Mapi3 Machine Learning : Optimisation pour le Machine Learning

ont cte : En apprentissage statistique, il faut souvent résoudre des problèmes de la forme

$$\hat{\theta} \approx \underset{\theta \in \Theta = \mathbb{R}^d}{\operatorname{argmin}} F(\theta).$$

Parfois : Pas de solution en forme close $\Longrightarrow$ Il faut utiliser un algorithme d'approximation, un algorithme d'optimisation !

Exemple : Minimisation du risque empirique

$$F(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f_\theta(x_i)) + \underbrace{\Omega(\theta)}_{\text{Régularisation.}}$$

Décomposition du risque

$$\mathcal{R}(f_{\hat{\theta}}) - \mathcal{R}(f_{\theta^*}) = \mathcal{R}(f_{\hat{\theta}}) - \hat{\mathcal{R}}(f_{\hat{\theta}}) + \hat{\mathcal{R}}(f_{\hat{\theta}}) - \hat{\mathcal{R}}(f_{\theta^*}) + \hat{\mathcal{R}}(f_{\theta^*}) - \mathcal{R}(f_{\theta^*})$$

$$\leq \underbrace{2 \sup_{\theta \in \Theta} \left| \mathcal{R}(f_\theta) - \hat{\mathcal{R}}(f_\theta) \right|}_{\text{cf leçons précédentes}} + \underbrace{\left(\hat{\mathcal{R}}(f_{\hat{\theta}}) - \inf_{\theta \in \Theta} (f_\theta) \right)}_{\text{Erreur d'optimisation}}.$$

Idée fondamentale de l'optimisation :

Remplacer F par un dvpt de Taylor, effectuer un pas en fonction de ce dvpt.

Exemple :

• <u>Dynamique d'ordre 1 :</u> $f(\theta+\delta\theta) = f(\theta) + \langle \nabla f(\theta), \delta\theta \rangle + o(\|\delta\theta\|)$

$\Rightarrow$ La direction de plus forte descente locale est $-\nabla f(\theta)$.

$\Rightarrow$ <u>Algorithme de descente de gradient</u>

<u>Entrée</u> : Un point de départ θ_0.

<u>Récurrence</u> : $\theta_t = \theta_{t-1} - \gamma_t \nabla f(\theta_{t-1})$, où $(\gamma_t)_{t\geq 1}$ est la suite des "pas" ou "learning rate".

• <u>Dynamique d'ordre 2 :</u> $\nabla f(\theta+\delta\theta) = \underbrace{\nabla f(\theta) + \langle \nabla^2 f(\theta)\delta\theta, \delta\theta \rangle}_{T_2(\theta)} + o(\|\delta\theta\|)$

un point critique satisfait $\nabla f(\theta) = 0$, et en remplaçant ∇f par T_2, on obtient

$$\delta\theta = -\nabla^2 f(\theta)^{-1} \nabla f(\theta).$$

$\Rightarrow$ <u>Algorithme de Newton</u>

<u>Entrée</u> : Un point de départ θ_0

<u>Récurrence</u> : $\theta_t = \theta_{t-1} - \gamma_t \nabla^2 f(\theta_{t-1})^{-1} \nabla f(\theta_{t-1})$

où (γ_t) est la suite des pas.

I. Le point de vue des systèmes dynamiques :

• <u>Descente de gradient</u> :

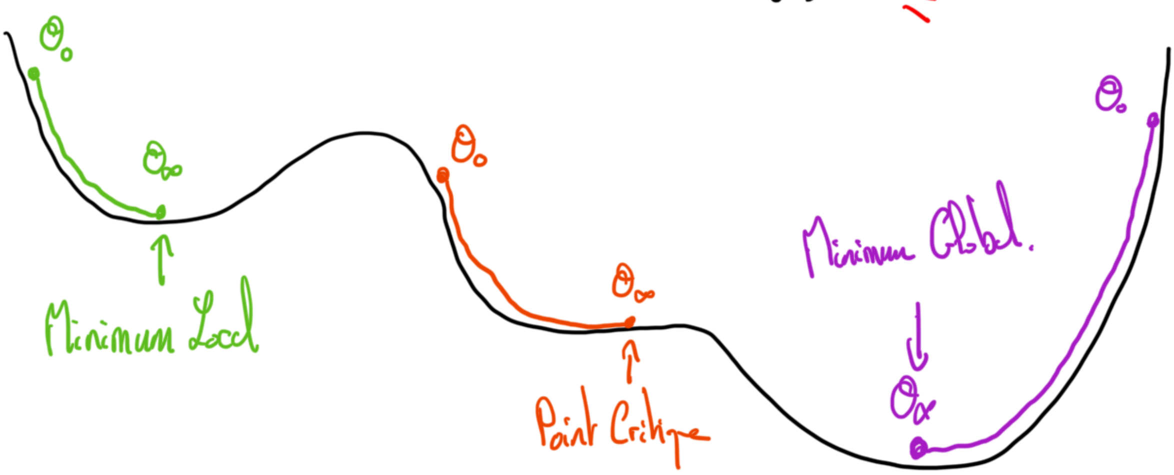
Lorsque $\sup_t \gamma_t \ll 1$, on peut voir la descente de gradient comme une discrétisation d'Euler de l'ODE

