TEHNICI DE OPTIMIZARE
Curs 3

Andrei Pătraşcu

Departament Informatică
Universitatea din Bucureşti

Definim generic un algoritm iterativ: iniţializăm $x^0 \in \mathbb{R}^n$ şi iterăm

$$x^{k+1} = T(x^k, x^{k-1}, \cdots, x^0)$$

până când criteriul de oprire ales este satisfăcut.

- un algoritm iterativ de optimizare primeşte informaţii precum: punctul de iniţializare x^0, funcţia obiectiv f (şi alte informaţii legate de f), acurateţea ϵ dorită, etc.

Definim generic un algoritm iterativ: iniţializăm $x^0 \in \mathbb{R}^n$ şi iterăm

$$x^{k+1} = T(x^k, x^{k-1}, \cdots, x^0)$$

până când criteriul de oprire ales este satisfăcut.

- un algoritm iterativ de optimizare primeşte informaţii precum: punctul de iniţializare x^0, funcţia obiectiv f (şi alte informaţii legate de f), acurateţea ϵ dorită, etc.
- datele se folosesc pentru a executa iteraţia T, care defineşte un set de operaţii asupra întregului istoric $\{x^0, \cdots, x^k\}$.

Definim generic un algoritm iterativ: iniţializăm $x^0 \in \mathbb{R}^n$ şi iterăm

$$x^{k+1} = T(x^k, x^{k-1}, \cdots, x^0)$$

până când criteriul de oprire ales este satisfăcut.

- un algoritm iterativ de optimizare primeşte informaţii precum: punctul de iniţializare x^0, funcţia obiectiv f (şi alte informaţii legate de f), acurateţea ϵ dorită, etc.
- datele se folosesc pentru a executa iteraţia T, care defineşte un set de operaţii asupra întregului istoric $\{x^0, \cdots, x^k\}$.
- deoarece nu se va executa un număr infinit de iteraţii, orice algoritm iterativ va răspunde la întrebarea: "Când poate fi considerat x^k o aproximare suficient de precisă a unui punct de optim?"

Algorithm 1: Algoritm de ordin I ($x^0, \epsilon, \ldots$):

Data: $k := 0$

while criteriu oprire = fals **do**

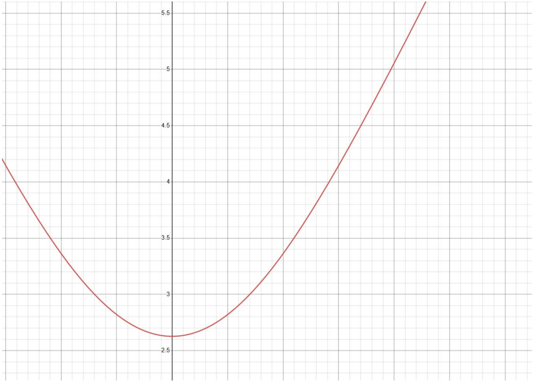
  Calculează: $d^k \in \mathrm{span}\{\nabla f(x^k), \nabla f(x^{k-1}), \cdots, \nabla f(x^0)\}$

  Actualizează x^{k+1} pe baza d^k şi $\{x^k, x^{k-1}, \cdots, x^0\}$ $k := k + 1$

end

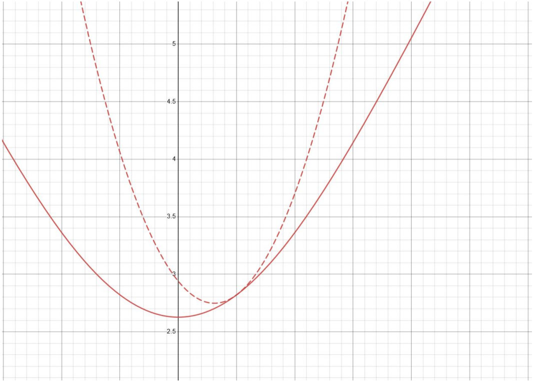
Majoritatea algoritmilor de ordin I realizează un model aproximativ local (pătratic) al funcţiei obiectiv $\mathcal{A}(x; x^k, \cdots, x^0; f)$. Aplicarea lui T reprezintă este echivalentă cu determinarea soluţiei optimă a acestuia, i.e

$$\min_x \ \mathcal{A}(x; x^k, \cdots, x^0; f).$$



Majoritatea algoritmilor de ordin I realizează un model aproximativ local (pătratic) al funcţiei obiectiv $\mathcal{A}(x; x^k, \cdots, x^0; f)$. Aplicarea lui T reprezintă este echivalentă cu determinarea soluţiei optimă a acestuia, i.e

$$\min_x \; \mathcal{A}(x; x^k, \cdots, x^0; f).$$



$$\min_{x \in \mathbb{R}^n} \; f(x)$$

Dacă f este diferenţiabilă şi $\nabla f(x) \neq 0$. Atunci:

$$\begin{aligned} f(x - \tau \nabla f(x)) &= f(x) - \tau \|\nabla f(x)\|^2 + o(\tau \nabla f(x)) \\ &= f(x) - \tau \left(\|\nabla f(x)\|^2 + \frac{1}{\tau} o(\tau) \right) < f(x), \end{aligned}$$

pentru $\tau > 0$ suficient de mic, prin definiţia lui $o(\tau)$. Ultima inegalitate intră în contradicţie cu presupunerea că x^* este punct de minim.

Aproximarea pătratică în x^k

$$f(x) \approx \mathcal{A}(x; x^k; f) := f(x^k) + \nabla f(x^k)^T (x - x^k) + \frac{1}{2}(x - x^k)^T H_k (x - x^k)$$

Alegerea matricii Hessiane H_k determină calitatea aproximării!
Pentru alegerea $H_k = \alpha_k I_n (\alpha_k > 0)$, modelul se simplifică :

$$f(x) \approx \mathcal{A}(x; x^k; f) := f(x^k) + \nabla f(x^k)^T (x - x^k) + \frac{1}{2\alpha_k} \|x - x^k\|^2$$

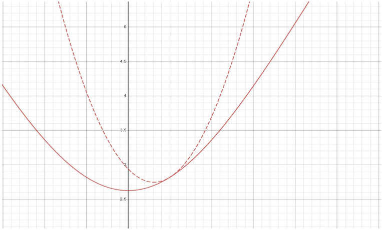


Figure: $f(x) = \ln\left(1 + e^{-2x+1}\right) + \ln\left(1 + e^{2x+1}\right)$; $x^0 = \frac{1}{2}$, $\alpha_0 = 1/4$

Considerăm:

$$x^{k+1} = \arg\min_x f(x^k) + \nabla f(x^k)^T(x - x^k) + \frac{1}{2\alpha_k}\|x - x^k\|^2$$

Din condiţiile de ordin I avem:

$$\begin{aligned}
\nabla f(x^k) + \frac{1}{\alpha_k}(x^{k+1} - x^k) &= 0 \\
\frac{1}{\alpha_k}(x^{k+1} - x^k) &= -\nabla f(x^k) \\
x^{k+1} &= x^k - \alpha_k \nabla f(x^k)
\end{aligned}$$

Intuim următorul algoritm: iniţializăm x^0

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

unde $\alpha_k \geq 0$ se numeşte **lungimea pasului** iteraţiei.

