TEHNICI DE OPTIMIZARE
Curs 4

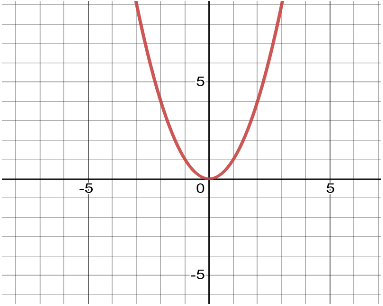
Andrei Pătraşcu

Departament Informatică
Universitatea din Bucureşti

Minimizare fără constrângeri:

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

- f convexă şi diferenţiabilă: $f(x) \geq f(y) + \nabla f(y)^T (x - y) \quad \forall x, y$
- Minime globale: $\nabla f(x^*) = 0, f(x^*) \leq f(x) \quad \forall x$



$$\min_{x \in \mathbb{R}^n} \; f(x) = \frac{1}{2}\|Ax - b\|_2^2, \;\; A \in \mathbb{R}^{m \times n}, m > n$$

La optimalitate: $A^T A x^* = A^T b$

$$\min_{x\in\mathbb{R}^n} \; f(x) = \frac{1}{2}\|Ax-b\|_2^2, \;\; A \in \mathbb{R}^{m\times n}, m > n$$

$$\begin{aligned} x^{k+1} &= x^k - \tau\nabla f(x^k) \\ &= x^k - \tau A^T(Ax^k - b) \\ &= \left(I - \tau A^T A\right) x^k + \tau A^T b \end{aligned}$$

Scădem x^* din ambele părţi:

$$\begin{aligned} x^{k+1} - x^* &= \left(I - \tau A^T A\right) x^k + \tau A^T b - x^* \\ &= \left(I - \tau A^T A\right)(x^k - x^*) + \tau A^T b - A^T A x^* \\ &= \left(I - \tau A^T A\right)(x^k - x^*). \end{aligned}$$

$$\min_{x \in \mathbb{R}^n} \quad f(x) = \frac{1}{2}\|Ax - b\|_2^2, \quad A \in \mathbb{R}^{m \times n}, m > n$$

Cheia convergenţei este $\|I - \tau A^T A\|$

$$\|x^{k+1} - x^*\| \leq \|I - \tau A^T A\| \|x^k - x^*\|$$

- O margine $\|I - \tau A^T A\|$ redusă implică o convergenţă mai bună!
- $\|I - \tau A^T A\|$ este minim pentru $\tau^* = \frac{2}{\lambda_{\max}(A^T A) + \lambda_{\min}(A^T A)}$ la valoarea

$$\|I - \tau^* A^T A\| = \frac{\lambda_{\max}(A^T A) - \lambda_{\min}(A^T A)}{\lambda_{\max}(A^T A) + \lambda_{\min}(A^T A)}$$

Convergenţă

$$\|x^k - x^*\| \leq \left(\frac{\lambda_{\max}(A^T A) - \lambda_{\min}(A^T A)}{\lambda_{\max}(A^T A) + \lambda_{\min}(A^T A)} \right)^k \|x^0 - x^*\|$$

$$\min_{x \in \mathbb{R}^n} \quad f(x) = \frac{1}{2}\|Ax - b\|_2^2, \quad A \in \mathbb{R}^{m \times n}, m > n$$

Convergenţa

$$\|x^k - x^*\| \leq \left(\frac{\lambda_{\max}(A^T A) - \lambda_{\min}(A^T A)}{\lambda_{\max}(A^T A) + \lambda_{\min}(A^T A)} \right)^k \|x^0 - x^*\|$$

Observăm că $\|x^k - x^*\|$ este descrescător şi:

$$\frac{\lambda_{\max}(A^T A) - \lambda_{\min}(A^T A)}{\lambda_{\max}(A^T A) + \lambda_{\min}(A^T A)} = \frac{1 - \frac{\lambda_{\min}(A^T A)}{\lambda_{\max}(A^T A)}}{1 + \frac{\lambda_{\min}(A^T A)}{\lambda_{\max}(A^T A)}} \approx 1 - \underbrace{\frac{\lambda_{\min}(A^T A)}{\lambda_{\max}(A^T A)}}_{1/\kappa(A)}$$

$$\kappa(A) = \frac{\lambda_{\max}(A^T A)}{\lambda_{\min}(A^T A)}$$ numărul de condiţionare al matricii $A^T A$

$$\min_{x \in \mathbb{R}^n} \quad f(x) = \frac{1}{2}\|Ax - b\|_2^2, \quad A \in \mathbb{R}^{m \times n}, m < n, A \text{ rang full pe linii}$$

$$\begin{aligned} x^{k+1} &= x^k - \tau \nabla f(x^k) \\ &= x^k - \tau A^T (Ax^k - b) \end{aligned}$$

Multiplicând ambele părţi cu A:

$$\begin{aligned} Ax^{k+1} - b &= Ax^k - b - \tau AA^T (Ax^k - b) \\ &= \left(I - \tau AA^T\right)(Ax^k - b) \end{aligned}$$

Convergenţa

$$\|Ax^k - b\| \leq \left(\frac{\lambda_{\max}(AA^T) - \lambda_{\min}(AA^T)}{\lambda_{\max}(AA^T) + \lambda_{\min}(AA^T)}\right)^k \|Ax^0 - b\|$$

$$\min_{x \in \mathbb{R}^n} \quad f(x) = \frac{1}{2}\|Ax - b\|_2^{1+\gamma}, \quad A \in \mathbb{R}^{m \times n}, \gamma \geq 0$$