$$\boxed{\frac{d}{dt}\theta = -\nabla f(\theta)}$$ (Gradient Flow)

dt

Alors $$\frac{dF(\theta)}{dt} \underset{\text{règle de la chaîne}}{=} \left\langle \nabla F(\theta), \frac{d}{dt}\theta \right\rangle \underset{\text{Gradient Flow}}{=} -\left\langle \nabla F(\theta), \nabla F(\theta) \right\rangle$$

$$= -\|\nabla F(\theta)\|^2 \leq 0$$

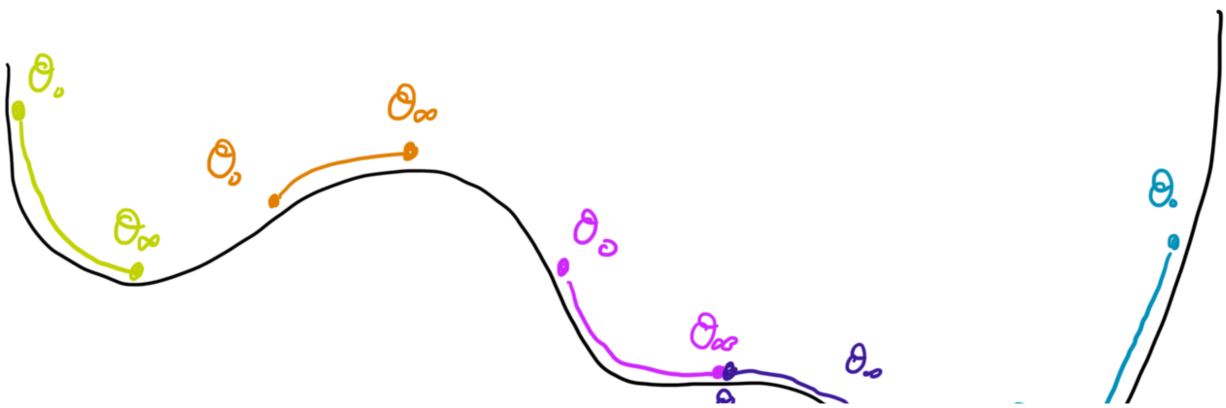


- Méthode de Newton :

$$\boxed{\frac{d}{dt}\theta = -\nabla^2 F(\theta)^{-1} \nabla F(\theta)}$$

Alors $$\nabla^2 F(\theta)\frac{d}{dt}\theta = -\nabla F(\theta) \underset{\text{règle de la chaîne}}{\Rightarrow} \frac{d}{dt}\nabla F(\theta) = -\nabla F(\theta)$$

donc $$\nabla F(\theta) = \nabla F(\theta_0)\, e^{-t}.$$



II. Convergence de la descente de gradient

1) Fonction lisse

Pour analyser la descente de gradient, il faut pouvoir "faire confiance" à l'approximation locale de la fonction par le polynôme de Taylor à l'ordre 1.

Définition : F est L-lisse si

$$\forall \theta, \forall \eta, \quad |F(\eta) - F(\theta) - \langle \nabla F(\theta), \eta - \theta \rangle| \leq \frac{L}{2} \|\theta - \eta\|_2^2$$

ou alors, de manière équivalente, si ∇f est L-Lipschitz pour $\|\cdot\|_2$.

Si F est C^2, c'est aussi équivalent à $\|\nabla^2 f(\theta)\|_{op} \leq L \ \forall \theta$.