Metoda Gradient a fost introdusă în 1847 de către Auguste Cauchy pentru rezolvarea unui sistem neliniar cu 6 necunoscute:

A. Cauchy, *Méthode générale pour la résolution des systèmes d'équations simultanées.* C.R. Acad. Sci. Paris, 25: 536-538, 1847.

Algorithm 2: Metoda Gradient ($x^0, \epsilon, \{\alpha_k\}_{k\geq 0}$):

Data: $k := 0$

while criteriu oprire = fals **do**

  Calculează: $\nabla f(x^k)$

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

  $k := k + 1$

end

- pas constant: $\alpha_k = \alpha$
- cea mai abruptă pantă: $\alpha_k = \arg\min_\alpha f(x^k - \alpha \nabla f(x^k))$
- adaptiv

Evaluarea calităţii unei iteraţii se poate realiza în mai multe moduri (pe baza acurateţii $\epsilon > 0$):

- $\|x^k - x^*\| \leq \epsilon \;\Leftrightarrow x^k \in B(x^*; \epsilon)$

Evaluarea calităţii unei iteraţii se poate realiza în mai multe moduri (pe baza acurateţii $\epsilon > 0$):

- $\|x^k - x^*\| \leq \epsilon \;\Leftrightarrow x^k \in B(x^*; \epsilon)$
- $f(x^k) - f^* \leq \epsilon \;\Leftrightarrow x^k \in S_f(f^* + \epsilon)$

Evaluarea calităţii unei iteraţii se poate realiza în mai multe moduri (pe baza acurateţii $\epsilon > 0$):

- $\|x^k - x^*\| \leq \epsilon \;\Leftrightarrow x^k \in B(x^*; \epsilon)$
- $f(x^k) - f^* \leq \epsilon \;\Leftrightarrow x^k \in S_f(f^* + \epsilon)$
- $\|\nabla f(x)\| \leq \epsilon$

Teoremă (Polyak)

Fie f diferenţiabilă pe $\mathbb{R}^n$ *cu gradientul* ∇f *continuu Lipschitz:*

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|.$$

De asemenea, presupunem $\min_x f(x) > -\infty$ *şi* $0 < \alpha < \frac{1}{2L}$. *Atunci şirul generat de Metoda Gradient* $x^{k+1} = x^k - \alpha \nabla f(x^k)$ *satisface:*

$$\lim_{k \to \infty} \nabla f(x^k) = 0,$$

şi $f(x^{k+1}) \leq f(x^k)$.

Demonstraţie pe scurt: Pentru simplitate $\alpha_k = \frac{1}{L}$. Din continuitatea Lipschitz avem

$$f(x^{k+1}) \leq f(x^k) + \nabla f(x^k)^T(x^{k+1} - x^k) + \frac{L}{2}\|x^{k+1} - x^k\|^2$$
$$\leq f(x^k) - \frac{1}{2L}\|\nabla f(x^k)\|^2.$$

Este evidentă descreşterea $f(x^{k+1}) \leq f(x^k)$. Trecând termenul normei în partea stângă avem:

$$\frac{1}{2L}\|\nabla f(x^k)\|^2 \leq f(x^k) - f(x^{k+1}) \quad \forall k \geq 0$$
$$\frac{1}{2L}\sum_{i=0}^{k-1}\|\nabla f(x^i)\|^2 \leq \sum_{i=0}^{k-1} f(x^i) - f(x^{i+1}) = f(x^0) - f(x^{k+1})$$
$$\leq f(x^0) - f^*.$$

Prin trecerea la limită $k \to \infty$ obţinem rezultatul.

Consecinţe şi limite

Teorema (Polyak). Fie f diferenţiabilă pe $\mathbb{R}^n$ cu gradientul ∇f continuu Lipschitz: $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$. De asemenea, presupunem $\min_x f(x) > -\infty$ şi $0 < \alpha < \frac{1}{2L}$. Atunci şirul generat de Metoda Gradient $x^{k+1} = x^k - \alpha \nabla f(x^k)$ satisface:

$$\lim_{k \to \infty} \nabla f(x^k) = 0,$$

şi $f(x^{k+1}) \leq f(x^k)$.

- Nu este necesară convexitatea (in acest caz, MG converge la un punct staţionar)

Teorema (Polyak). Fie f diferenţiabilă pe $\mathbb{R}^n$ cu gradientul ∇f continuu Lipschitz: $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$. De asemenea, presupunem $\min_x f(x) > -\infty$ şi $0 < \alpha < \frac{1}{2L}$. Atunci şirul generat de Metoda Gradient $x^{k+1} = x^k - \alpha \nabla f(x^k)$ satisface:

$$\lim_{k \to \infty} \nabla f(x^k) = 0,$$

şi $f(x^{k+1}) \leq f(x^k)$.

- Nu este necesară convexitatea (in acest caz, MG converge la un punct staţionar)
- Este necesară continuitatea Lipschitz (exemplu!) şi o aproximare a constantei L

Consecinţe şi limite

Teorema (Polyak). Fie f diferenţiabilă pe $\mathbb{R}^n$ cu gradientul ∇f continuu Lipschitz: $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$. De asemenea, presupunem $\min_x f(x) > -\infty$ şi $0 < \alpha < \frac{1}{2L}$. Atunci şirul generat de Metoda Gradient $x^{k+1} = x^k - \alpha \nabla f(x^k)$ satisface:

$$\lim_{k \to \infty} \nabla f(x^k) = 0,$$

şi $f(x^{k+1}) \leq f(x^k)$.

