Optimisation non-convexe
16 octobre 2020

$f : \mathbb{R}^n \to \mathbb{R}$

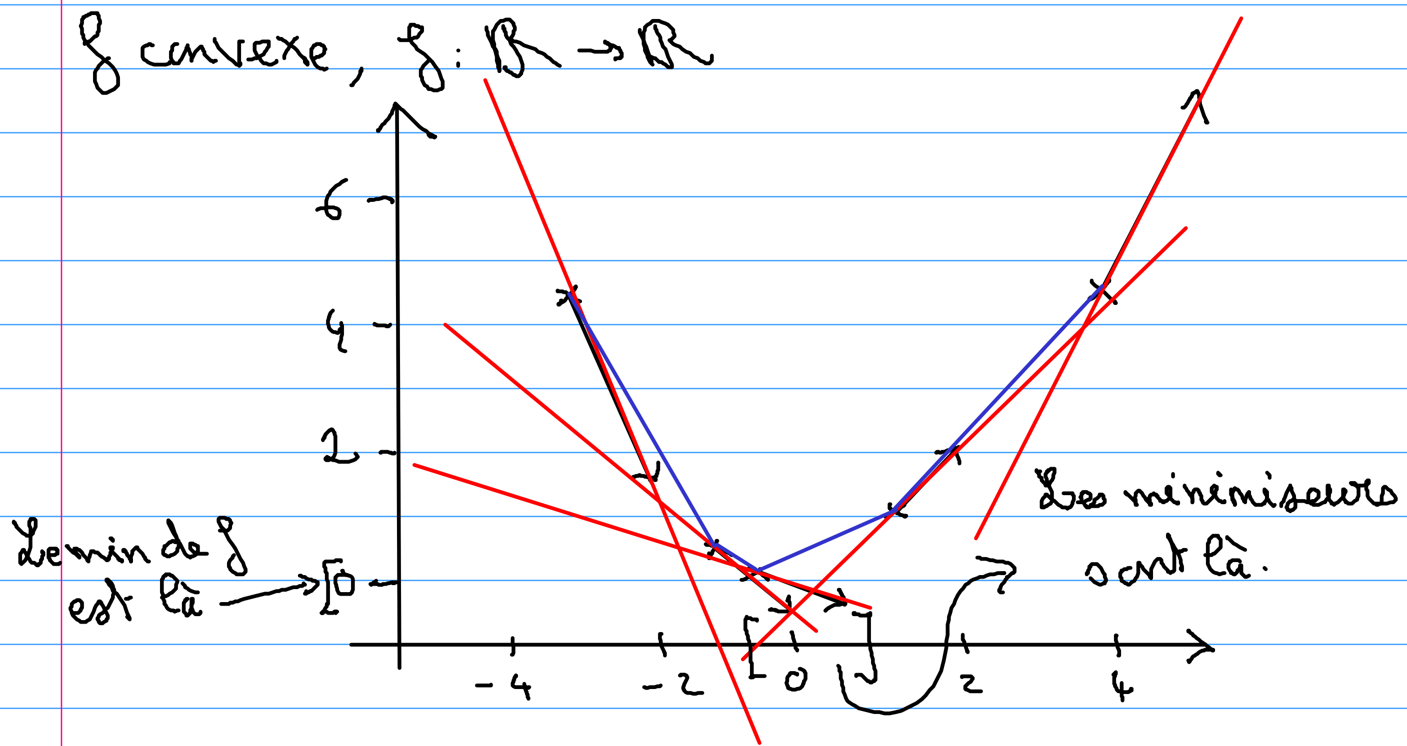
Objectif : trouver x_* tq $f(x_*) = \min_{x \in \mathbb{R}^n} f(x)$.

Hypothèses : - x_* existe ;
- f est $\mathcal{C}^\infty$.

