Optimisation pour le Machine Learning

Contexte : En apprentissage statistique, il est souvent utile de résoudre des problèmes de la forme

$$\hat{\theta} \approx \underset{\theta \in \Theta}{\operatorname{argmin}}\, F(\theta)$$

exemples :

- Minimisation du risque empirique

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

- Maximum de vraisemblance

$$f(\theta) = -\log\left(p(x_1, \dots, x_n \mid \theta)\right)$$

⚠ Parfois, il existe des solutions en forme close (moindres carrés, Ridge, ...), mais souvent ce n'est pas le cas et il faut alors utiliser un algorithme d'optimisation (qui approche la solution).
Il faut alors prendre en compte cette nouvelle source d'erreur.

Décomposition du risque :

$$\mathcal{R}(f_{\hat{\theta}}) - \mathcal{R}(f_{\theta^*}) = \underbrace{\mathcal{R}(f_{\hat{\theta}}) - \hat{\mathcal{R}}(f_{\hat{\theta}})}_{\text{déviation I}} + \underbrace{\hat{\mathcal{R}}(f_{\hat{\theta}}) - \hat{\mathcal{R}}(f_{\theta^*})}_{\leq \hat{\mathcal{R}}(f_{\hat{\theta}}) - \inf_\theta \hat{\mathcal{R}}(f_\theta) \,=\, \text{erreur d'optimisation}} + \underbrace{\hat{\mathcal{R}}(f_{\theta^*}) - \mathcal{R}(f_{\theta^*})}_{\text{déviation II}}$$

II. Descente de gradient

$\Theta = \mathbb{R}^d$, f différentiable.

↳ Localement, $f(\theta + \delta\theta) = f(\theta) + \langle \nabla f(\theta), \delta\theta \rangle + o(\|\delta\theta\|)$

Donc : Localement, pour faire décroître f le plus possible, il faut partir dans la direction $-\nabla f(\theta)$.

Algorithme de descente de gradient

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

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

où $(\gamma_t)_{t \geq 1}$ est une suite de tailles de pas.

Le point de vue des système dynamiques

Il est possible de voir la descente de gradient comme une discrétisation d'Euler explicite de l'équation

$$\dot{\theta} = -\nabla f(\theta) \quad \text{(Gradient flow)}$$

Alors $$\frac{df(\theta)}{dt} = \langle \nabla f(\theta), \dot{\theta} \rangle = -\langle \nabla f(\theta), \nabla f(\theta) \rangle$$
$$= -\|\nabla f(\theta)\|^2 \leq 0$$

↳ Tant que la dynamique n'est pas à un point dont le gradient est nul, elle décroît.

1) Fonctions lisses et fortement convexes

<u>Définition</u> : F est μ-fortement convexe ($\mu > 0$) si

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

<u>exemple</u> : Fonctions quadratiques à valeurs propres > 0.

<u>exemple</u> : $\|.\|_2^2$ régularisation

<u>Propriété (Inégalité de Łojasiewicz)</u> :

Si F est différentiable et μ-fortement convexe, avec pour minimiseur unique η_*, alors

$$\|\nabla F(\theta)\|_2^2 \geq 2\mu (F(\theta) - F(\eta_*)), \forall \theta$$

<u>démonstration</u> : dans $(*)$, le terme de droite est minimisé par $\eta = \theta - \frac{1}{\mu} \nabla f(\theta)$ ce qui donne le résultat.

<u>Conséquence d'un point de vue continu</u> :

$$\dot{\theta} = -\nabla F(\theta).$$

$$\Rightarrow \frac{dF(\theta)}{dt} = -\|\nabla F(\theta)\|^2 \leq -2\mu (F(\theta) - F(\eta_*))$$

$$\Rightarrow \frac{d(F(\theta) - F(\eta^*))}{dt} \leq -2\mu (F(\theta) - F(\eta_*))$$

4

(Grönwall) $\Rightarrow$ $F(\theta)-F(\eta_*) \leq \left[F(\theta_0)-F(\eta_*)\right]e^{-2t}$

Pour obtenir la convergence exponentielle des itérées discrètes, il faut avoir une forme de régularité des gradients pour assurer que l'approximation discrète se rapproche de la dynamique continue.

