Clasificare metode de optimizare

Informatia ce indica comportamentul unei functii $f \in \mathbb{R}^n \to \mathbb{R}$ intr-un punct $x \in \mathbb{R}^n$ se poate clasifica:

- Informatie de ordin 0: $f(x)$
- Informatie de ordin 1: $f(x), \nabla f(x)$
- Informatie de ordin 2: $f(x), \nabla f(x), \nabla^2 f(x)$
- ...

Fie algoritmul iterativ definit de $x_{k+1} = \mathcal{M}(x_k)$; in functie de ordinul informatiei utilizate in expresia lui $\mathcal{M}$:

- Metode de ordin 0: $f(x_k)$
- Metode de ordin 1: $f(x_k), \nabla f(x_k)$
- Metode de ordin 2: $f(x_k), \nabla f(x_k), \nabla^2 f(x_k)$
- ...

Istoric - Metode de ordinul I

Cea mai "simpla" metoda de ordinul I: Metoda Gradient

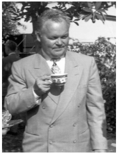
- Prima aparitie in lucrarea [1] a lui Augustin-Louis Cauchy, 1847
- Cauchy rezolva un sistem neliniar de ecuatii cu 6 necunoscute, utilizand Metoda Gradient

[1] A. Cauchy. *Methode generale pour la resolution des systemes dequations simultanees.* C. R. Acad. Sci. Paris, 25:536-538, 1847

Istoric - Metode de ordinul I

Rata de convergenta slaba a metodei gradient reprezinta motivatia dezvoltarii de alte metode de ordin I cu performante superioare

- Metoda de Gradienti Conjugati - autori independenti Lanczos, Hestenes, Stiefel (1952)
  - QP convex solutia in n iteratii



- Metoda de Gradient Accelerat - dezvoltata de Yurii Nesterov (1983)

E.g. metoda de gradient accelerat este cu un ordin mai rapida decat gradientul clasic in cazul problemelor convexe:

- $\mathcal{O}(\frac{LR^2}{k}) \rightarrow \mathcal{O}(\frac{LR^2}{k^2})$ (sublinear - gradient Lipschitz)
- $\mathcal{O}((\frac{L-\sigma}{L+\sigma})^k) \rightarrow \mathcal{O}((1-\sqrt{\frac{\sigma}{L}})^k)$ (liniar - tare convex + grad. Lip.)

Metoda Gradient

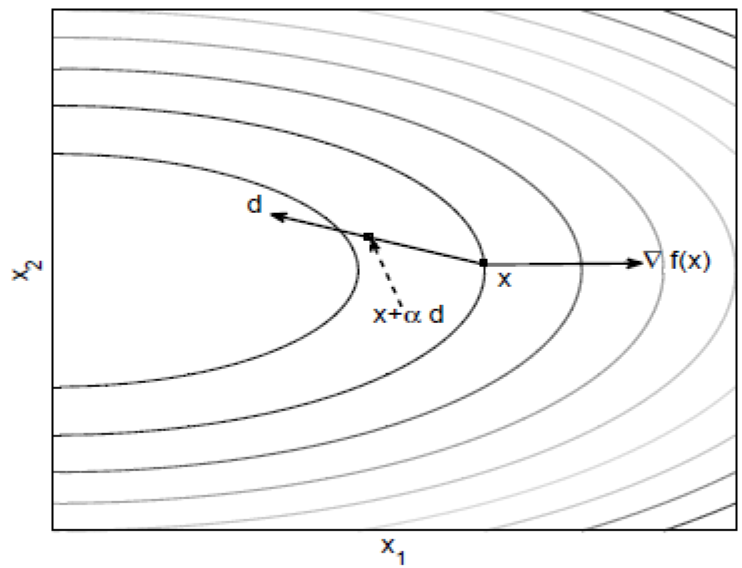
Fie functia $f : \mathbb{R}^n \to \mathbb{R}$ diferentiabila.

$$(UNLP): \quad \min_{x \in \mathbb{R}^n} f(x)$$

Iteratie Metoda Gradient:

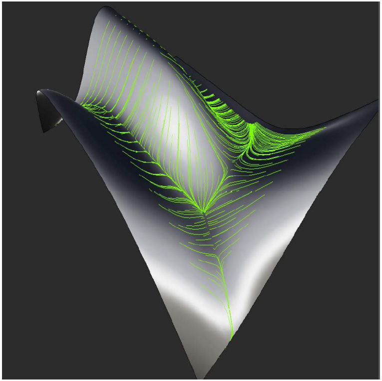
$x_{k+1} = x_k - \alpha_k \nabla f(x_k)$

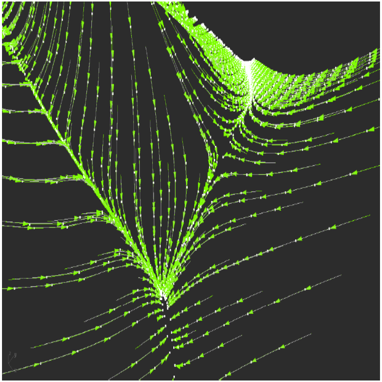
Complexitate pe iteratie $\mathcal{O}(n)$ daca evaluarea $\nabla f(x)$ este ieftina! Metoda gradient rezolva probleme de 10^9 variabile din motoare de cautare, procesarea de imagine



- **Interpretare**: metoda de descrestere cu directia $d = -\nabla f(x_k)$, deci $f(x_{k+1}) \leq f(x_k)$ pentru α_k suficient de mic
- Numeroase variante de alegere a pasului α_k: backtracking, conditii Wolfe, pas constant, pas ideal
- Punct initial x_0 arbitrar, criteriu de oprire e.g. $\|\nabla f(x_k)\| \leq \epsilon$

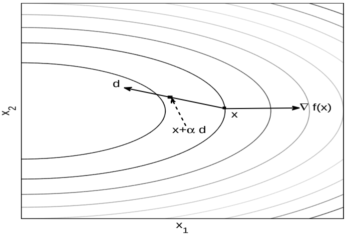
Metoda gradient





Metoda Gradient

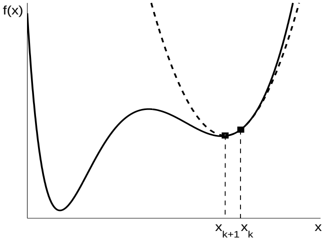
Iteratie Metoda Gradient:
$x_{k+1} = x_k - \alpha_k \nabla f(x_k)$



- **Interpretare**: iteratia metodei gradient se obtine din minimizarea unei aproximari patratice a functiei obiectiv f

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^n} f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2\alpha_k} \|x - x_k\|^2$$

- Aproximare patratica folosind numai $\nabla f(x)$, nu e nevoie de $f(x)$ (vezi asemanarea cu metoda falsei pozitii cazul scalar)



Metoda Gradient-Convergenta globala generala

Teorema 1: Daca urmatoarele conditii sunt satisfacute:

(i) f diferentiabila cu ∇f continuu.

(ii) multimea subnivel $S_{f(x_0)} = \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$ este compacta pentru orice punct initial x_0

(iii) lungimea pasului α_k satisface prima conditie Wolfe (W1),

atunci orice punct limita al sirului x_k generat de metoda gradient este punct stationar pentru problema (UNLP).