- Nu este necesară convexitatea (in acest caz, MG converge la un punct staţionar)
- Este necesară continuitatea Lipschitz (exemplu!) şi o aproximare a constantei L
- Fără o alegere limitată a pasului, MG poate diverge (exemplu!)

Consecinţe şi limite

Teorema (Polyak). Fie f diferenţiabilă pe $\mathbb{R}^n$ cu gradientul ∇f continuu Lipschitz: $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$. De asemenea, presupunem $\min_x f(x) > -\infty$ şi $0 < \alpha < \frac{1}{2L}$. Atunci şirul generat de Metoda Gradient $x^{k+1} = x^k - \alpha \nabla f(x^k)$ satisface:

$$\lim_{k \to \infty} \nabla f(x^k) = 0,$$

şi $f(x^{k+1}) \leq f(x^k)$.

- Nu este necesară convexitatea (in acest caz, MG converge la un punct staţionar)
- Este necesară continuitatea Lipschitz (exemplu!) şi o aproximare a constantei L
- Fără o alegere limitată a pasului, MG poate diverge (exemplu!)
- Când pasul este variabil, inegalitatea descreşterii devine:

$$f(x^{k+1}) \leq f(x^k) - \alpha_k \left(1 - \frac{L\alpha_k}{2}\right) \|\nabla f(x^k)\|^2.$$

Teorema (Polyak). Fie f diferenţiabilă pe $\mathbb{R}^n$ cu gradientul ∇f continuu Lipschitz: $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$. De asemenea, presupunem $\min_x f(x) > -\infty$ şi $0 < \alpha < \frac{1}{2L}$. Atunci şirul generat de Metoda Gradient $x^{k+1} = x^k - \alpha \nabla f(x^k)$ satisface:

$$\lim_{k \to \infty} \nabla f(x^k) = 0,$$

şi $f(x^{k+1}) \leq f(x^k)$.

- În cazul $S_f(f(x^0))$ mărginită, avem în plus convergenţa unui subşir al x^k la un punct staţionar al lui f (contrar, $f(x) = \frac{1}{1+\|x\|^2}$)

Consecinţe şi limite

Teorema (Polyak). Fie f diferenţiabilă pe $\mathbb{R}^n$ cu gradientul ∇f continuu Lipschitz: $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$. De asemenea, presupunem $\min_x f(x) > -\infty$ şi $0 < \alpha < \frac{1}{2L}$. Atunci şirul generat de Metoda Gradient $x^{k+1} = x^k - \alpha \nabla f(x^k)$ satisface:

$$\lim_{k \to \infty} \nabla f(x^k) = 0,$$

şi $f(x^{k+1}) \leq f(x^k)$.

- În cazul $S_f(f(x^0))$ mărginită, avem în plus convergenţa unui subşir al x^k la un punct staţionar al lui f (contrar, $f(x) = \frac{1}{1+\|x\|^2}$)
- Garanţii de convergenţă către un minim local/global nu există!

Consecinţe şi limite

Teorema (Polyak). Fie f diferenţiabilă pe $\mathbb{R}^n$ cu gradientul ∇f continuu Lipschitz: $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$. De asemenea, presupunem $\min_x f(x) > -\infty$ şi $0 < \alpha < \frac{1}{2L}$. Atunci şirul generat de Metoda Gradient $x^{k+1} = x^k - \alpha \nabla f(x^k)$ satisface:

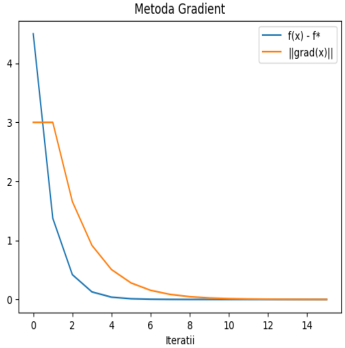
$$\lim_{k \to \infty} \nabla f(x^k) = 0,$$

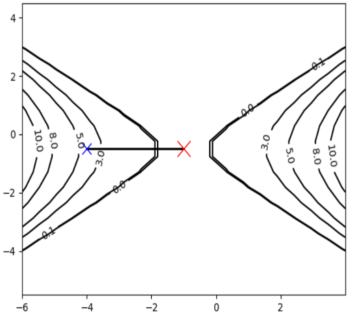
şi $f(x^{k+1}) \leq f(x^k)$.

- În cazul $S_f(f(x^0))$ mărginită, avem în plus convergenţa unui subşir al x^k la un punct staţionar al lui f (contrar, $f(x) = \frac{1}{1+\|x\|^2}$)
- Garanţii de convergenţă către un minim local/global nu există!
- Rata de convergenţă MG, în general, poate fi foarte pesimistă, e.g. pentru $f(x) = \frac{1}{x}$, cu $x \geq 1$, MG devine $x^{k+1} = x^k + \frac{1}{(x^k)^2}$, care implică $|f'(x^k)| = O(1/k^{2/3})$.

$$\min_{x \in \mathbb{R}^2} \; \frac{1}{2} x^T A x - b^T x$$

$$A = \begin{bmatrix} 1 & 0 \\ 0 & -2 \end{bmatrix}, \quad b = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$





Teoremă (Rată de convergenţă (convexitate))

Fie f convexă cu gradientul ∇f continuu Lipschitz:

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|.$$

De asemenea, presupunem $\min_x f(x) > -\infty$ *şi* $0 < \alpha < \frac{1}{2L}$. *Atunci şirul generat de Metoda Gradient* $x^{k+1} = x^k - \alpha \nabla f(x^k)$ *satisface:*

$$f(x^k) - f^* \leq \frac{L\|x^0 - x^*\|^2}{2k} \quad \forall k \geq 0.$$

Demonstraţie pe scurt: Pentru simplitate $\alpha_k = \frac{1}{L}$. Din continuitatea Lipschitz avem

$$f(x^{k+1}) \leq f(x^k) + \nabla f(x^k)^T (x^{k+1} - x^k) + \frac{L}{2}\|x^{k+1} - x^k\|^2$$
$$\leq f(x^k) - \frac{1}{2L}\|\nabla f(x^k)\|^2.$$

Este evidentă descreşterea $f(x^{k+1}) \leq f(x^k)$. Folosim următoarele observaţii:

(i) $x^k = x^0 - \sum_{i=1}^{k-1} \nabla f(x^i)$

(ii) $\frac{1}{2}\|\sum_i a^i\|^2 = \frac{1}{2}\sum_i \|a^i\|^2 + \sum_i (a^i)^T \left(\sum_{j=0}^{i-1} (a^j)^T\right)$

