Iterační metody numerické lineární algebry

- srovnání přímých metod (viz zimní semestr) a iteračních metod pro $Ax=b$

  <u>přímé metody</u> (LU rozklad, QR rozklad)

  + konečný počet kroků $\sim n^3$
  + opakovatelná využitelnost, např. máme-li více pravých stran
  − velká paměťová náročnost i pro řídké matice, neboť LU rozklad už nemusí být řídký

  <u>iterační metody</u>

  + mnohdy stačí jen výpočet (opakovaný) $A \cdot x$, pro řídké matice může být až $O(n)$, někdy není třeba ani matici A ukládat
  + může občas stačit jen několik iterací k dosažení požadované přesnosti ⇒ můžou být rychlejší než přímé metody
  − není zaručena konvergence pro obecnou soustavu ale často jen pro určitý typ matic
  − mnohdy nutné tzv. předpodmínění, aby byla konvergence dostatečně rychlá, často je „šité na míru" daného problému

Pozn: pracujeme-li s konečnou aritmetikou, pak kvůli zaokrouhlovacím chybám nedávají ani přímé metody přesné výsledky ⇒ iterační metody se používají k zpřesnění přímých metod

- budeme se zabývat
  <u>stacionárními</u> iteračními metodami, na nichž staví pokročilá efektivní metoda <u>multigridu</u>
  nestacionárními iter. metodami využívajícími tzv. <u>Krylovovy prostory</u>

• stacionární iterační metody

- základní myšlenka: rozložit $A = M - K$ tak, aby
  M byla regulární matice a šla snadno spočítat M^{-1}
  pak z $Ax = Mx - Kx = b$
  dostaneme $x = M^{-1}Kx + M^{-1}b = Rx + c$
  a hledáme x metodou prosté iterace
  $$\boxed{x_{m+1} = Rx_m + c}$$ s vhodnou volbou x_0
  pokud jsou R a c konstantní během iterace, jde o stacionární metodu

- <u>konvergence</u> závisí na vlastnostech matice R
  odečtením $x = Rx + c$, kde x je řešením $Ax=b$
  dostaneme $x_{m+1} - x = R(x_m - x)$
  $$\|x_{m+1} - x\| \leq \|R\| \, \|x_m - x\| \leq \|R\|^{m+1} \|x_0 - x\|$$
  pokud $\|R\| = \max\limits_{x \neq 0} \frac{\|Rx\|}{\|x\|}$ ~~je~~ operátorová norma pro danou vektorovou normu
  vidíme, že $x_m \to x$ a tedy $\|x_m - x\| \to 0$
  pokud $\underline{\|R\| < 1}$
  Navíc lze ukázat, že x_m konverguje k řešení $Ax = b$ pro lib. počáteční x_0 a pro lib. pravou stranu b, právě když <u>spektrální poloměr matice</u>
  $$\varrho(R) \equiv \max_{\substack{\text{vlastní}\\ \text{čísla } \lambda(R)}} |\lambda| < 1$$

- <u>rychlost konvergence</u> $r(R) = -\log_{10} \varrho(R)$
  zhruba udává, kolik správných desetinných míst přibude v jedné iteraci, neboť z $\underbrace{\|x_{m+1} - x\| \leq \varrho(R) \|x_m - x\|}_{\text{toto platí pro vhodnou vekt. normu}}$
  máme $\log_{10} \|x_m - x\| - \log_{10} \|x_{m+1} - x\| \geq r(R)$

- cíl iteračních metod:

  1) zvolit $M-K$ tak, aby šlo $M^{-1}Kx = Rx$ a $M^{-1}b = c$ snadno spočítat (M je typicky diagonální, nebo trojúhelníková matice)

  2) a zároveň tak, aby $\varrho(R)$ byl co nejmenší

Bohužel jde o konfliktní požadavky, neboť jsou

2 extrémy: 1) $M = I$ je ideální pro $M^{-1} = I$, ale $\varrho(R)$ může být mnohem větší než 1

2) $M = A$ a tedy $K = 0$ $\Rightarrow$ $\varrho(R) = 0$ ovšem typicky A^{-1} neumíme snadno

$\Rightarrow$ nutný kompromis

- Jacobiho metoda

myšlenka: vyjádříme x_j zodpovídající j-té rovnice přičemž za ostatní x_i dosadíme předchozí hodnoty

$$x_{m+1,j} = \frac{1}{a_{jj}}\Big(b_j - \sum_{k\neq j} a_{jk}x_{m,k}\Big)$$

(pokud $a_{jj} \neq 0$, pokud ne, tak nutno přeuspořádat řádky)