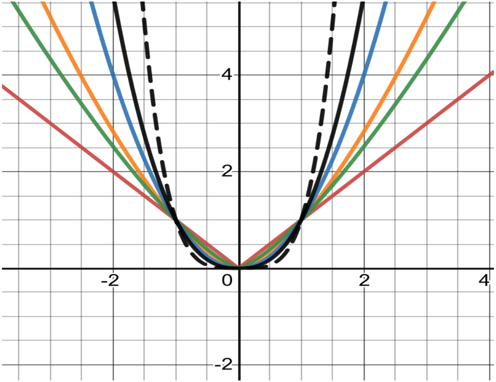


Figure: $|x|^\gamma, \gamma \in \{0, \frac{1}{2}, \frac{1}{3}, 1, \frac{3}{2}, 3\}$

$$\min_{x\in\mathbb{R}^n} \quad f(x) = \frac{1}{2}\|Ax-b\|_2^{1+\gamma}, \quad A\in\mathbb{R}^{m\times n}, \gamma\geq 0$$

$$\begin{aligned} x^{k+1} &= x^k - \tau\nabla f(x^k) \\ &= x^k - \tau(1+\gamma)\|Ax^k-b\|^{\gamma-1}A^T(Ax^k-b) \end{aligned}$$

Let $Ax^* = b, AA^T \succ 0, \underline{\lambda} = \lambda_{\min}(AA^T), \overline{\lambda} = \lambda_{\max}(AA^T)$

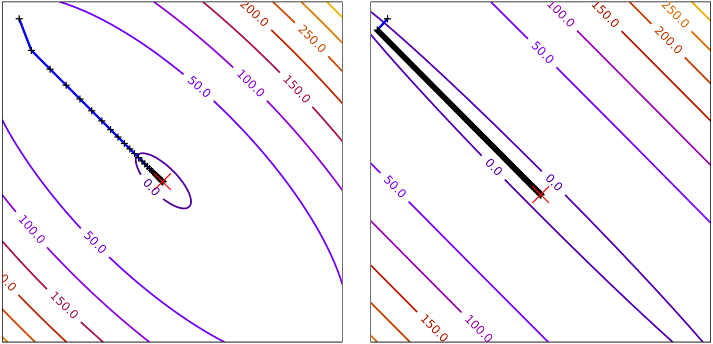
1. $\gamma\in(0,1)$ **["Not so smooth"]**

If $\|Ax^k-b\| < \left(\frac{\tau\underline{\lambda}^2}{2\overline{\lambda}}\right)^{\frac{1}{1-\gamma}}$ then $\|Ax^{k+1}-b\| > \|Ax^k-b\|$

2. $\gamma > 1$ **["Too smooth"]**

If $\|Ax^k-b\| > \left(\frac{2\overline{\lambda}}{\tau\underline{\lambda}^2}\right)^{\frac{1}{\gamma-1}}$ then $\|Ax^{k+1}-b\| > \|Ax^k-b\|$

$$\min_{x} \; \frac{1}{2} x^T Q x + q^T x$$



- Left: $\kappa = 7, K = 102, \epsilon = 10^{-7}, \tau = \frac{1}{L_f}$
- Right: $\kappa = 302, K = 3565, \epsilon = 10^{-7}, \tau = \frac{1}{L_f}$

Practical issues

- Stepsize:
  - ideal: $\tau_k = \min_{\tau > 0} \; f(x^k - \tau \nabla f(x^k))$
  - backtracking: decrease τ iteratively until

$$f(x^k - \tau \nabla f(x^k)) \leq f(x^k) - c\tau \|\nabla f(x^k)\|^2$$

- Unknown constants: L_f, σ_f
- Stopping criterion: $\|\nabla f(x^k)\| \leq \delta$ imply $\|x^k - x^*\| \leq \frac{\delta}{\sigma_f} (>> \delta)$

Informaţia asupra funcţiei obiectiv $\{f(x), \nabla f(x), \cdots\}$ este afectată de:

- erori de rotunjire (calcul), măsură, e.g $g(x^k) := \nabla f(x^k) + r^k$
- erori statistice (minimizarea riscului)

$$\min_x \; f(x) := \frac{1}{2}\|x\|^2 + \max_y F(x, y)$$

- Adesea $f, \nabla f$ sunt expresii ale soluţiei problemei de $\max$
- $\nabla f(x) = x + \nabla_x F(x, y(x))$

QP dimensiune n:

$$\min_x \quad \frac{1}{2}x^T Hx + c^T x \ \text{ s.l. } \ Ax \leq b.$$

Problema este convexă $H \succ 0$, m constrângeri liniare $\Rightarrow$ dualitate tare

$$\mathcal{L}(x, \lambda) = \frac{1}{2}x^T Hx + c^T x + \lambda^T (Ax - b)$$

$$\phi(\lambda) = \min_x \quad \frac{1}{2}x^T Hx + (c + A^T \lambda)^T x - \lambda^T b$$

Problema duală: QP dimensiune m

$$\max_{\lambda \geq 0} \quad \phi(\lambda)$$

Metoda Gradientului Dual:

$$\lambda^{k+1} = \pi_{\geq 0}(\lambda^k + \alpha \nabla \phi(\lambda^k))$$

- Calcularea $\nabla \phi(\lambda^k) = A \cdot \arg\min_x \mathcal{L}(x, \lambda^k)$ necesită soluţia unei probleme de minimizare auxiliare
- Aproximarea $s^k = \nabla \phi(\cdot) + r^k$ este cea mai realistă.
- Rămâne discutat calitatea necesară a aproximării.

Multe probleme de învăţare statistică se formulează:

$$f(x) = \mathbb{E}[F(x, \xi)] = \mathrm{int} F(x, \xi) dP(\xi),$$

unde $F(x, \xi)$ sunt funcţii cunoscute, dar distribuţia $P(\xi)$ este necunoscută.

