NUMERIČKA MATEMATIKA

1. RJEŠAVANJE SISTEMA LINEARNIH JEDNAČINA

Numeričke metode za rješavanje sistema ljnearnih jednačina dijele se u dvije klase: (a) direktne metode i (b) iterativne metode.

1.1. GAUSSOVA METODA ELIMINACIJE

Biće izložen uobičajeni Gaussov postupak uzastopne eliminacije nepoznatih za rješavanje sistema linearnih jednačina oblika $A\mathbf{x} = \mathbf{b}$.

Gaussovom metodom eliminacije za rješavanje sistema linearnih jednačina

$$A\mathbf{x} = \mathbf{b}, \tag{1}$$

gdje je A regularna kvadratna matrica dimenzije $n \times n$ i $\mathbf{b}$ n–dimenzioni vektor,

$$A = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix},$$

transformiše se u konačno mnogo koraka sistem (1) u sistem sa trougaonom matricom

$$U\mathbf{x} = \mathbf{c}, \qquad U = \begin{bmatrix} u_{11} & \dots & u_{1n} \\ & \ddots & \vdots \\ 0 & & u_{nn} \end{bmatrix}, \tag{2}$$

čije rješenje je identično rješenju polaznog sistema (1). Sistem (2) se, pod pretpostavkom da je $u_{ii} \neq 0$, $i = 1, \dots, n$, direktno može riješiti,

$$x_n = \frac{c_n}{u_{nn}}, \qquad x_i = \frac{1}{u_{ii}}\Big(c_i - \sum\nolimits_{j=i+1}^{n} u_{ij}x_j\Big), \qquad i = n-1, \dots, 1.$$

Prvi korak algoritma Gaussove eliminacije sastoji se u oduzimanju prve jednačine sistema (1), pomnožene odgovarajućim množiteljem, od svih ostalih jednačina. Množitelji se određuju tako da se anulira promjenljiva x_1 u svim jednačinama izuzev u prvoj, tako da je pri oduzimanju od i–te jednačine množitelj a_{i1}/a_{11}, $i = 2, \dots, n$. Očigledno je neophodno za realizaciju prvog koraka da bude $a_{11} \neq 0$. Ukoliko taj uslov u sistemu (1) nije zadovoljen, permutacijom jednačina sistema, tj. stavljanjem na prvo mjesto jednačine kod koje je $a_{p1} \neq 0$ ovaj uslov se može ispuniti. Element a_{p1} matrice A postoji, jer je po pretpostavci matrica regularna i ne može imati sve nula elemente u prvoj koloni.

Napišimo proširenu matricu datog sistema

$$\left[\, A \,\middle|\, \mathbf{b} \,\right] = \left[\begin{array}{ccc|c} a_{11} & \dots & a_{1n} & b_1 \\ \vdots & \ddots & \vdots & \vdots \\ a_{n1} & \dots & a_{nn} & b_n \end{array}\right]. \tag{3}$$

U prvom koraku transformiše se matrica (3) u matricu

$$\left[\begin{array}{c|c} A_1 & \mathbf{b}_1 \end{array}\right] = \left[\begin{array}{cccc|c} a_{11}^{(1)} & a_{12}^{(1)} & \dots & a_{1n}^{(1)} & b_1^{(1)} \\ 0 & a_{22}^{(1)} & \dots & a_{2n}^{(1)} & b_2^{(1)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & a_{n2}^{(1)} & \dots & a_{nn}^{(1)} & b_n^{(1)} \end{array}\right].$$

Pri tome se eventualno vrši permutacija prve i p–te vrste. Element a_{p1} naziva se glavni element ili pivot, a postupak nalaženja glavnog elementa naziva se izbor glavnog elementa ili pivotiranje. Zbog numeričke stabilnosti algoritma, obično se među svim nenula elementima bira najveći po modulu, $|a_{p1}| = \max_{1\le i\le n} |a_{i1}|$. Nalaženje pivota među elementima samo jedne kolone matrice naziva se djelimično pivotiranje. Radi jednostavnosti, pretpostavimo da nije neophodno vršiti permutacije vrsta matrice A.

Sljedeći korak eliminacije se sastoji u primjeni opisanog postupka na sistem dimenzije $(n-1)$, tj. na sistem sa proširenom matricom $i, j \ge 2$. Na taj način se u svakom koraku dimenzija sistema koji se transformiše smanjuje za jedan. U j–tom koraku proširena matrica sistema ima sljedeći oblik

$$\left[\begin{array}{c|c} A_j & \mathbf{b}_j \end{array}\right] = \left[\begin{array}{ccc|ccc|c} * & \dots & * & * & \dots & * & * \\ \vdots & \ddots & \vdots & \vdots & & \vdots & \vdots \\ 0 & \dots & * & * & \dots & * & * \\ \hline 0 & \dots & 0 & * & \dots & * & * \\ \vdots & & \vdots & \vdots & & \vdots & \vdots \\ 0 & \dots & 0 & * & \dots & * & * \end{array}\right] = \left[\begin{array}{c|c|c} A_{11}^{(j)} & A_{12}^{(j)} & \mathbf{b}_1^{(j)} \\ \hline 0 & A_{22}^{(j)} & \mathbf{b}_2^{(j)} \end{array}\right], \tag{4}$$

gdje $*$ označava u opštem slučaju nenulte elemente, $A_{11}^{(j)}$ je gornje trougaona matrica dimenzije $j \times j$, a dalje se transformiše proširena matrica $\left[\begin{array}{c|c} A_{22}^{(j)} & \mathbf{b}_2^{(j)} \end{array}\right]$ dimenzije $(n-j)\times(n-j+1)$. U $(n-1)$–vom koraku elementi matrice $\left[\begin{array}{c|c} A_{n-1} & \mathbf{b}_{n-1} \end{array}\right]$ su gornje trougaona matrica $A_{11}^{(n-1)}$ dimenzije $(n-1)\times(n-1)$, matrica $A_{12}^{(n-1)}$ dimenzije $1\times(n-1)$, matrica $A_{22}^{(n-1)}$ dimenzije 1×1 i vektori $\mathbf{b}_1^{(n-1)}$ dimenzije $(n-1)$ i $\mathbf{b}_2^{(n-1)}$ dimenzije 1. Dakle, matrica A_{n-1} je tražena gornje trougaona matrica U, a vektor $\mathbf{b}_{n-1}$ vektor c desne strane sistema (2). Tako smo dobili niz matrica oblika (4)

$$\left[\begin{array}{c|c} A & \mathbf{b} \end{array}\right] \;\to\; \left[\begin{array}{c|c} A_1 & \mathbf{b}_1 \end{array}\right] \;\to\; \dots \;\to\; \left[\begin{array}{c|c} A_{n-1} & \mathbf{b}_{n-1} \end{array}\right] = \left[\begin{array}{c|c} U & \mathbf{c} \end{array}\right].$$

Računske radnje koje služe da se izračuna proširena matrica $\left[\begin{array}{c|c} U & \mathbf{c} \end{array}\right]$ čine tzv. direktni hod algoritma, dok rješavanje sistema sa trougaonom matricom $U\mathbf{x} = \mathbf{c}$ čini njegov obrnuti hod. Time je Gaussov algoritam izložen u potpunosti.

Primjer za Gaussovu metodu ($n = 4$). U primjeru se pojavljuje proširena matrica sistema. U nastavku, "E" znači "equation" (jednačina). Gdje piše "E1", to znači "equation 1" i slično:

$$\begin{array}{l} x_1 + 2x_2 + 3x_3 + 4x_4 = 8 \\ 2x_1 + 5x_2 + 7x_3 + x_4 = 27 \\ 3x_1 + 8x_2 + 13x_3 + 2x_4 = 48 \\ -x_1 + x_2 + 4x_3 + 2x_4 = 10 \end{array} \qquad \left[\begin{array}{cccc|c} 1 & 2 & 3 & 4 & 8 \\ 2 & 5 & 7 & 1 & 27 \\ 3 & 8 & 13 & 2 & 48 \\ -1 & 1 & 4 & 2 & 10 \end{array}\right] \qquad \begin{array}{c} -2E1 + E2 \to E2 \\ -3E1 + E3 \to E3 \\ E1 + E4 \to E4 \end{array}$$

$$\left[\begin{array}{cccc|c} 1 & 2 & 3 & 4 & 8 \\ 0 & 1 & 1 & -7 & 11 \\ 0 & 2 & 4 & -10 & 24 \\ 0 & 3 & 7 & 6 & 18 \end{array}\right] \qquad \begin{array}{l} -2E2+E3 \to E3 \\ -3E2+E4 \to E4 \end{array} \qquad \left[\begin{array}{cccc|c} 1 & 2 & 3 & 4 & 8 \\ 0 & 1 & 1 & -7 & 11 \\ 0 & 0 & 2 & 4 & 2 \\ 0 & 0 & 4 & 27 & -15 \end{array}\right]$$

$$-2E3+E4 \to E4 \qquad \left[\begin{array}{cccc|c} 1 & 2 & 3 & 4 & 8 \\ 0 & 1 & 1 & -7 & 11 \\ 0 & 0 & 2 & 4 & 2 \\ 0 & 0 & 0 & 19 & -19 \end{array}\right] \qquad \begin{array}{l} x_1+2x_2+3x_3+4x_4=8 \\ x_2+x_3-7x_4=11 \\ 2x_3+4x_4=2 \\ 19x_4=-19 \end{array}$$

Sada, iz trougaonog oblika lako $x_4=-1$, $x_3=3$, $x_2=1$, $x_1=1$, čime smo došli do rješenja.

1.2. TRODIJAGONALNI SISTEM JEDNAČINA

Pri rješavanju različitih problema u numeričkoj matematici (interpolacija splajnovima, diferencijski granični zadaci, ...), javlja se potreba za rješavanjem sistema linearnih jednačina sa trodijagonalnom matricom

$$\begin{cases} c_1x_1+b_1x_2=d_1 \\ a_ix_{i-1}+c_ix_i+b_ix_{i+1}=d_i, \quad i=2,\dots,n-1 \\ a_nx_{n-1}+c_nx_n=d_n \end{cases} \tag{1}$$

Ovakav sistem se efikasno rješava upravo Gaussovom metodom eliminacije, jer je broj računskih operacija koje treba izvršiti asimptotski jednak $8n$. U prvom koraku se iz prve jednačine sistema (1), pod pretpostavkom da je $c_1 \neq 0$, izrazi x_1,

$$x_1=\alpha_2x_2+\beta_2, \quad \alpha_2=-\frac{b_1}{c_1}, \quad \beta_2=\frac{d_1}{c_1},$$

i eliminiše promjenljiva x_1 u drugoj jednačini sistema (1). Sada ova jednačina sadrži samo promjenljive x_2 i x_3, te se na isti način u drugom koraku izrazi x_2 pomoću x_3. U $(i-1)$–vom koraku se dobija veza

$$x_{i-1}=\alpha_ix_i+\beta_i, \tag{2}$$

pomoću koje eliminišemo promjenljivu x_{i-1} u i–toj jednačini sistema (1)

$$a_i(\alpha_ix_i+\beta_i)+c_ix_i+b_ix_{i+1}=d_i.$$

Ako napišemo ovu jednačinu u obliku (2),

$$x_i=-\frac{b_i}{c_i+a_i\alpha_i}x_{i+1}+\frac{d_i-a_i\beta_i}{c_i+a_i\alpha_i}=\alpha_{i+1}x_{i+1}+\beta_{i+1},$$

dobijamo rekurentne veze za izračunavanje koeficijenata α_i, β_i,

$$\alpha_{i+1}=-\frac{b_i}{c_i+a_i\alpha_i}, \quad \beta_{i+1}=\frac{d_i-a_i\beta_i}{c_i+a_i\alpha_i}, \quad i=1,\dots,n-1 \quad (a_1=0). \tag{3}$$

Poslije $(n-1)$ ovakvih koraka, sistem (1) se svodi na sistem

$$\begin{cases} x_{i-1} = \alpha_i x_i + \beta_i, & i = 2, \ldots, n \\ a_n x_{n-1} + c_n x_n = d_n \end{cases}$$

čije je rješenje neposredno određeno formulama

$$x_n = \frac{d_n - a_n \beta_n}{c_n + a_n \alpha_n}, \quad x_i = \alpha_{i+1} x_{i+1} + \beta_{i+1}, \quad i = n-1, \ldots, 1. \tag{4}$$

Stoga se Gaussova metoda eliminacije za rješavanje linearnog sistema jednačina sa trodijagonalnom matricom sistema svodi na izračunavanje α_i i β_i, $i = 2, \ldots, n$, po formulama (3) u direktnom hodu, i nalaženje rješenja sistema (4) u obrnutom hodu.

Primjer sa sistemom linearnih jednačina čija je matrica trodijagonalna $(n = 5)$:

$$\begin{array}{l} x_1 - x_2 = -1 \\ -x_1 + 2x_2 - x_3 = 0 \\ x_2 + 2x_3 + x_4 = 12 \\ 6x_3 + x_4 + x_5 = 27 \\ 6x_4 - 11x_5 = -31 \end{array} \quad \left(\begin{array}{c} E1 + E2 \to E2 \\ -E2 + E3 \to E3 \\ -2E3 + E4 \to E4 \\ 6E4 + E5 \to E5 \end{array} \right) \quad \begin{array}{l} x_1 - x_2 = -1 \\ x_2 - x_3 = -1 \\ 3x_3 + x_4 = 13 \\ -x_4 + x_5 = 1 \\ -5x_5 = -25 \end{array}$$

Iz trougaonog oblika lako $x_5 = 5$, $x_4 = 4$, $x_3 = 3$, $x_2 = 2$, $x_1 = 1$.

Za matricu se kaže da je trodijagonalna ako ona može imati nenula elemente samo na glavnoj dijagonali, na dijagonali neposredno iznad nje i na dijagonali neposredno ispod glavne dijagonale.

1.3. LU DEKOMPOZICIJA

Data kvadratna matrica A biće prikazana kao proizvod donje trougaone matrice L sa jedinicama na dijagonali i gornje trougaone matrice U, znači $A = LU$. Kaže se da je izvršena njena LU dekompozicija ili trougaona dekompozicija. Materijal koji će biti izložen u ovoj sekciji predstavlja direktan produžetak onoga o čemu u sekciji čiji je naslov Gaussova metoda eliminacije.

Pretnhodno opisane operacije sa jednačinama sistema $A\mathbf{x} = \mathbf{b}$ možemo prikazati matričnim operacijama. U prvom koraku transformiše se proširena matrica sistema $\left[\, A \,\middle|\, \mathbf{b} \,\right]$ u matricu $\left[\, A_1 \,\middle|\, \mathbf{b}_1 \,\right]$, što je ekvivalentno množenju proširene matrice sistema $\left[\, A \,\middle|\, \mathbf{b} \,\right]$ donje trougaonom matricom sa jedinicama na dijagonali

$$\left[\, A_1 \,\middle|\, \mathbf{b}_1 \,\right] = L_1 \left[\, A \,\middle|\, \mathbf{b} \,\right], \qquad L_1 = \begin{bmatrix} 1 & & & 0 \\ -l_{21} & 1 & & \\ & & \ddots & \\ -l_{n1} & 0 & & 1 \end{bmatrix}, \qquad l_{i1} = \frac{a_{i1}}{a_{11}}.$$

U j–tom koraku $\left[\begin{array}{c|c} A_j & \mathbf{b}_j \end{array}\right] = L_j \left[\begin{array}{c|c} A_{j-1} & \mathbf{b}_{j-1} \end{array}\right]$, gdje je

$$L_j = \begin{bmatrix} 1 & & & & & 0 \\ & \ddots & & & & \\ & & 1 & & & \\ & & -l_{j+1,j} & 1 & & \\ & & & & \ddots & \\ 0 & & -l_{nj} & 0 & & 1 \end{bmatrix}, \qquad l_{ij} = \frac{a_{ij}^{(j-1)}}{a_{jj}^{(j-1)}},, \qquad i > j.$$

Već je rečeno $j = 1, \ldots, n-1$. Isto tako $\left[\begin{array}{c|c} A_{n-1} & \mathbf{b}_{n-1} \end{array}\right] = \left[\begin{array}{c|c} U & \mathbf{c} \end{array}\right]$. Napomenimo da smo pretpostavili da se u toku realizacije algoritma ne vrše permutacije.

Stoga je $\left[\begin{array}{c|c} U & \mathbf{c} \end{array}\right] = L_{n-1}L_{n-2}\ldots L_1 \left[\begin{array}{c|c} A & \mathbf{b} \end{array}\right]$, tj. $L_1^{-1}\ldots L_{n-1}^{-1}\left[\begin{array}{c|c} U & \mathbf{c} \end{array}\right] = \left[\begin{array}{c|c} A & \mathbf{b} \end{array}\right]$.

S obzirom da je

$$L_j^{-1} = \begin{bmatrix} 1 & & & & & 0 \\ & \ddots & & & & \\ & & 1 & & & \\ & & l_{j+1,j} & 1 & & \\ & & & & \ddots & \\ 0 & & l_{nj} & 0 & & 1 \end{bmatrix}$$

to je

$$L = L_1^{-1}\ldots L_{n-1}^{-1} = \begin{bmatrix} 1 & & & & 0 \\ l_{21} & 1 & & & \\ & & \ddots & & \\ & & & \ddots & \\ l_{n1} & l_{n2} & & l_{n,n-1} & 1 \end{bmatrix}$$

pa je definitivno $A = LU$.

LU dekompoziciju $A = LU$ možemo izvršiti direktno, ne formirajući niz matrica. Iz $A = LU$ dobijamo n^2 veza između elemenata matrice A sa jedne strane i elemenata matrica L i U sa druge strane. Sistem od n^2 jednačina

$$a_{ij} = \sum_{k=1}^{\min(i,j)} l_{ik}u_{kj} \qquad (l_{ii} = 1)$$

ima $\frac{n(n-1)}{2}$ nepoznatih l_{ij}, $1 \le j < i \le n$, i $\frac{n(n+1)}{2}$ nepoznatih u_{ij}, $1 \le i \le j \le n$. Poredak izračunavanja l_{ij} i u_{ij} može biti različit. Može se računati po formulama

$$u_{1k} = a_{1k}, \qquad u_{ik} = a_{ik} - \sum_{j=1}^{i-1} l_{ij}u_{jk}, \qquad k = i, \ldots, n,$$

$$l_{k1} = \frac{a_{k1}}{u_{11}}, \qquad l_{ki} = \frac{1}{u_{ii}}\Big(a_{ki} - \sum_{j=1}^{i-1} l_{kj}u_{ji}\Big), \qquad k = i+1, \ldots, n, \qquad i = 2, \ldots, n. \tag{1}$$

Gaussova eliminacija i LU dekompozicija se razlikuju samo u redosljedu operacija. Kako je element $a_{ik}^{(j)} = a_{ik} - \sum_{s=1}^{j} l_{is}u_{sk}$ matrice A_j, date u (4), ustvari j–ta parcijalna suma prve od

formula (1), to znači da se u Gaussovoj eliminaciji skalarni proizvod (1) računa postepeno i međurezultati se memorišu, dok se LU dekompozicijom taj skalarni proizvod računa odjednom u cjelini, što može biti prednost u smislu smanjenja računske greške (ako se međurezultati ne memorišu u dvostrukoj tačnosti).

Na kraju, ponovimo jednu rečenicu iz sekcije na koju se pozivamo. Ako je poznata LU dekompozicija neke matrice A, tj. poznate su matrice L i U, imamo da je $A\mathbf{x} = LU\mathbf{x} = \mathbf{b}$, te se rješavanje sistema $A\mathbf{x} = \mathbf{b}$ svodi na rješavanje dva trougaona sistema: $L\mathbf{y} = \mathbf{b}$, $U\mathbf{x} = \mathbf{y}$.

Primjer za LU dekompoziciju date matrice A (primjer za $A = LU$ kada je $n = 2$):

$$\begin{bmatrix} 4 & 3 \\ 6 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1{,}5 & 1 \end{bmatrix} \cdot \begin{bmatrix} 4 & 3 \\ 0 & -1{,}5 \end{bmatrix}.$$

Primjer za LU dekompoziciju date matrice A (primjer za $A = LU$ kada je $n = 3$):

$$\begin{bmatrix} 3 & 1 & 6 \\ 1 & 1 & 1 \\ 2 & 1 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1/3 & 1 & 0 \\ 2/3 & 1/2 & 1 \end{bmatrix} \cdot \begin{bmatrix} 3 & 1 & 6 \\ 0 & 2/3 & -1 \\ 0 & 0 & -1/2 \end{bmatrix}.$$

1.4. CHOLESKY DEKOMPOZICIJA

Razmotrimo realnu kvadratnu matricu A dimenzije $n \times n$. Pretpostavimo da je A simetrična i pozitivna (pozitivno definitna). To znači da je $A^T = A > 0$. Detaljnije, $a_{ji} = a_{ij}$ i skalarni proizvod $\langle A\mathbf{x}, \mathbf{x}\rangle > 0$ kada $\mathbf{x} \neq 0$.

Postoji jedinstveno određena donje trougaona matrica L, koja je također realna,

$$L = \begin{bmatrix} l_{11} & & & 0 \\ l_{21} & l_{22} & & \\ \vdots & \vdots & \ddots & \\ l_{n1} & l_{n2} & \dots & l_{nn} \end{bmatrix}, \quad l_{ii} > 0, \quad i = 1, \dots, n,$$

takva da je

$$A = LL^T. \tag{1}$$

Iz relacije (1) se dobijaju veze između elemenata matrica A i L,

$$a_{ii} = |l_{i1}|^2 + \dots + |l_{ii}|^2, \quad a_{ij} = l_{i1}l_{j1} + \dots + l_{ij}l_{jj}, \quad j < i, \quad i = 1, \dots, n, \tag{2}$$

te se elemrnti matrice L računaju po formulama

$$l_{11} = \sqrt{a_{11}}, \quad l_{i1} = \frac{a_{i1}}{l_{11}}, \quad l_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} |l_{ik}|^2}, \quad l_{ij} = \frac{1}{l_{jj}}\Big(a_{ij} - \sum_{k=1}^{j-1} l_{ik}l_{jk}\Big), \quad 1 < j < i \leq n. \tag{3}$$

S obzirom na simetričnost i pozitivnu definitnost matrice A, potkorjene veličine u formulama (3) su pozitivne. Kako iz (2) slijedi da je $|l_{ij}| \leq \sqrt{a_{ii}}$, elementi matrice L ne mogu biti suviše veliki, pa je metoda stabilna u nekom smislu.

Kada je matrica L određena, rješenje sistema $A\mathbf{x} = \mathbf{b}$ se nalazi rješavanjem dva trougaona sistema $L\mathbf{y} = \mathbf{b}$, $L^T\mathbf{x} = \mathbf{y}$.

Iz (1) je $\det(A) = \det(L)\det(L^T) = (\det(L))^2$, te je $\det(A) = (l_{11} \cdot \ldots \cdot l_{nn})^2$.

Primjer za Cholesky dekompoziciju (primjer za $A = LL^T$):

$$\begin{bmatrix} 4 & 12 & -16 \\ 12 & 37 & -43 \\ -16 & -43 & 98 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 6 & 1 & 0 \\ -8 & 5 & 3 \end{bmatrix} \cdot \begin{bmatrix} 2 & 6 & -8 \\ 0 & 1 & 5 \\ 0 & 0 & 3 \end{bmatrix}.$$

Vremenski trošak ("run time") za realizaciju algoritma za rješavanje sistema linearnih jednačina $A\mathbf{x} = \mathbf{b}$ baziranog na Cholesky dekompoziciji iznosi $t_n \sim \frac{1}{6}n^3$ aritmetičkih operacija množenja i dijeljenja, dok je u slučaju primjene Gaussove metode eliminacije bilo $t_n \sim \frac{1}{3}n^3$.

1.5. JACOBIJEVA METODA

U ovoj sekciji biće izložena jedna iterativna metoda za rješavanje sistema linearnih jednačina i biće dokazana teorema o dovoljnim uslovima za njenu konvergenciju. Neka je A realna matrica dimenzije $n \times n$. U numeričkoj metodi, pretpostavlja se da je matrica sistema A dijagonalno dominantna. Za matricu A kaže se da je dijagonalno dominantna ako postoji broj $q < 1$ takav da za svako $i = 1, \ldots, n$ važi nejednakost $\sum_{j \neq i} |a_{ij}| \leq q|a_{ii}|$ (suma po $j = 1, \ldots, i-1, i+1, \ldots, n$), $a_{ii} \neq 0$. Iz linearne algebre je poznato da je takva matrica regularna.

Razmotrimo sistem linearnih jednačina $A\mathbf{x} = \mathbf{b}$, gdje je

$$A = \begin{bmatrix} a_{11} & \ldots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \ldots & a_{nn} \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}.$$

Matricu A rastavimo na zbir dijagonalne matrice P i matrice Q koja ima nule po glavnoj dijagonali, $A = P + Q$,

$$P = \begin{bmatrix} a_{11} & & 0 \\ & \ddots & \\ 0 & & a_{nn} \end{bmatrix}, \quad Q = \begin{bmatrix} 0 & a_{12} & \ldots \\ a_{21} & 0 & \ldots \\ \vdots & \vdots & \ddots \end{bmatrix}.$$

Pomnožimo relaciju $P\mathbf{x} + Q\mathbf{x} = \mathbf{b}$ sa P^{-1}: $\mathbf{x} = -P^{-1}Q\mathbf{x} + P^{-1}\mathbf{b}$ ili $\mathbf{x} = B\mathbf{x} + \mathbf{c}$,

$$B = \begin{bmatrix} 0 & -a_{12}/a_{11} & -a_{13}/a_{11} & \ldots \\ -a_{21}/a_{22} & 0 & -a_{23}/a_{22} & \ldots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \quad \mathbf{c} = \begin{bmatrix} b_1/a_{11} \\ \vdots \\ b_n/a_{nn} \end{bmatrix}.$$

Uzastopne aproksimacije računaju se po formuli $\mathbf{x}_{k+1} = B\mathbf{x}_k + \mathbf{c}$ $(k \geq 1)$. Da li niz uzastopnih aproksimacija konvergira kad $k \to \infty$ ka rješenju $\mathbf{x}$ sistema $A\mathbf{x} = \mathbf{b}$?

Teorema. Ako je matrica A dijagonalno dominantna (za svako i, $\sum_{j \neq i} |a_{ij}| \leq q|a_{ii}|$) onda važi $\lim_{k \to \infty} \mathbf{x}_k = \mathbf{x}$, bez obzira na izbor početne aproksimacije $\mathbf{x}_0$.

Dokaz. Mi ćemo dokazati teoremu svođenjem na princip kontrakcije, na Banachovu teoremu o fiksnoj tački u kojoj se posmatra preslikavanje φ: $X \to X$ koje je kontrakcija u kompletnom metričkom prostoru X. Razmotrimo vektorski prostor R^n i za proizvoljni vektor $\mathbf{x}$ iz

tog prostora stavimo $\|\mathbf{x}\| = \max_{1\le i\le n}|x_i|$, obično se označava kao $\|\mathbf{x}\|_\infty$. Poznato je da su sve aksiome norme ispunjene. Također je poznato da indukovana norma linearnog operatora B iznosi $\|B\| = \max_{1\le i\le n}\left(\sum_{j=1}^n |b_{ij}|\right)$. Kada se za dva proizvoljna vektora $\mathbf{x}$ i $\mathbf{y}$ iz prostora $X = R^n$ definiše rastojanje kao $d(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|$ onda imamo jedan kompletan metrički prostor (X,d).