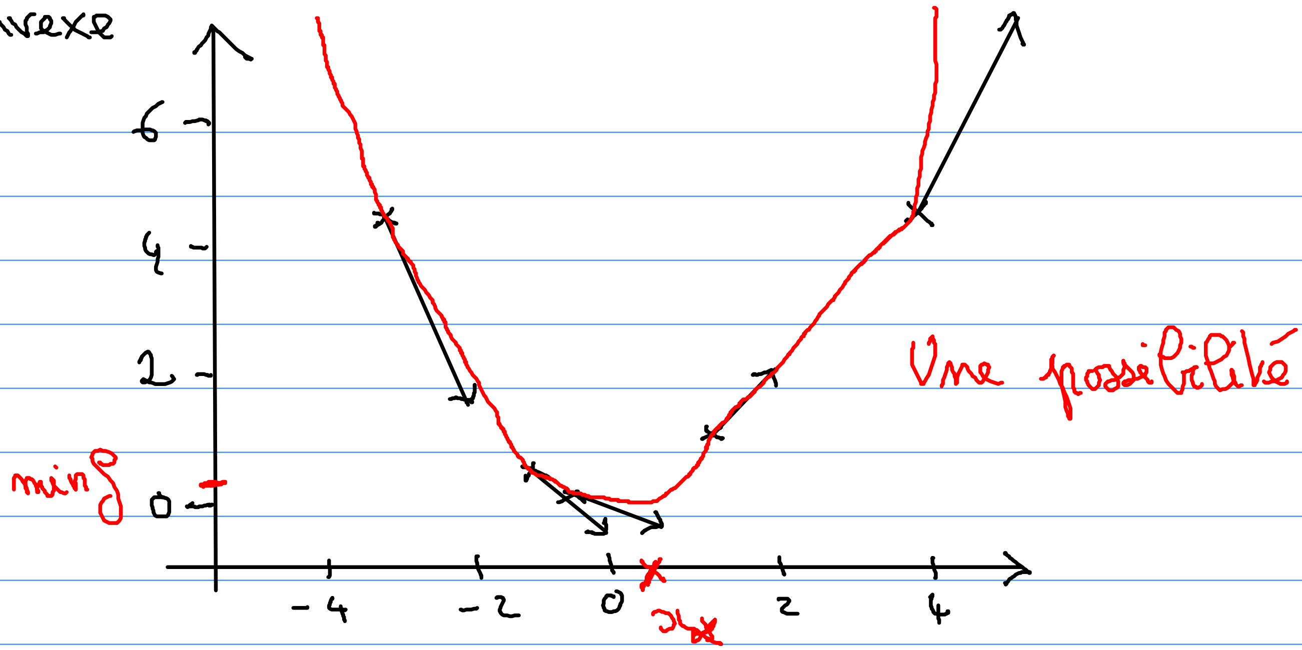
Aujourd'hui : f non-convexe.

1.1) Difficulté de l'optim. non-convexe

f convexe, $f : \mathbb{R} \to \mathbb{R}$



f non-convexe



Intuitivement, pour résoudre approximativement le problème il faut calculer la valeur de f sur un ensemble de points assez proches les uns des autres.

$\Rightarrow$ Inenvisageable en grande dimension.

Soyons plus modestes et demandons-nous quel problème plus simple on pourrait résoudre.

Première idée : minimum local ?
En fait, aussi très difficile.

II Points critiques

Prop: Soit x un minimiseur local de f. Alors :

(i) $\nabla f(x) = 0$;

(ii) $\mathrm{Hess}\, f(x) \succeq 0$.

Réciproquement, si $x \in \mathbb{R}^n$ est tel que
$\nabla f(x) = 0$ et $\mathrm{Hess}\, f(x) \succ 0$,
alors x est un minimiseur local de f.

Déf: soit $x \in \mathbb{R}^n$. On dit que x est

- un point critique d'ordre 1 si
$$\nabla f(x) = 0 \,;$$
- un point critique d'ordre 2 si
$$\nabla f(x) = 0 \text{ et } \mathrm{Hess}\, f(x) \succeq 0.$$

Ex: soit $f : \mathbb{R}^2 \to \mathbb{R}$
$$(x_1, x_2) \to x_1^2 - x_2^2.$$
Pour tout $x = (x_1, x_2)$, $\nabla f(x) = (2x_1, -2x_2)$.
Seul point critique d'ordre 1 : $(0,0)$.
Est-ce un point critique d'ordre 2 ? Non.

$$\mathrm{Hess}\, f(0,0) = \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix} \not\succeq 0$$

Motivation pour trouver un point critique de f : pour beaucoup de fonctions (pas toutes !), les points critiques d'ordre 2 sont des minimiseurs locaux ;
pour certaines fonctions, les points critiques d'ordre 2 sont exactement les minimiseurs, ou des minimiseurs approchés.

