IFT-7002 Fondements de l'apprentissage machine

Problèmes convexes et descente de gradient stochastique (DGS)

Shai Shalev-Shwartz
The Hebrew University of Jerusalem

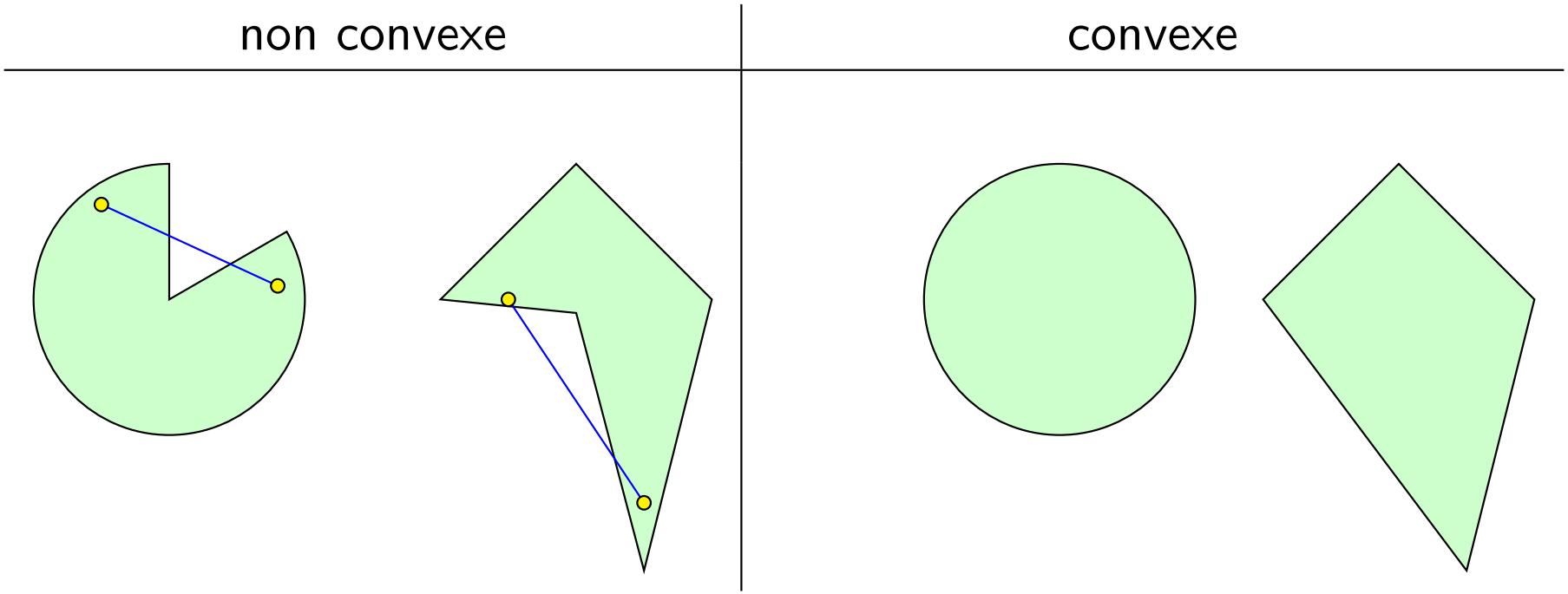
Traduit et adapté par Mario Marchand
Université Laval

Hiver 2024

Sommaire

Définition (Ensemble convexe)

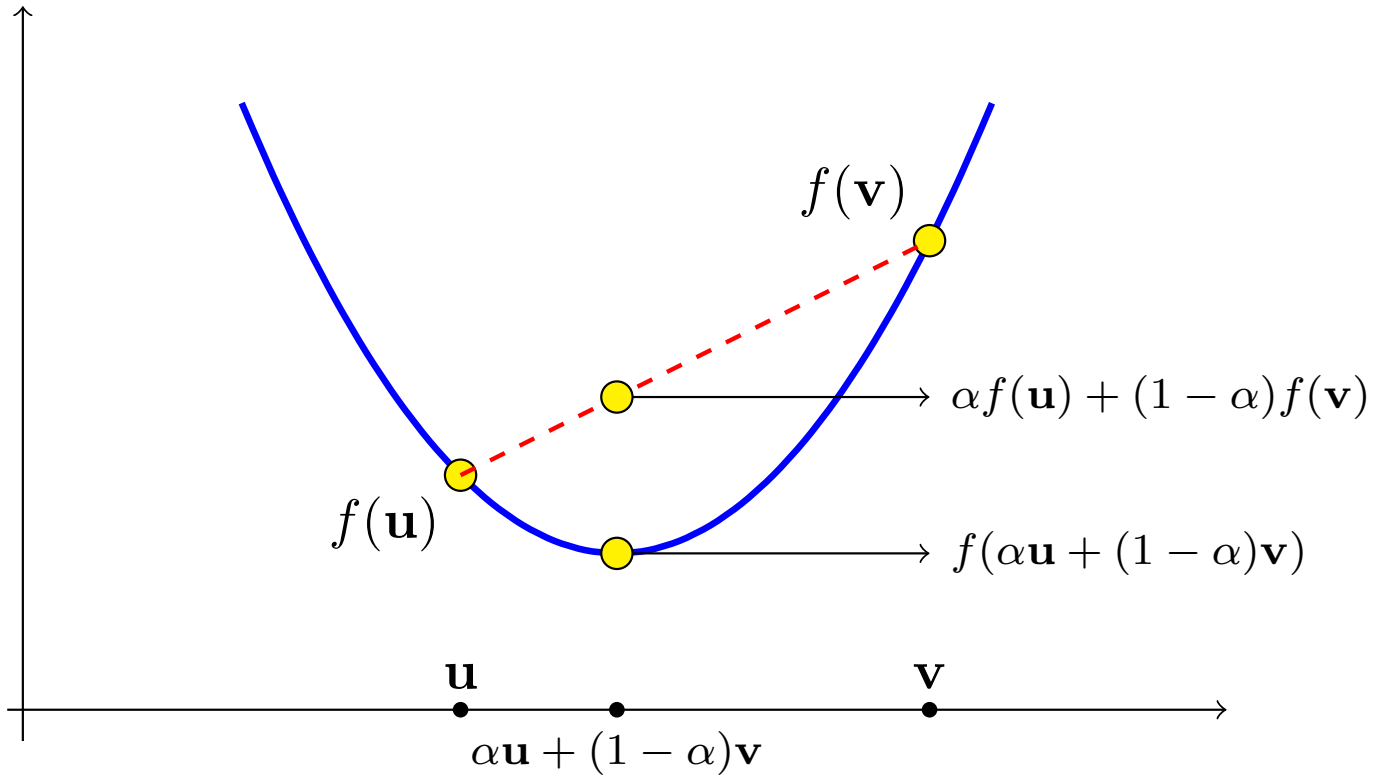
Un sous ensemble C d'un espace vectoriel est convexe si pour tout couple de vecteurs $\mathbf{u}, \mathbf{v}$ de C, le segment de droite entre $\mathbf{u}$ et $\mathbf{v}$ est contenu dans C. *i.e.*, pour tout $\alpha \in [0, 1]$, la combinaison convexe $\alpha\mathbf{u} + (1 - \alpha)\mathbf{v}$ appartient à C.



Définition (Fonction convexe)

Soit C un ensemble convexe. Une fonction $f : C \to \mathbb{R}$ est convexe si pour tout $\mathbf{u}, \mathbf{v} \in C$ et $\alpha \in [0, 1]$,

$$f(\alpha\mathbf{u} + (1-\alpha)\mathbf{v}) \;\leq\; \alpha f(\mathbf{u}) + (1-\alpha) f(\mathbf{v}) \;.$$



Inégalité de Jensen

Puisqu'une espérance est une combinaison convexe de plusieurs points, *e.g.*,

$$\mathbb{E}\,\mathbf{U} = \sum_{i=1}^{n} \mathbb{P}(\mathbf{U} = \mathbf{u}_i)\mathbf{u}_i\,,$$

on a l'inégalité suivante.

