TEHNICI DE OPTIMIZARE
Curs 2

Andrei Pătraşcu

Departament Informatică
Universitatea din Bucureşti

- **Modele de optimizare continuă fără constrângeri. Condiţii de extrem**
- Existenţă şi unicitate. Mulţimi izo(sub-)nivel. Convexitate tare
- Metoda Gradient

$$(MfC): \qquad \min_{x} \; f(x)$$

Aspecte în formularea şi rezolvarea (OfC):

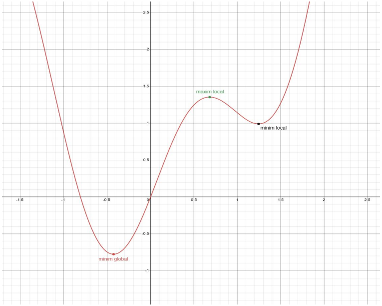
- definiţia (verificarea) unei soluţii
- existenţa unei soluţii
- unicitatea soluţiei
- calculul unei soluţii (algoritm iterativ)

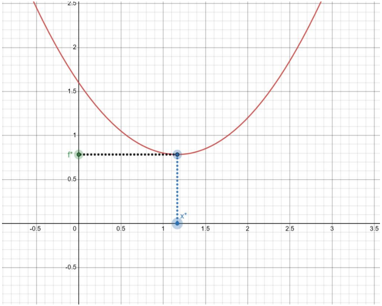
Definiţie (Minim local/global)

Punctul x^* *se numeşte* ***minim local*** *dacă: există* $\epsilon > 0$ *astfel încât*

$$f(x^*) \leq f(x) \qquad \forall x \in \{x : \|x - x^*\| \leq \epsilon\}$$

Mai mult, dacă $f(x^*) \leq f(x)$, $\forall x \in \mathrm{dom}(f)$, *atunci* x^* *este* ***minim global***.





Teoremă (Fermat)

Presupunem funcţia f diferenţiabilă. Fie x^* un punct de minim al $f(\cdot)$ din $\mathbb{R}^n$, atunci:

$$\nabla f(x^*) = 0.$$

Teoremă (Fermat)

Presupunem funcţia f diferenţiabilă. Fie x^* un punct de minim al $f(\cdot)$ din $\mathbb{R}^n$, atunci:

$$\nabla f(x^*) = 0.$$

Demonstraţie: Presupunem $\nabla f(x^*) \neq 0$. 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.

Teoremă (Fermat)

Presupunem funcţia f diferenţiabilă. Fie x^* un punct de minim al $f(\cdot)$ din $\mathbb{R}^n$, atunci:

$$\nabla f(x^*) = 0.$$

Observăm că demonstraţia teoremei de mai sus oferă o perspectivă asupra calculării unei direcţii de descreştere a funcţiei f.

Idee

Dacă $\nabla f(x) \neq 0$ (i.e. x nu este punct de extrem), atunci $x - \tau \nabla f(x)$ asigură descreşterea funcţiei f, pentru pasul τ suficient de mic.

Minimizare pătratică:

$$\min_{x \in \mathbb{R}^n} \frac{1}{2} x^T P x + q^T x + r$$

- Problema este rezolvată prin setul de ecuaţii: $Px^* + q = 0$
- $P \succ 0$, atunci avem soluţie unică $x^* = -P^{-1} q$;
- dacă $Px^* + q = 0$ nu are soluţie atunci f este nemărginită inferior.

Problema aproximării liniare:

$$\min_{x\in\mathbb{R}^n} \|Ax-b\|_2^2$$

$A\in\mathbb{R}^{m\times n}, b\in\mathbb{R}^m$ sunt datele problemei.

- Există cel puţin o soluţie $\nabla f(x^*) := A^T(Ax^*-b)=0$
- Dacă $b\in Im(A)$ atunci $Ax^*=b$ (interpolare)
- Dacă $b\notin Im(A)$ atunci $A^TAx^*=A^Tb$ (aproximare).