- Calculul $f(x), \nabla f(x)$ este imposibil!
- Aproximăm f şi $\nabla f(x)$ prin $\frac{1}{m}\sum\limits_{i=1}^{m} F(x, \xi_i)$ şi $\frac{1}{m}\sum\limits_{i=1}^{m} \nabla F(x, \xi_i)$

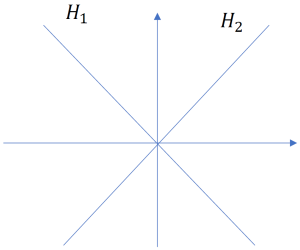
Exemple

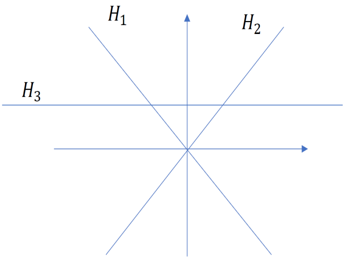
- Regresie liniară: $\min_x \mathbb{E}[(a_\xi x - b_\xi)^2]$;
  - Minimizarea riscului empiric: $\min_x \frac{1}{m}\sum_{\xi=1}^{m}(a_\xi x - b_\xi)^2$
- SVM: $\min_x \frac{1}{m}\sum_{\xi=1}^{m} \max\{0, a_\xi x - b_\xi\}^2 + \frac{\lambda}{2}\|x\|_2^2$

Când există o soluţie $\mathbb{E}[a_\xi x - b_\xi] = 0$ avem o problemă de interpolare liniară.

Interpolare

Definim $C = H_1 \cap \cdots \cap H_m \neq \emptyset$, unde $H_i = \{x : a_i^T x = b_i\}$. Determinaţi un punct din mulţimea C, i.e. $x \in C$.





$$\min_x f(x) := \mathbb{E}[F(x;\xi)],$$

$$f(x) \approx \frac{1}{N}\sum_i F(x;\xi_i) \;\; or \;\; F(x;\xi)$$

$$\nabla f(x) \approx \frac{1}{N}\sum_i \nabla F(x;\xi_i) \;\; or \;\; \nabla F(x;\xi)$$

Eroarea (varianţa): $V(x) = \mathbb{E}[\|\nabla f(x) - \nabla F(x;\xi)\|^2]$, $V^* = \mathbb{E}[\|\nabla F(x^*;\xi)\|^2]$
Metoda Gradient cu aproximare stohastică:

$$x^{k+1} = x^k - \tau_k \nabla F(x^{\cdot}\xi_k)$$

Zgomot stohastic - exemplu pătratic

$$\min_x f(x) := \mathbb{E}[(a_\xi x - b_\xi)^2]$$

- aproximări accesibile:

$$f(x) \approx \frac{1}{2N} \sum_i (a_{\xi_i} x - b_{\xi_i})^2 = \frac{1}{2N} \|Ax - b\|_2^2 \;\; sau \;\; \frac{1}{2}(a_\xi x - b_\xi)^2$$

$$\nabla f(x) \approx \frac{1}{N} A^T(Ax - b) \;\; sau \;\; a_\xi^T(a_\xi x - b_\xi)$$

- Cost: $\mathcal{O}(nN)$, în particular $\mathcal{O}(n)$

$$\min_x f(x) := \mathbb{E}[(a_\xi x - b_\xi)^2]$$

- aproximări accesibile:

$$f(x) \approx \frac{1}{2N}\sum_i (a_{\xi_i} x - b_{\xi_i})^2 = \frac{1}{2N}\|Ax - b\|_2^2 \;\; or \;\; \frac{1}{2}(a_\xi x - b_\xi)^2$$

$$\nabla f(x) \approx \frac{1}{N} A^T(Ax - b) \;\; or \;\; a_\xi^T(a_\xi x - b_\xi)$$

Metoda Gradient cu aproximare stohastică - MGS(SGD):

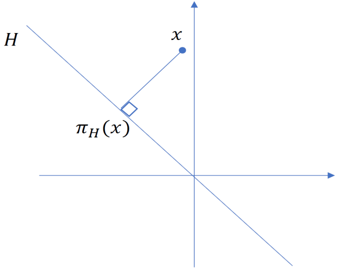
$$x^{k+1} = x^k - \tau_k a_{\xi_k}^T(a_{\xi_k} x^k - b_{\xi_k})$$

x_{SGD}^k este aproximarea stohastică a lui x_{GM}^k!

$$\min_x f(x) := \mathbb{E}[(a_\xi x - b_\xi)^2]$$

Proiecţia ortogonală pe hiperplanul $H = \{x : a^T x = b\}$ este punctul $\pi_H(x)$ din H cel mai "apropiat" de x.

$$\pi_H(x) = x - \frac{a^T x - b}{\|a\|^2} a \quad \left(= x - \frac{1}{\|a\|^2} a(a^T x - b) \right)$$



$$\min_{x \in \mathbb{R}^n} \quad \mathbb{E}[(a_\xi x - b_\xi)^2]$$

La optimalitate: $\mathbb{E}[a_\xi^T a_\xi] x^* = \mathbb{E}[a_\xi^T b_\xi]$

$$\begin{aligned} x^{k+1} &= x^k - \tau a_{\xi_k}^T (a_{\xi_k} x^k - b_{\xi_k}) \\ &= \left(I - \tau a_{\xi_k}^T a_{\xi_k}\right) x^k + \tau a_{\xi_k}^T b_{\xi_k} \end{aligned}$$

Prin expectanţă recuperăm Metoda Gradient! Scădem x^* din ambele părţi:

$$\begin{aligned} x^{k+1} - x^* &= \left(I - \tau a_{\xi_k}^T a_{\xi_k}\right) x^k + \tau a_{\xi_k}^T b_{\xi_k} - x^* \\ &= \left(I - \tau a_{\xi_k}^T a_{\xi_k}\right)(x^k - x^*) + \tau \underbrace{a_{\xi_k}^T (a_{\xi_k} x^* - b_{\xi_k})}_{\text{"zgomot interpolare"} = \nabla f(x^*) - \nabla F(x^*;\xi_k)} \end{aligned}$$

Varianţa la optim este crucială, indiferent de valoarea varianţei în alte puncte!

$$\min_{x \in \mathbb{R}^n} \quad \mathbb{E}[(a_\xi x - b_\xi)^2]$$

Interpolare: Pp că există soluţia $a_\xi x^* = b_\xi, \forall \xi$!

