Université Paris-Saclay

Stabilité des algorithmes d'apprentissage statistique et erreur de généralisation

Axel Forveille
Arnaud Gardille
Sofiane Dakhmouche

14 juin 2021

Introduction

Qu'est ce qui permet à un algorithme de faire un apprentissage effectif à partir d'un nombre de données restreint ? Il s'avère nécessaire que l'algorithme parvienne à généraliser, afin d'être efficace sur de nouvelles données. Ainsi, établir des bornes sur l'erreur de généralisation permet de démontrer la capacité d'apprentissage de l'algorithme. En plus de cela, une telle borne permet de contrôler l'erreur de l'algorithme avec grande probabilité, en fonction du nombre de données. Mais alors, comment garantir une bonne capacité de généralisation ? Nous démontrons dans ce mémoire qu'il suffit que l'algorithme vérifie une certaine forme de stabilité. C'est à dire qu'un léger changement du jeu de données ne modifie seulement légèrement le résultat de l'algorithme.

Ainsi, le présent travail, qui est principalement basé sur les articles de [Liu et al., 2017], [Wibisono et al., 2009] et [Hardt et al., 2016], est organisé comme suit. Dans une première partie, nous formalisons les notions de *uniform argument stability* et d'*erreur de généralisation*. Puis, nous démontrons que l'hypothèse renvoyée par un algorithme stable appartient à une boule centrée en son espérance avec grande probabilité. En mesurant alors la complexité de Rademacher de cette dernière, nous obtenons une borne probabiliste sur l'erreur de généralisation. Par la suite, nous introduisons l'erreur de généralisation déformée, afin d'affiner notre majoration. Enfin, nous considérons des applications de ce qui précède, en montrant que l'algorithme de minimisation du risque empirique est stable dès lors qu'il est régularisé ou entraîné par *stochastic gradient descent*.

Table des matières

1 Stabilité d'un algorithme d'apprentissage statistique

Dans cette section, nous définissons le cadre dans lequel nous travaillerons dans la suite de l'article, puis nous introduisons deux notions de stabilité d'un algorithme d'apprentissage. Soient $(\mathcal{X}, \mathfrak{X})$ et $(\mathcal{Y}, \mathfrak{Y})$ deux espaces mesurables. Dans un cadre d'apprentissage statistique, on considère $Z = (X, Y)$ un couple de variables aléatoires de loi jointe $\mathbb{P}$ sur l'espace produit que l'on note $(\mathcal{Z}, \mathfrak{Z})$.

Définition 1. (Algorithme d'apprentissage)
On appelle *algorithme d'apprentissage* une application $\mathcal{A}$ qui à un élément de $\mathcal{Z}^n$ associe une fonction mesurable de $\mathcal{X}$ dans $\mathcal{Y}$:

$$\mathcal{A} : \begin{cases} (\mathcal{X} \times \mathcal{Y})^n & \longrightarrow (\mathcal{X} \longrightarrow \mathcal{Y}) \\ S & \longmapsto h_{\mathcal{A},S} \end{cases}$$

On appelle $\mathrm{Im}(\mathcal{A})$ la *classe d'hypothèses* de $\mathcal{A}$ et on note $\mathcal{H} := \mathrm{Im}(\mathcal{A})$. On appelle *hypothèse* un élément de $\mathcal{H}$.

Si le cadre n'est pas ambigu, on omettra de mentionner l'algorithme et on notera simplement h_S les éléments de la *classe d'hypothèse*.

Une hypothèse $h \in \mathcal{H}$ associe à un élément x de l'espace des données une prédiction $h(x)$. On introduit alors la *fonction de perte* afin de quantifier la qualité d'une prédiction.

Définition 2. (Fonction de perte)
Dans le cadre introduit, une *fonction de perte* est une fonction mesurable positive :

$$\ell : \mathcal{H} \times \mathcal{Z} \longrightarrow \mathbb{R}_+$$

.

Définition 3. (Risque)
Étant donné une hypothèse $h \in \mathcal{H}$, on appelle *risque* de h la quantité notée $R(h)$ et définie par :

$$R(h) := \mathbb{E}_{Z \sim \mathbb{P}} \left[\ell(h, Z) \right].$$

L'objectif d'un problème d'apprentissage est de minimiser le risque pour prédire au mieux par rapport à la perte considérée. En pratique, il est souvent impossible de calculer directement le risque mais on peut l'estimer par sa moyenne empirique.

Définition 4. (Risque empirique)
Étant donné une hypothèse $h \in \mathcal{H}$ et un échantillon i.i.d $S = (Z_1, ..., Z_n)$ de loi $\mathbb{P}^{\otimes n}$, on appelle *risque empirique* de h sur l'échantillon S la quantité notée $R_S(h)$ et définie par :

$$R_S(h) := \frac{1}{n} \sum_{i=1}^{n} \ell(h, Z_i).$$

Ainsi, afin de minimiser le risque et si le risque empirique est une bonne approximation du risque, on peut chercher à minimiser le risque empirique directement. On introduit alors l'*erreur de généralisation* qui mesure l'écart entre le risque et le risque empirique.

Définition 5. (Erreur de généralisation)
Etant donné une hypothèse $h \in \mathcal{H}$ et un échantillon i.i.d $S = (Z_1, ..., Z_n)$ de loi $\mathbb{P}^{\otimes n}$, on appelle *erreur de généralisation* de h la quantité :

$$R(h) - R_S(h).$$

On dit qu'un algorithme généralise bien lorsque l'erreur de généralisation est petite. Étant donné un échantillon assez grand généré aléatoirement, on aimerait que le risque empirique de cet échantillon approche au mieux le risque réel et ce quelque soit l'échantillon. C'est la capacité d'un algorithme à améliorer ses performances sur des données qu'il ne connaît pas grâce à un échantillon de ces données, d'où le terme de "généralisation". De ce point de vue, un algorithme généralise bien lorsqu'il n' "overfit" pas.
Nous définissons maintenant des notions de stabilité d'un algorithme d'apprentissage. Qualitativement, un algorithme est stable lorsque une petite modification de l'échantillon à un petit impact sur le résultat de l'algorithme.

Formellement, considérons deux échantillons $S = (Z_1, ..., Z_n)$ et $S' = (Z'_1, ..., Z'_n)$ de loi $\mathbb{P}^{\otimes n}$. Notons, pour tout $i \in \{1, ..., n\}$, $S^i := (Z_1, ..., Z_{i-1}, Z'_i, Z_{i+1}, ..., Z_n)$, l'échantillon *i-altéré* de S.

Définition 6. (Stabilité uniforme)
Un algorithme d'apprentissage $\mathcal{A}$ est $\beta(n)$-*uniformément stable* par rapport à la perte ℓ si pour tout $i \in \{1, ..., n\}$ et tout échantillon S de loi $\mathbb{P}^{\otimes n}$:

$$|\ell(h_S, Z) - \ell(h_{S^i}, Z)| \leq \beta(n) \quad \mathbb{P}\text{-ps.},$$

où β est une fonction positive et décroissante sur $\mathbb{N}$.

Cette notion de stabilité est retenue dans certains ouvrages mais concerne la perte et les hypothèses. Il est intéressant d'introduire une notion de stabilité qui concerne uniquement les hypothèses et donc l'algorithme d'apprentissage. Munissons $\mathcal{H}$ d'une norme afin de mesurer l'écart entre les prédictions. Supposons $(\mathcal{X}, \mathfrak{X})$ munit d'une norme notée $\|.\|_*$ et $(\mathcal{Y}, \mathfrak{Y})$ munit d'une norme notée $\|.\|^*$, on munit alors $\mathcal{H}$ de la norme d'opérateur $\|.\|$ définie par :

$$\forall h \in \mathcal{H}, \ \|h\| = \sup_{x \in \mathcal{X}-0} \frac{\|h(x)\|^*}{\|x\|_*}.$$

Dans la suite on notera toutes les normes $\|.\|$.

Définition 7. (Stabilité uniforme des arguments)
Un algorithme d'apprentissage $\mathcal{A}$ est $\alpha(n)$-*argument uniformément stable* si pour tout $i \in \{1, ..., n\}$:

$$\|h_S - h_{S^i}\| \leq \alpha(n),$$

presque sûrement au sens de la loi jointe $\mathbb{P}^{\otimes(n+1)}$ de (S, Z'_i) où α est une fonction positive sur $\mathbb{N}$.

Définition 8. (Stabilité des arguments)
Un algorithme d'apprentissage $\mathcal{A}$ est $\alpha(n)$-*argument stable* si pour tout $i \in \{1, ..., n\}$:

$$\mathbb{E}[\|h_S - h_{S^i}\| \ |S] \leq \alpha(n) \text{ p.s},$$

presque sûrement au sens de la loi $\mathbb{P}^{\otimes n}$ de S où α est une fonction positive sur $\mathbb{N}$.

L'intérêt de la stabilité des arguments est d'établir pour presque tout échantillon S une stabilité en moyenne où l'aléa est porté sur la variable Z_i'. C'est une notion de stabilité plus faible que la stabilité uniforme des arguments, en effet il est clair que la stabilité uniforme des arguments implique la stabilité des arguments.

Définition 9. (Perte L-admissible)
Une fonction de perte l est dite *L-admissible*, avec $L > 0$ si :

$$\forall (h, h') \in \mathcal{H}, \forall (x, y) \in \mathcal{Z},\ |\ell(h, x, y) - \ell(h', x, y)| \leq L\|h(x) - h'(x)\|.$$

L'utilisation des inégalités portant sur le risque pour déterminer le nombre minimal d'exemples requis (complexité statistique) est formalisé dans la définition ci-dessous en considérant deux paramètres d'approximation, comme suit [Shalev-Shwartz and Ben-David, 2014]. Le paramètre de précision ϵ détermine jusqu'à quel niveau la sortie de l'algorithme peut être loin de la sortie optimale (ceci correspond au "approximately correct"). Le paramètre de confiance δ quant à lui, indique dans quelle mesure, il est probable que la fonction de prédiction sélectionnée vérifie le critère de précision (ce qui correspond au "probably").

Définition 10. (PAC Learnability [1])
Une classe d'hypothèses $\mathcal{H}$ est Probably Approximately Correctly learnable, s'il existe une fonction $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ et un algorithme d'apprentissage $\mathcal{A}$ vérifiant : pour tous ϵ, $\delta \in (0,1)$ et pour toute loi de probabilité $\mathcal{D}$ sur $(\mathcal{X}, \mathcal{Y})$, si l'algorithme est entraîné sur $m \geq m_{\mathcal{H}}(\delta, \epsilon)$ exemples i.i.d. générés par $\mathcal{D}$, alors l'algorithme retourne une fonction de prédiction h telle que :

$$\mathcal{R}(h) \leq \min_{h' \in \mathcal{H}} \mathcal{R}(h') + \epsilon$$

Remarque :
Par exemple, toute classe d'hypothèses finie $\mathcal{H}$ est learnable avec une complexité statistique :

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{\log(|\mathcal{H}|/\delta)}{\epsilon} \right\rceil.$$

Hypothèses pour la suite : on suppose dans la suite du mémoire qu'il existe $L > 0$ et $B > 0$ tels que :

- $\mathcal{X}$ est un espace de Banach séparable.
- $\mathcal{Y}$ est un espace de Banach.
- $\mathcal{H} \subseteq \{h : \mathcal{X} \longrightarrow \mathcal{Y} \mid h \text{ est linéaire continue } \} =: \mathfrak{B}$ (dual topologique de $\mathcal{X}$)
- La perte ℓ est L-admissible.
- $\|X\| \leq B$ presque sûrement.

On va montrer que si le Banach $(\mathfrak{B}, \|.\|)$ vérifie quelques hypothèses, alors si un algorithme est argument stable on peut contrôler l'erreur de généralisation qui lui est associé. C'est l'objet de la section suivante.

1. Cette définition correspond plus précisément à l'Agnostic PAC learnability. Le terme Agnosctic fait référence ici au fait de ne pas se restreindre au cas où $\min_{h \in \mathcal{H}} \mathcal{R}(h) = 0$. Il a été omis pour simplifier la terminologie.

2 Concentration de la prédiction d'un algorithme stable

L'objectif de cette partie est d'établir une inégalité de concentration pour la prédiction d'un algorithme stable autour de sa moyenne. Pour cela on commence par établir une inégalité de concentration pour des suites de variables aléatoires définies comme des différences de martingales. Sous des conditions favorables, il est possible de contrôler les sommes partielles de ces variables aléatoires. Nous commençons par introduire des espaces de Banach particulier qui vérifient une version plus générale de l'identité du parallélogramme. Les variables aléatoires que nous étudierons évolueront dans ces espaces.

2.1 Espaces de Banach et fonctions (2,D)-smooth

Définition 11.
Soit $(\mathfrak{B}, \|.\|)$ un Banach séparable, on dit que $(\mathfrak{B}, \|.\|)$ est $(2, D)$-*smooth* si :

$$\forall h, h' \in \mathfrak{B}, \|h + h'\|^2 + \|h - h'\|^2 \leq 2\|h\|^2 + 2D^2\|h'\|^2$$

En particulier, un espace de Hilbert vérifie l'identité du parallélogramme et est, par conséquent, $(2, 1)$-smooth.

Il existe une condition suffisante sur le différentielle seconde de sa norme pour qu'un Banach soit $(2, D)$-smooth, c'est l'objet de cette proposition. Notons $N := \|.\|^2$.

Proposition 1.
Soit $(\mathfrak{B}, \|.\|)$ un Banach séparable. Si pour tout $x, h \in \mathfrak{B}^2$,

$$d_x^2 N(h, h) \leq 2D^2 N(h).$$

Alors, $(\mathfrak{B}, \|.\|)$ est $(2, D)$-smooth.

Démonstration. En notant $|||.|||$ le norme subordonnée de $\|.\|$, on a que

$$\forall x \in \mathfrak{B}, \ \left|\left|\left|d_x^2 N\right|\right|\right| \leq 2D^2.$$

Par l'inégalité de Taylor Lagrange,

$$\begin{cases} N(h + h') \leq N(h) + d_h N(h') + D^2 N(h'), \\ N(h - h') \leq N(h) - d_h N(h') + D^2 N(h'). \end{cases}$$

D'où le résultat en sommant les deux inégalités. □

Ainsi, il suffit que la norme subordonnée sur $\mathfrak{B}$ de la différentielle seconde de N soit majorée par une constante M pour dire qu'un espace de Banach est $\left(2, \sqrt{\frac{M}{2}}\right)$-smooth. Dans la suite on introduit les fonctions $(2, D)$-smooth pour montrer qu'en fait cette condition caractérise tous les espaces de Banach $(2, D)$-smooth.

Définition 12.
Soit $(\mathfrak{B}, \|.\|)$ un Banach séparable, une fonction $\psi : \mathfrak{B} \longrightarrow [0, \infty[$ est $(2, D)$-smooth si :

1. $\psi(0) = 0$,

2. $\forall x, v \in \mathfrak{B}, |\psi(x+v) - \psi(x)| \leq \|v\|$,
3. $\forall x, v \in \mathfrak{B}, \psi^2(x+v) - 2\psi^2(x) + \psi^2(x-v) \leq 2D^2\|v\|^2$.

Proposition 2.
Un espace de Banach séparable est $(2, D)$-smooth si et seulement si sa norme est $(2, D)$-smooth.

Démonstration. Une norme vérifie toujours les points 1 et 2, le point 3 est la définition des espaces $(2, D)$-smooth. □

2.2 Inégalité de concentration pour des différences de martingales

Dans cette section, on considère $(\mathfrak{B}, \|.\|)$ un Banach séparable $(2, D)$-smooth et $(\Omega, A, \mathbb{P})$ un espace probabilisé. Soit $\mathbb{M} = (M_n)_{n\geq 0}$ une martingale par rapport à sa filtration naturelle $\mathbb{F} = (F_n)_{n\geq 0}$ définie sur $(\Omega, A, \mathbb{P})$ et à valeur dans $\mathfrak{B}$ où $\mathbb{F}$ est définie par $F_n = \sigma(M_0, ..., M_n)$, la plus petite tribue engendrée par la famille $(M_0, ..., M_n)$, pour tout $n \geq 0$.
On définit le processus $\mathbb{D} = (D_n)_{n\geq 0}$ par :

$$\begin{cases} D_0 = M_0 \\ D_n = M_n - M_{n-1} \text{ pour tout } n \geq 1 \end{cases}$$

De sorte que pour tout $n \geq 0$,

$$\sum_{i=0}^{n} D_i = M_n.$$

Par construction,

— le processus $\mathbb{D}$ est adapté à la filtration $\mathbb{F}$,

— pour tout $n \geq 0$, $\mathbb{E}[M_n] < \infty$,

— pour tout $n \geq 1$, $\mathbb{E}[M_n|F_{n-1}] = 0$.

Notons $\|.\|_\infty$, la norme sur $\mathfrak{B}$ définie par $\|f\|_\infty = \sup_{x\in D_f} \|f(x)\|^*$ pour toute fonction $f : D_f \to \mathfrak{B}$.

Le théorème suivant établit une inégalité de concentration pour les sommes partielles de différences de martingales lorsque les variables aléatoires évoluent dans un espace de Banach séparable $(2, D)$-smooth.

Proposition 3. (Inégalité de concentration pour les différences de martingales)[Pinelis, 1994]
Avec les notations introduites et en notant $D_*^2 := \max(1, D^2)$, si il existe $C^2 > 0$ tel que :

$$\sum_{i\geq 0} \|D_i\|_\infty^2 \leq C^2.$$

Alors, pour tout $r \geq 0$,

$$\mathbb{P}\left(\sup_{n\geq 1} \left\|\sum_{i=0}^{n} D_i\right\| \geq r\right) \leq 2\exp\left(-\frac{r^2}{2D_*^2C^2}\right).$$

Démonstration. Remarquons premièrement que $\sup_{n\geq 1}\|\sum_{i=0}^{n} D_i\| = \sup_{n\geq 1}\|M_n\|$ et notons cette quantité M^*. On va montrer que $\mathbb{P}(M^* \geq r) \leq 2\exp\left(-\frac{r^2}{2D^2C^2}\right)$.
Soient $x, v \in \mathfrak{B}$, posons $u(t) := \|x + tv\|$ pour tout $t \in \mathbb{R}$. Par le caractère $(2, D)$-smooth de $\|.\|$, pour tout $t \in \mathbb{R}$:

$$\begin{cases} u'(t) \leq \|v\|, \\ (u^2)''(t) \leq 2D^2\|v\|^2. \end{cases}$$

Notons, ch et sh les fonctions cosinus et sinus hyperbolique définies par $\mathrm{ch}(u) = \frac{\exp(u)+\exp(-u)}{2}$ et $\mathrm{sh}(u) = \frac{\exp(u)-\exp(-u)}{2}$ pour tout $u \in \mathbb{R}$.
Sur l'ensemble $(u''u > 0)$ on a la majoration suivante :

$$\begin{aligned} (\mathrm{ch}u)'' &= (u')^2\mathrm{ch}u + u''\mathrm{sh}u \\ &\leq (u')^2\mathrm{ch}u + u''u\mathrm{ch}u \\ &= \frac{1}{2}(u^2)''\mathrm{ch}u \\ &\leq D^2\|v\|^2\mathrm{ch}u. \end{aligned}$$

Où on a utilisé le lemme suivant.