Fie funcţia $f(x) = f(x_1, x_2) := \frac{1}{2}x_1^2 + \frac{1}{4}x_2^4 - \frac{1}{2}x_2^2$

Fie funcţia $f(x) = f(x_1, x_2) := \frac{1}{2}x_1^2 + \frac{1}{4}x_2^4 - \frac{1}{2}x_2^2$

$$\text{Puncte critice: } \nabla f(x^*) = \begin{bmatrix} x_1^* \\ (x_2^*)^3 - x_2^* \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

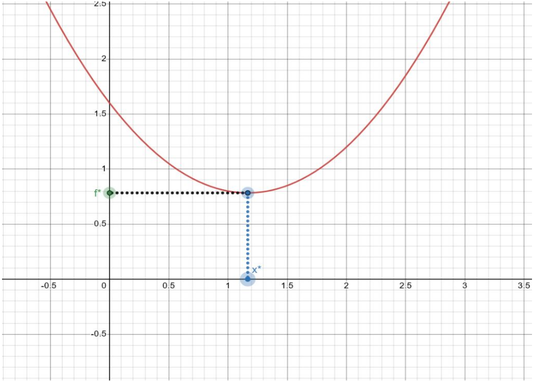
Fie funcţia $f(x) = f(x_1, x_2) := \frac{1}{2}x_1^2 + \frac{1}{4}x_2^4 - \frac{1}{2}x_2^2$

$$\text{Puncte critice: } \nabla f(x^*) = \begin{bmatrix} x_1^* \\ (x_2^*)^3 - x_2^* \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$x_{(1)}^* = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad x_{(2)}^* = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad x_{(3)}^* = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$$

Teoremă

Presupunem funcţia f convexă şi diferenţiabilă în x^*. Dacă $\nabla f(x^*) = 0$ atunci x^* este minim global al funcţiei f.



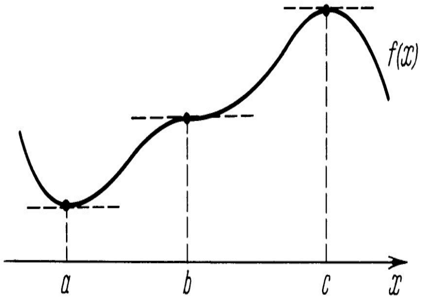
Teoremă

Presupunem funcţia f convexă şi diferenţiabilă în x^*. Dacă $\nabla f(x^*) = 0$ atunci x^* este minim global al funcţiei f.

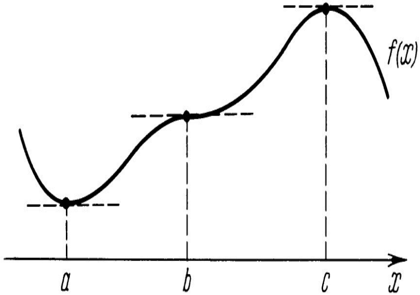
Demonstraţie: Din definţia funcţiilor convexe avem:

$$f(x) \geq f(x^*) + \langle \nabla f(x^*), x - x^* \rangle = f(x^*).$$

Condiţia $\nabla f(z) = 0$ este satisfăcută de orice punct staţionar:



Condiţii necesare de ordin II



- a: punct de minim (local)
- b: punct şa (inflexiune)
- c: punct de maxim (local)

Derivata de ordin I nu oferă informaţii despre natura punctului staţionar! Este necesar să apelăm la derivate de ordin superior.

Teoremă

Presupunem funcţia f este dublu diferenţiabilă. Fie x^* un punct de minim al $f(\cdot)$ din $\mathbb{R}^n$, atunci:

$$\nabla^2 f(x^*) \succeq 0.$$

Demonstraţie: Ştim că $\nabla f(x^*) = 0$, atunci deducem pentru orice $d \in \mathbb{R}^n$ şi $\tau > 0$ suficient de mic:

$$f(x^*) \leq f(x^* + \tau d) = f(x^*) + \frac{\tau^2}{2} \langle \nabla^2 f(x^*) d, d \rangle + o(\tau^2)$$

$$\langle \nabla^2 f(x^*) d, d \rangle \geq \frac{o(\tau^2)}{\tau^2}.$$

Luând $\tau \to 0$ atunci partea dreaptă se anulează şi obţinem $\nabla^2 f(x^*) \succeq 0$.

Teoremă

Presupunem funcţia f este dublu diferenţiabilă. Dacă $\nabla f(x^*) = 0$ şi:

$$\nabla^2 f(x^*) \succ 0$$

atunci x^* este minim local.

Teoremă

Presupunem funcţia f este dublu diferenţiabilă. Dacă $\nabla f(x^*) = 0$ şi:

$$\nabla^2 f(x^*) \succ 0$$

atunci x^* este minim local.

