i) = y_i$ ,  $0 \leq i \leq n$
- (ii)  $g'(x_0) = y'_0$
- (iii)  $g'(x_n) = y'_n$

Atunci:  $\inf_{g \in G} \int_a^b [g''(x)]^2 dx = \int_a^b [s''(x)]^2 dx$ .

**Demonstrație.**

Dacă notăm cu  $k(x) = s(x) - g(x)$ , unde  $g \in G$ , atunci

$$\int_a^b [g''(x)]^2 dx = \int_a^b [s''(x)]^2 dx - 2 \int_a^b s''(x)k''(x) dx + \int_a^b [k''(x)]^2 dx.$$

Mai departe avem

$$\begin{aligned} \int_a^b s''(x)k''(x) dx &= \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} s''(x)k''(x) dx = \\ &= \sum_{i=0}^{n-1} \left( s''(x)k'(x) \Big|_{x_i}^{x_{i+1}} - \int_{x_i}^{x_{i+1}} s'''(x)k'(x) dx \right). \end{aligned}$$

Deoarece  $s'''(x) = \alpha_i$  este o constantă pe  $[x_i, x_{i+1}]$  și  $k(x_i) = k(x_{i+1}) = 0$ ,  $i = \overline{0, n-1}$ , rezultă

$$\int_{x_i}^{x_{i+1}} s'''(x)k'(x) dx = 0$$

și mai departe

$$\int_a^b s''(x)k''(x)dx = \sum_{i=0}^{n-1} [s''(x_{i+1})k'(x_{i+1}) - s''(x_i)k'(x_i)] =$$

$$= s''(b)k'(b) - s''(a)k'(a) = 0 .$$

deoarece  $k'(a) = k'(b) = 0$ . Așadar

$$\int_a^b [g''(x)]^2 dx = \int_a^b [s''(x)]^2 dx + \int_a^b [k''(x)]^2 dx .$$

Rezultă  $\int_a^b [s''(x)]^2 dx \leq \int_a^b [g''(x)]^2 dx$  pentru orice  $g \in G$ .

Egalitatea are loc dacă  $\int_a^b [k''(x)]^2 dx = 0$ , deci dacă  $k''(x) = 0$ ,

$(\forall) x \in [a, b]$ . Așadar funcția  $k$  este liniară pe  $[a, b]$ .

Din condițiile de interpolare  $k(x_i) = 0$  pentru  $i = \overline{0, n}$ , rezultă  $k(x) = 0$ ,  $(\forall) x \in [a, b]$  și deci că

$$\int_a^b [s''(x)]^2 dx = \inf_{g \in G} \int_a^b [g''(x)]^2 dx . \quad \square$$

Se poate demonstra de asemenea următoarea teoremă.

**Teorema 3.** Fie  $f \in C^4[a, b]$  și  $M_4 = \sup\{|f^{(4)}(x)|; x \in [a, b]\}$  și fie  $x_i^{(n)} = a + ih$ ,  $i = \overline{0, n}$ , noduri echidistante, unde  $h = \frac{b-a}{n}$ . Dacă  $s_n$  este funcția spline cubică cu proprietățile:

$$s_n(x_i^{(n)}) = f(x_i^{(n)}), \quad i = \overline{0, n}; \quad s'_n(a) = f'(a) \quad \text{și} \quad s'_n(b) = f'(b),$$

atunci pentru orice  $x \in [a, b]$  avem:

$$|f(x) - s_n(x)| \leq \frac{5}{384} M_4 h^4, \quad |f'(x) - s'_n(x)| \leq \frac{1}{24} M_4 h^3, \quad |f''(x) - s''_n(x)| \leq \frac{3}{8} M_4 h^2 .$$

Așadar, din Teorema 3 rezultă că șirul funcțiilor spline cubice  $\{s_n\}$  care interpoaleză funcția  $f$  în nodurile echidistante  $\{x_i^{(n)}\}$  converge uniform pe intervalul  $[a, b]$  către funcția  $f$ . Mai mult:  $s'_n \xrightarrow{u} f'$  și  $s''_n \xrightarrow{u} f''$  pe intervalul  $[a, b]$ .

În continuare vom defini funcțiile *B-spline cubice* și vom arăta că orice funcție spline cubică care interpoaleză funcția  $f$  în nodurile  $x_0, x_1, \dots, x_n$  se reprezintă ca o combinație liniară unică de funcții B-spline cubice. Fie

$$\Delta: a = x_0 < x_1 < \dots < x_{i-1} < x_i < \dots < x_n = b$$

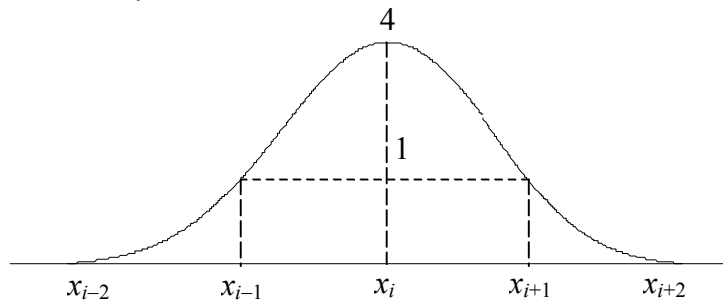
o diviziune a intervalului  $[a, b]$  cu noduri echidistante ( $x_i = x_0 + ih$ , unde  $h = \frac{b-a}{n}$ ). Asociem acestei diviziuni, diviziunea  $\tilde{\Delta}$  care are în plus șase noduri auxiliare, de asemenea echidistante.

$$\tilde{\Delta}: x_{-3} < x_{-2} < x_{-1} < x_0 < x_1 < \dots < x_n < x_{n+1} < x_{n+2} < x_{n+3}.$$

Definim pentru orice  $i = \overline{-1, n+1}$ , funcția B-spline cubică  $B_i$  astfel:

$$B_i(x) = \frac{1}{h^3} \begin{cases} (x - x_{i-2})^3 & x \in (x_{i-2}, x_{i-1}] \\ h^3 + 3h^2(x - x_{i-1}) + 3h(x - x_{i-1})^2 - 3(x - x_{i-1})^3 & x \in (x_{i-1}, x_i] \\ h^3 + 3h^2(x_{i+1} - x) + 3h(x_{i+1} - x)^2 - 3(x_{i+1} - x)^3 & x \in (x_i, x_{i+1}] \\ (x_{i+2} - x)^3 & x \in (x_{i+1}, x_{i+2}] \\ 0 & x \notin [x_{i-2}, x_{i+2}] \end{cases}$$

Graficul funcției  $B_i$  arată astfel:



Din definiția funcțiilor  $B_i$  rezultă că

$$B_i(y_j) = \begin{cases} 4 & \text{dacă } j=i \\ 1 & \text{dacă } j=i-1 \text{ sau } j=i+1 \\ 0 & \text{dacă } j \notin \{i-1, i, i+1\} \end{cases}$$

Se verifică ușor că  $B'_i(x_{i-1}) = \frac{3}{h}$ ,  $B'_i(x_{i+1}) = -\frac{3}{h}$  și  $B'_i(x_j) = 0$  dacă  $j \notin \{x_{i-1}, x_{i+1}\}$ . Se observă de asemenea că  $B_i \in C^2(\mathbb{R})$ , deci este o funcție spline cubică.

**Propoziția 2.** Funcțiile  $B_{-1}, B_0, \dots, B_{n+1}$  sunt liniar independente pe  $\mathbb{R}$ .

**Demonstrație.** Fie combinația liniară

$$\lambda_{-1}B_{-1}(x) + \lambda_0B_0(x) + \lambda_1B_1(x) + \dots + \lambda_{n+1}B_{n+1}(x) = 0, \quad x \in \mathbb{R}.$$

Dacă dăm lui  $x$  succesiv valorile  $x_{-1}, x_0, \dots, x_{n+1}$  obținem sistemul:

$$\begin{cases} 4\lambda_{-1} + \lambda_0 = 0 \\ \lambda_{-1} + 4\lambda_0 + \lambda_1 = 0 \\ \dots \\ \lambda_{n-1} + 4\lambda_n + \lambda_{n+1} = 0 \\ \lambda_n + 4\lambda_{n+1} = 0 \end{cases}$$

Deoarece matricea sistemului este strict diagonal dominantă, deci nesingulară, rezultă că sistemul admite numai soluția banală

$$\lambda_1 = \lambda_2 = \dots = \lambda_{n+1} = 0. \quad \square$$

În continuare notăm cu  $B_3(\Delta)$  spațiul liniar generat de funcțiile  $B_i$ ,  $i = \overline{-1, n+1}$ .

**Teorema 4.** Există o funcție unică  $B \in B_3(\Delta)$  care interpolează funcția  $f$  în nodurile  $x_i$ ,  $i = \overline{0, n}$ .

**Demonstrație.** Fie

$$B(x) = a_{-1}B_{-1}(x) + a_0B_0(x) + a_1B_1(x) + \dots + a_{n+1}B_{n+1}(x), \quad x \in \mathbb{R}. \quad (11)$$

Punând condiția ca  $B$  să interpoleze funcția  $f$  în nodurile  $x_0, x_1, \dots, x_n$  rezultă

$$B'(x_0) = a_{-1}B'_{-1}(x_0) + a_0B'_0(x_0) + \dots + a_{n+1}B'_{n+1}(x_0) = f'(x_0)$$

$$B(x_i) = a_{-1}B_{-1}(x_i) + a_0B_0(x_i) + \dots + a_{n+1}B_{n+1}(x_i) = f(x_i) \quad i = \overline{0, n}$$

$$B'(x_n) = a_{-1}B'_{-1}(x_n) + a_0B'_0(x_n) + \dots + a_{n+1}B'_{n+1}(x_n) = f'(x_n)$$

Se obține sistemul:

$$\begin{cases} -\frac{3}{h}a_{-1} + \frac{3}{h}a_1 = f'(x_0) \\ a_{-1} + 4a_0 + a_1 = f(x_0) \\ a_0 + 4a_1 + a_2 = f(x_1) \\ \dots \\ a_{n-1} + 4a_n + a_{n+1} = f(x_n) \\ -\frac{3}{h}a_{n-1} + \frac{3}{h}a_{n+1} = f'(x_n) \end{cases} \quad (12)$$

Sistemul (12) are  $(n+3)$  ecuații liniare și  $(n+3)$  necunoscute

$$a_{-1}, a_0, \dots, a_{n+1}.$$

Eliminând necunoscuta  $a_{-1}$  din primele două ecuații și necunoscuta  $a_{n+1}$  din ultimele două ecuații, obținem sistemul echivalent:

$$\begin{cases} 4a_0 + 2a_1 = f(x_0) + \frac{h}{3}f'(x_0) \\ a_0 + 4a_1 + a_2 = f(x_0) \\ \dots \\ a_{n-2} + 4a_{n-1} + a_n = f(x_{n-1}) \\ 2a_{n-1} + 4a_n = f(x_n) + \frac{h}{3}f'(x_n) \end{cases} \quad (13)$$

Matricea coeficienților sistemului (13) este strict diagonal dominantă, deci sistemul (13) are soluție unică. Așadar, sistemul (12) are soluție unică.

Înlocuind această soluție în (11) obținem funcția  $B$  căutată, care evident este unică.  $\square$

**Teorema 5.** Fie  $s : [a, b] \rightarrow \mathbb{R}$  o funcție spline cubică care interpolează funcția  $f$  în nodurile  $x_0, x_1, \dots, x_n$ . Atunci există  $a_{-1}, a_0, a_1, \dots, a_{n+1} \in \mathbb{R}$ , unic determinate, astfel încât

$$s(x) = a_{-1}B_{-1}(x) + a_0B_0(x) + \dots + a_{n+1}B_{n+1}(x), \quad (\forall) x \in [a, b].$$

**Demonstrație.**

Din Teorema 4 rezultă că există  $B \in \mathcal{B}_3(\Delta)$  astfel încât  $B$  interpolează funcția  $f$  în nodurile  $x_0, x_1, \dots, x_n$ . Funcția  $B$  este de clasă  $C^2$  pe  $\mathbb{R}$  și este polinomială de gradul trei pe porțiuni. Rezultă că restricția lui  $B$  la intervalul  $[a, b]$  este o funcție spline cubică, care interpolează  $f$  în nodurile  $x_0, x_1, \dots, x_n$ . Cum asemenea funcție este unică (conform Teoremei 1) rezultă

$$s(x) = B(x) = a_{-1}B_{-1}(x) + a_0B_0(x) + \dots + a_{n+1}B_{n+1}(x), \quad (\forall) x \in [a, b].$$

Unicitatea coeficienților

$a_{-1}, a_0, a_1, \dots, a_{n+1}$   
este asigurată de Teorema 4.  $\square$

Pachetul de programe MATLAB conține funcția *spline* care permite interpolarea unei funcții  $f$  în punctele  $x_1, x_2, \dots, x_n$ ,  $(\forall) n \in \mathbb{N}^*$ , finit, printr-o funcție spline cubică, dacă se cunosc valorile  $y_1, y_2, \dots, y_n$  ale funcției în nodurile  $x_1 < x_2 < \dots < x_n$ . Secvența de apelare este:  
 $y_i = \text{spline}(x, y, x_i)$ .

**Exemplu.** Să se determine valorile funcției  $f$  în punctele  $\frac{\pi}{12}, \frac{\pi}{8}, \frac{\pi}{5}$ , folosind interpolarea spline cubică, știind că:

$x_i$	0	$\frac{\pi}{6}$	$\frac{\pi}{4}$	$\frac{\pi}{3}$	$\frac{\pi}{2}$
$y_i = f(x_i)$	0	$\frac{1}{2}$	$\frac{1}{\sqrt{2}}$	$\frac{\sqrt{3}}{2}$	1

utilizând MATLAB.

În mediul MATLAB se scriu comenzile:

```

% interpolarea cu functii spline cubice folosind pachetul de programe
Matlab
function [x,y,xi,yi]=cub
% Nodurile
x=[0,pi/6,pi/4,pi/3,pi/2];
% Valorile functiei in noduri
y=[0,1/2,1/2^(1/2),3^(1/2)/2,1];
% Valorile in care se interpoleaza functia
xi=[pi/12,pi/8,pi/5];
% Apelarea functiei Matlab spline care face interpolarea
yi=spline(x,y,xi);
Funcția considerată este  $f(x)=\sin x$  și putem compara valorile de interpolare
cu cele "exacte":
zi=sin(xi);
Pentru reprezentarea grafică a funcției interpolate se poate folosi
următoarea secvența MATLAB.
% Reprezentarea grafica a functiei interpolate
plot(x,y,xi,yi,'*',xi,zi,'o');
axis([0,pi/2,0,1.2]); % se stabilesc intervalele de reprezentare pe
axe
title('Interpolarea cu spline cubice');
xlabel('Unghiul');
ylabel('Valorile functiei'); grid

```

### Exerciții

Folosind polinomul de interpolare a lui Lagrange să se determine valoarea aproximativă a funcțiilor date de tabelele următoare în punctele  $a$  menționate în fiecare caz.

1.

$x$	-3	-2	0	1	3
$y=f(x)$	91	23	1	-1	73

$a=-1$ .

R.

$$\begin{aligned}
 P_4(a) &= \sum_{i=1}^5 y_i \prod_{\substack{j=1 \\ j \neq i}}^5 \frac{a-x_j}{x_i-x_j} = y_1 \frac{a-x_2}{x_1-x_2} \cdot \frac{a-x_3}{x_1-x_3} \cdot \frac{a-x_4}{x_1-x_4} \cdot \frac{a-x_5}{x_1-x_5} + \\
 &+ y_2 \frac{a-x_1}{x_2-x_1} \cdot \frac{a-x_3}{x_2-x_3} \cdot \frac{a-x_4}{x_2-x_4} \cdot \frac{a-x_5}{x_2-x_5} + y_3 \frac{a-x_1}{x_3-x_1} \cdot \frac{a-x_2}{x_3-x_2} \cdot \frac{a-x_4}{x_3-x_4} \cdot \frac{a-x_5}{x_3-x_5} + \\
 &+ y_4 \frac{a-x_1}{x_4-x_1} \cdot \frac{a-x_2}{x_4-x_2} \cdot \frac{a-x_3}{x_4-x_3} \cdot \frac{a-x_5}{x_4-x_5} + y_5 \frac{a-x_1}{x_5-x_1} \cdot \frac{a-x_2}{x_5-x_2} \cdot \frac{a-x_3}{x_5-x_3} \cdot \frac{a-x_4}{x_5-x_4} = \\
 &= 91 \frac{-1+2}{-3+2} \cdot \frac{-1}{-3} \cdot \frac{-1-1}{-3-1} \cdot \frac{-1-3}{-3-3} + 23 \frac{-1+3}{-2+3} \cdot \frac{-1}{-2} \cdot \frac{-1-1}{-2-1} \cdot \frac{-1-3}{-2-3} + \\
 &+ \frac{-1+3}{3} \cdot \frac{-1+2}{2} \cdot \frac{-1-1}{-1} \cdot \frac{-1-3}{-3} - \frac{-1+3}{1+3} \cdot \frac{-1+2}{1+2} \cdot \frac{-1}{1} \cdot \frac{-1-3}{1-3} + \\
 &+ 73 \frac{-1+3}{3+3} \cdot \frac{-1+2}{3+2} \cdot \frac{-1}{3} \cdot \frac{-1-1}{3-1} = 5 .
 \end{aligned}$$

2.

$x$	0	$\frac{\pi}{6}$	$\frac{\pi}{4}$	$\frac{\pi}{3}$	$\frac{2\pi}{5}$	$\frac{\pi}{2}$
$y=f(x)$	0	0.5	1.70711	0.86603	0.95106	1

$$a = \frac{3\pi}{8} .$$

$$R. \quad P_5\left(\frac{3\pi}{8}\right) = 0.92388 .$$

Folosind metoda Aitken să se găsească valoarea aproximativă a funcțiilor date de tabelele următoare în punctele menționate în fiecare caz în parte.

3.

$x$	-2	-1	0	1	3
$y=f(x)$	-12	-5	-4	-3	23

$$a=0.5 .$$

R. Urmărind calculele ca în Teorema 1 §4.2 obținem tabelul următor în care ultima celulă dă valoarea aproximativă a funcției în punctul dat.

-2	-2.5	-12				
-1	-1.5	-5	5.5			
0	-0.5	-4	-2	-5.75		
1	0.5	-3	-4.5	-2	-3.875	
3	2.5	23	5.5	5.5	-3.875	-3.875

4.

$x$	0	30	45	60	90
$y=f(x)$	0	0.5	0.70710	0.86602	1

$a=36$  .

R.

0	-36	0				
30	-6	0.5	0.6			
45	9	0.70710	0.56568	0.58627		
60	24	0.86602	0.51961	0.58392	0.58768	
90	54	1	0.4	0.58	0.58752	0.58780

5. Să se determine valorile funcției  $f$  în punctul 0.5 folosind interpolarea spline cubică, știind că:

$x_i$	0	0.25	0.75	1
$y_i = f(x_i)$	1	0.96923	0.75484	0.60653

și că valorile lui  $f'$  în punctele  $x_0$  și  $x_3$  sunt:  $y'_0 = 0$ ,  $y'_3 = -0.60653$  .

R. Vom găsi funcția spline cubică  $s$  ce interpoalează funcția  $f$  , dacă determinăm coeficienții  $M_i$ ,  $i = \overline{0, 2}$  . Coeficienții  $M_i$  se determină prin rezolvarea sistemului de ecuații liniare  $AM = b$ , unde

$$A := \begin{bmatrix} \frac{h_0}{3} & \frac{h_0}{6} & 0 & 0 \\ \frac{h_0}{6} & \frac{h_0+h_1}{3} & \frac{h_1}{6} & 0 \\ 0 & \frac{h_1}{6} & \frac{h_1+h_2}{3} & \frac{h_2}{6} \\ 0 & 0 & \frac{h_2}{6} & \frac{h_2}{3} \end{bmatrix} \quad b := \begin{bmatrix} \frac{y_1-y_0}{h_0} - y'_0 \\ \frac{y_2-y_1}{h_1} - \frac{y_1-y_0}{h_0} \\ \frac{y_3-y_2}{h_2} - \frac{y_2-y_1}{h_1} \\ y'_3 - \frac{y_3-y_2}{h_2} \end{bmatrix}$$

iar  $h_i = x_{i+1} - x_i$ ,  $i = \overline{0, 2}$

$$\text{iar } M = \begin{pmatrix} M_0 \\ M_1 \\ M_2 \\ M_3 \end{pmatrix}. \text{ Obținem că: } M = \begin{pmatrix} -1.00712 \\ -0.93936 \\ -0.347 \\ 0.01396 \end{pmatrix}$$

Punctul  $0.5 \in [x_1, x_2]$ . Scriem funcția de interpolare  $s$  pe acest interval:

$$s(x) = \frac{(x_2 - x)^3 M_1 + (x - x_1)^3 M_2}{6h_1} + \frac{(x_2 - x)y_1 + (x - x_1)y_2}{h_1} - \frac{(x_2 - x)M_1 + (x - x_1)M_2}{6} h_1$$

Calculăm valoarea funcției  $s$  în punctul dat:  $s(0.5) = 0.882$ .

Deci, valoarea aproximativă a lui  $f$  în punctul  $0.5$  este  $0.882$ .

6. Să se determine valorile funcției  $f$  în punctul 3 folosind interpolarea spline cubică, știind că:

$x_i$	1	2	4	5
$y_i = f(x_i)$	0.5403	0.70121	0.80805	0.83382

și că valorile lui  $f'$  în punctele  $x_0$  și  $x_3$  sunt:  $y'_0 = 0.28049$ ,  $y'_3 = 0.02152$ .

$$\text{R. Obținem că: } M = \begin{pmatrix} -0.33459 \\ -0.0483 \\ -0.01028 \\ -7.60083 \cdot 10^{-3} \end{pmatrix}. \text{ Valoarea funcției } f \text{ în punctul } 3 \text{ este}$$

aproximată de  $s(3) = 0.76928$ .

## 5. Integrarea și derivarea numerică

Fie  $f: [a, b] \rightarrow \mathbb{R}$  o funcție integrabilă. Dacă putem găsi o primitivă  $F$  a funcției  $f$ , atunci conform *formulei Leibniz–Newton* avem

$$\int_a^b f(x)dx = F(b) - F(a).$$

Dacă nu putem găsi o primitivă  $F$  a funcției  $f$ , atunci pentru calculul integralei  $\int_a^b f(x)dx$  vom folosi o metodă numerică aproximativă.

O abordare firească este să aproximăm funcția  $f$  printr-un polinom, de exemplu prin polinomul de interpolare a lui Lagrange și să integrăm acest polinom.

Prin *formulă de integrare numerică (cuadratură numerică)* se înțelege o formulă de următorul tip

$$\int_a^b w(x)f(x)dx = A_1f(x_1) + \dots + A_n f(x_n) + R(f) \quad (1)$$

unde  $\{x_i\}$  se numesc noduri, iar  $\{A_i\}$  se numesc coeficienți.

Funcția  $w$  este o funcție fixată, integrabilă, care se numește *funcția pondere*. În multe cazuri  $w(x)=1$ ,  $(\forall) x \in [a, b]$ . Despre funcția  $f$  se presupune că este integrabilă pe  $[a, b]$  și definită în nodurile  $\{x_i\}$ .

Cu  $R(f)$  se notează *eroarea de aproximare a integralei*.

**Definiția 1.** Formula (1) se spune că este exactă pentru funcția  $f$  dacă  $R(f) = 0$ , deci dacă

$$\int_a^b w(x)f(x)dx = A_1f(x_1) + A_2f(x_2) + \dots + A_n f(x_n).$$

Formula (1) se spune că este de gradul  $m$  dacă:

- (i) este exactă pentru orice polinom de grad cel mult  $m$ ;
- (ii) există un polinom de gradul  $(m+1)$  pentru care formula (1)

nu este exactă.

**Teorema 1.** Pentru orice  $n$  noduri distincte,  $x_1, \dots, x_n$ , există  $n$  constante  $A_1, A_2, \dots, A_n$  (care nu depind de  $f$ ) astfel încât formula

$$\int_a^b w(x)f(x)dx = A_1f(x_1) + A_2f(x_2) + \dots + A_nf(x_n) + R(f)$$

este exactă pentru orice polinom de grad cel mult  $(n-1)$ .

**Demonstrație.** Fie

$$P_{n-1}(x) = \sum_{i=1}^n L_i(x)f(x_i)$$

polinomul lui Lagrange, care interpolează funcția  $f$  în nodurile  $\{x_i\}$ ,  $i = \overline{1, n}$ .  
Reamintim că

$$L_i(x) = \prod_{\substack{j=1 \\ j \neq i}}^n \frac{x - x_j}{x_i - x_j}.$$

Dacă notăm cu

$$E(f; x) = f(x) - P_{n-1}(x),$$

atunci

$$f(x) = \sum_{i=1}^n L_i(x)f(x_i) + E(f; x)$$

și mai departe

$$\int_a^b w(x)f(x)dx = \sum_{i=1}^n \left( \int_a^b w(x)L_i(x)dx \right) f(x_i) + \int_a^b w(x)E(f; x)dx.$$

Notăm cu

$$A_i = \int_a^b w(x)L_i(x)dx, \quad i = \overline{1, n}.$$

Evident  $A_i$  nu depind de  $f$ , ci de  $a, b$ , nodurile  $x_i$  și de ponderea  $w$ .

Dacă notăm cu

$$R(f) = \int_a^b w(x)E(f; x)dx,$$

atunci

$$\int_a^b w(x)f(x)dx = A_1f(x_1) + A_2f(x_2) + \dots + A_nf(x_n) + R(f).$$

Dacă  $f$  este un polinom de grad cel mult  $(n-1)$ , atunci  $E(f; x) = 0$  și deci  $R(f) = 0$  (Capitolul IV, §1, Observația 1).  $\square$

Există trei tipuri de formule de integrare numerică:

1. Formule *Newton-Côtes*,
2. Formule *Gauss*,
3. Formule *Romberg*.

În cele ce urmează vom prezenta primele două tipuri de formule.

### §5.1. Formule *Newton-Côtes*

Presupunem intervalul  $[a, b]$  finit, nodurile echidistante

$x_i = a + ih$ , unde  $h = \frac{b-a}{n}$  și  $w(x) = 1$ , pentru orice  $x \in [a, b]$ .

Fie  $L_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^n \frac{x - x_j}{x_i - x_j}$  și fie  $P_n$  – polinomul Lagrange care

interpolează funcția  $f$  în nodurile  $x_0, x_1, \dots, x_n$ . Avem

$$P_n(x) = \sum_{i=0}^n L_i(x) f(x_i).$$

Dacă facem schimbarea de variabilă  $x = a + th$ , atunci, așa cum s-a văzut în § 4.1, avem

$$\tilde{P}_n(t) = P_n(a + th) = \sum_{i=0}^n \frac{(-1)^{n-i} C_n^i}{n!} \frac{\pi_{n+1}(t)}{t-i} f(x_i) \quad \text{și}$$

$$\tilde{E}(f; t) = \frac{\pi_{n+1}(t)}{(n+1)!} h^{n+1} f^{(n+1)}(\xi_t),$$

unde

$$\pi_{n+1}(t) = t(t-1)(t-2)\dots(t-n).$$

În continuare avem

$$\int_a^b f(x) dx = \int_a^b P_n(x) dx + \int_a^b E(f; x) dx.$$

Cu schimbarea de variabilă  $x = a + th$ , obținem

$$\begin{aligned} \int_a^b f(x) dx &= \int_0^n \tilde{P}_n(t) h dt + \int_0^n \tilde{E}(f; t) h dt = \\ &= h \sum_{i=0}^n \left( \frac{(-1)^{n-i} C_n^i}{n!} \int_0^n \frac{\pi_{n+1}(t)}{t-i} dt \right) f(x_i) + \frac{h^{n+2}}{(n+1)!} \int_0^n \pi_{n+1}(t) f^{(n+1)}(\xi_t) dt. \end{aligned}$$

Introducem notațiile:

$$d_i = \frac{(-1)^{n-i} C_n^i}{n!} \int_0^1 \frac{\pi_{n+1}(t)}{t-i} dt, \quad i = \overline{0, n} \quad (2)$$

și

$$R(f) = \frac{h^{n+2}}{(n+1)!} \int_0^1 \pi_{n+1}(t) f^{(n+1)}(\xi_t) dt. \quad (3)$$

Numerele  $\{d_i\}$ ,  $i = \overline{0, n}$ , se numesc *coeficienții Newton-Côtes*. Prin urmare avem

$$\int_a^b f(x) dx = \sum_{i=0}^n A_i f(x_i) + R(f) \quad (4)$$

unde  $A_i = h d_i$ ,  $i = \overline{0, n}$ .

Primele trei formule Newton-Côtes au nume speciale.

Dacă  $n=1$  obținem *formula trapezelor*. În acest caz  $h = b-a$ ,

$$d_0 = \frac{(-1)^1 C_1^0}{1!} \int_0^1 \frac{t(t-1)}{t} dt = - \int_0^1 (t-1) dt = \frac{1}{2}, \quad d_1 = \frac{(-1)^0 C_1^1}{1!} \int_0^1 \frac{t(t-1)}{t-1} dt = \int_0^1 t dt = \frac{1}{2}$$

Rezultă

$$A_1 = A_2 = \frac{h}{2}.$$

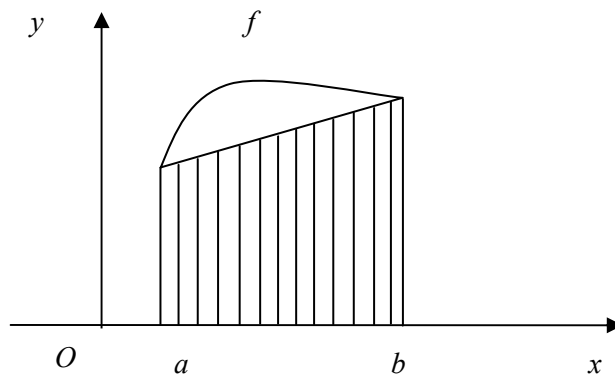
Formula trapezelor este

$$\int_a^b f(x) dx = \frac{b-a}{2} [f(a) + f(b)] + R(f) \quad (5)$$

$$R(f) = \frac{h^3}{2} \int_0^1 f''(\xi_t) t(t-1) dt.$$

Dacă  $f \in C^2[a, b]$  și notăm cu  $M_2 = \sup\{|f''(x)|; x \in [a, b]\}$  atunci

$$|R(f)| \leq \frac{(b-a)^3}{2} M_2 \int_0^1 |t(t-1)| dt = \frac{(b-a)^3}{12} M_2 \quad (6)$$



Din punct de vedere geometric formula trapezelor (5) revine la a aproxima aria mulțimii plane mărginită de graficul funcției  $f$ , axa  $Ox$  și dreptele  $x = a$ ,  $x = b$  cu aria trapezului hașurat în figură.

Dacă  $n=2$  obținem *formula Simpson*. În acest caz  $h = \frac{b-a}{2}$ ,

$$d_0 = \frac{(-1)^2 C_2^0}{2!} \int_0^2 \frac{t(t-1)(t-2)}{t} dt = \frac{1}{2} \int_0^2 (t^2 - 3t + 2) dt = \frac{1}{3}$$

$$d_1 = \frac{(-1)^1 C_2^1}{2!} \int_0^2 \frac{t(t-1)(t-2)}{t-1} dt = - \int_0^2 (t^2 - 2t) dt = \frac{4}{3}$$

$$d_2 = \frac{(-1)^0 C_2^2}{2!} \int_0^2 \frac{t(t-1)(t-2)}{t-2} dt = \frac{1}{2} \int_0^2 (t^2 - t) dt = \frac{1}{3}.$$

Formula Simpson este

$$\int_a^b f(x) dx = \frac{b-a}{6} \left[ f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right] + R(f) . \quad (7)$$

Se poate arăta că dacă  $f \in C^4[a, b]$  atunci

$$|R(f)| \leq \frac{M_4(b-a)^5}{2880} \quad (8)$$

unde  $M_4 = \sup\{|f^{(4)}(x)|; x \in [a, b]\}$ . În sfârșit, dacă  $n=3$ , atunci

$$h = \frac{b-a}{3}; d_0 = d_3 = \frac{3}{8}; d_1 = d_2 = \frac{9}{8}.$$

Se obține *formula 3/8 a lui Simpson*

$$\int_a^b f(x) dx = \frac{b-a}{8} [f(a) + 3f(a+h) + 3f(a+2h) + f(b)] + R(f) . \quad (9)$$

Pentru o mai bună aproximare a integralei se folosesc *formulele Newton-Côtes repetate*.

Dacă împărțim intervalul  $[a, b]$  în  $n$  subintervale egale și aplicăm formula (5) fiecărui interval  $[x_{i-1}, x_i]$ , obținem:

$$\int_a^b f(x) dx = \sum_{i=1}^n \int_{x_{i-1}}^{x_i} f(x) dx = \sum_{i=1}^n \frac{x_i - x_{i-1}}{2} [f(x_{i-1}) + f(x_i)] + R_i(f) .$$

Cum  $x_i - x_{i-1} = \frac{b-a}{n}$ ,  $x_0 = a$  și  $x_n = b$  avem

$$\int_a^b f(x) dx = \frac{b-a}{2n} \left[ f(a) + f(b) + 2 \sum_{i=1}^{n-1} f(x_i) \right] + R(f) \quad (10)$$

unde

$$|R(f)| \leq n \frac{M_2}{12} \left( \frac{b-a}{n} \right)^3 = \frac{M_2}{12n^2} (b-a)^3 . \quad (11)$$

Formula (10) se numește *formula trapezelor repetată*.

Dacă împărțim intervalul  $[a, b]$  în  $2n$  subintervale egale și aplicăm formula Simpson (7) fiecărui interval  $[x_{2i-2}, x_{2i}]$ , obținem *formula Simpson repetată*

$$\int_a^b f(x) dx = \frac{b-a}{6n} \left[ f(a) + f(b) + 4 \sum_{i=1}^n f(x_{2i-1}) + 2 \sum_{i=1}^{n-1} f(x_{2i}) \right] + R(f) \quad (12)$$

unde

$$|R(f)| \leq \frac{M_4}{2880n^4} (b-a)^5 . \quad (13)$$

Dacă notăm cu

$$\sigma_n^T = \frac{b-a}{2n} \left[ f(a) + f(b) + 2 \sum_{i=1}^{n-1} f(x_i) \right]$$

și cu

$$\sigma_n^S = \frac{b-a}{6n} \left[ f(a) + f(b) + 4 \sum_{i=1}^n f(x_{2i-1}) + 2 \sum_{i=1}^{n-1} f(x_{2i}) \right] ,$$

atunci se poate arăta că  $\sigma_n^T$  și  $\sigma_n^S$  sunt sume Riemann, sau combinații liniare de sume Riemann atașate funcției  $f$  și că

$$\lim_{n \rightarrow \infty} \sigma_n^T = \int_a^b f(x) dx, \quad \lim_{n \rightarrow \infty} \sigma_n^S = \int_a^b f(x) dx .$$

*Calculul aproximativ al integralelor definite folosind pachetul de programe MATLAB*

În MATLAB se află programate metodele trapezelor, Simpson și 3/8 Simpson (funcțiile `trapz`, `quad` și `quad8`).

Pentru calculul integralelor definite cu formula trapezelor trebuie să se cunoască  $X$ , vectorul absciselor și  $Y$ , vectorul (matricea) valorilor funcției (funcțiilor) corespunzătoare absciselor date de  $X$ .

Secvența de apelare este:  $Z = \text{trapz}(X)$ , când se calculează o singură integrală și  $Z$  este un număr, valoarea integralei, sau  $Z = \text{trapz}(X, Y)$ , când se calculează mai multe integrale deodată și  $Z$  este vectorul valorilor integralelor calculate. Această funcție consideră  $h=1$ , iar atunci când  $h \neq 1$ , se înmulțește cu  $h$  funcția `trapz`, așa cum se va vedea în exemplul următor.

**Exemplul 1.** Să se calculeze valoarea aproximativă a integralei  $\int_{0.5}^{2.5} \sin(x^2) dx$ ,

folosind funcția MATLAB `trapz`, și luând pasul  $h=0.1$ .

Fișierul de tip *m* cu care se realizează această cerință este:

```
% Calculul integralelor cu metoda trapezelor
function z=trapez
x=.5:0.1:2.5; %limita inferioara: pasul: limita superioara
y=sin(x.^2); % integrandul
z=0.1*trapz(y); % se inmulteste cu pasul, implicit pasul fiind 1
disp('Valoarea aproximativa a integralei');
```

Apelarea se face scriind *trapez*, iar valoarea afișată este 0.3924 .

Pentru calculul integralelor definite cu una din formulele Simpson trebuie să se cunoască *X*, vectorul absciselor și să se creeze un fișier de tip *m* care conține secvența de definire a funcției de integrat. Funcțiile *quad* și *quad8* se apelează cu una din formele:

<code>quad('f,a,b)</code>	<code>quad8('f,a,b)</code>
<code>quad('f,a,b,err)</code>	<code>quad8('f,a,b,err)</code>
<code>quad('f,a,b,err,urma)</code>	<code>quad8('f,a,b,err,urma)</code>

unde: *f* - este numele unui fișier funcție (de tip *m*) care conține descrierea funcției de integrat;

*a, b* - sunt limitele de integrare;

*err* - eroarea relativă admisă între doi pași consecutivi (implicit este  $10^{-3}$ );

*urma* - dacă este diferită de zero, se afișează pe ecran valorile intermediare.

Dacă nu se cunoaște expresia analitică a funcției, ci doar *X*, vectorul absciselor, și *Y*, vectorul valorilor funcției în aceste puncte, atunci se interpolează funcția și se consideră fișierul *f*, de tip *m*, care conține funcția de interpolare, după care se calculează integrala din această funcție.

Pentru exemplu de mai sus se creează fișierul *f* cu funcția de integrat:

```
% Fisierul cu functia de integrat numit f de tip m
```

```
function g=f(x)
```

```
g=sin(x.^2); % integrandul
```

după care se apelează cu secvența: *quad('f,0.5,2.5,0.00001)* și se va afișa valoarea 0.3890 . Am considerat eroarea relativă dintre doi pași consecutivi  $10^{-5}$ .

unde: *f* - este numele unui fișier funcție (de tip *m*) care conține descrierea funcției de integrat;

*a, b* - sunt limitele de integrare;

*err* - eroarea relativă admisă între doi pași consecutivi (implicit este  $10^{-3}$ );

*urma* - dacă este diferită de zero, se afișează pe ecran valorile intermediare.

Dacă nu se cunoaște expresia analitică a funcției, ci doar *X*, vectorul absciselor, și *Y*, vectorul valorilor funcției în aceste puncte, atunci se interpolează funcția și se consideră fișierul *f*, de tip *m*, care conține funcția de interpolare, după care se calculează integrala din această funcție.

Pentru exemplu de mai sus se creează fișierul *f* cu funcția de integrat:

```
% Fisierul cu functia de integrat numit f de tip m
```

```
function g=f(x)
g=sin(x.^2); % integrandul
```

după care se apelează cu secvența: `quad('f',0.5,2.5,0.00001)` și se va afișa valoarea 0.3890. Am considerat eroarea relativă dintre doi pași consecutivi  $10^{-5}$ .

## §5.2. Formule Gauss

Așa cum am văzut în Teorema 1, pentru orice  $n$  noduri distincte  $x_1, \dots, x_n$ , există  $n$  constante  $A_1, \dots, A_n$ , astfel încât formula

$$\int_a^b w(x)f(x)dx = A_1f(x_1) + A_2f(x_2) + \dots + A_nf(x_n) + R(f) \quad (1)$$

este exactă pentru orice polinom de grad cel mult  $(n-1)$ . În anumite cazuri, această formulă este exactă și pentru polinoame de grad mai mare.

De exemplu, formula Simpson, care corespunde la 3 noduri echidistante, este exactă pentru polinoame de grad cel mult 3.

În cele ce urmează vom stabili care este gradul maxim de exactitate al formulei (1). Vom arăta că pentru  $n$  noduri, gradul maxim de exactitate este  $(2n-1)$  și această se întâmplă pentru formula Gauss, când nodurile sunt rădăcinile unor polinoame ortogonale.

De acum înainte vom presupune că  $0 \leq w$ ,  $w$  este continuă pe  $[a,b]$  și  $w(x) = 0$  numai pentru un număr finit de puncte din  $[a,b]$ .

**Definiția 1.** Dacă

$$\int_a^b w(x)f(x)g(x)dx = 0$$

atunci spunem că funcțiile  $f$  și  $g$  sunt ortogonale pe  $[a, b]$  în raport cu ponderea  $w$ .

**Teorema 1.** Pentru orice  $n \in \mathbb{N}$ , există un polinom  $P_n^*$ , de gradul  $n$ , care este ortogonal pe  $[a, b]$  în raport cu ponderea  $w$ , pe orice polinom de grad cel mult  $(n-1)$ . Acest polinom este unic în afara unui factor constant de multiplicare nenul.

**Demonstrație.**

Demonstrația este constructivă. Vom arăta că se pot determina  $a_0, a_1, \dots, a_n$  astfel încât

$$\int_a^b w(x) (a_0 + a_1x + \dots + a_nx^n) Q(x) dx = 0$$

pentru orice polinom  $Q$  de grad cel mult  $(n-1)$ .