<u>Définition:</u> F est L-lisse si

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

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

<u>Théorème:</u> Si F est L-lisse et μ-fortement convexe, GD avec $\gamma_t = 1/L$ donne une suite $(\theta_t)_t$ satisfaisant

$$\cancel{F(\theta_t)-F(\theta_{t+1})}$$

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

où $\kappa = \frac{L}{\mu}$ est le conditionnement.

<u>démonstration</u>:

$$F(\theta_t) = F\left(\theta_{t-1} - \frac{1}{L}\nabla F(\theta_{t-1})\right)$$

$$\underset{L\text{-lisse}}{\leq} F(\theta_{t-1}) + \langle \nabla F(\theta_{t-1}), -\nabla F(\theta_{t-1})/L \rangle + \frac{L}{2}\|-\nabla F(\theta_{t-1})/L\|$$

$$= F(\theta_{t-1}) - \frac{1}{L}\|F'(\theta_{t-1})\|_2^2 + \frac{1}{2L}\|F'(\theta_{t-1})\|_2^2$$

donc

$$(**)\quad \left(F(\theta_t) - F(\eta_*)\right) \leq \left(F(\theta_{t-1}) - F(\eta_*)\right) - \frac{1}{2L}\|F'(\theta_{t-1})\|_2^2$$

Par μ-forte convexité et inégalité de Łojasiewicz :

$$F(\theta_t) - F(\eta_*) \leq \left(1 - \frac{\mu}{L}\right)\left(F(\theta_{t-1}) - F(\eta_*)\right).$$

$\square$

2) Le cas lisse et (simplement) convexe

Dans ce cas, l'analyse est légèrement plus complexe.

<u>Proposition (Co-coercivité)</u>:

si F est L-lisse et convexe,

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

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

démonstration :

Nous allons prouver la deuxième inégalité, la première n'étant que la somme de la deuxième et de la deuxième dans laquelle nous avons échangé θ et η.

Par convexité et ~~Lips~~ comme f est L-lisse,

$$f(\eta) + \langle \nabla f(\eta), \xi - \eta \rangle \leq f(\xi) \leq f(\theta) + \langle \nabla f(\theta), \xi - \theta \rangle + \frac{L}{2} \|\theta - \xi\|_2^2$$

et injecter dans cette inégalité le ξ qui minimise l'écart entre le terme de gauche et le terme de droite donne le résultat.

Corollaire : Si f est convexe et L-lisse, alors les itérés de GD satisfont

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

démonstration :

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

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

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

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

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

□

Théorème : Si F est convexe et L-lisse, alors les itérés $(\theta_t)_t$ de GD satisfont

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

démonstration :

$(**)$ est toujours valide, donc $\left(F(\theta_t) - F(\eta_*)\right)_t$ est décroissante.

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

De plus, par convexité, $\forall i$, $F(\theta_i) - F(\eta_*) \leq \langle \nabla F(\theta_i), \theta_i - \eta_* \rangle$

d'où, par le corollaire précédent,

$$F(\theta_t) - F(\eta_*) \leq \frac{L}{t+1}\sum_{i=0}^{t}\left(\|\theta_i - \eta_*\|_2^2 - \|\theta_{i+1} - \eta_*\|_2^2\right)$$

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

□

3) Le cas non-convexe :

Comme (**) reste valide,

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

même dans le cas non-convexe (mais L-lisse).