Demonstraţie: Pentru orice $d \in \mathbb{R}^n$ şi $\tau > 0$ suficient de mic:

$$\begin{aligned} f(x^* + \tau d) &= f(x^*) + \frac{\tau^2}{2}\langle \nabla^2 f(x^*)d, d\rangle + o(\tau^2\|d\|^2) \\ &\geq f(x^*) + \frac{\tau^2}{2}\lambda_{\min}\|d\|^2 + o(\tau^2\|d\|^2) \\ &= f(x^*) + \tau^2\|d\|^2\left(\frac{\lambda_{\min}}{2} + \frac{o(\tau^2\|d\|^2)}{\tau^2\|d\|^2}\right), \end{aligned}$$

unde $\lambda_{\min}$ reprezintă valoarea proprie minimă a matricii $\nabla^2 f(x^*)$. Pentru τ suficient de mic avem $\frac{\lambda_{\min}}{2} + \frac{o(\tau^2\|d\|^2)}{\tau^2\|d\|^2} \geq 0$.

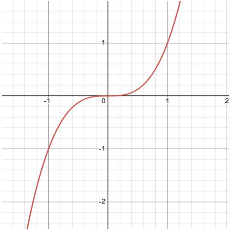
Observaţie

Sub condiţiile necesare de ordin I şi II ($\nabla f(x^*) = 0, \nabla^2 f(x^*) \succeq 0$), dacă cele suficiente nu au loc (matricea $\nabla^2 f(x^*)$ nu este pozitiv definită) atunci x^* nu este neapărat minim local.

Exemplu: Fie $f(x) = x^3$, atunci $x* = 0$ satisface condiţiile necesare, insă:

$$f''(x^*) = 0,$$

$x^* = 0$ este punct şa.



Fie funcţia $f(x) = f(x_1, x_2) := \frac{1}{2}x_1^2 + \frac{1}{4}x_2^4 - \frac{1}{2}x_2^2$

Fie funcţia $f(x) = f(x_1, x_2) := \frac{1}{2}x_1^2 + \frac{1}{4}x_2^4 - \frac{1}{2}x_2^2$

$$\text{Puncte critice: } \nabla f(x^*) = \begin{bmatrix} x_1^* \\ (x_2^*)^3 - x_2^* \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Fie funcţia $f(x) = f(x_1, x_2) := \frac{1}{2}x_1^2 + \frac{1}{4}x_2^4 - \frac{1}{2}x_2^2$

$$\text{Puncte critice:} \ \nabla f(x^*) = \begin{bmatrix} x_1^* \\ (x_2^*)^3 - x_2^* \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$x_{(1)}^* = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad x_{(2)}^* = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad x_{(3)}^* = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$$

Exemplu

Fie funcţia $f(x) = f(x_1, x_2) := \frac{1}{2}x_1^2 + \frac{1}{4}x_2^4 - \frac{1}{2}x_2^2$

$$\text{Puncte critice: } \nabla f(x^*) = \begin{bmatrix} x_1^* \\ (x_2^*)^3 - x_2^* \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\nabla^2 f(x) = \begin{bmatrix} 1 & 0 \\ 0 & 3x_2^2 - 1 \end{bmatrix}$$

Fie funcţia $f(x) = f(x_1, x_2) := \frac{1}{2}x_1^2 + \frac{1}{4}x_2^4 - \frac{1}{2}x_2^2$

$$\text{Puncte critice: } \nabla f(x^*) = \begin{bmatrix} x_1^* \\ (x_2^*)^3 - x_2^* \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\nabla^2 f(x^*_{(1)}) = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \qquad \nabla^2 f(x^*_{(2)}) = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \qquad \nabla^2 f(x^*_{(3)}) = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$$

Fie funcţia $f(x) = f(x_1, x_2) := \frac{1}{2}x_1^2 + \frac{1}{4}x_2^4 - \frac{1}{2}x_2^2$

$$\text{Puncte critice: } \nabla f(x^*) = \begin{bmatrix} x_1^* \\ (x_2^*)^3 - x_2^* \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\nabla^2 f(x^*_{(1)}) = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \qquad \nabla^2 f(x^*_{(2)}) = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \qquad \nabla^2 f(x^*_{(3)}) = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$$

punct staţionar — min. local nesingular — min. local nesingular

- În cazurile liniar-pătratice, condiţiile de optimalitate ajută direct la determinarea unei soluţii

