

Numerické metody v \mathbb{R}^n

```
$$ \xdef\scal#1#2{\langle #1, #2 \rangle} \xdef\norm#1{\left\| \right. #1 \right\|} \xdef\dist{\rho} \xdef\and{\&} \xdef\brackets#1{\left\{ #1 \right\}} \xdef\parc#1#2{\frac {\partial #1}{\partial #2}} \xdef\mtr#1{\begin{pmatrix} #1 \end{pmatrix}} \xdef\bm#1{\boldsymbol{#1}} \xdef\mc#1{\mathcal{#1}} \xdef\vv#1{\mathbf{#1}} \xdef\vvp#1{\pmb{#1}} \xdef\ve{\varepsilon} \xdef\l{\lambda} \xdef\th{\vartheta} \xdef\alpha{\alpha} \xdef\vf{\varphi} \xdef\Tagged#1{\text{#1}} \xdef\tagged*#1{\text{#1}} \xdef\tagEqHere#1#2{\href{#2#eq-#1}{\text{#1}}} \xdef\tagDeHere#1#2{\href{#2#de-#1}{\text{#1}}} \xdef\tagEq#1{\href{#eq-#1}{\text{#1}}} \xdef\tagDe#1{\href{#de-#1}{\text{#1}}} \xdef\T#1{\htmlid{eq-#1}{#1}} \xdef\D#1{\htmlid{de-#1}{\vv{#1}}} \xdef\conv#1{\mathrm{conv}}, #1 \xdef\cone#1{\mathrm{cone}}, #1 \xdef\aff#1{\mathrm{aff}}, #1 \xdef\lin#1{\mathrm{Lin}}, #1 \xdef\span#1{\mathrm{span}}, #1 \xdef\O{\mathcal O} \xdef\ri#1{\mathrm{ri}}, #1 \xdef\rd#1{\mathrm{r}\partial}, #1 \xdef\interior#1{\mathrm{int}}, #1 \xdef\proj{\Pi} \xdef\epi#1{\mathrm{epi}}, #1 \xdef\grad#1{\mathrm{grad}}, #1 \xdef\gradT#1{\mathrm{grad}}^T #1 \xdef\gradx#1{\mathrm{grad}}_x #1 \xdef\hess#1{\nabla^2}, #1 \xdef\hessx#1{\nabla^2_x #1} \xdef\jacobx#1{D_x #1} \xdef\jacob#1{D #1} \xdef\subdif#1{\partial #1} \xdef\co#1{\mathrm{co}}, #1 \xdef\iter#1{\wedge^{#1}} \xdef\str{\wedge^*} \xdef\spv{\mathcal V} \xdef\civ{\mathcal U} \xdef\knvxProg{\tagEqHere{4.1}{./nutne-a-postacujici-podminky-optimality} \, \, \text{and} \, \, \tagEqHere{4.2}{nutne-a-postacujici-podminky-optimality}} $$
```

Budeme se věnovat úlohám (přesněji numerickým metodám jejich řešení) typu $\min_{x \in \mathbb{R}^n} f(x)$ kde $f: \mathbb{R}^n \rightarrow \mathbb{R}$ je (jednou/dvakrát/třikrát) **spojitě diferencovatelná** funkce. Obecně jsou metody numerické optimalizace založeny na **minimalizační posloupnosti** $\{x^k\}$ definované jako $x^{k+1} = x^k + \alpha_k h_k$, kde $\alpha_k \in \mathbb{R}$ se nazývá **délka k -tého kroku** a vektor $h_k \in \mathbb{R}^n$ je **směr k -tého vektoru**.

“ Budeme uvažovat tzv. **přesnou minimalizaci**, kdy dílčí minimalizace řešíme přesně (nikoliv numerickými metodami)

“ Všechny následující metody jsou **spádové**

Metoda největšího spádu

U této metody volíme $h_k = -\frac{1}{\|\text{grad } f(x^{[k]})\|}$, přičemž délku kroku volíme **přesným řešením** úlohy $f(x^{[k+1]}) = f(x^{[k]} - \alpha_k \text{grad } f(x^{[k]})) = \min_{\alpha \geq 0} f(x^{[k]} - \alpha \cdot \text{grad } f(x^{[k]}))$. Dále bude platit, že vektory určené body $x^{[k+1]}$, $x^{[k]}$ a $x^{[k+2]}$, $x^{[k+1]}$ jsou na sebe **ortogonální**. Z tohoto dostáváme, že pro daný směr hledáme **nejblížší** vrstevnici, která bude **tečná** k tomuto vektoru.

