Venujeme sa numerickým metódam riešenia lineárnych systémov rovníc

$$Ax = b$$

Základné delenie numerických metód

1. Priame metódy: Gaussova eliminačná metóda, LU-rozklad, LTL^T-rozklad, Choleského rozklad
2. Iteračné metódy: Jacobiho, Gaussova-Siedelova, metóda najväčšieho spádu, metóda združených smerov, metóda združených gradientov,

Gaussova eliminačná metóda a LU rozklad

Máme daný systém rovníc v maticovom tvare

$$Ax = b$$

kde $A \in M_{nn}(\mathbb{R})$ je reálna $n \times n$ matica, x a b sú stĺpcové vektory dĺžky n. Postupným použitím elementárnych riadkových operácií potom eliminujeme prvky matice pod diagonálou. V prvom kroku teda postupne odpočítavaním $\frac{a_{i1}}{a_{11}}$ násobku prvého riadku od i-teho $(i = 2, 3, \ldots, n)$ vynulujeme prvky pod diagonálou v prvom stĺpci.

$$\underbrace{\left(\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} & b_n \end{array}\right)}_{(A|b)} \overset{1.\text{ krok}}{\sim} \underbrace{\left(\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ 0 & a_{22}^{(1)} & \cdots & a_{2n}^{(1)} & b_2^{(1)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & a_{n2}^{(1)} & \cdots & a_{nn}^{(1)} & b_n^{(1)} \end{array}\right)}_{(A^{(1)}|b^{(1)})}$$

Takto postupujeme aj v ďalších krokoch, teda v k-tom kroku odpočítavaním $m_{ik} = \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}}$ násobku k-teho riadku od i-teho $(i = k+1, k+2, \ldots, n)$ vynulujeme prvky pod diagonálou v k-tom stĺpci. Po $n-1$ krokoch získame systém $(A^{(n-1)}|b^{(n-1)})$, kde $A^{(n-1)}$ je horná trojuholníková matica. Z tohoto systému hľadané riešenie x získame spätným dosadením

$$x_n = \frac{b_n^{(n-1)}}{a_{nn}^{(n-1)}},\ x_{n-1} = \frac{b_{n-1}^{(n-2)} - a_{n-1\,n}^{(n-2)} x_n}{a_{n-1\,n-1}^{(n-2)}}, \ldots,$$

$$x_k = \frac{b_k^{(k-1)} - a_{k\,k+1}^{(k-1)} x_{k+1} - a_{k\,k+2}^{(k-1)} x_{k+2} - \cdots - a_{k\,n}^{(k-1)} x_n}{a_{kk}^{(k-1)}}$$

Operáciu pripočítania násobku jedného riadku matice k inému je možné reprezentovať maticovým násobením. Prenásobením matice A maticou

$$M_{ik} = \begin{pmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & 1 & \cdots & 0 & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & -m_{ik} & \cdots & 1 & \cdots & 0 \\ \vdots & & \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{pmatrix}$$

zľava dosiahneme odpočítanie m_{ik} násobku k-teho riadku matice A k i-temu riadku matice A. Ak teda máme maticu $A^{(k-1)}$ ktorá má nuly pod diagonálou v stĺpcoch 1 až $k-1$, k-ty stĺpec vynulujeme postupným násobením matice $A^{(k-1)}$ maticami M_{ik}, $i = k+1, \ldots, n$

$$A^{(k)} = M_{nk} M_{n-1\,k} \cdots M_{k+1\,k} A^{(k-1)}$$

Sučinom matíc $M_{nk} M_{n-1\,k} \cdots M_{k+1\,k}$ vznikne matica M_k, ktorá má 1 na diagonále a ktorej k-ty stĺpec je

$$(\underbrace{0 \ \cdots \ 0\ 1}_{k} \ -m_{k+1\,k} \ \cdots \ -m_{nk})^T.$$

Z matice A dostaneme potom hornú trojuholníkovú maticu $A^{(n-1)}$ nasledovne

$$A^{(n-1)} = M_{n-1} M_{n-2} \cdots M_2 M_1 A$$

Potom

$$A = M_1^{-1} M_2^{-1} \cdots M_{n-2}^{-1} M_{n-1}^{-1} A^{(n-1)}$$

Definujme vektor

$$m_k = (\underbrace{0 \ \cdots \ 0\ 0}_{k} \ m_{k+1\,k} \ \cdots \ m_{nk})^T.$$

potom maticu M_k môžeme zapísať ako

$$M_k = I - m_k e_k^T,$$

kde e_k je vektor, ktorý má na k-tom mieste 1 a ostatné prvky nulové. Z výpočtu

$$(I - m_k e_k^T)(I + m_k e_k^T) = I - m_k e_k^T m_k e_k^T = I - m_k \cdot 0 \cdot e_k^T = I,$$

máme hneď

$$M_k^{-1} = (I - m_k e_k^T)^{-1} = I + m_k e_k^T.$$

Hneď aj vidíme, že $M_1^{-1} M_2^{-1} \cdots M_{n-2}^{-1} M_{n-1}^{-1}$ je dolná trojuholníková matica, ktorú označíme L. Naviac ak $l \leq k$

$$(I + m_l e_l^T)(I + m_k e_k^T) = I + m_l e_l^T + m_k e_k^T$$

preto

$$L = M_1^{-1} \cdots M_{n-1}^{-1} = I + \sum_{k=1}^{n-1} m_k e_k^T$$

takže matica L obsahuje prvky m_{ik} na príslušných miestach pod diagonálou a 1 na diagonále. Dostaneme tak rozklad

$$A = LA^{(n-1)} = LU.$$

Na šetrenie pamäte, čo môže byť vhodné pri veľkých maticiach, je možné postupne prepisovať rozšírenú maticu $(A|b)$ tak, že prvkami pod diagonálou matice L prepíšeme prvky pod diagonálou matice A a horná trojuholníková časť matice U prepíše hornú trojuholníkovú časť matice A. Na záver možno vektor B prepísať vektrom riešenia. Týmto spôsobom vieme ušetriť pomerne veľa miesta, pretože namiesto alokovania miesta pre tri $n \times n$ matice stačí jedna $n \times n$ matica.

Praktický priebeh výpočtu si predvedieme na nasledovnom príklade.

Príklad: Nájdite LU rozklad matice

$$A = \begin{pmatrix} 2 & 1 & -1 & 3 \\ -4 & -3 & 3 & -9 \\ 6 & 1 & -4 & 1 \\ -2 & -3 & 9 & -2 \end{pmatrix}.$$

Princíp spočíva v postupnom dopĺňaní matíc L a U

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ & 1 & 0 & 0 \\ & & 1 & 0 \\ & & & 1 \end{pmatrix} \qquad U = \begin{pmatrix} 0 & & & \\ 0 & 0 & & \\ 0 & 0 & 0 & \end{pmatrix}.$$

Keďže prvý riadok matice L poznáme, ľahko dopočítame prvý riadok matice U,

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ & 1 & 0 & 0 \\ & & 1 & 0 \\ & & & 1 \end{pmatrix} \qquad U = \begin{pmatrix} 2 & 1 & -1 & 3 \\ 0 & & & \\ 0 & 0 & & \\ 0 & 0 & 0 & \end{pmatrix}.$$

Podobne vieme doplniť prvky v prvom stĺpci matice L

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ 3 & & 1 & 0 \\ -1 & & & 1 \end{pmatrix} \qquad U = \begin{pmatrix} 2 & 1 & -1 & 3 \\ 0 & & & \\ 0 & 0 & & \\ 0 & 0 & 0 & \end{pmatrix}.$$

Teraz doplníme prvky v druhom riadku matice U. Druhý riadok matice L môže mať nenulové len prvé dva prvky. Takže zo znalosti matice A a spôsobu násobenia matíc máme

$$-2 \cdot 1 + u_{22} = -3 \qquad -2 \cdot (-1) + u_{23} = 3 \qquad (-2) \cdot 3 + u_{24} = -9.$$

z čoho ľahko spočítame druhý riadok matice U

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ 3 & & 1 & 0 \\ -1 & & & 1 \end{pmatrix} \qquad U = \begin{pmatrix} 2 & 1 & -1 & 3 \\ 0 & -1 & 1 & -3 \\ 0 & 0 & & \\ 0 & 0 & 0 & \end{pmatrix}.$$

Následne dopočítame druhý stĺpec matice L zo vzťahov

$$3 \cdot 1 + (-1) \cdot l_{32} = 1 \qquad (-1) \cdot 1 + (-1) \cdot l_{42} = -3$$

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ 3 & 2 & 1 & 0 \\ -1 & 2 & & 1 \end{pmatrix} \qquad U = \begin{pmatrix} 2 & 1 & -1 & 3 \\ 0 & -1 & 1 & -3 \\ 0 & 0 & & \\ 0 & 0 & 0 & \end{pmatrix}.$$

Pre tretí riadok matice U máme analogicky

$$3 \cdot (-1) + 2 \cdot 1 + u_{33} = -4 \qquad 3 \cdot 3 + 2 \cdot (-3) + u_{34} = 1$$

preto

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ 3 & 2 & 1 & 0 \\ -1 & 2 & & 1 \end{pmatrix} \qquad U = \begin{pmatrix} 2 & 1 & -1 & 3 \\ 0 & -1 & 1 & -3 \\ 0 & 0 & -3 & -2 \\ 0 & 0 & 0 & \end{pmatrix}.$$

Podobne dopočítame prvky v treťom stĺpci L

$$(-1) \cdot (-1) + 2 \cdot 1 + (-3) \cdot l_{43} = 9$$

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ 3 & 2 & 1 & 0 \\ -1 & 2 & -2 & 1 \end{pmatrix} \qquad U = \begin{pmatrix} 2 & 1 & -1 & 3 \\ 0 & -1 & 1 & -3 \\ 0 & 0 & -3 & -2 \\ 0 & 0 & 0 & \end{pmatrix}$$

a vo štvrtom riadku matice U

$$(-1) \cdot 3 + 2 \cdot (-3) + (-2) \cdot (-2) + u_{44} = -2$$

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ 3 & 2 & 1 & 0 \\ -1 & 2 & -2 & 1 \end{pmatrix} \qquad U = \begin{pmatrix} 2 & 1 & -1 & 3 \\ 0 & -1 & 1 & -3 \\ 0 & 0 & -3 & -2 \\ 0 & 0 & 0 & 3 \end{pmatrix}$$

Formalizáciu tohoto postupu dosiahneme rozpísaním $A = LU$ po prvkoch a uvedomením si tvaru matíc L a U. Potom pre $A = (a_{ij})$, $L = (l_{ij})$ a $U = (u_{ij})$ máme z $A = LU$

$$a_{ik} = \sum_{j=1}^{n} l_{ij} u_{jk}$$

a keďže pre $i < k$ je $l_{ik} = 0$ a pre $i > k$ je $u_{ik} = 0$ môžeme písať

$$a_{ik} = \sum_{j=1}^{\min\{i,k\}} l_{ij} u_{jk}$$

Potom pre prípady $i = 1$, $k = 1, \ldots, n$ a $k = 1$, $i = 1, \ldots, n$ máme postupne

$$a_{1k} = u_{1k} \qquad\qquad a_{i1} = l_{i1} u_{11}$$

a následne pre $i = 2, \ldots, n$, keďže $l_{ii} = 1$

$$u_{ik} = a_{ik} - \sum_{j=1}^{i-1} l_{ij} u_{jk}, \ k = i, i+1, i+2, \ldots, n$$

a

$$l_{ki} = \frac{a_{ki} - \sum_{j=1}^{i-1} l_{kj} u_{ji}}{u_{ii}}, \ k = i, i+1, i+2, \ldots, n$$

LU rozklad matice nemusí existovať. Napríklad vyššie uvedený postup pre maticu

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 7 \\ 3 & 5 & 3 \end{pmatrix}$$

zlyhá pri

$$L = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & & 1 \end{pmatrix} \qquad\qquad U = \begin{pmatrix} 1 & 2 & 3 \\ 0 & 0 & 1 \\ 0 & 0 & \end{pmatrix}$$

pretože rovnica pre l_{32} je

$$(6 =)3 \cdot 2 + l_{32} \cdot 0 = 5$$

Veta: Ak Existuje LU rozklad matice A, tak je jednoznačný.

Dôkaz. Jendoznačnosť dostaneme klasickým spôsobom nasledovnou úvahou. Ak $L_1U_1 = A = L_2U_2$ sú dva LU rozklady matice A, tak $L_2^{-1}L_1 = U_2U_1^{-1}$. Matica $L_2^{-1}L_1$ je dolná trojuholníková s jednotkami na diagonále, $U_2U_1^{-1}$ je horná trojuholníková. Keďže sa rovnajú, musia byť obidve rovné identickej matici $L_2^{-1}L_1 = I = U_2U_1^{-1}$. Z toho potom $L_1 = L_2$ a $U_1 = U_2$.

Jeden problém, keď proces LU rozkladu (a teda aj Gaussova eliminačná metóda) zlyhá sme už videli, konkrétne, keď v $k-1$ kroku je $u_{kk} = a_{kk}^{(k-1)} = 0$. Z toho vieme odvodiť podmienku existencie LU rozkladu: Ak všetky hlavné minory matice A sú nenulové, tak existuje LU rozklad.

Druhým problémom je numerická stabilita, čo si ilustrujeme na nasledovnom príklade:
Príklad: Riešme systém

$$\begin{aligned} -10^{-4}x + y &= 1 \\ x + y &= 2 \end{aligned}$$

Systém prepíšeme do maticového tvaru a upravíme Gaussovou eliminačnou metódou

$$\left(\begin{array}{cc|c} -10^{-4} & 1 & 1 \\ 1 & 1 & 2 \end{array}\right) \sim \left(\begin{array}{cc|c} -10^{-4} & 1 & 1 \\ 0 & 10001 & 10002 \end{array}\right)$$

a spätnou substitúciou nájdeme

$$y = \frac{10002}{10001} \qquad\qquad x = \frac{10000}{10001}.$$

Toto je presné riešenie.

Ak však počítame v aritmetike s konečným počtom platných číslic môže nastať problém. Ak by sme napríklad v predošlej úlohe počítali s presnosťou na 3 platné číslice, tak $10000 + 1 = 10000$, $10000 + 2 = 10000$ a teda postup by vyzeral nasledovne

$$\left(\begin{array}{cc|c} -10^{-4} & 1 & 1 \\ 1 & 1 & 2 \end{array}\right) \sim \left(\begin{array}{cc|c} -10^{-4} & 1 & 1 \\ 0 & 1 & 1 \end{array}\right)$$

s riešením

$$y = 1 \qquad\qquad x = 0.$$

Gaussova eliminačná metóda slúži na riešenie systémov $Ax = b$. LU rozklad matice A je v princípe len iný zápis Gaussovej eliminačnej metódy. Ak poznáme LU rozklad, tak systém $Ax = b$ vieme prepísať na $LUx = b$ tento systém potom riešime tak, že najskôr nájdeme riešenie $Ly = b$ a následne riešenie $Ux = y$.

Videli sme, že LU rozklad a Gaussova eliminačná metóda pri použití na systém lineárnych rovníc je v princípe to isté. Je teda namieste otázka, prečo je dobré vôbec hľadať LU rozklad matice. Pokiaľ máme len jeden konkrétny systém, tak LU rozkladom naozaj nič nezískame. Avšak často sa stáva, že máme danú maticu systému A a rôzne pravé strany b_i a pre každé b_i chceme nájsť riešenie systému $Ax = b_i$. Časová zložitosť (t.j. počet operácii na výpočet) Gaussovej eliminačnej metódy pre maticu rozmeru $n \times n$ je $O(n^3)$ (presne $\frac{1}{3}n^3 + n^2 - \frac{n}{3}$). Časová zložitosť pre výpočet riešenia $Ux = b$ alebo $Lx = b$, kde U je horná trojuholníková a L je dolná trojuholníková matica je v oboch prípadoch $O(n^2)$. Takže pri viacerých pravých stranách LU rozkladom ušetrím čas výpočtu.

Príklad: Použitie LU rozkladu.

Predpokladajme, že máme danú $n \times n$ maticu A a dva n-vektory c a d. Chceme vypočítať $s = c^T A^{-1} d$. Prvá možnosť, ktorá sa ponúka je spočítať maticu A^{-1} a následne vypočítať uvedený výraz. Ekonomickejší spôsob je ale použiť LU rozklad. Výraz $A^{-1}d$ vlastne znamená nájsť x také, že $Ax = d$, čiže riešiť systém. Postupovať teda môžeme tak, že spočítame rozklad $PA = LU$, nájdeme y také, že $Ly = Pd$ a nájdeme x také, že $Ux = y$. Následne spočítame $s = c^T x$.

Poznámka: A^{-1} vo výrazoch zvyčajne znamená 'nájdi riešenie systému' a len málokedy 'nájdi inverznú maticu'.

V predchádzajúcej časti sme videli problémy, ktoré môžu nastať pri Gaussovej eliminačnej metóde – ak v niektorom kroku vznikla 0 v ľavom hornom rohu príslušnej podmatice, nemohli sme pokračovať v GEM. A ak na diagonále bolo číslo blízke 0, pri GEM v spojení s aritmetikou s konečným počtom platných číslic mohlo dôjsť k podstatným chybám. Tieto problémy z veľkej

časti rieši pivotovanie, t.j. výmena riadkov a stĺpcov matice, tak aby sme v príslušnom kroku v danej podmatici získali v ľavom hornom rohu vhodný prvok.

Je niekoľko typov pivotovania

1. Čiastočné pivotovanie s výberom najväčšieho prvku v stĺpci (riadku).
2. Úplné pivotovanie - výber najväčšieho prvku v príslušnej podmatici.
3. Vežové pivotovanie - výber najväčšieho prvku v príslušnom riadku alebo stĺpci (z šachovej terminológie veža=rook, rook pivoting).

Pozrieme sa najskôr na čiastočné pivotovanie s výberom prvku v stĺpci. Princíp spočíva v tom, že na začiatku kroku GEM vymeníme dva riadky tak, aby sme ako prvok v ľavom hornom rohu príslušnej podmatice dostali v absolútnej hodnote najväčší prvok v prvom stĺpci uvedenej podmatice. Tento postup je opäť výhodné zapísať pomocou matíc.

Výmenu dvoch (prípadne viacerých) riadkov matice A dosiahneme násobením matice A z ľava permutačnou maticou, t.j. maticou, ktorá vznikne z jednotkovej matice výmenou riadkov. Označme permutačnú maticu, ktorá vymieňa i-ty riadok za j-ty, $i < j$, ako Π_i. Potom postup GEM s čiastočným pivotovaním s výberom v stĺpci prebieha nasledovne:

$$M_{n-1}\Pi_{n-1}\cdots M_2\Pi_2 M_1\Pi_1 A = A^{(n-1)} = U \tag{1}$$

pričom $M_k = I - m_k e_k^T$. Pozrime sa, ako možno (1) upraviť ako LU rozklad.