Demonstratie: Demonstratia se bazeaza, in principal, pe *Teorema de Convergenta Generala* prezentata in cursul precedent.

Continuitate Lipschitz

Fie o functie continuu diferentiabila f (i.e. $f \in \mathcal{C}^1$), atunci gradientul ∇f este **continuu Lipschitz** cu parametrul $L > 0$ daca:

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \quad \forall x, y \in \mathrm{dom} f \tag{1}$$

Teorema 2: Relatia de Lipschitz (1) implica

$$|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \leq \frac{L}{2}\|x - y\|^2 \quad \forall x, y$$

Observatie: aceasta relatie este universal folosita in ratele de convergenta ale algoritmilor de ordinul I!

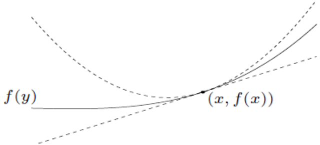
Teorema 3: In cazul functiilor de doua ori diferentiabile, relatia de Lipschitz (1) este echivalenta cu

$$\|\nabla^2 f(x)\| \leq L \quad \forall x \in \mathrm{dom} f$$

Continuitate Lipschitz - convexitate tare

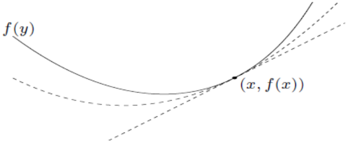
- Daca f are gradient Lipschitz atunci:

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



- Daca f este tare convexa atunci:

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\sigma}{2}\|y - x\|^2$$



- Daca f tare convexa si gradient Lipschitz atunci co-coercivitate:

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq \frac{\sigma L}{\sigma + L}\|x - y\|^2 + \frac{1}{\sigma + L}\|\nabla f(y) - \nabla f(x)\|^2$$

Continuitate Lipschitz - Exemplu 1

Fie $f : \mathbb{R}^n \to \mathbb{R}$ o functie patratica, i.e.

$$f(x) = \frac{1}{2}x^T Qx + \langle q, x\rangle.$$

Observam expresia gradientului $\nabla f(x) = Qx + q$.

Aproximam constanta Lipschitz a functiei f:

$$\|Qx + q - Qy - q\| = \|Q(x - y)\| \leq \|Q\|\|x - y\| = L\|x - y\|$$

In concluzie, pentru functiile patratice constanta Lipschitz este:

$$L = \|Q\| = \lambda_{\max}(Q)$$

Continuitate Lipschitz - Exemplu 2

Fie $f : \mathbb{R}^n \to \mathbb{R}$ definita de

$$f(x) = \log\left(1 + e^{a^T x}\right).$$

Observam expresia gradientului si a matricii Hessiene

$$\nabla f(x) = \frac{e^{a^T x}}{1 + e^{a^T x}} a \qquad \nabla^2 f(x) = \frac{e^{a^T x}}{(1 + e^{a^T x})^2} a a^T$$

Pentru orice constanta pozitiva $c > 0$ avem $\frac{c}{(1+c)^2} \leq \frac{1}{4}$, deci

$$\|\nabla^2 f(x)\| = \frac{e^{a^T x}}{(1 + e^{a^T x})^2} \|a a^T\| \leq \frac{\|a\|^2}{4} = L$$

Metoda Gradient-Convergenta globala sub Lipschitz

Teorema 4: Fie f diferentiabila cu ∇f *Lipschitz continuu* (constanta Lipschitz $L > 0$) si marginita inferior. Daca alegem lungimea pasului α_k astfel incat satisface conditiile Wolfe, atunci sirul x_k generat de metoda gradient satisface:

$$\lim_{k \to \infty} \nabla f(x_k) = 0.$$

Demonstratie: Se observa ca unghiul gradientului fata de directia metodei gradient (antigradient) este dat de $\theta_k = \pi$.
Din *Teorema de Convergenta Globala* a metodele de descrestere avem:

$$\sum_{k \geq 0} \cos^2 \theta_k \|\nabla f(x_k)\|^2 = \sum_{k \geq 0} \|\nabla f(x_k)\|^2 < \infty.$$

Rezulta: $\nabla f(x_k) \to 0$ cand $k \to \infty$.

Metoda Gradient-Rata de convergenta I (globala)

Teorema 5:

- Fie f diferentiabila cu ∇f *Lipschitz continuu* (constanta Lipschitz $L > 0$)
- Alegem lungimea pasului $\alpha_k = \frac{1}{L}$

Atunci rata de convergenta globala a sirului x_k generat de metoda gradient este subliniara, data de:

$$\min_{0 \leq i \leq k} \|\nabla f(x_i)\| \leq \frac{1}{\sqrt{k}} \sqrt{2L(f(x_0) - f^*)}$$

Observatie: daca dorim acuratete ϵ, i.e. $\|\nabla f(x)\| \leq \epsilon$ cate iteratii trebuie sa facem?

$$\frac{1}{\sqrt{k}} \sqrt{2L(f(x_0) - f^*)} \leq \epsilon \Longrightarrow k = \frac{2L(f(x_0) - f^*)}{\epsilon^2}$$

Spunem: rata de convergenta este de ordinul $\mathcal{O}(\frac{1}{\sqrt{k}})$ sau $\mathcal{O}(\frac{1}{\epsilon^2})$

Metoda Gradient-Rata de convergenta I (globala)

Demonstratie Teorema 5:

Sub presupunerea ca ∇f Lipschitz continuu avem:

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2 \quad x, y \in \mathrm{dom} f.$$

Considerand $x = x_k, y = x_{k+1} = x_k - (1/L)\nabla f(x_k)$ avem:

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

Insumam dupa $i = 0, \dots k-1$ si rezulta

$$\frac{1}{2L}\sum_{i=0}^{k-1}\|\nabla f(x_i)\|^2 \leq f(x_0) - f(x_k) \leq f(x_0) - f^*$$

In concluzie, observam

$$k \min_{0 \leq i \leq n}\|\nabla f(x_i)\|^2 \leq \sum_{i=0}^{k-1}\|\nabla f(x_i)\|^2 \leq 2L(f(x_0) - f^*)$$

Metoda Gradient-Rata de convergenta II (locala)

Teorema 6:

- Fie f diferentiabila cu ∇f *Lipschitz continuu* (constanta Lipschitz $L > 0$)
- Exista un punct de minim local x^*, astfel incat Hessiana in acest punct satisface

$$\sigma I_n \preceq \nabla^2 f(x^*) \preceq L I_n$$

- Punctul initial x_0 al iteratiei metodei gradient cu pas $\alpha_k = \frac{2}{\sigma + L}$ este suficient de aproape de punctul de minim, i.e.

$$\|x_0 - x^*\| \leq \frac{2\sigma}{L}$$

Atunci rata de convergenta locala a sirului x_k generat de metoda gradient este liniara (i.e de ordinul $\mathcal{O}\left(\log(\frac{1}{\epsilon})\right)$), data de:

$$\|x_k - x^*\| \leq \beta \left(1 - \frac{2\sigma}{L + 3\sigma}\right)^k \qquad \text{cu } \beta > 0$$

Metoda Gradient-Rata de convergenta III (convex)

