M5170: Matematické programování

Petr Zemánek (Masarykova Univerzita, Brno)

Kapitola 3: Numerické metody řešení úloh matematického programování I

(verze: 20. prosince 2024)

Nyní se již konečně dostáváme k úlohám

$$f(x) \to \min, \quad x \in X. \tag{3.1}$$

Kapitola 4: teoretický pohled $\rightsquigarrow$ nutné a postačující podmínky pro řešení (3.1).

Kapitola 3: numerické metody (=počátek MP)

Základní rozdělení úlohy (3.1):

(i) $X = [a, b]$, tj. jednorozměrná minimalizace,

(ii) $X = \mathbb{R}^n$, tj. nepodmíněná n-rozměrná optimalizace,

(iii) $X \subsetneq \mathbb{R}^n$, tj. podmíněná n-rozměrná optimalizace.

Cíl: vybudování některých (3) základních metod pro kategorii (ii) $\rightsquigarrow$ Proč skupina (i)? A skupina (iii)?

Klasické metody pro (ii) a (iii) jsou (obvykle) založeny na následujícím principu:

> při daném bodě $x_0 \in X$ (poč. aproximace) chceme zkonstruovat takovou posloupnost $\{x_k\}_{k=0}^{\infty}$, že
>
> $$\lim_{k\to\infty} f(x_k) = f^* = \inf_{x\in X} f(x)$$

Takovou posloupnost $\{x_k\}_{k=0}^{\infty}$ budeme nazývat minimalizující (ne nutně musí být *nekonečná*) $\rightsquigarrow$ algoritmy (zobrazení $x \in X \to$ podmnožina X) jsou iterativní procesy.

Základní otázkou je potom globální/lokální konvergence, tj. závislost konvergence MinPsl na volbě počáteční aproximace x_0.

Další důležitou charakteristikou iteračních procesů je *rychlost* jejich konvergence k limitní hodnotě. Máme-li minimalizující posloupnost $\{x_k\}_{k=0}^{\infty}$ splňující $x_k \in \mathbb{R}^n$ a současně $x_k \to x^*$, pak se rychlost konvergence měří vzhledem k posloupnosti odchylek (chyb) $\{e_k\}_{k=0}^{\infty}$, pro kterou $e_k \in [0,\infty)$ a $e_k \to 0$.

Obvykle se volí

$$e_k := \|x_k - x^*\|, \quad \text{kde } x^* \text{ je limitou } \{x_k\}, \text{ což nemusí být řešením (3.1),}$$
$$e_k := |f(x_k) - f(x^*)|, \quad \text{kde } x^* \text{ je opět limitou } \{x_k\},$$
$$e_k := \ell_k, \quad \text{kde } \ell_k \text{ je délka tzv. } \textit{intervalu lokalizace minima}, \text{ viz později.}$$

Posloupnost $\{e_k\}$ je pak srovnávána s geometrickou posloupností $\{h_k\}$ se členy

$$h_k := q\beta^k,$$

kde $q > 0$ a $\beta \in (0,1)$, a případně také s posloupností $\{h_k\}$ se členy

$$h_k := q\beta^{p^k},$$

kde $q > 0$, $\beta \in (0,1)$ a $p > 1$. Proč právě tyto posloupnosti?

Definice 3.1

Nechť jsou dány dvě posloupnosti $\{e_k\}_{k=0}^{\infty}$ a $\{h_k\}_{k=0}^{\infty}$ takové, že

$$e_k \in [0,\infty), \quad e_k \to 0 \quad \& \quad h_k \in [0,\infty), \quad h_k \to 0.$$

Řekneme, že posloupnost $\{e_k\}$ konverguje rychleji (pomaleji) než $\{h_k\}$, pokud existuje index $\tilde{k} \in \mathbb{N}_0 := \mathbb{N} \cup \{0\}$ takový, že

$$e_k \lessgtr h_k \quad \text{pro všechna } k \in [\tilde{k},\infty) \cap \mathbb{N}_0$$

Definice 3.2

Nechť je dána posloupnost $\{e_k\}_{k=0}^{\infty}$ splňující $e_k \in [0,\infty)$ a $e_k \to 0$. Řekneme, že posloupnost $\{e_k\}$ konverguje

(i) alespoň lineárně s rychlostí $\beta \in (0,1)$, pokud konverguje rychleji než geometrická posloupnost se členy tvaru $q\bar{\beta}^k$, kde $q > 0$ a $\bar{\beta} \in (\beta, 1)$;

(ii) nejvýše lineárně s rychlostí $\beta \in (0,1)$, pokud konverguje pomaleji než geometrická posloupnost se členy $q\bar{\beta}^k$, kde $q > 0$ a $\bar{\beta} \in (0, \beta)$;

(iii) lineárně s rychlostí $\beta \in (0,1)$, pokud konverguje nejvýše a současně alespoň lineárně s rychlostí β;

(iv) superlineárně (sublineárně), pokud konverguje rychleji (pomaleji) než libovolná geometrická posloupnost se členy tvaru $q\beta^k$, kde $q > 0$ a $\beta \in (0,1)$.

$\beta > 1$? *Lineární* vs. *geometrická* konvergence. Rychlost vs. β?

Příklady

(i) Posloupnosti konvergující lineárně s rychlostí β:

$$q\beta^k, \quad q(\beta + 1/k)^k, \quad q(\beta - 1/k)^k, \quad q\beta^{k+1/k},$$

kde $q > 0$ a $\beta \in (0, 1)$.

(ii) Nechť $0 < \beta_1 < \beta_2 < 1$ a uvažme posloupnost $\{e_k\}_{k=0}^{\infty}$ se členy

$$e_k = \begin{cases} \beta_1^\ell \, \beta_2^\ell, & k = 2\ell, \\ \beta_1^{\ell+1} \, \beta_2^\ell, & k = 2\ell + 1. \end{cases}$$

Pak tato posloupnost konverguje alespoň lineárně s rychlostí β_2 a nejvýše lineárně s rychlostí β_1. Dokonce se dá ukázat, že konverguje lineárně s rychlostí $\sqrt{\beta_1 \, \beta_2}$.

(iii) Posloupnost $\{e_k\}_{k=0}^{\infty}$ se členy ve tvaru

$$e_k = \begin{cases} 1/2^\ell, & k = 2\ell, \\ 1/(2^\ell + 1), & k = 2\ell + 1. \end{cases}$$

konverguje lineárně s rychlostí $1/\sqrt{2}$.

(iv) Posloupnost $\{1/k\}$ konverguje sublineárně a každá posloupnost se členy tvaru $q\beta^{p^k}$ konverguje superlineárně pro libovolné $q > 0$, $\beta \in (0, 1)$ a $p > 1$. Také např. $e_k = 1/k^k$ konverguje superlineárně.

Uvidíme, že většina algoritmů bude konvergovat lineárně a některé dokonce superlineárně. Není-li β *příliš* blízko 1, lze považovat lineární konvergenci za dostatečnou. Algoritmy se sublineární rychlostí konvergence nejsou obvykle uvažovány (a tak to učiníme i my), neboť nejsou z praktického hlediska příliš efektivní – konvergují příliš pomalu, např. $e_k = 1/(\ln(\ln k))$. Na druhou stranu je užitečné superlineární konvergenci klasifikovat podrobněji.

Definice 3.4

Nechť daná posloupnost $\{e_k\}_{k=0}^{\infty}$ splňuje $e_k \in [0,\infty)$ a $e_k \to 0$, přičemž konvergence je superlineární. Potom $\{e_k\}$ konverguje

(i) alespoň superlineárně s řádem $p > 1$, pokud konverguje rychleji než všechny posloupnosti se členy tvaru $q\beta^{\bar{p}^k}$, kde $q > 0$, $\beta \in (0,1)$ a $\bar{p} \in (1,p)$;

(ii) nejvýše superlineárně s řádem $p > 1$, pokud konverguje pomaleji než všechny posloupnosti se členy tvaru $q\beta^{\bar{p}^k}$, kde $q > 0$, $\beta \in (0,1)$ a $\bar{p} \in (p,\infty)$;

(iii) superlineárně s řádem $p > 1$, pokud konverguje nejvýše a současně alespoň superlineárně s řádem p;

(iv) superlineárně s řádem $p = 1$, pokud konverguje pomaleji než všechny posloupnosti se členy tvaru $q\beta^{\bar{p}^k}$, kde $q > 0$, $\beta \in (0,1)$ a $\bar{p} \in (1,\infty)$.

✎ Srovnání různých hodnot p.

Slůvko *superlineární* se obvykle vynechává a hovoří se pouze o *řádu konvergence*. Zejména *kvadratická konvergence* = superlineární konvergence s řádem $p = 2$.

Návod pro učení rychlosti/řádu konvergence: vypočteme

$$\gamma := \limsup_{k\to\infty} \frac{e_{k+1}}{e_k}$$

- je-li $\gamma \in (0,1)$ $\rightsquigarrow$ lineární konvergence $\rightsquigarrow$ rychlost
- je-li $\gamma = 1$ $\rightsquigarrow$ sublineární konvergence
- je-li $\gamma = 0$ $\rightsquigarrow$ superlineární konvergence $\rightsquigarrow$ řád

Je-li $e_k > 0$ pro všechna $k \in \mathbb{N}_0$, pak řád konvergence můžeme získat jako

$$p = \sup\left\{q \geqslant 1 \mid \limsup_{k\to\infty} e_{k+1}/e_k^q < \infty\right\}.$$

Je-li $\limsup_{k\to\infty} e_{k+1}/e_k^q = 0$ pro nějaké $q \geqslant 1$, pak existuje právě jeden exponent p takový, že

$$\limsup_{k\to\infty} \frac{e_{k+1}}{e_k^q} = \begin{cases} 0, & q < p, \\ \alpha, & q = p, \\ \infty, & q > p, \end{cases}$$

přičemž hodnoty $\alpha \in \{0,\infty\}$ nejsou vyloučeny.

Jinými slovy, najdeme-li $p \geqslant 1$ splňující

$$0 < \limsup_{k\to\infty} \frac{e_{k+1}}{e_k^p} = \alpha < \infty,$$

pak p je řád superlineární konvergence. Obzvláště v případě existence samotné limity platí

$$e_{k+1} \approx \alpha e_k^p,$$

tj. $e_{\ell+1} = \alpha e_\ell^p$, $e_{\ell+2} = \alpha^{p+1} e_\ell^{p^2}$, $e_{\ell+3} = \alpha^{p(p+1)+1} e_\ell^{p^3}$, ... pro dostatečně velké ℓ. Číslo α je tzv. asymptotická odchylka. Navíc v takovém případě

$$p = \lim_{k\to\infty} \frac{\ln e_{k+1}}{\ln e_k}. \qquad (\star)$$

Analogické úvahy lze provést také pro e_k^{1/p^k}. Zejména existuje-li $p \geqslant 1$ takové, že

$$\lim_{k\to\infty} e_k^{1/q^k} = \begin{cases} 0, & q < p, \\ \alpha, & q = p, \\ 1, & q > p, \end{cases}$$

pak p je řád superlineární konvergence, přičemž hodnoty $\alpha \in \{0, 1\}$ jsou povoleny. Jinými slovy,

$$p = \sup\{q \geqslant 1 \mid \lim_{k\to\infty} e_k^{1/q^k} < 1\}.$$

Pak

$$e_k \approx \alpha^{p^k} \quad \text{pro dostatečně velká } k.$$

Splývají tyto přístupy? Obecně nikoli. Uvažte např. posloupnost $\{e_k\}$ se členy

$$e_k = \begin{cases} (a/2)^{p^k}, & k \text{ sudé}, \\ (a^{q/p})^{p^k}, & k \text{ liché}, \end{cases}$$

kde $0 < a < 1$ a $1 < q < p$. Potom „odmocninový“ přístup dává řád konvergence p, zatímco „podílový“ přístup dává, že řád je nejvýše roven q. Nicméně v případě existence $0 < \lim_{k\to\infty} e_{k+1}/e_k^p < \infty$ pro nějaké $p \geqslant 1$ se obě hodnoty musí shodovat, a to nám *stačí*. Podobně k $(\star)$ v takovém případě platí

$$p = \lim_{k\to\infty} \sqrt[k]{|\ln e_k|}.$$

Příklad

Uvažme např. posloupnost $\{e_k\}$ se členy

$$e_k = (1/k)^k.$$

Pak platí

$$p = \lim_{k\to\infty} \frac{\ln(1/(k+1))^{k+1}}{\ln(1/k)^k} = \cdots = 1.$$

Je to skutečně řád? Ano, neboť

$$\lim_{k\to\infty} \frac{e_{k+1}}{e_k^p} = \cdots = \begin{cases} 0, & p < 1, \\ 0, & p = 1, \\ \infty, & p > 1. \end{cases}$$

Posloupnost tedy konverguje superlineárně s řádem $p = 1$.

Příklady

Uvažte sami posloupnost $\{e_k\}$ se členy

$$e_k = a^k$$

pro $0 < a < 1$ ($\rightsquigarrow p = 1$) a

$$e_k = A\, a^{p^k}$$

pro $0 < a < 1$ a $A > 0$ ($\rightsquigarrow$ řád p a asymptotická odchylka A^{1-p}).

Algoritmy lze kategorizovat několika způsoby:

(i) dle řádu konvergence minimalizující posloupnosti, kterou tímto způsobem získáme (nebo posloupnosti funkčních hodnot $\{f(x_k)\}$). Nicméně neměli bychom přecenit význam řádu konvergence a odpovídajícího porovnávání jednotlivých metod. Tento tradiční přístup totiž dává pouze jakýsi návod k orientaci mezi metodami — *nic víc*. Navíc se zdá, že ani neexistuje žádný čistě teoretický způsob k podrobné klasifikaci jednotlivých metod – vše závisí na konkrétních okolnostech a empirických zkušenostech. V praxi hraje velmi důležitou roli také např. *zaokrouhlování*, ale i některé další aspekty, které mohou ovlivnit užitečnost metody (zejména počet potřebných početních operací a jejich náročnost, např. potřeba výpočtu inverzní matice).

(ii) dle požadavků na funkci f, tj.
- metoda nultého řádu $\Rightarrow f$
- metoda prvního řádu $\Rightarrow f \,\&\, f'$
- metoda druhé řádu $\Rightarrow f \,\&\, f' \,\&\, f''$

Metody druhého řádu mají obvykle nejrychlejší konvergenci, ovšem nejsou vhodné např. ve chvíli, kdy výpočet f'' je složitý nebo zatížený numerickou chybou. V případě experimentálního určování funkčních hodnot je asi nejvýhodnější použití metod nultého řádu.

(iii) dle volby bodů $x_1, x_2, \ldots$, tj.
- závisí-li volba x_{k+1} na $x_0, \ldots, x_k \Rightarrow$ aktivní algoritmus
- nezávisí-li volba x_{k+1} na $x_0, \ldots, x_k \Rightarrow$ pasivní algoritmus

Dalším důležitým aspektem těchto algoritmů je jejich zastavení. V tomto ohledu můžeme najít (a vymyslet) širokou škálu pravidel, zejména

$$\|x_{k+1} - x_k\| < \varepsilon, \quad |f(x_{k+1}) - f(x_k)| < \varepsilon, \quad \|\operatorname{grad} f(x_k)\| < \varepsilon$$

(nebo jejich kombinace) apod. (vč. relativních změn – viz později).

V této části budeme řešit úlohu

$$f(x) \to \min, \quad x \in I := [a, b]. \tag{3.1.1}$$

Teoreticky sice víme, že stačí najít řešení úlohy $f'(x) = 0$ a odpovídající funkční hodnotu porovnat s $f(a)$ a $f(b)$, jenže řešení takové rovnice nemusí být zrovna snadné, např. $f(x) = (x-1)^2 + e^{x-1}$ vede na rovnici $2(x-1) + e^{x-1} = 0$...

My se omezíme pouze na jistou třídu funkcí, čímž zaručíme „hezké chování“ funkce f, a tudíž jednoznačnou řešitelnost úlohy (3.1.1) (příp. pro hledání hodnoty f^*).

Definice 3.1.1

Nechť je dán interval $I \subset \mathbb{R}$ a funkce $f : I \to \mathbb{R}$. Řekneme, že f je unimodální na I, jestliže existuje $x^* \in I$ takové, že

$$f(x_1) > f(x_2) \quad \text{pro libovolná } x_1, x_2 \in I \text{ splňující } x^* > x_2 > x_1,$$
$$f(x_1) < f(x_2) \quad \text{pro libovolná } x_1, x_2 \in I \text{ splňující } x^* < x_1 < x_2.$$

Jinými slovy...

Unimodalita vs. konvexnost/spojitost/diferencovatelnost.

Definice 3.1.1 dokonce dává přímý návod pro lokalizaci bodu x^*.

Lemma 3.1.2

Nechť $f : I \to \mathbb{R}$ je unimodální na I a $x_1, x_2 \in I$ jsou taková, že $x_1 < x_2$.

(i) Je-li $f(x_1) \leqslant f(x_2)$, pak $x^* \leqslant x_2$.

(ii) Je-li $f(x_1) \geqslant f(x_2)$, pak $x^* \geqslant x_1$.

Neohraničený interval?

Nyní si ukážeme několik numerických metod, s jejichž pomocí se budeme snažit najít přibližné řešení úlohy (3.1.1) pomocí **N hodnot** (= povolený počet vyčíslení funkce f). **Přesnost** těchto metod je dána hodnotou $|\bar{x} - x^*|$, kde x^* je (přesné) řešení úlohy (3.1.1) a $\bar{x}$ je jeho nalezená aproximace.

Tato metoda rozhodně není příliš intelektuálně náročná (**hrubá síla**) a vlastně ani efektivní (ale ideální pro začátek našich úvah). V závislosti na paritě N určíme dělící body intervalu I. Je-li N liché, pak

$$x_i := a + \frac{b-a}{N+1}\,i, \quad i = 1, \dots, N = 2k-1.$$

Je-li N sudé, pak

$$x_{2i} := a + \frac{b-a}{k+1}\,i \quad \& \quad x_{2i-1} := x_{2i} - \delta, \quad i = 1, \dots, k := N/2,$$

kde δ je vhodné malé kladné číslo. ✎Co znamená „vhodné“?

Vyčíslíme $f(x_1), \dots, f(x_N)$ (v případě $N = 2k$ & $\delta \in \{0, (b-a)/(k+1)\}$ máme de facto jenom k vyčíslení). Nechť v x_j nastává nejmenší funkční hodnota, tj.

$$f(x_j) = \min_{1 \leqslant i \leqslant N} f(x_i).$$

Položme $x_0 := a$ a $x_{N+1} := b$. ✎Potom z Lemma 3.1.2 plyne, že $x^* \in [x_{j-1}, x_{j+1}]$. Tento interval nazveme intervalem lokalizace minima (ILM) a za aproximaci x^* vezmeme střed ILM, tj. $\bar{x} := (x_{j-1} + x_{j+1})/2$.