Concluzii asupra condiţiilor de optimalitate

- În cazurile liniar-pătratice, condiţiile de optimalitate ajută direct la determinarea unei soluţii
- În general, rezolvarea condiţiilor de optimalitate (determinarea unui punct are satisface relaţiile de egalitate/inegalitate) are, adesea, aceeaşi dificultate ca şi problema originală

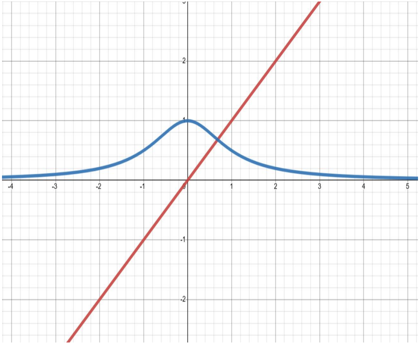
- În cazurile liniar-pătratice, condiţiile de optimalitate ajută direct la determinarea unei soluţii
- În general, rezolvarea condiţiilor de optimalitate (determinarea unui punct are satisface relaţiile de egalitate/inegalitate) are, adesea, aceeaşi dificultate ca şi problema originală
- Care este utilitatea acestora în aceste cazuri neliniare?
  - Condiţiile de optimalitate reflectă proprietăti ale punctelor de extrem
  - Aduc intuiţie asupra construţiei de algoritmi iterativi pentru rezolvarea problemei originale

- Modele de optimizare continuă fără constrângeri. Condiţii de extrem
- **Existenţă şi unicitate. Mulţimi izo(sub-)nivel. Convexitate tare**
- Metoda Gradient

Teoremă (Weierstrass)

Fie funcţia f continuă şi mulţimile sub-nivel $S_f(\alpha) = \{x : f(x) \leq \alpha\}$ nevide şi mărginite. Atunci există un punct de minim global al funcţiei f.

Condiţia mărginirii pare esenţială: e.g. $f(x) = x$, $f(x) = \frac{1}{1+x^2}$



www.desmos.com

Teoremă (Weierstrass)

Fie funcţia f continuă şi mulţimile sub-nivel $S_f(\alpha) = \{x : f(x) \leq \alpha\}$ nevide şi mărginite. Atunci există un punct de minim global al funcţiei f.

Cu toate acestea, dacă $f(x) = \|Ax - b\|_2^2$ atunci:

- f atinge minim-ul pe $\mathbb{R}^n$
- $S_f(\alpha)$ **nu este necesar mărginită!**

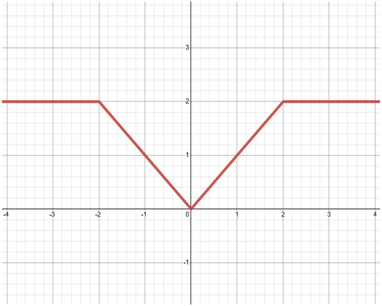
Definiţie

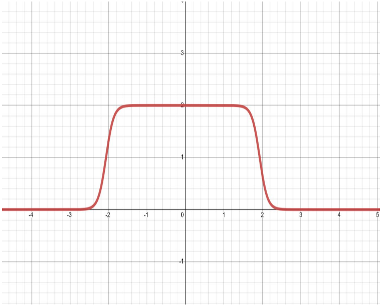
Un punct de optim este ***local unic*** *dacă într-o vecinătate a acestuia nu mai există ale puncte de optim.*

Definiţie

*Un punct de optim este **local unic** dacă într-o vecinătate a acestuia nu mai există ale puncte de optim.*

E.g. $\min\{2, |x|\}$, $\frac{1}{\frac{1}{2}+e^{10(x-2)}} - \frac{1}{\frac{1}{2}+e^{10(x+2)}}$





Definiţie

*Un punct de optim este **local unic** dacă într-o vecinătate a acestuia nu mai există ale puncte de optim.*

Definiţie

*Un punct de minim x^* este **nesingular** dacă:* $\nabla f(x^*) = 0, \nabla^2 f(x^*) \succ 0$.

Definiţie

*Un punct de optim este **local unic** dacă într-o vecinătate a acestuia nu mai există ale puncte de optim.*

Definiţie

*Un punct de minim x^* este **nesingular** dacă:* $\nabla f(x^*) = 0, \nabla^2 f(x^*) \succ 0$.

Teoremă (Unicitate)

Un punct de minim x^* nesingular este local unic.