Da izračunamo $\|B\|$:

$$\sum_{j=1}^n |b_{ij}| = \sum_{j\ne i}|a_{ij}|\Big/|a_{ii}| \quad \Rightarrow \quad \sum_{j=1}^n |b_{ij}| \le q \quad \Rightarrow \quad \|B\| \le q < 1.$$

Razmotrimo preslikavanje $\varphi(\mathbf{x}) = B\mathbf{x} + \mathbf{c}$ $(\mathbf{x}\in X)$. Za dva proizvoljna vektora $\mathbf{x}$ i $\mathbf{y}$:

$$\varphi(\mathbf{y}) = B\mathbf{y}+\mathbf{c}, \quad \varphi(\mathbf{x}) = B\mathbf{x}+\mathbf{c}, \quad \varphi(\mathbf{y})-\varphi(\mathbf{x}) = B\mathbf{y}+\mathbf{c}-B\mathbf{x}-\mathbf{c} = B(\mathbf{y}-\mathbf{x}) \quad \Rightarrow$$

$$\|\varphi(\mathbf{y})-\varphi(\mathbf{x})\| = \|B(\mathbf{y}-\mathbf{x})\| \le \|B\|\cdot\|\mathbf{y}-\mathbf{x}\| \le q\|\mathbf{y}-\mathbf{x}\| \quad \Rightarrow \quad d(\varphi(\mathbf{x}),\varphi(\mathbf{y})) \le q d(\mathbf{x},\mathbf{y}).$$

Mi smo pokazali da su ispunjeni svi uslovi principa kontrakcije, pa zato niz $\mathbf{x}_k = \varphi(\mathbf{x}_{k-1})$ konvergira ka jedinstvenom rješenju $\mathbf{x}$ jednačine $\varphi(\mathbf{x}) = \mathbf{x}$. Dokaz je završen.

Sa principom kontrakcije obično dolaze i dvije formule koje možemo koristiti za ocjenu greške k–te aproksimacije:

$$\|\mathbf{x}-\mathbf{x}_k\| \le \frac{q^k}{1-q}\|\mathbf{x}_1-\mathbf{x}_0\|, \qquad \|\mathbf{x}-\mathbf{x}_k\| \le \frac{q}{1-q}\|\mathbf{x}_k-\mathbf{x}_{k-1}\|.$$

Primjer za Jacobijevu metodu. Razmotrimo $A\mathbf{x} = \mathbf{b}$, tj. $\mathbf{x} = B\mathbf{x}+\mathbf{c} \Rightarrow \mathbf{x}_{k+1} = B\mathbf{x}_k + \mathbf{c}$ ili

$$\begin{bmatrix} 10 & -1 & -5 \\ 1 & 5 & -2 \\ 4 & -1 & 10 \end{bmatrix}\cdot\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -12 \\ 3 \\ 42 \end{bmatrix}, \quad \text{tj.} \quad \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 & 0{,}1 & 0{,}5 \\ -0{,}2 & 0 & 0{,}4 \\ -0{,}4 & 0{,}1 & 0 \end{bmatrix}\cdot\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} -1{,}2 \\ 0{,}6 \\ 4{,}2 \end{bmatrix}$$

$$\Rightarrow \quad \begin{bmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \\ x_3^{(k+1)} \end{bmatrix} = \begin{bmatrix} 0 & 0{,}1 & 0{,}5 \\ -0{,}2 & 0 & 0{,}4 \\ -0{,}4 & 0{,}1 & 0 \end{bmatrix}\cdot\begin{bmatrix} x_1^{(k)} \\ x_2^{(k)} \\ x_3^{(k)} \end{bmatrix} + \begin{bmatrix} -1{,}2 \\ 0{,}6 \\ 4{,}2 \end{bmatrix}.$$

Izaberimo početni vektor $\mathbf{x}_0 = \begin{bmatrix} x_1^{(0)} & x_2^{(0)} & x_3^{(0)} \end{bmatrix}^T = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}^T$. Pomoću kompjutera, izračunali smo $\mathbf{x}_1,\ldots,\mathbf{x}_{10}$, a rezultati su prikazani u narednoj tabeli:

i	$x_1^{(i)}$	$x_2^{(i)}$	$x_3^{(i)}$
1	−1,2	0,6	4,2
2	0,96	2,52	4,74
3	1,422	2,304	4,068
4	1,0644	1,9428	3,8616
5	0,92508	1,93176	3,96852
6	0,97744	2,00239	4,02314
7	1,01181	2,01377	4,00926
8	1,00601	2,00134	3,99665
9	0,99846	1,99746	3,99773
10	0,99861	1,99940	4,00036

Itd. Niz $\{\mathbf{x}_k\}_{k=0}^{\infty}$ konvergira ka $\mathbf{x} = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}^T$, gdje je $x_1 = 1$, $x_2 = 2$, $x_3 = 4$.

Dijagonalno dominantna $\sum\limits_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| < |a_{ii}|$.

1.6. METODA KONJUGOVANIH GRADIJENATA

Metoda konjugovanih gradijenata služi za rješavanje sistema linearnih jednačina $A\mathbf{x} = \mathbf{b}$ u slučaju simetrične i pozitivno definitne realne kvadratne matrice A ($A^T = A > 0$). Razmotrimo funkcional $f(\mathbf{x}) = \langle A\mathbf{x}, \mathbf{x}\rangle - 2\langle \mathbf{b}, \mathbf{x}\rangle$. Lako se pokazuje da se minimum funkcije $f = f(\mathbf{x})$ ostvaruje u tački $\mathbf{x} \in R^n$ koja predstavlja rješenje sistema. Tako da ćemo rješavati problem o minimumu funkcije f. Znamo da je f skalarno polje, a grad f vektorsko polje (druga oznaka ∇f). Znamo da je po definiciji grad $f = \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n}\right)$. Fiksirajmo za trenutak jednu tačku u R^n i označimo je kao $\mathbf{x}$ ili kao P. Iz matematičke analize poznato je sljedeće. Posmatrajmo tačku Q blizu tačke P i posmatrajmo vrijednosti $f(P)$ i $f(Q)$. Da li je $f(Q)$ veće, u kom smjeru $\overrightarrow{PQ}$ se postiže najbrži rast (povećanje) vrijednosti f? Naravno, to je smjer gradijenta. Slično, ako želimo smanjenje f onda treba izabrati smjer $-$ grad $f(P)$, ponekad se kaže da je to smjer anti–gradijenta. Lako je izračunati da je u razmatranom slučaju grad $f(\mathbf{x}) = 2A\mathbf{x} - 2\mathbf{b}$.

U numeričkoj metodi, na početku se izabere jedna proizvoljna tačka P_0. U prvoj iteraciji treba odrediti tačku P_1 u kojoj f ima manju vrijednost. Dogovorili smo se da ćemo od tačke P_0 preći određeni put po liniji $-$ grad $f(P_0)$ i tako stići do tačke P_1. Znači, $P_1 = P_0 - \frac{\alpha}{2}$ grad $f(P_0)$, za neko $\alpha > 0$. Treba se držati izabranog pravca sve dok f opada. Vidimo da smo postavili problem o minimumu u slučaju jedne promjenljive: naći $\alpha \in R$ za koje se ostvaruje najmanja moguća vrijednost izraza $f(\mathbf{x}_0 - \alpha(A\mathbf{x}_0 - \mathbf{b}))$. Mali problem se lako rješava i odgovor glasi $\alpha = \alpha_0 = \langle \mathbf{r}_0, \mathbf{r}_0\rangle / \langle A\mathbf{r}_0, \mathbf{r}_0\rangle$, gdje je $\mathbf{r}_0 = A\mathbf{x}_0 - \mathbf{b}$. Obično se za $\mathbf{r} = A\mathbf{x} - \mathbf{b}$ kaže da predstavlja "nepovezanost" ili "ostatak". Prema tome, imamo definitivno $P_1 = P_0 - \frac{\alpha_0}{2}$ grad $f(P_0) = P_0 - \alpha_0(A\mathbf{x}_0 - \mathbf{b})$ ili svejedno $\mathbf{x}_1 = \mathbf{x}_0 - \frac{\alpha_0}{2}$ grad $f(\mathbf{x}_0)$.

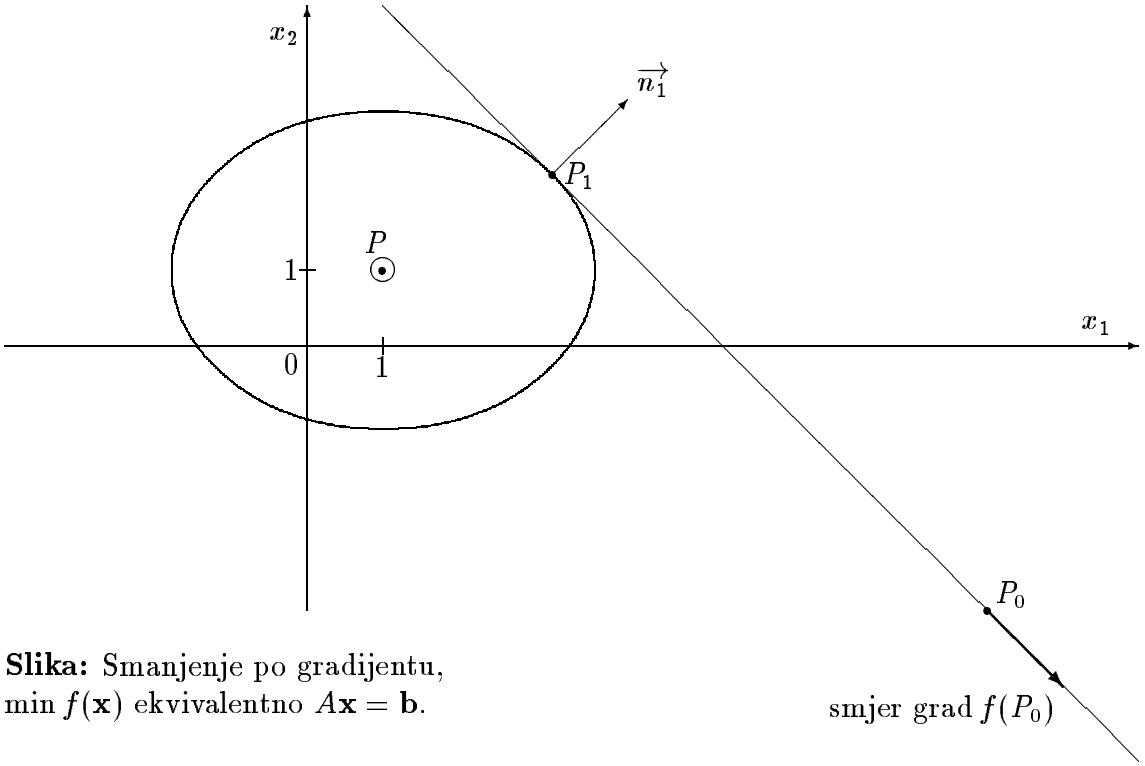
Oko ekvivalentnosti dva problema: važi $f(\mathbf{x}) - f(\mathbf{x}^*) = \langle A(\mathbf{x} - \mathbf{x}^*), \mathbf{x} - \mathbf{x}^*\rangle \geq 0$, gdje je $\mathbf{x}^* = A^{-1}\mathbf{b}$ (pa i tačka minimuma). Oko grad: treba primijeniti operaciju $\left(\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_n}\right)$ na $\langle A\mathbf{x}, \mathbf{x}\rangle - 2\langle \mathbf{x}, \mathbf{b}\rangle = \sum_{i,j=1}^{n} a_{ij}x_ix_j - 2\sum_{i=1}^{n} b_ix_i$, očito $\mathbf{x} = (x_1, \ldots, x_n)$, komponente vektora. Ili: $c_i = \lim_{\alpha\to 0} \frac{f(\mathbf{x}+\alpha\mathbf{e}_i) - f(\mathbf{x})}{\alpha} = 2\langle A\mathbf{x}, \mathbf{e}_i\rangle - 2\langle \mathbf{x}, \mathbf{e}_i\rangle$, grad $f(\mathbf{x}) = \sum_{i=1}^{n} c_i\mathbf{e}_i = 2A\mathbf{x} - 2\mathbf{b}$. Oko α: $f(\mathbf{x}_0 + \alpha\mathbf{c}) = \alpha^2\langle A\mathbf{c}, \mathbf{c}\rangle + 2\alpha\langle \mathbf{r}_0, \mathbf{c}\rangle + f(\mathbf{x}_0)$, već je bilo $\mathbf{r}_0 = A\mathbf{x}_0 - \mathbf{b}$. Zatim $\mathbf{c} = \mathbf{r}_0$ i $\frac{d}{d\alpha} f(\mathbf{x}_0 + \alpha\mathbf{c}) = 0$. Na taj način, $2\alpha\langle A\mathbf{r}_0, \mathbf{r}_0\rangle + 2\langle \mathbf{r}_0, \mathbf{r}_0\rangle = 0$. Zanimljivo je zapaziti da su vektori grad $f(\mathbf{x}_0)$ i $\mathbf{r}_0$ kolinearni.

Slično druga iteracija, itd.

Samo se napominje da je detaljno izlaganje metode konjugovanih gradijenata prilično komplikovano. Prva iteracija metode konjugovanih gradijenata poklapa se sa prvom iteracijom metode gradijentnog spusta.

Na redu je najjednostavniji mogući primjer. Neka je dato $A = \begin{bmatrix} 9 & 0 \\ 0 & 16 \end{bmatrix}$ i $b = \begin{bmatrix} 9 \\ 16 \end{bmatrix}$ (tako da je $(x_1, x_2) = (1, 1)$). Odgovarajući funkcional glasi $f(x_1, x_2) = 9x_1^2 + 16x_2^2 - 18x_1 - 32x_2$. Takođe je dato i $\mathbf{x}_0 = (9, -3{,}5)$, to je proizvoljno izabrana početna tačka P_0. Mi redom računamo, u cilju određivanja tačke P_1: grad $f(\mathbf{x}) = (18x_1 - 18, 32x_2 - 32)$, grad $f(\mathbf{x}_0) = (144, -144)$, $\mathbf{r}_0 = A\mathbf{x}_0 - \mathbf{b} = (72, -72)$, $A\mathbf{r}_0 = (648, -1152)$, $\alpha_0 = \frac{\langle \mathbf{r}_0, \mathbf{r}_0\rangle}{\langle A\mathbf{r}_0, \mathbf{r}_0\rangle} = \frac{2}{25}$, $\mathbf{x}_1 = \mathbf{x}_0 - \frac{\alpha_0}{2}$ grad $f(\mathbf{x}_0) = (3{,}24, 2{,}26)$. Time smo dobili odgovor. Broj $f(\mathbf{x}_1)$ manji je od broja $f(\mathbf{x}_0)$, rastojanje $d(P_1, P)$ manje je od rastojanja $d(P_0, P)$, gdje $P(1, 1)$ predstavlja cilj. V. sliku.

Na slici su prikazani: (1) tačka P koja odgovara minimumu funkcije $f(x_1, x_2)$, (2) polazna tačka P_0, (3) smjer grad $f(P_0)$ od P_0 SE, (4) suprotni smjer od P_0 NW prema P_1 i dalje, (5) iduća tačka P_1 i (6) nivo–linija $f(x_1, x_2) = \text{const}$, $f(\mathbf{x}) = f(\mathbf{x}_1)$, elipsa sa centrom u tački P, jednačina elipse $(x_1 - 1)^2/16 + (x_2 - 1)^2/9 = 0{,}49$. Zapaža se da prava linija tangira elipsu, tako da je prikazana i spoljašnja normala $\overrightarrow{n_1}$.



Slika: Smanjenje po gradijentu, $\min f(\mathbf{x})$ ekvivalentno $A\mathbf{x} = \mathbf{b}$.

2. METODE ZA RAČUNANJE SVOJSTVENIH VRIJEDNOSTI (SOPSTVENIH VRIJEDNOSTI) MATRICE

Riješiti potpuni problem svojstvenih vrijednosti za datu (realnu) kvadratnu matricu A reda n znači odrediti sve njene svojstvene vrijednosti i odgovarajuće svojstvene vektore. Riješiti djelimični problem znači odrediti neke svojstvene vrijednosti ili jednu svojstvenu vrijednost. Za male vrijednosti n, svojstvene vrijednosti mogu da budu određene iz uslova $\det(A - \lambda I) = 0$, gdje je I jedinična matrica.

2.1. METODA STEPENA

Na engleskom se kaže power method. Takođe se kaže i metoda skalarnog proizvoda. Numerička metoda služi za računanje najveće po modulu svojstvene vrijednosti date simetrične matrice A ($A^T = A$).

Razmotrimo realnu simetričnu matricu A dimenzije $n \times n$:

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \cdots & \cdots & \cdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix}$$

($a_{ji} = a_{ij}$). Označimo njene svojstvene vrijednosti kao $\lambda_1, \ldots, \lambda_n$, gdje $\lambda_i \in R$ ($i = 1, \ldots, n$), svaka svojstvena vrijednost broji se sa svojom višestrukošću. Neka je numeracija izvršena tako da važi

$$|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_n|.$$

Označimo odgovarajuće svojstvene vektore sa $\mathbf{e}_i$, gdje $\mathbf{e}_i \in R^n$ ($i = 1, \ldots, n$). tako da važi $A\mathbf{e}_i = \lambda_i \mathbf{e}_i$. Iz linearne algebre znamo: budući da je matrica A simetrična to važi $\mathbf{e}_i \perp \mathbf{e}_j$, tj. $\langle \mathbf{e}_i, \mathbf{e}_j \rangle = 0$ kada je $i \neq j$. Svakako da se skalarni proizvod vektora $\mathbf{x}$ čije su komponente $x_1, \ldots, x_n$ i vektora $\mathbf{y}$ čije su komponente $y_1, \ldots, y_n$ definiše kao $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^n x_i y_i$. Izaberimo svojstvene vektore tako da važi $\|\mathbf{e}_i\| = 1$. Svakako da se norma vektora definiše kao $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$. Sistem vektora $(\mathbf{e}_1, \ldots, \mathbf{e}_n)$ čini ortonormiranu bazu vektorskog prostora R^n.

Možemo pisati $\mathbf{x} = \sum_{i=1}^n \langle \mathbf{x}, \mathbf{e}_i \rangle \mathbf{e}_i$, kao i $A\mathbf{x} = \sum_{i=1}^n \lambda_i \langle \mathbf{x}, \mathbf{e}_i \rangle \mathbf{e}_i$.

Prelazimo na izgradnju numeričke metode. Biće konstruisan niz brojeva $\{\mu_k\}_{k=0}^{\infty}$ ($\mu_k \in R$) takav da je $\lim\limits_{k \to \infty} \mu_k = \lambda_1$, pod određenim pretpostavkama.

Izaberimo na proizvoljan način vektor $\mathbf{x}_0 \in R^n$, treba $\mathbf{x}_0 \neq (0, \ldots, 0)$. Vektor $\mathbf{x}_0$ razložimo po bazi $(\mathbf{e}_1, \ldots, \mathbf{e}_n)$. Imamo da je $\mathbf{x}_0 = \sum_{i=1}^n c_i \mathbf{e}_i$, gdje $c_i \in R$. Znamo da je $c_i = \langle \mathbf{x}_0, \mathbf{e}_i \rangle$.

Stavimo

$$\mathbf{x}_1 = A\mathbf{x}_0, \quad \mathbf{x}_2 = A\mathbf{x}_1, \quad \ldots, \quad \mathbf{x}_k = A\mathbf{x}_{k-1}, \quad \ldots.$$

Imamo da je $\mathbf{x}_1 = A\mathbf{x}_0 = A \sum_{i=1}^n c_i \mathbf{e}_i = \sum_{i=1}^n c_i A\mathbf{e}_i = \sum_{i=1}^n c_i \lambda_i \mathbf{e}_i$. Slično, $\mathbf{x}_2 = \sum_{i=1}^n c_i \lambda_i^2 \mathbf{e}_i$, ..., $\mathbf{x}_k = \sum_{i=1}^n c_i \lambda_i^k \mathbf{e}_i$, ... Izračunajmo skalarne proizvode $\langle \mathbf{x}_k, \mathbf{x}_k \rangle$ i $\langle \mathbf{x}_{k+1}, \mathbf{x}_k \rangle$. Imamo u vidu da je $\mathbf{e}_i \perp \mathbf{e}_j$ kada je $i \neq j$:

$$\langle \mathbf{x}_k, \mathbf{x}_k \rangle = \sum_{i,j=1}^n c_i c_j \lambda_i^k \lambda_j^k \langle \mathbf{e}_i, \mathbf{e}_j \rangle = \sum_{i=1}^n c_i^2 \lambda_i^{2k},$$

$$\langle \mathbf{x}_{k+1}, \mathbf{x}_k \rangle = \sum_{i,j=1}^n c_i c_j \lambda_i^{k+1} \lambda_j^k \langle \mathbf{e}_i, \mathbf{e}_j \rangle = \sum_{i=1}^n c_i^2 \lambda_i^{2k+1}.$$

Stavimo

$$\mu_k = \frac{\langle \mathbf{x}_{k+1}, \mathbf{x}_k \rangle}{\langle \mathbf{x}_k, \mathbf{x}_k \rangle}, \quad k \geq 0.$$

Sada je aproksimacioni niz $\{\mu_k\}_{k=0}^{\infty}$ definisan i proces računanja ili algoritam je definisan.

Mi ćemo pretpostaviti da je ispunjeno $|\lambda_1| > |\lambda_2|$. Drugim riječima, da je najveća po modulu svojstvena vrijednost jedinstvena. Tada se za nju kaže da je dominantna. Takođe ćemo pretpostaviti da je $c_1 = \langle \mathbf{x}_0, \mathbf{e}_1 \rangle \neq 0$, koeficijent uz $\mathbf{e}_1$ u razlaganju vektora $\mathbf{x}_0$.

Teorema. Neka je $|\lambda_1| > |\lambda_2|$ i $c_1 = \langle \mathbf{x}_0, \mathbf{e}_1 \rangle \neq 0$. Tada važi $\lim\limits_{k\to\infty} \mu_k = \lambda_1$. Pored toga, $|\lambda_1 - \mu_k| = O(q^k)$ kad $k \to \infty$, gdje je $q = |\lambda_2/\lambda_1|^2$.

Dokaz teoreme:

$$\mu_k = \frac{c_1^2\lambda_1^{2k+1} + \ldots + c_n^2\lambda_n^{2k+1}}{c_1^2\lambda_1^{2k} + \ldots + c_n^2\lambda_n^{2k}} \quad \text{i vidimo da je} \quad \lim_{k\to\infty} \mu_k = \lambda_1,$$

$$\mu_k - \lambda_1 = \frac{c_2^2(\lambda_2 - \lambda_1)\lambda_2^{2k} + \ldots + c_n^2(\lambda_n - \lambda_1)\lambda_n^{2k}}{c_1^2\lambda_1^{2k} + \ldots + c_n^2\lambda_n^{2k}},$$

$$|\mu_k - \lambda_1| \leq \frac{1}{c_1^2\lambda_1^{2k}}\Big(c_2^2 \cdot 2|\lambda_1|\lambda_2^{2k} + \ldots + c_n^2 \cdot 2|\lambda_1|\lambda_n^{2k}\Big) = O\Big(\Big|\frac{\lambda_2}{\lambda_1}\Big|^{2k}\Big).$$

Teorema je dokazana.

Zapaža se da $\mathbf{x}_k \sim c_1\lambda_1^k\mathbf{e}_1$ $(k \to \infty)$.

Primjer za pojam svojstvenih vrijednosti i vektora. Razmotrimo matricu $A = \begin{bmatrix} 3 & -1 \\ -2 & 2 \end{bmatrix}$. Može se lako izr. da je $\lambda_1 = 4$, $\lambda_2 = 1$ i da su odgovarajući vektori $\mathbf{v}_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$ i $\mathbf{v}_2 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$. Da provjerimo. Zaista, tačno je $\begin{bmatrix} 3 & -1 \\ -2 & 2 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 4 \\ -4 \end{bmatrix}$ i $\begin{bmatrix} 3 & -1 \\ -2 & 2 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$.

Primjer za metodu stepena. Razmotrimo matricu $A = \begin{bmatrix} 5 & 1 \\ 1 & 3 \end{bmatrix}$ za koju važi $\lambda_{1,2} = 4 \pm \sqrt{2}$. Izaberimo $\mathbf{x}_0 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$. Izračunajmo iteracije do μ_2 preko $\mathbf{x}_k = A\mathbf{x}_{k-1}$, $\mu_k = \frac{\langle \mathbf{x}_{k+1}, \mathbf{x}_k \rangle}{\langle \mathbf{x}_k, \mathbf{x}_k \rangle}$. Redom $\mathbf{x}_1 = \begin{bmatrix} 4 \\ -2 \end{bmatrix}$, $\mu_0 = \frac{6}{2} = 3$, $\mathbf{x}_2 = \begin{bmatrix} 18 \\ -2 \end{bmatrix}$, $\mu_1 = \frac{76}{20} = 3{,}8$, $\mathbf{x}_3 = \begin{bmatrix} 88 \\ 12 \end{bmatrix}$, $\mu_2 = \frac{1560}{328} = 4{,}8$.

2.2. QR ALGORITAM

QR ALGORITHM – FROM WIKIPEDIA, THE FREE ENCYCLOPEDIA

In numerical linear algebra, the QR algorithm is an eigenvalue algorithm: that is, a procedure to calculate the eigenvalues and eigenvectors of a matrix. The QR transformation was developed in the late 1950s by John G.F. Francis (England) and by Vera N. Kublanovskaya (USSR), working independently. The basic idea is to perform a QR decomposition, writing the matrix as a product of an orthogonal matrix and an upper triangular matrix, multiply the factors in the reverse order and iterate.

The practical QR algorithm

Formally, let A be a real matrix of which we want to compute the eigenvalues, and let $A_0 = A$. At the k–th step (starting with $k = 0$), we compute the QR decomposition $A_k = Q_k R_k$ where Q_k is an orthogonal matrix (i.e. $Q^T = Q^{-1}$) and R_k is an upper triangular matrix. We then form $A_{k+1} = R_k Q_k$. Note that

$$A_{k+1} = R_k Q_k = Q_k^{-1} Q_k R_k Q_k = Q_k^{-1} A_k Q_k = Q_k^T A_k Q_k,$$

so all the A_k are similar and hence they have the same eigenvalues. The algorithm is numerically stable because it proceeds by orthogonal similarity transforms.

Under certain conditions the matrices A_k converge to a triangular matrix, the Schur form of A. The eigenvalues of a triangular matrix are listed on the diagonal and the eigenvalue problem is solved. In testing for convergence it is impractical to require exact zeros, but the Gershgorin circle theorem provides a bound on the error.

In this crude form the iterations are relatively expensive. This can be mitigated by first bringing the matrix A to upper Hessenberg form (which costs $\frac{10}{3}n^3 + O(n^2)$ arithmetic operations using a technique based on Householder reduction), with a finite sequence of orthogonal similarity transforms, somewhat like a two–sided QR decomposition. (For QR decomposition, the Householder reflectors are multiplied only on the left, but for the Hessenberg case they are multiplied on both left and right.) Determining the QR decomposition of an upper Hessenberg matrix costs $6n^2 + O(n)$ arithmetic operations. Moreover, because the Hessenberg form is already nearly upper–triangular (it has just one nonzero entry below each diagonal), using it as a starting point reduces the number of steps required for convergence of the QR algorithm.

If the original matrix is symmetric, then the upper Hessenberg matrix is also symmetric and thus tridiagonal, and so are all the A_k. This procedure costs $\frac{4}{3}n^3 + O(n^2)$ arithmetic operations using a technique based on Householder reduction. Determining the QR decomposition of a symmetric tridiagonal matrix costs $O(n)$ operations.

The rate of convergence depends on the separation between eigenvalues, so a practical algorithm will use shifts, either explicit or implicit, to increase separation and accelerate convergence. A typical symmetric QR algorithm isolates each eigenvalue (then reduces the size of the matrix) with only one or two iterations, making it efficient as well as robust.

The implicit QR algorithm

In modern computational practice, the QR algorithm is performed in an implicit version which makes the use of multiple shifts easier to introduce. The matrix is first brought to upper Hessenberg form $A_0 = QAQ^T$ as in the explicit version; then, at each step, the first column of A_k is transformed via a small–size Householder similarity transformation to the first column of $p(A_k)$ (or $p(A_k)\mathbf{e}_1$), where $p(A_k)$, of degree r, is the polynomial that defines the shifting strategy (often $p(x) = (x - \lambda)(x - \bar{\lambda})$, where λ and $\bar{\lambda}$ are the two eigenvalues of the trailing 2×2 principal submatrix of A_k, the so–called "implicit double-shift"). Then successive Householder transformation of size $r + 1$ are performed in order to return the working matrix A_k to upper Hessenberg form. This operation is known as "bulge chasing", due to the peculiar shape of the non–zero entries of the matrix along the steps of the algorithm. As in the first version, deflation is performed as soon as one of the sub–diagonal entries of A_k is sufficiently small.