Teorema 7

- Fie f functie convexa, diferentiabila cu ∇f Lipschitz continuu (constanta Lipschitz $L > 0$). Daca alegem lungimea pasului constanta $\alpha_k = \frac{1}{L}$, atunci rata de convergenta globala a sirului x_k generat de metoda gradient este subliniara, data de:

$$f(x_k) - f^* \leq \frac{L\|x_0 - x^*\|^2}{2k}$$

- Daca in plus functia este tare convexa cu constanta $\sigma > 0$, atunci rata de convergenta globala a sirului x_k generat de metoda gradient cu pas $\alpha_k = \frac{1}{L}$ este liniara, data de:

$$\|x_k - x^*\|^2 \leq \left(\frac{L - \sigma}{L + \sigma}\right)^k \|x_0 - x^*\|^2$$

$$f(x_k) - f^* \leq \frac{L\|x_0 - x^*\|^2}{2}\left(\frac{L - \sigma}{L + \sigma}\right)^k$$

Metoda Gradient-Rata de convergenta III

Demonstratie Teorema 7: Daca ∇f Lipschitz continuu, atunci

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2 \quad \forall x, y$$

Considerand $x = x_k$ si $y = x_{k+1} = x_k - (1/L)\nabla f(x_k)$, avem:

$$\begin{aligned}
f(x_{k+1}) &\leq f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2 \\
&\leq f^* + \langle \nabla f(x_k), x_k - x^* \rangle - \frac{1}{2L}\|\nabla f(x_k)\|^2 \\
&= f^* + \frac{L}{2}\left(\|x_k - x^*\|^2 - \|x_k - x^* - \frac{1}{L}\nabla f(x_k)\|^2\right) \\
&= f^* + \frac{L}{2}\left(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2\right)
\end{aligned}$$

Prin insumare de la $k = 0, \ldots, N-1$ rezulta

$$\begin{aligned}
N(f(x_N) - f^*) &\leq \sum_{k=0}^{N-1} (f(x_{k+1}) - f^*) \\
&\leq \frac{L}{2}\sum_{k=0}^{N-1} \left(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2\right) \leq \frac{L}{2}\|x_0 - x^*\|^2.
\end{aligned}$$

Metoda Gradient-Rata de convergenta III

Demonstratie Teorema 7: Daca in plus f tare convexa, atunci avem relatia de coercivitate (vezi cursul V):

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{\sigma L}{\sigma + L}\|x-y\|^2 + \frac{1}{\sigma + L}\|\nabla f(x) - \nabla f(y)\|^2$$

Aceasta relatie conduce la:

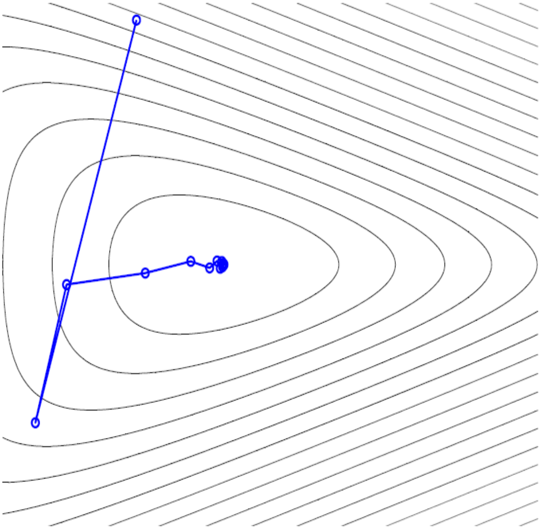
$$\begin{aligned}
\|x_{k+1} - x^*\|^2 &= \|x_k - 1/L\nabla f(x_k) - x^*\|^2 \\
&= \|x_k - x^*\|^2 - 2/L\langle \nabla f(x_k), x_k - x^* \rangle + 1/L^2\|\nabla f(x_k)\|^2 \\
&\underbrace{\leq}_{\text{coerciva}+\nabla f(x^*)=0} \left(1 - \frac{2\sigma}{\sigma + L}\right)\|x_k - x^*\|^2 + \underbrace{\left(\frac{1}{L^2} - \frac{2}{L(\sigma + L)}\right)}_{\leq 0}\|\nabla f(x_k)\|^2 \\
&\leq \left(\frac{L-\sigma}{L+\sigma}\right)\|x_k - x^*\|^2
\end{aligned}$$

Pe de alta parte, in valoarea functiei (cu gradient Lischitz) avem:

$$f(x_k) - f^* \leq \frac{L}{2}\|x_k - x^*\|^2 \leq \frac{L\|x_0 - x^*\|^2}{2}\left(\frac{L-\sigma}{L+\sigma}\right)^k$$

Metoda Gradient- Pas constant $\alpha = 1$

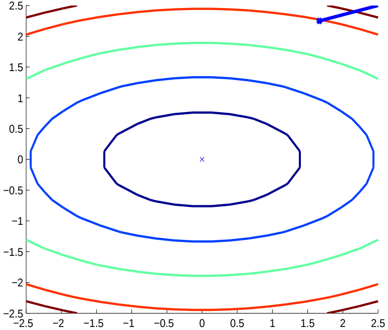
$$f(x) = \log(\exp(x_1 + 3x_2 - .1) + \exp(x_1 - 3x_2 - .1) + \exp(-x_1 - .1))$$



Metoda Gradient- Pas constant $\alpha = 1/L$

$$\min_{x \in \mathbb{R}^2} f(x) \quad \left(= \frac{1}{2}(0.5x_1^2 + \gamma x_2^2) \right), \qquad \text{cu } \gamma > 1$$

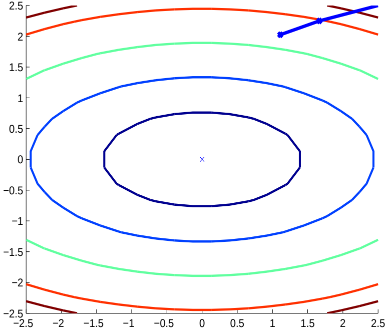
- Functia f tare convexa ($\sigma = 0.5$) si gradient Lipsctitz ($L = \gamma$) $\Longrightarrow$ convergenta liniara
- Metoda Gradient cu pas constant $\alpha = \frac{1}{L}$
- Punct initial $x_0 = \begin{bmatrix} \frac{3}{2}\gamma & 1 \end{bmatrix}$, $\gamma = 5/3$.



Metoda Gradient- Pas constant $\alpha = 1/L$

$$\min_{x \in \mathbb{R}^2} f(x) \quad \left(= \frac{1}{2}(0.5x_1^2 + \gamma x_2^2) \right), \qquad \text{cu } \gamma > 1$$

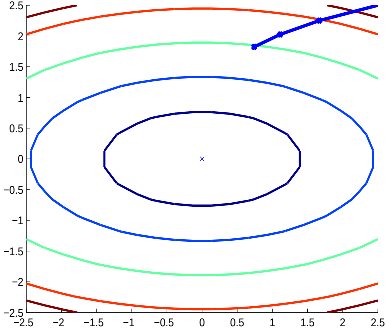
- Functia f tare convexa ($\sigma = 0.5$) si gradient Lipsctitz ($L = \gamma$) $\Longrightarrow$ convergenta liniara
- Metoda Gradient cu pas constant $\alpha = \frac{1}{L}$
- Punct initial $x_0 = \left[\frac{3}{2}\gamma \quad 1\right]$, $\gamma = 5/3$.