Fie funcţia $f(x) = f(x_1, x_2) := \frac{1}{2}x_1^2 + \frac{1}{4}x_2^4 - \frac{1}{2}x_2^2$

$$\text{Puncte critice: } \nabla f(x^*) = \begin{bmatrix} x_1^* \\ (x_2^*)^3 - x_2^* \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$x^*_{(1)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad x^*_{(2)} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad x^*_{(3)} = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$$

$$\nabla^2 f(x) = \begin{bmatrix} 1 & 0 \\ 0 & 3x_2^2 - 1 \end{bmatrix}$$

$$\nabla^2 f(x^*_{(1)}) = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \quad \nabla^2 f(x^*_{(2)}) = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \quad \nabla^2 f(x^*_{(3)}) = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$$

punct staţionar | min. local nesingular | min. local nesingular

Definiţie

Mulţimile subnivel şi izonivel ale funcţiei f de parametru $\alpha > 0$ sunt:

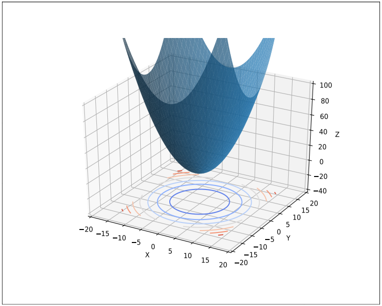
$$S_f(\alpha) := \{x \in dom\, f: \ f(x) \leq \alpha\}$$
$$S_f^\circ(\alpha) := \{x \in dom\, f: \ f(x) = \alpha\}.$$

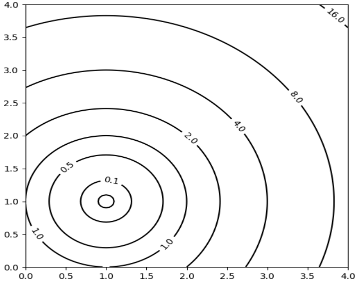
- $S_f^\circ(\alpha)$ reprezintă frontiera mulţimii $S_f(\alpha)$
- Observăm $S_f(\alpha) \subseteq S_f(\beta)$ pentru $\alpha \leq \beta$.
- Pentru funcţiile f convexe, $S_f(\alpha)$ sunt convexe.
- Nu sunt întodeauna mărginite.

Definiţie

Mulţimile subnivel şi izonivel ale funcţiei f de parametru $\alpha > 0$ *sunt:*

$$S_f(\alpha) := \{x \in dom\, f : \; f(x) \leq \alpha\}$$
$$S_f^\circ(\alpha) := \{x \in dom\, f : \; f(x) = \alpha\}.$$

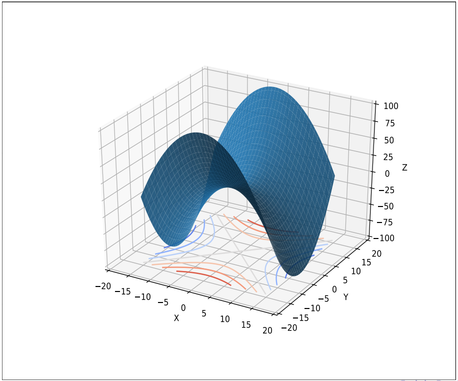




Definiţie

Mulţimile subnivel şi izonivel ale funcţiei f de parametru $\alpha > 0$ *sunt:*

$$S_f(\alpha) := \{x \in dom\, f : \; f(x) \leq \alpha\}$$
$$S_f^{\circ}(\alpha) := \{x \in dom\, f : \; f(x) = \alpha\}.$$



Remarcă

Dacă f este convexă, atunci $S_f(\alpha)$ este convexă pentru oricare $\alpha \geq f^*$.

- $\{x \in \mathbb{R}^n : \ \|Ax - b\|_2 \leq 1\}$
- $\{(x, t) \in \mathbb{R}^{n+1} : \ \|x\|_2 \leq t\}$
- $\{x \in \mathbb{R}^n_+ : \ \sum_i x_i \ln x_i \leq 1\}$

Denumim *lăţimea* unei mulţimi convexe $C \subseteq \mathbb{R}^n$ pe direcţia $v(\|v\| = 1)$:

$$W(C, v) = \sup_{z \in C} v^T z - \inf_{z \in C} v^T z.$$

Denumim *lăţimea* unei mulţimi convexe $C \subseteq \mathbb{R}^n$ pe direcţia $v(\|v\| = 1)$:

$$W(C, v) = \sup_{z \in C} v^T z - \inf_{z \in C} v^T z.$$

Lăţimea maximă şi minimă a lui C sunt date de:

$$W_{\max}(C) = \sup_{\|v\|=1} W(C, v) \qquad W_{\min}(C) = \inf_{\|v\|=1} W(C, v).$$