Interpretation and convergence

The QR algorithm can be seen as a more sophisticated variation of the basic "power" eigenvalue algorithm. Recall that the power algorithm repeatedly multiplies A times a single vector, normalizing after each iteration. The vector converges to an eigenvector of the largest eigenvalue. Instead, the QR algorithm works with a complete basis of vectors, using QR decomposition to renormalize (and orthogonalize). For a symmetric matrix A, upon convergence, $AQ = Q\Lambda$, where Λ is the diagonal matrix of eigenvalues to which A converged, and where Q is a composite of all the orthogonal similarity transforms required to get there. Thus the columns of Q are the eigenvectors.

History

The QR algorithm was preceded by the LR algorithm, which uses the LU decomposition instead of the QR decomposition. The QR algorithm is more stable, so the LR algorithm is rarely used nowadays. However, it represents an important step in the development of the QR algorithm.

NESTABILNOST?

$***$ Numerical stability. In the mathematical subfield of numerical analysis, numerical stability is a generally desirable property of numerical algorithms. The concern is the growth of round-off errors and/or initially small fluctuations in initial data which might cause a large deviation of final answer from the exact solution. Some numerical algorithms may damp out the small fluctuations (errors) in the input data; others might magnify such errors (Wikipedia).

$***$ Linearni sistem je stabilan ukoliko malim promjenama ulaznih parametara (elementi matrice sistema i vektora slobodnih članova) odgovaraju male promjene rješenja (Desanka Radunović: Numeričke metode).

Primjer sa greškom ulaznih podataka. Razmotrimo sistem $x + 0{,}99y = a$, $0{,}99x + 0{,}98y = b$. Ispostavlja se da i neznatne promjene Δa i Δb brojeva a i b prouzrokuju velike promjene Δx i Δy rješenja sistema x i y.

Primjer sa greškom računanja (sa računskom greškom). Neka je $a = 1{,}0001$, $b = 1$, $y = \sqrt{a} - \sqrt{b}$. Kompjuter saopštava $y = 5{,}000705 \cdot 10^{-5}$, što je prilično neprecizno. Naime, može se raditi po formuli $y = (a - b)/(\sqrt{a} + \sqrt{b})$, odnosno $y = 0{,}0001/(\sqrt{a} + \sqrt{b})$. Tako ćemo dobiti tačan rezultat $y = 4{,}999875 \cdot 10^{-5}$ (ima svih 7 sigurnih cifara).

3. METODA KONAČNIH RAZLKIKA ZA RJEŠAVANJE GRANIČNOG ZADATKA ZA OBIČNE DIFERENCIJALNE JEDNAČINE

Biće konstruisana numerička metoda za dobijanje približnog rješenja graničnog zadatka drugog reda. Biće dokazano da numerička metoda ima tzv. red aproksimacije h^2 i da je ona stabilna u odnosu na svoju desnu stranu i u odnosu na svoja dva granična uslova. Na kraju će biti dokazano da ona konvergira, s tim da je red konvergencije (red tačnosti) takođe h^2.

3.1. NUMERIČKI ALGORITAM

Prvo se daje postavka analitičkog zadatka koji treba da bude numerički riješen. Neka $p \in C^2[0,X]$, $p(x) \geq 0$ za $x \in [0,X]$, $f \in C^2[0,X]$. Razmotrimo jednačinu

$$-y'' + p(x)y(x) = f(x) \quad \text{za} \quad 0 \leq x \leq X \tag{1}$$

i granične uslove

$$y(0) = a, \quad y(X) = b. \tag{2}$$

Iz teorije običnih d. j. poznata su sljedeća svojstva graničnog zadatka (1)–(2). Zadatak (1)–(2) ima jedinstveno rješenje $y = y(x)$ i $y \in C^4[0,X]$. Može se posmatrati diferencijalni operator $Ly = -y'' + p(x)y$ čiji je domen skup $\mathcal{D} = \{y\colon\ y \in C^2[0,X],\ y(0) = y(X) = 0\}$. Samo se napominje da je operator L simetričan, tj. važi $\langle Ly_1, y_2\rangle = \langle y_1, Ly_2\rangle$ za sve $y_1, y_2 \in \mathcal{D}$ (za simetričnost nije potrebno $p(x) \geq 0$). Pored toga, operator je i pozirivan, tj. važi $\langle Ly, y\rangle > 0$ za sve $y \in \mathcal{D}$ osim za $y(x) \equiv 0$. U Lebesgueovom prostoru $L^2(0,X)$, skalarni proizvod dvije funkcije definiše se kao $\langle y_1, y_2\rangle = \int_0^X y_1(x)y_2(x)dx$.

Neka je N prirodan broj i neka je $h = X/N$. Uvedimo ekvidistantnu mrežu čvorova $x_n = nh$ za $0 \leq n \leq N$. Neka je $y = y(x)$ analitičko rješenje razmatranog zadatka (1)–(2), tako da je $y(x_n)$ njegova vrijednost u čvoru $x = x_n$. Neka su y_n odgovarajuće približne vrijednosti, biće dobijene kasnije. Uvedimo i oznaku za grešku (grešku metode): $R_n = y(x_n) - y_n$ za $0 \leq n \leq N$.

Neka je

$$\ell_0(y(x_n)) = -\frac{1}{h^2}\Big(y(x_{n+1}) - 2y(x_n) + y(x_{n-1})\Big)$$

i

$$r_n = -\ell_0(y(x_n)) - y''(x_n) = \frac{1}{h^2}\Big(y(x_{n+1}) - 2y(x_n) + y(x_{n-1})\Big) - y''(x_n).$$

Za r_n se kaže da je greška aproksimacije. Poznata je sljedeća formula:

$$r_n = \frac{1}{12}y^{IV}(\xi_n)h^2, \quad \text{gdje je} \quad x_{n-1} < \xi_n < x_{n+1}.$$

Dokažimo formulu korišćenjem Taylorovog razvoja

$$y(x_n + \alpha) = y(x_n) + \alpha y'(x_n) + \frac{1}{2!}\alpha^2 y''(x_n) + \frac{1}{3!}\alpha^3 y'''(x_n) + \frac{1}{4!}\alpha^4 y^{IV}(x_n + \theta\alpha)$$

$(0 < \theta < 1)$ kada je $\alpha = h$ i $\alpha = -h$:

$$r_n = \frac{1}{h^2}\Big(y(x_n) + \frac{1}{2!}y''(x_n)h^2 + \frac{1}{4!}y^{IV}(\xi_1)h^4 - 2y(x_n) +$$

$$y(x_n) + \frac{1}{2!}y''(x_n)h^2 + \frac{1}{4!}y^{IV}(\xi_2)h^4\Big) - 2y''(x_n) =$$

$$\frac{1}{4!h^2}\left(y^{IV}(\xi_1)+y^{IV}(\xi_2)\right)=\frac{1}{12}y^{IV}(\xi)h^2.$$

Tako da je greška aproksimacije reda h^2. Ili $|r_n|\le\frac{1}{12}M_4h^2$, gdje je $M_4=\max_{x\in[0,X]}|y^{IV}(x)|$, gdje je $y=y(x)$ tačno rješenje za (1)–(2).

Neka je $p_n=p(x_n)$ i $f_n=f(x_n)$. Napišimo jednačinu (1) kada je $x=x_n$:

$$-y''(x_n)+p_ny(x_n)=f_n.$$

Uvedimo oznaku

$$\ell(y(x_n))=-\frac{1}{h^2}\left(y(x_{n+1})-2y(x_n)+y(x_{n-1})\right)+p_ny(x_n).$$

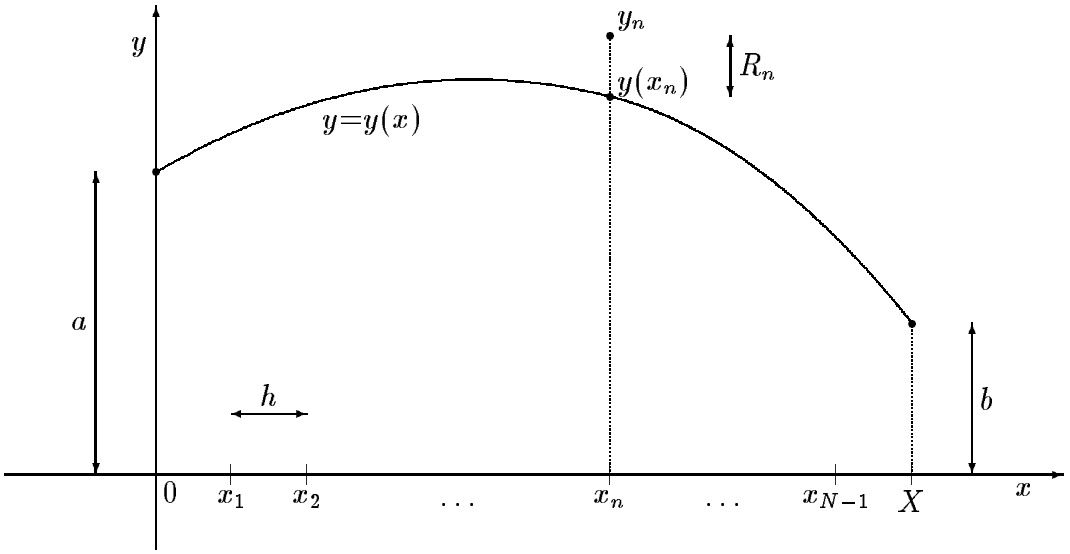
Imamo

$$\ell(y(x_n))=f_n-r_n.$$

U numeričkoj metodi, drugi izvod $y''(x)$ u čvoru $x=x_n$ biva zamijenjen podijeljenom razlikom koja je sastavljena od efektivno poznatih brojeva $\{y_n\}_{n=0}^N$ (biće dobijeni kasnije), tj. od y_{n-1}, y_n, y_{n+1}, a ne od brojeva $y(x_{n-1})$, $y(x_n)$, $y(x_{n+1})$. Mi ćemo saznati brojeve $\{y_n\}_{n=0}^N$ ako riješimo sljedeći sistem linearnih jednačina:

$$-\frac{1}{h^2}\left(y_{n+1}-2y_n+y_{n-1}\right)+p_ny_n=f_n\quad\text{ili}\quad\ell(y_n)=f_n,\quad 1\le n\le N-1,\tag{3}$$

$$y_0=a,\quad y_N=b.\tag{4}$$



3.2. TEOREMA O DOVOLJNIM USLOVIMA ZA KONVERGENCIJU NUMERIČKE METODE

Imamo da je $\ell(y(x_n))=f_n-r_n$ i $\ell(y_n)=f_n$. Oduzimanjem, uzimajući u obzir linearnost diferencnog izraza ℓ,

$$\ell(y(x_n))-\ell(y_n)=f_n-r_n-f_n\quad\text{ili}\quad\ell(y(x_n)-y_n)=-r_n\quad\text{ili}\quad\ell(R_n)=-r_n,\quad 1\le n\le N-1,$$

veza greške metode R_n i greške aproksimacije r_n.

Biće dokazana sljedeća teorema.

Teorema. Sistem linearnih jednačina (3)–(4) ima jedinstveno rješenje $\{y_n\}_{n=0}^N$ i važi sljedeća formula (za ocjenu greške)

$$\max_{0\le n\le N} |R_n| \le \frac{1}{96} X^2 M_4 h^2.$$

Mi ćemo ustvari dokazati nešto opštiju teoremu koja se odnosi na nešto opštiju situaciju. Prethodna teorema analizira samo grešku metode. Sljedeća teorema uzima u obzir i grešku računanja i grešku izazvanu približnošću ulaznih podataka; ulazni podaci su a i b. Neka veličine $\{y_n\}_{n=0}^N$ više ne zadovoljavaju sistem (3)–(4) nego odsad uzimamo da one zadovoljavaju sljedeći sistem:

$$-\frac{1}{h^2}\big(y_{n+1} - 2y_n + y_{n-1}\big) + p_n y_n = f_n + \delta_n \quad \text{ili} \quad \ell(y_n) = f_n + \delta_n, \quad 1 \le n \le N-1, \tag{5}$$

$$y_0 = a - R_0, \quad y_N = b - R_N. \tag{6}$$

Uslovi (5)–(6) bi se očito sveli na (3)–(4) da je $\delta_n = 0$ i $R_0 = R_N = 0$. Zašto su uvedeni δ_n? Kada se sistem linearnih jednačina riješi onda se njegovo rješenje radi provjere uvrsti u sami taj sistem. Lijeva i desna strana se ne poklope već se razlikuju za δ_n. Razlikuju se zato što se tokom rješavanja sistema akumulirala greška računanja. Još, brojevi δ_n odgovaraju i slučaju kada je desna strana jednačine $f = f(x)$ poznata samo približno, poznata sa nekom greškom. Zašto su potrebni R_0 i R_N? Moguće je da su brojevi a i b koji definišu par graničnih uslova samo približno poznate veličine.

Za mjeru greške računanja uzećemo $\max_{0<n<N} |\delta_n|$. Za mjeru greške ulaznih podataka uzećemo $\max(|R_0|, |R_N|)$. I dalje je $R_n = y(x_n) - y_n$. Sada R_n odražava sve tri komponente greške (metode, računanja i od ulaznih podataka).

Ranija veza $\ell(R_n) = -r_n$ sada u novoj situaciji očito postaje

$$\ell(R_n) = -r_n - \delta_n, \quad 1 \le n \le N-1.$$

Sada u novoj situaciji važi sljedeća teorema (koja će biti dokazana).

Teorema. Sistem linearnih jednačina (5)–(6) ima jedinstveno rješenje $\{y_n\}_{n=0}^N$ i važi sljedeća formula (za ocjenu greške)

$$\max_{0\le n\le N} |R_n| \le \frac{1}{96} X^2 M_4 h^2 + \frac{1}{8} X^2 \max_{0<n<N} |\delta_n| + \max(|R_0|, |R_N|).$$

Dokaz teoreme. Sistem linearnih jednačina (5) sastoji se od $N-1$ jednačina i ima $N-1$ nepoznatih $\{y_n\}_{n=1}^{N-1}$. Možemo pomnožiti jednačine sa $-h^2$. Označimo sa M matricu sistema, ona je oblika $(N-1)\times(N-1)$. Vidimo da je matrica M trodijagonalna. Prvo pitanje: pokazaćemo da je matrica M regularna, tj. da je $\det M \neq 0$; koristićemo sljedeće: $p(x) \ge 0$ za svako $x \in [0, X] \Rightarrow p_n \ge 0$ za svako $n \in \{1, \dots, N-1\}$. Mi pišemo

$$M = \begin{bmatrix} -2-p_1h^2 & 1 & \cdots & \cdots & \cdots & \cdots \\ 1 & -2-p_2h^2 & 1 & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \cdots & \cdots & \cdots & \cdots & 1 & -2-p_{N-1}h^2 \end{bmatrix}.$$

Recimo, u slučaju $p(x) \equiv 0$, $N = 6$ matrica glasi

$$M = \begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 1 & -2 & 1 \\ 0 & 0 & 0 & 1 & -2 \end{bmatrix}.$$

Posmatrajmo homogeni sistem $M\mathbf{x} = 0$, stavimo $x_l = \max\limits_{1 \le n \le N-1} |x_n|$ i dopustimo da je $|x_l| > 0$. Ne može biti $l = 1$, budući da prva jednačina sistema glasi $-2x_1 + x_2 = 0$ ili $-(2 + p_1 h^2)x_1 + x_2 = 0$. Na isti način, ne može biti $l = N - 1$. Prema tome, l-ta jednačina ima oblik $x_{l-1} - (2 + p_l h^2)x_l + x_{l+1} = 0$. Napisana relacija je održiva samo ako je $p_l = 0$ i $x_{l-1} = x_l = x_{l+1}$. Zatim gledamo $(l-1)$–vu jednačinu, itd. (zatim gledamo $(l+1)$–vu jednačinu, itd). Tako dobijamo $x_1 = \ldots = x_{N-1}$. Dobili smo kontradikciju jer smo dopustili da je $|x_l| > 0$. Ne može biti $|x_l| > 0$, mora biti $|x_l| = 0$. Znači $|x_n| = 0$ za svako $n = 1, \ldots, N-1$. Znači da homogeni sistem $M\mathbf{x} = 0$ ima samo trivijalno (nulto) rješenje. Mi smo pokazali da je matrica M regularna.

Zato sistem (5)–(6) ima jedinstveno rješenje. Prvo pitanje je završeno.

Mi raspolažemo sa

$$\ell(R_n) = -r_n - \delta_n, \quad 1 \le n \le N-1. \tag{7}$$

Bilo bi bolje da raspolažemo sa $R_n = \ldots$. Kao da bi se na relaciju $\ell(R_n) = -r_n - \delta_n$ primijenio operator ℓ^{-1}. O tome je ustvari riječ u nastavku.

Mi raspolažemo i sa

$$|r_n| \le \frac{1}{12} M_4 h^2, \quad 1 \le n \le N-1. \tag{8}$$

Dokažimo dvije leme.

Lema 1. Neka je $p_n \ge 0$ za $1 \le n \le N-1$. Neka je $\ell(z_n) = -\frac{1}{h^2}\left(z_{n+1} - 2z_n + z_{n-1}\right) + p_n z_n$. Neka je $\{z_n\}_{n=0}^N$ ma kakav konačan niz brojeva. Ako je $\ell(z_n) \ge 0$ za $1 \le n \le N-1$ i $z_0 \ge 0$, $z_N \ge 0$ onda je $z_n \ge 0$ za $1 \le n \le N-1$.

Dokaz. Uvedimo oznaku $d = \min\limits_{0 \le n \le N} z_n$ i dopustimo da je $d < 0$. Za koje n važi $z_n = d$? Ne može biti $z_0 = d$, niti $z_N = d$. Neka je q najmanji cio broj za koji je $z_q = d$. Imamo $z_{q-1} > d$ i $z_{q+1} \ge d$. Tako da je $-(z_{q+1} - 2z_q + z_{q-1}) < 0$. Još, $p_q \ge 0$ i $z_q = d < 0 \Rightarrow p_q z_q \le 0$. Sabiranjem

$$\ell(z_q) = -\frac{1}{h^2}\left(z_{q+1} - 2z_q + z_{q-1}\right) + p_q z_q < 0.$$

Po uslovu leme je $\ell(z_q) \ge 0$, tako da smo dobili kontradikciju. Dakle, ne može biti $d < 0$, nego mora biti $d \ge 0$. Lema je dokazana.

Lema 2. Neka je $p_n \ge 0$ za $1 \le n \le N-1$. Neka je $\ell(z_n) = -\frac{1}{h^2}\left(z_{n+1} - 2z_n + z_{n-1}\right) + p_n z_n$. Neka je $\{z_n\}_{n=0}^N$ ma kakav konačan niz brojeva. Važi nejednakost

$$\max_{0 \le n \le N} |z_n| \le \max(|z_0|, |z_N|) + \frac{1}{8} X^2 Z, \quad \text{gdje je} \quad Z = \max_{0 < n < N} |\ell(z_n)|.$$

Dokaz. Uvedimo u razmatranje niz brojeva

$$\omega_n = |z_0| \frac{X - nh}{X} + |z_N| \frac{nh}{X} + \frac{1}{2} Z(X - nh)nh.$$

Iz eksplicitnog izraza za ω_n je $\omega_n \geq 0$. Neposrednim računom nalazimo da je $-\frac{1}{h^2}\Big(\omega_{n+1} - 2\omega_n + \omega_{n-1}\Big) = Z$, tako da je $\ell(\omega_n) = Z + p_n\omega_n \geq Z$. Dalje, za $1 \leq n \leq N-1$ imamo $\ell(\omega_n \pm z_n) = \ell(\omega_n) \pm \ell(z_n) \geq Z \pm \ell(z_n) \geq 0$. Pored toga, $\omega_0 \pm z_0 = |z_0| \pm z_0 \geq 0$ i $\omega_N \pm z_N = |z_N| \pm z_N \geq 0$. Prema tome, niz brojeva $\{\omega_n \pm z_n\}_{n=0}^N$ zadovoljava sve uslove prethodne leme. Zato imamo $\omega_n \pm z_n \geq 0$ za $0 \leq n \leq N$. Napišimo odvojeno: $\omega_n + z_n \geq 0$ i $\omega_n - z_n \geq 0$. Znači da je $|z_n| \leq \omega_n$. Slijedi da je $|z_n| \leq \max\limits_{0\leq n\leq N} \omega_n$.

Ostaje samo da se izračuna $\max\limits_{0\leq n\leq N} \omega_n$. Imamo:

$$|z_0|\frac{X-nh}{X} + |z_N|\frac{nh}{X} \leq \max(|z_0|,|z_N|)\frac{X-nh}{X} + \max(|z_0|,|z_N|)\frac{nh}{X} = \max(|z_0|,|z_N|),$$

$$\frac{1}{2}Z(X-nh)nh \leq \frac{1}{2}Z\frac{1}{4}X^2 \quad \text{jer je} \quad x(1-x) \leq \frac{1}{4} \quad \text{za} \quad x \in [0,1],$$

$$\text{sabiranjem,} \quad \omega_n \leq \max(|z_0|,|z_N|) + \frac{1}{2}Z\frac{1}{4}X^2.$$

Znači da je $\max\limits_{0\leq n\leq N} \omega_n \leq \max(|z_0|,|z_N|) + \frac{1}{2}Z\frac{1}{4}X^2$. Lema je dokazana jer $(\forall n)\ |z_n| \leq x \Rightarrow \max\limits_{0\leq n\leq N} |z_n| \leq x$.

Lema 2 govori sljedeće. Neka $\{z_n\}_{n=0}^N$ predstavlja rješenje diferencnog graničnog zadatka $\ell(z_n) = f_n$ za $0 < n < N$, $z_0 = a$, $z_N = b$. Tada je ispunjen tzv. uslov stabilnosti, tj. tada važi nejednakost $\|\mathbf{z}\| \leq C_1\|\mathbf{f}\| + C_2\|\mathbf{g}\|$, gdje je $\mathbf{z} = (z_0, z_1, \ldots, z_N)$, $\mathbf{f} = (f_1, \ldots, f_{N-1})$ i $\mathbf{g} = (a, b)$. Imamo u vidu max–normu.

Slijedi završni dio dokaza teoreme. Primijenimo lemu 2 na niz brojeva $\{R_n\}_{n=0}^N$:

$$\max_{0\leq n\leq N} |R_n| \leq \max(|R_0|,|R_N|) + \frac{1}{8}X^2 \max_{0<n<N} |\ell(R_n)| = \qquad \text{po (7)}$$

$$\max(|R_0|,|R_N|) + \frac{1}{8}X^2 \max_{0<n<N} |-r_n - \delta_n| \leq$$

$$\max(|R_0|,|R_N|) + \frac{1}{8}X^2 \max_{0<n<N} |r_n| + \frac{1}{8}X^2 \max_{0<n<N} |\delta_n| \leq \qquad \text{po (8)}$$

$$\max(|R_0|,|R_N|) + \frac{1}{8}X^2\frac{1}{12}M_4h^2 + \frac{1}{8}X^2 \max_{0<n<N} |\delta_n|.$$

Teorema je dokazana.

U nastavku – dopuna.

Ako se ukine uslov $p(x) \geq 0$ onda posebno treba pretpostaviti da zadatak (1)–(2) ima jedinstveno rješenje. Osim toga, u iskazu teoreme treba dodati: za dovoljno male $h > 0$.

3.3. NEŠTO OPŠTIJI GRANIČNI ZADATAK

Kada se nešto promijeni u postavci graničnog problema, onda dolazi do manjih modifikacija u numeričkoj metodi. Pogledajmo prvo samu diferencijalnu jednačinu.

Ako diferencijalna jednačina ima oblik $y'' + p(x)y' + q(x)y = f(x)$ onda u sistemu linearnih jednačina imamo $\frac{1}{h^2}\Big(y_{n+1} - 2y_n + y_{n-1}\Big) + p_n\frac{1}{2h}\Big(y_{n+1} - y_{n-1}\Big) + q_ny_n = f_n$ za $1 \leq n \leq N-1$. Naime, poznate su formule $\lim_{h\to 0}\frac{y(x+h)-2y(x)+y(x-h)}{h^2} = y''(x)$, $\lim_{h\to 0}\frac{y(x+h)-y(x-h)}{2h} = y'(x)$.

Kao što je rađeno, ako u postavci zadatka figurišu uslovi (2) $y(0) = a$, $y(X) = b$ onda takvim uslovima odgovaraju jednačine (4) $y_0 = a$, $y_N = b$.

Ako granični uslovi glase $y'(0) = a$, $y'(X) = b$ onda na njihov račun formiramo diferencne uslove $\frac{1}{h}(y_1 - y_0) = a$, $\frac{1}{h}(y_N - y_{N-1}) = b$. Obrazloženje: poznate su formule $\lim_{h\to 0} \frac{y(x+h)-y(x)}{h} = y'(x)$, $\lim_{h\to 0} \frac{y(x)-y(x-h)}{h} = y'(x)$.

Slično, uzmimo da granični uslovi imaju oblik $\alpha_0 y(0) + \alpha_1 y'(0) = a$, $\beta_0 y(X) + \beta_1 y'(X) = b$. Tada, u numeričkoj metodi, njima odgovaraju uslovi $\alpha_0 y_0 + \alpha_1 \frac{y_1 - y_0}{h} = a$, $\beta_0 y_N + \beta_1 \frac{y_N - y_{N-1}}{h} = b$. Tako, prešli smo sva tri moguća tipa graničnih uslova.

Ako je osnovni interval $[A, B]$ (umjesto $[0, X]$) onda se oznake prilagođavaju.

Zaključak. Uvijek, iz d. j. se dobija $N-1$ diferencni uslov, a iz graničnih uslova se dobijaju dva diferencna uslova. Uvijek, u numeričkoj metodi, ukupan broj nepoznatih (nepoznate su $y_0, \ldots, y_N$) poklopiće se sa ukupnim brojem diferencnih uslova (to je $N + 1$). Kompjuter će riješiti sistem diferencnih uslova (linearnih jednačina).

4. METODA KONAČNIH ELEMENATA ZA RJEŠAVANJE GRANIČNOG ZADATKA ZA OBIČNE DIFERENCIJALNE JEDNAČINE

Biće izložena Ritzova metoda u opštem slučaju (A: $H \to H$) i na jednom konkretnom primjeru ($Ay = -y'' + p(x)y$), č. Ricova metoda.

4.1. PRIPREMA IZ FUNKCIONALNE ANALIZE

Definicija 1. Linearni operator A koji djeluje u realnom Hilbertovom prostoru H naziva se simetričnim ako je: 1) njegova oblast definisanosti $D(A)$ (svuda) gust skup u H i 2) $\langle Au, v\rangle = \langle u, Av\rangle$ za svako $u, v \in D(A)$. Definicija 2. Simetrični operator A naziva se pozitivnim ako za svako $u \in D(A)$ važi $\langle Au, u\rangle \geq 0$, s tim da znak jednakosti važi samo kada je $u = 0$.

Teorema 1. Ako je A pozitivan operator onda jednačina

$$Au = f \tag{1}$$

(f je dato, u je nepoznata) ima najviše jedno rješenje. Dokaz. Ako bi postojala dva rješenja u_1 i u_2 onda bismo pisali $Av = 0$, gdje je $v = u_2 - u_1$. Odavde je $\langle Av, v\rangle = \langle 0, v\rangle$, tj. $\langle Av, v\rangle = 0$, što treba uporediti sa: $v \neq 0 \Rightarrow \langle Av, v\rangle > 0$. Dakle, $v = 0$, čime je dokaz završen.

Teorema 2. Vektor u_0 je rješenje jednačine (1) (očito se rješenje traži u skupu $D(A)$) ako i samo ako se za $u = u_0$ realizuje minimum funkcionala

$$J(u) = \langle Au, u\rangle - 2\langle f, u\rangle \tag{2}$$

(očito je da se i taj minimum traži za $u \in D(A)$).

Dokaz. Neka je u_0 rješenje jednačine (1). Neka je v proizvoljni element iz $D(A)$. Vektor $\eta = v - u_0$ takođe pripada $D(A)$ jer je $D(A)$ linearan skup. Imamo:

$$J(v) = \langle Av, v\rangle - 2\langle f, v\rangle = \langle A(u_0 + \eta), u_0 + \eta\rangle - 2\langle f, u_0 + \eta\rangle =$$

(A je simetričan pa je $\langle A\eta, u_0\rangle = \langle \eta, Au_0\rangle$, skalarni proizvod je simetričan)

$$J(u_0) + 2\langle Au_0 - f, \eta\rangle + \langle A\eta, \eta\rangle \quad \Rightarrow$$

(pomoću $Au_0 - f = 0$ i pomoću $\langle A\eta, \eta\rangle > 0$ ako $\eta \neq 0$ jer je A pozitivan)

$$J(u) > J(u_0).$$

Dakle, u_0 ostvaruje minimum funkcionala $J(u)$ na $D(A)$.