(iii) $\max_z \; z^T a - \frac{\alpha}{2}\|z\|^2 = \frac{1}{2\alpha}\|a\|^2$

Demonstraţie pe scurt: din continuitatea Lipschitz avem pentru $k \geq 0$

$$
\begin{aligned}
f(x^{k+1}) &\leq f(x^k) + \nabla f(x^k)^T(x^{k+1} - x^k) + \frac{L}{2}\|x^{k+1} - x^k\|^2 \\
&\leq f(x^k) - \frac{1}{2L}\|\nabla f(x^k)\|^2 \\
&\leq f(x^*) + \nabla f(x^k)^T(x^k - x^*) - \frac{1}{2L}\|\nabla f(x^k)\|^2 \\
&\overset{(i)}{=} f(x^*) + \nabla f(x^k)^T(x^0 - x^*) - \frac{1}{L}\nabla f(x^k)^T \left(\sum_{j=0}^{k-1} \nabla f(x^j)\right) - \frac{1}{2L}\|\nabla f(x^k)\|^2.
\end{aligned}
$$

Însumăm inegalităţile cu indecşii $i = 0, \cdots, k$.

Demonstraţie pe scurt: Însumăm inegalităţile cu indecşii $i = 0, \cdots, k$:

$$
\begin{aligned}
\sum_{i=1}^{k-1} f(x^{i+1}) - f(x^*) &\leq \left[\sum_{i=1}^{k-1} \nabla f(x^i)\right]^T (x^0 - x^*) \\
&\quad - \frac{1}{L}\sum_{i=1}^{k-1} \nabla f(x^i)^T \left(\sum_{j=0}^{i-1} \nabla f(x^j)\right) - \frac{1}{2L}\sum_{i=1}^{k-1} \|\nabla f(x^i)\|^2 \\
&\overset{(ii)}{=} \left[\sum_{i=1}^{k-1} \nabla f(x^i)\right]^T (x^0 - x^*) - \frac{1}{2L}\|\sum_{i=1}^{k-1} \nabla f(x^i)\|^2 \\
&\overset{(iii)}{\leq} \frac{L}{2}\|x^0 - x^*\|^2.
\end{aligned}
$$

În final, observând $\sum_{i=1}^{k-1} f(x^{i+1}) - f(x^*) \geq k(f(x^k) - f^*)$ obţinem rata de mai sus.

- Rolul ratei de convergenţă: determinarea complexităţii rezolvării (OfC) până la o precizie fixată, e.g.

$$f(x^k) - f^* \leq \mathcal{O}\left((C/k)\right) < \epsilon \;\Rightarrow$$

$$\text{dupa } k \geq \mathcal{O}\left(\frac{C}{\epsilon}\right) \text{ atingem } f(x) - f^* \leq \epsilon.$$

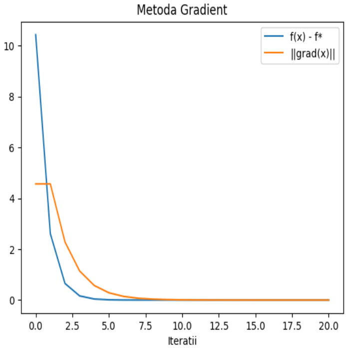
- Rolul ratei de convergenţă: determinarea complexităţii rezolvării (OfC) până la o precizie fixată, e.g.

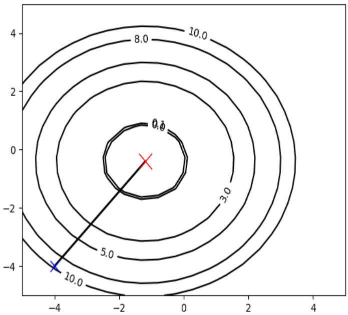
$$f(x^k) - f^* \leq \mathcal{O}\left((C/k)\right) < \epsilon \;\Rightarrow$$

$$\text{dupa } k \geq \mathcal{O}\left(\frac{C}{\epsilon}\right) \text{ atingem } f(x) - f^* \leq \epsilon.$$

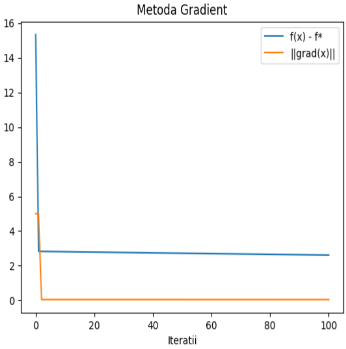
- Clase de rate de convergenţă:
  - subliniară: e.g $\mathcal{O}\left(C/k\right)$
  - liniară: e.g $\mathcal{O}\left(C \cdot (\frac{1}{2})^k\right)$
  - superliniară: e.g $\mathcal{O}\left(C \cdot (\frac{1}{2})^{k^2}\right)$
  - pătratică: e.g $\mathcal{O}\left(C \cdot (\frac{1}{2})^{2^k}\right)$

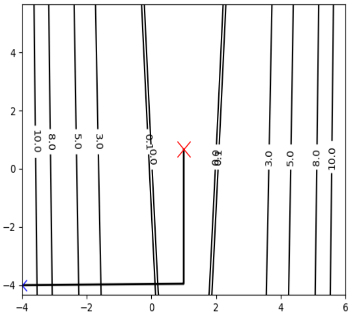
$$\min_{x \in \mathbb{R}^2} \frac{1}{2}\|Ax - b\|_2^2, \quad A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$





$$\min_{x \in \mathbb{R}^2} \frac{1}{2}\|Ax - b\|_2^2, \quad A = \begin{bmatrix} 1 & 0 \\ 0 & 0.02 \end{bmatrix}$$





Problema.

Fie funcţia $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x) = \frac{1}{2}\|Ax - b\|_2^2 + 2x^T x$, unde
$A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, b = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$.

a) Să se aducă la forma standard QP: $f(x) = \frac{1}{2}x^T Hx + q^T x + r$.

b) Fie $x^0 = [0 \;\; 1]^T$. Să se calculeze prima iteraţie a $MG(x^0, \epsilon)$ cu pas $\alpha_k = \frac{1}{L}$, unde L reprezintă constanta Lipschitz a $\nabla f(\cdot)$.

