

1 Iterační metody pro řešení soustav lineárních rovnic

Přímé metody řešení soustav lineárních rovnic (LU, LDLT, Choleského faktORIZACE atd.) vyžadují $\mathcal{O}(n^3)$ operací a velikost soustavy, kterou jimi dokážeme vyřešit je značně omezená. V případě husté matice $A \in \mathbb{R}^{n \times n}$ se přibližná velikost řešitelné soustavy historicky vyvíjela takto:

- 1950: $n = 20$ (Wilkinson)
- 1965: $n = 200$ (Forsythe a Moler)
- 1980: $n = 2000$ (LINPACK)
- 1995: $n = 20000$ (LAPACK)
- 2010: $n = 200000$ (HDSS)

Pracujeme-li s řídkými maticemi, jsme schopni řešit i systémy s mnohem větší dimenzí (miliony, desítky milionů neznámých) – zejména, pokud je řešič schopen pracovat paralelně. V případě použití přímého řešiče na řídkou matici však může dojít k jejímu zaplnění. Např. využitím prvního řádku následující matice k vynulování prvního sloupce dojde k zaplnění všech ostatních prvků v matici:

$$\begin{bmatrix} \times & \times & \times & \times & \times \\ \times & \times & 0 & 0 & 0 \\ \times & 0 & \times & 0 & 0 \\ \times & 0 & 0 & \times & 0 \\ \times & 0 & 0 & 0 & \times \end{bmatrix} \rightarrow \begin{bmatrix} \times & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times \end{bmatrix}$$

Tento problém lze částečně řešit vhodnou pivotizací.

Mezi další nevýhody přímých řešičů patří:

- Známe-li již přibližné řešení soustavy, nedokážeme tuto znalost využít ke snížení celkového počtu operací a zkrácení doby výpočtu.
- Naopak, pokud nám postačuje znalost pouze přibližného řešení, nemůžeme výpočet pomocí přímého řešiče ukončit předčasně.

Alternativou k přímým řešičům jsou iterační řešiče, které generují posloupnost přibližných řešení $\{\mathbf{x}^k\}$ a pracují téměř výhradně s násobením matice-vektor, které má náročnost $\mathcal{O}(n^2)$. Důležitou vlastností každé iterační metody je rychlost konvergence posloupnosti $\{\mathbf{x}^k\}$ k řešení. Může se totiž stát, že pro některé matice A iterační metoda konverguje velmi pomalu nebo vůbec.

1.1 Lineární iterační metody

Prvním typem iteračních metod, kterým se budeme zabývat, jsou tzv. lineární iterační metody. Ty hledají posloupnost řešení soustavy $A\mathbf{x} = \mathbf{b}$ ve tvaru

$$\mathbf{x}^{k+1} := M\mathbf{x}^k + N\mathbf{b}, \quad (1.1)$$

kde M a N jsou nějaké matice odpovídajících rozměrů¹.

Definice Lineární iterační metodu nazveme konzistentní, řeší-li rovnici $M\mathbf{x} + N\mathbf{b} = \mathbf{x}$ právě jeden vektor $\mathbf{x} = A^{-1}\mathbf{b}$. Je možné ukázat, že metoda je konzistentní právě tehdy, je-li splněno $M = I - NA$.

Definice Iterační metodu nazveme konvergentní platí-li $\mathbf{x}^k \rightarrow \mathbf{x} = A^{-1}\mathbf{b}$ pro $k \rightarrow \infty$. Je možné ukázat, že metoda je konvergentní, právě tehdy, je-li splněno $\|M\| < 1$, kde $\|M\| = \max_{\mathbf{v} \in \mathbb{R}^n} \frac{\|M\mathbf{v}\|_v}{\|\mathbf{v}\|_v}$ je maticová norma indukovaná vektorovou normou.