Jde o rozklad $A = M - K = \underset{\text{diagonální prvky}}{D} - (\tilde{L} + \tilde{U})$ prvky pod- a nad diagonálou vzaté s opačným znaménkem

a tedy $R_{Jac} = D^{-1}(\tilde{L} + \tilde{U})$, $c_{Jac} = D^{-1}b$

Př. Uvažujme matici $A = \underbrace{\begin{pmatrix} 2 & -1 & & 0 \\ -1 & 2 & -1 & \\ & -1 & 2 & \ddots & -1 \\ 0 & & & -1 & 2 \end{pmatrix}}_{n} \Big\} n$ jde o matici, která se objeví při řešení jednorozměrné Poisonovy rovnice na gridu (viz cvičení v Mathematice)

Pro tuto matici máme

$$D^{-1} = \begin{pmatrix} \frac{1}{2} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{2} \end{pmatrix} \quad \text{a} \quad R_{Jac} = \begin{pmatrix} 0 & \frac{1}{2} & & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & \\ & \frac{1}{2} & \ddots & \frac{1}{2} \\ 0 & & \frac{1}{2} & 0 \end{pmatrix}$$

Pro tuto matici lze ukázat, že $\varrho(R_{Jac}) = \cos\frac{\pi}{n+1} \approx 1 - \frac{\pi^2}{2(n+1)^2}$

a tedy pro $n \to \infty$ je $\varrho(R_{Jac}) \to 1$ a konverguje stále pomaleji.

např. pro $n = 100$ je $\varrho(R_{Jac}) = 0{,}999516\ldots$

- ## Gaussova - Seidelova metoda

modifikace Jacobiho metody: při výpočtu x_j použijeme už „updatované" hodnoty $x_1, \dots, x_{j-1}$

neboli
$$x_{m+1,j} = \frac{1}{a_{jj}} \Big(b_j - \sum_{k=1}^{j-1} \underbrace{a_{jk}\, x_{m+1,k}}_{\text{nové hodnoty}} - \sum_{k=j+1}^{n} \underbrace{a_{jk}\, x_{m,k}}_{\text{staré hodnoty}} \Big)$$

lze přepsat maticově
$$(D - \tilde{L})\, x_{m+1} = \tilde{U} x_m + b$$

a tedy $\quad R_{GS} = (D-\tilde{L})^{-1} \tilde{U} \quad , \quad c_{GS} = (D-\tilde{L})^{-1} b$

na rozdíl od Jacobiho metody záleží na pořadí, v jakém počítáme x_j

někdy vhodné přeuspořádání může zrychlit konvergenci

Př. opět uvažujme
$$A = \underbrace{\begin{pmatrix} 2 & -1 & & 0 \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & 2 \end{pmatrix}}_{n} \Rightarrow R_{GS} = \overbrace{\begin{pmatrix} 2 & & & 0 \\ -1 & 2 & & \\ & \ddots & \ddots & \\ 0 & & -1 & 2 \end{pmatrix}}^{D-\tilde{L}}{}^{-1} \overbrace{\begin{pmatrix} 0 & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & 0 \end{pmatrix}}^{\tilde{U}} = \begin{pmatrix} \frac{1}{2} & & & 0 \\ \frac{1}{4} & \ddots & & \\ \vdots & \ddots & \ddots & \\ \frac{1}{2^n} & \cdots & \frac{1}{4} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 0 & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & 0 \end{pmatrix} =$$

$$= \begin{pmatrix} 0 & \frac{1}{2} & & 0 \\ \vdots & \frac{1}{4} & \ddots & \\ \vdots & \vdots & \ddots & \frac{1}{2} \\ 0 & \frac{1}{2^n} & \cdots & \frac{1}{4} \end{pmatrix}$$

a pro $n = 100$ nyní dostaneme $\varrho(R_{GS}) = 0{,}999033$

sice $< \varrho(R_{Jac})$, ale stále pomalé

- ## SOR(ω) = successive overrelaxation (superrelaxační metoda)

modifikace G-S metody: místo $x_{m+1,j}$ z G-S metody použijeme vážený průměr $(1-\omega)\, x_{m,j} + \omega\, x_{m+1,j}$

neboli
$$x_{m+1,j} = (1-\omega)\, x_{m,j} + \frac{\omega}{a_{jj}} \Big(b_j - \sum_{k=1}^{j-1} a_{jk} x_{m+1,k} - \sum_{k=j+1}^{n} a_{jk} x_{m,k} \Big)$$