Metoda Gradient- Pas constant $\alpha = 1/L$

$$\min_{x \in \mathbb{R}^2} f(x) \quad \left(= \frac{1}{2}(0.5x_1^2 + \gamma x_2^2) \right), \qquad \text{cu } \gamma > 1$$

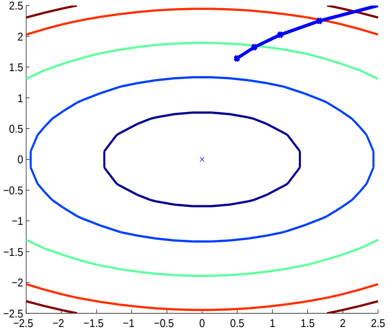
- Functia f tare convexa ($\sigma = 0.5$) si gradient Lipsctitz ($L = \gamma$) $\Longrightarrow$ convergenta liniara
- Metoda Gradient cu pas constant $\alpha = \frac{1}{L}$
- Punct initial $x_0 = \left[\frac{3}{2}\gamma \quad 1\right]$, $\gamma = 5/3$.



Metoda Gradient- Pas constant $\alpha = 1/L$

$$\min_{x \in \mathbb{R}^2} f(x) \quad \left(= \frac{1}{2}(0.5x_1^2 + \gamma x_2^2) \right), \qquad \text{cu } \gamma > 1$$

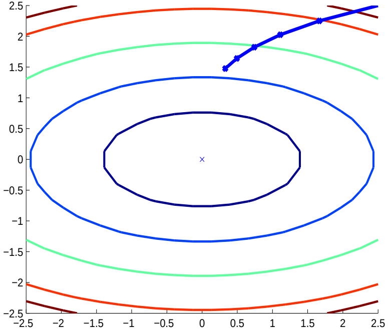
- Functia f tare convexa ($\sigma = 0.5$) si gradient Lipsctitz ($L = \gamma$) $\Longrightarrow$ convergenta liniara
- Metoda Gradient cu pas constant $\alpha = \frac{1}{L}$
- Punct initial $x_0 = \left[\frac{3}{2}\gamma \quad 1\right]$, $\gamma = 5/3$.



Metoda Gradient- Pas constant $\alpha = 1/L$

$$\min_{x \in \mathbb{R}^2} f(x) \quad \left(= \frac{1}{2}(0.5x_1^2 + \gamma x_2^2) \right), \qquad \text{cu } \gamma > 1$$

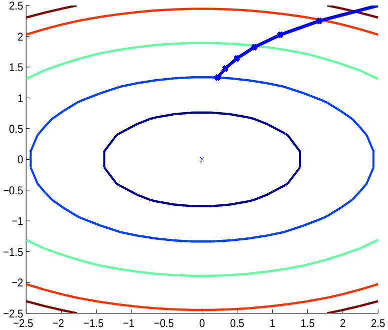
- Functia f tare convexa ($\sigma = 0.5$) si gradient Lipsctitz ($L = \gamma$) $\Longrightarrow$ convergenta liniara
- Metoda Gradient cu pas constant $\alpha = \frac{1}{L}$
- Punct initial $x_0 = \left[\frac{3}{2}\gamma \quad 1\right]$, $\gamma = 5/3$.



Metoda Gradient- Pas constant $\alpha = 1/L$

$$\min_{x \in \mathbb{R}^2} f(x) \quad \left(= \frac{1}{2}(0.5x_1^2 + \gamma x_2^2) \right), \qquad \text{cu } \gamma > 1$$

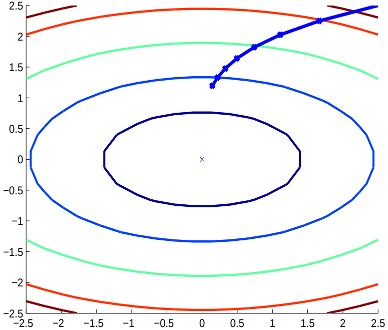
- Functia f tare convexa ($\sigma = 0.5$) si gradient Lipsctitz ($L = \gamma$) $\Longrightarrow$ convergenta liniara
- Metoda Gradient cu pas constant $\alpha = \frac{1}{L}$
- Punct initial $x_0 = \left[\frac{3}{2}\gamma \quad 1\right]$, $\gamma = 5/3$.



Metoda Gradient- Pas constant $\alpha = 1/L$

$$\min_{x\in\mathbb{R}^2} f(x) \quad \left(= \frac{1}{2}(0.5x_1^2 + \gamma x_2^2)\right), \qquad \text{cu } \gamma > 1$$

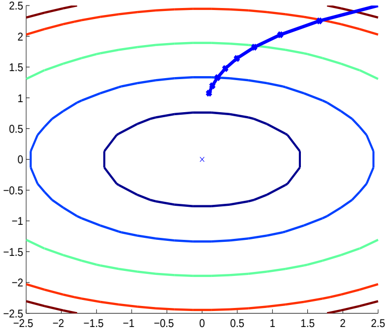
- Functia f tare convexa ($\sigma = 0.5$) si gradient Lipsctitz ($L = \gamma$) $\Longrightarrow$ convergenta liniara
- Metoda Gradient cu pas constant $\alpha = \frac{1}{L}$
- Punct initial $x_0 = \left[\frac{3}{2}\gamma \quad 1\right]$, $\gamma = 5/3$.



Metoda Gradient- Pas constant $\alpha = 1/L$

$$\min_{x \in \mathbb{R}^2} f(x) \quad \left(= \frac{1}{2}(0.5x_1^2 + \gamma x_2^2) \right), \qquad \text{cu } \gamma > 1$$

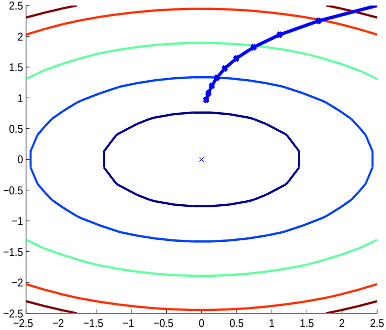
- Functia f tare convexa ($\sigma = 0.5$) si gradient Lipsctitz ($L = \gamma$) $\Longrightarrow$ convergenta liniara
- Metoda Gradient cu pas constant $\alpha = \frac{1}{L}$
- Punct initial $x_0 = \left[\frac{3}{2}\gamma \quad 1\right]$, $\gamma = 5/3$.



Metoda Gradient- Pas constant $\alpha = 1/L$

$$\min_{x \in \mathbb{R}^2} f(x) \quad \left(= \frac{1}{2}(0.5x_1^2 + \gamma x_2^2) \right), \qquad \text{cu } \gamma > 1$$

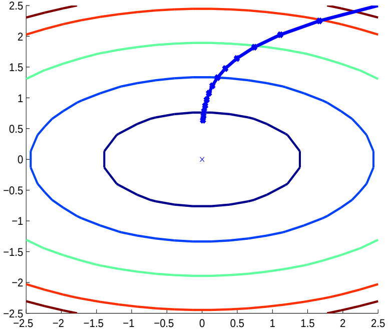
- Functia f tare convexa ($\sigma = 0.5$) si gradient Lipsctitz ($L = \gamma$) $\Longrightarrow$ convergenta liniara
- Metoda Gradient cu pas constant $\alpha = \frac{1}{L}$
- Punct initial $x_0 = \left[\frac{3}{2}\gamma \quad 1\right]$, $\gamma = 5/3$.