Při odvozování následujících iteračních metod budeme využívat rozkladu matice A na součet dolní trojúhelníkové, diagonální a horní trojúhelníkové matice, tedy $A = L + D + U$ (pozor, nepleťte si matice L, D, U se stejně nazvanými maticemi, které se vyskytovaly u přímých řešičů). Např. matici

$$A = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix}$$

rozložíme na

$$L = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, D = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}, U = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}.$$

1.1.1 Jacobiho metoda

Vyděme z rovnice $A\mathbf{x} = \mathbf{b}$. Dosazením $A = L + D + U$ dostaneme

$$(L + D + U)\mathbf{x} = \mathbf{b}.$$

Roznásobme a přezávorkujme výraz na levé straně

$$D\mathbf{x} + (L + U)\mathbf{x} = \mathbf{b}$$

Jacobiho metodu odvodíme tak, že přidáme indexy $k + 1$ a k k příslušným vektorům \mathbf{x}

$$D\mathbf{x}^{k+1} + (L + U)\mathbf{x}^k = \mathbf{b}.$$

Osamostatněním \mathbf{x}^{k+1} dostaneme předpis pro $k + 1$ aproximaci vektoru \mathbf{x}

$$\mathbf{x}^{k+1} := D^{-1}(\mathbf{b} - (L + U)\mathbf{x}^k) = \underbrace{-D^{-1}(L + U)}_{=M} \mathbf{x}^k + \underbrace{D^{-1}}_{=N} \mathbf{x}^k.$$

¹Index k označuje číslo aktuální iterace.

Jednotlivé složky vektoru \mathbf{x}^{k+1} můžeme vyjádřit jako

$$(\mathbf{x}^{k+1})_i := \frac{1}{a_{i,i}} \left(b_i - \sum_{j=1, j \neq i}^n a_{i,j}(\mathbf{x}^k)_j \right) = \quad (1.2)$$

$$= \frac{1}{a_{i,i}} \left(b_i - \sum_{j=1}^{i-1} a_{i,j}(\mathbf{x}^k)_j - \sum_{j=i+1}^n a_{i,j}(\mathbf{x}^k)_j \right) \quad (1.3)$$

Pro snadnější a přehlednější zápis budeme i -tý prvek vektoru \mathbf{x}^{k+1} také značit jako x_i^{k+1} (horní index tedy označuje číslo iterace, dolní index značí pořadí prvku ve vektoru).

Konzistence metody vyplývá z jejího odvození, můžeme však ještě ověřit, že $M = I - NA$. V případě Jacobiho metody je $M = -D^{-1}(L + U)$ a $N = D^{-1}$ (viz výše). Platí tedy

$$\begin{aligned} I - NA &= I - D^{-1}A = I - D^{-1}(L + D + U) = \\ &= I - D^{-1}(L + U) - D^{-1}D = -D^{-1}(L + U) = M. \end{aligned}$$

Metoda je tedy konzistentní.

Lze dokázat, že metoda je konvergentní právě tehdy, když $\|M\| = \|D^{-1}(L + U)\| < 1$. To splňují např. striktně diagonálně dominantní matice (tedy matice, pro které platí $\forall i : |a_{i,i}| > \sum_{j=1, j \neq i}^n |a_{i,j}|$). Konvergenci metody pro diagonálně dominantní matice se dá poměrně snadno dokázat. Vyjděme ze vztahu pro i -tý prvek aproximovaného vektoru v iteraci $k + 1$:

$$x_i^{k+1} = \frac{1}{a_{i,i}} \left(b_i - \sum_{j=1, j \neq i}^n a_{i,j} x_j^k \right).$$

Jelikož je metoda konzistentní, musí tuto rovnost splňovat i prvky vektoru přesného řešení \mathbf{x} :

$$x_i = \frac{1}{a_{i,i}} \left(b_i - \sum_{j=1, j \neq i}^n a_{i,j} x_j \right).$$

Odečteme-li od první rovnosti druhou, dostaneme

$$\underbrace{x_i^{k+1} - x_i}_{=e_i^{k+1}} = -\frac{1}{a_{i,i}} \sum_{j=1, j \neq i}^n a_{i,j} \underbrace{(x_j^k - x_j)}_{=e_j^k},$$

kde e^k je vektor chyby v k -tém kroce. Chybu v kroce $k + 1$ můžeme tedy odhadnout pomocí vlastností striktně diagonálně dominantní matice:

$$\begin{aligned} |e_i^{k+1}| &\leq \frac{1}{|a_{i,i}|} \sum_{j=1, j \neq i}^n |a_{i,j}| |e_j^k| \leq \frac{1}{|a_{i,i}|} \sum_{j=1, j \neq i}^n |a_{i,j}| \max_{j=1, \dots, n, j \neq i} |e_j^k| = \\ &= \max_{j=1, \dots, n, j \neq i} |e_j^k| \underbrace{\frac{1}{|a_{i,i}|} \sum_{j=1, j \neq i}^n |a_{i,j}|}_{<1} < \max_{j=1, \dots, n, j \neq i} |e_j^k| \end{aligned}$$

Každý prvek vektoru chyby v kroce $k + 1$ je tedy v absolutní hodnotě menší než maximální prvek vektoru chyby v předchozím kroce. Vektor chyby tedy konverguje k nulovému vektoru.