Tato metoda je **prvního řádu** (stačí nám pouze gradient).

“ V některých případech dochází k tzv. "cik-cak" efektu (klikatění), kdy se minimalizující posloupnost dostává k optimu velmi pomalu. Toto se děje například u **Rosenbrockovy (banánové) funkce** (jedna z testovacích funkcí)

Kvadratické funkce

Nechť f je kvadratická funkce tvaru $f(x) = \frac{1}{2} x^T Q x - x^T b$, kde $Q = Q^T > 0$ je **symetrická** $n \times n$ matice a $b \in \mathbb{R}^n$. Taková funkce f je **ostře** (i *silně*) konvexní. Z pozitivní definitnosti Q dostáváme, že vlastní čísla matice Q jsou **kladné** a můžeme je uspořádat následovně $0 < \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$. Díky tomu můžeme úlohu $\tag{3.2.1}$ vyřešit "přímo" jako $x^* = Q^{-1} b$, nicméně počítání inverze může být **velice náročné** (výpočetně).

V tomto případě je gradient g funkce f dán jako $g(x) := \text{grad } f(x) = Qx - b$, tedy v jednotlivých iteracích dostáváme $g_k := Qx^{[k]} - b$ a α_k můžeme určit jako $\alpha_k = \frac{g_k^T g_k}{g_k^T Q g_k} \in [1/\lambda_n, 1/\lambda_1]$

Nyní se zaměříme na konvergenci metody, kterou můžeme zkoumat pomocí $E(x) := f(x) - f(x^*) = \frac{1}{2} (x - x^*)^T Q (x - x^*)$

Lemma 3.2.1 (Konvergence metody největšího spádu)

Platí $E(x^{[k+1]}) = \left(1 - \frac{(g_k^T g_k)^2}{(g_k^T Q g_k)(g_k^T Q^{-1} g_k)}\right) E(x^{[k]})$

Z Lemmatu 3.1.2 okamžitě plyne, že pokud pro nějaké $k \in \mathbb{N}$ nastane $1 = \frac{(g_k^T g_k)^2}{(g_k^T Q g_k)(g_k^T Q^{-1} g_k)}$, $\tag{3.2.1-a}$ tak v $k+1$ metoda největšího spádu nalezne řešení **přesně**. V opačném případě je metoda **nekonečně-kroková**.

Rovnost $\tag{3.2.1-a}$ nastane v případě, že g_k je **vlastním vektorem** matice Q , jinak řečeno gradient musí mířit **do středu** elipsy (*elipsoidu*).

V případě, že $\lambda_1 = \dots = \lambda_n$ je konvergence **superlineární**. Naopak pokud $\lambda_1 = \dots = \lambda_{n-1} \neq \lambda_n$, tak může být konvergence **velice pomalá**. Ve skutečnosti ještě rychlost

konvergence závisí na počátečním $x^{[0]}$

Nekvadratické funkce

V případě nekvadratické funkce je metoda největšího spádu schopna nalézt **pouze stacionární body**.

Lemma 3.2.2iii (Lokální konvergence)

Nechť $f: \mathbb{R}^n \rightarrow \mathbb{R}$ je **spojitě diferencovatelná**. Jestliže bod $x^{[0]} \in \mathbb{R}^n$ je takový, že množina $\{x \in \mathbb{R}^n \mid f(x) \leq f(x^{[0]})\}$ je **ohraničená**, pak posloupnost $\{x^{[k]}\}$ generovaná metodou největšího spádu konverguje k bodu x^* , kde $\nabla f(x^*) = 0$.

Pokud konverguje $\{x^{[k]}\}$ k bodu x^* a funkce f je **dvakrát** spojitě diferencovatelná **na okolí** x^* a platí $\alpha I \preceq \nabla^2 f(x^*) \preceq A I$, kde $\alpha, A > 0$ (tedy f je v okolí x^* **silně konvexní**), pak metoda konverguje (**alespoň**) s rychlostí $\left(\frac{A - \alpha}{A + \alpha}\right)^2$

“ Tedy i pro nekvadratické funkce hraje velkou roli podmíněnost matice $\nabla^2 f(x^*)$

“ Metoda největšího spádu se nejčastěji využívá v jiných metodách jako pomocné, když ony metody samotné v tu chvíli neposkytnou dostatečné zlepšení

Celkem můžeme *metodu největšího spádu* shrnout:

- **globální** konvergence (pro nekvadratické metody za dalších předpokladů)
- **pomalá** konvergence
 - mnohdy numericky ani nekonverguje
- je základem pro další (lepší) metody

Newtonova metoda

Hlavní myšlenkou Newtonovy metody je, že v $(k+1)$ -kroku, kde $k \in \mathbb{N}_0$, funkci f aproximujeme **Taylorovým polynomem druhého řádu** se středem v bodě $x^{[k]}$ a jako $x^{[k+1]}$ volíme bod, ve kterém tento polynom nabývá svého minima.