Teoremă (Rată de convergenţă (convexitate tare))

Fie f σ-tare convexă. Sub presupunerile teoremei precedente, şirul generat de Metoda Gradient $x^{k+1} = x^k - \frac{1}{L}\nabla f(x^k)$ satisface:

$$f(x^k) - f^* \leq \left(1 - \frac{\sigma}{L}\right)^k (f(x^0) - f^*)$$

$$\|x^k - x^*\|^2 \leq \left(1 - \frac{\sigma}{L}\right)^k \frac{f(x^0) - f^*}{\sigma} \quad \forall k \geq 0.$$

Observaţie: Nu este necesar ca MG să cunoască constanta σ!

Demonstraţie pe scurt: Considerăm $x^{k+1} = x^k - \alpha \nabla f(x^k)$. O remarcă importantă în cazul tare convex este:

$$f(x) - f^* \leq \frac{\sigma}{2} \|\nabla f(x)\|^2. \quad (1)$$

Din continuitatea Lipschitz avem

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) - \alpha \left(1 - \frac{L\alpha}{2}\right) \|\nabla f(x^k)\|^2. \\ &\leq f(x^k) - \sigma\alpha\,(2 - L\alpha)\,(f(x^k) - f^*). \end{aligned}$$

Scădem f^* în ambele părţi ale inegalităţii şi, pentru $\alpha = 1/L$, obţinem primul rezultat. Al doilea rezultă din relaţia de creştere pătratică a funcţiilor $\sigma-$tari convexe.

- Complexitate în cazul convergenţei liniare:

$$f(x^k) - f^* \leq q^k C < \epsilon \;\Rightarrow$$

$$\text{dupa } k \geq \frac{1}{\log(q^{-1})} \log \frac{C}{\epsilon} \text{ atingem } f(x^k) - f^* \leq \epsilon.$$

- Complexitate în cazul convergenţei liniare:

$$f(x^k) - f^* \leq q^k C < \epsilon \ \Rightarrow$$

$$\text{dupa } k \geq \frac{1}{\log(q^{-1})} \log \frac{C}{\epsilon} \text{ atingem } f(x^k) - f^* \leq \epsilon.$$

- Pentru $q = 1 - \sigma/L$:

$$\text{După } k \geq \frac{L}{\sigma} \log \frac{C}{\epsilon} > \frac{1}{\log(q^{-1})} \log \frac{C}{\epsilon} \text{ atingem } f(x^k) - f^* \leq \epsilon.$$

Putem estima performanţa MG utilizând numărul de condiţionare: $\kappa = \frac{L}{\sigma}$.

Problema.

2. Fie funcţia $f : \mathbb{R}^2 \to \mathbb{R}, f(x) = x_1^2 + 5x_2^2$ şi $x_0 = \begin{bmatrix} 5 \\ 1 \end{bmatrix}$.

a) Să se construiască $\phi(\alpha) = f(x^0 - \alpha\nabla f(x^0))$.

b) Aflaţi minimul funcţiei $\alpha^* = \arg\min_{\alpha>0} \ \phi(\alpha)$ şi generaţi o nouă iteraţie x^1 a MG cu pas α^*.

- Iteraţie simplă, bazată doar pe estimarea constantei L

- Iteraţie simplă, bazată doar pe estimarea constantei L
- În caz tare convex, fără a fi necesară estimarea lui σ, atinge ϵ acurateţe dupa $\mathcal{O}\left(\frac{L}{\sigma}\log\frac{1}{\epsilon}\right)$

- Iteraţie simplă, bazată doar pe estimarea constantei L
- În caz tare convex, fără a fi necesară estimarea lui σ, atinge ϵ acurateţe dupa $\mathcal{O}\left(\frac{L}{\sigma}\log\frac{1}{\epsilon}\right)$
- Se aplică în aceeaşi formă şi pentru funcţii neconvexe.

- **Metode de ordin I: Metode Multi-pas**
  - Metoda Heavy-Ball
  - Metoda Gradienţilor Conjugaţi
  - Metoda Gradientului Accelerat
- Metode de ordin II: Metoda Newton
- Metode de ordin I perturbate

Metode multi-pas = Algoritmi în care noua iteraţie se calculează pe baza unei fracţiuni a istoricului iteraţiilor precendente:

$$x^{k+1} = \Phi(x^k, x^{k-1}, \cdots, x^{k-s+1})$$

Metode multi-pas = Algoritmi în care noua iteraţie se calculează pe baza unei fracţiuni a istoricului iteraţiilor precendente:

$$x^{k+1} = \Phi(x^k, x^{k-1}, \cdots, x^{k-s+1})$$

Metoda Heavy-Ball:

$$x^{k+1} = x^k - \alpha\nabla f(x^k) + \beta(x^k - x^{k-1}),$$

unde $\alpha > 0, \beta \geq 0$ sunt parametri.

- iteraţia este inspirată din modelul de mişcare a unui corp sub forţa de frecare; datorită frecării, corpul pierde din energie şi atinge minimum-ul energiei potenţiale.
- termenul inerţial $\beta(x^k - x^{k-1})$ poate spori viteza de convergenţă

Metoda Heavy-Ball:

$$x^{k+1} = x^k - \alpha \nabla f(x^k) + \beta(x^k - x^{k-1}),$$

unde $\alpha > 0, \beta \geq 0$ sunt parametri.

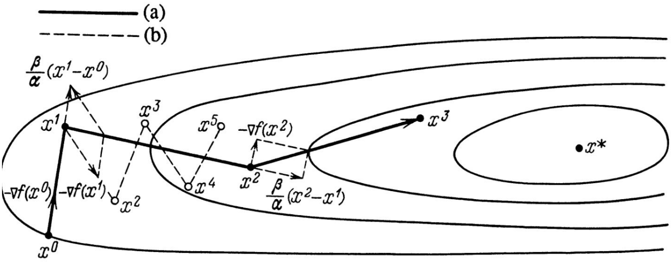


Fig. 6 (a) The heavy-ball method; (b) the gradient method.

Metoda Heavy-Ball:

$$x^{k+1} = x^k - \alpha \nabla f(x^k) + \beta(x^k - x^{k-1}),$$

unde $\alpha > 0, \beta \geq 0$ sunt parametri.