U drugom smjeru. Neka je u_0 rješenje zadatka "$J \to \min$". Tada za svako $\eta \in D(A)$ i $t \in R$ važi

$$J(u_0 + t\eta) \geq J(u_0) \quad \text{ili} \quad \langle A(u_0 + t\eta), u_0 + t\eta\rangle - 2\langle f, u_0 + t\eta\rangle \geq \langle Au_0, u_0\rangle - 2\langle f, u_0\rangle \quad \Rightarrow$$

(jer je A simetričan)

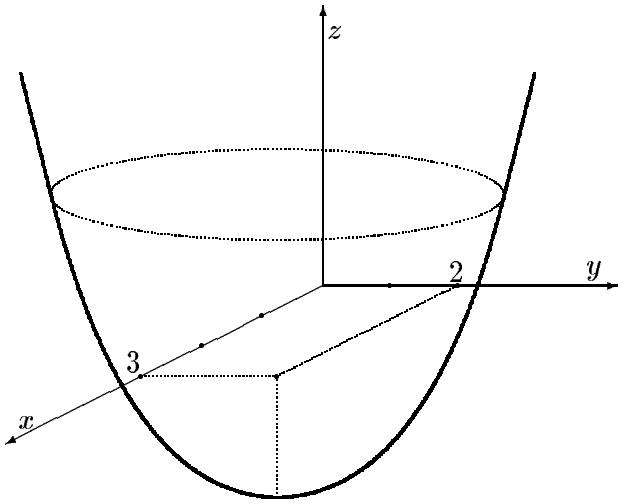
$$2t\langle Au_0 - f, \eta\rangle + t^2\langle A\eta, \eta\rangle \geq 0.$$

Na lijevoj strani je napisan kvadratni izraz po t koji je za svako realno t nenegativan pa je zato njegova diskriminanta ≤ 0. Dakle, $\langle Au_0 - f, \eta\rangle^2 \leq 0 \Rightarrow \langle Au_0 - f, \eta\rangle = 0$, tj. $Au_0 - f \perp \eta$.

Posljednje važi za svako $\eta \in D(A)$, $D(A)$ je gust skup pa je konačno $Au_0 - f = 0$, čime je dokaz završen.

Funkcional $J(u)$ naziva se energetskim funkcionalom zadatka (1).

Primjer: $H = R^2$, $A = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$, $\mathbf{u} = \begin{bmatrix} x \\ y \end{bmatrix}$, $\mathbf{f} = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$, $\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$, $J(\mathbf{u}) = x^2 + 2y^2 - 6x - 8y$. V. sliku 1. Na slici je prikazan grafik funkcije $z(x,y) = x^2 + 2y^2 - 6x - 8y$. Funkcija dostiže minimum za $(x,y) = (3,2)$. Također, $(x,y) = (3,2)$ je rješenje jednačine $A\mathbf{u} = \mathbf{f}$. Zapaziti da je A pozitivan.



Slika 1: $z = x^2 + 2y^2 - 6x - 8y$

4.2. RITZOVA METODA

Ideja o metodi varijacionog tipa i pojam Ritzove metode

U skupu $D(A)$ uvodi se tzv. energetski skalarni proizvod po formuli $\langle u, v\rangle_A = \langle Au, v\rangle$. Pokazuje se da su sve aksiome skalarnog proizvoda ispunjene jer je A pozitivan. Odgovarajuća norma je $\|u\|_A = \sqrt{\langle u, u\rangle_A}$. Oko oznaka: $\|\ \|$ – norma u H, $\|\ \|_A$ – energetska. Slično za skalarni proizvod.

Sada se izabere sistem vektora $\{\varphi_1, \varphi_2, \ldots\}$ (tzv. koordinatni vektori) koji ispunjavaju sljedeća tri uslova: 1) svaki φ_n pripada $D(A)$, 2) vektori $\varphi_1, \ldots, \varphi_n$ su linearno nezavisni za svako n i 3) sistem $\{\varphi_n\}_{n=1}^{\infty}$ je kompletan u smislu energetske norme, tj. za svako $u \in D(A)$ i svako $\varepsilon > 0$ postoji (konačna) linearna kombinacija vektora tog sistema $\{\varphi_n\}_{n=1}^{\infty}$, označimo je sa $v = \sum_{i=1}^{\ell} c_i v_i$, takva da je $\|u - v\|_A < \varepsilon$.

Sada se opredjeljujemo za određeno n (jasno je da kasnije n može da uzme i neku drugu vrijednost). Time smo se ograničili na vektore $\varphi_1, \ldots, \varphi_n$.

Nas interesuje vektor koji minimizuje funkcional $J(u)$. Neka nam kao njegova aproksimacija služi neki vektor koji je linearna kombinacija vektora $\varphi_1, \ldots, \varphi_n$ na koje smo se ograničili. Tj. ograničili smo se na potprostor čiju bazu čini prvih n koordinatnih vektora.

Neka $u^* \in D(A)$ označava rješenje zadatka "$\min J(u)$", samim tim i jednačine (1). Neka u_n označava rješenje zadatka "$\min J(u)$, u se traži u spomenutom potprostoru". Drugim riječima, u^* je tačno rješenje za "$J \to \min$", dok je u_n samo približno rješenje za taj zadatak. Može se pokazati da $\|u^* - u_n\|_A \to 0$ kad $n \to \infty$ (pod pretpostavkom da se sva računanja izvode tačno). Iz ovoga slijedi da je i $\lim_{n\to\infty} \|u^* - u_n\| = 0$. Kako saznati u_n?

Dakle, n smo fiksirali i tražimo vektor $c_1\varphi_1+\ldots+c_n\varphi_n$ (numerički odgovor), odnosno tražimo $c_1,\ldots,c_n$, da se postigne najmanja moguća vrijednost za $J(c_1\varphi_1+\ldots+c_n\varphi_n)$. Uvedimo oznaku $F(c_1,\ldots,c_n)=J(c_1\varphi_1+\ldots+c_n\varphi_n)$. Neposrednim računom nalazimo da je:

$$F(c_1,\ldots,c_n)=c_1^2\langle A\varphi_1,\varphi_1\rangle+c_1c_2\langle A\varphi_1,\varphi_2\rangle+\ldots+c_1c_n\langle A\varphi_1,\varphi_n\rangle+\ldots\ldots\ldots+$$

$$c_nc_1\langle A\varphi_n,\varphi_1\rangle+c_nc_2\langle A\varphi_n,\varphi_2\rangle+\ldots+c_n^2\langle A\varphi_n,\varphi_n\rangle-2c_1\langle f,\varphi_1\rangle-2c_2\langle f,\varphi_2\rangle-\ldots-2c_n\langle f,\varphi_n\rangle.$$

Određivanje brojeva c_k svodi se na rješavanje sistema $\dfrac{\partial F}{\partial c_1}=0,\quad\ldots,\quad\dfrac{\partial F}{\partial c_n}=0$ (stacionarna tačka za F), ovo je tzv. **Ritzov sistem**. Neposrednim računom vidimo da je taj sistem linearan po nepoznatim c_k, a u razvijenom obliku Ritzov sistem glasi

$$\begin{cases}\langle A\varphi_1,\varphi_1\rangle c_1+\langle A\varphi_1,\varphi_2\rangle c_2+\ldots+\langle A\varphi_1,\varphi_n\rangle c_n=\langle f,\varphi_1\rangle\\ \ldots\qquad\ldots\qquad\ldots\\ \langle A\varphi_n,\varphi_1\rangle c_1+\langle A\varphi_n,\varphi_2\rangle c_2+\ldots+\langle A\varphi_n,\varphi_n\rangle c_n=\langle f,\varphi_n\rangle\end{cases}\tag{3}$$

Matrica ovog sistema je $M=[\langle A\varphi_i,\varphi_j\rangle]_{i,j=1}^n=[\langle\varphi_i,\varphi_j\rangle_A]_{i,j=1}^n$ pa je njena determinanta $\neq 0$, jer je to Gramova matrica sistema $\varphi_1,\ldots,\varphi_n$ međusobno linearno nezavisnih vektora.

Pojedini element matrice sistema jednak je $m_{ij}=\langle A\varphi_i,\varphi_j\rangle=\langle\sqrt{A}\varphi_i,\sqrt{A}\varphi_j\rangle$. Imamo u vidu da je operator A pozitivan.

Šablon Ritzove metode

Dosad izložena metoda zove se Ritzovom.

- Prvi korak: zadatak o rješavanju jednačine (1) (obično je to neki granični zadatak) svodi se na zadatak o minimumu funkcionala (2).
- Drugi korak: treba se opredijeliti za jedan konkretan sistem $\{\varphi_n\}_{n=1}^\infty$ koji je kompletan u smislu energetske norme.
- Treći korak: opredjeljujemo se za jedno konkretno n.
- I četvrti korak: efektivno računanje – rješavanje sistema linearnih jednačina (3).

Ilustrujemo Ritzovu metodu na jednom konkretnom primjeru

Neka se razmatra granični zadatak

$$-y''+p(x)y(x)=f(x),\quad 0\le x\le 1,\quad y(0)=0,\quad y(1)=0,\quad p\text{ i }f\text{ neprekidne},\quad p(x)\ge 0.\tag{4}$$

Neka je $H=L^2(0,1)$ (Lebesgue). Neka je $D(A)=\{y\colon y''\in L^2(0,1),\,y(0)=0,\,y(1)=0\}$ i $Ay=-y''+p(x)y$. Poznato je da zadatak (4) ima jedinstveno rješenje. Poznato je da je operator A simetričan i pozitivan. Znamo da se u prostoru $L^2(0,1)$ skalarni proizvod definiše kao $\langle u,v\rangle=\int_0^1 u(x)v(x)dx$.

Po teoremi 2, postavljeni granični problem ekvivalentan je sljedećem varijacionom zadatku: naći $y\in D(A)$ za koje se dostiže minimum funkcionala

$$I(y)=\int_0^1\Big[(-y''+p(x)y)y-2fy\Big]\,dx.$$

Pomoću jednostavne transformacije sa parcijalnom integracijom i uzimajući u obzir granične uslove:

$$\int_0^1 \Big[-y''(x)y(x)\Big]dx = -\int_0^1 y(x)dy'(x) = -y(x)y'(x)\Big|_{x=0}^{x=1} + \int_0^1 y'(x)dy(x) =$$

$$0 + \int_0^1 (y'(x))^2 dx \qquad \Rightarrow$$

$$I(y) = \int_0^1 \Big[(y'(x))^2 + p(x)y^2(x) - 2f(x)y(x)\Big]dx. \tag{5}$$

Time je završen prvi korak. Prelazimo na drugi korak. Energetska norma u ovom slučaju glasi $\|u\|_A = \sqrt{\int_0^1 [(u'(x))^2 + p(x)u^2(x)]dx}$.

(Zapaziti da u izrazu za energetsku normu $\|u\|_A$ sada figuriše samo prvi izvod funkcije $u(x)$, a više ne figuriše drugi izvod. Traži se manja glatkost, odnosno oblast definisanosti se ustvari proširuje.)

Postavlja se pitanje: koji su primjeri kompletnih sistema po ovoj normi?

- Prvi odgovor: trigonometrijski sistem, $\varphi_1(x) = \sin \pi x$, $\varphi_2(x) = \sin 2\pi x$, $\varphi_3(x) = \sin 3\pi x$, ...
- Drugi odgovor: polinomski sistem, $\varphi_1(x) = x(1-x)$, $\varphi_2(x) = x^2(1-x)$, $\varphi_3(x) = x^3(1-x)$, ... Ponovimo da će približno rješenje imati oblik $v(x) = \sum_{i=1}^n c_i\varphi_i(x)$.
- Treći odgovor: sistem $\{\varphi_{n,k}(x),\ k = 0, \ldots, n,\ n = 1, 2, 3, \ldots\}$, v. u nastavku.

Ritzova metoda u formi metode konačnih elemenata na konkretnom primjeru

Ako bismo se opredijelili za prvi ili drugi odgovor onda bi matrica sistema (3) bila "puna", tj. skoro svi njeni elementi bili bi različiti od nule.

Ponekad se, kod varijacionih metoda, uzima da $\varphi_1, \ldots, \varphi_n$ zavise od konkretnog n pa se detaljnije piše $\varphi_{n,1}, \ldots, \varphi_{n,n}$.

Treći odgovor: za određeno n imamo $\varphi_{n,0}, \varphi_{n,1}, \ldots, \varphi_{n,n}$, tj. sada je došlo do male modifikacije oznaka. Funkcija $\varphi_{n,k}(x)$ ima svojstvo da je $\neq 0$ samo za $x \in [x_{k-1}, x_{k+1}]$, gdje je $x_j = jh$ i $h = 1/n$ pa se kaže da je skup $[x_{k-1}, x_{k+1}]$ nosač funkcije $\varphi_{n,k}$. Grafik funkcije $\varphi_{n,k}$ sastoji se praktično od dvije (ili jedne) kose linije.

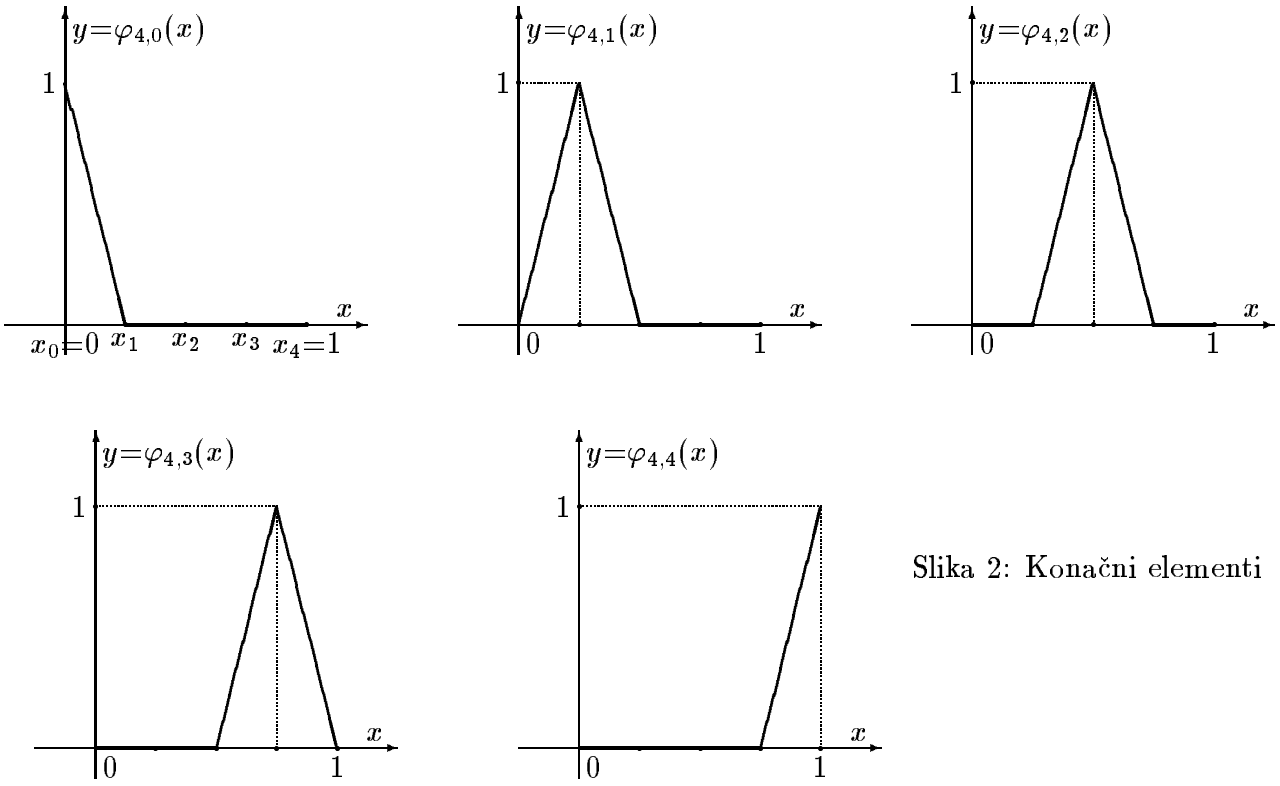
Na slici 2 prikazani su grafici konačnih elemenata $\varphi_{n,k}(x)$, $k = 0, \ldots, n$ kada je $n = 4$. Linearna kombinacija prikazanih konačnih elemenata je izlomljena linija. Traži se linearna kombinacija $\sum_{k=0}^n c_k\varphi_{n,k}(x)$ za koju se postiže minimum funkcionala $I = I(y)$.

Formule u opštem slučaju (ovdje je $nh = 1$, kao i $x_k = kh$):

$$\varphi_{n,k}(x) = \begin{cases} (x - x_{k-1})/h & \text{za} \quad x_{k-1} \le x \le x_k \\ (x_{k+1} - x)/h & \text{za} \quad x_k \le x \le x_{k+1} \\ 0 & \text{inače} \end{cases} \qquad (k = 1, \ldots, n-1)$$

$$\varphi_{n,0}(x) = \begin{cases} (x_1 - x)/h & \text{za} \quad 0 \le x \le x_1 \\ 0 & \text{inače} \end{cases} \qquad \varphi_{n,n}(x) = \begin{cases} (x - x_{n-1})/h & \text{za} \quad x_{n-1} \le x \le 1 \\ 0 & \text{inače} \end{cases}$$

Može se pokazati da je razmatrani sistem funkcija kompletan u smislu energetske norme. Misli se na sistem $\varphi_{n,k}(x)$, $k = 0, \ldots, n$, $n = 1, 2, 3, \ldots$. Posmatrajmo sistem $\{\varphi_{n,k}(x)\}_{k=0}^{n}$. Kakva svojstva aproksimacije u smislu energetske norme ima posmatrani sistem? Kako n raste, to su ta svojstva sve bolja i bolja, tj. postižu se proizvoljno dobre aproksimacije. Mi se opredjeljujemo za treći odgovor.



Slika 2: Konačni elementi

Što se tiče sistema (3), u ovom primjeru njegova matrica će biti trodijagonalna, jer ako dvije koordinatne funkcije imaju disjunktne nosače onda je odgovarajući element matrice jednak nuli.

`Uopšte, ako varijaciona metoda ima svojstvo da pojedini` φ_k `tj.` $\varphi_{n,k}$ `ima "mali" nosač onda se (bez obzira da li se radi o Ritzovoj ili nekoj drugoj metodi) kaže da je to metoda konačnih elemenata.`

Produžetak prethodnog podnaslova – završni dio: efektivno računanje

Približno rješenje tražimo u obliku $v(x) = c_0\varphi_{n,0}(x) + c_1\varphi_{n,1}(x) + \ldots + c_n\varphi_{n,n}(x)$. Detaljnije zapisivanje: umjesto c_k pisali bismo $c_{n,k}$. Da bi funkcija $v(x)$ zadovoljavala dva granična uslova iz (4) mora da bude $c_0 = c_n = 0$, a to je i dovoljno. Dakle, $v(x) = c_1\varphi_{n,1}(x) + \ldots + c_{n-1}\varphi_{n,n-1}(x)$. Sistem (3) je u ovom slučaju veličine $(n-1) \times (n-1)$. Izvedimo sada eksplicitni izraz za taj sistem. Neka je sistem označen kao $M\mathbf{c} = \mathbf{b}$, gdje je $M = [m_{ij}]$. Već smo rekli da je to jedan trodijagonalni sistem, tako da će se lako doći do njegovog rješenja $\mathbf{c} = (c_1, \ldots, c_{n-1})$.

Jasno, u matričnom obliku sistem glasi

$$\begin{bmatrix} m_{11} & \cdots & m_{1,n-1} \\ \vdots & \ddots & \vdots \\ m_{n-1,1} & \cdots & m_{n-1,n-1} \end{bmatrix} \cdot \begin{bmatrix} c_1 \\ \vdots \\ c_{n-1} \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_{n-1} \end{bmatrix}.$$

Iskoristimo poznate relacije: $m_{ij} = \langle A\varphi_{n,i}, \varphi_{n,j}\rangle = \langle \sqrt{A}\varphi_{n,i}, \sqrt{A}\varphi_{n,j}\rangle$ i $b_i = \langle f, \varphi_{n,i}\rangle$, tj. $m_{ij} = \int_0^1 \big[\varphi'_{n,i}(x)\varphi'_{n,j}(x) + p(x)\varphi_{n,i}(x)\varphi_{n,j}(x)\big]dx$ i $b_i = \int_0^1 f(x)\varphi_{n,i}(x)dx$. Dobijamo:

$$m_{k,k-1} = \int_0^1 \big[\varphi'_{k-1}\varphi'_k + p\varphi_{k-1}\varphi_k\big]dx, \quad m_{kk} = \int_0^1 \big[(\varphi'_k)^2 + p\varphi_k^2\big]dx,$$

$$m_{k,k+1} = \int_0^1 \big[\varphi'_k\varphi'_{k+1} + p\varphi_k\varphi_{k+1}\big]dx, \quad \text{a inače je } m_{ij} = 0, \quad b_k = \int_0^1 f(x)\varphi_k(x)dx.$$

Dalje:

$$m_{kk} = \int_{x_{k-1}}^{x_k} \big[1 + p(x)(x - x_{k-1})^2\big]dx\big/h^2 + \int_{x_k}^{x_{k+1}} \big[1 + p(x)(x_{k+1} - x)^2\big]dx\big/h^2, \quad k = 1, \ldots, n-1,$$

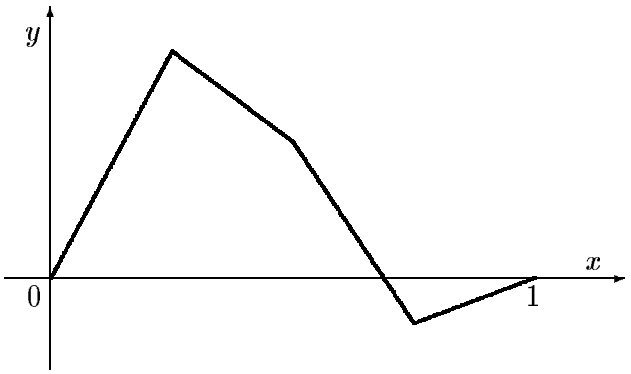
$$m_{k,k+1} = \int_{x_k}^{x_{k+1}} \big[-1 + p(x)(x_{k+1} - x)(x - x_k)\big]dx\big/h^2, \quad k = 1, \ldots, n-2,$$

$$m_{k+1,k} = m_{k,k+1}, \quad k = 1, \ldots, n-2,$$

$$b_k = \int_{x_{k-1}}^{x_k} f(x)(x - x_{k-1})dx\big/h + \int_{x_k}^{x_{k+1}} f(x)(x_{k+1} - x)dx\big/h, \quad k = 1, \ldots, n-1.$$

Algoritam za (4): formiraj M i $\mathbf{b}$, riješi $M\mathbf{c} = \mathbf{b}$, napiši $v(x) = \sum_k c_k\varphi_{n,k}(x)$. V. sliku 3.

Greška $y(x) - v(x)$ ove konkretne metode? Može se pokazati da greška ima red veličine h^2 u tački na x–osi koja je čvor, a ima red veličine h u bilo kojoj tački sa x–ose (između 0 i 1).



Slika 3: Približno rješenje
$v(x) = \sum_{i=1}^{n-1} c_i\varphi_{n,i}(x)$
kada je data d. j. $-y'' + p(x)y = f(x)$
sa homogenim g. u. $y(0) = 0, y(1) = 0$

Slučaj nehomogenih graničnih uslova

Razmotrimo granični zadatak

$$-y'' + p(x)y = f(x), \quad y(0) = a, \quad y(1) = b,$$

gdje je i dalje $p(x) \geq 0$. Može se pokazati da razmatrani zadatak ima jedinstveno rješenje $y = y(x)$. Takođe, može se pokazati da se to rješenje poklapa sa funkcijom $y = y(x)$ na kojoj se dostiže minimum funkcionala $I = I(y)$, pri čemu je $I(y) = \int_0^1 [(y'(x))^2 + p(x)y^2(x) - 2f(x)y(x)]dx$. Vidimo da je ostao isti funkcional. Dakle, kao i u dosad razmatranom slučaju nultih (homogenih) graničnih uslova $y(0) = 0, y(1) = 0$, rješavanje graničnog problema ekvivalentno je rješavanju problema o minimumu funkcionala. Kako da nađemo numeričko rješenje graničnog problema sa nehomogenim graničnim uslovima $y(0) = a$, $y(1) = b$? Postoje male razlike u odnosu na ono što je rečeno u prethodnom tekstu.

Razmotrimo funkciju $\varphi_0(x) = a + (b-a)x$. Razmotrimo funkcije $\varphi_n(x) = x^n(1-x)$ za $n = 1, 2, 3, \ldots$. Fiksirajmo jedan prirodan broj n. To znači da smo se ograničili na funkcije $\varphi_0, \varphi_1, \ldots, \varphi_n$. Napišimo linearnu kombinaciju $v(x) = \varphi_0(x) + \sum_{i=1}^{n} c_i\varphi_i(x)$, gdje $c_i \in R$ za $i = 1, \ldots, n$. Dakle, u slučaju primjene Ritzove metode treba odrediti $c_1, \ldots, c_n$. Drugim riječima, treba izabrati jednu funkciju oblika $v = v(x)$ iz klase funkcija na koju smo se ograničili. Naravno, traži se ona funkcija na kojoj se dostiže najmanja moguća vrijednost funkcionala $I = I(y)$ u toj klasi.

Ako se primjenjuje Ritzova metoda u formi metode konačnih elemenata onda se posmatraju funkcije $v(x) = a\varphi_{n,0}(x) + \sum_{i=1}^{n-1} c_i\varphi_{n,i}(x) + b\varphi_{n,n}(x)$. Naravno, treba izabrati $c_1, \ldots, c_{n-1}$ tako da se minimizuje vrijednost funkcionala $I = I(y)$.

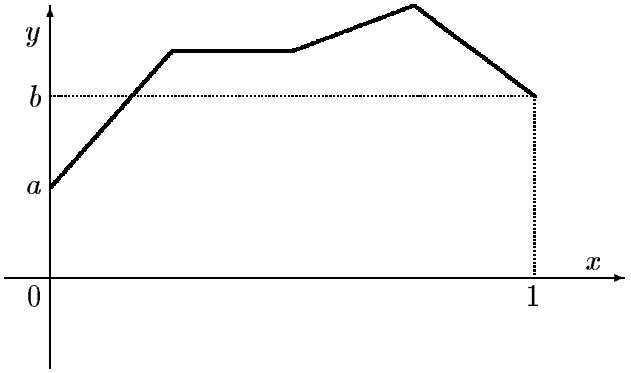
Za vježbu, formirajte Ritzov sistem (ranije je bio označen kao $M\mathbf{c} = \mathbf{b}$). Sada je to jedan sistem od $n-1$ linearnih jednačina sa $n-1$ nepoznatih $c_1, \ldots, c_{n-1}$. Kada se sistem riješi, onda i saopštavamo numerički rezultat $v(x) = a\varphi_{n,0}(x) + \sum_{i=1}^{n-1} c_i\varphi_{n,i}(x) + b\varphi_{n,n}(x)$. V. sliku 4.

Prilikom računanja: parcijalna integracija:

$$\int_0^1 [(-y''(x))z(x)]dx = -\int_0^1 z(x)dy'(x) =$$

$$(-y'(x)z(x))\Big|_{x=0}^{x=1} + \int_0^1 y'(x)dz(x) = -y'(1)z(1) + y'(0)z(0) + \int_0^1 y'(x)z'(x)dx,$$

gdje npr. može da bude $y(x) = \varphi_{n,i}(x)$, $z(x) = \varphi_{n,j}(x)$.



