Premična pika

$$x = (-1)^S m \cdot b^e$$
$$m = c_0 + c_1 b^{-1} + c_2 b^{-2} + \cdots + c_t b^{-t}$$

S	...	predznak
b	...	baza, ponavadi 2
m	...	vrednost mantise
t	...	dolžina mantise
e	...	vrednost eksponenta $L \leq e \leq U$
c_i	...	števke v mejah $0 \leq c_i \leq b-1$

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**:

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

Sicer pa **normiran zapis**:

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

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 $x = \text{NaN}$.

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

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

$$x = (c_1 b^{-1} + c_2 b^{-2} + \cdots + c_t b^{-t} + c_{t+1} b^{-t-1}) + \ldots b^e$$

Kandidata za približek $\text{fl}(x)$ sta:

$$x_- = (c_1 b^{-1} + c_2 b^{-2} + \cdots + c_t b^{-t}) b^e$$
$$x_+ = (c_1 b^{-1} + c_2 b^{-2} + \cdots + c_t b^{-t} + b^{-t}) b^e$$

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

Osnovna zaokrožitvena napaka

$$u = \frac{1}{2} b^{-t}$$
$$\text{fl}(x) = x(1+\delta) \quad \text{za} \quad |\delta| \leq u$$
$$\frac{|\text{fl}(x) - x|}{|x|} \leq u$$

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 g.

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

- **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 odvedljiva funkcija in δx majhna motnja.

- **Absolutna občutljivost** f v točki x:

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

- **Relativna občutljivost** 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.

$$|\text{direktna napaka}| \lesssim \text{občutljivost} \cdot |\text{obratna napaka}|$$

Nelinearne enačbe

Iščemo ničle funkcije f.

- **Enostavne ničle**:

$$f(\alpha) = 0 \quad \text{in} \quad f'(\alpha) \neq 0$$

- **m-kratne ničle**:

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

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:

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

Bisekcija

```
vhod: funkcija f : [a, b] → ℝ, f(a)f(b) < 0, natancnost ε
izhod: nicla funkcije f

dokler |b − a| > ε:
    c ← (a+b)/2
    ce sign(f(c)) = sign(f(a)):
        a ← c
    sicer :
        b ← c
vrni c
```

c je približek ničle α. Velja

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

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

Navadna iteracija

Rešujemo $f(x) = 0$. Enačbo pretvorimo v $x = g(x)$. 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:

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

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:

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

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 $|g'(\alpha)| < 1$, 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:

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

Vsak korak se št. decimalk pomnoži s p.

Naj bo g v okolici α p-krat zvezno odvedljiva 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

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

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

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

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

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

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

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

$$\begin{aligned} \|x\|_p &= \sqrt[p]{|x_1|^p + \cdots + |x_n|^p} && \ldots \quad p\text{-norma} \\ \|x\|_\infty &= \max_{1\leq i\leq n} |x_i| && \ldots \quad \text{max norma} \end{aligned}$$

Cauchy-Schwartzova neenakost: $|x^T y| \leq \|x\|_2 \ \|y\|_2$

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

- $\|A\| \geq 0$ in $\|A\| = 0 \iff A = 0$
- $\|\alpha A\| = |\alpha|\|A\|$, $\alpha \in \mathbb{R}$
- $\|A+B\| \leq \|A\| + \|B\|$
- $\|AB\| \leq \|A\|\|B\|$ *submultiplikativnost*

$$\|A\|_1 = \max_{x\neq 0} \frac{\|Ax\|_1}{\|x\|_1} = \max_{1\leq j\leq n} \overbrace{\left(\sum_{k=1}^{m} |a_{kj}|\right)}^{\|a_j\|_1}$$

$$\|A\|_2 = \max_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \sqrt{\rho(A^T A)}$$

$$\|A\|_\infty = \max_{x\neq 0} \frac{\|Ax\|_\infty}{\|x\|_\infty} = \max_{1\leq i\leq m} \left(\sum_{k=1}^{n} |a_{ik}|\right)$$

$$\|A\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$$

$$\|A\|_{op} = \max_{\|c\|_v = 1} \|Ax\|_v \qquad \|\cdot\|_v \ \textit{vekt. norma}$$

spektralni radij $\rho(A) = \max\{|\lambda_1|, \ldots, |\lambda_n|\}$, 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

$$\|\cdot\|_\alpha, \|\cdot\|_\beta \implies C_1\|A\|_\beta \leq \|A\|_\alpha \leq C_2\|A\|_\beta$$