Pak pro délku ILM platí

$$\ell_N := \max_{1\leqslant i\leqslant N}(x_{i+1}-x_{i-1}) = \begin{cases} 2\frac{b-a}{N+1}, & N=2k-1, \\ \frac{b-a}{(N/2)+1}+\delta, & N=2k, \end{cases}$$

což můžeme vyjádřit jako $\frac{b-a}{\lceil (N+1)/2\rceil}+\delta$ (přičemž pro liché N bereme $\delta=0$). Polovina tohoto čísla udává přesnost MPD. Jaký je řád/rychlost konvergence MPD? Vezmeme-li posloupnost délek ILM pro jednotlivé počty vyčíslení, pak pro liché dostaneme

$$\lim_{N\to\infty}\frac{\ell_{N+2}}{\ell_N}=1,$$

tj. máme sublineární konvergenci, tj. chová se jako $1/N$. Tento algoritmus je *pasivní*, neboť volba bodů $x_1,\ldots,x_N$ závisí pouze na počtu vyčíslení (a na δ).

✎ Počet vyčíslení vs. délka ILM.

✎ Příklad.

Nechť nyní $N = 2k$. ✎ Položme $x_1^- := \frac{1}{2}(a+b) - \delta$ a $x_1^+ := \frac{1}{2}(a+b) + \delta$, kde $\delta > 0$ je dostatečně malé číslo (co to znamená?). Porovnáme funkční hodnoty $f(x_1^-)$ a $f(x_1^+)$.

- Jestliže $f(x_1^-) < f(x_1^+)$, pak podle Lemma 3.1.2 je ILM $[a, x_1^+]$, neboť $x_1^- < x_1^+$. (Neboli: pokud funkční hodnoty rostou $\Longrightarrow$ můžeme ignorovat body za x_1^+.)
- Jestliže $f(x_1^-) > f(x_1^+)$, pak podle Lemma 3.1.2 je ILM $[x_1^-, b]$, neboť $x_1^- < x_1^+$. (Neboli: pokud funkční hodnoty klesají $\Longrightarrow$ můžeme ignorovat body před x_1^-.)

✎ Dodatek.

Tento nově získaný interval opět rozdělíme na polovinu a ve vzdálenosti δ od středu nalezneme body x_2^- a x_2^+. Následným porovnáním obou funkčních hodnot nalezneme nový ILM. Tento postup opakujeme celkem k-krát.

Algoritmický popis MPI

Zadání: Funkce f, interval $[a, b]$, přesnost $\varepsilon > 0$ nebo číslo $N \geqslant 2$. Stanovení čísel $N/2$ a δ.

Krok 1: (*Inicializace*). Položíme $a_0 := a$, $b_0 := b$, $k = 1$ a vypočteme

$$x_1^- := \frac{a_0 + b_0}{2} - \delta, \quad x_1^+ := \frac{a_0 + b_0}{2} + \delta.$$

Krok 2: Vyčíslíme $f(x_k^-)$ a $f(x_k^+)$. Jestliže $f(x_k^-) \geqslant f(x_k^+)$, pokračujeme Krokem 3. V opačném případě následuje Krok 4.

Krok 3: Položíme $a_k := x_k^-$, $b_k := b_{k-1}$ a pokračujeme Krokem 5.

Krok 4: Položíme $a_k := a_{k-1}$, $b_k := x_k^+$ a pokračujeme Krokem 5.

Krok 5: Je-li $k = N/2$ pokračujeme Krokem 6. Je-li $k < N/2$, vypočteme

$$x_{k+1}^- := \frac{a_k + b_k}{2} - \delta \quad \text{a} \quad x_{k+1}^+ := \frac{a_k + b_k}{2} + \delta.$$

Položíme $k := k + 1$ a pokračujeme Krokem 2.

Krok 6: Stanovíme poslední ILM jako $[a_k, b_k]$ a vypočteme $\bar{x} := \frac{a_k + b_k}{2}$. KONEC.

V prvním kroku je délka ILM

$$\ell_1 = b - x_1^- = x_1^+ - a = \tfrac{b-a}{2} + \delta.$$

Ve druhém kroku je délka ILM

$$\ell_2 = \tfrac{1}{2}(\tfrac{b-a}{2} + \delta) + \delta = \frac{b-a}{4} + \frac{3\delta}{2},$$

ve třetím

$$\ell_3 = \tfrac{1}{2}(\tfrac{b-a}{4} + \tfrac{3\delta}{2}) + \delta = \frac{b-a}{8} + \frac{7\delta}{4},$$

..., v i-tém kroku

$$\ell_i = \tfrac{1}{2}\ell_{i-1} + \delta = \tfrac{1}{2}(\tfrac{1}{2}\ell_{i-2} + \delta) + \delta = \cdots = \frac{b-a}{2^i} + \frac{(2^i-1)\delta}{2^{i-1}}.$$

Tedy pro $N = 2k$ vyčísleních budeme mít ILM délky

$$\ell_k = \frac{b-a}{2^k} + \frac{(2^k-1)\delta}{2^{k-1}}.$$

Pro $k \to \infty$ potom dostáváme

$$\ell_k = \tfrac{1}{2^k}(b-a) + (2 - \tfrac{1}{2^{k-1}})\delta \overset{k\to\infty}{\longrightarrow} 2\delta.$$

Jaký je řád/rychlost konvergence MPI? Platí (uvažuji posloupnost $\{\ell_k - 2\delta\}$ – proč?)

$$\lim_{k\to\infty} \frac{\ell_{k+1} - 2\delta}{\ell_k - 2\delta} = \lim_{k\to\infty} \frac{\ell_0/2^{k+1} - \delta/2^k}{\ell_0/2^k - \delta/2^{k-1}} = 1/2,$$

tj. konvergence je lineární s rychlostí $1/2$ (délka ILM klesá exponenciálně s exponentem $1/2$, tj. jako $(1/2)^k = (1/\sqrt{2})^N$).

Jako *aproximaci* $\bar{x}$ bodu x^* bereme střed k-tého ILM.

Přesnost této metody je tedy nejméně $\frac{1}{2}\ell_k \rightsquigarrow$ čím menší δ, tím lepší – vskutku?

Algoritmus je *aktivní*, neboť rozložení bodů x_i^- a x_i^+ závisí na předchozích bodech a odpovídajících funkčních hodnotách.

✎ Volba δ?

✎ Příklad.

Skutečné půlení intervalu? A lichý počet vyčíslení?

„Cesta“ k MZŘ

V MPI vyjadřuje číslo δ absolutní vzdálenost bodů $x_i^{\pm}$ od středu ILM. Ovšem δ by mohlo také vyjadřovat relativní vzdálenost bodů bodů $x_i^{\pm}$ od středu intervalu $I_{i-1} = [a_{i-1}, b_{i-1}]$, tj.

$$x_i^- = \frac{1}{2}(a_{i-1} + b_{i-1}) - \delta\,(b_{i-1} - a_{i-1}) = a_{i-1} + (1/2 - \delta)\,(b_{i-1} - a_{i-1}),$$
$$x_i^+ = \frac{1}{2}(a_{i-1} + b_{i-1}) + \delta\,(b_{i-1} - a_{i-1}) = a_{i-1} + (1/2 + \delta)\,(b_{i-1} - a_{i-1}).$$

Pak z intervalu I_{i-1} délky ℓ_{i-1} dostaneme interval délky

$$\ell_i = (1/2 + \delta)\,\ell_{i-1},$$

takže z výchozího intervalu $I = [a, b]$ budeme mít po k krocích (tj. $N = 2k$ vyčíslení) ILM délky

$$\ell_k = (1/2 + \delta)^k\,(b - a).$$

Mohli bychom tuto myšlenku ještě vylepšit? Ano, snížením potřebného počtu vyčíslení nebo lépe: při stejném počtu vyčíslení získat více ILM.

„Ideální“ případ: v každém kroku (vyjma prvního) budeme potřebovat pouze jedno nové vyčíslení. To je základní myšlenka následujících dvou metod.

„Cesta“ k MZŘ (pokrač.)

Z ILM $I_{i-1} = [a_{i-1}, b_{i-1}]$ chceme pomocí bodů $\lambda_i (= x_i^-)$ a $\mu_i (= x_i^+)$ (a Lemmatu 3.1.2) získat nový ILM $I_i = [a_i, b_i]$, tj. (obecně)

$$\lambda_i = a_{i-1} + \alpha\,\ell_{i-1}, \qquad \mu_i = a_{i-1} + \beta\,\ell_{i-1},$$

kde $\ell_{i-1} \coloneqq b_{i-1} - a_{i-1}$ je délka intervalu I_{i-1}, $\alpha, \beta \in (0,1)$ a $\alpha < \beta$. Potom unimodalita funkce f (a Lemma 3.1.2) opět dává

(i) $f(\lambda_i) < f(\mu_i) \implies x^* \in [a_{i-1}, \mu_i] =: [a_i, b_i]$,
(ii) $f(\lambda_i) \geqslant f(\mu_i) \implies x^* \in [\lambda_i, b_{i-1}] =: [a_i, b_i]$.

Jak ale najít vhodná α, β? My ještě požadujeme, aby λ_{i+1} nebo μ_{i+1} odpovídalo některému z předchozích bodů, konkrétně (šlo by i jinak?): v případě (i) $\mu_{i+1} = \lambda_i$ a (ii) $\lambda_{i+1} = \mu_i$, tj. v případě (i) máme

$$\mu_{i+1} = a_{i-1} + \beta^2\,\ell_{i-1} \quad \& \quad \lambda_i = a_{i-1} + \alpha\,\ell_{i-1} \quad \implies \quad \alpha = \beta^2,$$

a v případě (ii) dostaneme

$$\lambda_{i+1} = a_{i-1} + (2\alpha - \alpha^2)\,\ell_{i-1} \quad \& \quad \mu_i = a_{i-1} + \beta\,\ell_{i-1} \quad \implies \quad \beta = 2\alpha - \alpha^2.$$

Odtud získáme rovnici $2\beta^2 - \beta^4 = \beta$ neboli $\beta\,(\beta - 1)\,(\beta^2 + \beta - 1) = 0$, která má v intervalu $(0,1)$ jediný kořen

$$\beta = \frac{\sqrt{5}-1}{2} = \frac{1}{\tau} \approx 0{,}618 \qquad \implies \qquad \alpha = \beta^2 = 1 - \beta = \frac{1}{\tau^2} = \frac{3-\sqrt{5}}{2} \approx 0{,}382.$$

„Cesta“ k MZŘ (pokrač.)

Číslo τ je tzv. *zlaté číslo*. Platí

$$\tau^2 - \tau - 1 = 0.$$

Výpočet ILM

V intervalu lokalizace minima $I_{i-1} = [a_{i-1}, b_{i-1}]$ určíme body

$$\lambda_i = a_{i-1} + \frac{1}{\tau^2}(b_{i-1} - a_{i-1}),$$
$$\mu_i = a_{i-1} + \frac{1}{\tau}(b_{i-1} - a_{i-1}).$$

S jejich pomocí (a s využitím Lemmatu 3.1.2) získáme nový interval lokalizace minima $I_i = [a_i, b_i]$, jehož jeden krajní bod je právě jeden z bodů λ_i, μ_i, přičemž druhý z těchto bodů leží uvnitř tohoto intervalu (a stane se z něj μ_{i+1} nebo λ_{i+1}).

Průběh MZŘ s výchozím intervalem $I = [a_0, b_0]$ s délkou $\ell_0 := b_0 - a_0$.

i ($i+1$ vyčíslení)	vzdálenost od a_{i-1} (b_{i-1})		délka ILM
	λ_i (μ_i)	μ_i (λ_i)	
1	ℓ_0/τ^2	ℓ_0/τ	ℓ_0/τ
2	ℓ_0/τ^3	ℓ_0/τ^2	ℓ_0/τ^2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$N-1$	ℓ_0/τ^N	ℓ_0/τ^{N-1}	ℓ_0/τ^{N-1}

Algoritmický popis MZŘ

Zadání: Funkce f, interval $[a, b]$, přesnost $\varepsilon > 0$ nebo číslo $N \geqslant 2$.

Krok 1: (*Inicializace*) Položíme $a_0 := a$, $b_0 := b$ a $k := 1$. Vypočteme

$$\lambda_1 := a_0 + (b_0 - a_0)/\tau^2 \quad \text{a} \quad \mu_1 := a_0 + (b_0 - a_0)/\tau.$$

Krok 2: Je-li $k = N$, pokračujeme Krokem 7. Jinak následuje Krok 3.

Krok 3: Vyčíslíme $f(\lambda_k)$ a $f(\mu_k)$. Jestliže $f(\lambda_k) \geqslant f(\mu_k)$, pokračujeme Krokem 4. V opačném případě následuje Krok 5.

Krok 4: Položíme $a_k := \lambda_k$, $b_k := b_{k-1}$, $\lambda_{k+1} := \mu_k$ a

$$f(\lambda_{k+1}) := f(\mu_k), \quad \mu_{k+1} := a_k + (b_k - a_k)/\tau$$

a pokračujeme Krokem 6.

Krok 5: Položíme $a_k := a_{k-1}$, $b_k := \mu_k$, $\mu_{k+1} := \lambda_k$ a

$$f(\mu_{k+1}) := f(\lambda_k), \quad \lambda_{k+1} := a_k + (b_k - a_k)/\tau^2$$

a pokračujeme Krokem 6.

Krok 6: Položíme $k := k + 1$ a pokračujeme Krokem 2.

Krok 7: Stanovíme poslední ILM jako $[a_{k-1}, b_{k-1}]$ a vypočteme $\bar{x} := \frac{a_{k-1} + b_{k-1}}{2}$. KONEC.

Vlastnosti MZŘ

Rychlost konvergence:

$$\lim_{i\to\infty} \frac{\ell_{i+1}}{\ell_i} = \lim_{i\to\infty} \frac{\frac{\ell_0}{\tau^{i+1}}}{\frac{\ell_0}{\tau^i}} = \lim_{i\to\infty} \frac{1}{\tau} = \frac{1}{\tau}$$

tj. lineární konvergence s rychlostí $1/\tau \approx 0,618 \rightsquigarrow$ pomalejší než MPI??

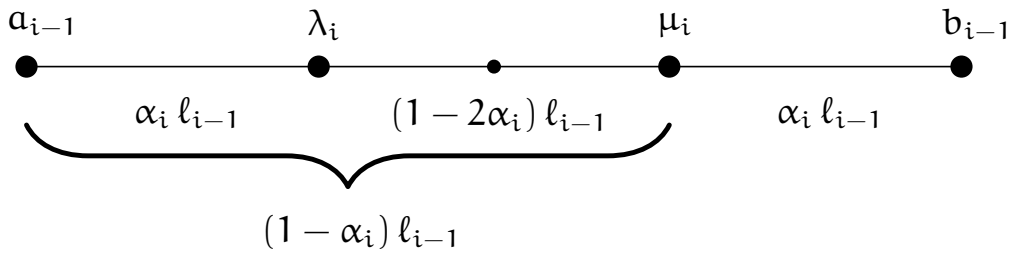
Geometrický popis MZŘ („nasazení" λ_1).

Příklad

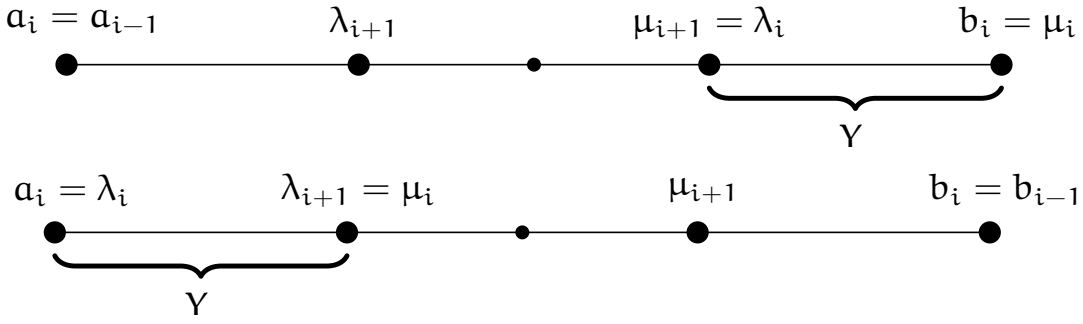
Metodou zlatého řezu s $N = 5$ najděme přibližně bod, ve kterém nastává minimum funkce $f(x) = 5x^2 - 4x + 2$ na intervalu $I = [0, 1]$.

„Cesta“ k FM

Nyní připusťme, že v každém kroku je možné mít jiné δ určující „relativní zkrácení“ ILM. Ovšem místo $1/2-\delta_i$ uvažme $\alpha_i \in (0,1/2)$ (číslo β_i zavádět nebudeme). Potom v i-té iteraci (tj. pro výpočet I_i) máme



z čehož podle toho, zda $f(\lambda_i) < f(\mu_i)$ nebo $f(\lambda_i) \geqslant f(\mu_i)$ dostaneme v dalším kroku (stejnou úvahou jako pro MZŘ)



„Cesta“ k FM (pokrač.)

Takže v obou případech musí platit

$$Y = (1-2\alpha_i)\,\ell_{i-1} \qquad \& \qquad Y = \alpha_{i+1}\,\ell_i = \alpha_{i+1}\,(1-\alpha_i)\,\ell_{i-1},$$

z čehož plyne

$$\alpha_{i+1} = 1 - \frac{\alpha_i}{1-\alpha_i} \tag{$*$}$$

(kdyby α bylo opět konstantní, dostaneme rovnici $\alpha^2-3\alpha+1=0$ s kořenem $\alpha = 1/\tau^2$, viz MZŘ). Budeme-li takto postupovat dále, bude po M krocích (tj. $M+1$ vyčíslení) délky ILM

$$\ell_M = (1-\alpha_1)\,(1-\alpha_2)\,\cdots\,(1-\alpha_M)\,\ell_0.$$

Vztah $(*)$ vlastně ukazuje, jako volit $\alpha_1, \alpha_2, \ldots$, aby celý algoritmus probíhal dle našich představ, tj. zvolíme $\alpha_1 \in (0,1/2)$ a v tu chvíli jsou již hodnoty $\alpha_2, \alpha_3, \ldots$ jasně dány. V takovém případě bude apriorní odhad chyby roven $\ell_M/2$. Doplníme-li přirozený požadavek, aby tato chyba byla nejmenší možná dostaneme úlohu

$$(1-\alpha_1)\,(1-\alpha_2)\cdots(1-\alpha_M) \to \min,$$

$$\alpha_{i+1} = 1 - \frac{\alpha_i}{1-\alpha_i}, \quad i = 1,2,\ldots,M-1,$$

$$0 \leqslant \alpha_i \leqslant 1/2, \quad i = 1,\ldots,M,$$

Povolujeme i „extrémní“ hodnoty $\alpha_i \in \{0,1/2\}$, co kdyby náhodou... (ale je jasné, že $\alpha_1 \in \{0,1/2\}$ k řešení nepovede).