a maticově
$$(D - \omega\tilde{L})\, x_{m+1} = \big[(1-\omega) D + \omega \tilde{U}\big] x_m + \omega b$$

$$R_{SOR(\omega)} = (D - \omega\tilde{L})^{-1} \big[(1-\omega) D + \omega\tilde{U}\big]$$

$$c_{SOR(\omega)} = (D - \omega\tilde{L})^{-1} \omega b$$

pro $\omega > 1$ jde o tzv. superrelaxaci ← toto se používá

pro $\omega = 1$ (G-S metoda) o relaxaci a pro $\omega < 1$ o „underrelaxation"

- pro metodu SOR(ω) lze ukázat (viz Demmel), že

1) obecně $\varrho(R_{SOR(\omega)}) \geq |\omega - 1|$ a tedy, aby metoda konvergovala, musí být $\underline{0 < \omega < 2}$.

2) pokud je A symetrická ($y^T A x = x^T A y$), pozitivně definitní matice ($x^T A x > 0$ pro $\forall x \neq 0$), pak $\varrho(R_{SOR(\omega)}) < 1$ pro lib. $0 < \omega < 2$

a tedy SOR(ω) konverguje (a tedy i G-S metoda ($\omega = 1$))

3) optimální volba $\omega_{opt} = \dfrac{2}{1 + \sqrt{1 - \varrho(R_{Jac})^2}}$ závisí na spektrální poloměru R_{Jac}!

a v tomto případě platí

$$\varrho(R_{SOR(\omega)}) = \left(\frac{\varrho(R_{Jac})}{1 + \sqrt{1 - \varrho(R_{Jac})^2}} \right)^2$$

Př. pro náš standardní problém

je $\varrho(R_{Jac}) \approx 1 - \dfrac{\pi^2}{2(n+1)^2}$ a tedy $\omega_{opt} \approx \dfrac{2}{1 + \frac{\pi}{n+1}}$ pro $n \to \infty$ $\omega \to 2$

a $\varrho(R_{SOR(\omega)}) \approx 1 - \dfrac{2\pi}{n+1}$

pro $n = 100$ je $\varrho(R_{SOR(\omega)}) \approx 0{,}94$ ve srovnání s $\varrho(R_{GS}) = 0{,}999$

takže by měla superrelax. metoda konvergovat mnohem rychleji

- obecně ale neznáme $\varrho(R_{Jac})$, takže musíme ω odhadovat a SOR(ω) bohužel funguje rychle jen pro hodnoty ω blízké ω_{opt}

odhady lze dělat např. na hrubším gridu (pokud matice A pochází z diskretizace určité diferenciální rovnice

nebo z odhadu rychlosti konvergence Jacobiho metody

Konvergence Jacobiho, G-S a SOR(ω) metody

- pro obecný problém (matici A) záleží na vlastnostech a struktuře A

Def: A je silně (řádkově) diagonálně dominantní, pokud $|a_{ii}| > \sum_{j \neq i} |a_{ij}|$ pro $\forall i$.

Věta: Pokud je A silně diagonálně dominantní, pak Jacobiho i Gaussova-Seidelova metoda konvergují a navíc $\|R_{GS}\|_\infty \le \|R_{Jac}\|_\infty < 1$.

Pozn: Nerovnost $\|R_{GS}\|_\infty \le \|R_{Jac}\|$ nutně neznamená, že G-S ~~je vždy rychlejší~~ konverguje vždy rychleji než Jacobi, neboť Jacobiho metoda může mít občas „náhodou" menší chybu v jistém kroku.

Dk konvergence: stačí ukázat, že

Př. Ani 1D, ani 2D elipt. problém nedává A silně diagonálně dominantní avšak jsou slabě diagonálně dominantní

Def: A je slabě diag. dominantní, pokud pro $\forall i$ platí $|a_{ii}| \ge \sum_{k \neq i} |a_{ik}|$ a v alespoň jednom případě platí ostrá nerovnost.

- ovšem toto samo o sobě nestačí na důkaz konvergence Jac. a G-S. metody je zapotřebí ještě tzv. ireducibilita matice A.