Lemme (Inégalité de Jensen)

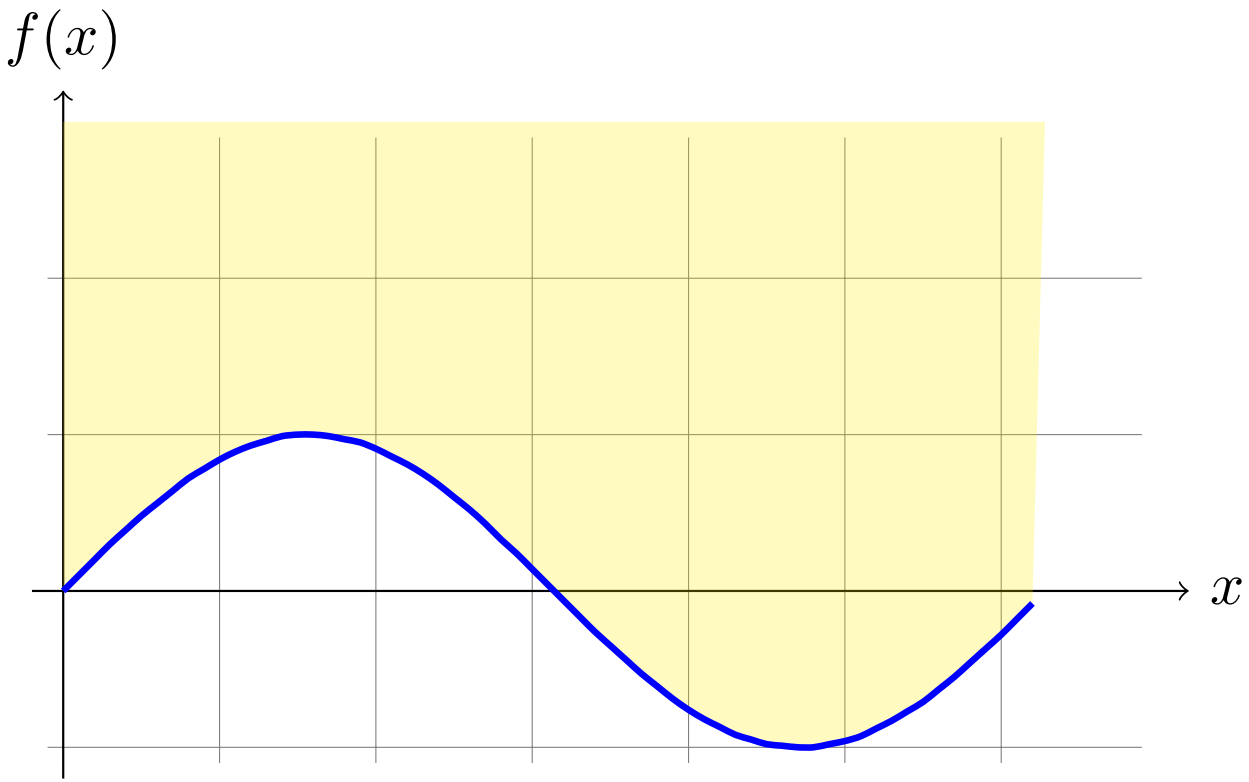
Soit $f : C \to \mathbb{R}$ une fonction convexe sur un domaine convexe. Soit une variable aléatoire $\mathbf{U}$ prenant des valeurs dans C. Alors

$$f\left(\mathbb{E}\,\mathbf{U}\right) \leq \mathbb{E}\, f(\mathbf{U})\,.$$

Epigraphe

Une fonction f est convexe si et seulement si son *épigraphe* est un ensemble convexe :

$$\text{epigraph(f)} \stackrel{\text{def}}{=} \{(\mathbf{x}, \beta) : f(\mathbf{x}) \leq \beta\} \ .$$



Les minimums locaux sont des minimums globaux

Si f est convexe alors tout minimum local de f est également un minimum global.

- Soit $B(\mathbf{u}, r) \stackrel{\text{def}}{=} \{\mathbf{v} : \|\mathbf{v} - \mathbf{u}\| \leq r\}$; avec $\|\mathbf{v}\| \stackrel{\text{def}}{=} \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle}$.
- Nous disons que $\mathbf{u}$ est un minimum local de f si $\exists r > 0$ tel que $f(\mathbf{v}) \geq f(\mathbf{u}),\ \forall \mathbf{v} \in B(\mathbf{u}, r)$.
- Alors, pour tout $\mathbf{v}$ (pas nécessairement dans B), il existe $0 \leq \alpha < 1$ suffisamment près de 1 tel que $\alpha\mathbf{u} + (1-\alpha)\mathbf{v} \in B(\mathbf{u}, r)$ et, alors

$$f(\mathbf{u}) \leq f(\alpha\mathbf{u} + (1-\alpha)\mathbf{v}) \ .$$

- Puisque f est convexe, nous avons également que

$$f(\alpha\mathbf{u} + (1-\alpha)\mathbf{v}) \leq \alpha f(\mathbf{u}) + (1-\alpha) f(\mathbf{v}) \ .$$

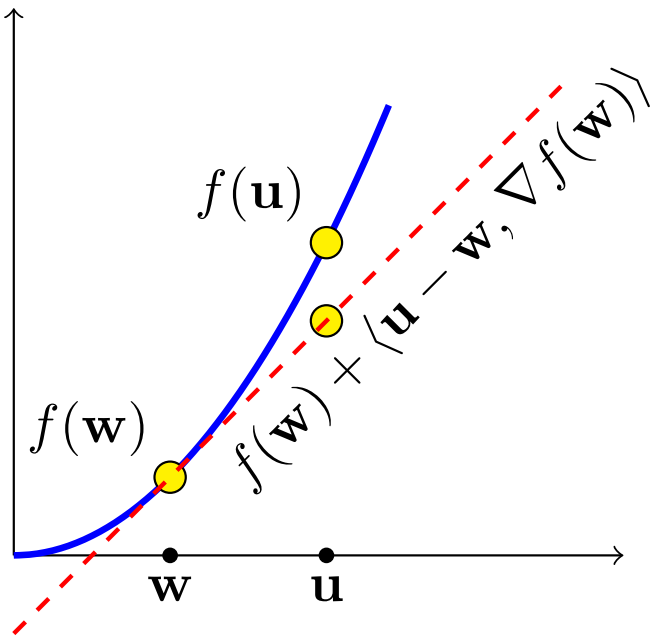
- En combinant, cela implique que $(1-\alpha) f(\mathbf{u}) \leq (1-\alpha) f(\mathbf{v}),\ \forall \mathbf{v}$.
- $\alpha < 1 \implies \mathbf{u}$ est également un minimum global de f.

Les tangentes de f convexe sont toujours sous f

Si f est convexe et différentiable en $\mathbf{w}$, alors

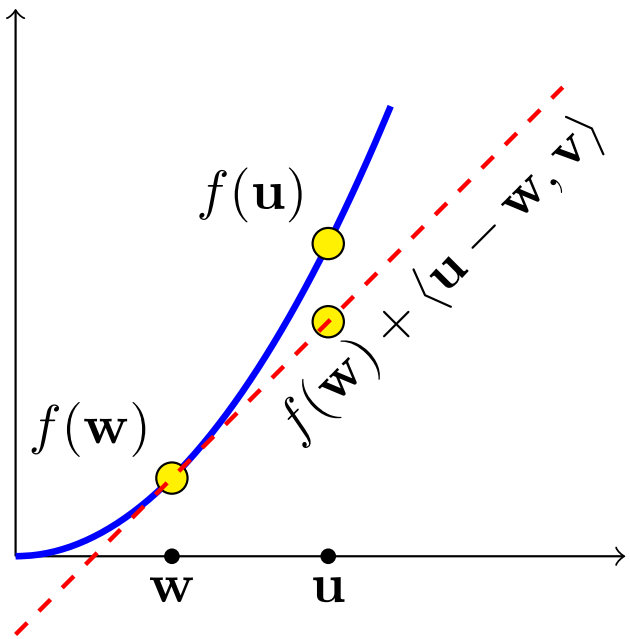
$$\forall \mathbf{u}, \quad f(\mathbf{u}) \geq f(\mathbf{w}) + \langle \nabla f(\mathbf{w}), \mathbf{u} - \mathbf{w} \rangle$$

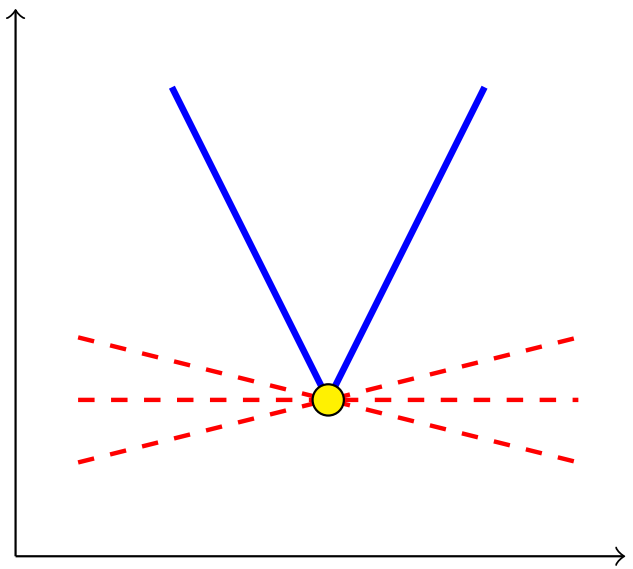
(Rappel : $\nabla f(\mathbf{w}) \stackrel{\text{def}}{=} \left(\frac{\partial f(\mathbf{w})}{\partial w_1}, \ldots, \frac{\partial f(\mathbf{w})}{\partial w_d} \right)$ est le gradient de f évalué à $\mathbf{w}$)



Sous-gradient d'une fonction

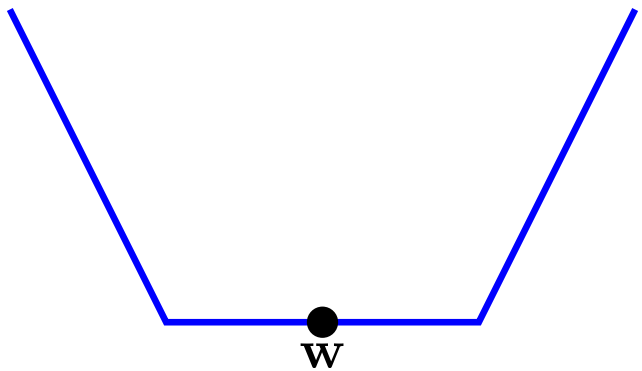
- Soit f une fonction sur un domaine convexe C.
- $\mathbf{v}$ est un sous-gradient de f à $\mathbf{w}$ si $\forall \mathbf{u}, \quad f(\mathbf{u}) \geq f(\mathbf{w}) + \langle \mathbf{v}, \mathbf{u} - \mathbf{w} \rangle$.
- Le sous-diférentiel de f à $\mathbf{w}$, $\partial f(\mathbf{w})$, est l'ensemble de tous les sous-gradients de f à $\mathbf{w}$.
  - Si f est différentiable à $\mathbf{w}$, alors $\partial f(\mathbf{w}) = \{\nabla f(\mathbf{w})\}$.
- Lemme : f est convexe ssi pour tout $\mathbf{w}$, $\partial f(\mathbf{w}) \neq \emptyset$.