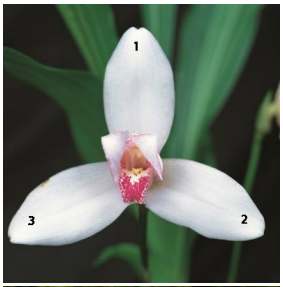
„Cesta“ k FM (pokrač.)

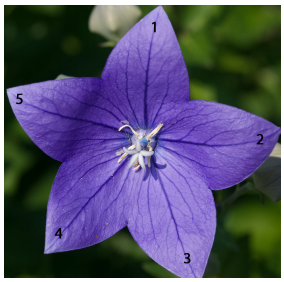
Řešení této úlohy úzce souvisí s tzv. Fibonacciho čísly F_i, kde

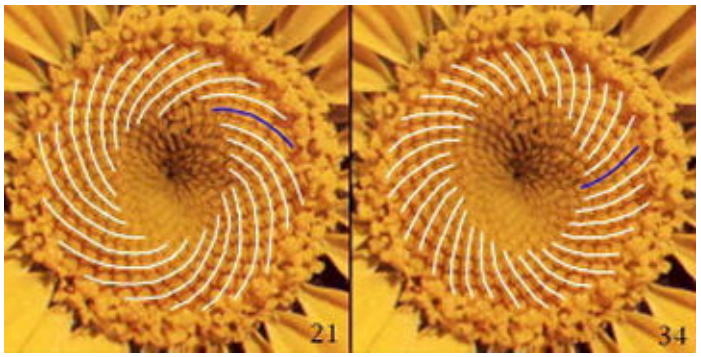
$$F_{i+2} = F_{i+1} + F_i, \quad F_0 = 1, \quad F_1 = 1,$$

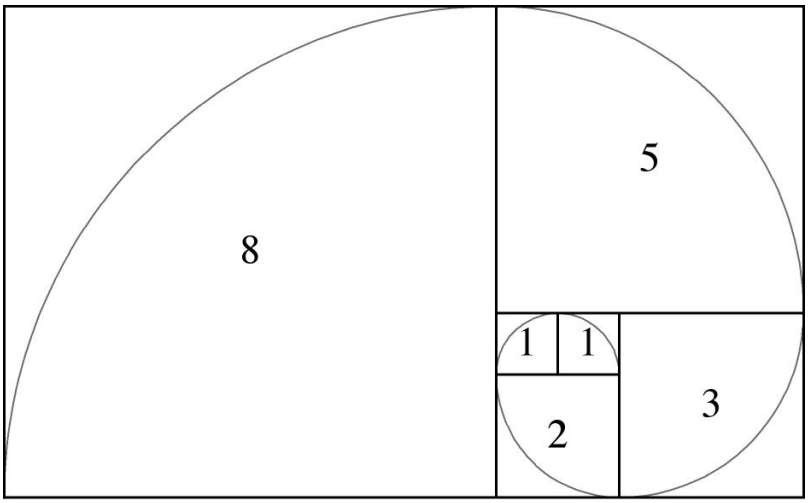
tj. máme posloupnost $1, 1, 2, 3, 5, 8, 13, 21, 34, \ldots$. Tato čísla lze také vyjádřit pomocí zlatého čísla τ, konkrétně

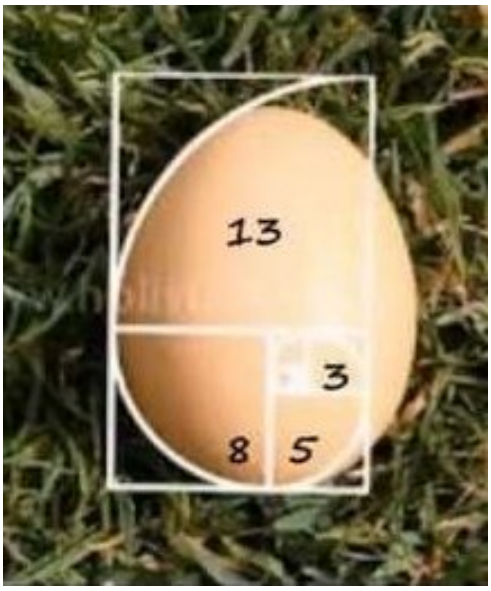
$$F_i = \frac{1}{\sqrt{5}} \left[\tau^{i+1} - \left(-\frac{1}{\tau} \right)^{i+1} \right]$$

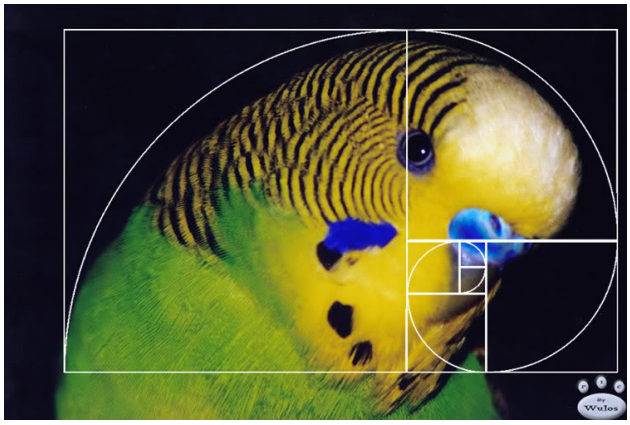


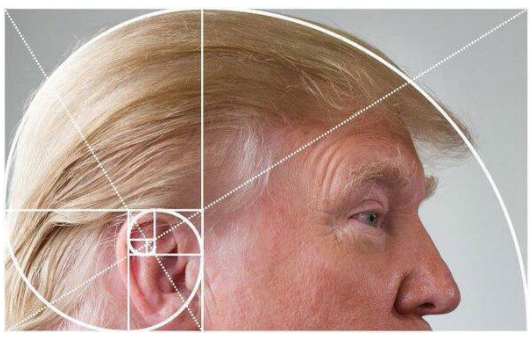












Řešením naší minimalizační úlohy jsou čísla

$$\alpha_1 = \frac{F_{M-1}}{F_{M+1}}, \quad \alpha_2 = \frac{F_{M-2}}{F_M}, \quad \dots \quad , \quad \alpha_i = \frac{F_{M-i}}{F_{M-i+2}}, \quad \dots \quad , \quad \alpha_M = \frac{1}{2}$$

a délka intervalu I_M (po $M+1$ vyčísleních) je

$$\ell_M = \frac{\ell_0}{F_{M+1}}.$$

Výpočet FM

Máme povoleno N vyčíslení, takže $M = N - 1$ a

$$\lambda_i = a_{i-1} + \frac{F_{N-i-1}}{F_{N-i+1}}\,\ell_{i-1} = b_{i-1} - \frac{F_{N-i}}{F_{N-i+1}}\,\ell_{i-1},$$
$$\mu_i = a_{i-1} + \frac{F_{N-i}}{F_{N-i+1}}\,\ell_{i-1} = b_{i-1} - \frac{F_{N-i-1}}{F_{N-i+1}}\,\ell_{i-1}.$$

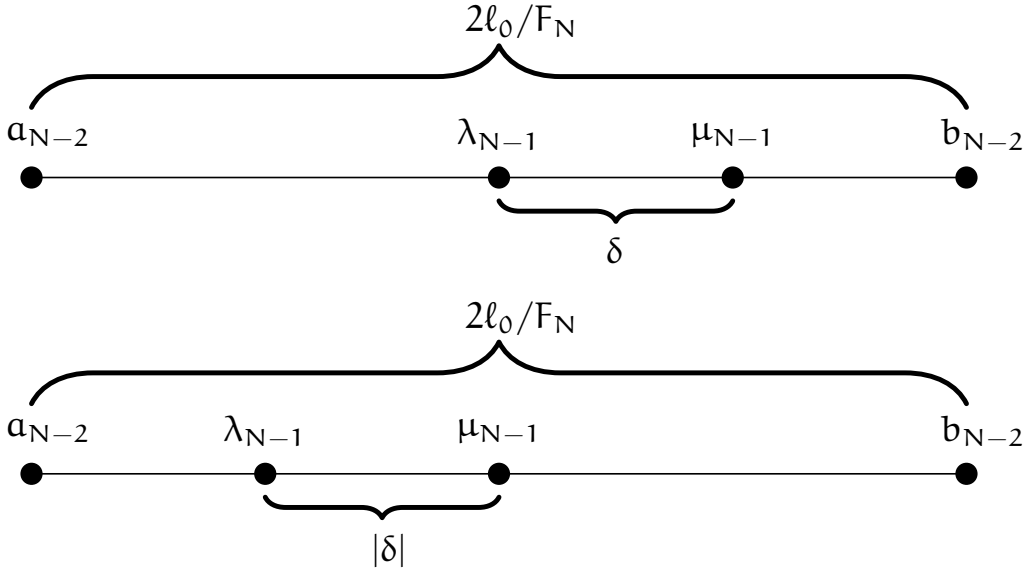
Ovšem pro $i = N - 1$ (tj. po vyčerpání $N - 1$ vyčíslení) je

$$\lambda_{N-1} = a_{N-2} + \frac{1}{2}\ell_{N-2} = \mu_{N-1},$$

takže v posledním kroku nedostaneme žádný nový bod. To je vlastně daň za to, že jsme povolili $\alpha_i \in \{0, 1/2\}$, jenže v opačném případě by minimalizační úlohy neměla řešení.

Poslední krok FM

Co s tím?? Jedno vyčíslení nevyužijeme? Zvolíme-li číslo $0 < |\delta| < \ell_0/F_N$, potom v závislosti na $\delta \in (0, \ell_0/F_N)$ nebo $\delta \in (-\ell_0/F_N, 0)$ dostaneme



Opět vyčíslíme a porovnáme hodnoty $f(\lambda_{N-1})$ a $f(\mu_{N-1})$, čímž získáme poslední ILM I_{N-1} s délkou ℓ_0/F_N nebo $\ell_0/F_N + |\delta|$.

Průběh FM s výchozím intervalem $I = [a_0, b_0]$ s délkou $\ell_0 := b_0 - a_0$.

i ($i+1$ vyčíslení)	vzdálenost od a_{i-1} (b_{i-1})		délka ILM
	λ_i (μ_i)	μ_i (λ_i)	
1	$\frac{F_{N-2}}{F_N}\,\ell_0$	$\frac{F_{N-1}}{F_N}\,\ell_0$	$\frac{F_{N-1}}{F_N}\,\ell_0$
2	$\frac{F_{N-3}}{F_N}\,\ell_0$	$\frac{F_{N-2}}{F_N}\,\ell_0$	$\frac{F_{N-2}}{F_N}\,\ell_0$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	$\frac{F_{N-i-1}}{F_N}\,\ell_0$	$\frac{F_{N-i}}{F_N}\,\ell_0$	$\frac{F_{N-i}}{F_N}\,\ell_0$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$N-2$	$\frac{1}{F_N}\,\ell_0$	$\frac{2}{F_N}\,\ell_0$	$\frac{2}{F_N}\,\ell_0$
$N-1$ & $\delta > 0$	$\frac{1}{F_N}\,\ell_0$	$\frac{1}{F_N}\,\ell_0 \overset{+}{(-)}\delta$	$\frac{1}{F_N}\,\ell_0$ nebo $\frac{1}{F_N}\,\ell_0 + \delta$
$N-1$ & $\delta < 0$	$\frac{1}{F_N}\,\ell_0 \overset{+}{(-)}\delta$	$\frac{1}{F_N}\,\ell_0$	$\frac{1}{F_N}\,\ell_0$ nebo $\frac{1}{F_N}\,\ell_0 + \lvert\delta\rvert$

Algoritmický popis FM

Zadání: Funkce f, interval $[a, b]$, přesnost $\varepsilon > 0$ nebo číslo $N \geqslant 2$. Stanovení δ splňujícího $0 < |\delta| < \ell_0/F_N$.

Krok 1: (*Inicializace*) Položíme $a_0 := a$, $b_0 := b$ a $k := 1$. Vypočteme

$$\lambda_1 := a_0 + \frac{F_{N-2}}{F_N}(b_0 - a_0) \quad \text{a} \quad \mu_1 := a_0 + \frac{F_{N-1}}{F_N}(b_0 - a_0).$$

Krok 2: Je-li $k = N - 1$, pokračujeme Krokem 7. Jinak následuje Krok 3.

Krok 3: Vyčíslíme $f(\lambda_k)$ a $f(\mu_k)$. Jestliže $f(\lambda_k) > f(\mu_k)$, pokračujeme Krokem 4. V opačném případě následuje Krok 5.

Krok 4: Položíme $a_k := \lambda_k$, $b_k := b_{k-1}$, $\lambda_{k+1} := \mu_k$ a

$$f(\lambda_{k+1}) := f(\mu_k), \quad \mu_{k+1} := a_k + \frac{F_{N-k-1}}{F_{N-k}}(b_k - a_k)$$

a pokračujeme Krokem 6.

Krok 5: Položíme $a_k := a_{k-1}$, $b_k := \mu_k$, $\mu_{k+1} := \lambda_k$ a

$$f(\mu_{k+1}) := f(\lambda_k), \quad \lambda_{k+1} := a_k + \frac{F_{N-k-2}}{F_{N-k}}(b_k - a_k)$$

a pokračujeme Krokem 6.

Algoritmický popis FM (pokr.)

Krok 6: Položíme $k := k+1$ a pokračujeme Krokem 2.

Krok 7: Položíme $\sigma := (b_k + a_k)/2 + \delta$. Vyčíslíme $f\big((b_k + a_k)/2\big)$ a $f(\sigma)$.

(a) Je-li $\sigma < (b_k + a_k)/2$ a $f(\sigma) < f\big((b_k + a_k)/2\big)$, pak $a_{N-1} := a_{N-2}$ a $b_{N-1} := (b_k + a_k)/2$.

(b) Je-li $\sigma < (b_k + a_k)/2$ a $f(\sigma) \geqslant f\big((b_k + a_k)/2\big)$, pak $a_{N-1} := \sigma$ a $b_{N-1} := b_{N-2}$.

(c) Je-li $(b_k + a_k)/2 < \sigma$ a $f\big((b_k + a_k)/2\big) < f(\sigma)$, pak $a_{N-1} := a_{N-2}$ a $b_{N-1} := \sigma$.

(d) Je-li $\sigma < (b_k + a_k)/2$ a $f\big((b_k + a_k)/2\big) \geqslant f(\sigma)$, pak $a_{N-1} := (b_k + a_k)/2$ a $b_{N-1} := b_{N-2}$.

Posledním ILM je $[a_{k-1}, b_{k-1}]$ a vypočteme $\bar{x} := \frac{a_{k-1}+b_{k-1}}{2}$.
KONEC.

Vlastnosti FM

Rychlost konvergence:

$$\lim_{N\to\infty} \frac{\ell_{N+1}(-\delta)}{\ell_N(-\delta)} = \lim_{N\to\infty} \frac{\frac{\ell_0}{F_{N+1}}}{\frac{\ell_0}{F_N}} = \lim_{N\to\infty} \frac{F_N}{F_{N+1}} = \frac{1}{\tau}$$

tj. lineární konvergence s rychlostí $1/\tau \approx 0,618$ $\rightsquigarrow$ stejné jako MZŘ. A v jiných ohledech??

Geometrický popis FM („nasazení“ λ_1).

Příklad

Fibonacciho metodou najděme přibližně bod, ve kterém nastává minimum funkce $f(x) = 7x^2 - 6x + 2$ na intervalu $I = [0, 1]$ s přesností nejméně $\varepsilon = 0,14$.

Další numerické metody

Pro diferencovatelné funkce můžeme také použít metody známé z kurzu numerických metod pro hledání kořenů funkce $f(x)$, tj. řešení rovnice $f(x) = 0$, např.

- metoda půlení intervalu,
- Newtonova metoda,
- metoda sečen.

Ovšem vše aplikujeme pro hledání stacionární bodu funkce $f(x)$, tj. řešení rovnice $f'(x) = 0$.

Nyní se budeme věnovat některým numerickým metodám pro řešení úlohy

$$f(x) \to \min, \qquad x \in \mathbb{R}^n, \tag{3.2.1}$$

kde $f : \mathbb{R}^n \to \mathbb{R}$ je (jednou/dvakrát/třikrát) spojitě diferencovatelná funkce. Ukážeme si dvě metody 1. řádu

- metoda největšího spádu (MNS),
- metoda sdružených gradientů (MSG),

a metodu 2. řádu

- Newtonova metoda (NM).

Tyto metody jsou založeny (přeneseně i NM) na konstrukci tzv. *minimalizující posloupnosti* $\{x^{[k]}\}$, kde

$$x^{[k+1]} = x^{[k]} + \alpha_k h_k,$$

přičemž $\alpha_k \in \mathbb{R}$ se nazývá *délka* k*-tého kroku* a vektor $h_k \in \mathbb{R}^n$ je *směr* k*-tého kroku*.

Jednotlivé metody se zásadně liší volbou vektorů h_k.

Volba délky kroku α_k může být stejná — jen je nutné zaručit, aby kroky nebyly příliš dlouhé ani příliš krátké. My se budeme držet tzv. přesné minimalizace, tj. α_k určíme jako řešení jisté jednorozměrné minimalizační úlohy (tedy pomocí derivování). Toto samozřejmě není z praktického pohledu příliš efektivní, neboť nalezení přesného řešení může vyžadovat nemalé úsilí (a v praxi to může být i nemožné). Proto se místo přesné minimalizace používají jiné nástroje: (i) konstantní volba, (ii) klesající volba, (iii) podmíněná minimalizace, (iv) řešení pomocí jednorozměrných numerických metod, (v) zpětná minimalizace (drobení kroku).

Volba h_k a metody nultého řádu

Navzdory různým přístupům k volbě h_k mají všechny metody jednu společnou vlastnost: jedná se o spádové metody. To znamená, že volbou h_k a α_k dostaneme minimalizující posloupnost $\{x^{[k]}\}$, která *generuje klesající posloupnost funkčních hodnot* $\{f(x^{[k]})\}$.

Nejdříve můžeme naznačit dvě metody 0. řádu.

