Premična pika

$$x = (-1)^{S} m \cdot b^{e}$$

$$m = c_{0} + c_{1}b^{-1} + c_{2}b^{-2} + \dots + c_{t}b^{-t}$$

S ... predznak

b ... baza, ponavadi 2

m ... vrednost mantise

t ... dolžina mantise

e ... vrednost eksponenta $L \le e \le U$

 c_i ... števke v mejah $0 < c_i < b-1$

Sistem premične pike označimo z P(b, t, L, U).

Standard IEEE

Eksponent je zapisan z odmikom:

$$E = e + \text{odmik}$$

Če je E=0, uporabimo **denormiran zapis**:

$$x = (-1)^S (c_1 b^{-1} + c_2 b^{-2} + \dots + c_t b^{-t}) \cdot b^{e+1}$$

Sicer pa **normiran zapis**:

$$x = (-1)^{S} (1 + c_1 b^{-1} + c_2 b^{-2} + \dots + c_t b^{-t}) \cdot b^{e}$$

Če so vsi biti eksponenta 1 in vsi biti mantise 0, je $x = (-1)^S \infty$.

Če so vis biti eksponenta 1 in vsi biti mantise niso 0, je x = NaN.

• Single precision b = 2, t = 23, L = -126, U = 127, odmik: 127

preazitak i exportent o mantisa 25	predznak 1	exponent 8	mantisa 23
--	------------	------------	------------

• Double precision b = 2, t = 52, L = -1022, U = 1023, odmik: 1023

predznak 1	exponent 11	mantisa 52

Zaokroževanje

Naj bo x pozitivno število z neskončnim zapisom $x = (c_1b^{-1} + c_2b^{-2} + \cdots + c_tb^{-t} + c_{t+1}b^{-t-1}) + \cdots b^e$

Kandidata za približek fl(x) sta:

$$x_{-} = (c_1b^{-1} + c_2b^{-2} + \dots + c_tb^{-t})b^e$$

$$x_{+} = (c_1b^{-1} + c_2b^{-2} + \dots + c_tb^{-t} + b^{-t})b^e$$

Vzamemo tistega, ki je bljižje. Če sta enako blizu, izberemo tistega, ki ima zadnjo števko sodo.

Osnovna zaokrožitvena napaka

$$u = \frac{1}{2}b^{-t}$$

$$fl(x) = x(1+\delta) \quad \text{za} \quad |\delta| \le u$$

$$\frac{|fl(x) - x|}{|x|} \le u$$

Napake pri numeričnem računanju

• Neodstranljiva napaka Namesto x imamo približek \bar{x} .

$$D_n = f(x) - f(\bar{x})$$

• Napaka metode Namesto funkcije f imamo približek a.

$$D_m = f(\bar{x}) - g(\bar{x})$$

• Zaokrožitvena napaka Pri računanju $\tilde{y} = f(\bar{x})$ se pri vsaki operaciji se pojavi zaokrožitvena napaka. Namesto \tilde{y} dobimo \hat{y} .

$$D_z = \tilde{y} - \hat{y}$$

Celotna napaka je $D = |D_n| + |D_m| + |D_z|$.

Stopnja občutljivosti

Razmerje velikosti spremembe podatkov in spremembe rezultata.

Naj bo $f: \mathbb{R} \to \mathbb{R}$ zvezno odvedlijva funkcija in δx majhna motnja.

• Absolutna občutljivost f v točki x:

$$|f(x + \delta x) - f(x)| \approx |f'(x)| \cdot |\delta x|$$

• Relativna občutliivost f v točki x:

$$\frac{|f(x+\delta x) - f(x)|}{|f(x)|} \approx \frac{|f'(x)| \cdot |\delta x|}{|f(x)|}$$

Obratna in direktna stabilnost

- Direktna stabilnost: za vsak x direktna napaka $(|f(x) f(x + \Delta x)|)$ majhna (absolutno oz. relativno).
- Obratna stabilnost: za vsak x razlika Δx , ki bi nam dala pravi rezultat majhna.

|direktna napaka| ≤ občutlijvost · |obratna napaka|

Nelinearne enačbe

Iščemo ničle funkcije f.