1.1.2 Gaussova-Seidelova metoda

Všimněme si, že při výpočtu x_i^{k+1} využíváme v sumě $\sum_{j=1}^{i-1} a_{i,j} x_j^k$ ve výrazu (1.3) pouze prvky x_1^k, \dots, x_{i-1}^k . Tyto prvky tedy můžeme nahradit již vypočtenými prvky aktuálními iterace $x_1^{k+1}, \dots, x_{i-1}^{k+1}$. Dostaneme tak předpis Gaussovy-Seidelovy metody:

$$x_i^{k+1} := \frac{1}{a_{i,i}} \left(b_i - \sum_{j=1}^{i-1} a_{i,j} x_j^{k+1} - \sum_{j=i+1}^n a_{i,j} x_j^k \right) \quad (1.4)$$

Podobně jako v předchozím případě můžeme metodu odvodit, nahradíme-li v soustavě $A\mathbf{x} = \mathbf{b}$ matici A součtem $L + D + U$. Tentokrát ovšem vektor s indexem $k + 1$ ponecháme u součtu $L + D$

$$(L + D)\mathbf{x}^{k+1} = \mathbf{b} - U\mathbf{x}^k, \quad (1.5)$$

tedy

$$\mathbf{x}^{k+1} = \underbrace{-(L + D)^{-1}U}_{=M} \mathbf{x}^k + \underbrace{(L + D)^{-1}\mathbf{b}}_{=N}.$$

Vztah mezi maticovým zápisem a zápisem po prvcích (1.4) je nejlépe vidět na rovnosti (1.5). Jedná se o soustavu rovnic s dolní trojúhelníkovou maticí $L + D$, vektorem pravé strany $\mathbf{b} - U\mathbf{x}^k$ a neznámým vektorem \mathbf{x}^{k+1} . Všimněte si, že výraz (1.4) pak přesně odpovídá algoritmu pro dopřednou substituci pro řešení takovéto soustavy.

Podobně jako v případě Jacobiho metody můžeme ověřit, zda je metoda konzistentní porovnáním $I - NA$ a M .

$$\begin{aligned} I - NA &= I - (L + D)^{-1}A = I - (L + D)^{-1}(L + D + U) = \\ &= I - (L + D)^{-1}(L + D) - (L + D)^{-1}U = -(L + D)^{-1}U = M. \end{aligned}$$

Metoda je tedy konzistentní.

Metoda je konvergentní, právě když $\|(L + D)^{-1}U\| < 1$.

1.1.3 Richardsonova metoda

Iterace Richardsonovy metody je dána předpisem

$$\mathbf{x}^{k+1} := \mathbf{x}^k + \omega \mathbf{r}^k,$$

kde $\omega \in \mathbb{R}_+$ a $\mathbf{r}^k = \mathbf{b} - \mathbf{A}\mathbf{x}^k$ je reziduum, které určuje, jak dobře je splněna původní rovnice. Vztah mezi reziduem a chybou $\mathbf{e}^k = \mathbf{x}^k - \mathbf{x}$ lze odvodit přenásobením definice chyby maticí \mathbf{A}

$$\mathbf{A}\mathbf{e}^k = \mathbf{A}\mathbf{x}^k - \mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{x}^k - \mathbf{b} = -\mathbf{r}^k.$$

Studujme konvergenci metody pro symetrickou pozitivně definitní matici \mathbf{A} . V takovém případě vlastní čísla λ_i a vlastní vektory \mathbf{v}_i (tedy skaláry a vektory, pro které platí $\mathbf{A}\mathbf{v}_i = \lambda_i\mathbf{v}_i$, $\|\mathbf{v}_i\| = 1$) splňují

$$0 < \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$$

a

$$\text{span}\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\} = \mathbb{R}^n.$$

Zde span značí lineární obal. vlastní vektory tvoří ortonormální bázi \mathbb{R}^n .