- *Metoda náhodného hledání*: na základě jistého výběrového kritéria vybereme kandidáty pro vyčíslení a za člena minimalizující posloupnosti vezmeme bod s nejmenší funkční hodnotou (?? ruční výpočet ??).
- *Metoda souřadnicového spádu*: minimalizace funkce podél souřadných os, tj. ve směru vektorů $e_1, \ldots, e_n$, přičemž v každém kroku řešíme n jednorozměrných minimalizačních úloh. Z výchozího bodu $x^{[0]}$ „vyrazíme“ ve směru e_1 a hledáme bod s nejmenší funkční hodnotou. Z tohoto bodu vyrazíme ve směru e_2, atd. až po e_n, čímž získáme bod $x^{[1]}$ a celý postup opakuje s tímto výchozím bodem (?? nalezení řešení ??).

Volba h_k a α_k

Nejpřirozenější volba směru je

$$h_k = -\operatorname{grad} f(x^{[k]})$$

(poprvé použito asi Cauchym v roce 1847). Jedná se tedy o tzv. *gradientní metodu*, přičemž MNS se od ostatních liší právě tím, že délku kroku volíme jako přesné řešení úlohy

$$f(x^{[k+1]}) = f(x^{[k]} - \alpha_k \operatorname{grad} f(x^{[k]})) = \min_{\alpha \geqslant 0} f(x^{[k]} - \alpha \operatorname{grad} f(x^{[k]})). \quad (*)$$

Díky této volbě je MNS spádovou metodou, tj. v případě $\operatorname{grad} f(x^{[k]}) \neq 0$ platí $f(x^{[k+1]}) < f(x^{[k]})$.

Konec MNS

Z předpisu pro výpočet α_k je jasné, že v okamžiku $\operatorname{grad} f(x^{[k]}) = 0$ se minimalizující posloupnost stane konstantní (vlastně nalezneme stacionární bod). Toto ale v praxi není úplně vhodné ukončovací kritérium, obvykle se volí kritéria typu (absolutní vs. relativní)

$$\|\operatorname{grad} f(x^{[k]})\| < \varepsilon, \quad |f(x^{[k+1]}) - f(x^{[k]})| < \varepsilon, \quad \|x^{[k+1]} - x^{[k]}\| < \varepsilon,$$
$$\frac{|f(x^{[k+1]}) - f(x^{[k]})|}{|f(x^{[k]})|} < \varepsilon, \quad \frac{\|x^{[k+1]} - x^{[k]}\|}{\|x^{[k]}\|} < \varepsilon.$$

Ortogonalita v MNS

Vektor $\operatorname{grad} f(x^{[k]})$ je normálovým vektorem vrstevnice funkce f na úrovni $f(x^{[k]})$. Navíc díky přesné minimalizaci v $(*)$ dostáváme, že pro $\alpha = \alpha_k \neq 0$ musí platit

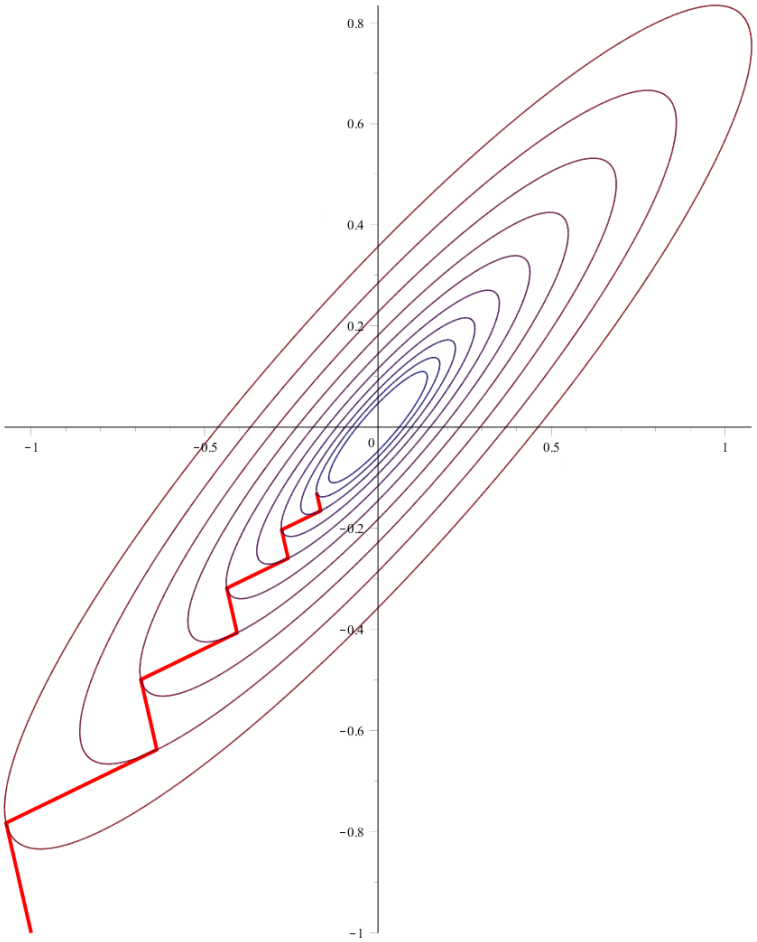
$$0 = \frac{d}{d\alpha} f(x^{[k]} - \alpha \operatorname{grad} f(x^{[k]}))\Big|_{\alpha=\alpha_k} = \frac{d}{d\alpha} f(x^{[k+1]}(\alpha))\Big|_{\alpha=\alpha_k} =$$
$$= \operatorname{grad}^\top f(x^{[k+1]}(\alpha))\Big|_{\alpha=\alpha_k} \frac{d}{d\alpha} x^{[k+1]}(\alpha)\Big|_{\alpha=\alpha_k} =$$
$$= -\operatorname{grad}^\top f(x^{[k+1]}) \operatorname{grad} f(x^{[k]}),$$

tedy směry největšího spádu funkce f v bodech $x^{[k]}$ a $x^{[k+1]}$ (tj. vektory $\operatorname{grad} f(x^{[k]})$ a $\operatorname{grad} f(x^{[k+1]})$) jsou ortogonální. Proto platí

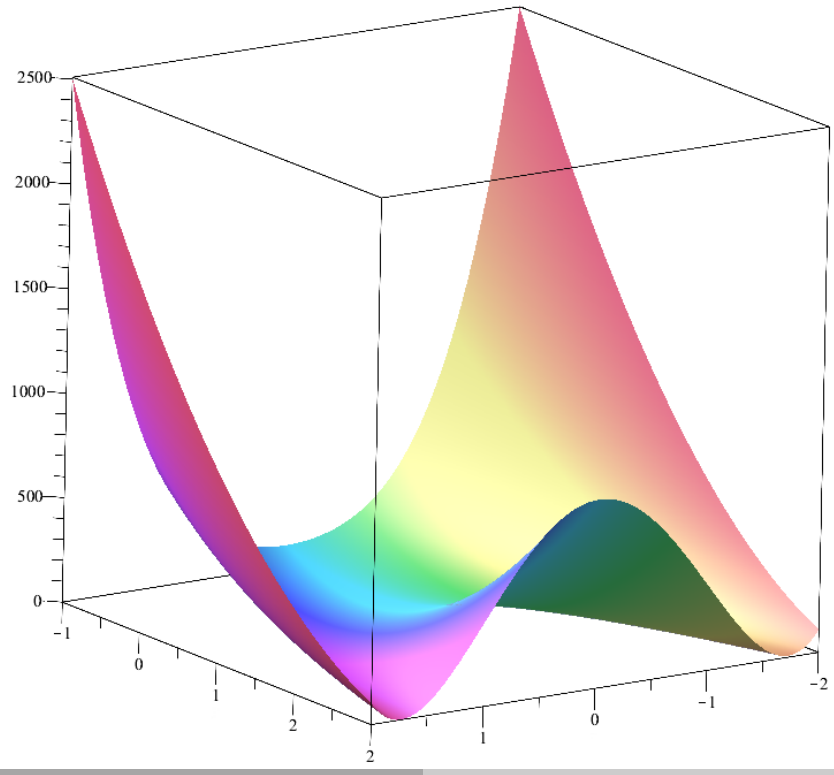
$$\left\langle x^{[k+2]} - x^{[k+1]}, x^{[k+1]} - x^{[k]} \right\rangle = \alpha_k \alpha_{k+1} \left\langle \operatorname{grad} f(x^{[k+1]}), \operatorname{grad} f(x^{[k]}) \right\rangle = 0,$$

tedy vektory určené body $x^{[k+1]}$, $x^{[k]}$ a $x^{[k+2]}$, $x^{[k+1]}$ jsou také ortogonální. To nás přivádí k typickému obrázku ilustrujícím MNS a jejímu jednoduchému popisu: „vyjdeme“ z bodu $x^{[k]}$ ve směru vektoru $-\operatorname{grad} f(x^{[k]})$ a hledáme nejbližší vrstevnici, pro kterou bude tato polopřímka tečnou.

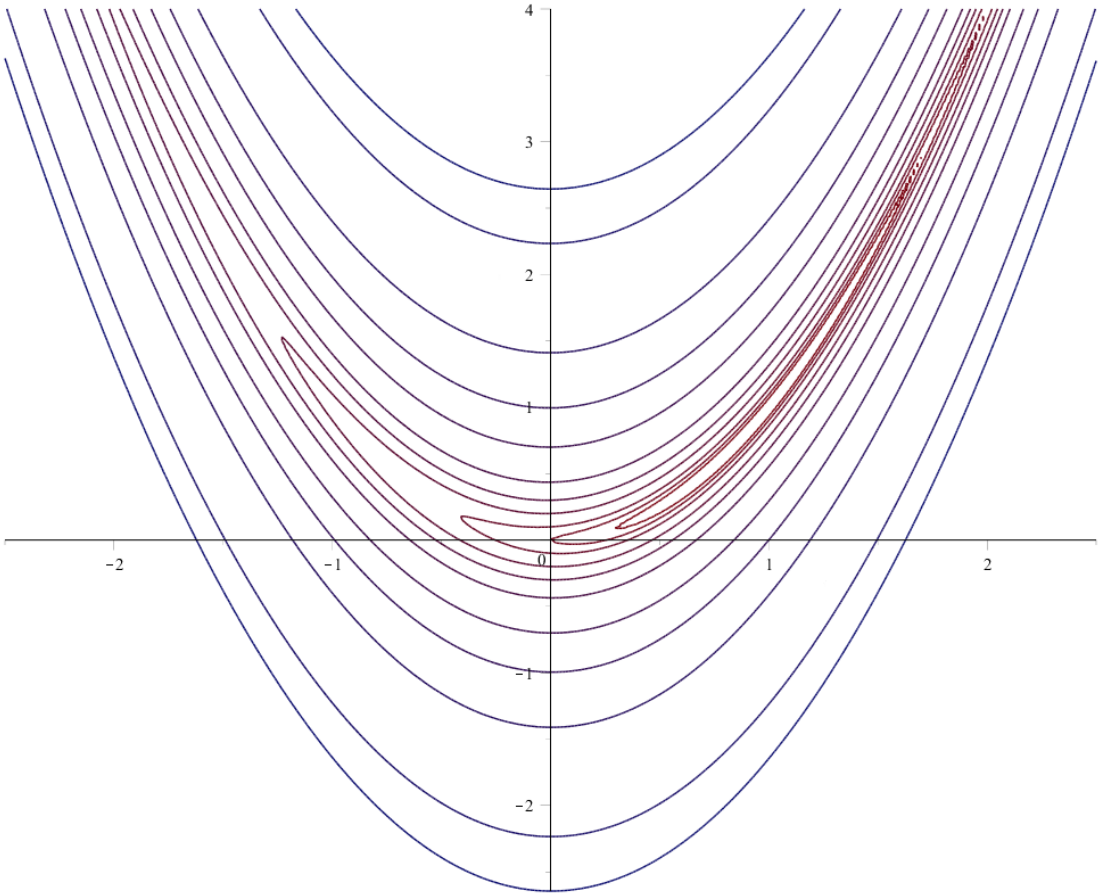
MNS pro $f(x_1, x_2) = 3x_1^2 - 7x_1x_2 + 5x_2^2$ s $x^{[0]} = [-1, -1]$



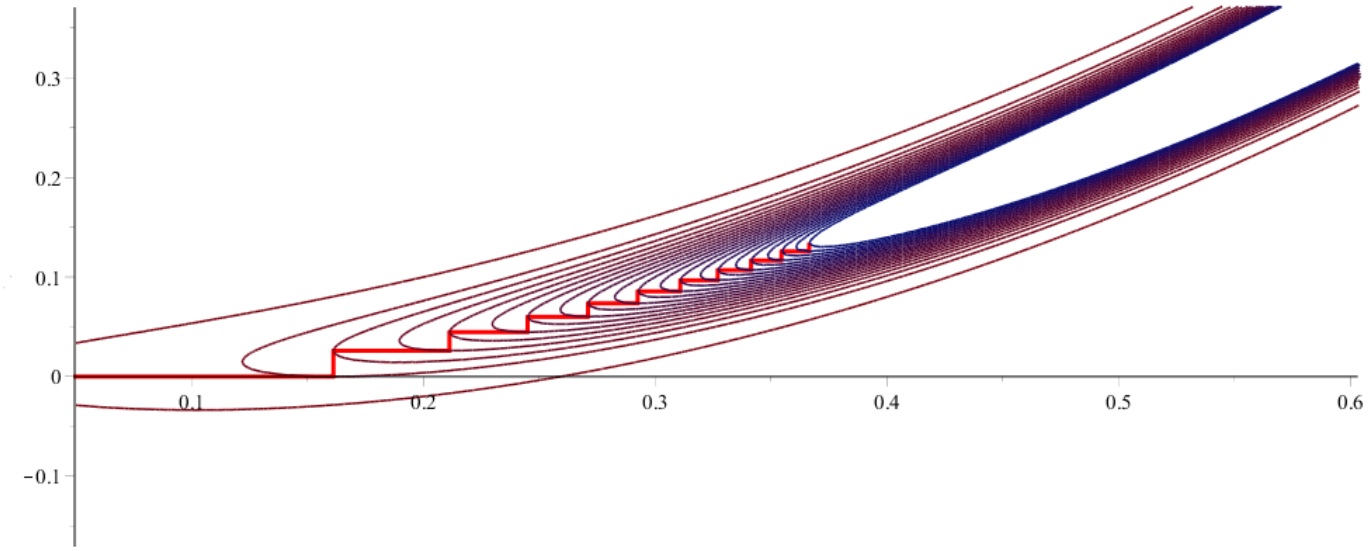
Avšak v některých případech může být „pohyb“ určený MNS velmi zdlouhavý a vede ke „klikatění“ (tzv. cik-cak efekt). — např. pro *Rosenbrockovu funkci* (též Rosenbrockovo údolí, Rosenbrockova banánová funkce) $f(x_1, x_2) = (a - x_1)^2 + b(x_2 - x_1^2)^2$ s globálním minimem v bodě $[a, a^2]$. Jedná se o nekonvexní funkci používanou k testování optimalizačních algoritmů (jedna z tzv. *testovacích funkcí*). Nejčastěji se volí $a = 1$ a $b = 100$, viz obrázek.



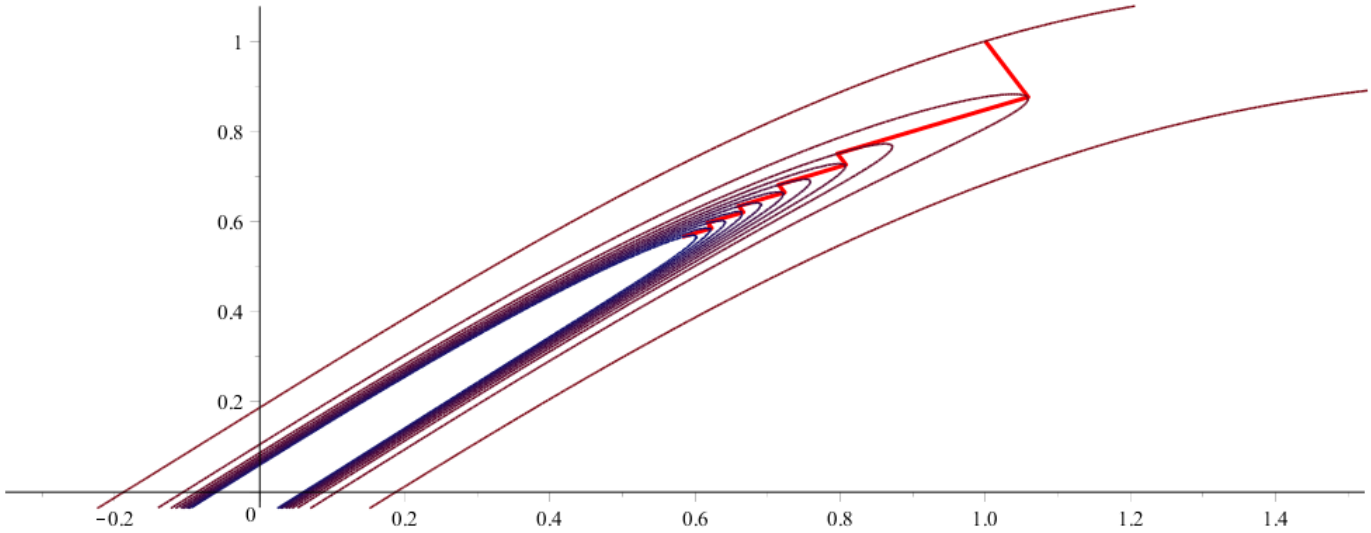
Vrstevnice Rosenbrockovy funkce $f(x_1, x_2) = (1 - x_1)^2 + 100(x_2 - x_1^2)^2$.