Lemme 1.
Soient $a, b \in \mathbb{R}$ tels que $ab > 0$, alors $a\mathrm{sh}(b) \leq ab\mathrm{ch}(b)$.

Démonstration. Comme $ab > 0$, alors, en notant $s(x) = 1$ si $x \geq 0$, -1 sinon, on a $s(a) = s(b)$. Donc $ab\mathrm{ch}(b) - a\mathrm{sh}(b) = |a||b|\mathrm{ch}|b| - |a|\mathrm{sh}|b|$ par symétrie de ch et antisymétrie de sh. Supposons alors $a \geq 0$ et $b \geq 0$. Il suffit donc de montrer que $\psi(b) := b\mathrm{ch}(b) - \mathrm{sh}(b) \geq 0$.
Or $\psi'(b) = b\mathrm{sh}(b) \geq 0$, donc ψ est croissante sur $\mathbb{R}_+^*$ et $\lim_{u\longrightarrow 0}\psi(u) = 0$, d'où le résultat. □

Sur l'ensemble $(u''u \leq 0)$ on a la majoration suivante :

$$\begin{aligned} (\mathrm{ch}u)'' &= (u')^2\mathrm{ch}u + u''\mathrm{sh}u \\ &\leq (u')^2\mathrm{ch}u \text{ car } u''\mathrm{sh}u \leq 0, \\ &\leq \|v\|^2\mathrm{ch}u. \end{aligned}$$

Donc pour tout $t \in \mathbb{R}$,

$$(\mathrm{ch}u(t))'' \leq D_*^2\|v\|^2\mathrm{ch}u(t). \tag{1}$$

Soient $\lambda \in \mathbb{R}_+$ et $j \geq 1$. Introduisons la fonction ϕ définie sur $[-1, 1]$ par

$$\phi(t) := \mathbb{E}\left[\mathrm{ch}(\lambda\|M_{j-1} + tD_j\|)|F_{j-1}\right] = \mathbb{E}\left[\mathrm{ch}(u(t))|F_{j-1}\right],$$

où $u(t) = \|\lambda M_{j-1} + t\lambda D_j\|$.
ϕ est deux fois dérivable de dérivées en $t \in [-1, 1]$:

$$\phi'(t) = \mathbb{E}\left[\mathrm{ch}(u(t))'|F_{j-1}\right] = \lambda\mathbb{E}\left[\mathrm{sh}(u(t))D_j d_{M_{j-1}+tD_j}\|.\|(1)|F_{j-1}\right],$$

$$\phi''(t) = \mathbb{E}\left[\mathrm{ch}(u(t))''|F_{j-1}\right].$$

Donc $\phi'(0) = \lambda\mathbb{E}\left[\mathrm{sh}(M_{j-1})D_j d_{M_{j-1}}\|.\|(1)|F_{j-1}\right] = \lambda\mathrm{sh}(M_{j-1})d_{M_{j-1}}\|.\|(1)\mathbb{E}\left[D_j|F_{j-1}\right] = 0$.
Et par (1), pour tout $t \in [-1, 1]$,

$$\begin{aligned}\phi''(t) &\leq D_*^2\mathbb{E}[\|\lambda D_j\|^2\mathrm{ch}u(t)|F_{j-1}] \\ &\leq \lambda^2 D_*^2\|D_j\|_\infty^2\mathbb{E}[\mathrm{ch}u(t)|F_{j-1}] \\ &= \lambda^2 D_*^2\|D_j\|_\infty^2\phi(t).\end{aligned}$$

Lemme 2.
Soit $f : \mathbb{R} \longrightarrow [0, \infty[$ de classe C^2 telle que $f'(0) = 0$ et $f'' \leq R^2 f$ où $R \in \mathbb{R}$, alors

$$\forall t \in \mathbb{R}, f(t) \leq f(0)\mathrm{ch}(Rt) \leq f(0)\exp\left(\frac{(Rt)^2}{2}\right).$$

Démonstration. Quitte à changer d'échelle, on peut supposer que $R = 1$ et $f(0) = 1$. On veut résoudre le problème :

$$\begin{cases} f'' = f + g, \\ f(0) = 1, \\ f'(0) = 0. \end{cases}$$

$f(t) = ch(t) + \int_0^t g(s)sh(t-s)\mathrm{d}s$ est solution et par $f'' \leq f$, $g \leq 0$. Donc, par positivité de l'intégrale, $f(t) \leq \mathrm{ch}(t) \leq \exp\left(\frac{t^2}{2}\right)$ où la dernière inégalité s'obtient par un développement de Taylor du cosinus hyperbolique. □

Donc, puisque $\phi'(0) = 0$ et $\phi'' \leq \lambda^2 D_*^2\|D_j\|_\infty\phi(t)$, par le lemme 2, et en particulier au point 1 :

$$\phi(1) \leq \exp\left(\frac{\lambda^2 D_*^2\|D_j\|_\infty^2}{2}\right)\phi(0).$$

Soit en remplaçant par les valeurs de ϕ,

$$\mathbb{E}[\mathrm{ch}(\lambda\|M_j\|)|F_{j-1}] \leq \exp\left(\frac{\lambda^2 D_*^2\|D_j\|_\infty^2}{2}\right)\mathrm{ch}(\lambda\|M_{j-1}\|). \tag{2}$$

Cette inégalité suggère de construire la surmartingale suivante $\mathbb{G}$ définie pour $n \geq 0$ par :

$$G_n = \exp\left(-\frac{\lambda^2 D_*^2 s_n^2}{2}\right)\mathrm{ch}(\lambda\|M_n\|),$$

où $s_n = \sqrt{\sum_{i=0}^n \|D_i\|_\infty^2}$.
On vérifie en effet avec l'équation 2 que

$$\begin{aligned}\mathbb{E}[G_n|F_{n-1}] &= \exp\left(-\frac{\lambda^2 D_*^2 s_n^2}{2}\right)\mathbb{E}[\mathrm{ch}(\lambda\|M_n\|)|F_{n-1}] \\ &\leq \exp\left(-\frac{\lambda^2 D_*^2 s_n^2}{2} + \frac{\lambda^2 D_*^2\|D_n\|_\infty^2}{2}\right)\mathrm{ch}(\lambda\|M_{n-1}\|) \\ &= G_{n-1}\end{aligned}$$

et $\mathbb{E}[G_0] = 1$.
Notons $\tau_r := \inf\{i \in \mathbb{N}|\|M_i\| \geq r\}$, τ_r est un temps d'arrêt de la filtration $\mathbb{F}$ car $[r, \infty[$ est un

borélien de $\mathbb{R}$. Ainsi, puisque $\mathbb{G}$ est une surmartingale, $(G_{n\wedge\tau_r})_n$ est une surmartingale. En particulier, si τ_r est fini presque sûrement, alors

$$\mathbb{E}[G_{\tau_r}] \leq \mathbb{E}[G_0] = 1.$$

$(G_{n\wedge\tau_r})_n$ est une surmatingale positive donc converge presque sûrement vers G_∞ qui vérifie $\mathbb{E}[G_\infty] \leq \mathbb{E}[G_0] = 1$, ainsi, que τ soit fini presque sûrement où non, on a dans tous les cas :

$$\mathbb{E}[G_{\tau_r}] \leq \mathbb{E}[G_0] = 1.$$

Finalement, par croissance du cosinus hyperbolique, inégalité de Markov et l'inégalité $\mathrm{ch}u > \frac{e^u}{2}$, pour tout $\lambda \geq 0$

$$\begin{aligned}
\mathbb{P}(M^* \geq r) &\leq \mathbb{P}\left(G_{\tau_r} \geq \exp\left(-\frac{\lambda^2 D_*^2 s_{\tau_r}^2}{2}\right)\mathrm{ch}(\lambda r)\right) \\
&\leq \frac{\exp\left(\frac{\lambda^2 D_*^2 s_{\tau_r}^2}{2}\right)}{\mathrm{ch}(\lambda r)}\mathbb{E}[G_{\tau_r}] \\
&\leq 2\exp\left(-\lambda r + \frac{\lambda^2 D_*^2 C^2}{2}\right) \\
&\leq 2\exp\left(-\frac{r^2}{D_*^2 C^2}\right).
\end{aligned}$$

Où la dernière inégalité est obtenue en minimisant $f : \lambda \to -\lambda r + \frac{\lambda^2 D_*^2 C^2}{2}$ sur $\mathbb{R}_+$.

□

2.3 Inégalité de concentration pour la prédiction d'un algorithme argument stable

Comme énoncé plus tôt, on va utiliser l'inégalité de concentration des martingales pour contrôler l'écart entre la prédiction d'un algorithme stable et sa moyenne.

Lemme 3. (Concentration de la prédiction)
Dans le cadre de la section 1, si de plus $(\mathfrak{B}, \|.\|)$ est $(2, D)$-smooth et soit $\mathcal{A}$ un algorithme $\alpha(n)$-argument stable. Alors pour tout échantillon S et pour tout $\delta > 0$, avec probabilité au moins $1 - \delta$:

$$\|h_S - \mathbb{E}[h_S]\| \leq D_*\alpha(n)\sqrt{2Bn\log(2\delta^{-1})}.$$

Démonstration. Soient $(Z_t)_{t\geq 1}$ une suite de variables aléatoire i.i.d de loi $\mathbb{P}$, $n \in \mathbb{N}$ et $t \geq 2$, notons $S = (Z_1, ..., Z_n)$ et :

$$D_t = \mathbb{E}[h_S|Z_1, ..., Z_t] - \mathbb{E}[h_S|Z_1, ..., Z_{t-1}] \text{ et } D_1 = \mathbb{E}[h_S].$$

On peut construire tout échantillon de Z^n pour toute valeur de n de cette façon.

Alors $(D_t)_{t\geq 1}$ est une suite de différences de martingales qui vérifie :

$$h_S - \mathbb{E}[h_S] = \sum_{t=1}^{n} D_t$$

car $\mathbb{E}[h_S|Z_1, ..., Z_n] = \mathbb{E}[h_S|S] = h_S$.
Or on a :

$$\begin{aligned}
\sum_{t\geq 1} \|D_t\|_\infty^2 &= \sum_{t=1}^{n} \|\mathbb{E}[h_S|Z_1, ..., Z_t] - \mathbb{E}[h_S|Z_1, ..., Z_{t-1}]\|_\infty^2 \text{ car } D_t = 0 \text{ pour tout } t \geq n+1 \\
&= \sum_{t=1}^{n} \|\mathbb{E}[h_S - h_{S^t}|Z_1, ..., Z_t]\|_\infty^2 \text{ car } \mathbb{E}[h_S|Z_1, ..., Z_{t-1}] = \mathbb{E}[h_{S^t}|Z_1, ..., Z_t] \\
&\leq \sum_{t=1}^{n} \mathbb{E}[\|h_S - h_{S^t}\|_\infty|Z_1, ..., Z_t]^2 \\
&= \sum_{t=1}^{n} \mathbb{E}[\mathbb{E}[\|h_S - h_{S^t}\|_\infty|S]|Z_1, ..., Z_t]^2 \\
&\leq Bn\alpha(n)^2.
\end{aligned}$$

Car $\|h_S - h_{S^t}\|_\infty \leq B\|h_S - h_{S^t}\|$. Donc, par la proposition 3, pour tout $r > 0$:

$$\mathbb{P}\left(\|h_S - \mathbb{E}[h_S]\| \leq r\right) \geq \mathbb{P}\left(\sup_{k\geq 1} \left\|\sum_{t=1}^{k} D_t\right\| \leq r\right) \geq 1 - 2\exp\left(-\frac{r^2}{2D_*^2 Bn\alpha(n)^2}\right).$$

On conclut en inversant la borne, c'est-à-dire en prenant $\delta = 2\exp(-\frac{r^2}{2D_*^2 Bn\alpha(n)^2})$. □

3 Un contrôle de l'erreur de généralisation pour les algorithmes stables

3.1 Erreur de généralisation classique

Le résultat de concentration du lemme 3 motive la définition suivante de "classe algorithmique d'hypothèses" : puisque, avec grande probabilité, h_S est concentrée autour de son espérance $\mathbb{E}h_S$, c'est la complexité de la boule centrée en $\mathbb{E}h_S$ qui compte, *et non pas celle de la classe d'hypothèses entière* $\mathcal{H}$. Cette remarque peut amener à une amélioration significative des garanties de performance.

Définition 13. (Classe algorithmique d'hypothèses)
Pour une taille d'échantillon n et un paramètre de confiance $\delta > 0$, on pose $r = r(n, \delta) = D\alpha(n)\sqrt{2n\log(2/\delta)}$. On définit la classe algorithmique d'hypothèses d'un algorithme d'apprentissage stable par :

$$B_r = \{h \in \mathcal{H} \,|\quad \|h - \mathbb{E}h_S\| \leq r(n, \delta)\}\,.$$

Notons que par le lemme 3, $h_S \in B_r$ avec probabilité au moins $1 - \delta$.
À présent, on majore le risque de généralisation en fonction de la complexité de Rademacher de la classe algorithmique d'hypothèses [Bartlett and Mendelson, 2002]. La complexité de Rademacher d'une classe d'hypothèses $\mathcal{H}$ sur un espace de features $\mathcal{X}$ est définie par :

$$\mathcal{R}(\mathcal{H}) = \mathbb{E}\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \langle h,\, X_i\rangle,$$

où, $\sigma_1, \cdots, \sigma_n$ sont des variables aléatoires de Rademacher, c'est-à-dire i.i.d. uniformément distribuées sur $\{-1, +1\}$.
Le théorème suivant fournit une borne sur la complexité de Rademacher de la classe algorithmique d'hypothèses. Cette borne dépend du *type* de l'espace des features $\mathcal{X}$. Rappelons que l'espace de Banach $(\mathcal{X}, \|.\|)$ est de type $p \geq 1$, s'il existe une constante $C_p > 0$ telle que pour tous $x_1, \cdots, x_n \in \mathcal{X}$,

$$\mathbb{E}\left\|\sum_{i=1}^{n} \sigma_i x_i\right\| \leq C_p \left(\sum_{i=1}^{n} \|x_i\|^p\right)^{1/p}.$$

Dans le cas particulier important où $\mathcal{X}$ est un espace de Hilbert, il est de type 2 avec une constante $C_2 = 1$.

Théorème 1.
Supposons que $\mathcal{X}$ soit un espace de Banach séparable type p et que son dual soit $(2, D)$-smooth. Supposons, de plus, que la loi de X_i soit telle que $\|X_i\| \leq B$ avec probabilité 1, pour une certaine constante $B > 0$. Si un algorithme d'apprentissage statistique est $\alpha(n)$-argument stable, alors la complexité de Rademacher de la classe algorithmique d'hypothèses B_r vérifie :

$$\mathcal{R}(B_r) \leq DC_pB\sqrt{2\log(2/\delta)}\alpha(n)n^{-1/2+1/p}.$$

En particulier, si $\mathcal{B}$ est un espace de Hilbert la borne se simplifie :

$$\mathcal{R}(B_r) \leq B\sqrt{2\log(2/\delta)}\alpha(n).$$

Démonstration.
On a :

$$\begin{aligned}\mathcal{R}(B_r) &= \mathbb{E}\sup_{h\in B_r}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\langle h, X_i\rangle \\ &= \mathbb{E}\sup_{h\in B_r}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\langle h, X_i\rangle - \sigma_i\mathbb{E}[\langle h_{S^i}, X_i\rangle|X_i] + \sigma_i\mathbb{E}[\langle h_{S^i}, X_i\rangle|X_i]\end{aligned}$$

où S^i est l'échantillon S où l'on a remplacé X_i par une variable aléatoire X_i' indépendante de S et de même loi que X_i, ainsi (puisque $\mathbb{P}(\sigma_i = \pm 1) = 1/2$),

$$\begin{aligned}\mathcal{R}(B_r) &= \mathbb{E}\sup_{h\in B_r}\frac{1}{n}\sum_{i=1}^{n}\sigma_i(\langle h, X_i\rangle - \mathbb{E}[\langle h_{S^i}, X_i\rangle|X_i]) \\ &\quad \text{d'où par indépendance entre } S^i \text{ et } X_i, \\ &= \mathbb{E}\sup_{h\in B_r}\frac{1}{n}\sum_{i=1}^{n}\sigma_i(\langle h, X_i\rangle - \langle \mathbb{E}h_{S^i}, X_i\rangle) \\ &= \mathbb{E}\sup_{h\in B_r}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\langle h - \mathbb{E}h_{S^i}, X_i\rangle \\ &\quad \text{d'où, puisque } \mathbb{E}h_{S^i} = \mathbb{E}h_S, \\ &\leq \mathbb{E}\sup_{h\in B_r}\frac{1}{n}\|h - h_S\|.\left\|\sum_{i=1}^{n}\sigma_i X_i\right\| \\ &\leq \frac{r}{n}\mathbb{E}\left\|\sum_{i=1}^{n}\sigma_i X_i\right\| \\ &\leq \frac{1}{n}\alpha(n)D\sqrt{2n\log(2/\delta)}C_p\left(\sum_{i=1}^{n}\|X_i\|^p\right)^{1/p} \\ &\leq DC_pB\sqrt{2\log(2/\delta)}\alpha(n)n^{-1/2+1/p},\end{aligned}$$

d'où le résultat. □

Ce théorème permet d'obtenir facilement une borne sur la performance d'un algorithme $\alpha(n)$-argument stable. Ce que donne le corollaire suivant :

Corollaire 1.
Avec les hypothèses du théorème 1, en supposant de plus que la fonction de perte ℓ soit bornée et L-admissible, i.e. $\ell(h, Z) \leq M$ avec probabilité 1, pour une certaine constante $M > 0$ et $|\ell(h, z) - \ell(h', z)| \leq L|\langle h, x\rangle - \langle h', x\rangle|$ pour tous $z \in \mathcal{Z}$ et $h, h' \in H$. Si un algorithme d'apprentissage statistique est $\alpha(n)$-argument uniformément stable, alors son erreur

de généralisation est bornée comme suit. Avec probabilité au moins $1-2\delta$,

$$R(h_S) - R_S(h_S) \leq 2LDC_pB\sqrt{2\log(2/\delta)}\alpha(n)n^{-1/2+1/p} + M\sqrt{\frac{\log(1/\delta)}{2n}}.$$

En particulier, si $\mathcal{X}$ est un espace de Hilbert, la borne se simplifie :

$$R(h_S) - R_S(h_S) \leq 2L\sqrt{2\log(2/\delta)}\alpha(n) + M\sqrt{\frac{\log(1/\delta)}{2n}}$$

Démonstration.
On note, tout d'abord, que d'après le lemme 1, avec probabilité au moins $1-\delta$,

$$R(h_S) - R_S(h_S) \leq \sup_{h\in B_r}(R(h) - R_S(h)).$$

d'autre part, par la bornétude de la fonction de perte, et par l'inégalité des différences majorées [2], avec probabilité au moins $1-\delta$,

$$\begin{aligned}\sup_{h\in B_r}(R(h) - R_S(h)) &\leq \mathbb{E}\sup_{h\in B_r}(R(h) - R_S(h)) + M\sqrt{\frac{\log(1/\delta)}{2n}} \\ &\leq 2\mathcal{R}(\ell\circ B_r) + M\sqrt{\frac{\log(1/\delta)}{2n}},\end{aligned}$$

où $\ell \circ H$ représente l'ensemble des compositions de fonctions entre ℓ et $h \in \mathcal{H}$. Par la propriété d'admissibilité de ℓ, et par un argument de contraction standard [3], on a,

$$\begin{aligned}\mathcal{R}(\ell\circ B_r) &\leq L.\mathcal{R}(B_r) \\ &\leq LDC_pB\sqrt{2\log(2/\delta)}\alpha(n)n^{-1/2+1/p}.\end{aligned}$$

d'où le résultat. □

3.2 Erreur de généralisation déformée

Lorsqu'il est raisonnable de s'attendre à un risque petit, des bornes sur le risque par excès sont utiles. C'est ce qui est obtenu au théorème 3 ci-dessous, dont la preuve repose sur le lemme et la partie technique -dont le résultat qui nous intéresse est donné par le corollaire 4- suivants.

