Zapis števil in napake

Števila predstavimo kot elemente $P(b,t,L,U)$, to so vsa decimalna števila $0.c_1c_2\dots c_t \cdot b^e$, $L \leq e \leq U, c_1 \neq 0$. Osnovna zaokrožitvena napaka je $u = \frac{1}{2}b^{-t}$.
Standard IEEE single: $\boxed{s|e|f}$, s predznak, 1 bit, e je eksponent, 8 bitov, f je mantisa, 23 bitov. Število x zapišemo kot $x = (-1)^s(1+f)2^{e-127}$. Denormalizirano število: $e = 0, f \neq 0, x = (-1)^s(0+f)2^{-126}$
Za elementarne operacije velja $\mathrm{fl}(a \oplus b)$ se v praksi izračuna z relativno napako $|\delta| < u$ v $(a \oplus b)(1+\delta)$. Za zaporednje n operacij je napaka manjša od nu.
Direktna slabilnost: vedno majhna relativna napaka.
Obratna stabilnost: izračunan rezultat je točen rezultat malo spremenjenih začetnih vrednosti.

Nelinearne enačbe

Iščemo ničle α funkcije f. Občutljivost $\frac{1}{f'(\alpha)}$, za dvojno ničlo $\sqrt{\frac{2}{f''(x)}}$.
Bisekcija: razpolavljamo interval, na katerem imamo ničlo. Št korakov za natančnost ε: $k \geq \log\left(\frac{|b-a|}{\varepsilon}\right)$.
Navadna iteracija: Iščemo fiksno točno $g(\alpha) = \alpha$. Metoda: $x_{r+1} = g(x_r)$. Če je $|g'(\alpha)| < 1$ je točka privlačna, če $|g'(\alpha)| > 1$ je odbojna. Red konvergence je p, če je α p-kratna ničla g. Ocene za napako: $|x_r - \alpha| \leq m^r|x_0 - \alpha|$, $|x_{r+1} - \alpha| \leq \frac{m}{1-m}|x_r - x_{r-1}|$, kje je m Lipscitzeva konstanta za g ($m = \max g'$).
Tangentna metoda: $x_{r+1} = x_r - \frac{f(x_r)}{f'(x_r)}$. Konvergenca je za enojne ničle kvadratična, za večkratne ničle linearna. Če za enostavno ničlo velja $f''(\alpha) = 0$ je konvergenca kubična, itn... Vse ničle so privlačne.
Sekantna metoda: $x_{r+1} = x_r - \frac{f(x_r)(x_r - x_{r-1})}{f(x_r) - f(x_{r-1})}$. Red konvergence: $\frac{1+\sqrt{5}}{2}$.
Laguerrova metoda za iskanje ničel polinomov: $z_{r+1} = z_r - \frac{np(z_r)}{p'(z_r) \pm \sqrt{(n-1)((n-1)p'^2(z_r) - np(z_r)p''(z_r))}}$
Pri stabilni metodi izberemo predznak tako, da je absolutna vrednost imenovalca največja. Če izbiramo vedno $-$ ali $+$ skonvergiramo k levi oz. desni ničli, če so vse ničle realne. Konvergenca v bližini enostavne ničle je kubična. Metoda najde tudi kompleksne ničle.
Redukcija polinoma: Imamo eno ničlo, radi bi jo faktorizirali ven. Poznamo obratno in direktno redukcijo, pri katerih je stabilno izločati ničle v padajočem in naraščajočem vrstnem redu po absolutni vrednosti. V praksi uporabimo kombinirano metodo: do nekega r uporabimo z ene strani obratno, z druge pa direktno. Ta r izberemo tako, da je $|\alpha^r a_{n-r}|$ maksimalen.
Durand-Kernerjeva metoda: Iščemo vse ničle naenkrat: $x_k^{(r+1)} = x_k^{(r)} - \frac{p(x_k^{(r)})}{\prod_{\substack{j=1 \\ j\neq k}}^{n}(x_k^{(r)} - x_j^{(r)})}$. Kvadratična konvergenca. Za kompleksne ničle je treba začeti s kompleksnimi približki.