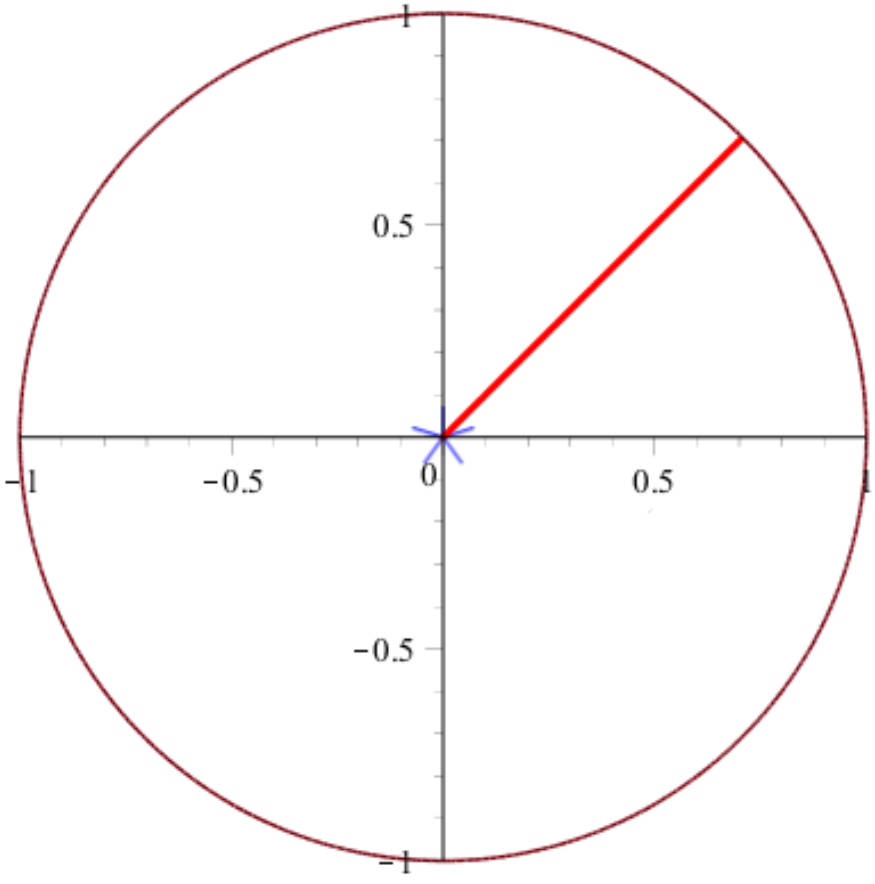
MNS pro Rosenbrockovu funkci $f(x_1, x_2) = (1 - x_1)^2 + 100(x_2 - x_1^2)^2$ s $x^{[0]} = [0, 0]$ a $x^* = [1, 1]$.



MNS pro funkci $f(x_1, x_2) = 10(x_2 - \sin x_1)^2 + x_1^2/10$ s $x^{[0]} = [1, 1]$



MNS pro $f(x_1, x_2) = (x_1^2 + x_2^2)^{3/2}$ s $x^{[0]} = [1/\sqrt{2}, 1/\sqrt{2}]$



MNS pro kvadratické funkce

Nyní se podíváme blíže na MNS pro kvadratické funkce

$$f(x) = \frac{1}{2}\langle Qx, x\rangle - x^\top b, \tag{3.2.2}$$

kde $Q^\top = Q > 0$ je $n \times n$ matice a $b \in \mathbb{R}^n$. Pak funkce f je ostře (a současně i silně) konvexní. Vlastní hodnoty matice Q jsou kladné a můžeme je uspořádat jako

$$0 < \lambda_1 \leqslant \lambda_2 \leqslant \cdots \leqslant \lambda_n.$$

V tomto případě můžeme vyřešit úlohu (3.2.1) „přímo“ — řešením je bod x^* splňující

$$Qx^* = b, \quad \text{tj.} \quad x^* = Q^{-1}b$$

s hodnotou $f(x^*) = -\frac{1}{2}b^\top Q^{-1}b$ (ale i v tomto „jednoduchém“ případě může být řešení soustavy nebo výpočet matice Q^{-1} početně velmi náročné).

MNS pro kvadratické funkce (pokr.)

Gradient funkce $f(x)$ je

$$g(x) \coloneqq \operatorname{grad} f(x) = Qx - b,$$

takže jednotlivé iterační body jsou tvaru ($g_k \coloneqq Qx^{[k]} - b$)

$$x^{[k+1]} = x^{[k]} - \alpha_k\, g_k = x^{[k]} - \alpha_k(Qx^{[k]} - b),$$

přičemž α_k můžeme explicitně určit jako

$$\alpha_k = \frac{g_k^\top g_k}{g_k^\top Q g_k} \quad \text{s} \quad \alpha_k \in [1/\lambda_n, 1/\lambda_1].$$

Tudíž

$$x^{[k+1]} = x^{[k]} - \frac{g_k^\top g_k}{g_k^\top Q g_k}\, g_k.$$

V tomto případě jsou vrstevnice n-dimenzionální elipsoidy s osami ve směru vlastních vektorů, které jsou vzájemně ortogonální. Délka poloosy odpovídající i-tému vlastnímu vektoru je násobkem čísla $1/\sqrt{\lambda_i}$ (viz první obrázek pro MNS; ale pro $b \neq 0$ bude střed posunutý mimo počátek).

Nyní se podíváme na konvergenci MNS. Ovšem místo funkce $f(x)$ se zaměříme na funkci

$$E(x) := \frac{1}{2}(x - x^*)^\top Q(x - x^*) = f(x) - f(x^*).$$

Lemma 3.2.1

Platí

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

Odtud vidíme, že $E(x^{[k+1]}) = 0$ neboli $f(x^{[k+1]}) = f(x^*)$ právě tehdy, když platí

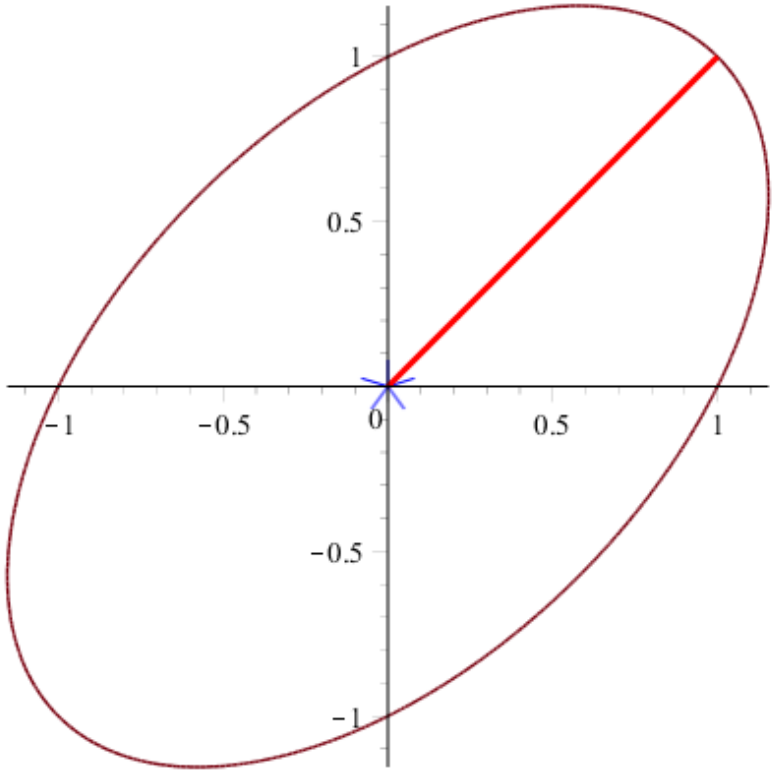
$$1 = \frac{(g_k^\top g_k)^2}{(g_k^\top Q g_k)\,(g_k^\top Q^{-1} g_k)}. \qquad (**)$$

Jelikož $f(x)$ je ostře konvexní funkce, platí v takovém případě $x^{[k+1]} = x^*$, a MNS dává přesné řešení po $k+1$ krocích.

?? Kdy nastane (**) ??

Pokud pro žádné $k \in \mathbb{N}$ nenastane rovnost (**), tak MNS je nekonečněkrokovým iteračním procesem.

MNS pro $f(x_1, x_2) = x_1^2 - x_1 x_2 + x_2^2$ s $x^{[0]} = [1, 1]$



Lemma 3.2.2

(Kantorovičova nerovnost) Nechť $A = A^\top > 0$ je $n \times n$ matice. Pak pro každý vektor $x \in \mathbb{R}^n \setminus \{0\}$ platí

$$\frac{(x^\top x)^2}{(x^\top Ax)(x^\top A^{-1}x)} \geqslant \frac{4\lambda_{\min}\lambda_{\max}}{(\lambda_{\min}+\lambda_{\max})^2},$$

kde $\lambda_{\min}$ je nejmenší vlastní číslo matice A a $\lambda_{\max}$ je největší vlastní číslo matice A.

Kombinací Lemma 3.2.1 a 3.2.2 dostaneme hlavní výsledek o konvergenci MNS.

Věta 3.2.3

Nechť platí (3.2.2). Pak pro libovolné $x^{[0]} \in \mathbb{R}^n$ posloupnost $\{x^{[k]}\}$ generovaná MNS konverguje k jedinému řešení x^* úlohy (3.2.1). Navíc pro funkci $E(x)$ platí

$$E(x^{[k+1]}) \leqslant \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right)^2 E(x^{[k]}).$$

Tedy $E(x^{[k]}) \leqslant C\beta^k$, kde $k \in \mathbb{N}$, $C = f(x^{[0]}) - f(x^*)$ a $\beta := \left(\frac{\lambda_n-\lambda_1}{\lambda_n+\lambda_1}\right)^2$. To znamená, že konvergence je alespoň lineární s rychlostí nejméně β. Nicméně toto je pouze horní odhad a skutečná rychlost konvergence závisí na volbě $x^{[0]}$.

V případě $\lambda_1 = \lambda_n$ (tj. $\lambda_1 = \cdots = \lambda_n$) bude konvergence dokonce alespoň superlineární. Ovšem již v případě $\lambda_1 = \lambda_2 = \cdots = \lambda_{n-1} \neq \lambda_n$ může být konvergence velmi pomalá. Dá se také ukázat, že pro „téměř všechny“ výchozí body $\{x^{[0]}\}$ bude konvergence lineární s rychlostí velmi blízkou číslu β.

Hodnota β závisí na rozdílu $\lambda_n - \lambda_1$, což úzce souvisí s tzv. *excentricitou* elipsoidů, které jsou vrstevnicemi funkce f (jde o „míru zploštění“ elipsoidů, větší excentricita znamená větší hodnotu β, viz vrstevnice pro Rosenbrockovu funkci).

Číslo β můžeme také vyjádřit jako

$$\beta = \left(\frac{\kappa - 1}{\kappa + 1}\right)^2,$$

kde $\kappa := \lambda_n/\lambda_1$ vyjadřuje tzv. podmíněnost matice Q („citlivost“; hodnota $1/\kappa$ vyjadřuje „vzdálenost“ od nejbližší singulární matice). Čím větší je hodnota κ, tím horší je podmíněnost matice Q. Uvažte např. $f(x_1, x_2) = x_1^2 + \alpha x_2^2$ s $\alpha = 2$ a $\alpha = 1/100$ (potom $\kappa = 2$ a $\kappa = 100$).

Příklad

Pomocí MNS určeme první iteraci $x^{[1]}$ pro řešení úlohy

$$f(x_1, x_2) = x_1^2 + 2x_1x_2 + 2x_2^2 \to \min$$

při počáteční aproximaci $x^{[0]} = [1, 1]$.

Vlastnosti MNS pro nekvadratické funkce?

MNS pro nekvadratické funkce (i)

Nechť funkce $f : \mathbb{R}^n \to \mathbb{R}$ je spojitě diferencovatelná na $\mathbb{R}^n$, zdola ohraničená na $\mathbb{R}^n$ a gradient funkce f je lipschitzovsky spojitý, tj. existuje číslo $L > 0$ takové, že

$$\|\operatorname{grad} f(x) - \operatorname{grad} f(y)\| \leqslant L\|x - y\| \quad \text{pro každé} \quad x, y \in \mathbb{R}^n.$$

Nechť $\{x^{[k]}\}$ je posloupnost vytvořena pomocí MNS. Potom posloupnost $\{f(x^{[k]})\}$ je nerostoucí a pro $k \in \mathbb{N} \cup \{0\}$ dokonce platí $f(x^{[k+1]}) < f(x^{[k]})$ v případě $\operatorname{grad} f(x^{[k]}) \neq 0$ (tj. $\{x^{[k]}\}$ je minimalizující posloupnost), přičemž

$$\lim_{k\to\infty} \operatorname{grad} f(x^{[k]}) = 0,$$

tj. každý hromadný bod posloupnosti $\{x^{[k]}\}$ je současně stacionárním bodem funkce f.

MNS pro nekvadratické funkce (ii)

Konverguje-li dokonce posloupnost $\{x^{[k]}\}$ k bodu x^* a funkce f je navíc dvakrát spojitě diferencovatelná v okolí x^* s

$$aI \leqslant \nabla^2 f(x^*) \leqslant AI$$

kde $a, A > 0$ (tedy f je v okolí bodu x^* silně konvexní a $\nabla^2 f(x^*) > 0$, tj. x^* je nedegenerované lokální minimum funkce f), pak platí

$$f(x^{[k+1]}) - f(x^*) \leqslant \left(\frac{A-a}{A+a}\right)^2 (f(x^{[k]}) - f(x^*)),$$

tj. konvergence je (alespoň) lineární s rychlostí (nejméně) $\left(\frac{A-a}{A+a}\right)^2$.

Tedy i v případě nekvadratické funkce hraje velkou roli podmíněnost matice $\nabla^2 f(x^*)$. Např. pro Rosenbrockovu funkci je $A/a = 2504$.

Pozor v případě nekvadratických funkcí:

- samotná spojitá diferencovatelnost funkce f ale nestačí pro konvergenci MNS, uvažte např. $f(x_1, x_2) = x_1^3/3 + x_2^2/2$ s výchozím bodem $x^{[0]}$ splňujícím $x_1^{[0]} > 1$;
- a ani v případě konvergence MNS nemusí platit, že v limitním bodě nastává globální/lokální minimum, uvažte stejný příklad s volbou $0 < x_1^{[0]} < 1$.

Při vhodném oslabení požadavků předchozí věty může zachovat alespoň lokální konvergenci MNS.

MNS pro nekvadratické funkce — lokální konvergence

Nechť $f : \mathbb{R}^n \to \mathbb{R}$ je spojitě diferencovatelná. Jestliže $x^{[0]} \in \mathbb{R}^n$ je takové, že množina

$$\left\{x \in \mathbb{R}^n \mid f(x) \leqslant f(x^{[0]})\right\}$$

je ohraničená, pak posloupnost $\{x^{[k]}\}$ generovaná MNS konverguje k bodu x^* s $\operatorname{grad} f(x^*) = 0$.

Souhrn o MNS

- globální konvergence (pro nekvadratické funkce za dodatečných předpokladů)
- konvergence je velmi (velmi) pomalá (obvykle)
- mnohdy numericky ani nekonverguje
- řada metod ale využívá MNS ve chvíli, kdy samy neposkytují dostatečné zlepšení (alespoň v jednom kroku)
- princip je základem pro mnoho velmi efektivních metod

Problémy s rychlostí konvergence lze částečně vyřešit např. pomocí *metody paralelních tečen*:

> pro dané $x^{[0]}$ určíme $x^{[1]}$ a $x^{[2]}$ pomocí MNS, ale $x^{[3]}$ získáme pomocí $h_2 = x^{[2]} - x^{[0]}$ a odpovídajícího α_2. Body $x^{[4]}$ a $x^{[5]}$ opět pomocí MNS ...

Toto v případě kvadratické funkce vede k tzv. *metodě sdružených gradientů*, kterou si ukážeme později.

Základní popis NM

Jedná se o metodu druhého řádu, tj. požadujeme $f \in C^2$ na $\mathbb{R}^n$.

Hlavní myšlenka: v $(k+1)$-ním kroku ($k \in \mathbb{N} \cup \{0\}$) funkci f aproximujeme Taylorovým polynomem druhého řádu se středem v bodě $x^{[k]}$ a za $x^{[k+1]}$ zvolíme bod, ve kterém nabývá tento polynom nabývá svého minima.

To znamená, že místo funkce f uvažujeme

$$T_k(x) := f(x^{[k]}) + \operatorname{grad}^\top f(x^{[k]})\,(x - x^{[k]}) + \frac{1}{2}(x - x^{[k]})^\top \nabla^2 f(x^{[k]})\,(x - x^{[k]}) (\approx f(x))$$

Jelikož hledáme řešení úlohy $T_k(x) \to \min$, zderivováním $T_k(x)$ dostaneme

$$\operatorname{grad} T_k(x) = \operatorname{grad} f(x^{[k]}) + \nabla^2 f(x^{[k]})\,(x - x^{[k]}),$$

což v případě regulární matice $\nabla^2 f(x^{[k]})$ vede díky nutnosti $\operatorname{grad} T_k(x) = 0$ k bodu

$$x^{[k+1]} = x^{[k]} - \left[\nabla^2 f(x^{[k]})\right]^{-1} \operatorname{grad} f(x^{[k]}). \tag{3.2.3}$$

První pozorování: v případě, kdy funkce f je dokonce kvadratická, je aproximace Taylorovým polynomem přesná, tj. $f(x) = T_k(x)$. V takovém případě nalezneme řešení úlohy (3.2.1) pomocí (3.2.3) právě v JEDNOM kroku („konvergence je superlineární s řádem ∞“).

Absence regulárnosti matice $\nabla^2 f(x^{[k]})$ celou situaci velmi zkomplikuje, takže tento požadavek nevypustíme. Bez dalších požadavků získáme tvrzení o lokální konvergenci NM.

Věta 3.2.5

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

Pozor: Věta 3.2.5 platí dokonce v případě $\det f(x^*) \neq 0$, takže je-li např. x^* lokální maximum, pak při dostatečně blízké počáteční aproximaci $x^{[0]} \in \mathbb{R}^n$, bude posloupnost $\{x^{[k]}\}$ generovaná NM konvergovat k tomuto maximu x^*.

Co znamená *dostatečně blízko* ve Větě 3.2.5? Z praktického pohledu vlastně nemáme vůbec žádnou informaci, jak zvolit $x^{[0]}$ — ani z důkazu tuto informaci nezískáme, neboť bychom museli znát x^*. Nicméně při zesílení požadavků na funkci f dostaneme následující tvrzení zpřesňující „oblast lokální konvergence“. (Všimněte si, že za uvedených předpokladů bude bod $x^{[k+1]}$ dokonce globálním minimem $T_k(x)$.)

Věta 3.2.6

Nechť $f \in C^2$ je silně konvexní s konstantou silné konvexnosti $\vartheta > 0$ a funkce $\nabla^2 f(x)$ lipschitzovsky spojitá na $\mathbb{R}^n$ s konstantou M, tj.