Observăm că dacă  $f$  este ortogonal pe  $g$  atunci  $\lambda f$  este ortogonal pe  $g$ . Rezultă că putem presupune  $a_n = 1$ . Observăm de asemenea că dacă  $f$  este ortogonal pe  $g_1$  și  $g_2$ , atunci  $f$  este ortogonal pe  $\alpha_1 g_1 + \alpha_2 g_2$ , oricare ar fi constantele  $\alpha_1$  și  $\alpha_2$ . Rezultă că este suficient să determinăm  $a_0, a_1, \dots, a_{n-1}$  astfel încât pentru orice  $m \in \{0, 1, \dots, n-1\}$  să avem:

$$\int_a^b w(x) (a_0 + a_1 x + \dots + a_{n-1} x^{n-1} + x^n) x^m dx = 0 \quad (2)$$

Dacă introducem notația

$$\int_a^b w(x) x^k dx = c_k \quad (3)$$

atunci, dându-i lui  $m$  pe rând valorile  $0, 1, \dots, (n-1)$  în relația (2) obținem:

$$\begin{cases} a_0 c_0 + a_1 c_1 + \dots + a_{n-1} c_{n-1} = -c_n \\ a_0 c_1 + a_1 c_2 + \dots + a_{n-1} c_n = -c_{n+1} \\ \vdots \\ a_0 c_{n-1} + a_1 c_n + \dots + a_{n-1} c_{2n-2} = -c_{2n-1} \end{cases} \quad (4)$$

Problema ar fi rezolvată dacă am arăta că  $\det C \neq 0$ , unde

$$C = \begin{pmatrix} c_0 & c_1 & \dots & c_{n-1} \\ c_1 & c_2 & \dots & c_n \\ \vdots & \vdots & \vdots & \vdots \\ c_{n-1} & c_n & \dots & c_{2n-2} \end{pmatrix}.$$

Presupunem prin absurd că  $\det C = 0$ . Atunci există  $\lambda_0, \lambda_1, \dots, \lambda_{n-1}$  nu toți nuli, astfel încât

$$\begin{cases} \lambda_0 c_0 + \lambda_1 c_1 + \dots + \lambda_{n-1} c_{n-1} = 0 \\ \lambda_0 c_1 + \lambda_1 c_2 + \dots + \lambda_{n-1} c_n = 0 \\ \vdots \\ \lambda_0 c_{n-1} + \lambda_1 c_n + \dots + \lambda_{n-1} c_{2n-2} = 0 \end{cases} \quad (5)$$

Înlocuind expresiile coeficienților  $c_k$  dați de relațiile (3) în (5) rezultă:

$$\begin{cases} \int_a^b w(x) (\lambda_0 + \lambda_1 x + \dots + \lambda_{n-1} x^{n-1}) dx = 0 \\ \vdots \\ \int_a^b w(x) (\lambda_0 x^{n-1} + \lambda_1 x^n + \dots + \lambda_{n-1} x^{2n-2}) dx = 0 \end{cases} \quad (6)$$

Amplificând pe rând, prima relație cu  $\lambda_0$ , a doua cu  $\lambda_1$ , și așa mai departe, și adunându-le, rezultă

$$\int_a^b w(x) [\lambda_0 + \lambda_1 x + \dots + \lambda_{n-1} x^{n-1}]^2 dx = 0 \quad (7)$$

Din (7) rezultă:

$$w(x) [\lambda_0 + \lambda_1 x + \dots + \lambda_{n-1} x^{n-1}]^2 = 0$$

pentru orice  $x \in [a, b]$ . Cum  $w$  se anulează numai într-un număr finit de puncte din  $[a, b]$ , rezultă  $\lambda_0 + \lambda_1 x + \dots + \lambda_{n-1} x^{n-1} = 0$ . Am ajuns astfel la o contradicție. Așadar,  $\det C \neq 0$  deci sistemul (4) admite soluție unică.

Dacă notăm cu  $a_0^*, a_1^*, \dots, a_{n-1}^*$  soluția sistemului (4), atunci

$$P_n^*(x) = a_0^* + a_1^* x + \dots + a_{n-1}^* x^{n-1} + x^n$$

și cu aceasta teorema este demonstrată.  $\square$

Polinoamele ortogonale  $\{P_n^*\}$  au diferite denumiri în funcție de ponderea  $w$  și de intervalul  $[a, b]$ .

Intervalul	Ponderea	Numele polinomului $P_n^*$
$[-1, 1]$	$w(x)=1$	Polinomul Legendre
$[-1, 1]$	$w(x) = \frac{1}{\sqrt{1-x^2}}$	Polinomul Cebîșev de speța I
$[-1, 1]$	$w(x) = \sqrt{1-x^2}$	Polinomul Cebîșev de speța II
$(-\infty, \infty)$	$w(x) = e^{-x^2}$	Polinomul Hermite
$[0, \infty)$	$w(x) = e^{-x}$	Polinomul Laguerre
$[-1, 1]$	$w(x) = (1-x)^\alpha (1+x)^\beta$ , $\alpha, \beta > -1$	Polinomul Jacobi

În continuare vom exemplifica cum se pot construi polinoamele ortogonale  $\{P_n^*\}$  cu ajutorul Teoremei 1.

Fie  $[a, b] = [-1, 1]$  și  $w(x)=1$ .

$$c_k = \int_{-1}^1 x^k dx = \frac{x^{k+1}}{k+1} \Big|_{-1}^1 = \begin{cases} \frac{2}{k+1} & \text{pentru } k \text{ par} \\ 0 & \text{pentru } k \text{ impar} \end{cases} \quad (8)$$

Să construim de exemplu polinomul Legendre de gradul 3. Sistemul (4) devine:

$$\begin{cases} c_0 a_0 + c_1 a_1 + c_2 a_2 = -c_3 \\ c_1 a_0 + c_2 a_1 + c_3 a_2 = -c_4 \\ c_2 a_0 + c_3 a_1 + c_4 a_2 = -c_5 \end{cases} \quad (9)$$

Înlocuind (8) în (9) rezultă

$$\begin{cases} 2a_0 + \frac{2}{3}a_2 = 0 \\ \frac{2}{3}a_1 = -\frac{2}{5} \\ \frac{2}{3}a_0 + \frac{2}{5}a_2 = 0 \end{cases}$$

care admite soluția:  $a_0 = a_2 = 0$ ;  $a_1 = -\frac{3}{5}$ . Așadar,  $P_3^*(x) = x^3 - \frac{3}{5}x$ .

Primele șase polinoame Legendre sunt:

$$1; x; x^2 - \frac{1}{3}; x^3 - \frac{3}{5}x; x^4 - \frac{6}{7}x^2 + \frac{3}{35}; x^5 - \frac{10}{9}x^3 + \frac{15}{63}x$$

Fie  $[a, b] = [-1, 1]$  și  $w(x) = \frac{1}{\sqrt{1-x^2}}$ .

$$c_0 = \int_{-1}^1 \frac{1}{\sqrt{1-x^2}} dx = \arcsin x \Big|_{-1}^1 = \pi; \quad c_1 = \int_{-1}^1 \frac{x}{\sqrt{1-x^2}} dx = 0.$$

Pentru calculul coeficienților  $c_k = \int_{-1}^1 \frac{x^k}{\sqrt{1-x^2}} dx$ , facem schimbarea de

variabilă  $x = \sin t$  și obținem

$$c_k = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \sin^k x dx = \frac{k-1}{k} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \sin^{k-2} x dx = \frac{k-1}{k} c_{k-2}.$$

Rezultă  $c_k = 0$  pentru orice  $k$  impar și

$$c_2 = \frac{\pi}{2}; \quad c_4 = \frac{3}{4} \cdot \frac{\pi}{2}; \quad c_6 = \frac{5}{6} \cdot \frac{3}{4} \cdot \frac{\pi}{2} \quad \text{etc.}$$

Să determinăm polinomul Cebîșev de speța I,  $T_3^*$ . Sistemul (4) devine

$$\begin{cases} \pi a_0 + \frac{\pi}{2} a_2 = 0 \\ \frac{\pi}{2} a_1 = -\frac{3\pi}{8} \\ \frac{\pi}{2} a_0 + \frac{\pi}{4} a_2 = 0 \end{cases}$$

care admite soluția:  $a_0 = a_2 = 0$ ,  $a_1 = -\frac{3}{4}$ . Așadar  $T_3^* = x^3 - \frac{3}{4}x$ .

În Capitolul IV, §3 am arătat că polinomul Cebîșev de gradul trei este:  $T_3(x) = 4x^3 - 3x$ . După cum am precizat în Teorema 1, polinomul ortogonal se determină în afara unui factor constant de multiplicare.

Observăm că avem  $T_3 = 4T_3^*$  și în general  $T_n = 2^{n-1}T_n^*$ ,  $n \in \mathbb{N}$ .

**Teorema 2.** Fie  $P_n^*$  un polinom de gradul  $n$  care este ortogonal pe intervalul  $[a, b]$  în raport cu ponderea  $w$ , pe orice polinom de grad mai mic ca  $n$ . Atunci zerourile polinomului  $P_n^*$  sunt reale, simple și aparțin intervalului  $[a, b]$ .

**Demonstrație.**

Fie  $x_i$ ,  $i = \overline{1, k}$ , zerourile polinomului  $P_n^*$  și fie  $m_i$  ordinul de multiplicitate al zeroului  $x_i$ . Rezultă

$$P_n^*(x) = (x - x_1)^{m_1} \cdot \dots \cdot (x - x_k)^{m_k},$$

unde  $m_1 + m_2 + \dots + m_k = n$ . Presupunem că numerotarea zerourilor s-a făcut astfel încât  $x_1, \dots, x_l$ ,  $l \leq k$  sunt zerouri reale, aparțin intervalului  $[a, b]$  și au ordinele de multiplicitate impare.

Dacă  $l = n$ , teorema este demonstrată.

Să presupunem că  $l < n$ . Considerăm atunci polinomul

$$Q_l(x) = (x - x_1)(x - x_2) \dots (x - x_l).$$

(Dacă  $l = 0$ , atunci alegem  $Q_0 = 1$ ). Rezultă că polinomul produs  $P_n^*(x)Q_l(x)$  păstrează semn constant pe  $[a, b]$ , deci

$$\int_a^b w(x)P_n^*(x)Q_l(x)dx \neq 0,$$

ceea ce contrazice faptul că  $P_n^*$  este ortogonal pe  $Q_l$ .  $\square$

Prin formulă de cuadratură Gauss se înțelege orice formulă

$$\int_a^b w(x)f(x)dx = A_1f(x_1) + A_2f(x_2) + \dots + A_nf(x_n) + R(f), \quad (1)$$

unde nodurile  $x_1, x_2, \dots, x_n$  sunt zerourile polinomului  $P_n^*$  care este ortogonal pe  $[a, b]$ , în raport cu ponderea  $w$ , pe orice plinom de grad mai mic ca  $n$ . Vom avea astfel formule de cuadratură de tip Gauss–Legendre, Gauss–Cebîșev, Gauss–Hermite etc.

**Teorema 3.** Orice formulă de cuadratură de tip Gauss are gradul de exactitate  $2n-1$ .

**Demonstrație.** Din Teorema 1 din introducerea în Capitolul V, știm că formula (1) este exactă pentru polinoame de grad mai mic ca  $n$ . Fie  $Q_m$  un polinom de gradul  $m$ , unde  $m \in \{n, n+1, \dots, (2n-1)\}$ . Din teorema împărțirii avem

$$Q_m(x) = Q_{m-n}(x) \cdot P_n^*(x) + R_{n-1}(x),$$

unde  $R_{n-1}$  este un polinom de gradul  $n-1$ . Fie  $x_k, k = \overline{1, n}$ , zerourile polinomului  $P_n^*$ . Evident, rezultă

$$Q_m(x_k) = R_{n-1}(x_k), \quad k = \overline{1, n}. \quad (10)$$

În continuare avem:

$$\int_a^b w(x) Q_m(x) dx = \int_a^b w(x) P_n^*(x) Q_{m-n}(x) dx + \int_a^b w(x) R_{n-1}(x) dx.$$

Deoarece  $P_n^*$  este ortogonal pe  $Q_{m-n}$  rezultă:

$$\int_a^b w(x) Q_m(x) dx = \int_a^b w(x) R_{n-1}(x) dx.$$

Ținând seama că formula (1) este exactă pentru  $R_{n-1}$  obținem

$$\int_a^b w(x) Q_m(x) dx = A_1 R_{n-1}(x_1) + \dots + A_n R_{n-1}(x_n).$$

În sfârșit, ținând seama și de (10) rezultă

$$\int_a^b w(x) Q_m(x) dx = A_1 Q_m(x_1) + \dots + A_n Q_m(x_n),$$

deci formula (1) este exactă pentru orice polinom de grad mai mic ca  $2n$ .

Pe de altă parte dacă notăm cu  $g = (P_n^*)^2$ , atunci  $g$  este un polinom de grad  $2n$  și restul

$$R(g) = \int_a^b w(x) g(x) dx - \sum_{i=1}^n A_i g(x_i) = \int_a^b w(x) g(x) dx > 0.$$

Așadar, formula (1) nu este exactă pentru  $g$ , deci gradul de precizie al acestei formule este  $2n-1$ .  $\square$

**Observația 1.** Orice formulă de cuadratură care are gradul de precizie  $(2n-1)$  este o formulă Gauss.

Într-adevăr, considerăm formula

$$\int_a^b w(x) f(x) dx = A_1 f(z_1) + A_2 f(z_2) + \dots + A_n f(z_n) + R(f) \quad (11)$$

cu gradul de exactitate  $(2n-1)$ . Notăm cu

$$U_n(x) = (x-z_1) \dots (x-z_n).$$

Dacă  $Q$  este un polinom de grad mai mic ca  $n$ , atunci polinomul produs  $U_n Q$ , are gradul cel mult  $2n-1$  și formula (11) este exactă pentru acest polinom. Rezultă

$$\int_a^b w(x) U_n(x) Q(x) dx = \sum_{i=1}^n A_i U_n(z_i) Q(z_i) = 0.$$

Rezultă că  $U_n$  este un polinom de gradul  $n$ , care este ortogonal, în raport cu ponderea  $w$ , pe orice polinom de grad cel mult  $(n-1)$ . Din Teorema 1 rezultă că  $U_n = c P_n^*$  și deci că  $z_1, \dots, z_n$  sunt zerourile polinomului  $P_n^*$ . Așadar, formula (11) este o formulă Gauss.

**Observația 2.** Coeficienții  $A_k$  din formula Gauss sunt pozitivi.

Într-adevăr, fie  $x_k, k = \overline{1, n}$ , zerourile polinomului  $P_n^*$ , și fie

$$Q_j(x) = \prod_{\substack{i=1 \\ i \neq j}}^n (x - x_i)^2.$$

Evident,  $\text{grad} Q_j = 2n-2$  și  $Q_j(x_i) = 0$  dacă  $i \neq j$ . Deoarece  $Q_j \geq 0$  și formula (1) este exactă pentru  $Q_j$  rezultă

$$0 < \int_a^b w(x) Q_j(x) dx = \sum_{i=1}^n A_i Q_j(x_i) = A_j Q_j(x_j),$$

deci  $A_j > 0$ .

În particular, obținem

$$A_j = \frac{\int_a^b w(x) Q_j(x) dx}{Q_j(x_j)}, \quad j = \overline{1, n}. \quad (12)$$

Formula (12) permite calculul coeficienților din formula Gauss.

**Teorema 4.** Fie  $x_i^{(n)}, i = \overline{1, n}$ , zerourile polinomului  $P_n^*$  și fie  $A_i^{(n)}, i = \overline{1, n}$ , coeficienții din formula Gauss corespunzătoare.

Dacă  $f: [a, b] \rightarrow \mathbb{R}$  este continuă și  $I_n = \sum_{i=1}^n A_i^{(n)} f(x_i^{(n)})$ , atunci

$$\lim_{n \rightarrow \infty} I_n = \int_a^b w(x) f(x) dx.$$

**Demonstrație.** Din teorema Weierstrass (Capitolul IV, Teorema 4) rezultă că pentru orice  $\varepsilon > 0$  există un polinom  $P_\varepsilon$  astfel încât

$$|f(x) - P_\varepsilon(x)| < \varepsilon \text{ pentru orice } x \in [a, b].$$

Fie  $n > \text{grad} P_\varepsilon$ . Atunci formula Gauss este exactă pentru  $P_\varepsilon$  și avem:

$$\begin{aligned}
& \left| \int_a^b w(x)f(x)dx - I_n \right| = \left| \int_a^b w(x)f(x)dx - \sum_{i=1}^n A_i^{(n)} f(x_i^{(n)}) \right| \leq \\
& \leq \left| \int_a^b w(x)[f(x) - P_\varepsilon(x)]dx + \sum_{i=1}^n A_i^{(n)} [P_\varepsilon(x_i^{(n)}) - f(x_i^{(n)})] \right| \leq \\
& \leq \int_a^b w(x)|f(x) - P_\varepsilon(x)|dx + \sum_{i=1}^n A_i^{(n)} |P_\varepsilon(x_i^{(n)}) - f(x_i^{(n)})| < \\
& < \varepsilon \left( \int_a^b w(x)dx + \sum_{i=1}^n A_i^{(n)} \right) = 2\varepsilon \int_a^b w(x)dx .
\end{aligned}$$

Am folosit faptul că  $\int_a^b w(x)dx = \sum_{i=1}^n A_i^{(n)}$ .

Așadar,

$$\lim_{n \rightarrow \infty} I_n = \int_a^b w(x)f(x)dx$$

și cu aceasta teorema este demonstrată.  $\square$

**Exemplu.** Formula Gauss–Legendre cu trei noduri

Polinomul Legendre de gradul trei este

$$P_3^*(x) = x^3 - \frac{3}{5}x$$

și are zerourile

$$x_1 = -\sqrt{\frac{3}{5}}; \quad x_2 = 0; \quad x_3 = \sqrt{\frac{3}{5}}.$$

Vom avea

$$\int_{-1}^1 f(x)dx = A_1 f\left(-\sqrt{\frac{3}{5}}\right) + A_2 f(0) + A_3 f\left(\sqrt{\frac{3}{5}}\right) + R(f) . \quad (13)$$

Punând condiția ca formula (13) să fie exactă pentru 1, x și x<sup>2</sup> obținem sistemul:

$$\begin{cases} A_1 + A_2 + A_3 = 2 \\ -\sqrt{\frac{3}{5}}A_1 + \sqrt{\frac{3}{5}}A_3 = 0 \\ \frac{3}{5}A_1 + \frac{3}{5}A_3 = \frac{2}{3} \end{cases}$$

care admite soluția:  $A_0 = A_3 = \frac{5}{9}$  și  $A_1 = \frac{8}{9}$ .

Formula Gauss–Legendre de ordinul trei este

$$\int_{-1}^1 f(x) dx = \frac{5}{9} f\left(-\sqrt{\frac{3}{5}}\right) + \frac{8}{9} f(0) + \frac{5}{9} f\left(\sqrt{\frac{3}{5}}\right) + R(f).$$

### §5.3. Integrarea numerică a integralelor duble

Fie  $f: D \subset \mathbb{R}^2 \rightarrow \mathbb{R}$  o funcție continuă, unde  $D = [a, b] \times [c, d]$  este un dreptunghi. Atunci:

$$\iint_D f(x, y) dx dy = \int_a^b \left( \int_c^d f(x, y) dy \right) dx \quad (1)$$

Pentru fiecare integrală simplă putem aplica o formulă de integrare numerică. De exemplu, dacă aplicăm formula trapezelor obținem

$$\begin{aligned} \iint_D f(x, y) dx dy &\cong \int_a^b \frac{d-c}{2} [f(x, c) + f(x, d)] dx = \\ &= \frac{b-a}{2} \frac{d-c}{2} [f(a, c) + f(a, d) + f(b, c) + f(b, d)] \end{aligned}$$

Așadar, *formula trapezelor* pentru integrala (1) este

$$\iint_D f(x, y) dx dy = \frac{(b-a)(d-c)}{4} [f(a, c) + f(a, d) + f(b, c) + f(b, d)] + R(f). \quad (2)$$

În mod asemănător, *formula Simpson* va fi

$$\begin{aligned} \iint_D f(x, y) dx dy &= \frac{(b-a)(d-c)}{36} [f(a, c) + f(a, d) + f(b, c) + f(b, d) + \\ &4[f(a, y_1) + f(b, y_1) + f(x_1, c) + f(x_1, d) + 16f(x_1, y_1)] + \tilde{R}(f) \end{aligned} \quad (3)$$

$$\text{unde } x_1 = \frac{a+b}{2}, \quad y_1 = \frac{c+d}{2}.$$

Pentru o mai bună aproximare a integralei se folosesc formulele repetate.

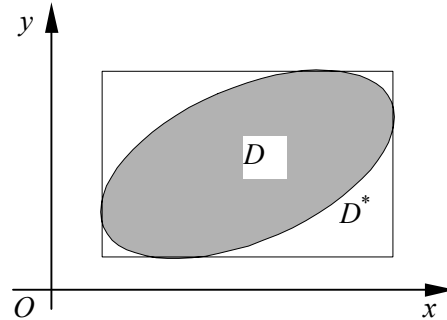
Dacă domeniul de integrare nu este un dreptunghi, atunci se construiește un dreptunghi  $D^*$ , cu laturile paralele cu axele de coordonate și care include dreptunghiul  $D$ . Considerăm funcția auxiliară

$$f^*(x, y) = \begin{cases} f(x, y) & \text{dacă } (x, y) \in D \\ 0 & \text{dacă } (x, y) \in D^* \setminus D \end{cases}$$

Integrala pe  $D$  se va aproxima cu

$$\iint_D f(x, y) dx dy \cong \iint_{D^*} f^*(x, y) dx dy,$$

iar ultima integrală se calculează cu una din formule (2) sau (3).



#### §5.4. Diferențe divizate. Polinomul de interpolare al lui Newton

Fie  $P_n(x; x_0, \dots, x_n)$  polinomul Lagrange care interpoalează funcția  $f: [a, b] \rightarrow \mathbb{R}$  în nodurile  $x_0, \dots, x_n$  și fie

$$Q(x) = P_n(x; x_0, \dots, x_{n-1}, x_n) - P_{n-1}(x; x_0, \dots, x_{n-1}) \quad (1)$$

Evident,  $Q$  este un polinom de gradul  $n$ , care se anulează în nodurile  $x_0, x_1, \dots, x_{n-1}$ , deoarece  $Q(x_i) = f(x_i) - f(x_i) = 0$ , pentru orice  $i = \overline{0, n-1}$ . Rezultă că polinomul  $Q$  este de forma:

$$Q(x) = a(x-x_0) \dots (x-x_{n-1}). \quad (2)$$

Coeficientul  $a$  se numește *diferența divizată de ordinul  $n$*  corespunzătoare nodurilor  $x_0, x_1, \dots, x_n$  și funcției  $f$  și se notează cu  $f[x_0, x_1, \dots, x_n]$ .

Așadar, avem:

$$Q(x) = (x-x_0) \dots (x-x_{n-1}) f[x_0, x_1, \dots, x_n] \quad (3)$$

Din (1) rezultă, pe de o parte, că  $f[x_0, x_1, \dots, x_n]$  este coeficientul lui  $x^n$  în polinomul lui Lagrange  $P_n(x; x_0, x_1, \dots, x_n)$ , iar pe de altă parte că avem relația:

$$P_n(x; x_0, x_1, \dots, x_n) = P_{n-1}(x; x_0, x_1, \dots, x_{n-1}) + (x-x_0) \dots (x-x_{n-1}) f[x_0, \dots, x_n] \quad (4)$$

Particularizându-l pe  $n$  obținem:

$$P_0(x; x_0) = f(x_0)$$

$$P_1(x; x_0, x_1) = f(x_0) + (x-x_0)f[x_0, x_1]$$

$$P_2(x; x_0, x_1, x_2) = f(x_0) + (x-x_0)f[x_0, x_1] + (x-x_0)(x-x_1)f[x_0, x_1, x_2]$$

.....

$$P_n(x) = f(x_0) + (x-x_0)f[x_0, x_1] + \dots + (x-x_0)\dots(x-x_{n-1})f[x_0, x_1, \dots, x_n] \quad (5).$$

Forma (5) a polinomului de interpolare poartă numele de *polinomul de interpolare al lui Newton*.

În continuare prezentăm principalele proprietăți ale diferențelor divizate.

1) Diferența divizată de ordinul  $n$  este invariantă la permutarea nodurilor, adică:

$$f[x_0, x_1, \dots, x_n] = f[x_{i_0}, x_{i_1}, \dots, x_{i_n}].$$

Într-adevăr știm că polinomul de interpolare al lui Lagrange are forma:

$$P_n(x; x_0, \dots, x_n) = \frac{(x-x_1)\dots(x-x_n)}{(x_0-x_1)\dots(x_0-x_n)}f(x_0) + \dots + \frac{(x-x_0)\dots(x-x_{n-1})}{(x_n-x_0)\dots(x_n-x_{n-1})}f(x_n) \quad (6)$$

Egalând coeficientul lui  $x^n$  din (3) și (6) obținem:

$$f[x_0, x_1, \dots, x_n] = \frac{f(x_0)}{(x_0-x_1)\dots(x_0-x_n)} + \dots + \frac{f(x_n)}{(x_n-x_0)\dots(x_n-x_{n-1})} \quad (7)$$

Cum expresia din membrul drept al relației (7) este simetrică în raport cu cu nodurile  $x_0, x_1, \dots, x_n$ , rezultă că diferența divizată dordinul  $n$ ,  $f[x_0, x_1, \dots, x_n]$  este invariantă în raport cu permutarea nodurilor.

$$2) f[x_0, \dots, x_n] = \frac{f[x_1, \dots, x_n] - f[x_0, \dots, x_{n-1}]}{x_n - x_0}$$

Pentru a demonstra această proprietate observăm pentru început că polinomul lui Lagrange verifică următoarea relație:

$$P_n(x; x_0, \dots, x_n) = \frac{(x-x_0)P_{n-1}(x; x_1, \dots, x_n) - (x-x_n)P_{n-1}(x; x_0, x_1, \dots, x_{n-1})}{x_n - x_0} \quad (8)$$

Într-adevăr, dacă notăm cu  $R(x)$  membrul drept al relației (6) obținem:

$$R(x_0) = -\frac{(x_0-x_n)}{x_n-x_0}f(x_0) = f(x_0)$$

$$R(x_n) = -\frac{(x_n - x_0)}{x_n - x_0} f(x_n) = f(x_n)$$

Pentru nodurile  $x_i, i = \overline{1, n-1}$  avem

$$R(x_i) = -\frac{(x_i - x_0)f(x_i) - (x_i - x_n)f(x_i)}{x_n - x_0} = f(x_i)$$

Așadar,  $R(x_i) = f(x_i), i = \overline{0, n}$ , deci  $R(x) \equiv P_n(x; x_0, \dots, x_n)$  conform unicității polinomului de interpolare Lagrange. Egalând coeficientul lui  $x^n$  din membrul stâng al relației (8) cu coeficientul lui  $x^n$  din membrul drept al acestei relații obținem  $f[x_0, \dots, x_n] = \frac{f[x_1, \dots, x_n] - f[x_0, \dots, x_{n-1}]}{x_n - x_0}$ .

3) Dacă  $f$  este de clasă  $C^{n+1}$ , atunci pentru orice  $t \in [a, b]$ ,  $t \neq x_i$ ,  $i = \overline{0, n}$  avem  $f[x_0, x_1, \dots, x_n, t] = \frac{f^{(n+1)}(\xi_t)}{(n+1)!}$ , unde  $\xi_t \in [a, b]$ .

Într-adevăr, fie  $P_{n+1}(x) = P_{n+1}(x; x_0, \dots, x_n, t)$  polinomul Lagrange care interpoalează funcția  $f$  în nodurile  $x_0, \dots, x_n, t$ . Atunci avem

$$P_{n+1}(x) = P_n(x) + (x-x_0) \dots (x-x_n) f[x_0, x_1, \dots, x_n, t].$$

Deoarece  $P_{n+1}(t) = f(t)$  rezultă că eroarea în punctul  $t$  este

$$E(f; t) = f(t) - P_n(t) = (t-x_0) \dots (t-x_n) f[x_0, x_1, \dots, x_n, t]. \quad (9)$$

Pe de altă parte din Teorema 2, §4.1 știm că dacă  $f$  este de clasă  $C^{n+1}$ , atunci există  $\xi_t \in [a, b]$  astfel încât

$$E(f; t) = \frac{f^{(n+1)}(\xi_t)}{(n+1)!} (t-x_0) \dots (t-x_n) \quad (10)$$

Din (9) și (10) rezultă

$$f[x_0, x_1, \dots, x_n, t] = \frac{f^{(n+1)}(\xi_t)}{(n+1)!}, \text{ unde } \xi_t \in [a, b]. \quad (11)$$

4) **Teorema 1 (Hermite–Genocchi).** Fie  $x_0, x_1, \dots, x_n$ ,  $(n+1)$  puncte distincte din intervalul  $(a, b)$  și fie  $f \in C^{(n)}(a, b)$ . Atunci:

$$f[x_0, x_1, \dots, x_n] = \int_{T_n} \dots \int f^{(n)}(t_0 x_0 + t_1 x_1 + \dots + t_n x_n) dt_1 \dots dt_n$$

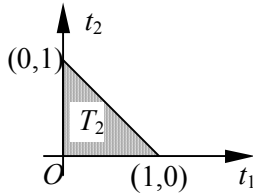
unde  $T_n = \left\{ (t_1, \dots, t_n) \in \mathbb{R}^n \mid t_i \geq 0, i=1, \dots, n, \sum_{i=1}^n t_i \leq 1 \right\}$ , iar  $t_0 = 1 - \sum_{i=1}^n t_i$ .

**Demonstrație.** Fie  $T_1 = [0, 1]$  iar  $t_0 = 1 - t_1$ .

$$\begin{aligned} \int_0^1 f'(t_0 x_0 + t_1 x_1) dt_1 &= \int_0^1 f'[x_0 + t_1(x_1 - x_0)] dt_1 = \frac{1}{x_1 - x_0} f[x_0 + t_1(x_1 - x_0)] \Big|_0^1 = \\ &= \frac{f(x_1) - f(x_0)}{x_1 - x_0} = f[x_0, x_1] \end{aligned}$$

$T_2 = \{ (t_1, t_2) \mid t_1 + t_2 \leq 1, t_1 \geq 0, t_2 \geq 0 \}$ , iar  $t_0 = 1 - t_1 - t_2$ .

$$\begin{aligned} \iint_{T_2} f''(x_0 t_0 + x_1 t_1 + x_2 t_2) dt_1 dt_2 &= \int_0^1 \left( \int_0^{1-t_1} f''(x_0 + t_1(x_1 - x_0) + t_2(x_2 - x_0)) dt_2 \right) dt_1 = \\ &= \int_0^1 \left( \frac{1}{x_2 - x_0} f'(x_0 + t_1(x_1 - x_0) + t_2(x_2 - x_0)) \Big|_0^{1-t_1} \right) dt_1 = \\ &= \frac{1}{x_2 - x_0} \left[ \int_0^1 f'(x_2 + t_1(x_1 - x_2)) dt_1 - \int_0^1 f'(x_0 + t_1(x_1 - x_0)) dt_1 \right] = \\ &= \frac{1}{x_2 - x_0} (f[x_1, x_2] - f[x_0, x_1]) = f[x_0, x_1, x_2] \end{aligned}$$



În continuare demonstrația se face prin inducție matematică.

Trecerea de la  $T_n$  la  $T_{n+1}$  este asemănătoare cu trecerea de la  $T_1$  la  $T_2$ .  $\square$

**Corolarul 1.** Dacă  $f \in C^{n+2}(a, b)$ . Atunci există

$$\frac{d}{dx} f[x_0, \dots, x_n, x] = f[x_0, \dots, x_n, x, x].$$

**Demonstrație.** Din Teorema 1 rezultă

$$f[x_0, x_1, \dots, x_n, x] = \int_{T_{n+1}} \dots \int f^{(n+1)}(t_0 x_0 + t_1 x_1 + \dots + t_{n+1} x) dt_1 \dots dt_n dt_{n+1} \quad (12)$$

Membrul drept este o integrală cu parametru și anume cu parametrul  $x$ . Deoarece  $f^{(n+1)}$  este de clasă  $C^1$  rezultă că această integrală este derivabilă, deci că există  $\frac{d}{dx} f[x_0, \dots, x_n, x]$ . Mai departe, ținând seama de (7) și (9) rezultă:

$$\begin{aligned} \frac{d}{dx} f[x_0, \dots, x_n, x] &= \lim_{h \rightarrow 0} \frac{f[x_0, \dots, x_n, x+h] - f[x_0, \dots, x_n, x]}{h} = \\ &= \lim_{h \rightarrow 0} \frac{f[x_0, \dots, x_n, x+h] - f[x, x_0, \dots, x_n]}{h} = \\ &= \lim_{h \rightarrow 0} f[x, x_0, \dots, x_n, x+h] = f[x, x_0, \dots, x_n, x] = \\ &= f[x_0, \dots, x_n, x, x] \end{aligned}$$

□

### §5.5. Derivarea numerică

Aproximarea numerică a derivatelor se folosește de regulă în două situații. Prima situație se referă la calculul derivatelor unei funcții dată printr-un tabel de valori. A doua situație se referă la aproximarea derivatelor în cadrul metodelor numerice de rezolvare a ecuațiilor diferențiale sau cu derivate parțiale.

O cale firească de abordare a derivării numerice, este aceea de a aproxima derivata funcției  $f$  prin derivata polinomului Lagrange  $P_n(x; x_0, \dots, x_n)$  care interpolează funcția  $f$  în nodurile  $x_0, x_1, \dots, x_n$ . În continuare notăm derivata numerică a funcției  $f$  cu  $D_h f$  și o definim prin:

$$D_h f(x) \stackrel{\text{def}}{=} P'_n(x; x_0, \dots, x_n) \quad (1)$$

Așadar folosim aproximarea  $f'(x) \approx D_h f(x)$ .

Pentru  $n=1$  avem  $P_1(x; x_0, x_1) = f(x_0) + (x - x_0)f[x_0, x_1]$  deci

$$f'(x) \approx P'_1(x; x_0, x_1) = f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0} = \frac{f(x_1) - f(x_0)}{h}, \forall x$$

Așadar, pentru două noduri avem aproximarea:

$$f'(x_0) \approx \frac{f(x_1) - f(x_0)}{h} \quad (2)$$

Pentru  $n = 2$  avem

$$P_2(x; x_0, x_1, x_2) = f(x_0) + (x - x_0)f[x_0, x_1] + (x - x_0)(x - x_1)f[x_0, x_1, x_2] \text{ deci}$$

$$P_2'(x; x_0, x_1, x_2) = f[x_0, x_1] + [2x - (x_0 + x_1)]f[x_0, x_1, x_2], \quad \forall x$$

Dacă presupunem în plus că nodurile sunt echidistante rezultă:

$$\begin{aligned} P_2'(x_0; x_0, x_1, x_2) &= f(x_0, x_1) + (x_0 - x_1)f[x_0, x_1, x_2] = \\ &= \frac{f(x_1) - f(x_0)}{x_1 - x_0} - (x_1 - x_0) \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0} = \\ &= \frac{f(x_1) - f(x_0)}{h} - h \frac{\frac{f(x_2) - f(x_1)}{h} - \frac{f(x_1) - f(x_0)}{h}}{2h} = \\ &= \frac{-3f(x_0) + 4f(x_1) - f(x_2)}{2h} \end{aligned}$$

Așadar, pentru trei noduri echidistante avem aproximarea:

$$f'(x_0) \approx \frac{-3f(x_0) + 4f(x_1) - f(x_2)}{2h} \quad (3)$$

În continuare ne propunem să evaluăm eroarea la derivarea numerică. Dacă notăm cu  $U_n(x) = (x - x_0) \dots (x - x_n)$ , atunci din relațiile (10) și (11) de la §5.4 rezultă  $E(f; x) = f(x) - P_n(x; x_0, \dots, x_n) = U_n(x)f[x_0, \dots, x_n, x]$ .

Ținând seama acum și de Corolarul 1 obținem:

$$f'(x) - D_h f(x) = U_n'(x)f[x_0, \dots, x_n, x] + U_n(x)f[x_0, \dots, x_n, x, x].$$

Aplicând din nou relația (10) din §5.4 rezultă

$$f'(x) - D_h f(x) = U_n'(x) \frac{f^{(n+1)}(\xi_x)}{(n+1)!} + U_n(x) \frac{f^{(n+2)}(\tilde{\xi}_x)}{(n+2)!} \quad (4)$$

Pentru  $n = 1$ ,  $U_1(x) = (x - x_0)(x - x_1)$  și  $U_1'(x) = 2x - (x_0 + x_1)$ . Așadar în acest caz

$$f'(x_0) - D_h f(x_0) = U_1'(x_0) \frac{f''(\xi_{x_0})}{2!} + 0 \frac{f'''(\tilde{\xi}_{x_0})}{3!} = -\frac{h}{2} f''(\xi_{x_0}) \quad (5)$$

În concluzie, în cazul a două noduri avem  $f'(x_0) \approx \frac{f(x_1) - f(x_0)}{h}$  și eroarea este dată de relația

$$f'(x_0) - \frac{f(x_1) - f(x_0)}{h} = -\frac{h}{2} f''(\xi_{x_0}) \quad (6)$$

unde  $\xi_{x_0} \in (x_0, x_1)$ .

În cazul a 3 noduri echidistante  $n = 2$  avem:

$$f'(x_0) - P'_2 f(x_0; x_0, x_1, x_2) = \frac{f'''(\xi_{x_0})}{3!} U'_2(x_0) + \frac{f^{iv}(\tilde{\xi}_{x_0})}{4!} U_2(x_0)$$

cum  $U_2(x) = (x - x_0)(x - x_1)(x - x_2)$  rezultă pe de o parte că  $U_2(x_0) = 0$  iar pe de altă parte că  $U'_2(x) = (x_0 - x_1)(x_0 - x_2) = 2h^2$ .

Așadar eroarea de aproximare a derivatei va fi:

$$f'(x_0) - P'_2 f(x_0; x_0, x_1, x_2) = \frac{h^2}{3} f'''(\xi_{x_0}), \text{ unde } \xi_{x_0} \in (x_0, x_2).$$

În concluzie, în cazul a 3 noduri echidistante derivata se aproximează cu expresia:

$$f'(x_0) \approx \frac{-3f(x_0) + 4f(x_1) - f(x_2)}{2h} = D_h f(x_0)$$

iar eroarea care se face este  $f'(x_0) - D_h f(x_0) = \frac{h^2}{3} f'''(\xi_{x_0})$  (7)

În sfârșit să vedem cum se poate aproxima derivata de ordinul 2.

În cazul a 3 noduri echidistante avem

$$\begin{aligned} f''(x_0) &= P''_2(x_0; x_0, x_1, x_2) = 2f[x_0, x_1, x_2] = \\ &= \frac{f(x_2) - 2f(x_1) + f(x_0)}{h^2} \end{aligned} \quad (8)$$

Se poate arăta că eroarea în acest caz este

$$f''(x_0) - P''_2(x_0) = -\frac{h^2}{12} f^{IV}(\xi_{x_0}), \text{ unde } \xi_{x_0} \in (x_0, x_2) \quad (9)$$

*Derivarea numerică în MATLAB.*

MATLAB permite aproximarea derivatei numerice a unei funcții folosind diferențele divizate, prin intermediul funcției *diff*.

**Exemplu 1.**

Să se aproximeze derivata funcției  $f(x) = \ln(x^4+x^2+1)$  pe intervalul  $[0.5, 2.5]$  în puncte echidistante ( $h = 0.1$ ) folosind MATLAB. Secvența care realizează această aproximare este:

```
% Calculul derivatei numerice
x = 0.5:0.1:2.5; % punctele in care se face aproximarea
f(x) = log(x.^4+x.^2+1); % functia a carei derivata se doreste
disp('Valorile derivatei');
df = diff(log(x.^4+x.^2+1))./diff(x)
```

Se va afișa:

Valorile derivatei

```
1.2657  1.4967  1.6947  1.8499  1.9597  2.0270  2.0579  2.0600  2.0406
2.0060  1.9612  1.9100  1.8552  1.7989  1.7423  1.6866  1.6323  1.5798
1.5293  1.4809
```

Pentru calculul derivatei folosind diferențele centrate se poate scrie secvența MATLAB:

```
% Calculul derivatei numerice folosind diferentele divizate
x=0.55:0.1:2.5; % punctele in care se face aproximare
g=log(x.^4+x.^2+1); % functia a carei derivata se doreste
dg=g(3:length(g))-g(1:length(g)-2);
dx=x(3:length(x))-x(1:length(x)-2);
disp('Valorile derivatei');
dy=dg./dx % derivata in punctele x=0.6 , 0.7 pana la 2.4
```

Se afișează rezultatele:

Valorile derivatei

1.3812 1.5957 1.7723 1.9048 1.9934 2.0424 2.0589 2.0503 2.0233  
 1.9836 1.99356 1.8826 1.8270 1.7706 1.7145 1.6595 1.6060 1.5545  
 1.5051

Pentru a folosi polinomul de interpolare Newton pentru calculul derivatelor de ordinul întâi și doi pentru funcția de mai sus, se poate scrie secvența MATLAB:

```
% Calculul derivatelor de ordinul intai si doi
% folosind polinomul de interpolare Newton in x(1)
x=0.5:0.1:1.3;
h=x(2)-x(1);
y=log(x.^4+x.^2+1); % functia ale carei derivate se aproximate
se calculeaza
d1y=diff(y);
d2y=diff(d1y);
d3y=diff(d2y);
d4y=diff(d3y);
d1f=(1/h)*(d1y(1)-d2y(1)/2+d3y(1)/3-d4y(1)/4),
d2f=(1/h^2)*(d2y(1)-d3y(1)+(11/12)*d4y(1));
disp('Derivata de ordinul intai in x(1)');
disp(d1f);
disp('Derivata de ordinul doi in x(1)');
disp(d2f);
```

Se afișează rezultatele:

Derivata de ordinul intai in x(1)

1.1416

Derivata de ordinul doi in x(1)

2.5519

**Exemplul 2.** Fie funcția dată prin tabelul de valori:

$x_i$	0	2
$f(x_i)$	1	5

Să se calculeze  $f'(0)$  folosind derivata polinomului de interpolare al lui

Lagrange.

Aproximând funcția cu polinomul de interpolare al lui Lagrange, conform (2) rezultă

$$f'(x_0) \approx \frac{f(x_1) - f(x_0)}{h}, \quad f'(x_0) \approx \frac{5-1}{2} = 2.$$

**Exemplul 3.** Fie funcția dată prin tabelul de valori:

$x_i$	2	4	6
$f(x_i)$	3	11	27

Să se calculeze  $f'(2)$  și  $f''(2)$ .

Aproximând funcția cu polinomul de interpolare al lui Lagrange, conform

(3) rezultă

$$f'(x_0) \approx \frac{-3f(x_0) + 4f(x_1) - f(x_2)}{2h}, \quad f'(2) = \frac{-3 \cdot 3 + 4 \cdot 11 - 27}{2 \cdot 2} = 2.$$

De asemenea conform (8)

$$f''(x_0) = \frac{f(x_2) - 2f(x_1) + f(x_0)}{h^2}, \quad f''(2) = \frac{27 - 2 \cdot 11 + 3}{4} = 2.$$

**Exerciții**

Folosind metoda trapezelor să se calculeze valoarea aproximativă a următoarelor integrale:

$$1. \int_{\frac{\pi}{12}}^{\frac{\pi}{2}} \frac{\sin x}{\sqrt{x}} dx, \text{ considerând } n = 5 \text{ subintervale egale.}$$

$$R. \quad h = \frac{\pi}{12}, \quad x_i = \frac{\pi}{12} + i \cdot \frac{\pi}{12}, \quad i = \overline{1,4}, \quad a = \frac{\pi}{12}, \quad b = \frac{\pi}{2}, \quad I = 1.003.$$

$$2. \int_{\frac{\pi}{10}}^{\frac{\pi}{2}} \frac{\cos x}{\sqrt{x}} dx, \text{ considerând } n = 4 \text{ subintervale egale.}$$

$$R. \quad h = \frac{\pi}{10}, \quad x_i = \frac{\pi}{10} + i \cdot \frac{\pi}{10}, \quad i = \overline{1,3}, \quad a = \frac{\pi}{10}, \quad b = \frac{\pi}{2}, \quad I = 0.86398.$$

$$3. \int_0^3 e^{\frac{x^2}{2}} dx, \text{ considerând } n = 4 \text{ subintervale egale.}$$

$$R. \quad h = \frac{3}{4} = 0.75, \quad x_i = 3 \cdot i, \quad i = \overline{1,3}, \quad a = 0, \quad b = 3, \quad I = 2.24845.$$

Pentru fiecare din cele trei exerciții de mai sus să se calculeze valoarea aproximativă a integralei dublând valoarea lui  $n$ .