Slika 4: Približno rješenje
$v(x) = a\varphi_{n,0}(x) + \sum_{i=1}^{n-1} c_i\varphi_{n,i}(x) + b\varphi_{n,n}(x)$
kada je data d. j. $-y'' + p(x)y = f(x)$
sa nehomogenim g. u. $y(0) = a$, $y(1) = b$

Napomena o slučaju graničnih uslova druge i treće vrste

Razmotrimo granični zadatak

$$-y'' + p(x)y = f(x), \quad \alpha_0 y(0) + \alpha_1 y'(0) = \alpha, \quad \beta_0 y(1) + \beta_1 y'(1) = \beta,$$

gdje je $\alpha_1 \neq 0$, $\beta_1 \neq 0$. Tada, rješavanje postavljenog zadatka svodi se na rješavanje zadatka o određivanju funkcije koja minimizuje jedan određeni funkcional $I = I(y)$. Dakle, prvi zadatak se svodi na drugi zadatak i obrnuto. Drugim riječima, zadaci su ekvivalentni. Funkcional glasi $I(y) = \int_0^1 [(y'(x))^2 + p(x)y^2(x) - 2f(x)y(x)]dx + \frac{1}{\alpha_1}[-\alpha_0 y^2(0) + 2\alpha y(0)] + \frac{1}{\beta_1}[\beta_0 y^2(1) - 2\beta y(1)]$. Ova formula preuzeta je iz knjige o numeričkim metodama čiji su autori Berezin i Židkov.

4.3. PAR REČENICA O VARIJACIONOM RAČUNU

Varijacioni račun predstavlja jednu matematičku teoriju (oblast matematike). Ne ulazeći u sve detalje, pogledajmo kako glasi jedan tipičan problem iz varijacionog računa i skicirajmo kako

se problem rješava. Neka je F funkcija od tri promjenljive i neka je K jedna klasa funkcija. Za funkciju $y = y(x)$ iz K definiše se funkcional $J = J(y)$ relacijom $J(y) = \int_a^b F(x, y(x), y'(x))dx$. Treba riješiti zadatak o minimumu funkcionala, u konciznom zapisu zadatak "$J \to \min$". Drugim riječima, treba naći funkciju $y = y(x)$ koja realizuje minimum.

Datom funkcionalu $J = J(y)$ pridružuje se njegova tzv. **Eulerova jednačina**, č. Ojler. Ona glasi: $F'_y(x, y, y') - \frac{d}{dx}F'_{y'}(x, y, y') = 0$. Kada funkcional stacionira onda je moguća njegova ekstremna vrijednost. Pod određenim pretpostavkama, važi sljedeće tvrđenje: ako $y = y(x)$ minimizuje funkcional onda ta ista funkcija $y = y(x)$ predstavlja rješenje napisane d. j. (Eulerove jednačine), kao i obrnuto. Kratko, dva zadatka su međusobno ekvivalentna.

Dalje, pogledajmo jedan konkretan slučaj funkcionala. Stavimo $J(y) = \int_0^1 [(y'(x))^2 + p(x)y^2(x) - 2f(x)y(x)]dx$. Kako glasi njegova Eulerova jednačina? Ispostavlja se da je to upravo $-y'' + p(x)y = f(x)$. Zaista, imamo redom: $F(x, y, y') = (y')^2 + p(x)y^2 - 2f(x)y$, $F'_y = 2p(x)y - 2f(x)$, $F'_{y'} = 2y'$, $\frac{d}{dx}F'_{y'} = \frac{d}{dx}(2y') = 2y''$, ukupno $F'_y - \frac{d}{dx}F'_{y'} = 2p(x)y - 2f(x) - 2y''$, $F'_y - \frac{d}{dx}F'_{y'} = 0$, $-y'' + p(x)y = f(x)$.

4.4. POJAM O METODI GALERKINA

Razmotrimo granični zadatak sa homogenim graničnim uslovima $-y'' + p(x)y = f(x)$, $y(0) = 0$, $y(1) = 0$. Više se ne traži da je $p(x) \geq 0$. Razmotrimo sistem funkcija $\{\varphi_n(x)\}_{n=1}^{\infty}$ kao gore, npr. $\varphi_n(x) = x^n(1 - x)$. Naravno da funkcije $\varphi_1, \ldots, \varphi_n$ treba da budu nezavisne. Potrebno je da sistem $\{\varphi_n\}_{n=1}^{\infty}$ bude kompletan u razmatranom Hilbertovom prostoru. Na početku algoritma, mi fiksiramo jedan prirodan broj n. Približno rješenje $v = v(x)$ ima oblik $v(x) = \sum_{i=1}^{n} c_i\varphi_i(x)$. Sada se $c_1, \ldots, c_n$ određuju iz uslova $Lv(x) - f(x) \perp \varphi_k(x)$ za $k = 1, \ldots, n$. Drukčije zapisano, $\int_0^1 [-v''(x) + p(x)v(x) - f(x)]\varphi_k(x)dx = 0$, $k = 1, \ldots, n$. Uvedena je oznaka $Ly(x)$ za diferencijalni izraz $Ly = -y'' + p(x)y$.

Jasno, treba izračunati napisane integrale, čime će se dobiti jedan sistem linearnih jednačina po nepoznatim $c_1, \ldots, c_n$. Kada se taj sistem linearnih jednačina riješi, onda ćemo biti u mogućnosti da saopštimo numerički odgovor $v(x)$.

Obrazloženje za predloženi postupak: jedino je nula–vektor ortogonalan na sve elemente jednog kompletnog sistema. Ako je ortogonalan na njih nekoliko, onda je za očekivati da će biti blizak nuli.

Zanimljivo je da se primjenom metode Galerkina na način kako je dosad izloženo dolazi do istog numeričkog odgovora koji daje Ritzova metoda. U nastavku ćemo izložiti drugu varijantu metode Galerkina, nešto opštiju varijantu.

Razmotrimo zadatak sa graničnim vrijednostima koji se sastoji od diferencijalne jednačine $-y'' + p(x)y = f(x)$ i graničnih uslova $y(0) = a$, $y(1) = b$. Vidi se da su sada postavljeni nehomogeni granični uslovi ($a, b \in R$). Uvode se dva sistema funkcija. Neka je $\varphi_0(x) = a + (b - a)x$ i $\varphi_1(x), \varphi_2(x), \ldots$ kao maločas. Drugi sistem funkcija označen je kao $\psi_1(x), \psi_2(x), \ldots$. Stavimo $v(x) = \varphi_0(x) + \sum_{i=1}^{n} c_i\varphi_i(x)$. Po kom kriterijumu se određuju $c_1, \ldots, c_n$? Mi tražimo da bude zadovoljeno n uslova: $Lv(x) - f(x) \perp \psi_k(x)$. Drukčije zapisano, treba da važi relacija $\int_0^1 [-v'' + p(x)v - f(x)]\psi_k(x)dx = 0$ za $k = 1, \ldots, n$.

Ostali elementi postupka numeričke metode su isti kao u prethodnoj varijanti, pa ponovimo ukratko. Treba izračunati napisane integrale, čime će se dobiti jedan sistem linearnih jednačina po nepoznatim $c_1, \ldots, c_n$. Kada se taj sistem linearnih jednačina riješi onda ćemo biti u mogućnosti da saopštimo numerički odgovor $v(x)$.

U zaključku, metoda Galerkina je pogodna u slučaju kada linearni diferencijalni operator nije pozitivan. Ustvari, metoda Galerkina i ne pripada klasi varijacionih metoda.

5. METODA KONAČNIH RAZLIKA ZA PARCIJALNU DIFERENCIJALNU JEDNAČINU ELIPTIČKOG TIPA

5.1. OZNAKE I NUMERIČKI ALGORITAM

Neka je $\Omega = (0,1) \times (0,1)$, $\overline{\Omega} = [0,1] \times [0,1]$ i $\partial\Omega = \overline{\Omega} \setminus \Omega$. Razmotrimo **granični zadatak** koji je sastavljen od Poissonove jednačine

$$Lu = -\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) = f(x,y) \quad \text{za} \quad (x,y) \in \Omega \tag{1}$$

i graničnog uslova

$$u(x,y) = g(x,y) \quad \text{za} \quad (x,y) \in \partial\Omega. \tag{2}$$

Poznato je da taj granični zadatak ima jedinstveno rješenje klase $u(x,y) \in C(\overline{\Omega}) \cap C^2(\Omega)$ ako je $f(x,y) \in C^1(\overline{\Omega})$ i $g(x,y) \in C(\partial\Omega)$.

Izaberimo $n \geq 1$ i stavimo $h = 1/n$. Čvorovi mreže su tačke $(x,y) = (ih, jh)$, gdje je $i,j = 0,1,\ldots,n$. Uvedimo oznake za skup unutrašnjih čvorova, skup svih čvorova i skup graničnih čvorova: neka bude $\omega = \{(ih,jh) \mid 0 < i < n, 0 < j < n\}$, $\overline{\omega} = \{(ih,jh) \mid 0 \leq i \leq n, 0 \leq j \leq n\}$ i $\partial\omega = \overline{\omega} \setminus \omega$. Vidi se da unutrašnjih čvorova ima $(n-1)^2$, svih čvorova ima $(n+1)^2$ i graničnih čvorova ima $4n$. V. sliku 1. Uvedimo sljedeće diskretne norme (tipa C) za funkcije $u = u_{ij}$ koje su definisane na mreži:

$$\|u\|_\omega = \max_{0<i,j<n} |u_{ij}|, \qquad \|u\|_{\overline{\omega}} = \max_{0\leq i,j\leq n} |u_{ij}|, \qquad \|u\|_{\partial\omega} = \max_{(i,j)\in\partial\omega} |u_{ij}|.$$

Svakoj tački mreže pridružuje se jedna jednačina (jedan uslov), tako da će se broj jednačina izjednačiti sa brojem uslova. Što se tiče unutrašnjih tačaka mreže, poznata je formula za aproksimaciju drugog izvoda funkcije $f = f(x)$ od jedne promjenljive: $f''(x) \approx (f(x+h) - 2f(x) + f(x-h))/h^2$ i poznato je da greška te aproksimacije iznosi $L - R = -h^2 f^{IV}(\xi)/12$, gdje $\xi \in (x-h, x+h)$. Što se tiče graničnih tačaka mreže $i,j = 0,n$, vrijednosti u_{ij} date su graničnim uslovom (2). Tako nastaje **diferencni granični zadatak**:

$$\ell(u) = -(u_{i+1,j} - 2u_{ij} + u_{i-1,j} + u_{i,j+1} - 2u_{ij} + u_{i,j-1})/h^2 = f_{ij} \quad \text{za} \quad (i,j) \in \omega, \tag{3}$$

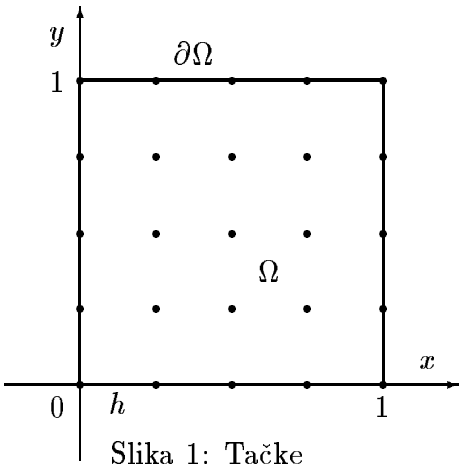
$$u_{ij} = g_{ij} \quad \text{za} \quad (i,j) \in \partial\omega. \tag{4}$$

Jasno, $f_{ij} = f(ih,jh)$ i $g_{ij} = g(ih,jh)$. U jednačini oblika (3) učestvuju vrijednosti funkcije $u = u_{ij}$ u pet čvorova: (i,j) i $(i \pm 1, j \pm 1)$. Tih pet čvorova obrazuju šablon ("krst"). V. sliku 2. Sistem linearnih jednačina (3)–(4) ima $(n+1)^2$ jednačina i isto toliko nepoznatih u_{ij} $(0 \leq i,j \leq n)$. Matrica sistema je trakasta (njeni članovi $\neq 0$ nalaze se svi u jednoj traci oko dijagonale). Ta okolnost koristi se da sistem bude riješen efikasno.

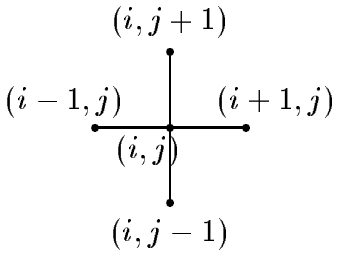
Označimo sa u^t_{ij} vrijednost tačnog rješenja $u = u(x,y)$ graničnog zadatka (1)–(2) u čvoru (i,j), tj. u tački $(x,y) = (ih,jh)$. Isto tako, neka bude $f^t_{ij} = f(ih,jh)$ i $g^t_{ij} = g(ih,jh)$.

Dopustimo da je prisutna određena greška u saznavanju funkcija $f = f(x,y)$ i $g = g(x,y)$. Neka su f^p_{ij} i g^p_{ij} njihove raspoložive približne vrijednosti u čvorovima. Ako je u jednačinama (3) $f_{ij} = f^p_{ij}$ i ako je u jednačinama (4) $g_{ij} = g^p_{ij}$, tada označimo sa $u^p = u^p_{ij}$ rješenje sistema (3)–(4). Znači, $\ell(u^p) = f^p$, $u^p = g^p$.

U nastavku će biti pokazano da sistem (3)–(4) ima jedinstveno rješenje $u^p = u^p_{ij}$ i biće pokazano da je $\lim_{n\to\infty} u^p_{ij} = u^t_{ij}$ (da je $\lim_{h\to 0} u^p_{ij} = u^t_{ij}$).



Slika 1: Tačke



Slika 2: Šablon

5.2. ŠEMA DOKAZA

Kasnije će biti dokazano da za ma koju dovoljno glatku funkciju $u = u(x, y)$ važi relacija (uslov aproksimacije)

$$\lim_{h \to 0} \|Lu - \ell(u)\|_\omega = 0, \tag{5}$$

odnosno relacija

$$\|Lu - \ell(u)\|_\omega \leq Ch^2 \quad (h > 0). \tag{6}$$

Ovdje je C konstanta koja ne zavisi od h. Vidi se da (6) $\Rightarrow$ (5).

Takođe, kasnije će biti dokazano da za rješenje $u = u_{ij}$ zadatka (3)–(4) važi nejednakost (uslov stabilnosti)

$$\|u\|_{\overline{\omega}} \leq C_1 \|f\|_\omega + C_2 \|g\|_{\partial\omega}. \tag{7}$$

Ovdje su C_1 i C_2 komstante koje ne zavise od h. Nejednakost (7) govori da je diferencna šema (3)–(4) stabilna u odnosu na svoju desnu stranu $f = f_{ij}$ i svoj granični uslov $g = g_{ij}$.

Upotrebimo (7) u slučaju $f_{ij} = 0$ za $(i, j) \in \omega$ i $g_{ij} = 0$ za $(i, j) \in \partial\omega$ (slijedi $\|f\| = 0$ i $\|g\| = 0$). Tada (7) povlači $\|u\| = 0$ (slijedi $u_{ij} = 0$ za $(i, j) \in \overline{\omega}$). S druge strane, tada je sistem (3)–(4) homogen. Dakle, sistem (3)–(4), u slučaju kada je homogen, ima samo trivijalno rješenje. Prema tome, sistem (3)–(4) uvijek ima jedinstveno rješenje. Jedinstvenost rješenja je dokazana!

Rekli smo da $u^p = u^p_{ij}$ zadovoljava

$$\ell(u^p) = f^p \quad \text{za} \quad (i, j) \in \omega, \qquad u^p = g^p \quad \text{za} \quad (i, j) \in \partial\omega.$$

Kako je diferencni operator $\ell = \ell(u)$ linearan i kako je $u^t = g^t$ za $(i, j) \in \partial\omega$ to slijedi

$$\ell(u^t - u^p) = \ell(u^t) - f^p \quad \text{za} \quad (i, j) \in \omega, \qquad u^t - u^p = g^t - g^p \quad \text{za} \quad (i, j) \in \partial\omega.$$

Dakle, funkcija $u^t - u^p$ zadovoljava jedan sistem oblika (3)–(4), pa na tu funkciju može da bude primijenjena nejednakost (7). Tako imamo

$$\|u^t - u^p\|_{\overline{\omega}} \leq C_1 \|\ell(u^t) - f^p\|_\omega + C_2 \|g^t - g^p\|_{\partial\omega}.$$

Dodamo i oduzmemo f^t:

$$\|u^t - u^p\|_{\overline{\omega}} \leq C_1 \|\ell(u^t) - f^t\|_\omega + C_1 \|f^t - f^p\|_\omega + C_2 \|g^t - g^p\|_{\partial\omega}.$$

Funkcija $u = u(x,y)$ je rješenje jednačine (1) (funkcija $u = u(x,y)$ je rješenje jednačine $Lu = f$):

$$\|u^t - u^p\|_{\overline{\omega}} \le C_1\|\ell(u^t) - Lu\|_\omega + C_1\|f^t - f^p\|_\omega + C_2\|g^t - g^p\|_{\partial\omega}.$$

Znamo da je u^t_{ij} samo druga oznaka za $u(ih, jh)$:

$$\|u^t - u^p\|_{\overline{\omega}} \le C_1\|\ell(u) - Lu\|_\omega + C_1\|f^t - f^p\|_\omega + C_2\|g^t - g^p\|_{\partial\omega}. \qquad (*)$$

Iskoristimo sada relaciju (5), pretpostavljajući dopunski da je $\lim\limits_{h\to 0}\|f^t - f^p\|_\omega = 0$ i $\lim\limits_{h\to 0}\|g^t - g^p\|_{\partial\omega} = 0$ (približne vrijednosti ulaznih podataka treba da konvergiraju ka odgovarajućim tačnim vrijednostima). Tako imamo

$$\lim_{h\to 0}\|u^t - u^p\|_{\overline{\omega}} = 0. \qquad (8)$$

Dobili smo da $\rho \to 0$, gdje je ρ oznaka za grešku ($\rho = \|u^t - u^p\|_{\overline{\omega}}$). Greška teži ka nuli. Dokazano je da razmatrana numerička metoda konvergira!

Možemo se osloniti na (6) umjesto na (5), što je sada na redu.

Pretpostavimo dopunski da je $\|f^t - f^p\|_\omega = O(h^2)$ za $h > 0$ i da je $\|g^t - g^p\|_{\partial\omega} = O(h^2)$ za $h > 0$. Iz $(*)$ i (6) imamo

$$\|u^t - u^p\|_{\overline{\omega}} \le C_1 C h^2 + C_1 O(h^2) + C_2 O(h^2),$$

$$\|u^t - u^p\|_{\overline{\omega}} \le C_3 h^2 \quad (C_3 = \text{const}) \qquad \text{ili} \qquad \|u^t - u^p\|_{\overline{\omega}} = O(h^2). \qquad (9)$$

Dokazano je da numerička metoda ima drugi stepen konvergencije! Dokaz je završen!

Aproksimacija + stabilnost $\Rightarrow$ konvergencija. Drugi stepen aproksimacije + stabilnost $\Rightarrow$ drugi stepen konvergencije.

Samo se napominje da je proučavanje odstupanja $f^t - f^p$ i odstupanja $g^t - g^p$ manje značajno i da se ponekad uzima prosto da je $\|f^t - f^p\|_\omega = 0$ i $\|g^t - g^p\|_{\partial\omega} = 0$.

5.3. DOKAZ ZA APROKSIMACIJU, TJ. DOKAZ ZA RELACIJU (6)

Polazimo od:

$$Lu(x,y) = -\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) \qquad \text{i}$$

$$\ell(u(x,y)) = -(u(x+h,y) - 2u(x,y) + u(x-h,y) + u(x,y+h) - 2u(x,y) + u(x,y-h))/h^2.$$

Poznate su formule:

$$(u(x+h,y) - 2u(x,y) + u(x-h,y))/h^2 - \frac{\partial^2 u}{\partial x^2}(x,y) = (h^2/12)\frac{\partial^4 u}{\partial x^4}(\xi,y), \quad \xi \in (x-h, x+h),$$

$$(u(x,y+h) - 2u(x,y) + u(x,y-h))/h^2 - \frac{\partial^2 u}{\partial y^2}(x,y) = (h^2/12)\frac{\partial^4 u}{\partial y^4}(x,\eta), \quad \eta \in (y-h, y+h).$$

Saberemo dvije posljednje formule:

$$\ell(u(x,y)) - Lu(x,y) = O(h^2).$$

Dopunski smo pretpostavili da analitičko rješenje graničnog zadatka (1)–(2) (dopunski smo pretpostavili da funkcija $u = u(x,y)$) ispunjava uslov $u \in C^4(\overline{\Omega})$ ($\Rightarrow$ njeni četvrti parcijalni izvodi su ograničeni na skupu $\overline{\Omega}$).

Time je (6) dokazano.

5.4. DOKAZ ZA STABILNOST, TJ. DOKAZ ZA NEJEDNAKOST (7)

U okviru dokaza za stabilnost govorimo samo o diskretno definisanim funkcijama (o funkcijama koje su definisane u čvorovima mreže). Znamo da v_{ij} označava vrijednost funkcije $v = v_{ij}$ u čvoru $(i,j) \in \overline{\omega}$.

Lema 1. Neka je funkcija $v = v_{ij}$ definisana na skupu $\overline{\omega}$ i neka je $\Delta v \geq 0$ za $(i,j) \in \omega$. Tada se najveća vrijednost funkcije $v = v_{ij}$ dostiže bar u jednoj tački skupa $\partial\omega$ (bar u jednom graničnom čvoru).

Ovdje je diskretni Laplaceov operator

$$\Delta v_{ij} = \frac{v_{i+1,j} + v_{i-1,j} + v_{i,j+1} + v_{i,j-1} - 4v_{ij}}{h^2}, \qquad \text{tj.} \qquad \Delta v = -\ell(v).$$

Dokaz. Dopustimo suprotno, da je $v_{ij} < V$ za svako $(i,j) \in \partial\omega$, gdje smo označili $V = \max\limits_{(i,j)\in\overline{\omega}} v_{ij}$. Tada se vrijednost V dostiže u jednoj ili u više unutrašnjih tačaka. Uočimo među tim tačkama onu koja ima najveću apscisu, odnosno (ako takvih tačaka ima više) uočimo jednu takvu tačku; neka je njen indeks (k,ℓ). Tako da je $v_{k\ell} = V$ i $v_{k+1,\ell} < V$. Što se tiče ostale tri tačke iz šablona, važi $v_{k-1,\ell} \leq V$, $v_{k,\ell+1} \leq V$ i $v_{k,\ell-1} \leq V$. Sabiranjem: $-4v_{k\ell} + v_{k+1,\ell} + v_{k-1,\ell} + v_{k,\ell+1} + v_{k,\ell-1} < 0$, slijedi (kada se podijeli sa h^2) $\Delta v_{k\ell} < 0$. Nejednakost $\Delta v_{k\ell} < 0$ protivrječi uslovu leme $\Delta v \geq 0$. Dokaz je završen.

Lema 2. Neka je funkcija $v = v_{ij}$ definisana na skupu $\overline{\omega}$ i neka je $\Delta v \leq 0$ za $(i,j) \in \omega$. Tada se najmanja vrijednost funkcije $v = v_{ij}$ dostiže bar u jednoj tački skupa $\partial\omega$.

Dokazuje se analogno.

Teorema 1 (diskretni princip maksimuma modula). Neka je funkcija $v = v_{ij}$ definisana na skupu $\overline{\omega}$ i neka je $\Delta v = 0$ za $(i,j) \in \omega$. Tada se najveća po modulu vrijednost funkcije $v = v_{ij}$ dostiže bar u jednoj tački skupa $\partial\omega$.

Dokaz. Iz $\Delta v = 0$ imamo $\Delta v \geq 0$ i $\Delta v \leq 0$ pa razmatrana funkcija $v = v_{ij}$ zadovoljava uslove i prve i druge leme. Zato se $\max\limits_{(i,j)\in\overline{\omega}} v_{ij}$ dostiže na skupu $\partial\omega$, a isto važi i za $\min\limits_{(i,j)\in\overline{\omega}} v_{ij}$. Jedan od dva broja ($\max\limits_{(i,j)\in\overline{\omega}} v_{ij}$ i $\min\limits_{(i,j)\in\overline{\omega}} v_{ij}$) jednak je upravo najvećoj po modulu vrijednosti funkcije $v = v_{ij}$ (jednak je $\max\limits_{(i,j)\in\overline{\omega}} |v_{ij}|$). Dokaz je završen.

U sljedećoj teoremi razmatraju se dva sistema linearnih jednačina. Jedan i drugi sistem su oblika (3)–(4). Pretpostavlja se da su podaci $f = f_{ij}$ i $g = g_{ij}$ drugog sistema nadređeni odgovarajućim podacima prvog sistema.

Teorema 2 (teorema o upoređivanju). Neka je v_1 (neka je $v_1 = (v_1)_{ij}$) rješenje sistema linearnih jednačina

$$\ell(v_1) = f_1 \quad \text{za} \quad (i,j) \in \omega, \qquad v_1|_{\partial\omega} = g_1 \quad (\text{ili svejedno} \quad v_1 = g_1 \quad \text{za} \quad (i,j) \in \partial\omega).$$

Neka je v_2 rješenje sistema linearnih jednačina

$$\ell(v_2) = f_2 \quad \text{za} \quad (i,j) \in \omega, \qquad v_2|_{\partial\omega} = g_2.$$

Neka je $|f_1| \leq f_2$ za $(i,j) \in \omega$ i $|g_1| \leq g_2$ za $(i,j) \in \partial\omega$. Tada za svako $(i,j) \in \overline{\omega}$ važi nejednakost $|v_1| \leq v_2$.

Dokaz. Znamo da je $\ell = -\Delta$. Za funkciju $v_2 - v_1$ važe jednakosti

$$\Delta(v_2 - v_1) = -(f_2 - f_1) \quad \text{za} \quad (i,j) \in \omega \qquad \text{i} \qquad (v_2 - v_1)|_{\partial\omega} = g_2 - g_1.$$

Po uslovu teoreme je $-(f_2 - f_1) \leq 0$ pa je za funkciju $v_2 - v_1$ ispunjen uslov leme 2. Zato možemo pisati

$$(v_2 - v_1)_{ij} \geq \min_{(i,j)\in\overline{\omega}}(v_2 - v_1) = \min_{(i,j)\in\partial\omega}(v_2 - v_1) = \min_{(i,j)\in\partial\omega}(g_2 - g_1) \geq 0,$$

jer je $(v_2 - v_1)|_{\partial\omega} = g_2 - g_1 \geq 0$. Dobili smo:

$$(v_2 - v_1)_{ij} \geq 0 \quad \text{za} \quad (i,j) \in \overline{\omega}. \tag{10}$$

Slično, za funkciju $v_2 + v_1$ je $\Delta(v_2 + v_1) = -(f_2 + f_1) \leq 0$ pa možemo pisati (opet po lemi 2)

$$(v_2 + v_1)_{ij} \geq \min_{(i,j)\in\partial\omega}(v_2 + v_1) = \min_{(i,j)\in\partial\omega}(g_2 + g_1) \geq 0.$$

Dobili smo:

$$(v_2 + v_1)_{ij} \geq 0 \quad \text{za} \quad (i,j) \in \overline{\omega}. \tag{11}$$

Iz (10) i (11) ($v_1 \leq v_2$ i $-v_1 \leq v_2$) slijedi $|v_1| \leq v_2$ u svakom čvoru mreže. Dokaz je završen.

Tvrđenje teoreme 2 ($|v_1| \leq v_2$ za $(i,j) \in \overline{\omega}$) povlači da je $\|v_1\|_{\overline{\omega}} \leq \|v_2\|_{\overline{\omega}}$.