Díky předchozímu poznatku můžeme reziduum vyjádřit jako lineární kombinaci prvků báze tvořené vlastními vektory, tedy $\mathbf{r}^{k+1} = \sum_{i=1}^n \alpha_i^{k+1} \mathbf{v}_i$. Studujme nyní, jak se chová reziduum (a tedy i chyba) v jednotlivých iteracích. Na základě toho se později pokusíme odvodit optimální hodnotu ω pro co nejrychlejší konvergenci.

$$\begin{aligned} \sum_{i=1}^n \alpha_i^{k+1} \mathbf{v}_i &= \mathbf{r}^{k+1} = \mathbf{b} - \mathbf{A}\mathbf{x}^{k+1} = \underbrace{\mathbf{b} - \mathbf{A}(\mathbf{x}^k + \omega \mathbf{r}^k)}_{=\mathbf{r}^k} = \\ &= \mathbf{r}^k - \mathbf{A}\omega \mathbf{r}^k = (1 - \omega \mathbf{A}) \underbrace{\mathbf{r}^k}_{=\sum_{i=1}^n \alpha_i^k \mathbf{v}_i} = (1 - \omega \mathbf{A}) \sum_{i=1}^n \alpha_i^k \mathbf{v}_i = \\ &= \sum_{i=1}^n \alpha_i^k \mathbf{v}_i - \sum_{i=1}^n \alpha_i^k \omega \underbrace{\mathbf{A}\mathbf{v}_i}_{=\lambda_i \mathbf{v}_i} = \sum_{i=1}^n \alpha_i^k \mathbf{v}_i - \sum_{i=1}^n \alpha_i^k \omega \lambda_i \mathbf{v}_i = \\ &= \sum_{i=1}^n (1 - \omega \lambda_i) \alpha_i^k \mathbf{v}_i. \end{aligned}$$

Všimněme si, že jsme vyjádřili koeficienty rozvoje rezidua \mathbf{r}^{k+1} v bázi $\{\mathbf{v}_i\}_{i=1}^n$ pomocí násobků koeficientů v předchozím kroce (viz podtržené části předchozího výrazu). Koeficienty se tedy budou zmenšovat (a jednotlivé složky vektoru rezidua budou konvergovat k nule) právě tehdy, když $|1 - \omega \lambda_i| < 1$ pro všechna $i = 1, 2, \dots, n$. Rychlost konvergence bude záviset na největší hodnotě $|1 - \omega \lambda_i|$. Shrňme tento poznatek do následující věty.

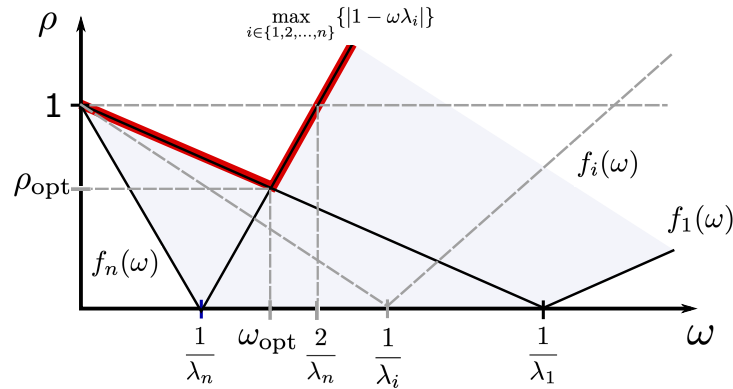


Figure 1.1: Konvergenční faktor Richardsonovy metody v závislosti na ω .

Věta Richardsonova metoda konverguje, právě když $\forall i \in \{1, 2, \dots, n\} : |1 - \omega\lambda_i| < 1$. Konvergenční faktor $\rho = \max_{i \in \{1, 2, \dots, n\}} \{|1 - \omega\lambda_i|\}$ určuje rychlost konvergence: $\|\mathbf{r}^{k+1}\| \leq \rho \|\mathbf{r}^k\|$.

