1 Układy równań liniowych

1.1 Podstawowe pojęcia

• $Ax = b$ ma rozwiązanie gdy: $det(A) \neq 0$, $b = 0$ to $x = 0$, $ker(A) = \{0\}$, A nie ma zerowej wartości własnej, A jest odwracalna.

• Macierz z diagonalą rozwiązuje się trywialnie: $x_i = \frac{b_i}{a_{ii}}$. Koszt $O(n)$.

• FLOP - FLoating point OPeration - dodawanie, mnożenie, dzielenie, odejmowanie, sqrt to jeden flop.

• Macierz trójkątna dolna, układ równań rozwiązany w przód: $x_i = \frac{b_i - \Sigma_{j=1}^{i-1} a_{ji}x_j}{a_{ii}}$. Koszt $O(n^2)$.

• Macierz trójkątna górna rozwiązanie analogiczne (w tył), koszt $O(n^2)$.

• Macierz ortogonalna Q spełnia $Q^TQ = I$, $Q^{-1} = Q^T$.

• Gdy A ortogonalna, to $Ax = b$ ma rozwiązanie $x = A^Tb$. Koszt $O(n^2)$.

1.2 Rozkład LU

• $PA = LU$ (z wyborem) lub $A = LU$ (bez wyboru), P - permutacja, L - trójkątna dolna, U - trójkątna górna.

• Standardowy algorytm rozkładu LU z bez wyboru elementu głównego: $A = LU$. $PA = \left[\begin{array}{c|c} a_{1,1} & a_{1,2}^T \\ \hline a_{2,1} & A_{2,2}\end{array}\right] = \left[\begin{array}{c|c} 1 & 0^T \\ \hline l_{2,1} & L_{2,2}\end{array}\right] \cdot \left[\begin{array}{c|c} u_{1,1} & u_{1,2}^T \\ \hline 0 & U_{2,2}\end{array}\right]$

$u_{1,1} = a_{1,1}$, $u_{1,2} = a_{1,2}$, $l_{2,1} = \frac{a_{2,1}}{u_{1,1}}$, aktualizuj $A_{2,2} = A_{2,2} - l_{2,1}u_{1,2}^T$, wyznacz $A_{2,2} = L_{2,2}U_{2,2}$.
Działa gdy wszystkie miniory główne macierzy A są niezerowe.

• GEPP - Gauss Elimination with Partial Pivoting - GE z częściowym wyborem elementu głównego.
$PA = LU$, P - permutacja, L - trójkątna dolna, U - trójkątna górna.
$PAx = LUx = Pb$, $\tilde{b} = Pb$, $Lx = \tilde{b}$, $Uy = y$. Koszt $O(n^3)$.

2 Arytmetyka zmiennopozycyjna

• $6.63 \cdot 10^{-34}$ - 6.63 - mantysa, -34 - cecha, 10 - podstawa

• $x = (-1)^s \cdot m \cdot \beta^e$ - s - znak, $m = (f_0.f_1f_2...f_{p-1})_2$ - mantysa, b - podstawa, e - cecha

• W liczbach maszynowych $\beta = 2$

• $(1 - e_{max} \leqslant e \leqslant e_{max})$

• Liczby maszynowe są normalizowane $f_0 = 1$ i nie jest zapisywane.

• W 6-bitowej arytmetyce zmiennopozycyjnej liczby subnormalne zachodzą dla $e = -2$ i $f_0 = 0$.

• Metody aproksymacji liczb maszynowych:
RN - do najbliższej (domyślnie)
RD - w dół, tzn w stone -∞
RU - w górę, tzn w stronę ∞
RZ - do zera $RZ(x) = RD(x)$ gdy $x \geqslant 0$,$RU(x)$ gdy $x \leqslant 0$

• Jeśli $|x| \in [realmin, realmax]$ to $\frac{|x - RN(x)|}{|x|} \leqslant \frac{1}{2^p} = v$

• $fl(x) = x(1 \cdot \epsilon)$, gdy $|\epsilon| \leqslant v$.