Un fonction convexe est “localement platte” au minimum global

Soit f une fonction convexe. $\mathbf{w}$ est un minimum global de f ssi $\mathbf{0}$ est un sous-gradient de f en $\mathbf{w}$.



Propriété de Lipschitz

Rappel : pour tout vecteur $\mathbf{v}$, $\|\mathbf{v}\|$ désigne la norme Euclidienne de $\mathbf{v}$. *i.e.*, $\|\mathbf{v}\| = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle}$.

Définition (Fonction Lipschitzienne)

Une fonction $f : C \to \mathbb{R}$ est ρ-Lipschitzienne sur C si pour tout $\mathbf{w}_1, \mathbf{w}_2 \in C$ nous avons $|f(\mathbf{w}_1) - f(\mathbf{w}_2)| \leq \rho \|\mathbf{w}_1 - \mathbf{w}_2\|$.

Lemme

Si f est convexe sur un domaine C convexe et ouvert, alors f est ρ-Lipschitzienne sur C ssi pour tout $\mathbf{w} \in C$ et pour tout $\mathbf{v} \in \partial f(\mathbf{w})$, on a que $\|\mathbf{v}\| \leq \rho$. (i.e., *la norme de tous les sous gradients de f est au plus ρ).*

Preuve:

Supposons que pour tout $\mathbf{v} \in \partial f(\mathbf{w})$ et $\forall \mathbf{w} \in C$, on a que $\|\mathbf{v}\| \leq \rho$.

- $\mathbf{v} \in \partial f(\mathbf{w})$ implique $f(\mathbf{w}) - f(\mathbf{u}) \leq \langle \mathbf{v}, \mathbf{w} - \mathbf{u} \rangle$ pour tout $\mathbf{u} \in C$.
- Selon Cauchy-Schwarz : $\langle \mathbf{v}, \mathbf{w} - \mathbf{u} \rangle \leq \|\mathbf{v}\| \, \|\mathbf{w} - \mathbf{u}\| \leq \rho \|\mathbf{w} - \mathbf{u}\|$. Donc $f(\mathbf{w}) - f(\mathbf{u}) \leq \rho \|\mathbf{w} - \mathbf{u}\|$.
- $\mathbf{v} \in \partial f(\mathbf{u})$ implique $f(\mathbf{u}) - f(\mathbf{w}) \leq \langle \mathbf{v}, \mathbf{u} - \mathbf{w} \rangle$. Alors on a aussi $f(\mathbf{u}) - f(\mathbf{w}) \leq \rho \|\mathbf{u} - \mathbf{w}\|$.
- Donc $|f(\mathbf{w}) - f(\mathbf{u})| \leq \rho \|\mathbf{w} - \mathbf{u}\|$. Donc f est ρ-Lipschitzienne.

Supposons que f est ρ-Lipschitzienne.

- Choisissons $\mathbf{w} \in C$ et $\mathbf{v} \in \partial f(\mathbf{w})$.
- C ouvert implique $\exists \epsilon > 0 : \mathbf{u} = \mathbf{w} + \epsilon \mathbf{v} / \|\mathbf{v}\|$ appartient à C.
- Alors, $\langle \mathbf{u} - \mathbf{w}, \mathbf{v} \rangle = \epsilon \|\mathbf{v}\|$ et $\|\mathbf{u} - \mathbf{w}\| = \epsilon$.
- Puisque $\mathbf{v} \in \partial f(\mathbf{w})$, on a $f(\mathbf{u}) - f(\mathbf{w}) \geq \langle \mathbf{u} - \mathbf{w}, \mathbf{v} \rangle = \epsilon \|\mathbf{v}\|$.
- f est ρ-Lipschitzienne implique $\rho \epsilon = \rho \|\mathbf{u} - \mathbf{w}\| \geq f(\mathbf{u}) - f(\mathbf{w})$.
- En combinant ces 2 inégalités on a que $\|\mathbf{v}\| \leq \rho$.

Exemple des fonctions linéaires

- Pour $(\mathbf{x}, y)$ donné, considérez la fonction linéaire de $\mathbf{w}$:

$$f(\mathbf{w}) \stackrel{\text{def}}{=} \langle \mathbf{w}, \mathbf{x} \rangle - y \,.$$

- On a : $\nabla f(\mathbf{w}) = \mathbf{x}$, alors $\partial f(\mathbf{w}) = \{\mathbf{x}\}$.
- Le lemme précédent implique que f est $\|\mathbf{x}\|$-Lipschitzienne.
  - Donc lorsque $\|\mathbf{x}\| \leq R \; \forall \mathbf{x} \in \mathcal{X}$, on a que f est R-Lipschitzienne quelque soit $(\mathbf{x}, y)$.
- Plusieurs fonctions de perte sont des fonctions de $\langle \mathbf{w}, \mathbf{x} \rangle - y$.
  - Exemple 1 : $|\langle \mathbf{w}, \mathbf{x} \rangle - y|$.
  - Exemple 2 : $(\langle \mathbf{w}, \mathbf{x} \rangle - y)^2$.
- Que pouvons nous dire de la propriété de Lipschitz de $g(\langle \mathbf{w}, \mathbf{x} \rangle - y)$?

Propriété de Lipschitz pour composition de fonctions

Lemme

Si g_1 est ρ_1-Lipschitzienne et que g_2 est ρ_2-Lipschitzienne, alors $g_1 \circ g_2$ est $\rho_1\rho_2$-Lipschitzienne.

Démonstration.

$$\begin{aligned} |g_1(g_2(\mathbf{w}_a)) - g_1(g_2(\mathbf{w}_b))| &\leq \rho_1 \|g_2(\mathbf{w}_a) - g_2(\mathbf{w}_b)\| \\ &\leq \rho_1\rho_2 \|\mathbf{w}_a - \mathbf{w}_b\| \end{aligned}$$

□

Propriété de Lipschitz de fonctions de perte

- $|x|$ est 1-Lipschitzienne car tous ses sous gradients ont une valeur absolue ≤ 1.
- Donc, pour un $(\mathbf{x}, y)$ donné, $|\langle \mathbf{w}, \mathbf{x} \rangle - y|$ est $\|\mathbf{x}\|$-Lipschitzienne.
- x^2 n'est pas Lipschitzienne car sa dérivée n'est pas bornée.
- Mais x^2 est 2ρ-Lipschitzienne sur $C \stackrel{\text{def}}{=} \{x : |x| < \rho\}$ car la valeur absolue de sa dérivée sur C est $< 2\rho$.
- Considérons $(\langle \mathbf{w}, \mathbf{x} \rangle - y)^2$ pour un $(\mathbf{x}, y)$ donné.
  - Nous venons de voir qu'il faut que $|\langle \mathbf{w}, \mathbf{x} \rangle - y|$ soit borné.
  - Or, $|\langle \mathbf{w}, \mathbf{x} \rangle - y| \leq |\langle \mathbf{w}, \mathbf{x} \rangle| + |y|$.
  - Considérons que $\|\mathbf{w}\| < B$.
  - Inégalité de Schwarz : $-\|\mathbf{w}\| \, \|\mathbf{x}\| \leq \langle \mathbf{w}, \mathbf{x} \rangle \leq \|\mathbf{w}\| \, \|\mathbf{x}\|$.
  - Dans ce cas, $|\langle \mathbf{w}, \mathbf{x} \rangle| + |y| < B\|\mathbf{x}\| + |y|$.
- Donc $(\langle \mathbf{w}, \mathbf{x} \rangle - y)^2$ est $2(B\|\mathbf{x}\| + |y|)$-Lipschitzienne dans un domaine pour $\mathbf{w}$ satisfaisant $\|\mathbf{w}\| < B$.