Linearni sistemi

Norme: $\|A\|_1 = \max_{j\in\{1..n\}}\left(\sum_{i=1}^n |a_{ij}|\right) =$ največji stolpec, $\|A\|_\infty = \|A^T\|_1 =$ največja vrstica
$\|A\|_2 = \sigma_1 = \sqrt{\lambda_{max}(A^HA)} =$ največja singularna vrednost, $\|A\|_F = \sqrt{\sum_{ij} a_{ij}^2} =$ gledamo kot vektor
Operatorska norma: $\|A\| = \max_{x\neq 0}\frac{\|Ax\|}{\|x\|}$. Neenakosti: $\lambda \leq \|A\|$. $\|Ax\| \leq \|A\|\|x\|$.

$$\begin{aligned}
\frac{1}{\sqrt{n}}\|A\|_F &\leq \|A\|_2 \leq \|A\|_F \\
\frac{1}{\sqrt{n}}\|A\|_1 &\leq \|A\|_2 \leq \sqrt{n}\|A\|_1 \\
\frac{1}{\sqrt{n}}\|A\|_\infty &\leq \|A\|_2 \leq \sqrt{n}\|A\|_\infty \\
N_\infty(A) &\leq \|A\|_2 \leq nN_\infty(A) \\
&\leq \|A\|_2 \leq \sqrt{\|A\|_1\|A\|_\infty} \\
\|a_i\|_2, \|\alpha_i\|_2 &\leq \|A\|_2
\end{aligned}$$

Rešujemo sistem $Ax = b$. Za napako x velja ocena:

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

Količina $\kappa(A)$ se imenuje občutljivost matrike. $\kappa(A) = \|A\|\|A^{-1}\|$. Velja $\kappa_2(A) = \frac{\sigma_1(A)}{\sigma_n(A)} \geq 1$.
LU razcep s kompletnim pivotiranjem: matriko A zapišemo kot $PAQ = UL$, L sp. trikotna z 1 na diagnoali in U zg. trikotna, ter P, Q permutacijski matriki stolpcev in vrstic. Algoritem:

```
Q = I, P = I
for j = 1 to n:
    r, q taka, da a_rq največji v podmatriki A(j+1:n)
    zamenjaj vrstici r in j v A, L, P // za delno pivotiranje
    zamenjaj stolpca q in j v A, L, Q // za kompletno pivotiranje
```

```
    for i  = j+1 to n:
        l_ij = a_ij / a_jj
        for k = j+1 to n:
            a_ik = a_ik - l_ij * a_jk
```

Postopek na roke:

1. * Če delamo pivotiranje zamenjamo primerne vrstice in stolpce v A, P, Q, da je a_{00} največji.
2. Prvi stolpec delimo z a_{00}, razen a_{00}, ki ga pustimo na miru.
3. Za vsak element v podmatriki $A(2:n, 2:n)$: $a_{ij} = a_{ij} - a_{i1} \cdot a_{1j}$ (odštejemo produkt $\leftarrow$ in $\uparrow$).
4. Ponovimo postopek na matriki $A(2:n, 2:n)$.

Delno pivotiranje uporablja samo matriko P, za LU razcep brez pivotiranja pa preskočimo 1.
Skalarni produkt potrebuje $2n$ operacij. Reševanje s premimi substitucijami potrebuje n^2, z obratnimi $n^2 + n$. Reševanje z LU razcepom (brez pivotiranja) potrebuje $\frac{2}{3}n^3 + \frac{3}{2}n^2 + \frac{5}{6}n$ operacij. Za izračun Lx potrebujemo n^2 operacij, za Ax potrebujemo $2n^2$ operacij. Za izračun inverza matrike preko LU razcepa potrebujemo $2n^3$ operacij.
Za izračunani LU razcep $\hat{L}\hat{U} = A + E$ velja $|E| \leq nu|\hat{L}||\hat{U}|$.
Pivotna rast: $g = \frac{\max u_{ij}}{\max a_{ij}}$. Pri delnem pivotiranju $g < 2^n$.
Razcep Choleskega: Za spd matriko A obstaja razcep $A = VV^T$.

```
for k = 1 to n:
    v_kk = sqrt(a_kk - sum(v_kj^2, j=1 to k))
    for i = k+1 to n:
        v_ik = 1/v_kk * (a_ik - sum(v_ij * v_kj, j = 1 to k))
```

Postopek na roke po stolpcih:

1. Če sem diagonalen element: odštejem od sebe skalarni produkt vrstice na levo same s sabo in se korenim.
2. Če nisem diagonalni: od sebe odštejem skalarni produkt vrstice levo od sebe z vrstico levo od mojega diagonalnega. Nato se delim z diagonalnim.