• dla float32 $v = 6.0 \cdot 10^{-8}$ oraz dla float64 $v = 1.1 \cdot 10^{-16}$

3 Uwarunkowanie zad. i numeryczna poprawnosc

• $\|\tilde{x} - x\|$ - błąd bezwzględny, $\|\tilde{x} - x\| / \|x\|$ - błąd względny ($x \neq 0$)

• Wskaźnik uwarunkowania (bezwzględny) na poziomie ϵ
$cond_{abs}(P, x, \epsilon) = \sup_{\|\Delta\| \leqslant \epsilon} \frac{\|P(x+\Delta) - P(x)\|}{\|\Delta\|}$
$\|P(\tilde{x}) - P(x)\| \leqslant cond_{abs}(P, x, \epsilon) \cdot \|\tilde{x} - x\|$, dla $\|\tilde{x} - x\| \leqslant \epsilon$

• Idealizacja punktowy wskaźnik uwarunkowania
$cond_{abs}(P, x) = \lim_{\epsilon \to 0} cond_{abs}(P, x, \epsilon)$

• Gdy P różniczkowalna to $cond_{abs}(P, x) = \|P'(x)\|$

• Wrażliwość dla błędu względnego
$cond_{rel}(P, x, \epsilon) = \sup_{\|\Delta\| \leqslant \epsilon} \frac{\|P(x+\Delta) - P(x)\|}{\|P(x)\|} / \frac{\|\Delta\|}{\|x\|} = cond_{abs}(P, x, \epsilon\|x\|)\frac{\|x\|}{\|P(x)\|}$
Punktowo:
$cond_{rel}(P, x) = \lim_{\epsilon \to 0} cond_{rel}(P, x, \epsilon)$
Gdy P jest różniczkowalna to $cond_{rel}(P, x) = \frac{\|P'(x)\|}{\|P(x)\|}$

• Zadanie dobrze uwarunkowanie: $cond(P, x)$ nieduże
Zadanie źle uwarunkowane: $cond(P, x)$ bardzo duże

• Normy wektorowe
$\|x\|_1 = \sum_{i=1}^n |x_i|$, $\|x\|_2 = \sqrt{\sum_{i=1}^n |x_i|^2}$, $\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$, $\|x\|_\infty = \max_{1 \leqslant i \leqslant n} |x_i|$

• $\|x\|_\infty \leqslant \|x\|_1 \leqslant N\|x\|_\infty$, $\|x\|_\infty \leqslant \|x\|_2 \leqslant \sqrt{n}\|x\|_\infty$, $\|x\|_2 \leqslant \|x\|_1 \leqslant \sqrt{n}\|x\|_2$

• $\|A\| = max_{\|x\| \neq 0}\|Ax\| / \|x\| = max_{\|x\|=1}\|Ax\| = max_{\|x\| \leqslant 1}\|Ax\|$

• $\|Ax\| \leqslant \|A\|\|x\|$, $\|AB\| \leqslant \|A\|\|B\|$, $\|I\| = 1$, $\|A\|_1 = max_j\Sigma_i|a_{ij}|$, $\|A\|_\infty = max_i\Sigma_j|a_{ij}|$, $\|A\|_2 = max\{\sqrt{\mu} : \mu$ jest w. wł. $A^TA\}$

• $cond(A) = \|A\|\|A^{-1}\|$

• Ay = b. Jeśli $\epsilon cond(A) \leqslant 1/2$ to $\|\tilde{y} - y\| / \|y\| \leqslant 4cond(A) \cdot \epsilon$

• Jeśli $\|\Delta\| < 1$ to $I + \Delta$ nieosobliwa i $1/(1 + \|\Delta\|) \leqslant \|(I + \Delta)^{-1}\| \leqslant 1/(1 - \|\Delta\|)$

• Algorytm poprawnie numeryczny - dla każdego $x \in X$ wynik algorytmu A zrealizowanego w fl $fl(A(fl(x)))$ jest dokładnym rozwiązaniem zadania dla danych x zaburzonych na poziomie błędu reprezentacji.

• Algorytm NP daje wynik, którego błąd można oszacować na podstawie własności zadania obliczeniowego:
$\|\tilde{y} - y\| / \|y\| = \|P(\tilde{x}) - P(x)\| / \|P(x)\| \lesssim cond_{rel}(P, x)\|\tilde{x} - x\| / \|x\| \leqslant K \cdot cond_{rel}(P, x) \cdot v$

4 LZNK

$A \in \mathbb{R}^{m \times n}, m \geqslant n, \text{rank}(A) = n$, $\vec{b} \in \mathbb{R}^m$, znaleźć $\vec{x} \in \mathbb{R}^n$ taki, że $\|\vec{b} - A\vec{x}\|_2 \to min$.
$\|\vec{b} - A\vec{x}\| \leqslant \|\vec{b} - A\vec{y}\| \forall y$.