Za vsako operatorsko normo nad kvadratnimi matrikami velja:

$$|\lambda| \leq \|A\| \qquad \lambda \text{ lastna vrednost matrike A}$$

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

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

$$\begin{aligned} \text{Im}A &= \{Ax : x \in \mathbb{R}^n\} = \text{Lin}\{a_1, \ldots, a_n\} \\ \text{Ker}A &= \{x \in \mathbb{R}^n : Ax = 0\} \end{aligned}$$

$$\text{rang}A = n - \dim(\text{Ker}A)$$

Če je $m = n$, so naslednje izjave ekvivalentne:

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

Lastnosti matrike A:

$$\begin{aligned} A &= A^T && \ldots \quad \text{simetrična} \\ A &= A^H && \ldots \quad \text{hermitska} \\ A^T A &= I && \ldots \quad \text{ortogonalna} \end{aligned}$$

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

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

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

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

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

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

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

Za nesingularno matriko $A \in \ltimes \times \ltimes$ 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 $\underset{i=k\ldots n}{\text{argmax}}|a_{k,i}^{(k)}|$-to vrstico matrike A^k.

Občutlivost sistemov linearnih enačb

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

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

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

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 $\max\limits_{i,j} |a_{ij}| = a_{kk}$

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:

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

vhod: matrika A z elementi a_{ij}
izhod: matrika V z elementi v_{ij}
za vsak $j = 1 \dots n$:
  $v_{jj} \leftarrow \sqrt{a_{jj} - \sum_{k=1}^{j-1} v_{jk}^2}$
  za vsak $i = j+1 \dots n$:
    $v_{ij} \leftarrow \frac{1}{v_{jj}} \left(a_{ij} - \sum_{k=1}^{j-1} v_{ik} v_{jk} \right)$
vrni V

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 $F(x) = 0$

Posplošitev navadne iteracije

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

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

Jakobijeva matrika

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

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

- $x \in \Omega \implies G(x) \in \Omega$
- $x \in \Omega \implies \rho(J_G(x)) \leq q < 1$

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.

Variacijske metode

Ekstremi funkcije $G : \mathbb{R}^n \to \mathbb{R}$:

Potreben pogoj za ekstrem je $\nabla G(\alpha) = 0$.

Karakterizacija s Hessejevo matriko:

$$H_G(\alpha) = \left[\frac{\partial^2 G}{\partial x_i \partial x_j} \right]_{i,j=1}^n$$

- $H_G(\alpha)$ poz. definitna $\implies$ min
- $H_G(\alpha)$ neg. definitna $\implies$ max
- $H_G(\alpha)$ nedefinitna $\implies$ ni ekstrema
- $H_G(\alpha)$ semidefinitna $\implies$ ne vemo

Metoda za iskanje minimuma

Naj bo $x^{(r)}$ tekoči približek za minimum.

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

Izberemo tako smer spusta v_r in velikost koraka λ_r, da je $G(x^{(r+1)}) < G(x^{(r)})$.

$v_r = -\nabla G(x^{(r)})$ ... Metoda najhitrejšega spusta

Za določitev λ gledamo $g_r(\lambda) = G(x^{(r)} + \lambda v_r)$:

- *Največji spust*: λ_r je rešitev $g_r'(\lambda) = 0$
- *Tangentni spust*: $\lambda_r = \frac{-g_r(0)}{g_r'(0)}$
  Če je $g_r(\lambda_r) > g_r(0)$, razpolavljamo λ_r, dokler ni $g_r(\lambda_r) < g_r(0)$

Metoda najmanjših kvadratov

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

Predpostavimo, da je A polnega ranga. Iščemo rešitev v smislu najmanjših kvadratov:

$$x^* = \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} \|Ax - b\|_2$$

Geometrijsko je x^* pravokotna projekcija b na $\operatorname{Im} A$:

$$Ax^* - b \perp \operatorname{Im} A \iff \forall i : \; a_i^T (Ax^* - b) = 0$$

Rešitev lahko dobimo z **normalnim sistemom**:

$$A^T A x^* = A^T b$$

QR razcep

$$A \in \mathbb{R}^{m \times n} \quad m > n \quad \operatorname{rang} A = n$$

Obstaja enoličen razcep $A = QR$

$Q \in \mathbb{R}^{m \times n}$ z ortonormiranimi stolpci
$R \in \mathbb{R}^{n \times n}$ zgornje trikotna s poz. diag. el.

za vsak $k = 1, \dots, n$:
  $q_k \leftarrow a_k$
  za vsak $i = 1, \dots, k-1$:
    ce modificiran nacin:
      $r_{ik} \leftarrow q_i^T q_k$
      $q_k \leftarrow q_k - r_{ik} q_i$
    sicer :
      $r_{ik} \leftarrow q_i^T a_k$
  $r_{kk} \leftarrow \|q_k\|_2$
  $q_k \leftarrow \frac{q_k}{r_{kk}}$