Să se calculeze valoarea aproximativă a următoarelor integrale folosind metoda lui Simpson, considerând  $m$ , numărul de subintervale egale, specificat în fiecare caz în parte:

$$5. \int_{\frac{\pi}{10}}^{\frac{\pi}{2}} \frac{\cos x}{\sqrt{x}} dx \quad (m = 8)$$

$$R. \quad n = 4, \quad h = \frac{\pi}{20}, \quad x_i = \frac{\pi}{10} + i \cdot \frac{\pi}{20}, \quad i = \overline{1,7}, \quad a = \frac{\pi}{10}, \quad b = \frac{\pi}{2}, \\ I = 0.8452.$$

6. 
$$\int_{\frac{\pi}{12}}^{\frac{\pi}{2}} \frac{\sin x}{\sqrt{x}} dx \quad (m = 8)$$

R.  $n = 4, h = \frac{5\pi}{96}, x_i = \frac{\pi}{12} + i \cdot \frac{\pi}{96}, i = \overline{1,7}, a = \frac{\pi}{12}, b = \frac{\pi}{2},$   
 $I = 1.00966.$

7. 
$$\int_0^3 e^{-\frac{x^2}{2}} dx, \quad (m = 8)$$

R.  $h = \frac{3}{8} = 0.375, x_i = \frac{3}{8} \cdot i, i = \overline{1,7}, a = 0, b = 3, I = 2.24991.$

8. Fie funcția dată prin tabelul de valori:

$x_i$	0	2
$f(x_i)$	-1	3

Să se calculeze  $f'(0)$  folosind derivata polinomului de interpolare al lui

Lagrange.

R.  $f'(0) = 2$

9. Fie funcția dată prin tabelul de valori:

$x_i$	0	1	2
$f(x_i)$	1	4	15

Să se calculeze  $f'(0)$  și  $f''(0)$  folosind derivata polinomului de interpolare al lui Lagrange.

R.  $f'(0) = -1, f''(0) = 8$

10. Fie funcția dată prin tabelul de valori:

$x_i$	0	2	4
$f(x_i)$	1	9	65

Să se calculeze  $f'(0)$  și  $f''(0)$  folosind derivata polinomului de interpolare al lui .

R.  $f'(0) = -8, f''(0) = 12$

11. Folosind formula Gauss–Legendre de ordinul 4, să se calculeze valoarea

aproximativă a integralei  $\int_0^3 e^{-\frac{x^2}{2}} dx$ .

R. În general pentru calculul aproximativ al integralelor  $\int_a^b f(x)dx$  se face

schimbarea de variabilă  $x = \frac{b-a}{2}t + \frac{b+a}{2}$ , pentru a avea limitele de integrare  $-1$  și  $1$  și astfel se obține formula

$$\int_a^b f(x)dx = \frac{b-a}{2} \sum_{i=1}^n A_i f(x_i) \quad (2)$$

Se ajunge la integrala  $\frac{3}{2} \int_{-1}^1 e^{-\frac{1}{2}\left(\frac{3t+3}{2}\right)^2} dt$  căreia i se poate aplica formula

Gauss–Legendre de ordinul 4. Polinomul Legendre de gradul 4 este

$$P_4(x) = x^4 - \frac{6}{7}x^2 + \frac{3}{35}, \text{ are rădăcinile } x_1 = -0.8611, x_2 = -0.34, x_3 = 0.34,$$

$x_4 = 0.8611$ . Coeficienții  $A_i$  formula Gauss–Legendre de ordinul 4 sunt:

$$A_1 = 0.34785 = A_4, A_2 = 0.65215 = A_3, I = 1.25018.$$

12. Folosind formula Gauss–Legendre de ordinul 5, să se calculeze valoarea

aproximativă a integralei  $\int_0^2 \cos x^2 dx$ .

R. Se face schimbarea de variabilă  $x = t+1$ , pentru a avea limitele de integrare  $-1$  și  $1$  și astfel se ajunge la integrala  $\int_{-1}^1 \cos(t+1)^2 dt$  căreia i se

poate aplica formula Gauss–Legendre de ordinul 5. Polinomul Legendre de gradul

$$5 \text{ este } P_5(x) = x^5 - \frac{10}{9}x^3 + \frac{15}{63}x, \text{ are rădăcinile}$$

$$x_1 = -0.90618, x_2 = -0.53847, x_3 = 0, x_4 = 0.53847, x_5 = 0.90618.$$

Coeficienții  $A_i$  formula Gauss–Legendre de ordinul 5 sunt:

$$A_1 = 0.23693 = A_5, A_2 = 0.47863 = A_4, A_3 = 0.56889, I = 0.46123.$$

13. Să se determine valoarea aproximativă a integralei  $\int_0^1 \sin x^2 dx$  folosind

formula Gauss–Legendre de ordinul 3.

R. Înlocuind  $x$  cu  $t+1$  integrala devine  $\int_{-1}^1 \sin(t+1)^2 dx$  și acestea i se poate aplica formula Gauss-Legendre de ordinul 3. Ținând seama de Exemplul din Capitolul 5 §2 se obține  $I=0.31028$ .

Să se calculeze valoarea aproximativă a următoarelor integrale duble:

14.  $\iint_D e^{-(x^2+y^2)} dx dy$ , mulțimea de integrare fiind precizată de

a)  $D = \{ (x,y) \in \mathbb{R}^2 / |x| \leq 0.5, |y| \leq 1 \}$ ,

b)  $D = \{ (x,y) \in \mathbb{R}^2 / x^2+y^2 \leq 4, x \geq 0, y \geq 0 \}$

folosind formula trapezelor cu  $n = 4$  subintervale egale pe axa  $Ox$  și  $m = 4$  subintervale egale pe axa  $Oy$ .

R. Pentru calculul integralei  $I = \iint_D f(x,y) dx dy = \int_a^b \left( \int_c^d f(x,y) dy \right) dx$  unde  $D = [a, b] \times [c, d]$  se folosește formula trapezelor repetată

$$I = \frac{h \cdot k}{4} \left\{ f(a,c) + f(a,d) + f(b,c) + f(b,d) + \right. \\ \left. + 2 \left[ \sum_{j=1}^{m-1} f(a, y_j) + \sum_{j=1}^{m-1} f(b, y_j) + \sum_{i=1}^{n-1} f(x_i, c) + \sum_{i=1}^{n-1} f(x_i, d) \right] + \right. \\ \left. + 4 \sum_{i=1}^{n-1} \sum_{j=1}^{m-1} f(x_i, y_j) \right\} \quad (1)$$

unde

$$h = \frac{b-a}{n}, \quad k = \frac{d-c}{m}, \quad x_i = a + ih, \quad i = \overline{1, n-1}, \quad y_j = c + jk, \quad j = \overline{1, m-1}$$

a)  $a = -0.5, b = 0.5, c = -1, d = 1, h = 0.25, k = 0.5, I = 1.33754$

b) Se trece la coordonate polare pentru a transforma sfertul de disc de rază 2 din cadranul întâi într-un dreptunghi

$$\begin{cases} x = \rho \cdot \cos \theta & \theta \in \left[ 0, \frac{\pi}{2} \right] \\ y = \rho \cdot \sin \theta & \rho \in [0, 2] \end{cases}, \text{ iar iacobianul este } J = \rho$$

și se aplică formula de mai sus integralei

$$I = \iint_D e^{-(x^2+y^2)} dx dy = \int_0^{\frac{\pi}{2}} \left( \int_0^2 \rho \cdot e^{-\rho^2} d\rho \right) d\theta$$

și rezultă  $I = 0.73332$  .

15.  $\iint_D \frac{\ln(x^2 + y^2)}{x^2 + y^2} dx dy$  , mulțimea de integrare fiind precizată de

a)  $D = \{ (x,y) \in \mathbb{R}^2 / 1 \leq x \leq 2, 0 \leq y \leq 2 \}$  ,

b)  $D = \{ (x,y) \in \mathbb{R}^2 / 1 \leq x^2 + y^2 \leq 4, 0 \leq x, 0 \leq y \}$

folosind formula trapezelor cu  $n = 2$  subintervale egale pe axa Ox și  $m = 4$  subintervale egale pe axa Oy.

R. a) Înlocuind în formula (1)

$$a = 1, b = 2, c = 0, d = 2, h = 0.5 = k \quad \text{și} \quad f(x, y) = \frac{\ln(x^2 + y^2)}{x^2 + y^2}$$

se obține  $I = 0.636$  .

b) Trecând la coordonate polare se aplică formula (1) pentru

$$a = 1, b = 2, c = 0, d = \frac{\pi}{2}, f(\rho, \theta) = \frac{2 \ln \rho}{\rho}$$

și se obține  $I = 0.73976$  .

## 6. Rezolvarea numerică a problemei Cauchy pentru ecuații diferențiale

### §6.1. Generalități

Ecuatiile diferențiale reprezintă unul dintre cele mai importante instrumente matematice, necesar pentru înțelegerea unor rezultate din mecanică, fizică, etc.

În acest capitol prezentăm metode numerice pentru rezolvarea problemei Cauchy pentru ecuații diferențiale.

Fie  $D = [a, b] \times J \subset \mathbb{R}^2$ ,  $f: D \rightarrow \mathbb{R}$  și  $(x_0, y_0) \in D$ . *Problema Cauchy* pentru ecuația diferențială

$$y' = f(x, y), \quad (1)$$

constă în determinarea unei soluții a ecuației (1), adică a unei funcții derivabile  $y: I \subset [a, b] \rightarrow \mathbb{R}$  astfel ca pentru orice  $x \in I$ ,  $(x, y(x)) \in D$  și  $y' = f(x, y(x))$ ,  $(\forall) x \in I$  care satisface *condiția inițială*

$$y(x_0) = y_0. \quad (2)$$

După cum este cunoscut, găsirea soluției exacte a problemei (1)-(2) nu este posibilă decât în anumite cazuri. De exemplu, determinarea soluției exacte, prin tehnici clasice, a ecuației aparent simple

$$y' = x^2 + y^2, \quad y(0) = 1,$$

nu este posibilă. Se justifică astfel necesitatea recurgerii la metode aproximative pentru rezolvarea problemei Cauchy.

Reamintim, pentru început, câteva rezultate privind existența, unicitatea și stabilitatea soluției acestei probleme.

**Definiție.** Funcția  $f: D \rightarrow \mathbb{R}$  se numește *lipschitziană în raport cu  $y \in J$* , dacă există o constantă  $L > 0$  astfel ca pentru orice  $(x, y) \in D$  și  $(x, z) \in D$  are loc *inegalitatea*

$$|f(x, y) - f(x, z)| \leq L|y - z|. \quad (3)$$

**Observație.** Dacă  $\frac{\partial f}{\partial y}$  există și este mărginită pe  $D$ , atunci  $f$  este lipschitziană pe  $D$ . Într-adevăr, din Teorema lui Lagrange rezultă

$$f(x, y) - f(x, z) = \frac{\partial f}{\partial y}(x, c)(y - z),$$

pentru un anumit  $c$  între  $y$  și  $z$ , deci putem alege

$$L = \max_{(x, y) \in D} \left| \frac{\partial f}{\partial y}(x, y) \right|. \quad (4)$$

În ce privește existența și unicitatea soluției problemei (1)-(2), are loc următoarea teoremă.

**Teorema 1.** Presupunem că sunt îndeplinite condițiile:

- (i)  $f$  este continuă pe  $[a, b]$  în raport cu  $x$ ;
- (ii)  $f$  este lipschitziană pe  $D$  în raport cu  $y$ ;
- (iii)  $(x_0, y_0)$  este punct interior lui  $D$ .

Atunci pentru un  $\alpha > 0$  convenabil, există o soluție unică pe  $I = [x_0 - \alpha, x_0 + \alpha]$  a problemei (1) - (2).

**Exemplul 1.** Fie ecuația  $y' = 1 + \sin(xy)$ ,  $D = [0, 1] \times \mathbb{R}$ . Deoarece  $\frac{\partial f}{\partial y} = x \cos xy$ , conform observației de mai sus, putem lua  $L = 1$ . Atunci pentru orice  $(x_0, y_0)$  cu  $0 < x_0 < 1$  există o soluție a problemei Cauchy pentru ecuația dată pe un anumit interval  $[x_0 - \alpha, x_0 + \alpha] \subset [0, 1]$ .

După cum se știe, soluția problemei (1)-(2) se află cu metoda aproximațiilor succesive.

Fie

$$y_0(x) = y_0,$$

$$y_n(x) = y_0 + \int_{x_0}^x f(t, y_{n-1}(t)) dt, \quad x \in I, \quad n = 1, 2, \dots$$

Atunci șirul de funcții  $(y_n)_n$  este uniform convergent pe intervalul  $I$  și limita sa  $y = \lim_{n \rightarrow \infty} y_n$  este soluție unică a problemei (1)-(2). Mai mult, are loc

$$|y(x) - y_n(x)| \leq \frac{ML^n C^{n+1}}{(n+1)!} e^{LC}, \quad (5)$$

unde

$$M = \sup_{x \in I} |f(x, y_0)|, \quad C = \max(|a - x_0|, |b - x_0|).$$

*Metoda aproximațiilor succesive (Picard)* este o metodă aproximativă de rezolvare a problemei Cauchy. Se aproximează soluția  $y(x)$  cu  $y_n(x)$  și se cunoaște o evaluare a erorii.

**Exemplul 2.** Fie ecuația  $y' = y$ ,  $y(0) = 1$ ,  $D = [-1, 1] \times \mathbb{R}$ . Atunci  $M = 1$ ,  $L = 1$ . Șirul aproximațiilor succesive este:

$$y_1(x) = 1 + \int_0^x 1 \cdot dt = 1 + x,$$

$$y_2(x) = 1 + \int_0^x (1+t) dt = 1 + x + \frac{x^2}{2},$$

$$y_3(x) = 1 + \int_0^x \left(1+t + \frac{t^2}{2}\right) dt = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!},$$

$$\dots\dots\dots$$

$$y_n(x) = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots + \frac{x^n}{n!}.$$

Este clar că  $y_n \rightarrow y = e^x$ , care este soluția exactă a problemei.

Metoda aproximațiilor succesive are dezavantajul că presupune calculul unor integrale, lucru dificil de realizat. Din această cauză, această metodă este mai puțin folosită în practică, metoda având o importanță deosebită, mai ales din punct de vedere teoretic.

În ceea ce privește stabilitatea soluției problemei (1)-(2), ne interesează comportamentul soluției la modificări mici ale funcției  $f(x,y)$  și ale datei inițiale  $y_0$ . Considerăm, deci, problema perturbată

$$y' = f(x, y) + \delta(x), \quad (6)$$

$$y(x_0) = y_0 + \varepsilon, \quad (7)$$

cu aceleași ipoteze asupra funcției  $f$  ca în Teorema 1. Presupunem, în plus, că  $\delta(x)$  este continuă pe  $[a, b]$ . Atunci problema (6)-(7) are soluție unică, notată  $y(x; \delta, \varepsilon)$ .

**Teorema 2.** Presupunem satisfăcute ipotezele Teoremei 1 și că funcția  $\delta(x)$  este continuă pe  $[a, b]$ . Atunci problema (6)-(7) va avea soluție unică  $y(x; \delta, \varepsilon)$  pe un interval  $[x_0 - \alpha, x_0 + \alpha]$ ,  $\alpha > 0$ , uniform pentru toate perturbările  $\varepsilon$  și  $\delta(x)$  ce satisfac:

$$|\varepsilon| \leq \varepsilon_0, \quad \|\delta\|_{\infty} \leq \varepsilon_0,$$

cu  $\varepsilon_0$  suficient de mic. În plus, dacă  $y(x)$  este soluția problemei neperturbate, atunci

$$\max_{|x-x_0| \leq \alpha} |y(x) - y(x; \delta, \varepsilon)| \leq k(|\varepsilon| + \alpha \|\delta\|_\infty), \quad (8)$$

cu  $k = 1/(1 - \alpha L)$ .

Utilizând acest rezultat, se poate afirma că problema (1)-(2) este *corect pusă* sau *stabilă*. Deci dacă se fac mici modificări în ecuația diferențială sau în data inițială, atunci soluția nu se modifică semnificativ.

Soluția  $y$  depinde continuu de datele problemei, anume funcția  $f$  și data inițială  $y_0$ . Din punct de vedere fizic, semnificația Teoremei 2 constă în faptul că pentru fenomene fizice descrise de ecuații diferențiale, mici abateri sau erori în condițiile inițiale sau în însăși legea de evoluție, nu deformează prea puternic procesul. Rezultatul este important cu atât mai mult cu cât asemenea perturbații sau erori sunt întotdeauna inevitabile. Se poate întâmpla ca o problemă să fie stabilă, dar prost condiționată în raport cu calculul numeric, deși asemenea situații nu apar prea des în practică.

Pentru a înțelege mai bine când aceasta se poate produce, vom estima perturbările soluției  $y(x)$ , datorate perturbărilor în problemă. Vom simplifica discuția considerând numai perturbările  $\varepsilon$  în data inițială  $y_0$ ; perturbările  $\delta(x)$  intervin în răspunsul final conform (8).

Perturbăm deci, valoarea inițială  $y_0$  ca în (7). Fie  $y(x; \varepsilon)$  soluția perturbată. Atunci

$$y'(x; \varepsilon) = f(x, y(x; \varepsilon)), \quad x_0 - \alpha \leq x \leq x_0 + \alpha,$$

$$y(x_0; \varepsilon) = y_0 + \varepsilon$$

Dacă  $y(x)$  este soluția problemei neperturbate (1)-(2) și  $z(x) = y(x; \varepsilon) - y(x)$  este eroarea, atunci

$$z'(x; \varepsilon) = f(x, y(x; \varepsilon)) - f(x, y(x)) \approx \frac{\partial f(x, y(x))}{\partial y} \cdot z(x; \varepsilon), \quad (9)$$

$$z(x_0; \varepsilon) = \varepsilon.$$

Aproximarea (9) este valabilă când  $y(x; \varepsilon)$  este suficient de aproape de  $y(x)$ , ceea ce se întâmplă pentru valori mici ale lui  $\varepsilon$  și intervale mici  $[x_0 - \alpha, x_0 + \alpha]$ .

Ecuația diferențială aproximativă (9) se poate integra ușor. Se obține

$$z(x; \varepsilon) \approx \varepsilon \exp \left[ \int_{x_0}^x \frac{\partial f(t, y(t))}{\partial y} dt \right].$$

Dacă derivata parțială satisface

$$\frac{\partial f}{\partial y}(t, y(t)) \leq 0, \quad |x_0 - t| \leq \alpha,$$

atunci  $z(x; \varepsilon)$  rămâne mărginită de  $\varepsilon$  când  $x$  crește. În acest caz, se spune că problema Cauchy este *bine condiționată*. Ca exemplu de comportare opusă, considerăm problema

$$y' = \lambda y + g(x), \quad y(0) = y_0, \quad (10)$$

cu  $\lambda > 0$ . Cum  $\frac{\partial z}{\partial y} = \lambda$ , putem calcula exact  $z(x; \varepsilon) = \varepsilon e^{\lambda x}$ .

Atunci perturbarea lui  $y(x)$  se mărește când  $x$  crește.

**Exemplul 3.** Ecuația diferențială

$$y' = 100y - 101e^{-x}, \quad y(0) = 1, \quad (11)$$

are soluția  $y(x; \varepsilon) = e^{-x} + \varepsilon e^{100x}$ , care se depărtează rapid de soluția exactă. Spunem că problema (11) este *prost condiționată*.

Revenim acum la problema (1)-(2). Din considerente practice, se presupune că  $x_0 = a$ , adică se înlocuiește condiția (2) cu

$$y(a) = y_0. \quad (12)$$

Această presupunere nu este restrictivă, pentru că dacă am găsit un algoritm care rezolvă problema (1)-(12), atunci cu acest algoritm putem rezolva și problema (1)-(2). Într-adevăr, fie  $I_1 = [a, x_0]$  și  $I_2 = [x_0, b]$ .

Pe intervalul  $I_1$  facem schimbarea de variabilă  $X = x_0 - x$ . Atunci  $y(x) = y(x_0 - X) = Y(X)$  și ecuația (1) devine

$$Y'(X) = -f(X, Y), \quad X \in [0, x_0 - a], \quad (13)$$

iar condiția inițială (12) devine

$$Y(0) = y_0. \quad (14)$$

Metodele numerice pentru rezolvarea problemei (1)-(12) constau în alegerea unor noduri (de obicei, echidistante)  $x_k = a + kh$ ,  $k \in \mathbb{N}$  și determinarea unor valori aproximative ale soluției exacte  $y(x)$  în aceste noduri, valori pe care le notăm cu  $y_k$ . Așadar  $y_k \cong y(x_k)$ .

Se cunosc două clase importante de metode numerice pentru rezolvarea problemei Cauchy.

1. *Metode directe (uni-pas)* în care  $y_k$  este calculat, printr-o relație de recurență, în funcție numai de valoarea  $y_{k-1}$  calculată anterior. În această categorie intră metoda Taylor și metodele Runge-Kutta.

2. *Metode indirecte (cu mai mulți pași)* în care  $y_k$  se calculează printr-o relație de recurență în funcție de valorile precedente  $y_{k-m}, \dots, y_{k-2}, y_{k-1}$ .

În această categorie intră metodele Adams-Bashforth, Adams-Moulton și metoda predictor-corector.

## §6.2. Metode directe

**Metoda lui Taylor** (1685-1731). Fie nodurile echidistante  $x_n = x_0 + nh$ ,  $x_0 = a$  și  $y = y(x)$  soluția exactă a problemei (1)-(2).

Așadar  $y'(x) = f(x, y(x))$ ,  $y(x_0) = y_0$ .

Presupunem că  $f$  este diferențiabilă de un număr suficient de ori. Cum  $x_1 = x_0 + h$ , din formula lui Taylor rezultă

$$y(x_1) = y(x_0 + h) = y(x_0) + \frac{h}{1!} y'(x_0) + \frac{h^2}{2!} y''(x_0) + \dots + \frac{h^p}{p!} y^{(p)}(x_0) + R_{p+1},$$

unde

$$R_{p+1} = \frac{y^{(p+1)}(\xi)}{(p+1)!} h^{p+1}, \quad \xi \in (x_0, x_1).$$

Din  $y'(x) = f(x, y(x))$  rezultă succesiv

$$\begin{aligned} y''(x) &= \frac{\partial f}{\partial x}(x, y(x)) + \frac{\partial f}{\partial y}(x, y(x)) y'(x) = \\ &= \frac{\partial f}{\partial x}(x, y(x)) + \frac{\partial f}{\partial y}(x, y(x)) f(x, y(x)), \\ y'''(x) &= \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial x \partial y} \cdot y' + \left( \frac{\partial^2 f}{\partial x \partial y} + \frac{\partial^2 f}{\partial y^2} y' \right) f + \frac{\partial f}{\partial y} \left( \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y} \cdot y' \right) = \\ &= \frac{\partial^2 f}{\partial x^2} + 2 \frac{\partial^2 f}{\partial x \partial y} f + \frac{\partial^2 f}{\partial y^2} f^2 + \frac{\partial f}{\partial y} \cdot \frac{\partial f}{\partial x} + \left( \frac{\partial f}{\partial y} \right)^2 \cdot f \quad \text{etc.} \end{aligned}$$

Atunci:

$$\begin{aligned} y'(x_0) &= f(x_0, y(x_0)) = f(x_0, y_0), \\ y''(x_0) &= \frac{\partial f}{\partial x}(x_0, y_0) + \frac{\partial f}{\partial y}(x_0, y_0) f(x_0, y_0), \\ y'''(x_0) &= \frac{\partial^2 f}{\partial x^2}(x_0, y_0) + 2 \frac{\partial^2 f}{\partial x \partial y}(x_0, y_0) f(x_0, y_0) + \frac{\partial^2 f}{\partial y^2}(x_0, y_0) f^2(x_0, y_0) + \\ &\quad + \frac{\partial f}{\partial x}(x_0, y_0) \cdot \frac{\partial f}{\partial y}(x_0, y_0) + \left( \frac{\partial f}{\partial y}(x_0, y_0) \right)^2 \cdot f(x_0, y_0) \quad \text{etc.} \end{aligned}$$

Pentru  $p = 3$  obținem

$$y(x_1) = y_0 + \frac{h}{1!} y'(x_0) + \frac{h^2}{2!} y''(x_0) + \frac{h^3}{3!} y'''(x_0) + R_4.$$

Aproximăm soluția exactă în  $x_1$ , deci  $y(x_1)$ , cu

$$y_1 = y_0 + \frac{h}{1!} y'(x_0) + \frac{h^2}{2!} y''(x_0) + \frac{h^3}{3!} y'''(x_0),$$

eroarea fiind dată de

$$|y(x_1) - y_1| = |R_4| \leq \frac{1}{4!} M_4 \cdot h^4, \quad \text{unde } M_4 = \sup_{x \in [a, x_1]} |y^{IV}(x)|,$$

unde  $y^{IV}(x)$  se calculează ca mai sus.

În continuare, considerând soluția problemei (1) ce satisface  $y(x_1) = y_1$ , deci pornind cu punctul  $(x_1, y_1)$ , se determină  $y_2$  care aproximează pe  $y(x_2)$ , ș.a.m.d. În general,  $y(x_n)$  se aproximează cu  $y_n$ , dat de

$$y_n = y_{n-1} + \frac{h}{1!} y'(x_{n-1}) + \frac{h^2}{2!} y''(x_{n-1}) + \frac{h^3}{3!} y'''(x_{n-1}).$$

Evident, erorile se acumulează.

**Exemplul 4.** Fie ecuația  $y' = 1 - \frac{y}{x}$ ,  $y(1) = \frac{3}{2}$ . Alegem  $h = 0.1$ . În acest caz

$$f(x, y) = 1 - \frac{y}{x}, \quad \frac{\partial f}{\partial x} = \frac{y}{x^2}, \quad \frac{\partial f}{\partial y} = -\frac{1}{x}, \quad \frac{\partial^2 f}{\partial x^2} = -\frac{2y}{x^3}, \quad \frac{\partial^2 f}{\partial x \partial y} = \frac{1}{x^2}, \quad \frac{\partial^2 f}{\partial y^2} = 0.$$

$$\text{Atunci } y'(1) = f\left(1, \frac{3}{2}\right) = -\frac{1}{2}, \quad y''(1) = 2, \quad y'''(1) = -6.$$

Deci  $x_0 = 1$ ,  $x_1 = 1.1$ . Aproximăm  $y(1.1)$  cu  $y_1$  dat de

$$y_1 = \frac{3}{2} + \frac{0.1}{1!} \left(-\frac{1}{2}\right) + \frac{(0.1)^2}{2!} \cdot 2 + \frac{(0.1)^3}{3!} (-6).$$

Obținem  $y_1 = 1.459$ . Pe de altă parte, soluția exactă a problemei este  $y = \frac{x}{2} + \frac{1}{x}$ , deci  $y(1.1) = 1.459090$ .

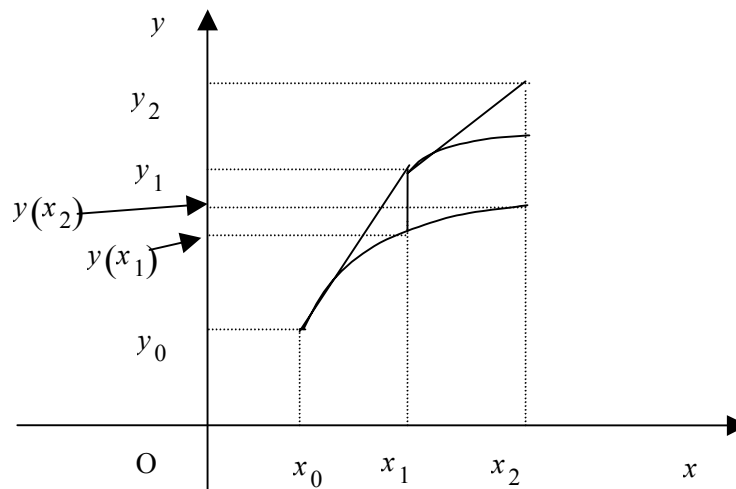
Dezavantajul metodei Taylor constă în faptul că presupune calculul derivatelor  $y^{(2)}, y^{(3)}, \dots, y^{(k)}, \dots$ , la fiecare pas, ceea ce este dificil de realizat. Această deficiență este înlăturată de metodele care urmează.

Dacă  $p = 1$  se obține **metoda lui Euler** (1707-1783). În acest caz valorile aproximative  $y_n$  ale lui  $y(x_n)$  sunt date de

$$y_n = y_{n-1} + h f(x_{n-1}, y_{n-1}), \quad n \geq 1, \quad (15)$$

unde, evident,  $y_0 = y(x_0)$ .

Metoda lui Euler are o interpretare geometrică foarte simplă: dacă s-a determinat valoarea  $y_{n-1}$ , pentru a determina  $y_n$ , se consideră soluția ecuației (1) care trece prin  $(x_{n-1}, y_{n-1})$  (deci care satisface  $y(x_{n-1})=y_{n-1}$ ); se duce apoi tangenta la graficul acestei soluții, în punctul  $(x_{n-1}, y_{n-1})$ ; se intersectează această tangentă cu dreapta  $x=x_n$ , obținându-se  $y_n$ . Din acest motiv, metoda lui Euler se mai numește și *metoda liniilor poligonale*. După cum se observă din figura de mai jos, erorile se acumulează.



**Exemplul 5.** Folosind metoda Euler să se determine soluția aproximativă a următoarei probleme Cauchy

$$\begin{cases} y' &= y^2 - \frac{y}{x} - \frac{1}{4x^2} \\ y(1) &= 0.5 \end{cases}$$

în punctul  $x=2$  în doi pași.

În acest caz

$$x_0=1, \quad y_0=0.5, \quad n=2, \quad h=0.5, \quad x_1=1+0.5=1.5, \quad x_2=2.$$

Atunci:

$$y_1 = y_0 + h \cdot f(x_0, y_0) = 0.5 + 0.5 \left( 0.5^2 - \frac{0.5}{1} - \frac{1}{4 \cdot 1^2} \right) = 0.25,$$

$$y_2 = y_1 + h \cdot f(x_1, y_1) = 0.14236.$$

**Metodele Runge-Kutta** se deosebesc de metoda lui Taylor prin faptul că înlocuiesc calculul derivatelor funcției  $f$ , prin evaluări ale lui  $f$  în diverse puncte. Metoda a fost introdusă de matematicianul german Carl David Runge în 1895 și dezvoltată de un alt matematician german, Wilhelm Kutta, în 1901. Vom analiza în detaliu

metoda Runge-Kutta de ordinul doi. Valorile aproximative  $y_n$  ale lui  $y(x_n)$  sunt date de

$$y_n = y_{n-1} + a_1 h f(x_{n-1}, y_{n-1}) + a_2 h f(x_{n-1} + b_1 h, y_{n-1} + b_2 h f(x_{n-1}, y_{n-1})) , \quad n \geq 1 , \quad (16)$$

iar  $y_0 = y(x_0)$ . Constantele  $a_1, a_2, b_1, b_2$  urmează a fi determinate. Dacă notăm

$$f = f(x_{n-1}, y_{n-1}), f_x = \frac{\partial f}{\partial x}(x_{n-1}, y_{n-1}), f_y = \frac{\partial f}{\partial y}(x_{n-1}, y_{n-1}), \text{ atunci, } \quad \text{din}$$

formula lui Taylor pentru funcții de două variabile, obținem

$$y_n = y_{n-1} + a_1 h f + a_2 h \left( f + b_1 h f_x + b_2 h f_y \right) + O(h^2) = \quad (17)$$

$$= y_{n-1} + (a_1 + a_2) f \cdot h + (a_2 b_1 f_x + a_2 b_2 f f_y) \cdot h^2 + O(h^3) .$$

Pe de altă parte, din metoda lui Taylor, avem

$$y_n = y_{n-1} + \frac{h}{1!} f + \frac{h^2}{2!} (f_x + f_y \cdot f) + O(h^3) . \quad (18)$$

Identificând coeficienții lui  $h$  și  $h^2$  din (17) și (18), rezultă

$$\begin{cases} a_1 + a_2 = 1 \\ a_2 b_1 = \frac{1}{2} \\ a_2 b_2 = \frac{1}{2} \end{cases} . \quad (19)$$

Deoarece sistemul (19) are 3 ecuații și 4 necunoscute, una din necunoscute poate fi aleasă arbitrar. De exemplu, dacă alegem  $b_2 = \alpha$ , atunci  $b_1 = \alpha$ , deci

$$\begin{cases} a_1 + a_2 = 1 \\ a_2 \alpha = \frac{1}{2} \\ b_2 = \alpha \end{cases} ,$$

astfel că formulele (16) se mai scriu

$$\begin{cases} y_n = y_{n-1} + h(a_1 g_1 + a_2 g_2) , n \geq 1 \\ g_1 = f(x_{n-1}, y_{n-1}) \\ g_2 = f(x_{n-1} + \alpha h, y_{n-1} + \alpha h g_1) \\ a_1 + a_2 = 1 \\ a_2 \alpha = \frac{1}{2} . \end{cases} \quad (20)$$

Pentru  $a_1 = a_2 = \frac{1}{2}$  ,  $\alpha = 1$  , se obține *metoda Euler îmbunătățită*

$$y_n = y_{n-1} + \frac{h}{2} [f(x_{n-1}, y_{n-1}) + f(x_{n-1} + h, y_{n-1} + h f(x_{n-1}, y_{n-1}))], n \geq 1. \quad (21)$$

Pentru  $a_1 = 0$  ,  $a_2 = 1$  ,  $\alpha = \frac{1}{2}$  , se obține *metoda Euler modificată*

$$y_n = y_{n-1} + h f\left(x_{n-1} + \frac{h}{2}, y_{n-1} + \frac{h}{2} f(x_{n-1}, y_{n-1})\right), n \geq 1. \quad (22)$$

**Exemplul 6.** Folosind metoda Euler îmbunătățită să se determine soluția aproximativă a următoarei probleme Cauchy

$$\begin{cases} y' = y^2 - \frac{y}{x} - \frac{1}{4x^2} \\ y(1) = 0.5 \end{cases}$$

în punctul  $x=2$  în doi pași.

În acest caz  $x_0=1$  ,  $y_0=0.5$  ,  $h=0.5$  ,  $x_1=1.5$  ,  $x_2=2$ .

Cum  $y_1 = y_0 + \frac{h}{2} \cdot [f(x_0, y_0) + f(x_0 + h, y_0 + h \cdot f(x_0, y_0))]$ , prin calcul

se obține  $y_1 = 0.32118$ .

Similar

$$\begin{aligned} y_2 &= y_1 + \frac{h}{2} \cdot [f(x_1, y_1) + f(x_1 + h, y_1 + h \cdot f(x_1, y_1))] = \\ &= 0.32118 + 0.25 \cdot (-0.22207 - 0.12341) = 0.23481. \end{aligned}$$

**Exemplul 7.** Folosind metoda Euler modificată să se determine soluția aproximativă a problemei Cauchy din Exemplul 6, în punctul  $x=2$  în doi pași.

În acest caz  $y_1 = y_0 + h \cdot f\left(x_0 + \frac{h}{2}, y_0 + \frac{h}{2} f(x_0, y_0)\right)$ , deci  $y_1 = 0.34031$ , iar

$$y_2 = y_1 + h \cdot f\left(x_1 + \frac{h}{2}, y_1 + \frac{h}{2} f(x_1, y_1)\right) = 0.25868.$$

În continuare prezentăm metoda Runge-Kutta de ordinul patru în forma particulară sub care este cel mai des utilizată (W.Kutta - 1901).

$$y_n = y_{n-1} + \frac{h}{6}(g_1 + 2g_2 + 2g_3 + g_4), \quad n \geq 1, \quad (23)$$

unde

$$\begin{aligned} g_1 &= f(x_{n-1}, y_{n-1}), \\ g_2 &= f\left(x_{n-1} + \frac{h}{2}, y_{n-1} + \frac{h}{2} g_1\right), \\ g_3 &= f\left(x_{n-1} + \frac{h}{2}, y_{n-1} + \frac{h}{2} g_2\right), \\ g_4 &= f(x_{n-1} + h, y_{n-1} + h g_3), \end{aligned}$$

și  $y_0 = y(x_0)$ .

**Exemplul 8.** Folosind metoda Runge-Kutta de ordinul 4 să se determine soluția aproximativă a problemei Cauchy din Exemplul 6, în punctul  $x=2$  în doi pași.

Folosind notațiile din Exemplul 6 se obține:

$$\begin{aligned} g_1 &= f(x_0, y_0) = -0.5, \\ g_2 &= f\left(x_0 + \frac{h}{2}, y_0 + \frac{h}{2} \cdot g_1\right) = f(1.25, 0.5 - 0.25 \cdot 0.5) = -0.31937, \\ g_3 &= f\left(x_0 + \frac{h}{2}, y_0 + \frac{h}{2} \cdot g_2\right) = f(1.25, 0.5 - 0.25 \cdot 0.31937) = -0.31959, \\ g_4 &= f(x_0 + h, y_0 + h \cdot g_3) = f(1.5, 0.5 - 0.5 \cdot 0.31959) = -0.22218, \end{aligned}$$

deci

$$y_1 = y_0 + \frac{h}{6} \cdot [g_1 + 2 \cdot (g_2 + g_3) + g_4] = 0.33332.$$

Pentru  $y_2$  calculăm mai întâi

$$\begin{aligned} g_1 &= f(x_1, y_1) = -0.22222, \\ g_2 &= f\left(x_1 + \frac{h}{2}, y_1 + \frac{h}{2} \cdot g_1\right) = -0.1632, \\ g_3 &= f\left(x_1 + \frac{h}{2}, y_1 + \frac{h}{2} \cdot g_2\right) = -0.16322, \\ g_4 &= f(x_1 + h, y_1 + h \cdot g_3) = -0.125. \end{aligned}$$

În consecință

$$y_2 = y_1 + \frac{h}{6} \cdot [g_1 + 2 \cdot (g_2 + g_3) + g_4] = 0.24999.$$

Această metodă se poate aplica și pentru sisteme de ecuații diferențiale.

Fie sistemul de ecuații diferențiale

$$\begin{cases} \frac{dy_1}{dx} = f_1(x, y_1, y_2) \\ \frac{dy_2}{dx} = f_2(x, y_1, y_2), \end{cases} \quad (24)$$

cu condiția inițială

$$\begin{cases} y_1(x_0) = y_{10} \\ y_2(x_0) = y_{20} \end{cases} \quad (25)$$

Formulele (23) se scriu în acest caz astfel

$$\begin{pmatrix} y_{1n} \\ y_{2n} \end{pmatrix} = \begin{pmatrix} y_{1,n-1} \\ y_{2,n-1} \end{pmatrix} + \frac{h}{6} \begin{pmatrix} g_{11} \\ g_{12} \end{pmatrix} + \frac{h}{3} \begin{pmatrix} g_{21} \\ g_{22} \end{pmatrix} + \frac{h}{3} \begin{pmatrix} g_{31} \\ g_{32} \end{pmatrix} + \frac{h}{6} \begin{pmatrix} g_{41} \\ g_{42} \end{pmatrix}, \quad n \geq 1,$$

unde:

$$\begin{aligned} g_{11} &= f_1(x_{n-1}, y_{1,n-1}, y_{2,n-1}) \\ g_{12} &= f_2(x_{n-1}, y_{1,n-1}, y_{2,n-1}) \end{aligned} \quad (27)$$

$$g_{21} = f_1\left(x_{n-1} + \frac{h}{2}, y_{1,n-1} + \frac{h}{2}g_{11}, y_{2,n-1} + \frac{h}{2}g_{12}\right)$$

$$g_{22} = f_2\left(x_{n-1} + \frac{h}{2}, y_{1,n-1} + \frac{h}{2}g_{11}, y_{2,n-1} + \frac{h}{2}g_{12}\right)$$

$$g_{31} = f_1\left(x_{n-1} + \frac{h}{2}, y_{1,n-1} + \frac{h}{2}g_{21}, y_{2,n-1} + \frac{h}{2}g_{22}\right)$$

$$g_{32} = f_2\left(x_{n-1} + \frac{h}{2}, y_{1,n-1} + \frac{h}{2}g_{21}, y_{2,n-1} + \frac{h}{2}g_{22}\right)$$

$$g_{41} = f_1(x_{n-1} + h, y_{1,n-1} + hg_{31}, y_{2,n-1} + hg_{32})$$

$$g_{42} = f_2(x_{n-1} + h, y_{1,n-1} + hg_{31}, y_{2,n-1} + hg_{32})$$

și

$$y_{1,0} = y_1(x_0)$$

$$y_{2,0} = y_2(x_0)$$

**Asupra existenței, stabilității și convergenței metodelor directe**

Orice metodă directă pentru rezolvarea problemei Cauchy poate fi scrisă sub forma generală

$$y_n = y_{n-1} + h\Phi(x_{n-1}, y_{n-1}, h), \quad n \geq 1, \quad (28)$$

$$y_0 = y(x_0),$$

unde funcția  $\Phi(x, y, h)$  se numește *funcție de creștere*. Pentru  $h \neq 0$ , relația (28) se scrie sub forma

$$\frac{y_n - y_{n-1}}{h} = \Phi(x_{n-1}, y_{n-1}, h). \quad (29)$$

**Definiție.** Dacă  $y(x)$  este soluția exactă a problemei Cauchy (1)-(2), se numește eroare de trunchiere a metodei, funcția

$$t(x, h) = \frac{y(x+h) - y(x)}{h} - \Phi(x, y(x), h), \quad (30)$$

unde  $x \in [x_0, b)$  și  $h > 0$  astfel ca  $x+h \leq b$ .

**Definiție.** Se spune că formula (28) dă o aproximare consistentă a problemei (1), (2), dacă  $t(x, h) \rightarrow 0$  când  $h \rightarrow 0$ , uniform în raport cu  $x \in [x_0, b)$ .

Din (30) se vede că, pentru o metodă consistentă, avem

$$y'(x) = \Phi(x, y(x), 0) = f(x, y(x)).$$

**Definiție.** Se spune că formula (28) dă o aproximare consistentă de ordin  $p$ , dacă există  $N \geq 0$ ,  $h_0 > 0$  și un întreg pozitiv  $p$  astfel ca  $\sup_{x_0 \leq x \leq b} |t(x, h)| \leq N \cdot h^p$ ,

pentru orice  $h \in (0, h_0]$ .

**Propoziția 1.** Metoda lui Taylor de ordin  $p$  este consistentă de ordin  $p$ . Metodele Runge-Kutta de ordinul  $p$  dat sunt metode consistente de ordinul  $p$ .