Súčin matíc na ľavej strane (1) môžeme prepísať nasledovne

$$\cdots(\Pi_{n-1}\cdots\Pi_3 M_2\Pi_3\cdots\Pi_{n-1})(\Pi_{n-1}\cdots\Pi_2 M_1\Pi_2\cdots\Pi_{n-1})(\Pi_{n-1}\cdots\Pi_1)A \tag{2}$$

Označme $\hat{M}_k = \Pi_{n-1}\cdots\Pi_{k+1}M_k\Pi_{k+1}\cdots\Pi_{n-1}$. Potom

$$\begin{aligned}\hat{M}_k &= \Pi_{n-1}\cdots\Pi_{k+1}M_k\Pi_{k+1}\cdots\Pi_{n-1}\\ &= \Pi_{n-1}\cdots\Pi_{k+1}(I - m_k e_k^T)\Pi_{k+1}\cdots\Pi_{n-1}\\ &= I - \Pi_{n-1}\cdots\Pi_{k+1}m_k e_k^T\Pi_{k+1}\cdots\Pi_{n-1}.\end{aligned}$$

Keď si uvedomíme, že $e_k^T\Pi_{k+1}\cdots\Pi_{n-1}$ znamená permutáciu $k+1$-ho až $n-1$-ho prvku vektora e_k^T, čo sú ale 0, tak $e_k^T\Pi_{k+1}\cdots\Pi_{n-1} = e_k^T$. Podobne $\Pi_{n-1}\cdots\Pi_{k+1}m_k$ je vektor, ktorý dostaneme nejakou permutáciou $k+1$-ho až $n-1$-ho prvku vektora m_k, a teda prvých k-súradníc tohoto vektora bude

stále rovných 0. Označme preto $\hat{m}_k = \Pi_{n-1} \cdots \Pi_{k+1} m_k$, z čoho dostávame, že

$$\hat{M}_k = I - \hat{m}_k e_k^T$$

je dolná trojuholníková matica, ktorá vznikne z matice M_k permutovaním prvkov v k-tom stĺpci pod hlavnou diagonálou. V každom prípade (2) prejde na tvar

$$\hat{M}_{n-1} \cdots \hat{M}_1 (\Pi_{n-1} \cdots \Pi_1) A = U$$

a po preznačení $P = \Pi_{n-1} \cdots \Pi_1$ a $(\hat{M}_{n-1} \cdots \hat{M}_1)^{-1} = L$ dostaneme LU rozklad s čiastočným pivotovaním s výberom hlavného prvku v stĺpci

$$PA = LU$$

Príklad: Nájdite LU rozklad matice s čiastočným pivotovaním v stĺpci

$$\begin{pmatrix} 1 & 1 & -1 & 2 \\ 0 & 2 & 0 & 1 \\ 2 & 0 & 2 & 0 \\ 1 & 3 & 2 & -1 \end{pmatrix}$$

Riešenie: Budeme počítať GEM hornú trojuholníkovú maticu, pričom multiplikátory m_k budeme zapisovať pod hlavnú diagonálu namiesto núl. Zaroveň si budeme zapisovať transpozície pre príslušné výmeny riadkov.

V prvom kroku teda najskôr nájdeme v absolútnej hodnote najväčší prvok v prvom stĺpci, čo je 2 v treťom riadku, a preto vymeníme prvý a tretí riadok, takže $\Pi_1 = (1\,3)$.

$$\begin{pmatrix} 2 & 0 & 2 & 0 \\ 0 & 2 & 0 & 1 \\ 1 & 1 & -1 & 2 \\ 1 & 3 & 2 & -1 \end{pmatrix}$$

Multiplikátory m_{i1} sú po rade 0, 1/2, 1/2, takže maticu upravíme GEM a miesto núl v prvom stĺpci zapíšeme tieto multiplikátory

$$\begin{pmatrix} 2 & 0 & 2 & 0 \\ 0 & 2 & 0 & 1 \\ 1/2 & 1 & -2 & 2 \\ 1/2 & 3 & 1 & -1 \end{pmatrix}$$

V druhom kroku pokračujeme s hlavnou 3×3 podmaticou. Najväčší prvok v druhom stĺpci je 3 zo štvrtého riadku, takže vymeníme druhý a štvrtý riadok $\Pi_2 = (24)$

$$\begin{pmatrix} 2 & 0 & 2 & 0 \\ 1/2 & 3 & 1 & -1 \\ 1/2 & 1 & -2 & 2 \\ 0 & 2 & 0 & 1 \end{pmatrix}$$

a použijeme GEM na 3×3 podmaticu (pozor, prvky prvého stĺpca neupravujeme, tam sú v skutočnosti 0). Multiplikátory m_{i2} sú po rade 1/3 a 2/3 a po úprave dostaneme

$$\begin{pmatrix} 2 & 0 & 2 & 0 \\ 1/2 & 3 & 1 & -1 \\ 1/2 & 1/3 & -7/3 & 7/3 \\ 0 & 2/3 & -2/3 & 5/3 \end{pmatrix}$$

V treťom, poslednom kroku riadky nemeníme, lebo $|-7/3| > |-2/3|$. Multiplikátor je $m_{i3} = (-2/3)/(-7/3) = 2/7$ a poslednou GEM dostaneme

$$\begin{pmatrix} 2 & 0 & 2 & 0 \\ 1/2 & 3 & 1 & -1 \\ 1/2 & 1/3 & -7/3 & 7/3 \\ 0 & 2/3 & 2/7 & 1 \end{pmatrix}$$

Z toho potom máme

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1/2 & 1 & 0 & 0 \\ 1/2 & 1/3 & 1 & 0 \\ 0 & 2/3 & 2/7 & 1 \end{pmatrix} \qquad U = \begin{pmatrix} 2 & 0 & 2 & 0 \\ 0 & 3 & 1 & -1 \\ 0 & 0 & -7/3 & 7/3 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Permutácia $P = (24)(13)$ vznikne z I postupnou výmenou riadkov 1–3 a 2–4,

$$P = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}.$$

Tým sme získali hľadaný LU rozklad.

Ak máme LU rozklad s čiastočným pivotovaním matice A a potrebujeme nájsť riešenie systému $Ax = b$, rozpíšeme $PAx = LUx = Pb$. Najskôr teda nájdeme y také, že $Ly = Pb$ a potom x je riešením systému $Ux = y$.

Poznámky:

1. LU rozklad s čiastočným pivotovaním regulárnej matice A vždy existuje.

2. Prvky matice L sú v absolútnej hodnote nanajvýš 1.

3. Dá sa teoreticky ukázať, že aj GEM s čiastočným pivotovaním je nestabilná metóda. Avšak v praktických úlohách sa nestabilita tejto metódy prejavuje veľmi zriedka, preto ju z praktického hľadiska považujeme za stabilnú.

Prípad úplného pivotovania už teraz predstavíme veľmi rýchlo. V k-tom kroku nájdeme najväčší prvok príslušnej hlavnej podmatice a výmenou riadka a stĺpca ho presunieme do ľavého horného rohu. Následne eliminujeme prvky v určenom stĺpci pod hlavnou diagonálou. Takže maticovo môžeme postup zapísať ako

$$M_{n-1}\Pi_{n-1}\cdots M_1\Pi_1 A\Gamma_1\cdots\Gamma_{n-1} = U.$$

Z toho podobnými úvahami ako v prípade čiastočného pivotovania dostaneme rozklad

$$PAQ^T = LU.$$

Riešenie rovnice $Ax = b$ potom prebieha nasledovne:

1. Nájsť riešenie $Lz = Pb$ s neznámou z.

2. Nájsť riešenie $Uy = z$ s neznámou y.

3. Spočítať $x = Q^T y$.

GEM s úplným pivotovaním už možno považovať za stabilnú metódu avšak vzhľadom na zriedkavú nestabilitu GEM s čiastočným pivotovaním sa často nevyužíva aj vzhľadom na veľký počet operácií porovnávania.

Medzi novšie metódy patrí GEM s vežovým pivotovaním. Postup pri tomto pivotovaní je nasledovný. V k-tom kroku GEM v príslušnej hlavnej podmatici nájdeme prvok s najväščou absolútnou hodnotou v prvom stĺpci tejto podmatice a následne zistíme, či je tento prvok aj najväčší v tom riadku v ktorom leží. Ak je, stane sa pivotom. Ak nie, vezmeme najväčší prvok z toho istého riadka a zistíme, či je najväčší aj v stĺpci, v ktorom leží. Ak áno, stane sa pivotom, ak nie, vezmeme najväčší prvok z toho stĺpca a zistíme, či je najväčší aj v riadku v ktorom leží,. . . .

Tento postup nakoniec nájde nejaký pivot.

Výhodou tejto metódy je, že počet porovnaní, kým nájdeme pivot býva zvyčajne omnoho menší ako počet porovnaní pre GEM s úplným pivotovaním, ale napriek tomu má táto metóda porovnateľnú numerickú stabilitu ako GEM s úplným pivotovaním.

V tomto prípade dostávame tiež rozklad

$$PAQ^T = LU$$

pretože po nájdení pivota musíme zvyčajne meniť riadok aj stĺpec.

V prípade nedourčených systémov, t.j. systémov s maticou A s počtom riadkov menším ako počet stĺpcov a plnou hodnosťou je zrejme možné spočítať LU rozklad (s úplným alebo vežovým pivotovaním) v tvare

$$PAQ^T = L[U_1|U_2]$$

kde P a Q sú permutácie, L je dolná trojuholníková matica a U_1 je regulárna, horná trojuholníková.

Ak by sme chceli nájsť nejaké riešenie takého systému (ktorých je samozrejme nekonečne veľa), môžeme postupovať nasledovne. Zo systému $Ax = b$ dostaneme

$$(PAQ^T)(Qx) = Pb \Rightarrow L[U_1|U_2](Qx) = Pb$$

Vektor Qx napíšme ako $[z_1|z_2]^T$. Potom

$$L[U_1|U_2](Qx) = Pb \Rightarrow L[U_1|U_2][z_1|z_2]^T = Pb \Rightarrow L(U_1z_1 + U_2z_2) = Pb$$

Potom riešenie nájdeme nasledovne

1. Riešme $Ly = Pb$ s neznámou y.
2. Zvoľme z_2 a riešme $U_1z_1 = y - U_2z_2$ s neznámou z_1.
3. Položme $x = Q^T[z_1, z_2]^T$.

Symetrické matice

Pri aplikácii GEM na symetrické matice si môžeme všimnúť, že po vynulovaní k-teho stĺpca v k-tom kroku zvyšná hlavná podmatica zostáva symetrická a navyše v LU rozklade sú prvky v k-tom riadku matice U napravo od diagonály násobkami prvkov v k-tom stĺpci matice L pod diagonálou.

Veta. (LDL^T rozklad)

Nech A je $n \times n$ regulárna matica, taká, že každá jej hlavná $k \times k$ podmatica je regulárna. Potom existuje dolná trojuholníková matica L s jednotkami na diagonále a diagonálna matica $D = \mathrm{diag}(d_1, \ldots, d_m)$ taká, že

$$A = LDL^T,$$

Pričom tento rozklad je jednoznačný.

Dôkaz. Z daného predpokladu dostávame, že existuje LU rozklad matice A, $A = LU$. Tento rozklad môžeme upraviť nasledovne

$$L^{-1}AL^{-T} = UL^{-T}.$$

Na ľavej strane rovnosti je symetrická matica, na pravej strane je horná trojuholníková matica, takže táto matica musí byť diagonálna, preto

$$UL^{-T} = D \Rightarrow U = DL^T.$$

Z jednoznačnosti LU rozkladu potom dostávame jednoznačnosť LDL^T rozkladu.

Systém $Ax = b$ potom riešime v troch krokoch: $Ax = LDL^Tx = b$

1. $Lz = b$
2. $Dy = z$
3. $L^Tx = y$.

Pokiaľ by sme chceli použiť čiastočné pivotovanie, porušíme symetriu. Ak by sme teda pri výmenách riadkov zachovať symetriu, musíme meniť aj príslušné stĺpce, čo však spôsobí, že na mieste pivota môže byť len prvok z diagonály. Prvky na diagonále však môžu byť všetky nulové aj pri regulárnej matici, ako ukazuje príklad

$$\begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}.$$

V prípade kladne definitných matíc však môžeme použiť nasledovný fakt

Veta. Ak symetrická matica A je kladne definitná, tak všetky jej hlavné podmatice získané pri GEM sú kladne definitné.

Dôkaz. Vezmime kladne definitnú symetrickú maticu

$$A = \begin{pmatrix} a_{11} & a^T \\ a & B \end{pmatrix}.$$

Po prvom kroku GEM získame maticu

$$A = \begin{pmatrix} a_{11} & a^T \\ 0 & B - \frac{1}{a_{11}} a a^T \end{pmatrix}.$$

Vezmime ľubovoľný nenulový vektor z tvaru $z = (x|y^T)^T$ kde $x \in \mathbb{R}$ a y je $n-1$-vektor. Potom z kladnej definitnosti matice A máme

$$0 < z^T A z = x^2 a_{11} + 2a^T yx + y^T By$$

Po úprave GEM pre príslušnú podmaticu máme

$$y^T \left(B - \frac{aa^T}{a_{11}} \right) y = y^T By - \frac{(a^T y)^2}{a_{11}}$$

Potom môžeme odhadnúť

$$y^T By - \frac{(a^T y)^2}{a_{11}} > -\frac{(a^T y)^2}{a_{11}} - x^2 a_{11} - 2a^T yx = -\frac{1}{a_{11}} \left(a_{11} x + a^T y \right)^2$$

Táto nerovnosť platí pre ľubovoľné x, teda aj pre $x = -\frac{a^T y}{a_{11}}$, pre ktoré je výraz n zátvorke rovný 0. T.j. $B - \frac{aa^T}{a_{11}}$ je kladne definitná. Ďalej indukcia.

Vďaka tejto vete vieme, že v diagonálnej matici D sú všetky prvky na diagonále kladné. Môžeme preto D odmocniť, $D = \sqrt{D}\sqrt{D}$ a prepísať

$$A = LDL^T = L\sqrt{D}\sqrt{D}L^T = \bar{L}\bar{L}^T$$

Tým dostávame nový typ rozkladu pre symetrické kladne definitné matice, ktorý voláme Choleského rozklad symetrickej kladne definitnej matice A.

$$A = \bar{L}\bar{L}^T$$

Pokiaľ by sme chceli použiť čiastočné pivotovanie, porušíme symetriu. Ak by sme teda pri výmenách riadkov zachovať symetriu, musíme meniť aj príslušné stĺpce, čo však spôsobí, že na mieste pivota môže byť len prvok z diagonály. Prvky na diagonále však môžu byť všetky nulové aj pri regulárnej matici, ako ukazuje príklad

$$\begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}.$$

V prípade kladne definitných matíc však môžeme použiť nasledovný fakt

Veta. Ak symetrická matica A je kladne definitná, tak všetky jej hlavné podmatice získané pri GEM sú kladne definitné.

Dôkaz. Vezmime kladne definitnú symetrickú maticu

$$A = \begin{pmatrix} a_{11} & a^T \\ a & B \end{pmatrix}.$$

Po prvom kroku GEM získame maticu

$$A = \begin{pmatrix} a_{11} & a^T \\ 0 & B - \frac{1}{a_{11}} a a^T \end{pmatrix}.$$

Vezmime ľubovoľný nenulový vektor z tvaru $z = (x|y^T)^T$ kde $x \in \mathbb{R}$ a y je $n-1$-vektor. Potom z kladnej definitnosti matice A máme

$$0 < z^T A z = x^2 a_{11} + 2a^T y x + y^T B y$$

Po úprave GEM pre príslušnú podmaticu máme

$$y^T \left(B - \frac{a a^T}{a_{11}} \right) y = y^T B y - \frac{(a^T y)^2}{a_{11}}$$

Potom môžeme odhadnúť

$$y^T B y - \frac{(a^T y)^2}{a_{11}} > -\frac{(a^T y)^2}{a_{11}} - x^2 a_{11} - 2a^T y x = -\frac{1}{a_{11}} \left(a_{11} x + a^T y \right)^2$$

Táto nerovnosť platí pre ľubovoľné x, teda aj pre $x = -\frac{a^T y}{a_{11}}$, pre ktoré je výraz n zátvorke rovný 0. T.j. $B - \frac{a a^T}{a_{11}}$ je kladne definitná. Ďalej indukcia.

Vďaka tejto vete vieme, že v diagonálnej matici D sú všetky prvky na diagonále kladné. Môžeme preto D odmocniť, $D = \sqrt{D}\sqrt{D}$ a prepísať

$$A = LDL^T = L\sqrt{D}\sqrt{D}L^T = \bar{L}\bar{L}^T$$

Tým dostávame nový typ rozkladu pre symetrické kladne definitné matice, ktorý voláme Choleského rozklad symetrickej kladne definitnej matice A.

$$A = \bar{L}\bar{L}^T \tag{3}$$

pričom $\bar{L}$ je dolná trojuholníková matica s kladnými prvkami na diagonále.

Aj keď Choleského rozklad dokážeme získať pomocou LDL rozkladu matice A, efektívnejšie je počítať ho porovnaním stĺpcov matíc v rovnosti (3). Pre $n \times n$ maticu A a $1 \leq j \leq n$ máme

$$\begin{aligned} A(:,j) &= \sum_{k=1}^{n} \bar{L}(:,k)\bar{L}^T(k,j) = \sum_{k=1}^{n} \bar{L}(:,k)\bar{L}(j,k) \\ &= \sum_{k=1}^{j} \bar{L}(j,k)\bar{L}(:,k) = \bar{L}(j,j)\bar{L}(:,j) + \sum_{k=1}^{j-1} \bar{L}(j,k)\bar{L}(:,k) \end{aligned}$$

z čoho dostávame

$$\bar{L}(j,j)\bar{L}(:,j) = A(:,j) - \sum_{k=1}^{j-1} \bar{L}(j,k)\bar{L}(:,k)$$

označme pravú stranu ako vektor v

$$A(:,j) - \sum_{k=1}^{j-1} \bar{L}(j,k)\bar{L}(:,k) = v$$

Potom

$$\bar{L}(j,j)\bar{L}(j,j) = v(j) \Rightarrow \bar{L}(j,j) = \sqrt{v(j)}$$

a

$$\bar{L}(j:n,j) = \frac{1}{\sqrt{v(j)}} \left[A(j:n,j) - \sum_{k=1}^{j-1} \bar{L}(j,k)\bar{L}(j:n,k) \right] = \frac{v(j:n)}{v(j)}$$

Opäť je možné ušetriť miesto prepisovaním dolnej trojuholníkovej časti matice A. **Príklad**: Nájdite Choleského rozklad matice

$$A = \begin{pmatrix} 4 & -2 & 4 & 2 \\ -2 & 2 & -5 & -1 \\ 4 & -5 & 22 & 8 \\ 2 & -1 & 8 & 9 \end{pmatrix}$$

Riešenie: Začneme výpočtom $L(1,1)$, čo je zrejme $\sqrt{A(1,1)} = 2$ a prvý stĺpec matice $\bar{L}$ bude $\frac{1}{2}A(1:n,1)$. Prepísaním prvého stĺpca matice A dostaneme

$$\begin{pmatrix} 2 & -2 & 4 & -2 \\ -1 & 2 & -5 & -1 \\ 2 & -5 & 22 & 8 \\ 1 & -1 & 8 & 9 \end{pmatrix}$$

Ďalej pokračujeme výpočtom druhého stĺpca.

$$L(2,2) = \sqrt{A(2,2) - L(2,1)^2} = \sqrt{2 - (-1)^2} = 1$$

a

$$L(2:4,2) = A(2:4,2) - L(2,1)L(2:n,1) = \begin{pmatrix} 2 \\ -5 \\ -1 \end{pmatrix} - (-1)\begin{pmatrix} -1 \\ 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ -3 \\ 0 \end{pmatrix}$$

Dostávame tak

$$\begin{pmatrix} 2 & -2 & 4 & 2 \\ -1 & 1 & -5 & -1 \\ 2 & 3 & 22 & 8 \\ 1 & 0 & 8 & 9 \end{pmatrix}$$

Pre tretí stĺpec máme

$$L(3,3) = \sqrt{A(3,3) - L(3,1)^2 - L(3,2)^2} = \sqrt{22 - 4 - 9} = 3$$

a

$$L(3:4,3) = \frac{1}{3}\left[\begin{pmatrix} 22 \\ 8 \end{pmatrix} - 2\begin{pmatrix} 2 \\ 1 \end{pmatrix} - 3\begin{pmatrix} 3 \\ 0 \end{pmatrix}\right] = \begin{pmatrix} 3 \\ 2 \end{pmatrix}$$

čiže

$$\begin{pmatrix} 2 & -2 & 4 & 2 \\ -1 & 1 & -5 & -1 \\ 2 & 3 & 3 & 8 \\ 1 & 0 & 2 & 9 \end{pmatrix}$$

Nakoniec už len stačí spočítať

$$L(4,4) = \sqrt{A(4,4) - L(4,1)^2 - L(4,2)^2 - L(4,3)^2} = \sqrt{9 - 4 - 1} = 2$$

a máme výsledok

$$\begin{pmatrix} 2 & -2 & 4 & 2 \\ -1 & 1 & -5 & -1 \\ 2 & 3 & 3 & 8 \\ 1 & 0 & 2 & 2 \end{pmatrix}$$

teda rozklad

$$A = \begin{pmatrix} 4 & -2 & 4 & 2 \\ -2 & 2 & -5 & -1 \\ 4 & -5 & 22 & 8 \\ 2 & -1 & 8 & 9 \end{pmatrix} = \begin{pmatrix} 2 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 2 & 3 & 3 & 0 \\ 1 & 0 & 2 & 2 \end{pmatrix} \begin{pmatrix} 2 & -1 & 2 & 1 \\ 0 & 1 & 3 & 0 \\ 0 & 0 & 3 & 2 \\ 0 & 0 & 0 & 2 \end{pmatrix}$$

V prípade presnej aritmetiky vieme, že kladne definitná štvorcová matica má Choleského rozklad. Navyše platí aj opačné tvrdenie v nasledujúcej forme. Ak vo vyššie uvedenom algoritme sú všetky odmocniny kladné reálne čísla, tak daná matica je kladne definitná. Choleského rozklad je preto možné v pricípe použiť aj na zistenie, či daná matica je kladne definitná. Z hľadiska numerickej stability je zaujímavá nerovnosť

$$a_{ii} = \sum_{k=1}^{i} l_{ik}^2 \geq l_{ik}^2$$

Takže prvky matice $\bar{L}$ sú pekne ohraničené.