Jinak řečeno místo **tečné nadroviny** k funkci konstruuujeme **tečnou n -rozměrnou parabolou**

Tedy místo funkce f uvažujeme v k -tém kroku $T_k(x) := f(x^{[k]}) + \nabla f(x^{[k]})^T (x - x^{[k]}) + \frac{1}{2} (x - x^{[k]})^T \nabla^2 f(x^{[k]}) (x - x^{[k]}) \approx f(x)$

Jelikož hledáme řešení $T_k(x) \rightarrow \min$, tak výraz výše první zderivujeme, z čehož dostaneme $\nabla T_k(x) = \nabla f(x^{[k]}) + \nabla^2 f(x^{[k]}) (x - x^{[k]})$, což v případě **regulární** matice $\nabla^2 f(x^{[k]})$ vede na $x^{[k+1]} = x^{[k]} - (\nabla^2 f(x^{[k]}))^{-1} \nabla f(x^{[k]})$

“ Pro $\nabla^2 f(x) > 0$ je funkce **konvexní** a nalezneme **minimum**

Nicméně je důležité podotknout, že výpočet $(\nabla^2 f(x^{[k]}))^{-1}$ je **velmi výpočetně náročný**. Avšak v případě **kvadratické funkce** Newtonova metoda nalezne řešení **v jednom kroku**, tedy její rychlost konvergence je **superlineární** s řádem ∞ .

Regularita matice $\nabla^2 f(x^{[k]})$ je velmi důležitá pro konvergenci, viz následující věta.

Věta $D\{3.2.5\}$

Nechť $f \in C^3$ v okolí bodu $x^* \in \mathbb{R}^n$, který je **nedegenerovaným minimem**, tj. $\nabla f(x^*) = 0$ a $\nabla^2 f(x^*) > 0$. Potom pro $x^{[0]} \in \mathbb{R}^n$ dostatečně blízko x^* konverguje $\{x^{[k]}\}$ generovaná Newtonovou metodou k bodu x^* **superlineárně** s řádem (alespoň) $p = 2$ (tj. *kvadraticky*).

“ Zde určit, co znamená "dostatečně blízko x^* " je obtížné

Celkem můžeme *Newtonovu metodu* shrnout jako:

- **velmi rychlá** konvergence
- nutnost **dostatečně blízké počáteční aproximace**
- velmi velká výpočetní náročnost při velkém n (počtu dimenzí)

Metoda sdružených gradientů

Uvažujme situaci v úloze $\text{tagEq}\{3.2.1\}$, kdy máme funkci $f(x) = \frac{1}{2} x^T Q x - b^T x$ $\text{tag}\{T\{MSG\}\}$, kde $Q = Q^T \in \mathbb{R}^{n \times n}$, $Q > 0$ je symetrická matice a $b \in \mathbb{R}^n$.

Pak nalezení úlohy $\text{tagEq}\{3.2.1\}$ a $\text{tagEq}\{MSG\}$ je ekvivalentní s řešením úlohy $Qx = b$ $\text{tag}\{T\{MSGa\}\}$, kterou umíme řešit například Gaussovou eliminací.

Metoda sdružených gradientů je v případě $Q > 0$ přímou metodou, která dojde k řešení $\text{tagEq}\{\text{MSGa}\}$ po n krocích. Nicméně tento fakt lze brát také jako, že je to iterační metoda s velmi rychlou konvergencí v případě pozitivně definitní matice.

Definice $\text{D}\{3.2.7\}$ (Q -sdružené vektory)

Nechť $Q = Q^T \in \mathbb{R}^{n \times n}$ je **pozitivně definitní**. Vektory $h_1, h_2 \in \mathbb{R}^n \setminus \{0\}$ se nazývají **Q -sdružené** (nebo také **Q -ortogonální**), jestliže $\text{scal}\{Qh_1\} \{h_2\} = h_1^T Q h_2 = 0$. Systém vektorů $\{h_0, \dots, h_{m-1}\} \in \mathbb{R}^n \setminus \{0\}$ pro $m \in \{2, \dots, n\}$ se nazývá **Q -sdružený**, jestliže $\text{scal}\{Q h_i\} \{h_j\} = 0$ pro $i \neq j$.