Metoda Gradient- Pas constant $\alpha = 1/L$

$$\min_{x \in \mathbb{R}^2} f(x) \quad \left(= \frac{1}{2}(0.5x_1^2 + \gamma x_2^2) \right), \qquad \text{cu } \gamma > 1$$

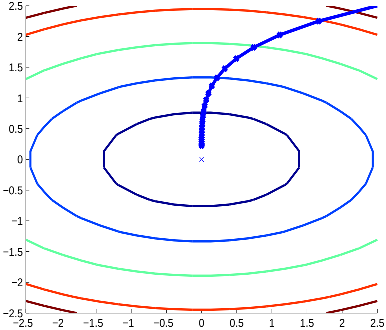
- Functia f tare convexa ($\sigma = 0.5$) si gradient Lipsctitz ($L = \gamma$) $\Longrightarrow$ convergenta liniara
- Metoda Gradient cu pas constant $\alpha = \frac{1}{L}$
- Punct initial $x_0 = \left[\frac{3}{2}\gamma \quad 1 \right]$, $\gamma = 5/3$.



Metoda Gradient- Pas constant $\alpha = 1/L$

$$\min_{x \in \mathbb{R}^2} f(x) \quad \left(= \frac{1}{2}(0.5x_1^2 + \gamma x_2^2) \right), \qquad \text{cu } \gamma > 1$$

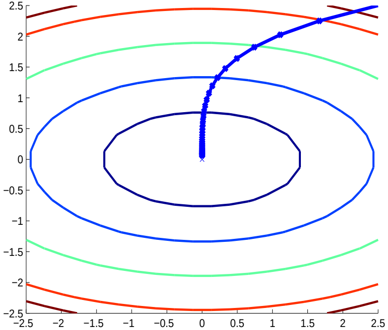
- Functia f tare convexa ($\sigma = 0.5$) si gradient Lipsctitz ($L = \gamma$) $\Longrightarrow$ convergenta liniara
- Metoda Gradient cu pas constant $\alpha = \frac{1}{L}$
- Punct initial $x_0 = \left[\frac{3}{2}\gamma \quad 1\right]$, $\gamma = 5/3$.



Metoda Gradient- Pas constant $\alpha = 1/L$

$$\min_{x \in \mathbb{R}^2} f(x) \quad \left(= \frac{1}{2}(0.5x_1^2 + \gamma x_2^2) \right), \qquad \text{cu } \gamma > 1$$

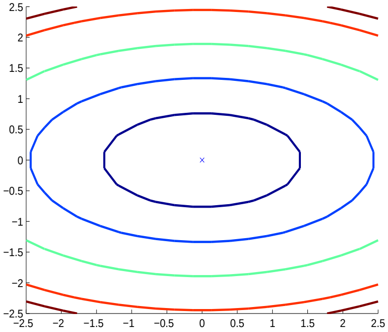
- Functia f tare convexa ($\sigma = 0.5$) si gradient Lipsctitz ($L = \gamma$) $\Longrightarrow$ convergenta liniara
- Metoda Gradient cu pas constant $\alpha = \frac{1}{L}$
- Punct initial $x_0 = \left[\frac{3}{2}\gamma \quad 1\right]$, $\gamma = 5/3$.



Metoda Gradient- Pas ideal

$$\min_{x \in \mathbb{R}^2} f(x) \quad \left(= \frac{1}{2}(0.5x_1^2 + \gamma x_2^2) \right)$$

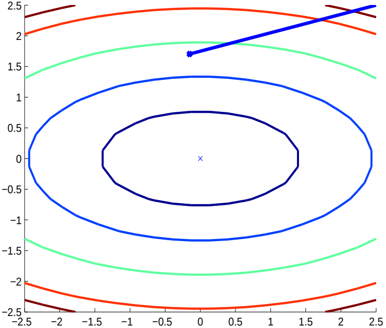
- Metoda Gradient cu pas ideal $\alpha_k = \arg\min_{\alpha \geq 0} f(x_k - \alpha \nabla f(x_k))$
- Punct initial $x_0 = \left[\frac{3}{2}\gamma \quad 1\right]$, $\gamma = 5/3$.



Metoda Gradient- Pas ideal

$$\min_{x \in \mathbb{R}^2} f(x) \quad \left(= \frac{1}{2}(0.5x_1^2 + \gamma x_2^2) \right)$$

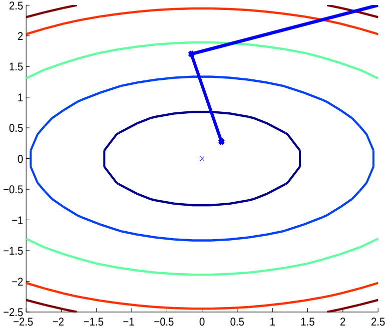
- Metoda Gradient cu pas ideal $\alpha_k = \arg\min_{\alpha \geq 0} f(x_k - \alpha \nabla f(x_k))$
- Punct initial $x_0 = \left[\frac{3}{2}\gamma \quad 1\right]$, $\gamma = 5/3$.



Metoda Gradient- Pas ideal

$$\min_{x \in \mathbb{R}^2} f(x) \quad \left(= \frac{1}{2}(0.5x_1^2 + \gamma x_2^2) \right)$$

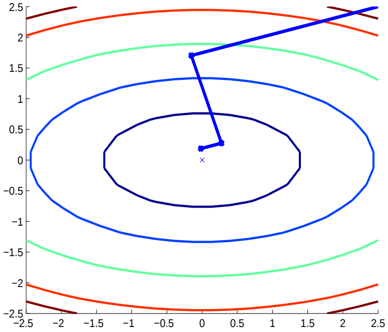
- Metoda Gradient cu pas ideal $\alpha_k = \arg\min_{\alpha \geq 0} f(x_k - \alpha \nabla f(x_k))$
- Punct initial $x_0 = \begin{bmatrix} \frac{3}{2}\gamma & 1 \end{bmatrix}$, $\gamma = 5/3$.



Metoda Gradient- Pas ideal

$$\min_{x \in \mathbb{R}^2} f(x) \quad \left(= \frac{1}{2}(0.5x_1^2 + \gamma x_2^2) \right)$$

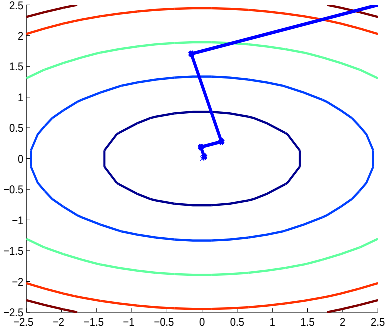
- Metoda Gradient cu pas ideal $\alpha_k = \arg\min_{\alpha \geq 0} f(x_k - \alpha \nabla f(x_k))$
- Punct initial $x_0 = \left[\frac{3}{2}\gamma \quad 1\right]$, $\gamma = 5/3$.