Sommaire

La descente de gradient

- Débuter avec $\mathbf{w}^{(1)}$ (habituellement $\mathbf{w}^{(1)} = \mathbf{0}$)
- À l'itération t, mettre à jour

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \nabla f(\mathbf{w}^{(t)}) \ ,$$

  où $\eta > 0$ est un paramètre.
- Intuition :
  - Approximation par série de Taylor : si $\mathbf{w}$ est près de $\mathbf{w}^{(t)}$, alors

$$f(\mathbf{w}) \approx f(\mathbf{w}^{(t)}) + \langle \mathbf{w} - \mathbf{w}^{(t)}, \nabla f(\mathbf{w}^{(t)}) \rangle \ .$$

  - On désire minimiser l'approximation tout en demeurant près de $\mathbf{w}^{(t)}$:

$$\mathbf{w}^{(t+1)} = \operatorname*{argmin}_{\mathbf{w}} \frac{1}{2} \|\mathbf{w} - \mathbf{w}^{(t)}\|^2 + \eta \left(f(\mathbf{w}^{(t)}) + \langle \mathbf{w} - \mathbf{w}^{(t)}, \nabla f(\mathbf{w}^{(t)}) \rangle \right) \ ,$$

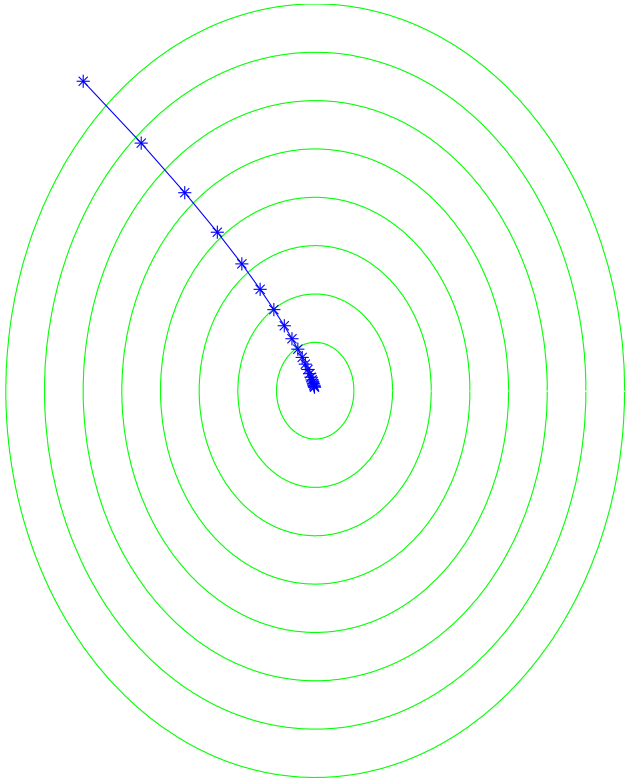
  - ce qui donne notre règle de mise à jour.

La descente de gradient

- Initialiser $\mathbf{w}^{(1)} = \mathbf{0}$
- Effectuer les mises à jour

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \nabla f(\mathbf{w}^{(t)}) \,.$$

- Sortie : $\bar{\mathbf{w}} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{w}^{(t)}$



La descente de sous-gradient

Il suffit de remplacer les gradients par des sous-gradients :

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \mathbf{v}_t \ ,$$

avec $\mathbf{v}_t \in \partial f(\mathbf{w}^{(t)})$.

- C'est une généralisation de la descente de gradient pour les fonctions f convexes, mais non différentiables

Analyse de la descente de sous-gradient

Théorème (convergence de la descente de sous-gradient)

Soit f une fonction convexe ρ-Lipschitzienne et

$$\mathbf{w}^\star \stackrel{def}{=} \operatorname*{argmin}_{\mathbf{w}:\|\mathbf{w}\|\leq B} f(\mathbf{w})\,.$$

Si nous exécutons la descente de sous-gradient durant T itérations avec $\eta = B/(\rho\sqrt{T})$, alors le vecteur $\bar{\mathbf{w}}$ produit par cet algorithme satisfait

$$f(\bar{\mathbf{w}}) - f(\mathbf{w}^\star) \leq \frac{B\rho}{\sqrt{T}}\,.$$

De plus, pour tout $\epsilon > 0$, pour obtenir $f(\bar{\mathbf{w}}) - f(\mathbf{w}^\star) \leq \epsilon$, il suffit d'exécuter la descente de sous-gradient pour un nombre T d'itérations satisfaisant

$$T \geq \frac{B^2\rho^2}{\epsilon^2}\,.$$

Preuve du théorème de convergence

La preuve s'appuie sur 2 lemmes. Voici le premier.

Lemme

$$f(\bar{\mathbf{w}}) - f(\mathbf{w}^\star) \leq \frac{1}{T}\sum_{t=1}^{T}\langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathbf{v}_t\rangle \quad ; \quad avec\ \mathbf{v}_t \in \partial f(\mathbf{w}^{(t)})\,.$$

Preuve: L'inégalité de Jensen nous donne

$$\begin{aligned} f(\bar{\mathbf{w}}) - f(\mathbf{w}^\star) &= f\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{w}^{(t)}\right) - f(\mathbf{w}^\star) \\ &\leq \frac{1}{T}\sum_{t=1}^{T}\Big(f(\mathbf{w}^{(t)})\Big) - f(\mathbf{w}^\star) \\ &= \frac{1}{T}\sum_{t=1}^{T}\Big(f(\mathbf{w}^{(t)}) - f(\mathbf{w}^\star)\Big) \end{aligned}$$

Preuve du théorème de convergence

- Puisque f est convexe et que $\mathbf{v}_t \in \partial f(\mathbf{w}^{(t)})$, nous avons pour tout t

$$f(\mathbf{w}^\star) \geq f(\mathbf{w}^{(t)}) + \langle \mathbf{w}^\star - \mathbf{w}^{(t)}, \mathbf{v}_t \rangle \, .$$

- De manière équivalente, on a

$$f(\mathbf{w}^{(t)}) - f(\mathbf{w}^\star) \leq \langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathbf{v}_t \rangle \, .$$

- Le lemme est obtenu en combinant cela avec le résultat au bas de la page précédente.

■

Preuve du théorème de convergence

Voici le deuxième lemme.

Lemme

Soit $\mathbf{v}_1, \ldots, \mathbf{v}_t$ *une séquence arbitraire de vecteurs. La règle de mise à jour* $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \mathbf{v}_t$, *avec* $\mathbf{w}^{(1)} = \mathbf{0}$, *satisfait*

$$\sum_{t=1}^{T} \langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathbf{v}_t \rangle \leq \frac{\|\mathbf{w}^\star\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|\mathbf{v}_t\|^2, \quad \forall \mathbf{w}^\star \, .$$

Preuve: En complétant les carrés nous obtenons

$$\begin{aligned}
\langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathbf{v}_t \rangle &= \frac{1}{\eta} \langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \eta \mathbf{v}_t \rangle \\
&= \frac{1}{2\eta} (-\|\mathbf{w}^{(t)} - \mathbf{w}^\star - \eta \mathbf{v}_t\|^2 + \|\mathbf{w}^{(t)} - \mathbf{w}^\star\|^2 + \eta^2 \|\mathbf{v}_t\|^2) \\
&= \frac{1}{2\eta} (-\|\mathbf{w}^{(t+1)} - \mathbf{w}^\star\|^2 + \|\mathbf{w}^{(t)} - \mathbf{w}^\star\|^2) + \frac{\eta}{2} \|\mathbf{v}_t\|^2
\end{aligned}$$

Preuve du théorème de convergence

En sommant la dernière égalité sur t, nous obtenons

$$\sum_{t=1}^{T}\langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathbf{v}_t\rangle =$$

$$\frac{1}{2\eta}\sum_{t=1}^{T}(-\|\mathbf{w}^{(t+1)} - \mathbf{w}^\star\|^2 + \|\mathbf{w}^{(t)} - \mathbf{w}^\star\|^2) + \frac{\eta}{2}\sum_{t=1}^{T}\|\mathbf{v}_t\|^2$$

$$= \frac{1}{2\eta}(\|\mathbf{w}^{(1)} - \mathbf{w}^\star\|^2 - \|\mathbf{w}^{(T+1)} - \mathbf{w}^\star\|^2) + \frac{\eta}{2}\sum_{t=1}^{T}\|\mathbf{v}_t\|^2$$

$$\leq \frac{1}{2\eta}\|\mathbf{w}^{(1)} - \mathbf{w}^\star\|^2 + \frac{\eta}{2}\sum_{t=1}^{T}\|\mathbf{v}_t\|^2$$

$$= \frac{1}{2\eta}\|\mathbf{w}^\star\|^2 + \frac{\eta}{2}\sum_{t=1}^{T}\|\mathbf{v}_t\|^2$$

■

Preuve du théorème de convergence

Preuve: (du théorème)

- La combinaison des deux lemmes nous donne

$$f(\bar{\mathbf{w}}) - f(\mathbf{w}^\star) \leq \frac{1}{T}\frac{1}{2\eta}\|\mathbf{w}^\star\|^2 + \frac{1}{T}\frac{\eta}{2}\sum_{t=1}^{T}\|\mathbf{v}_t\|^2$$