Notation
Pour $Z = (X, Y)$ vecteur aléatoire à valeurs dans $\mathcal{X} \times \mathcal{Y}$, posons :

$$\mathcal{G}_r(Z) := \left\{\frac{r}{\max\{r, \mathbb{E}\ell(h, Z)\}}, h \in B_r\right\}$$

Lemme 4. [Bartlett et al., 2005]
En posant $V_r := \sup_{g\in\mathcal{G}_r}(\mathbb{E}g(Z) - \frac{1}{n}\sum_{i=1}^n g(Z_i))$, on a, pour tous $r > 0$ et $a > 1$, si $V_r \leq \frac{r}{a}$ alors

$$\forall h \in B_r, \quad \mathbb{E}\ell(h, Z) \leq \frac{a}{a-1}\frac{1}{n}\sum_{i=1}^n \ell(h, Z_i) + \frac{r}{a}.$$

2. Qui relie le risque empirique au risque de généralisation.
3. Talagrand Contraction Lemma (Ledoux & Talagrand, 2013)

Démonstration. Notons g la fonction $(x, y) \longmapsto r\ell(h, (x, y))$. On a, par définition de V_r,

$$\mathcal{E}g(Z) \leq \frac{1}{n}\sum_{i=1}^{n} g(Z_i) + V_r$$

autrement dit,

$$\frac{r}{max\left\{r, \mathbb{E}\ell(h, Z)\right\}}\mathbb{E}\ell(h, Z) \leq \frac{r}{max\left\{r, \mathbb{E}\ell(h, Z)\right\}}\frac{1}{n}\sum_{i=1}^{n}\ell(h, Z_i) + V_r \tag{3}$$

ainsi, si $\max\left\{r, \mathbb{E}\ell(h, Z)\right\} = r$ alors on a l'inégalité :

$$\begin{aligned}\mathbb{E}\ell(h, Z) &\leq \frac{1}{n}\sum_{i=1}^{n}\ell(h, Z_i) + \frac{r}{a} \\ &\leq \frac{a}{a-1}\frac{1}{n}\sum_{i=1}^{n}\ell(h, Z_i) + \frac{r}{a},\end{aligned}$$

et si $\max\left\{r, \mathbb{E}\ell(h, Z)\right\} = \mathbb{E}\ell(h, Z)$, alors l'inégalité (3) se réécrit :

$$\begin{aligned}r\mathbb{E}\ell(h, Z) &\leq r\frac{1}{n}\sum_{i=1}^{n}\ell(h, Z_i) + \mathbb{E}\ell(h, Z)V_r \\ &\leq r\frac{1}{n}\sum_{i=1}^{n}\ell(h, Z_i + \mathbb{E}\ell(h, Z)\frac{r}{a}\end{aligned}$$

d'où

$$a\mathbb{E}\ell(h, Z) \leq \frac{a}{n}\sum_{i=1}^{n}\ell(h, Z_i) + \mathbb{E}\ell(h, Z)$$

d'où

$$\begin{aligned}\mathbb{E}\ell(h, Z) &\leq \frac{a}{a-1}\frac{1}{n}\sum_{i=1}^{n}\ell(h, Z_i) \\ &\leq \frac{a}{a-1}\frac{1}{n}\sum_{i=1}^{n}\ell(h, Z_i) + \frac{r}{a}.\end{aligned}$$

□

3.2.1 Théorème de concentration

Établissons maintenant une proposition qui donne, avec grande probabilité, une majoration uniforme sur l'erreur de généralisation d'une famille de fonctions. C'est l'objet du Corollaire 4 qui suit. Ce résultat est une conséquence d'un théorème plus général qui nous permet de contrôler l'écart entre une variable aléatoire et sa moyenne sous de bonnes hypothèses.

Soit $(\Omega, A, \mathbb{P})$ un espace probabilisé où Ω est séparable. Soient $(X_1, \dots, X_n)$ n variables aléatoires i.i.d de loi $\mathbb{P}$ et à valeurs dans $(\mathfrak{X}, B)$ où $\mathfrak{X}$ est séparable. Soit $F : \mathfrak{X}^n \to \mathbb{R}$ mesurable. Nous allons étudier la concentration de $Z := F(X_1, \dots, X_n)$ par rapport à sa moyenne. Pour cela on introduit les notations suivantes :

— $\mathcal{A} := \sigma(X_1, \dots, X_n)$ la tribu engendrée par la famille $(X_1, \dots, X_n)$,
— $\forall k \in \{1, \dots, n\}, \mathcal{A}_k := \sigma(X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_n)$,
— $\forall k \in \{1, \dots, n\}, \mathbb{E}_k$ l'espérance par rapport à $\mathbb{P}$ conditionnellement à $\mathcal{A}_k$,
— $\mathrm{h} : x \in]-1, \infty[\mapsto (1+x)\log(1+x) - x$,
— $\psi : x \in \mathbb{R} \mapsto \mathrm{e}^{-x} - 1 + x$,
— $\phi : x \in \mathbb{R} \mapsto 1 - (1+x)\mathrm{e}^{-x}$.

Théorème 2. (Théorème de concentration)[Bousquet, 2001]
Dans le cadre introduit, considérons $(Z'_1, \dots, Z'_n)$ des variable aléatoires $\mathcal{A}$-mesurable et $(Z_k)_{1 \le k \le n}$ des variables aléatoires respectivement $\mathcal{A}_k$-mesurables définies par :

$$Z_k = F(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$$

. Supposons que pour tout $k \in \{1, \dots, n\}$,

$$\begin{cases} Z'_k \le Z - Z_k \le 1 \text{ p.s.} \\ \sum_{k=1}^n Z - Z_k \le Z \text{ p.s.} \\ \mathbb{E}_k[Z'_k] \ge 0 \text{ p.s.} \\ Z'_k \le u \text{ p.s.} \\ \frac{1}{n}\sum_{k=1}^n \mathbb{E}_k[(Z'_k)^2] \le \sigma^2 \text{ p.s.} \end{cases}$$

où u et σ^2 sont des constantes strictement positives.
Alors, en notant $v := n\sigma^2 + (1+u)\mathbb{E}[Z]$, on a pour tout $\lambda \ge 0$:

$$\log \mathbb{E}\left[\exp(\lambda(Z - \mathbb{E}[Z]))\right] \le v\psi(-\lambda).$$

Démonstration. La démonstration de ce théorème fait appel à la notion probabiliste d'entropie et à de nombreux résultats la concernant, elle se trouve en annexe. □

De cette majoration sur la log-transformée de Laplace de Z, on en déduit une inégalité de concentration sur Z autour de sa moyenne par la méthode de Chernov.

Corollaire 2.
Dans le cadre du théorème précédent, on a pour tout $t \ge 0$:

$$\mathbb{P}(Z \ge \mathbb{E}[Z] + t) \le \exp\left(-v\mathrm{h}\left(\frac{t}{v}\right)\right),$$

$$\mathbb{P}\left(Z \ge \mathbb{E}[Z] + \sqrt{2vt} + \frac{t}{3}\right) \le \exp(-t).$$

Démonstration. En notant $F_Z(\lambda) := \mathbb{E}[\exp(\lambda Z)]$ la transformée de Laplace de Z, $\Psi_Z(\lambda) := \log(F_Z(\lambda))$ et $\Psi_Z^*(t) := \sup_{\lambda\in\mathbb{R}}(\lambda t - \Psi_Z(\lambda))$ alors pour tout $t \geq 0$:

$$\mathbb{P}(Z \geq \mathbb{E}[Z] + t) \leq \exp -\Psi_Z^*(t + \mathbb{E}[Z]).$$

Or $\sup_{\lambda\in\mathbb{R}}(\lambda t - \Psi_Z(\lambda)) \geq \sup_{\lambda\geq 0}(\lambda t - \Psi_Z(\lambda))$,
et $\lambda\mathbb{E}[Z] - \Psi_Z(\lambda) = -\log\mathbb{E}\left[\exp(\lambda(Z - \mathbb{E}[\mathbb{Z}]))\right] \geq -v\psi(-\lambda)$.
Donc $\sup_{\lambda\in\mathbb{R}}(\lambda t - \Psi_Z(\lambda)) \geq \sup_{\lambda\geq 0}(\lambda t - v\psi(-\lambda))$.
Donc,

$$\mathbb{P}(Z \geq \mathbb{E}[Z] + t) \leq \exp -\sup_{\lambda\geq 0}(\lambda t - v\psi(-\lambda)).$$

Notons $g : \lambda \in \mathbb{R}_+ \mapsto \lambda t - v\psi(-\lambda)$, g est de classe C^1 et concave comme somme de deux fonctions concaves parce que ψ est convexe sur $\mathbb{R}$. Et puisque $\lim_{\lambda\to\infty}\psi(-\lambda) = \infty$, les points critiques de g sont ses maxima.

Soit $\lambda \geq 0$, $g'(\lambda) = t - v(\exp(\lambda) - 1)$ donc l'unique point critique de g est $\lambda_* = \log(1 + t/v)$.
Puis $g(\lambda_*) = v\mathrm{h}(\frac{t}{v})$ est le maximum de g sur $\mathbb{R}_+$ donc

$$\mathbb{P}(Z \geq \mathbb{E}[Z] + t) \leq \exp\left(-v\mathrm{h}\left(\frac{t}{v}\right)\right).$$

La seconde inégalité s'obtient en résolvant l'équation $x = v\mathrm{h}\left(\frac{t}{v}\right)$ d'inconnue t. □

Ce corollaire nous dit que, étant donnée une variable aléatoire Z, si on parvient à construire des variables aléatoires $(Z'_k)_k$ et $(Z_k)_k$ qui vérifient les conditions du théorème 1, alors on peut contrôler l'écart entre Z et sa moyenne. Ce théorème couvre de nombreux cas, en particulier un important dans le cadre de l'apprentissage statistique que nous étudions dans la section suivante.

3.2.2 Application : majoration uniforme de l'erreur de généralisation d'une classe de fonctions

Le théorème 1 donne une inégalité de concentration pour une famille de variable aléatoires, en particulier si $\mathcal{F}$ est une famille de fonctions de $\mathfrak{X} \to \mathbb{R}$, il permet d'établir une inégalité de concentration pour la variable aléatoire :

$$Z = \sup_{f\in\mathcal{F}} \sum_{k=1}^{n} f(X_k),$$

si la classe $\mathcal{F}$ vérifie quelques conditions.

Corollaire 3.
Soit $\mathcal{F}$ une classe de fonctions de $\mathfrak{X} \to \mathbb{R}$ telle que pour tout $f \in \mathcal{F}$:

$$\begin{cases} (\mathcal{F}, \|.\|_\infty) \text{ est séparable} \\ \forall i \in \{1, \ldots, n\}, \mathbb{E}[f(X_i)] = 0 \\ \forall i \in \{1, \ldots, n\}, \mathbb{V}(f(X_i)) \leq \sigma^2 \\ \|f\|_\infty \leq c \end{cases}$$

Alors, avec $v = n\sigma^2 + 2c\mathbb{E}[Z]$, pour tout $t \geq 0$:

$$\mathbb{P}(Z \geq \mathbb{E}[Z] + t) \leq \exp\left(-v\mathrm{h}\left(\frac{t}{cv}\right)\right);$$

$$\mathbb{P}\left(Z \geq \mathbb{E}[Z] + \sqrt{2tv} + \frac{ct}{3}\right) \leq \exp(-t).$$

Démonstration. Pour se ramener au cadre du théorème, quitte à considérer f/c pour normaliser, on suppose que $c = 1$. Notons,

— $s_n : f \in \mathcal{F} \mapsto \sum_{i=1}^n f(X_i)$;

— Pour tout $i \in \{1, ..., n\}$, $s_n^i : f \in \mathcal{F} \mapsto \sum_{\substack{k=1 \\ k\neq i}}^n f(X_k)$.

Commençons par supposer $\mathcal{F}$ dénombrable et notons ses éléments $\mathcal{F} = (f_i)_{i\geq 1}$ puis posons $\mathcal{F}_N = \{f_1, ..., f_N\}$. Définissons $Z_i = \sup_{f\in\mathcal{F}} \sum_{\substack{k=1 \\ k\neq i}}^n f(X_k)$, alors Z_i existe car $\mathcal{F}$ est borné et

$$Z = \lim_{N\longrightarrow\infty} \uparrow \sup_{f\in\mathcal{F}_\mathcal{N}} s_n(f) \text{ et } Z_i = \lim_{N\longrightarrow\infty} \uparrow \sup_{f\in\mathcal{F}_\mathcal{N}} s_n^i(f).$$

Pour tout $N \in \mathbb{N}$ et tout $i \in \{1, ..., n\}$, $\mathcal{F}_N$ est fini donc il existe f_0^N et f_i^N dans $\mathcal{F}_N$ tels que :

$$\sup_{f\in\mathcal{F}_\mathcal{N}} s_n(f) = s_n(f_0^N) \text{ et } \sup_{f\in\mathcal{F}_\mathcal{N}} s_n^i(f) = s_n^i(f_i^N).$$

Alors, pour tout $N \in \mathbb{N}$ et tout $i \in \{1, \ldots, n\}$:

— $s_n(f_0^N) - s_n^i(f_i^N) = \sum_{\substack{k=1 \\ k\neq i}}^n f_0^N(X_k) - \sup_{f\in\mathcal{F}_N} \sum_{\substack{k=1 \\ k\neq i}}^n f(X_k) + f_0^N(X_i) \leq f_0^N(X_i) \leq 1$ p.s.

— $f_i^N(X_i) + s_n^i(f_i^N) = \sum_{k=1}^n f_i^N(X_k) \leq \sup_{f\in\mathcal{F}_N} \sum_{k=1}^n f(X_k) \leq \sup_{f\in\mathcal{F}} \sum_{k=1}^n f(X_k) = Z$ p.s.

— $\sum_{i=1}^n s_n(f_0^N) - s_n^i(f_i^N) \leq \sum_{i=1}^n f_0^N(X_i) \leq Z$ p.s.

— $-1 \leq s_n^i(f_i^N) \leq 1$ p.s.

— $\frac{1}{n}\sum_{k=1}^n \mathbb{E}_k[f_k^N(X_k)^2] = \frac{1}{n}\sum_{k=1}^n \mathbb{E}[(f_k^N(X_k))^2] \leq \sigma^2$

Quitte à considérer une suite extraite de la suite $(f_i^N(X_i))_N$ bornée presque sûrement dans $\mathbb{R}$ et faire de même pour les suites $(s_n(f_0^N))_N$ et $(s_n(f_i^N))_N$ définies plus haut, qui elles sont dans tous les cas convergentes, supposons que $\lim_{N\longrightarrow\infty} f_i^N(X_i)$ existe et posons $Z_i' = \lim_{N\longrightarrow\infty} f_i^N(X_i)$. Alors en passant à la limite quand $N \longrightarrow \infty$ dans les inégalités précédentes et en utilisant le théorème de convergence monotone pour le dernier point on construit $(Z_k)_{1\leq k\leq n}$ et $(Z_k')_{1\leq k\leq n}$ qui vérifient les hypothèses du théorème de concentration. Donc celui ci s'applique à Z et on obtient le résultat dans le cas où $\mathcal{F}$ est dénombrable.

Supposons maintenant $(\mathcal{F}, \|.\|_\infty)$ séparable, alors $\mathcal{F}$ admet un sous ensemble $\widetilde{\mathcal{F}}$ dénombrable et dense. Les résultats précédents s'appliquent à $\widetilde{F}$, il suffit de montrer que :

$$\sup_{f\in\widetilde{\mathcal{F}}} s_n(f) = \sup_{f\in\mathcal{F}} s_n(f).$$

La preuve pour les fonctions $(s_n^i)_i$ est identique.

Supposons que cette égalité soit fausse, alors par inclusion on a $\sup_{f\in\widetilde{\mathcal{F}}} s_n(f) < \sup_{f\in\mathcal{F}} s_n(f)$, donc il existe $f \in \mathcal{F}$ telle que :

$$\sup_{f\in\widetilde{\mathcal{F}}} s_n(f) < s_n(f) < \sup_{f\in\mathcal{F}} s_n(f). \tag{4}$$

f n'est pas dans $\widetilde{\mathcal{F}}$ par définition de la borne supérieure et il existe une suite $(f_k)_k$ de $\widetilde{\mathcal{F}}$ telle que $\|f_k - f\|_\infty \longrightarrow_{k\longrightarrow\infty} 0$. Or $|s_n(f_k) - s_n(f)| \leq n\|f_k - f\|_\infty$, donc

$$|s_n(f_k) - s_n(f)| \longrightarrow_{k\longrightarrow\infty} 0. \tag{5}$$

Puis, par 4, il existe $\epsilon > 0$ tel que $\sup_{f\in\widetilde{\mathcal{F}}} s_n(f) + \epsilon \leq s_n(f)$ et par 5 il existe un entier K tel que $s_n(f) \leq s_n(f_k) + \epsilon/2$. En combinant les deux inégalités on obtient :

$$\sup_{f\in\widetilde{\mathcal{F}}} s_n(f) < s_n(f_k),$$

ce qui est absurde.

□

On peut en déduire une inégalité de concentration sur V^+ définie par :

$$V^+ := \sup_{f\in\mathcal{F}} \left[\mathbb{E}[f(X_i)] - \frac{1}{n}\sum_{k=1}^{n} f(X_i) \right].$$

Nous pouvons établir le corollaire suivant qui constitue une borne supérieure de l'erreur de généralisation.