Čím menší bude konvergenční faktor $\rho = \max_{i \in \{1, 2, \dots, n\}} \{|1 - \omega\lambda_i|\}$, tím rychleji bude metoda konvergovat. Vzhledem k tomu, že vlastní čísla matice A jsou daná, můžeme konvergenční faktor ovlivnit pouze vhodnou volbou parametru ω . Odvození ideální hodnoty ω ilustrujeme na Obrázku 1.1. Jsou na něm znázorněny funkce $f_1(\omega) = |1 - \omega\lambda_1|$ a $f_n(\omega) = |1 - \omega\lambda_n|$. Protože směrnice funkcí $f_i(\omega) = |1 - \omega\lambda_i|$ jsou určeny vlastními čísly matice A a ta jsou seřazena od nejmenšího po největší, budou grafy všech funkcí $f_i, i = 2, \dots, n-1$, ležet "mezi" grafy f_1 a f_n (v šedě vyznačené oblasti). Funkci

$$\max_{i \in \{1, 2, \dots, n\}} \{|1 - \omega\lambda_i|\}$$

tedy můžeme vykreslit jako červeně zvýrazněnou lomenou čáru tvořenou částí funkce f_1 a částí funkce f_n . Z grafu této funkce tedy rovnou můžeme odvodit:

1. Interval, ve kterém musí ω ležet. Aby metoda konvergovala, musí platit $\rho = \max_{i \in \{1, 2, \dots, n\}} \{|1 - \omega\lambda_i|\} < 1$. Červená funkce tedy musí ležet pod zakreslenou konstantní funkcí $\rho = 1$. Levý krajní bod intervalu je 0, pravý určíme jako průsečík příslušné části funkce f_n s konstantní funkcí 1:

$$-(1 - \omega\lambda_n) = 1 \quad \Rightarrow \quad \omega = \frac{2}{\lambda_n}.$$

Metoda tedy konverguje pro $\omega \in (0, 2/\lambda_n)$.

2. Optimální ω je bod, ve kterém červeně vyznačená funkce dosahuje minima. Tento bod dostaneme jako průsečík funkcí f_1 a f_n :

$$1 - \omega_{\text{opt}}\lambda_1 = -(1 - \omega_{\text{opt}}\lambda_n) \quad \Rightarrow \quad \omega_{\text{opt}} = \frac{2}{\lambda_1 + \lambda_n}.$$

Zjistili jsme tedy, že nejlepší konvergence dosáhneme, zvolíme-li $\omega_{\text{opt}} = \frac{2}{\lambda_1 + \lambda_n}$. Konvergenční faktor bude v tomto případě

$$\begin{aligned}\rho_{\text{opt}} &= 1 - \omega_{\text{opt}}\lambda_1 = 1 - \frac{2\lambda_1}{\lambda_1 + \lambda_n} = \frac{\lambda_1 + \lambda_n - 2\lambda_1}{\lambda_1 + \lambda_n} = \\ &= \frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1} \frac{\frac{1}{\lambda_1}}{\frac{1}{\lambda_1}} = \frac{\frac{\lambda_n}{\lambda_1} - 1}{\frac{\lambda_n}{\lambda_1} + 1} = \frac{\kappa(\mathbf{A}) - 1}{\kappa(\mathbf{A}) + 1},\end{aligned}$$

kde $\kappa(\mathbf{A}) = \lambda_n/\lambda_1$ je číslo podmíněnosti matice \mathbf{A} .

Můžeme také odvodit, kolik iterací je třeba, abychom dosáhli požadované relativní změny normy rezidua. Hledáme tedy k , pro které platí

$$\frac{\|\mathbf{r}^k\|}{\|\mathbf{r}^0\|} \leq \varepsilon, \quad \text{tedy} \quad \|\mathbf{r}^k\| \leq \varepsilon\|\mathbf{r}^0\|$$

Využijme toho, že $\|\mathbf{r}^k\| \leq \rho_{\text{opt}}^k \|\mathbf{r}^0\|$ a přepišme nerovnici na

$$\rho_{\text{opt}}^k \|\mathbf{r}^0\| \leq \varepsilon \|\mathbf{r}^0\|.$$

Vykrácením normy a zlogaritmováním obou stran nerovnice dostaneme řešení $k \geq \frac{\log \varepsilon}{\log \rho_{\text{opt}}}$ (nezapomeňte, že protože $\rho_{\text{opt}} \in (0, 1)$, je třeba otočit znaménko nerovnosti).