- En utilisant $\|\mathbf{w}^\star\| \leq B$ et $\|\mathbf{v}_t\| \leq \rho$, nous avons

$$f(\bar{\mathbf{w}}) - f(\mathbf{w}^\star) \leq \frac{B^2}{2\eta T} + \frac{\eta\rho^2}{2}\,.$$

- La plus petite valeur de la borne supérieure est obtenue avec $\eta = B/(\rho\sqrt{T})$. Ce qui donne

$$f(\bar{\mathbf{w}}) - f(\mathbf{w}^\star) \leq \frac{B\rho}{\sqrt{T}}\,.$$

Exemple : Trouver un hyperplan séparateur

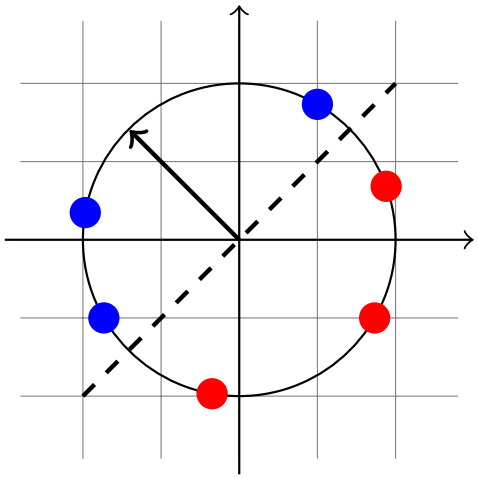
Soit $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$ avec $y_i \in \{-1, +1\}$. On désire trouver $\mathbf{w}$ séparant les exemples :

$$\forall i, \quad y_i \langle \mathbf{w}, \mathbf{x}_i \rangle > 0 \ .$$

- Cherchons un hyperplan séparateur tel que $\|\mathbf{w}\| = 1$.
- s.p.d.g. supposons que $\|\mathbf{x}_i\| \leq 1$ pour tout i.
- Nous désirons trouver l'hyperplan séparateur $\mathbf{w}^*$ qui maximise la marge minimale de séparation, *i.e.* :

$$\mathbf{w}^* = \underset{\mathbf{w}:||\mathbf{w}||=1}{\operatorname{argmax}} \left(\min_i y_i \langle \mathbf{w}, \mathbf{x}_i \rangle \right)$$

- $\gamma \stackrel{\text{def}}{=} \min_i y_i \langle \mathbf{w}^*, \mathbf{x}_i \rangle$ est la marge obtenue par $\mathbf{w}^*$.



Exemple : Trouver un hyperplan séparateur

Utilisons les égalités suivantes :

$$\begin{aligned} \min_i y_i \langle \mathbf{w}, \mathbf{x}_i \rangle &= -\max_i -y_i \langle \mathbf{w}, \mathbf{x}_i \rangle \\ \max_{\mathbf{w}:||\mathbf{w}||=1} g(\mathbf{w}) &= -\min_{\mathbf{w}:||\mathbf{w}||=1} -g(\mathbf{w}) \,. \end{aligned}$$

Cela nous donne

$$\begin{aligned} \max_{\mathbf{w}:||\mathbf{w}||=1} \min_i y_i \langle \mathbf{w}, \mathbf{x}_i \rangle &= \max_{\mathbf{w}:||\mathbf{w}||=1} -\max_i -y_i \langle \mathbf{w}, \mathbf{x}_i \rangle \\ &= -\min_{\mathbf{w}:||\mathbf{w}||=1} \underbrace{\max_i -y_i \langle \mathbf{w}, \mathbf{x}_i \rangle}_{f(\mathbf{w})} \,. \end{aligned}$$

Exemple : Trouver un hyperplan séparateur

Notre problème de maximisation est donc équivalent au problème de minimisation :

$$\min_{\mathbf{w}} f(\mathbf{w}) \qquad \text{avec} \qquad f(\mathbf{w}) \;=\; \max_{i} -y_i \langle \mathbf{w}, \mathbf{x}_i \rangle$$

Observations :

- f est convexe (car le max de fonctions convexes est convexe).
- Un sous gradient de f à $\mathbf{w}$ est $-y_i \mathbf{x}_i$ pour $i \in \operatorname{argmax} -y_i \langle \mathbf{w}, \mathbf{x}_i \rangle$.
- f est 1-Lipschitz et $||\mathbf{w}^*|| = 1 \;\Rightarrow\; f(\bar{\mathbf{w}}) - f(\mathbf{w}^*) \leq \frac{1}{\sqrt{T}}$.
- Or, selon la page précédente, $f(\mathbf{w}^*) = -\gamma$.
- Donc, après $T > \frac{1}{\gamma^2}$ itérations. on aura $f(\bar{\mathbf{w}}) < -\gamma + \gamma = 0$.
- Alors, $\bar{\mathbf{w}}$ est un hyperplan séparateur après ce nombre d'itérations.

Descente de sous gradient vs. le perceptron

Algorithme de descente de sous-gradient :

- Initialiser $\mathbf{w}^{(1)} = \mathbf{0}$.
- Pour $t = 1, \ldots, T$:
  - Soit $i = \operatorname{argmax} -y_i \langle \mathbf{w}^{(t)}, \mathbf{x}_i \rangle = \operatorname{argmin} y_i \langle \mathbf{w}^{(t)}, \mathbf{x}_i \rangle$.
    - Mettre à jour : $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + \eta y_i \mathbf{x}_i$
- Retourner $\bar{\mathbf{w}} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{w}^{(t)}$.

Remarque : Le théorème de convergence fixe $\eta = 1/\sqrt{T}$ dans notre cas. Mais ici η ne fait que modifier la norme de $\bar{\mathbf{w}}$ (et n'a pas d'influence sur sa direction). On peut donc choisir $\eta = 1$.

En comparaison, voici **l'algorithme du perceptron en mode batch :**

- Initialiser $\mathbf{w}^{(1)} = \mathbf{0}$
- Tant qu'il existe i tel que $y_i \langle \mathbf{w}^{(t)}, \mathbf{x}_i \rangle \leq 0$, mettre à jour :

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + y_i \mathbf{x}_i$$

Sommaire

Problèmes d'apprentissage convexes

- Notre formalization du problème d'apprentissage concerne un ensemble $Z = \mathcal{X} \times \mathcal{Y}$ de paires (instance, étiquette) observables, une classe $\mathcal{H}$ d'hypothèses et une fonction de perte $\ell : \mathcal{H} \times Z \to \mathbb{R}_+$.
- Dans ce qui suit, $\mathcal{H}$ est un sous ensemble de $\mathbb{R}^d$ et nous désignerons par $\mathbf{w}$ les éléments de $\mathcal{H}$.
- $\forall z \in Z$, $\ell(\cdot, z)$ est la fonction $f : \mathcal{H} \to \mathbb{R}_+$ telle que $f(\mathbf{w}) = \ell(\mathbf{w}, z)$.

Définition (problème d'apprentissage convexe)

Un problème d'apprentissage $(\mathcal{H}, Z, \ell)$ est dit convexe si la classe $\mathcal{H}$ est un ensemble convexe et pour tout $z \in Z$, la fonction de perte $\ell(\cdot, z)$ est une fonction convexe.

- L'algorithme $\mathrm{ERM}_{\mathcal{H}}$ pour un problème d'apprentissage convexe nous donne un problème d'optimisation convexe : $\min_{\mathbf{w} \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \ell(\mathbf{w}, z_i)$
- Exemple – moindres carrés : $\mathcal{H} = \mathbb{R}^d$, $Z = \mathbb{R}^d \times \mathbb{R}$, $\ell(\mathbf{w}, (\mathbf{x}, y)) = (\langle \mathbf{w}, \mathbf{x} \rangle - y)^2$

Apprenabilité des problèmes d'apprentissage convexes

- Affirmation : Ce n'est pas tous les problèmes convexes qui sont apprenables (voir exemples, section 12.2).
- Mais nous obtenons l'apprenabilité en introduisant ces deux conditions :
  - $\mathcal{H}$ est borné.
  - La fonction de perte est Lipschitzienne.

Problème d'apprentissage convexe-Lipschitzien-borné

Définition (problème d'apprentissage convexe-Lipschitzien-borné)

Un problème d'apprentissage $(\mathcal{H}, Z, \ell)$ est convexe-Lipschitzien-borné avec paramètres ρ, B si les propriétés suivantes sont satisfaites :

- $\mathcal{H}$ est un ensemble convexe et pour tout $\mathbf{w} \in \mathcal{H}$, nous avons $\|\mathbf{w}\| \leq B$.
- Pour tout $z \in Z$, la fonction de perte $\ell(\cdot, z)$ est convexe et ρ-Lipschitzienne.

Exemple :

- $\mathcal{H} = \{\mathbf{w} \in \mathbb{R}^d : \|\mathbf{w}\| \leq B\}$
- Avec $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| \leq \rho\}$, $\mathcal{Y} = \mathbb{R}$,
- Avec $\ell(\mathbf{w}, (\mathbf{x}, y)) = |\langle \mathbf{w}, \mathbf{x} \rangle - y|$.
- Exercice : qu'en est-il pour $(\langle \mathbf{w}, \mathbf{x} \rangle - y)^2$?