Def: A je ireducibilní matice, pokud neexistuje permutační matice P taková, že

$$PAP^T = \left(\begin{array}{c|c} A_{11} & A_{12} \\ \hline 0 & A_{22} \end{array}\right)$$

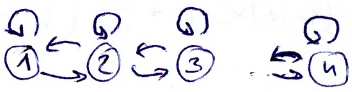
- vidíme, že při aplikaci na vektor x

$$\left(\begin{array}{c|c} A_{11} & A_{12} \\ \hline 0 & A_{22} \end{array}\right)\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} A_{11}x_1 + A_{22}x_2 \\ A_{22}x_2 \end{pmatrix}$$

nejsou hodnoty na pozicích x_2 ovlivněny hodnotami na x_1 a navíc vekt. $\begin{pmatrix} x_1 \\ 0 \end{pmatrix}$ tvoří „inv. podprostor" při aplikaci A

- ireducibilita souvisí s grafem matice

Def: Orientovaný graf matice A, značený $G(A)$, je graf s uzly $1, \dots, n$ a hranami z i do j, právě když $a_{ij} \neq 0$.

Př. $A = \begin{pmatrix} 2 & -1 & & 0 \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & 2 \end{pmatrix} \longrightarrow G(A) =$  silně souvislý

Def: Orient. graf se nazývá silně souvislý, pokud existuje cesta z lib. (i) do každého (j). Silně souvislá komponenta (část) orient. grafu je podgraf, který je silně souvislý a nemůže být rozšířen na větší silně souvislý podgraf.

Př. $\left(\begin{array}{c|c} A_{11} & A_{12} \\ \hline O & A_{22} \end{array}\right)$ prvky v A_{11} nejsou dosažitelné, startujeme-li z A_{22}, ale A_{22} či A_{11} můžou být silně souvislé komponenty.

Věta: A je ireducibilní, právě když $G(A)$ je silně souvislý

Dk. $\Rightarrow$ jasné z Př.

$\Leftarrow$ pokud $G(A)$ není silně souvislý, lze přerovnat řádky a sloupce tak, že silně souvislá komponenta bude na konci a dostaneme $\left(\begin{array}{c|c} A & A \\ \hline O & A \end{array}\right)$

Věta: Pokud je A ireducibilní a slabě diagonálně dominantní, pak obě Jacobiho a G-S metoda konvergují a navíc $\rho(R_{GS}) < \rho(R_J) < 1$

Dk: viz R.S. Varga: Matrix Iterative Analysis, Prentice-Hall 1962

Př. 1D a 2D modelové problémy vedou na ireduc. matice a jsou slabě diag. dominantní $\Rightarrow$ Jacobi i G-S jsou konvergentní.

- i přes tyto výsledky obecně neplatí, že Jacobi a G-S jsou konvergentní pro lib. A, vskutku jsou nesymetrické matice, pro něž Jacobi nefunguje, ale funguje G-S a naopak

---

- pokud jde o SOR(ω) metodu, tak platí

Věta: $\rho(R_{SOR(\omega)}) \geq |\omega - 1|$ a proto musí být vždy $0 < \omega < 2$, aby byla metoda konvergentní

Dk: $R_{SOR(\omega)} = (I - \omega L)^{-1}\left[(1-\omega)I + \omega U\right]$

Nechť $\varphi(\lambda) = \det(\lambda I - R_{SOR(\omega)}) = \det\left[\overbrace{(I-\omega L)}^{\text{toto má det} = 1 \text{ tj. vsunuta jednotka}}(\lambda I - R_{SOR(\omega)})\right] =$

$= \det\left[(\lambda + \omega - 1)I - \omega\lambda L - \omega U\right]$ je charakteristický polynom

a tedy $\varphi(\lambda) = \prod_{i=1}^{n}\left(\lambda - \lambda_i(R_{SOR(\omega)})\right) \Rightarrow |\varphi(0)| = \prod_{i=1}^{n}|\lambda_i| = \left|\det((\omega-1)I - \omega U)\right| =$

$= |(\omega - 1)^n|$

a tedy $\rho(R_{SOR(\omega)}) = \max_{i=1,\dots,n} |\lambda_i(R_{SOR(\omega)})| \geq |\omega - 1|$

Věta: Pokud je A symetrická pozitivně definitní matice, pak $\rho(R_{SOR(\omega)}) < 1$ pro $0 < \omega < 2$, a tedy SOR(ω) konverguje. Pro $\omega = 1$ máme G-S metodu, která tedy pro tuto A též konverguje.

Def: A je pozitivně definitní symetrická matice, pokud $x^T A x > 0$ pro $\forall x \neq 0$. (pro reálný případ)

A je — " — hermitovská matice, pokud $x^* A x > 0$ pro $\forall x \neq 0$.