Teoremă (Polyak)

Fie x^* minim nesingular al lui f, atunci pentru $\beta \in [0, 1), \ \ \alpha \in \left(0, \frac{2(1+\beta)}{L}\right), \ \ \sigma I \preceq \nabla^2 f(x^*) \preceq LI$ există $\delta > 0$ astfel încât pentru oricare x^0, x^1 şi $\max\{\|x^0 - x^*\|, \|x^1 - x^*\|\} \leq \delta$ MHB converge cu rată liniară:

$$\|x^k - x^*\| \leq (q + \delta)^k c(\delta), \quad 0 \leq q < 1, \ \ 0 < \delta < 1 - q.$$

Factorul $q^* = \frac{\sqrt{L}-\sqrt{\sigma}}{\sqrt{L}+\sqrt{\sigma}}$ este minimal pentru $\alpha = \frac{4}{(\sqrt{L}+\sqrt{\sigma})^2}, \beta = \left(\frac{\sqrt{L}-\sqrt{\sigma}}{\sqrt{L}+\sqrt{\sigma}}\right)^2$.

- Pentru o convergenţă bună, este necesară estimarea constantelor σ, L
- Complexitate mai bună decât în cazul MG

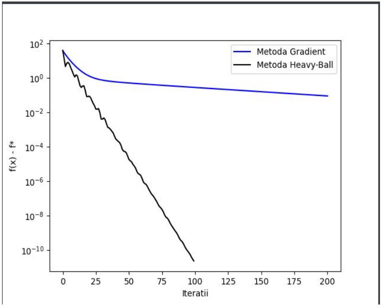
$$\text{MG}: x^{k+1} = x^k - \alpha \nabla f(x^k) \text{ complexitate } \quad \mathcal{O}\left(\frac{L}{\sigma}\log(1/\epsilon)\right)$$

vs.

$$\text{MHB}: x^{k+1} = x^k - \alpha \nabla f(x^k) + \beta(x^k - x^{k-1}) \text{ complexitate } \quad \mathcal{O}\left(\sqrt{\frac{L}{\sigma}}\log(1/\epsilon)\right)$$

MHB este de $\frac{\kappa \log(1/\epsilon)}{\sqrt{\kappa}\log(1/\epsilon)} = \sqrt{\kappa}$ ori mai rapidă!

- Pentru o convergenţă bună, este necesară estimarea constantelor σ, L
- Rata de convergenţă mai bună decât în cazul MG



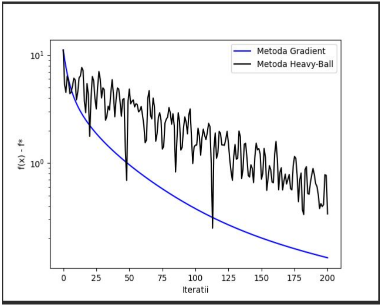


Figure: $\sigma \approx 0.5$(stânga) $\sigma \approx 10^{-3}$(dreapta)

- **Metode de ordin I: Metode Multi-pas**
  - Metoda Heavy-Ball
  - Metoda Gradienţilor Conjugaţi
  - Metoda Gradientului Accelerat
- Metode de ordin II: Metoda Newton
- Metode de ordin I perturbate

Fie $f(x) := \frac{1}{2}x^T Hx - q^T x$

Definiţie

Se numesc direcţii conjugate $\{p^1, \cdots, p^n\}$ *asociate matricii* $A \succ 0$ *dacă:*

$$(p^i)^T A p^k = 0 \qquad \forall i \neq k.$$

- Exemplu: vectorii proprii

Fie $f(x) := \frac{1}{2}x^T Hx - q^T x$

Definiţie

Se numesc direcţii conjugate $\{p^1, \cdots, p^n\}$ *asociate matricii* $A \succ 0$ *dacă:*

$$(p^i)^T A p^k = 0 \qquad \forall i \neq k.$$

- Exemplu: vectorii proprii
- Dacă H este matrice diagonală atunci rezolvăm problema prin n subprobleme 1D

$$\min_x \frac{1}{2}x^T Hx - q^T x = \min_x \frac{1}{2}\sum_{i=1}^n h_i x_i^2 - q_i x_i$$
$$= \sum_{i=1}^n \min_{x_i} \frac{1}{2}h_i x_i^2 - q_i x_i.$$

Fie $f(x) := \frac{1}{2}x^T Hx - q^T x$

Definiţie

Se numesc direcţii conjugate $\{p^1, \cdots, p^n\}$ *asociate matricii* $A \succ 0$ *dacă:*

$$(p^i)^T A p^k = 0 \qquad \forall i \neq k.$$

- Exemplu: vectorii proprii
- Dacă H este matrice diagonală atunci rezolvăm problema prin n subprobleme 1D

$$\begin{aligned}\min_x \ \frac{1}{2}x^T Hx - q^T x &= \min_x \ \frac{1}{2}\sum_{i=1}^n h_i x_i^2 - q_i x_i \\ &= \sum_{i=1}^n \min_{x_i} \ \frac{1}{2}h_i x_i^2 - q_i x_i.\end{aligned}$$

- Direcţiile $\{p^1, \cdots, p^n\}$ diagonalizează Hessiana

Dacă dispunem de $\{p^1, \cdots, p^n\}$, un algoritm de direcţii conjugate care rezolvă problema pătratică în n paşi este:

$$x^{k+1} = x^k + \alpha_k p^k$$
$$\alpha_k = \arg\min_{\alpha} f(x^k + \alpha p^k)$$

Algoritm direcţii conjugate: $x^{k+1} := x^k + \alpha_k p^k, \quad \alpha_k = \arg\min_\alpha f(x^k + \alpha p^k)$

Teoremă

Şirul $\{x^k\}$ *satisface*

$$(p^i)^T r^k = 0 \quad \forall i < k.$$

Mai mult, $x^k = \arg\min \; f(x)$ *s.l.* $x \in x^0 + span\{p^0, \cdots, p^{k-1}\}$

- Corolar: Şirul direcţiilor conjugate converge la x^* după cel mult n paşi
- Problema: cum calculăm $\{p^1, \cdots, p^n\}$? Pentru calculul p^k sunt necesari vectorii precedenţi?

Metoda Gradienţilor Conjugaţi (MGC):

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k) + \beta_k (x^k - x^{k-1})$$
$$(\alpha_k, \beta_k) = \arg\min_{\alpha,\beta} f(x^k - \alpha \nabla f(x^k) + \beta (x^k - x^{k-1}))$$

- Metoda Hestenes-Stiefel ('50). MGC propune alegera optimală a paşilor (α_k, β_k).
- Presupune rezolvarea unei probleme 2D
- Se foloseşte în programarea pătratică, echivalent $Ax = b, A \succ 0$.
- Extensiile neliniare sunt variate

Metoda Gradienţilor Conjugaţi (neliniar):

$$
\begin{aligned}
x^{k+1} &= x^k + \alpha_k p^k, \qquad \alpha_k = \arg\min_{\alpha \geq 0} \; f(x^k + \alpha p^k) \\
p^k &= -r^k + \beta_k p^{k-1}, \qquad \beta_k = \|r^k\|^2 / \|r^{k-1}\|^2 \\
r^k &= \nabla f(x^k), \quad \beta_0 = 0
\end{aligned}
$$

- Pentru f pătratic cele două forme sunt echivalente.