III) Descente de gradient

Hypothèse : f est L-lisse pour un $L>0$
$(\forall x,y,\ \|\nabla f(x)-\nabla f(y)\| \leq L\|x-y\|)$

Descente de gradient à pas constant $1/L$:
$$x_{t+1} = x_t - \frac{1}{L}\nabla f(x_t).$$

3.1) Points critiques d'ordre 1

Thm : Soit $T\in\mathbb{N}$. On calcule T étapes de descente de gradient : $x_0, x_1, \dots, x_T$.
On définit T_{min} tq $\|\nabla f(x_{T_{min}})\| = \min_{k\leq T}\|\nabla f(x_k)\|$.
On fait renvoyer $x_{T_{min}}$ à la descente.

Alors $\|\nabla f(x_{T_{min}})\| \leq \sqrt{\dfrac{2L(f(x_0)-f(x_*))}{T}}$. $\Big)\ O\left(\frac{1}{\sqrt{T}}\right)$

Dém : $\forall t,\ f(x_{t+1}) \leq f(x_t) - \frac{1}{2L}\|\nabla f(x_t)\|^2$ (vu le 15 octobre)
$$\sum_{t=0}^{T-1}\|\nabla f(x_t)\|^2 \leq \sum_{t=0}^{T-1} 2L\left(f(x_t)-f(x_{t+1})\right)$$
$$= 2L\left(f(x_0)-f(x_T)\right)$$

$$\leq 2L(f(x_0) - f(x_*)).$$

$$\geq T \times \min_{t \leq T} \|\nabla f(x_t)\|^2 = T \|\nabla f(x_{T_{min}})\|^2.$$

Donc $\|\nabla f(x_{T_{min}})\|^2 \leq \dfrac{2L(f(x_0) - f(x_*))}{T}$

$$\Rightarrow \|\nabla f(x_{T_{min}})\| \leq \sqrt{\frac{2L(f(x_0) - f(x_*))}{T}}. \qquad \square$$

3.2) Points critiques d'ordre 2

Rq: la descente de gradient ne converge pas nécessairement vers un point critique d'ordre 2.

Ex: Si x_0 est un point critique d'ordre 1 mais pas d'ordre 2,
$x_1 = x_0 - \frac{1}{L}\nabla f(x_0) = x_0$
$x_2 = x_0$, $x_3 = x_0$, ...

Thm [Lee, Simchowitz, Jordan, Recht, 2016]
Hypothèses: • f a un nombre fini de points critiques d'ordre 1;
• pour tout $M \in \mathbb{R}$, $\{x, f(x) \leq M\}$ est borné.

On considère la descente de gradient à pas constant $\alpha \in]0; 1/L[$.

Pour presque tout $x_0 \in \mathbb{R}^n$, la descente de gradient converge vers un point critique d'ordre 2.

→ Pour tout $x_0 \in \mathbb{R}^n \setminus E$ où E est un ensemble de mesure de Lebesgue nulle.

Idée de dém: $(x_t)_{t\in\mathbb{N}}$ converge vers un point critique d'ordre 1. (Admis)

À montrer: si z est un point critique d'ordre 1 mais pas 2, pour presque tout x_0, $(x_t)_{t\in\mathbb{N}}$ ne converge pas vers z.

Fixons z critique d'ordre 1 mais pas 2.

Hypothèse très simplificatrice : f est quadratique sur $B(\cancel{z}\,0, r_0)$ pour un $r_0>0$.

$$f : x \longrightarrow \frac{1}{2}\langle x, Mx\rangle + \langle \cancel{b}, x\rangle.$$

$z=0$ critique d'ordre 1

$$\Rightarrow \nabla f(0) = 0 \quad \Rightarrow \quad M\times 0 + b\,(=b) = 0.$$

$z=0$ pas critique d'ordre 2

$$\Rightarrow \mathrm{Hess}\, f(0) \not\succeq 0 \quad \Rightarrow \quad M \not\succeq 0$$
$$(\lambda_{min}(M)<0).$$

Diagonalisons M : $\quad M = U^T \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} U$

avec U orthonormale et $\lambda_1 \geq \dots \geq \lambda_n$ les valeurs propres.
Quitte à changer de coordonnées, on peut supposer $U = \mathrm{Id}$ donc

$$M = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}.$$

Pour tout t, on note $x_t = (x_{t,1}, \dots, x_{t,n})$. On a

$$\begin{aligned} x_{t+1} = x_t - \alpha \nabla f(x_t) &= x_t - \alpha M x_t \\ &= (x_{t,1}, \dots, x_{t,n}) - (\alpha\lambda_1 x_{t,1}, \dots \alpha\lambda_n x_{t,n}) \\ &= \big((1-\alpha\lambda_1) x_{t,1}, \dots, \\ &\qquad (1-\alpha\lambda_n) x_{t,n} \big). \end{aligned}$$