Corollaire 4.
Soient $\mathcal{F}$ une classe de fonctions de $\mathfrak{X} \to [0, M]$ et $\rho > 0$ tels que pour tout $f, i \in \mathcal{F} \times \{1, \ldots, n\}$, $\mathbb{V}\mathrm{ar}(f(X_i)) \leq \rho$ et $(\mathcal{F}, \|.\|_\infty)$ est séparable. Alors, pour tout $\delta \in]0, 1]$, avec probabilité au moins $1 - \delta$:

$$\sup_{f\in\mathcal{F}} \left[\mathbb{E}[f(X_i)] - \frac{1}{n}\sum_{k=1}^{n} f(X_i) \right] \leq 4\mathcal{R}_n(\mathcal{F}) + \sqrt{\frac{2\rho\log(\delta^{-1})}{n}} + \frac{4M\log(\delta^{-1})}{3n}$$

Démonstration. Toujours en notant $V^+ := \sup_{f\in\mathcal{F}} \left[\mathbb{E}[f(X_i)] - \frac{1}{n}\sum_{k=1}^{n} f(X_i)\right]$, d'après le corollaire 2, pour tout $x \geq 0$, avec probabilité au moins $1 - \exp(-x)$:

$$V^+ \leq \mathbb{E}[V^+] + \sqrt{\frac{2\rho x}{n} + \frac{4M\mathbb{E}[V^+]x}{n}} + \frac{Mx}{3n}.$$

Or pour tout $u, v \in \mathbb{R}$: $2\sqrt{uv} \leq u + v$ et $\sqrt{u+v} \leq \sqrt{u} + \sqrt{v}$, donc avec probabilité au moins $1 - \exp(-x)$:

$$\begin{aligned} V^+ &\leq \mathbb{E}[V^+] + \sqrt{\frac{2\rho x}{n}} + 2\sqrt{\frac{M\mathbb{E}[V^+]x}{n}} + \frac{Mx}{3n} \\ &\leq 2\mathbb{E}[V^+] + \sqrt{\frac{2\rho x}{n}} + \frac{4Mx}{3n}. \end{aligned}$$

On utilise ensuite le résultat suivant pour majorer l'espérance de V^+ par la complexité de Rademacher de $\mathcal{F}$:

Lemme 5.

$$\mathbb{E}[V^+] \leq 2\mathcal{R}(\mathcal{F}).$$

Démonstration. Soit $S' = (X'_1, \ldots, X'_n)$ un échantillon i.i.d de loi $\mathbb{P}$ et indépendant de l'échantillon $S = (X_1, \ldots, X_n)$. Notons $\mathcal{E}(f) = \mathbb{E}[f(X_1)]$ et $\hat{\mathcal{E}}(f) = \frac{1}{n}\sum_{k=1}^n f(X_i)$ et $\mathcal{E}'(f) = \mathbb{E}[f(X'_1)]$ et $\hat{\mathcal{E}}'(f) = \frac{1}{n}\sum_{k=1}^n f(X'_i)$, alors $\mathbb{E}_{S'}\left[\hat{\mathcal{E}}'(f)\right] = \mathcal{E}'(f) = \mathcal{E}(f)$, donc :

$$\begin{aligned}
\mathbb{E}_S\left[\sup_{f\in\mathcal{F}}\left(\mathcal{E}(f) - \hat{\mathcal{E}}(f)\right)\right] &= \mathbb{E}_S\left[\sup_{f\in\mathcal{F}}\left(\mathbb{E}_{S'}\left[\hat{\mathcal{E}}'(f)\right] - \hat{\mathcal{E}}(f)\right)\right] \\
&= \mathbb{E}_S\left[\sup_{f\in\mathcal{F}}\mathbb{E}_{S'}\left[\hat{\mathcal{E}}'(f) - \hat{\mathcal{E}}(f)\right]\right] \text{ par indépendance de } S \text{ et } S', \\
&\leq \mathbb{E}_{S,S'}\left[\sup_{f\in\mathcal{F}}\left(\hat{\mathcal{E}}'(f) - \hat{\mathcal{E}}(f)\right)\right] \text{ par } \sup\mathbb{E}[.] \leq \mathbb{E}\sup[.],
\end{aligned}$$

Soit σ un vecteur de Rademacher de taille n indépendant de S et S'. Alors, par symétrie de la loi de Rademacher, puis en intégrant par rapport à cette loi et parce que le terme de gauche est indépendant de σ :

$$\begin{aligned}
\mathbb{E}_{S,S'}\left[\sup_{f\in\mathcal{F}}\left(\hat{\mathcal{E}}'(f) - \hat{\mathcal{E}}(f)\right)\right] &= \mathbb{E}_{S,S'}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n \sigma_i(f(X'_i) - f(X_i))\right] \\
&= \mathbb{E}_{S,S',\sigma}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n \sigma_i(f(X'_i) - f(X_i))\right] \\
&\leq \mathbb{E}_{S,S',\sigma}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n \sigma_i f(X'_i) + \sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n -\sigma_i f(X_i)\right] \\
&= \mathbb{E}_{S,\sigma}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n \sigma_i f(X'_i)\right] + \mathbb{E}_{S',\sigma}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n \sigma_i f(X'_i)\right] \\
&= 2\mathcal{R}_n(\mathcal{F}).
\end{aligned}$$

□

On conclut en posant $\delta = \exp(-x)$. □

3.2.3 Borne sur l'erreur de généralisation déformée

On peut maintenant établir le théorème de majoration de l'erreur de généralisation déformée.

Théorème 3.
Supposons que $\mathcal{X}$ est un espace de Hilbert séparable, que la loi de probabilité de X_i est telle que $\|X_i\| \leq B$ avec probabilité 1, pour une certaine constante $B > 0$ et que la fonction de perte est bornée et L-admissible ; c'est à dire $\ell(h, Z) \leq M$ avec probabilité 1, pour une certaine constante $M > 0$ et $|\ell(h, z) - \ell(h', z)| \leq L|\langle h, x\rangle - \langle h', x\rangle|$ pour tout $z \in \mathcal{Z}$ et

$h, h' \in \mathcal{H}$. Soit $a > 1$. Si un algorithme d'apprentissage est $\alpha(n)$-argument stable, alors, avec probabilité au moins $1 - 2\delta$,

$$R(h_S) - \frac{a}{a-1} R_S(h_S) \leq 8LB\sqrt{2\log(2/\delta)}\alpha(n) + \frac{(6a+8)M\log(1/\delta)}{3n}.$$

Démonstration.
Tout d'abord, on introduit une inégalité pour expliciter la relation entre la stabilité algorithmique et la complexité de la classe d'hypothèses. D'après le lemme 3, pour tous $a > 1$ et $\delta > 0$, avec probabilité au moins $1 - \delta$, on a :

$$R(h_S) - \frac{a}{a-1} R_S(h_S) \leq \sup_{h \in B_r} (R(h) - \frac{a}{a-1} R_S(h)).$$

Ensuite, on va majorer le terme $\sup_{h \in B_r} (R(h) - \frac{a}{a-1} R_S(h))$ avec grande probabilité. On note que pour tout $g \in \mathcal{G}_r$, $\mathbb{E}g(Z) \leq r$ et $g(Z) \in [0, M]$. D'où

$$var(g(Z)) \leq \mathbb{E}(g(Z))^2 \leq M\mathbb{E}g(Z) \leq Mr.$$

D'où, par le corollaire 4,

$$V_r \leq 4\mathcal{R}(\mathcal{G}_r) + \sqrt{\frac{2Mr\log(1/\delta)}{n}} + \frac{4M}{3}\frac{\log(1/\delta)}{n}$$

. Cherchons s qui vérifie :

$$\mathcal{R}(\mathcal{G}_s) + \sqrt{\frac{2Ms\log(1/\delta)}{n}} + \frac{4M}{3}\frac{\log(1/\delta)}{n} = \frac{s}{a}$$

En posant $X = \sqrt{s}$, cette équation se réécrit :

$$X^2 - a\sqrt{\frac{2M\log(1/\delta)}{n}} X - \frac{4Ma}{3}\frac{\log(1/\delta)}{n} - 4a\mathbb{R}(\mathcal{G}_r) = 0 \tag{6}$$

ce qui est équivalent à :

$$X = \frac{2aM\log(1/\delta)}{2n} \pm \frac{\sqrt{\Delta}}{2}$$

où

$$\Delta = \frac{2aM\log(1/\delta)}{n} + 4\frac{4Ma}{3}\frac{4\log(1/\delta)}{n} + 4a\mathbb{R}(\mathcal{G}_r)$$

est le discriminant de l'équation quadratique (6).
Ainsi, on peut choisir :

$$\begin{aligned} s = X^2 &= \left(\tfrac{2a^2M\log(1/\delta)}{2n} - \tfrac{\sqrt{\Delta}}{2}\right)^2 \\ &= \tfrac{2a^2M\log(1/\delta)}{4n} + \tfrac{\delta}{4} - \tfrac{2a\sqrt{\Delta}}{4}\sqrt{\tfrac{2M\log(1/\Delta}{n}} \\ &= \tfrac{2a^2M\log(1/\delta)}{4n} + \tfrac{2a^2M\log(1/\delta)}{4n} + \tfrac{4Ma}{3}\tfrac{\log(1/\delta)}{n} + 4a\mathbb{R}(\mathcal{G}_r) - \tfrac{2a\sqrt{\Delta}}{4}\sqrt{\tfrac{2M\log(1/\Delta}{n}} \end{aligned}$$

d'où,

$$\begin{aligned} s &\leq \frac{2a^2 M \log(1/\delta)}{4n} + \frac{2a^2 M \log(1/\delta)}{4n} + \frac{4Ma}{3}\frac{\log(1/\delta)}{n} + 4a\mathbb{R}(\mathcal{G}_r) \\ &\leq \frac{2a^2 M \log(1/\delta)}{n} + \frac{4Ma}{3}\frac{2\log(1/\delta)}{n} + 8a\mathbb{R}(\mathcal{G}_r) \end{aligned}$$

Ce qui veut dire qu'il existe un $r^* \leq \frac{2a^2 M \log(1/\delta)}{n} + \frac{4Ma}{3}\frac{2\log(1/\delta)}{n} + 8a\mathbb{R}(\mathcal{G}_r)$ tel que $V_{r^*} \leq r^*/a$. Ainsi, d'après le lemme 4, pour tout $h \in B_{r^*}$, avec probabilité au moins $1-\delta$, on a :

$$\begin{aligned} \mathbb{E}\ell(h,Z) &\leq \frac{a}{a-1}\frac{1}{n}\sum_{i=1}^n \ell(h,Z_i) + \frac{r^*}{a} \\ &\leq \frac{a}{a-1}\frac{1}{n}\sum_{i=1}^n \ell(h,Z_i) + \frac{2a^2 M \log(1/\delta)}{n} + \frac{4Ma}{3}\frac{2\log(1/\delta)}{n} + 8a\mathbb{R}(\mathcal{G}_r^*). \end{aligned}$$

A présent, comme $\mathcal{G}_r \subset \{\alpha\ell \circ h | \, h \in B_r, \alpha \in [0,1]\} \subset conv(\ell \circ B_r)$. Par les propriétés élémentaires de la complexité de Rademacher, si $\mathcal{H} \subset \mathcal{H}'$ alors $\mathcal{R}(\mathcal{H}) \subset \mathcal{R}(\mathcal{H}')$, d'où avec probabilité $1-\delta$, on a

$$\sup_{h \in B_r}(R(h) - \frac{a}{a-1}R_S(h)) \leq \frac{2a^2 M \log(1/\delta)}{n} + \frac{4Ma}{3}\frac{2\log(1/\delta)}{n} + 8a\mathcal{R}(\ell \circ B_{r^*}).$$

Finalement, par le théorème 1, le théorème 2 et le lemme de contraction de Talagrand, on obtient l'inégalité recherchée. □

4 Application aux variantes de l'ERM

La méthode ERM, pour *empirical risk minimisation*, consiste à modifier les paramètres de façon à minimiser l'erreur empirique de l'algorithme : $R_S(h) = \frac{1}{n}\sum_i \ell(h, Z_i)$. Il est alors nécessaire de stabiliser l'algorithme, afin d'éviter un sur-apprentissage. En stabilisant l'algorithme, nous atténuons le biais induit par les données, et on s'assure alors que l'algorithme minimise effectivement l'erreur globale.

La stabilité de l'ERM peut s'obtenir soit par la régularisation, soit par un entraînement par *stochastic gradient descent*

4.1 Stabilité de l'ERM régularisée

Dans cette section, nous nous plaçons dans le cas où $\mathcal{B}$ est un espace de Hilbert séparable et montrons que, sous de bonnes hypothèses de régularité sur les fonctions de perte et de pénalisation, l'algorithme de minimisation du risque empirique régularisé

$$\underset{h\in\mathcal{H}}{argmin}\, \frac{1}{n}\sum_{i=1}^{n} \ell(h, Z_i) + \lambda N(h)$$

est non seulement uniformément stable mais aussi argument uniformément stable [proposition 4], ce qui permet de déduire [théorème 4] par le théorème 3, une borne sur le risque de généralisation pour un tel algorithme.
Pour cela, nous aurons besoin des lemmes suivants.

Lemme 6. [Wibisono et al., 2009]
Pour $\lambda > 0$, il existe $\tau(\lambda) \geq 0$ tel que

$$N(h_S) \leq \tau(\lambda)$$

Démonstration. Comme h_S minimise le risque empirique régularisé on a,

$$\begin{aligned} \lambda N(h_S) &\leq \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \ell(h_S, Z_i) + \lambda N(h_S) \\ &\leq \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \ell(h_S, Z_i) + \lambda N(0) \\ &\leq B + \lambda N(0). \end{aligned}$$

d'où le résultat (en divisant les deux membres de l'inégalité par λ et) en prenant $\tau(\lambda) = \frac{B}{\lambda} + N(0)$. □

Pour le second lemme, introduisons les notations suivantes, pour $h \in \mathcal{H}$, $z_1, \cdots, z_n, z^{'} \in \mathcal{X} \times \mathcal{Y}$ et $\lambda > 0$, posons :

- $R_r(h) := \frac{1}{n}\sum_{i=1}^{n} \ell(h, z_i) + \lambda N(h)$,
- $R_{emp} := \frac{1}{n}\sum_{i=1}^{n} \ell(h, z_i)$
- $R_r^{\setminus j}(h) := \frac{1}{n}\sum_{i\neq j}\ \ell(h, z_i) + \frac{1}{n}\ell(h, z^{'}) + \lambda N(h)$
- $R_{emp}^{\setminus j}(h) := \frac{1}{n}\sum_{i\neq j}\ \ell(h, z_i) + \frac{1}{n}\ell(h, z^{'})$

— $\Delta h := h^{\setminus j} - h$

Lemme 7. [Bousquet and Elisseeff, 2002]
Soit N une fonctionnelle définie sur $\mathcal{B}$, telle que pour tout échantillon (d'entraînement) S, R_r et $R_r^{\setminus j}$ admettent des minima. Notons h un minimiseur de R_r et $h^{\setminus j}$ un minimiseur de $R_r^{\setminus j}$. Si la fonction de perte ℓ est convexe et L-Lipschitz en sa première variable alors

$$N(h) - N(h + t\Delta h) + N(h^{\setminus j} - t\Delta h) - N(h^{\setminus j}) \leq \frac{2tL}{\lambda n}|\Delta h(x_j)|$$

Démonstration.
On a, par convexité de ℓ (et donc de $R_{emp}^{\setminus j}(h)$),

$$R_{emp}^{\setminus j}(h + t\Delta h) - R_{emp}^{\setminus j}(h) \leq t(R_{emp}^{\setminus j}(h^{\setminus j}) - R_{emp}^{\setminus j}(h)$$

et en échangeant le rôle de h et $h^{\setminus j}$,

$$R_{emp}^{\setminus j}(h^{\setminus j} - t\Delta h) - R_{emp}^{\setminus j}(h^{\setminus j}) \leq t(R_{emp}^{\setminus j}(h) - R_{emp}^{\setminus j}(h^{\setminus j})$$

d'où par sommation,

$$R_{emp}^{\setminus j}(h + t\Delta h) - R_{emp}^{\setminus j}(h) + R_{emp}^{\setminus j}(h^{\setminus j} - t\Delta h) - R_{emp}^{\setminus j}(h^{\setminus j}) \leq 0 \tag{7}$$

d'autre part, par hypothèse,

$$R_r(h) - R_r(h + t\Delta h) \leq 0 \quad \text{et} \quad R_r^{\setminus j}(h^{\setminus j}) - R_r^{\setminus j}(h^{\setminus j} - t\Delta h) \leq 0$$

ainsi, en utilisant l'inégalité (7), on trouve :

$$\begin{aligned} &n\lambda(N(h) - N(h + t\Delta h) + N(h^{\setminus j}) - N(h^{\setminus j} - t\Delta h)) \\ &\leq \ell(h, z_j) - \ell(h + t\Delta h, z_j) + \ell(h^{\setminus j}, z') - \ell(h - t\Delta h, z') \\ &\leq 2tL\|h_{s^j} - h_s\| \end{aligned}$$

et ce par la propriété de Lipschitz de ℓ. □

Lemme 8. [Wibisono et al., 2009]
Pour $1 < p \leq 2$, la régularisation avec la norme ℓ_p, vérifie le critère suivant,

$$\|h_S\|_{\ell_p} + \|h_{S^i}\|_{\ell_p} - 2\left\|\frac{h_S + h_{S^i}}{2}\right\|_{\ell_p} \geq C\|h_S - h_{S^i}\|_{\ell_2}^{\xi}.$$

avec $\xi = 2$ et $C = \frac{1}{4}p(p-1)\left(\frac{B}{\lambda}\right)^{\frac{p-2}{p}}$.