Metoda Gradient- Pas ideal

$$\min_{x \in \mathbb{R}^2} f(x) \quad \left(= \frac{1}{2}(0.5x_1^2 + \gamma x_2^2) \right)$$

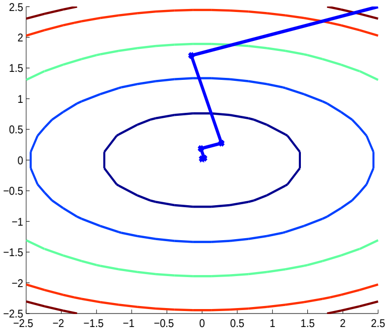
- Metoda Gradient cu pas ideal $\alpha_k = \arg\min_{\alpha \geq 0} f(x_k - \alpha \nabla f(x_k))$
- Punct initial $x_0 = \left[\frac{3}{2}\gamma \quad 1 \right]$, $\gamma = 5/3$.



Metoda Gradient- Pas ideal

$$\min_{x\in\mathbb{R}^2} f(x) \quad \left(= \frac{1}{2}(0.5x_1^2 + \gamma x_2^2)\right)$$

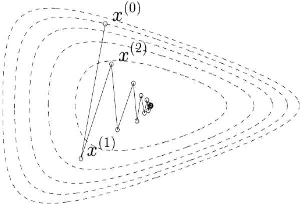
- Metoda Gradient cu pas ideal $\alpha_k = \arg\min_{\alpha\geq 0} f(x_k - \alpha\nabla f(x_k))$
- Punct initial $x_0 = \left[\frac{3}{2}\gamma \quad 1\right]$, $\gamma = 5/3$.



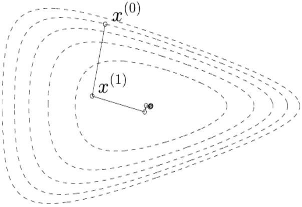
Metoda Gradient - Exemplu nepatratic

$$\min_{x \in \mathbb{R}^2} f(x) \quad \left(= e^{x_1+3x_2-0.1} + e^{x_1-3x_2-0.1} + e^{-x_1-0.1}\right)$$

- functie obiectiv convexa (nu este tare convexa, nu are gradient Lipschitz pe $\mathbb{R}^2$)
- Metoda Gradient cu pas backtraking / ideal



backtracking line search



exact line search

Alte metode de ordinul I

Rata de convergenta slaba a metodei gradient reprezinta motivatia dezvoltarii de metode cu performante superioare

- Metoda de Gradient Accelerat (Nesterov 1983) - cu un ordin mai rapida decat gradientul clasic in cazul problemelor convexe
- Metoda de Gradienti Conjugati (Lanczos, Hestenes, Stiefel 1952) - pentru QP convex solutia in n iteratii

Metoda de Gradient Accelerat

$$
\begin{aligned}
x_{k+1} &= y_k - \frac{1}{L}\nabla f(y_k) \\
y_{k+1} &= x_{k+1} + \beta_k(x_{k+1} - x_k)
\end{aligned}
$$

unde se iau punctele initiale $x_0 = y_0$ si β_k ales in mod adecvat: e.g. sub convexitate tare putem alege $\beta_k = \frac{\sqrt{L}-\sqrt{\sigma}}{\sqrt{L}+\sqrt{\sigma}}$

Observatie: Costul iteratiei similar cu cel al metodei gradient clasice!

Metoda Gradient Accelerat

Teorema 8

- Fie f o functie convexa, diferentiabila cu ∇f Lipschitz continuu (constanta Lipschitz $L > 0$). Rata de convergenta globala a sirului x_k generat de metoda gradient accelerat este subliniara, data de:

$$f(x_k) - f^* \leq \frac{4L\|x_0 - x^*\|^2}{k^2}$$

- Daca in plus functia este tare convexa cu constanta $\sigma > 0$, atunci rata de convergenta este liniara, data de:

$$f(x_k) - f^* \leq L\|x_0 - x^*\|^2 \left(1 - \sqrt{\frac{\sigma}{L}}\right)^k$$

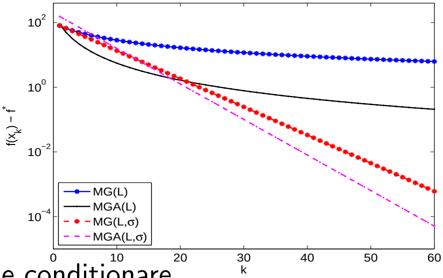
Metoda gradient accelerat este cu un ordin mai rapida decat gradientul clasic in cazul problemelor convexe ($R = \|x_0 - x^*\|$):

- $\mathcal{O}(\frac{LR^2}{k}) \rightarrow \mathcal{O}(\frac{LR^2}{k^2})$ (sublinear - gradient Lipschitz)
- $\mathcal{O}((\frac{L-\sigma}{L+\sigma})^k) \rightarrow \mathcal{O}((1 - \sqrt{\frac{\sigma}{L}})^k)$ (liniar - tare convex + grad. Lip.)

Metoda Gradient Accelerat

- $\mathcal{O}(\frac{LR^2}{k}) \rightarrow \mathcal{O}(\frac{LR^2}{k^2})$ versus $\mathcal{O}((\frac{L-\sigma}{L+\sigma})^k) \rightarrow \mathcal{O}((1-\sqrt{\frac{\sigma}{L}})^k)$

Figura corespunde constantelor: $R = 10, L = 2$ si $\sigma = 0.1$



Se observa ca numarul de conditionare

$$\kappa = \frac{L}{\sigma}$$

este relativ mic (i.e. 20). Raportul $\kappa = \frac{L}{\sigma}$ reprezinta numarul de conditionare al problemei de optimizare convexe (UNLP) datorita similitudinii cu definitia numarului de conditionare al unei matrici

$$\min_x f(x) \quad (= 0.5 x^T Q x)$$

atunci $L = \lambda_{\max} = \|Q\|$ si $\sigma = \lambda_{\min} = 1/\|Q^{-1}\|$

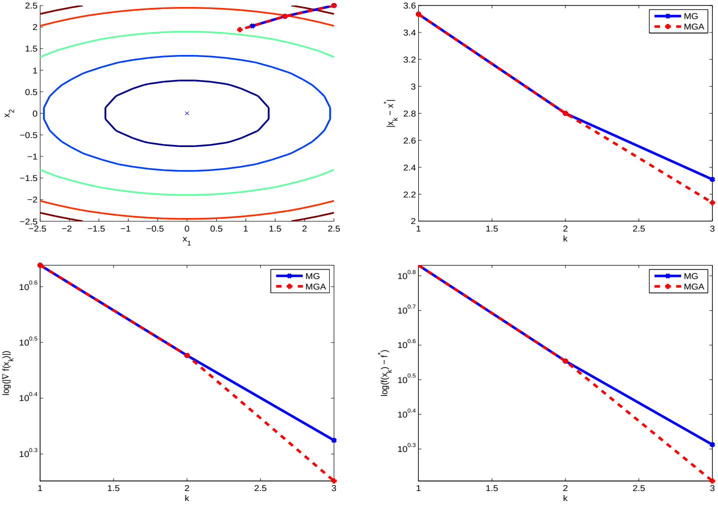
Gradient - Gradient Accelerat (exemplu 1)

$$\min_{x\in\mathbb{R}^2} f(x) \quad \left(= \frac{1}{2}(0.5x_1^2 + \gamma x_2^2)\right), \qquad \text{cu } \gamma > 1$$

- Functia obiectiv f tare convexa ($\sigma = 0.5$) si gradient Lipschitz ($L = \gamma$)
- Metoda Gradient cu pas constant $\alpha_k = 1/L$
- Metoda Gradient Accelerat cu pas constant $\alpha_k = 1/L$ si $\beta_k = (\sqrt{L} - \sqrt{\sigma})/(\sqrt{L} + \sqrt{\sigma})$
- Ambele metode de ordinul I converg liniar
- Punct initial $x_0 = \left[\frac{3}{2}\gamma \quad 1\right]$, $\gamma = 5/3$.