Razcep stane $\frac{1}{3}n^3$ operacij. Je obratno stabilno. Je enoličen.

Nelinearni sistemi

Jacobijeva iteracija: Posplošitev navadne iteracije. Naj velja $G(\alpha) = \alpha$. Metoda: $x^{(r+1)} = G(x^{(r)})$. Točka α je privlačna, če velja $\rho(DG(\alpha)) < 1$. Dovolj je $\|DG(\alpha)\| < 1$. Konvergenca je linearna.
Newtonova metoda: Posplošitev tangentne metode. Metoda: reši sistem $DF(x^{(r)})\Delta x^{(r)} = -F(x^{(r)})$.
$x^{(r+1)} = x^{(r)} + \Delta x^{(r)}$. Konvergenca je kvadratična.

Problem najmanjših kvadratov

Reševanje predoločenih sistemov: Za dan predoločen sistem $Ax = b$ rešujemo normalni sistem $A^TAx = A^Tb$. Če je A polnega ranga, je x enoličen. Rešujemo z razcepom Choleskega. Število operacij: $n^2m + \frac{1}{3}n^3$.
QR razcep je bolj stabilen. Za $A \in \mathbb{R}^{m\times n}$ obstaja enoličen razcep $A = QR$, $Q^TQ = I$ in R zg. trikotna s pozitivnimi diagonalci. Za predoločen sistem rešimo $Rx = Q^Tb$.
CGS in MGS Klasična GS ortogonalizacija. Od vsakega stolpca a_k odštejemo pravokotne projekcije $a_i, i < k$. Algoritem ni najbolj stabilen.

```
for k = 1 to n:
    q_k = a_k
    for i = 1 to k-1:
        r_ik = q_i' * a_k (CGS) ALI = q_i' * q_k (MGS)
        q_k = q_k - r_ik q_i
    r_kk = ||q_k||
    q_k = q_k / r_kk
```

Za večjo natančnost izračunamo $[Ab] = [Qq_{n+1}][Rz; 0p]$ in rešimo $Rx = z$. Porabi $2nm^2$ operacij.
Razširjeni QR razcep: $A = \tilde{Q}\tilde{R}$, $Q \in \mathbb{R}^{m\times m}$ ortogonalna, R zgornje trapezna. $\tilde{Q} = [Q\ Q_1]$, $\tilde{R} = [R; 0]$.
Givensove rotacije
Elemente v A po stolpcih enega po enega ubijamo z rotacijami. Rotacija, ki ubije element a_{ki} je $R_{ik}^T([ik],[i,k]) = [c\ s; -s\ c]$, in ostalo identiteta. Parametre nastavimo: $c = x_{ii}/r$, $s = x_{ki}/r$, $r = \sqrt{x_{ii}^2 + x_{ki}^2}$. $\tilde{Q}$ dobimo kot prokdukt vseh rotacij, potrebnih za genocid elementov A. Rotacija spremeni samo i-to in k-to vrstico.
Število operacij: $3mn^2 - n^3$. Če potrebujemo $\tilde{Q}$, potem rabimo še dodatnih $6m^2n - 3mn^2$ operacij.

```
Q = I_m
for i = 1 to n:
    for k = i+1 to m:
      r = sqrt(a_ii^2 + a_ki^2)
      c = a_ii/r, s = a_ki/r
      A([i,k], i:n) = [c s; -s c] A([i k], i:n)
      b([i, k]) = [c s; -s c] b([i, k]) // za predoločen sistem
      Q(i, [i k]) = Q(i, [i k]) [c -s; s c] // za matriko Q
Q = Q'
```