$$x^{k+1} - x^* = \left(I - \tau a_{\xi_k}^T a_{\xi_k}\right)(x^k - x^*) + \tau \underbrace{a_{\xi_k}^T (b_{\xi_k} - a_{\xi_k} x^*)}_{=0}$$

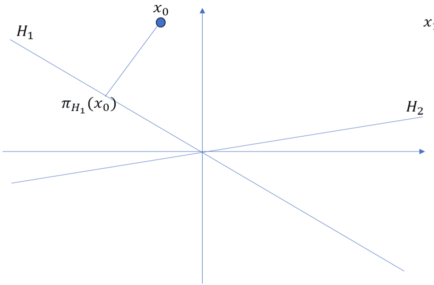
Pentru $\tau = \frac{1}{\|a_\xi\|^2}$, MGS se transformă în algoritmul Kaczmarz !

$$\min_{x \in \mathbb{R}^n} \quad \mathbb{E}[(a_\xi x - b_\xi)^2]$$

Interpolare: Pp că există soluţia $a_\xi x^* = b_\xi, \forall \xi$!

$$x^{k+1} - x^* = \left(I - \tau a_{\xi_k}^T a_{\xi_k}\right)(x^k - x^*) + \tau \underbrace{a_{\xi_k}^T (b_{\xi_k} - a_{\xi_k} x^*)}_{=0}$$

Pentru $\tau = \frac{1}{\|a_\xi\|^2}$, MGS se transformă în algoritmul Kaczmarz !



$$x_1 = x_0 - \frac{a_{(1)}^T x_0 - b_1}{\|a_{(1)}\|^2} a_{(1)}.$$

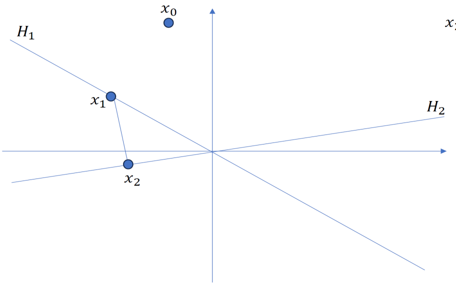
GM - caz patratic

$$\min_{x \in \mathbb{R}^n} \quad \mathbb{E}[(a_\xi x - b_\xi)^2]$$

Interpolare: Pp că există soluţia $a_\xi x^* = b_\xi, \forall \xi$!

$$x^{k+1} - x^* = \left(I - \tau a_{\xi_k}^T a_{\xi_k}\right)(x^k - x^*) + \tau \underbrace{a_{\xi_k}^T (b_{\xi_k} - a_{\xi_k} x^*)}_{=0}$$

Pentru $\tau = \frac{1}{\|a_\xi\|^2}$, MGS se transformă în algoritmul Kaczmarz !



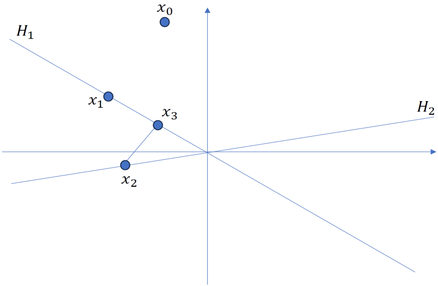
$$x_2 = x_1 - \frac{a_{(2)}^T x_1 - b_2}{\left\|a_{(2)}\right\|^2} a_{(2)}.$$

$$\min_{x \in \mathbb{R}^n} \quad \mathbb{E}[(a_\xi x - b_\xi)^2]$$

Interpolare: Pp că există soluţia $a_\xi x^* = b_\xi, \forall \xi$!

$$x^{k+1} - x^* = \left(I - \tau a_{\xi_k}^T a_{\xi_k}\right)(x^k - x^*) + \tau \underbrace{a_{\xi_k}^T (b_{\xi_k} - a_{\xi_k} x^*)}_{=0}$$

Pentru $\tau = \frac{1}{\|a_\xi\|^2}$, MGS se transformă în algoritmul Kaczmarz !



$$x_3 = x_2 - \frac{a_{(1)}^T x_2 - b_1}{\left\|a_{(1)}\right\|^2} a_{(1)}.$$

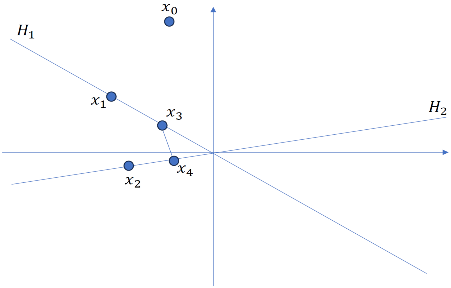
GM - caz patratic

$$\min_{x \in \mathbb{R}^n} \quad \mathbb{E}[(a_\xi x - b_\xi)^2]$$

Interpolare: Pp că există soluţia $a_\xi x^* = b_\xi, \forall \xi$!

$$x^{k+1} - x^* = \left(I - \tau a_{\xi_k}^T a_{\xi_k}\right)(x^k - x^*) + \tau \underbrace{a_{\xi_k}^T (b_{\xi_k} - a_{\xi_k} x^*)}_{=0}$$

Pentru $\tau = \frac{1}{\|a_\xi\|^2}$, MGS se transformă în algoritmul Kaczmarz !