Denumim *lăţimea* unei mulţimi convexe $C \subseteq \mathbb{R}^n$ pe direcţia $v(\|v\| = 1)$:

$$W(C, v) = \sup_{z \in C} v^T z - \inf_{z \in C} v^T z.$$

Lăţimea maximă şi minimă a lui C sunt date de:

$$W_{\max}(C) = \sup_{\|v\|=1} W(C, v) \qquad W_{\min}(C) = \inf_{\|v\|=1} W(C, v).$$

Numărul de condiţionare al mulţimii C:

$$\kappa(C) = \frac{W_{\max}(C)^2}{W_{\min}(C)^2},$$

redă o măsură a excentricităţii mulţimii C.

Fie $E = \{v : (x-c)P^{-1}(x-c) \leq 1\}, P \succ 0$, atunci lăţimea mulţimii E în direcţia v este

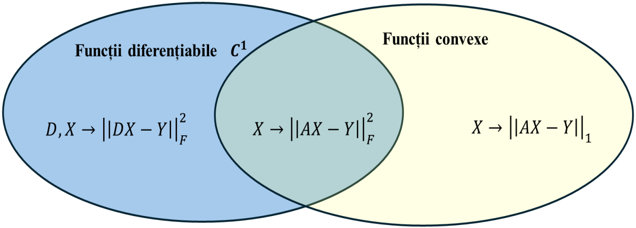
$$\sup_{z\in E} v^T z - \inf_{z\in E} v^T z = (\|P^{1/2}v\| + v^T c) - (-\|P^{1/2}v\| + v^T c)$$
$$= 2\|P^{1/2}v\|.$$

Atunci

$$W_{\max}(E) = 2\lambda_{\max}(P)^{1/2} \qquad W_{\min}(E) = 2\lambda_{\min}(P)^{1/2}.$$

şi

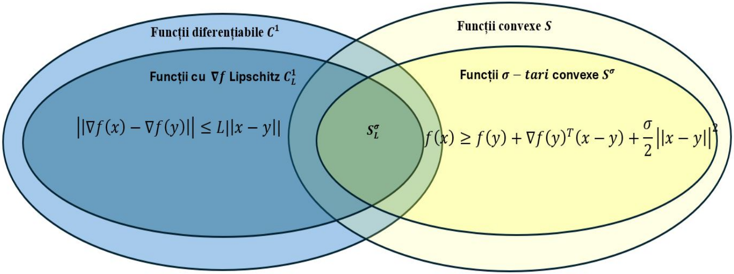
$$\kappa(E) = \frac{\lambda_{\max}(P)}{\lambda_{\min}(P)}.$$



- Funcţii diferenţiabile (∇f Lipschitz); e.g. $D, X \to \|DX - Y\|_F^2$
- Funcţii convexe diferenţiabile (∇f Lipschitz); e.g. $X \to \|AX - Y\|_F^2$
- Funcţii convexe nediferenţiabile; e.g. $X \to \|AX - Y\|_1$

Există clase mai ”puternice”?

Clase de funcţii



Reamintim definiţiile (echivalente) ale funcţiilor convexe $\mathcal{S}$:

- Ordin 0: $f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y), \quad \forall \alpha \in [0,1], x, y$

Reamintim definiţiile (echivalente) ale funcţiilor convexe $\mathcal{S}$:

- Ordin 0: $f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y), \;\; \forall \alpha \in [0,1], x, y$
- Ordin 1: $f(x) \geq f(y) + \nabla f(y)^T (x - y), \;\; \forall x, y$

Reamintim definiţiile (echivalente) ale funcţiilor convexe $\mathcal{S}$:

- Ordin 0: $f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y), \;\; \forall \alpha \in [0,1], x, y$
- Ordin 1: $f(x) \geq f(y) + \nabla f(y)^T(x-y), \;\; \forall x, y$
- Ordin 2: $\nabla^2 f(x) \succeq 0 \;\; \forall x$

Definiţie

O funcţie $f \in \mathcal{S}^{\sigma}$ este $\sigma-$tare convexă dacă satisface:

- *Ordin* 0*:*
  $f(\alpha x+(1-\alpha)y) \leq \alpha f(x)+(1-\alpha)f(y)-\alpha(1-\alpha)\frac{\sigma}{2}\|x-y\|^2, \;\; \forall \alpha \in [0,1], x, y$
- *Ordin* 1*:* $f(x) \geq f(y) + \nabla f(y)^T(x-y) + \frac{\sigma}{2}\|x-y\|^2, \;\; \forall x, y$
- *Ordin* 2*:* $\nabla^2 f(x) \succeq \sigma I_n \;\; \forall x$

Exemple:

- $x \to (c^T x)^2 + \frac{1}{2}\|x\|_2^2$
- $x \to \|Ax - b\|_2^2 + \|x\|_2^2$
- $x \to \frac{1}{2}x^T H x + q^T x, \;\; H \succ 0$

Considerăm $y = x^*$ (punct de optim):

$$f(x) \geq f(x^*) + \nabla f(x^*)^T (x - x^*) + \frac{\sigma}{2}\|x - x^*\|^2, \ \ \forall x$$
$$= f^* + \frac{\sigma}{2}\|x - x^*\|^2, \ \ \forall x.$$

Funcţia $f \in \mathcal{S}^{\sigma}$ are *creştere pătratică*:

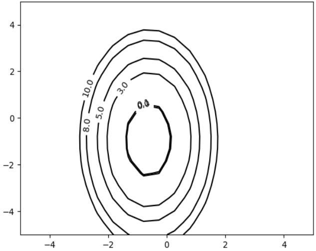
$$f(x) - f^* \geq \frac{\sigma}{2}\|x - x^*\|^2.$$

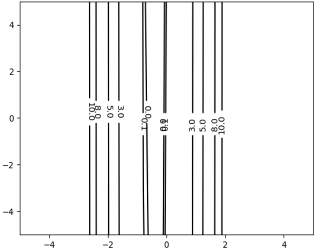
Dacă $f \in \mathcal{S}_L^{\sigma}$ atunci:

$$\frac{L}{2}\|x - x^*\|^2 \geq f(x) - f^* \geq \frac{\sigma}{2}\|x - x^*\|^2, \ \forall x.$$

Convexitate tare

$$H_1 = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix} \qquad H_2 = \begin{bmatrix} 4 & 0 \\ 0 & 10^{-}4 \end{bmatrix}$$





- Modele de optimizare continuă fără constrângeri. Condiţii de extrem
- Existenţă şi unicitate. Mulţimi izo(sub-)nivel. Convexitate tare
- **Metoda Gradient**

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?"

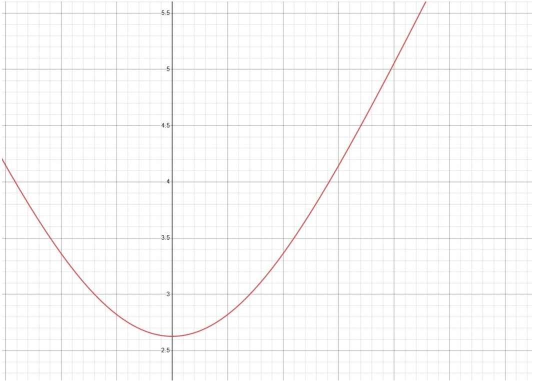
Algorithm 1: Algoritm de ordin I ($x^0, \epsilon, \ldots$):

Data: $k := 0$

```
while criteriu oprire = fals do
    Calculează: d^k ∈ span{∇f(x^k), ∇f(x^{k-1}), ⋯ , ∇f(x^0)}
    Actualizează x^{k+1} pe baza d^k şi {x^k, x^{k-1}, ⋯ , x^0}
    k := k + 1
end
```

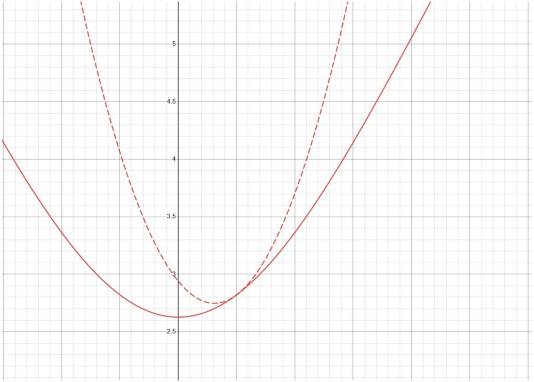
Majoritatea algoritmilor de ordin I realizează un model aproximativ local (pătratic) al funcţiei obiectiv $\mathcal{A}(x; x^k, \cdots, x^0; f)$. Aplicarea lui T reprezintă este echivalentă cu determinarea soluţiei optimă a acestuia, i.e