**Demonstrație.** Pentru metoda dată de formula lui Taylor de ordinul  $p$ , avem

$$t(x, h) = \frac{h^p}{(p+1)!} y^{(p+1)}(\xi), \quad \text{unde } x < \xi < x+h.$$

Alegând  $N = \sup_{x_0 \leq x \leq b} \frac{1}{(p+1)!} |y^{(p+1)}(x)|$ , rezultă existența de ordinul  $p$ .  $\square$

Spre exemplu, metoda lui Euler este consistentă de ordinul 1. Pentru metodele Runge-Kutta este dificil să obținem o formă explicită a lui  $N$ .

Până acum am considerat erorile de trunchiere, adică erorile ce apar prin discretizarea ecuației diferențiale. Ne interesează însă, în ce măsură soluția ecuației

cu diferențe, adică a ecuației obținută prin discretizare (deci șirul  $(y_n)_n$ ), aproximează soluția  $y(x)$  a ecuației diferențiale.

Așadar, vom aborda problema "convergenței" soluției ecuației cu diferențe la soluția ecuației diferențiale. Această convergență trebuie definită cu atenție. De exemplu, dacă analizăm comportarea șirului  $(y_n)_n$ , când  $h \rightarrow 0$ , pentru  $n$  fixat, nu vom obține un concept util, deoarece, în acest caz  $x_n = x_0 + nh \rightarrow x_0$ , iar pe noi ne interesează ce se întâmplă pentru  $x \neq x_0$ . Deci trebuie să considerăm comportarea șirului  $(y_n)_n$  când  $h \rightarrow 0$ , cu  $x = x_n = x_0 + nh$  fixat. Pentru a obține o soluție pentru valoarea  $x \neq x_0$  fixată, trebuie să mărim numărul de pași ceruți pentru a ajunge la  $x$  din  $x_0$ , dacă pasul  $h$  descrește.

**Definiție.** Metoda numerică directă se numește convergentă dacă pentru orice  $x \in [x_0, b]$ , avem

$$\lim_{h \rightarrow 0} y_n = y(x) \quad (31)$$

$$x = x_n = x_0 + nh.$$

Această definiție se datorează lui G. Dahlquist.

În studiul convergenței metodelor directe, este utilă următoarea lemă.

**Lema 1.** Au loc inegalitățile:

$$1 + x \leq e^x, \quad (\forall) x \in \mathbf{R}, \quad (32)$$

$$0 \leq (1+x)^m \leq e^{mx}, \quad (\forall) x \geq -1, \quad m \in \mathbf{N}. \quad (33)$$

**Demonstrație.** Cum (33) este consecință imediată a lui (32), este suficient să justificăm (32). Dar (32) este consecință imediată a formulei lui Taylor, deoarece

$$e^x = 1 + x + \frac{x^2}{2} e^\xi,$$

unde  $\xi$  este între 0 și  $x$ .  $\square$

În continuare, vom analiza convergența metodelor directe.

Fie  $e_n = y(x_n) - y_n$ ,  $n \geq 0$ . Deci  $e_n$  reprezintă eroarea globală dintre soluția exactă și soluția aproximativă în nodurile  $x_n$ .

Din (30), obținem

$$y(x_n) = y(x_{n-1}) + h\Phi(x_{n-1}, y(x_{n-1}), h) + ht(x_{n-1}, h).$$

Scăzând (28) din această egalitate, avem

$$e_n = e_{n-1} + h[\Phi(x_{n-1}, y(x_{n-1}), h) - \Phi(x_{n-1}, y_{n-1}, h)] + ht(x_{n-1}, h) \quad (33')$$

Evident  $e_0 = 0$ .

Din nefericire, faptul că  $t(x_{n-1}, h)$  este mic, nu este suficient pentru a asigura că  $e_n$  este mic. Ar trebui să arătăm că

$$\max_n |e_n| \leq C \max_n |t(x_n, h)|,$$

unde constanta  $C$  este independentă de  $h$ ; este ceea ce numim *stabilitatea metodei de aproximare*.

În cele ce urmează, presupunem că funcția de creștere  $\Phi$  satisface condiția lui Lipschitz

$$|\Phi(x, y, h) - \Phi(x, z, h)| \leq K|y - z|, \quad x \in [x_0, b], \quad y, z \in \mathbb{R}, \quad h > 0. \quad (34)$$

și că metoda este consistentă de ordin  $p$  cu constanta  $N$ .

Considerăm mai întâi cazul  $K > 0$ . Atunci din (33') rezultă că există  $h_0 > 0$  astfel ca pentru  $h \in (0, h_0]$  să avem

$$|e_n| \leq |e_{n-1}|(1 + hK) + Nh^{p+1}.$$

Aplicând această inegalitate recursiv rezultă

$$|e_n| \leq (1 + hK)^n |e_0| + Nh^{p+1} [1 + (1 + hK) + \dots + (1 + hK)^{n-1}].$$

Deoarece  $e_0 = 0$ , din Lema 1, rezultă că

$$|e_n| \leq Nh^p \frac{(1 + hK)^n - 1}{K} \leq Nh^p \frac{e^{nhK} - 1}{K} = Nh^p \frac{e^{(x_n - x_0)K} - 1}{K}.$$

Pentru  $K = 0$ , se obține imediat

$$|e_n| \leq (x_n - x_0)Nh^p.$$

Am demonstrat deci următoarea teoremă.

**Teorema 3.** Dacă metoda (28) este consistentă de ordin  $p$ , cu constanta  $N$ , iar funcția  $\Phi$  satisface condiția lui Lipschitz (34), atunci există  $h_0 > 0$ , astfel ca pentru  $h \in (0, h_0]$  avem

$$|y(x_n) - y_n| \leq \begin{cases} \frac{e^{(x_n - x_0)K} - 1}{K} Nh^p, & \text{dacă } K \neq 0 \\ (x_n - x_0)Nh^p, & \text{dacă } K = 0, \end{cases} \quad (35)$$

unde  $y(x)$  este soluția exactă a problemei Cauchy. Deci metoda este convergentă.

Această teoremă s-a obținut pornind de la principiu important al analizei numerice, care se poate enunța astfel:

CONSISTENȚĂ + STABILITATE  $\Rightarrow$  CONVERGENȚĂ

**Definiție.** O metodă numerică directă se numește convergentă de ordinul  $p \in \mathbb{N}$  dacă există  $C > 0$  astfel încât  $|y(x_n) - y_n| \leq Ch^p$ , oricare ar fi  $x_n = x \in [x_0, b]$ .

Teorema 3 spune că o metodă numerică consistentă de ordinul  $p \in \mathbb{N}$  și pentru care funcția de creștere satisface condiția Lipschitz (34) este convergentă de ordinul  $p$ . Numărul  $p$  introdus de Teorema 3 nu este unic determinat (dacă

există): dacă  $h < 1$  și  $p$  este ordin de convergență, atunci și  $p'$  cu  $0 < p' < p$  este un ordin de convergență pentru această metodă. Se poate pune problema determinării unui ordin de convergență maximal pentru o metodă dată. În practică este suficient să se determine un ordin de convergență convenabil.

Metodele prezentate mai sus au ordine de convergență diferite.

Spre exemplu se poate arăta că metoda lui Euler îmbunătățită are ordin de convergență 2, dacă există  $L > 0$  astfel încât

$$\left| \frac{\partial f}{\partial y}(x, y) \right| \leq L, \quad (\forall)(x, y) \in [a, b] \times J.$$

În funcție de  $L$ , se poate determina  $K$  din (34).

Ținând seama de (21) pentru această metodă avem

$$\Phi(x, y, h) = \frac{1}{2} [f(x, y) + f(x + h, y + hf(x, y))].$$

Atunci

$$\begin{aligned} & |\Phi(x, y, h) - \Phi(x, z, h)| \leq \frac{1}{2} |f(x, y) - f(x, z)| + \\ & + \frac{1}{2} |f(x + h, y + hf(x, y)) - f(x + h, z + hf(x, z))| \leq \\ & \leq \frac{1}{2} L |y - z| + \frac{1}{2} L [|y - z| + h |f(x, y) - f(x, z)|] \leq (L + \frac{L^2 h_0}{2}) |y - z|, \quad h \leq h_0, \end{aligned}$$

$$\text{deci } K = L + \frac{L^2 h_0}{2}.$$

În acest caz, din formula Taylor rezultă

$$t(x, h) = h^2 (R_3(x) + Q_3(x, h)),$$

unde  $R_3(x)$  este eroarea de la formula Taylor, iar  $Q_3(x, h)$  eroarea care se face oprind termenii de ordinul doi din formula Runge-Kutta. Dacă derivatele lui  $y$  și ale lui  $f$  sunt mărginite, atunci metoda Euler îmbunătățită are ordinul de consistență 2.

În încheiere, menționăm că metoda folosită în studiul stabilității soluției problemei Cauchy se poate aplica și în cazul metodei Euler.

Considerăm metoda numerică (analoagă problemei Cauchy):

$$z_n = z_{n-1} + h [f(x_{n-1}, z_{n-1}) + \delta(x_{n-1})], \quad n \leq 1 \quad (36)$$

$$z_0 = y_0 + \varepsilon.$$

Comparăm cele două soluții numerice  $(z_n)_n$ ,  $(y_n)_n$ , când  $h \rightarrow 0$ .

Fie  $e_n = z_n - y_n$ ,  $n \geq 0$ . Atunci  $z_0 = \varepsilon$ . Scăzând (15) din (36), obținem

$$e_n = e_{n-1} + h [f(x_{n-1}, z_{n-1}) - f(x_{n-1}, y_{n-1})] + h \delta(x_{n-1}),$$

care are aceeași formă ca (23). Utilizând același procedeu ca îndemonstrația Teoremei 3, rezultă

$$\max_n |z_n - y_n| \leq e^{(b-x_0)} |\varepsilon| + \frac{e^{(b-x_0)L} - 1}{L} \cdot \|\delta\|_\infty .$$

În consecință, există constantele  $k_1, k_2$ , independente de  $h$ , astfel ca

$$\max_n |z_n - y_n| \leq k_1 |\varepsilon| + k_2 \|\delta\|_\infty . \quad (37)$$

Această inegalitate este analoagă inegalității (8) din cazul problemei Cauchy inițiale. Așadar metoda Euler este stabilă numeric. De altfel, toate metodele numerice pentru problema Cauchy au această formă de stabilitate, imitând stabilitatea problemei inițiale. Analiza se poate simplifica, luând  $\delta(x) = 0$  și considerând numai efectul perturbării inițiale  $y_0$ . Nu vom analiza aici problema erorilor de rotunjire.

### §6.3. Metode indirecte (cu mai mulți pași)

**Metoda Adams-Bashforth.** Să presupunem că printr-o metodă directă (de exemplu, de tip Runge-Kutta) s-au determinat valorile  $y_1, \dots, y_n$  în nodurile  $x_1, \dots, x_n$ , unde  $y_k \approx y(x_k)$ . Se pune problema determinării unei valori aproximative  $y_{n+1}$  pentru  $y(x_{n+1})$  ( $y(x)$  este soluția exactă a problemei Cauchy). Integrând (1) pe intervalul  $[x_n, x_{n+1}]$ , obținem:

$$y(x_{n+1}) - y(x_n) = \int_{x_n}^{x_{n+1}} f(x, y(x)) dx. \quad (38)$$

Pentru a calcula integrala, folosim o metodă numerică, de exemplu o metodă Newton-Côtes pentru nodurile echidistante  $x_{n-m}, \dots, x_i, \dots, x_n$ ,  $m \leq n$ ,  $x_i = x_n + (i-n)h$ ,  $i = n-m, n$ .

Dacă  $x \in [x_n, x_{n+1}]$ , atunci există  $t \in [0, 1]$  astfel încât

$$x = x_n + th. \quad (39)$$

Fie  $P_m$  polinomul de interpolare Lagrange corespunzător tabelului

$x$	$x_{n-m}$	$\dots$	$x_i$	$\dots$	$x_n$
$y$	$f_{n-m}$	$\dots$	$f_i$	$\dots$	$f_n$

unde  $f_i = f(x_i, y_i)$ ,  $i = n-m, n$ .

Vom aproxima valoarea exactă  $y(x_{n+1})$  prin

$$y_{n+1} = y_n + \int_{x_n}^{x_{n+1}} P_m(x) dx. \quad (40)$$

După cum se știe  $P_m(x) = \sum_{i=n-m}^n L_i(x) f(x_i, y_i)$ , unde

$$L_i(x) = \prod_{\substack{j=n-m \\ j \neq i}}^n \frac{x-x_j}{x_i-x_j} = \prod_{\substack{j=n-m \\ j \neq i}}^n \frac{(t-j+n)h}{(i-j)h} =$$

$$= \frac{(t+m)\dots(t+n-i+1)(t+n-i-1)\dots t}{(i-n+m)\dots 1(-1)\dots[-(n-i)]} = \frac{(-1)^{n-i} \prod_{k=0}^m (t+k)}{(i-n+m)!(n-i)!(t+n-i)}.$$

Făcând schimbarea de variabilă (39) în (40), obținem:

$$y_{n+1} = y_n + \int_0^1 \left( \sum_{i=n-m}^n \frac{(-1)^{n-i} \prod_{k=0}^m (t+k)}{(i-n+m)!(n-i)!(t+n-i)} \cdot h f_i \right) dt$$

Dacă notăm cu

$$A_i = \frac{(-1)^{n-i} \cdot h}{(i-n+m)!(n-i)!} \cdot \int_0^1 \frac{\prod_{k=0}^m (t+k)}{t+n-i} dt, \quad (41)$$

obținem

$$y_{n+1} = y_n + \sum_{i=n-m}^n A_i f_i, \quad (42)$$

cunoscută sub numele de *formula Adams-Bashforth*.

În continuare vom explicita formula (42) pentru valori particulare ale lui  $m$ . Astfel pentru  $m = 1$ , deci când se folosesc nodurile  $x_n$  și  $x_{n-1}$ , formula (42) devine

$$y_{n+1} = y_n + A_{n-1} f_{n-1} + A_n f_n,$$

unde conform (41)

$$A_{n-1} = \frac{(-1)^{-1} h}{0!1!} \cdot \int_0^1 \frac{t(t+1)}{t+1} dt = -h \frac{t^2}{2} \Big|_0^1 = -\frac{h}{2},$$

$$A_n = \frac{(-1)^0 h}{1! \cdot 0!} \cdot \int_0^1 \frac{t(t+1)}{t} dt = h \left( \frac{t^2}{2} + t \right) \Big|_0^1 = \frac{3}{2} h.$$

În consecință pentru  $m = 1$ , formula lui Adams-Bashforth se scrie

$$y_{n+1} = y_n + \frac{h}{2} (3f_n - f_{n-1}). \quad (43)$$

Similar, pentru  $m = 2$  se obține

$$y_{n+1} = y_n + \frac{h}{12} (23f_n - 16f_{n-1} + 5f_{n-2}), \quad (44)$$

iar pentru  $m = 3$

$$y_{n+1} = y_n + \frac{h}{24} (55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3}). \quad (45)$$

În ceea ce privește evaluarea erorii, scăzând (40) din (38) obținem

$$y(x_{n+1}) - y_{n+1} = y(x_n) - y_n + \int_{x_n}^{x_{n+1}} [f(x, y(x)) - P_m(x)] dx,$$

deci

$$|y(x_{n+1}) - y_{n+1}| \leq |y(x_n) - y_n| + \int_{x_n}^{x_{n+1}} |f(x, y(x)) - P_m(x)| dx.$$

Deci eroarea din metoda lui Adams-Bashforth este mai mică decât suma dintre eroarea din metoda Runge-Kutta folosită în calculul lui  $y_n$  și eroarea de la integrarea numerică. În cazul  $m = 1$ , se obține folosind schimbarea de variabilă (39) :

$$\begin{aligned} \int_{x_n}^{x_{n+1}} |f(x, y(x)) - P_m(x)| dx &\leq \frac{M_2}{2} \int_{x_n}^{x_{n+1}} (x - x_n)(x - x_{n-1}) dx = \frac{M_2}{2} \int_0^1 h^3 t(t+1) dt = \\ &= \frac{5}{12} h^3 M_2, \end{aligned}$$

unde

$$M_2 = \sup_{x \in [a, b]} |f''(x, y(x))|.$$

Deci eroarea de integrare în acest caz este de ordinul  $h^3$ .

Să menționăm acum că integrând (1) pe  $[x_{n-1}, x_{n+1}]$ , în locul lui (38) se poate considera

$$y(x_{n+1}) = y(x_{n-1}) + \int_{x_{n-1}}^{x_{n+1}} f(x, y(x)) dx. \quad (46)$$

Se poate proceda apoi ca mai sus. Această metodă este atribuită lui E. J. Nyström (1925). Pentru  $m = 1$ , de exemplu, se obține

$$y_{n+1} = y_{n-1} + 2hf(x_n, y_n). \quad (47)$$

Metodele Adams-Bashforth și Nyström sunt cunoscute ca metode explicite, deoarece relația de recurență (42) sau cea corespunzătoare pentru metoda Nyström nu conțin  $f(x_{n+1}, y_{n+1})$ ; ele exprimă explicit  $y_{n+1}$  în funcție de  $y_n, y_{n-1}, \dots, y_{n-m}$ .

**Exemplul 1.** Folosind metoda Adams-Bashforth de ordin trei să se determine soluția aproximativă a următoarei probleme Cauchy

$$\begin{cases} y' &= y^2 - \frac{y}{x} - \frac{1}{4x^2} \\ y(1) &= 0.5 \end{cases}$$

în punctul  $x=2.25$ , determinând soluția în  $x=2$  cu metoda Runge-Kutta de ordinul patru în patru pași.

Pentru a aplica metoda Runge-Kutta de ordin patru luăm:  $x_0=1$ ,  $y_0=0.5$ ,  $x=2$ ,  $n=4$ ,  $h = \frac{x-x_0}{n} = 0.25$ ,  $x_1=x_0+h=1+0.25=1.25$ ,  $x_2=1.5$ ,  $x_3=1.75$ ,  $x_4=2$  și obținem  $y_1=0.4$ ,  $y_2=0.33333$ ,  $y_3=0.28571$ ,  $y_4=0.25$ .

Pentru a determina valoarea aproximativă a soluției în  $x=2.25$  folosim metoda Adams-Bashforth de ordin trei

$$y_5 = y_4 + \frac{h}{12}(55f_4 - 59f_3 + 37f_2 - 9f_1),$$

unde

$f_2 = f(x_2, y_2) = -0.22222$ ,  $f_3 = f(x_3, y_3) = -0.16327$ ,  $f_4 = f(x_4, y_4) = -0.125$ , obținându-se  $y(2.25) \approx y_5 = 0.22307$  (soluția exactă fiind  $y(2.25) = 0.22222$ ).

**Metoda Adams-Moulton.** Presupunem că printr-o metodă directă am determinat valorile aproximative  $y_1, \dots, y_n$  în nodurile  $x_k = x_0 + kh$ ,  $k = \overline{1, n}$  și că  $x_{n+1} < b$ . Fie  $P_{m+1}$  polinomul de interpolare Lagrange corespunzător tabelului

$x$	$x_{n-m}$	...	$x_n$	$x_{n+1}$
$f$	$f_{n-m}$	...	$f_n$	$f_{n+1}$

Formula corespunzătoare lui (40) este:

$$y_{n+1} = y_n + \int_{x_n}^{x_{n+1}} P_{m+1}(x) dx. \quad (48)$$

Procedând ca mai sus, obținem

$$y_{n+1} = y_n + \sum_{i=n-m}^{n+1} B_i f_i, \quad (49)$$

unde

$$B_i = \frac{(-1)^{n+1-i} h}{(i-n+m)!(n+1-i)!} \cdot \int_0^1 \frac{\prod_{k=1}^m (t+k)}{t+n-i} dt, \quad i = \overline{n-m, n+1}, \quad (50)$$

cunoscută sub numele de *formula Adams-Moulton*.

Vom particulariza acum această formulă. Pentru  $m=0$  se obține

$$y_{n+1} = y_n + \frac{h}{2}(f_{n+1} + f_n), \quad (51)$$

pentru  $m=1$

$$y_{n+1} = y_n + \frac{h}{12}(5f_{n+1} + 8f_n - f_{n-1}), \quad (52)$$

pentru  $m=2$

$$y_{n+1} = y_n + \frac{h}{24}(9f_{n+1} + 19f_n - 5f_{n-1} + f_{n-2}). \quad (53)$$

Deoarece  $f_{n+1} = f(x_{n+1}, y_{n+1})$ , necunoscuta  $y_{n+1}$  apare și în membrul drept, deci, în general nu se poate explicita. De aceea metoda Adams-Moulton este o *metodă implicită*.

De obicei (49) trebuie rezolvată ca o ecuație algebrică printr-o metodă iterativă. Se alege  $y_{n+1}^{(0)}$ , apoi se calculează

$$y_{n+1}^{(1)} = F(y_{n+1}^{(0)}), \quad y_{n+1}^{(2)} = F(y_{n+1}^{(1)}) \text{ etc.},$$

unde  $F$  apare din scrierea convenabilă a lui (49) sub forma  $y_{n+1} = F(y_{n+1})$ .

Pentru a calcula o aproximație bună  $y_{n+1}^{(0)}$ , se poate utiliza o formulă explicită, de exemplu, Adams-Bashforth.

Se poate demonstra următoarea teoremă.

**Teorema 4.** Fie șirul recurent

$$y_{n+1}^{(k+1)} = y_n + \sum_{i=n-m}^n B_i f_i + B_{n+1} f(x_{n+1}, y_{n+1}^{(k)}), \quad k \in \mathbf{N}. \quad (54)$$

Dacă funcția  $f$  satisface condițiile Teoremei 1 și  $h$  este ales astfel încât  $|B_{n+1}| \cdot L < 1$ ,  $L$  fiind constanta lui Lipschitz, atunci șirul  $y_{n+1}^{(k)}$  este convergent și

limita sa  $y_{n+1} = \lim_{k \rightarrow \infty} y_{n+1}^{(k)}$  satisface ecuația  $y_{n+1} = y_n + \sum_{i=n-m}^{n+1} B_i f_i$ .

Eroarea din metoda Adams-Moulton se poate estima ca și în cazul metodei Adams-Bashforth.

**Exemplul 2.** Folosind metoda Adams-Moulton de ordin unu să se determine soluția aproximativă a următoarei probleme Cauchy

$$\begin{cases} y' &= y^2 - \frac{y}{x} - \frac{1}{4x^2} \\ y(1) &= 0.5 \end{cases}$$

în punctul  $x=1.5$ , considerând soluția  $y^{(0)}(1.5)$  obținută cu metoda Euler modificată, cu  $h = 0.05$ . Atunci  $y^{(0)}(1.5) = 0.333406$  și cum

$$y_{n+1}^{(k+1)} = y_n + \frac{h}{12}(5f_{n+1} + 8f_n - f_{n-1}), \quad k=0,1,2,\dots$$

obținem

$$y_5^{(1)} = 0.333403, \quad y_5^{(2)} = 0.333331$$

**Metoda predictor-corector (Adams-Bashforth-Moulton)**

Presupunem că printr-o metodă directă am determinat valorile aproximative  $y_1, \dots, y_n$  în nodurile  $x_1, \dots, x_n$ .

Fie  $m_1, m_2 \leq n$  și  $x_{n+1} = x_n + h \leq b$ .

În prima etapă (*etapa predictor*) se determină valoarea aproximativă  $y_{n+1}$  cu metoda Adams-Bashforth, pentru  $m = m_1$ .

Valoarea astfel determinată se notează cu  $y_{n+1}^{(0)}$ , și este folosită în continuare în etapa a doua (*etapa corector*) pentru determinarea valorii  $y_{n+1}$  cu metoda Adams-Moulton cu  $m = m_2$ .

Cele mai utilizate metode predictor-corector sunt:

$$\begin{aligned}
 1) \quad & \begin{cases} y_{n+1}^{(0)} = y_n + \frac{h}{2}(3f_n - f_{n-1}) & (m_1 = 1) \\ y_{n+1}^{(k+1)} = y_n + \frac{h}{2} [f(x_{n+1}, y_{n+1}^{(k)}) + f_n] & (m_2 = 0) \end{cases} \\
 2) \quad & \begin{cases} y_{n+1}^{(0)} = y_n + \frac{h}{12}(23f_n - 16f_{n-1} + 5f_{n-2}) & (m_1 = 2) \\ y_{n+1}^{(k+1)} = y_n + \frac{h}{12} (5f(x_{n+1}, y_{n+1}^{(k)}) + 8f_n - f_{n-1}) & (m_2 = 1) \end{cases} \\
 3) \quad & \begin{cases} y_{n+1}^{(0)} = y_n + \frac{h}{24}(55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3}) & (m_1 = 3) \\ y_{n+1}^{(k+1)} = y_n + \frac{h}{24} (9f(x_{n+1}, y_{n+1}^{(k)}) + 19f_n - 5f_{n-1} + f_{n-2}) & (m_2 = 2) \end{cases} .
 \end{aligned}$$

**Exemplul 3.** Folosind metoda Adams-Bashforth-Moulton de ordin (3,2) să se determine soluția aproximativă a următoarei probleme Cauchy

$$\begin{cases} y' = y^2 - \frac{y}{x} - \frac{1}{4x^2}, \\ y(1) = 0.5 \end{cases}$$

în punctul  $x=2.25$ , considerând soluția  $y^{(0)}(2.25)$  obținută în exemplul 1.

Ca și în exemplul 1 avem:  $x_0=1$ ,  $y_0=0.5$ ,  $x=2$ ,  $n=4$ ,  $h=0.25$ ,  $x_1=x_0+h=1+0.25=1.25$ ,  $x_2=1.5$ ,  $x_3=1.75$ ,  $x_4=2$  și obținem  $y_1=0.4$ ,  $y_2=0.33333$ ,  $y_3=0.28571$ ,  $y_4=0.25$ .

Pentru a determina valoarea aproximativă a soluției în  $x=2.25$  folosim metoda Adams-Bashforth de ordin trei

$$y_5 = y_4 + \frac{h}{12}(55f_4 - 59f_3 + 37f_2 - 9f_1),$$

și obținem  $y(2.25) \approx y_5 = 0.22307$ . Notăm  $y_5^{(0)} = 0.22307$  și aplicăm în continuare metoda Adams-Moulton de ordin 2. Obținem:  $y_5^{(1)} = 0.22219$ ;  $y_5^{(2)} = 0.22219$ .

În continuare vom aborda un anumit tip de stabilitate pentru a ilustra anumite idei, care nu pot fi prezentate în cazul metodei Euler. Am arătat la metoda Euler, că metodele numerice pentru problema Cauchy au o anumită formă de stabilitate, imitând stabilitatea problemei Cauchy. În particular acest tip de stabilitate caracterizează și metoda (47) dată de

$$y_{n+1} = y_{n-1} + 2hf(x_n, y_n), \quad n \geq 1.$$

Din păcate, această stabilitate nu este satisfăcătoare pentru scopuri practice. Vom arăta această metodă nu este convenabilă în raport cu un anumit sens de stabilitate pe care o vom defini. Deoarece relația de recurență depinde de  $f(x, y)$  este greu să dăm rezultate generale privind stabilitatea numerică a unei astfel de metode. Este instructiv să căutăm, cu metoda de mai sus, soluția numerică a problemei

$$y' = \lambda y, \quad y(0) = 1, \quad \lambda \in \mathbf{R}, \quad (55)$$

a cărei soluție este  $y(x) = e^{\lambda x}$ . Această problemă o vom utiliza ca *problemă model*. Dacă o metodă numerică se comportă rău cu o problemă atât de simplă ca (55), este puțin probabil ca aceasta să fie bună pentru ecuații diferențiale mai complicate. În acest caz (47) devine

$$y_{n+1} = y_{n-1} + 2h\lambda y_n, \quad n \geq 1. \quad (56)$$

Vom calcula soluția exactă a acestei ecuații și o vom compara cu soluția exactă a ecuației (55),  $y(x) = e^{\lambda x}$ . Ecuația (56) este un exemplu de ecuație liniară cu diferențe de ordin 2. Există o teorie generală pentru ecuații liniare cu diferențe de ordin  $p$ . Multe metode pentru rezolvarea ecuațiilor diferențiale au un analog în rezolvarea ecuațiilor cu diferențe, fiind un ghid în a rezolva (56). Vom începe căutând soluții liniar independente pentru ecuații cu diferențe. Acestea sunt combinate sub forma soluției generale.

Similar cu soluțiile exponențiale ale ecuațiilor diferențiale liniare, căutăm soluții pentru (56) de forma

$$y_n = r^n, \quad n \geq 0, \quad (57)$$

pentru un anumit  $r$  necunoscut. Înlocuind în (56), pentru a găsi condiții necesare pentru  $r$ , obținem

$$r^{n+1} = r^{n-1} + 2h\lambda r^n.$$

Împărțind cu  $r^{n-1}$ , rezultă

$$r^2 = 1 + 2\lambda hr. \quad (58)$$

Este valabilă și reciproca. Dacă  $r$  satisface (58), atunci  $y_n$  dat de (57) satisface (56). Ecuația (58) se numește *ecuație caracteristică* pentru metoda (47). Rădăcinile sale sunt

$$r_0 = h\lambda + \sqrt{1 + h^2 \lambda^2}, \quad r_1 = h\lambda - \sqrt{1 + h^2 \lambda^2}. \quad (59)$$

Soluția generală a lui (56) este

$$y_n = \beta_0 r_0^n + \beta_1 r_1^n, \quad n \geq 0. \quad (60)$$

Coeficienții  $\beta_0$  și  $\beta_1$  din (60) se determină din condițiile ca  $y_0$  și  $y_1$  să coincidă cu ce se obține din (60) pentru  $n = 0$  și  $n = 1$ .

$$\begin{cases} \beta_0 + \beta_1 = y_0 \\ \beta_0 r_0 + \beta_1 r_1 = y_1 \end{cases}.$$

Soluția acestui sistem este

$$\beta_0 = \frac{y_1 - r_1 y_0}{r_0 - r_1}, \quad \beta_1 = \frac{y_0 r_0 - y_1}{r_0 - r_1}.$$

Dar  $y_0 = 1$ ,  $y_1 = e^{\lambda h}$  (acestea sunt valorile soluției exacte). Atunci, folosind formula lui Taylor, obținem

$$\beta_0 = \frac{e^{\lambda h} - r_1}{2\sqrt{1 + h^2 \lambda^2}} = 1 + O(h^2 \lambda^2)$$

$$\beta_1 = \frac{r_0 - e^{\lambda h}}{2\sqrt{1 + h^2 \lambda^2}} = O(h^3 \lambda^3).$$

Pentru aceste valori,  $\beta_0 \rightarrow 1$  și  $\beta_1 \rightarrow 0$ , când  $h \rightarrow 0$ . În consecință  $\beta_1 r_1^n \rightarrow 0$  când  $h \rightarrow 0$ , deci din (60) rezultă că termenul  $\beta_0 r_0^n$  ar trebui să corespundă soluției exacte  $e^{\lambda x_n}$ . De fapt

$$r_0^n = e^{\lambda x_n} [1 + O(h^2)].$$

Într-adevăr

$$r_0 = \lambda h + 1 + \frac{1}{2} \lambda^2 h^2 + O(h^4),$$

$$e^{\lambda h} = 1 + \lambda h + \frac{1}{2} \lambda^2 h^2 + O(h^3).$$

Atunci

$$r_0 = e^{\lambda h} + O(h^3) = e^{\lambda h} [1 + O(h^3)],$$

deoarece  $e^{\lambda h} = 1 + O(h)$ .

În consecință

$$r_0^n = e^{\lambda x_n} [1 + nO(h^3)] = e^{\lambda x_n} (1 + O(h^2)).$$

Pentru a vedea dificultatea utilizării formulei (60) în rezolvarea numerică a ecuației (55), să examinăm cu atenție, valorile relative ale lui  $r_0$  și  $r_1$ . Pentru  $0 < \lambda < \infty$  are loc  $r_0 > |r_1| > 0, (\forall)h$ .

Atunci termenul  $r_1^n$  va crește mai puțin rapid decât  $r_0^n$  și termenul corect în (60),  $\beta_0 r_0^n$  va domina. Totuși pentru  $\lambda < 0$ , vom avea  $0 < r_0 < 1$ ,  $r_1 < -1$ ,  $h > 0$ . În consecință,  $\beta_1 r_1^n$  va domina  $\beta_0 r_0^n$  când  $n$  crește pentru  $h$  fixat, necontând cât de mic este  $h$  ales inițial. Termenul  $\beta_0 r_0^n \rightarrow 0$  când  $n \rightarrow \infty$ , pe când termenul  $\beta_1 r_1^n$  crește în magnitudine, alternând ca semn când  $n$  crește. Termenul  $\beta_1 r_1^n$  se numește *soluție parazită* a metodei numerice (56), deoarece nu corespunde unei soluții a ecuației diferențiale originale  $y' = \lambda y$ . Ecuația originală are o familie de soluții cu un parametru, depinzând de valoarea inițială  $y_0$ , dar aproximația (56) are familia de soluții (60), cu doi parametri, care depinde de  $y_0$  și  $y_1$ . Noua soluție  $\beta_1 r_1^n$  este o creație a metodei numerice; pentru problema (55) cu  $\lambda < 0$  ea face ca soluția numerică să se depărteze de soluția corectă când  $x_n \rightarrow +\infty$ . Din cauza acestei comportări, spunem că metoda (47) este *slab stabilă*.

**Exemplul 4.** Fie problema model  $y' = -y$  cu  $y(0) = 1$  și  $h = 0.25$ . Se aplică metoda (47) cu  $y_0 = 1$  și  $y_1$  determinat cu metoda Euler. Pentru  $x_n = 2.25$  soluția  $y_n$  devine negativă și alternează ca semn la fiecare pas.

Se constată că dacă  $\frac{\partial f}{\partial y}$  are semn negativ, atunci instabilitatea slabă apare

uzual în rezolvarea problemei Cauchy prin metoda (47).

$x_k$	$y_k$	$y(x_k)$	$x_k$	$y_k$	$y(x_k)$
0	1	1	1.75	0.89844	0.173774
0.25	0.75	0.77880	2	0.244141	0.135353
0.5	0.625	0.606531	2.25	-0.32227	0.105399
0.75	0.4375	0.472367	2.5	0.260254	0.082085
1	0.40625	0.367879	2.75	-0.162354	0.063928
1.25	0.234375	0.286505	3	0.341431	0.049787
1.5	0.289063	0.223130			

**Exemplul 5.** Fie problema  $y' = x - y^2$ ,  $y(0) = 0$ . Soluția acestei ecuații diferențiale este strict crescătoare pentru  $x \geq 0$ . Dar  $f(x, y) = x - y^2$ , deci  $\frac{\partial f}{\partial y} = -2y < 0$  pentru  $y > 0$ . Ne așteptăm la o anumită instabilitate. Luând  $h = 0.25$  se constată că de la  $x_n = 2.25$  soluția numerică începe să descrească, ajungând în  $x_n = 3.25$  să fie negativă.

$x_k$	$y_k$	$x_k$	$y_k$

0	0	2	1.2914
0.25	0	2.25	1.145864
0.5	0.125	2.5	1.759889
0.75	0.242188	2.75	0.847244
1	0.470673	3	2.775987
1.25	0.631421	3.25	-1.505808
1.5	0.896326	3.5	3.267258
1.75	0.979721	3.75	-5.093296

### **Integrarea numerică a ecuațiilor diferențiale de ordinul întâi în MATLAB**

În MATLAB funcțiile  $ode23(fxy,x0,x,y0)$  și  $ode45(fxy,x0,x,y0)$  sau  $ode23(fxy,x0,x,y0,err,urma)$  și  $ode45(fxy,x0,x,y0,err,urma)$  rezolvă ecuații diferențiale de ordinul întâi

$$y' = f(x,y),$$

$$y(x_0) = y_0,$$

prin metoda Runge-Kutta de ordinul doi, respectiv patru, parametrii având următoarele semnificații:

$fxy$  este numele fișierului de tip  $m$  care conține funcția  $f(x,y)$ ,  
 $(x_0,y_0)$  sunt coordonatele punctului inițial, iar  
 $x$  este punctul în care se cere valoarea aproximativă a soluției  $y$ ,  
 $err$  este precizia soluției, implicit  $10^{-3}$ , respectiv  $10^{-6}$ ,  
 $urma$  atunci când are valoare diferită de zero se tipăresc rezultatele intermediare.

**Exemplu.** Să se determine valoarea aproximativă a soluției următoarei probleme Cauchy

$$y' = xy^2 + x^3 + 1,$$

$$y(0) = 1,$$

în punctul  $x = 2$ , pasul fiind stabilit în mod automat de funcția  $ode23$  ( $ode45$ ).

Se creează fișierul de tip  $m$  numit  $fxy$  care conține  $f(x,y)$  cu secvența

```
% Fișierul cu funcția f(x,y) este de tip m
```

```
function f=fxy(x,y)
```

```
f=x*y^2+x^3+1;
```

după care se apelează funcția  $ode45$  astfel

```
[x,y]=ode45('fxy',0,2,1,0.0001,1);
```

```
disp('Solutia aproximativa intre x0 si x');
```

```
disp(x);
```

```
disp(y);
```

**Exerciții**

Folosind metoda Taylor de ordinul 3 să se găsească soluția aproximativă a următoarelor probleme Cauchy în punctele menționate.

$$1. \quad \begin{cases} y' = xy \\ y(0) = 1 \end{cases} \text{ în } x = 1 \quad \text{în 4 pași.}$$

$$R. \quad h = 0.25$$

$x$	$x_0=1$	$x_1=1.25$	$x_2=1.5$	$x_3=1.75$	$x_4=2$
$y'$	0	0.25781	0.56772	1.00432	
$y''$	1	1.16106	1.70316	2.84558	
$y'''$	0	1.3374	3.26439	6.7164	
$y$	1	1.03125	1.13544	1.3391	1.69659

$$2. \quad \begin{cases} y' = \frac{4x}{y} \\ y(1) = 2 \end{cases} \text{ în } x = 2 \quad \text{în 4 pași.}$$

$$R. \quad h = 0.25$$

$x$	$x_0=1$	$x_1=1.25$	$x_2=1.5$	$x_3=1.75$	$x_4=2$
$y'$	2	2.05128	2.07279	1.00432	
$y''$	-2	-1.81149	-1.5867	2.84558	
$y'''$	0	0.335864	0.3667	6.7164	
$y$	2	2.4375	2.89465	3.36421	3.84191

Să se determine soluția aproximativă a ecuațiilor diferențiale următoare folosind metoda Euler și Euler îmbunătățită.

$$3. \quad \begin{cases} y' = y - \frac{2x}{y} \\ y(0) = 1 \end{cases} \text{ în } x = 1 \quad \text{cu pasul } h = 0.2 .$$

R. Cu metoda Euler pentru

$$x_i = x_0 + ih, \quad y_{i+1} = y_i + hf(x_i, y_i), \quad i=0, 1, \dots, n, \quad n=5,$$

obținem:

$i$	$x_i$	$y_i$	$f(x_i, y_i)$
0	0	1	1
1	0.2	1.2	0.8667
2	0.4	1.3733	0.7805
3	0.6	1.5294	0.7458
4	0.8	1.6786	0.7254
5	1	1.8237	

Cu metoda Euler îmbunătățită

$$y_{i+1} = y_i + \frac{h}{2} [f(x_{i-1}, y_{i-1}) + f(x_{i-1} + h, y_{i-1} + hf(x_{i-1}, y_{i-1}))]$$

obținem:

$x_i$	0	0.2	0.4	0.6	0.8	1
$y_i$	1	1.1867	1.3484	1.4938	1.6272	1.7542

$$4. \quad \begin{cases} y' = -y - \frac{2}{x^2} \\ y(1.5) = \frac{2}{3} \end{cases} \text{ în } x = 2.5 \text{ cu pasul } h = 0.2 .$$

R. Cu metoda Euler pentru

$$x_i = x_0 + ih, \quad y_{i+1} = y_i + hf(x_i, y_i), \quad i=0, 1, \dots, n, \quad n=5,$$

obținem:

$x_i$	1.5	1.7	1.9	2.1	2.3	2.5
$y_i$	0.66667	0.75556	0.77979	0.76898	0.74142	0.70709

Cu metoda Euler îmbunătățită

$$y_i = y_{i-1} + \frac{h}{2} [f(x_{i-1}, y_{i-1}) + f(x_{i-1} + h, y_{i-1} + hf(x_{i-1}, y_{i-1}))]$$

obținem:

$x_i$	1.5	1.7	1.9	2.1	2.3	2.5
$y_i$	0.66667	0.72323	0.73822	0.72971	0.70865	0.68148

Folosind metoda Runge-Kutta de ordinul patru să se determine soluția aproximativă a următoarelor ecuații diferențiale de ordinul întâi în condițiile precizate în fiecare caz în parte.

$$5. \quad \begin{cases} y' = \frac{y}{x} + y^2 - \frac{8}{x^2} \\ y(1) = 2 \end{cases} \text{ în } 5 \text{ pași.}$$

R.  $h = 0.2$

$x$	$x_0=1$	$x_1=1.2$	$x_2=1.4$	$x_3=1.6$	$x_4=1.8$	$x_5=2$
$g_1$	-2	-1.39534	-1.03432	-0.80563	-0.65645	
$g_2$	-1.73521	-1.23279	-0.92905	-0.73552	-0.60971	
$g_3$	-1.61511	-1.17043	-0.89411	-0.71505	-0.59759	
$g_4$	-1.34582	-1.01161	-0.7942	-0.65032	-0.55593	
$y$	2	1.66512	1.42467	1.24218	1.09694	0.97604

$$6. \quad \begin{cases} y' = 0.25y^2 + x^2 \\ y(0) = -1 \end{cases} \text{ în } 4 \text{ pași.}$$

R.  $h = 0.25$

$x$	$x_0=0$	$x_1=0.25$	$x_2=0.5$	$x_3=0.75$	$x_4=1$
$g_1$	0.25	0.28158	0.43039	0.68916	
$g_2$	0.25024	0.34354	0.54889	0.86348	
$g_3$	0.25023	0.34007	0.54305	0.85678	
$g_4$	0.2822	0.43109	0.68984	1.0619	
$y$	-1	-0.93612	-0.84946	-0.71178	-0.49547

7. Folosind metoda Adams-Bashforth de ordin 3 să se determine soluția aproximativă a următoarei probleme Cauchy :

$$\begin{cases} y' = \frac{y^2}{3} - \frac{y}{x} - \frac{3}{x^2} \\ y(3) = 1 \end{cases}$$

în punctul  $x=3.5$ , luând  $h=0.1$  și determinând cu metoda Runge-Kutta de ordin 4 valoarea lui  $y(3.4)$ .