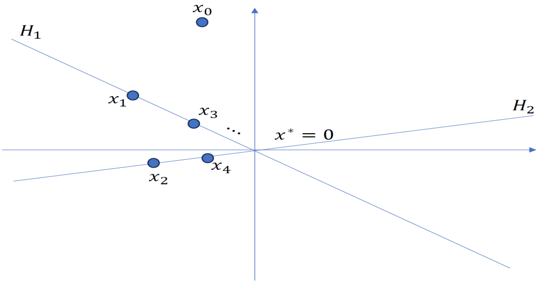
$$x_4 = x_3 - \frac{a_{(2)}^T x_3 - b_2}{\left\|a_{(2)}\right\|^2} a_{(2)}.$$

$$\min_{x \in \mathbb{R}^n} \quad \mathbb{E}[(a_\xi x - b_\xi)^2]$$

Interpolare: Pp că există soluţia $a_\xi x^* = b_\xi, \forall \xi$!

$$x^{k+1} - x^* = \left(I - \tau a_{\xi_k}^T a_{\xi_k}\right)(x^k - x^*) + \tau \underbrace{a_{\xi_k}^T (b_{\xi_k} - a_{\xi_k} x^*)}_{=0}$$

Pentru $\tau = \frac{1}{\|a_\xi\|^2}$, MGS se transformă în algoritmul Kaczmarz !



$$\min_{x \in \mathbb{R}^n} \quad \mathbb{E}[(a_\xi x - b_\xi)^2]$$

Interpolare: Pp că există soluţia $a_\xi x^* = b_\xi, \forall \xi$!

$$x^{k+1} - x^* = \left(I - \tau a_{\xi_k}^T a_{\xi_k}\right)(x^k - x^*) + \tau \underbrace{a_{\xi_k}^T (b_{\xi_k} - a_{\xi_k} x^*)}_{=0}$$

Pentru $\tau = \frac{1}{\|a_\xi\|^2}$, MGS se transformă în algoritmul Kaczmarz !

Algoritmul Kaczmarz

1. Alegem ξ_k cu probabilităţile $p_{\xi_k} = \dfrac{\|a_{\xi_k}\|^2}{\|A\|_F^2}$

2. $x^{k+1} = x^k - \dfrac{1}{\|a_{\xi_k}\|^2} a_{\xi_k}^T (a_{\xi_k} x^k - b_{\xi_k}) := \pi_{H_k}(x^k)$

3. $k = k + 1$

$$\min_{x \in \mathbb{R}^n} \quad \mathbb{E}[(a_\xi x - b_\xi)^2]$$

Interpolare: Pp că există soluţia $a_\xi x^* = b_\xi, \forall \xi$!

Algoritmul Kaczmarz

1. Alegem ξ_k cu probabilităţile $p_{\xi_k} = \dfrac{\|a_{\xi_k}\|^2}{\|A\|_F^2}$

2. $x^{k+1} = x^k - \dfrac{1}{\|a_{\xi_k}\|^2} a_{\xi_k}^T (a_{\xi_k} x^k - b_{\xi_k}) := \pi_{H_k}(x^k)$

3. $k = k + 1$

OBS1: $x^k - x^{k+1} [\in \mathrm{span}(a_{\xi_k}^T)] \perp x^{k+1} - x^* [\in \mathrm{span}(a_{\xi_k}^T)^\perp]$

$$\|x^{k+1} - x^*\|^2 = \|x^k - x^*\|^2 - \|x^{k+1} - x^k\|^2$$

OBS2: $\mathbb{E}_{\xi_k}[\|x^{k+1} - x^k\|^2] = \mathbb{E}_{\xi_k}[\|a_{\xi_k}^T (a_{\xi_k} x^k - b_{\xi_k})\|^2] \geq \frac{1}{\kappa(A)^2} \|x^k - x^*\|^2$

$$\min_{x \in \mathbb{R}^n} \quad \mathbb{E}[(a_\xi x - b_\xi)^2]$$

By taking expectation:

$$\mathbb{E}[\|x^{k+1} - x^*\|^2] \leq \mathbb{E}[\|x^k - x^*\|^2] - \frac{1}{\kappa(A)^2}\mathbb{E}[\|x^k - x^*\|^2]$$

Sthromer & Vershynin, 2009

Fie $Ax^* = b$, atunci algoritmul Kaczmarz converge conform:

$$\mathbb{E}[\|x^{k+1} - x^*\|^2] \leq \left(1 - \frac{1}{\kappa(A)^2}\right)^k \|x^0 - x^*\|^2$$

- rata de convergenţă: $\sqrt{1 - \frac{1}{\kappa(A)^2}}$ $\left(vs. \approx 1 - \frac{1}{\kappa(A)} \text{ for GM} \right)$
- Observăm convergenţa în expectanţă (vs. convergenţa deterministă pentru MG)

$$\min_{x \in \mathbb{R}^n} \quad \|Ax - b\|_2^2, \;\; m = 700$$

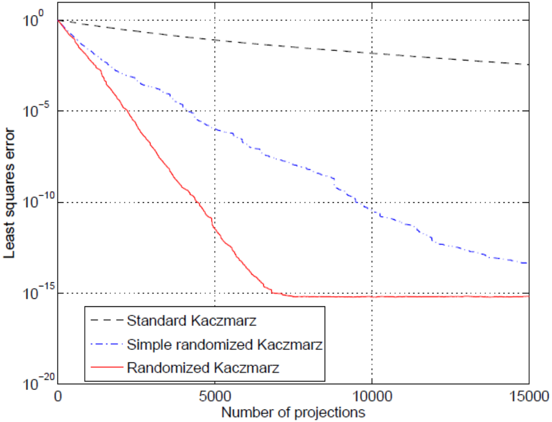


Figure: Comparison from Sthromer & Vershynin, 2009

$$\min_x \; f(x) := \mathbb{E}[F(x;\xi)]$$

Algoritm SGD: pentru $k \geq 1$

1. Alege aleator $\xi_k \in \Omega$
2. $x^{k+1} = x^k - \tau_k \nabla F(x^k;\xi_k)$

Teoremă (Ma et.al, 2018, Needle et al., 2015)

Fie $F(\cdot;\xi)$ funcţii σ_f-tare convexe, $L_{F,\xi}-$netedă şi presupunem $\nabla f(x^*;\xi) = 0, \forall \xi$. Atunci SGD satisface:

$$\mathbb{E}[\|x^k - x^*\|^2] \leq \left(1 - \frac{\sigma_f}{\sup_\xi L_{F,\xi}}\right)^k \|x^0 - x^*\|^2$$

- rata de convergenţă: $1 - \frac{\sigma_f}{\sup_\xi L_{F,\xi}}$ $\left(vs. \approx 1 - \frac{\sigma_f}{L_f} \text{ for GM}\right)$
- Pe funcţii pătratice şi $\xi \in \{1,\cdots,m\}$: $\mathcal{O}(n)$ per iteraţie MGS (vs. $\mathcal{O}(mn)$ per iteraţie MG)
- Convergenţa în expectanţă (vs. convergenţa deterministă pentru MG)

$$\min_{x\in\mathbb{R}^n} \quad \mathbb{E}[(a_\xi x - b_\xi)^2]$$

Presupunem că relaţia de interpolare nu are loc!