Proposition (Lemme de descente) Si F est L-lisse, GD avec $\gamma_t = \frac{1}{L}$ donne

$$F(\theta_t) - F(\eta_*) \leq F(\theta_{t-1}) - F(\eta_*) - \frac{1}{2L} \|\nabla F(\theta_{t-1})\|_2^2$$

Preuve :

$$\begin{aligned} F(\theta_t) &= F\left(\theta_{t-1} - \frac{1}{L} \nabla F(\theta_{t-1})\right) \\ &\leq F(\theta_{t-1}) + \left\langle \nabla F(\theta_{t-1}), -\frac{1}{L} \nabla F(\theta_{t-1}) \right\rangle + \frac{L}{2} \|-\nabla f(\theta_{t-1})/L\|_2^2 \\ &= F(\theta_{t-1}) - \frac{1}{L} \|\nabla F(\theta_{t-1})\|_2^2 + \frac{1}{2L} \|\nabla F(\theta_{t-1})\|_2^2 \end{aligned}$$

2) Le cas non-convexe et lisse

Si F est L-lisse, un télescopage dans le lemme de descente donne

$$\frac{1}{2Lt} \sum_{i=1}^{t} \|\nabla F(\theta_{i-1})\|_2^2 \leq \frac{F(\theta_0) - F(\eta_*)}{t}$$

La trajectoire converge en <u>Moyenne de Césaro quadratique</u> vers un point critique.

3) Le cas convexe et lisse

Proposition (Co-coercivité) Si F est <u>L-lisse</u> et <u>convexe</u>

$$\forall \theta, \forall \eta, \quad \frac{1}{L}\|\nabla F(\theta)-\nabla F(\eta)\|_2^2 \leq \langle \nabla F(\theta)-\nabla F(\eta), \theta-\eta \rangle$$

et $F(\theta) \geq F(\eta) + \langle \nabla F(\eta), \theta-\eta \rangle + \frac{1}{2L}\|\nabla F(\theta)-\nabla F(\eta)\|_2^2$.

Preuve

$$\underbrace{F(\eta)+\langle \nabla F(\eta), \xi-\eta \rangle}_{(G)} \underbrace{\leq}_{\text{Convexité}} F(\xi) \underbrace{\leq}_{\text{L-lisse}} \underbrace{F(\theta)+\langle \nabla F(\theta), \xi-\theta \rangle + \frac{L}{2}\|\theta-\xi\|_2^2}_{(D)}$$

$$(D)-(G) = F(\theta)-F(\eta)+\langle \nabla F(\theta), \xi-\theta \rangle - \langle \nabla F(\eta), \xi-\eta \rangle + \frac{L}{2}\|\theta-\xi\|_2^2$$

donc $(D)-(G)$ est minimum pour ξ tq

$$\nabla F(\theta)-\nabla F(\eta)+L(\theta-\xi)=0 \Rightarrow \xi = \theta + \frac{1}{L}\left(\nabla F(\theta)-\nabla F(\eta)\right)$$

ce qui donne la deuxième inégalité du résultat en réinjectant. $\square$

Corollaire: Si F est convexe et L-lisse, alors les itérés de GD avec $\gamma_t = \frac{1}{L}$ satisfont

$$\|\theta_t - \eta_*\|_2^2 \leq \|\theta_t - \eta_*\|_2^2 - \frac{1}{L}\langle \nabla F(\theta_{t-1}), \theta_t - \eta_* \rangle$$

Preuve $\|\theta_t - \eta_*\|_2^2 = \|\theta_{t-1} - \frac{1}{L}\nabla F(\theta_{t-1}) - \eta_*\|_2^2$

$$= \underbrace{\|\theta_{t-1} - \eta_*\|_2^2 - \frac{2}{L}\langle \nabla F(\theta_{t-1}), \theta_{t-1} - \eta_* \rangle}_{} + \frac{1}{L^2}\|\nabla F(\theta_{t-1})\|_2^2$$

$$= \quad '' \quad + \frac{1}{L^2}\|\nabla F(\theta_{t-1}) - \underbrace{\nabla F(\eta_*)}_{=0}\|_2^2$$

$$\underset{\text{Co-Coercivité}}{\leq} \quad '' \quad + \frac{1}{L}\langle \nabla F(\theta_{t-1}) - \nabla F(\eta_*), \theta_{t-1} - \eta_* \rangle$$

$$= \|\theta_{t-1} - \eta_*\|_2^2 - \frac{1}{L}\langle \nabla F(\theta_{t-1}), \theta_{t-1} - \eta_* \rangle \qquad \square$$

Théorème : Si F est convexe et L-lisse, alors les itérés $(\theta_t)_t$ de GD avec $\gamma_t = \frac{1}{L}$ satisfont

$$F(\theta_t) - F(\eta_*) \leq \frac{L}{t+1}\|\theta_0 - \eta_*\|_2^2$$

Preuve : D'après le lemme de descente, $(F(\theta_t))_t$ est décroissante.