Apprenabilité

- Nous allons démontrer plus loin que les problèmes d'apprentissage convexes-Lipschitziens-bornés sont apprenables avec une complexité d'échantillon qui dépends (uniquement) de $(\epsilon, \delta, B, \rho)$.
  - Ils sont donc agnostiquement PAC apprenables.
- Nous présenterons deux algorithmes permettant d'apprendre ces problèmes :
  - Le descente de gradient stochastique (ce chapitre).
  - La minimisation du risque empirique régularisé (prochain chapitre).
- Dans le manuel, il est également démontré que les problèmes d'apprentissage convexes-harmonieux-bornés sont apprenables mais nous le démontrerons pas en classe.
  - Une fonction de perte est harmonieuse lorsque son gradient possède la propriété de Lipschitz.

Sommaire

Fonction de perte substitut ("surrogate")

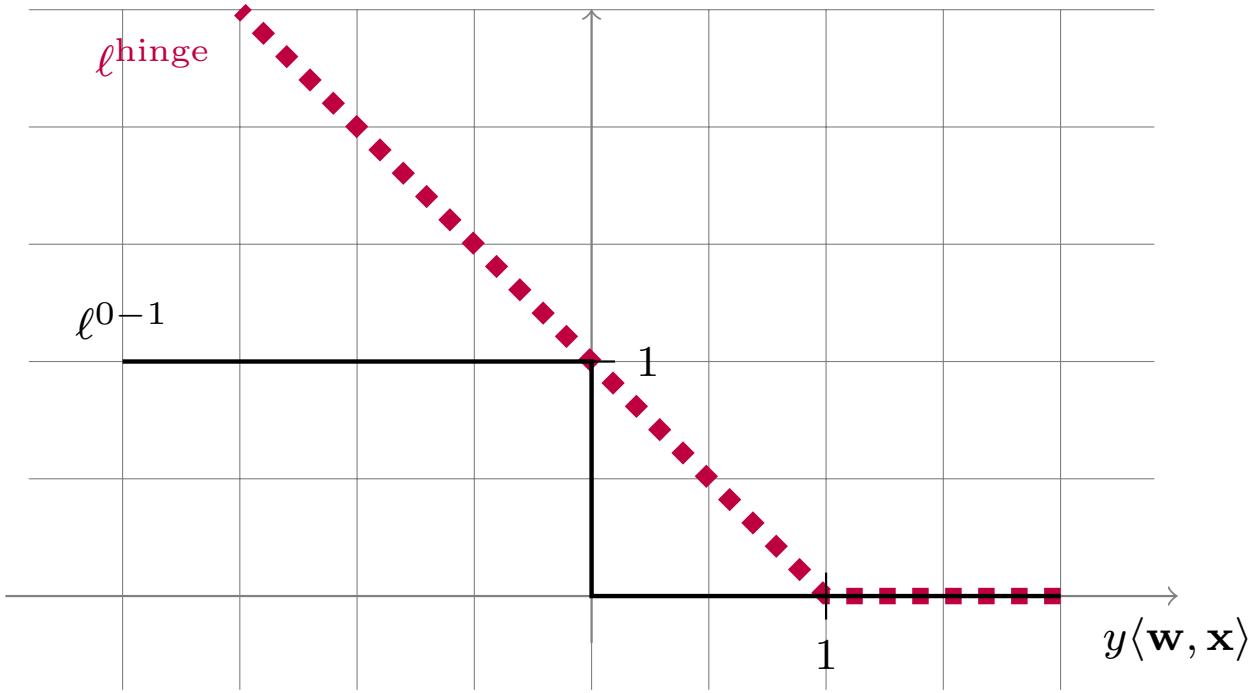
- Souvent, la fonction de perte que nous désirons utiliser n'est pas convexe.
- Par exemple, la perte $0-1$ avec des demi-espaces.

$$\ell^{0-1}(\mathbf{w},(\mathbf{x},y)) = \mathbb{1}_{[y \neq \text{sign}(\langle \mathbf{w},\mathbf{x}\rangle)]} \leq \mathbb{1}_{[y\langle \mathbf{w},\mathbf{x}\rangle \leq 0]} \ .$$

- Si la fonction de perte est non-convexe, $\text{ERM}_{\mathcal{H}}(S)$ est habituellement un problème d'optimisation NP-difficile.
- Approche très utilisée : contourner la difficulté du problème en utilisant une fonction de perte substitut qui doit :
  - être convexe,
  - borner supérieurement la fonction de perte désirée.

Fonction de perte charnière ("hinge-loss")

$$\ell^{\text{hinge}}(\mathbf{w}, (\mathbf{x}, y)) \stackrel{\text{def}}{=} \max\{0, 1 - y\langle \mathbf{w}, \mathbf{x} \rangle\} \ .$$



Nouvelle décomposition de l'erreur

- Supposons qu'un apprenant A utilisant ℓ^{hinge} nous garantit :

$$L_{\mathcal{D}}^{\text{hinge}}(A(S)) \;\leq\; \min_{\mathbf{w}\in\mathcal{H}} L_{\mathcal{D}}^{\text{hinge}}(\mathbf{w}) + \epsilon \;,$$

- Puisque ℓ^{hinge} borne supérieurement ℓ^{0-1}, nous avons,

$$L_{\mathcal{D}}^{0-1}(A(S)) \;\leq L_{\mathcal{D}}^{\text{hinge}}(A(S)) \;\leq\; \min_{\mathbf{w}\in\mathcal{H}} L_{\mathcal{D}}^{\text{hinge}}(\mathbf{w}) + \epsilon \;.$$

- La borne supérieure peut donc s'écrire :

$$\begin{aligned} L_{\mathcal{D}}^{0-1}(A(S)) &\leq \min_{\mathbf{w}\in\mathcal{H}} L_{\mathcal{D}}^{0-1}(\mathbf{w}) + \left(\min_{\mathbf{w}\in\mathcal{H}} L_{\mathcal{D}}^{\text{hinge}}(\mathbf{w}) - \min_{\mathbf{w}\in\mathcal{H}} L_{\mathcal{D}}^{0-1}(\mathbf{w}) \right) + \epsilon \\ &= \epsilon_{\text{approximation}} + \epsilon_{\text{optimization}} + \epsilon_{\text{estimation}} \end{aligned}$$

- L'erreur d'optimization résulte de l'utilisation de ℓ^{hinge} comme substitut à ℓ^{0-1} (suite à notre incapacité d'optimiser à l'aide de ℓ^{0-1}).

Sommaire

Apprendre avec la descente de gradient stochastique

- Considérons un problème d'apprentissage convexe-Lipschitzien-borné.
- Rappel : notre but est de résoudre (probablement et approximativement) :

$$\min_{\mathbf{w} \in \mathcal{H}} L_{\mathcal{D}}(\mathbf{w}) \quad \text{où} \quad L_{\mathcal{D}}(\mathbf{w}) \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{z \sim \mathcal{D}} [\ell(\mathbf{w}, z)]$$

- Jusqu'à maintenant l'apprenant se concentrait sur $L_S(\mathbf{w})$.
- Considérons la possibilité de minimiser directement $L_{\mathcal{D}}(\mathbf{w})$.

Descente de gradient stochastique pour minimiser $L_{\mathcal{D}}(\mathbf{w})$

$$\min_{\mathbf{w}\in\mathcal{H}} L_{\mathcal{D}}(\mathbf{w}) \quad \text{où} \quad L_{\mathcal{D}}(\mathbf{w}) = \mathop{\mathbb{E}}_{z\sim\mathcal{D}}[\ell(\mathbf{w}, z)]$$

- Descente de gradient pour $L_{\mathcal{D}}(\mathbf{w})$: on débute avec $\mathbf{w}^{(1)} = \mathbf{0}$ et on met à jour $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \nabla L_{\mathcal{D}}(\mathbf{w})$.
- Il est impossible de calculer $\nabla L_{\mathcal{D}}(\mathbf{w})$ car nous ne connaissons pas $\mathcal{D}$.
- Mais nous pouvons l'estimer avec $\nabla \ell(\mathbf{w}, z)$ pour $z \sim \mathcal{D}$.
- En fait, $\nabla \ell(\mathbf{w}, z)$ est un estimateur non biaisé de $\nabla L_{\mathcal{D}}(\mathbf{w})$ car

$$\mathop{\mathbb{E}}_{z\sim\mathcal{D}}[\nabla \ell(\mathbf{w}, z)] \;=\; \nabla L_{\mathcal{D}}(\mathbf{w})\,.$$

- Lorsque l'on bouge dans la direction de $= -\nabla \ell(\mathbf{w}, z)$ alors, en espérance, on bouge dans la direction de $-\nabla L_{\mathcal{D}}(\mathbf{w})$.
- Nous allons voir que cela est suffisant.