LTL^T rozklad

V prípade symetrických indefinitných matíc sme videli, že LDL^T rozklad existovať nemusí. A aj v prípade, že existuje, zlepšenie stability pivotovaním bez straty symetrie môžeme len výberom pivota z diagonály. Jedna z možností, ako umožniť pivotovanie aj inými prvkami ako diagonálnymi je LTL^T rozklad matice

$$PAP^T = LTL^T$$

kde P je permutačná matica, L je dolná trojuholníková s jednotkami na diagonále a T je trojdiagonálna matica (matica v ktorej sa nenulové prvky môžu nachádzať len na diagonále a nad a pod diagonálou).

Pri použití LTL^T rozkladu na riešenie systému rovníc $Ax = b$ sa postupuje nasledovne

$$Lz = Pb,\ Tw = z,\ L^T y = w,\ x = P^T y.$$

Uvedieme dva algoritmy na výpočet tohoto rozkladu. Prvý je založený na klasickej GEM s pivotovaním.

Parlettov-Riedov algoritmus

Nech A je matica $n \times n$. Pomocou Gaussových operácií budeme postupne vytvárať trojdiagonálnu maticu. V prvom kroku vezmeme časť $A(2:n,1)$ prvého stĺpca matice A a nájdeme permutačnú maticu $\tilde{P}$ takú že vektor $\tilde{v} = \tilde{P}A(2:n,2)$ bude mať na prvom mieste v absolútnej hodnote najväčší prvok z prvkov vektora $A(2:n,1)$

$$|v(1)| = \max_{2 \le i \le n} \{|A(i,1)|\}.$$

Potom zostrojíme permutačnú maticu $P_1 = \mathrm{diag}\{1, \tilde{P}\}$ a maticu $M_1 = I - m_1 e_2^T$, kde m_1 je vektor multiplikátorov $m_1 = (0 \; v(2)/v(1) \; v(3)/v(1) \; \cdots)$. Potom vytvoríme klasicky maticu

$$A^{(1)} = M_1 P_1 A P_1^T M_1^T = \begin{pmatrix} \alpha_1 & \beta_1 & 0 & \cdots \\ \beta_1 & \alpha_2 & v_3 & \cdots \\ 0 & u_3 & \cdots & \cdots \\ 0 & u_4 & \cdots & \cdots \\ \vdots & \vdots & & \end{pmatrix}$$

Tento postup opakujeme pre podmaticu $A(2:n,2:n)$. Po $n-2$ krokoch dostaneme

$$T = A^{(n-2)} = M_{n-2} P_{n-2} \cdots M_1 P_1 A P_1^T M_1^T \cdots P_{n-2}^T M_{n-2}^T$$

z čoho ľahko odvodíme hľadaný rozklad

$$PAP^T = LTL^T$$

kde $P = P_{n-2} \cdots P_1$ a $L = (M_{n-2} P_{n-2} \cdots M_1 P_1 P^T)^{-1}$. Môžeme si všimnúť, že matica L má prvý stĺpec e_1.

Aasenova metóda

Budeme hľadať rozklad (zatiaľ bez pivotovania)

$$A = LTL^T$$

kde L je dolná trojuholníková a $L(:,1) = e_1$ a

$$\begin{pmatrix} \beta_1 & \alpha_1 & \cdots & \cdots & 0 \\ \alpha_1 & \beta_1 & & & \\ \vdots & \ddots & \ddots & \ddots & \\ & & & & \beta_{n-1} \\ 0 & & & \beta_{n-1} & \alpha_n \end{pmatrix}$$

Položme $H = TL^T$. To je tzv. horná Hessenbergova matica (matica, ktorá môže mať nenulové prvky najviac jeden krok pod diagonálou). Potom $A = LH$ a

$$A(:,j) = LH(:,j) = \sum_{k=1}^{j+1} L(:,k)H(k,j) = \sum_{k=1}^{j+1} L(:,k)h(k)$$

pričom sme označili vektor $h(1:j+1) = H(1:j+1,j)$ a $j \leq n-1$. Potom

$$\begin{aligned} A(j+1:n,j) &= \sum_{k=1}^{j+1} L(j+1:n,k)h(k) \\ &= L(j+1:n,j+1)h(j+1) + \sum_{k=1}^{j} L(j+1:n,k)h(k) \\ &= L(j+1:n,j+1)h(j+1) + L(j+1:n,1:j)h(1:j) \end{aligned}$$

teda

$$h(j+1)L(j+1:n,j+1) = A(j+1:n,j) - L(j+1:n,1:j)h(1:j)$$

Táto rovnica umožňuje spočítať $j+1$-vý stĺpec matice L ak poznáme všetky predchádzajúce stĺpce matice L a j-ty stĺpec matice H. Ak označíme

$$v(j+1:n) = A(j+1:n,j) - L(j+1:n,1:j)h(1:j)$$

dostaneme $v(j+1) = h(h+1)$, lebo $L(j+1,j+1) = 1$. Z toho potom dostávame

$$L(j+2:n,j+1) = \frac{v(j+2:n)}{v(j+1)}$$

V tomto kroku je z hľadiska stability vhodné, aby $v(j+1)$ bolo čo najväčšie, čo vieme dosiahnuť vhodnou permutáciou. Z predpisu pre v vidíme, že túto permutáciu je potrebné aplikovať na A aj na L.

Pozrime sa teraz, ako v j-tom kroku získať prvky $h(1:j+1)$. Na to využijeme vzťah $H = TL^T$.

$$\begin{pmatrix} h(1) \\ h(2) \\ \vdots \\ h(j+1) \end{pmatrix} = \begin{pmatrix} \alpha_1 & \beta_1 & & & & & \\ \beta_1 & \alpha_2 & \beta_2 & & & & \\ 0 & \beta_2 & \alpha_3 & \beta_3 & & & \\ & & & & & & \\ & & & & \beta_{j-2} & \alpha_{j-1} & \beta_{j-1} \\ & & & & & \beta_{j-1} & \alpha_j \\ & & & & & & \beta_j \end{pmatrix} \begin{pmatrix} 0 \\ l_{j2} \\ \vdots \\ l_{j\,j-1} \\ 1 \end{pmatrix}$$

Na začiatku kroku j poznáme $\alpha(1:j-1)$, $\beta(1:j-1)$ a $L(:,1:j)$. Poznáme teda prvých $j-2$ riadkov (časti) matice T. Z toho dostávame vzťahy pre $h(k)$, $1 \leq k \leq j-1$

$$\begin{aligned} h(1) &= \beta_1 l_{j2} \\ h(k) &= \beta_{k-1} l_{j\,k-1} + \alpha_k l_{j\,k} + \beta_k l_{j\,k+1}, \qquad 2 \leq k \leq j-1 \end{aligned}$$

Hodnotu $h(j)$ získame z $A = LH$

$$h(j) = A(j,j) - \sum_{k=1}^{j-1} L(j,k)h(k)$$

následne vieme spočítať

$$\begin{aligned} \alpha_j &= h(j) - \beta_{j-1} l_{j\,j-1} \\ \beta_j &= h(j+1) = v(j+1). \end{aligned}$$

Maticové normy-opakovanie

Označme K pole reálnych čísel a $K^{m\times n}$ vektorový priestor $m \times n$ matíc nad poľom K. Maticová norma je zobrazenie $||\cdot|| : K^{m\times n} \to \mathbb{R}$, ktorá spĺňa nasledovné vlastnosti:

Pre ľubovoľné $\alpha \in K$ a $A, B \in K^{m\times n}$,

1. $||\alpha A|| = |\alpha|\,||A||$
2. $||A + B|| \leq ||A|| + ||B||$
3. $||A|| \geq 0$ a $||A|| = 0 \Leftrightarrow A = 0$.
4. Pre štvorcové matice A, B platí $||AB|| \leq ||A||\,||B||$

Lema. Označme $\rho(A)$ spektrálny polomer (štvorcovej) matice A, t.j. najväčšie vlastné číslo v absolútnej hodnote. Potom

$$\rho(A) \leq ||A|| \tag{4}$$

pre každú maticovú normu.

Dôkaz. Označme $B = (v\ v\ \cdots v)$ maticu, ktorá má stĺpce rovné vlastnému vektoru v matice A. Potom $AB = \lambda B$, kde λ je vlastná hodnota prislúchajúca vlastnému vektoru v. Z toho

$$|\lambda|\,||B|| = ||\lambda B|| = ||AB|| \leq ||A||\,||B||$$

takže $|\lambda| \leq ||A||$ pre ľubovoľnú vlastnú hodnotu λ, teda aj pre maximálnu.

Príklady maticových noriem

Operátorová norma Ak $||\cdot||$ označuje vektorová norma na K^m a na K^n, pre maticu $A \in K^{m\times n}$ definujeme

$$||A|| = \sup\left\{\frac{||Ax||}{||x||}; x \in K^n, x \neq 0\right\}$$

Špeciálne pri použití p-noriem na K^n aj K^M, $(1 \leq p \leq \infty)$ dostávame

$$||A||_p = \sup_{x\neq 0}\frac{||Ax||}{||x||} = \sup_{||x||_p=1}||Ax||$$

Dá sa ukázať

1. $||A||_1$ je maximum zo súčtov absolútnych hodnôt prvkov v stĺpcoch matice A.
2. $||A||_\infty$ je maximum zo súčtov absolútnych hodnôt prvkov v riadkoch matice A.
3. $||A||_2 = \sqrt{\lambda_{max}(AA^*)}$

Frobéniová norma Pre maticu $A = (a_{ij}) \in K^{m\times n}$

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

Stabilita riešenia regulárnej sústavy

Začneme dvoma príkladmi

Príklad 1: Majme systém $Ax = b$

$$\left(\begin{array}{cc|c} 1/1000 & 1 & 1+1/1000 \\ 1 & 1 & 2 \end{array}\right)$$

Jeho riešením je zrejme $(1,1)^T$. Pokiaľ trochu pozmeníme pravú stranu, $A\tilde{x} = \tilde{b}$

$$\left(\begin{array}{cc|c} 1/1000 & 1 & 1 \\ 1 & 1 & 2 \end{array}\right)$$

získame riešenie $(1000/999, 998/999)^T$.

Relatívna chyba pravej strany je

$$|\Delta b|/|b| = |b - \tilde{b}|/|b| \approx 0.447 \cdot 10^{-3}$$

a relatívna chyba riešenia

$$|\Delta x|/|x| = |x - \tilde{x}|/|x| \approx 0.1 \cdot 10^{-2},$$

takže pomer relatívnej chyby riešenia ku relatívnej chybe pravej strany je

$$\frac{|\Delta x|/|x|}{|\Delta b|/|b|} \approx 2.24$$

Vidíme, že malá relatívna chyba pravej strany sa prejaví malou relatívnou chybou v riešení.

Príklad 2: Majme systém $Ax = b$

$$\left(\begin{array}{cc|c} 1 + 1/1000 & 1 & 2 + 1/1000 \\ 1 & 1 & 2 \end{array}\right)$$

Jeho riešením je zrejme $(1, 1)^T$. Ak však pozmeníme pravú stranu

$$\left(\begin{array}{cc|c} 1 + 1/1000 & 1 & 2 \\ 1 & 1 & 2 \end{array}\right)$$

dostaneme ako riešenie $(0, 2)^T$. Podobnou analýzou ako v predošlom príklade dostaneme

$$|\Delta b|/|b| = |b - \tilde{b}|/|b| \approx 0.353 \cdot 10^{-3}$$
$$|\Delta x|/|x| = |x - \tilde{x}|/|x| = 1$$
$$\frac{|\Delta x|/|x|}{|\Delta b|/|b|} \approx 2829.1$$

takže relatívna chyba riešenia je v tomto prípade pri veľmi malej zmene pravej strany skoro 3000-násobkom relatívnej chyby pravej strany.

Pripomeňme, že dobre podmienený systém rovníc je taký systém rovníc, v ktorom malá zmena matice koeficientov a/alebo pravej stany spôsobí len malú zmenu v riešení. Naopak zle podmienený systém rovníc je taký systém rovníc, v ktorom malá zmena koeficientov matice a/alebo pravej strany spôsobí veľkú zmenu riešenia systému. Podmienenosť systému nám teda hovorí, nakoľko môžeme veriť získanému riešeniu.

Ideu toho, čo sa vlastne deje nám môže dať nasledovná úvaha.

Majme systémy rovníc $A\bar{x} = b$ a $A\tilde{x} = \tilde{b}$, kde $\tilde{b} = b + \Delta b$. Potom

$$\tilde{x} = A^{-1}\tilde{b} = A^{-1}(b + \Delta b) = A^{-1}b + A^{-1}\Delta b = \bar{x} + A^{-1}\Delta b.$$

Takže ak A^{-1} bude mať veľké prvky, tak $A^{-1}\Delta b$ môže byť veľké aj pre malé Δb. V príkladoch vyššie máme

$$\begin{pmatrix} 1/1000 & 1 \\ 1 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} -\frac{1000}{999} & \frac{1000}{999} \\ \frac{1000}{999} & -\frac{1}{999} \end{pmatrix}$$

a

$$\begin{pmatrix} 1 + 1/1000 & 1 \\ 1 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 1000 & -1000 \\ -1000 & 1001 \end{pmatrix}$$

Pozrime sa teda podrobnejšie na vzťah sústavy riešenia s presnou a pozmenenou maticou sústavy resp. pravou stranou. Máme teda systémy

$$A\bar{x} = b \tag{5}$$

$$(A + \Delta A)\tilde{x} = b + \Delta b \tag{6}$$

Pri pohľade na (6) vidíme problém. A je regulárna, ale $A + \Delta A$ už regulárna byť nemusí. Aby sme mohli pokračovať, potrebujeme zabezpečiť regularitu tejto pozmenenej matice. Platia však nasledujúce ekvivalencie

$$\begin{aligned} A + \Delta A \text{ je regulárna } &\Leftrightarrow A(I + A^{-1}\Delta A) \text{ je regulárna} \\ &\Leftrightarrow I + A^{-1}\Delta A \text{ je regulárna} \end{aligned}$$

To nám umožňuje využiť nasledovné tvrdenie

Veta. Ak $||X|| < 1$, tak $I - X$ je regulárna a $||I - X|| \leq \frac{1}{1-||X||}$

Dôkaz. Podľa predpokladu je $1 > ||X|| \geq \rho(X)$. Potom ale $I - X$ je regulárna, pretože X je podobná so svojim Jordanovým kanonickým tvarom $J(X) = PXP^{-1}$, ktorý má vlastné hodnoty na diagonále. Potom ale

$$I - X = I - P^{-1}J(X)P = P^{-1}(I - J(X))P$$

pričom P je regulárna a $I - J(X)$ je regulárna, pretože je horná trojuholníková s nenulovými prvkami na diagonále, keďže $\rho(X) < 1$.

Ďalej máme

$$\begin{aligned} (I - X)^{-1}(I - X) &= I \\ (I - X)^{-1} - (I - X)^{-1}X &= I \\ (I - X)^{-1} &= I + (I - X)^{-1}X \end{aligned}$$

Normovaním poslednej rovnosti dostaneme

$$||(I-X)^{-1}|| = || = I + (I-X)^{-1}X|| \leq ||I|| + ||(I-X)^{-1}||\,||X||$$

a úpravou

$$||(I-X)^{-1}|| - ||(I-X)^{-1}||\,||X|| \leq 1$$
$$||(I-X)^{-1}|| \leq \frac{1}{1-||X||}$$

dostávame tvrdenie.

Predpokladajme teda, že $||A^{-1}\Delta A|| \leq 1$. Potom

$$\Delta x = \tilde{x} - \bar{x} = (A+\Delta A)^{-1}(b+\Delta b) - \bar{x} = (A+\Delta A)^{-1}[b+\Delta b - (A+\Delta A)\bar{x}]$$
$$= (A+\Delta A)^{-1}(\Delta b - \Delta A\bar{x}) = (I + A^{-1}\Delta A)^{-1}A^{-1}(\Delta b - \Delta A\bar{x})$$

Po aplikácii normy a predošlého tvrdenia dostaneme

$$||\Delta x|| \leq \frac{||A^{-1}||}{1-||A^{-1}\Delta A||}(||\Delta b|| + ||\Delta A||\,||\bar{x}||)$$

Označme relatívne chyby

$$R(A) = \frac{||\Delta A||}{||A||}, \qquad R(b) = \frac{||\Delta b||}{||b||}.$$

Potom

$$||\Delta x|| \leq \frac{||A^{-1}||}{1-||A^{-1}\Delta A||}(R(b)||b|| + R(A)||A||\,||\bar{x}||)$$
$$= \frac{||A||\,||A^{-1}||}{1-||A^{-1}\Delta A||}\left(R(b)\frac{||b||}{||A||} + R(A)||\bar{x}||\right)$$

Z toho dostávame odhad pre relatívnu chybu riešenia

$$R(\bar{x}) = \frac{||\Delta x||}{||\bar{x}||} = \frac{||A||\,||A^{-1}||}{1-||A^{-1}\Delta A||}\left(R(b)\frac{||b||}{||A||\,||\bar{x}||} + R(A)\right)$$
$$\leq \frac{||A||\,||A^{-1}||}{1-||A^{-1}\Delta A||}\,(R(b)+R(A))$$

Ak sprísnime podmienku regularity $||A^{-1}\Delta A|| \leq ||A^{-1}||\,||\Delta A|| < 1$, tak

$$R(\bar{x}) \leq \frac{||A||\,||A^{-1}||}{1-||A^{-1}||\,||\Delta A||}\,(R(b)+R(A)) = \frac{||A||\,||A^{-1}||}{1-||A^{-1}||\,||\Delta A||}\,(R(b)+R(A))$$

Označme $c(A) = ||A||\,||A^{-1}||$. Potom

$$R(\bar{x}) \leq \frac{c(A)}{1 - c(A)R(A)}\,(R(b) + R(A))$$

Dosiahnutý výsledok zhrnieme v nasledujúcej vete

Veta a definícia. (Odhad relatívnej chyby riešenia) Nech A je regulárna matica a nech pre maticu ΔA platí $||A||\,||\Delta A|| < 1$. Potom $A + \Delta A$ je regulárna a pre relatívnu chybu riešenia sústavy $(A + \Delta A)\tilde{x} = (b + \Delta b)$ vzhľadom na sústavu $A\bar{x} = b$ platí

$$R(x) \leq \frac{c(A)}{1 - c(A)R(A)}\,(R(b) + R(A))$$

Číslo $c(A)$ nazývame číslo podmienenosti matice A.

Poznámka:

- $c(A)$ závisí od použitej normy.
- V prípade veľkého rozdielu v prvkoch A a A^{-1} bude $c(A)$ veľké v ľubovoľnej norme.
- $c(A) \gg 1$ implikuje zlú podmienenosť.

Pre relatívnu chybu riešenia systému lineárnych rovníc $Ax = b$ so štvorcovou regulárnou maticou A sme našli odhad

$$R(x) \leq \frac{c(A)}{1 - c(A)R(A)}\,(R(b) + R(A)) \tag{7}$$

Prirodzená otázka samozrejme je, nakoľko je tento odhad dobrý a či ho nie je možné zlepšiť. Nasledujúci príklad ukazuje, že odpoveď na druhú otázku je nie,

Príklad: Pozrime sa na systém $Ax = b$ s kde

$$A = \begin{pmatrix} 1 & 0.99 \\ 0.99 & 0,98 \end{pmatrix} \qquad \text{a} \qquad b = \begin{pmatrix} 1.99 \\ 1,97 \end{pmatrix}.$$

Hneď vidíme, že presné riešenie je $\bar{x} = (1,\ 1)^T$.