Hausholderjeva zrcaljenja Definiramo $P = I - \frac{2}{w^T w} ww^T$. P je zrcaljenje prek ravnine z normalo w. $Px = x - \frac{1}{m}(x^T w)w$, $m = \frac{1}{2} w^T w$.
Da vektor x prezrcalimo tako, da mu uničimo vse razen prve komponente, uporabimo $w = [x_1 + \text{sign}(x_1)\|x\|_2; x_2; \dots x_n]$ in $m = \|x\|_2(\|x\|_2 + |x_1|)$. Število operacij za Pz je $4nm$ za w in m pa potrebujemo $2n$ operacij.

```
Q = I_m
for i = 1 to n:
    w_i iz R^{m-i+1}, ki prezrcali A(i:m, i) v +-k e_1
    A(i:m,i:n) = P_i * A(i:m, i:n)
    b(i:m) = P_i * b(i:m)           // za predoločen sistem
    Q(i:m, 1:n) = P_i * Q(i:m, 1:n) // za matriko Q
Q = Q'
```

Reševanje predoločenega sistema tako stane $2mn^2 - \frac{2}{3}n^3$. Za $\tilde{Q}$ potrebujemo še $4m^2 n - 2mn^2$ operacij. Za kvadratne sisteme je stabilnejši, a rabimo $\frac{4}{3}n^3$ operacij.
Za napako pri reševanju predoločenega sistema velja: $\frac{\|\Delta x\|}{\|x\|} \leq \frac{\varepsilon \kappa_2(A)}{1 - \varepsilon \kappa_2(A)} \left(2 + (\kappa_2(A) + 1) \frac{\|r\|}{\|A\| \|x\|} \right), r = Ax - b$.

Lastne vrednosti

Gerschgorinov izrek: Naj bo $A \in \mathbb{C}^{n \times n}$, $C_i = \overline{K}(a_{ii}, r = \sum_{j=1, j \neq i}^{n} |a_{ij}|), i = 1,2,\dots,n$. Potem vsaka lastna vrednost leži v vsaj enem Gerschgorinovem krogu. Če m krogov C_i sestavlja povezano množico, ločeno od ostalih $n - m$ krogov, potem ta množica vsebuje natanko m lastnih vrednosti.
Diagonalno dominantna matrika ($|a_{ij}| > \sum_{j=1, j \neq i}^{n} |a_{ij}|$) je obrnljiva.

Interpolacija

Lagrangeev interpolacijski polinom:

$$l_{n,j}(x) = \frac{\prod_{i=0, i \neq j}^{n} (x - x_i)}{\prod_{i=0, i \neq j}^{n} (x_j - x_i)}$$

Polinom: $p(x) = \sum_{i=0}^{n} f(x_i) l_{n,i}(x)$
Deljene diference:

- Če so točke paroma različne: $D_{i,0} = y_i$, ostalo izračunamo po rekurzivni formuli: $D_{i,j} = \frac{D_{i,j-1} - D_{i-1,j-1}}{x_i - x_{i-j}}$. Če sta dve točki na j-tem koraku enaki, je $D_{i,j} = \frac{f^{(j)}(x_i)}{j!}$.
  Polinom: $p(x) = D_{1,1} + D_{2,2}(x - x_0) + D_{3,3}(x - x_0)(x - x_1) + \cdots D_{n,n}(x - x_0) \cdots (x - x_{n-1})$

Integriranje

Ekvidistančne točke $a = x_0 < x_1 < \cdots < x_n = b$, $x_i = x_0 + ih$.
Sest. trapezno pravilo: $\int_a^b f(x)dx = \frac{h}{2}(f(x_0) + 2f(x_1) + 2f(x_2) + \cdots + 2f(x_{m-1}) + f(x_m)) - \frac{h^2}{12}(b - a) f''(\xi)$
Sest. Simpsonovo: $\int_a^b f(x)dx = \frac{h}{3}(f(x_0) + 4f(x_1) + 2f(x_2) + \cdots + 2f(x_{2m-2}) + 4f(x_{2m-1}) + f(x_{2m})) - \frac{h^4}{180}(b - a) f^{(4)}(\xi)$
3/8 pravilo: $\int_a^b f(x)dx = \frac{3}{8}h(f(x_0) + 3f(x_1) + 3f(x_2) + f(x_3)) - \frac{3}{80}h^5 f^{(4)}(\xi), \quad \xi \in (x_0, x_3)$