Sada ćemo uz pomoć posljednje teoreme dokazati nejednakost (7). Treba pogodno izabrati dva diferencna granična zadatka oblika (3)–(4). Neka podaci zadatka sa indeksom 1 budu podaci koji figurišu u nejednakosti (7). Dakle, neka je $f_1 = f$ i $g_1 = g$ pa je samim tim $v_1 = u$. Što se tiče zadatka čiji je indeks 2 (zadatak koji majorira), polazi se od njegovog rješenja $v_2 = (v_2)_{ij}$. Kada se rješenje uvrsti u lijevu stranu diferencne jednačine onda se dobije odgovarajuće f_2. Slično za g_2. Dakle, neka bude

$$(v_2)_{ij} = \frac{1}{4}\left[\frac{1}{2} - \left(x - \frac{1}{2}\right)^2 - \left(y - \frac{1}{2}\right)^2\right]\|f\|_\omega + \|g\|_{\partial\omega}, \qquad (x,y) = (ih, jh). \tag{12}$$

Neposredni račun daje $\ell(v_2) = \|f\|_\omega$. Tako da je $f_2 = \|f\|_\omega$ (funkcija f_2 je konstanta). Vidimo da je ispunjen uslov $|f_1| \leq f_2$ ($|f_{ij}| \leq \|f\|_\omega$). Uslov $|g_1| \leq g_2$ svodi se na $|g_{ij}| \leq \frac{1}{4}\alpha_{ij}\|f\|_\omega + \|g\|_{\partial\omega}$, gdje je $\alpha_{ij} \geq 0$. Vidimo da je i uslov $|g_1| \leq g_2$ ispunjen. Ispunjena su oba uslova teoreme 2. Mi smo dokazali da je $|v_1| \leq v_2$ za $(i,j) \in \overline{\omega}$. Mi smo dokazali da je $\|v_1\|_{\overline{\omega}} \leq \|v_2\|_{\overline{\omega}}$, odnosno da je $\|u\|_{\overline{\omega}} \leq \|v_2\|_{\overline{\omega}}$. Jedino još ostaje da se izračuna $\|v_2\|_{\overline{\omega}}$. Da izračunamo. V. (12). Lako se vidi da je $0 \leq \frac{1}{2} - \left(x - \frac{1}{2}\right)^2 - \left(y - \frac{1}{2}\right)^2 \leq \frac{1}{2}$ za $(x,y) \in \overline{\Omega} = [0,1] \times [0,1]$, a tim prije za $(x,y) = (ih, jh)$. Zato je $\|v_2\|_{\overline{\omega}} \leq \frac{1}{8}\|f\|_\omega + \|g\|_{\partial\omega}$. Mi smo dokazali nejednakost:

$$\|u\|_{\overline{\omega}} \leq \frac{1}{8}\|f\|_\omega + \|g\|_{\partial\omega}.$$

Time smo dokazali (7). Može se staviti $C_1 = \frac{1}{8}$ i $C_2 = 1$.

Zaključak: izložena numerička metoda konvergira, s tim da je brzina konvergencije drugog reda. Iteracije za (3)–(4): $u_{ij}^{(k+1)} = \frac{1}{4}\left[u_{i+1,j}^{(k)} + u_{i-1,j}^{(k)} + u_{i,j+1}^{(k)} + u_{i,j-1}^{(k)}\right] + h^2 f_{ij}$, $k \geq 0$, tzv. metoda relaksacije.

6. METODA KONAČNIH RAZLIKA ZA RJEŠAVANJE PARCIJALNE DIFERENCIJALNE JEDNAČINE PARABOLIČKOG TIPA

6.1. NUMERIČKA METODA I NJENA SVOJSTVA

Kako glasi analitički problem

Razmotrimo problem paraboličkog tipa koji ima jednu prostornu promjenljivu x. Taj zadatak opisuje, između ostalog, proces hlađenja odnosno zagrijavanja tankog metalnog štapa koji je postavljen po x–osi i čija se temperatura $u(x,t)$ u pojedinoj tački mijenja kako protiče vrijeme t. Na engleskom se kaže thin metallic rod. Drukčije se kaže jednačina provođenja toplote.

Nepoznata funkcija $u = u(x,t)$ razmatra se za $(x,t) \in \overline{\Omega} = [0,1] \times [0,T]$, gdje je T pozitivna konstanta. Ta funkcija zadovoljava jednačinu

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(x,t) \quad \text{za} \quad 0 < x < 1,\ 0 < t \leq T, \tag{1}$$

početni uslov

$$u(x,0) = \alpha(x) \quad \text{za} \quad 0 \leq x \leq 1, \tag{2}$$

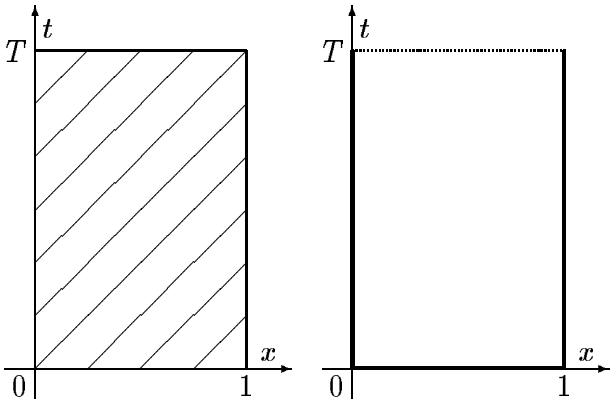
kao i dva granična uslova

$$u(0,t) = 0, \quad u(1,t) = 0 \quad \text{za} \quad 0 \leq t \leq T. \tag{3}$$

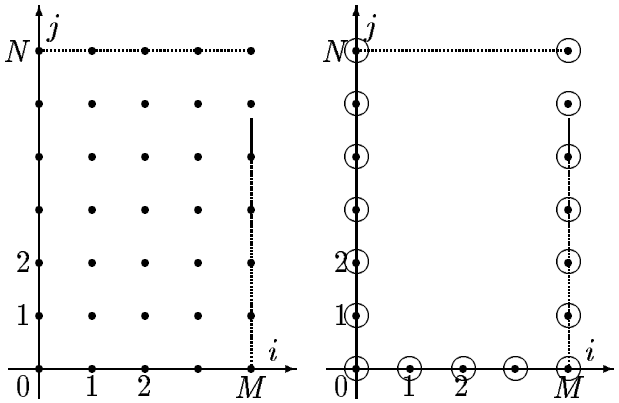
Zadatak sadrži i početne i granične uslove pa se zato naziva mješovitim.

Svuda dalje ćemo smatrati da su date funkcije $f(x,t)$ i $\alpha(x)$ dovoljno glatke i takve da postoji jedinstveno rješenje razmatranog zadatka $u = u(x,t)$ koje je dovoljno glatka funkcija. Time smo identifikovali prvi objekat: tačno rješenje. Na slici 1 prikazana je u (x,t) ravni oblast u kojoj se odvijaju dešavanja.

U opštem slučaju, granični uslovi ne moraju da budu homogeni, već mogu da glase $u(0,t) = \mu(t)$, $u(1,t) = \nu(t)$ za $0 \leq t \leq T$. Ovakvi granični uslovi se lako svode na homogene, pomoću smjene nepoznate funkcije po formuli $v(x,t) = u(x,t) - u_0(x,t) = u(x,t) - [(1-x)\mu(t) + x\nu(t)]$, gdje $v(x,t)$ zadovoljava nulte granične uslove, ako $u(x,t)$ zadovoljava opšte granične uslove. Dakle, ako je u početku postavljen zadatak sa opštim graničnim uslovima onda se on odmah svede, posredstvom ukazane smjene, na zadatak tipa (1)–(3). Pri sprovođenju smjene dođe do promjene desne strane (f) i početnog uslova (α). Ono što bude rečeno o zadatku (1)–(3), skoro bez promjene važi i za opšti zadatak.



Slika 1: Skup $\overline{\Omega}$ i unaprijed određeni dio



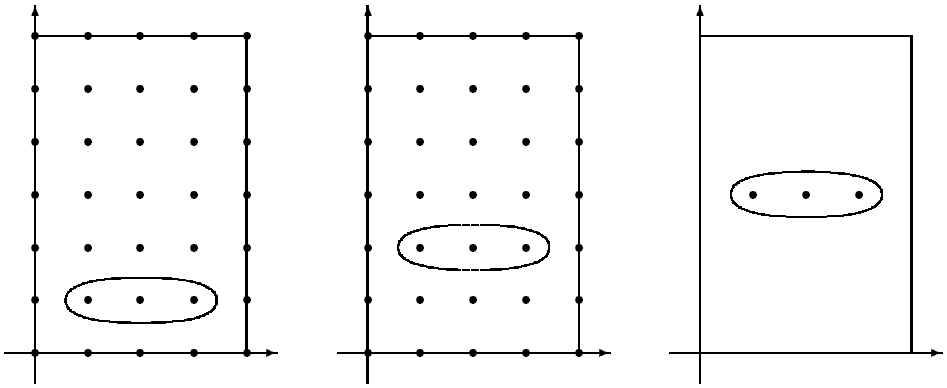
Sl. 2: Čvorovi (i,j) i unaprijed određeni dio

Uvod o numeričkoj metodi

Prelazimo na izlaganje numeričke metode za koju se još koristi i naziv **metoda presjeka**.

Opredijelimo se za prostorni, tj. dužinski korak h, neka je $h = 1/M$, neka je $x_m = mh$ za $m = 0, 1, \ldots, M$. Takođe, izaberimo vremenski korak τ, gdje je $N\tau = T$ i neka je $t_n = n\tau$. Čvorovi su tačke oblika (x_m, t_n). Skup svih čvorova (za $0 \leq m \leq M$, $0 \leq n \leq N$) čini mrežu, a označava se kao $\overline{\omega}$ ili $\overline{\omega}_{h,\tau}$. Ponekad će se za čvor (x_m, t_n) koristiti njegova cjelobrojna oznaka (m, n). Vrijednost približnog rješenja u pojedinom čvoru označavaćemo sa u_m^n, tj. $u(x_m, t_n) \approx u_m^n$; $u(x,t)$ – tačno rješenje za (1)–(3). Svi čvorovi koji imaju jedan te isti gornji indeks (drugi indeks) čine jedan vremenski sloj ili jedan presjek. Tako se početni ili nulti presjek sastoji od $(0,0), (1,0), \ldots, (M,0)$. Sve približne vrijednosti koje odgovaraju jednom određenom vremenskom sloju čine jedan vektor koji će biti označavan kao $\mathbf{u}^n$, $\mathbf{u}^n = (u_0^n, u_1^n, \ldots, u_M^n)$. Na slici 2 prikazana je ukupna mreža tačaka (čvorova).

Budući da je zadatak (1)–(3) mješovitog tipa (po promjenljivoj t je to jedan početni zadatak), to se približne vrijednosti mogu postepeno određivati (može se po t postepeno napredovati). Ovo je očito povoljna okolnost koja utiče da se približno rješenje lako izračuna. Tako su vrijednosti $\mathbf{u}^0$ koje se odnose na početni vremenski sloj – date početnim uslovom $\alpha(x)$. Približne vrijednosti sa prvog vremenskog sloja ($t = \tau$) određuju se prve, to je vektor $\mathbf{u}^1$. Zatim se određuje vektor $\mathbf{u}^2$, odgovara mu $t = 2\tau$. Itd. Sve dok se ne dođe do najgornjeg sloja $t = T$. Pri ovome se, kod računanja $\mathbf{u}^j$, oslanjamo na već poznate vrijednosti $\mathbf{u}^{j-1}$. Na slici 3 prikazan je redosljed napredovanja po vremenskim nivoima, sa jednog nivoa na drugi.



Slika 3: Redosljed računanja: zaokruženi su čvorovi za pojedini korak

Neka smo došli do n–tog vremenskog sloja, tj. već imamo približne vrijednosti $\mathbf{u}^0, \ldots \mathbf{u}^{n-1}$, a želimo da odredimo probližne vrijednosti $\mathbf{u}^n$. Za njihovo računanje koristimo već dobijene $\mathbf{u}^{n-1}$ pa se za šemu kaže da je dvoslojna. Prvo se radi za $n = 1$, a tada nam je $\mathbf{u}^0$ poznato jer je $\mathbf{u}^0 = \alpha$, detaljnije $u_m^0 = \alpha(mh)$ za svako m od 0 do M. Zatim $\mathbf{u}^2$ pomoću $\mathbf{u}^1$. Itd.

U jednačini se pojavljuje drugi izvod (u_{xx}). Za njegovu aproksimaciju koristimo običnu formulu $y''(x) \approx \frac{1}{h^2}\big(y(x+h) - 2y(x) + y(x-h)\big)$, napisano u opštim oznakama. U jednačini se pojavljuje i prvi izvod (u_t). Opet, da bi se formirao diferencni zadatak, i prvi izvod treba da bude zamijenjen nekom konačnom razlikom (da bude zamijenjen izrazom u kome se pojavljuju

konačne razlike). Za ovo ćemo koristiti jednostavan izraz $\frac{1}{h}\big(y(x+h)-y(x)\big)$, što će nam služiti kao približna vrijednost za $y'(x)$ ili za $y'(x+h)$ ili za $y'(x+h/2)$. Naša numerička metoda imaće tri varijante.

U numeričkoj metodi postoje tri opcije

Na fiksiranom presjeku ima $M+1$ čvorova odnosno $M+1$ nepoznatih približnih vrijednosti. Za $m=0$ i $m=M$ lako formiramo uslov: $u_0^n=0$, $u_M^n=0$, očito po (3). Kako formirati uslove u unutrašnjim čvorovima? Tri varijante koje se pojavljuju biće označene redom kao: slučajevi $\sigma=0$, $\sigma=1$ i $\sigma=1/2$, gdje σ ima smisao težinskog faktora (udio donjeg sloja je $1-\sigma$, a udio gornjeg sloja je σ), što će biti razjašnjeno u daljem tekstu.

Prva varijanta, $\sigma=0$. Jednačina (1) važi za svako (x,t) pa važi i kada je $(x,t)=(x_m,t_{n-1})=(mh,(n-1)\tau)$; napisati. U toj jednačini se pojavljuju u_t i u_{xx} evaluirani u toj tački (x_m,t_{n-1}). Drugi izvod će biti zamijenjen na običan način, dok će prvi izvod biti aproksimiran preko $\frac{\partial u}{\partial t}\approx\frac{1}{\tau}\big(u(x_m,t_n)-u(x_m,t_{n-1})\big)$. Tako dobijamo sljedeće jednačine:

$$\frac{1}{\tau}\big(u_m^n-u_m^{n-1}\big)=\frac{1}{h^2}\big(u_{m+1}^{n-1}-2u_m^{n-1}+u_{m-1}^{n-1}\big)+f(x_m,t_{n-1})\quad\text{za}\quad 0<m<M \tag{4}$$

ili

$$Lu_m^n=\Lambda u_m^{n-1}+f(x_m,t_{n-1})\qquad(\text{ovdje je }\sigma=0),$$

gdje je L diferencni operator definisan sa $L\mathbf{v}^n=(\mathbf{v}^n-\mathbf{v}^{n-1})/\tau$, dok je Λ diferencni operator koji djeluje kao $\Lambda\mathbf{u}_m=(\mathbf{u}_{m+1}-2\mathbf{u}_m+\mathbf{u}_{m-1})/h^2$. U (4) su za drugi izvod upotrebljene poznate vrijednosti sa prethodnog sloja, sa sloja $n-1$.

Druga varijanta, $\sigma=1$. Mogu da se za drugi izvod upotrebe (zasad nepoznate) vrijednosti sa aktuelnog sloja, sa sloja n. Drugim riječima, napisati jednačinu (1) za $(x,t)=(x_m,t_n)$. Ako se tako uradi, nastaje sljedeći sistem uslova:

$$\frac{1}{\tau}\big(u_m^n-u_m^{n-1}\big)=\frac{1}{h^2}\big(u_{m+1}^{n}-2u_m^{n}+u_{m-1}^{n}\big)+f(x_m,t_{n})\quad\text{za}\quad 0<m<M \tag{5}$$

ili

$$Lu_m^n=\Lambda u_m^{n}+f(x_m,t_{n})\qquad(\text{ovdje je }\sigma=1).$$

Treća varijanta, slučaj $\sigma=1/2$ (ili slučaj opšteg σ). Može se pokušati sa kombinacijom prethodne dvije šeme (4) i (5). Nećemo se opredijeliti za Λu_m^{n-1} niti za Λu_m^n već za njihovu kombinaciju $(1-\sigma)\Lambda u_m^{n-1}+\sigma\Lambda u_m^n$, obično se parametar σ izabere tako da pripada intervalu $[0,1]$. Kaže se: šema sa težinama odnosno šema sa težinskim faktorima. Najčešće je $\sigma=1/2$, jer se dobijaju dobra svojstva aproksimacije pa dalje (što se tiče treće varijante) govorimo samo o tom slučaju $\sigma=1/2$. Riječima, prethodni i aktuelni sloj su istog značaja. Koristiće se $(\Lambda u_m^{n-1}+\Lambda u_m^n)/2$. Govoreći drukčije, zapišimo jednačinu (1) za $(x,t)=(x_m,t_n-\tau/2)$. Zatim stavimo $\frac{\partial u(x_m,t_{n-1/2})}{\partial t}\approx\frac{1}{\tau}\big(u(x_m,t_n)-u(x_m,t_{n-1})\big)$,

$$\frac{\partial^2 u(x_m,t_{n-1/2})}{\partial x^2}\approx\frac{1}{2}\left(\frac{\partial^2 u(x_m,t_{n-1})}{\partial x^2}+\frac{\partial^2 u(x_m,t_n)}{\partial x^2}\right).$$

Tako se dobijaju sljedeći uslovi:

$$\frac{1}{\tau}\left(u_m^n - u_m^{n-1}\right) = \frac{1}{2}\left[\frac{1}{h^2}\left(u_{m+1}^{n-1} - 2u_m^{n-1} + u_{m-1}^{n-1}\right) + \frac{1}{h^2}\left(u_{m+1}^{n} - 2u_m^{n} + u_{m-1}^{n}\right)\right] + f(x_m, t_{n-1/2}) \quad (6)$$

za $0 < m < M$ ili

$$Lu_m^n = \frac{1}{2}\left[\Lambda u_m^{n-1} + \Lambda u_m^n\right] + f(x_m, t_{n-1/2}) \qquad \text{(ovdje je } \sigma = 1/2\text{)}.$$

Numerički algoritam

Dalje, napišimo još dvije jednačine koje proističu iz graničnih uslova, a zajedničke su za sve tri varijante:

$$u_0^n = 0, \quad u_M^n = 0. \quad (7)$$

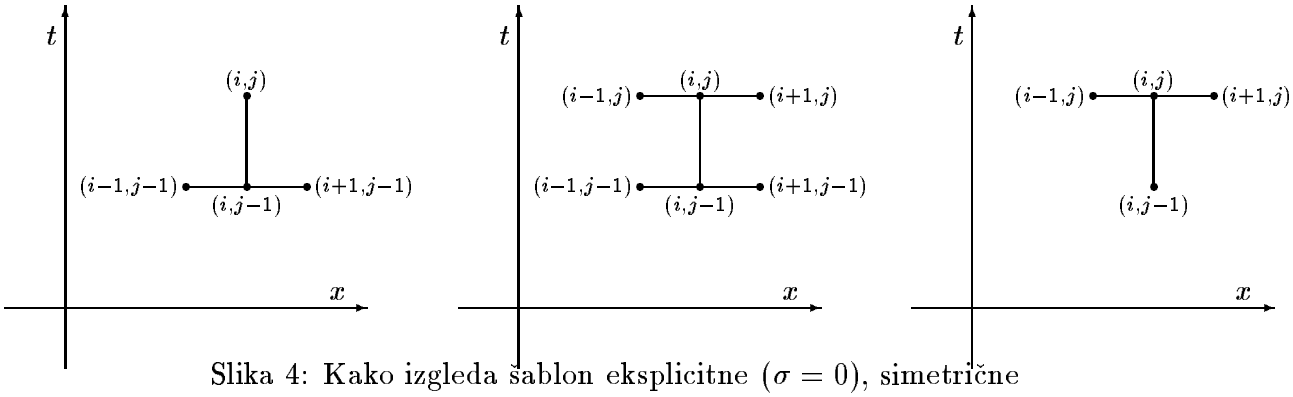
Vidimo da je (4), (7) jedan sistem linearnih jednačina, jednačina ima $M+1$ a i nepoznatih ima $M+1$, to su $u_0^n, u_1^n, \ldots, u_M^n$. Isto važi za (5), (7). A isto važi i za sistem (6), (7).

Već je rečeno da je u sistemima $n = 1$, zatim $n = 2$, itd. na kraju $n = N$. Po n se napreduje postepeno, uvijek se oslanjajući na prethodni sloj. U sve tri opcije, za $n = 0$ imamo, na osnovu datog početnog uslova (2):

$$u_m^0 = \alpha(mh) \quad \text{za} \quad 0 \le m \le M. \quad (8)$$

Time je definisan numerički algoritam za nalaženje svih približnih vrijednosti u_m^n, $0 \le m \le M$, $0 \le n \le N$ (za bilo koju od tri varijante $\sigma = 0$, $\sigma = 1$, $\sigma = 1/2$). Naravno da svaka varijanta produkuje svoje približne vrijednosti. Dakle, definisali smo tri posebne odnosno različite diferencne šeme. Treba ispitati svojstva svake od njih.

Šemi $\sigma = 0$ odgovara šablon od četiri tačke: $(m-1, n-1)$, $(m, n-1)$, $(m+1, n-1)$, (m, n). Šemi $\sigma = 1$ odgovara šablon od sljedeće četiri tačke: $(m, n-1)$, $(m-1, n)$, (m, n), $(m+1, n)$, dok šema u kojoj je $\sigma = 1/2$ ima šablon koji uključuje sljedećih šest čvorova: $(m \pm 1, n-1)$, $(m, n-1)$, $(m \pm 1, n)$, (m, n). Ove okolnosti prikazane su na slici 4.



Slika 4: Kako izgleda šablon eksplicitne ($\sigma = 0$), simetrične ($\sigma = 1/2$), odnosno čisto implicitne šeme ($\sigma = 1$)?

Za šemu (4) kaže se da je **eksplicitna**. Zaista, pojedina jednačina sistema sadrži jedino nepoznatu u_m^n (rekli smo već da su nepoznate $u_0^n, u_1^n, \ldots, u_M^n$), tako da se sve nepoznate u_m^n (za $0 \le m \le M$) mogu direktno izračunati. I dalje ćemo govoriti da je to – jedan sistem linearnih

jednačina, mada bi se moglo reći i da je to sistem eksplicitnih formula $u_m^n = \ldots$. Tako da se u slučaju (4) uopšte ne postavlja pitanje o postojanju približnog rješenja $\{u_m^n\}$. Šeme (5) i (6) su implicitne jer nepoznate $u_0^n, u_1^n, \ldots, u_M^n$ ne mogu da budu neposredno izračunate. Šema (5) se naziva **čisto implicitnom**. Šema (6), u kojoj je $\sigma = 1/2$, zove se **simetrična**.

Za svaku od tri šeme, ispitivanje se sastoji od nekoliko koraka, a biće rađeno "zajedno".

Da li je linearni sistem određen?

To znači – da li ima jedinstveno rješenje. Drugim riječima, postoji li uopšte približno rješenje $\{u_m^n\}$ i da li je ono jedinstveno? Za eksplicitnu šemu je očito da je odgovor potvrdan. Isti odgovor važi i za druge dvije šeme. Zaista, u slučaju (5), (7) matrica sistema je oblika $(M+1) \times (M+1)$ i glasi:

$$\begin{bmatrix} 1 & 0 & 0 & 0 & \ldots & 0 \\ -1 & 2+h^2/\tau & -1 & 0 & \ldots & 0 \\ 0 & -1 & 2+h^2/\tau & -1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \end{bmatrix}$$

(pomnoženo je sa h^2). Razmatrana matrica je očito dijagonalno–dominantna (element sa glavne dijagonale nadmašuje zbir svih ostalih elemenata u njegovom redu). Poznato je da je takva matrica – regularna. Dakle, sistem (5), (7) je jednoznačno rješiv. Slično važi i za treću (simetričnu) šemu. Naime, matrica sistema je sličnog oblika (napisati je za vježbu) i odmah se vidi da je i ona dijagonalno–dominantna. Tako da je i sistem (6), (7) jednoznačno rješiv. Prema tome, identifikovali smo i drugi objekat – približno rješenje (važi za sve tri šeme). Zato se sada može postaviti pitanje o rastojanju ta dva objekta kad $(h, \tau) \to (0, 0)$; da li rastojanje analitičkog i približnog teži ka nuli, kojom brzinom ono teži ka nuli, šta je to rastojanje (po kojoj normi se mjeri). O tome uskoro.

U slučaju čisto implicitne šeme, a i u slučaju simetrične šeme, matrica sistema je očito trodijagonalna pa se taj linearni sistem može efikasno riješiti metodom Thomasa. Poznato je da metoda Thomasa zahtijeva $O(M)$ aritmetičkih operacija. Zanimljivo je da $O(M)$ aritmetičkih operacija treba i da se saznaju $u_0^n, u_1^n, \ldots, u_M^n$ po eksplicitnoj šemi. Dakle, eksplicitna šema ima isti red veličine potrebnog računarskog vremena (da bi se saznale približne vrijednosti) kao i bilo koja od dvije implicitne šeme. Drugim riječima, eksplicitna šema nije znatno brža (efikasnija), barem u razmatranom jednodimenzionom slučaju.

Ispitivanje aproksimacije

Koliko diferencijalni operator $u_t - u_{xx}$ odstupa od operatora koji je poslužio za dobijanje diferencne šeme? Zanima nas – koliko je odstupanje u jednom proizvoljnom čvoru, a takođe – ocjena odstupanja koja važi u bilo kom čvoru.

Neka je $u = u(x,t)$ bilo koja dovoljno glatka funkcija. Uvedimo oznake:

$$(\sigma = 0) \qquad r_m^n = \{u_t - u_{xx}\}(x_m, t_{n-1}) - \{Lu_m^n - \Lambda u_m^{n-1}\}$$

ili (pišući detaljnije)

$$(\sigma = 0) \qquad r_m^n = \Big\{u_t(x_m, t_{n-1}) - u_{xx}(x_m, t_{n-1})\Big\} -$$

$$\left\{\frac{1}{\tau}\big(u(x_m,t_n)-u(x_m,t_{n-1})\big)-\frac{1}{h^2}\big(u(x_{m+1},t_n)-2u(x_m,t_n)+u(x_{m-1},t_n)\big)\right\},$$

slično

$$(\sigma=1)\qquad r_m^n=\{u_t-u_{xx}\}(x_m,t_n)-\{Lu_m^n-\Lambda u_m^n\}\qquad u_m^n \text{ se odnosi na } u(x,t),$$

$$(\sigma=1/2)\qquad r_m^n=\{u_t-u_{xx}\}(x_m,t_{n-1/2})-\left\{Lu_m^n-\frac{1}{2}\big[\Lambda u_m^{n-1}+\Lambda u_m^n\big]\right\}.$$

Izračunajmo za $\sigma=0$. Treba funkciju $u=u(x,t)$ razviti po Taylorovoj formuli u tački $(x,t-\tau)=(x_m,t_{n-1})$. Imamo da je

$$\frac{\partial u(x,t-\tau)}{\partial t}-\frac{1}{\tau}\big[u(x,t)-u(x,t-\tau)\big]=-\frac{\tau}{2}\cdot\frac{\partial^2u(x,t-\tau+\xi)}{\partial t^2},\quad 0<\xi<\tau,$$

$$\frac{\partial^2u(x,t-\tau)}{\partial x^2}-\frac{1}{h^2}\big[u(x+h,t-\tau)-2u(x,t-\tau)+u(x-h,t-\tau)\big]=-\frac{h^2}{12}\cdot\frac{\partial^4u(x+\eta,t-\tau)}{\partial x^4},$$

$-h<\eta<h$. Sabirajući dvije relacije:

$$(\text{kada je } \sigma=0)\qquad r_m^n=O(h^2+\tau).$$

U slučaju $\sigma=1$ razvijanje se vrši u tački $(x,t)=(x_m,t_n)$. U slučaju $\sigma=1/2$ razvijanje se vrši u tački $(x_m,t_{n-1/2})$. Zbog sličnosti postupka, izostavljamo odgovarajući jednostavni račun. Tako se dobija:

$$(\text{kada je } \sigma=1)\qquad r_m^n=O(h^2+\tau),$$

$$(\text{kada je } \sigma=1/2)\qquad r_m^n=O(h^2+\tau^2).$$

Tri navedene ocjene važe pod uslovom da su parcijalni izvodi

$$\frac{\partial^3u}{\partial x^2\partial t},\quad \frac{\partial^3u}{\partial t^3}\quad \text{i}\quad \frac{\partial^4u}{\partial x^4}$$

ograničene funkcije na skupu $\overline{\Omega}=[0,1]\times[0,T]$. Do sada je $u=u(x,t)$ bila proizvoljna funkcija koja zadovoljava maločas nabrojane uslove glatkosti (ograničenosti). Ipak, mi ćemo formulu za r_m^n koristiti samo u slučaju kada je $u=u(x,t)$ – rješenje zadatka (1)–(3). Drugim riječima, dovoljno je da uslov aproksimacije bude ispunjen na rješenju (mada je on ispunjen i "šire"). Navedeni uslovi glatkosti (ograničenosti) će očito biti ispunjeni ako je tačno rješenje $u\in C^4(\overline{\Omega})$. Vidimo da simetrična šema $\sigma=1/2$ ima bolja svojstva aproksimacije (jer τ^2).