Donc, $$F(\theta_t) - F(\eta_*) \leq \frac{1}{t+1}\sum_{i=0}^{t}\left(F(\theta_i) - F(\eta_*)\right)$$

$$\underset{\text{Convexité}}{\leq} \frac{1}{t+1}\sum_{i=0}^{t}\langle \nabla F(\theta_i), \theta_i - \eta_* \rangle$$

$$\underset{\text{Corr. prec.}}{\leq} \frac{L}{t+1}\sum_{i=0}^{t}\left(\|\theta_i - \eta_*\|_2^2 - \|\theta_{i+1} - \eta_*\|_2^2\right)$$

$$\underset{\text{Telescopage}}{\leq} \frac{L}{t+1}\|\theta_0 - \eta_*\|_2^2$$

4) Fonctions fortement convexes

💡 Gradient Flow : $\frac{dF}{dt} = -\|\nabla F\|^2$ donc, si $2\mu(F - F(\theta_*)) \leq \|\nabla F\|^2$,

$$\Rightarrow \frac{d(F - F(\theta_*))}{dt} \leq -2\mu\,(F - F(\theta_*))$$

$$\frac{d}{dt} \ldots \leq \ldots$$

$$\underset{\text{Gronwall}}{\Rightarrow} \quad F - F(\theta_*) \leq e^{-2\mu t}\left(F(\theta_0) - F(\theta_*)\right).$$

<u>Définition</u>: F est μ-fortement convexe si

$$\forall \eta, \theta, \quad F(\eta) \geq F(\theta) + \langle \nabla F(\theta), \eta - \theta \rangle + \frac{\mu}{2} \|\eta - \theta\|_2^2 .$$

Si F est C^2, c'est équivalent à $\nabla^2 F(\theta) \succcurlyeq \mu I \quad \forall \theta$.

<u>Proposition</u> Si F est différentiable, μ fortement convexe et que η_* est son minimiser unique,

$$\|\nabla F(\theta)\|_2^2 \geq 2\mu\left(F(\theta) - F(\eta_*)\right), \quad \forall \theta \quad \text{(Inégalité de Łojasiewicz)}.$$

<u>Preuve</u>: Minimiser en η le terme de droite dans la déf de μ-fortement convexe.

<u>Théorème</u>: Si F est L-lisse et μ fortement convexe, GD avec $\gamma_t = \frac{1}{L}$ satisfait

$$F(\theta_t) - F(\eta_*) \leq \underbrace{\left(1 - \frac{1}{\kappa}\right)^t}_{\leq e^{-t/\kappa}} \left(F(\theta_0) - F(\eta_*)\right)$$

$\kappa = \frac{L}{\mu}$: conditionnement.

<u>Preuve</u>: Inégalité de Łojasiewicz dans le lemme de descente.

III. Descente de gradient stochastique

💡 Remplacer $\nabla F(\theta_{t-1})$ par un estimateur non-biaisé $g_t(\theta_{t-1})$

$$\left(\text{i.e. } \mathbb{E}\left(g_t(\theta_{t-1}) \mid \underset{\substack{\uparrow \\ \text{information au temps } t-1.}}{\theta_{t-1}}\right) = \nabla F(\theta_{t-1})\right)$$

<u>Pourquoi ?</u> Car $g_t(\theta_{t-1})$ peut être bien plus rapide à calculer que $\nabla F(\theta_{t-1})$.

<u>Exemple :</u> $F(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f_\theta(x_i))$.

On choisit $B_t \subseteq \{1, \dots, n\}$ au hasard, puis

$$g_t(\theta) = \frac{1}{|B_t|} \sum_{i \in B_t} \ell(y_i, f_\theta(x_i)).$$

Complexité : $O(n) \longrightarrow O(|B_t|)$.

<u>Algorithme (Stochastic Gradient Descent)</u>

<u>Entrée :</u> Un point de départ θ_0.

<u>Récurrence :</u> $\theta_t = \theta_{t-1} - \gamma_t\, g_t(\theta_{t-1})$. (γ_t) : suite de "pas"

<u>H1 :</u> $\mathbb{E}\left(g(\theta_{t-1}) \mid \theta_{t-1}\right) = \nabla F(\theta_{t-1})$ (Estimateur sans biais)

<u>H2 :</u> $\|g_t(\theta_{t-1})\|_2^2 \leq B^2$ p.s. (Variance contrôlée).