R. Folosind metoda Runge-Kutta de ordin 4 se determină  $y(3.4)=0.88235$ , și aplicând formula Adams Bashforth pentru  $m=3$  găsim  $y(3.5)=0.85714$

8. Folosind metoda Adams-Bashforth de ordin 3 să se determine soluția aproximativă a următoarei probleme Cauchy :

$$\begin{cases} y' = -\frac{y}{x} \\ y(2) = 0.5 \end{cases}$$

în punctul  $x=2.5$ , luând  $h=0.1$  și determinând cu metoda Runge-Kutta de ordin 4 valoarea lui  $y(2.4)$ .

R. Folosind metoda Runge-Kutta de ordin 4 se determină  $y(2.4)=0.41666$ , și aplicând formula Adams Bashforth pentru  $m=3$  găsim  $y(2.5)=0.40000$ .

9. Folosind metoda Adams-Moulton de ordin 2 să se determine soluția problemei

$$\text{Cauchy } \begin{cases} y' = y^2 - \frac{2y}{x} - \frac{2}{x^2} \\ y(1) = 2 \end{cases}$$

în punctul  $x=2.25$ , luând  $h=0.25$  și determinând cu metoda Runge-Kutta de ordin 4 valoarea lui  $y(2.25)$ .

R. Cu metoda Runge-Kutta de ordin 4 se determină valoarea aproximativă  $y(2.25)=0.88717$ , iar după 3 iterații cu metoda Adams-Moulton se obține rezultatul  $y(2.25)=0.88704$ .

10. Folosind metoda Adams-Moulton de ordin 2 să se determine soluția problemei Cauchy

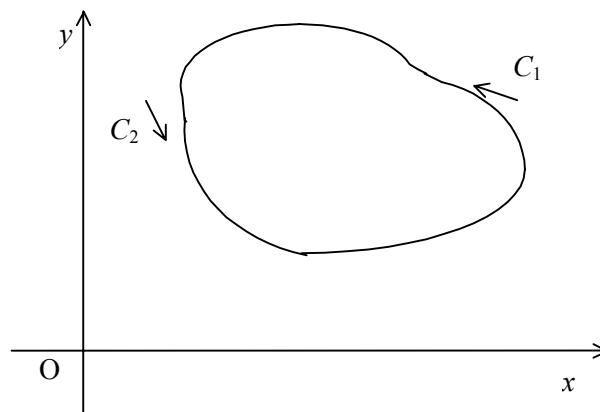
$$\begin{cases} y' = \frac{y^2}{3} - \frac{y}{x} - \frac{3}{x^2} \\ y(3) = 1 \end{cases}$$

în punctul  $x=3.5$ , luând  $h=0.1$  și determinând cu metoda Euler îmbunătățită valoarea lui  $y(3.4)$ .

11. Folosind metoda Adams-Bashforth-Moulton pentru  $m_1=3$  și  $m_2=2$  să se determine soluția aproximativă după 3 iterații a problemei Cauchy de la exercițiul 9.

## 7. Rezolvarea numerică a problemelor la limită pentru ecuații cu derivate parțiale de tip eliptic

Fie  $G \subset \mathbb{R}^2$  o mulțime deschisă, conexă și mărginită, a cărei frontieră  $C$  este netedă pe porțiuni. Așadar  $C$  poate fi o juxtapunere de mai multe curbe. În continuare vom presupune că  $C$  este juxtapunerea a două curbe  $C_1$  și  $C_2$  și este orientată în sens trigonometric.



Considerăm ecuația cu derivate parțiale de ordinul al doilea

$$\Delta u + p(x, y)u = f(x, y), \quad (x, y) \in G, \quad (1)$$

unde

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}, \quad p \in C^{(2)}(G),$$

și  $f$  este continuă pe porțiuni pe  $G$ . Considerăm de asemenea următoarele condiții la limită:

$$u|_{C_1} = \varphi, \quad \text{unde } \varphi \in C^{(0)}(C_1) \text{ este cunoscută}, \quad (2)$$

$$\frac{\partial u}{\partial n} + \alpha \cdot u|_{C_2} = \gamma, \quad (3)$$

unde  $\alpha, \gamma \in C^0(C_2)$  sunt cunoscute, iar  $\frac{\partial u}{\partial n}$  este derivata după normala

exterioară la  $C_2$ .

Problema la limită pentru ecuația (1) constă în determinarea unei funcții  $u \in C^2(G) \cap C^1(\bar{G})$ , care verifică ecuația (1) și condițiile la limită (2) și (3).

Dacă  $p=f=0$  și  $C=C_1$ , obținem *problema Dirichlet* pentru *ecuația Laplace*

$$\begin{cases} \Delta u = 0, \\ u|_C = \varphi. \end{cases} \quad (4)$$

Dacă  $p=f=0$ ,  $C=C_2$  și  $\alpha=0$ , obținem *problema Neumann* pentru *ecuația Laplace*

$$\begin{cases} \Delta u = 0, \\ \frac{\partial u}{\partial n}|_C = \gamma. \end{cases} \quad (5)$$

Dacă  $p=0$  și  $f \neq 0$  obținem *ecuația Poisson*

$$\Delta u = f. \quad (6)$$

Evident și pentru ecuația Poisson putem considera problema Dirichlet sau problema Neumann

$$\begin{cases} \Delta u = f \\ u|_C = \varphi \end{cases}, \text{ respectiv } \begin{cases} \Delta u = f \\ \frac{\partial u}{\partial n}|_C = \gamma \end{cases}.$$

Dacă  $f=0$  și  $p \neq 0$  se obține *ecuația vibrațiilor*  $\Delta u + p(x,y)u=0$ .

Așadar, problema la limită (1)+(2)+(3), deși nu reprezintă cazul cel mai general, este suficient de generală pentru a acoperi cazurile uzuale de probleme la limită pentru ecuații cu derivate parțiale de tip eliptic în plan.

În continuare, notăm cu

$$D = \left\{ u \in C^2(G) \cap C^1(\bar{G}) ; u|_{C_1} = \varphi \right\} \quad (7)$$

și cu

$$\begin{aligned} J(u) = & \iint_G \left[ \frac{1}{2} (\text{grad } u)^2 - \frac{1}{2} p(x,y)u^2 + f(x,y)u \right] dx dy + \\ & + \int_{C_2} \left[ \frac{1}{2} \alpha(s)u^2 - \gamma(s)u \right] ds \end{aligned} \quad (8)$$

și ne punem următoarea *problemă variațională*:

*să se minimizeze funcționala  $J$  pe mulțimea  $D$ .*

Pentru ca această problemă să aibă soluție trebuie mai întâi ca mulțimea  $D$  să fie nevidă. Vom presupune așadar, că există cel puțin o funcție  $\bar{u} \in D$ . Atunci  $D = \bar{u} + D_0$ , unde

$$D_0 = \left\{ h \in C^2(G) \cap C^1(\bar{G}) ; h|_{C_1} = 0 \right\} .$$

Se observă că  $D$  nu depinde de funcția  $\bar{u}$ , în sensul că oricare ar fi  $u_0 \in D$  avem  $D = u_0 + D_0$ .

**Teorema 1.** Dacă există  $u_0 \in D$  astfel încât  $J(u_0) = \min\{J(u) ; u \in D\}$ , atunci  $u_0$  este soluție pentru problema la limită (1)+(2)+(3). Dacă  $p(x, y) \leq 0$  pentru orice  $(x, y) \in G$  și  $\alpha(s) \geq 0$  pentru orice  $s \in C_2$ , atunci are loc și afirmația reciprocă și anume: dacă  $u_0$  este soluție a problemei la limită (1)+(2)+(3), atunci  $J(u_0) = \min\{J(u) ; u \in D\}$  și  $u_0$  este singurul punct de minim al funcției  $J$  pe  $D$ .

**Demonstrație.** Fie  $u_0 \in D$  astfel încât  $J(u_0) = \min\{J(u) ; u \in D\}$ ,

$h \in C^{(1)}(\bar{G})$  cu proprietatea  $h|_{C_1} = 0$  și  $\varphi : [-a, a] \rightarrow \mathbb{R}$ ,  $a > 0$ , definită astfel  $\varphi(t) = J(u_0 + th)$ . Cum

$$\varphi(t) = J(u_0 + th) \geq J(u_0) = \varphi(0),$$

rezultă că  $t = 0$  este un punct de minim pentru  $\varphi$  și deci  $\varphi'(0) = 0$ . Pe de altă parte, ținând seama de (8) avem

$$\begin{aligned} \varphi(t) = & \iint_G \left\{ \frac{1}{2} \left[ \left( \frac{\partial u_0}{\partial x} + t \frac{\partial h}{\partial x} \right)^2 + \left( \frac{\partial u_0}{\partial y} + t \frac{\partial h}{\partial y} \right)^2 \right] \right\} dx dy + \\ & + \iint_G \left( -\frac{1}{2} p(x, y)(u_0 + th)^2 + f(x, y)(u_0 + th) \right) dx dy + \\ & + \int_0^S \left[ \frac{1}{2} \alpha(s)(u_0 + th)^2 - \gamma(s)(u_0 + th) \right] ds , \end{aligned}$$

unde  $S$  este lungimea curbei  $C_2$ , iar

$$\begin{cases} x = x(s) \\ y = y(s) \end{cases} , \quad s \in [0, S]$$

este reprezentarea sa normală.

Un calcul direct ne conduce la

$$\begin{aligned} \varphi'(0) = & \iint_G \left[ \left( \frac{\partial u_0}{\partial x} \frac{\partial h}{\partial x} + \frac{\partial u_0}{\partial y} \frac{\partial h}{\partial y} \right) - p(x, y)u_0 h + f(x, y)h \right] dx dy + \\ & + \int_0^S [\alpha(s)u_0 - \gamma(s)]h ds . \end{aligned} \quad (9)$$

Din formula Green rezultă

$$\iint_G \left( \frac{\partial u_0}{\partial x} \frac{\partial h}{\partial x} + \frac{\partial u_0}{\partial y} \frac{\partial h}{\partial y} \right) dx dy = \iint_G \left[ \frac{\partial}{\partial x} \left( h \frac{\partial u_0}{\partial x} \right) + \frac{\partial}{\partial y} \left( h \frac{\partial u_0}{\partial y} \right) \right] dx dy -$$

$$- \iint_G h \left( \frac{\partial^2 u_0}{\partial x^2} + \frac{\partial^2 u_0}{\partial y^2} \right) dx dy = \oint_C h \frac{\partial u_0}{\partial y} dx + h \frac{\partial u_0}{\partial x} dy - \iint_G h \Delta u_0 dx dy .$$

Deoarece  $h|_{C_1} = 0$ , rezultă că avem

$$\iint_G \left( \frac{\partial u_0}{\partial x} \frac{\partial h}{\partial x} + \frac{\partial u_0}{\partial y} \frac{\partial h}{\partial y} \right) dx dy = \int_{C_2} h \left[ -\frac{\partial u_0}{\partial y} dx + \frac{\partial u_0}{\partial x} dy \right] -$$

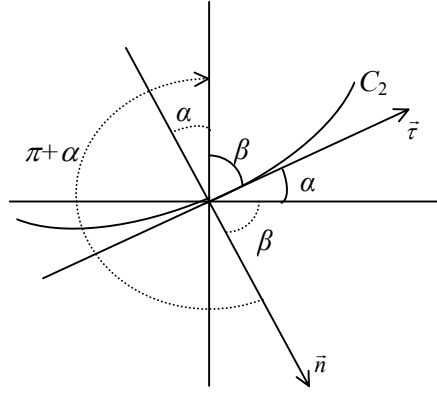
$$- \iint_G h \Delta u_0 dx dy . \quad (10)$$

Dacă notăm cu  $\vec{\tau}$  versorul tangentei la curba  $C_2$ , atunci

$$\vec{\tau} = \frac{dx}{ds} \vec{i} + \frac{dy}{ds} \vec{j} = \cos \alpha \vec{i} + \sin \beta \vec{j} .$$

Pe de altă parte, versorul normalei exterioare la curba  $C_2$  este  $\vec{n} = \cos \beta \vec{i} + \sin(\pi + \alpha) \vec{j}$ .

Ținând seama de aceste observații, mai departe avem



$$\oint_{C_2} h \frac{\partial u_0}{\partial y} dx + h \frac{\partial u_0}{\partial x} dy = \int_0^S h \left( -\frac{\partial u_0}{\partial y} \frac{dx}{ds} + \frac{\partial u_0}{\partial x} \frac{dy}{ds} \right) ds =$$

$$= \int_0^S h \left( \frac{\partial u_0}{\partial x} \cos \beta + \frac{\partial u_0}{\partial y} \cos(\alpha + \pi) \right) ds = \int_0^S h \frac{\partial u_0}{\partial n} ds = \int_{C_2} h \frac{\partial u_0}{\partial n} ds . \quad (11)$$

Cum  $\varphi'(0) = 0$ , din (9), (10) și (11) rezultă

$$0 = \varphi'(0) = \iint_G [-\Delta u_0 - p(x, y)u_0 + f(x, y)] \cdot h dx dy +$$

$$+ \int_{C_2} \left[ \frac{\partial u_0}{\partial n} + \alpha(s)u_0 - \gamma(s) \right] \cdot h ds . \quad (12)$$

Egalitatea (12) are loc pentru orice  $h \in C^{(1)}(\bar{G})$  cu proprietatea  $h|_{C_1} = 0$ ;

în particular și pentru o funcție  $h \in C^{(1)}(\bar{G})$ , nulă pe  $C$ . Atunci obținem

$$\iint_G [-\Delta u_0 - p(x, y)u_0 + f(x, y)] \cdot h dx dy = 0 , \quad (13)$$

pentru orice  $h \in C^{(1)}(\overline{G})$ ,  $h|_{C_1} = 0$ . Dintr-o cunoscută lemă de calcul variațional rezultă

$$-\Delta u_0 - p(x, y)u_0 + f(x, y) = 0,$$

adică ecuația (1).

Înlocuind acum în (12), rezultă

$$\int_{C_2} \left[ \frac{\partial u_0}{\partial n} + \alpha \cdot u_0 - \gamma \right] \cdot h ds = 0,$$

pentru orice  $h \in C^{(1)}(\overline{G})$ ,  $h|_{C_1} = 0$ .

Printr-un raționament asemănător cu cel precedent deducem

$$\frac{\partial u_0}{\partial n} + \alpha \cdot u_0|_{C_2} = \gamma,$$

adică (3).

În continuare demonstrăm afirmația reciprocă. Fie  $u_0$  o soluție a problemei la limită (1)+(2)+(3),  $v \in D$  și  $h = v - u_0$ . Evident  $h \in D_0$ .

Ținând seama de definiția funcționalei  $J$  dată de (8) rezultă

$$\begin{aligned} J(v) - J(u_0) &= J(u_0 + h) - J(u_0) = \\ &= \iint_G \left[ \text{grad } u_0 \text{ grad } h + \frac{1}{2} (\text{grad } h)^2 - pu_0 h^2 - \frac{1}{2} ph^2 + fh \right] dx dy + \\ &\quad + \int_{C_2} \left( \alpha u_0 h + \frac{1}{2} \alpha h^2 - \gamma h \right) ds. \end{aligned}$$

Cum  $f = \Delta u_0 + pu_0$  (deoarece  $u_0$  satisface (1)), mai departe avem

$$\begin{aligned} J(v) - J(u_0) &= \iint_G \left[ \text{grad } u_0 \text{ grad } h + \frac{1}{2} (\text{grad } h)^2 - \frac{1}{2} ph^2 + h\Delta u_0 \right] dx dy + \\ &\quad + \int_{C_2} \left( \alpha u_0 h + \frac{1}{2} \alpha h^2 - \gamma h \right) ds. \end{aligned}$$

Ținând seama de (10) și (11) rezultă

$$\begin{aligned} J(v) - J(u_0) &= \iint_G \left[ \frac{1}{2} (\text{grad } h)^2 - \frac{1}{2} ph^2 \right] dx dy + \\ &\quad + \int_{C_2} \left[ \left( \frac{\partial u_0}{\partial n} + \alpha u_0 - \gamma \right) \cdot h + \frac{1}{2} \alpha h^2 \right] ds. \end{aligned} \quad (14)$$

Deoarece  $u_0$  satisface (3), egalitatea (14) devine

$$J(v) - J(u_0) = \iint_G \left[ \frac{1}{2} (\text{grad } h)^2 - \frac{1}{2} ph^2 \right] dx dy + \int_{C_2} \frac{1}{2} \alpha h^2 ds. \quad (15)$$

Cum  $p(x, y) \leq 0$  pe  $G$  și  $\alpha(s) \geq 0$  pe  $C_2$ , din (15) rezultă

$$J(v) - J(u_0) \geq \frac{1}{2} \iint_G (\text{grad } h)^2 dx dy \geq 0.$$

Observăm că dacă  $h = v - u_0 \neq 0$  atunci  $\iint_G (\text{grad } h)^2 dx dy > 0$ . Într-adevăr, în caz contrar, ar rezulta  $\text{grad } h = 0$  și deci că  $h$  este constantă pe  $\overline{G}$ . Cum  $h|_{C_1} = 0$ , rezultă  $h = 0$ , ceea ce contrazice ipoteza făcută. Așadar  $J(v) > J(u_0)$  dacă  $v \neq u_0$ .

Rămâne să demonstrăm unicitatea elementului  $u_0$ . Fie  $u_1 \in D$ ,  $u_1 \neq u_0$ , astfel încât  $J(u_1) = \min\{J(u) ; u \in D\}$ . Conform celor demonstrate mai înainte rezultă  $J(u_1) > J(u_0)$ .

Analog avem  $J(u_0) > J(u_1)$ . Rezultă astfel o contradicție și cu aceasta teorema este demonstrată.  $\square$

**Observația 1.** Funcționala (8) are sens și pentru funcții  $u$  dintr-o clasă mai largă și anume funcții de clasă  $C^1$ .

Analizând demonstrația Teoremei 1 constatăm că dacă funcția  $u_0$  minimizează funcționala  $J$  pe mulțimea

$$\tilde{D} = \{u \in C^1(G) \cap C^0(\overline{G}) ; u|_{C_1} = \varphi\},$$

atunci  $u_0$  satisface ecuația ce derivă din  $\varphi'(0) = 0$  și anume

$$\iint_G [grad u_0 grad h - pu_0 h + fh] dx dy + \int_{C_2} (cu_0 h - \gamma \cdot h) ds = 0, \quad (16)$$

pentru orice  $h \in C^1(G) \cap C^0(\overline{G})$  cu proprietatea  $h|_{C_1} = 0$ .

Evident  $u_0$  nu mai este soluție (clasică) a problemei la limită (1)+(2)+(3). O funcție din  $\tilde{D}$  care verifică (16) poartă numele de *soluție slabă* a problemei la limită (1) + (2) + (3).

Pentru rezolvarea numerică a problemei la limită (1)+(2)+(3), se consideră o rețea pătratică de drepte paralele cu axele de coordonate:

$$\begin{aligned} x &= x_i = a + ih, \quad i = \overline{1, m}, \\ y &= y_j = b + jh, \quad j = \overline{1, n}, \end{aligned}$$

care acoperă întreg domeniul  $G$ . Punctele  $M_{ij}(x_i, y_j)$  se numesc *nodurile rețelei*, iar  $h$ , *pasul rețelei*.

O primă metodă de rezolvare numerică a problemei la limită (1)+(2)+(3) constă în discretizarea ecuației (1) și a condițiilor la limită (2)+(3) în nodurile rețelei, obținându-se un sistem de ecuații liniare.

Soluția acestui sistem aproximează, în nodurile rețelei, soluția problemei la limită (1)+(2)+(3).

În continuare vom numi această metodă, *metoda rețelelor*. O a doua metodă constă în discretizarea integralei (8) din problema variațională, rezultând o

funcție pătratică, care apoi se minimizează. Vom numi această metodă, *metoda energiei*.

Prezentăm cele două metode pe următorul exemplu.

**Exemplul 1.** Fie  $G$  dreptunghiul  $ABCD$  de laturi  $AB = 5$  și  $AD = 4$ .

Se cere să se determine o funcție  $u \in C^{(2)}(G) \cap C^{(1)}(\bar{G})$  care este soluție pentru ecuația Poisson

$$\Delta u = f(x,y), \quad (x,y) \in G, \quad (17)$$

unde

$$f(x,y) = \begin{cases} 4 & \text{dacă } (x,y) \in G_1 \\ 0 & \text{dacă } (x,y) \notin G_1 \end{cases}$$

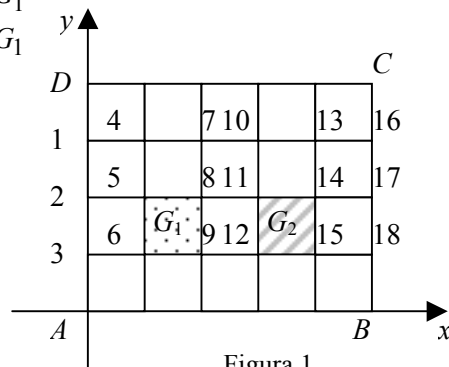


Figura 1

și care verifică condițiile la limită

$$u|_{AB} = u|_{DC} = 0, \quad (18)$$

$$\frac{\partial u}{\partial x} \Big|_{AD} = 0, \quad (19)$$

$$\frac{\partial u}{\partial x} + u|_{BC} = 0. \quad (20)$$

Interpretarea fizică este următoarea: o membrană elastică are marginile  $AB$  și  $CD$  fixe, marginea  $AD$  liberă, iar marginea  $BC$  este rezemată elastic. Funcția căutată  $u = u(x,y)$  reprezintă deplasarea membranei sub acțiunea unei încărcări continue  $f = f(x,y)$ , care este aplicată perpendicular pe membrană.

### §7.1. Metoda rețelelor (a diferențelor finite)

Pentru discretizarea problemei (17)+(18)+(19)+(20) considerăm o rețea pătratică de pas  $h = 1$ . Deoarece  $u = 0$  pe  $AB$  și  $CD$ , nodurile de pe aceste laturi nu prezintă interes. Cele 18 noduri în care urmează să determinăm funcția  $u$

sunt notate în figură. Valorile funcției  $u = u(x,y)$  și ale funcției  $f$  în aceste noduri le vom nota cu  $u_1, u_2, \dots, u_{18}$ , respectiv  $f_1, f_2, \dots, f_{18}$ .

Pentru discretizarea ecuației (17) va trebui să aproximăm derivatele  $\frac{\partial^2 u}{\partial x^2}$

și  $\frac{\partial^2 u}{\partial y^2}$  în nodurile rețelei. Ne propunem să facem această aproximare în nodul

11. Așadar, aproximăm derivatele  $\frac{\partial^2 u}{\partial x^2}$  și  $\frac{\partial^2 u}{\partial y^2}$  în nodul 11 cu derivatele lor numerice în acest nod. Conform (19), §5.4, rezultă

$$\left(\frac{\partial^2 u}{\partial x^2}\right)_{11} = \frac{u_8 - 2u_{11} + u_{14}}{h^2} \quad \text{și} \quad \left(\frac{\partial^2 u}{\partial y^2}\right)_{11} = \frac{u_{10} - 2u_{11} + u_{12}}{h^2}.$$

Înlocuind în ecuația Poisson (17), obținem

$$\frac{u_8 - 2u_{11} + u_{14}}{h^2} + \frac{u_{10} - 2u_{11} + u_{12}}{h^2} = f_{11}$$

și mai departe

$$4u_{11} - u_8 - u_{14} - u_{10} - u_{12} + h^2 f_{11} = 0. \quad (21)$$

În fiecare din cele 12 noduri interioare vom obține câte o ecuație liniară de tipul (21).

Modul de alcătuire al ecuației de tip (21) este pus simbolic în evidență de figura 2.

Dacă nodul este interior, dar este de tipul 4, atunci va trebui să ținem seama că  $u|_{CD} = 0$ .

Ecuația corespunzătoare nodului 4 va fi

$$4u_4 - u_5 - u_7 - u_1 + h^2 f_4 = 0. \quad (21')$$

Așadar, celor 12 noduri interioare le corespund 12 ecuații liniare cu 18 necunoscute  $u_1, u_2, \dots, u_{18}$ . Cele 6 ecuații liniare care lipsesc se obțin din condițiile la limită (19) și (20).

În nodurile de pe laturile  $AD$  și  $BC$  aproximăm derivata  $\frac{\partial u}{\partial x}$  cu derivata numerică dată de (17), §5.4.

De exemplu în nodul 1 avem  $\left(\frac{\partial u}{\partial x}\right)_1 = \frac{-3u_1 + 4u_4 - u_7}{2h}$ .

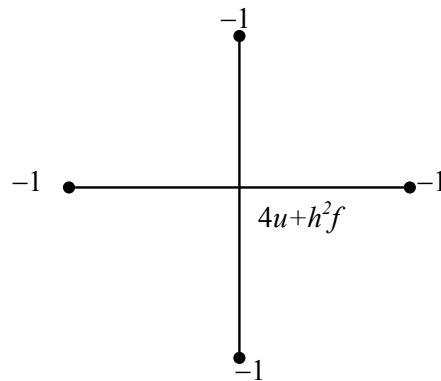


Figura 2

Cum  $\frac{\partial u}{\partial x}|_{AC} = 0$ , rezultă ecuația liniară

$$-3u_1 + 4u_4 - u_7 = 0 . \quad (22)$$

Ecuatii asemănătoare se obțin datorită nodurilor 2 și 3.

Deoarece pe latura  $BC$  avem condiția la limită  $\frac{\partial u}{\partial x} + u|_{BC} = 0$ , în nodul 16

obținem

$$\frac{-3u_{16} + 4u_{13} - u_{10}}{2h} + u_{16} = 0$$

și mai departe

$$-3u_{16} + 4u_{13} - u_{10} + 2hu_{16} = 0 . \quad (23)$$

Ecuatii asemănătoare se obțin datorită nodurilor 17 și 18.

În final se obține un sistem de 18 ecuații liniare cu 18 necunoscute  $u_1, u_2, \dots, u_{18}$ . Rezolvând acest sistem se obțin valorile aproximative ale funcției  $u$  în nodurile rețelei.

## §7.2. Metoda energiei

Problema la limită din Exemplul 1 este un caz particular al problemei la limită (1)+(2)+(3) și anume:

$G$  este dreptunghiul  $ABCD$ ,  $C_1 = AB \cup CD$ ,  $C_2 = AD \cup BC$ ,

$$p(x,y)=0, \quad (x,y) \in G, \quad \varphi|_{C_1} = 0, \quad \alpha|_{AD} = 0, \quad \alpha|_{BC} = 1, \quad \gamma|_{C_2} = 0 .$$

Funcționala  $J$  asociată acestei probleme la limită, va fi un caz particular al funcționalei (8) și anume

$$J(u) = \iint_G \left[ \frac{1}{2} (\text{grad } u)^2 + f(x,y)u \right] dx dy + \frac{1}{2} \int_{BC} u^2 dy . \quad (24)$$

Conform Teoremei 1, problema la limită (17)+(18)+(19)+(20) este echivalentă cu următoarea problemă variațională:

să se găsească o funcție  $u \in C^2(G) \cap C^1(\bar{G})$  care satisface condiția

$$u|_{AB} = u|_{DC} = 0 \quad (25)$$

și care minimizează funcționala (24).

Introducem notațiile:

$$J_1(u) = \iint_G \frac{1}{2} (\text{grad } u)^2 dx dy = \iint_G \frac{1}{2} \left[ \left( \frac{\partial u}{\partial x} \right)^2 + \left( \frac{\partial u}{\partial y} \right)^2 \right] dx dy, \quad (26)$$

$$J_2(u) = \iint_G f(x,y)u(x,y) dx dy, \quad (27)$$

$$J_3(u) = \int_{BC} \frac{1}{2} u^2 dy \quad (28)$$

și considerăm rețeaua pătratică de pas  $h = 1$  din figura 1. Metoda energiei constă în aproximarea integralei  $J = J_1 + J_2 + J_3$  în nodurile rețelei, aproximare în urma căreia se obține o funcție pătratică  $F = F(u_1, u_2, \dots, u_{18})$ .

Problema minimizării funcționalei  $J$  cu condiția la limită (25), se va înlocui cu problema minimizării funcției pătratice  $F$  cu aceeași condiție la limită. Condiția necesară de minim pentru  $F$  și anume,  $\text{grad}F = 0$ , ne conduce la un sistem de 18 ecuații liniare în necunoscutele  $u_1, u_2, \dots, u_{18}$ .

**Observația 1.** Există două avantaje majore ale metodei energiei în raport cu metoda rețelelor. Primul constă în faptul că în metoda energiei nu este necesară discretizarea condițiilor la limită (19) și (20) și nici a derivatelor parțiale de ordinul doi. Al doilea constă în aceea că, termenii pătratici din expresia lui  $F$ , constituie o formă pătratică pozitiv definită. Drept urmare, sistemul de ecuații liniare, care rezultă din minimizarea funcției pătratice  $F$ , este simetric și pozitiv definit și astfel avem acces la metodele de relaxare pentru rezolvarea sa.

Pentru discretizarea integralei  $J_1$  va trebui să aproximăm derivatele  $\frac{\partial u}{\partial x}$  și  $\frac{\partial u}{\partial y}$ . Ne propunem să facem această aproximare în pătratul hașurat,  $G_2$ . Pe segmentul orizontal determinat de nodurile 11 și 14 avem

$$\frac{\partial u}{\partial x} \approx \frac{u_{14} - u_{11}}{h},$$

iar pe segmentul orizontal determinat de nodurile 12 și 15 avem

$$\frac{\partial u}{\partial x} \approx \frac{u_{15} - u_{12}}{h}.$$

Media aritmetică a pătratelor acestor expresii constituie o aproximare bună pentru derivata  $\left(\frac{\partial u}{\partial x}\right)^2$  în centrul pătratului  $G_2$ . Așadar, avem:

$$\left(\frac{\partial u}{\partial x}\right)^2 \approx \frac{1}{2h^2} \left[ (u_{14} - u_{11})^2 + (u_{15} - u_{12})^2 \right]. \quad (29)$$

Un rezultat asemănător obținem pentru  $\left(\frac{\partial u}{\partial y}\right)^2$  și anume:

$$\left(\frac{\partial u}{\partial y}\right)^2 \approx \frac{1}{2h^2} \left[ (u_{14} - u_{15})^2 + (u_{11} - u_{12})^2 \right]. \quad (30)$$

Ținând seama că aria pătratului  $G_2$  este  $h^2$ , din teorema de medie pentru integrala dublă rezultă

$$J_1 \approx \frac{1}{4} \left[ (u_{14} - u_{11})^2 + (u_{15} - u_{12})^2 + (u_{14} - u_{15})^2 + (u_{11} - u_{12})^2 \right]. \quad (31)$$

Pentru aproximarea integralei  $J_2$  folosim metoda trapezelor pentru integrala dublă (§5.3, (1)). Avem

$$J_2 \approx \frac{h^2}{4} [f_{11}u_{11} + f_{14}u_{14} + f_{15}u_{15} + f_{12}u_{12}]. \quad (32)$$

Pentru aproximarea integralei  $J_3$  se folosește metoda dreptunghiurilor și se obține

$$J_3 \approx \frac{h}{2} (u_{16}^2 + u_{17}^2 + u_{18}^2). \quad (33)$$

Așadar, integrala  $J = J_1 + J_2 + J_3$  se va aproxima cu o funcție pătratică  $F = F(u_1, u_2, \dots, u_{18})$ , care provine din adunarea expresiilor (31)+(32)+(33). Funcția pătratică  $F$  se compune dintr-o formă pătratică pozitiv definită provenind din discretizarea integralelor  $J_1$  și  $J_3$  și o formă liniară provenind din discretizarea integralei  $J_2$ . Pentru ca  $F$  să fie minimă trebuie ca  $\text{grad}F = 0$ .

Contribuția celulei  $G_2$  în  $\frac{\partial F}{\partial u_{11}}$  va fi

$$\frac{1}{2} [(u_{11} - u_{12}) - (u_{14} - u_{11})] + \frac{h^2}{4} f_{11} = \frac{1}{2} (-u_{14} + 2u_{11} - u_{12}) + \frac{h^2}{4} f_{11}. \quad (34)$$

Datorită celorlalte trei celule vecine, care au în comun cu  $G_2$ , nodul 11, în  $\frac{\partial F}{\partial u_{11}}$  vor apare și expresiile

$$\frac{1}{2} (-u_8 + 2u_{11} - u_{12}) + \frac{h^2}{4} f_{11}, \quad \frac{1}{2} (-u_{10} + 2u_{11} - u_{14}) + \frac{h^2}{4} f_{11},$$

$$\frac{1}{2} (-u_8 + 2u_{11} - u_{10}) + \frac{h^2}{4} f_{11}.$$

Deoarece variabila  $u_{11}$  intervine în expresia lui  $F$ , numai datorită celulelor care au comun nodul 11, rezultă că  $\frac{\partial F}{\partial u_{11}}$  se compune din suma celor patru expresii de mai sus. Rezultă că

$$\frac{\partial F}{\partial u_{11}} = 4u_{11} - u_{14} - u_{10} - u_8 - u_{12} + h^2 f_{11} = 0. \quad (35)$$

Contribuția nodului 11 în  $\text{grad}F = 0$  este pusă în evidență de schema (în cruce) din Figura 3. Deoarece  $u|_{CD} = 0$ , un nod interior de tipul nodului 4 ne conduce la ecuația

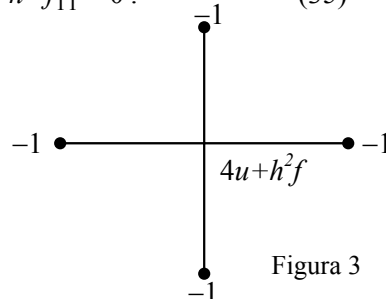


Figura 3

$$\frac{\partial F}{\partial u_4} = 4u_4 - u_5 - u_7 - u_1 + h^2 f_4 = 0. \quad (35')$$

Să analizăm acum contribuția în  $\text{grad}F = 0$  a unui nod de pe  $AD$ , de exemplu a nodului 2. În acest caz sunt numai două celule vecine care au în comun nodul 2. Expresia cu care intervine  $u_2$  în  $F$  va fi

$$\frac{1}{4}[(u_5 - u_2)^2 + (u_1 - u_2)^2] + \frac{h^2}{4} f_2 u_2 + \left[ \frac{1}{4}(u_5 - u_2)^2 + (u_2 - u_3)^2 \right] + \frac{h^2}{4} f_2 u_2.$$

Contribuția nodului 2 în  $\text{grad}F = 0$  revine la

$$\frac{1}{2}[-(u_5 - u_2) - (u_1 - u_2) - (u_5 - u_2) + (u_2 - u_3)] + \frac{h^2}{2} f_2 = 0.$$

Se obține astfel ecuația

$$2u_2 - u_5 - \frac{1}{2}u_1 - \frac{1}{2}u_3 + \frac{h^2}{2} f_2 = 0. \quad (36)$$

Modul de alcătuire a ecuației (36) este pus în evidență de schema din Figura 4.

În cazul nodului 1 vom avea

$$2u_1 - u_4 - \frac{1}{2}u_2 + \frac{h^2}{2} f_1 = 0,$$

deoarece  $u|_{CD} = 0$ . În mod analog, nodului 3 îi

corespunde ecuația  $2u_3 - u_6 - \frac{1}{2}u_2 + \frac{h^2}{2} f_3 = 0$ .

În sfârșit, rămâne să analizăm nodurile de pe latura  $BC$ , de exemplu nodul 17. Și în acest caz avem numai două celule vecine.

Contribuția nodului 17 în  $\text{grad}F$  datorită discretizării integralelor  $J_1$  și  $J_2$  va fi

$$2u_{17} - u_{14} - \frac{1}{2}u_{16} - \frac{1}{2}u_{18} + \frac{h^2}{2} f_{17} = 0.$$

La această expresie trebuie să adăugăm și termenul  $\frac{h}{2} \cdot 2u_{17} = hu_{17}$  care

provine din discretizarea integralei  $J_3$ . Așadar, nodului 17 îi corespunde ecuația

$$(2 + h)u_{17} - u_{14} - \frac{1}{2}u_{16} - \frac{1}{2}u_{18} + \frac{h^2}{2} f_{17} = 0. \quad (37)$$

Modul de alcătuire al ecuației (37) este pus în evidență de schema din Figura 5.

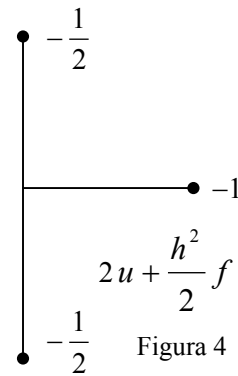


Figura 4

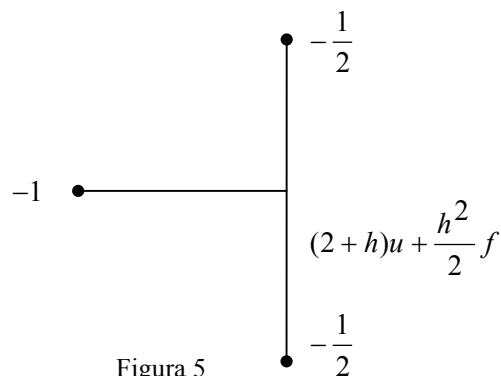


Figura 5



Observăm că matricea sistemului este simetrică, ireductibilă și slab diagonal dominantă. Conform Teoremei 2 din §1.2, rezultă că această matrice este pozitiv definită. Numărul ecuațiilor sistemului liniar, este egal cu numărul nodurilor. Pentru o rețea cu un număr mic de noduri, sistemul se poate rezolva cu metoda Cholesky. De asemenea se poate folosi metode relaxării simple sau metoda Gauss - Seidel.

Dacă rețeaua se alege mai fină, numărul ecuațiilor crește rapid. Cea mai indicată metodă de rezolvare în acest caz este metoda suprarelaxării.

Noi știm să determinăm parametrul optim de relaxare pentru o matrice simetrică, pozitiv definită, diagonal bloc tridiagonală (§1.11, Teorema 2).

În cazul de față, matricea sistemului este simetrică, pozitiv definită și bloc tridiagonală, dar nu este diagonal bloc tridiagonală. Să observăm însă că structura matricei coeficienților este legată de numerotarea nodurilor. Pentru același număr de noduri, dacă se schimbă numerotarea, se schimbă și matricea sistemului. Să considerăm din nou o rețea formată din 18 noduri pe care le împărțim în două părți. O jumătate din noduri au culoarea neagră, iar cealaltă jumătate au culoarea albă.

Numerotarea o facem astfel încât orice segment paralel cu axele unește noduri de culori diferite (vezi Figura 6). Pentru nodul 6, schema în cruce din Figura 3 ne conduce la ecuația

$$4u_6 - u_{13} - u_{14} - u_{15} - u_{16} + h^2 f_6 = 0.$$

Pentru nodul 10, schema din Figura 4 ne conduce la ecuația

$$2u_{10} - \frac{1}{2}u_1 - \frac{1}{2}u_2 - u_3 + \frac{h^2}{2} f_{10} = 0 \text{ etc.}$$

Se obține următorul sistem

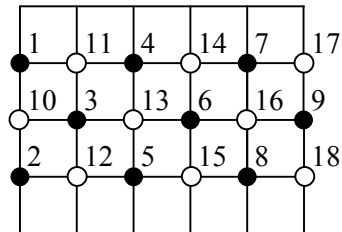


Figura 6

$u_1$	$u_2$	$u_3$	$u_4$	$u_5$	$u_6$	$u_7$	$u_8$	$u_9$	$u_{10}$	$u_{11}$	$u_{12}$	$u_{13}$	$u_{14}$	$u_{15}$	$u_{16}$	$u_{17}$	$u_{18}$	$b$
									$-\frac{1}{2}$	-1								0
	2								$-\frac{1}{2}$	0	-1							0
		4							-1	-1	-1	-1						1
			4							-1	0	-1	-1					0
				4							-1	0	-1	-1				1
					4							-1	0	-1	-1			0
						4							-1	0	-1	-1		0
							4							-1	-1	0	-1	0
								3							-1	$-\frac{1}{2}$	$-\frac{1}{2}$	0

$-\frac{1}{2}$	$-\frac{1}{2}$	-1								2									0
-1	0	-1	-1								4								0
	-1	-1	0	-1								4							1
		-1	-1	-1	-1								4						1
			-1	0	-1	-1								4					0
				-1	-1	0	-1								4				0
					-1	-1	-1	-1								4			0
						-1	0	$-\frac{1}{2}$									3		0
							-1	$-\frac{1}{2}$										3	0

(39)

Matricea acestui sistem este simetrică, pozitiv definită și diagonal bloc tridiagonală.

**Definiție.** Fie  $M = \{1, 2, 3, \dots, n\}$ . O matrice pătratică  $A \in M_n(\mathbb{R})$  se numește de tip (A), dacă există două submulțimi  $S$  și  $T$  ale lui  $M$ , nevide, cu proprietățile: (i)  $S \cup T = M$  (ii)  $S \cap T = \emptyset$  (iii) Dacă  $a_{ij} \neq 0$ , atunci sau  $i = j$  sau  $i \in S$  și  $j \in T$ .

Să observăm că matricea (38) de tipul (A). Într-adevăr, submulțimile  $S = \{1, 3, 5, 7, 9, 11, 13, 15, 17\}$  și  $T = \{2, 4, 6, 8, 10, 12, 14, 16, 18\}$  satisfac proprietățile (i)-(iii).

Se poate arăta că o matrice este de tipul (A), dacă prin permutări simultane de linii și coloane, poate fi adusă la forma diagonal bloc tridiagonală.

O matrice de tipul (A) poate avea mai multe reprezentări diagonal bloc tridiagonale. Matricea (39) este una dintre aceste reprezentări ale matricei (38).

1	2	4	7	10	13
3	5	8	11	14	16
6	9	12	15	17	18

Figura 7

Numerotarea din Figura 6 este tipică pentru aducerea unei matrice de tip (A) la forma diagonal bloc tridiagonală.

O numerotare a nodurilor pe diagonală, ca în Figura 7, conduce de asemenea la o matrice diagonal bloc tridiagonală. Lăsăm în seama cititorului deducerea matricei sistemului de ecuații liniare ce corespunde acestei numerotări.

**Definiție.** Pentru o matrice de tip (A), un sistem de numerotare a nodurilor, căruia îi corespunde o matrice diagonal bloc tridiagonală, se numește consistent.

Reamintim că pentru o matrice simetrică, pozitiv definită, diagonal bloc tridiagonală, parametrul optim de relaxare este

$$\omega_{opt} = \frac{2}{1 + \sqrt{1 - \lambda_1^2}},$$

unde  $\lambda_1$  este cea mai mare valoare proprie a matricei  $-D^{-1}(E+F)$  (Teorema 2, §1.11).

Se poate demonstra că, pentru o matrice de tipul (A), parametrul optim de relaxare este independent de sistemul particular consistent de numerotare a nodurilor.