Alors, par téléscopage

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

---

4) Aller plus loin :

- Optimisation non lisse : Il est possible d'obtenir une convergence à vitesse $\frac{1}{\sqrt{t}}$ lorsque F n'est pas lisse (mais convexe) si cette dernière est Lipschitz (voir Bach)
- Accélération de Nesterov : Dans les cas fortement convexes, permet de passer de κ à $\sqrt{\kappa}$. Dans les cas convexes, lisses, vitesse en $\frac{1}{t^2}$.
- Méthode du gradient proximal : Permet d'optimiser

$$\underset{\text{Lisse}}{\underset{\uparrow}{F}} + \underset{\text{Non\_lisse (norme, indicatrice, ...)}}{\underset{\uparrow}{G}}$$

II Descente de gradient stochastique :

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

$$\left(\text{i.e } \mathbb{E}\left(g_t(\theta_{t-1}) \,\middle|\, \underbrace{\theta_{t-1}}_{\text{information au temps } t-1}\right) = \nabla F(\theta_{t-1})\right)$$

<u>Pourquoi ?</u> Car calculer un estimateur peut être beaucoup plus rapide !

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

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

<u>Récurrence :</u> $\theta_t = \theta_{t-1} - \gamma_t \, g_t(\theta_{t-1})$

où $(\gamma_t)_t$ est une suite de tailles de pas.

<u>Exemples (SGD en ML) :</u>

- <u>Empirical Risk minimization</u>

$$F(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f_\theta(x_i))$$

à chaque itération, un ou plusieurs indices sont choisis au hasard (mini-batch), l'estimateur du gradient est calculé sur ces indices.

- <u>Expected Risk minimization</u>

$$F(\theta) = \mathbb{E}\left(\ell\left(y, f_\theta(x)\right)\right)$$

À chaque étape, un couple (x, y) est échantillonné et l'estimateur du gradient est calculé sur ce couple.

Problème : On ne peut théoriquement voir qu'une seule fois chaque point du dataset même si en pratique c'est ce que font les gens.

~~II) Analyse dans le cas [illegible]~~

Hypothèse 1 : $\mathbb{E}\left(g_t(\theta_{t-1}) \mid \theta_{t-1}\right) = \nabla F(\theta_{t-1})$

(Estimateur sans biais)

Hypothèse 2 : $\| g_t(\theta_{t-1}) \|_2^2 \leq B^2$ presque sûrement $(\forall t)$.

(Peut être remplacé par une hypothèse de type Lipschitz sur les gradients).

Théorème : Si F est convexe et B-Lipschitz, admet un minimiseur θ_* tq $\|\theta_0 - \theta_*\|_2 \leq D$, et si H1 et H2 sont satisfaites, alors si $\gamma_t = \left(\frac{D}{B}\right)\frac{1}{\sqrt{t}}$, les itérées $(\theta_t)_{t \geq 0}$ de SGD satisfont

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

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

<u>démonstration:</u>

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

de plus,

$$\mathbb{E}\left(\langle y_t(\theta_{t-1}),\theta_{t-1}-\theta_*\rangle\right)=\mathbb{E}\left(\mathbb{E}\left(\text{''}\,\middle|\,\theta_{t-1}\right)\right)$$
$$=\mathbb{E}\left(\langle\mathbb{E}(y_t(\theta_{t-1})\,|\,\theta_{t-1}),\theta_{t-1}-\theta_*\rangle\right)$$
$$=\mathbb{E}\left(\langle\nabla F(\theta_{t-1}),\theta_{t-1}-\theta_*\rangle\right)$$

donc

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

de plus, par convexité, $F(\theta_{t-1})-F(\theta_*)\leq\langle\nabla F(\theta_{t-1}),\theta_{t-1}-\theta_*\rangle$,

donc

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

Donc, en sommant :

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

Ensuite, le terme de gauche provient de Jensen, et le terme de droite résulte de calculs simples (comparaison séries-intégrales) □

Remarque : Le cas non-stochastique et convexe découle de cette preuve.

Aller plus loin

- **Le cas fortement convexe**

Il est possible d'obtenir une vitesse en $\frac{1}{t}$ dans le cas fortement convexe

- **La réduction de variance**

Garder en mémoire des gradients peut permettre de réduire la variance et de gagner en vitesse de convergence

- **Méthodes adaptatives**

En pratique, des méthodes adaptatives (Adagrad, RMSProp, Adam, ...) sont utilisées pour trouver un bon pré-conditionnement au problème.