Luăm $\tau = \frac{1}{\|a_\xi\|^2}$ atunci *blue* $\perp$ *red*:

$$x^{k+1} - x^* = \left(I - \frac{1}{\|a_{\xi_k}\|^2} a_{\xi_k}^T a_{\xi_k}\right)(x^k - x^*) + \frac{1}{\|a_{\xi_k}\|^2} a_{\xi_k}^T (a_{\xi_k} x^* - b_{\xi_k})$$

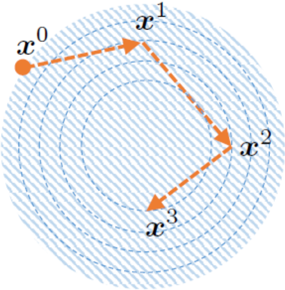
Atunci:

$$\|x^{k+1} - x^*\|^2 = \|\left(I - \frac{1}{\|a_{\xi_k}\|^2} a_{\xi_k}^T a_{\xi_k}\right)(x^k - x^*)\|^2 + \underbrace{(a_{\xi_k} x^* - b_{\xi_k})^2}_{V^*}$$

Convergenţă doar către o vecinătate a mulţimii optime!

$$\min_{x\in\mathbb{R}^n} \quad \mathbb{E}[(a_\xi x - b_\xi)^2]$$

$$\|x^{k+1} - x^*\|^2 = \left\| \left(I - \frac{1}{\|a_{\xi_k}\|^2} a_{\xi_k}^T a_{\xi_k} \right) (x^k - x^*) \right\|^2 + \underbrace{(a_{\xi_k} x^* - b_{\xi_k})^2}_{V^*}$$



Remediu: alegerea τ_k descrecător implică deplasarea arbitrar de aproape de optim!

$$\min_x \; f(x) := \mathbb{E}[F(x;\xi)]$$

Algoritm SGD: for $k \geq 1$

1. Alege aleator $\xi_k \in \Omega$
2. $x^{k+1} = x^k - \tau_k \nabla F(x^k;\xi_k)$

Teoremă (Nguyen et.al, 2018)

Fie $F(\cdot;\xi)$ funcţii σ_f-tare convexe, $L_{F,\xi}-$netedă. Atunci SGD satisface:

$$\mathbb{E}[\|x^k - x^*\|^2] \leq \mathcal{O}\left(\frac{V^*}{\sigma_f^2 k}\right)$$

- rată subliniară: $\mathcal{O}\left(\frac{V^*}{\sigma_f^2 k}\right)$ $\left(vs. \approx 1 - \frac{\sigma_f}{L_f} \text{ for GM}\right)$
- Pe funcţii pătratice şi $\xi \in \{1,\cdots,m\}$:
  $\mathcal{O}(n)$ per iteraţie MGS (vs. $\mathcal{O}(mn)$ per iteraţie MG)
- Convergenţa în expectanţă (vs. convergenţa deterministă pentru MG)
- Convergence in expectation (vs. deterministic convergence for GM)

$$\min_x \; f(x) := \mathbb{E}[F(x;\xi)]$$

unde f este $\sigma-$tare convexă, cu ∇f $L-$continuu Lipschitz.

- $F(\cdot;\xi)$ funcţii convexe diferenţiabile
- Notăm $X^* = \{x^* : f(x^*) = \min_x f(x)\}$
- În general $\nabla F(x^*;\xi) \neq 0$
- Exemplu: alegem aleator uniform ξ şi aproximăm $g(x) := \nabla F(x;\xi)$.

$$MGS(SGD): \quad x^{k+1} := x^k - \alpha_k g(x^k) = x^k - \alpha_k \nabla F(x;\xi)$$

$$\min_x \; f(x) := \mathbb{E}[F(x;\xi)]$$

unde f este $\sigma-$tare convexă, cu ∇f $L-$continuu Lipschitz.
Zgomot aleator:

- absolut: $\mathbb{E}[r^k] = 0, E\|r^k\|^2 \leq \Sigma^2$
- absolut la optim: $\mathbb{E}[r^k] = 0, E\|g(x^*)\|^2 \leq \Sigma^2$
- relativ: $\mathbb{E}[r^k] = 0, E\|r^k\|^2 \leq \tau\|\nabla f(x^k)\|^2$

$$MG-S: \quad x^{k+1} := x^k - \alpha_k g(x^k), \qquad g(x^k) := \nabla f(x^k) + r^k$$

Zgomot aleator

Algorithm 1: Metoda Gradient Stohastic (x^0, ϵ, N):

Data: $k := 0, \{\alpha_k\}_{k \geq 0}$

while criteriu oprire = fals **do**

Alege aleator $\{\xi_1^k, \cdots, \xi_N^k\}$

Calculează: $\{\nabla F(x^k; \xi_1^k), \nabla F(x^k; \xi_2^k), \cdots, \nabla F(x^k; \xi_N^k)\}$

Actualizează $x^{k+1} = x^k - \alpha_k \frac{1}{N} \sum_{i=1}^{N} \nabla F(x^k; \xi_i)$

$k := k + 1$

end

$$\min_x \; \frac{1}{2}\|Ax - b\|_2^2 = \mathbb{E}\left[\frac{1}{2}(a_i^T x - b_i)^2\right], \; A \in \mathbb{R}^{70\times 50}$$