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leqslant M\|x - y\| \quad \text{pro všechna} \quad x, y \in \mathbb{R}^n.$$

Je-li počáteční aproximace $x^{[0]}$ taková, že

$$\|\operatorname{grad} f(x^{[0]})\| < \frac{8\vartheta^2}{M},$$

pak posloupnost $\{x^{[k]}\}$ generovaná NM konverguje k x^* (což je jediné minimum funkce f na $\mathbb{R}^n$, tj. řešení (3.2.1)) superlineárně s (alespoň) řádem $p = 2$, tj.

$$\|x^{[k]} - x^*\| \leqslant \frac{4\vartheta}{M}\beta^{2^k}, \quad \text{kde} \quad \beta = \frac{M\|\operatorname{grad} f(x^{[0]})\|}{8\vartheta^2} \in (0, 1).$$

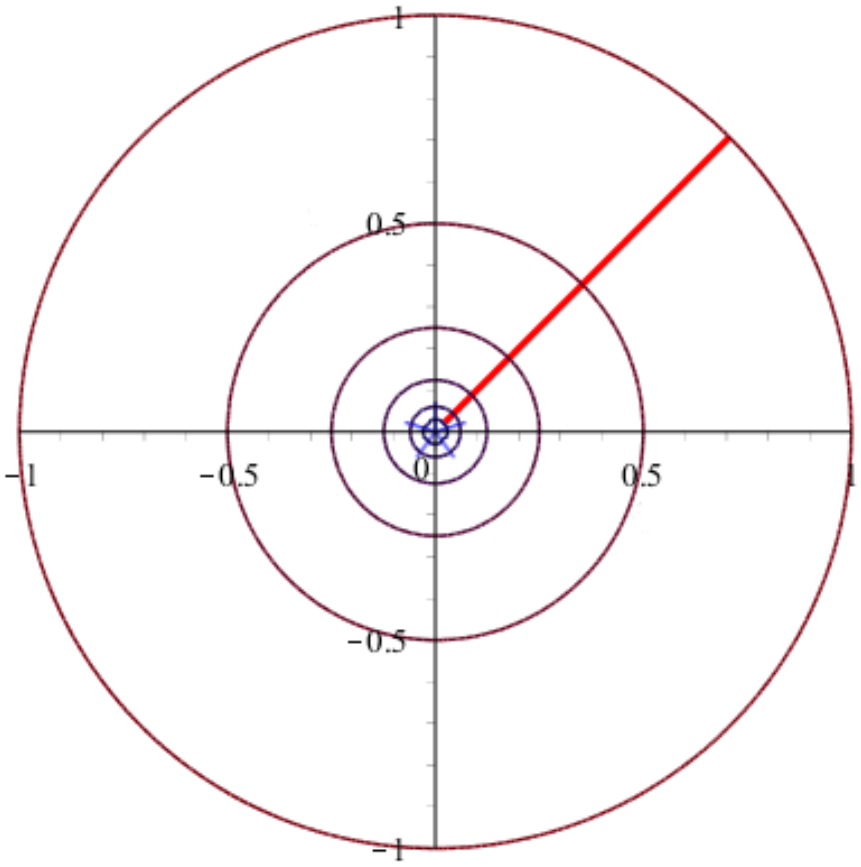
Příklad

Pomocí NM určeme první iteraci $x^{[1]}$ pro řešení úlohy

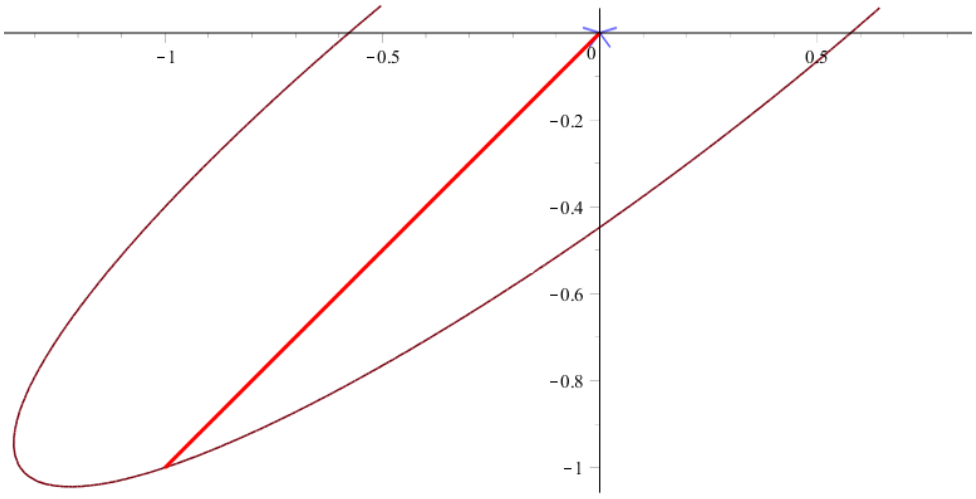
$$f(x_1, x_2) = (x_1^2 + x_2^2)^{3/2} \to \min$$

při počáteční aproximaci $x^{[0]} = [1/\sqrt{2}, 1/\sqrt{2}]$.

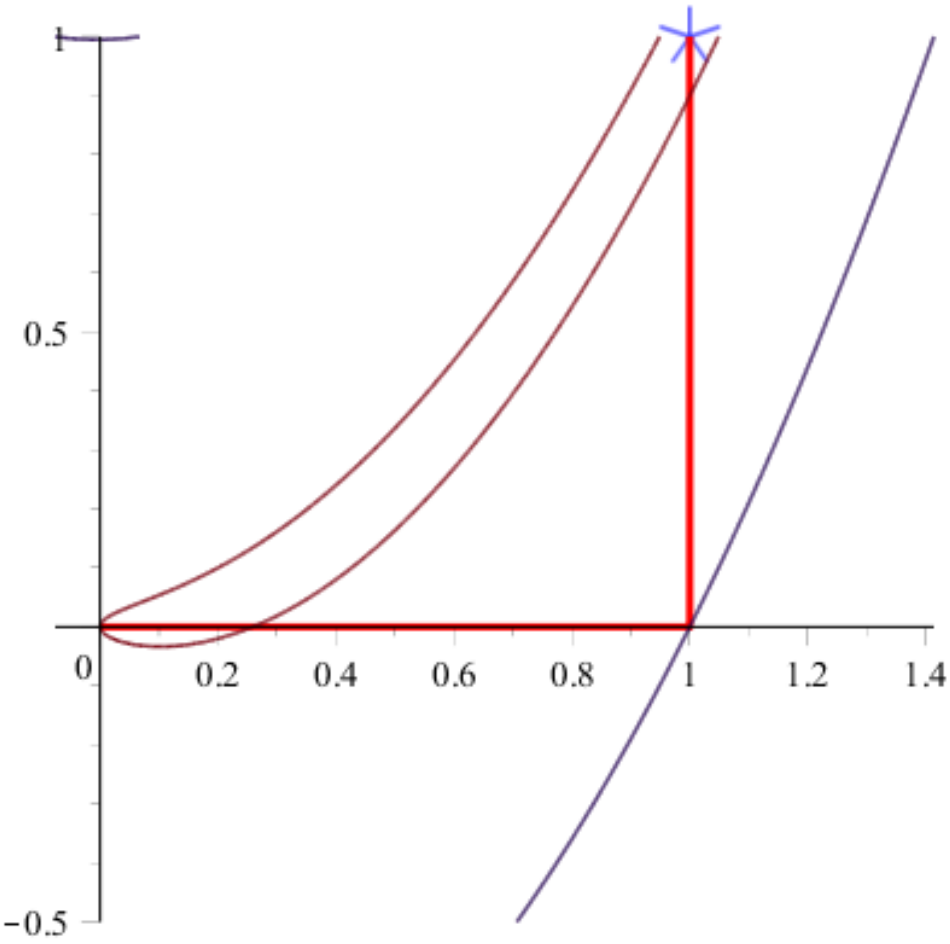
NM pro $f(x_1, x_2) = (x_1^2 + x_2^2)^{3/2}$ s $x^{[0]} = [1/\sqrt{2}, 1/\sqrt{2}]$ (viz také MNS)



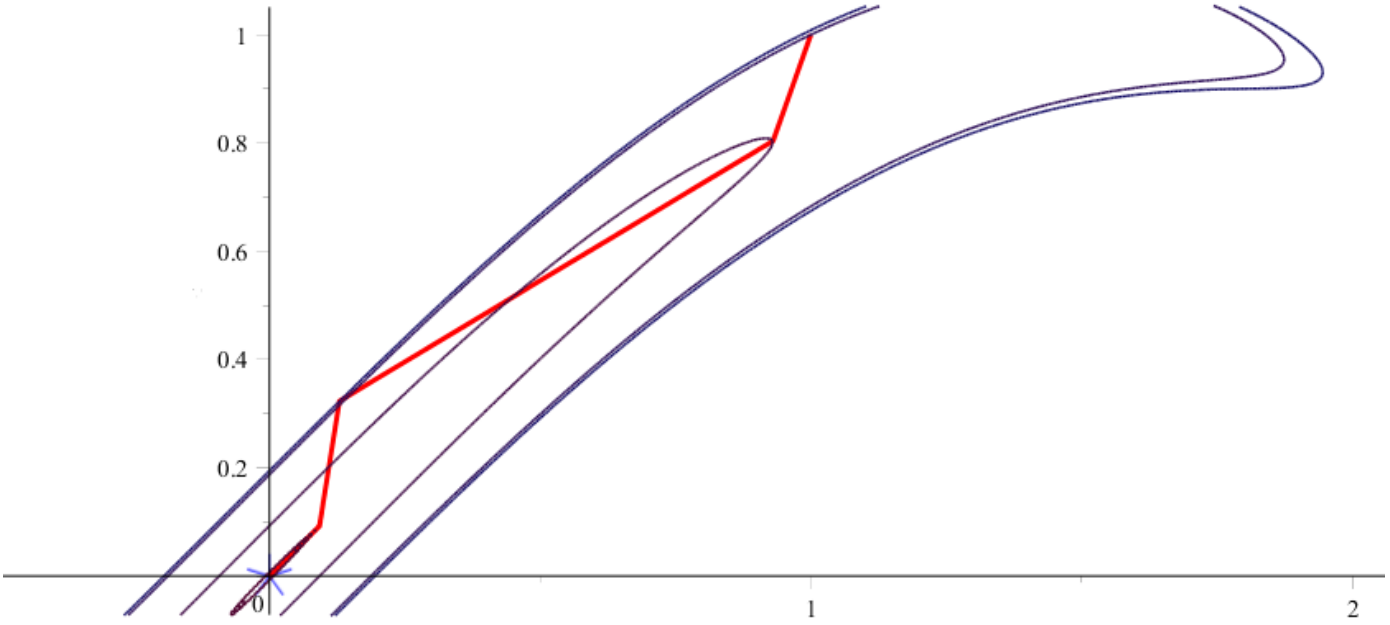
NM pro $f(x_1, x_2) = 3x_1^2 - 7x_1x_2 + 5x_2^2$ s $x^{[0]} = [-1, -1]$



NM pro Rosenbrockovu funkci $f(x_1, x_2) = (1 - x_1)^2 + 100(x_2 - x_1^2)^2$ s $x^{[0]} = [0, 0]$



NM pro funkci $f(x_1, x_2) = 10(x_2 - \sin x_1)^2 + x_1^2/10$ s $x^{[0]} = [1, 1]$



Souhrn o NM

- velmi rychlá konvergence;
- nutnost dostatečně blízké počáteční aproximace (existují silně konvexní funkce, pro které je při nevhodně zvoleném $x^{[0]}$ posloupnost $\{f(x^{[k]})\}$ neklesající, např. $f(x_1, x_2) = -1/(x_1^2 + x_2^2 + 1)$ s počáteční aproximací $x^{[0]} = [a, 0]$ pro $a > 0$ dostatečně velké nebo viz příklad níže);
- velmi vysoká početní náročnost pro výpočet $[\nabla^2 f(x^{[k]})]^{-1}$ při velkých hodnotách n (řádově jde o n^3 aritmetických operací) a zatížení tohoto výpočtu numerickými chybami.

Poslední dva nedostatky odstraňují (alespoň částečně) různé modifikace NM.

Příklad

Pomocí NM určeme první iteraci $x^{[1]}$ pro řešení úlohy

$$f(x_1, x_2) = \sqrt{1 + x_1^2 + x_2^2} \to \min$$

při počáteční aproximaci $x^{[0]} = [1/2, 1/2]$. A co při volbě $x^{[0]} = [1, 1]$? Kdy se „pokazí“ konvergence?

Modifikace NM

NM můžeme také chápat jako iterační proces podobný MNS, tj. $x^{[k+1]} = x^{[k]} + \alpha_k h_k$, kde tentokrát

$$h_k = -\left[\nabla^2 f(x^{[k]})\right]^{-1} \operatorname{grad} f(x^{[k]}) \quad \text{a} \quad \alpha_k = 1$$

První přirozenou modifikací této metody pak je změna volby α_k jako

$$\alpha_k = \operatorname{argmin}_{\alpha \geqslant 0}\{f(x^{[k]} - \alpha h_k)\},$$

čímž dostaneme tzv. spádovou NM. Samozřejmě můžeme α_k volit i jiným vhodným způsobem např. popsaným na začátku.

Jenže při těchto modifikacích NM můžeme přijít o tu největší výhodu NM — o kvadratickou konvergenci. Navíc stále ještě musíme počítat $[\nabla^2 f(x^{[k]})]^{-1}$.

Poslední zmíněný nedostatek odstraňují tzv. kvazinewtonovské metody, kdy matice $[\nabla^2 f(x^{[k]})]^{-1}$ je nahrazena jistou maticí A_k, tj.

$$x^{[k+1]} = x^{[k]} - \alpha_k A_k g_k,$$

kde $A_k \in \mathbb{R}^{n \times n}$, $\alpha_k > 0$ a $g_k = \operatorname{grad} f(x^{[k]})$. Volbou $A_k = I$ dostaneme MNS a volba $A_k = [\nabla^2 f(x^{[k]})]^{-1}$ dává NM. Toto nás přivádí k následující metodě, kterou lze považovat za rozumný kompromis mezi MNS a NM.

Cesta k MSG

Vraťme se opět k situaci, kdy v úloze (3.2.1) máme pouze funkci

$$f(x) = \frac{1}{2}x^\top Qx - b^\top x, \tag{$\triangle$}$$

přičemž $Q = Q^\top \in \mathbb{R}^{n\times n}$, $Q > 0$ a $b \in \mathbb{R}^n$.

Potom nalezení řešení úlohy (3.2.1) & ($\triangle$) je ekvivalentní s řešením soustavy

$$Qx = b. \tag{$*$}$$

Úlohu ($*$) umíme řešit např. Gaussovou eliminací. Ovšem v roce 1952 publikovali Hestenes a Stiefel přesnou metodu pro řešení nehomogenní soustavy rovnic s pozitivně definitní maticí jakožto alternativu ke známé Gaussově eliminaci. Tato metoda nalezne řešení $n \times n$ soustavy v nejvýše n krocích. Tuto metodu pojmenovali the conjugate gradient method. Nicméně tato metoda zůstala delší dobu nepovšimnuta, neboť vyžadovala $2n^3 + \mathcal{O}(n^2)$ aritmetických operací, zatímco Gaussova metoda pouze $n^3/3 + \mathcal{O}(n^2)$ operací pro $n \to \infty$. Tuto metodu lze použít i v případě pozitivně semidefinitní matice, ovšem nemusí nás přivést k výsledku (pozn.: v roce 1945 trval na „automatickém počítači“ výpočet inverze matice 10×10 přibližně 15 člověko-dnů, v roce 1949 to údajně bylo /z dnešního pohledu stále nekonečných/ 8 hodin).

Ačkoli se jedná o přímou metodu (řešení po n krocích), lze ji chápat také jako iterativní proces s velmi rychlou konvergencí v případě pozitivně definitní matice. Toho si všimli o 15 let později a MSG jakožto metodu pro řešení velkých nelineárních optimalizačních problému publikovali Fletcher a Reeves. Později byla tato metoda revidována a upravena pro úlohu (3.2.1) & ($\triangle$), což si nyní představíme.

Klíčovým pojmem celé metody je tzv. Q-sdruženost vektorů.

Definice 3.2.7

Nechť $Q = Q^\top \in \mathbb{R}^{n\times n}$ je pozitivně definitní. Vektory $h_1, h_2 \in \mathbb{R}^n \setminus \{0\}$ se nazývají Q-*sdružené* (též Q-ortogonální), jestliže

$$\langle Qh_1,\, h_2\rangle = h_1^\top Q h_2 = 0.$$

Systém vektorů $\{h_0, \ldots, h_{m-1}\}$ v $\mathbb{R}^n \setminus \{0\}$ pro $m \in \{2, \ldots, n\}$ se nazývá Q-*sdružený*, jestliže

$$\langle Qh_i,\, h_j\rangle = 0 \quad \text{pro} \quad i \neq j.$$

Věta 3.2.8

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

Následující tvrzení ukazuje, jak je užitečné (= důležité) míti systém Q-sdružených vektorů.

Věta 3.2.9

Nechť $m \in \{2, \ldots, n\}$ a mějme systém $\{h_0, \ldots, h_{m-1}\}$ Q-sdružených vektorů v $\mathbb{R}^n$. Nechť dále $x^{[0]} \in \mathbb{R}^n$ je dáno a body $x^{[1]}, \ldots, x^{[m]}$ jsou dány jako

$$x^{[k+1]} = x^{[k]} + \alpha_k h_k = x^{[0]} + \sum_{i=0}^{k} \alpha_i h_i, \quad k \in \{0, \ldots, m-1\}, \tag{3.2.8}$$

kde α_k jsou volena tak, že $f(x^{[k]} + \alpha_k h_k) = \min_{\alpha\in\mathbb{R}} f(x^{[k]} + \alpha h_k)$ pro $k \in \{0, \ldots, m-1\}$. Pak pro kvadratickou funkci f definovanou v ($\triangle$) platí

$$f(x^{[m]}) = \min_{x\in X_m} f(x),$$

kde $X_m := x^{[0]} + \mathrm{Lin}\{h_0, \ldots, h_{m-1}\}$. Zejména pro $m = n$ dostáváme

$$f(x^{[n]}) = \min_{x\in\mathbb{R}^n} f(x),$$

tj. $x^{[n]}$ je řešením úlohy (3.2.1) & ($\triangle$).

Vzhledem k volbě funkce f lze snadno odvodit explicitní předpis pro délku k-tého kroku. Chceme totiž

$$\frac{1}{2}(x^{[k]} + \alpha h_k)^\top Q(x^{[k]} + \alpha h_k) - b^\top(x^{[k]} + \alpha h_k) \to \min,$$

což po zderivování vzhledem k α dává

$$\alpha_k = -\frac{h_k^\top \operatorname{grad} f(x^{[k]})}{h_k^\top Q h_k}. \tag{3.2.9}$$

Metoda popsaná ve Větě 3.2.9 se nazývá MSG a pro její použití zbývá vhodně zvolit vektory $h_0, \ldots, h_{n-1}$. Z LAaG známe Gramův–Schmidtův ortogonalizační proces, což je vlastně určení Q-sdružených vektorů pro $Q = I$. Tento postup můžeme zobecnit na libovolné $Q > 0$: pro systém LNZ vektorů $u_0, \ldots, u_{n-1}$ položíme

$$h_0 := u_0 \quad \& \quad h_i := u_i + \sum_{j=0}^{i-1} \beta_{ij} h_j,$$

přičemž h_1 konstruujeme tak, aby byl Q-sdružený s h_0; h_2 tak, aby byl Q-sdružený s h_0 a h_1 atd. Jinými slovy, v i-tém kroku vezmeme u_i a odečteme ty „složky“, které nejsou Q-sdružené k $h_0, \ldots, h_{i-1}$. To nás přivádí k volbě

$$\beta_{ij} = -\frac{u_i^\top Q h_j}{h_j^\top Q h_j}, \quad i > j.$$

Nevýhodou tohoto procesu je fakt, že si musíme „pamatovat“ všechny předchozí vektory pro konstrukci dalšího a celkově potřebujeme $\mathcal{O}(n^3)$ aritmetických operací.