Descente de sous-gradient stochastique

- Démontrons que tout sous-gradient de $\ell(\mathbf{w}, z)$ au point $\mathbf{w}$ est un estimateur non biaisé d'un sous-gradient de $L_{\mathcal{D}}(\mathbf{w})$ au point $\mathbf{w}$.
  - **Preuve:** Si $\mathbf{v}(\mathbf{w}, z)$ est un sous-gradient de $\ell(\mathbf{w}, z)$ au point $\mathbf{w}$, alors pour tout ℓ convexe et pour tout $\mathbf{u}$ on a

$$\ell(\mathbf{u}, z) - \ell(\mathbf{w}, z) \geq \langle \mathbf{u} - \mathbf{w}, \mathbf{v}(\mathbf{w}, z) \rangle .$$

  - Alors en prenant l'espérance sur $z \sim \mathcal{D}$ de chaque côté on obtient

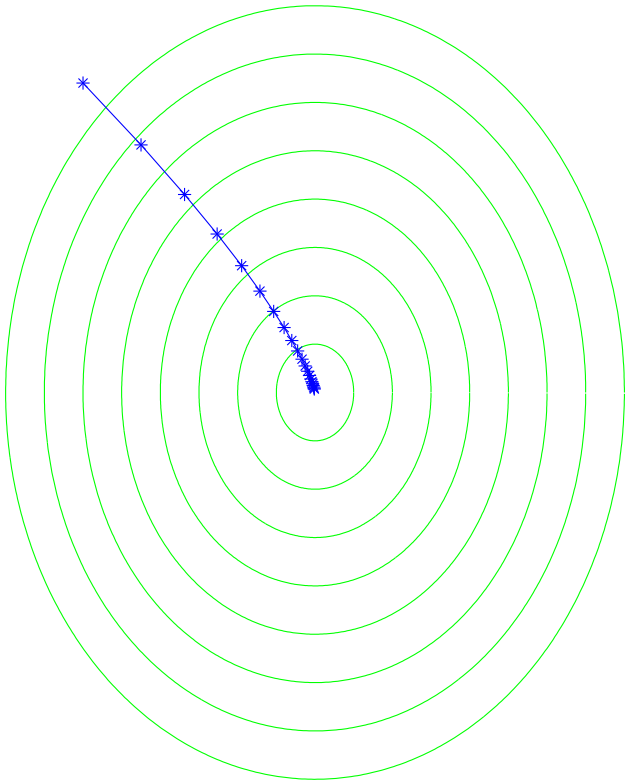
$$L_{\mathcal{D}}(\mathbf{u}) - L_{\mathcal{D}}(\mathbf{w}) \geq \langle \mathbf{u} - \mathbf{w}, \mathop{\mathbb{E}}_{z} \mathbf{v}(\mathbf{w}, z) \rangle .$$

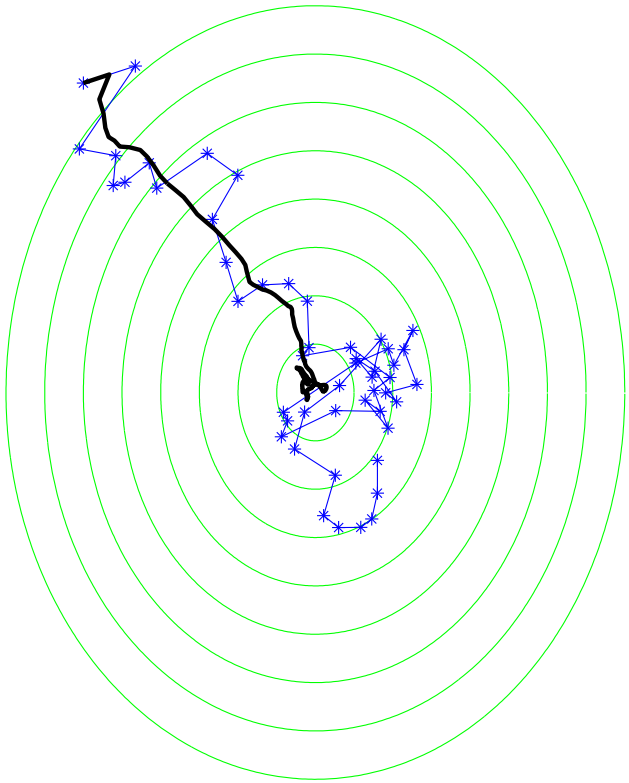
  - Donc, par définition du sous-gradient, cela implique que $\mathbb{E}_z \, \mathbf{v}(\mathbf{w}, z)$ est un sous-gradient de $L_{\mathcal{D}}(\mathbf{w})$ évalué au point $\mathbf{w}$ ■

.

Descente de gradient stochastique pour minimiser $L_{\mathcal{D}}(\mathbf{w})$

- **initialiser :** $\mathbf{w}^{(1)} = \mathbf{0}$
- **pour** $t = 1, 2, \ldots, T$
  - tirer $z_t \sim \mathcal{D}$
  - soit $\mathbf{v}_t \in \partial \ell(\mathbf{w}^{(t)}, z_t)$
  - mise à jour : $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \mathbf{v}_t$
- **sortie :** $\bar{\mathbf{w}} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{w}^{(t)}$





DGS pour problèmes convexes-Lipschitziens-bornés

Revenons au lemme suivant :

Lemme

Soit $\mathbf{v}_1, \ldots, \mathbf{v}_t$ *une séquence arbitraire de vecteurs. La règle de mise à jour* $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \mathbf{v}_t$, *avec* $\mathbf{w}^{(1)} = \mathbf{0}$, *satisfait*

$$\sum_{t=1}^{T} \langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathbf{v}_t \rangle \leq \frac{\|\mathbf{w}^\star\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|\mathbf{v}_t\|^2, \quad \forall \mathbf{w}^\star.$$

Supposons que $\|\mathbf{v}_t\| \leq \rho$ pour tout t et que $\|\mathbf{w}^\star\| \leq B$, alors on a

$$\sum_{t=1}^{T} \langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathbf{v}_t \rangle \;\leq\; \frac{B^2}{2\eta} + \frac{\eta\, \rho^2\, T}{2} \;=\; B\, \rho \sqrt{T}\ ,$$

lorsque $\eta = \sqrt{\frac{B^2}{\rho^2\, T}}$.

DGS pour problèmes convexes-Lipschitziens-bornés

En prenant l'espérance sur $z_1, \ldots, z_T$, on a

$$\mathop{\mathbb{E}}_{z_1,\ldots,z_T}\left[\sum_{t=1}^{T}\langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathbf{v}_t\rangle\right] \;\leq\; B\,\rho\,\sqrt{T}\;.$$

Considérons d'abord le cas $\mathbf{v}_t = \nabla\ell(\mathbf{w}^{(t)}, z_t)$. Puisque chaque $z_t \sim \mathcal{D}$, et que $\mathbf{w}^{(t)}$ ne dépends que de $z_1, \ldots, z_{t-1}$, on a

$$\begin{aligned}
\mathop{\mathbb{E}}_{z_1,\ldots,z_T}[\langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathbf{v}_t\rangle] &= \mathop{\mathbb{E}}_{z_1,\ldots,z_t}[\langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \nabla\ell(\mathbf{w}^{(t)}, z_t)\rangle] \\
&= \mathop{\mathbb{E}}_{z_1,\ldots,z_{t-1}}[\langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathop{\mathbb{E}}_{z_t}\nabla\ell(\mathbf{w}^{(t)}, z_t)\rangle] \\
&= \mathop{\mathbb{E}}_{z_1,\ldots,z_{t-1}}[\langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \nabla L_{\mathcal{D}}(\mathbf{w}^{(t)})\rangle] \\
&= \mathop{\mathbb{E}}_{z_1,\ldots,z_T}[\langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \nabla L_{\mathcal{D}}(\mathbf{w}^{(t)})\rangle]\;.
\end{aligned}$$

DGS pour problèmes convexes-Lipschitziens-bornés

Nous avons alors

$$\mathop{\mathbb{E}}_{z_1,\ldots,z_T}\left[\sum_{t=1}^{T}\langle \mathbf{w}^{(t)} - \mathbf{w}^{\star}\ ,\ \nabla L_{\mathcal{D}}(\mathbf{w}^{(t)})\rangle\right] \ \leq\ B\,\rho\,\sqrt{T}\ .$$

De plus, la convexité de $\ell(\mathbf{w}, z)$ implique que l'on a

$$L_{\mathcal{D}}(\mathbf{w}^{(t)}) - L_{\mathcal{D}}(\mathbf{w}^{\star}) \leq \langle \mathbf{w}^{(t)} - \mathbf{w}^{\star}\ ,\ \nabla L_{\mathcal{D}}(\mathbf{w}^{(t)})\rangle\,.$$

Donc, en combinant, nous avons

$$\mathop{\mathbb{E}}_{z_1,\ldots,z_T}\left[\sum_{t=1}^{T}(L_{\mathcal{D}}(\mathbf{w}^{(t)}) - L_{\mathcal{D}}(\mathbf{w}^{\star}))\right] \ \leq\ B\,\rho\,\sqrt{T}\,.$$