• Enostavne ničle:

$$f(\alpha) = 0$$
 in $f'(\alpha) \neq 0$

• m-kratne ničle:

$$f(\alpha) = f'(\alpha) = \dots = f^{(m-1)}(\alpha) = 0$$

Občutljivost ničle

Naj bo α *m*-kratna ničla $\hat{\alpha}$ približek, da je $f(\hat{\alpha}) = \varepsilon$.

Če f razvijemo v Taylorjevo vrsto okoli α in vzamemo prvih m+1 členov dobimo:

$$\varepsilon \doteq \frac{f^{(m)}(\alpha)}{m!} (\hat{\alpha} - \alpha)^m \quad |\hat{\alpha} - \alpha| \doteq \sqrt[m]{\frac{\varepsilon \cdot m!}{|f^{(m)}(\alpha)|}}$$

Bisekcija

 $\textit{vhod}\colon$ funkcija $f:[a,b]\to\mathbb{R},\,f(a)f(b)<0,$ natancnost ε $\textit{izhod}\colon$ nicla funkcije f

$$\begin{aligned} & \textit{dokler} \quad |b-a| > \varepsilon \colon \\ & c \leftarrow \frac{a+b}{2} \\ & c \quad \text{sign}(f(c)) = \text{sign}(f(a)) \colon \\ & a \leftarrow c \\ & \textit{sicer} \colon \\ & b \leftarrow c \end{aligned}$$

c je približek ničle α . Velja

$$|\alpha - c| \le \frac{b - a}{2^m} \le \varepsilon$$

Za natančnost ε potrebujemo $\log_2\left(\frac{b-a}{\varepsilon}\right)$ korakov.

Navadna iteracija

Rešujemo f(x) = 0. Enačbo pretvorimo vx = g(x). Načinov je veliko:

- g(x) = f(x) + x
- g(x) = cf(x) + x
- g(x) = h(x)f(x) + x kjer je h(x) funkcija, ki nima ničle v α .

Izrek o konvergenci navadne iteracije

Naj bo α negibna točka za g in naj g na intervalu $[\alpha - d, \alpha + d]$ (d > 0) zadošča Lipschitzovemu pogoju:

$$\exists m \in [0,1) \ \forall x, y \in I : |g(x) - g(y)| \le m|x - y|$$

tedaj je g skrčitev na I.

Potem za vsak $x_0 \in I$ zaporedje $x_{r+1} = g(x_r)$ konvergira k α in velja:

$$|x_r - \alpha| \le \frac{m}{1 - m} |x_r - x_{r-1}|$$

Posledica: Če je g zvezno odvedljiva v α in velja $g'(\alpha) < 1$, obstaja interval I, ki vsebuje α , da za vsak $x_0 \in I$ zaporedje konvergira k α .

- Če je $|q'(\alpha)| < 1$, je α privlačna negibna točka
- Če je $|g'(\alpha)| > 1$, je α odbojna negibna točka

Če je α odbojna za g, je privlačna za g^{-1} : $g(x) = x \implies x = g^{-1}(x)$

Hitrost konvergence

p>0 je red konvergence, če $\exists C_1,C_2>0$, da za vse dovolj pozne člene zaporedja $x_{r+1}=g(x_r)$ velja:

$$C_1|x_r - \alpha|^p \le |x_{r+1} - \alpha| \le C_2|x_r - \alpha|^p$$

 $Vsak\ korak\ se\ \check{s}t.\ decimalk\ pomno\check{z}i\ s\ p.$

Naj bo g v okolici α p-krat zvezno odvedlijva in naj velja $g(\alpha) = \alpha$, $g'(\alpha) = \cdots = g^{(p-1)}(\alpha) = 0$ in $g^{(p)}(\alpha) \neq 0$. Tedaj je red konvergence enak p.