În cazul exemplului nostru, pentru rețeaua cu 18 noduri avem:

$$\lambda_1 = 0.837319 \text{ și } \omega_{opt} = 1.29306.$$

Pentru a obține o soluție cu 6 zecimale exacte, sunt necesare 50 de iterații cu metoda Gauss-Seidel și numai 16 cu metoda suprarelaxării.

Pentru o rețea cu 77 de noduri,  $\lambda_1 = 0.957686$  și  $\omega_{opt} = 1.5530$ . Pentru a obține o soluție cu 6 zecimale exacte sunt necesare 200 de iterații cu metoda Gauss-Seidel și numai 35 cu metoda suprarelaxării.

În încheierea acestui capitol vom analiza pe scurt cazul când frontiera domeniului  $G$  este o curbă  $C$  (Figura 8).

Considerăm o rețea pătratică, de pas  $h$ , care acoperă domeniul  $G$  și notăm cu  $G'$  domeniul hașurat (format din celulele conținute în interiorul domeniului  $G$ ).

Să presupunem că se cunosc valorile funcției  $u$  pe curba  $C$ .

Notăm cu  $a$  distanța de la nodul 1 la punctul  $R_1$ . Dreapta determinată de punctele  $(0, u_5)$  și  $(a+h, u(R_1))$  are ecuația

$$y - u_5 = \frac{u(R_1) - u_5}{a + h} x.$$

Punând condiția ca pentru  $x = h$ , să rezulte  $y = u_1$ , obținem

$$u_1 = \frac{au_5 + hu(R_1)}{a + h}. \quad (40)$$

Pe parcursul discretizării, variabila  $u_1$  se va înlocui cu expresia din membrul drept al relației (40), astfel încât variabila  $u_1$  va dispărea din sistemul final. Nodul 1 se numește nod eliminat, iar nodul 5 se numește *nod auxiliar*. Nu același lucru se întâmplă cu nodul  $u_2$ . Acesta este nod eliminat din punct de vedere al punctului  $R_2$  de pe frontieră, dar, în același timp este nod auxiliar pentru punctul  $R_3$ . Rezultă că variabila  $u_2$  nu va dispărea din sistemul final.

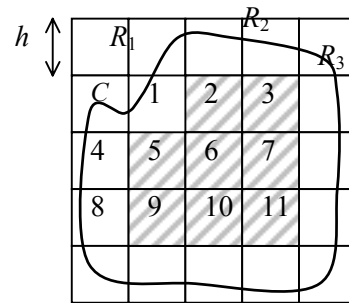


Figura 8

## Exerciții

1. Aplicând metoda rețelei pentru  $h = k = \frac{1}{4}$  să se determine soluția ecuației lui

Laplace,  $\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$ , într-un pătrat cu vârfurile  $A(0,0)$ ,  $B(0,1)$ ,  $C(1,1)$ ,  $D(1,0)$

și cu condițiile la limită următoare:  $u|_{AB} = y$ ,  $u|_{BC} = 30(1-x^2)$ ,  $u|_{CD} = 0$ ,  
 $u|_{AD} = 0$ .

R. Facem următoarele notații:

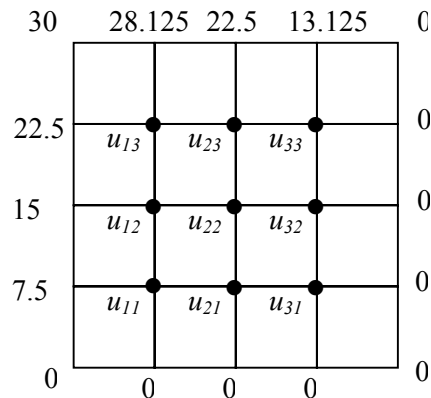
$x_0 = 0$ ,  $y_0 = 0$ ,  $x_i = x_0 + ih$ ,  $i = \overline{0,4}$ ,  $y_j = y_0 + jk$ ,  $j = \overline{0,4}$ ;  $u_{ij} = u(x_i, y_j)$

Determinăm valorile la limită ale funcției  $u$  și obținem:

$$u(0, \frac{1}{4}) = 7.5, \quad u(0, \frac{1}{2}) = 15 \quad (1), \quad u(0, \frac{3}{4}) = 22.5, \quad u(0,1) = 30$$

$$u(\frac{1}{4},1) = 28.125, \quad u(\frac{1}{2},1) = 22.5, \quad u(\frac{3}{4},1) = 13.125.$$

Vom numerota nodurile rețelei ca în figura de mai jos:



Vom înlocui derivatele parțiale de ordin 2 în ecuația lui Laplace, pentru nodurile interioare rețelei.

Obținem astfel ecuațiile:

$$u_{21} + u_{01} + u_{12} + u_{10} - 4u_{11} = 0$$

$$u_{22} + u_{02} + u_{13} + u_{11} - 4u_{12} = 0$$

$$u_{23} + u_{03} + u_{14} + u_{12} - 4u_{13} = 0$$

$$u_{31} + u_{11} + u_{22} + u_{20} - 4u_{21} = 0$$

$$u_{32} + u_{12} + u_{23} + u_{21} - 4u_{22} = 0$$

$$u_{33} + u_{13} + u_{24} + u_{22} - 4u_{23} = 0$$

$$u_{41} + u_{21} + u_{32} + u_{30} - 4u_{31} = 0$$

$$\begin{aligned} u_{42} + u_{22} + u_{33} + u_{31} - 4u_{32} &= 0 \\ u_{43} + u_{23} + u_{34} + u_{32} - 4u_{33} &= 0 \end{aligned}$$

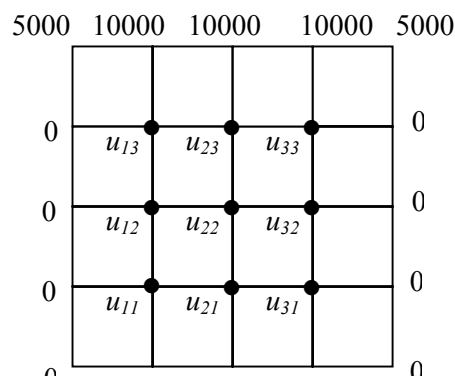
Înlocuind în ecuațiile de mai sus valorile la limita specificate, obținem sistemul de ecuații:

$$\begin{cases} u_{21} + u_{12} - 4u_{11} = -7.5 \\ u_{22} + u_{13} + u_{11} - 4u_{12} = -15 \\ u_{23} + u_{12} - 4u_{13} = -22.5 - 28.125 \\ u_{31} + u_{11} + u_{22} - 4u_{21} = 0 \\ u_{32} + u_{12} + u_{23} + u_{21} - 4u_{22} = 0 \\ u_{33} + u_{13} + u_{22} - 4u_{23} = -22.5 \\ u_{21} + u_{32} - 4u_{31} = 0 \\ u_{22} + u_{33} + u_{31} - 4u_{32} = 0 \\ u_{23} + u_{32} - 4u_{33} = -13.125 \end{cases}$$

Rezolvând acest sistem se obțin valorile:  $u_{11} = 6.077$ ,  $u_{12} = 12.422$ ,  $u_{13} = 19.47$ ,  $u_{21} = 4.386$ ,  $u_{22} = 9.141$ ,  $u_{23} = 14.833$ ,  $u_{31} = 2.327$ ,  $u_{32} = 4.922$ ,  $u_{33} = 8.22$ .

2. Aplicând metoda rețelei pentru  $h = k = \frac{1}{4}$  să se determine soluția ecuației lui

Laplace cu condițiile la limită specificate în figura de mai jos, dacă vârful din stânga jos al plăcii are coordonatele  $(0,0)$ .



R. Din cauza simetriei condițiilor la limită față de axa  $x = \frac{1}{2}$ , vom avea:

$$u_{11} = u_{31}, u_{12} = u_{32}, u_{13} = u_{33} \quad (1)$$

Aceste relații reduc numărul valorilor necunoscute ale funcției  $u$ , în punctele interioare rețelei, la 6.

Vom înlocui derivatele parțiale de ordin 2 în ecuația lui Laplace, pentru nodurile (1,1), (1,2), (1,3), (2,1), (2,2), (2,3).

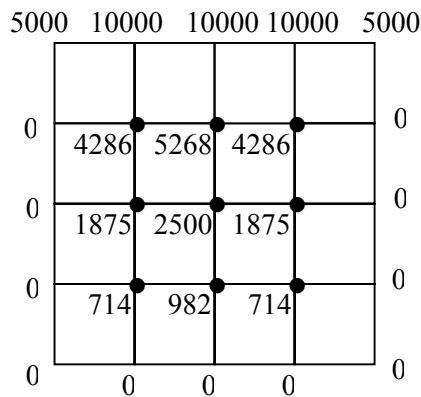
Obținem astfel ecuațiile:

$$\begin{aligned} u_{21} + u_{01} + u_{12} + u_{10} - 4u_{11} &= 0 \\ u_{22} + u_{02} + u_{13} + u_{11} - 4u_{12} &= 0 \\ u_{23} + u_{03} + u_{14} + u_{12} - 4u_{13} &= 0 \\ u_{31} + u_{11} + u_{22} + u_{20} - 4u_{21} &= 0 \\ u_{32} + u_{12} + u_{23} + u_{21} - 4u_{22} &= 0 \\ u_{33} + u_{13} + u_{24} + u_{22} - 4u_{23} &= 0 \end{aligned}$$

Ținând cont că  $u_{i0} = 0, i = \overline{1,3}$ ,  $u_{0j} = 0, j = \overline{1,3}$ ,  $u_{14} = u_{24} = u_{34} = 10000$ , și înlocuind în ecuațiile de mai sus valorile la limita specificate, obținem sistemul de ecuații:

$$\begin{cases} u_{21} + u_{12} - 4u_{11} = 0 \\ u_{22} + u_{13} + u_{11} - 4u_{12} = 0 \\ u_{23} + u_{12} - 4u_{13} = -10000 \\ 2u_{11} + u_{22} - 4u_{21} = 0 \\ 2u_{12} + u_{23} + u_{21} - 4u_{22} = 0 \\ 2u_{13} + u_{22} - 4u_{23} = -10000 \end{cases}$$

Rezolvând acest sistem se obțin valorile:  $u_{11} = 714$ ,  $u_{12} = 1875$ ,  $u_{13} = 4286$ ,  $u_{21} = 982$ ,  $u_{22} = 2500$ ,  $u_{23} = 5268$ , după cum se observă și în figura de mai jos.



3. Problema deformării elastice a unei plăci pătrate sub acțiunea unei forțe constante se reduce la rezolvarea ecuației  $\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 1$ , cu valori la limită egale cu 0. Să se determine soluția ecuației, folosind metoda rețelelor, dacă latura plăcii pătrate se ia egală cu 1, iar distanța  $h = \frac{1}{4}$ .

$$R. u_{11} = u_{13} = u_{33} = u_{31} = 0.0429, \quad u_{12} = u_{32} = u_{21} = u_{23} = 0.0547, \\ u_{22} = 0.0703.$$

4. Fie  $G$  domeniul plan  $ABCDE$  din fig.1,  $AB = 1.5$ ,  $BC = 0.5$ ,  $DE = 1$ ,  $AE = 1$ . Formulăți problema la limită corespunzătoare problemei de minim pentru funcționala

$$J(u) = \frac{1}{2} \iint_G (\text{grad } u)^2 dx dy - \int_{CD} u ds,$$

definită pe mulțimea funcțiilor  $u \in C^{(2)}(G) \cap C^{(1)}(\bar{G})$ , care satisfac:  $u = 0$  pe  $BC$ ,  $u = 1$  pe  $DE$  și  $EA$ .

R. Procedând ca în demonstrația teoremei 1, se obține problema la limită

$$(1) \begin{cases} \Delta u = 0 \text{ în } G \\ \frac{\partial u}{\partial n} = 0 \text{ pe } AB \\ \frac{\partial u}{\partial n} = 1 \text{ pe } CD \\ \frac{\partial u}{\partial n} = 0 \text{ pe } BC, u = 1 \text{ pe } DE \text{ și } EA. \end{cases}$$

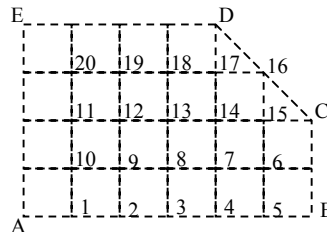


Fig. 1.

5. Fie  $G$  trapezul  $ABCD$ ,  $AB = 1.5$ ,  $DA = 1$ ,  $DC = 0.5$  (fig.2). Să se determine funcționala asociată problemei la limită:

$$(2) \begin{cases} \Delta u + 2 = 0 \text{ în } G \\ u = 0 \text{ pe } AB, CD \text{ și } DA \\ \frac{\partial u}{\partial n} = 0 \text{ pe } BC \end{cases}$$

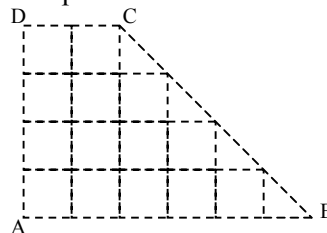


Fig. 2.

$$R. J(u) = \frac{1}{2} \iint_G (\text{grad } u)^2 dx dy - 2 \iint_G u dx dy.$$

6. Fie  $G$  trapezul  $ABCD$ ,  $AB = 1.5$ ,  $DA = 1$ ,  $DC = 0.5$  (fig.2). Să se determine funcționala asociată problemei la limită:

$$(3) \begin{cases} \Delta u = 1 \text{ în } G \\ u = 1 \text{ pe } AB \\ \frac{\partial u}{\partial n} = 0 \text{ pe } BC \\ u = 0 \text{ pe } CD \\ 2u + \frac{\partial u}{\partial n} = 1 \text{ pe } DA \end{cases} .$$

$$R. J(u) = \frac{1}{2} \iint_G (\text{grad } u)^2 dx dy + \iint_G u dx dy + \int_{DA} (u^2 - u) ds .$$

7. Să se discretizeze problema la limită (1), folosind metoda energiei pentru funcționala asociată acestei probleme, alegând rețeaua și numerotarea nodurilor ca în fig.1.

R. Folosim notațiile și tehnica prezentată la metoda energiei. Fie

$$J_1(u) = \frac{1}{2} \iint_G (\text{grad } u)^2 dx dy , \quad J_2(u) = - \int_{CD} u ds .$$

Integrala  $J_1 + J_2$  se va aproxima cu o funcție pătratică  $F = F(u_1, u_2, \dots, u_{19}, u_{20})$ .

Condiția  $\text{grad} F = 0$ , conduce în cazul unui nod interior, de exemplu nodul 7, la ecuația (vezi (35)):

$$\frac{\partial F}{\partial u_7} = 4u_7 - u_4 - u_6 - u_8 - u_{14} = 0 .$$

În cazul nodurilor de pe  $AB$ , cum este cazul nodului 2, ținând seama că nodul 2 este comun la 2 celule, rezultă:

$$\frac{\partial F}{\partial u_2} = 2u_2 - u_9 - \frac{1}{2}u_1 - \frac{1}{2}u_3 = 0 .$$

Ținând seama că  $u = 1$  pe  $AE$  și  $u = 0$  pe  $BC$ , obținem:

$$\frac{\partial F}{\partial u_1} = 2u_1 - u_{10} - \frac{1}{2} - \frac{1}{2}u_2 = 0 ,$$

$$\frac{\partial F}{\partial u_5} = 2u_5 - u_6 - \frac{1}{2}u_4 = 0 .$$

În cazul nodului 17, se procedează ca în cazul nodului 7, ținând seama că  $u = 1$  pe  $DE$ . În consecință:

$$\frac{\partial F}{\partial u_{17}} = 4u_{17} - u_{18} - u_{16} - u_{14} - 1 = 0 .$$

Să analizăm acum cazul nodului 16. Pentru a aproxima  $J_1$  pe celula  $D$ , 16, 17, ținem seama că:

$$\left(\frac{\partial u}{\partial x}\right)^2 \approx \left(\frac{u_{16} - u_{17}}{h}\right)^2, \quad \left(\frac{\partial u}{\partial y}\right)^2 \approx \left(\frac{1 - u_{17}}{h}\right)^2.$$

Deci, pe această celulă,

$$J_1(u) \approx \frac{1}{2} \left[ (u_{16} - u_{17})^2 + (1 - u_{17})^2 \right].$$

Similar, pe celula C,15,16

$$J_1(u) \approx \frac{1}{2} \left[ u_{15}^2 + (u_{16} - u_{15})^2 \right].$$

Pentru a aproxima  $J_2$ , cum ecuația dreptei CD este  $y = -x + 8h$ ,

$ds = \sqrt{2}dx$ , deci

$$J_2(u) = \int_{4h}^{6h} u(x, -x + 8h) \sqrt{2} dx \approx h\sqrt{2}(1 + 0 + 2u_{16}),$$

conform formulei trapezelor.

Ținând seama și de aceasta și aportul celulelor vecine, rezultă:

$$\frac{\partial F}{\partial u_{16}} = 3u_{16} - \frac{3}{2}u_{15} - \frac{3}{2}u_{17} + 2\sqrt{2}hu_{16} = 0,$$

$$\frac{\partial F}{\partial u_{15}} = 5u_{15} - \frac{3}{2}u_{16} - u_{14} - u_6 = 0.$$

Similar obținem:

$$5u_{17} - \frac{3}{2}u_{16} - u_{14} - u_{18} - 1 = 0,$$

$$4u_{18} - u_{17} - u_{13} - u_{19} - 1 = 0$$

$$4u_{20} - u_{11} - u_{19} - 2 = 0.$$

În final se obține următorul sistem de ecuații liniare:

$$AU + b = 0,$$

unde  $A$  este matricea coeficienților necunoscutelor  $u_1, u_2, \dots, u_{18}$ ,

$U = (u_1, u_2, \dots, u_{18})^T$ , iar  $b = (b_1, b_2, \dots, b_{18})^T$ ,

Matricea  $A$  arată astfel:

$$\begin{pmatrix}
 2 & -\frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\
 -\frac{1}{2} & 2 & -\frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\
 0 & -\frac{1}{2} & 2 & -\frac{1}{2} & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\
 0 & 0 & -\frac{1}{2} & 2 & -\frac{1}{2} & 0 & -1 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & -\frac{1}{2} & 2 & -1 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & -1 & 4 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\
 0 & 0 & 0 & -1 & 0 & -1 & 4 & -1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\
 0 & 0 & -1 & 0 & 0 & 0 & -1 & 4 & -1 & 0 & 0 & 0 & -1 & 0 & 0 \\
 0 & -1 & 0 & 0 & 0 & 0 & 0 & -1 & 4 & -1 & 0 & -1 & 0 & 0 & 0 \\
 -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 4 & -1 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 4 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 4 & -1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & -1 & 4 & -1 & 0 & 0 & -1 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & -1 & 4 & -1 & 0 & -1 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 5 & -\frac{3}{2} & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{3}{2} & (3+2\sqrt{2})h & -\frac{3}{2} & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & -\frac{3}{2} & 5 & -1 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & -1 & 4 & -1 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & -1 & 4 & -1 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 4 & -1
 \end{pmatrix}$$

$$b = \left( -\frac{1}{2}, 0, 0, 0, 0, 0, 0, 0, 0, -1, -1, 0, 0, 0, 0, -1, -1, -1, -2 \right)^T .$$

8. Să se discretizeze problema la limită (2), folosind metoda energiei pentru funcționala asociată acestei probleme, alegând rețeaua și numerotarea nodurilor ca în fig.2.

9. Să se discretizeze problema la limită (3), folosind metoda energiei pentru funcționala asociată acestei probleme, alegând rețeaua și numerotarea nodurilor ca în fig.2.

## 8. Introducere în metoda elementului finit

Formularea variațională a diferitelor probleme la limită împreună cu cerințele mai slabe de regularitate conduc în mod natural la metode aproximative de rezolvare numite, de obicei, *metode directe*. Aplicarea acestor metode transformă problema în găsirea punctelor staționare ale unei funcții de un număr finit de variabile reale.

Rezolvarea aproximativă a problemelor la limită pentru ecuații diferențiale și cu derivate parțiale s-a dezvoltat pe trei direcții principale:

- a) *metoda diferențelor finite*,
- b) *metoda elementului finit*,
- c) *metoda elementului de frontieră*.

În metoda diferențelor finite, sistemul de ecuații diferențiale sau cu derivate parțiale valabil pentru orice punct al domeniului de analiză se transformă într-un sistem de ecuații valabile numai pentru anumite puncte ale domeniului, puncte ce definesc rețeaua de discretizare a domeniului.

Dezavantajul principal al acestei metode îl constituie utilizarea unei rețele rectangulare de discretizare a domeniului de analiză. Deci folosirea ei pe domenii cu contururi sau suprafețe curbe introduce o serie de dificultăți și de artificii de calcul. Totodată apar numeroase probleme de stabilitate și de convergență a soluțiilor, fapt ce impune determinarea condițiilor specifice de apariție și respectiv, de evitare a lor, pentru fiecare clasă de probleme.

În metoda elementului finit, se utilizează, ca punct de plecare un model integral al fenomenului studiat. Acest model poate fi obținut, de exemplu, cu ajutorul calculului variațional. Această metodă se bazează pe aproximarea locală pe porțiuni sau subdomenii. Datorită folosirii unui model integral ca bază de plecare și a unor seturi de funcții continue pe porțiuni, metoda elementului finit nu mai este condiționată de existența unei rețele rectangulare. Cu ajutorul ei se pot discretiza practic corpuri geometrice oarecare. Datorită performanțelor sale ridicate, metoda elementului finit a devenit aproape o metodă standard de analiză și proiectare în ingineria construcțiilor și alte domenii.

În acest capitol vom studia *metoda elementului finit*.

### §8.1. Spații Hilbert

Spațiul euclidian  $\mathbb{R}^n$  se distinge printre toate spațiile de dimensiune finită  $n$ , prin faptul că în el este definit un *produs scalar* legat de normă printr-o relație simplă: *pătratul normei unui element este produsul scalar al acestui element cu el însuși*. De aceea este natural să se considere spații în care este definit un produs scalar și norma să fie definită de produsul scalar ca mai sus.

**Definiția 1.** *Spațiul vectorial real  $H$  se numește spațiu prehilbertian dacă pentru fiecare pereche de elemente  $x, y$  din  $H$  este definit un număr real  $\langle x, y \rangle$ , numit produs scalar al elementului  $x$  cu elementul  $y$ , astfel încât sunt îndeplinite următoarele condiții:*

- (i)  $\langle x, y \rangle = \langle y, x \rangle, \quad (\forall) x, y \in H$
- (ii)  $\langle \alpha \cdot x + \beta \cdot y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle, \quad (\forall) x, y, z \in H, \alpha, \beta \in \mathbb{R}.$
- (iii)  $\langle x, x \rangle \geq 0, \quad \langle x, x \rangle = 0 \Leftrightarrow x = \theta_H$

Din definiția produsului scalar rezultă imediat:

- a)  $\langle x, \alpha \cdot y + \beta \cdot z \rangle = \alpha \langle x, y \rangle + \beta \langle x, z \rangle, \quad (\forall) x, y, z \in H, \alpha, \beta \in \mathbb{R}.$
- b)  $\langle x, \theta_H \rangle = \langle \theta_H, x \rangle = 0.$

Ca și în cazul spațiilor euclidiene se poate demonstra

- c)  $|\langle x, y \rangle| \leq \sqrt{\langle x, x \rangle} \cdot \sqrt{\langle y, y \rangle}, \quad (\forall) x, y \in H$  (inegalitatea Cauchy-Buniakowski-Schwarz)

Într-un *spațiu prehilbertian*  $H$  se definește

$$\|x\| = \sqrt{\langle x, x \rangle}, \quad x \in H. \quad (1)$$

Din (iii) și (1) se obține:

- d)  $\|x\| \geq 0, \quad (\forall) x \in H; \quad \|x\| = 0 \Leftrightarrow x = \theta_H$
- e)  $\|\alpha x\| = |\alpha| \|x\|, \quad (\forall) x \in H, \alpha \in \mathbb{R}.$

Totodată din c) rezultă

- f)  $\|x + y\| \leq \|x\| + \|y\|, \quad (\forall) x, y \in H$  (inegalitatea triunghiului).

În concluzie, (1) definește o normă pe  $H$ , deci  $(H, \|\cdot\|)$  este un *spațiu normat*.

**Definiția 2.** *Un șir  $(x_n)_n$  din  $H$  converge la elementul  $x$  din  $H$  și vom nota  $x_n \rightarrow x$ , dacă șirul numeric  $(\|x_n - x\|)_n$  converge la 0, deci dacă pentru orice  $\varepsilon > 0$ , există  $n_\varepsilon \in \mathbb{N}^*$  astfel încât  $\|x_n - x\| < \varepsilon, \quad (\forall) n \geq n_\varepsilon$ . Un șir  $(x_n)_n$  din  $H$*

se numește șir fundamental (Cauchy) dacă pentru orice  $\varepsilon > 0$  există  $n_\varepsilon \in \mathbb{N}^*$  astfel ca  $\|x_n - x_m\| < \varepsilon$ ,  $(\forall) n, m \geq n_\varepsilon$ .

Evident, orice șir convergent este șir Cauchy, afirmația reciprocă nefiind, în general, adevărată.

**Definiția 3.** Un spațiu normat în care orice șir Cauchy este convergent se numește complet (Banach). Un spațiu prehilbertian complet se numește spațiu Hilbert (de la numele matematicianului german D. Hilbert).

Se poate arăta ușor că orice șir convergent este mărginit.

Propoziția următoare semnalează proprietăți simple specifice spațiilor prehilbertiene.

**Propoziția 1.** Fie  $H$  un spațiu prehilbertian

(i)  $\|x + y\|^2 + \|x - y\|^2 = 2(\|x\|^2 + \|y\|^2)$  (identitatea paralelogramului)

(ii) Dacă  $x_n \rightarrow x$  și  $y_n \rightarrow y$ , atunci  $\langle x_n, y_n \rangle \rightarrow \langle x, y \rangle$  (continuitatea produsului scalar)

**Demonstrație.**

(i) Din definiția normei se obține

$$\|x + y\|^2 = \langle x + y, x + y \rangle = \|x\|^2 + 2\langle x, y \rangle + \|y\|^2$$

$$\|x - y\|^2 = \langle x - y, x - y \rangle = \|x\|^2 - 2\langle x, y \rangle + \|y\|^2.$$

Aducând cele două egalități obținem (i). De remarcat că această identitate este generalizarea următoarei proprietăți din geometria elementară: suma pătratelor diagonalelor unui paralelogram este egală cu suma pătratelor laturilor sale.

(ii) Folosind inegalitatea lui Cauchy-Buniakowski-Schwarz obținem

$$|\langle x, y \rangle - \langle x_n, y_n \rangle| \leq |\langle x, y - y_n \rangle| + |\langle x - x_n, y_n \rangle| \leq \|x\| \cdot \|y - y_n\| + \|x - x_n\| \cdot \|y_n\|.$$

Cum șirul  $(y_n)_n$  este mărginit, rezultă că membrul drept al inegalității converge la 0, deci  $\langle x_n, y_n \rangle \rightarrow \langle x, y \rangle$ .  $\square$

Un interes fundamental îl reprezintă *spațiile Hilbert*. Acestea reprezintă generalizarea imediată a spațiilor euclidiene deoarece „geometria” lor este mai apropiată de geometria euclidiană decât geometria oricăror alte spații Banach. Spațiile Hilbert au numeroase proprietăți specifice spațiilor euclidiene care nu sunt generice spațiilor Banach (de exemplu, identitatea paralelogramului). În continuare, vom da un exemplu de spațiu Hilbert, important în teoria ecuațiilor diferențiale și cu derivate parțiale.

**Exemplul 1.** Fie  $G$  o submulțime deschisă, conexă și mărginită a lui  $\mathbb{R}^n$ . Notăm cu  $L^2(G) = \{u : G \rightarrow \mathbb{R}; u \text{ măsurabilă și } \int_G u^2(x) dx < \infty\}$ .

Vom identifica în  $L^2(G)$ , orice două funcții care coincid aproape peste tot (a.p.t) pe  $G$ . Este clar că dacă  $\lambda \in \mathbb{R}$  și  $u \in L^2(G)$ , atunci  $\lambda u \in L^2(G)$ . Fie acum  $u, v \in L^2(G)$ . Din inegalitatea

$$[u(x) + v(x)]^2 \leq 2[u^2(x) + v^2(x)], \quad (\forall) x \in G,$$

obținem că

$$\int_G [u(x) + v(x)]^2 dx \leq 2 \left[ \int_G u^2(x) dx + \int_G v^2(x) dx \right] < \infty.$$

În consecință,  $u+v \in L^2(G)$ , deci  $L^2(G)$  este un spațiu vectorial real. Totodată pentru orice  $u, v \in L^2(G)$ , avem

$$(|u(x)| - |v(x)|)^2 \geq 0, \quad (\forall) x \in G,$$

de unde

$$[u(x)v(x)]^2 \leq \frac{1}{2}[u^2(x) + v^2(x)], \quad (\forall) x \in G,$$

deci are sens numărul real

$$\langle u, v \rangle \stackrel{def}{=} \int_G u(x)v(x) dx. \quad (2)$$

Se verifică ușor că  $L^2(G)$  este un spațiu prehilbertian. Conform (1), pentru orice  $u \in L^2(G)$  se poate defini norma

$$\|u\|_2 = \left( \int_G u^2(x) dx \right)^{1/2}. \quad (3)$$

Se poate demonstra

**Teorema 1.**  $L^2(G)$  este un spațiu Hilbert.

**Definiția 4.** Fie  $H$  un spațiu Hilbert. O mulțime  $D \subset H$  se numește densă în  $H$ , dacă pentru orice  $x \in H$  există un șir  $(x_n)_n$  în  $D$  astfel ca  $x_n \rightarrow x$ .

De remarcat că dacă  $D \subset D' \subset H$  și  $D$  este densă în  $H$ , atunci și  $D'$  este densă în  $H$ .

**Exemplul 2.** Dacă  $G \subset \mathbb{R}^n$  și  $u : G \rightarrow \mathbb{R}$ , definim suportul lui  $u$  ca  $\text{suppu} = \overline{\{x \in G; u(x) \neq 0\}}$ . Considerăm mulțimea  $C_0^\infty(G)$ , a funcțiilor reale  $\varphi$ , cu suport compact în  $G$  (adică anulându-se în afara unei mulțimi compacte din  $G$ , ce depinde de funcția considerată), indefinit derivabile. Aceste funcții vor fi numite *funcții test*. Evident că în raport cu adunarea funcțiilor test și înmulțirea

cu numere reale a funcțiilor test,  $C_0^\infty(G)$  este un spațiu vectorial. Există foarte multe funcții test.

De exemplu, se poate arăta că, pentru orice funcție continuă  $f$ , cu suport compact, există totdeauna o funcție test  $\varphi$  ce o aproximează oricât de bine, adică pentru orice  $\varepsilon > 0$  există  $\varphi$  astfel ca, pentru orice  $x$ ,  $|f(x) - \varphi(x)| < \varepsilon$ .

Vom admite fără demonstrație că mulțimea  $C_0^\infty(G)$  este densă în  $L^2(G)$ .

**Observația 1.** Dacă  $D$  este densă în  $H$  și  $\langle x, y \rangle = 0$ , pentru orice  $y$  din  $D$ , atunci  $x = \theta_H$ .

Într-adevăr, dacă  $z \in H$ , atunci există  $(y_n)_n$  din  $D$  astfel ca  $y_n \rightarrow z$ . Ținând seama de continuitatea produsului scalar, rezultă  $\langle x, z \rangle = 0$ . Alegând  $z = x$ , se obține că  $x = \theta_H$ .

**Definiția 5.** Fie  $H$  un spațiu Hilbert. Elementele  $x, y, \in H$  se numesc ortogonale și se notează  $x \perp y$ , dacă  $\langle x, y \rangle = 0$ . Elementul  $x \in H$  este ortogonal pe mulțimea  $E \subset H$  și se notează  $x \perp E$ , dacă  $x$  este ortogonal pe fiecare element din  $E$ . Mulțimea tuturor elementelor ortogonale pe o mulțime dată  $E$  formează un subspațiu vectorial închis al lui  $H$ , numit complementul ortogonal al mulțimii  $E$  și se notează  $E^\perp$ .

Teorema următoare este fundamentală în teoria spațiilor Hilbert și în rezolvarea aproximativă a unor probleme la limită pentru ecuații diferențiale și cu derivate parțiale.

**Teorema 2.** Fie  $H_0$  un subspațiu închis al spațiului Hilbert  $H$  și  $H_0^\perp$  complementul ortogonal al lui  $H_0$ . Orice element  $x \in H$  se poate reprezenta în mod unic sub forma

$$x = x' + x'' , \quad x' \in H_0 , \quad x'' \in H_0^\perp \quad (4)$$

Mai mult, în  $x'$  se atinge distanța dintre  $x$  și  $H_0$ , adică

$$\|x - x'\| = \min_{y \in H_0} \|x - y\| . \quad (5)$$

**Demonstrație.** Notăm cu  $d = \inf_{y \in H_0} \|x - y\|$  și alegem elementele  $x_n \in H_0$  astfel

ca

$$\|x - x_n\|^2 < d^2 + \frac{1}{n^2} , \quad n = 1, 2, \dots \quad (6)$$

Din identitatea paralelogramului se obține

$$\|x_n - x_m\|^2 + \|(x - x_m) + (x - x_n)\|^2 = 2\left(\|x - x_m\|^2 + \|x - x_n\|^2\right) . \quad (7)$$

Dar

$$\|(x - x_m) + (x - x_n)\|^2 = 4\|x - (x_m + x_n)/2\|^2 .$$

Deoarece  $\frac{x_m + x_n}{2} \in H_0$ , rezultă

$$\|(x - x_m) + (x - x_n)\|^2 \geq 4d^2 . \quad (8)$$

Ținând seama de (6) și (8), din (7) obținem

$$\|x_n - x_m\|^2 \leq 2\left(d^2 + \frac{1}{n^2} + d^2 + \frac{1}{m^2}\right) - 4d^2 = \frac{2}{n^2} + \frac{2}{m^2} .$$

Deci șirul  $(x_n)_n$  este șir Cauchy și cum  $H$  este complet, există  $x' = \lim_{n \rightarrow \infty} x_n$ . Deoarece  $H_0$  este închis, rezultă  $x' \in H_0$ . Trecând la limită în

(6), găsim  $\|x - x'\| \leq d$ , iar din  $x' \in H_0$ , avem  $\|x - x'\| \geq d$ , deci

$$\|x - x'\| = d . \quad (9)$$

Vom arăta acum că  $x'' = x - x'$  este ortogonal pe  $H_0$  și deci  $x'' \in H_0^\perp$ . Fie  $y$  un element nenul din  $H_0$ . Pentru orice  $\lambda \in \mathbb{R}$ , avem  $x' + \lambda y \in H_0$ , deci

$$\|x'' - \lambda y\|^2 = \|x - (x' + \lambda y)\|^2 \geq d^2 ,$$

adică

$$\|x - x'\|^2 - 2\lambda \langle x'', y \rangle + \lambda^2 \|y\|^2 \geq d^2 .$$

Având în vedere (9), obținem

$$-2\lambda \langle x'', y \rangle + \lambda^2 \|y\|^2 \geq 0 .$$

În particular, pentru  $\lambda = \frac{\langle x'', y \rangle}{\|y\|^2}$ , se deduce  $\langle x'', y \rangle^2 \leq 0$ , deci

$\langle x'', y \rangle = 0$ , adică  $x'' \perp y$ . Astfel reprezentarea (4) și relația (5) sunt stabilite.

Rămâne să arătăm unicitatea reprezentării (4). Fie

$x = x'_1 + x''_1$ ,  $x'_1 \in H_0$ ,  $x''_1 \in H_0^\perp$ . Din (4) se obține  $x' - x'_1 = x''_1 - x''$ .

Cum  $x' - x'_1 \in H_0$ , iar  $x''_1 - x'' \in H_0^\perp$ , rezultă  $(x' - x'_1) \perp (x''_1 - x'')$ , deci

$\langle x' - x'_1, x''_1 - x'' \rangle = 0$ . În consecință  $x' = x'_1$ ,  $x''_1 = x''$ . Teorema este demonstrată.  $\square$

**Definiția 6.** Elementele  $x'$  și  $x''$  unic definite de elementul  $x$  se numesc proiecțiile elementului  $x$  pe subspațiul  $H_0$ , respectiv  $H_0^\perp$ .

După cum este cunoscut nu orice șir Cauchy de numere raționale are limită în  $\mathbb{Q}$ , ci în  $\mathbb{R}$ . De fapt în  $\mathbb{R}$  noțiunile de șir convergent și șir Cauchy sunt echivalente (altfel spus,  $\mathbb{R}$  este complet). În mod similar orice spațiu prehilbertian

poate fi inclus într-un spațiu Hilbert (deci complet). Se numește *completatul* unui spațiu prehilbertian  $H$ , cel mai mic spațiu Hilbert care îl conține pe  $H$  ca subspațiu. Un rezultat cunoscut de analiză funcțională precizează că orice spațiu prehilbertian admite un completat. În spațiul completat, vom face distincție între elementele „vechi” din  $H$  și elementele „noi” sau ideale obținute prin completare.

Din teoreme cunoscute ale analizei funcționale rezultă că dacă  $u$  este un element ideal din completat, atunci există un șir de elemente  $(u_n)_n \subset H$ , ce converge la  $u$ , deci  $H$  este dens în completat.

## §8.2. Teorema variațională fundamentală

Fie  $H$  un spațiu Hilbert real,  $D \subset H$  un subspațiu dens și  $A : D \rightarrow H$  un operator liniar.

**Definiția 7.** Operatorul  $A$  se numește strict pozitiv dacă  $\langle Au, u \rangle > 0$ , oricare ar fi  $u \neq \theta_H$ . Operatorul  $A$  se numește simetric dacă  $\langle Au, v \rangle = \langle u, Av \rangle$ , pentru orice  $u, v \in D$ .

În cele ce urmează vom presupune că operatorul  $A$  este simetric și strict pozitiv. Fie  $f \in H$ . Funcționala pătratică

$$F(u) = \langle Au, u \rangle - 2\langle f, u \rangle, \quad u \in D, \quad (10)$$

se numește *funcționala energetică* a operatorului  $A$ . Are loc

**Teorema 3.** Pentru ca  $u_0 \in D$  să realizeze minimul funcționalei energetice este necesar și suficient ca acesta să satisfacă

$$Au_0 = f. \quad (11)$$

Dacă un astfel de element există, el este unic.

**Demonstrație. Necesitatea.** Presupunem că  $u_0 \in D$  realizează minimul funcționalei (10). Fie  $h$  un element arbitrar din  $D$  și  $t$  un număr real arbitrar. Atunci

$$F(u_0 + th) \geq F(u_0). \quad (12)$$

Rezultă că funcția reală  $\varphi(t) = F(u_0 + th)$  își atinge minimul pentru  $t=0$ , deci dacă  $\varphi$  este derivabilă în  $0$ , atunci  $\varphi'(0) = 0$ . Cum  $A$  este simetric, un calcul direct conduce la

$$t^{-1}[\varphi(t) - \varphi(0)] = 2\langle Au_0 - f, h \rangle + t\langle Ah, h \rangle, \quad (\forall) h \in D.$$

Trecând la limită cu  $t \rightarrow 0$ , obținem

$$\langle Au_0 - f, h \rangle = 0, \quad (\forall) h \in D$$

și cum  $D$  este dens în  $H$ , rezultă  $Au_0 = f$  (Observația 1).

*Suficiența*. Să presupunem acum că  $u_0$  satisface ecuația (11). Dacă  $u \in D, u \neq u_0$ , atunci  $u = u_0 + v, v \neq \theta_H$ . Atunci, cum  $A$  este simetric, prin calcul obținem

$$F(u) = F(u_0) + 2\langle Au_0 - f, v \rangle + \langle Av, v \rangle .$$

Dar  $u_0$  satisface ecuația (11), deci

$$F(u) = F(u_0) + \langle Av, v \rangle .$$

Operatorul  $A$  fiind strict pozitiv și  $v \neq \theta_H$ , rezultă că  $\langle Av, v \rangle > 0$  și în consecință  $F(u) > F(u_0)$ . Aceasta înseamnă că în punctul  $u_0$  funcționala (10) își atinge minimumul.

Pentru unicitate, să presupunem că există încă un element  $u_1$  în care  $F$  își atinge minimumul. Conform celor de mai sus  $F(u_1) > F(u_0)$ . În același mod ca mai sus, se poate arăta că  $F(u_0) > F(u_1)$ . Din contradicția obținută rezultă că funcționala (10) își poate atinge minimumul într-un singur punct și teorema este demonstrată.  $\square$

**Observația 2.** Teorema stabilește echivalența între problema rezolvării ecuației  $Au = f$  și aceea a aflării minimumului funcționalei energetice (10); dacă una din aceste probleme este rezolvabilă, atunci și cealaltă este rezolvabilă și soluția uneia dintre ele este și soluția celeilalte. Teorema nu stabilește dacă aceste probleme au soluție. Mai mult, este posibil să nu avem soluție pentru problema formulată.

**Exemplul 3.** Să considerăm următoarea ecuație diferențială foarte simplă

$$-\frac{d^2u}{dx^2} = f(x), \quad x \in (0,1), \quad u(0) = u(1) = 0 . \quad (13)$$

Fie

$$H = L^2((0,1)), \quad Au = -u'', \quad D = \{u \in C^2((0,1)) \cap C^1([0,1]) ; u(0) = u(1) = 0\} .$$

Cum  $C_0^\infty((0,1)) \subset D$ , rezultă că  $D$  este subspațiu dens în  $H$ . Să arătăm că operatorul  $A$  este simetric. Fie  $u, v \in D$ . Atunci, integrând prin părți, obținem

$$\langle Au, v \rangle = -\int_0^1 u''(x)v(x)dx = -u'(x)v(x)\Big|_0^1 + \int_0^1 u'(x)v'(x)dx = \int_0^1 u'(x)v'(x)dx = \langle u, Av \rangle .$$

Totodată  $\langle Au, u \rangle = \int_0^1 u'^2(x)dx > 0$ , deoarece în caz contrar  $u$  este

constantă și din condițiile la limită, rezultă că  $u = 0$ . Conform Teoremei 3, pentru ca  $u_0 \in D$  să fie o soluție a problemei (13), este necesar și suficient ca  $u_0$  să realizeze pe  $D$  minimumul următoarei funcționale

$$F(u) = \int_0^1 \left[ u'^2(x) - 2f(x)u(x) \right] dx . \quad (14)$$

În plus, dacă un astfel de element există, el este unic.

**Exemplul 4.** Fie ecuația diferențială

$$-x''(t) + \sigma(t)x(t) = f(t), \quad t \in (0,1), \quad (15)$$

cu condițiile la limită

$$x(0) = x(1) = 0, \quad (16)$$

iar  $\sigma(t)$  este o funcție pozitivă neidentică nulă și continuă pe  $[0, 1]$ .