Vezmime teraz systém $Ax = b+\Delta b$, pričom $\Delta b = (10^{-4},\ 10^{-4})^T$. Riešením takto pozmeneného systému je $x = (-0.97,\ 2.99)^T$. Z toho spočítame $\Delta x = x - \bar{x} = (-1.97,\ 1.99)^T$ a relatívnu chybu riešenia

$$R_\infty(x) = \frac{||\Delta x||_\infty}{||x||_\infty} = 1.99$$

Aby sme mohli použiť nájdený všeobecný odhad, potrebujeme spočítať číslo podmienenosti $c(A)$. Nájdeme preto inverznú maticu k matici A a normy matíc A a A^{-1}

$$A^{-1} = \begin{pmatrix} -9800 & 9900 \\ 9900 & -10000 \end{pmatrix}, \qquad ||A||_\infty = 1.99, \qquad ||A^-1||_\infty = 19900.$$

Z toho spočítame

$$c_\infty(A) = 1.99 \cdot 10^4$$

V našej situácii máme $R(A) = 0$ a $R_\infty(b) = 10^-4/1.99$, takže dostávame

$$1.99 = R_\infty(x) \leq c(A)R(b) = 1.99^2 \cdot 10^4 \cdot 10^4/1.99 = 1.99$$

Takže v tejto situácii sa uvedený odhad dosahuje.

Z uvedeného príkladu vidno, že získaný odhad relatívnej chyby riešenia nie je možné vylepšiť. Na druhej strane však nemôžeme povedať, že veľké číslo podmienenosti automaticky znamená, že malé chyby v pravej strany (alebo matice) sa automaticky prejavia vo veľkej relatívnej chybe riešenia, čo ukazuje nasledujúci príklad.

Príklad: Vezmime systém $Ax = b$ s kde

$$A = \begin{pmatrix} 1 & 0.99 \\ 0.99 & 0,98 \end{pmatrix} \qquad \text{a} \qquad b = \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

Opäť vezmime $\Delta b = (10^{-4},\ 10^{-4})^T$. Potom riešenie systému $Ax = b + \Delta b$ môžeme nájsť ako

$$x = A^{-1}b + A^{-1}\Delta b = (-1.97 \cdot 10^4, 1.99 \cdot 10^4)^T + A^{-1}\Delta b$$

Potom $A^{-1}b$ je zrejme presné riešenie $\bar{x}$ a $A^{-1}\Delta b = \Delta x$. Z toho pre relatívnu chybu riešenia máme

$$\begin{aligned} R_\infty(x) = \frac{||x||_\infty}{||\bar{x}||_\infty} = \frac{||A^{-1}\Delta b||_\infty}{||\bar{x}||_\infty} \leq \frac{||A^{-1}||_\infty||\Delta b||_\infty}{||\bar{x}||_\infty} &= \frac{1.99 \cdot 10^4}{1.99 \cdot 10^4}||\Delta b||_\infty \\ &= ||\Delta b||_\infty = \frac{||\Delta b||_\infty}{||b||_\infty} = R_\infty(b) \end{aligned}$$

čiže

$$R_\infty(x) \leq R_\infty(b)$$

Takže v tomto prípade chyba riešenia nepresiahne chybu pravej strany. Avšak odhad chyby z podmienenosti matice dáva

$$R_\infty(x) \leq c_\infty(A) R_\infty(b) = 1.99 \cdot 10^4 ||b||_\infty$$

Z druhého príkladu vidno, že napriek zlej podmienenosti matice chyba riešenia môže byť malá. Číslo podmienenosti teda nehovorí veľa o presnosti získaného riešenia, ale skôr o tom, nakoľko môžeme veriť, že získané riešenie je blízko presnému riešeniu. Veľké číslo podmienenosti teda znamená veľkú neistotu v presnosti riešenia.

Ako sme videli, voľba pravej strany má vplyv na relatívnu chybu riešenie a skutočná relatívna chyba riešenia sa môže od odhadu (7) výrazne líšiť. Prirodzená otázka teda je, ako veľmi sa skutočná relatívna chyba líši od odhadu (7) a ako na ňu vplýva pravá strana.

Odpoveď na túto otázku získame použitím singulárneho rozkladu. Z rozkladu $n \times n$ matice A máme $A = U\Sigma V^T$, kde $\Sigma = \mathrm{diag}(d_1, \ldots, d_n)$, $d_1 \geq d_2 \geq \cdots \geq d_n$. Potom $AV = \Sigma U$, preto $Av_i = d_i u_i$, pre $i = 1, \ldots, n$.

Uvažujme systémy

$$Ax = b, \qquad\qquad Ax = b + \Delta b. \tag{8}$$

a) Zvoľme $b = u_1$ a $\Delta b = \varepsilon u_n$. Potom riešenia sústav

$$U\Sigma V^T x = u_1, \qquad\qquad U\Sigma V^T x = u_1 + \varepsilon u_n$$

sú

$$x = V\Sigma^{-1}U^T u_1 = V\Sigma^{-1}e_1 = \frac{1}{d_1}Ve_1 = \frac{1}{d_1}v_1$$

a

$$x + \Delta x = V\Sigma^{-1}U^T(u_1 + \varepsilon u_n) = \frac{1}{d_1}v_1 + \varepsilon V\Sigma^{-1}U^T u_n = \frac{1}{d_1}v_1 + \varepsilon\frac{1}{d_n}v_n.$$

Z toho ľahko získame Δx a po aplikácii 2-normy dostaneme

$$\|\Delta x\|_2 = \varepsilon\frac{1}{d_n}\|v_n\|_2 = \frac{\varepsilon}{d_n}$$
$$\|x\|_2 = \frac{1}{d_1},$$

a preto relatívna chyba riešenia v 2-norme je

$$R_2(x) = \frac{\|\Delta x\|_2}{\|x\|_2} = \varepsilon \frac{d_1}{d_n} = \varepsilon c_2(A).$$

Ak si uvedomíme, že $R_2(b) = \frac{\|\Delta b\|_2}{\|b\|_2} = \varepsilon$, relatívna chyba riešenia bude

$$R_2(x) = \varepsilon c_2(A) = c_2(A) R_2(b).$$

b) Zvoľme teraz $b = u_n$ a Δb ľubovoľné také, aby $R(b) = \varepsilon < 1$. Potom riešenia sústav (8) sú

$$x = \frac{1}{d_n} v_n, \qquad x + \Delta x = \frac{1}{d_n} v_n + V \Sigma^{-1} U^T \Delta b.$$

Preto

$$\|x\|_2 = \frac{1}{d_n}, \qquad \text{a} \qquad \|\Delta x\|_2 = \|V \Sigma^{-1} U^T \Delta b\|_2 \leq \frac{1}{d_n} \|\Delta b\|_2.$$

Potom pre relatívnu chybu riešenia platí

$$R_2(x) \leq \frac{1}{d_n} \|\Delta b\|_2 \Big/ \frac{1}{d_n} = \|\Delta b\|_2 < 1.$$

Ak zvolíme $\Delta b = \varepsilon u_1$, tak

$$R_2(x) = \|\frac{\varepsilon v_1}{d_1}\|_2 \Big/ \|\frac{v_n}{d_n}\|_2 = \varepsilon \frac{d_n}{d_1} = \frac{\varepsilon}{c_2(A)}.$$

Z týchto úvah vidno, že relatívna chyba riešenia môže byť za vhodných okolností omnoho menšia než chyba pravej strany. Tieto dva prípady sú v skutočnosti hraničné, teda máme odhady

$$\frac{R_2(b)}{c_2(A)} \leq R_2(x) \leq c_2(A) R(b). \tag{9}$$

Číslo podmienenosti symetrických matíc

Pre 2-normu regulárnej matice A platí

$$\|A\|_2 = \left[\rho(A^T A)\right]^{\frac{1}{2}},$$

kde ρ je spektrálny polomer. Pre symetrickú maticu A potom dostaneme

$$\rho(A^T A) = \rho(A^2) = |\lambda|^2_{max}$$

t.j. druhú mocninu najväčšej vlastnej hodnoty matice A, preto $\|A\|_2 = |\lambda|_{max}$.

Podobne sa odvodí aj $\|A^{-1}\|_2 = \left|\frac{1}{\lambda}\right|_{max} = \frac{1}{|\lambda|_{min}}$. Číslo podmienenosti symetrickej matice A v 2-norma sa teda spočíta ako

$$c_2(A) = \frac{|\lambda|_{max}}{|\lambda|_{min}}.$$

Pre nesymetrické matice a ľubovoľnú normu platí vždy odhad

$$\|A\| \geq \rho(A) = |\lambda|_{max}$$

a

$$\|A^{-1}\| \geq \rho(A^{-1}) = \frac{1}{|\lambda|_{min}}.$$

Preto pre číslo podmienenosti v ľubovoľnej norme dostávame dolný odhad

$$c(A) \geq \frac{|\lambda|_{max}}{|\lambda|_{min}}.$$

Teda ak podiel absolútnych hodnôt najväčšieho a najmenšieho vlastného čísla matice A je veľký, matica je zle podmienená.

Poznámka: Ak riešime systém $Ax = b$ použitím LU-rozkladu, môže sa stať, že chyba riešenia bude väčšia než $c(A)R(b)$, pretože matice L a U môžu mať horšie čísla podmienenosti. Preto pri zle podmienených maticiach je vhodnešie použiť QR-rozklad.

Hranice rezidua ako kritérium presnosti

Klasický postup pre overenie, či nájdene riešenie systému rovníc je naozaj riešením je dosadenie riešenia do systému (tzv. skúška správnosti). Čiže pre systém $Ax = b$ nájdeme riešenie x a toto riešenie spĺňa $Ax = b$. Samozrejme pri vzniku chyby v riešení (napr. zaokrúhľovaním, alebo inou stratou presnosti) nájdené riešenie x nebude vyhovovať danému systému. Čo ale vieme zistiť je, ako veľmi sa líši hodnota $A\bar{x}$ od pravej strany b. Tento rozdiel nazveme reziduum a označíme ho r, teda máme

$$r = b - Ax$$

Prirodzene vyvstáva otázka, či reziduum podáva dobrú informáciu o presnosti riešenia. Ukazuje sa, že nie

Príklad: Uvažujme systém $Ax = b$ v maticovom tvare

$$\left(\begin{array}{cc|c} 1 + \frac{1}{1000} & 1 & 2.001 \\ 1 & 1 & 2 \end{array}\right).$$

Presné riešenie tohoto systému je $\bar{x} = (1,1)^T$.

Vezmime $x^{(1)} = (1.5, 0.5)^T$ a $x^{(2)} = (0.99, 0.99)^T$ a spočítajme pre ne rezidua

$$r^{(1)} = b - Ax^{(1)} = \begin{pmatrix} 2.001 \\ 2 \end{pmatrix} - \begin{pmatrix} 2,0015 \\ 2 \end{pmatrix} = \begin{pmatrix} -0.0005 \\ 0 \end{pmatrix}$$
$$r^{(2)} = b - Ax^{(2)} = \begin{pmatrix} 2.001 \\ 2 \end{pmatrix} - \begin{pmatrix} 1.98099 \\ 1,98 \end{pmatrix} = \begin{pmatrix} 0.02001 \\ 0.02 \end{pmatrix}.$$

Potom pre veľkosti reziduií v 2-norme platí

$$\|r^{(1)}\|_2 < \|r^{(2)}\|_2.$$

Takže ak by sme reziduum brali ako kritérium presnosti riešenia, v tomto prípade by nám indikovalo, že $x^{(2)}$ je menej presné riešenie než $x^{(1)}$, napriek tomu, že $x^{(2)}$ je zjavne v 2-norme bližšie k presnému riešeniu než $x^{(1)}$.

Že reziduum nie je dobré kritérium presnosti môžeme nahliadnuť aj z nasledovných úvah. Ak označíme x vypočítané riešenie a $\bar{x}$ presné riešenie systému $Ax = b$, tak máme

$$r = b - Ax = A\bar{x} - Ax = A(\bar{x} - x) = A\Delta x,$$

teda $\Delta x = A^{-1}b$ a Δx môže byť veľké aj pre malé r.

Zo znalosti rezidua môžeme dostať nasledovné odhady relatívnej chyby riešenia.

$$R(x) = \frac{\|\Delta x\|}{\|\bar{x}\|} = \frac{\|A^{-1}r\|}{\|\bar{x}\|} \leq \frac{\|A^{-1}\|\|r\|}{\|\bar{x}\|} = \frac{\mathrm{c}(A)}{\|A\|}\frac{\|r\|}{\|\bar{x}\|} \leq \frac{\mathrm{c}(A)\|r\|}{\|A\bar{x}\|} = \frac{\mathrm{c}(A)\|r\|}{\|b\|}$$
$$R(x) = \frac{\|\Delta x\|}{\|\bar{x}\|} = \frac{\|A\|\|\Delta x\|}{\|A\|\|\bar{x}\|} \geq \frac{\|A\Delta x\|}{\|A\|\|\bar{x}\|} = \frac{\|r\|}{\|A\|\|A^{-1}b\|} \geq \frac{\|r\|}{\|A\|\|A^{-1}\|\|b\|} = \frac{\|r\|}{\mathrm{c}(A)\|b\|}.$$

Teda

$$\frac{1}{\mathrm{c}(A)}\frac{\|r\|}{\|b\|} \leq R(x) \leq \mathrm{c}(A)\frac{\|r\|}{\|b\|}. \tag{10}$$

Iteračné spresnenie riešenia

Napriek tomu že reziduum nemožno vždy dobre využiť na určenie presnosti riešenia, je možné ho výhodne použiť na spresnenie už nájdeného približného riešenia. Idea je nasledovná.

Predpokladajme, že sme našli približné riešenie $x^{(1)}$ systému $Ax = b$. Potom presné riešenie $\bar{x}$ je možné zapísať ako $\bar{x} = x^{(1)} + \Delta x^{(1)}$. Potom

$$b = A\bar{x} = Ax^{(1)} + A\Delta x^{(1)},$$

a teda

$$A\Delta x^{(1)} = b - Ax = r^{(1)}, \tag{11}$$

teda $\Delta x^{(1)}$ je riešením systému s maticou A a pravou stranou rovnou reziduu $r^{(1)}$. Numerickým riešením tohoto systému získame približne hodnotu $\Delta\tilde{x}^{(1)}$.Potom môžeme položiť $x^{(2)} = x^{(1)} + \Delta\tilde{x}^{(1)}$, pričom nové reziduum bude $r^{(2)} = b - Ax^{(2)}$ a celý postup môžeme zopakovať. Celý algoritmus môžeme zhrnúť v do niekoľkých krokov.

- Zvoľme $x^{(0)} = 0$ a $r^{(0)} = b$.
- Pre $k = 1, 2, \ldots$
  a) Vypočítame $\Delta\tilde{x}^{(k)}$ ako riešenie sústavy $Ax = r^{(k)}$.
  b) Vypočítame $x^{(k+1)} = x^{(k)} + \Delta\tilde{x}^{(k)}$
  c) Vypočítame $r^{(k+1)} = b - Ax^{(k+1)}$.

V tomto algoritme sa dá výhodne využiť LU rozklad, keďže v kroku a) riešime pre každé k systém s tou istou maticou A.

Iteračné metódy riešenia sústavy rovníc

Vyššie uvedené metódy (GEM, LU,...) riešenia lineárnych systémov rovníc zaraďujeme medzi priame metódy. To znamená, že po konečnom počte krokov bez zaokrúhľovacích chýb dostaneme presné riešenie. Iteračné metódy sú založené na vytvorení postupnosti, od ktorej očakávame, že konverguje k hľadanému riešeniu. V každom kroku iteračnej metódy teda vytvoríme nový člen postupnosti a okrem konvergencie požadujeme, aby v sa každom kroku chyba podstatne zmenšila a aby každý krok nebol príliš drahý v zmysle zložitosti.

Ako prvé uvedieme metódy založené na myšlienke rozkladu matice

$$A = M_k - N_k, \qquad \text{k=1, 2,...}$$

Potom systém $Ax = b$ prepíšeme do tvaru

$$M_k x = N_k x + b$$

a zvolíme iterácie

$$M_k x^{(k)} = N_k x^{(k-1)} + b.$$

Potom ak M_k je regulárna, dostaneme

$$x^{(k)} = M_k^{-1} N_k x^{(k-1)} + M_k^{-1} b.$$

Označme $Q_k = M_k^{-1} N_k$ a $d^{(k)} = M_k^{-1} b$. Dostaneme tak postupnosť systémov

$$x^{(k)} = Q_k x^{(k-1)} + d^{(k)}$$

Zrejme ak postupnosť $\{x^k\}$ konverguje, tak konverguje k riešeniu systému $Ax = b$.

Ak rozklad $A = M - N$ nezávisí od k, hovoríme o stacionárnej metóde a máme

$$x^{(k)} = Qx^{(k-1)} + d$$

Na konvergenciu potrebujeme, aby odchýlka od presného riešenia sa v limite blížila k nule. Spočítajme teda

$$x^{(k)} - \bar{x} = Q_k x^{(k-1)} + d^{(k)} - (Q_k \bar{x} + d^{(k)}) = Q_k (x^{(k-1)} - \bar{x}) \tag{12}$$

$$= Q_k Q_{k-1} (x^{(k-2)} - \bar{x}) = \cdots = Q_k \cdots Q_1 (x^0 - \bar{x}). \tag{13}$$

Preto postupnosť $\{x^k\}$ bude pre ľubovoľné x^0 konvergovať práve vtedy, keď

$$Q_k \cdots Q_1 \to 0$$

Ak normujeme (13) dostaneme

$$\|x^{(k)} - \bar{x}\| \leq \|Q_k\| \cdots \|Q_1\| \|(x^0 - \bar{x})\|.$$

Potom ak existuje $c < 1$ také, že $\|Q_i\| \leq c$ pre každé i, tak

$$\|x^{(k)} - \bar{x}\| \leq c^k \|(x^0 - \bar{x})\| \xrightarrow{k \to \infty} 0$$

Tým dostávame postačujúcu podmienku konvergencie

$$\|Q_k\| \leq c < 1.$$

Pre prípad stacionárnych metód samozrejme na konvergenciu stačí

$$\|Q\| < 1.$$

Avšak v prípade stacionárnych metód máme dokonca nutnú a postačujúcu podmienku konvergencie

$$\rho(Q) < 1,$$

kde $\rho(Q)$ je spektrálny polomer matice Q.

Veta. Postupnosť $x^{(k)}$ riešení systémov $x^{(k)} = Qx^{(k-1)} + d$ konverguje k riešeniu $Ax = b$ práve vtedy, keď $\rho(Q) < 1$.

Dôkaz. Najskôr ukážeme, že pre ľubovoľné $\epsilon > 0$ a ľubovoľnú maticu R existuje operátorová norma $||\cdot||_*$ taká, že $||R||_* \leq \rho(R) + \epsilon$. Nech $J = S^{-1}RS$ je Jordanov kanonický tvar matice R. Označme $D_\epsilon = \mathrm{diag}(1, \epsilon, \epsilon^2, \ldots, \epsilon^{n-1})$. Potom matica $(SD_\epsilon)^{-1}R(SD_\epsilon) = D_\epsilon^{-1}JD_\epsilon$ je "Jordanova matica" s ϵ nad diagonálou. Definujme vektorovú normu $||x||_* = ||(SD_\epsilon)^{-1}x||_\infty$. Potom

$$\begin{aligned} ||R||_* &= \max_{x\neq 0} \frac{||Rx||_*}{||x||_*} = \max_{x\neq 0} \frac{||(SD_\epsilon)^{-1}Rx||_\infty}{||(SD_\epsilon)^{-1}x||_\infty} \\ &= \max_{y\neq 0} \frac{||(SD_\epsilon)^{-1}R(SD_\epsilon)y||_\infty}{||y||_\infty} = ||(SD_\epsilon)^{-1}R(SD_\epsilon)||_\infty \\ &\leq \max_i |\lambda_i| + \epsilon = \rho(R) + \epsilon. \end{aligned}$$

Teraz dokážeme vetu. Ak $\rho(R) \geq 1$, zvoľme $x^0 - x$ vlastný vektor s vlastnou hodnotou λ takou, že $|\lambda| = \rho(R)$. Potom

$$(x^{(m+1)} - x) = R(x^{(m)} - x) = \cdots = R^{m+1}(x^{(0)} - x) = \lambda^{m+1}(x^{(0)} - x) \quad (14)$$

nekonverguje k 0.