Standardni redi konvergence

p=1 ... linearna konvergenca p=2 ... kvadratična konvergenca

Tangentna (Newtonova) metoda

$$x_{r+1} = x_r - \frac{f(x_r)}{r}$$

Red konvergence:

- 2. če je α enkratna ničla $f'(\alpha) \neq 0$
- 3, če je $f'(\alpha) \neq 0$ in $f''(\alpha) = 0$
- 1, če je α večkratna ničla $f'(\alpha) = 0$

Sekantna metoda

$$x_{r+1} = x_r - \frac{f(x_r)}{\frac{f(x_r) - f(x_{r-1})}{x_r - x_{r-1}}}$$

Red konvergence: $\frac{\sqrt{5}+1}{2} \approx 1.62$

Metoda (f, f', f'')

$$x_{r+1} = x_r - \frac{f(x_r)}{f'(x_r)} - \frac{f''(x_r)f^2(x_r)}{2f'^3(x_r)}$$

Red konvergence: 3 (pri predpostavkah)

Müllerjeva metoda

Na x_r, x_{r-1}, x_{r-2} napnemo parabolo, ničla parabole je naslednji približek.

$$p(x) = a(x - x_r)^2 + b(x - x_r) + c$$

 $x_{r+1} = x_r - \frac{2c}{b + \text{sign}(b)\sqrt{b^2 - 4ac}}$

Hallejeva metoda

$$x_{r+1} = x_r - \frac{2f(x_r)f'(x_r)}{2f'(x_r)^2 - f(x)f''(x)}$$

Iskanje ničel polinoma

$$p_n(x) = a_n x^n + a_{n-1} x^{n-1} + \dots + a_0, \quad a_n \neq 0$$

Durand-Kernerjeva metoda

Naj bodo z_1, \ldots, z_n približki za $\alpha_1, \ldots, \alpha_n$.

$$z_i^{(r+1)} = z_i^{(r)} - \frac{p(z_i^{(r)})}{\prod\limits_{\substack{k=1\\k\neq i}}^{n}(z_i^{(r)} - z_k^{(r)})}$$

Metoda pridužene matrike

$$C_{p_n} = \begin{bmatrix} 0 & 1 & \dots & 0 & 0 \\ & 0 & & & & \\ \vdots & & \ddots & & \vdots \\ & & 1 & 0 \\ & & & 1 & 0 \\ & & & 0 & 1 \\ -\frac{a_0}{a_n} & -\frac{a_1}{a_n} & \dots & -\frac{a_{n-2}}{a_n} -\frac{a_{n-1}}{a_n} \end{bmatrix}$$

Lastne vrednosti (ničle $\det(C - \lambda I)$) matrike C_{p_n} so ravno ničle polinoma p_n .

Vektorske in matrične norme

 $\|\cdot\|:\mathbb{R}^n\to\mathbb{R}_+$ je vektorska norma, če

- $||x|| \ge 0$ in $||x|| = 0 \iff x = 0$
- $\|\alpha x\| = |\alpha| \|x\|, \ \alpha \in \mathbb{R}$
- $||x + y|| \le ||x|| + ||y||$

$$||x||_p = \sqrt[p]{|x_1|^p + \dots + |x_n|^p}$$
 ... p-norma $||x||_\infty = \max_{1 \le i \le n} |x_i|$... max norma

Cauchy-Schwartzova neenakost: $|x^Ty| \le ||x||_2 ||y||_2$

 $\|\cdot\|:\mathbb{R}^{m\times n}\to\mathbb{R}_+$ je matrična norma, če

- $||A|| \ge 0$ in $||A|| = 0 \iff A = 0$
- $\|\alpha A\| = |\alpha| \|A\|, \ \alpha \in \mathbb{R}$
- $||A + B|| \le ||A|| + ||B||$
- ||AB|| < ||A|| ||B|| submultiplikativnost

$$\begin{split} \|A\|_1 &= \max_{x \neq 0} \frac{\|Ax\|_1}{\|x\|_1} = \max_{1 \leq j \leq n} \left(\sum_{k=1}^m |a_{kj}| \right) \\ \|A\|_2 &= \max_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \sqrt{\rho(A^T A)} \\ \|A\|_{\infty} &= \max_{x \neq 0} \frac{\|Ax\|_{\infty}}{\|x\|_{\infty}} = \max_{1 \leq i \leq m} \left(\sum_{k=1}^n |a_{ik}| \right) \\ \|A\|_F &= \sqrt{\sum_{i,j} a_{ij}^2} \\ \|A\|_{op} &= \max_{\|c\|_v = 1} \|Ax\|_v \qquad \|\cdot\|_v \ vekt. \ norma \end{split}$$

spektralni radij $\rho(A) = \max\{|\lambda_1|, \dots, |\lambda_n|\}$, kjer so λ lastne vrednosti matrike A.