Neka je dalje

$$R=\max_{0\le m\le M,0\le n\le N}|r_m^n|.$$

Jasno je da pod navedenom pretpostavkom $u\in C^4(\overline{\Omega})$, R ima isti red veličine kao i r_m^n, tj. da je ocjena uniformna po (h,τ). Ostaje samo da se formuliše u obliku teoreme.

Teorema 1. Važi $R=O(h^2+\tau)$ u slučajevima $\sigma=0$ i $\sigma=1$. Važi $R=O(h^2+\tau^2)$ u slučaju $\sigma=1/2$.

Time je pokazano da je uslov aproksimacije ispunjen (kad $h\to0$ i $\tau\to0$ onda greška aproksimacije teži ka nuli) i izračunat je stepen (red) aproksimacije.

Ispitivanje stabilnosti

Stabilnost se ispituje u odnosu na desnu stranu i početni uslov, a opredjeljujemo se za diskretnu normu tipa C. Znamo da je svojstvo stabilnosti vezano samo za diferencni zadatak. Uzmimo da se radi o šemi $\sigma = 0$. Dati su $\alpha_m = \alpha(mh)$ za $0 \le m \le M$, kao i $f_m^{n-1} = f(x_m, t_{n-1})$ za $0 < m < M$, $1 \le n \le N$. Posmatrajmo ukupni linearni sistem koji se sastoji od (4), (7) za $1 \le n \le N$ i (8). Ako su dati svi ulazni podaci $\{f_m^{n-1}\}$ i $\{\alpha_m\}$ onda razmatrani ukupni linearni sistem determiniše sve izlazne podatke $\{u_m^n\}$, gdje je $0 \le m \le M$, $0 \le n \le N$. Može li se desiti da su ulazni podaci "mali", a da izlazni podaci budu – "veliki"? Ako ovo može da se desi onda to znači da numerička metoda (diferencna šema) nije otporna na grešku ulaznih podataka i grešku računanja. Ako ovo ne može da se desi onda se za razmatranu diferencnu šemu kaže da je stabilna. Naravno da ista ovakva deskripcija važi i za druge dvije šeme $\sigma = 1$ i $\sigma = 1/2$. Uvedimo potrebne oznake.

Razmotrimo linearne sisteme:

$$v_m^0 = \beta_m, \quad 0 \le m \le M; \quad Lv_m^n = (1-\sigma)\Lambda v_m^{n-1} +$$

$$\sigma \Lambda v_m^n + g_m^n, \quad 0 < m < M, \quad v_0^n = 0, \quad v_M^n = 0, \quad n = 1, 2, \ldots, N. \tag{9}$$

Napišimo sljedeći uslov (uslov stabilnosti):

$$\|\mathbf{v}^j\| \le \|\beta\| + \tau \sum_{i=1}^{j} \|\mathbf{g}^i\| \quad (\text{za} \quad j = 0, \ldots, N), \tag{10}$$

gdje je $\|\mathbf{x}\|_C = \|\mathbf{x}\| = \|(x_0, x_1, \ldots, x_M)\| = \max_{0 \le k \le M} |x_k|$. Vidimo da se u (10) broj sabiraka oblika $\|\mathbf{g}^i\|$ povećava (kad $\tau \to 0$), ali se to u potpunosti kompenzuje istovremenim smanjivanjem faktora τ.

Može se dokazati (korišćenjem odgovarajućeg principa maksimuma) da važi sljedeća teorema.

Teorema 2. Eksplicitna diferencna šema ($\sigma = 0$, tj. (4), (7)) je stabilna ako važi $\tau \le h^2/2$. Za čisto implicitnu šemu ($\sigma = 1$) uslov stabilnosti (10) je ispunjen za ma kakve h i τ. Simetrična šema ($\sigma = 1/2$) je stabilna pod uslovom da je $\tau \le h^2$.

Ako (za neku šemu) uslov stabilnosti važi za ma kakve h i τ onda se za tu šemu kaže da je **bezuslovno stabilna**. A ako važi samo kada su h i τ vezani nekim uslovom onda se kaže da je ona **uslovno stabilna**. Tako su eksplicitna i simetrična šema – uslovno stabilne. Kod njihove upotrebe, treba voditi računa da h i τ zadovoljavaju odgovarajući uslov. Dok je čisto implicitna šema $\sigma = 1$ – primjer jedne bezuslovno stabilne šeme. Ako se ona upotrebljava onda se po t–osi (po vremenu) može brže napredovati (krupnijim koracima).

Dokažimo teoremu 2 u slučaju šeme $\sigma = 0$. Za ostale dvije šeme $\sigma = 1$ i $\sigma = 1/2$ dokaz je analogan.

Budući da je $\sigma = 0$, to jednačina iz (9) glasi:

$$Lv_m^n = \Lambda v_m^{n-1} + g_m^n \qquad \text{ili} \qquad \frac{1}{\tau}\left(v_m^n - v_m^{n-1}\right) = \frac{1}{h^2}\left(v_{m+1}^{n-1} - 2v_m^{n-1} + v_{m-1}^{n-1}\right) + g_m^n,$$

što se može prepisati u obliku, stavljeno je $\rho = \tau/h^2$:

$$v_m^n = \rho v_{m+1}^{n-1} + (1 - 2\rho)v_m^{n-1} + \rho v_{m-1}^{n-1} + \tau g_m^n.$$

Kako se $\max_m |v_m^n|$ dostiže u unutrašnjoj tački (m_0, n), jer je $v_0^n = v_M^n = 0$, to je

$$\max_m |v_m^n| = \max_m |\rho v_{m+1}^{n-1} + (1-2\rho)v_m^{n-1} + \rho v_{m-1}^{n-1} + \tau g_m^n| \le \max_m |\rho v_{m+1}^{n-1}| + \max_m |(1-2\rho)v_m^{n-1}| +$$

$$\max_m |\rho v_{m-1}^{n-1}| + \max_m |\tau g_m^n| \le \rho\|\mathbf{v}^{n-1}\| + (1-2\rho)\|\mathbf{v}^{n-1}\| + \rho\|\mathbf{v}^{n-1}\| + \tau\|\mathbf{g}^n\| = \|\mathbf{v}^{n-1}\| + \tau\|\mathbf{g}^n\|,$$

gdje je očito uzeta u obzir definicija norme. Tako smo dobili sljedeću nejednakost:

$$\|\mathbf{v}^n\| \le \|\mathbf{v}^{n-1}\| + \tau\|\mathbf{g}^n\|,$$

koja povezuje norme vektora na susjednim presjecima. Posljednja nejednakost važi za $n = 1, 2, \ldots$. Napišimo tu nejednakost za $n = 1, \ldots, n = j$. Ostaje samo da se ove nejednakosti saberu. Umjesto v_m^0 pišemo β_m, tj. umjesto $\|\mathbf{v}^0\|$ pišemo $\|\beta\|$. Tako:

$$\|\mathbf{v}^j\| \le \|\beta\| + \tau(\|\mathbf{g}^1\| + \ldots + \|\mathbf{g}^j\|).$$

Tako je dokazana relacija (10), i to za svako j od 0 do N. Zapaziti da je bitno što je $1-2\rho \ge 0$. **Dokaz je završen.**

Ispitivanje konvergencije

Kao u prethodnom poglavlju (šablon dokaza je isti), iz aproksimacije i stabilnosti slijedi konvergencija. Više od toga, stepen konvergencije poklapa se sa stepenom aproksimacije. Ovdje jedino treba dodati: pod uslovom da se početni uslov $\alpha(x_m)$ i desna strana f ne računaju sa manjom preciznošću od preciznosti aproksimacije. Ili: pod uslovom da se računaju sasvim tačno.

Kod ispitivanja konvergencije, interesuje nas da li se približne vrijednosti neograničeno približavaju odgovarajućim tačnim i kojim se tempom ostvaruje približavanje kada prostorni korak h i vremenski korak τ istovremeno teže ka nuli. Već su pripremljene sve oznake. Rečeno je da $u(x,t)$ znači – tačno rješenje zadatka (1)–(3) pa samim tim $u(x_m,t_n)$ znači vrijednost tačnog rješenja u čvoru (x_m,t_n), tj. cjelobrojno numerisano u čvoru (m,n). S druge strane, odgovarajuća približna vrijednost je u_m^n. Recimo, za šemu $\sigma = 0$, ona je definisana pomoću (4), (7), (8). Za druge dvije šeme, brojevi $\{u_m^n\}$ biće neki drugi. Tako da greška u čvoru iznosi $u(x_m,t_n) - u_m^n$. Interesuje nas i najveća greška, po svim čvorovima, pa uvodimo u razmatranje i veličinu

$$E = \max\{|u(x_m,t_n) - u_m^n|,\ \ 0 \le m \le M,\ \ 0 \le n \le N\}.$$

Teorema 3. Za eksplicitnu šemu ($\sigma = 0$) važi da je $E = O(h^2+\tau)$ kad $(h,\tau) \to (0,0)$, ako je zadovoljen uslov $\tau \le h^2/2$. Ili: postoji konstanta C (ne zavisi ni od h ni od τ) takva da je $E \le C(h^2+\tau)$ za sve (h,τ) takve da je $\tau \le h^2/2$. Ili: $|u(x_m,t_n) - u_m^n| \le C(h^2+\tau)$. Za čisto implicitnu šemu ($\sigma = 1$): $E = O(h^2+\tau)$ bez obzira na odnos između h i τ. Za simetričnu šemu ($\sigma = 1/2$): ako je $\tau \le h^2$ onda je $E = O(h^2+\tau^2)$.

Dokaz teoreme se izostavlja.

Time je kompletirana analiza predložene numeričke metode.

6.2. JEDAN PRIMJER SA SIMETRIČNOM ŠEMOM

Ponovimo ukratko o simetričnoj šemi. Ograničimo se na slučaj $f \equiv 0$, tako da je sada jednačina homogena i ona glasi $u_t = u_{xx}$. Ostao je bez promjene početni uslov $u(x,0) = \alpha(x)$, kao i granični uslovi $u(0,t) = u(1,t) = 0$. Ostale su i oznake $h = 1/M$, $\tau = T/N$, $x_i = ih$, $t_j = j\tau$, u_{ij} – približna vrijednost za $u(x,t)$ kada je $(x,t) = (x_i,t_j)$. Kako funkcioniše metoda presjeka? Na nultom vremenskom presjeku ($t = 0$, $j = 0$) nepoznate su $u_{00}, u_{10}, \ldots, u_{M0}$. Njihove vrijednosti definisane su početnim uslovom: $u_{i0} = \alpha(x_i)$. Neka je $j \ge 1$. Na j–tom

presjeku ($t = j\tau$) nepoznate su $u_{0j}, u_{1j}, \ldots, u_{Mj}$. Njihove vrijednosti saznaćemo kada riješimo sistem linearnih jednačina

$$\frac{1}{\tau}\big(u_{ij} - u_{i,j-1}\big) = \frac{1}{2}\Big(\frac{1}{h^2}\big(u_{i+1,j-1} - 2u_{i,j-1} + u_{i-1,j-1}\big) +$$

$$\frac{1}{h^2}\big(u_{i+1,j} - 2u_{ij} + u_{i-1,j}\big)\Big), \quad 1 \le i \le M-1, \quad u_{0j} = u_{Mj} = 0.$$

Naravno, prvo je u pitanju $j = 1$, zatim $j = 2$, itd. Mi postepeno napredujemo sve do finalnog $j = N$ (do $t = N\tau = T$). Svojstva? Simetrična šema je stabilna ako je $\tau \le h^2$. Njena greška iznosi $E = O(h^2 + \tau^2)$, gdje smo definisali $E = \max|u(x_i, t_j) - u_{ij}|$ (po mreži).

Možda je preglednije da se piše samo jedan indeks. U ton cilju. označimo sa a_i poznate vrijednosti sa $(j-1)$–vog presjeka. Nepoznate označavamo kao b_i. Tada sistem linearnih jednačina dobija oblik ($1 \le i \le M-1$)

$$\frac{1}{\tau}\big(b_i - a_i\big) = \frac{1}{2}\Big(\frac{1}{h^2}\big(a_{i+1} - 2a_i + a_{i-1}\big) + \frac{1}{h^2}\big(b_{i+1} - 2b_i + b_{i-1}\big)\Big), \qquad b_0 = b_M = 0.$$

Ilustrujmo na jednom konkretnom zadatku kako djeluje simetrična šema. Na redu je postavka.

U ravni (x,t), razmotrimo skup $\overline{\Omega} = [0,1] \times [0, \frac{6}{16}]$. Razmotrimo zadatak sa jednačinom paraboličkog tipa:

$$u_t = u_{xx}, \qquad u(x,0) = \sin \pi x, \qquad u(0,t) = u(1,t) = 0.$$

Samo se napominje da se fizički proces provođenja toplote može posmatrati od $t = 0$ pa naprijed (nema gornjeg ograničenja za t). Primjenom metode presjeka, odredite približno rješenje u_{ij}. Izaberite $M = 4$ (tako da je $h = 1/4$) i $N = 6$ (tako da je $\tau = 1/16$).

U prvoj iteraciji imamo $a_0 = 0$, $a_1 = \sqrt{2}/2$, $a_2 = 1$, $a_3 = \sqrt{2}/2$, $a_4 = 0$. Numeričke vrijednosti (rezultati) koje je kompjuter saopštio prikazani su na slici 5.

t					
$t = 6/16$	0	0,0189	0,0268	0,0189	0
$t = 5/16$	0	0,0346	0,0489	0,0346	0
$t = 4/16$	0	0,0633	0,0895	0,0633	0
$t = 3/16$	0	0,1157	0,1636	0,1157	0
$t = 2/16$	0	0,2115	0,2991	0,2115	0
$t = 1/16$	0	0,3867	0,5469	0,3867	0
$t = 0$	0	0,7071	1	0,7071	0
Heat equation	$x = 0$	$x = 1/4$	$x = 2/4$	$x = 3/4$	$x = 1$

x

Slika 5: Numerički primjer sa simetričnom šemom: prikazane su približne vrijednosti u_{ij} ($0 \le i \le 4$, $0 \le j \le 6$) koje služe kao aproksimacija za tačne vrijednosti $u(x_i, t_j)$, gdje je $x_i = ih$, $t_j = j\tau$, s tim da je $h = 1/4$, $\tau = 1/16$.

Svakako da je primjer odabran tako da raspolažemo i sa njegovim analitičkim rješenjeem, da bismo imali kompletan uvid u grešku numeričkog odgovora. Analitičko rješenje glasi $u(x,t) = e^{-\pi^2 t}\sin \pi x$.

Na redu je tabela koja se odnosi na neke tačke mreže. Ona se odnosi na tačke (x_1,t_1), (x_1,t_2), $\ldots$, (x_1,t_6). Za pojedinu tačku, prikazani su redom podaci kako slijedi. Gdje piše "analitičko" data je vrijednost analitičkog rješenja. Gdje piše "numeričko" data je približna vrijednost koju je kompjuter saopštio. Najzad, gdje piše "razlika" prikazana je razlika prethodna dva broja (prikazana je greška): $x_1 = \frac{1}{4}$:

analitičko	0,3816	0,2059	0,1111	0,0600	0,0323	0,0175
numeričko	0,3867	0,2115	0,1157	0,0633	0,0346	0,0189
razlika	−0,0051	−0,0056	−0,0046	−0,0033	−0,0023	−0,0014

.

Sada dolazi druga tabela. Ona se takođe odnosi na neke tačke (x,t) ravni. To su sada tačke $(x_2,t_1),(x_2,t_2),\ldots,(x_2,t_6)$. Smisao prikazanih podataka je isti kao maločas: $x_2 = \frac{2}{4}$:

analitičko	0,5396	0,2912	0,1572	0,0848	0,0458	0,0247
numeričko	0,5469	0,2991	0,1636	0,0895	0,0489	0,0268
razlika	−0,0073	−0,0079	−0,0064	−0,0047	−0,0031	−0,0021

.

Umjesto simetrična šema, kaže se i engl. Crank–Nicolson method.

6.3. ISPITIVANJE STABILNOSTI U DRUGIM NORMAMA

Norma vektora se može definisati na različite načine. U zavisnosti od izbora norme, može i da se izmijeni odgovor na pitanje – da li je šema stabilna, pod kojim uslovima je stabilna. Izbor norme utiče i na ispitivanje aproksimacije i konvergencije. Za praktičnu upotrebu jedne metode konačnih razlika, obično je dovoljno da ona bude stabilna po jednoj normi. Izložimo ukratko i bez dokaza neke činjenice koje se odnose na ovo pitanje.

Kako definisati normu?

Prva mogućnost. Norma tipa C.

O ovom slučaju je bilo riječi u dosadašnjem tekstu. Za vektor $\mathbf{u} = (u_0,\ldots,u_M)$ predložena norma je po definiciji jednaka $\|\mathbf{u}\|_C = \max_{0\le i\le M}|u_i|$. Ova norma je povezana, tj. usaglašena sa normom funkcije f koja je neprekidna na intervalu $[0,1]$, tj. sa normom u prostoru $C[0,1]$: $\|f\|_C = \max_{0\le x\le 1}|f(x)|$ (tzv. maksimum–norma ili supremum–norma). Norma vektora $\|\mathbf{u}\|_C$ i norma funkcije $\|f\|_C$ usaglašene su u sljedećem smislu: ako je $u_i = f(i/M)$ onda važi $\lim_{M\to\infty}\|\mathbf{u}\|_C = \|f\|_C$. Riječima, ako su u_i vrijednosti funkcije f u čvorovima onda (kad broj čvorova teži ka beskonačnosti) diskretna norma teži ka kontinualnoj normi.

U teoremi 2 je ispitana stabilnost (u odnosu na početni uslov i desnu stranu) šema $\sigma = 0$ $\tau \le h^2/2$, $\sigma = 1/2$ $\tau \le h^2$, $\sigma = 1$ bez uslova po ovoj diskretnoj normi tipa C. Zapaziti da je greška aproksimacije (pokazatelj R) u teoremi 1 mjerena po ovoj normi (R je definisano kao maksimum greški aproksimacija u pojedinim čvorovima).

Druga mogućnost. Norma tipa L^2.

Poznato je da se u Hilbertovom prostoru $L^2 = L^2(0,1)$ koji se sastoji od svih (realnih) funkcija koje su integrabilne sa kvadratom (u Lebesgueovom smislu) norma definiše kao $\|f\|_2 = \left(\int_0^1 |f(x)|^2 dx\right)^{1/2}$. Ovo je tzv. srednje–kvadratna norma. Ona proističe iz sljedećeg skalarnog proizvoda: $\langle f,g\rangle = \int_0^1 f(x)g(x)dx$, $\|f\|_2 = \sqrt{\langle f,f\rangle}$.

Usaglašena, tj. odgovarajuća norma vektora i skalarni proizvod dva vektora u ovom slučaju glase:

$$\langle \mathbf{u}, \mathbf{v} \rangle = \sum_{i=0}^{M} h u_i v_i, \qquad \|\mathbf{u}\|_2 = \langle u, u \rangle^{1/2} = \Big(\sum_{i=0}^{M} h |u_i|^2 \Big)^{1/2},$$

gdje je $\mathbf{u} = (u_0, \ldots, u_M)$, $\mathbf{v} = (v_0, \ldots, v_M)$, $h = 1/M$.

Može se ispitati stabilnost metode (bilo koje od tri metode $\sigma = 0$, $\sigma = 1/2$, $\sigma = 1$) u odnosu na predloženu normu. Što se tiče predložene norme, opredjeljujemo se za ispitivanje stabilnosti **u odnosu na početni uslov**. Ovo ispitivanje koristi za slučaj da je jednačina (1) homogena, tj. da u njoj nema sabirka $f(x,t)$. Sada se linearni sistem (9) mijenja utoliko što u njemu nema $g = g_m^n$, tj. svi g_m^n su jednaki nuli. Da li to što su veličine $\{\beta_m\}$ "male" (po predloženoj normi) povlači da je i rješenje $\{v_m^n\}$ sistema (9) $v_m^0 = \beta_m$, $0 \le m \le M$; $Lv_m^n = (1-\sigma)\Lambda v_m^{n-1} + \sigma \Lambda v_m^n$, $0 < m < M$, $v_0^n = 0$, $v_M^n = 0$, $n = 1, 2, \ldots, N$ "malo"? Dakle, uslov stabilnosti (u odnosu na početni uslov) glasi:

$$\|\mathbf{v}^j\|_2 \le C \|\beta\|_2 \qquad (\text{za} \qquad j = 0, \ldots, N).$$

Važe sljedeća tri iskaza. Kada je $\sigma = 0$ (eksplicitna šema): uslov važi ako je $\tau \le h^2/2$. Kada je $\sigma = 1/2$ (simetrična šema): uslov važi za ma kakve h i τ. Kada je $\sigma = 1$ (čisto implicitna šema): uslov važi za ma kakve h i τ.

Treća mogućnost. Energetska norma.

Stabilnost diferencne šeme se ispituje u tzv. energetskoj normi.

Stabilnost se ispituje u odnosu na početni uslov i desnu stranu. Definicija skalarnog proizvoda i norme:

$$\langle \mathbf{u}, \mathbf{v} \rangle_A = \langle A\mathbf{u}, \mathbf{v} \rangle_2 = \sum_{i=1}^{M-1} h (Au)_i v_i, \qquad \text{gdje je} \qquad (Au)_i = -\frac{1}{h^2} \Big(u_{i+1} - 2u_i + u_{i-1} \Big)$$

(ispunjeno je $A^* = A > 0$),

$$\|\mathbf{u}\|_A = \left(\sum_{i=1}^{M-1} h \left(-\frac{1}{h^2} \right) \Big[u_{i+1} - 2u_i + u_{i-1} \Big] u_i \right)^{1/2}$$

(uvijek se norma koja proističe iz skalarnog proizvoda definiše kao $\|u\|^2 = \langle u, u \rangle$). Ovdje su $\langle \ , \ \rangle_A$ i $\| \ \|_A$ definisani za vektore koji za $i = 0$ i $i = M$ imaju vrijednost nula; ako je $\mathbf{u} = (u_0, u_1, \ldots, u_M)$ onda je $u_0 = u_M = 0$.

Uslov stabilnosti (po β i g) odnosi se na linearni sistem (9), a glasi:

$$(\exists C_1 > 0)\ (\exists C_2 > 0) \quad \|\mathbf{v}^j\|_A^2 \le C_1 \|\beta\|_A^2 + C_2 \sum_{i=1}^{j} \tau \|\mathbf{g}^i\|_C^2 \qquad (\text{za} \qquad j = 0, \ldots, N).$$

Za $\mathbf{g}^i$ je upotrebljena norma tipa C. Važe sljedeća tri iskaza.

Šema $\sigma = 0$: uslov je ispunjen ako je $\tau \le h^2/2$.

Šema $\sigma = 1/2$ ili metoda Cranka–Nicolsonove: uslov je ispunjen za ma kakve h i τ.

Posljednji napisani uslov stabilnosti je realniji pokazatelj od ranijeg uslova stabilnosti (10) koji se odnosio na normu tipa C. Norma tipa C je "gruba".

Šema $\sigma = 1$: uslov je ispunjen za ma kakve h i τ.

7. ZADATAK O NAJBOLJOJ APROKSIMACIJI DATE FUNKCIJE

U ovom poglavlju biće izloženo nekoliko metoda za aproksimaciju funkcija od jedne promjenljive $y = y(x)$ i funkcija od dvije promjenljive $z = z(x, y)$.

7.1. LINEARNI TREND (LINEARNI MODEL)

Na redu je problem u kome se polazi od jedne date funkcije $y = y(x)$. Treba odrediti njenu dobru aproksimaciju. Za aproksimaciju će poslužiti jedna funkcija iz klase svih linearnih funkcija.

Neka je $n \geq 2$. Neka su $x_1, \ldots, x_n$ međusobno različiti brojevi ($x_j \neq x_i$ za $j \neq i$). Neka su dati i brojevi $y_1, \ldots, y_n$. Time su definisane tačke (x_i, y_i) u ravni. Možemo smatrati da te tačke predstavljaju vrijednosti jedne funkcije $y = y(x)$, da je $y_i = y(x_i)$. Navedene tačke čine set ulaznih podataka. Ove okolnosti prikazane su na slici 1.

Prelazimo na postavku problema. Traži se funkcija iz određene klase funkcija koja najbolje aproksimira navedene ulazne podatke. Treba se izjasniti – koju klasu funkcija imamo u vidu. Izbor klase utiče na postupak rješavanja. U ovom naslovu razmatra se najjednostavniji slučaj. Mi se opredjeljujemo za klasu linearnih funkcija, to su funkcije oblika $y = ax + b$ ($a, b \in R$). Treba odrediti a i b da se ostvari najbolja aproksimacija.

Dalje, treba dati precizan smisao riječima "najbolja aproksimacija", treba se izraziti preko formule. Posmatrajmo tačku $x = x_i$. Obično se za y_i kaže da je vrijednost dobijena mjerenjem. Za njenu aproksimaciju poslužiće veličina ax_i+b. Obično se za nju kaže da je teorijska vrijednost ili vrijednost po modelu. Ponovimo da a i b još nisu izračunati. Uvedimo u razmatranje razliku dva broja $r_i = ax_i + b - y_i$. Poželjno je da razlika (odstupanje) bude što manje. Kao mjera za odstupanje u pojedinoj tački uzima se $|r_i|^2 = (ax_i + b - y_i)^2$. Ukupno rastojanje (ukupno odstupanje) dobija se sabiranjem pojedinačnih. Mi pišemo $r^2 = \sum_{i=1}^n |r_i|^2 = \sum_{i=1}^n (ax_i+b-y_i)^2$. Prirodno je da se veličina r nazove ukupnim rastojanjem. Možemo pisati $r = r(a, b)$, r zavisi od a i b. Mi tražimo (a, b) za koje se dostiže globalni minimum funkcije $r = r(a, b)$. Neformalno se zapisuje da želimo da riješimo problem "$r \to \min$". Sada je problem koji predstavlja predmet razmatranja dobio formalan i precizan matematički izraz (oblik)!

Uvedimo oznaku $F(a, b) = r^2$. Jasno, za iste (a, b) se dostiže minimum i funkcije $r = r(a, b)$ i funkcije $F = F(a, b)$. Ubuduće radimo sa F, jer je tako pogodnije, da ne bismo pisali $\sqrt{\ }$. Prepišimo formulu $F(a, b) = \sum_{i=1}^n (ax_i + b - y_i)^2$.

Parcijalni izvodi $\frac{\partial F}{\partial a} = 2\sum_{i=1}^n x_i(ax_i + b - y_i)$, $\frac{\partial F}{\partial b} = 2\sum_{i=1}^n (ax_i + b - y_i)$. Želimo da nađemo stacionarnu tačku, pa zato $\frac{\partial F}{\partial a} = 0$, $\frac{\partial F}{\partial b} = 0 \Rightarrow a\sum_{i=1}^n x_i^2 + b\sum_{i=1}^n x_i - \sum_{i=1}^n x_i y_i = 0$, $a\sum_{i=1}^n x_i + nb - \sum_{i=1}^n y_i = 0$. Dobili smo sistem linearnih jednačina po nepoznatim a i b. Možemo ga zapisati u matričnom obliku:

$$\begin{bmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & n \end{bmatrix} \cdot \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n y_i \end{bmatrix}. \qquad (*)$$

Matricu sistema označimo kao $M = [m_{ij}]_{i,j=1}^2$. Lako se vidi da je $\det M \neq 0$. Zaista, primijenimo poznatu Cauchy–Schwarzovu nejednakost: $|\langle \mathbf{u}, \mathbf{v} \rangle|^2 \leq \|\mathbf{u}\|^2 \|\mathbf{v}\|^2$, znak jednakosti važi samo ako su vektori $\mathbf{u}$ i $\mathbf{v}$ kolinearni. Naravno da je skalarni proizvod $\langle \mathbf{u}, \mathbf{v} \rangle = \sum_{i=1}^n u_i v_i$ i norma $\|\mathbf{u}\| = \sqrt{\sum_{i=1}^n |u_i|^2}$. Očito, $\mathbf{u} = (u_1, \ldots, u_n) \in R^n$. Nama služi nejednakost kada je $\mathbf{u} = (x_1, \ldots, x_n)$, $\mathbf{v} = (1, \ldots, 1)$. Dakle, sistem ima jedinstveno rješenje (a, b). U daljem tekstu sa (a, b) označavamo rješenje sistema linearnih jednačina.