Naopak, ak $\rho(R) < 1$, tak vieme nájsť $\epsilon > 0$, také, že $\rho(R) + \epsilon < 1$, a teda aj normu $||\cdot||*$ takú, že $||R||_* < 1$, čo zaručuje konvergenciu. Pri vytváraní matíc M, N treba brať do úvahy

- Matica M musí byť ľahko invertovateľná, prípadne systémy s maticou M sa musia dať ľahko riešiť.
- Mali by sme vedieť ľahko overiť niektorú z podmienok konvergencie, napríklad pre stacionárne metódy $\rho(M^{-1}N) < 1$.

Jacobiho metóda

Jacobiho metódu dostaneme pre rozklad

$$A = D - (D - A),$$

kde D je diagonálna matica obsahujúca diagonálu matice A, pričom o A predpokladáme, že má nenulovú diagonálu. Získame tak rekurentný vzťah

$$\begin{aligned} x^{(k)} &= D^{-1}(D-A)x^{(k-1)} + D^{-1}b = -D^{-1}(A-D)x^{(k-1)} + D^{-1}b \\ &= Q_J x^{(k-1)} + d. \end{aligned}$$

Matica Q_J má tým pádom tvar

$$Q_J = \begin{pmatrix} 0 & \frac{a_{12}}{a_{11}} & \cdots & \frac{a_{1n}}{a_{11}} \\ \frac{a_{21}}{a_{11}} & 0 & \cdots & \frac{a_{1n}}{a_{11}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{a_{n1}}{a_{nn}} & \frac{a_{n2}}{a_{nn}} & \cdots & 0 \end{pmatrix}$$

Pozrime sa teraz na konvergenciu riešenia pre túto maticu v prípade jednoducho spočítateľných noriem. V prípade riadkovej normy máme

$$\|Q_J\|_\infty = \max_i \sum_{j \neq i} \left| \frac{a_{ij}}{a_{ii}} \right|,$$

z čoho okamžite dostaneme, že ak $|a_{ii}| > \sum_{j \neq i} |a_{ij}|$ (hovoríme, že A má má po riadkoch prevládajúcu diagonálu), tak $\|Q_j\| < 1$.

Ak A má prevládajúcu diagonálu po stĺpcoch, tak analogicky

$$\|(A-D)D^{-1}\| < 1.$$

Keďže platí

$$-Q_J = D^{-1}(A-D) = D^{-1}[(A-D)D^{-1}]D,$$

tak

$$\rho(D^{-1}(A-D)) = \rho((A-D)D^{-1}) \leq \|(A-D)D^{-1}\|_1 < 1$$

Zhrnutím uvedených úvah dostávame, že ak A má prevládajúcu diagonálu (po riadkoch alebo po stĺpcoch) tak Jacobiho metóda konverguje.

Gaussova-Seidlova metóda

Pri rozklade matice A na tvar

$$A = D + L + U,$$

kde D je diagonálna matica, L je striktne dolná trojuholníková (t.j. dolná trojuholníková s nulami na diagonále) a U je striktne horná trojuholníková, voľbou

$$M = D + L, \qquad N = -U$$

získame iteračnú rovnicu

$$x^{(k)} = -(D+L)^{-1}Ux^{(k-1)} + (D+L)^{-1}b.$$

Ukazuje sa, že ako postačujúca podmienka konvergencia Gaussovej-Seidlovej metódy je diagonálna dominancia rovnako dobrá ako v prípade Jacobiho metódy.

Veta 1. *Ak matica A má prevládajúcu diagonálu po riadkoch alebo stĺpcoch, tak Gaussova-Seidlova metóda konverguje*

Dôkaz: Predpokladajme riadkovú diagonálnu dominanciu matice $A = (a_{ij})$. Označme

$$c = \max_i \sum_{j \neq i} \frac{|a_{ij}|}{|a_{ii}|}.$$

Zrejme $c < 1$, lebo podľa predpokladu A má prevládajúcu diagonálu po riadkoch. Označme

$$Q_{GS} = (D+L)^{-1}U.$$

Potrebujeme ukázať, že $\|Q_{GS}\|_\infty < c$.

Nech x je vektor s vlastnosťou $\|x\|_\infty = 1$. Potom stačí ukázať, že

$$\|Q_{GS}x\|_\infty < c.$$

Označme $y = Q_{GS}x = (D+L)^{-1}Ux$, čo vieme prepísať na tvar $(D+L)y = -Ux$, čiže

$$\begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ a_{21} & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 0 & -a_{12} & \cdots & -a_{1n} \\ 0 & 0 & \cdots & -a_{2n} \\ \vdots & \vdots & \ddots & -a_{n-1\,n} \\ 0 & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

Tento tvar naznačuje možnosť pokračovať indukciou.

V prvom kroku dostávame

$$|y_1| \leq \sum_{j=2}^{n} \frac{|a_{ij}||x_j|}{|a_{11}|} \leq \sum_{j=2}^{n} \frac{|a_{ij}|}{|a_{11}|} \max_k |x_k| = \sum_{j=2}^{n} \frac{|a_{ij}|}{|a_{11}|} \|x_k\|_\infty = \sum_{j=2}^{n} \frac{|a_{ij}|}{|a_{11}|} \leq c.$$

Predpokladajme, že $|y_k| \leq c$ pre $k = 1, \ldots, i-1$. Potom

$$|y_i| \leq \frac{1}{|a_{ii}|} \left(\sum_{j=1}^{i-1} |a_{ij}||y_j| + \sum_{j=i+1}^{n} |a_{ij}||x_j| \right)$$

V prvej sume v zátvorke môžeme použiť indukčný predpoklad a v druhej vlastnosť $\|x\|_\infty = 1$, z ktorej dostávame $|x_j| \leq 1$ pre každé $j = 1, \ldots, n$. Teda máme nerovnosť

$$|y_i| \leq \frac{1}{|a_{ii}|} \left(\sum_{j=1}^{i-1} |a_{ij}|c + \sum_{j=i+1}^{n} |a_{ij}| \right)$$

a keďže $c < 1$,

$$|y_i| \leq \frac{1}{|a_{ii}|} \left(\sum_{j=1}^{i-1} |a_{ij}| + \sum_{j=i+1}^{n} |a_{ij}| \right) = \sum_{j \neq i} \frac{|a_{ij}|}{|a_{ii}|} \leq c.$$

Takže $|y_i| < c$ pre každé $i = 1, \ldots, n$ a teda aj $\|Q_{GS}x\|_\infty = \|y\|_\infty = \max_i |y_i| < c$.

Za predpokladu stĺpcovej dominancie matice A dostaneme podobne ako pri Jacobiho metóde

$$\begin{aligned} \rho((D+L)^{-1}U) = \rho(U(D+L)^{-1}) &\leq \|U(D+L)^{-1}\|_1 \\ &= \|(D+L)^{-T}U^T\|_\infty = \|Q_{GS}^T\|. \end{aligned}$$

Keďže stĺpcová dominancia v A znamená riadkovú dominanciu v A^T, použitím časti dôkazu pre riadkovú dominanciu dostávame $\|Q_{GS}^T\| < c^T$, kde $c^T = \max_j \sum_{j \neq i} \frac{|a_{ij}|}{|a_{jj}|}$

Pre istú triedu matíc však máme konvergenciu zaručenú vždy.

Veta 2. *Ak A je symetrická kladne definitná matica, tak Gaussova-Seidlova metóda konverguje.*

Dôkaz: Pre prípad symetrickej matice dostávame rozklad

$$A = D + L + L^T,$$

kde D je diagonálna matica obsahujúca diagonálne prvky matice A a L a L^T sú zvyšné dolné a horné trojuholníkové časti matice $A - D$. Keďže A je kladne definitná, diagonála matice D obsahuje len kladné prvky, a teda je tiež

kladne definitná (ak by $a_{ii}<0$, tak $e_i^T A e_i = a_{ii} < 0$, kde $e_i = (0, \ldots, 1, \ldots, 0)$ je vektor kanonickej bázy).

Na odhad pre konvergenciu použijeme 2-normu, čím zredukujeme problém na odhad veľkosti absolútnych hodnôt vlastných čísel, keďže pre symetrickú maticu A je $\|A\|_2 = |\lambda|_{\max}$. Najskôr zjednodušíme maticu $Q_{GS} = -(D + L)^{-1}L^T$.

$$Q_{GS} = -(D+L)^{-1}L^T = -D^{-\frac{1}{2}}(I + D^{-\frac{1}{2}}LD^{-\frac{1}{2}})^{-1}D^{-\frac{1}{2}}L^T D^{-\frac{1}{2}}D^{\frac{1}{2}}.$$

Označme $L_1 = D^{-\frac{1}{2}}LD^{-\frac{1}{2}}$. Potom

$$Q_{GS} = -D^{-\frac{1}{2}}(I + L_1)^{-1}L_1^T D^{\frac{1}{2}}.$$

Matica $-(I + L_1)^{-1}L_1^T$ je teda podobná s maticou Q_{GS}, a tým pádom má rovnaké vlastné čísla.

Ak x označuje jednotkový vlastný vektor, t.j. $\|x\|_2 = \sqrt{x^*x} = 1$ a $-(I + L_1)^{-1}L_1^T x = \lambda x$ (x môže byť komplexný vektor a λ komplexná vlastná hodnota, * znamená hermitovské združenie), máme

$$\begin{aligned} -(I + L_1)^{-1}L_1^T x &= \lambda x \\ -L_1^T x &= \lambda(I + L_1)x \end{aligned}$$

a prenásobením posledného vzťahu zľava x^* a úpravou

$$-x^*L_1^T x = \lambda x^*(I + L_1)x = \lambda(I + x^*L_1 x).$$

Označme $z = x^*L_1^T x$. Potom platí

$$\begin{aligned} -z &= \lambda(1 + \bar{z}) \\ \lambda &= -\frac{z}{1+\bar{z}}. \end{aligned}$$

Z toho dostaneme, že

$$|\lambda|^2 = \lambda\bar{\lambda} = \frac{z}{1+\bar{z}} \cdot \frac{\bar{z}}{1+z} = \frac{|z|^2}{1 + z + \bar{z} + |z|^2}$$

a navyše

$$\begin{aligned} 1 + z + \bar{z} &= 1 + x^*L_1^T x + x^*L_1 x = x^*(I + L_1 + L_1^T)x \\ &= x^*D^{-\frac{1}{2}}(D + L + L^T)D^{-\frac{1}{2}} \\ &= x^*D^{-\frac{1}{2}}AD^{-\frac{1}{2}}x > 0. \end{aligned}$$

Preto

$$|\lambda|^2 = \frac{|z|^2}{1 + z + \bar{z} + |z|^2} < 1,$$

a teda aj $\|Q\|_2 = |\lambda|_{\max} < 1$.

Poznámka: Ak sa na výpočet Jacobiho metódou pozrieme po súradniciach uvidíme, že j-ta súradnica riešenia v x^m sa spočíta ako

$$x_j^{(m)} = \frac{1}{a_{jj}}(b_j - \sum_{k \neq j} a_{jk} x_k^{(m-1)}). \tag{15}$$

To ale znamená, že v čase, keď počítame $x_j^{(m)}$ už poznáme $x_l^{(m)}$ pre $l < j$, čo sú súradnice zlepšeného riešenia. Ak tieto nové súradnice použijeme vo výpočte (15), t.j. rozdelíme sumu v zátvorke na $k < j$ a $k > j$ a v $k < j$ nahradíme $x_k^{(m-1)}$ novými súradnicami $x_k^{(m)}$, dostaneme

$$x_j^{(m)} = \frac{1}{a_{jj}} \left(b_j - \sum_{k<j} a_{jk} x_k^{(m)} - \sum_{k>j} a_{jk} x_k^{(m-1)} \right),$$

čo nie je ale nič iné ako Gaussova-Seidelova metóda.

Ako si možno všimnúť v prípade že niektorý prvok diagonály matice A je nulový, metódy zlyhajú (delenie prvkom a_{ii}). Tento problém možno obísť výmenou riadkov alebo stĺpcov, avšak za cenu straty istoty konvergencie.

Ak by sme chceli všeobecný prípad $Ax = b$ previesť na symetrický $A^T Ax = A^T b$, môže sa stať, že $A^T A$ je zle podmienená ($c_2(A^T A) = c_2(A)^2$).

Metódy založené na minimalizácii kvadratickej formy

Pri týchto metódach hľadáme riešenie systému $Ax = b$ nepriamo ako minimum nejakej vhodne definovanej funkcie $F(x)$. Ako príklady takých funkcií môžeme uviesť

$$F_1(x) = (b - Ax)^T (b - Ax),$$

alebo ak $r = b - Ax$ je reziduum a $\bar{x}$ presné riešenie, tak

$$F_2(x) = (\bar{x} - x)^T (\bar{x} - x) = \Delta x^T \Delta x = (A^{-1} r)^T (A^{-1} r).$$

Symetrické, kladne definitné matice

Pre prípad systému $Ax = b$ so symetrickou a kladne definitnou maticou systému A môžeme definovať funkciu F ako

$$F(x) = \Delta x^T A \Delta x = (\bar{x} - x)^T A(\bar{x} - x).$$

Vďaka kladnej definitnosti máme $F(x) \geq 0$ a $F(x) = 0$ práve vtedy keď $x = \bar{x}$.

Keďže

$$\begin{aligned} F(x) = (\bar{x} - x)^T A(\bar{x} - x) &= \bar{x}^T \underbrace{A\bar{x}}_{b} - x^T \underbrace{A\bar{x}}_{b} - \underbrace{\bar{x}^T A}_{b^T} x + x^T A x \\ &= x^T A x - 2b^T x + \bar{x}^T b \end{aligned}$$

a $\bar{x}^T b$ nezávisí od x, stačí minimalizovať funkciu

$$f(x) = x^T A x - 2b^T x$$

Samotný iteračný krok potom prebieha tak, že k už získanému riešeniu pripočítame vhodný $\alpha \in \mathbb{R}$ násobok nejakého vhodne zvoleného vektora v, čiže $x + \alpha v$ bude nové zlepšené odhad riešenia. To vedie k otázke, ako voliť α a v.

Najskôr predpokladajme, že v sme zvolili. Potom α je zrejme vhodné voliť tak, aby $x + \alpha v$ minimalizovalo funkciu f. Po dosadení $x - \alpha v$ dostaneme

$$\begin{aligned} g(\alpha) = f(x + \alpha v) &= (x + \alpha v)^T A(x + \alpha v) - 2b^T(x + \alpha v) = \\ &= \alpha^2 v^T A v + 2\alpha(v^T A x - b^T v) + x^T A x - 2b^T x \\ &= \alpha^2 v^T A v + 2\alpha v^T (Ax - b) + f(x) \\ &= \alpha^2 v^T A v - 2\alpha v^T r + f(x) \end{aligned}$$

čo je kvadratická funkcia premennej α, pričom $v^T A v > 0$, takže minimum g získame pre α spĺňajúce

$$g'(\alpha) = 2\alpha v^T A v - 2v^T r = 0,$$

teda pre

$$\alpha = \frac{v^T r}{v^T A v}$$

Iteračný proces teda môžeme zapísať ako

$$x^{(k)} = x^{(k-1)} + \alpha_k v^{(k)} = x^{(k-1)} + \frac{v^{(k)T} r^{(k-1)}}{v^{(k)T} A v^{(k)}} v \tag{16}$$

Ostáva už len určiť vektory $v^{(k)}$. Na zlepšovanie riešenia stačí, aby $\alpha_k \neq 0$, čo znamená, že

$$v^{(k)T} r^{(k-1)} \neq 0.$$

Vektory $v^{(k)}$ treba teda voliť tak, aby neboli kolmé na reziduá. Rôznym spôsobom voľby vektorov $v^{(k)}$ potom dostávame rôzne metódy.

Metóda najväčšieho spádu

Prirodzená voľba vektora v je zrejme smer, v ktorom sa funkcia $f(x)$ mení najviac. Keďže veľkosť zmeny meriame deriváciou, vektor v hľadáme tak, aby číslo

$$\lim_{t\to\infty} \frac{f(x+tv) - f(x)}{t}$$

bolo v absolútnej hodnote čo najväčšie. Počítajme teda:

$$\lim_{t\to\infty} \frac{f(x+tv) - f(x)}{t} = \left.\frac{df(x+tv)}{dt}\right|_{t=0} = g'(0) = -2v^T r = -2\|v\|_2 \|r\|_2 \cos\varphi.$$

Pri voľbe v tak, aby $\|v\|_2 = 1$ je toto číslo najväčšie pre $\varphi = 0$ (prípadne $\varphi = \pi$, ale v tom prípade α bude mať len opačné znamienko). Zvolíme teda $v = r$, takže v iteráciách $v^{(k)} = r^{(k-1)}$ a

$$x^{(k)} = x^{(k-1)} + \frac{r^{(k-1)T} r^{(k-1)}}{r^{(k-1)T} A r^{(k-1)}} r^{(k-1)}$$

Samotný algoritmus prebieha sa dá zhrnúť nasledovne

- $x = x^0$
- $r^{(k)} = b - Ax^{(k)}$; $\alpha_{k+1} = \dfrac{r^{(k)T} r^{(k)}}{r^{(k)T} A r^{(k)}}$; $x^{(k+1)} = x^{(k)} + \alpha_{k+1} r^{(k)}$.