In acest caz

$$\nabla F(x^k; \xi_i) = a_{\xi_i}(a_{\xi_i}^T x - b_{\xi_i})$$

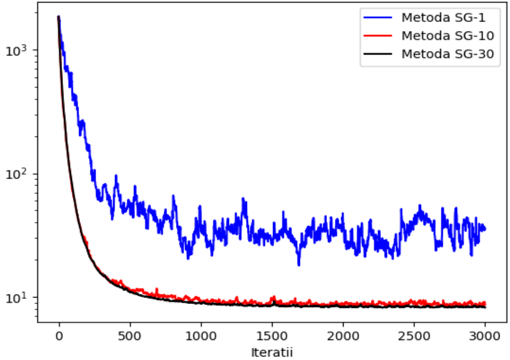
$$MGS(x^0, \epsilon, 1): \; x^{k+1} := x^k - \alpha a_\xi (a_\xi^T x^k - b_\xi)$$

$$MGS(x^0, \epsilon, N): \; x^{k+1} := x^k - \alpha \frac{1}{N}\sum_{i=1}^{N} a_{\xi_i}(a_{\xi_i}^T x - b_{\xi_i})$$

Zgomot aleator

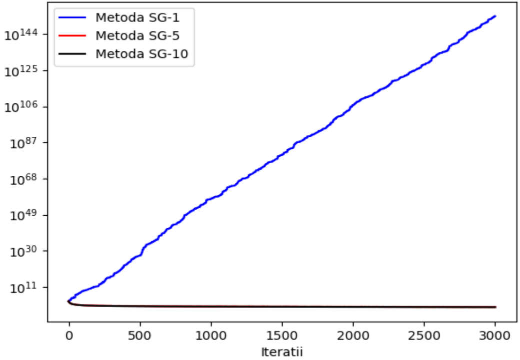
$$\min_x \ \frac{1}{2}\|Ax - b\|_2^2 = \mathbb{E}\left[\frac{1}{2}(a_i^T x - b_i)^2\right], \ A \in \mathbb{R}^{70\times 50}$$

MGS : Alegem aleator $\{\xi_1, \cdots, \xi_N\}$, $\quad x^{k+1} := x^k - \underbrace{\frac{1}{3L}}_{\alpha} \underbrace{\frac{1}{N}\sum_{i=1}^{N} \nabla F(x^k; \xi_i)}_{g(x^k)}$



$$\min_x \frac{1}{2}\|Ax - b\|_2^2 = \mathbb{E}\left[\frac{1}{2}(a_i^T x - b_i)^2\right]$$

$$MGS: \ \text{Alegem aleator } \{\xi_1, \cdots, \xi_N\}, \quad x^{k+1} := x^k - \sqrt{\frac{2}{L}}\left(\frac{1}{N}\sum_{i=1}^{N} \nabla F(x^k; \xi_i)\right)$$



Ipoteza zgomot absolut (IZA): $\mathbb{E}[r^k] = 0, E\|r^k\|^2 \leq \Sigma^2$.
IZA implică $\mathbb{E}[g(x^k)] = \nabla f(x^k)$.

Teoremă

Sub IZA, considerăm $\alpha_k := \alpha$ *cu* $0 < \alpha < \frac{1}{2L}$. *Atunci* $\{x^k\}_{k\geq 0}$ *satisface:*

$$\mathbb{E}[f(x^k) - f^*] \leq (1 - \alpha\sigma(1/2 - L\alpha))^k \left(f(x^0) - f^*\right) + \frac{\Sigma^2}{\sigma}\frac{L\alpha}{1/2 - L\alpha}$$

$$\mathbb{E}[\|x^k - x^*\|^2] \leq (1 - \alpha\sigma(1/2 - L\alpha))^k \frac{2(f(x^0) - f^*)}{\sigma} + \frac{\Sigma^2}{\sigma^2}\frac{L\alpha}{1/2 - L\alpha},$$

unde x^* *punctul de minim al funcţiei* f.

- MGS nu converge la minim cu α_k constant!
- Observăm convergenţa într-o vecinătate (sau $S_f(f^* + \frac{\delta^2}{2\sigma})$) a minim-ului, a cărei dimensiune depinde de δ

Demonstraţie: Două inegalităţi utile:

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

$$\mathbb{E}[\|g(x^k)\|^2] \leq 2\mathbb{E}[\|\nabla f(x^k)\|^2] + 2\mathbb{E}[\|r^k\|^2]. \tag{2}$$

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

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

Evaluăm expectanţa în ambele părţi:

$$\begin{aligned}
\mathbb{E}[f(x^{k+1})] &\leq \mathbb{E}[f(x^k)] - \alpha \mathbb{E}[\nabla f(x^k)^T g(x^k)] + \frac{L\alpha^2}{2}\mathbb{E}[\|g(x^k)\|^2] \\
&= \mathbb{E}[f(x^k)] - \alpha \frac{1}{2}\|\nabla f(x^k)\|^2 + \frac{L\alpha^2}{2}\mathbb{E}[\|g(x^k)\|^2] \\
&\overset{(2)}{\leq} \mathbb{E}[f(x^k)] - \alpha \frac{1}{2} E[\|\nabla f(x^k)\|^2] + L\alpha^2 \mathbb{E}[\|\nabla f(x^k)\|^2] + L\alpha^2 \mathbb{E}[\|r^k\|^2]
\end{aligned}$$

$$MG: \quad x^{k+1} := x^k - \alpha_k g(x^k), \qquad g(x^k) := \nabla f(x^k) + r^k$$

Demonstraţie(continuare): folosim marginea absolută a zgomotului r^k

$$\begin{aligned}
\mathbb{E}[f(x^{k+1})] &\leq \mathbb{E}[f(x^k)] + (L\alpha^2 - \alpha/2)\mathbb{E}[\|\nabla f(x^k)\|^2] + L\alpha^2\mathbb{E}[\|r^k\|^2] \\
&\leq \mathbb{E}[f(x^k)] + (L\alpha^2 - \alpha/2)\mathbb{E}[\|\nabla f(x^k)\|^2] + L\alpha^2\Sigma^2.
\end{aligned}$$