Démonstration.
Considérons la fonction $g : \mathbb{R} \longmapsto \mathbb{R}$ définie par : $g(\theta) = |\theta|^p$. Notons que g est dérivable avec $g'(\theta) = p\, sign(\theta)|\theta|^{p-1}$. Et g' est dérivable sur $\mathbb{R}^*$ avec $g''(\theta) = p(p-1)|\theta|^{p-2}$. Quand au comportement de g' au voisinage de 0, nous avons :

$$g''(0) = \lim_{\epsilon \to 0} \frac{g(\epsilon) - 2g(0) + g(-\epsilon)}{\epsilon^2} = \lim_{\epsilon \to 0} \frac{2|\epsilon|^p}{\epsilon^2} = \begin{cases} 2, & \text{si } p = 2, \\ +\infty & \text{si } p < 2. \end{cases}$$

donc, par la formule de Taylor-Lagrange, nous avons pour tous $a, b \in \mathbb{R}$,

$$|a|^p = \left|\frac{a+b}{2} + \frac{a-b}{2}\right|^p = \left|\frac{a+b}{2}\right|^p + sign\left(\frac{a+b}{2}\right) p(a-b) \left|\frac{a+b}{2}\right|^{p-1} + \frac{1}{2}\left(\frac{a-b}{2}\right)^2 p(p-1)|c_1|^{p-2}$$

pour un certain c_1 compris entre a et $\frac{a+b}{2}$.
En inversant les rôles de a et b, on obtient la formule analogue suivante :

$$|b|^p = \left|\frac{a+b}{2} + \frac{b-a}{2}\right|^p = \left|\frac{a+b}{2}\right|^p + sign(\frac{a+b}{2}) p(b-a) \left|\frac{a+b}{2}\right|^{p-1} + \frac{1}{2}\left(\frac{b-a}{2}\right)^2 p(p-1)|c_2|^{p-2}$$

pour un certain c_2 compris entre $\frac{a+b}{2}$ et b.
Ainsi, en posant $c = (|\frac{c_1}{2}|^{p-2} + |\frac{c_2}{2}|^{p-2})^{\frac{1}{p-2}}$, et par sommation, nous obtenons :

$$|a|^p + |b|^p - 2\left|\frac{a+b}{2}\right|^p = \frac{1}{4}(b-a)^2 p(p-1)|c|^{p-2}$$

où $\min\{a, b\} \leq c \leq \max\{a, b\}$.
En particulier, comme $\mathcal{B}$ est un Hilbert séparable, nous pouvons prendre pour a et b les composantes $\alpha_{z,\gamma}$ et $\alpha_{z^j,\gamma}$ pour tout $\gamma \in \mathbb{N}$, de h_S et h_{S^j} respectivement dans une base hilbertienne de $\mathcal{B}$, pour obtenir :

$$|\alpha_{z,\gamma}|^p + |\alpha_{z^j,\gamma}|^p - 2\left|\frac{\alpha_{z,\gamma} + \alpha_{z^j,\gamma}}{2}\right|^p = \frac{1}{4}(\alpha_{z,\gamma} - \alpha_{z^j,\gamma})^2 p(p-1)|c_\gamma|^{p-2} \tag{8}$$

où c_γ compris entre $\alpha_{z,\gamma}$ et $\alpha_{z^j,\gamma}$.
A présent, par le lemme 6 avec $\tau(\lambda) = (B/\lambda)^{1/p}$, nous obtenons la borne pour c_γ suivante :

$$|c_\gamma| \leq \max\{|\alpha_{z,\gamma}|, |\alpha_{z^j,\gamma}|\} \leq \max\{\|\alpha_z\|_{l_2}, ||\alpha_{z^j}||_{l_2}\} \leq \left(\frac{B}{\lambda}\right)^{1/p}.$$

De plus, puisque $1 < p \leq 2$, ceci implique

$$|c_\gamma|^{p-2} \geq \left(\frac{B}{\lambda}\right)^{\frac{p-2}{p}}.$$

En remplaçant dans l'équation (8), nous obtenons

$$|\alpha_{z,\gamma}|_{\ell_p}^p + |\alpha_{z^j,\gamma}|^p - 2\left|\frac{\alpha_{z,\gamma} + \alpha_{z^j,\gamma}}{2}\right|^p \geq \frac{1}{4}p(p-1)\left(\frac{B}{\lambda}\right)^{\frac{p-2}{p}} |\alpha_{z,\gamma} - \alpha_{z^j,\gamma}|^2$$

Finalement, par sommation sur $\gamma \in \Gamma$,

$$\|h_S\|_{\ell_p}^p + \|h_{S^j}\|_{\ell_p}^p - 2\left\|\frac{h_S + h_{S^j}}{2}\right\|_{\ell_p}^p \geq \frac{1}{4}p(p-1)\left(\frac{B}{\lambda}\right)^{\frac{p-2}{p}} \|h_S - h_{S^j}\|_{\ell_2}^2$$

□

Proposition 4. [Wibisono et al., 2009]
Supposons que la loi marginale de X_i est telle que $\|X_i\| \leq B$ avec probabilité 1, pour une certaine constante $B > 0$ et que la fonction de perte est convexe en sa première variable h, bornée par M et L-Lipschitz. Supposons de plus qu'il existe des constantes $C > 0$ et $\xi > 1$, telles que la fonction de pénalité N vérifie :

$$N(h_S) + N(h_{S^i}) - 2N\left(\frac{h_S + h_{S^i}}{2}\right) \geq C\|h_S - h_{S^i}\|_{\ell_2}^{\xi}. \tag{9}$$

Alors, l'algorithme ERM régularisé est $\beta(n)$-uniformément stable avec :

$$\beta(n) = \left(\frac{B^{\xi}L^{\xi}}{C\lambda n}\right)^{\frac{1}{\xi-1}}$$

et il est $\alpha(n)$-argument uniformément stable avec :

$$\alpha(n) = \left(\frac{BL}{C\lambda n}\right)^{\frac{1}{\xi-1}}.$$

Démonstration.
On a, tout d'abord, en se restreignant à $\mathbf{B}(0, B) \subset \mathcal{X}$, puisque p.s. $\|X_i\| \leq B$:

$$\|h_S - h_{S^j}\|_{\infty} = \sup_{x \in \mathbf{B}(0,B)} |h_S(x) - h_{S^j}(x)| \leq B\|h_S - h_{S^j}\|_{\ell_2}.$$

d'où, par le lemme 7 avec $t = 1/2$:

$$N(h_S) + N(h_{S^j}) - 2N\left(\frac{h_S + h_{S^j}}{2}\right) \leq \frac{L}{n\lambda}\|h_S - h_{S^j}\|_{\infty}$$

En combinant ceci avec l'hypothèse 9 sur N, on obtient :

$$\begin{aligned}\|h_S - h_{S^j}\|_{\ell_2}^{\xi} &\leq \tfrac{L}{n\lambda C}\|h_S - h_{S^j}\|_{\infty} \\ &\leq \tfrac{LB}{n\lambda C}\|h_S - h_{S^j}\|_{\ell_2}\end{aligned}$$

d'où

$$\|h_S - h_{S^j}\|_{\ell_2} \leq \left(\frac{LB}{n\lambda C}\right)^{\frac{1}{\xi-1}}$$

de plus, puisque ℓ est supposée $L-$admissible, on déduit aussi que :

$$\begin{aligned}\forall i \in \{1, \cdots, n\}, \quad |\ell(h - S, Z_i) - \ell(h_{S^j}, Z)| &\leq L\|h_S - h_{S^j}\|_{\infty} \\ &\leq LB\|h_S - h_{S^j}\|_{\ell_2}\end{aligned}$$

d'où,

$$\forall i \in \{1, \cdots, n\}, \quad |\ell(h - S, Z_i) - \ell(h_{S^j}, Z)| \leq \left(\frac{L^{\xi}B^{\xi}}{n\lambda C}\right)^{\frac{1}{\xi-1}}$$

□

Théorème 4.
Avec les hypothèses de la proposition 4, pour tout $\delta > 0$, et $a > 1$, si h_S est la sortie de la R-ERM, alors avec probabilité au moins $1 - 2\delta$,

$$R(h_S) - \frac{a}{a-1} R_S(h_S) \leq 8LB \left(\frac{LB}{C\lambda n} \right)^{\frac{1}{\xi - 1}} \sqrt{2\log(2/\delta)} + \frac{(6a+8)M \log(1/\delta)}{3n}$$

En particulier, lorsque $N(h) = \|h\|^2$, la condition (9) est vérifiée pour $\xi = 2$ et $C = \frac{1}{2}\left(\frac{M}{\lambda}\right)^{\frac{1}{2}}$.

Démonstration. La preuve est une conséquence directe du théorème 3, de la proposition 4 et du lemme 8. □

4.2 Stabilité de l'ERM entraîné par *stochastic gradient descent*

Si le jeu d'entraînement est très volumineux, une seule passe sur le jeux de donnée peux s'avérer très coûteuse. C'est pourquoi on préfère parfois minimiser cette fonction par descente de gradient stochastique. Nous démontrons ici que cette méthode d'optimisation a aussi le grand avantage de stabiliser notre algorithme.

Définition 14. (descente de gradient stochastique)
A chaque étape, on choisit un exemple au hasard, Z_{i_t}, puis on effectue un pas dans la direction de son gradient.
$h_{t+1} = h_t - \alpha_t \nabla_h \ell(h_t, Z_{i_t})$
avec α_t strictement positif et décroissant.

En prenant l'espérance sur i_t, qui suit une loi uniforme sur $\{1, ..., n\}$, on obtient $\mathbb{E}_{i_t}[\nabla_h \ell(h_t, Z_{i_t})] = \nabla_h R_S(h)$, ce qui justifie l'emploi de cette technique d'optimisation.

4.2.1 Régularité de la fonction de perte

Définition 15. (Smooth)
Une fonction de perte différentiable l est *s-smooth* si son gradient est s-Lipschitz :

$$\forall h, h' \in H, ||\nabla_h \ell(h, \cdot) - \nabla_{h'} \ell(h', \cdot)|| \leq s||h - h'||$$

Définition 16. (Fortement Convexe)
Une fonction de perte différentiable l est γ-*fortement convexe*, avec $\gamma > 0$, si :

$$(\nabla_h \ell(h, \cdot) - \nabla_{h'} \ell(h', \cdot))^T (h - h') \geq \gamma ||h - h'||^2$$

Lemme 9. f est γ-*fortement convexe* $\iff$ $\phi : x \mapsto f(x) - \frac{\gamma}{2}||x||^2$ est convexe

4.2.2 Propriétés de la règle de mise à jour du gradient.

On note $G_t(h) = h - \alpha \nabla l(h, Z_{i_t})$ la *règle de mise à jour du gradient* à la t^{ieme} itération.

Définition 17. (η-expansive)
La règle de mise à jour du gradient est dite η-expansive si :

$$\sup_{h,h' \in \mathcal{H}} \frac{||G(h) - G(h')||}{||h - h'||} \leq \eta$$

Définition 18. (σ-bornée)
La règle de mise à jour du gradient est dite σ-bornée si :

$$\sup_{h \in \mathcal{H}} ||h - G(h)|| \leq \sigma$$

Lemme 10. (Évolution de l'écart maximal entre hypothèses)
Soit $G_1, ..., G_T$ et $G'_1, ..., G'_T$ deux séquences de mise a jours, pour un même jeu de données. On effectue les deux SGD indépendemment : $h_{t+1} = G_t(h_t)$ et $h'_{t+1} = G_t(h_t)'$ On définit $\delta_t = ||h'_t - h_t||$, l'écart entre nos deux hypothèses respectives à l'instant t.
Si $G_t = G'_t$, et est η-expansive, alors :

$$\delta_{t+1} \leq \eta \delta_t$$

Si G_t et G'_t sont σ-bornées, et que G_t est η-expansive, alors :

$$\delta_{t+1} \leq \min(\eta, 1)\delta_t + 2\sigma_t$$

Démonstration. Le premier cas découle immédiatement de la définition de l'η-expansivité. Supposons à présent que G_t et G'_t sont σ-bornées, et que G_t est η-expansive. D'une part,

$$\begin{aligned}
\delta_{t+1} &= ||G_t(h_t) - G'_t(h'_t)|| \\
&\leq ||G_t(h_t) - h_t + h'_t + G'_t(h'_t)|| + ||h_t - h'_t|| \\
&\leq \delta_t + ||G_t(h_t) - h_t|| + ||G_t(h'_t) - h'_t|| \\
&\leq \delta_t + 2\sigma
\end{aligned}$$

D'autre part,

$$\begin{aligned}
\delta_{t+1} &= ||G(h_t) - G'(h'_t)|| \\
&= ||G(h_t) - G_t(h'_t) + G_t(h'_t) + G'(h'_t)|| \\
&\leq ||G_t(h_t) - G_t(h'_t)|| + ||G_t(h'_t) - G'_t(h'_t)|| \\
&\leq ||G_t(h_t) - G_t(h'_t)|| + ||h'_t - G_t(h'_t)|| + ||h'_t - G'_t(h'_t)|| \\
&\leq \eta\delta_t + 2\sigma
\end{aligned}$$

□

Notations :

— Soit $S = \{Z_1, ..., Z_i, ..., Z_n\}$ et $S^i = \{Z_1, ..., Z'_i, ..., Z_n\}$, deux exemples de d'entraînement qui ne diffèrent que d'une seule donnée.

— Soit $G_1, ..., G_T$ et $G'_1, ..., G'_T$ deux séquences de mise a jours, pour les jeux de données S et S' respectivement.

— On effectue les deux SGD indépendemment : $h_{t+1} = G_t(h_t)$ et $h'_{t+1} = G_t(h_t)'$. Ainsi, h_T et h_T^i sont les hypothèses respectivement retournées à partir d'un entraînement sur S et S^i après T pas de SGD .

— Les espérances concernent l'aléa du choix de l'exemple dans S ou dans S'.

4.2.3 Expansivité de la mise a jour du gradient

Nous montrons le lemme suivant dans le cas où la dimension est finie. Il est cependant possible de le montrer dans un cadre plus général.

Lemme 11. (Baillon-Haddad) Si f est convexe et s-smooth, alors $\langle \nabla f(h') - \nabla f(h), h - h' \rangle \geq \frac{1}{s}||\nabla f(h') - \nabla f(h)||^2$

Démonstration. — Supposons que f est C^2

Par convexité et s-coercivité de f, on a $0 \leq D^2 f \leq LI$. Alors

$$\begin{aligned}\langle \nabla f(h') - \nabla f(h), h - h' \rangle &= \int_0^1 \langle D^2 f(y + s(h - h'))(h - h'), h - h' \rangle \\ &= \langle A(h - h'), h - h' \rangle\end{aligned}$$

Avec $A := \int_0^1 \langle D^2 f(y + s(h - h'))ds$, qui est symétrique. De plus, $0 \leq A \leq LI$. Ainsi,

$$\begin{aligned}||\nabla f(h) - \nabla f(h')||^2 &= ||A(h - h')||^2 \\ &= \langle AA^{1/2}(h - h'), A^{1/2}(h - h') \rangle \\ &\leq s\langle A^{1/2}(h - h'), A^{1/2}(h - h') \rangle \\ &\leq s\langle A(h - h'), h - h' \rangle\end{aligned}$$

D'où le résultat.

— Cas général

Si f n'est pas C^2, on la régularise en la convoluant avec une approximation de l'unité ρ_n qui est C^2. On a alors que $f * \rho_n \xrightarrow{L^1} f$, et que $\forall n \in \mathbb{N}, f * \rho_n \in L^2$. Ainsi le résultat est vrai pour $f * \rho_n$, et donc pour f par passage à la limite.

□

Quand la perte est régulière, la mise a jours du gradient n'est pas trop expansive. C'est ce que formalise le lemme suivant, due à Polyak et Nesterov.

Lemme 12. Si f est s-smooth, alors :

1. **Cas général**
   $G_{f,\alpha}$ est $(1 + \alpha s)$-expansive.

2. **Cas convexe**
   Si de plus, f est convexe, alors pour tout $\alpha \leq 2/s$, alors $G_{f,\alpha}$ est 1-expansive.

3. **Cas strictement convexe**
   Si de plus, f est γ-fortement convexe, alors pour tout $\alpha \leq \frac{2}{s+\gamma}$, alors $G_{f,\alpha}$ est $\left(1 - \frac{\alpha s \gamma}{s+\gamma}\right)$-expansive.

Démonstration. Supposons que f est s-smooth

1. **Cas général**

   Comme f est s-smooth, et par inégalité triangulaire,

$$\begin{aligned} ||G_{f,\alpha}(h) - G_{f,\alpha}(h')|| &\leq ||h - h'|| + \alpha||\nabla f(h') - \nabla f(h)|| \\ &\leq ||h - h'|| + \alpha s||h - h'|| \\ &= (1 + \alpha s)||h - h'||. \end{aligned}$$

2. **Cas convexe**

   Supposons de plus que f est convexe. D'après le lemme de Baillon-Haddad, $\langle \nabla f(h) - \nabla f(h'), h - h' \rangle \geq \frac{1}{s}||\nabla f(h) - \nabla f(h')||^2$. Ainsi,

$$\begin{aligned} ||G_{f,\alpha}(h) - G_{f,\alpha}(h')||^2 &= ||h - h'||^2 - 2\alpha\langle \nabla f(h) - \nabla f(h'), h - h' \rangle + \alpha^2||\nabla f(h) - \nabla f(h')||^2 \\ &\leq ||h - h'||^2 - \left(\frac{2\alpha}{s} - \alpha^2\right)||\nabla f(h) - \nabla f(h')||^2 \\ &\leq ||h - h'||^2, \end{aligned}$$

   d'où le résultat.

3. **Cas strictement convexe**

   D'après le lemme 1, $\phi : h \mapsto f(h) - \frac{\gamma}{2}||h||^2$ est convexe, donc $\langle \nabla\phi(h) - \nabla\phi(h'), h - h' \rangle \geq 0$. De plus, $\nabla\phi(h) = \nabla f(h) - \gamma h$, d'où $\langle \nabla f(h) - \nabla f(h'), h - h' \rangle \geq \gamma\|h - h'\|^2$. Puis,

$$\begin{aligned} ||\nabla\phi(h) - \nabla\phi(h')||^2 &= ||\nabla f(h) - \nabla f(h')||^2 - 2\gamma\langle \nabla f(h) - \nabla f(h'), h - h' \rangle + \gamma^2||h - h'||^2 \\ &\leq ||\nabla f(h) - \nabla f(h')||^2 - \gamma^2||h - h'||^2 \\ &\leq (s^2 - \gamma^2)||h - h'||^2 \\ &\leq (s - \gamma)^2||h - h'||^2. \end{aligned}$$

   Ainsi ϕ est $(s - \gamma)$-smooth. D'après le lemme de Baillon-Haddad,

$$\langle \nabla\phi(h) - \nabla\phi(h'), h - h' \rangle \geq \frac{1}{s - \gamma}||\nabla\phi(h) - \nabla\phi(h')||^2.$$

Ainsi,

$$\begin{aligned}\langle \nabla f(h) - \nabla f(h'), h - h' \rangle &= \langle \nabla \phi(h) - \nabla \phi(h'), h - h' \rangle + \gamma ||h - h'||^2 \\ &\geq \frac{1}{s-\gamma} ||\nabla \phi(h) - \nabla \phi(h')||^2 + \gamma ||h - h'||^2 \\ &= \frac{1}{s-\gamma} (||\nabla f(h) - \nabla f(h')||^2 - 2\gamma \langle \nabla f(h) - \nabla f(h'), h - h' \rangle \\ &+ \gamma^2 ||h - h'||^2) + \gamma ||h - h'||^2 \\ &\geq \frac{1}{s-\gamma} ||\nabla f(h) - \nabla f(h')||^2 - \frac{2\gamma}{s-\gamma} \langle \nabla f(h) - \nabla f(h'), h - h' \rangle \\ &+ \frac{s\gamma}{s-\gamma} ||h - h'||^2,\end{aligned}$$

$$\frac{s+\gamma}{s-\gamma} \langle \nabla f(h) - \nabla f(h'), h - h' \rangle \geq \frac{1}{s-\gamma} ||\nabla f(h) - \nabla f(h')||^2 + \frac{s\gamma}{s-\gamma} ||h - h'||^2,$$

$$\langle \nabla f(h) - \nabla f(h'), h - h' \rangle \geq \frac{1}{s+\gamma} ||\nabla f(h) - \nabla f(h')||^2 + \frac{s\gamma}{s+\gamma} ||h - h'||^2.$$

On en déduit :

$$\begin{aligned}||G_{f,\alpha}(h) - G_{f,\alpha}(h')||^2 &= ||h - h'||^2 - 2\alpha \langle \nabla f(h) - \nabla f(h'), h - h' \rangle + \alpha^2 ||\nabla f(h) - \nabla f(h')||^2 \\ &\leq \left(1 - 2\frac{s\alpha\gamma}{s+\gamma}\right) ||h - h'||^2 - \alpha \left(\frac{2}{s+\gamma} - \alpha\right) ||\nabla f(h) - \nabla f(h')||^2.\end{aligned}$$

En utilisant que $\alpha \leq \frac{2}{s+\gamma}$, on obtient :

$$||G_{f,\alpha}(h) - G_{f,\alpha}(h')|| \leq \sqrt{1 - 2\frac{s\alpha\gamma}{s+\gamma}}\ ||h - h'||.$$

Or $\forall x \in [0,1]$, $\sqrt{1-x} \leq 1 - \frac{x}{2}$, d'où le résultat.