Věta $\text{D}\{3.2.8\}$

Nechť systém vektorů $\{h_0, \dots, h_{m-1}\} \in \mathbb{R}^n \setminus \{0\}$ s $m \in \{2, \dots, n\}$ je **Q -sdružený**. Potom jsou tyto vektory **lineárně nezávislé**.

Věta $\text{D}\{3.2.9\}$

Nechť $m \in \{2, \dots, n\}$ a mějme systém **Q -sdružených** vektorů $\{h_0, \dots, h_{m-1}\} \in \mathbb{R}^n$. Nechť $x_{\text{iter } 0}$ je **dáno** a body $x_{\text{iter } 1}, \dots, x_{\text{iter } m}$ jsou dány jako $x_{\text{iter } k+1} = x_{\text{iter } k} + \alpha_k h_k = x_{\text{iter } 0} + \sum_{i=0}^k \alpha_i h_i$, $\forall k \in \{0, \dots, m-1\}$, $\text{tag}\{\text{T}\{3.2.8\}\}$ kde α_k jsou volena tak, že $f(x_{\text{iter } k} + \alpha_k h_k) = \min_{\alpha \in \mathbb{R}} f(x_{\text{iter } k} + \alpha h_k)$ pro $k \in \{0, \dots, m-1\}$ (tj. jsou volena **přesnou minimalizací**). Pak pro kvadratickou funkci f definovanou v $\text{tagEq}\{\text{MSG}\}$ platí $f(x_{\text{iter } m}) = \min_{x \in X_m} f(x)$, kde $X_m := \text{lin}\{h_0, \dots, h_{m-1}\}$ (viz **Definice** $\text{tagDeHere}\{2.1.6\}$./konvexni-mnoziny) - lineární obal). Zejména pro $m = n$ dostáváme $f(x_{\text{iter } n}) = \min_{x \in \mathbb{R}^n} f(x)$, tj. $x_{\text{iter } n}$ je řešením úlohy $\text{tagEq}\{3.2.1\}$ a $\text{tagEq}\{\text{MSG}\}$.

“ Nalezení **Q -sdružených** vektorů lze provést zobecněným **Gram-Schmidtovým ortogonalizačním procesem** (ten je uveden v Lin. Alg. ve speciálním tvaru pro $Q = I$).

Explicitně můžeme odvodit délku α_k -tého kroku jako $\alpha_k = - \frac{h_k^T \text{grad } f(x_{\text{iter } k})}{h_k^T Q h_k}$ $\text{tag}\{\text{T}\{3.2.9\}\}$

Celkem můžeme metodu popsat následovně $h_0 := -\text{grad } f(x_{\text{iter } 0})$, $\forall k \quad h_k := -\text{grad } f(x_{\text{iter } k}) + \beta_{k-1} h_{k-1}$ $\text{tag}\{3.2.10\}$ $\beta_{k-1} := \frac{h_{k-1}^T \text{grad } f(x_{\text{iter } k})}{h_{k-1}^T Q h_{k-1}}$ $\text{tag}\{3.2.11\}$, přičemž body minimalizující posloupnosti jsou počítány podle **Věty** $\text{tagDe}\{3.2.9\}$.

“ Tento výpočet lze "zjednodušit", viz $\text{Tagged}\{3.2.12\}$ v přednášce.

Hlavní výhodou **metody sdružených gradientů** je její **snadná implementace**, naopak nevýhodou citlivost na podmíněnost matice Q . Také se daří říct, že metoda sdružených gradientů **konverguje nejrychleji** z metod založených pouze na *maticovém násobení*.

Pro nekvadratické funkce používáme stejný algoritmus jako doted' až na volbu β_k , ale metodu restartujeme po n krocích

Věta 3.2.13 (Rychlost konvergence)

Nechť $f \in C^3$ na \mathbb{R}^n , $x^0 \in \mathbb{R}^n$ a x^* je **nedegenerované lokální minimum**, tj. $\nabla f(x^*) = 0$ a $\nabla^2 f(x^*) > 0$. Nechť x^k je výsledek **metody sdružených gradientů** s cyklem délky n a výchozím bodem x^{k-1} a nechť $x^k \rightarrow x^*$ pro $k \rightarrow \infty$. Potom **minimalizující posloupnost** $\{x^k\}$ konverguje **superlineárně** s řádem **alespoň** $p = 2$.

“ MSG souvisí s metodami *Krylovových podprostorů*.

Revision #13

Created 1 January 2023 14:43:01 by Sceptri

Updated 5 January 2023 12:18:01 by Sceptri