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



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 :

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

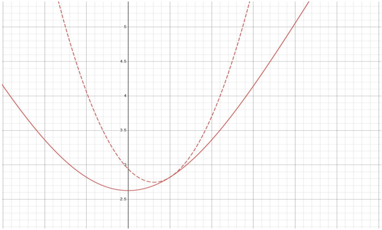


Figure: $f(x) = \ln\left(1 + e^{-2x+1}\right) + \ln\left(1 + e^{2x+1}\right)$; $x^0 = \frac{1}{2}$, $\alpha_0 = 1/4$

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 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 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 şi o aproximare a constantei L
- Fără o alegere limitată a pasului, MG poate diverge (exemplu!)
- Când pasul este variabil, inegalitatea descreşterii devine:

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

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, $f(x) = \frac{1}{1+\|x\|^2}$)

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!

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:

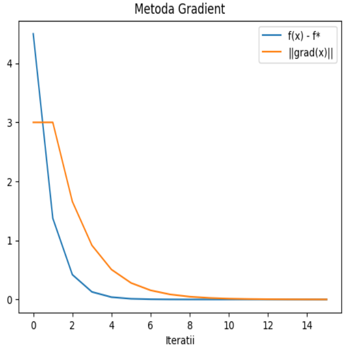
$$\lim_{k \to \infty} \nabla f(x^k) = 0,$$

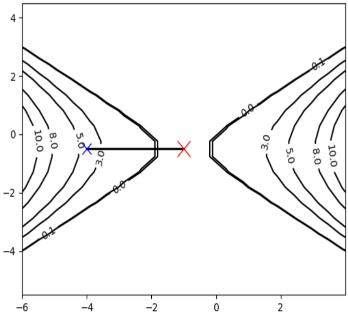
şi $f(x^{k+1}) \leq f(x^k)$.

- 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.

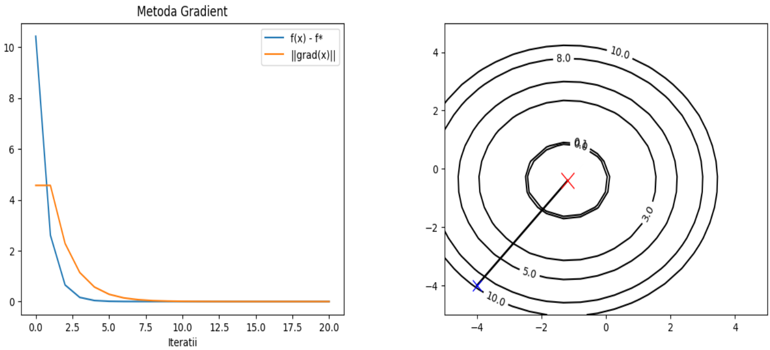
$$f(x^k) - f^* \leq \mathcal{O}\left((C/k)\right) < \epsilon \;\Rightarrow$$

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

- 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 $\mathcal{O}\left(C \cdot (\frac{1}{2})^{2^k}\right)$

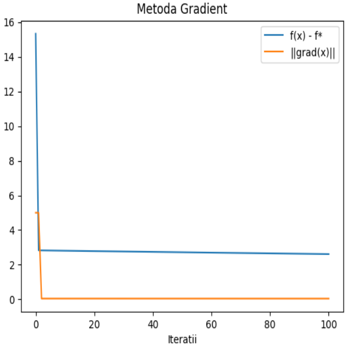
$$\min_{x \in \mathbb{R}^2} \frac{1}{2}\|Ax - b\|_2^2$$

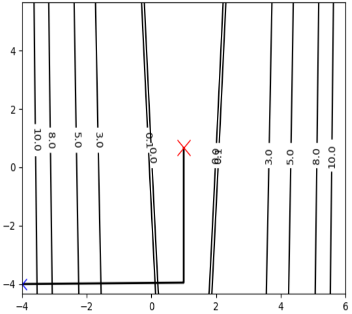
$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



$$\min_{x \in \mathbb{R}^2} \frac{1}{2}\|Ax - b\|_2^2$$

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 0.02 \end{bmatrix}$$





- 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.
- www.desmos.com