(si x_t appartient à $B(0,r_0)$)

Supposons que $(x_t)_{t\in\mathbb{N}}$ tend vers $z=0$.
Soit N tq $x_t \in B(0,r_0)$ pour tout $t \geq N$.
Pour tout $t\geq N$,

$$x_t = \Big((1-\alpha\lambda_1)^{t-N} x_{N,1}, \dots, \underbrace{(1-\alpha\lambda_n)}_{>1}{}^{t-N} x_{N,n} \Big)$$

car $\lambda_n<0$ et $\alpha>0$

$$(1-\alpha\lambda_n)^{t-N} \xrightarrow{t\to+\infty} +\infty$$

Pour que $x_t \xrightarrow{t\to+\infty} 0$, il faut que $(1-\alpha\lambda_n)^{t-N} x_{N,n} \xrightarrow{t\to+\infty} 0$

donc que $x_{N,n} = 0$.

Pour que $x_t \xrightarrow{t\to+\infty} z$, il faut que

$$\underbrace{(Id-\alpha\nabla f)^N}_{\text{opérateur qui décrit une étape de descente}}(x_0) = x_N \in \underbrace{\{x \in \mathbb{R}^n,\ x_n = 0\}}_{\overset{\text{déf}}{=} \mathcal{E}} \quad \text{pour un certain } N \in \mathbb{N}.$$

<u>càd</u> $\quad x_0 \in (Id-\alpha\nabla f)^{-N}(\mathcal{E})$ pour un certain $N \in \mathbb{N}$

<u>càd</u> $\quad x_0 \in \underbrace{\bigcup_{N\in\mathbb{N}} \underbrace{(Id-\alpha\nabla f)}_{\text{difféomorphisme}}{}^{-N} \underbrace{(\mathcal{E})}_{\text{mesure de Lebesgue nulle}}}_{\text{mesure de Lebesgue nulle}}$

donc pour que $(x_t)_{t\in\mathbb{N}}$ converge vers $z=0$, il faut que x_0 appartienne à un ensemble de mesure nulle.

IV) <u>Méthode de second ordre</u>

Descente de gradient : pas de vitesse de convergence vers un point critique d'ordre 2.

$\longrightarrow$ dans le pire cas, très lente

Autre algo ? Perturbed Gradient Descent [Jin, Ge, Netrapalli, Kakade, Jordan, 17]

Voyons un algorithme d'ordre 2 :

Prop : $\forall x \in \mathbb{R}^n$,

$$\forall h \in \mathbb{R}^n \quad f(x+h) = f(x) + \langle \nabla f(x), h \rangle + \frac{1}{2}\langle \mathrm{Hess} f(x) h, h \rangle + o(\|h\|^2).$$

Définition de x_{t+1} à partir de x_t :

$$x_{t+1} = x_t + h_t$$

où $$h_t = \underset{\|h\| \leq R_t}{\operatorname{argmin}} \; f(x_t) + \langle \nabla f(x_t), h \rangle + \frac{1}{2}\langle \mathrm{Hess} f(x_t) h, h \rangle$$

(méthode des régions de confiance, Trust Region)

Thm : Soit $L_2 > 0$. On suppose que $\mathrm{Hess} f$ est L_2-lipschitzienne.

Soit $\varepsilon > 0$. On fixe $R_t = \frac{\sqrt{\varepsilon}}{L_2}$ pour tout t.

On arrête l'algo si $\frac{\|\nabla f(x_t) + \mathrm{Hess} f(x_t) h_t\|}{\|h_t\|} \leq \sqrt{\varepsilon}$.

On renvoie x_{t+1}.