En divisant par T et en exploitant la convexité de $L_{\mathcal{D}}(\mathbf{w})$, on a finalement

$$\mathop{\mathbb{E}}_{z_1,\ldots,z_T}\left[L_{\mathcal{D}}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{w}^{(t)}\right)\right] \ \leq\ L_{\mathcal{D}}(\mathbf{w}^{\star}) + \frac{B\,\rho}{\sqrt{T}}\,.$$

DGS pour problèmes convexes-Lipschitziens-bornés

Lorsque $\mathbf{v}_t(\mathbf{w}^{(t)}, z_t)$ est un sous-gradient de $\ell(\mathbf{w}^{(t)}, z_t)$ au point $\mathbf{w}^{(t)}$, nous avons

$$\begin{aligned}\mathop{\mathbb{E}}_{z_1,\ldots,z_T}[\langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathbf{v}_t(\mathbf{w}^{(t)}, z_t)\rangle] &= \mathop{\mathbb{E}}_{z_1,\ldots,z_t}[\langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathbf{v}_t(\mathbf{w}^{(t)}, z_t)\rangle] \\ &= \mathop{\mathbb{E}}_{z_1,\ldots,z_{t-1}} \langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathop{\mathbb{E}}_{z_t} \mathbf{v}_t(\mathbf{w}^{(t)}, z_t)\rangle \\ &= \mathop{\mathbb{E}}_{z_1,\ldots,z_T} \langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathop{\mathbb{E}}_{z_t} \mathbf{v}_t(\mathbf{w}^{(t)}, z_t)\rangle \ .\end{aligned}$$

Donc, similairement au cas différentiable, nous avons

$$\mathop{\mathbb{E}}_{z_1,\ldots,z_T}\left[\sum_{t=1}^{T}\langle \mathbf{w}^{(t)} - \mathbf{w}^\star \ , \ \mathop{\mathbb{E}}_{z_t} \mathbf{v}_t(\mathbf{w}^{(t)}, z_t)\rangle\right] \ \leq \ B\,\rho\,\sqrt{T} \ .$$

Par le fait que $\mathbb{E}_{z_t}\, \mathbf{v}_t(\mathbf{w}^{(t)}, z_t)$ est un sous-gradient de $L_{\mathcal{D}}(\mathbf{w}^{(t)})$, la convexité de $\ell(\mathbf{w}, z)$ implique

$$L_{\mathcal{D}}(\mathbf{w}^{(t)}) - L_{\mathcal{D}}(\mathbf{w}^\star) \leq \langle \mathbf{w}^{(t)} - \mathbf{w}^\star \ , \ \mathop{\mathbb{E}}_{z_t} \mathbf{v}_t(\mathbf{w}^{(t)}, z_t)\rangle \ .$$

DGS pour problèmes convexes-Lipschitziens-bornés

Donc, en combinant, nous obtenons également pour le cas non-différentiable que

$$\mathop{\mathbb{E}}_{z_1,\ldots,z_T}\left[\sum_{t=1}^{T}(L_{\mathcal{D}}(\mathbf{w}^{(t)})-L_{\mathcal{D}}(\mathbf{w}^{\star}))\right] \leq B\,\rho\,\sqrt{T}\,.$$

En divisant par T et en exploitant la convexité de $L_{\mathcal{D}}(\mathbf{w})$, on a également pour le cas non-différentiable que

$$\mathop{\mathbb{E}}_{z_1,\ldots,z_T}\left[L_{\mathcal{D}}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{w}^{(t)}\right)\right] \leq L_{\mathcal{D}}(\mathbf{w}^{\star})+\frac{B\,\rho}{\sqrt{T}}\,.$$

Nous avons donc démontré, dans les deux cas, le théorème suivant.

DGS pour problèmes convexes-Lipschitziens-bornés

Théorème (Garantie pour la DGS minimisant $L_{\mathcal{D}}$)

*Considérons un problème d'apprentissage convexe-Lipschitzien-borné avec les paramètres ρ et B. Alors, pour tout $\epsilon > 0$, l'algorithme de descente de (sous-) gradient stochastique pour minimiser $L_{\mathcal{D}}(\mathbf{w})$ avec un nombre T d'itérations (*i.e., *un nombre d'exemples) tel que*

$$T \geq \frac{B^2 \rho^2}{\epsilon^2}$$

avec $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$, retournera un prédicteur $\bar{\mathbf{w}}$ tel que

$$\mathbb{E}\left[L_{\mathcal{D}}(\bar{\mathbf{w}})\right] \leq \min_{\mathbf{w} \in \mathcal{H}} L_{\mathcal{D}}(\mathbf{w}) + \epsilon \ .$$

Notez que la DGS, telle que formulée, ne restreint pas la norme de $\bar{\mathbf{w}}$, mais la garantie sur $\mathbb{E}\left[L_{\mathcal{D}}(\bar{\mathbf{w}})\right]$ est exprimée à l'aide d'une classe comparative $\mathcal{H}$ contenant uniquement des prédicteurs dont la norme est au plus B.

Choisir η et T

- Notez que les valeurs pour η et de T dépendent de B.
- Plus que B est élevé, plus faible sera l'erreur d'approximation $\min_{\mathbf{w}:||\mathbf{w}||\leq B} L_{\mathcal{D}}(\mathbf{w})$.
- Si on choisi $T = B^2\rho^2/\epsilon^2$, on obtient $\eta = \epsilon/\rho^2$.
- Puisque ρ est connu, η est fixé par le choix de l'erreur d'estimation ϵ.
- Lorsque ϵ est choisi, la valeur de T sera déterminée par le choix de B.
- Donc plus que le nombre T d'itérations (et d'exemples) est élevé et plus on contribue à diminuer l'erreur d'approximation.
- On a donc intérêt à choisir ϵ petit (donc η petit) et à choisir T le plus élevé possible.
- Si $T = m$ (le nombre d'exemples), on a alors $B = \frac{\epsilon}{\rho}\sqrt{m}$ et l'erreur d'approximation augmente lorsque ϵ diminue. En choisissant ϵ (donc η) sur un ensemble de validation, on détermine alors le bon compromis entre l'erreur d'approximation et l'erreur d'estimation.

DGS pour problèmes convexes-Lipschitziens-bornés

- Notez que le théorème précédent est formulé en espérance.
- Pour obtenir un résultat formulé en probabilité (et conforme au critère PAC agnostique), il suffit d'utiliser le lemme suivant.

Lemme (Une borne sur l'espérance du risque implique le critère PAC)

Soit une classe $\mathcal{H}$ de fonctions et un algorithme d'apprentissage A. Si pour $m \geq m_{\mathcal{H}}(\epsilon)$, on a

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))] \leq \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \epsilon\,,$$

alors pour $m \geq m_{\mathcal{H}}(\epsilon\delta)$, on a que

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\left(L_{\mathcal{D}}(A(S)) \leq \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \epsilon\right) \geq 1-\delta\,.$$

Inégalité de Markov

Pour démontrer ce lemme, il faudra utiliser l'inégalité de Markov.

Théorème (Inégalité de Markov)

Pour toute variable aléatoire X non négative et d'espérance μ et pour tout $t > 0$,

$$\mathbb{P}\left(X \geq t\mu\right) \leq \frac{1}{t}\,.$$

Preuve pour le cas discret.

$$\mathbb{P}(X \geq t\mu) = \sum_{x \geq t\mu} \mathbb{P}(X = x) \leq \sum_{x \geq t\mu} \frac{x}{t\mu}\,\mathbb{P}(X = x)$$

$$\leq \frac{1}{t\mu} \sum_{x} x\,\mathbb{P}(X = x) = \frac{1}{t}$$

□

Preuve du lemme précédent

Preuve:

- Par hypothèse, pour $m \geq m_{\mathcal{H}}(\epsilon\delta)$, on a

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))] \leq \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \epsilon\delta\,.$$

- Soit la variable non négative $X \stackrel{\text{def}}{=} L_{\mathcal{D}}(A(S)) - \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h)$.
- Soit $\mu \stackrel{\text{def}}{=} \mathbb{E}\, X$. Alors $\mu \leq \epsilon\delta$.
- L'inégalité de Markov implique : $\mathbb{P}(X \geq \frac{\mu}{\delta}) \leq \delta$.
- Puisque $\epsilon \geq \frac{\mu}{\delta}$, on a $\mathbb{P}(X \geq \epsilon) \leq \delta$.
- Alors, $\mathbb{P}(X \leq \epsilon) \geq 1 - \delta$. Ce qui donne le lemme. ■

Remarque : le nombre d'exemples requis $m_{\mathcal{H}}(\epsilon\delta)$ peut sembler excessif, mais l'exercice 13.1 démontre qu'il est possible de satisfaire le critère PAC agnostique en n'utilisant que $O(m_{\mathcal{H}}(\epsilon/2)\log(1/\delta))$ exemples.

Résumé

- Notions : convexité, fonctions Lipschitziennes, et sous-gradients.
- Algorithme de la descente de (sous) gradient avec ses garanties.
- Problèmes d'apprentissage convexes.
- Problèmes d'apprentissage convexes-Lipschitziens-bornés.
- L'algorithme de la descente de (sous) gradient stochastique pour minimiser le vrai risque apprends, au sens PAC agnostique, les problèmes convexes-Lipschitziens-bornés.