Metoda Gradienţilor Conjugaţi (neliniar):

$$x^{k+1} = x^k + \alpha_k p^k, \qquad \alpha_k = \arg\min_{\alpha \geq 0} \; f(x^k + \alpha p^k)$$
$$p^k = -r^k + \beta_k p^{k-1}, \qquad \beta_k = \|r^k\|^2 / \|r^{k-1}\|^2$$
$$r^k = \nabla f(x^k), \quad \beta_0 = 0$$

Considerăm: $x^1 = x^0 - \frac{\|r^0\|^2}{(r^0)^T A r^0}$. Atunci în cazul pătratic:

$$(r^i)^T r^k = 0 \quad \forall i < k.$$

- Vectorii $\{r^i\}$ sunt ortogonali
- În $\mathbb{R}^n$ nu putem crea mai mult de n vectori r^k
- Concluzie: numărul de iteraţii este maxim n
- Mai mult, $x^k = \arg\min \; f(x)$ s.l. $x \in x^0 + \text{span}\{r^0, \cdots, r^{k-1}\}$

Metoda Gradienţilor Conjugaţi:

$$
\begin{aligned}
x^{k+1} &= x^k + \alpha_k p^k, \qquad \alpha_k = \arg\min_{\alpha \geq 0} \; f(x^k + \alpha p^k) \\
p^k &= -r^k + \beta_k p^{k-1}, \qquad \beta_k = \|r^k\|^2 / \|r^{k-1}\|^2 \\
r^k &= \nabla f(x^k), \quad \beta_0 = 0
\end{aligned}
$$

Teoremă

Dacă H are r valori proprii distincte, atunci MGC converge în cel mult r iteraţii.

Metoda Gradienţilor Conjugaţi:

$$x^{k+1} = x^k + \alpha_k p^k, \qquad \alpha_k = \arg\min_{\alpha \geq 0} \; f(x^k + \alpha p^k)$$

$$p^k = -r^k + \beta_k p^{k-1}, \qquad \beta_k = \|r^k\|^2 / \|r^{k-1}\|^2$$

$$r^k = \nabla f(x^k), \quad \beta_0 = 0$$

Teoremă

Dacă H are r valori proprii distincte, atunci MGC converge în cel mult r iteraţii.

Teoremă

Fie $\lambda_n(H) \leq \cdots \leq \lambda_1(H)$, atunci:

$$\|x^{k+1} - x^*\|_H^2 \leq \left(\frac{\lambda_{k+1} - \lambda_n}{\lambda_{k+1} + \lambda_n} \right)^2 \|x^k - x^*\|_H^2$$

- **Metode de ordin I: Metode Multi-pas**
  - Metoda Heavy-Ball
  - Metoda Gradienţilor Conjugaţi
  - Metoda Gradientului Accelerat
- Metode de ordin II: Metoda Newton
- Metode de ordin I perturbate

Metoda de Gradient Accelerat (Nesterov):

$$x^{k+1} = y^k - \frac{1}{L}\nabla f(y^k)$$
$$y^{k+1} = x^{k+1} + \beta_k (x^{k+1} - x^k)$$

- În plus faţă de MG, execută un pas de extrapolare la fiecare iteraţie:
  - convex: $\beta_k = \frac{\theta_k - 1}{\theta_{k+1}},\ \theta_{k+1} = \frac{1+\sqrt{1+4\theta_k^2}}{2}$
  - $\sigma-$tare convex: $\beta_k = \frac{(\theta_k - 1)(L_f - \theta_{k+1}\sigma_f)}{\theta_{k+1}(L_f - \sigma_f)},\ \theta_{k+1}$ rădăcina a ec. : $\theta_{k+1}^2 - \theta_{k+1} = \left(1 - \frac{\theta_{k+1}\sigma_f}{L_f}\right)\theta_k^2$
- Este cu un ordin de mărime mai rapidă
- Nu este metodă de descreştere!

Metoda de Gradient Accelerat (Nesterov):

$$x^{k+1} = y^k - \frac{1}{L}\nabla f(y^k)$$
$$y^{k+1} = x^{k+1} + \beta_k(x^{k+1} - x^k)$$

Teoremă

Fie f funcţie convexă cu gradient $L-$continuu Lipschitz, atunci şirul $\{x^k\}_{k\geq 0}$ generat de MGA satisface:

$$f(x^k) - f^* \leq \frac{2L\|x^0 - x^*\|^2}{k^2}.$$

Dacă, f este $\sigma-$tare convexă atunci:

$$f(x^k) - f^* \leq \left(1 - \sqrt{\frac{\sigma}{L}}\right)^k (f(x^0) - f^*).$$

Metoda de Gradient Accelerat (Nesterov):

$$x^{k+1} = y^k - \frac{1}{L}\nabla f(y^k)$$
$$y^{k+1} = x^{k+1} + \beta_k(x^{k+1} - x^k)$$

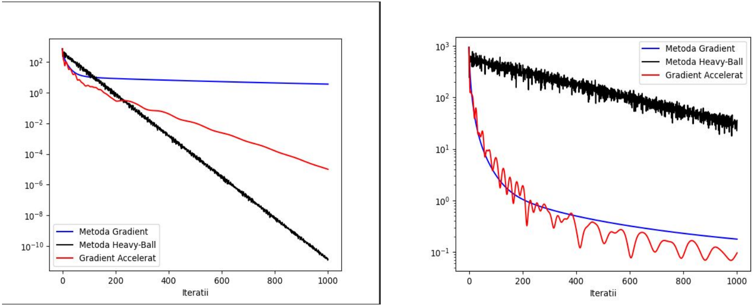


Figure: $\kappa \approx 1.6 \cdot 10^4$(stânga) $\kappa \approx 2 \cdot 10^6$(dreapta)

- Metode de ordin I: Metode Multi-pas
- **Metode de ordin II: Metoda Newton**
- Metode de ordin I perturbate

Aproximarea pătratică în x^k

$$f(x) \approx f(x^k) + \nabla f(x^k)^T (x - x^k) + \frac{1}{2}(x - x^k)^T H_k (x - x^k)$$

Alegerea $H_k = \alpha I$ stă la baza MG.