Pour tout $x_0 \in \mathbb{R}^n$, l'algo s'arrête après au plus $O\left(\frac{L_2^2 (f(x_0) - f(x_*))}{\varepsilon^{3/2}}\right)$ et

$$\|\nabla f(x_{t+1})\| \lesssim \frac{\varepsilon}{L_2} \text{ et } \lambda_{min}(\mathrm{Hess}(f(x_{t+1})) \gtrsim -\sqrt{\varepsilon}.$$

V) Exemple : reconstruction de phase

Définition d'un pb de reconstruction de phase : on veut reconstruire $x_* \in \mathbb{C}^n$

Pour cela, on a accès à

$$|L_1(x_*)|, |L_2(x_*)|, \dots, |L_m(x_*)|$$

où $L_1, \dots, L_m : \mathbb{C}^n \to \mathbb{C}$ sont linéaires et connues.

Rq : $\forall \alpha \in \mathbb{R}$, $|L_1(x_* e^{i\alpha})| = |e^{i\alpha} L_1(x_*)|$
$= |L_1(x_*)|$

De même pour $L_2, \dots, L_m$.

Donc on ne peut espérer reconstruire x_* qu'à « phase globale près ».

Thm : si $m \geq 4n$, x_* est uniquement déterminé par les mesures $|L_1(x_*)|, \dots, |L_m(x_*)|$ pour presque toutes les $L_1, \dots, L_m$.

Comment reconstruire ?

On veut trouver x tq

$|L_1(x)| = y_1$ où $y_1 = |L_1(x_*)|$ est la 1ère mesure

...

$|L_m(x)| = y_m$ où y_m est la m-ème mesure

$$\Leftrightarrow \quad |L_1(x)|^2 = y_1^2$$
$$\dots$$
$$|L_m(x)|^2 = y_m^2$$

Définissons $F : x \to \sum_{k=1}^{m} \left(|L_k(x)|^2 - y_k^2 \right)^2$.

On veut trouver un minimiseur de F.

F non-convexe. $\Rightarrow$ on ne peut pas la minimiser à coup sûr.

On peut en revanche trouver un point critique d'ordre 2.

Cas « simple » : $L_1, \dots, L_m$ choisies au hasard, indépendamment selon une loi normale.

Thm [Sun, Qu, Wright, 18]:
si $m \geq C n \log^3(n)$ pour une constante C assez grande
avec grande probabilité $1-O(1/m)$,
les points critiques d'ordre 2 de F sont les minimiseurs globaux, $\{x_* e^{i\alpha}, \alpha \in \mathbb{R}\}$.

(Ce qui suit est un calcul fait pendant le TP ; ce n'est pas une partie du cours)

Soit $L : U \to \sum_{(i,j)\in S} \left((UU^T)_{i,j} - M_{ij} \right)^2$

Soit U. Pour $H \in Sym(n)$,

$$
\begin{aligned}
L(U+H) &= ? \\
&= \sum_{(i,j)} \left(\left((U+H)(U+H)^T\right)_{ij} - M_{ij} \right)^2 \\
&= \sum_{(i,j)} \Big((UU^T)_{ij} + (HU^T)_{ij} + (UH^T)_{ij} + O(\|H\|^2) - M_{ij} \Big)^2 \\
&= \sum_{(i,j)\in S} \Big((UU^T)_{ij} - M_{ij} \Big)^2 + 2\left((HU^T)_{ij} + (UH^T)_{ij} \right) \times \left((UU^T)_{ij} - M_{ij} \right) + O(\|H^2\|) \\
&= L(U) + \underbrace{\sum_{(i,j)\in S} 2\left((HU^T)_{ij} + (UH)^T_{ij} \right) \times \left((UU^T)_{ij} - M_{ij} \right)}_{(*)} + O(\|H\|^2)
\end{aligned}
$$

À mettre sous la forme

$$\sum_{i,j} H_{ij} G_{ij} = \langle H, \underbrace{G}_{= \nabla L(U)} \rangle$$

$$(*) = \sum_{(i,j)} 2\, Mask_{ij} \left((HU^T)_{ij} + (UH)^T_{ij} \right) \left((UU^T)_{ij} - M_{ij} \right)$$

$$= 2 \left\langle HU^T + UH^T , \overbrace{((UU^T) - M) \otimes Mask}^{\overset{def}{=} E} \right\rangle$$

$$= 2 \langle HU^T + UH^T , E \rangle$$

$$= 2(\langle HU^T , E \rangle + \langle UH^T , E \rangle)$$

$$= 2(\langle H , EU \rangle + \langle E^T U , \cancel{E} H \rangle)$$

$$= 2 \langle H , EU + E^T U \rangle$$

$$= \langle H , \underbrace{2(E + E^T)}_{\nabla L(U)} U \rangle$$