1) Le cas général lisse

<u>Lemme (de descente)</u> Si F est L-lisse et sous H_1 :

$$\mathbb{E}(F(\theta_t) \mid \theta_{t-1}) \leq F(\theta_{t-1}) - \gamma_t \|\nabla F(\theta_{t-1})\|_2^2 + \frac{L}{2} \gamma_t^2\, \mathbb{E}\left(\|g_t(\theta_{t-1})\|_2^2 \mid \theta_{t-1}\right)$$

Preuve : Par L-lissitude

$$F(\theta_t) \leq F(\theta_{t-1}) - \gamma_t \langle \nabla F(\theta_{t-1}), g_t(\theta_{t-1})\rangle + \frac{L}{2}\gamma_t^2 \|g_t(\theta_{t-1})\|_2^2$$

et l'espérance conditionnelle donne le résultat. □

Donc sous H_2, une somme telescopique donne $(\gamma_t = \gamma)$

$$\frac{1}{t}\sum_{i=0}^{t-1} \mathbb{E}\left(\|\nabla F(\theta_i)\|_2^2\right) \leq \frac{F(\theta_0) - F(\theta_*)}{\gamma t} + \frac{L}{2}\gamma B^2$$

et $\gamma = \frac{1}{\sqrt{t}}$ donne la vitesse finale.

2) Le cas convexe

Théorème : Si F est convexe et B-Lipschitz, admet un minimiseur θ_* tq $\|\theta_0 - \theta_*\|_2 \leq D$ et si H_1 et H_2 sont satisfaites, si $\gamma_t = \left(\frac{D}{B}\right)\frac{1}{\sqrt{t}}$, les itérées de SGD satisfont

$$\mathbb{E}\left(F(\bar{\theta}_t) - F(\theta_*)\right) \leq DB\,\frac{2 + \log(t)}{2\sqrt{t}}$$

où $\bar{\theta}_t = \left(\sum_{i=1}^{t} \gamma_i \theta_{i-1}\right) / \left(\sum_{i=1}^{t}\gamma_i\right)$.

démonstration : $\mathbb{E}\left(\|\theta_t - \theta_*\|_2^2\right) = \mathbb{E}\left(\|\theta_{t-1} - \gamma_t g_t(\theta_{t-1}) - \theta_*\|_2^2\right)$

$$= \mathbb{E}\left(\|\theta_{t-1} - \theta_*\|_2^2\right) - 2\gamma_t \mathbb{E}\left(\langle g_t(\theta_{t-1}), \theta_{t-1} - \theta_*\rangle\right) + \gamma_t^2 \mathbb{E}\left(\|g_t(\theta_{t-1})\|_2^2\right)$$

de plus, $\mathbb{E}\left(\langle g_t(\theta_{t-1}), \theta_{t-1} - \theta_*\rangle\right) = \mathbb{E}\left(\mathbb{E}\left(" \mid \theta_{t-1}\right)\right)$

$$= \mathbb{E}\left(\left\langle \mathbb{E}\left(g(\theta_{t-1}) \mid \theta_{t-1}\right), \theta_{t-1} - \theta_* \right\rangle \right)$$

$$= \mathbb{E}\left(\left\langle \nabla f(\theta_{t-1}), \theta_{t-1} - \theta_* \right\rangle \right)$$

$$\text{donc } \mathbb{E}\left(\|\theta_t - \theta_*\|_2^2\right) \leq \mathbb{E}\left(\|\theta_{t-1} - \theta_*\|_2^2\right) - 2\gamma_t \, \mathbb{E}\left(\overbrace{\left\langle \nabla f(\theta_{t-1}), \theta_{t-1} - \theta_* \right\rangle}^{\geq \langle \nabla f(\theta_{t-1}), \theta_{t-1} - \theta_* \rangle} \right) + \gamma_t^2 B^2$$

$$\text{donc : } \gamma_t \, \mathbb{E}\left(f(\theta_{t-1}) - f(\theta_*)\right) \leq \frac{1}{2}\left(\mathbb{E}\left(\|\theta_{t-1} - \theta_*\|_2^2\right) - \mathbb{E}\left(\|\theta_t - \theta_*\|_2^2\right) \right) + \frac{1}{2}\gamma_t^2 B^2$$

$$\Rightarrow \frac{1}{\sum_{i=1}^{t} \gamma_i} \sum_{i=1}^{t} \gamma_i \, \mathbb{E}\left(f(\theta_{i-1}) - f(\theta_*)\right) \leq \frac{\|\theta_0 - \theta_*\|_2^2}{2 \sum_{i=1}^{t} \gamma_i} + B^2 \frac{\sum_{i=1}^{t} \gamma_i^2}{\sum_{i=1}^{t} \gamma_i}$$

Puis on peut minorer le terme de gauche par Jensen et majorer le terme de droite par comparaison série intégrale.

$\square$