Pozrime sa teraz na rýchlosť konvergencie tejto metódy. Pomerne jednoducho sa dá ukázať, že

$$\frac{F(x^{(k+1)})}{F(x^{(k)})} = 1 - \frac{(r^{(k)T} r^{(k)})^2}{(r^{(k)T} A r^{(k)})(r^{(k)T} A^{-1} r^{(k)})}. \tag{17}$$

V predchádzajúcom výraze označme $r = r^{(k)}$ a zapíšme r v báze ortonormálnych vlastných vektorov $(x_1, \ldots, x_n)$, kde $Ax_i = \lambda_i x$. Potom $r = \sum_{i=1}^n \beta_i x_i$

a

$$(r^T r)^2 = \left[\left(\sum_{i=1}^{n} \beta_i x_i^T\right)\left(\sum_{i=1}^{n} \beta_i x_i\right)\right]^2 = \left(\sum_{i=1}^{n} \beta_i^2\right)^2,$$
$$(r^T A r)^2 = \left[\left(\sum_{i=1}^{n} \beta_i x_i^T\right)\left(\sum_{i=1}^{n} \beta_i \lambda_i x_i\right)\right] = \sum_{i=1}^{n} \beta_i^2 \lambda_i,$$
$$(r^T A^{-1} r)^2 = \left[\left(\sum_{i=1}^{n} \beta_i x_i^T\right)\left(\sum_{i=1}^{n} \beta_i \frac{1}{\lambda_i} x_i\right)\right] = \sum_{i=1}^{n} \frac{\beta_i^2}{\lambda_i}.$$

Po dosadení do (17) máme

$$\frac{F(x^{(k+1)})}{F(x^{(k)})} = 1 - \frac{\left(\sum_{i=1}^{n} \beta_i^2\right)^2}{\left(\sum_{i=1}^{n} \beta_i^2 \lambda_i\right)\left(\sum_{i=1}^{n} \frac{\beta_i^2}{\lambda_i}\right)} = 1 - \frac{1}{H},$$

kde

$$H = \frac{\left(\sum_{i=1}^{n} \beta_i^2 \lambda_i\right)\left(\sum_{i=1}^{n} \frac{\beta_i^2}{\lambda_i}\right)}{\left(\sum_{i=1}^{n} \beta_i^2\right)\left(\sum_{i=1}^{n} \beta_i^2\right)}$$

a pri označení $s = \sum_{i=1}^{n} \beta_i^2$

$$H = \left(\sum_{i=1}^{n} \frac{\beta_i^2 \lambda_i}{s}\right)\left(\sum_{i=1}^{n} \frac{\beta_i^2}{s\lambda_i}\right).$$

Toto číslo je možné odhadnúť použitím Kantorovičovej nerovnosti

Veta 3. *Ak $x_1, \ldots, x_n$ sú kladné reálne čísla a $\gamma_1, \ldots, \gamma_n \geq 0$, také, že $\sum_{j=1}^{n} \gamma_i = 1$, tak*

$$\left(\sum_{i=1}^{n} \gamma_i x_i\right)\left(\sum_{i=1}^{n} \gamma_i x_i^{-1}\right) \leq \frac{(x_1 + x_n)^2}{4 x_1 x_2}.$$

Z toho dostávame odhad

$$H = \left(\sum_{i=1}^{n} \frac{\beta_i^2 \lambda_i}{s}\right)\left(\sum_{i=1}^{n} \frac{\beta_i^2}{s\lambda_i}\right) \leq \frac{(\lambda_1 + \lambda_n)^2}{4\lambda_1 \lambda_n},$$

a potom

$$\frac{F(x^{(k+1)})}{F(x^{(k)})} \leq 1 - \frac{4\lambda_1 \lambda_n}{(\lambda_1 + \lambda_n)^2} = \left[\frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n}\right].$$

Dá sa ukázať, že potom

$$\|\Delta x^{(k)}\|_2 \leq \left[\frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n}\right]^k \|x^{(0)}\|_2 \sqrt{\frac{\lambda_1}{\lambda_n}}.$$

Takže pokiaľ λ_1 a λ_n sú rádovo blízke, metóda konverguje celkom rýchlo. Pokiaľ ale $\lambda_1 \gg \lambda_n$, čiže pre zle podmienené matice, môže byť konvergencia pomalá.
Poznámky: Rýchlosť algoritmu možno zlepšiť

a) $r^{(k+1)} = b - Ax^{(k+1)} = b - A(x^{(k)} + \alpha_{k+1} r^k) = (b - Ax^{(k)}) - \alpha_{k+1} Ar^{(k)}$ $= r^{(k)} - \alpha_{k+1} Ar^{(k)}$, pričom $Ar^{(k)}$ počítame už pri α_{k+1}, čiže nemusíme počítať $Ax^{(k+1)}$.

b) $x^{(k+1)} = x^{(k)} + \omega_{k+1}\alpha_{k+1} x^{(k)}$, kde $\omega_{k+1} = \frac{\alpha_k}{\alpha_{k+1}}$.

Čo si môžeme všimnúť je, že

$$\begin{aligned} r^{(k)T} r^{(k+1)} &= r^{(k)T}(r^{(k)} - \alpha_{k+1} Ar^{(k)}) = r^{(k)T} r^{(k)} - r^{(k)T} \frac{r^{(k)T} r^{(k)}}{r^{(k)T} Ar^{(k)}} Ar^{(k)} \\ &= r^{(k)T} r^{(k)} - r^{(k)T} r^{(k)} \frac{r^{(k)T} Ar^{(k)}}{r^{(k)T} Ar^{(k)}} = 0. \end{aligned}$$

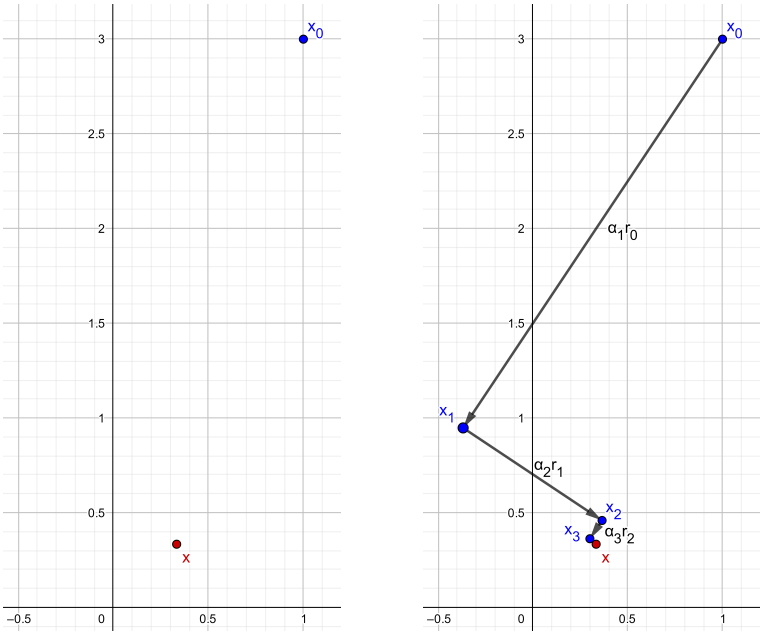
Takže $r^{(k)T} \perp r^{(k+1)}$.

Nižšie je animácia a obrázok zobrazujúci tri iterácie prezentovanej metódy pre systém

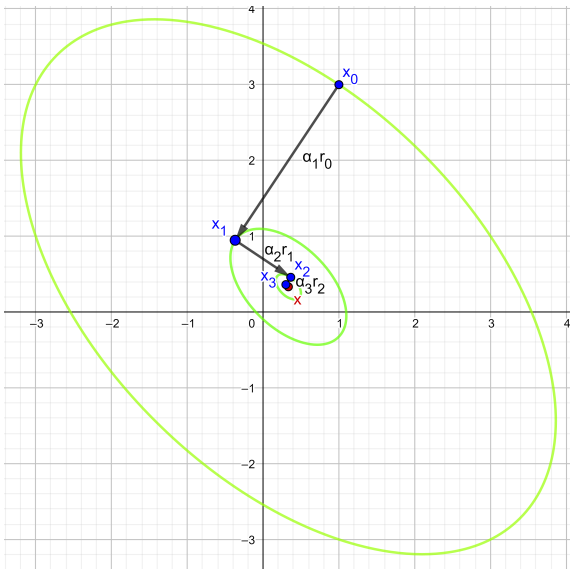
$$\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Metóda združených smerov

Predchádzajúcu metódu môžeme v rovine geometrický opísať nasledovne. Vrstevnice kvadratickej funkcie F sú elipsy (lebo A je kladne definitná symetrická matica). Bod $x^{(0)}$ v ktorom začíname algoritmus teda leží na nejakej elipse $e^{(0)}$. Zostrojíme kolmicu na túto elipsu v bode $x^{(0)}$. To bude priamka ktorej smerový vektor je vektor rezidua $r^{(0)}$, čo je mimochodom aj nejaký násobok gradientu f v bode $x^{(0)}$. Táto priamka je dotyčnicou nejakej ďalšej vrstevnice $e^{(1)}$ a bod dotyku je nové približné riešenie $x^{(1)}$. Ďalšie priblíženie k riešeniu dostaneme opakovaním tejto "konštrukcie"(obr. 2). V prípade viacrozmerného prípadu je postup podobný. Konštruujeme kolmice



Obr. 1: Vľavo je animácia postupu výpočtu metódou najväčšieho spádu, vpravo sú zobrazené riešenia získané z troch iterácií.



Obr. 2: Metóda najväčšieho spádu s vrstevnicami

na vrstevnice podobným spôsobom, a tým konštruujeme postupne minimá kvadratickej funkcie v daných smeroch.

Čo si môžeme všimnúť z vyššie uvedenej konštrukcie je, že bod $x^{(2)}$ nie

je minimálny vzhľadom na smer určený vektorom $r^{(0)}$, pretože posunom v smere $r^{(2)}$, čo je len nejaký násobok vektora $r^{(0)}$ dostaneme $x^{(3)}$, v ktorom má f menšiu hodnotu ako v predošlom $x^{(2)}$. Ponúka sa otázka, či nevieme vektory $v^{(k)}$ v iteračnom procese (16) voliť tak, aby sme v danom iteračnom kroku získali bod minima danej kvadratickej funkcie, ktorý bude minimálny vzhľadom na všetky predchádzajúce smery. Poďme túto myšlienku formalizovať.

V k-tom kroku minimalizácie dostávame

$$x^{(k)} = x^{(k-1)} + \alpha_k v^{(k)}.$$

Ak zvolíme niektorý z predchádzajúcich smerov $v^{(i)}$, $i < k$, ako nový smer posunu, tak

$$x^{(k+1)} = x^{(k)} + \alpha v^{(i)},$$

pričom

$$\alpha = \frac{v^{(i)T} r^{(k)}}{v^{(i)T} A v^{(i)}}.$$

Ak ale predpokladáme, že v $x^{(k)}$ je minimum vzhľadom na predchádzajúce smery, musí byť $\alpha = 0$, a teda

$$\begin{aligned} 0 &= v^{(i)T} r^{(k)} = v^{(i)T}(b - Ax^{(k)}) = v^{(i)T}(b - A(x^{(k-1)} + \alpha_k v^{(k)})) \\ &= v^{(i)T}((b - A(x^{(k-1)})) + \alpha_k A v^{(k)}) = v^{(i)T}(r^{(k-1)} + \alpha_k A v^{(k)}) \\ &= v^{(i)T} r^{(k-1)} + v^{(i)T} \alpha_k A v^{(k)} = v^{(i)T} \alpha_k A v^{(k)}. \end{aligned}$$

Posledná rovnosť vyplýva z toho, že $v^{(i)T} r^{(k-1)}$, pretože sme predpokladali, že $x^{(k)}$ je minimum vzhľadom na predchádzajúce smery. Preto pre nulovosť α potrebujeme $v^{(i)T} A v^{(k)} = 0$.

Keďže A je symetrická, kladne definitná, výraz $\langle x, y \rangle_A = y^T A x$ má všetky vlastnosti skalárneho súčinu. Tým pádom vzťah $v^{(i)T} A v^{(k)} = 0$ znamená, že vektory $v^{(i)}$ a $v^{(k)}$ sú A-ortogonálne.

Teraz môžeme zostrojiť algoritmus na riešenie systému $Ax = b$ s maticou rozmeru $n \times n$.

- Zvolíme A-ortogonálnu bázu $v^{(1)}, \ldots, v^{(n)}$.
- Zvolíme $x^{(0)}$.
- $r^{(0)} = b - Ax^{(0)}$.

- Pre $k = 1, \ldots, n$ spočítame

$$\begin{aligned} \alpha_k &= \frac{v^{(k)T} r^{(k-1)}}{v^{(k)T} A v^{(k)}} \\ x^{(k)} &= x^{(k-1)} + \alpha_k v^{(k)} \\ r^{(k)} &= b - Ax^{(k)} = b - A(x^{(k-1)} + \alpha_k v^{(k)}) = r^{(k-1)} - \alpha_k a v^{(k)} \end{aligned}$$

Po n krokoch dostávame minimum funkcie v n rôznych smeroch, čo znamená, že sme získali presné riešenie.

A-ortogonálnu bázu $\{v^{(i)}\}$ je možné získať z ľubovoľnej bázy $\{u^{(i)}\}$ Gramovou-Schmidtovou A-ortogonalizáciou.

$$\begin{aligned} v^{(1)} &= u^{(1)} \\ v^{(k+1)} &= u^{(k+1)} - \sum_{j=1}^{k} \frac{\langle u^{(k+1)}, v^{(j)} \rangle_A}{\langle v^{(j)}, v^{(j)} \rangle_A} v^j. \end{aligned}$$

Nevýhodou je že na výpočet $v^{(k+1)}$ potrebujeme mať uložené všetky vektory $v^{(1)}, \ldots, v^{(k)}$, čo podstatne zvyšuje nároky na pamäť. Taktiež samotný výpočet bázy má zložitosť $O(n^3)$, čo nie je rýchlejšie ako GEM.

Miernou modifikáciou predošlého postupu získame nasledujúcu metódu.

Metóda združených gradientov

Idea tejto metódy je taká, A-ortogonálnu bázu z predošlého postupu počítame priebežne z reziduí metódy najväčšieho spádu.

- Zvolíme $x^{(0)}$, $r^{(0)} = b - Ax^{(0)}$, $v^{(1)} = r^{(0)}$.
- Pre $k = 1, \ldots, n$

$$\begin{aligned} \alpha_k &= \frac{v^{(k)T} r^{(k-1)}}{v^{(k)T} A v^{(k)}} \\ x^{(k)} &= x^{(k-1)} + \alpha_k v^{(k)} \\ r^{(k)} &= r^{(k-1)} - \alpha_k A v^{(k)} \\ v^{(k+1)} &= r^{(k)} - \sum_{j=1}^{k} \frac{\langle r^{(k)}, v^{(j)} \rangle_A}{\langle v^{(j)}, v^{(j)} \rangle_A} v^{(j)} \end{aligned}$$

Z predchádzajúcej časti máme $v^{(i)T} r^{(k)} = 0$ pre $i \leq k$, teda $v^{(i)} \perp r^{(k)}$ pre $i \leq k$. Taktiež vektory $\{v^{(1)}, \ldots, v^{(k)}\}$ generujú ten istý priestor ako

$\{r^{(1)}, \ldots, r^{(k-1)}\}$. Z toho dostávame, že $r^{(k)} \perp r^{(i)}$ pre každé $i < k$. Potom

$$\langle r^{(k)}, v^{(j)} \rangle_A = v^{(j)T} A r^{(k)} = r^{(k)T} A v^{(j)} = r^{(k)T} \cdot \frac{r^{(j-1)} - r^{(j)}}{\alpha_j} = 0$$

pre $j < k$. Teda $r^{(k)}$ je A-ortogonálny na všetky v^j okrem posledného. Tým dostaneme modifikáciu predošlého algoritmu.

- Zvolíme $x^{(0)}$, $r^{(0)} = b - Ax^{(0)}$, $v^{(1)} = r^{(0)}$.
- Pre $k = 1, \ldots, n$

$$\alpha_k = \frac{v^{(k)T} r^{(k-1)}}{v^{(k)T} A v^{(k)}} = \frac{r^{(k-1)T} r^{(k-1)}}{v^{(k)T} A v^{(k)}} \qquad \text{(alpha)}$$

$$x^{(k)} = x^{(k-1)} + \alpha_k v^{(k)}$$

$$r^{(k)} = r^{(k-1)} - \alpha_k A v^{(k)}$$

$$v^{(k+1)} = r^{(k)} - \sum_{j=1}^{k} \frac{\langle r^{(k)}, v^{(j)} \rangle_A}{\langle v^{(j)}, v^{(j)} \rangle_A} v^{(j)} = r^{(k)} + \frac{r^{(k)T} r^{(k)}}{r^{(k-1)T} r^{(k-1)}} v^{(j)} \qquad \text{(vector)}$$

QR-rozklad

Riešenie systémov $Ax = b$ použitím LU-rozkladu má nevýhodu v tom, že matice L a U z rozkladu môžu mať horšie číslo podmienenosti ako samotná matica A. Preto je často lepšie použiť iný typ rozkladu. Maticu systému rozložíme na súčin ortogonálnej matice Q a stupňovitej matice R (v prípade že A je štvorcová, tak R je horná trjuholníková). Pre maticu A rozmeru $m \times n$ dostávame rozklad

$$A = QR,$$

kde Q je ortogonálna matica rozmeru $m \times m$ a R je stupňovitá rozmeru $m \times n$.

Čo sa týka výpočtovej zložitosti hľadania riešenia systému $Ax = QRx = b$, máme

$$x = R^{-1} Q^T b,$$

čo vieme urobiť v $O(n^2)$. Teda celkovú rýchlosť určuje rýchlosť výpočtu QR-rozkladu, ktorý je $O(n^3)$. Uvedieme niekoľko algoritmov na výpočet QR-rozkladu matice A.

QR-rozklad pomocou GSO

Nech A je matica rozmeru $m \times n$, pričom $m \geq n$. Potom A a jej QR-rozklad vieme zapísať ako

$$A = QR = \begin{pmatrix} Q_1 & | & Q_2 \end{pmatrix} \begin{pmatrix} R_1 \\ O \end{pmatrix}$$

kde Q_1 je $m \times n$ matica s ortonormálnymi stĺpcami, R_1 je horná trojuholníková matica a O je nulová matica. Potom $Q_1 R_1 = A$, čo znamená, že stĺpce matice A sú lineárne kombinácie stĺpcov matice Q_1. Ak A má stĺpcovú hodnosť n, tak R_1 je regulárna a $Q_1 = AR_1^{-1}$. V tom prípade ale aj stĺpce matice Q_1 sú lineárne kombinácie stĺpcov matice A, čo spolu s predošlým faktom implikuje, že A a Q_1 majú rovnaké stĺpcové priestory. Maticu Q_1 teda môžeme ľahko získať pomocou Gramovej-Schmidtovej ortogonalizácie. Ak i-ty stĺpec matice A označíme a_i a i-ty stĺpec matice Q_1 označíme q_i,

$$\begin{aligned}
q_1 &= \frac{a_1}{\|a_1\|} && \Rightarrow \quad a_1 = q_1\|a_1\| \\
q_2 &= \frac{a_2 - \langle a_2, q_1\rangle q_1}{\|a_2 - \langle a_2, q_1\rangle q_1\|} && \Rightarrow \quad a_2 = q_2\|z_2\| + q_1\langle a_2, q_1\rangle \\
q_3 &= \frac{a_3 - \langle a_3, q_1\rangle q_1 - \langle a_3, q_2\rangle q_2}{\|a_3 - \langle a_3, q_1\rangle q_1 - \langle a_3, q_2\rangle q_2\|} && \Rightarrow \quad a_3 = q_3\|z_3\| + q_2\langle a_3, q_2\rangle + q_1\langle a_3, q_1\rangle \\
&\vdots
\end{aligned}$$

pričom z_i označuje čitateľa vo vyjadrení q_i.

Z vyjadrení stĺpcov a_i napravo, potom môžeme vyskladať matice Q_1 a R_1.

$$Q_1 = \begin{pmatrix} q_1 & q_2 & q_3 & \cdots & q_n \end{pmatrix}, \qquad R_1 = \begin{pmatrix} \|a_1\| & \langle a_2, q_1\rangle & \langle a_3, q_1\rangle & \cdots \\ 0 & \|z_2| & \langle a_3, q_2\rangle & \cdots \\ 0 & 0 & \|z_3\| & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Modifikovaná GSO

Zapíšme

$$A = QR = \begin{pmatrix} q_1 & q_2 & \cdots & q_n \end{pmatrix} \begin{pmatrix} r_1^T \\ r_2^T \\ \vdots \\ r_n^T \end{pmatrix} = q_1 r_1^T + q_2 r_2^T + \cdots + q_n r_n^T,$$

kde q_i sú stĺpce matice Q a r_i^T sú riadky matice R.

Potom $q_1^T A = r_1^T$ takže pomocou q_1 vieme spočítať celý prvý riadok matice R. Vo všeobecnosti ak poznáme q_i máme $q_i^T A = r_i^T$. Ďalej máme

$$A - q_1 r_1^T = \begin{pmatrix} 0 & a_2^{(1)} & \ldots & a_n^{(1)} \end{pmatrix} = A^{(1)} = q_2 r_2^T + \cdots + q_n r_n^T$$

Stĺpec $a_2^{(1)}$ je druhým stĺpcom matice $q_2 r_2^T$, pretože všetky ostatné matice $q_i r_i^T$ majú prvé dva stĺpce nulové. Takže $r_{22} = \|a_2^{(1)}\|$, $q_2 = \dfrac{a_2^{(1)}}{\|a_2^{(1)}\|}$ a

$$r_2^T = q_2^T A = q_2^T (A^{(1)} + q_1 r_1^T) = q_2^T A^{(1)}.$$

Následne pokračujeme s $A^{(2)} = A - q_1 r_1^T - q_2 r_2^T$ a celý postup opakujeme.

Zhrňme celý algoritmus:

- $A^0 = A$,
- Pre k=1,. . . ,n
  a) $r_{kk} = \|a_k^{(k-1)}\|_2$,
  b) $q_k = \dfrac{a_k^{(k-1)}}{r_{kk}}$,
  c) $r_k^T = q_k^T A^{(k-1)}$,
  d) $A^{(k)} = A^{(k-1)} - q_k r_k^T$.