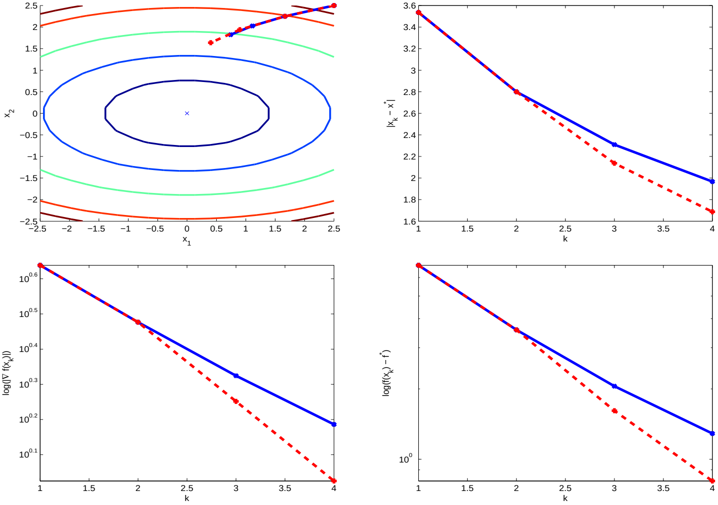
Gradient - Gradient Accelerat (exemplu 1)

- Metoda Gradient cu pas constant $\alpha_k = 1/L$
- Metoda Gradient Accelerat cu pas constant $\alpha_k = 1/L$ si $\beta_k = (\sqrt{L} - \sqrt{\sigma})/(\sqrt{L} + \sqrt{\sigma})$



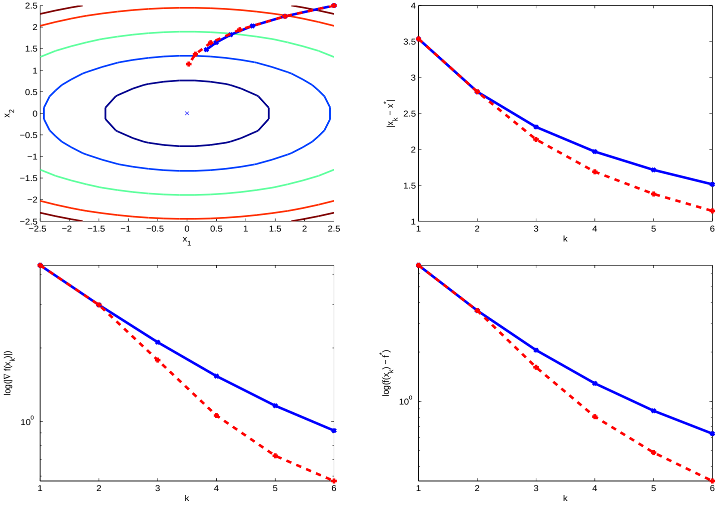
Gradient - Gradient Accelerat (exemplu 1)

- Metoda Gradient cu pas constant $\alpha_k = 1/L$
- Metoda Gradient Accelerat cu pas constant $\alpha_k = 1/L$ si $\beta_k = (\sqrt{L} - \sqrt{\sigma})/(\sqrt{L} + \sqrt{\sigma})$



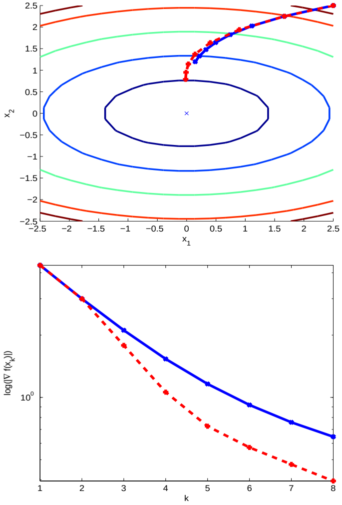
Gradient - Gradient Accelerat (exemplu 1)

- Metoda Gradient cu pas constant $\alpha_k = 1/L$
- Metoda Gradient Accelerat cu pas constant $\alpha_k = 1/L$ si $\beta_k = (\sqrt{L} - \sqrt{\sigma})/(\sqrt{L} + \sqrt{\sigma})$



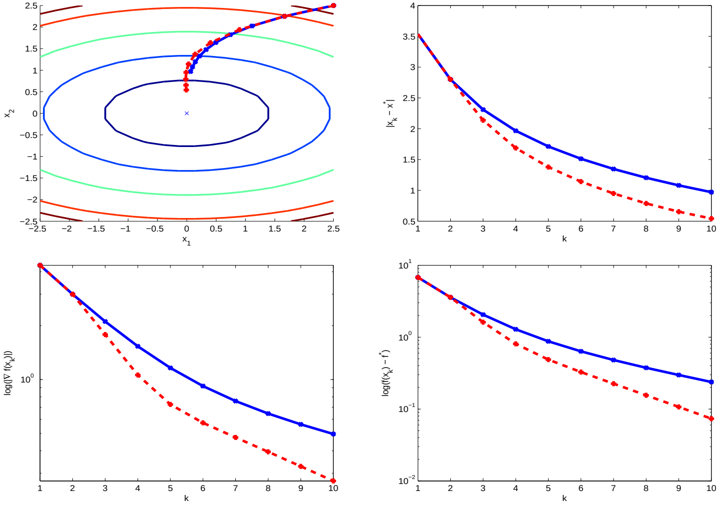
Gradient - Gradient Accelerat (exemplu 1)

- Metoda Gradient cu pas constant $\alpha_k = 1/L$
- Metoda Gradient Accelerat cu pas constant $\alpha_k = 1/L$ si $\beta_k = (\sqrt{L} - \sqrt{\sigma})/(\sqrt{L} + \sqrt{\sigma})$