Matrična norma $\|\cdot\|_m$ je **usklajena** z vektorsko normo $\|\cdot\|_v$, če je

$$\|Ax\|_v \leq \|A\|_m \|x\|_v$$

Na končno dimenzijskih prostorih so vse norme ekvivalentne

$$\|\cdot\|_{\alpha}, \|\cdot\|_{\beta} \implies C_1 \|A\|_{\beta} \le \|A\|_{\alpha} \le C_2 \|A\|_{\beta}$$

Za vsako operatorsko normo nad kvadratnimi matrikami velja:

$$|\lambda| \le ||A||$$
 λ lastna vrednost matrike A

Reševanje sistemov linearnih enačb

Imamo m enačb in n neznank:

- m < n: **nedoločen**, ∞ rešitev
- m = n: kvadraten, ponavadi 1 rešitev
- m > n: **predoločen**, ponavadi 0 rešitev

$$Ax = b$$
 $A \in \mathbb{R}^{m \times n}$ $b \in \mathbb{R}^m$ $x \in \mathbb{R}^n$

$$\operatorname{Im} A = \{Ax : x \in \mathbb{R}^n\} = \operatorname{Lin}\{a_1, \dots, a_n\}$$
$$\operatorname{Ker} A = \{x \in \mathbb{R}^n : Ax = 0\}$$
$$\operatorname{rang} A = n - \operatorname{dim}(\operatorname{Ker} A)$$

Če je m = n, so naslednje izjave ekvivalentne:

- $\exists A^{-1}: A^{-1}A = I = A^{-1}A$
- $det(A) \neq 0$
- $\operatorname{rang} A = n$
- $Ker A = \{0\}$

Lastnosti matrike A:

$$A = A^T$$
 ... simetrična
$$A = A^H$$
 ... hermitska
$$A^T A = I$$
 ... ortogonalna

LU razcep

Matriko A zapišemo kot produkt LU, kjer je L spodnje trikotna z 1 na diagonali, U pa zgornje trikotna.

Sistem enačb Ax = b potem rešujemo kot:

$$L(Ux) = b \implies Ly = b \implies Ux = y$$

Elementarne eliminacije

Matriko A množimo z elementarnimi matrikami L_k

$$A^{(k)} = L_{k-1} \dots L_1 A$$

$$L_k = I - l_k e_k^T \qquad L_k^{-1} = I + l_k e_k^T$$

$$l_k = \begin{bmatrix} 0 & \dots & 0 & \frac{a_{k+1,k}^{(k)}}{a_{k,k}^{(k)}} & \dots & \frac{a_{n,k}^{(k)}}{a_{k,k}^{(k)}} \end{bmatrix}^T$$

$$e_k = \begin{bmatrix} 0 & \dots & 0 & 1 & 0 & \dots & 0 \end{bmatrix}^T$$

$$L = L_1^{-1} \dots L_{n-1}^{-1}$$

$$U = L_{n-1} \dots L_1 A$$

Za nesingularno matriko $A \in \mathbb{K} \times \mathbb{K}$ obstaja enolični LU razcep \iff vse vodilne podmatrike $(A_k = A(1:k,1:k))$ nesingularne.

LU razcep z delnim pivotiranjem

Preden dolčimo L_k zamenjamo k-to in $\underset{i-k}{\operatorname{argmax}}|a_{k,i}^{(k)}|$ -to vrstico matrike A^k .

Občutlivost sistemov linearnih enačb

$$\kappa(A) = ||A^{-1}|| ||A||$$
 število občutlivosti

$$(A+\delta A)(x+\delta x)=(b+\delta b)$$

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

Približno velja $\kappa(A)=10^e \implies$ relativna napaka rešitve reda 10^{e-16}

Razcep Choleskega

Matrika A je simetrična pozitivno definitna (SPD), če je $A = A^T$ in $\forall x \neq 0 : x^T(Ax) > 0$.