□

Lemme 13. Si f est L-Lipschitz, Alors $G_{f,\alpha}$ est (αL)-borné.

Démonstration.

$$||h - G_{f,\alpha}(h)|| = ||\alpha \nabla f(h)|| \leq \alpha L.$$

□

Lemme 14. Supposons que

— La perte l est s-Smooth et L-Admissible

— Le pas vérifie $\alpha_t \leq c/t$ pour une constante c

Alors $\forall z \in \mathcal{Z}, \forall t_0 \in \{0, ..., n\}$,

$$\mathbb{E}[\|h_T - h'_T\||S] \leq \frac{2cLt_0}{n} + \mathbb{E}[\delta_T|S, \delta_{t_0} = 0]$$

Démonstration. On note $\mathcal{E}$ l'évènement $(\delta_{t_0} = 0)$.

$$\mathbb{E}[\|h_T - h'_T\|] = \mathbb{P}(\mathcal{E})\mathbb{E}[\|h_T - h'_T\||\mathcal{E}] + \mathbb{P}(\mathcal{E}^c)\mathbb{E}[\|h_T - h'_T\|\mathcal{E}^c]$$

D'après les lemmes 13 et 10, $\mathbb{E}[\|h_T - h'_T\||S, \mathcal{E}^c] \leq 2T\alpha_T L \leq 2cL$ (car $\alpha_T \leq c/T$)

$$\mathbb{E}[\|h_T - h'_T\|] \leq \mathbb{E}[\|h_T - h'_T\|\mathcal{E}] + 2cL\ \mathbb{P}(\mathcal{E}^c)$$

Soit $i^* \in \{1, ..., n\}$ l'indice où S et S' diffèrent. Soit I la variable aléatoire correspondant à l'instant de la première utilisation de l'exemple z_{i^*}. Si $I > t_0$, alors $\delta_{t_0} = 0$. Ainsi,

$$\mathbb{P}(\mathcal{E}^c) = \mathbb{P}(\delta_{t_0} \neq 0) \leq \mathbb{P}(I \leq t_0)$$

Par borne d'union, $\mathbb{P}(I \leq t_0) = \frac{t_0}{n}$ d'où le résultat. □

Lemme 15. Soit $a > 1$. Alors

$$\sum_{k=n}^{p} \frac{1}{k^\alpha} \leq \frac{1}{(\alpha - 1)(n-1)^{\alpha-1}}$$

Démonstration. Soit $f(x) = \frac{-1}{(\alpha-1)x^{\alpha-1}}$ Cette fonction est dérivable, de dérivée : $f'(x) = \frac{1}{x^\alpha}$
En appliquant l'inégalité des accroissements finis sur l'intervalle $[k-1, k]$, on obtient : $\forall k > 1, \frac{1}{k^\alpha} \leqslant f(k) - f(k-1)$

et en sommant les membres de $k = n > 1$ jusqu'à $k = p$: $\sum_{k=n}^{p} \frac{1}{k^\alpha} \leqslant \sum_{k=n}^{p} (f(k) - f(k-1))$

Par télescopage, il nous reste : $\sum_{k=n}^{p} \frac{1}{k^\alpha} \leqslant f(p) - f(n-1)$

Comme f est négative sur $\mathbb{R}^+$, on obtient le résultat voulu. □

La proposition suivante permet d'obtenir la *Argument stability* de notre ERM.

Proposition 5. (*Argument stability* des algorithmes entraînés par SGD)

1. **Cas général :**
   Si :
   — La perte l est s-Smooth et L-Admissible
   — Le pas vérifie $\alpha_t \leq c/t$ pour une constante c
   — $\mathcal{H}$ est borné par une constante B.

Alors :

$$\mathbb{E}[\|h_T - h'_T\||S] \leq (c + \frac{1}{s})\frac{2LT^{\frac{sc}{sc+1}}}{n+1}$$

2. **Cas convexe**
   Si :
   — La perte l est s-Smooth, L-Admissible et convexe
   — Le pas vérifie $\alpha_t \leq 2/s$
   Alors :

$$\mathbb{E}[\|h_T - h'_T\||S] \leq \frac{2BL}{n}\sum_{t=1}^{T}\alpha_t$$

3. **Cas strictement convexe avec projection sur un convexe compacte**
   On considère ici la la SGD projetée sur un convexe compacte Ω : $h_{t+1} = \Pi_\Omega(h_t - \alpha_t \nabla_h \ell(h_t, Z_{i_t}))$
   Si
   — La perte l est s-Smooth, L-Admissible sur Ω et γ-fortement convexe
   — Le pas vérifie $\alpha_t \leq 1/s$
   Alors :

$$\mathbb{E}[\|h_T - h'_T\||S] \leq \frac{2BL}{\gamma n}$$

Remarque :

— Dans le cas strictement convexe, la borne ne dépend même plus du nombre d'itérations.
— De façon surprenante, la non convexité de la fonction de perte n'empêche pas d'obtenir la stabilité de la SGD, ce qui explique en partie l'excellente capacité de généralisation des réseaux de neurones.
— Notre borne concernant le cas général diffère de l'article d'origine. Ce dernier affirme qu'il est possible de généraliser la stabilité uniforme en stabilité des arguments dans l'article de Hardt (2015). Nous somme parvenu à une preuve de cette généralisation pour les deux autres cas, mais avons du procéder autrement pour le cas général. Néanmoins, les bornes obtenus sont équivalentes en T, le nombres d'étapes, et en n, le nombre d'exemples d'apprentissage.

Démonstration. La preuve du cas général s'avère être la plus délicate. Il s'agit d'utiliser une majoration probabiliste du temps à partir duquel la SGD fait usage de l'exemple qui diffère. Dans cette preuve, nous travaillons conditionnellement à S, nos données d'entraînement.

1. **Cas général**
   D'après le lemme précédent, $\forall z \in \mathcal{Z}, \forall t_0 \in \{0, ..., n\}$,

$$\mathbb{E}[\ell(h_T, z) - \ell(h'_T, z)|S] \leq \frac{t_0}{n}\sup_{h,z}\ell(h, z) + L\ \mathbb{E}[\delta_T|S, \delta_{t_0} = 0]$$

   avec $\delta_t = \|h_t - h'_t\|$. On pose $\Delta_t = \mathbb{E}[\delta_T|S, \delta_{t_0} = 0]$. A l'étape t, on a deux possibilités :

— Soit avec probabilité $1 - 1/n$, l'exemple sélectionné par la SGD est le même dans S et dans S'.

Dans ce cas, $G_t = G'_t$ et d'après le lemme 2, la mise à jour est $(1 + \alpha_t s)$-expansive.

— Soit avec probabilité $1/n$, l'exemple sélectionné est différent.

Dans ce cas, le lemme 5 implique que G_t et G'_t sont $(\alpha_t L)$-bornés. Le lemme 2 implique alors que $\delta_t \leq \delta_t + 2\alpha_t L$. Ainsi, par linéarité de l'espérance,

$$\begin{aligned}
\Delta_{t+1} &\leq \left(1 - \frac{1}{n}\right)(1 + \alpha_t s)\Delta_t + \frac{1}{n}\Delta_t + \frac{2\alpha_t L}{n} \\
&\leq \left(\frac{1}{n} + (1 - 1/n)(1 + cs/t)\right)\Delta_t + \frac{2cL}{tn} \\
&= \left(1 + (1 + 1/n)\frac{cs}{t}\right)\Delta_t + \frac{2cL}{tn} \\
&= \exp\left((1 + 1/n)\frac{cs}{t}\right)\Delta_t + \frac{2cL}{tn}
\end{aligned}$$

car $\forall x, 1 + x \leq \exp(x)$

Comme $\Delta_{t_0} = 0$, on obtient par récurrence :

$$\begin{aligned}
\Delta_T &\leq \sum_{t=t_0+1}^{T} \left(\prod_{k=t+1}^{T} \exp((1 - \frac{1}{n})\frac{sc}{k}\right)\frac{2cL}{tn} \\
&= \sum_{t=t_0+1}^{T} \left(\exp((1 - \frac{1}{n})sc\sum_{k=t+1}^{T}\frac{1}{k}\right)\frac{2cL}{tn} \\
&\leq \sum_{t=t_0+1}^{T} \left(\exp((1 - \frac{1}{n})sc\ln(\frac{T}{t})\right)\frac{2cL}{tn} \\
&= \frac{2cL}{n}T^{sc(1-1/n)}\sum_{t=t_0+1}^{T} t^{sc(1-1/n)-1}
\end{aligned}$$

En utilisant le lemme 15, applicable car $sc(1 - 1/n) > 0$, on obtient :

$$\sum_{t=t_0+1}^{T} t^{sc(1-1/n)-1} \leq \frac{1}{sc(1 - 1/n)t_0^{sc(1-1/n)-1}}$$

D'où

$$\begin{aligned}
\Delta_T &\leq \frac{1}{sc(1 - 1/n)}\frac{2cL}{n}\left(\frac{T}{t_0}\right)^{sc(1-1/n)} \\
&\leq \frac{2L}{s(n-1)}\left(\frac{T}{t_0}\right)^{sc}
\end{aligned}$$

On applique le lemme 14 :

$$\begin{aligned}\mathbb{E}[\|h_T - h'_T\||S] &\leq \frac{2cLt_0}{n} + \Delta_t \\ &\leq \frac{2cLt_0}{n} + \frac{2L}{s(n-1)}\left(\frac{T}{t_0}\right)^{sc}\end{aligned}$$

Soit $\phi : t \mapsto \frac{2cLt}{n} + \frac{2L}{s(n-1)}\left(\frac{T}{t}\right)^{sc}$

$$\phi'(t) = \frac{2cL}{n} - \frac{2Lc}{(n-1)}\frac{T^{sc}}{t^{sc+1}} \phi''(t) = \frac{2Lc(sc+1)}{n-1}\frac{T^{sc}}{t^{sc+2}}$$

$\forall t \geq 0, \phi''(t) \geq 0$ donc ϕ est convexe.

$$\phi'(t) = 0 \iff \frac{1}{n} = \frac{1}{n-1}\frac{T^{sc}}{t^{sc+1}} \iff t = \left(\frac{n}{n-1}T^{sc}\right)^{\frac{1}{sc+1}}$$

Qui correspond au minimum de ϕ. On choisit le minorant plus grossier $t_0 = T^{\frac{sc}{sc+1}}$
Ce qui nous donne

$$\begin{aligned}\mathbb{E}[\|h_T - h'_T\||S] &\leq \frac{2cLT^{\frac{sc}{sc+1}}}{n} + \frac{2L}{s(n-1)}T^{\frac{sc}{sc+1}} \\ &\leq (c + \frac{1}{s})\frac{2LT^{\frac{sc}{sc+1}}}{n+1}\end{aligned}$$

2. **Cas convexe**
   A l'étape t, on a deux possibilités :
   — Soit avec probabilité $1 - 1/n$, l'exemple sélectionné par la SGD est le même dans S et dans S'.

   Dans ce cas, $G_t = G'_t$ et d'après le lemme 2, la mise à jour est 1 expansive.
   — Soit avec probabilité $1/n$, l'exemple sélectionné est différent.

   Dans ce cas, le lemme 5 implique que G_t et G'_t sont $(\alpha_t L)$-bornés. Le lemme 2 implique alors que $\delta_t \leq \delta_t + 2\alpha_t L$.
   Ainsi, par linéarité de l'espérance, on a $\forall t \geq t_0$,

$$\mathbb{E}[\delta_{t+1}|S] \leq \left(1 - \frac{1}{n}\right)\mathbb{E}[\delta_t|S] + \frac{1}{n}\mathbb{E}[\delta_t|S] + \frac{2\alpha_t L}{n} = \mathbb{E}[\delta_t|S] + \frac{2\alpha_t L}{n}$$

   Comme $\delta_0 = 0$, on obtient par téléscopage :

$$\mathbb{E}[\delta_T|S] = \sum_{t=1}^{T}\mathbb{E}[\delta_t|S] \leq \frac{2L}{n}\sum_{t=1}^{T}\alpha_t$$

3. **Cas strictement convexe avec projection sur un convexe compacte**
   Le gradient de l est continu sur le compact Ω, donc borné par une constante M. On considère un exemple $z \in \mathcal{Z}$. A l'étape t, on a deux possibilités :

— Soit avec probabilité $1 - 1/n$, l'exemple sélectionné par la SGD est le même dans S et dans S'.

Dans ce cas, $G_t = G'_t$. Comme la projection euclidienne est 1-lipschitzienne,

$$\delta_t \leq ||h_{t+1} - \alpha\nabla l(h_t, z) - h'_{t-1} + \alpha\nabla\ell(f'_t, z)||$$

et d'après le lemme 2, la mise à jour est 1 expansive.
Si $\alpha \leq 1/s$, comme $\frac{2\alpha s\gamma}{s+\gamma} \geq \alpha\gamma$ et $\alpha\gamma \leq 1$, le lemme 2 implique que $G_{f,\alpha}$ est $(1 - \alpha\gamma)$-expansif.

— Soit avec probabilité $1/n$, l'exemple sélectionné est différent.

Dans ce cas, le lemme 5 implique que G_t et G'_t sont $(\alpha_t L)$-bornés. Le lemme 2 implique alors que $\delta_t \leq \delta_t + 2\alpha_t L$.
Ainsi, par linéarité de l'espérance,

$$\begin{aligned}\mathbb{E}[\delta_{t+1}|S] &\leq \left(1 - \frac{1}{n}\right)(1 - \alpha\gamma)\mathbb{E}[\delta_t|S] + \frac{1}{n}(1 - \alpha\gamma)\mathbb{E}[\delta_t|S] + \frac{2\alpha_t L}{n} \\ &= (1 - \alpha\gamma)\mathbb{E}[\delta_t|S] + \frac{2\alpha_t L}{n}\end{aligned}$$

Comme $\delta_0 = 0$, on obtient par téléscopage :

$$\mathbb{E}[\delta_T|S] = \sum_{t=1}^{T}\mathbb{E}[\delta_t|S] \leq \frac{2L\alpha}{n}\sum_{t=1}^{T}(1 - \alpha\gamma)^t \leq \frac{2L}{\gamma n}$$

□

A l'aide de la proposition précédente, on peut obtenir la *Argument stability* des algorithmes entraînés par ERM. Grâce aux précédents résultats de l'article (?), On peut alors contrôler finement l'erreur de généralisation déformée de ces algorithmes.

Théorème 5. (Borne probabiliste de *l'erreur de généralisation déformée* des algorithmes entraînés par SGD)
On suppose que nos données sont bornées : $||X|| \leq B$ p.s.

1. **Cas général :**

   Si

   — La perte l est s-Smooth, L-Admissible et bornée par M.

   — Le pas vérifie $\alpha_t \leq c/t$ pour une constante c

   Alors $\forall\delta > 0, \forall a > 1$, on a avec probabilité au moins $1 - 2\delta$:

$$R(h_T) - \frac{a}{a-1}R_S(h_T) \leq 16(c + \frac{1}{s})\frac{BLT^{1+\frac{sc}{sc+1}}}{n+1}\sqrt{2\ln(2/\delta)} + \frac{(6a+8)M\ln(1/\delta)}{3n}$$

2. **Cas convexe**

   Si :

— La perte l est s-Smooth, L-Admissible, convexe et bornée par M.

— Le pas vérifie $\alpha_t \leq 2/s$

Alors $\forall \delta > 0, \forall a > 1$, on a avec probabilité au moins $1 - 2\delta$:

$$R(h_T) - \frac{a}{a-1} R_S(h_T) \leq \frac{16 B^2 L^2}{n} \sum_{t=1}^{T} \alpha_t \sqrt{2 \ln(2/\delta)} + \frac{(6a+8) M \ln(1/\delta)}{3n}$$

3. **Cas strictement convexe avec projection sur un convexe compact**

On considère ici la la SGD projetée sur un convexe compacte Ω : $h_{t+1} = \Pi_\Omega(h_t - \alpha_t \nabla_h \ell(h_t, Z_{i_t}))$
Si

— La perte l est s-Smooth, L-Admissible sur Ω, γ-fortement convexe et bornée par M.

— Le pas vérifie $\alpha_t \leq 1/s$

Alors :

$$R(h_T) - \frac{a}{a-1} R_S(h_T) \leq \frac{16 B^2 L^2}{\gamma n} \sum_{t=1}^{T} \alpha_t \sqrt{2 \ln(2/\delta)} + \frac{(6a+8) M \ln(1/\delta)}{3n}$$

Démonstration. On travaille dans R^d, qui est bien un espace de Hilbert.
Le théorème découle directement de la proposition précédente. □

Remarque :
il serait intéressant d'étudier si l'utilisation simultané de la SGD et de régularisation permet d'accroître encore la régularité.

4.3 Des algorithmes stables pour l'ERM régularisée

Nous avons établit dans la partie 4.1 des bornes sur l'erreur de généralisation dans la cadre d'algorithmes visant à minimiser le risque empirique régularisé. Dans cette section nous introduisons des algorithmes de minimisation du risque empirique régularisé dont l'erreur de généralisation converge en $O(1/n)$ sans avoir à vérifier la condition (9) du théorème 4. Il se trouve qu'en utilisant le théorème 3, le second terme du majorant est toujours en $O(1/n)$, on ne peut donc pas faire mieux qu'une convergence d'ordre 1 de l'erreur en utilisant ce théorème. L'algorithme de descente de gradient stochastique propose bien une borne en $O(1/n)$ mais nécessite la différentiabilité et le caractère smooth du risque empirique. Dans cette section nous utiliserons des méthodes d'approximation proximale pour établir des bornes sur l'erreur de généralisation en $O(1/n)$ sans avoir à vérifier ni la condition (9), ni la différentiabilité du risque empirique.

Dans ces deux premières sous-sections nous introduisons les algorithmes proximaux qui permettent de minimiser des fonctions convexes non différentiables sous quelques hypothèses. Nous citons sans les démontrer les propositions déjà connues qui seront nécessaires.

4.3.1 Opérateur proximal

Définition 19.
Soit $\mathbb{X}$ un espace topologique, on dit que $f : \mathbb{X} \to]-\infty, \infty]$ est *semi-continue inférieurement* en $x \in \mathbb{X}$, abrégé en s.c.i, si pour tout $\epsilon > 0$, il existe un ouvert U contenant x tel que :

$$\inf_U f \geq f(x) - \epsilon.$$

Dans la suite, on suppose que $\mathcal{H}$ est un espace de Hilbert séparable, notons $\Gamma_0(\mathcal{H})$ l'ensemble des fonctions $f : \mathcal{H} \to]-\infty, \infty]$ convexe, s.c.i et propre (de domaine non vide).