Nevýhoda týchto dvoch metód je, že nie sú numericky stabilné. Ak počítame s konečnou presnosťou, nemôžeme očakávať, že získaná matica Q bude skutočne ortogonálna. Preto získanú maticu Q budeme považovať za ortogonálnu, ak $\|I - Q^T Q\| < \varepsilon$, pre dostatočne malé $\varepsilon > 0$. V prípade, že stĺpce matice A sú skoro lineárne závisle, tak pri počítaní s konečnou presnosťou matica Q môže byť veľmi zle ortogonálna, čo znamená, že $\|I - Q^T Q\|$ omnoho väčšie ako zvolené ϵ.

QR-rozklad pomocou matíc odrazu

Matica odrazu (niekedy tiež reflexia) je matica tvaru $P = I - 2uu^T$, kde $\|u\|_2 = 1$. Pomerne jednoducho sa ukáže, že takáto matica P je ortogonálna a symetrická.

Pre daný vektor x skúsme nájsť maticu odrazu P tak, aby

$$Px = \begin{pmatrix} c & 0 & \cdots & 0 \end{pmatrix} = ce_1.$$

Z definíce P máme rovnosť

$$x - 2u(u^T x) = ce_1,$$

čo vieme upraviť na

$$u = \frac{1}{2u^T x}(x - ce_1).$$

Takže u je lineárnou kombináciou x a e_1. Navyše $\|x\|_2 = \|Px\|_2 = |c|$ takže vektor u je násobkom vektora $\tilde{u} = x \pm \|x\|_2 e_1$. Potom ale $u = \dfrac{\tilde{u}}{\|\tilde{u}\|_2}$.

Ľahko sa ukáže, že znamienko $\pm$ v $\tilde{u}$ možno voliť ľubovoľne (ak $\tilde{u} \neq 0$). Položme teda $\tilde{u} = x + \operatorname{sign}(x_1)\|x\|_2 e_1$. Teda $\tilde{u}$ má tvar

$$\tilde{u} = \begin{pmatrix} x_1 + \operatorname{sign}(x_1)\|x\|_2 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

Samotný výpočet QR-rozkladu matice $A_0 = A$ prebieha nasledovne:

Vezmeme prvý stĺpec a_1 matice A_0 a nájdeme maticu odrazu P_1 tak, aby $P_1 a_1 = c_1 e_1$. Potom $P_1 A_0 = A_1$, kde prvý stĺpec matice A_1 je $c_1 e_1$.

Vo všeobecnosti ak máme maticu A_i, ktorá má každý j-ty stĺpec, $j \leq i$ tvaru $c_j e_j$, tak maticu A_{i+1} získame z A_i vynásobením maticou P_{i+1}, ktorá má tvar

$$P_{i+1} = \begin{pmatrix} I_i & 0 \\ 0 & \widetilde{P} \end{pmatrix},$$

kde I_i je jednotková matica rozmeru $i \times i$ a $\tilde{P}$ je matica odrazu rozmeru $n-i \times n-i$, úprave príslušnej časti $i+1$-stĺpca matice A_i, schématicky

$$A_2 = \begin{pmatrix} * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & * & * & * \end{pmatrix} \qquad P_3 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & p_{33} & p_{34} & p_{35} & p_{36} \\ 0 & 0 & p_{43} & p_{44} & p_{45} & p_{46} \\ 0 & 0 & p_{53} & p_{54} & p_{55} & p_{56} \\ 0 & 0 & p_{63} & p_{64} & p_{65} & p_{66} \end{pmatrix}$$

kde podmatica $\widetilde{P} = (p_{ij})$ matice P je matica odrazu pre časť stĺpca vyznačného červenou. Potom

$$P_3 A_2 = \begin{pmatrix} * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & 0 & * & * \\ 0 & 0 & 0 & * & * \\ 0 & 0 & 0 & * & * \end{pmatrix} = A_3$$

Matica Q je potom súčinom matíc P_i, $i = 1, \ldots, n-1$.

V praktickej implementácii netreba počítať matice P_i, pretože

$$\widetilde{P}_{i+1}\widetilde{A}_i = (I - 2u_{i+1}u_{i+1}^T)\widetilde{A}_i = \widetilde{A}_{i+1} - 2u_{i+1}(u_{i+1}^T\widetilde{A}_i)$$

kde $\widetilde{A}_i$ je prislušná časť matice A_i transformovaná maticou P_{i+1} a u_{i+1} je príslušná časť určeného stĺpca, ktorá sa nuluje, až na prvý prvok. Takže namiesto matíc P_i si stačí pamätať vektory u_i, pričom dĺžka každého vektora u_i je $n - (i-1)$.

QR-rozklad pomocou Givensových matíc rotácií

Pre daný vektor $x = (x_1, x_2)^T$ vieme nájsť maticu rotácie R tak, aby $Rx = (u, 0)^T$. Keďže R zachováva dĺžky, prvá súradnica výsledného vektora je $u = \sqrt{x_1^2 + x_2^2}$, takže

$$\begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \sqrt{x_1^2 + x_2^2} \\ 0 \end{pmatrix}$$

Z toho pomerne jednoducho získame tvar matice R, čiže hodnoty $\cos\varphi$ a $\sin\varphi$

$$\cos\varphi = \frac{x_1}{\sqrt{x_1^2 + x_2^2}} \qquad\qquad \sin\varphi = -\frac{x_2}{\sqrt{x_1^2 + x_2^2}}$$

Givensova matica rotácie o uhol θ súradníc i, j je matica tvaru

$$R(\theta, i, j) = \begin{array}{c} \\ \\ \\ \\ i \\ \\ j \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{ccccccccc} & & & & i & & j & & \end{array} \\ \begin{pmatrix} 1 & & & & & & & & \\ & 1 & & & & & & & \\ & & \ddots & & & & & & \\ & & & & \cos\theta & & -\sin\theta & & \\ & & & & & \ddots & & & \\ & & & & \sin\theta & & \cos\theta & & \\ & & & & & & & \ddots & \\ & & & & & & & & 1 \\ & & & & & & & & & 1 \end{pmatrix} \end{array}$$

Pre daný vektor x a súradnice i a j je možné vyššie uvedenou maticou $R(\theta, i, j)$ vynulovať j tu súradnicu vektora x.

Postup na získanie QR-rozkladu je potom zrejmý. Danú maticu A rozmeru $m \times n$ upravíme na stupňovitý tvar postupným násobením zľava Givensovými maticami. Začneme prvým stĺpcom A a postupne z ľava násobime A maticami $R(1, j)$, $j = 2, \ldots, m$. Tým vynulujeme všetky prvky pod diagonálou. Následne postup zopakujeme pre ďalšie stĺpce. Takto získame maticu R z QR-rozkladu.

Maticu Q by sme získali vynásobením zostrojených matíc $R(\theta, i, j)$. Pri praktickej implementácii môžeme ale postupovať nasledovne. Pre ušetrenie pamäte môžeme využiť upravovanú maticu A, kde na pozíciu prvku, ktorý vynulujeme maticou $R(\theta, i, j)$ zapíšeme buď $\operatorname{sign}(\cos\theta)\sin\theta$ ak $|\sin\theta| < |\cos\theta|$, alebo $\frac{\operatorname{sign}(\sin\theta)}{\cos\theta}$ v opačnom prípade. Ak označíme uloženú hodnotu ako p, tak $\sin\theta$ a $\cos\theta$ vieme zrekonštruovať nasledovne: Ak $|p| < 1$, tak $\sin\theta = p$ a $\cos\theta = \sqrt{1 - \sin^2\theta}$. V opačnom prípade $\cos\theta = \frac{1}{p}$ a $\sin\theta = \sqrt{1 - \cos^2\theta}$.

Tento spôsob zaručuje numerickú stabilitu výpočtu, pretože zabraňuje situácii, keď by $\sin\theta$ alebo $\cos\theta$ boli blízko 1 a teda problému pri počítaní $1 - \cos^2\theta$ alebo $1 - \sin^2\theta$.

Zložitosť QR-rozkladu použitím Givensových matíc rotácií je dvojnásobná oproti zložitosti QR-rozkladu použitím matíc odrazu.

Problém najmenších štvorcov

Sústavy $Ax = b$ s maticou $A \in M_{m\times n}$, ktoré majú viac riadkov ako stĺpcov ($m > n$) sa nazývajú preurčené sústavy. Takéto sústavy zvyčajne nemajú riešenie. Môžeme ale miesto presného riešenia hľadať také riešenie, ktoré minimalizuje reziduum v nejakej norme,

$$|r| = \|b - Ax\|.$$

Ak uvedená norma je euklidovská, hovoríme o probléme najmenších štvorcov (PNŠ), čiže hľadáme

$$\min_{x\in\mathbb{R}^n} \|b - Ax\|_2$$

čo pri prepísaní po súradniciach $Ax = (y_1, y_2, \ldots, y_m)^T$, $b = (b_1, b_2, \ldots, b_m)^T$ znamená hľadať x minimalizujúce

$$(b_1 - y_1)^2 + (b_2 - y_2)^2 + \cdots + (b_m - y_m)^2$$

Príklad: Typická aplikácia PNŠ je hľadanie krivky aproximujúcej nejakú sadu bodov. Majme m dvojíc čísel (súradnice bodov) $(y_1, b_1), \ldots, (y_m, b_m)$ a

chceme nájsť kubický polynóm najlepšie aproximujúci (v nejakom zmysle) tieto body. Čiže hľadáme koeficienty $x_1, \ldots, x_4$ polynómu

$$p(y) = x_1 + x_2 y + x_3 y^2 + x_4^3$$

tak, aby tento polynóm minimalizoval reziduum $r = (r_1, \ldots, r_m)^T$, kde

$$r_i = p(y_i) - b_i$$

V maticovom tvare dostávame

$$\begin{aligned} r &= \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_m \end{pmatrix} = \begin{pmatrix} p(y_1) \\ p(y_2) \\ \vdots \\ p(y_m) \end{pmatrix} - \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix} \\ &= \begin{pmatrix} 1 & y_1 & y_1^2 & y_1^3 \\ 1 & y_2 & y_2^2 & y_2^3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & y_m & y_m^2 & y_m^3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} - \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix} = Ax - b \end{aligned}$$

Ak r minimalizujeme v 2-norme, dostávame PNŠ.

Na riešenie problému najmenších štvorcov sa štandardne využívajú tri metódy:

- Normálne rovnice,
- QR-rozklad,
- singulárny rozklad.

Normálne rovnice

Označme $a_1, \ldots, a_n$ stĺpce matice A. Pri riešení PNŠ hľadáme minimum funkcie

$$\begin{aligned} F(x) &= \|b - Ax\|_2^2 = (b - Ax)^T (b - Ax) = b^T b - b^T Ax - x^T A^T b + x^T A^T Ax \\ &= b^T b - 2b^T Ax + x^T A^T Ax. \end{aligned}$$

Na určenie stacionárnych bodov potrebujeme deriváciu F,

$$\frac{\partial F}{\partial x_i} = -2b^T a_i + a_i^T Ax + x^T A^T a_i = -2b^T a_i + 2a_i^T Ax = 2a_i^T (Ax - b).$$

Potom stacionárny bod určíme riešením sústavy

$$\frac{\partial F}{\partial x_i} = 0, \qquad i = 1, \ldots, n,$$

ktorej maticový tvar je

$$2A^T(Ax - b) = 0 \qquad \Leftrightarrow \qquad A^T Ax = A^T b$$

Systém

$$A^T Ax = A^T b$$

nazývame sústava normálnych rovníc.

Takto nájdený stacionárny bod je minimum pretože matica druhých derivácií F je $2A^T A$, ktorá je pozitívne semidefinitná.

Ak A má plnú stĺpcovú hodnosť (h(A)=n), tak $A^T A$ je regulárna a jediné riešenie je

$$x = (A^T A)^{-1} A^T b.$$

V prípade, že A nemá plnú stĺpcovú hodnosť nájdeme (jednoznačne určené) riešenie s najmenšou normou, ktoré je geometricky ortogonálnym priemetom bodu O (stred súradnicovej sústavy) do afinného priestoru riešení normálových rovníc. Toto riešenie možno nájsť použitím SVD nasledovne.

Nech $h(A) < n$. Vezmime singulárny rozklad $A = UDV^T$. Potom

$$\|b - Ax\|_2 = \|b - UDV^T x\|_2 = \|U^T b - DV^T x\|_2.$$

Označme $V^T x = y$ a $U^T b = f$. Potom $Ax = b$ môžeme prepísať na sústavu $Dy = f$ v tvare

$$\begin{pmatrix} D_r & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}$$

Potom dostávame

$$\begin{aligned} \|Dy - f\|_2^2 &= \left\| \begin{pmatrix} D_r y_1 \\ 0 \end{pmatrix} - \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} \right\|_2 = \left\| \begin{pmatrix} D_r y_1 - f_1 \\ -f_2 \end{pmatrix} \right\|_2 \\ &= \|D_r y_1 - f_1\|_2^2 + \|f_2\|_2^2 \end{aligned}$$

a minimum teda získame pre $y_1 = D_r^{-1} f_1$ a y_2 ľubovoľné. Riešenie s najmenšou normou vieme získať ako

$$y = \begin{pmatrix} D_r^{-1} f_1 \\ 0 \end{pmatrix} = \begin{pmatrix} D_r^{-1} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} = \begin{pmatrix} D_r^{-1} & 0 \\ 0 & 0 \end{pmatrix} f = \begin{pmatrix} D_r^{-1} & 0 \\ 0 & 0 \end{pmatrix} U^T b$$

a keďže $x = Vy$, dostaneme hľadané minimum pre

$$x = V \begin{pmatrix} D_r^{-1} & 0 \\ 0 & 0 \end{pmatrix} U^T b$$

Označme $D^+ = \begin{pmatrix} D_r^{-1} & 0 \\ 0 & 0 \end{pmatrix}$ Matica $A^+ = VD^+U$ sa nazýva pseudoinverzná matica k matici A a má nasledovné vlastnosti

- A^+ je jediná matica, ktorá spĺňa tzv. Moorove-Penroseove podmienky

$$AXA = A \qquad XAX = X \qquad (AX)^T = AX \qquad (XA)^T = (XA)$$

- $\min_{A \in \mathbb{R}^{m,n}} \|AX - I\|_F$ je dosiahnuté pre A^+
- Ak A je regulárna, tak $A^+ = A^{-1}$.
- Vo všeobecnosti $(AB)^+ \neq B^+A^+$.
- Závislosť A^+ od koeficientov A nie je spojitá.
- $(A^+)^+ = A$, $(A^T)^+ = (A^+)^T$, $(\alpha A)^+ = \frac{1}{\alpha}A^+$.

PNŠ pre matice s plnou hodnosťou

Budeme predpokladať, že $m \times n$ matica A, $m > n$, má hodnosť $h(A) = n$. Vtedy A^TA je regulárna matica, takže môžeme systém $A^TAx = A^Tb$ riešiť ako klasické systémy so symetrickou, kladne definitnou maticou (napr. Choleskeho rozkladom).

Problémom je že tento nový systém môže mať horšie numerické vlastnosti ako pôvodný

Príklad: Ak

$$A = \begin{pmatrix} 1 & 1 \\ 0 & a \\ a & 0 \end{pmatrix}$$

tak

$$A^TA = \begin{pmatrix} 1 + a^2 & 1 \\ 1 & 1 + a^2 \end{pmatrix}. \tag{18}$$

ak teda napríklad $a = 10^{-9}$, tak A^TA môže vyjsť pri počítaní s konečnou presnosťou $\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$, čo je singulárna matica, pričom pôvodná matica A mala hodnosť 2.

Namiesto riešenia normálnych rovníc môžeme použiť QR-rozklad $A = QR$, kde Q má rozmer $m \times m$. Myšlienka je podobná ako pri použití SVD na minimalizáciu vyššie.

Minimalizujme

$$F(x) = \|b - Ax\|_2^2 = \|b - QRx\|_2^2 = \|Q^Tb - Rx\|_2^2 = \|f - Rx\|_2^2,$$

pričom sme označili $f = Q^Tb$. Potom pre

$$R = \begin{pmatrix} R \\ 0 \end{pmatrix}$$

máme

$$f - Rx = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} - \begin{pmatrix} R_1x \\ 0 \end{pmatrix}$$

takže minimum $F(x)$ sa zjavne nadobúda pre $x = R_1^{-1}f_1$, a vtedy $F(x) = \|f_2\|_2^2$.

Pre matice s neúplnou hodnosťou budeme potrebovať nasledujúcu konštrukciu

QR-rozklad s pivotizáciou

Ak matica A nemá plnú stĺpcovú hodnosť, je vhodné QR-rozklad počítať tak, aby prvky na uhlopriečke v R boli usporiadané podľa absolutnej hodnoty. Tým pádom (numericky) nulové r_{kk} dostaneme až na konci výpočtu. Na to potrebujeme z A vyberať vždy ten stĺpec, ktorý má najväčšiu normu. Celý postup, môžeme zhrnúť nasledovne.

Začneme s maticou A. V nej nájdeme stĺpec s najväčšou normou a vymeníme tento stĺpec s prvým pomocou permutačnej matice Π_1. Následne použitím vhodnej matice odrazu vynulujeme časť pod diagonálou.

$$A \to A\Pi_1 \to P_1A\Pi_1$$

Matica P_1 nemení normu stĺpcov $A\Pi_1$, pretože P_1 je ortogonálna, teda jej riadky tvoria ortonormálnu bázu a súčin vektora (so súradnicami vyjadrenými v nejakej inej ortonormálnej báze) s vektormi ortonormálnej bázy

(riadky P_1) vracia súradnice toho istého vektora vyjadrené v tej ortonormálnej báze (riadky P_1). Jedná sa teda len o zmenu súradníc z jednej ortonormálnej bázy do inej ortonormálnej bázy, a takáto zmena súradníc zachováva dĺžky vektorov (t.j. 2-normu vektorov).

Opakovaním vyššie uvedeného postupu dostaneme teda postupne

$$A \to A\Pi_1 \to P_1 A\Pi_1 \to P_2 P_1 A\Pi_1\Pi_2 \to \cdots \to P_l \cdots P_2 P_1 A\Pi_1\Pi_2 \cdots \Pi_l = R.$$

Označme potom $Q^T = P_l \cdots P_2 P_1$, $\Pi = \Pi_1\Pi_2 \cdots \Pi_l$. Dostaneme tak $Q^T A\Pi = R$, čiže $A\Pi = QR$, čo je QR rozklad s pivotizáciou.