My zvolíme vektory $u_k := -\operatorname{grad} f(x^{[k]})$ pro $k \in \{0, \ldots, n-1\}$, čímž dostaneme

$$h_0 := -\operatorname{grad} f(x^{[0]}), \quad h_k := -\operatorname{grad} f(x^{[k]}) + \beta_{k-1}\, h_{k-1}, \tag{3.2.10}$$

$$\beta_{k-1} := \frac{\operatorname{grad}^\top f(x^{[k]})\, Q h_{k-1}}{h_{k-1}^\top Q h_{k-1}}, \tag{3.2.11}$$

přičemž body $x^{[k]}$ jsou počítány dle (3.2.8). Vlastně volíme $\beta_{ij} = \beta_{k,k-1}$, což zaručuje, že vektory h_{k-1} a h_k jsou Q-sdružené. Je to skutečně dobrá volba?

Věta 3.2.10

Nechť $x^{[0]} \in \mathbb{R}^n$ je libovolný a $x^{[1]}, \ldots, x^{[n-1]}$, $h_0, \ldots, h_{n-1}$ jsou určeny vztahy (3.2.8), (3.2.9), (3.2.10) a (3.2.11). Potom systém vektorů $\{h_0, \ldots, h_{n-1}\}$ je Q-sdružený a $\operatorname{grad} f(x^{[0]}), \ldots, \operatorname{grad} f(x^{[n-1]})$ jsou ortogonální.

Navíc místo (3.2.9), (3.2.10) a (3.2.11) můžeme pro $g_k := \operatorname{grad} f(x^{[k]}) = Qx^{[k]} - b$ brát

$$\left.\begin{aligned} \alpha_k &= \frac{g_k^\top g_k}{h_k^\top Q h_k}, \qquad \beta_{k-1} = \frac{g_k^\top g_k}{g_{k-1}^\top g_{k-1}}, \\ h_0 &= -g_0, \qquad h_k = -g_k + \beta_{k-1} h_{k-1}. \end{aligned}\right\} \tag{3.2.12}$$

Ovšem pozor: v MNS je možné při řešení příkladů „zkracovat“ vektory h_k, abychom si zjednodušili výpočty (to se pouze projeví změnou α_k). Toto je možné také pro MSG, ale pouze při použití vzorců (3.2.9) a (3.2.10). Při použití (3.2.12) to již fungovat nebude! Nicméně pomocí obou přístupů musí vyjít tytéž hodnoty $x^{[1]}, x^{[2]}, \ldots$

Příklad

Pomocí MSG určeme první iteraci $x^{[1]}$ a směr h_1 další iterace pro řešení úlohy

$$f(x_1, x_2) = x_1^2 + x_1 x_2 + x_2^2$$

při počáteční aproximaci $x^{[0]} = [0, 1]$.

Numerické vlastnosti MSG (i)

Z teoretického pohledu musí MSG pro úlohu (3.2.1) & (△) najít řešení v nejvýše n krocích (dle Věty 3.2.9). V praxi (při numerickém výpočtu) však dochází k zaokrouhlovacím chybám a může se stát, že toto řešení nenalezneme. Nalezneme pouze nějakou jeho aproximaci, přičemž „kvalita“ této aproximace závisí právě na zaokrouhlovacích chybách, a tedy zejména na podmíněnosti matice Q vyjádřené podílem $\lambda_{\max}/\lambda_{\min}$.

V praktických implementacích MSG se tato metoda používá v cyklech délky n, tj. po n krocích se celý proces restartuje a za výchozí bod nového cyklu se bere $x^{[n]}$, tj.

$$\beta_{k-1} = \begin{cases} \beta_{k-1}, & k \neq n, 2n, 3n, \ldots, \\ 0, & k = n, 2n, \ldots, \end{cases}$$

($n \rightsquigarrow$ Fletcher & Reeves). Je také možné zvolit nějaké ukončovací pravidlo, zejména např. $g_k^\top g_k = \|\operatorname{grad} f(x^{[k]})\| < \varepsilon$.

Numerické vlastnosti MSG (ii)

Hlavní výhodou MSG je její *snadná implementace* (podobně jako pro MNS). Početní náročnost jednoho kroku je

$$\mathcal{O}(n) + [\text{počet aritmetických operací pro násobení } h \rightsquigarrow Qh].$$

To znamená, že pro řídkou matici Q (tj. počet nenulových prvků $N \ll n^2$) bude počet operací malý (pro hustou matici je počet aritmetických operací při násobení matice–vektor roven $(2n-1)n$, což v případě řídké matice bude menší než $2N \ll (2n-1)n$). V praxi se mnohdy jedná o systémy, které mají velmi řídké matice (počet nenulových prvků je asi 0,01 %–1 %) nebo jsou součinem dvou řídkých matic (což je početně stejně náročné jako kdybychom měli jednu řídkou matici). Je-li řád matice opravdu velký (desítky tisíc, např. v tomografii se jedná o 10^5 až 10^6) a není zde žádná „řídkost“, tak v podstatě není způsob, jak soustavu $Qx = b$ vyřešit (v řídkém případě lze použít Gaussovu eliminaci s vhodnou modifikací pro zachování řídkosti). V takovém případě tedy nelze použít přímé metody lineární algebry a nezbývá nám nic jiného než skutečně použít nějaký iterační proces, jako např. MNS nebo MSG. Nicméně stále je aritmetická náročnost relativně malá (závislost je polynomiální) – totéž platí i pro MNS, ale MSG má v jistém smyslu nejlepší řád konvergence mezi všemi iteračními procesy založenými na maticovém násobení. Nevýhodou je citlivost na podmíněnost matice Q.

Počet kroků MSG

Má-li matice Q pouze r různých vlastních hodnot, potom MSG nalezne řešení v nejvýše r krocích. Navíc platí

$$\|x^{[k+1]} - x^*\|_Q^2 \leqslant \left(\frac{\lambda_{n-k} - \lambda_1}{\lambda_{n-k} + \lambda_1}\right)^2 \|x^{[0]} - x^*\|_Q^2.$$

Toto můžeme využít pro předpověď chování MSG: uvažme matici Q s m velkými vlastními hodnotami a zbývajícími $n - m$ menšími „nahromaděnými“ kolem 1, potom dostaneme

$$\|x^{[m+1]} - x^*\|_Q^2 \approx \varepsilon \|x^{[0]} - x^*\|_Q^2, \quad \varepsilon \coloneqq \lambda_{n-m} - \lambda_1,$$

tedy pro velmi malé ε, můžeme říci, že $x^{[m+1]}$ získané po pouhých $m+1$ krocích MSG je velmi dobrou aproximací x^*.

MSG lze aplikovat také v případě $Q \geqslant 0$.

Věta 3.2.11

Nechť matice $Q \geqslant 0$ s hodností $r \coloneqq \operatorname{rank} Q \leqslant n$. Pak MSG buď zajistí nalezení minima funkce f dané jako v (△) v nejvýše r krocích (v takovém případě pro nějaké $k \in \{0, \dots, r\}$ nastane $\operatorname{grad} f(x^{[k]}) = 0$ a $h_0, \dots, h_{k-1} \neq 0$) nebo signalizuje, že toto minimum neexistuje (v takovém případě $h_k^\top Q h_k = 0$, přičemž $\operatorname{grad} f(x^{[0]}), \dots, \operatorname{grad} f(x^{[k-1]}) \neq 0$ pro nějaké $k \in \{0, \dots, r\}$).

MSG pro nekvadratické funkce

MSG lze aplikovat také pro nekvadratické funkce. Algoritmus funguje stejně jako pro kvadratické funkce pouze se liší volbou β_k, přičemž obvykle dochází k restartu MSG po n krocích. Existuje několik přístupů, zejména (Fletcher–Reeves) β_{k-1}^{FR} stejně jako v (3.2.12) pro $g_k := \operatorname{grad} f(x^{[k]})$ nebo (Polak–Ribiére)

$$\beta_{k-1}^{PR} := \frac{(g_k - g_{k-1})^\top g_k}{g_{k-1}^\top g_{k-1}}$$

nebo (Hestenes–Stiefel)

$$\beta_{k-1}^{HS\#1} := \frac{(g_k - g_{k-1})^\top g_k}{h_{k-1}^\top (g_k - g_{k-1})}, \quad \beta_{k-1}^{HS\#2} := \frac{(g_k - g_{k-1})^\top g_k}{g_{k-1}^\top (g_k - g_{k-1})}.$$

Dokonce lze najít strategii

$$\beta_k^{FR-PR} = \begin{cases} -\beta_k^{FR}, & \beta_k^{PR} < -\beta_k^{FR}, \\ \beta_k^{PR}, & |\beta_k^{PR}| \leqslant \beta_k^{FR}, \\ \beta_k^{FR}, & \beta_k^{PR} > \beta_k^{FR}. \end{cases}$$

Obecně platí: různé volby β_k jsou výhodné v různých situacích. Např. je známo, že volba β_k^{PR} dává mnohem lepší výsledky než β_k^{FR} v některých případech – ale již ne v jiných. Existují dokonce případy, kdy g_k jsou odraženy od nuly v případech volby β_k^{PR} $\rightsquigarrow$ v tomto článku je na základě analýzy globální konvergence MSG upřednostňována volba β_k^{FR}. Autor téhož článku ale také navrhuje volbu $\beta_k = \max\{0, \beta_k^{PR}\}$.

Následující dvě tvrzení dávají informace o „oblasti“ lokální konvergence a o řádu konvergence posloupnosti výsledků MSG s cykly délky n (ovšem nikoli o konvergenci samotné MSG).

Věta 3.2.12

Nechť $f \in C^1$ na $\mathbb{R}^n$ a $x^{[0]} \in \mathbb{R}^n$ je takové, že množina

$$\left\{x \in \mathbb{R}^n \mid f(x) \leqslant f(x^{[0]})\right\}$$

je ohraničená. Nechť $x^{[1]}$ je výsledek MSG s β^{FR} nebo β^{PR} po n-krocích (tj. cyklu délky n), $x^{[2]}$ výsledek dalšího cyklu s výchozím bodem $x^{[1]}$ atd. Potom posloupnost $\{x^{[k]}\}$ je ohraničená a její hromadné body jsou stacionárními body funkce f, tj. $\lim_{j\to\infty} \operatorname{grad} f(x^{[j]}) = 0$ pro každou podposloupnost $\{x^{[j]}\} \subseteq \{x^{[k]}\}$.

A jak je to s rychlostí konvergence?

Věta 3.2.13

Nechť $f \in C^3$ na $\mathbb{R}^n$, $x^{[0]} \in \mathbb{R}^n$ a x^* je nedegenerované lokální minimum, tj. $\operatorname{grad} f(x^*) = 0$ a $\nabla^2 f(x^*) > 0$. Nechť $x^{[k]}$ je výsledek MSG s cyklem délky n a výchozím bode $x^{[k-1]}$ a nechť $x^{[k]} \to x^*$ pro $k \to \infty$. Potom posloupnost $\{x^{[k]}\}$ konverguje k x^* superlineárně s řádem alespoň $p = 2$ (tj. kvadraticky), tj.

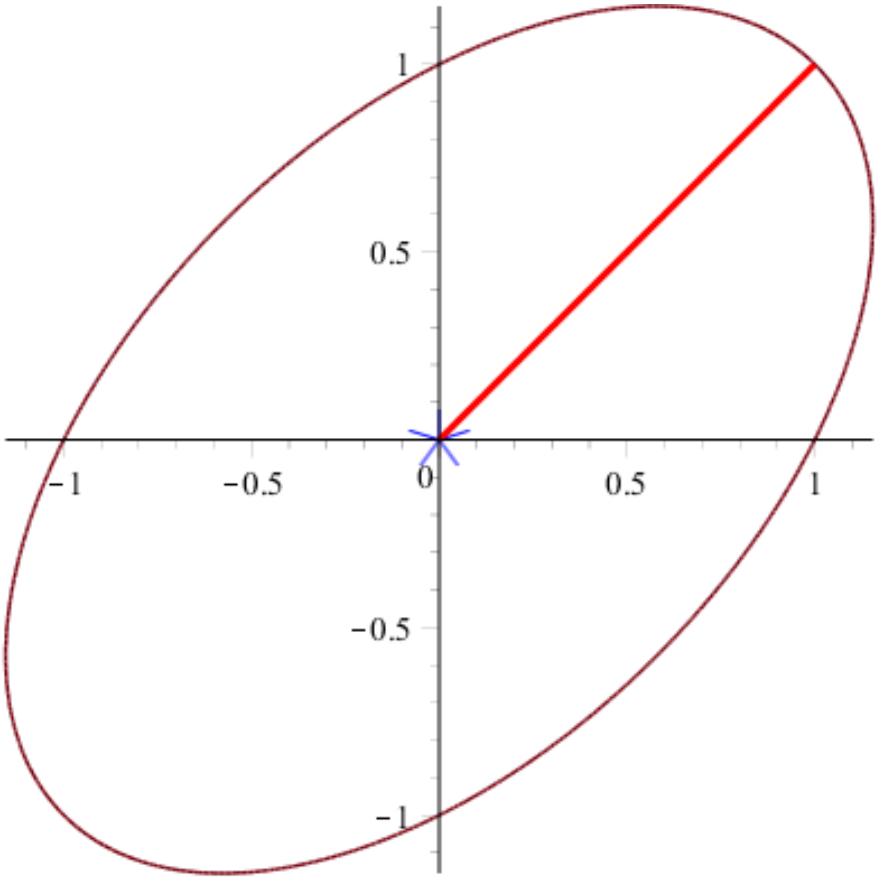
$$\|x^{[k+1]} - x^*\| \leqslant C\|x^{[k]} - x^*\|^2$$

pro nějaké $C < \infty$ a všechna $k \in \mathbb{N} \cup \{0\}$.

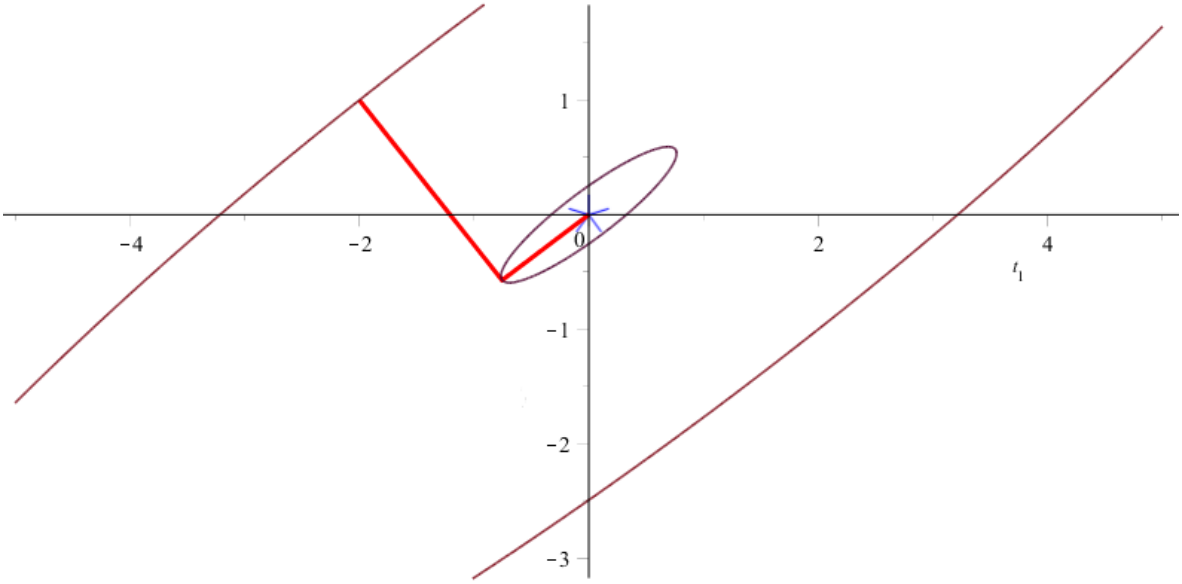
Naší původní motivací bylo nalezení metody, která je lepší než MNS ale méně náročná než NM. Můžeme říci, že MSG toto splňuje, neboť předchozí věta ukazuje, že výsledek jednoho cyklu délky n MSG „udělá stejný pokrok“ jako 1 krok NM.

MSG úzce souvisí s tzv. metodami *Krylovových podprostorů*, které byly zařazena mezi 10 algoritmů, které nejvíce ovlivnili rozvoj a užití vědy a inženýrství ve 20. století.