Fie

$$H = L^2(0, 1), \quad D = \{u \in C^2((0,1)) \cap C^1([0,1]) ; u(0) = u(1) = 0\} \quad \text{și}$$

$$A : D \rightarrow H, \quad Ax(t) = -x''(t) + \sigma(t)x(t).$$

Pentru  $x, y \in D$ , integrând prin părți avem

$$\langle Ax, y \rangle = \int_0^1 (-x''(t) + \sigma(t)x(t))y(t)dt = \int_0^1 x'(t)y'(t)dt + \int_0^1 \sigma(t)x(t)y(t)dt = \langle x, Ay \rangle,$$

deci  $A$  este simetric.

Totodată

$$\langle Ax, x \rangle = \int_0^1 x'^2(t)dt + \int_0^1 \sigma(t)x^2(t)dt \geq \int_0^1 x'^2(t)dt > 0$$

pentru  $x(t) \neq 0$  (vezi Exemplul 3). Așadar  $A$  este strict pozitiv. Aplicând Teorema 3, rezultă că, pentru ca  $x_0 \in D$  să fie o soluție a problemei (15), (16), este necesar și suficient ca  $x_0$  să realizeze pe  $D$  minimumul funcționalei

$$F(u) = \int_0^1 x'^2(t)dt + \int_0^1 \sigma(t)x^2(t)dt - 2 \int_0^1 f(t)x(t)dt. \quad (17)$$

De asemenea, dacă există un astfel de element, el este unic.

**Exemplul 5.** Fie  $G \subset \mathbb{R}^2$ , o mulțime deschisă, conexă și mărginită și  $C$  frontiera sa. Se caută  $u \in C^2(G) \cap C^1(\bar{G})$ , care satisface

$$-\Delta u = -\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = f \in L^2(G) \quad \text{în } G \quad (18)$$

și condiția la limită

$$u|_C = 0. \quad (19)$$

În acest caz,

$$H = L^2(G), \quad A = -\Delta, \quad D = \{u \in C^2(G) \cap C^1(\bar{G}) ; u|_C = 0\}.$$

Cum  $D$  conține  $C_0^\infty(G)$ , rezultă că  $D$  este densă în  $H$ .

De asemenea, pentru  $u, v \in D$ , conform formulei Green-Riemann, rezultă

$$\langle -\Delta u, v \rangle = \iint_G \left( \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right) v \, dx \, dy = - \oint_C \frac{\partial u}{\partial n} \cdot v \, ds + \iint_G \left( \frac{\partial u}{\partial x} \frac{\partial v}{\partial x} + \frac{\partial u}{\partial y} \frac{\partial v}{\partial y} \right) dx \, dy,$$

$n$  fiind versorul normalei exterioare la  $C$ .

Cum  $v \in D$ ,  $v|_C = 0$ , deci

$$\langle -\Delta u, v \rangle = \iint_G \left( \frac{\partial u}{\partial x} \frac{\partial v}{\partial x} + \frac{\partial u}{\partial y} \frac{\partial v}{\partial y} \right) dx dy = \langle u, -\Delta v \rangle,$$

adică operatorul  $A$  este simetric. În plus

$$\langle -\Delta u, u \rangle = \iint_G \left[ \left( \frac{\partial u}{\partial x} \right)^2 + \left( \frac{\partial u}{\partial y} \right)^2 \right] dx dy > 0, \quad (\forall) u \neq 0,$$

deci  $A$  este strict pozitiv.

Conform Teoremei 3,  $u_0 \in D$  este o soluție a problemei (18), (19) dacă și numai dacă  $u_0$  realizează pe  $D$  minimul funcționalei

$$F(u) = \iint_G \left[ \left( \frac{\partial u}{\partial x} \right)^2 + \left( \frac{\partial u}{\partial y} \right)^2 \right] dx dy - 2 \iint_G f(x, y) u(x, y) dx dy. \quad (20)$$

**Observația 3.** Problema clasică nu are sens decât pentru funcții  $u$  care sunt de clasă  $C^2(G) \cap C^1(\bar{G})$ . Funcționala corespunzătoare are sens pentru funcții de clasă  $C^1(G) \cap C^0(\bar{G})$ . Deci prin trecerea de la problema clasică la funcționala energetică, condițiile de regularitate pot fi slăbite. Altfel spus, problema de minim pentru funcționala energetică se poate pune pe o clasă mai largă de funcții. Spre exemplu, funcționala (20) are sens dacă  $\frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}, f \in L^2(G)$ , deci nu mai este

necesar ca  $u$  să admită derivate parțiale de ordinul al doilea, iar derivatele parțiale de ordinul întâi nu trebuie să fie neapărat continue. Totodată, existența și unicitatea soluției clasice, (adică a problemei (18), (19)) nu se poate garanta dacă funcția  $f$  nu este regulată. Se poate arăta că orice soluție a problemei de minim care este de clasă  $C^2(G)$ , este soluție clasică a problemei considerate.

Altfel spus, problema de minim pentru funcționala energetică se poate pune pe o clasă mai largă de funcții.

### §8.3. Metoda Ritz

Creatorul metodei directe clasice este considerat matematicianul elvețian W. Ritz (1878-1909). Vom considera o funcțională  $F$ , definită pe un spațiu corespunzător  $H$ , de funcții admisibile. Se caută o funcție  $u_0$  astfel ca

$$F(u_0) = \min_{u \in H} F(u) = d. \quad (21)$$

Funcția  $u_0$  care minimizează funcționala se aproximează cu o funcție  $u$  dintr-un subspațiu  $n$ -dimensional oarecare  $K_n \subset H$ . Evident  $F(u) \geq d$ , pentru orice  $u \in K_n$ . Așadar, dacă funcțiile  $\varphi_i(x)$ ,  $i = \overline{1, n}$ , formează o bază a subspațiului  $K_n$ , atunci vom căuta soluția aproximativă sub forma

$$u(x) = \sum_{i=1}^n c_i \varphi_i(x), \quad (22)$$

numerele reale  $c_1, c_2, \dots, c_n$ , urmând a fi determinate. Înlocuind  $u$  dat de (22) în funcționala  $F$ , rezultă  $F(u) = \Phi(c_1, c_2, \dots, c_n)$  și deci problema minimizării funcționalei  $F$  este înlocuită cu problema determinării extremelor funcției  $\Phi: \mathbb{R}^n \rightarrow \mathbb{R}$ . De remarcat că cele două probleme nu sunt echivalente, deoarece s-a trecut de la funcționala  $F$  la funcția  $\Phi$ , prin intermediul funcțiilor  $\varphi_1, \varphi_2, \dots, \varphi_n$ , iar alegerea acestora este la dispoziția noastră; eficiența acestei metode, care se mai numește și metoda Rayleigh-Ritz, depinde în mare măsură de alegerea funcțiilor  $\varphi_1, \varphi_2, \dots, \varphi_n$ . Valorile parametrilor  $c_1, c_2, \dots, c_n$  se determină, după cum se cunoaște, din sistemul de ecuații

$$\frac{\partial \Phi}{\partial c_i} = 0, \quad i = \overline{1, n}, \quad (23)$$

adică

$$\frac{\partial}{\partial c_i} F\left(\sum_{j=1}^n c_j \varphi_j\right) = 0, \quad i = \overline{1, n}.$$

În secțiunile următoare, vom arăta pe exemple concrete, cum se aleg funcțiile  $\varphi_1, \dots, \varphi_n$ .

În continuare, vom prezenta metoda Rayleigh-Ritz ca metodă de cea mai bună aproximare „în energie”. Fie  $D$  un subspațiu dens al unui spațiu Hilbert  $H$ , iar  $A: D \rightarrow H$  un operator liniar, simetric și pozitiv definit. Presupunem că pentru un  $f \in H$  dat, ecuația  $Au = f$  admite o soluție unică  $u_0 \in D$ . Fiind dat un subspațiu  $n$ -dimensional  $K_n \subset D$ , vrem să aproximăm soluția prin  $u_n \in K_n$ ,  $K_n = Sp(\{\varphi_1, \dots, \varphi_n\})$ . Deci căutăm  $u_n = c_1 \varphi_1 + \dots + c_n \varphi_n$  astfel ca  $\|u_0 - u_n\|$  să fie mică. În ipotezele formulate asupra operatorului  $A$ , vom defini un produs scalar, numit „produs energie” în  $D$ , astfel

$$\langle u, v \rangle_A = \langle Au, v \rangle, \quad \text{iar } \|u\|_A = \sqrt{\langle Au, u \rangle}.$$

Vom nota cu  $H_A$  completatul lui  $D$  în raport cu norma  $\|\cdot\|_A$ . Spațiul  $H_A$  se numește spațiul energetic al operatorului  $A$ . Conform Teoremei 2, există și este unic un element  $u_n \in K_n$ , element de cea mai bună aproximare, adică

$$\|u_0 - u_n\|_A = \min_{v \in K_n} \|u_0 - v\|_A. \quad (24)$$

**Definiția 8.** Vectorul unic  $u_n$  cu proprietatea (24) se numește aproximanta Rayleigh-Ritz a soluției  $u_0$  după subspațiul finit dimensional  $K_n$ .

Dacă  $K_n = Sp(\{\varphi_1, \dots, \varphi_n\})$ , atunci aproximanta Rayleigh-Ritz a soluției  $u_0$  a ecuației  $Au = f$  este dată de

$$u_n = c_1 \varphi_1 + \dots + c_n \varphi_n.$$

Fie funcția

$$g(c_1, c_2, \dots, c_n) = \|u_n - u_0\|_A^2.$$

Determinăm  $(c_1, \dots, c_n)$  punctul de minim al funcției  $g$ . Deoarece

$$g(c_1, \dots, c_n) = \langle A(u_n - u_0), u_n - u_0 \rangle = \sum_{i=1}^n \left[ \sum_{j=1}^n c_i c_j \langle A\varphi_i, \varphi_j \rangle - 2c_i \langle f, \varphi_i \rangle \right] + \langle Au_0, u_0 \rangle,$$

din condițiile

$$\frac{\partial g}{\partial c_i} = 0, \quad i = \overline{1, n},$$

rezultă că  $c_1, \dots, c_n$  sunt soluții ale sistemului liniar

$$\sum_{j=1}^n \langle \varphi_i, \varphi_j \rangle_A c_j = \langle f, \varphi_i \rangle, \quad i = \overline{1, n}, \quad (25)$$

sistem care s-ar putea obține direct din (23), ținând seama de formula (10), care dă funcționala energetică a operatorului  $A$ .

### §8.4. Metoda lui Kantorovici (metoda semidiscretă)

Metoda constă în căutarea soluției aproximative sub forma

$$u = \sum_{i=1}^n \alpha_i \cdot \varphi_i,$$

unde coeficienții  $\alpha_i$ ,  $i = \overline{1, n}$ , nu mai sunt scalari, ci funcții de una din variabilele independente, de exemplu  $x_1$ , iar funcțiile  $\varphi_i$  sunt funcții de variabilele rămase,  $x_2, \dots, x_m$ , adică

$$u(x_1, \dots, x_m) = \sum_{i=1}^n \alpha_i(x_1) \varphi_i(x_2, \dots, x_m).$$

Această metodă se leagă de numele matematicianului rus L. V. Kantorovici și stă la baza metodei elementului finit de rezolvare a problemelor nestaționare (dependente de timp).

**Exemplul 6.** Fie  $G = \{(x, y); -\frac{\pi}{2} < x, y < \frac{\pi}{2}\}$ . Să aplicăm metoda lui Kantorovici la rezolvarea aproximativă a ecuației

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 2, \quad (26)$$

care satisface condițiile la limită

$$\begin{aligned}
 u\left(x, \pm \frac{\pi}{2}\right) &= 0, \quad |x| \leq \frac{\pi}{2}, \\
 u\left(\pm \frac{\pi}{2}, y\right) &= 0, \quad |y| \leq \frac{\pi}{2}.
 \end{aligned}
 \tag{27}$$

Se alege ca subspațiu aproximant  $K_n$ , subspațiul funcțiilor de un singur argument  $y$ , care conform (27) satisfac

$$\varphi_i\left(-\frac{\pi}{2}\right) = \varphi_i\left(\frac{\pi}{2}\right) = 0, \quad i = \overline{1, n}.$$

Soluția aproximativă se caută de forma

$$u(x, y) = \sum_{i=1}^n \alpha_i(x) \varphi_i(y), \tag{28}$$

unde funcțiile  $\alpha_i$ ,  $i = \overline{1, n}$ , se determină astfel ca  $u$  să minimizeze funcționala  $F$  corespunzătoare problemei date. În acest caz

$$F(u) = \iint_G \left[ \left( \frac{\partial u}{\partial x} \right)^2 + \left( \frac{\partial u}{\partial y} \right)^2 + 4 \right] dx dy.$$

Ținând seama de (28), avem

$$F(u) = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \Phi(\alpha_1, \alpha_2, \dots, \alpha_n) dx = J(\alpha_1, \alpha_2, \dots, \alpha_n),$$

unde

$$\begin{aligned}
 \Phi(\alpha_1, \alpha_2, \dots, \alpha_n) &= \sum_{i=1}^n \sum_{j=1}^n \left\{ \alpha_i \alpha_j \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{d\varphi_i}{dy} \frac{d\varphi_j}{dy} dy + \frac{d\alpha_i}{dx} \frac{d\alpha_j}{dx} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \varphi_i \varphi_j \cdot dy \right\} + \\
 &+ \sum_{i=1}^n \alpha_i \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} 4\varphi_i dy.
 \end{aligned}
 \tag{29}$$

Deoarece funcțiile  $\varphi_i$ ,  $i = \overline{1, n}$ , sunt cunoscute, integralele în (29) se pot calcula exact. Se pune deci problema determinării extremalelor funcționalei  $J(\alpha_1, \alpha_2, \dots, \alpha_n)$ . Conform unui rezultat clasic de calcul variațional, coeficienții  $\alpha_i$ ,  $i = \overline{1, n}$  sunt dați de sistemul Euler-Lagrange

$$\frac{\partial \Phi}{\partial \alpha_i} - \frac{d}{dx} \left( \frac{\partial \Phi}{\partial \alpha_i'} \right) = 0, \quad i = \overline{1, n}.$$

În consecință funcțiile necunoscute  $\alpha_i$ ,  $i = \overline{1, n}$ , care apar în soluția aproximativă (28), se obțin din sistemul de ecuații diferențiale

$$\sum_{j=1}^n (\alpha_j c_{ij} - \alpha_j'' d_{ij}) = b_i, \quad i = \overline{1, n}, \quad (30)$$

unde

$$c_{ij} = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{d\varphi_i}{dy} \frac{d\varphi_j}{dy} dy, \quad d_{ij} = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \varphi_i \varphi_j dy, \quad b_i = - \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} 2\varphi_i dy,$$

cu condițiile

$$\alpha_i \left( \frac{\pi}{2} \right) = \alpha_i \left( -\frac{\pi}{2} \right) = 0, \quad i = \overline{1, n}.$$

În general, metoda semidiscretă se poate aplica cu condiția ca problema unidimensională să poată fi rezolvată nemijlocit și exact.

### §8.5. Metoda lui Galerkin

Am prezentat în §8.3 metoda Ritz pentru determinarea soluției aproximative a ecuației

$$Au = f. \quad (31)$$

În ipotezele formulate acolo, soluția ecuației  $Au = f$  minimizează funcționala energetică

$$F(u) = \langle Au, u \rangle - 2\langle f, u \rangle. \quad (32)$$

Astfel (vezi Exemplul 3), dacă  $A = -\Delta$ , funcționala energetică este

$$F(u) = \iint_G \left[ \left( \frac{\partial u}{\partial x} \right)^2 + \left( \frac{\partial u}{\partial y} \right)^2 \right] dx dy - 2 \iint_G f u dx dy. \quad (33)$$

Utilizarea integrării prin părți, adică a formulei Green, pentru transformarea funcționalei (32) într-o formă care cere o regularitate mai slabă a funcțiilor admisibile (cum se întâmplă de exemplu în (33)), este unul din succesele

de bază ale metodei elementului finit. Aproximanta Ritz  $u = \sum_{i=1}^n c_i \varphi_i$  a soluției problemei variaționale satisface

$$\frac{\partial}{\partial c_i} F(u) = 0, \quad i = \overline{1, n},$$

adică

$$\langle Au, \varphi_i \rangle - \langle f, \varphi_i \rangle = 0, \quad i = \overline{1, n}, \quad (34)$$

numai când avem un operator  $A$  de forma cerută (de remarcat, că, sistemul (34) nu este altceva decât o rescriere a sistemului (25)).

*Ideea metodei lui Galerkin este de a considera soluții aproximative pentru ecuația (31), de forma de mai sus, unde coeficienții ei se determină din sistemul*

$$\langle Au - f, \varphi_i \rangle = 0, \quad i = \overline{1, n}, \quad (35)$$

*chiar dacă  $A$  nu satisface condițiile din §8.3. Astfel de soluții aproximative au fost considerate de matematicianul rus B. G. Galerkin (1878-1945).*

*Așadar sistemul (35) se poate utiliza chiar dacă operatorul  $A$  este neliniar.*

*În consecință, metoda elementului finit se poate utiliza pentru rezolvarea unei clase largi de probleme, mult mai interesante decât clasa problemelor care provin din probleme variaționale. Totuși, este de dorit ca (35) să poată fi integrat prin părți, pentru a slăbi regularitatea cerută funcțiilor  $\varphi_i$ .*

Deci metoda lui Galerkin este absolut generală. Ea se poate aplica cu succes la ecuații de tipuri diferite: eliptice, hiperbolice, parabolice, chiar dacă ele nu sunt legate de probleme variaționale, ceea ce reprezintă un avantaj față de metoda lui Ritz. Totuși, pentru probleme legate de probleme variaționale, ea se găsește într-o interdependență strânsă cu metoda lui Ritz, iar în multe cazuri este echivalentă cu aceasta din urmă, în sensul că ambele conduc la aceeași soluție aproximativă.

**Exemplul 7.** *Vom prezenta acum o problemă de tip Neumann-Dirichlet pentru operatorul lui Laplace în dimensiune 2. Fie  $G$  o mulțime deschisă și conexă din  $\mathbb{R}^2$ , cu frontiera  $C$  netedă pe porțiuni.*

*De asemenea, fie  $C_1, C_2$  o partiție a lui  $C$ , lungimea lui  $C_1$  fiind strict pozitivă și  $\vec{n}$  versorul normalei exterioare la  $C$ .*

*Să considerăm acum problema clasică următoare:*

*să se găsească funcția  $u \in C^2(G) \cap C^1(\overline{G})$  astfel ca*

$$-\Delta u = f \text{ în } G, \quad (36)$$

$$u = 0 \text{ pe } C_1, \quad (37)$$

$$\frac{\partial u}{\partial n} = g \text{ pe } C_2, \quad (38)$$

*$f$  fiind o funcție reală definită și continuă pe  $G$ , iar  $g$  o funcție reală definită și continuă pe  $C_2$ .*

*De remarcat că este vorba de ecuația lui Poisson, cu membrul secund  $-f$  (semnul "−" se introduce din motive tehnice). Condiția Dirichlet pe  $C_1$  este omogenă, ceea ce nu este restrictiv. Într-adevăr, dacă  $u$  satisface  $-\Delta u = f$  în*

*$G$ ,  $u = h$  pe  $C_1$  și  $\frac{\partial u}{\partial n} = g$  pe  $C_2$  și dacă știm să găsim o funcție  $u_0$  suficient de regulată, care ia valorile  $h$  pe  $C_1$ , atunci funcția  $\tilde{u} = u - u_0$  verifică (36),*

(37), (38) cu  $f$  înlocuit cu  $f + \Delta u_0$  și  $g$  înlocuit cu  $g - \frac{\partial u_0}{\partial n}$ . A trece de la  $h$  pe  $C_1$  la  $u_0$  pe  $G$  înseamnă că se face o prelungire,  $u_0$  fiind o prelungire a lui  $h$  (există o infinitate).

Datorită prelungirii  $h \rightarrow u_0$  și apoi a translației  $u \rightarrow \tilde{u} = u - u_0$ , rezultă că ipoteza condiției Dirichlet omogene pe  $C_1$  nu este o restricție.

În continuare, vom multiplica ecuația cu derivate parțiale cu o funcție test, apoi vom integra pe  $G$ , utilizând formula lui Green și ținând seama de condițiile la limită (37), (38).

Fie acum spațiile vectoriale reale  $W$  și  $V$ , definite astfel:

$$W = \left\{ v: G \rightarrow \mathbb{R} ; v \in C^2(G) \cap C^1(\bar{G}), v = 0 \text{ pe } C_1 \right\},$$

$$V = \left\{ v: G \rightarrow \mathbb{R} ; v \in C^1(G) \cap C^0(\bar{G}), v = 0 \text{ pe } C_1, \text{grad } v \text{ marginit pe } G \right\}.$$

Problema clasică (36)-(38) se poate formula astfel:

găsiți  $u \in W$  care verifică (36) și (38).

De remarcat că integrând prin părți, dacă  $u \in W$  și  $v \in V$ , are loc prima formulă Green:

$$-\iint_G \Delta u \cdot v \, dx dy = \iint_G \text{grad}(u) \cdot \text{grad}(v) \, dx dy - \int_{C_2} \frac{\partial u}{\partial n} v \, ds. \quad (39)$$

Din această identitate, ținând seama de (36) și (38), rezultă că

$$a(u, v) = \iint_G f \cdot v \, dx dy + \int_{C_2} g \cdot v \, ds, \quad (\forall) v \in V, \quad (40)$$

unde

$$a(u, v) = \iint_G \text{grad } u \cdot \text{grad } v \, dx dy.$$

Așadar,  $a(\cdot, \cdot)$  este o formă biliniară și simetrică.

Mulți autori prezintă ca problemă variațională asociată problemei (36)-(38) următoarea problemă:

găsiți  $u \in V$  astfel încât să aibă loc (40).

Problema clasică nu are sens decât pentru funcții având regularitatea lui  $W$ . Pentru problema variațională, este suficientă regularitatea lui  $V$ . Așadar trecând de la problema clasică la problema variațională, condițiile de regularitate au fost slăbite. Se poate arăta că orice soluție  $u$  a problemei variaționale care este în  $W$ , este soluție a problemei clasice.

Totodată  $u \in V$  este soluție a problemei variaționale dacă și numai dacă minimizează pe  $V$  funcționala

$$F(v) = \frac{1}{2} a(v, v) - \left[ \iint_G f v \, dx dy + \int_{C_2} g v \, ds \right]$$

(comparați cu (33)). Formularea variațională permite introducerea, explicarea și justificarea metodelor numerice. Pentru a discretiza problema clasică în elemente finite, este nevoie să punem în prealabil problema sub forma variațională. Nu este

deloc necesară funcționala  $F$ . Ea este introdusă din simplul motiv că în numeroase probleme similare, dar de interes fizic sau mecanic, funcționala  $F(v)$  are o interpretare mecanică interesantă, legea fizică corespunzătoare scriindu-se adesea sub forma unei probleme de minim.

În concluzie sistemul (35) se scrie, după integrarea prin părți, astfel:

$$a(u, \varphi_i) = \iint_G f \cdot \varphi_i \, dx dy + \int_{C_2} g \cdot \varphi_i \, ds, \quad i = \overline{1, n}. \quad (41)$$

## §8.6. Aproximarea funcțiilor

În mod obișnuit, elementele finite se definesc în cadrul procesului de discretizare ca rezultat al descompunerii unui domeniu de studiu în mai multe subdomenii cu interior disjunct. Conexiunea acestor domenii se face prin intermediul nodurilor, care nu sunt altceva decât puncte selectate în domeniul considerat în care se specifică valorile funcției studiate sau ale funcției și ale derivatelor sale până la un anumit ordin. Într-un sens mai larg, elementul finit apare ca un model de aproximare cu proprietăți fizice, geometrice și funcționale. Geometric, elementul finit reproduce într-o formă idealizată părți dintr-un corp supus analizei.

În problemele în care funcția este dată implicit de o ecuație (diferențială, integrală, etc.) valorile funcției sunt parametrii necunoscuți ai problemei. În problemele de interpolare, valorile funcției sunt cunoscute de la început.

Așadar, funcțiile de interpolare permit aproximarea funcțiilor având ca puncte de reper valorile nodale ale funcției sau valorile nodale ale funcției și ale derivatelor sale până la un anumit ordin. Deoarece structura acestor funcții de interpolare depinde de structura nodală a elementului, respectiv de forma lui, ele se mai numesc și funcții de formă. Aceste funcții de formă vor juca rolul funcțiilor coordonate  $\varphi_i$ ,  $i = \overline{1, n}$ , din metodele prezentate în secțiunile anterioare. Deși se pot concepe multe tipuri de funcții de interpolare, se folosesc aproape în exclusivitate funcțiile polinomiale, datorită ușurinței relative cu care acestea pot fi derivate, respectiv integrate.

### 8.6.1. Aproximarea prin polinoame pe porțiuni. Cazul unidimensional

Așadar, ne propunem să aproximăm o funcție de o variabilă reală  $f$ , pe un interval finit  $[a, b]$ . Vom considera o diviziune a acestui interval

$$\Delta: a = x_0 < x_1 < \dots < x_n = b.$$

Se obțin astfel  $n$  subintervale  $[x_i, x_{i+1}]$ ,  $i = \overline{0, n-1}$ . Mai întâi vom aborda problema aproximării prin polinoame liniare pe porțiuni. Funcția de interpolare liniară pe porțiuni depinde de valorile funcției  $f$  în nodurile  $x_i$ . Aceste valori le notăm cu  $f_i = f(x_i)$ ,  $i = \overline{0, n}$ . Pe fiecare subinterval  $[x_i, x_{i+1}]$ , funcția de interpolare este un polinom de forma  $\varphi_i(x) = a_i \cdot x + b_i$ , unde  $a_i$  și  $b_i$  se determină în mod unic din condițiile  $\varphi_i(x_j) = \delta_{ij}$ ,  $0 \leq i, j \leq n$  ( $\delta_{ij}$  - simbolul lui Kronecker).

Astfel din  $\varphi_0(x_0) = 1$ ,  $\varphi_0(x_i) = 0$ ,  $i = \overline{1, n}$ , obținem

$$\varphi_0(x) = \begin{cases} \frac{x_1 - x}{x_1 - x_0}, & x \in [x_0, x_1] \\ 0 & , x \in [x_1, x_n]. \end{cases} \quad (42)$$

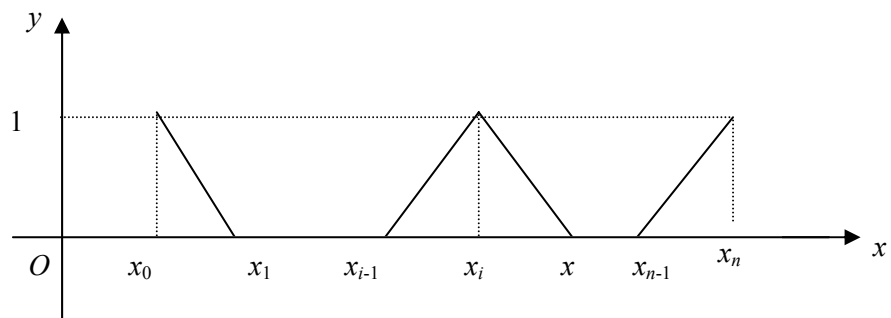
Totodată, pentru  $i$  fixat,  $1 \leq i \leq n-1$ , din  $\varphi_i(x_i) = 1$ ,  $\varphi_i(x_j) = 0$ ,  $j \neq i$ , rezultă

$$\varphi_i(x) = \begin{cases} \frac{x - x_{i-1}}{x_i - x_{i-1}}, & x \in [x_{i-1}, x_i] \\ \frac{x_{i+1} - x}{x_{i+1} - x_i}, & x \in [x_i, x_{i+1}] \\ 0 & , \text{ în rest.} \end{cases} \quad (43)$$

În sfârșit din  $\varphi_n(x_n) = 1$ ,  $\varphi_n(x_i) = 0$ ,  $i = \overline{1, n-1}$ , avem

$$\varphi_n(x) = \begin{cases} 0 & , x \in [x_0, x_{n-1}] \\ \frac{x - x_{n-1}}{x_n - x_{n-1}}, & x \in [x_{n-1}, x_n] \end{cases} \quad (44)$$

Funcțiile  $\varphi_i$ ,  $i = \overline{0, n}$ , reprezintă cel mai simplu tip de funcții de formă și se reprezintă astfel



Cu ajutorul acestor funcții *acoperiș*, funcția de interpolare este dată de formula

$$p_1(x) = \sum_{i=0}^n f_i \varphi_i(x). \quad (45)$$

Se arată ușor că funcțiile  $\varphi_i$ ,  $i = \overline{0, n}$  sunt liniar independente, adică din  $\alpha_0\varphi_0 + \alpha_1\varphi_1 + \dots + \alpha_n\varphi_n = 0$ , rezultă  $\alpha_0 = \alpha_1 = \dots = \alpha_n = 0$  (este suficient să scriem relația dată în nodurile  $x_i$ ,  $i = \overline{0, n}$ ). Vom nota cu  $L(\Delta)$ , spațiul vectorial real de dimensiune  $n+1$ , generat de funcțiile  $\varphi_i$ ,  $i = \overline{0, n}$ , deci al funcțiilor continue de forma

$$g(x) = \sum_{i=0}^n c_i \varphi_i(x), \quad c_i \in \mathbb{R}, \quad i = \overline{0, n}. \quad (46)$$

Să remarcăm, în particular că funcțiile de formă  $\varphi_i(x)$ ,  $i = \overline{1, n-1}$ , sunt nule în afara intervalului  $[x_{i-1}, x_{i+1}]$ , deci au suport compact. Funcția  $p_1$  este locală în sensul că dacă  $x \in [x_i, x_{i+1}]$ ,  $i = \overline{0, n-1}$ , depinde numai de  $f_i$  și  $f_{i+1}$ .

Se poate arăta că dacă funcția  $f$  pe care vrem s-o aproximăm este suficient de netedă (de exemplu, admite derivată de ordinul al doilea), atunci interpolarea prin polinoame liniare pe porțiuni, ne dă o aproximație de ordinul al doilea atât în norma spațiului  $L^2[a, b]$ , dată de (3), cât și în norma Cebîșev. Așadar avem

$$\|f - p_1\|_2 \leq kh^2 \|f''\|_2, \quad (k > 0), \quad (47)$$

respectiv

$$\|f - p_1\|_\infty = \sup_{x \in [a, b]} |f(x) - p_1(x)| \leq k_1 h^2 \|f''\|_\infty, \quad (k_1 > 0). \quad (48)$$

În continuare menționăm că pentru aproximarea soluțiilor problemelor bilocale pentru ecuații diferențiale se pot folosi funcțiile B-spline (§4.4).

Vom avea însă nevoie ca funcțiile generatoare să se anuleze în extremitățile intervalului pe care căutăm soluția ecuației diferențiale considerate.

De remarcat că funcțiile  $B_2, \dots, B_{n-2}$  se anulează în  $x_0, \dots, x_n$ , iar funcțiile  $B_{-1}, B_0, B_1, B_{n-1}, B_n, B_{n+1}$  nu se anulează. De aceea vom proceda după cum urmează. Vom determina constantele  $a, b, c, d$  și  $\alpha, \beta, \gamma, \delta$  astfel ca funcțiile:

$$\begin{aligned} \tilde{B}_0(x) &= aB_{-1}(x) + bB_0(x), & \tilde{B}_1(x) &= cB_0(x) + dB_1(x), \\ \tilde{B}_{n-1}(x) &= \alpha B_{n-1}(x) + \beta B_n(x), & \tilde{B}_n(x) &= \gamma B_n(x) + \delta B_{n+1}(x), \end{aligned}$$

să satisfacă condițiile:

$$\begin{aligned} \tilde{B}_0(x_0) &= 0; & \tilde{B}_0(x_1) &= 1; & \tilde{B}_1(x_0) &= 0; & \tilde{B}_1(x_{-1}) &= 1; \\ \tilde{B}_{n-1}(x_{n+1}) &= 1; & \tilde{B}_{n-1}(x_n) &= 0; & \tilde{B}_n(x_{n-1}) &= 1; & \tilde{B}_n(x_n) &= 0. \end{aligned}$$

De aici rezultă

$$\begin{aligned} \tilde{B}_0(x) &= B_0(x) - 4B_{-1}(x); & \tilde{B}_1(x) &= B_0(x) - 4B_1(x); \\ \tilde{B}_{n-1}(x) &= B_n(x) - 4B_{n-1}(x); & \tilde{B}_n(x) &= B_n(x) - 4B_{n+1}(x). \end{aligned}$$

Deci funcțiile generatoare  $\tilde{B}$ -spline le vom considera

$$\tilde{B}_0, \tilde{B}_1, \tilde{B}_2, \dots, \tilde{B}_{n-1}, \tilde{B}_n, \quad \text{unde } \tilde{B}_i(x) = B_i(x), \quad i = \overline{2, n-2}.$$

### 8.6.2. Aproximarea prin polinoame pe porțiuni. Cazul bidimensional

Vom aborda problema aproximării unei funcții reale definită pe un domeniu mărginit  $G \subset \mathbb{R}^2$ , de frontieră  $C$ . Ne vom ocupa întâi de interpolarea prin *elemente finite triunghiulare*.

Să presupunem, pentru simplitate, că frontiera  $C$  a domeniului  $G$  este o linie frântă. Atunci, este totdeauna posibil să acoperim foarte exact  $G$  cu o mulțime de triunghiuri  $T_k$ ,  $k = \overline{1, p}$ , numite *elemente finite* și care constituie (abstracție făcând de frontiera lor) o partiție a lui  $G$ .

Se realizează astfel o triangularizare a domeniului. Să notăm cu  $A_i$ ,  $i = \overline{1, s}$ , vârfurile triunghiurilor. Unele vor fi în interiorul lui  $G$ , altele pe  $C$ . Aceste vârfuri se mai numesc *noduri de discretizare*.

Triunghiurile se aleg astfel ca:

- un vârf al unui triunghi  $T_k$  nu trebuie să fie niciodată interior unei laturi a altui triunghi, dar poate fi comun mai multor triunghiuri;
- nici un triunghi să nu fie *plat*; este de dorit să se evite unghiurile foarte apropiate de  $0^\circ$  sau  $180^\circ$  (vom vedea de ce).

De remarcat că, o triangularizare de tip *elemente finite* este mult mai suplă decât o rețea de tip *diferențe finite* și permite urmărirea mai fină a frontierei domeniului. Totodată, contrar diferențelor finite, căutăm o aproximare a soluției  $u$ , nu numai în nodurile  $A_i$ , ci peste tot în  $G$ .

Vom căuta soluția problemei de interpolare ca o funcție polinomială pe porțiuni, înțelegând prin porțiuni triunghiurile. Mai precis, în fiecare triunghi  $T_k = A_p A_q A_r$ , vom căuta o aproximație de primul grad în  $x$  și  $y$ , deci o funcție  $u(x, y) = Ax + By + C$ . Constantele  $A, B, C$  se pot determina în funcție de valorile lui  $u$  în cele trei vârfuri  $A_p(x_p, y_p)$ ,  $A_q(x_q, y_q)$ ,  $A_r(x_r, y_r)$ , notate  $u_p, u_q, u_r$ , respectiv.

Așadar  $A, B, C$  reprezintă soluția sistemului liniar

$$\begin{cases} Ax_p + By_p + C = u_p \\ Ax_q + By_q + C = u_q \\ Ax_r + By_r + C = u_r, \end{cases} \quad (49)$$

de determinant

$$\Delta = \begin{vmatrix} x_p & y_p & 1 \\ x_q & y_q & 1 \\ x_r & y_r & 1 \end{vmatrix} = \begin{vmatrix} x_q - x_p & y_q - y_p \\ x_r - x_p & y_r - y_p \end{vmatrix} = \pm 2S, \quad (50)$$

$S$  fiind aria triunghiului  $A_p A_q A_r$ . Deci, dacă nici un triunghi nu este aplatizat, determinantul  $\Delta$  este nenul. Așadar o funcție  $u$  ca mai sus, este definită în mod unic, pe fiecare triunghi, prin valorile sale în cele trei vârfuri.

Să considerăm acum urma funcției  $u(x, y)$  pe una din laturile triunghiului  $A_p A_q A_r$ , să zicem  $A_p A_q$ . Este o funcție de gradul întâi de abscisă (oblică)  $\xi$  de-a lungul lui  $A_p A_q$  (se scrie sub forma  $D\xi + E$ ); această funcție este deci unic determinată de valorile sale în cele două vârfuri.

Presupunând că funcția de interpolare, definită pe întreg  $G$ , ia aceeași valoare în fiecare nod, comun mai multor triunghiuri, rezultă că este continuă de la un triunghi la triunghiul vecin, de-a lungul laturii comune. Deci, în condițiile impuse, funcția de interpolare este continuă (aceasta justifică cerința ca un vârf al unui triunghi să nu fie interior unei laturi a altui triunghi).

Așadar fiind dată o funcție  $u$  continuă pe  $G$ , vom numi *funcție de interpolare pe porțiuni* a lui  $u$ , funcția continuă pe  $G$ , luând aceleași valori ca și  $u$  în toate nodurile (vârfurile) de triangularizare și polinomială de gradul unu în fiecare triunghi.

Vom începe cu construirea unei *baze canonice* într-un triunghi. Pentru comoditate, să notăm triunghiul  $A_p A_q A_r$  cu  $A_1 A_2 A_3$  și să-l studiem deocamdată independent de alte triunghiuri.

Considerăm funcțiile  $\lambda_1(x, y)$ ,  $\lambda_2(x, y)$ ,  $\lambda_3(x, y)$  afine, care satisfac  $\lambda_i(A_j) = \delta_{ij}$  (simbolul lui Kronecker),  $i, j = 1, 2, 3$ . Conform celor de mai sus fiecare funcție există și este unică și

$$\begin{aligned}\lambda_1(x, y) &= \frac{1}{\Delta} [(y_2 - y_3)x + (x_3 - x_2)y + x_2 y_3 - x_3 y_2], \\ \lambda_2(x, y) &= \frac{1}{\Delta} [(y_3 - y_1)x + (x_1 - x_3)y - x_1 y_3 + x_3 y_1], \\ \lambda_3(x, y) &= \frac{1}{\Delta} [(y_1 - y_2)x + (x_2 - x_1)y + x_1 y_2 - x_2 y_1],\end{aligned}\quad (51)$$

$$\Delta = (x_2 - x_1)(y_3 - y_1) - (x_3 - x_1)(y_2 - y_1).$$

Aceste funcții sunt liniar independente, deoarece

$$\alpha_1 \lambda_1(x, y) + \alpha_2 \lambda_2(x, y) + \alpha_3 \lambda_3(x, y) = 0 \Rightarrow \alpha_1 = \alpha_2 = \alpha_3 = 0,$$

după cum se vede ușor, scriind relația în punctele  $A_1, A_2, A_3$ .

În consecință funcțiile  $\lambda_1(x, y)$ ,  $\lambda_2(x, y)$ ,  $\lambda_3(x, y)$  formează o bază, numită *baza canonică*, în spațiul vectorial al polinoamelor de gradul întâi, relativ la triunghiul  $A_1 A_2 A_3$ .

În plus, dacă se caută funcția polinomială de gradul întâi, care ia în  $A_1, A_2, A_3$  valorile impuse  $u_1, u_2, u_3$ , răspunsul este simplu

$$u(x, y) = u_1 \lambda_1(x, y) + u_2 \lambda_2(x, y) + u_3 \lambda_3(x, y), \quad (52)$$

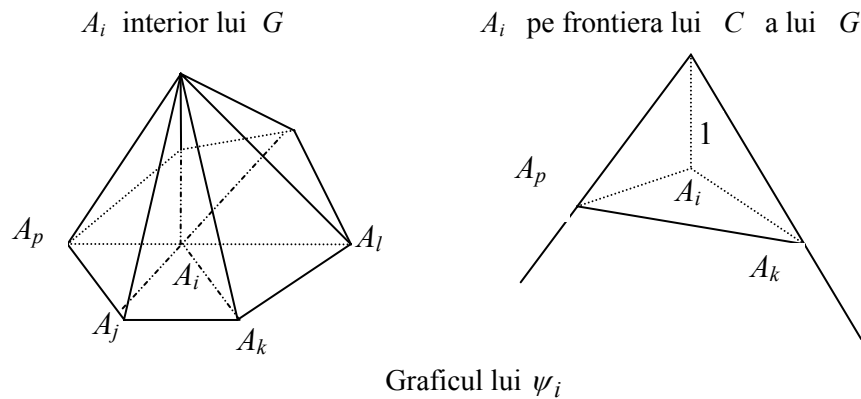
ceea ce justifică adjectivul *canonic*.

Fie acum  $P$  spațiul vectorial real al funcțiilor continue și afine pe porțiuni, pe  $\overline{G}$ . Este clar că dimensiunea lui  $P$  coincide cu numărul vârfurilor,  $\dim P = s$  ( $P$  depinde evident de triangularizarea domeniului aleasă). Baza canonică a lui  $P$  este dată de funcțiile  $\psi_i(x, y)$ ,  $i = \overline{1, s}$ , unde  $\psi_i(A_j) = \delta_{ij}$ ,  $i, j = \overline{1, s}$ .

Atunci pentru orice  $u \in P$  are loc

$$u(x, y) = \sum_{i=1}^s u_i \psi_i(x, y), \quad u_i = u(A_i), \quad i = \overline{1, s}.$$

Pentru aceasta este necesară definirea unei corespondențe biunivoce între numărul global al nodului și numărul triunghiului din care acesta face parte și numărul local al nodului în triunghi. Ce putem spune despre o funcție  $\psi_i(x, y)$ ? Suportul lui  $\psi_i$  este format numai din triunghiurile care îl au pe  $A_i$  ca vârf; el este deci mic, cu atât mai mic cu cât triangularizarea este mai fină, aceasta fiind una dintre caracteristicile metodei elementului finit. Dacă punctul  $A_i$  este interior lui  $G$ ,  $\psi_i$  este nulă pe  $C$ , dar dacă  $A_i \in C$ ,  $\psi_i$  este nenulă în segmentele de frontieră care ajung în  $A_i$ . Dacă se reprezintă pe axa  $Oz$  valorile lui  $\psi_i$  se obține graficul lui  $\psi_i$ . Este o suprafață poliedrală (adică formată din fețe plane). Această suprafață este o piramidă a cărei înălțime este verticala din  $A_i$ , adică 1 și a cărei bază se întinde până la punctele imediat vecine  $A_j, A_k, \dots$ , piramidă prelungită de planul orizontal.



În continuare, ne vom ocupa de interpolarea prin *elemente finite dreptunghiulare*. Domeniile de tip dreptunghiular, adică domeniile cu laturile paralele cu axele de coordonate apar în multe probleme ale fizicii și tehnicii. Prin urmare elementul dreptunghiular are mare importanță.