Dacă funcţia este dublu diferenţiabilă atunci calitatea maximă aproximării se obţine prin alegerea:

$$H_k := \nabla^2 f(x^k).$$

O nouă iteraţie (Newton):

$$x^{k+1} := \arg\min_{x} \; f(x^k) + \nabla f(x^k)^T(x - x^k) + \frac{1}{2}(x - x^k)^T \nabla^2 f(x^k)(x - x^k)$$

La optim avem:

$$\begin{aligned} \nabla^2 f(x^k)(x^{k+1} - x^k) + \nabla f(x^k) &= 0 \\ \nabla^2 f(x^k)(x^{k+1} - x^k) &= -\nabla f(x^k) \\ x^{k+1} &= x^k - [\nabla^2 f(x^k)]^{-1} \nabla f(x^k) \end{aligned}$$

Fie funcţia pătratică:

$$f(x) = \frac{1}{2}x^T A x - b^T x, \quad A \succ 0.$$

Metoda Newton converge într-un singur pas!

$$\begin{aligned} x^1 &= x^0 - [\nabla^2 f(x^0)]^{-1} \nabla f(x^0) \\ &= x^0 - A^{-1}(Ax^0 - b) = A^{-1} b =: x^* \end{aligned}$$

Cu cât f este mai aproape de o funcţie pătratică, cu atât MN converge mai rapid.

Metoda Newton

Algorithm 3: Metoda Newton $(x^0, \epsilon, \{\alpha_k\}_{k\geq 0})$:

Data: $k := 0$

while criteriu oprire = fals **do**

    Calculează: $d^k = [\nabla^2 f(x^k)]^{-1} \nabla f(x^k)$

    $x^{k+1} = x^k - \alpha_k d^k$

    $k := k + 1$

end

Teoremă (Polyak, caz convex)

Fie f dublu diferenţiabilă şi $\nabla^2 f$ *continuu Lipschitz cu constanta L:*

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L\|x - y\|.$$

De asemenea, presupunem f tare convexă cu constanta σ*. Atunci, dacă iteraţia iniţială satisface:*

$$q = \frac{L}{2\sigma^2}\|\nabla f(x^0)\| < 1,$$

atunci x^k *generat de MN cu pas constant* $\alpha_k = 1$ *converge pătratic la optimul global* x^**, i.e.*

$$\|x^k - x^*\| \leq \frac{2\sigma}{L} q^{2^k}$$

Teoremă (Polyak, caz general neconvex)

Fie f dublu diferenţiabilă într-o vecinuătate U a unui minim local nesingular x^*. De asemenea, presupunem $\nabla^2 f$ continuu Lipschitz cu constanta L în U. Atunci, există δ astfel încât pentru:

$$\|x^0 - x^*\| < \delta,$$

x^k generat de MN cu $\alpha_k = 1$ converge pătratic la optimul local x^*.

- Ipoteza $\frac{L}{2\sigma^2}\|\nabla f(x^0)\| < 1$ pretinde iniţializarea lui x^0 într-o vecinătate a lui x^* suficient de mică

- Ipoteza $\frac{L}{2\sigma^2}\|\nabla f(x^0)\| < 1$ pretinde iniţializarea lui x^0 într-o vecinătate a lui x^* suficient de mică
- convexitatea tare asigură existenţa $\nabla^2 f(x^k) \succ 0$

- Ipoteza $\frac{L}{2\sigma^2}\|\nabla f(x^0)\| < 1$ pretinde iniţializarea lui x^0 într-o vecinătate a lui x^* suficient de mică
- convexitatea tare asigură existenţa $\nabla^2 f(x^k) \succ 0$
- Exemplu 1: fie $f(x) = |x|^{5/2}$, atunci pentru $x^0 > 0$ MN are forma: $x^{k+1} = \frac{x^k}{3}$.

- Ipoteza $\frac{L}{2\sigma^2}\|\nabla f(x^0)\| < 1$ pretinde iniţializarea lui x^0 într-o vecinătate a lui x^* suficient de mică
- convexitatea tare asigură existenţa $\nabla^2 f(x^k) \succ 0$
- Exemplu 1: fie $f(x) = |x|^{5/2}$, atunci pentru $x^0 > 0$ MN are forma: $x^{k+1} = \frac{x^k}{3}$.
- Exemplu 2: rezolvaţi $\frac{x}{\sqrt{x^2+1}} = 0$ folosind MN.

$$\min_{x} \quad x - \sum_{i=1}^{n} \ln\left(x - s_i\right)$$

$$\min_{x} \quad x - \sum_{i=1}^{n} \ln\left(x - s_i\right)$$

Condiţii optimalitate: $\sum_{i=1}^{n} \frac{1}{x^* - s_i} = 1 \quad \Rightarrow \| \sum_{i=1}^{n} \frac{1}{x - s_i} - 1 \| \leq \epsilon$

$$\min_{x} \quad x - \sum_{i=1}^{n} \ln(x - s_i)$$

Condiţii optimalitate: $\sum_{i=1}^{n} \frac{1}{x^* - s_i} = 1 \quad \Rightarrow \| \sum_{i=1}^{n} \frac{1}{x - s_i} - 1 \| \leq \epsilon$

MG ($\alpha = 1/100$)

0 :	1.9289682539682538
1 :	1.899511692860917
2 :	1.8713323892897624
3 :	1.8443349259131798
...	
998 :	0.13231868385201828
999 :	0.1321305240489103
1000 :	0.131942700534587

MN ($\alpha = 1$)

1 :	1.9289682539682538
2 :	0.8941238647842491
3 :	0.3198233564591495
4 :	0.0677706555611941
5 :	0.0043675601235591
6 :	$2.0421919943114375e-05$
7 :	$4.503926120946744e-10$

- B. Polyak, Introduction to Optimization, Optimization Software Inc., New York, 1987
- D. Bertsekas, Nonlinear Programming, Third Edition. Athena Scientic, 2016.
- Y. Nesterov, Introductory Lectures on Convex Optimization, Kluwer, 2004.
- Hiriart-Urruty, Jean-Baptiste, and Claude Lemaréchal. *Convex analysis and minimization algorithms I: Fundamentals.* Vol. 305. Springer science & business media, 1996.