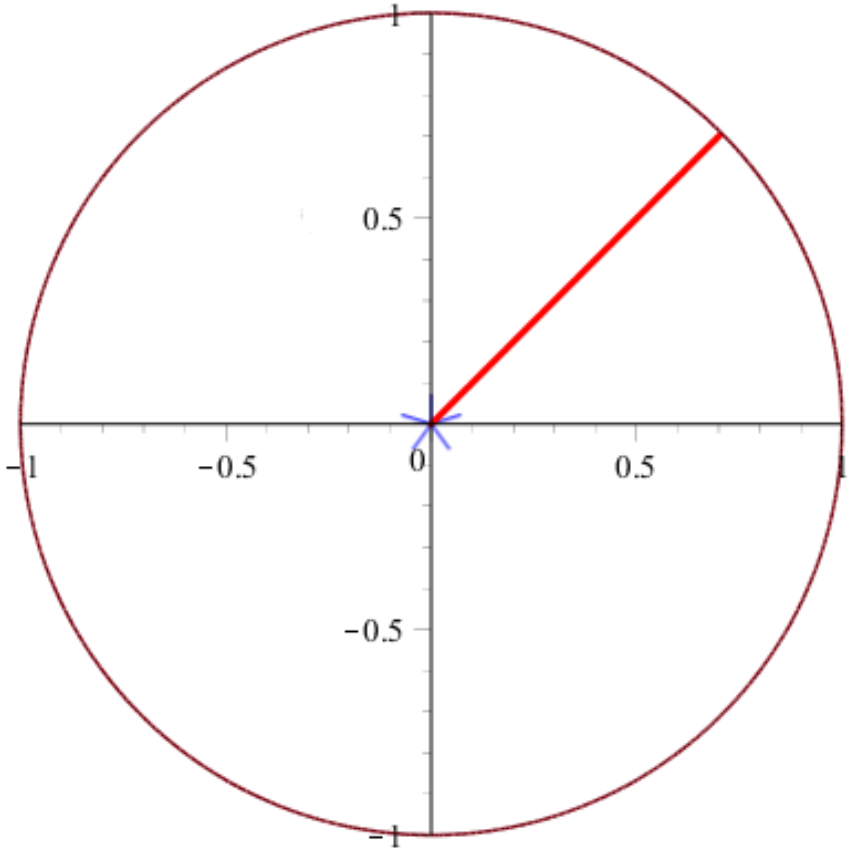
MSG pro funkci $f(x_1, x_2) = x_1^2 - x_1x_2 + x_2^2$ s $x^{[0]} = [1, 1]$



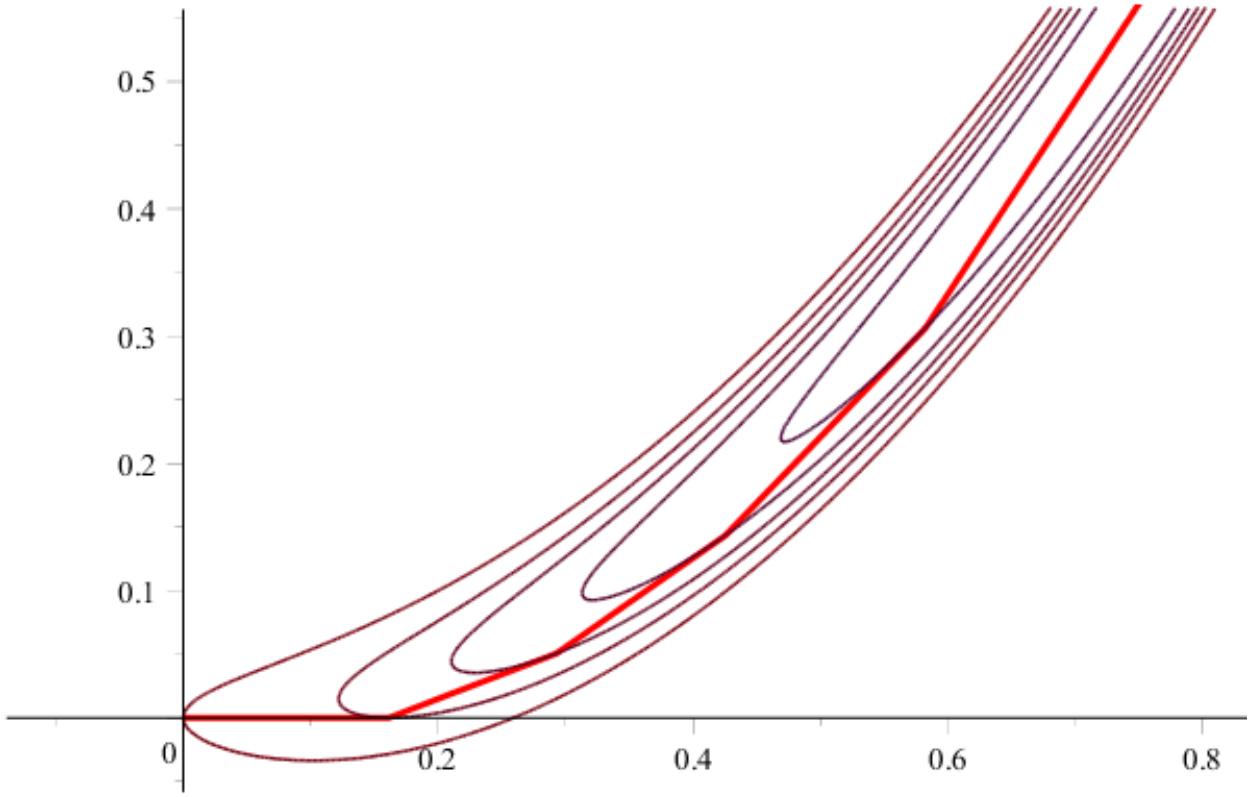
MSG pro funkci $f(x_1, x_2) = 3x_1^2 - 7x_1x_2 + 5x_2^2$ s $x^{[0]} = [-2, 1]$



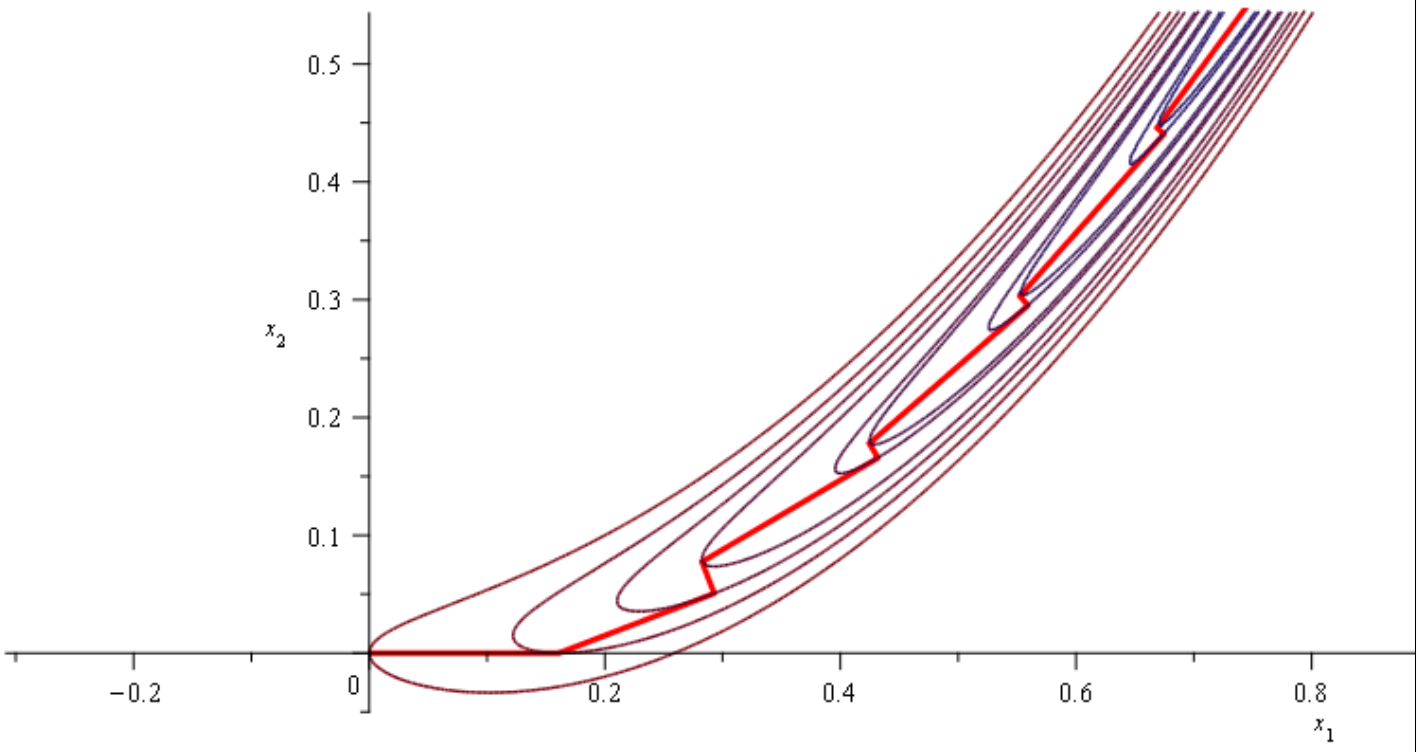
MSG pro funkci $f(x_1, x_2) = (x_1^2 + x_2^2)^{3/2}$ s $x^{[0]} = [1/\sqrt{2}, 1/\sqrt{2}]$



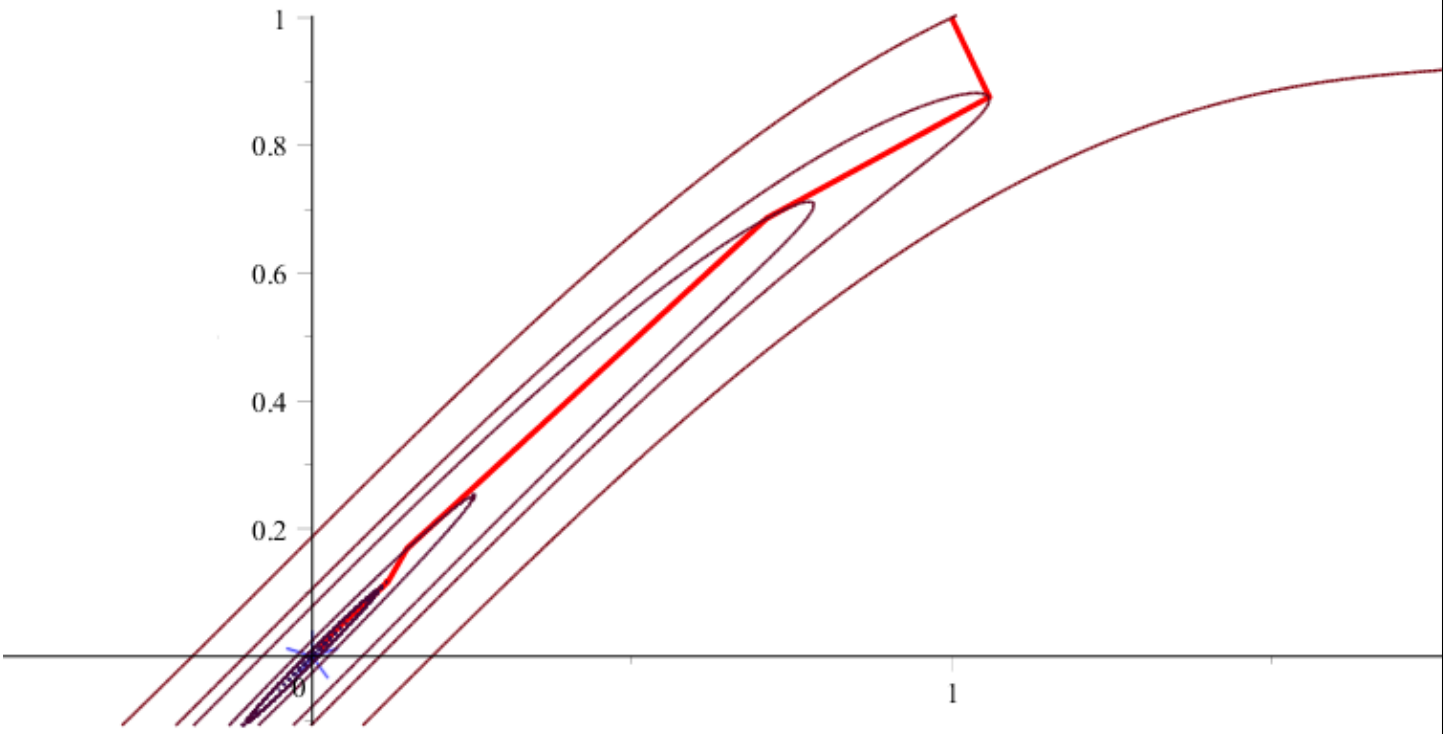
MSG pro Rosenbrockovu funkci $f(x_1, x_2) = (1 - x_1)^2 + 100(x_2 - x_1^2)^2$ s $x^{[0]} = [0, 0]$ a bez resetu β



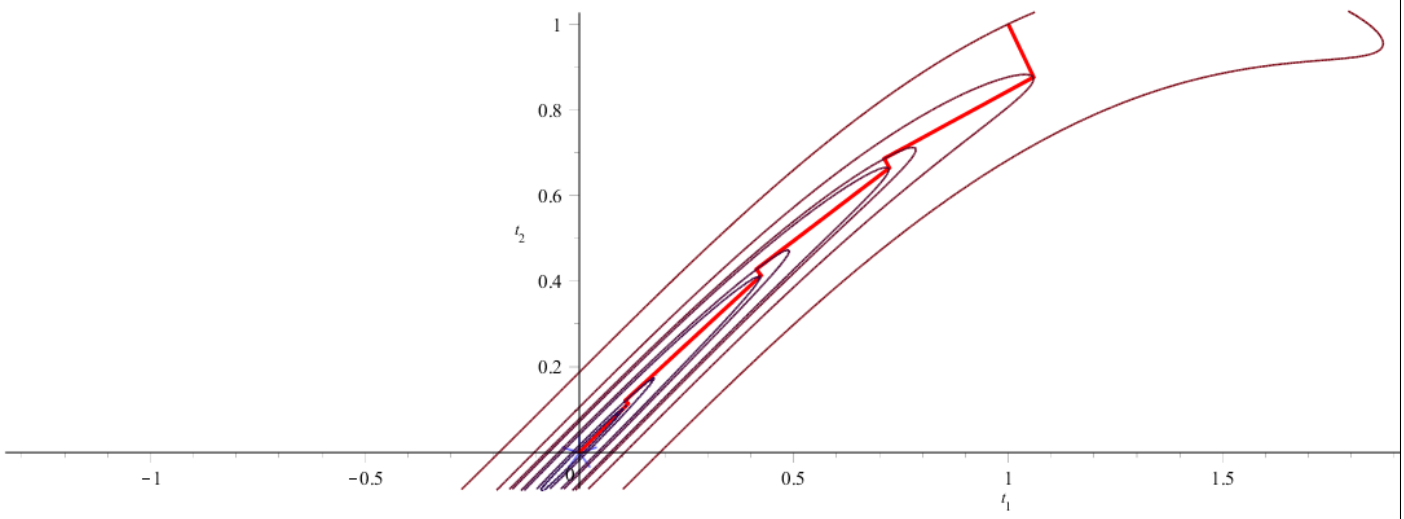
MSG pro Rosenbrockovu funkci $f(x_1, x_2) = (1 - x_1)^2 + 100(x_2 - x_1^2)^2$ s $x^{[0]} = [0, 0]$ a resetem β



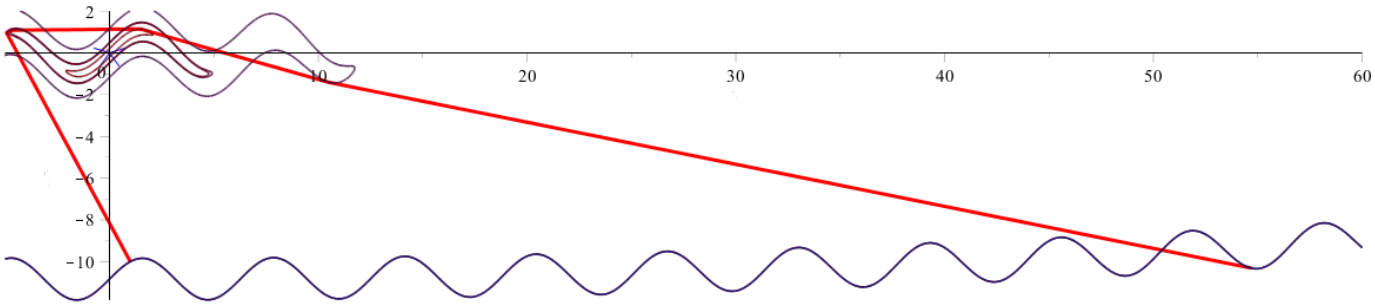
MSG pro funkci $f(x_1, x_2) = 10(x_2 - \sin x_1)^2 + x_1^2/10$ s $x^{[0]} = [1, 1]$ a bez resetu β



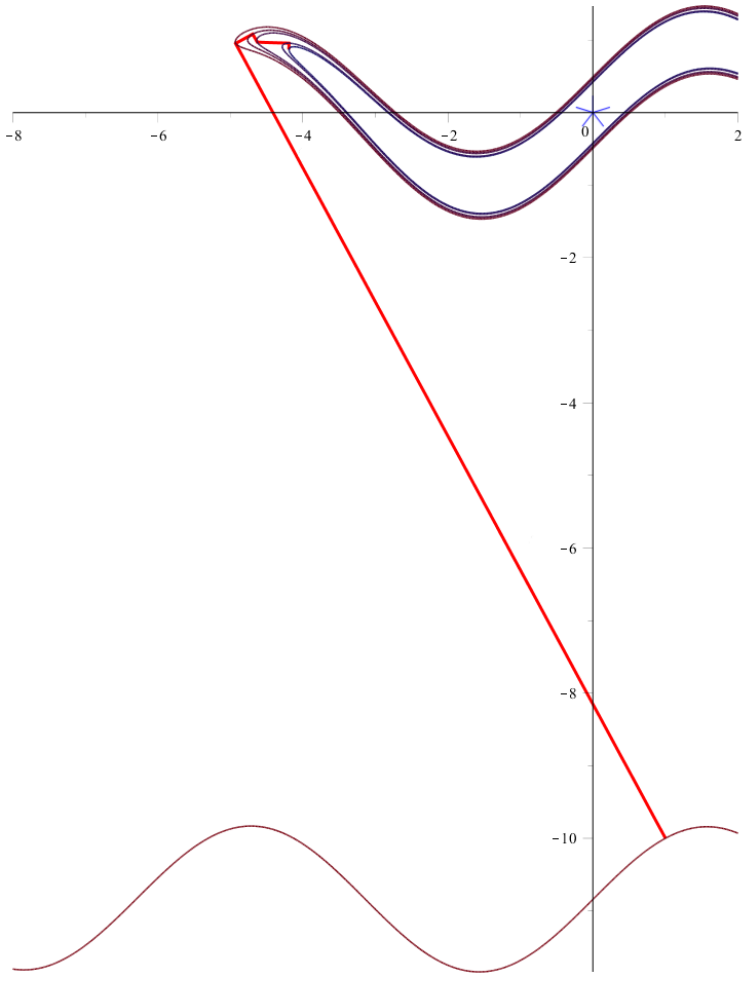
MSG pro funkci $f(x_1, x_2) = 10(x_2 - \sin x_1)^2 + x_1^2/10$ s $x^{[0]} = [1, 1]$ a resetem β



MSG pro funkci $f(x_1, x_2) = 10(x_2 - \sin x_1)^2 + x_1^2/10$ s $x^{[0]} = [1, -10]$ a bez resetu β



MSG pro funkci $f(x_1, x_2) = 10(x_2 - \sin x_1)^2 + x_1^2/10$ s $x^{[0]} = [1, -10]$ a resetem β



Konec.