1.2 Gradientní iterační metody

Věta Řešení soustavy $\mathbf{Ax} = \mathbf{b}$ se symetrickou pozitivně definitní maticí \mathbf{A} je ekvivalentní s minimalizací kvadratické formy

$$f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T \mathbf{A} \mathbf{x} - \mathbf{b}^T \mathbf{x}.$$

Důkaz Dokažme nejdříve implikaci $\mathbf{Ax} = \mathbf{b} \Rightarrow \mathbf{x} = \arg \min_{\mathbf{v} \in \mathbb{R}^n} f(\mathbf{v})$. Podíváme se, jak se změní funkční hodnota f , posuneme-li se z bodu \mathbf{x} o nějaký nenulový vektor \mathbf{c} :

$$\begin{aligned}f(\mathbf{p}) &= f(\mathbf{x} + \mathbf{c}) = \frac{1}{2} (\mathbf{x} + \mathbf{c})^T \mathbf{A} (\mathbf{x} + \mathbf{c}) - \mathbf{b}^T (\mathbf{x} + \mathbf{c}) = \\ &= \frac{1}{2} \mathbf{x}^T \mathbf{A} \mathbf{x} + \mathbf{c}^T \underbrace{\mathbf{A} \mathbf{x}}_{=\mathbf{b}} + \frac{1}{2} \mathbf{c}^T \mathbf{A} \mathbf{c} - \mathbf{b}^T \mathbf{x} - \mathbf{b}^T \mathbf{c} = \\ &= \underbrace{\frac{1}{2} \mathbf{x}^T \mathbf{A} \mathbf{x} - \mathbf{b}^T \mathbf{x}}_{=f(\mathbf{x})} + \underbrace{\mathbf{c}^T \mathbf{b} - \mathbf{b}^T \mathbf{c}}_{=0} + \frac{1}{2} \mathbf{c}^T \mathbf{A} \mathbf{c} = f(\mathbf{x}) + \underbrace{\frac{1}{2} \mathbf{c}^T \mathbf{A} \mathbf{c}}_{>0}.\end{aligned}$$

Díky pozitivní definitnosti \mathbf{A} je výraz $\mathbf{c}^T \mathbf{A} \mathbf{c}$ kladný. Posuneme-li se tedy z bodu \mathbf{x} v libovolném směru, hodnota funkce f se zvětší. V bodě \mathbf{x} tedy nastává minimum.

K důkazu opačné implikace $\mathbf{x} = \arg \min_{\mathbf{v} \in \mathbb{R}^n} f(\mathbf{v}) \Rightarrow \mathbf{Ax} = \mathbf{b}$ je třeba si uvědomit nutnou podmínku minima funkce $f : \mathbb{R}^n \rightarrow \mathbb{R}$, tedy nulovost gradientu:

$$\mathbf{x} = \arg \min_{\mathbf{v} \in \mathbb{R}^n} f(\mathbf{v}) \Rightarrow \nabla f(\mathbf{x}) = \mathbf{o}. \quad (1.6)$$

Lze ukázat, že pro gradient funkce f platí

$$\nabla f(\mathbf{x}) = \left[\frac{\partial f}{\partial x_1}(\mathbf{x}), \dots, \frac{\partial f}{\partial x_n}(\mathbf{x}) \right]^T = \frac{1}{2} \mathbf{A}^T \mathbf{x} + \frac{1}{2} \mathbf{Ax} - \mathbf{b} = \mathbf{Ax} - \mathbf{b}.$$

Z podmínky (1.6) tedy vyplývá $\mathbf{Ax} - \mathbf{b} = \mathbf{o}$. \square

References

- [1] Trefethen, L. N, Bau, D. Numerical Linear Algebra. SIAM. 1997.
- [2] Schewchuk, J. R. An Introduction to the Conjugate Gradient Method Without the Agonizing Pain. 1994. Dostupné z <https://www.cs.cmu.edu/~quake-papers/painless-conjugate-gradient.pdf>
- [3] Lukáš, D. Zápisky z přednasek. Dostupné z <https://home1.vsb.cz/~luk76>