Ne propunem să aproximăm o funcție  $u$  definită pe domeniul dreptunghiular  $G = [a, b] \times [c, d]$ . Fie  $\Delta_x : a = x_0 < x_1 < \dots < x_n = b$  o diviziune a

lui  $[a, b]$  cu  $n+1$  puncte,  $\Delta_y: c=y_0 < y_1 < \dots < y_m$  o diviziune a lui  $[c, d]$  cu  $m+1$  puncte și  $\Delta = \Delta_x \times \Delta_y$  diviziunea lui  $G$ .

Elementul finit dreptunghiular tipic este  $[x_i, x_{i+1}] \times [y_j, y_{j+1}]$ . Putem considera polinoamele liniare pe porțiuni din cazul unidimensional,  $\varphi_i$ , date de (42)-(44). Atunci funcția de interpolare este dată de

$$p(x, y) = \sum_{i=0}^n \sum_{j=0}^m u_{i,j} \varphi_i(x) \varphi_j(y), \quad (53)$$

unde

$$u_{i,j} = u(x_i, y_j), \quad i = \overline{0, n}, \quad j = \overline{0, m}.$$

Pe elementul dreptunghiular  $[x_i, x_{i+1}] \times [y_j, y_{j+1}]$ , funcția de interpolare are forma

$$p_{i,j}(x, y) = u_{i,j} \varphi_i(x) \varphi_j(y) + u_{i+1,j} \varphi_{i+1}(x) \varphi_j(y) + u_{i,j+1} \varphi_i(x) \varphi_{j+1}(y) + u_{i+1,j+1} \varphi_{i+1}(x) \varphi_{j+1}(y). \quad (54)$$

Baza canonică în spațiul funcțiilor de forma

$$p(x, y) = \sum_{i=0}^n \sum_{j=0}^m c_{i,j} \varphi_i(x) \varphi_j(y), \quad c_{i,j} \in \mathbb{R}, \quad i = \overline{0, n}, \quad j = \overline{0, m},$$

este dată de funcțiile

$$\psi_{ij}(x, y) = \varphi_i(x) \varphi_j(y), \quad i = \overline{0, n}, \quad j = \overline{0, m}. \quad (55)$$

### §8.7. Metoda elementului finit pentru probleme bilocale

Se consideră următoarea ecuație diferențială foarte simplă

$$-u''(x) = f(x), \quad 0 < x < 1, \quad u(0) = u(1) = 0, \quad f \in L^2(0,1) \quad (56)$$

(Exemplul 3).

Problema rezolvării acestei ecuații este echivalentă cu cea a minimizării funcționalei

$$F(u) = \int_0^1 u'^2(x) dx - 2 \int_0^1 f(x) u(x) dx$$

pe mulțimea

$$W = \{ u \in C^2([0, 1]) \cap C^1([0, 1]) ; u(0) = u(1) = 0 \}.$$

Pentru orice  $u, v \in W$ , are loc

$$\langle u, v \rangle_A = \int_0^1 u'(x) v'(x) dx.$$

Vrem să aproximăm soluția problemei bilocale (56) folosind polinoamele liniare pe porțiuni (§8.6.1). Aceste funcții sunt continue, dar nu sunt derivabile, deci nu aparțin lui  $W$ . În consecință,  $W$  nu este „spațiul bun” pentru rezolvarea aproximativă a acestei probleme, folosind polinoamele liniare pe porțiuni. Pentru a stabili spațiul convenabil vom proceda după cum urmează. Introducem o noțiune nouă. Funcția  $g$  se numește *derivata în sensul distribuțiilor* a funcției  $w$  și se notează  $g = w'$  dacă și numai dacă satisface

$$\int_0^1 w\varphi' dx = -\int_0^1 g\varphi dx, \quad (\forall)\varphi \in C^1, \quad \varphi(0) = \varphi(1) = 0. \quad (57)$$

Dacă funcția  $w$  are derivată continuă  $w'$ , atunci aceasta coincide cu derivata în sensul distribuțiilor a lui  $w$ . Bineînțeles că, derivata în sensul distribuțiilor poate exista fără ca derivata în sens clasic să existe.

De exemplu, o funcție  $w$  continuă, care are derivată mărginită cu excepția unui număr finit de puncte, are derivată în sensul distribuțiilor. În punctele în care derivata în sens clasic există, cele două derivate coincid.

Astfel pe intervalul  $[-1, 1]$ , funcția  $w(x) = |x|$  nu este derivabilă în  $x = 0$ , dar admite derivată în sensul distribuțiilor funcția  $g(x) = \text{sign}(x)$ .

Fie

$$H_0^1((0, 1)) = \{u \in L^2((0, 1)); u'(x) \in L^2((0, 1)); u(0) = u(1) = 0\}$$

(este vorba de derivata în sensul distribuțiilor).

Spațiul  $W$  este dens în  $H_0^1((0, 1))$ . De fapt,  $H_0^1((0, 1))$  este chiar spațiul energetic al operatorului  $Ax(t) = -x''(t)$ .

Așadar, pentru construirea aproximantelor Ritz, vom folosi polinoamele liniare pe porțiuni, construite în §8.6.1.

Fie  $\pi : 0 = x_0 < x_1 < x_2 < \dots < x_n < x_{n+1} = 1$ , o diviziune cu noduri echidistante a intervalului  $[0, 1]$ ,  $x_i = ih$ ,  $h = \frac{1}{n+1}$  și  $\varphi_i$ ,  $i = \overline{1, n}$ , funcțiile

$$\varphi_i(x) = \begin{cases} \frac{x - x_{i-1}}{h}, & x_{i-1} \leq x \leq x_i \\ \frac{x_{i+1} - x}{h}, & x_i \leq x \leq x_{i+1}, \\ 0, & \text{în rest} \end{cases} \quad i = \overline{1, n}. \quad (58)$$

Funcțiile  $\varphi_i \in H_0^1((0, 1))$ ,  $i = \overline{1, n}$ . Căutăm soluția sub forma

$\tilde{u} = \sum_{j=1}^n c_j \varphi_j$ , constantele  $c_j$  determinându-se din sistemul (25), unde

$$\langle \varphi_i, \varphi_j \rangle_A = \int_0^1 \varphi_i'(x) \varphi_j'(x) dx, \quad \text{iar} \quad \langle f, \varphi_i \rangle_A = \int_0^1 f(x) \varphi_i(x) dx$$

(derivatele sunt în sensul distribuțiilor).

Este un sistem de  $n$  ecuații, numite *nodale*, de forma  $Bx = d$ , unde elementele matricei  $B$  sunt

$$b_{ij} = \int_0^1 \varphi_i'(x) \varphi_j'(x) dx, \quad i, j = \overline{1, n}, \quad x = (c_1, c_2, \dots, c_n)^T, \quad d = (d_1, d_2, \dots, d_n)^T,$$

$$d_i = \int_0^1 f(x) \varphi_i(x) dx, \quad i = \overline{1, n}.$$

Matricea  $B$  se numește *matrice de rigiditate* și este simetrică și pozitiv definită.

În consecință, sistemul  $Bx = d$  admite soluție unică. Prin calcul obținem

$$b_{ij} = \begin{cases} \frac{2}{h}, & \text{daca } i = j \\ -\frac{1}{h}, & \text{daca } |i-j| = 1 \\ 0, & \text{în rest.} \end{cases}$$

De asemenea

$$d_i = \frac{1}{h} \int_{x_{i-1}}^{x_i} f(x)(x - x_{i-1}) dx + \frac{1}{h} \int_{x_i}^{x_{i+1}} f(x)(x_{i+1} - x) dx.$$

Aproximând aceste integrale cu formula trapezului, obținem  $\tilde{d}_i = hf(x_i)$ ,

deci în locul sistemului  $Bx = d$ , avem de rezolvat sistemul  $Bx = \tilde{d}$ , cu membrul drept obținut prin aplicarea unei formule de integrare numerică.

Matricea de rigiditate  $B$  s-a calculat exact, având polinoame pe porțiuni și integrarea făcându-se ușor. În alte cazuri matricea  $B$  se obține tot prin calculul aproximativ al unor integrale.

Se pune problema alegerii acestor formule, în sensul că trebuie arătat că formulele de cuadratură aplicate ne dau o convergență *bună*, deci o compatibilitate în rezolvarea problemelor puse de metodele variaționale.

Practic, în cele mai multe cazuri, alegând ca funcții de bază polinoamele pe porțiuni, matricea  $B$  se calculează exact, integrarea făcându-se exact. Cum însă funcția  $f$  nu este, de obicei, un polinom, problema este de a calcula aproximativ  $d_i$ .

În cazul de mai sus, avem

$$B = \frac{1}{h^2} \begin{pmatrix} 2 & -1 & 0 & \dots & 0 & 0 \\ -1 & 2 & -1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 2 & -1 \\ 0 & 0 & 0 & \dots & -1 & 2 \end{pmatrix}, \quad d = \begin{pmatrix} f(x_1) \\ f(x_2) \\ \dots \\ f(x_{n-1}) \\ f(x_n) \end{pmatrix}$$

Aplicând ecuației  $-u''(x) = f(x)$ , metoda standard cu diferențe finite, se obține exact același sistem de ecuații.

În cazul metodei elementului finit, soluția aproximativă găsită, aproximează soluția exactă în orice punct al intervalului  $[0, 1]$ .

Să considerăm acum problema mai generală

$$-x''(t) + \sigma(t)x(t) = f(t), \quad 0 \leq t \leq 1,$$

$$x(0) = x(1) = 0,$$

(59)

$\sigma(t) > 0$  și continuă pe  $[0, 1]$ .

Problema este abordată variațional în Exemplul 4. Considerăm deci operatorul

$$Ax(t) = -x''(t) + \sigma(t)x(t).$$

Dacă funcția  $f$  este continuă pe  $[0, 1]$ , atunci există soluție unică a problemei (59) și  $x \in C^2([0, 1])$ . Pentru determinarea aproximației element finit  $\tilde{x}(t)$  a soluției  $x(t)$  vom folosi, pentru început, polinoamele *spline* cubice.

Fie  $\pi : 0 = t_0 < t_1 < \dots < t_n = 1$ , o diviziune cu noduri echidistante și funcțiile  $\tilde{B}_0, \tilde{B}_1, \dots, \tilde{B}_n$ , din §8.6.2. În acest caz  $\varphi_i(t) = \tilde{B}_i(t)$ ,  $i = \overline{0, n}$ . Aceste funcții aparțin domeniului de definiție al operatorului  $A$ .

Aproximația element finit  $\tilde{x}(t)$  a soluției exacte  $x(t)$ , va fi

$$\tilde{x}(t) = \sum_{i=0}^n c_i \tilde{B}_i(t), \quad \text{unde } c = (c_0, c_1, \dots, c_n)^T \text{ este soluția sistemului algebric liniar}$$

$$\sum_{j=0}^n \langle A\tilde{B}_i, \tilde{B}_j \rangle c_j = \langle f, \tilde{B}_i \rangle, \quad i = \overline{0, n}.$$

În acest caz, elementele matricei de rigiditate sunt

$$b_{i,j} = \langle A\tilde{B}_i, \tilde{B}_j \rangle = \int_0^1 [\tilde{B}_i'(t)\tilde{B}_j'(t) + \sigma(t)\tilde{B}_i(t)\tilde{B}_j(t)] dt, \quad i, j = \overline{0, n},$$

iar

$$d_i = \int_0^1 f(t)\tilde{B}_i(t) dt, \quad i = \overline{0, n}.$$

Matricea de rigiditate sau matricea *energie* va fi o matrice bandă de tip 7. Matricea energie și termenul liber se pot calcula folosind, de exemplu, metoda lui Gauss cu două noduri.

În ceea ce privește eroarea, se poate demonstra

**Teorema 4.** În ipotezele de mai sus și dacă  $f \in C^2[0,1]$ , atunci există o constantă  $K$ , independentă de  $n$ , astfel încât

$$\|x - \tilde{x}\|_{\infty} = \sup_{x \in [0,1]} |x(t) - \tilde{x}(t)| \leq Kh^3$$

(unde  $h$  este pasul rețelei).

Să analizăm acum situația în care funcțiile de formă nu aparțin domeniului operatorului  $A$ , cum este cazul când acestea sunt polinoamele liniare pe porțiuni date de (58). Vom proceda ca în exemplul de la începutul secțiunii. Deci

$\tilde{x}(t) = \sum_{i=1}^n c_i \varphi_i(t)$ , cu  $\varphi_i$  date de (58). Acum

$$b_{ij} = \int_0^1 (\varphi_i' \varphi_j' + \sigma \varphi_i \varphi_j) dt, \quad i, j = \overline{1, n}$$

(derivatale sunt luate în sensul distribuțiilor).

Matricea de rigiditate va fi o matrice bandă de tip 3.

Se poate demonstra

**Teorema 5.** În ipotezele de mai sus și dacă  $f \in C[0,1]$ , există o constantă  $K$  independentă de  $n$ , astfel ca

$$\|x - \tilde{x}\|_{\infty} \leq Kh.$$

Se observă că dacă în aproximare folosim funcții de bază mai netede, obținem o aproximare mai bună. Se ajunge însă la un sistem liniar algebric cu mai multe elemente nenule. În cazul polinoamelor cubice spline se obține o matrice bandă de tip 7, în timp ce în cazul polinoamelor pe porțiuni de grad întâi, se obține o matrice bandă de tip 3.

### §8.8. Metoda elementului finit pentru probleme la limită pentru ecuația lui Laplace în plan

Fie  $G = \left(-\frac{\pi}{2}, \frac{\pi}{2}\right) \times \left(-\frac{\pi}{2}, \frac{\pi}{2}\right) \subset \mathbf{R}^2$  și  $C$  frontiera sa. Se cere să se găsească soluția ecuației

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 2, \quad (60)$$

cu condițiile la limită

$$\begin{aligned} u\left(x, \pm \frac{\pi}{2}\right) &= 0, \quad |x| \leq \frac{\pi}{2} \\ u\left(\pm \frac{\pi}{2}, y\right) &= 0, \quad |y| \leq \frac{\pi}{2}. \end{aligned} \quad (61)$$

Problema este abordată variațional în Exemplul 5, rezolvarea sa fiind echivalentă cu cea a minimizării funcționalei

$$F(u) = \iint_G \left[ \left( \frac{\partial u}{\partial x} \right)^2 + \left( \frac{\partial u}{\partial y} \right)^2 \right] dx dy + 2 \iint_G u(x, y) dx dy$$

pe mulțimea

$$W = \{ u \in C^2(G) \cap C^1(\bar{G}) ; u = 0 \text{ pe } C \} .$$

Pentru orice  $u, v \in W$ , are loc

$$(u, v)_{-\Delta} = \iint_G \left( \frac{\partial u}{\partial x} \cdot \frac{\partial v}{\partial x} + \frac{\partial u}{\partial y} \cdot \frac{\partial v}{\partial y} \right) dx dy .$$

Pentru găsirea soluției aproximative a problemei (60)-(61), vom folosi elementele finite dreptunghiulare construite în secțiunea 6.2, care nu aparțin lui  $W$ . În consecință  $W$  nu este „spațiul bun” pentru rezolvarea acestei probleme cu elemente finite dreptunghiulare.

Pentru a depăși această dificultate, să constatăm, mai întâi, că noțiunea de derivată în sensul distribuțiilor, introdusă în secțiunea anterioară pentru funcții de o variabilă, se extinde, în mod corespunzător la funcții de mai multe variabile.

De exemplu, funcția  $g(x, y)$  este  $\frac{\partial u}{\partial x}$ , în sensul distribuțiilor, dacă satisface

$$\iint_G g \cdot \varphi dx dy = - \iint_G u \frac{\partial \varphi}{\partial x} dx dy, \quad (\forall) \varphi \in C^1, \quad \varphi = 0 \text{ pe } C .$$

Fie

$$H_0^1(G) = \{ u \in L^2(G) ; u, \frac{\partial u}{\partial x}, \frac{\partial u}{\partial y} \in L^2(G), u = 0 \text{ pe } C \}$$

(derivatele parțiale sunt luate în sensul distribuțiilor).

Spațiul  $W$  este dens în  $H_0^1(G)$ ,  $H_0^1(G)$  fiind spațiul energetic al operatorului  $-\Delta$ . Spațiul  $H_0^1(G)$  este spațiu Hilbert în raport cu produsul scalar

$$\langle u, v \rangle = \iint_G \left( uv + \frac{\partial u}{\partial x} \cdot \frac{\partial v}{\partial x} + \frac{\partial u}{\partial y} \cdot \frac{\partial v}{\partial y} \right) dx dy .$$

Pentru găsirea soluției aproximative, partiționăm domeniul  $G$  în  $(n+1)^2$  pătrate, folosind  $2n$  paralele (echidistante) cu axele de coordonate.

Funcțiile de bază utilizate sunt  $\psi_{ij}$ ,  $i, j = \overline{1, n}$ , date de (55) și satisfac condițiile la limită (61). Căutăm soluția aproximativă de forma

$$\tilde{u}(x, y) = \sum_{i=1}^n \sum_{j=1}^n c_{ij} \psi_{ij}(x, y), \quad (62)$$

constantele  $c_{ij}$ ,  $i, j = \overline{1, n}$ , urmând a fi determinate din următorul sistem obținut din (25)

$$\sum_{k=l=1}^n \sum_{k=l=1}^n c_{kl} \cdot \iint_G \left[ \frac{\partial \psi_{kl}}{\partial x} \frac{\partial \psi_{ij}}{\partial x} + \frac{\partial \psi_{kl}}{\partial y} \frac{\partial \psi_{ij}}{\partial y} \right] dx dy + 2 \iint_G \psi_{ij} dx dy = 0, \quad (63)$$

$$i, j = \overline{1, n}$$

sau încă

$$\sum_{k=1}^n \sum_{l=1}^n a_{ijkl} c_{kl} + 2 \iint_G \psi_{ij} dx dy = 0, \quad i, j = \overline{1, n}$$

unde

$$a_{ijkl} = \iint_G \left[ \frac{\partial \varphi_k}{\partial x}(x) \frac{\partial \varphi_l}{\partial x}(x) \varphi_l(y) \varphi_j(y) + \frac{\partial \varphi_l}{\partial y}(y) \frac{\partial \varphi_j}{\partial y}(y) \varphi_k(x) \varphi_i(x) \right] dx dy,$$

funcțiile  $\varphi_i$ ,  $i = \overline{1, n}$ , fiind date de (43).

Mai întâi să calculăm

$$\iint_G \psi_{ij} dx dy = \iint_G \varphi_i(x) \varphi_j(y) dx dy = \int_{x_{i-1}}^{x_{i+1}} \varphi_i(x) dx \cdot \int_{y_{j-1}}^{y_{j+1}} \varphi_j(y) dy = h^2.$$

Ținând seama de suportul funcțiilor  $\psi_{ij}(x, y)$ , sistemul (63) se mai scrie sub forma

$$\sum_{k=i-1}^{i+1} \sum_{l=j-1}^{j+1} a_{ijkl} c_{kl} + 2h^2 = 0, \quad i, j = \overline{1, n}.$$

Calcul elementare arată că:

$$\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \varphi_{i-1}(t) \varphi_i(t) dt = \frac{h}{6}, \quad \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \varphi_i^2(t) dt = \frac{2h}{3}, \quad \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \varphi'_{i-1}(t) \varphi_i(t) dt = -\frac{1}{h}, \quad \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \varphi_i^2(t) dt = \frac{2}{h}.$$

În consecință  $a_{ijkl} = -\frac{1}{3}$ ,  $k = \overline{i-1, i+1}$ ,  $l = \overline{j-1, j+1}$ , cu excepția

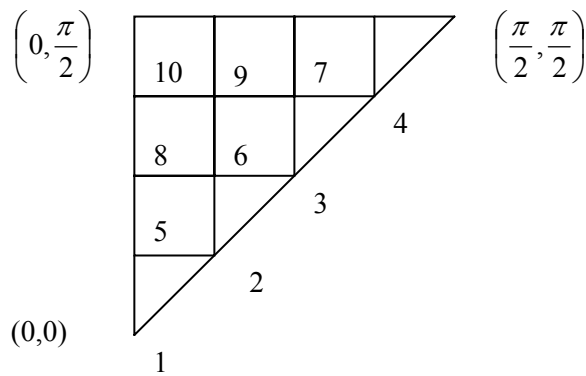
elementului  $a_{ijij} = \frac{8}{3}$ , deci sistemul (64) devine

$$3c_{ij} - \frac{1}{3} \sum_{k=i-1}^{i+1} \sum_{l=j-1}^{j+1} c_{kl} + 2h^2 = 0, \quad i, j = \overline{1, n}. \quad (65)$$

Soluția exactă a problemei considerate este

$$u(x, y) = -\left(\frac{\pi}{2}\right)^2 + x^2 + \frac{8}{\pi} \cdot \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{(2k-1)^3} \frac{ch(2k-1)y}{ch((2k-1)\pi/2)} \cos(2k-1)x.$$

Valorile soluției aproximative în nodurile din figura :



sunt date în următorul tabel :

Nodul	$n = 3$	$n = 7$	$N = 15$	Soluția exactă
1	- 1,534	- 1,473	- 1,459	- 1,454
2		- 1,321	- 1,308	- 1,304
3	- 0,950	- 0,907	- 0,897	- 0,894
4		- 0,370	- 0,362	- 0,359
5		- 1,394	- 1,381	- 1,376
6		-1,089	- 1,078	- 1,075
7		- 0,566	- 0,559	- 0,556
8	- 1,278	- 1,146	- 1,135	- 1,132
9		- 0,666	- 0,660	- 0,658
10		- 0,698	- 0,692	- 0,690

În continuare, reluăm problema Dirichlet-Neumann din Exemplul 7, pe care o vom concretiza. Considerăm pătratul  $G = (0,1) \times (0,1) \subset \mathbb{R}^2$ , din figură, cu frontiera  $C_1 \cup C_2$ , unde  $C_1 = OC \cup CB \cup BA$  și  $C_2 = OA$ .

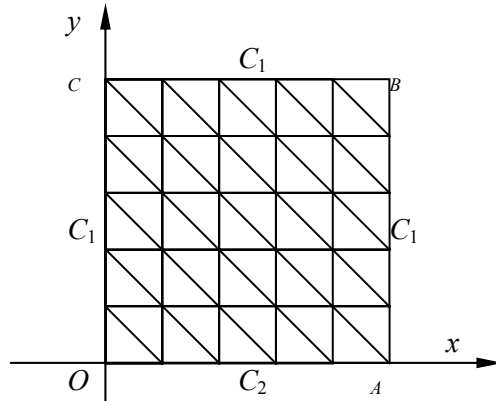
Problema (36)-(38) se scrie

$$-\Delta u = f \quad \text{în } G, \quad (66)$$

$$u = 0 \quad \text{pe } OC, CB \text{ și } BA, \quad (67)$$

$$\frac{\partial u}{\partial x} = -\frac{\partial u}{\partial y} = g \text{ pe } OA \tag{68}$$

Vom considera o rețea de tip *diferențe finite*, cu pasul  $h = k = \frac{1}{n}$  pe  $OC$ , respectiv  $OA$ . Decupăm apoi fiecare pătrat astfel obținut în două triunghiuri și obținem o triangularizare de tip *elemente finite*. Fie  $\tilde{u}(x, y)$  soluția aproximativă definită pe  $\overline{G}$  și nu numai în puncte izolate, cum se întâmplă în cazul diferențelor finite. Vom utiliza pentru triangularizarea de mai sus, spațiul  $P$  al funcțiilor continue pe  $\overline{G}$  și affine pe porțiuni construite în secțiunea 6.2. Aceste funcții nu aparțin însă spațiului  $V = C^1(G) \cap C^0(\overline{G})$  din exemplul 7. În consecință,  $V$  nu este „spațiul bun”, alegerea *bună* fiind spațiul



$$H^1(G) = \{u \in L^2(G) ; u, \frac{\partial u}{\partial x}, \frac{\partial u}{\partial y} \in L^2(G)\}$$

(derivatele sunt considerate în sensul distribuțiilor), pentru care  $P$  este subspațiu.

Căutăm soluția  $\tilde{u}$  de forma

$$\tilde{u}(x, y) = \sum_{i=1}^s c_i \psi_i(x, y), \tag{69}$$

$s$  fiind numărul vârfurilor, funcțiile  $\psi_i, i = \overline{1, s}$  sunt funcțiile de bază corespunzătoare triangularizării, iar coeficienții  $c_i, i = \overline{1, s}$ , urmează a fi determinați. Pentru simplitatea expunerii, vom presupune că vârfurile de pe  $C_1$  ocupă ultimele poziții,  $n+1, n+2, \dots, s$ . Așadar vom avea  $c_i = 0, i = \overline{n+1, s}$ , necunoscutele propriu-zise fiind  $c_i, i = \overline{1, n}$ . În definitiv

$$\tilde{u}(x, y) = \sum_{i=1}^n c_i \psi_i(x, y), \tag{70}$$

necunoscutele  $c_i$  trebuind determinate din condiția ca  $\tilde{u}$  să satisfacă sistemul (41), adică

$$\sum_{j=1}^n a(\psi_j, \psi_i) c_j = \iint_G f \psi_i dx dy + \int_{C_2} g \psi_i dx, \quad i = \overline{1, n}, \tag{71}$$

unde

$$a(\psi_j, \psi_i) = \iint_G \text{grad} \psi_i \cdot \text{grad} \psi_j dx dy.$$

Notând

$$a_{ij} = a(\psi_i, \psi_j), \quad i, j = \overline{1, n}, \quad b_i = \iint_G f \psi_i dx dy + \int_{C_2} g \psi_i dx, \quad i = \overline{1, n}, \quad (72)$$

sistemul (71) se scrie

$$\sum_{j=1}^n a_{ij} c_j = b_i, \quad i = \overline{1, n}. \quad (73)$$

Așadar se pune problema rezolvării unui sistem de  $n$  ecuații liniare, numite *nodale*. Dacă  $x$  este vectorul coloană al necunoscutelor  $c_j$  și  $b$  vectorul coloană termen liber din (73), atunci sistemul (73) se scrie sub forma

$$Ax = b, \quad (74)$$

unde matricea  $A = (a_{ij})$ ,  $i, j = \overline{1, n}$ , se numește „matrice de rigiditate” și este evident simetrică. Mai mult, matricea  $A$  este și pozitiv definită.

Într-adevăr

$$x^T Ax = \sum_{i=1}^n \sum_{j=1}^n a(\psi_j, \psi_i) c_i c_j = a \left( \sum_{j=1}^n c_j \psi_j, \sum_{i=1}^n c_i \psi_i \right) = \iint_G (\text{grad } \tilde{u})^2 dx dy \geq 0.$$

Totodată  $x^T Ax = 0$ , implică  $\text{grad } \tilde{u} = 0$  în mulțimea conexă  $G$ , deci  $\tilde{u}$  este constantă pe  $G$ . Dar  $\tilde{u}$  se anulează pe  $C_1$ , deci  $\tilde{u} = 0$  în  $G$  și în consecință  $x = 0$ , adică  $A$  este pozitiv definită. Prin urmare matricea  $A$  este nesingulară, deci sistemul (73) admite soluție unică.

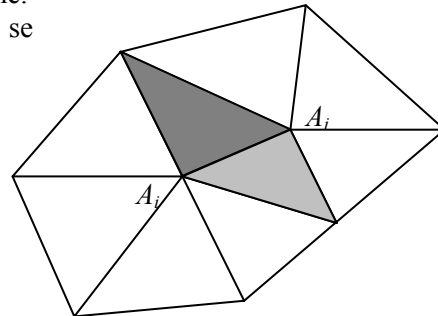
Matricea  $A$  nu este numai simetrică ci și *rară* (are multe zerouri). Într-adevăr  $\text{supp } \psi_i$  este constituit din mulțimea triunghiurilor care îl au pe  $A_i$  ca vârf. Deci elementul  $a_{ij}$  are șanse să fie nenul, dacă și numai dacă  $A_i$  și  $A_j$  sunt vârfuri ale cel puțin unui același triunghi.

Elementul  $a_{ij}$  este integrala pe  $\text{supp } \psi_i \cap \text{supp } \psi_j$ , adică pe o reuniune de triunghiuri. Integrala fiind „funcție aditivă de mulțime”, se va calcula pe fiecare triunghi și adunând rezultatele obținute.

Spre exemplu, în figura alăturată  $a_{ij}$  este integrala din  $\text{grad } \psi_i \cdot \text{grad } \psi_j$  pe cele două triunghiuri marcate, integrandul (constant) având o expresie diferită pe fiecare din cele două triunghiuri în chestiune.

În practică se procedează astfel: se inițializează coeficienții  $a_{ij}$  cu zero. Se

trece în revistă fiecare triunghi, adunând valorile care reprezintă contribuția acestui triunghi la coeficienții  $a_{ij}$  corespunzători (până în acest moment aceștia reprezintă contribuțiile aduse de triunghiurile precedente). Vectorul coloană  $b$  se poate calcula în același timp cu matricea  $A$ .



De remarcat că  $a_{ii} > 0$ ,  $(\forall) i = \overline{1, n}$ . Matricea  $A$  nu este totdeauna diagonal dominantă.  $A$  fiind simetrică și pozitiv definită pentru rezolvarea sistemului (74) se poate aplica o factorizare de tip Cholesky. Bineînțeles că se pot aplica și metode iterative.

Pentru calculul termenului liber în (73), vom utiliza o formulă de cuadratură numerică și anume:

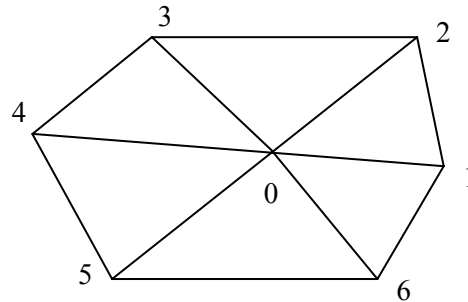
$$\iint_T h(x, y) dx dy \cong S(T) \frac{h(A_p) + h(A_q) + h(A_r)}{3}, \quad (75)$$

dacă  $T$  este triunghiul  $A_p, A_q, A_r$ , respectiv

$$\int_a^b v(x) dx \cong (b - a) \frac{v(a) + v(b)}{2}. \quad (76)$$

Formula (76) nu este altceva decât formula trapezului; are o precizie de ordinul 2: dacă  $f$  este regulată (local de clasă  $C^2$ ) și  $b - a = h \rightarrow 0$ , atunci integrala pe  $(a, b)$ , care este de ordin  $h$ , este aproximată de ordinul  $O(h^3)$ . În ceea ce privește (75), este formula echivalentă formulei (76) în dimensiune 2; are o precizie de ordinul 2.

Având în vedere simplitatea și repetitivitatea rețelei, este suficient să scriem (71) într-un nod interior lui  $G$  și într-un nod pe  $OA$  (extremitățile se exclud dacă  $\tilde{u} = 0$  în  $O$  și  $A$ ). Să analizăm mai întâi cazul unui nod interior lui  $G$ , ca în figura alăturată. Vom utiliza o numerotare locală, nodul fiind notat cu indicele „0”, iar celelalte cu 1, 2, 3, 4, 5, 6. Funcția  $\psi_0$ , reprezentată pe axa  $Oz$  este o piramidă de înălțime 1 pe verticala din nodul 0 și a cărei bază se întinde până la nodurile imediat vecine, 1, 2, 3, 4, 5, 6, prelungită în planul orizontal.



Așadar  $\text{supp } \psi_0$  este hexagonul 123456. Din (73) se obține

$$\sum_{i=0}^6 a_{0i} c_i = b_0, \quad (77)$$

cu  $a_{0i}$  și  $b_0$  date de (72). Prin translație și simetrie se constată că  $a_{01} = a_{04}$ ,  $a_{02} = a_{05}$ ,  $a_{03} = a_{06}$ .

Pe fiecare triunghi din  $\text{supp } \psi_0$ ,  $\text{grad } \psi_0$  este un vector constant, deoarece în fiecare triunghi  $\psi_0$  are forma

$$\psi_0(x, y) = Ax + By + C, \text{ deci } \text{grad } \psi_0 = A\vec{i} + B\vec{j}.$$

De exemplu, dacă nodul 0 are coordonatele  $(ih, jh)$ ,  $h$  fiind pasul rețelei, atunci pe triunghiul 012, impunând condițiile

$$\psi_0(ih, jh) = 1, \psi_0((i+1)h, jh) = 0, \psi_0(ih, (j+1)h) = 0,$$

se obține

$$A = -\frac{1}{h}, B = -\frac{1}{h}, \text{ deci } \text{grad}\psi_0 = \left(-\frac{1}{h}, -\frac{1}{h}\right).$$

De remarcat că  $\text{grad}\psi_0$  nu depinde de  $i$  și  $j$ . În mod asemănător se poate calcula  $\text{grad}\psi_0$  și pe celelalte triunghiuri care compun  $\text{supp}\psi_0$  și de asemenea  $\text{grad}\psi_0, i = \overline{1,6}$ .

În consecință, integrala pe fiecare triunghi este egală cu produsul dintre aria triunghiului  $\frac{h^2}{2}$  cu un produs scalar obișnuit.

Să calculăm acum  $a_{0i}, i = \overline{1,6}$ .  $\text{Supp}\psi_0$  este hexagonul 123456. Avem:

$$\text{pe } 012, \text{ grad}\psi_0 = \left(-\frac{1}{h}, -\frac{1}{h}\right); \text{ pe } 045, \text{ grad}\psi_0 = \left(\frac{1}{h}, \frac{1}{h}\right);$$

$$\text{pe } 023, \text{ grad}\psi_0 = \left(0, -\frac{1}{h}\right); \text{ pe } 056, \text{ grad}\psi_0 = \left(0, \frac{1}{h}\right);$$

$$\text{pe } 034, \text{ grad}\psi_0 = \left(\frac{1}{h}, 0\right); \text{ pe } 061, \text{ grad}\psi_0 = \left(-\frac{1}{h}, 0\right);$$

$$\text{supp}\psi_0 \cap \text{supp}\psi_1 = 012 \cup 061:$$

$$\text{pe } 012, \text{ grad}\psi_1 = \left(\frac{1}{h}, 0\right); \text{ pe } 061, \text{ grad}\psi_1 = \left(\frac{1}{h}, \frac{1}{h}\right).$$

$$\text{supp}\psi_0 \cap \text{supp}\psi_2 = 012 \cup 023;$$

$$\text{pe } 012, \text{ grad}\psi_2 = \left(0, \frac{1}{h}\right); \text{ pe } 023, \text{ grad}\psi_2 = \left(\frac{1}{h}, \frac{1}{h}\right).$$

$$\text{supp}\psi_0 \cap \text{supp}\psi_3 = 023 \cup 034:$$

$$\text{pe } 023, \text{ grad}\psi_3 = \left(-\frac{1}{h}, 0\right); \text{ pe } 034, \text{ grad}\psi_3 = \left(0, \frac{1}{h}\right);$$

În consecință:

$$a_{00} = \iint_S (\text{grad}\psi_0)^2 dx dy, S = 012 \cup 023 \cup 034 \cup 045 \cup 056 \cup 061,$$

deci

$$a_{00} = \frac{h^2}{2} \left( \frac{2}{h^2} + \frac{1}{h^2} + \frac{1}{h^2} + \frac{2}{h^2} + \frac{1}{h^2} + \frac{1}{h^2} \right) = 4$$

$$a_{01} = a_{04} = \iint_S \text{grad}\psi_0 \cdot \text{grad}\psi_1 dx dy, S = 012 \cup 061, \text{ deci}$$

$$a_{01} = a_{04} = -1$$

$$a_{02} = a_{05} = \iint_S \text{grad } \psi_0 \cdot \text{grad } \psi_2 \, dx dy, \quad S = 012 \cup 023, \text{ deci}$$

$$a_{02} = a_{05} = -1$$

$$a_{03} = a_{06} = \iint_S \text{grad } \psi_0 \cdot \text{grad } \psi_3 \, dx dy, \quad S = 023 \cup 034, \text{ deci}$$

$$a_{03} = a_{06} = 0$$

Deoarece  $\varphi_0=0$  pe  $C_2$ , conform (72),  $b_0 = \iint_S f \psi_0 \, dx dy$ ,

$$S = 012 \cup 023 \cup 034 \cup 045 \cup 056 \cup 061, \text{ deci}$$

$$b_0 \cong \frac{h^2}{2} \left( \frac{1}{3} f_0 \cdot 1 + \frac{1}{3} f_0 \cdot 1 + \frac{1}{3} f_0 \cdot 1 + \frac{1}{3} f_0 \cdot 1 + \frac{1}{3} f_0 \cdot 1 + \frac{1}{3} f_0 \cdot 1 \right) = h^2 f_0,$$

conform formulei (75) și ținând seama că  $\psi_0$  se anulează în nodurile 1, 2, 3, 4, 5, 6 și ia valoarea 1 în nodul 0. S-a notat cu  $f_0$  valoarea funcției  $f$  în nodul 0.

Așadar ecuația nodală (77) se scrie

$$4c_0 - c_1 - c_4 - c_2 - c_5 = h^2 f_0,$$

sau

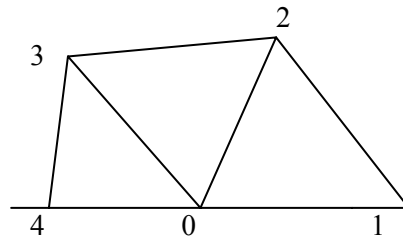
$$-\frac{c_1 - 2c_0 + c_4}{h^2} - \frac{c_2 - 2c_0 + c_5}{h^2} = f_0. \tag{78}$$

Se regăsește discretizarea ecuației  $-\Delta u = f$ , cu diferențe finite, după binecunoscuta schemă „în cruce”.

Să analizăm acum cazul unui nod pe  $OA$ , numerotarea locală fiind cea din figură.

Aceste noduri sunt similare cazului nodului interior fiind evident de două ori mai puține.

În consecință  $a_{00} = 2$ ,  $a_{02} = -1$ ,  $a_{03} = 0$  (ultimele două valori fiind ca acolo), iar  $a_{01} = a_{04} = -\frac{1}{2}$  (deoarece se



ia în calcul un singur triunghi). Din aceleași motive  $\iint_G f \psi_0 \, dx dy \approx \frac{h^2}{2} f_0$

(3 triunghiuri în loc de 6) și  $\int_{C_2} g \psi_0 \, dx = \int_0^{\frac{1}{2}} + \int_{\frac{1}{2}}^1 \approx h \cdot \left( \frac{1}{2} g_0 \cdot 1 + \frac{1}{2} g_0 \cdot 1 \right)$ , unde  $g_0$

,  $f_0$  sunt valorile lui  $g$  respectiv  $f$  în nodul 0. S-au utilizat iarăși (75) și (76) și faptul că  $\psi_0$  este 1 în nodul 0 și se anulează în celelalte.

Așadar, în acest caz  $b_0 = \frac{h^2}{2} f_0 + h g_0$ . Ecuația nodală corespunzătoare

este

$$c_2 - 2c_0 + \frac{1}{2}(c_4 + c_1) = -hg_0 - \frac{h^2}{2} f_0. \quad (79)$$

În ceea ce privește convergența metodei, dacă  $h$  este mărimea elementelor finite (de exemplu diametrul), atunci *lema lui Bramble și Hilbert* precizează că

$$\|u - \tilde{u}\| = O(h),$$

unde

$$\|u\| = \left( \|u\|_{L^2(G)}^2 + \left\| \frac{\partial u}{\partial x} \right\|_{L^2(G)}^2 + \left\| \frac{\partial u}{\partial y} \right\|_{L^2(G)}^2 \right)^{1/2},$$

deci aproximarea este de ordinul 1. Pentru obținerea unei aproximări de ordin mai mare, se impune utilizarea unor polinoame de grad mai mare (nu de gradul unu), deci introducerea unor elemente finite noi care comportă mai multe noduri.

### Exerciții

Să se determine funcționalele asociate problemelor la limită:

1. Se caută  $u \in C^2[a, b]$  care satisface:  $-\frac{d}{dx} \left[ p(x) \frac{du}{dx} \right] + q(x)u(x) = f(x)$ ,  
 $u(a) = u(b) = 0$ , unde  $p \in C^1([a, b])$ ,  $q \in C([a, b])$ ,  $f \in L^2(a, b)$ ,  $p(x) \geq p_0 > 0$ ,  
 $q(x) \geq 0$ .

$$R. F(u) = \int_a^b [p(x)u'(x)^2 + q(x)u^2(x) - 2f(x)u(x)] dx.$$

2. Fie  $G \subset \mathbb{R}^2$  un domeniu mărginit de curba  $C$ . Se caută  $u \in C^2(G)$  care satisface:

$$\Delta u = -\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = f, \quad f \in L^2(G), \text{ în } G,$$

$$\frac{\partial u}{\partial n} + \sigma(P)u|_C = 0, \quad \sigma \in C^0(C), \text{ unde } \sigma(P) \geq \sigma_0 > 0.$$

$$R. F(u) = \iint_G (\text{grad } u)^2 dx dy + \int_C \sigma(P)u^2 ds - \iint_G f(x)u(x) dx dy.$$

3. Fie  $G \subset \mathbb{R}^2$  un domeniu mărginit de curba  $C$ . Se caută  $u \in C^4(G)$  care satisface:

$$\begin{cases} \Delta(\Delta u) = f, & f \in L^2(G) \text{ în } G \\ u|_C = \frac{\partial u}{\partial n}|_C = 0 \end{cases} .$$

$$R. F(u) = \iint_G \left[ \left( \frac{\partial^2 u}{\partial x^2} \right)^2 + \left( \frac{\partial^2 u}{\partial y^2} \right)^2 + 2 \frac{\partial^2 u}{\partial x^2} \frac{\partial^2 u}{\partial y^2} \right] - 2 \iint_G f(x)u(x) dx dy .$$

4. Fie triunghiul  $A_1A_2A_3$ , unde  $A_i(x_i, y_i)$ ,  $i=1, 2, 3$ . Dacă  $P_j = A_j$ ,  $j=1, 2, 3$ , iar  $P_4, P_5, P_6$  sunt mijloacele laturilor  $A_1A_2, A_2A_3, A_3A_1$  respectiv, să se arate că funcția polinomială de gradul al doilea care ia în punctele  $P_j$  valorile  $u_j$ ,  $j=1, \overline{6}$  este

$$u(x, y) = \sum_{j=1}^6 u_j \mu_j(x, y), \text{ unde } \mu_j = \lambda_j(2\lambda_j - 1), \quad j=1, 2, 3, \quad \mu_4 = 4\lambda_1\lambda_2, \\ \mu_5 = 4\lambda_2\lambda_3, \quad \mu_6 = 4\lambda_3\lambda_1, \text{ funcțiile } \lambda_i \text{ fiind date de (51).}$$

5. Folosind elemente finite triunghiulare, să se găsească soluția aproximativă a

problemei la limită:  $\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 4$ ,  $(x, y) \in G$ ,  $u(x, y) = x^2 + y^2$  pe frontiera

lui  $G$ ,  $G = \{(x, y) \in \mathbb{R}^2 \mid |x| \leq 1, |y| \leq 1\}$ .