• QR — $A = QR = \begin{bmatrix}\hat{Q} & \tilde{Q}\end{bmatrix}\begin{bmatrix}\hat{R} \\ 0\end{bmatrix} \implies$
$\|\vec{b} - A\vec{x}\|_2^2 = \|\hat{Q}^T\vec{b} - \hat{R}\vec{x}\|_2^2 + \|\tilde{Q}^T\vec{b}\|_2^2 = \min \iff \hat{R}\vec{x} = \hat{Q}\vec{b}$

• A^TA — mała, symetryczna, dodatnio określona $A^TA\vec{x} = A^T\vec{b}$

• Algorytm: oblicz $B = A^TA$ $\to$ wyznacz Cholesky'ego $B = LL^T$ $\to$ rozwiąż $LL^T\vec{x} = A^T\vec{b}$

• SVD — $A \in \mathbb{R}^{m \times n}, m \geqslant n$, istnieje $A = U\Sigma V^T$, że
$U \in \mathbb{R}^{m \times n} : U^TU = I, V \in \mathbb{R}^{n \times n} : V^TV = I, \Sigma \in \mathbb{R}^{n \times n} : \Sigma = \text{diag}(\sigma_1, .., \sigma_n), \sigma_1 \geqslant .. \geqslant \sigma_n \geqslant 0$
$\lambda(A^TA) = (\sigma_1^2, .., \sigma_n^2), \sigma_1 \geqslant .. \geqslant \sigma_r > \sigma_{r+1} = .. = \sigma_n = 0 \implies \text{rank}A = r$
$A = \sum_{i=1}^n \sigma_i\vec{u_i}\vec{v_i}^T$ gdzie $\vec{u_i}, \vec{v_i}$ — kolumny U, V
$\|\vec{b} - A\vec{x}\|_2^2 = \sum_{i=1}^r \left(\sigma_iy_i - \vec{u_i}^T\vec{b}\right)^2 + \sum_{i=r+1}^m \left(\vec{u_i}^T\vec{b}\right)^2 \to \min$, gdy $y_i = \frac{\vec{u_i}^T\vec{b}}{\sigma_i}$,
To oznacza, że rozwiązanie LZNK jest jednoznacznie określone $\vec{x^*} = \sum_{i=1}^r \frac{\vec{u_i}^T\vec{b}}{\sigma_i}\vec{v_i}$

5 Metody iteracyjne

• Ciąg $x_{k+1} = Bx_k + c$ jest zbieżny do x^* dla każdego x_0 wtedy i tylko wtedy, gdy $\rho(B) < 1$. gdzie $\rho(B)$ to promień spektralny macierzy B ($\rho(B) = max\{|\lambda| : \lambda$ jest w. wł. B$\}$)

• Jeśli $\|B\| < 1$ to ciąg $x_{k+1} = Bx_k + c$ jest zbieżny do x^* dla każdego x_0. oraz $\|x^* - x_k\| \leqslant \|B\| \cdot \|x_k - x^*\|$

• Niech $A = M - Z$ oraz A, M nieosobliwe. $Ax^* = b$. Jeśli $\rho(M^{-1}Z) < 1$ to metoda iteracyjna $Mx_{k+1} = Zx_k + b$ jest zbież-

na do x^* dla każdego x_0.
Jeśli dodatkowo $\gamma = \left\| M^{-1}Z \right\|_\infty < 1$ to $\|x^* - x_k\| \leqslant \gamma \cdot \|x_k - x^*\|$

• Niech $A = L + D + U$, $D = diag(A)$, L (odp. U) - dolna (odp. górna) trójkątna z zerową diagonalą.
$x_{k+1} = x_k + M^{-1}(b - Ax_k)$
Metoda Jacobiego: $M = D$, Metoda Gaussa-Seidela: $M = D + L$, Metoda SOR: $M = 1/\omega D + L$.

• Jeśli A jest diagonalnie dominująca to m. Jacobiego jest zbieżna do x^* dla każdego x_0.

• $x_{k+1} = x_k + \delta_k$. Jak wybrać δ_k?. Idealna poprawka δ_k^*:
$A\delta_k^* = r_k \rightarrow x_{k+1} = x^*$
Wyznaczamy idealną poprawkę $\delta_k^* = V_k a_k$, gdzie $V_k, U_k \in \mathbb{R}^{N \times r}$ max rzędu t. że $a_k \in \mathbb{R}^r$ spełnia:
$U_k^T A V_k a_k = U_k^T r_k$