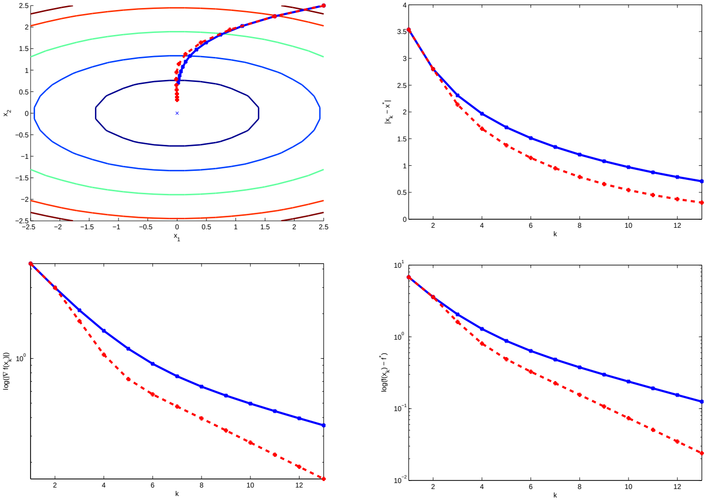
Gradient - Gradient Accelerat (exemplu 1)

- Metoda Gradient cu pas constant $\alpha_k = 1/L$
- Metoda Gradient Accelerat cu pas constant $\alpha_k = 1/L$ si $\beta_k = (\sqrt{L} - \sqrt{\sigma})/(\sqrt{L} + \sqrt{\sigma})$



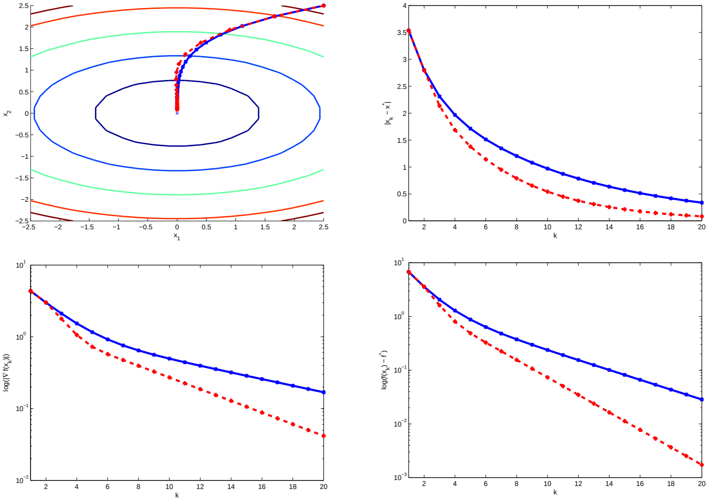
Gradient - Gradient Accelerat (exemplu 1)

- Metoda Gradient cu pas constant $\alpha_k = 1/L$
- Metoda Gradient Accelerat cu pas constant $\alpha_k = 1/L$ si $\beta_k = (\sqrt{L} - \sqrt{\sigma})/(\sqrt{L} + \sqrt{\sigma})$



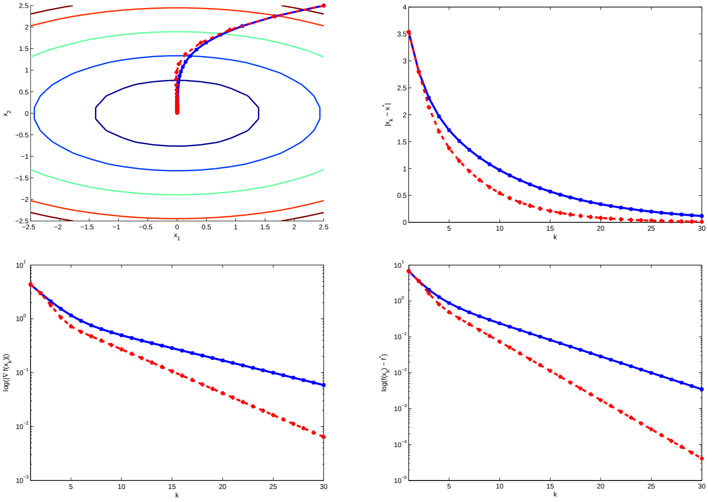
Gradient - Gradient Accelerat (exemplu 1)

- Metoda Gradient cu pas constant $\alpha_k = 1/L$
- Metoda Gradient Accelerat cu pas constant $\alpha_k = 1/L$ si $\beta_k = (\sqrt{L} - \sqrt{\sigma})/(\sqrt{L} + \sqrt{\sigma})$



Gradient - Gradient Accelerat (exemplu 1)

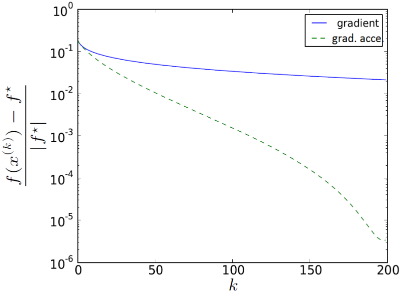
- Metoda Gradient cu pas constant $\alpha_k = 1/L$
- Metoda Gradient Accelerat cu pas constant $\alpha_k = 1/L$ si $\beta_k = (\sqrt{L} - \sqrt{\sigma})/(\sqrt{L} + \sqrt{\sigma})$

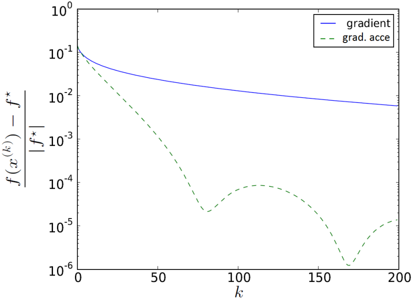


Gradient - Gradient Accelerat (exemplu 2)

$$\min_{x \in \mathbb{R}^n} f(x) \quad \left(= \log \sum_{j=1}^{m} e^{a_i^T x + b_i} \right)$$

- Functia obiectiv are gradient Lipschitz, dar nu e tare convexa
- Date generate aleator cu $m = 10^4$ si $n = 10^3$
- Metoda Gradient si Gradient Accelerat cu pas $\alpha_k = 1/L$





Metoda Gradient accelerat nu este o metoda de descrestere!

Program Matlab - metoda gradient ideal

Algoritmul MG. (Se da punctul de start x_0 si acuratetea ϵ. Se calculeaza o ϵ-solutie optima pentru problema de optimizare $\min_x f(x) \quad (= 10x_1^6 + 30x_2^6 + x_1^2 + 50x_2^2)$ cu MG-ideal.)

0. function $[\cdot] = \text{MG-ideal}(x0, \epsilon)$
1. obj $= @(x)\ 10 * x(1)^6 + 30 * x(2)^6 + x(1)^2 + 50 * x(2)^2$
2. grad $= @(x)\ [60 * x(1)^5 + 2 * x(1);\ 180 * x(2)^5 + 100 * x(2)]$
3. $x = x0, tg = x0$
4. while(norm(grad(x)) $> \epsilon$)
    1. $\text{obj}_\alpha = @(\alpha)\ \text{obj}(x - \alpha \text{grad}(x))$
    2. $\alpha^* = \text{fminbnd}(\text{obj}_\alpha, 0, 1)$
    3. $x = x - \alpha^* * \text{grad}(x); \quad tg = [tg; x]$
5. end while
6. $x = -0.2 : 0.1 : 0.2;\ y = -0.2 : 0.1 : 0.2;\ [X, Y] = \text{meshgrid}(x, y)$
7. $Z = 10 * X.^6 + 30 * Y.^6 + X.^2 + 50 * Y.^2$
8. figure; plot($tg(1, :), tg(2, :)$); hold on; contour(X, Y, Z)