PNŠ pre matice s neplnou hodnosťou

Namiesto QR-rozkladu $n \times m$ vezmeme QR-rozklad s pivotizáciou $A\Pi = QR$. Potom miesto $Ax = b$ riešime sústavu $QR\Pi^T x = b$, pričom

$$R = \begin{pmatrix} R_{11} & R_{12} \\ 0 & 0 \end{pmatrix}$$

kde R_{11} má rozmer $r \times r$ (R má rozmer $n \times m$). Pomocou jednoduchých úprav $\|Ax - b\|_2$ dostaneme

$$\|Ax - b\|_2^2 = \|R\Pi^T x - Q^T b\|_2^2$$

Označme $\Pi^T x = \begin{pmatrix} y \\ z \end{pmatrix}$ tak že má zmysel násobiť R_{11} s y a R_{12} so z a $Q^T b = \begin{pmatrix} c \\ d \end{pmatrix}$, takže

$$\begin{aligned} \|R\Pi^T x - Q^T b\|_2^2 &= \left\| \begin{pmatrix} R_{11}y + R_{12}z \\ 0 \end{pmatrix} - \begin{pmatrix} c \\ d \end{pmatrix} \right\|_2^2 \\ &= \|R_{11}y - (c - R_{12}z)\|_2^2 + \|d\|_2^2. \end{aligned}$$

Teda x pre ktoré je $\|Ax - b\|_2^2$ minimálna musí byť

$$x = \Pi \begin{pmatrix} y \\ z \end{pmatrix} = \begin{pmatrix} R_{11}^{-1}(c - R_{12}z) \\ z \end{pmatrix}$$

Takzvané základné riešenie dostaneme pre $z = 0$, vtedy

$$x = \Pi \begin{pmatrix} R_{11}^{-1}c \\ 0 \end{pmatrix},$$

ktoré však nemusí mať najmenšiu normu. Riešenie s najmenšou normou môžeme nájsť tak že ešte upravíme maticu R pomocou matíc odrazu. Maticu

$$R = \begin{pmatrix} R_{11} & R_{12} \\ 0 & 0 \end{pmatrix}$$

upravíme na tvar

$$\tilde{R} = RU = \begin{pmatrix} \tilde{R}_{11} & 0 \\ 0 & 0 \end{pmatrix}$$

Maticu U získame ako súčin vhodných matíc odrazu $U = P_1 P_2 \cdots P_r$, schematicky pre maticu R (prázdne miesta označujú 0)

$$\begin{pmatrix} x & x & x & x & x & x & x \\ & x & x & x & x & x & x \\ & & x & x & x & x & x \\ & & & x & x & x & x \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \end{pmatrix} \xrightarrow{\cdot P_1} \begin{pmatrix} x & x & x & x & x & x & x \\ & x & x & x & x & x & x \\ & & x & x & x & x & x \\ & & & x & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \end{pmatrix} \xrightarrow{\cdot P_2}$$

$$\begin{pmatrix} x & x & x & x & x & x & x \\ & x & x & x & x & x & x \\ & & x & x & & & \\ & & & x & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \end{pmatrix} \xrightarrow{\cdot P_3} \begin{pmatrix} x & x & x & x & x & x & x \\ & x & x & x & & & \\ & & x & x & & & \\ & & & x & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \end{pmatrix} \xrightarrow{\cdot P_4} \cdots$$

Potom $R = \tilde{R}U^T$.

$$\|Ax - b\|_2^2 = \|QR\Pi^T x - b\|_2^2 = \|R\Pi^T x - Q^T b\|_2^2 = \|\tilde{R}U^T\Pi^T x - Q^T b\|_2^2$$

Podobne ako vyššie označme

$$U^T \Pi^T x = \begin{pmatrix} y \\ z \end{pmatrix}$$

takže

$$\|\tilde{R} U^T \Pi^T x - Q^T b\|_2^2 = \left\| \begin{pmatrix} \tilde{R}_{11} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} y \\ z \end{pmatrix} - \begin{pmatrix} c \\ d \end{pmatrix} \right\|_2^2 = \|\tilde{R}_{11} y - c\|_2^2 + \|d\|_2^2$$

Teda $\|Ax - b\|_2^2$ sa opäť minimalizuje voľbou $y = \tilde{R}_{11}^{-1} c$ a ľubovoľného z. Vtedy

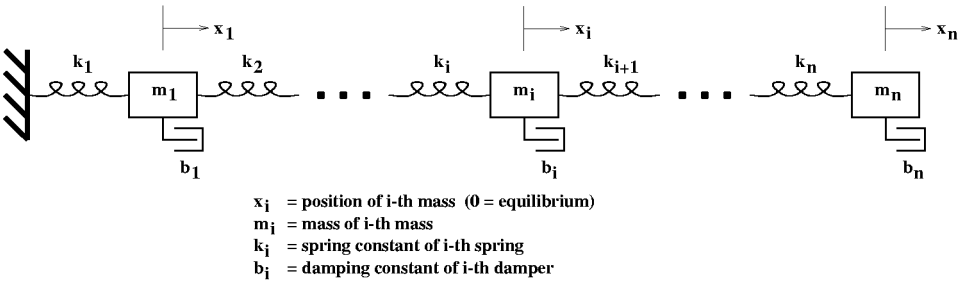
$$U^T \Pi^T x = \begin{pmatrix} \tilde{R}_{11}^{-1} c \\ z \end{pmatrix},$$

čiže

$$x = U \Pi \begin{pmatrix} \tilde{R}_{11}^{-1} c \\ z \end{pmatrix},$$

a takéto x má minimálnu normu, pre $z = 0$.

Problém vlastných hodnôt



Obr. 3: Systém mechanických oscilátorov (J. W. Demmel, Applied numerical linear algebra, Society for industrial and applied mathematics, 1997)

Skúmajme mechanický systém z obr. 1. Použitím Newtonovho pohybového zákona a Hookovho zákona dostávame systém diferenciálnych rovníc druhého rádu

$$m_i \ddot{x}_i(t) = k_i(x_{i-1}(t) - x_i(t)) + k_{i+1}(x_{i+1}(t) - x_i(t)) - b_i \dot{x}_i(t)$$

Systém n takých rovníc môžeme zapísať v maticovom tvare ako

$$M\ddot{x}(t) = -B\dot{x}(t) - Kx(t),$$

kde $M = \mathrm{diag}(m_1, \ldots, m_n)$, $B = \mathrm{diag}(b_1, \ldots, b_n)$ a

$$K = \begin{pmatrix} k_1 + k_2 & -k_2 & & & \\ -k_2 & k_2 + k_3 & -k_3 & & \\ & \ddots & \ddots & \ddots & \\ & & -k_{n-1} & k_{n-1} + k_n & -k_n \\ & & & -k_{n-1} & k_n \end{pmatrix}$$

Tento systém ODR druhého rádu transformujeme na systém ODR prvého rádu substitúciou

$$\dot{y}(t) = \begin{pmatrix} \ddot{x}(t) \\ \dot{x}(t) \end{pmatrix}$$

Tým dostávame

$$\begin{aligned} \dot{y}(t) = \begin{pmatrix} \ddot{x}(t) \\ \dot{x}(t) \end{pmatrix} = \begin{pmatrix} -M^{-1}B\dot{x}(t) - M^{-1}Kx(t) \\ \dot{x}(t) \end{pmatrix} = \\ = \begin{pmatrix} -M^{-1}B & -M^{-1}K \\ I & 0 \end{pmatrix} \begin{pmatrix} \dot{x}(t) \\ x(t) \end{pmatrix} = Ay(t) \end{aligned}$$

Na vyriešenie tohoto systému potrebujeme poznať začiatočnú polohu a rýchlosť závaží, teda vektor $y(0)$. V teórii obyčajných diferenciálnych rovníc sa ukazuje, že takýto systém s danými začiatočnými podmienkami má riešenie $y(t) = e^{At}y(0)$. Ak A je diagonalizovateľná, t.j. $A = S\Lambda S^{-1}$, kde $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, dostávame

$$\dot{y}(t) = S\Lambda S^{-1}y(t) \Leftrightarrow S^{-1}\dot{y}(t) = \Lambda S^{-1}y(t)$$

takže po substitúcii $z(t) = S^{-1}y(t)$ dostávame systém $\dot{z} = \Lambda z(t)$, ktorý má riešenie $z(t) = \mathrm{diag}(e^{\lambda_1 t}, \ldots, e^{\lambda_n t})z(0)$.

Vidíme, na nájdenie riešenia takéhoto dynamického systému je dobré poznať vlastné čísla matice asociovanej s týmto systémom. O vlastných číslach nejakej matice A vieme, že sú to riešenia charakteristického polynómu matice A, $\chi(\lambda) = \det(A - \lambda I)$, preto pre prípad veľkých matíc nemáme inú možnosť ako použiť numerické metódy. Zrejme výpočet koreňov charakteristického

polynómu nebude veľmi efektívny, pretože by sme potrebovali spočítať determinant matice, čo je časovo drahá operácia.

Každú štvorcovú maticu A je možné upraviť na Jordanov kanonický tvar $J(A) = S^{-1}AS$, čo je blokovo diagonálna matica, pozostávajúca z blokov, ktoré majú na diagonále vlastné čísla matice A a nad diagonálou jednotky. Mohli by sme sa teda pokúsiť numericky hľadať tento kanonický tvar. Problémom je však to, že zobrazenie $A \to J(A)$ nie je spojité.

$$A = \begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{pmatrix}$$

je matica v Jordanovom tvare. Ak zvolíme ľubovoľné ϵ a nahradíme prvok na $a_{ii} = 0$ na diagonále prvkom $i \cdot \epsilon$, dostaneme maticu, ktorej Jordanov kanonický tvar bude $\mathrm{diag}(1 \cdot \epsilon, \ldots, n \cdot \epsilon)$. Takže pri napríklad pri zaokrúhľovacej chybe by sme mohli dostať uplne iný výsledok.

Taktiež výpočet Jordanovej kanonickej formy nemusí byť stabilný, teda ak by sme aj spočítali S a J pre maticu A, tak nemáme garanciu , že $S^{-1}(A+\delta A)S = J$ pre nejaké malé δA. Totiž, ak vieme, že $S^{-1}AS = J$, kde S je veľmi zle podmienená, a spočítali sme S presne a J s nejakou malou chybou δJ, $||\delta J|| = O(\epsilon)||A||$, tak vieme odhadnúť, ako veľmi máme "opraviť" A, aby J bolo presné riešenie, t.j. hľadáme odhad pre δA, s $S^{-1}(A + \delta A)S = J + \delta J$. Po jednoduchej úprave dostávame $\delta A = S\delta J S^{-1}$ a odhad

$$||\delta A|| \leq ||S|| \cdot ||\delta J|| \cdot ||S^{-1}|| = c(S)O(\epsilon)||A||,$$

takže $||\delta A||$ môže byť veľké.

Posledný príklad naznačuje, že bude dobré kontrolovať podmienenosť matice S. Môžeme sa teda skúsiť obmedziť na matice ortogonálne matice S. O tých vieme, že majú číslo podmienenosti $c(S) = 1$. Tieto síce nestačia na získanie Jordanovho kanonického tvaru, avšak máme

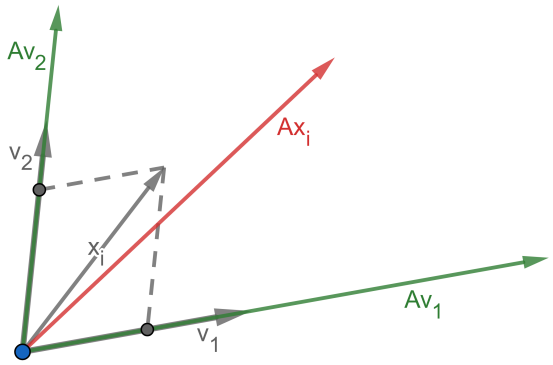
Veta 4 (Shurov kanonický tvar)**.** *Pre danú maticu A existuje unitárna matica Q a horná trojuholníková matica T taká, že $Q^*AQ = T$ a vlastné hodnoty matice A sú prvky diagonály matice T.*

V tejto vete vo všeobecnosti predpokladáme, že tam uvedené matice sú komplexné. V prípade, že A je reálna matica ale preferujeme kanonickú formu, ktorá bude obsahovať len reálne prvky. V takom prípade ale musíme obetovať "trojuholníkovosť"matice.

Veta 5 (Reálny Shurov kanonický tvar). *Ak A je reálna matica, tak existuje reálna ortogonálna matica V taká, že $V^T AV = T$ je blokovo horná trojuholníková matica, s blokmi rozmeru 1×1 alebo 2×2 na diagonále. Bloky rozmeru 1×1 zodpovedajú reálnym vlastným hodnotám matice A a bloky rozmeru 2×2 dvojici komplexne združených komplexných vlastných čísel matice A.*

Algoritmus pre všeobecný problém vlastných čísel

Najjednoduchšia metóda je **mocninová metóda**. Táto metóda vo všeobecnosti nájde najväčšie vlastné číslo a príslušný vlastný vektor. Ideu tejto metódy možno geometricky opísať obrázkom Ak v_1 a v_2 sú vlastné vektory ma-



tice A prislúchajúce rôznym vlastným hodnotám, a x_i je ľubovoľný vektor, ktorý nie je vlastným vektorom, tak vektor Ax_i sa prikloní k násobku vlastného vektora, ktorý prislúcha k väčšej vlastnej hodnote. Príslušnú vlastnú hodnotu potom dostaneme normovaním a projekciou. Celý postup sa dá zhrnúť do algoritmu

- Zvoľ vektor x_0, $i = 0$
- opakuj
  - $y_{i+1} = Ax_i$
  - $x_{i+1} = y_{i+1}/||y_{i+1}||_2$
  - $\tilde{\lambda}_{i+1} = x_{i+1}^T Ax_{i+1}$
  - $i = i + 1$
- kým nedôjde ku konvergencii.

Ukážme, že táto metóda naozaj konverguje. Predpokladajme, že A je diagonalizovateľná, $A = S\Lambda S^{-1}$, $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, pričom $|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_n|$. Označme s_i i-ty stĺpec matice S. s_i je vlastný vektor prislúchajúci λ_i a $||s_i||_2 = 1$. Potom $x_0 = S(S^{-1}x_0) = S(\alpha_1, \ldots, \alpha_n)^T$. Tiež $A^i = S\Lambda^i S^{-1}$. Potom

$$A^i x_0 = (S\Lambda^i S^{-1})S \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{pmatrix} = S \begin{pmatrix} \lambda_1^i \alpha_1 \\ \lambda_2^i \alpha_2 \\ \vdots \\ \lambda_n^i \alpha_n \end{pmatrix} = \alpha_1 \lambda_1^i S \begin{pmatrix} 1 \\ \frac{\alpha_2}{\alpha_1}\left(\frac{\lambda_2}{\lambda_1}\right)^i \\ \vdots \\ \frac{\alpha_n}{\alpha_1}\left(\frac{\lambda_n}{\lambda_1}\right)^i \end{pmatrix}$$

a vektor v zátvorkách zrejme konverguje k e_1, takže $A^i x_0$ konverguje k násobku $Se_1 = s_1$, čo je vlastný vektor zodpovedajúci vlastnej hodnote λ_1. Takže $\tilde{\lambda}_i = x_i^T A x_i$ konverguje k $s_1^T A s^1 = s_1^T \lambda_1 s_1 = \lambda_1$. Problémom tejto metódy je, (zamlčaný) predpoklad $\alpha_1 \neq 0$. A pomalá konvergencia v prípade, že pomer $|\lambda_2|/|\lambda_1|$ je blízky 1. Napríklad, A je reálna matica ktorej v absolútnej hodnote najväčšie vlastné číslo $|\lambda_1|$ je komplexné, tak $|\lambda_1| = |\lambda_2|$ a vyššie uvedená analýza zlyhá. Extrémny prípad je ortogonálna matica, ktorej všetky vlastné hodnoty sú rovné 1.

Uvedené nedostatky však možno obísť keď uvedenú metódu aplikujeme namiesto matice A na maticu $(A - \sigma I)^{-1}$. Číslo σ nazývame posun. Táto modifikácia umožní konvergenciu k vlastnej hodnote, ktorá je najbližšia k σ. Tejto metóde sa hovorí **inverzná mocninová metóda**, alebo **metóda inverznej iterácie**.

- Zvoľ vektor x_0, $i = 0$
- opakuj
  - $y_{i+1} = (A - \sigma I)^{-1} x_i$
  - $x_{i+1} = y_{i+1}/||y_{i+1}||_2$
  - $\tilde{\lambda}_{i+1} = x_{i+1}^T A x_{i+1}$
  - $i = i + 1$
- kým nedôjde ku konvergencii.

Ďalšie vylepšenie algoritmu umožní spočítať $(p > 1)$-invariantný podpriestor. Tento spôsob nazveme **ortogonálna iterácia**

- Zvoľ $n \times p$ ortogonálnu maticu Z_0, $i = 0$

- opakuj
  - $Y_{i+1} = AZ_i$
  - $Y_{i+1} = Z_{i+1}R_{i+1}$ pomocou QR-rozkladu, Z_{i+1} približne aproximuje p-invariantný podpriestor.
  - $i = i + 1$
- kým nedôjde ku konvergencii.

Predpokladajme, že $|\lambda_p| > |\lambda_{p+1}|$. Pre $p = 1$ dostávame klasickú mocninovú metódu. Ak $p > 0$, tak máme

$$\text{span}(Z_{i+1}) = \text{span}(Y_{i+1}) = \text{span}(AZ_i).$$

Preto

$$\text{span}(Z_i) = \text{span}(A^i Z_0) = \text{span}(S\Lambda S^{-1} Z_0)$$

Rozpísaním $S\Lambda S^{-1}Z_0$ máme

$$S\Lambda S^{-1}Z_0 = S\text{diag}(\lambda_1^i, \ldots, \lambda_n^i)S^{-}1Z_0 = \lambda_p^i S\text{diag}((\lambda_1/\lambda_p)^i, \ldots, (\lambda_1/\lambda_n)^i)SZ_0$$

Keďže $|\lambda_j/\lambda_p| \geq 1$, pre $j \leq p$ a $|\lambda_j/\lambda_p| < 1$ pre $j > p$ dostávame

$$\text{diag}((\lambda_1/\lambda_p)^i, \ldots, (\lambda_1/\lambda_n)^i)SZ_0 = \begin{pmatrix} X_i^{p\times p} \\ Y_i^{(n-p)\times p} \end{pmatrix}$$

pričom zjavne $Y_i^{(n-p)\times p}$ konverguje k 0 ako $(\lambda_{p+1}/\lambda_p)^i$ a $X_i^{p\times p}$ nekonverguje k 0.

Označme S_p maticu obsahujúcu prvých p stĺpcov matice S a $\hat{S}_p$ maticu zvyšných stĺpcov.

Potom

$$\text{span}(Z_i) = \text{span}(S\Lambda^i S^{-1}Z_0) = \text{span}(S_p X_i + \hat{S}_p Y_i)$$

a keďže Y_i konverguje k 0 tak $\text{span}(S_p X_i + \hat{S}_p Y_i)$ konverguje ku $\text{span}(S_p X_i) = \text{span}(S_p)$

Veta 6. *Pre danú $n \times n$ maticu A, nech $p = n$ a $Z_0 = I$. Ak všetky vlastné hodnoty matice A sú v absolútnych hodnotách navzájom rôzne, a všetky hlavné podmatice matice S sú regulárne, tak $A_i = Z_i^T AZ_i$ konverguje ku Schurovmu kanonickému tvaru matice A. Vlastné hodnoty budú v tomto tvare usporiadané zostupne podľa ich absolútnej hodnoty.*

Ďalšie modfikácia umožnia odstrániť predpoklad rôznosti absolútnych hodnôt vlastných čísel. Dosiahneme to pomocou posunu a invertovania. Najskôr jemne modifikujeme predošlý algoritmus. Dostaneme algoritmus QR iterácie.

- Daná je matica A_0, $i = 0$
- opakuj
  - $A_i = Q_i R_i$, pomocou QR-rozkladu.
  - $A_{i+1} = R_i Q_i$.
  - $i = i + 1$
- kým nedôjde ku konvergencii.

V tomto prípade matice A_i sú presne matice $Z_i^T A T_i$ z predošlého algoritmu.

Pridaním posunu získame algoritmus

- Daná je matica A_0, $i = 0$
- opakuj
  - zvoľ σ_i blízko vlastnej hodnoty matice A_i.
  - $A_i - \sigma_i I = Q_i R_i$, pomocou QR-rozkladu.
  - $A_{i+1} = R_i Q_i + \sigma_i I$.
  - $i = i + 1$
- kým nedôjde ku konvergencii.

Potom nie je ťažké ukázať, že A_i a A_{i+1} sú ortogonálne podobné.

Ak σ_i je vlastná hodnota, tak algoritmus skonverguje po prvej iterácii. Ak σ_i nie je vlastná hodnota, tak zastavovacia podmienka je aby $A(n, 1 : n-1)$ bola dostatočne malá.

Ešte ostáva vyriešiť problém, ako voliť σ_i blízko vlastných hodnôt, keď tieto vlastné hodnoty hľadáme. Ako vhodná voľba sa ukazuje $\sigma_i = A_i(n, n)$, ktorá umožní kvadratickú konvergenciu.

Poznámky: Časová zložitosť uvedeného algoritmu je $O(n^4)$. Dá sa vylepšiť na $O(n^3)$ redukciou matice A na horný Hessenbergov tvar. Problém komplexných vlastných hodnôt je možné vyriešiť dvojitým posunom σ_i a $\bar{\sigma}_i$.

Literatúra

[1] T. Bušinská, *Numerické metódy lineárnej algebry*, skriptá, Univerzita Komenského, 1993

[2] G. H. Golub, C. F. Van Loan, *Matrix Computations*, John's Hopkins University Press, 1996

[3] J. W. Demmel, *Applied Numerical Linear Algebra*, Society for Indrustrial and Applied Mathemaitcs, 1997

[4] C. D. Meyer, *Matrix Analysis and Applied Linear Algebra*, SIAM, 2000