Les éléments de $\Gamma_0(\mathcal{H})$ ne sont pas tous différentiables, ceci motive la définition suivante qui généralise la notion de différentielle.

Définition 20.
Soit $f \in \Gamma_0(\mathcal{H})$, on appelle *sous-différentielle* de f la fonction $\partial f : \mathcal{H} \to \mathcal{P}(\mathcal{H})$ définie par :

$$\partial f : h \longmapsto \{u \in \mathcal{H} | \forall y \in \mathcal{H}, f(y) \geq f(h) + \langle u, y - h \rangle\}.$$

Les éléments de ∂f sont les *sous-gradients* de f.

Définition 21.
Soient $f \in \Gamma_0(\mathcal{H})$ et $\mu > 0$, on appelle *enveloppe de Moreau* de f d'indice μ la fonction $\tilde{f}_\mu : \mathcal{H} \to \mathbb{R}$ définie par :

$$\tilde{f}_\mu : y \longmapsto \inf_{h \in \mathcal{H}} \left(f(h) + \frac{\|y - h\|^2}{2\mu} \right).$$

On peut monter que l'enveloppe de Moreau d'une fonction de $\Gamma_0(\mathcal{H})$ est aussi élément de $\Gamma_0(\mathcal{H})$ qui est de plus C^1.

Proposition 6.
Soient $f \in \Gamma_0(\mathcal{H})$ et $y \in \mathcal{H}$. Alors il existe un unique point noté $\text{prox}_{\mu f}(y) \in \mathcal{H}$, appelé *point proximal* de y relativement à μf tel que :

$$\tilde{f}_\mu(y) = f(\text{prox}_{\mu f}(y)) + \frac{\|y - \text{prox}_{\mu f}(y)\|^2}{2\mu}.$$

La notation μf en indice du point proximal vient du fait que le point proximal de y relativement à l'enveloppe de Moreau de f d'indice μ est le même que le point proximal de y relativement à l'enveloppe de Moreau de μf d'indice 1.

Proposition 7.
Soient $f \in \Gamma_0(\mathcal{H})$ et $x, y \in \mathcal{H}$, alors :

$$f(y) - f(\text{prox}_f(x)) \geq \langle y - \text{prox}_f(x), x - \text{prox}_f(x) \rangle. \qquad \text{(Inégalité proximale)}$$

$$\|\text{prox}_f(x) - \text{prox}_f(y)\| \leq \|x - y\|. \qquad \text{(Contraction fermée)}$$

Dans un cadre de minimisation, l'opérateur proximal peut permettre de trouver les minimum. C'est l'objet de la proposition suivante.

Proposition 8.
Soient $f \in \Gamma_0(\mathcal{H})$ et $\mu > 0$, alors

$$x \text{ est un minimum de } f \Leftrightarrow x = \text{prox}_{\mu f}(x).$$

4.3.2 Algorithme du gradient proximal

Nous introduisons un algorithme qui minimise $F = f + g$ lorsque f est convexe différentiable, g est convexe et F est coercive.

Proposition 9.
Dans ce cadre, pour tout $\mu > 0$,

$$h \text{ est un minimum de } F \Leftrightarrow h = \text{prox}_{\mu g}(h - \mu \nabla f(h)).$$

Ainsi, dans l'objectif de minimiser F on peut considérer les itérées :

$$h_{T+1} = \text{prox}_{\mu_T f}(h_T - \mu_T \nabla f(h_T)),$$

où nous discuterons du choix de μ_T. On appelle l'algorithme définie comme tel l'*algorithme de gradient proximal* sur g. Il est important de comprendre que l'itérée h_{T+1} minimise la fonction :

$$Q^{\mu_T}(., h_T) : h \mapsto f(h_T) + \langle \nabla f(h_T), h - h_T \rangle + \frac{1}{2\mu_T}\|h - h_T\|^2 + g(h).$$

$Q^{\mu_T}(h, h_T)$ est une approximation de $F(h)$ obtenue en faisant un développement de Taylor à l'ordre 2 de f autour de h_T.

Choix du pas :
On construit la suite $(\mu_T)_T$ de la façon suivante : supposons μ_T construit, alors $\mu_{T+1} = \beta^p \mu_T$ où $\beta \in]0, 1[$ et p est le premier entier non nul tel que

$$F(\text{prox}_{\beta^p \mu_T f}(h_T - \beta^p \mu_T \nabla f(h_T))) \leq Q^{\mu_T}(\text{prox}_{\beta^p \mu_T f}(h_T - \beta^p \mu_T \nabla f(h_T)), h_T).$$

Ainsi, à chaque étape, h_T minimise une approximation majorante de F.

Théorème 6. (Convergence du gradient proximal)
Soient $f, g \in \Gamma_0(\mathcal{H})$ avec f différentiable et s-smooth et $F = f + g$ coercive, soient $h_0 \in \mathcal{H}$, $\beta \in]0, 1[$ et μ_0 tel que $\mu_0 s \geq 1$. Alors, en notant h_T la T-ème itération de l'algorithme du gradient proximal sur g et h_* un minimum de F :

$$\forall T \geq 0,\ F(h_T) - F(h_*) \leq \frac{s}{2\beta T}\|h_0 - h_*\|^2.$$

Ce théorème assure la convergence de l'algorithme du gradient proximal et nous donne une majoration du nombre d'itérations nécessaires pour obtenir une approximation du minimum satisfaisante.

4.3.3 Application à la RERM : fonction de pénalité non différentiable

Revenons dans un cadre d'apprentissage statistique. On aimerait utiliser l'algorithme du gradient proximal pour minimiser en h le risque empirique régularisé :

$$R_{S,\lambda}(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(h, X_i) + \lambda N(h).$$

Dans la suite de cette section nous supposons :

— Les données sont presque sûrement bornées par $B > 0$;

— $N \in \Gamma_0(\mathcal{H})$;

— $l \in \Gamma_0(\mathcal{H})$, différentiable, L-admissible, s-smooth et bornée par $M > 0$;

— $R_{S,\lambda}$ est coercive au sens de la norme $\|.\|$.

Notons que l'hypothèse supplémentaire de coercivité n'est pas trop restrictive, en général la fonction de pénalité N est elle même coercive au sens de $\|.\|$ et la perte est bornée.

Pour minimiser $R_{S,\lambda}$, nous allons utilisons une version stochastique de l'algorithme du gradient proximal. Considérons :

$$h_{S,T+1} = \text{prox}_{\mu_T N}(h_{S,T} - \mu_T \nabla_h \ell(h_{S,T}, X_{i_T})), \tag{10}$$

où i_T est un indice choisi uniformément aléatoirement dans $\{1, ..., n\}$.

Proposition 10. (Convergence en moyenne de l'algorithme)
Sous les hypothèses précédentes, soient $h_{S,0} \in \mathcal{H}$, $\beta \in]0, 1[$ et μ_0 tel que $\mu_0 \geq 1/s$, alors, en notant h_* un minimum de $R_{S,\lambda}$:

$$\forall T \geq 0,\ \mathbb{E}[R_{S,\lambda}(h_{S,T}) - R_{S,\lambda}(h_*)] \leq \frac{s}{2\beta T}\|h_{S,0} - h_*\|^2.$$

Démonstration. La preuve utilise les même argument que celle du théorème 6, il suffit d'établir les inégalités en espérance pour s'affranchir de la difficulté supplémentaire due au gradient stochastique. □

Remarque : μ_T sont des variables aléatoires avec cette méthode puisque qu'ils sont définies par une inégalité aléatoire, qui dépend de la donnée X_{i_T} choisie à l'étape T.

Proposition 11. (Argument stability)
Sous les hypothèses précédente et si $\mu_0 \leq 2/s$, alors pour tout $T \geq 0$:

$$\mathbb{E}\left[\|h_{S,T} - h_{S^i,T}\| \mid S\right] \leq \frac{2L}{n} \sum_{t=0}^{T} \mu_T \text{ p.s.}$$

Démonstration. Reprenons les notations introduite dans la partie 4.2 et notons $G_{S,T}$ la fonction de mise à jour du gradient stochastique à l'étape T sur l'échantillon S. Alors

$$h_{S,T+1} = \text{prox}_{\mu_T N}(G_{S,T}(h_{S,T})) \text{ et } h_{S^i,T+1} = \text{prox}_{\mu_T N}(G_{S^i,T}(h_{S,T})).$$

Sur l'évènement $\{G_{S,T} = G_{S^i,T}\}$:

$$\begin{aligned}\delta_{T+1} &\leq \|G_{S,T}(h_{S,T}) - G_{S,T}(h_{S^i,T})\| \text{ par concentration fermée} \\ &\leq \delta_T \text{ car } G_{S,T} \text{ est 1-expansive d'après 12.}\end{aligned}$$

Sur l'évènement $\{G_{S,T} \neq G_{S^i,T}\}$:

$$\begin{aligned}\delta_{T+1} &\leq \|G_{S,T}(h_{S,T}) - G_{S,T}(h_{S^i,T})\| \text{ par concentration fermée} \\ &\leq \|G_{S,T}(h_{S,T}) - h_{S,T}\| + \|h_{S,T} - h_{S^i,T}\| + \|h_{S^i,T} - G_{S,T}(h_{S^i,T})\| \\ &\leq \delta_T + 2\mu_T L \text{ car } G_{S,T} \text{ est } \mu_T L\text{-bornée.}\end{aligned}$$

Donc, parce que $\mathbb{P}(G_{S,T} = G_{S^i,T}|S) = 1 - \frac{1}{n}$ p.s. :

$$\mathbb{E}[\delta_{T+1}|S] \leq \mathbb{E}[\delta_T|S] + \frac{2\mu_T L}{n} \text{ p.s.}$$

Finalement, par téléscopage, on obtient le résultat.

□

Théorème 7.
Supposons que $\mathcal{H}$ est un Hilbert séparable. Si $\mu_0 \leq 2/s$, alors pour tout $\delta > 0$ et tout $a > 0$ et tout $T \geq 0$ avec probabilité au moins $1 - 2\delta$:

$$R(h_{S,T}) - \frac{a}{a-1} R_S(h_{S,T}) \leq 16L^2 B\sqrt{2\log(2/\delta)} \frac{\sum_{t=0}^{T} \mu_T}{n} + \frac{(6a+8)M\log(1/\delta)}{3n}.$$

Démonstration. C'est une application directe du théorème 3 et de la proposition 11. □

Remarques :

— Par construction des pas μ_T, $\mu_T \leq \beta\mu_{T-1} \leq \beta^T \mu_0$ p.s., donc :

$$\sum_{t=0}^{T} \mu_t \leq \frac{\mu_0}{1-\beta} \text{ p.s.}$$

— On peut s'affranchir de l'hypothèse $\mu_0 \leq 2/s$ et établir la proposition pour tout $T \geq T_0$ où $\mu_{T_0} \leq 2/s$ p.s. Un tel rang existe puisque $\mu_T \to_{T\to\infty} 0$ p.s. Ceci a un intérêt en pratique, si on veut avoir à la fois convergence de l'algorithme et stabilité, on doit vérifier respectivement $1/s \leq \mu_0 \leq 2/s$. Mais si on ne connaît pas précisément la constante s, on choisit un μ_0 assez grand pour assurer $\mu_0 \geq 1/s$ et le corollaire suivant.

Corollaire 5.
Dans le cadre du théorème 7 et en notant T_0 le premier indice tel que $\mu_{T_0} \leq 2/s$ p.s. Alors pour tout $\delta > 0$, tout $a > 0$ et tout $T \geq T_0$ avec probabilité au moins $1 - 2\delta$:

$$R(h_{S,T}) - \frac{a}{a-1} R_S(h_{S,T}) \leq 8LB\sqrt{2\log(2/\delta)} \left[\frac{2L\sum_{t=0}^{T} \mu_T}{n} + (1 - (1-1/n)^{T_0})\mathbb{E}[\delta_{T_0}|S] \right] + \frac{(6a+8)M\log(1/\delta)}{3n}.$$

Remarque : Le terme résiduel tend à disparaître quand l'échantillon devient suffisamment grand mais grandit quand T_0 augmente, c'est-à-dire quand $\mu_0 >> 1/s$.

Démonstration. Comme dans la preuve de la proposition 11, on peut établir pour tout $T \geq T_0$:

$$\mathbb{E}[\delta_{T+1}|S] \leq \mathbb{E}[\delta_T|S] + \frac{2\mu_T L}{n} \text{ p.s.} \tag{11}$$

Puis, parce que $\mathbb{P}(\delta_{T_0} = 0) = \mathbb{P}(\forall t \leq T_0, G_{S,t} = G_{S^i,t}) = (1-1/n)^{T_0}$, en distinguant deux cas on a :

$$\mathbb{E}[\delta_T|S] = (1-1/n)^{T_0}\mathbb{E}[\delta_T|S, \delta_{T_0} = 0] + (1 - (1-1/n)^{T_0})\mathbb{E}[\delta_T|S, \delta_{T_0} \neq 0] \text{ p.s.} \tag{12}$$

En combinant 11 et 12, puis en sommant, sachant que presque sûrement $\mathbb{E}[\delta_{T_0}|S, \delta_{T_0} = 0] = 0$ et $\mathbb{E}[\delta_{T_0}|S, \delta_{T_0} \neq 0] = \mathbb{E}[\delta_{T_0}|S]$, par téléscopage on obtient le résultat. □

Le choix du pas initial est important, ainsi avant l'application de l'algorithme il peut être intéressant d'estimer la valeur de s.

4.3.4 Application à la RERM : fonction de perte non différentiable

Dans cette section, on souhaite toujours minimiser le risque empirique régularisé mais sous des hypothèses différentes. On suppose dans la suite :

— Les données sont presque sûrement bornées par $B > 0$;

— $N \in \Gamma_0(\mathcal{H})$, différentiable, K-Lipschitz et s-smooth ;

— $l \in \Gamma_0(\mathcal{H})$, L-admissible et bornée par $M > 0$;

— $R_{S,\lambda}$ est coercive au sens de la norme $\|.\|$.

Nous considérons les itérés :

$$h_{S,T+1} = \text{prox}_{\mu_T l_{i_T}}(h_{S,T} - \mu_T \nabla N(h_{S,T})),$$

où $l_{i_T} : h \in \mathcal{H} \longmapsto \ell(h, X_{i_T})$ et i_T est un indice choisi uniformément aléatoirement dans $\{1, ..., n\}$. On note aussi $l^i_{i_T} : h \in \mathcal{H} \longmapsto \ell(h, X^i_{i_T})$ où $S^i = (X^i_1, ..., X^i_n)$.

Proposition 12. (Convergence en moyenne de l'algorithme)
Sous les hypothèses précédentes, soient $h_{S,0} \in \mathcal{H}$, $\beta \in]0, 1[$ et μ_0 tel que $\mu_0 \geq 1/s$, alors, en notant h_* un minimum de $R_{S,\lambda}$:

$$\forall T \geq 0, \mathbb{E}[R_{S,\lambda}(h_{S,T}) - R_{S,\lambda}(h_*)] \leq \frac{s}{2\beta T}\|h_{S,0} - h_*\|^2.$$

Démonstration. La preuve utilise les même argument que celle du théorème 6, il suffit d'établir les inégalités en espérance pour s'affranchir de la difficulté supplémentaire due à l'opérateur proximal aléatoire. □

Remarque : L'intérêt de calculer l'opérateur proximal sur une fonction l_{i_T} aléatoire est d'assurer la stabilité. Sur deux échantillons S et S^i, l'opérateur proximal sera le même à chaque itération avec grande probabilité si l'échantillon est grand. En contre-partie, on n'a plus qu'une convergence en moyenne vers le minimum.

Proposition 13.
Sous les hypothèses précédente et si $\mu_0 \leq 2/s$, alors pour tout $T \geq 0$:

$$\mathbb{E}\left[\|h_{S,T} - h_{S^i,T}\||S\right] \leq \frac{2\sqrt{KL}}{n} \sum_{t=0}^{T} \mu_T \text{ p.s.}$$

Démonstration. Avec les mêmes notations que dans la proposition précédente,

$$h_{S,T+1} = \text{prox}_{\mu_T l_{i_T}}(G_{S,T}(h_{S,T})) \text{ et } h_{S^i,T+1} = \text{prox}_{\mu_T l^i_{i_T}}(G_{S,T}(h_{S^i,T})).$$

Sur l'évènement $\{X_{i_T} = X^i_{i_T}\}$:

$$\begin{aligned}\delta_{T+1} &\leq \|G_{S,T}(h_{S,T}) - G_{S,T}(h_{S^i,T})\| \text{ par concentration fermée} \\ &\leq \delta_T \text{ car } G_{S,T} \text{ est 1-expansive d'après 12.}\end{aligned}$$

Sur l'évènement $\{X_{i_T} \neq X^i_{i_T}\}$:

$$\begin{aligned}\delta_{T+1} &\leq \|h_{S,T+1} - G_{S,T}(h_{S,T})\| + \|G_{S,T}(h_{S,T}) - G_{S,T}(h_{S^i,T})\| + \|G_{S,T}(h_{S^i,T}) - h_{S^i,T+1}\| \\ &\leq \delta_T + \|h_{S,T+1} - G_{S,T}(h_{S,T})\| + \|G_{S,T}(h_{S^i,T}) - h_{S^i,T+1}\| \\ &\leq \delta_T + \sqrt{\mu_T \left|l_{i_T}(h_{S,T+1}) - l_{i_T}(G_{S,T}(h_{S,T}))\right|} + \sqrt{\mu_T \left|l^i_{i_T}(h_{S^i,T+1}) - l^i_{i_T}(G_{S,T}(h_{S^i,T}))\right|},\end{aligned}$$

par l'inégalité proximale appliquée en $y = x = G_{S,T}(h_{S,T})$ et $y = x = G_{S,T}(h_{S^i,T})$. Puis par admissibilité de la perte et parce $G_{S,T}$ est $\mu_T K$-bornées :

$$\delta_{T+1} \leq \delta_T + 2\sqrt{KL}\mu_T.$$

On conclut comme dans la proposition précédente, en prenant l'espérance et en sommant les T premiers termes. □

Théorème 8.
Supposons que $\mathcal{H}$ est un Hilbert séparable. Si $\mu_0 \leq 2/s$, alors pour tout $\delta > 0$, tout $a > 0$ et tout $T \geq 0$ avec probabilité au moins $1 - 2\delta$:

$$R(h_{S,T}) - \frac{a}{a-1} R_S(h_{S,T}) \leq 16L\sqrt{KL}B\sqrt{2\log(2/\delta)}\frac{\sum_{t=0}^T \mu_T}{n} + \frac{(6a+8)M\log(1/\delta)}{3n}.$$

Démonstration. C'est une application directe du théorème 3 et de la proposition 13. □

Comme dans le cas précédent on peut établir le corollaire suivant pour s'affranchir de l'hypothèse $\mu_0 \leq 2/s$.