Da li je (a,b) tačka minimuma funkcije $F = F(a,b)$? Po Sylvesterovom kriterijumu, dovoljno je da početni glavni minori matrice M budu pozitivni. Drugim riječima, ako je $m_{11} > 0$ i $\det M > 0$ onda imamo tačku minimuma. Obe nejednakosti su ispunjene. Zaista, $\sum_{i=1}^{n} x_i^2 > 0$ jer je $n \geq 2$. Isto tako, $\det M > 0$ jer je u našem slučaju $|(\mathbf{u},\mathbf{v})|^2 < \|\mathbf{u}\|^2\|\mathbf{v}\|^2$. Znači, jeste minimum.

Najzad, lokalni minimum naše funkcije je i njen globalni minimum, budući da ona drugih ekstrena nema. Time je postavljeni zadatak riješen u potpunosti!

Mogli smo u uvodu kazati kako slijedi. Mi unaprijed znamo (mi na bazi teorije znamo) da je zavisnost x i y linearna, ona ima oblik $y = ax + b$, samo ne znamo čemu su jednaki a i b. U cilju njihovog određivanja, izvršili smo n mjerenja, čime smo dobili ulazne podatke kako je rečeno. Na osnovu tih podataka, treba odrediti a i b koji najbolje naliježu na izmjerene vrijednosti.

Za izloženu metodu koristi se naziv METODA NAJMANJIH KVADRATA, od izgleda funkcije F i od "$F \to \min$". Za izloženu numeričku metodu koristi se i naziv fitovanje, od engleskog "best fit". Takođe se koristi i naziv linearna regresija, dolazi iz matematičke statistike.

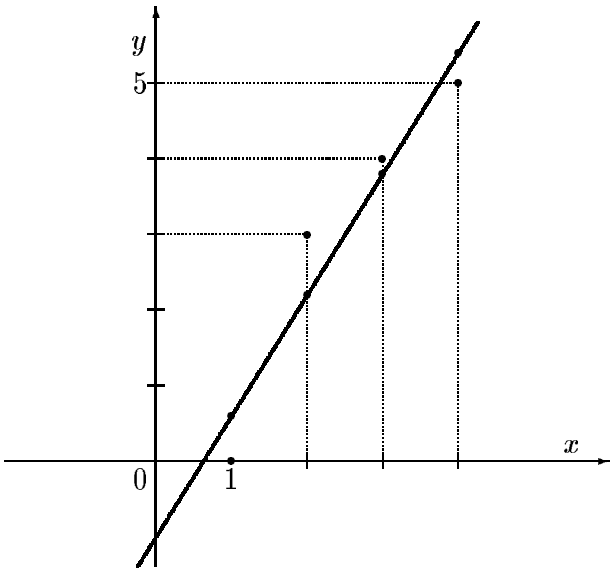
Primjer. Naći najbolju aproksimaciju oblika $y = ax + b$ po metodi najmanjih kvadrata za podatke

x	1	2	3	4
y	0	3	4	5

. Izrada. Imamo redom $\sum x_i^2 = 30$, $\sum x_i = 10$, $\sum x_i y_i = 38$, $\sum y_1 = 12$. Sistem $\begin{cases} 30a + 10b = 38 \\ 10a + 4b = 12 \end{cases}$ Njegovo rješenje $a = 1{,}6$, $b = -1$. Odgovor: $y = 1{,}6x - 1$. V. sliku 2.

i	x_i	y_i
1	x_1	y_1
2	x_2	y_2
...	...	...
...	...	...
...	...	...
n	x_n	y_n

Slika 1: Ulazni podaci za zadatak o aproksimaciji



Slika 2: Linearni trend $y = ax + b$ ($y = 1{,}6x - 1$)

Primjedba. Za sistem $(*)$ se kaže da predstavlja NORMALNI SISTEM jednačina. Ako uvedemo oznake $\mathbf{x} = (x_1, \ldots, x_n)$, $\mathbf{y} = (y_1, \ldots, y_n)$, $\mathbf{1} = (1, \ldots, 1)$, tada sistem glasi $\langle \mathbf{y} - a\mathbf{x} - b\mathbf{1}, \mathbf{x} \rangle = 0$, $\langle \mathbf{y} - a\mathbf{x} - b\mathbf{1}, \mathbf{1} \rangle = 0$. Ako uvedemo oznake $\mathbf{z} = a\mathbf{x} + b\mathbf{1}$ za aproksimacioni vektor i $\mathbf{r} = \mathbf{y} - \mathbf{z}$ za vektor razlike, tada sistem glasi $\langle \mathbf{y} - \mathbf{z}, \mathbf{x} \rangle = 0$, $\langle \mathbf{y} - \mathbf{z}, \mathbf{1} \rangle = 0$ ili $\langle \mathbf{r}, \mathbf{x} \rangle = 0$, $\langle \mathbf{r}, \mathbf{1} \rangle = 0$ ili $\mathbf{r} \perp \mathbf{x}$, $\mathbf{r} \perp \mathbf{1}$. Zapažamo da je vektor razlike $\mathbf{r}$ ortogonalan na dvodimenzioni potprostor čiju bazu čine vektori $\mathbf{x}$ i $\mathbf{1}$. To nije slučajno. Naime, vektor $\mathbf{z}$ predstavlja projekciju

vektora $\mathbf{y} \in R^n$ na navedeni potprostor. Mi se nalazimo u euklidskom prostoru R^n, skalarni proizvod $\langle \mathbf{x}, \mathbf{y} \rangle$. Bilo je: u potprostoru, naći vektor $\mathbf{z}$ koji je najbliži po kriterijumu $\min \|\mathbf{r}\|$.

Dokaz za CAUCHY–SCHWARZ nejednakost, $L \leq R$, $L = |\langle \mathbf{u}, \mathbf{v} \rangle|$, $R = \|\mathbf{u}\| \cdot \|\mathbf{v}\|$. Ako je $\mathbf{v} = 0$ onda $L = R = 0$. U daljem, $\mathbf{v} \neq 0$. Razmotrimo kvadratni trinom $p(\lambda) = a\lambda^2 + b\lambda + c = \sum_{i=1}^{n} (u_i + \lambda v_i)^2$. Iz $a > 0$ i $p(\lambda) \geq 0$ za svako $\lambda \in R \Rightarrow b^2 - 4ac \leq 0$, $L \leq R$. Dalje, ako $L = R$, $b^2 - 4ac = 0$ onda $p(\lambda_0) = 0$, $\mathbf{u} + \lambda_0 \mathbf{v} = 0$ ($\lambda_0 = -\frac{b}{2a}$). Dokaz je završen. Trivijalno: $\mathbf{u}$ i $\mathbf{v}$ kolinearni $\Rightarrow L = R$.

7.2. MODEL KOJI SE SVODI NA LINEARNI SLUČAJ

Sada ulogu prave linije $y = ax + b$ preuzima teorijski model $y = ae^{bx}$. Navedeni model se često pojavljuje u primjenama, budući da je to tzv. prirodna funkcija rasta.

Razmotrimo ulazne podatke $(x_1, y_1), \ldots, (x_n, y_n)$ kao u prethodnom naslovu. Obično se kaže da su ulazni podaci dobijeni mjerenjem. Treba naći najbolju aproksimaciju oblika $y = ae^{bx}$. Drugim riječima, treba naći $a \in R$, $b \in R$. Kako da riješimo postavljeni problem? Do rješenja se dolazi u jednom potezu, rješenje se može opisati jednom rečenicom. Naime, lako se svodi na problem razmatran u prethodnom naslovu. Upravo, dovoljno je da se primijeni logaritam na lijevu i desnu stranu relacije $y = ae^{bx}$. Tako, $\ln y = \ln a + bx$. Vidimo da treba primijeniti model prave linije na podatke $(x_i, \ln y_i)$, $i = 1, \ldots, n$ i da ćemo kao rezultat imati b i $\ln a$. Još samo ostaje da se logaritam razduži.

Primjer. Po metodi najmanjih kvadrata, naći najbolju aproksimaciju oblika $y = ae^{bx}$ (svođenjem na linearni slučaj) za podatke

x	1	2	3	4
y	1	4	10	20

. Izrada. Smjena $\ln y = Y$, podaci

x	1	2	3	4
Y	0	1,39	2,30	2,99

, pomoćni model $Y = Ax + B$, sistem $\begin{cases} 30A + 10B = 21{,}64 \\ 10A + 4B = 6{,}68 \end{cases}$ njegovo rješenje $A = 0{,}99$, $B = -0{,}80$. Odgovor: $y = 0{,}45e^{0{,}99x}$ (jer je $e^{-0{,}80} = 0{,}45$).

U zaključku, vidimo da smo riješili model $y = ae^{bx}$ time što smo izvršili njegovu linearizaciju. Slično se i neki drugi modeli mogu redukovati na linearni slučaj.

7.3. METODA INVERZNIH DISTANCI

Na redu je jedna metoda koja služi za aproksimaciju funkcije od dvije promjenljive $z = z(x, y)$. Koordinate tačke u ravni označavamo kao (x, y), a zavisno promjenljivu kao z. Ukratko, mjerenjem smo saznali vrijednosti zavisno promjenljive u n tačaka. Na osnovu raspoloživih podataka, treba naći dobru procjenu za $z(x_0, y_0)$, gdje je (x_0, y_0) neka tačka u ravni.

Uvedimo potrebne oznake. Razmotrimo funkciju od dvije promjenljive $z = z(x, y)$. Neka je $n \geq 1$. Neka su $(x_1, y_1), \ldots, (x_n, y_n)$ tačke u kojima raspolažemo sa $z_i = z(x_i, y_i)$. Sve zajedno čini set ulaznih podataka, što je prikazano na slici 3.

Prelazimo na postavku problema. Fiksirajmo za trenutak jednu tačku u ravni (x_0, y_0). Naravno da $z(x_0, y_0)$ označava vrijednost razmatrane funkcije u toj tački. Pisaćemo kraće $z_0 = z(x_0, y_0)$. Tačna vrijednost z_0 je van domašaja. Treba izmisliti metodu da se dođe do njene dobre aproksimacije. Tu aproksimaciju (biće izračunata) označavaćemo kao $\hat{z}_0$, č. kapa ili engl. hat. Možemo pisati neformalno $z_0 = \hat{z}_0 + \varepsilon$.

Tačke $(x_1, y_1) \ldots, (x_n, y_n)$ možemo da označimo kao $P_1, \ldots, P_n$. Na slici će one biti označene kao $P, Q, \ldots$. Za (x_0, y_0) oznaka P_0. Na slici će za nju biti oznaka X. Uvedimo oznaku d_i za rastojanje između P_0 i P_i. Znači, neka je $d_i = \sqrt{(x_i - x_0)^2 + (y_i - y_0)^2}$ za $i = 1, \ldots, n$.

Kako da dobijemo $\hat{z}_0$? Mi ćemo uzeti jednu linearnu kombinaciju datih vrijednosti funkcije $c_1 z_1 + \ldots + c_n z_n$. Lako se razumije da linearna kombinacija treba da bude konveksna. To znači da treba da bude ispunjeno $0 \leq c_i \leq 1$ i $c_1 + \ldots + c_n = 1$. Što je c_i veći to i–ta tačka više utiče na formiranje rezultata $\hat{z}_0$. Drugim riječima, značaj takvih tačaka, tj. njihov težinski faktor biće veći od onog udaljenih tačaka.

Logično, tačka (x_i, y_i) treba da ima veći značaj ukoliko je bliža tački (x_0, y_0) o kojoj se radi. Drugim riječima, značaj i–te tačke obrnuto je srazmjeran njenom odstojanju d_i. U numeričkoj metodi, uzima se da je koeficijent c_i proporcionalan recipročnoj vrijednosti rastojanja d_i. Automatski, $c_i = (1/d_i)/(\sum_{j=1}^{n}(1/d_j))$. Time je numerička metoda konstruisana. Možemo napisati definitivno

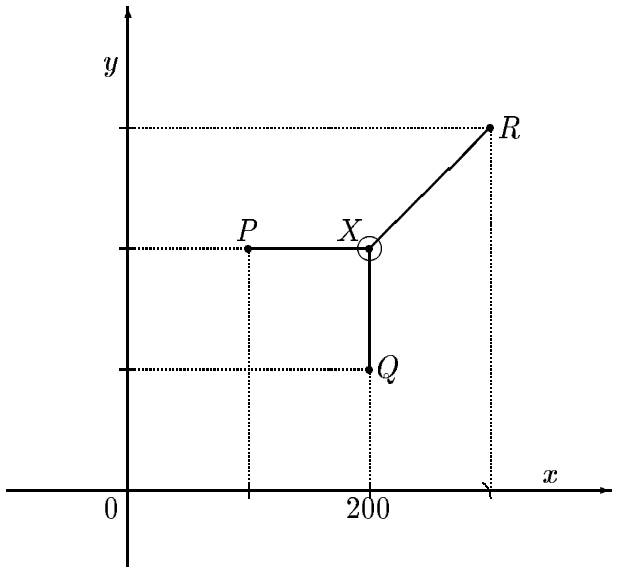
$$\hat{z}_0 = \sum_{i=1}^{n} c_i z_i, \qquad c_i = \frac{1/d_i}{1/d_1 + \ldots + 1/d_n}.$$

Primjer. Dato je $(x_1, y_1) = (100, 200)$, $(x_2, y_2) = (200, 100)$, $(x_3, y_3) = (300, 300)$, $(x_0, y_0) = (200, 200)$, $z_1 = 10$, $z_2 = 11$, $z_3 = 15$. V. sliku 4.

Izrada. Mi računamo $d_1 = 100$, $d_2 = 100$, $d_3 = 141$, $1/d_1 = 0{,}01$, $1/d_2 = 0{,}01$, $1/d_3 = 0{,}007$, $1/d_1 + 1/d_2 + 1/d_3 = 0{,}027$, $c_1 = 1/d_1/0{,}027 = 0{,}37$, $c_2 = 1/d_2/0{,}027 = 0{,}37$, $c_3 = 1/d_3/0{,}027 = 0{,}26$. Tako da je $\hat{z}_0 = 0{,}37 z_1 + 0{,}37 z_2 + 0{,}26 z_3$ i nastavljamo da računamo $\hat{z}_0 = 0{,}37 \cdot 10 + 0{,}37 \cdot 11 + 0{,}26 \cdot 15$. Rezultat glasi $\hat{z}_0 = 11{,}67$.

i	(x_i, y_i)	z_i
1	(x_1, y_1)	z_1
2	(x_2, y_2)	z_2
...	...	...
...	...	...
...	...	...
n	(x_n, y_n)	z_n

Slika 3: Ulazni podaci (zatim $z(x_0, y_0) = ?$)



Slika 4: Rastojanja $d_1 = d_2 = 100$, $d_3 = 141{,}4$

7.4. METODA KVADRATNIH INVERZNIH DISTANCI

U slučaju metode kvadratnih inverznih distanci stavlja se $c_i = (1/d_i^2)/(1/d_1^2 + \ldots + 1/d_n^2)$ za $1 \leq i \leq n$, što predstavlja jedinu razliku. Ostalo sve isto kao kod metode inverznih distanci.

7.5. JEDNOSTAVNI KRIGING (SIMPLE KRIGING)

Kriging je tehnika interpolacije čiji je autor D. Krige. Za razmatranu tehniku (za razmatranu metodu) se kaže i engl. Gaussian process regression. Takođe se kaže i Wiener–Kolmogorov predikcija.

Neka je u (x,y) ravni data zavisna veličina $z = z(x,y)$. Zavisna veličina može da bude elevacija terena ili koncentracija zlata. Neka je u $N \geq 1$ tačaka (x,y) izmjerena veličina z. Označimo sa m srednju vrijednost veličine z po svih N tačaka. S druge strane, razmotrimo tačku P_0 čije su koordinate (x_0, y_0) a koja nije čvor. Treba naći dobru procjenu, u oznaci $\hat{z}_0$, za veličinu $z_0 = z(x_0, y_0)$. Od ranijih N tačaka, izaberimo $n \geq 1$ tačaka i to one najbliže P_0. Označimo ih kao P_i, a njihove koordinate kao (x_i, y_i), gdje je $1 \leq i \leq n$. Takođe, $z_i = z(x_i, y_i)$.

Znamo da se rastojanje dvije tačke (x_1, y_1) i (x_2, y_2) u ravni izražava sa

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}.$$

Neka je d_{ij} rastojanje između P_i i P_j, gdje je $1 \leq i \leq n$, $1 \leq j \leq n$. Takođe, d_{i0} neka označava rastojanje između P_i i P_0, gdje je $1 \leq i \leq n$. Prirodno, stepen korelacije vrijednosti z u dvije tačke obrnuto je srazmjeran njihovom rastojanju. Treba definisati funkciju da odražava inverznu proporciju. Razmotrimo funkciju koja broju $h \geq 0$ pridružuje broj $1 - \Gamma(h) = 1 - 1{,}5h + 0{,}5h^3$, a mogući su i neki drugi oblici. Za Γ se kaže da je variogram. Od $h = 0$ do $h = 1$ funkcija opada od vrijednosti 1 do vrijednosti 0. Dodefinišimo funkciju da je $= 0$ za $h > 1$. Ako h posmatramo kao rastojanje u metrima onda se obično uzima funkcija $1 - 1{,}5(h/5000) + 0{,}5(h/5000)^3$ ili slično. Najzad, stavimo

$$\gamma(h) = 0{,}8\Big(1 - 1{,}5(h/5000) + 0{,}5(h/5000)^3\Big) \text{ za } 0 \leq h \leq 5000 \text{ i } \gamma(h) = 0 \text{ za } h \geq 5000.$$

Koeficijent 0,8 (ili sličan broj) ima određeni fizički smisao.

Stavimo $a_{ij} = \gamma(d_{ij})$, $1 \leq i \leq n$, $1 \leq j \leq n$, izražava mjeru korelacije među čvorovima. Stavimo $b_i = \gamma(d_{i0})$, $i = 1, \ldots, n$, uzajamna povezanost čvora P_i i tačke P_0 za koju se aproksimacija vrši. Neka je A matrica, $A = [a_{ij}]_{i,j=1}^n$. Vidimo da je matrica A simetrična i da je $a_{ii} = 0{,}8$. Neka je $\mathbf{b}$ vektor kolona, $\mathbf{b} = [b_i]_{i=1}^n$. Neka je $\mathbf{c}$ vektor kolona, $\mathbf{c} = [c_i]_{i=1}^n$, vektor nepoznatih. Napišimo standardni sistem linearnih jednačina

$$A\mathbf{c} = \mathbf{b}.$$

Veličina c_i pokazuje udio (udio uticaja) broja z_i na formiranje broja $\hat{z}_0$, tj. broja koji se traži (procjena za z_0). Tačnije, procjena se formira po formuli

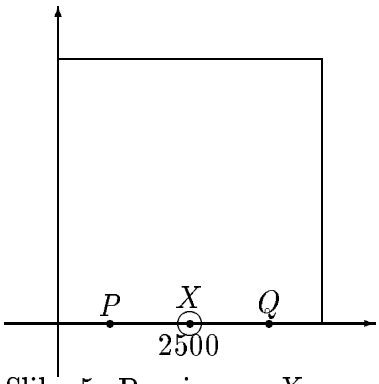
$$\hat{z}_0 = m + c_1(z_1 - m) + \ldots + c_n(z_n - m).$$

Može se izvršiti procjena za veliki broj tačaka tipa tačke P_0 i onda se svi rezultati mogu prikazati u obliku mape. Iz mape se sa dobrom preciznošću može vidjeti gdje veličina z dostiže maksimum. Kriging predstavlja veoma moćno i korisno sredstvo za interpolaciju (aproksimaciju). Na primjer, ako su dva čvora (od n čvorova) blizu onda to neće narušiti kvalitet numeričkog odgovora.

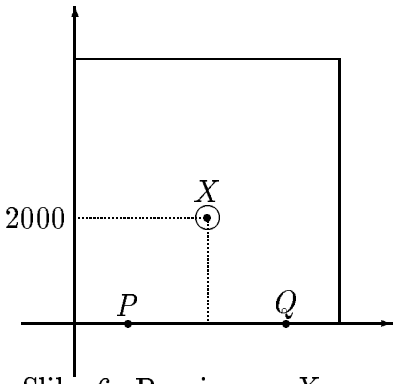
Primjer. Neka je $\gamma(h)$ kao u tekstu. Vrijednost z poznata je u tačkama P_1 i P_2, a nije poznata u tački P_0, traži se procjena. V. sliku 5. Na slici su tačke označene kao P, Q, X. Ulazni podaci: $P_1(1000, 0)$, $P_2(4000, 0)$, $P_0(2500, 0)$, $z_1 = 800$, $z_2 = 900$, $m = 1000$. Među–rezultati: sistem linearnih jednačina $0{,}8c_1 + 0{,}17c_2 = 0{,}45$, $0{,}17c_1 + 0{,}8c_2 = 0{,}45$, njegovo rješenje $c_1 = c_2 = 0{,}46$. Rezultat: $\hat{z}_0 = m + c_1(z_1 - m) + c_2(z_2 - m) = 860$.

Nastavak primjera. Samo je promijenjeno $P_0(2500, 2000)$. V. sliku 6. Sada sistem linearnih jednačina glasi $0{,}8c_1 + 0{,}17c_2 = 0{,}25$, $0{,}17c_1 + 0{,}8c_2 = 0{,}25$, a njegovo rješenje $c_1 = c_2 = 0{,}26$. Rezultat glasi $\hat{z}_0 = 920$.

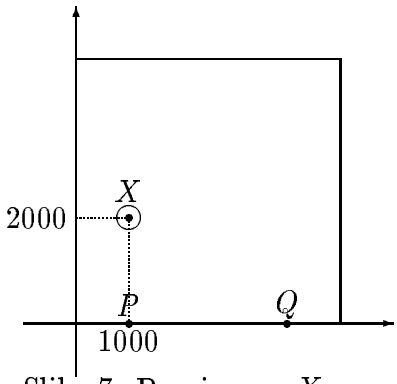
Nastavak primjera. Opet je sve ostalo isto, a jedino je izmijenjena pozicija tačke P_0. Neka je sada $P_0(1000, 2000)$. V. sliku 7. Imamo sistem linearnih jednačina $0{,}8c_1 + 0{,}17c_2 = 0{,}34$, $0{,}17c_1 + 0{,}8c_2 = 0{,}12$, a rješenje $c_1 = 0{,}4$, $c_2 = 0{,}06$. Sada odgovor glasi $\hat{z}_0 = 910$.



Slika 5: Procjena u X



Slika 6: Procjena u X



Slika 7: Procjena u X

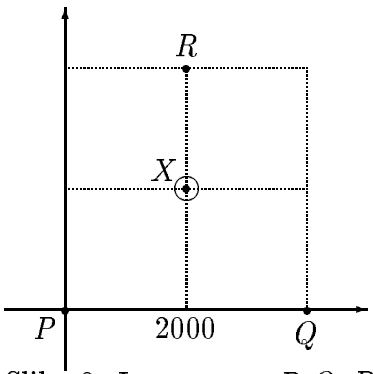
Primjer. Neka je $\gamma(h)$ kao u tekstu. V. sliku 8. Neka je dato

$$P_1(0,0), \quad P_2(4000,0), \quad P_3(2000,4000), \quad P_0(2000,2000).$$

Sistem linearnih jednačina glasi

$$\begin{cases} 0{,}8c_1 + 0{,}05c_2 + 0{,}01c_3 = 0{,}18 \\ 0{,}05c_1 + 0{,}8c_2 + 0{,}01c_3 = 0{,}18 \\ 0{,}01c_1 + 0{,}01c_2 + 0{,}8c_3 = 0{,}34 \end{cases}$$

Njegovo rješenje je $c_1 = 0{,}18$, $c_2 = 0{,}18$, $c_3 = 0{,}43$. Znamo da se aproksimacija vrši po formuli $\hat{z}_0 = m + c_1(z_1 - m) + c_2(z_2 - m) + c_3(z_3 - m)$.



Slika 8: Imamo z u P, Q, R

7.6. OBIČNI KRIGING (ORDINARY KRIGING)

Predstavlja usavršeni oblik jednostavnog kriginga (i služi za rješavanje istog zadatka), pa navedimo samo u čemu se sastoje razlike. U računu više ne učestvuje m, a sada važi relacija $c_1 + \ldots + c_n = 1$. Sada matrica A ima oblik $(n+1) \times (n+1)$, a vektori $\mathbf{b}$ i $\mathbf{c}$ imaju dužinu $n+1$.

Veličine a_{ij} $(1 \leq i \leq n,\ 1 \leq j \leq n)$ poklapaju se sa onima kod jednostavnog kriginga. Isto tako, brojevi b_i $(1 \leq i \leq n)$ poklapaju se sa onima u slučaju jednostavnog kriginga. Stavlja se $a_{i,n+1} = 1$ $(1 \leq i \leq n)$, $a_{n+1,j} = 1$ $(1 \leq j \leq n)$, $a_{n+1,n+1} = 0$, $b_{n+1} = 1$. Tako da matrica A ostaje simetrična. Treba riješiti sistem linearnih jednačina $A\mathbf{c} = \mathbf{b}$. Dobićemo $c_1, \ldots, c_{n+1}$. Rezultat glasi, odnosno za aproksimaciju služi veličina $\hat{z}_0 = c_1 z_1 + \ldots + c_n z_n$.

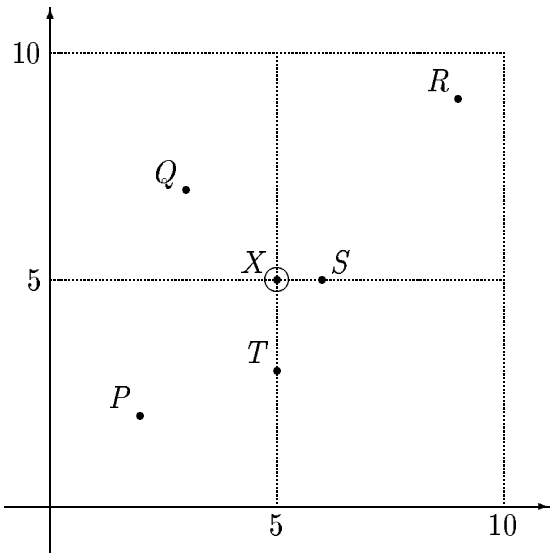
Primjer za obični kriging iz knjige Burrough, p. 139. Izabrana je funkcija $\gamma = \gamma(h)$ na odgovarajući način. Ulazni podaci prikazani su u tabeli $(n = 5)$:

i	1	2	3	4	5
x_i	2	3	9	6	5
y_i	2	7	9	5	3
z_i	3	4	2	4	6

.

Vrši se aproksimacija za tačku $P_0(x_0, y_0)$, gdje je $x_0 = 5$, $y_0 = 5$. Ulazni podaci prikazani su i na slici 9, barem što se tiče (x, y).

Među–rezultate izostavljamo, jer su obimni. Tokom računanja, rješava se jedan sistem linearnih jednačina oblika 6×6. Kao rezultat imamo sljedeće brojeve: $c_1 = 0{,}02$, $c_2 = 0{,}23$, $c_3 = -0{,}09$, $c_4 = 0{,}64$, $c_5 = 0{,}20$. Tako da odgovor glasi $\hat{z}_0 = c_1 z_1 + c_2 z_2 + c_3 z_3 + c_4 z_4 + c_5 z_5 = 0{,}02 \cdot 3 + 0{,}23 \cdot 4 - 0{,}09 \cdot 2 + 0{,}64 \cdot 4 + 0{,}20 \cdot 6 = 4{,}56$.



Slika 9: Raspored tačaka $P_1, \ldots, P_5$ i P_0, tj. $P, \ldots, T$ i X u (x, y) ravni.

NUMERIČKA MATEMATIKA
Sadržaj predavanja

Fajlovi (gsview): matea.tex 10 pages, mateb.tex 4 pages, matec.tex 6 pages, mated.tex 8 pages, matee.tex 5 pages, matef.tex 11 pages, mateg.tex 7 pages. $\sum = 7$ files & 51 pages