În final, scădem f^* şi deducem primul rezultat:

$$\begin{aligned}
\mathbb{E}[f(x^{k+1}) - f^*] &\overset{(1)}{\leq} \mathbb{E}[f(x^k) - f^*] - \sigma\alpha(1/2 - L\alpha)\mathbb{E}[f(x^k) - f^*]) + L\alpha^2\Sigma^2 \\
&\leq [1 - \sigma\alpha(1/2 - L\alpha)]\mathbb{E}[f(x^k) - f^*] + L\alpha^2\Sigma^2.
\end{aligned}$$

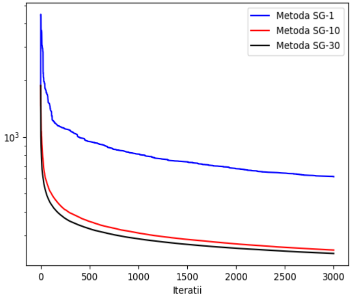
Al doilea reiese din creşterea pătratică a funcţiei f.

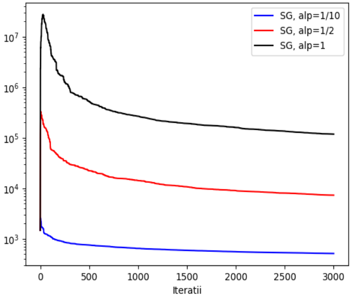
Teoremă

Sub IZA, considerăm $\alpha_k := \frac{\alpha}{k}$ *Atunci* $\{x^k\}_{k\geq 0}$ *satisface:*

$$\mathbb{E}[f(x^k) - f^*] \leq \mathcal{O}\left(\frac{V^*}{k}\right) \qquad \mathbb{E}[\|x^k - x^*\|^2] \leq \mathcal{O}\left(\frac{V^*}{k}\right),$$

unde x^* *punctul de minim al funcţiei* f.





$$MG: \quad x^{k+1} := x^k - \alpha_k g(x^k), \qquad g(x^k) := \nabla f(x^k) + r^k$$

Ipoteza zgomot relativ (IZR): $\mathbb{E}[r^k] = 0, \mathbb{E}[\|r^k\|^2] \leq \delta \mathbb{E}[\|\nabla f(x^k)\|^2], \quad \delta < 1.$

Teoremă

Sub IZR, alegem $\alpha_k := \alpha$. *Atunci pentru* $0 < \alpha < \frac{1}{2L(1+\delta)}$, *şirul* x^k *converge către* x^* *cu rată de convergenţă liniară:*

$$\|x^k - x^*\|^2 \leq (1 - \tilde{\alpha}\sigma)^k \frac{2(f(x^0) - f^*)}{\sigma},$$

unde $\tilde{\alpha} = 2\alpha\left[\frac{1}{2} - L\alpha(1+\delta)\right]$.

Problemă interpolare:

$$\min_x f(x) := \mathbb{E}[F(x;\xi)]$$

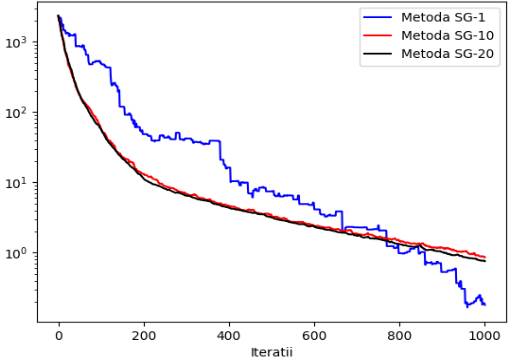
- $F(\cdot;\xi)$ funcţii convexe diferenţiabile
- $\nabla F(x^*;\xi) = 0 \quad \forall \xi$

Exemple:

- $[m \leq n] \quad \min_x \frac{1}{m}\sum_{i=1}^{m} |a_i^T x - b_i|^p$
- $[\exists x^* : Ax^* \leq b] \quad \min_x \frac{1}{m}\sum_{i=1}^{m}([a_i^T x - b_i]_+)^2$
- Pentru funcţii reziduale generale, $\min_x \mathcal{L}(x) := \mathbb{E}[\mathcal{L}(x; a_\xi, b_\xi)]$, unde $\mathcal{L}^* = 0$.

$$\min_x \frac{1}{2}\max\{0, Ax - b\}^2 = \mathbb{E}\left[\frac{1}{2}([a_i^T x - b_i]_+)^2\right]$$

MGS : Alegem aleator $\{\xi_1, \cdots, \xi_N\}$, $\quad x^{k+1} := x^k - \sqrt{\frac{1}{2L}}\left(\frac{1}{N}\sum_{i=1}^{N} \nabla F(x^k; \xi_i)\right)$



$$\min_x \frac{1}{2}\max\{0, Ax - b\}^2 = \mathbb{E}\left[\frac{1}{2}([a_i^T x - b_i]_+)^2\right]$$

$$MGS: \text{ Alegem aleator } \{\xi_1, \cdots, \xi_N\}, \quad x^{k+1} := x^k - \sqrt{\frac{1}{2L}}\left(\frac{1}{N}\sum_{i=1}^{N} \nabla F(x^k; \xi_i)\right)$$

- MGS converge la optim cu pas constant (limitat)!
- Rata de convergenţă este liniară (funcţie de κ)

Metoda Gradient - (Stohastic)

Concluzii:

- MGS cu pas constant converge într-o vecinătate a optimului
- În general, dimensiunea vecinătăţii de convergenţă depinde de Σ, α şi κ.
- Pentru convergenţă la optim: (*i*) pas descrescător $\mathcal{O}\left(1/k\right)$; (*ii*) ipoteza de interpolare
- Compromis: alegerea pasului α vs. acurateţea estimării ∇f