Lastnosti SPD matrik:

- vse vodilne podmatrike (A(1:k,1:k)) so SPD
- $A([i_1,\ldots,i_k],[i_1,\ldots,i_k])$, kjer je 1 $\leq i_1 < \cdots < i_k \leq n$ je SPD
- A je SPD \iff vse lastne vrednosti pozitivne
- Naj bo C obrnljiva, potem je CAC^T SPD
- Če je A SPD, je $A_{ii} > 0$ in $\max_{i,j} |a_{ij}| = a_{kk}$

Matrika $A \in \mathbb{R}^{n \times n}$, det $A \neq 0$ je SPD \iff obstaja nesingularna spodnje trikotna matrika V s pozitivnimi diagonalnimi elementi, da je $A = VV^T$

Algoritem za razcep Cholskega

Prvi stolpec matrike V izračunamo:

$$\begin{bmatrix} a_{11} \\ \vdots \\ a_{n1} \end{bmatrix} = v_{11} \begin{bmatrix} v_{11} \\ \vdots \\ v_{n1} \end{bmatrix}$$

j-ti stolpec izračunamo:

$$\begin{bmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{nj} \end{bmatrix} = v_{j1} \begin{bmatrix} v_{11} \\ v_{21} \\ \vdots \\ v_{n1} \end{bmatrix} + v_{j2} \begin{bmatrix} 0 \\ v_{22} \\ \vdots \\ v_{n2} \end{bmatrix} + \dots + v_{jj} \begin{bmatrix} 0 \\ \vdots \\ v_{jj} \\ \vdots \\ v_{nj} \end{bmatrix}$$

$$F : \mathbb{R}^n \to \mathbb{R}^n, \ F = (f_1, \dots, f_n)^T, \ x = (x_1, \dots, x_n)^T \text{ rešujemo } F(x) = 0$$

$$F(x) = 0 \implies G(x) = x$$

$$x^{(r+1)} = G(x^r)$$

$$\begin{split} &\textit{vhod} : \text{ matrika } A \text{ z elementi } a_{ij} \\ &\textit{izhod} : \text{ matrika } V \text{ z elementi } v_{ij} \\ &\textit{za } \textit{vsak } j = 1 \dots n \text{:} \\ &v_{jj} \leftarrow \sqrt{a_{jj} - \sum\limits_{k=1}^{j-1} v_{jk}^2} \\ &\textit{za } \textit{vsak } i = j+1 \dots n \text{:} \\ &v_{ij} \leftarrow \frac{1}{v_{jj}} \left(a_{ij} - \sum\limits_{k=1}^{j-1} v_{ik} v_{jk}\right) \end{split}$$

Reševanje sistemov nelinearnih enačb

$$F: \mathbb{R}^n \to \mathbb{R}^n, F = (f_1, \dots, f_n)^T, x = (x_1, \dots, x_n)^T$$
 rešujemo $F(x) = 0$

$$F(x) = 0 \implies G(x) = x$$

$$x^{(r+1)} = G(x^r)$$

Jakobijeva matrika

$$J_F = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \dots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \dots & \frac{\partial f_n}{\partial x_n} \end{bmatrix}$$

Če obstaja odprta množica $\Omega \subseteq \mathbb{R}^n$ z lastnostmi:

•
$$x \in \Omega \implies G(x) \in \Omega$$

•
$$x \in \Omega \implies \rho(J_G(x)) \le q < 1$$

potem ima G na Ω natanko eno negibno točko α , ki jo dobimo kot:

$$\lim_{r \to \infty} x^{(r)} \quad \text{za poljuben } x^{(0)} \in \Omega$$

Newtnova metoda na več dimenzijah

$$G(x) = x - J_F^{-1}(x)F(x)$$

Broydenova metoda

Spremembo Δx izračunamo iz enačbe:

$$B_r \Delta x^{(r)} = -F(x^{(r)})$$

$$x^{(r+1)} = x^{(r)} + \Delta x^{(x)}$$

$$B_{r+1} = B_r + \frac{F(x^{(r)})(\Delta x^{(r)})^T}{(\Delta x^{(r)})^T \Delta x^{(x)}}$$

Za B_0 vzamemo $J_F(x^{(0)})$ ali pa kar I.