Corollaire 6.
Dans le cadre du théorème 7 et en notant T_0 le premier indice tel que $\mu_{T_0} \leq 2/s$ p.s. Alors pour tout $\delta > 0$, tout $a > 0$ et tout $T \geq T_0$ avec probabilité au moins $1 - 2\delta$:

$$\begin{aligned}R(h_{S,T}) - \frac{a}{a-1} R_S(h_{S,T}) \leq 8LB\sqrt{2\log(2/\delta)} &\left[\frac{2L\sqrt{K}\sum_{t=0}^T \mu_T}{n} + (1-(1-1/n)^{T_0})\mathbb{E}[\delta_{T_0}|S]\right] \\ &+ \frac{(6a+8)M\log(1/\delta)}{3n}.\end{aligned}$$

Démonstration. La preuve est la même quand dans la partie précédente □

Il existe des améliorations de l'algorithme du gradient proximal qui, en particulier, convergent plus vite. On pourrait imaginer établir des résultats similaires en "randomisant" ces algorithmes comme on l'a fait ici avec l'algorithme du gradient proximal.

5 Discusion des résultats

Nous commençons par une courte comparaison entre différentes notions de stabilité -non abodées précédemment- et les résultats qui en découlent. [Zhang, 2003] a montré que des algorithmes à noyaux (avec régularisation) sont leave-one-out cross-validation stables -au sens de la définition ci-dessous- et en a déduit des bornes de généralisation en $O(\frac{1}{n^{1/2}})$ et des bornes leave-one-out en variance en $O(\frac{1}{\lambda^2 n})$.

Définition 22. Un algorithme d'apprentissage $\mathcal{A}$ est dit leave-one-out cross-validation stable, si pour tout $n \in \mathbb{N}$, il existe $\beta^{(n)} \in \mathbb{R}^+$ et $\delta^{(n)} \in (0,1)$ tels que,

$$\forall i \in \{1, \cdots, n\}, \quad \mathbb{P}_S(|\ell(h_{S^{|i}}, Z_i) - \ell(h_S, Z_i)| \leq \beta^{(n)}) \geq 1 - \delta^{(n)}$$

avec $\lim_{n\to\infty}(\beta^{(n)}, \delta^{(n)}) = (0,0)$.

De plus, [Mukherjee et al., 2006] ont montré que la leave-one-out stabilité -au sens de la défintion ci-dessous- est une condition nécessaire et suffisante pour la généralisation et la learnability, par ERM, de familles de problèmes bornés.

Définition 23. Un algorithme $\mathcal{A}$ est dit leave-one-out stable s'il est leave-one-out cross-validation stable et si pour tout $n \in \mathbb{N}$, il existe $\beta^{(n)} \in \mathbb{R}^+$ et $\delta^{(n)} \in (0,1)$ tels que,

$$\forall i \in \{1, \cdots, n\}, \quad \mathbb{P}_S(|\mathbb{E}(h_S, Z') - \frac{1}{n}\sum_{i=1}^{n} \ell(h_{S^{|i}}, Z_i)| \leq \beta^{(n)}) \geq 1 - \delta^{(n)}$$

avec $\lim_{n\to\infty}(\beta^{(n)}, \delta^{(n)}) = (0,0)$.

Quant à [Shalev-Shwartz et al., 2010], ils ont introduit la notion de stabilité au sens de remplacement-d'un-exemple-en-moyenne (On-Average-Replace-One-Stability) rappelée ci-dessous -qui est une notion plus faible que les précédentes- et ont montré qu'elle est nécessaire et suffisante pour la généralisation et la learnability (par ERM régularisé) d'une famille de problèmes convexes (comme illustré par les théorèmes ci-dessous).

Définition 24. Soit $\epsilon : \mathbb{N} \longrightarrow \mathbb{R}$ une fonction décroissante. On dit qu'un algorithme $\mathcal{A}$ est On-Average-Replace-One ϵ-stable si, pour tout $n \in \mathbb{N}$

$$\mathop{\mathbb{E}}_{i\sim Unif(1,n)}[\ell(h_{S^i}, Z_i) - \ell(h_S, Z_i)] \leq \epsilon(n).$$

Remarque :
Cette notion de stabilité est clairement plus faible que celles de la section 1.

Théorème 9. (Stabilité et généralisation)
Pour tout algorithme $\mathcal{A}$, on a

$$\mathbb{E}_S\left[\mathbb{E}_Z[\ell(h_S, Z)] - \frac{1}{n}\sum_{i=1}^{n} \ell(h_S, Z_i)\right] = \mathop{\mathbb{E}}_{i\sim Unif(1,n)}[\ell(h_{S^i}, Z_i) - \ell(h_S, Z_i)]$$

Théorème 10. (Learnability)
Supposons que la fonction de perte ℓ convexe et L-admissible. Alors, l'ERM régularisé est On-Average-Replace-One stable de rapport $\frac{2L^2}{\lambda n}$. Il s'ensuit que,

$$\mathbb{E}_S\left[\mathbb{E}_Z[\ell(h_S, Z)] - \frac{1}{n}\sum_{i=1}^{n}\ell(h_S, Z_i)\right] \leq \frac{2L^2}{\lambda n}$$

Remarque :
En particulier, la borne ci-dessus permet de majorer le risque de généralisation avec grande probabilité, par l'inégalité de Markov.

Ainsi, comme tous ces résultats conduisent à des bornes avec grande probabilité en $O(1/n^{1/2})$ (si l'on ne se restreint pas à un cadre plus particulier), nous remarquons que la notion de stabilité la plus pertinente pour obtenir des telles bornes dépend des hypothèses faites sur la famille de problèmes considérée.

A présent, nous allons discuter les résultats étudiés dans le présent travail en les comparant à d'autres rencontrés dans la littérature. Le résultat principal donné dans le théorème 1, fournit une nouvelle façon d'obtenir une borne sur le risque de généralisation. Ceci est fait grâce à l'inégalité obtenue au lemme 3, qui permet de se restreindre à l'estimation de la complexité de Rademacher d'une boule centrée en $\mathbb{E}h_S$. En termes d'hypothèses, cependant, l'espace des features $\mathcal{X}$ est supposé de Banach, avec une classe d'hypothèses $\mathcal{H}$ incluse dans le dual topologique de $\mathcal{X}$. De plus, si certaines hypothèses (bornétudes des X_i, fonction de perte Lipschitz) sont fortes, les bornes obtenues sont relativement meilleures que celles rencontrées précédemment. En effet, [Bousquet and Elisseeff, 2002] obtiennent des bornes en $O(\sqrt{\beta_1(n)})$ et $O(\sqrt{n}\beta_2(n))$ données par :

$$R(h_S) \leq R_S(h_S) + \sqrt{\frac{M^2 + 12Mn\beta_1(n)}{2n\delta}}$$

et

$$R(h_S) \leq R_S(h_S) + 2\beta(n) + (4n\beta_2(n) + M)\sqrt{\frac{\log(1/\delta)}{2n}}$$

où, M est un majorant de la fonction de perte ℓ, $\beta_1(n)$ le coefficient de stabilité et $\beta_2(n)$ celui d'uniforme stabilité. Celles-ci ayant été obtenues grâce à des inégalités de concentration [McDiarmid, 1989] et de variance [Steele, 1986].
Par ailleurs, le théorème 3 (et par conséquent, les théorèmes 4 et 5), qui majorent le risque par excès, se basent sur des techniques développées par [Bartlett et al., 2005] mais qui furent utilisées par eux pour établir des bornes dans le cadre (différent) de la classification binaire uniquement, (en supposant que la classe $\mathcal{H}$ des classifieurs soit de dimension de Vapnik-Chervonenkis finie). En outre, les résultats de [Bartlett et al., 2005] se basent sur la détermination d'un point fixe d'une fonction (sub-root [4]) majorant la complexité de Rademacher empirique locale, i.e. une fonction ψ vérifiant une inégalité de la forme :

$$\psi(r) \geq \frac{c_1}{n} + c_2\mathbb{E}_\sigma \sup_{h\in\mathcal{F}_n}\frac{1}{n}\sum_{i=1}^{n}\sigma_i h(X_i)$$

4. Une fonction $\psi : [0, +\infty) \longrightarrow [0, +\infty)$ est dite sub-root si elle est positive, croissante et que $r \longmapsto \psi(r)/\sqrt{r}$ est croissante sur $(0, +\infty)$

où

$$\mathcal{F}_n = \left\{ \alpha h \,|\quad \frac{1}{n}\sum_{i=1}^{n} h^2(X_i) \leq 2r,\ \alpha \in [0,1] \right\},$$

c_1, c_2 des constantes et $(\sigma_i)_{1\leq i\leq n}$ des variables aléatoires de Rademacher.
Les techniques étudiées ici fournissent donc, il nous semble, une façon plus directe d'obtenir des bornes sur le risque par excès.
Cela, ayant été illustré, par l'obtention de bornes sur l'erreur par excès pour l'algorithme de l'ERM régularisé ainsi que pour un ERM entraîné par *descente de gradient stochastique*, y compris dans un cadre non-convexe pour ce dernier, ce qui qui explique en partie l'excellente capacité de généralisation des réseaux de neurones. Enfin, ces résultats ont permis la construction de deux algorithmes stables pour l'ERM régularisé en absence d'hypothèses de différentiabilité, établis en randomisant *l'algorithme du gradient proximal*, permettant d'élargir les bornes sur l'erreur de généralisation à de nouveaux cas.
Parmi les problèmes intéressant qui n'ont pas été abordés, nous notons celui d'explorer s'il y a des propriétés algorithmiques autres que la stabilité, qui puissent permettre de construire une classe algorithmique d'hypothèses, dont il est possible de contrôler la complexité.

Annexes

A Intégrale de fonctions à valeurs dans un Banach

Nous rappelons dans ce paragraphe la notion d'intégrale de Bochner ou intégrale de fonctions à valeurs dans un Banach [Mikusiński, 1978], ainsi que certaine de ses propriétés, puisque le cadre d'apprentissage dans lequel nous nous plaçons est celui de sélection d'une fonction $h \in \mathcal{H} \subset \mathcal{B}$ où $\mathcal{B}$ est un Banach et qu'il sera donc nécessaire de disposer de notion d'espérance d'une variable aléatoire à valeurs dans $\mathcal{B}$.

Définition 25.
Soit (X, Σ, μ) un espace mesuré et $\mathcal{B}$ muni de sa tribu de Borel.
Si $f : X \longrightarrow \mathcal{B}$ est une fonction mesurable étagée avec $f = \sum_{i=1}^{n} a_i \mathbb{1}_{A_i}$, $a_i \in \mathcal{B}$ et $A_i \in \Sigma$, telle que $\mu(A_i)$ fini pour les $a_i \neq 0$, alors f est intégrable et son intégrale de Bochner est définie comme suit :

$$\int_X f \, \mathrm{d}\mu = \sum_{i=1}^{n} a_i \mu(A_i).$$

Ainsi, de manière générale,

Définition 26.
Une fonction mesurable $f : X \longrightarrow \mathcal{B}$ est intégrable au sens de Bochner, s'il existe une suite de fonctions mesurables étagées intégrables $(f_n)_{\mathbb{N}}$ telle que

$$\lim_{n \to \infty} \int_X \|f - f_n\|_{\mathcal{B}} \, \mathrm{d}\mu = 0$$

Dans ce cas, l'intégrale de Bochner est définie par :

$$\int_X f \, \mathrm{d}\mu = \lim_{n \to \infty} \int_X f_n \, \mathrm{d}\mu$$

Nous rappelons quelques propriétés dans la proposition suivante,

Proposition 14.

1. Une fonction mesurable $f : X \longrightarrow \mathcal{B}$ est intégrable au sens de Bochner si, et seulement si, $\int_X \|f\|_{\mathcal{B}} \, \mathrm{d}\mu < +\infty$.
2. Si $T : \mathcal{B} \longrightarrow \mathcal{B}$ est un opérateur linéaire continu, alors
$$\int_X Tf \, \mathrm{d}\mu = T \int_X f \, \mathrm{d}\mu.$$
3. *Théorème de convergence dominée*
Supposons que l'espace mesuré X soit complet.
Soit $f_n : X \longrightarrow \mathcal{B}$ une suite de foncions mesurables convegeant presque partout vers une fonction limite f, telle qu'il existe $g : X \longrightarrow \mathbb{R}_+$ intégrable vérifiant :
$$\|f_n\|_{\mathcal{B}} \leq g \quad \mu - \text{presque partout}.$$

Alors,

$$\lim_{n\to\infty} \int_X \|f - f_n\|_{\mathcal{B}} \, \mathrm{d}\mu = 0$$

d'où,

$$\lim_{n\to\infty} \int_X f_n \, \mathrm{d}\mu = \int_X f \, \mathrm{d}\mu.$$

B Démonstration du théorème de concentration

Les résultats qui suivent sont essentiellement inspirés de [Massart, 2000], [Bousquet, 2003] et [Chafai, 2005]

Définition 27. (Entropie)
Notons $\Phi : \mathbb{R}_+ \to \mathbb{R}$ définie par $\Phi(u) = u\log(u)$. Soit (Ω, A) un espace mesurable. Soient g une fonction mesurable positive sur (Ω, A) et $\mathbb{P}$ une mesure de probabilité telle que g soit intégrable par rapport à la mesure $\mathbb{P}$. On appelle entropie de g par rapport à $\mathbb{P}$ la quantité notée $\mathbb{H}_{\mathbb{P}}(g)$ définie par :

$$\mathbb{H}_{\mathbb{P}}(g) = \mathbb{E}_{\mathbb{P}}[\Phi(g)] - \Phi[\mathbb{E}_{\mathbb{P}}[g]].$$

Si il n'y a pas d'ambïguité sur la probabilité concernée, on note simplement $\mathbb{H}$ au lieu de $\mathbb{H}_{\mathbb{P}}$.

Proposition 15. (Formules variationnelles de l'entropie)
Soit $(\Omega, A, \mathbb{P})$ un espace probabilisé. Soit g une fonction mesurable positive sur (Ω, A) telle que $\Phi(g)$ soit intégrable par rapport à $\mathbb{P}$. Alors,

$$\mathbb{H}_{\mathbb{P}}(g) = \inf_{a\in\mathbb{R}_+^*} \mathbb{E}_{\mathbb{P}}[g(\log g - \log a) - (g - a)],$$

Et,

$$\mathbb{H}_{\mathbb{P}}(g) = \sup_{f \text{ mes. tq } \mathbb{E}_{\mathbb{P}}[\exp f]=1} \mathbb{E}_{\mathbb{P}}\left[gf\right] = \sup_{u} \mathbb{E}_{\mathbb{P}}\left[g \log \frac{u}{\mathbb{E}_{\mathbb{P}}[u]}\right],$$

où le supremum dans la seconde égalité est pris sur les variables aléatoires u mesurables positives sur (Ω, A) et telles que $\log \frac{u}{\mathbb{E}[u]}$ est définit.

Démonstration. Notons d'abord que si $\mathbb{E}_{\mathbb{P}}[g] = 0$, alors $g = 0$ $\mathbb{P}$-p.s. donc $\mathbb{H}_{\mathbb{P}}(g) = 0$. Supposons donc que $\mathbb{E}_{\mathbb{P}}[g] > 0$. Soit $a > 0$,

$$\mathbb{E}_{\mathbb{P}}[(\log g - \log a)g - g + a] - \mathbb{H}_{\mathbb{P}}(g) = \mathbb{E}_{\mathbb{P}}[(\log \mathbb{E}_{\mathbb{P}}[g] - \log a)g - g + a]$$

. Par l'inégalité $\log(1+x) \leq x$ vraie sur $]-1;\infty[$:

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}}[(\log \mathbb{E}_{\mathbb{P}}[g] - \log a)g - g + a] &= \mathbb{E}_{\mathbb{P}}\left[g \log \frac{\mathbb{E}_{\mathbb{P}}[g]}{a} - g + a\right] \\
&= \mathbb{E}_{\mathbb{P}}\left[g \log \frac{\mathbb{E}_{\mathbb{P}}[g]}{\mathbb{E}_{\mathbb{P}}[g] + a - \mathbb{E}_{\mathbb{P}}[g]} - g + a\right] \\
&= \mathbb{E}_{\mathbb{P}}\left[-g \log\left(1 + \frac{a - \mathbb{E}_{\mathbb{P}}[g]}{\mathbb{E}_{\mathbb{P}}[g]}\right) - g + a\right] \\
&\geq \mathbb{E}_{\mathbb{P}}\left[-g\left(\frac{a - \mathbb{E}_{\mathbb{P}}[g]}{\mathbb{E}_{\mathbb{P}}[g]}\right) - g + a\right] \\
&= 0.
\end{aligned}$$

Donc $\mathbb{E}_{\mathbb{P}}[(\log g - \log a)g - g + a] \geq \mathbb{H}_{\mathbb{P}}(g)$, on obtient l'égalité en $a = \mathbb{E}_{\mathbb{P}}[g]$.

Soit $f : \Omega \to \mathbb{R}$ mesurable telle que $\mathbb{E}_{\mathbb{P}}[\exp f] = 1$, quitte à renormaliser supposons aussi que $\mathbb{E}_{\mathbb{P}}[g] = 1$, ainsi $\mathbb{E}_{\mathbb{P}}[g] - \mathbb{E}_{\mathbb{P}}[\exp f] = 0$. Par l'inégalité $uv \leq u \log u - u + \exp(v)$, vraie sur $\mathbb{R}_+ \times \mathbb{R}$, on a $\mathbb{E}_{\mathbb{P}}[gf] \leq \mathbb{E}_{\mathbb{P}}[g \log g] = \mathbb{H}_{\mathbb{P}}(g)$. Puis l'égalité est atteinte en $f = \log g - \log \mathbb{E}_{\mathbb{P}}[g]$. La dernière égalité s'obtient en changeant la variable u par $\exp u$. □

La première formule nous permet d'établir, lorsque l'espace probabilisé est un espace produit, une majoration de l'entropie par le somme des moyennes des entropies par rapport aux lois qui constituent le produit. C'est l'objet de la proposition suivante.

Proposition 16. (Inégalité de tensorisation)
Soient $(\Omega_i, A_i, \mu_i)_{1 \leq i \leq n}$ des espaces probabilisés. Considérons l'espace probabilisé produit,

$$(\Omega, A, \mathbb{P}) := \left(\prod_{i=1}^{n} \Omega_i, \bigotimes_{i=1}^{n} A_i, \bigotimes_{i=1}^{n} \mu_i\right).$$

Soit g une fonction mesurable positive sur (Ω, A) telle que $\Phi(g)$ soit intégrable par rapport à $\mathbb{P}$. Notons, pour $x = (x_1, \ldots, x_n) \in \Omega$ et $1 \leq i \leq n$, $g_{i,x}$ la fonction définie sur Ω_i par,

$$\forall y \in \Omega_i, g_{i,x}(y) = g(x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_n).$$

Alors,

$$\mathbb{H}_{\mathbb{P}}(g) \leq \sum_{i=1}^{n} \mathbb{E}_{X \sim \mathbb{P}}\left[\mathbb{H}_{\mu_i}(g_{i,X})\right].$$

Démonstration. Par la proposition précédente,

$$\mathbb{H}_{\mathbb{P}}(g) = \sup_{f \text{ mes. tq } \mathbb{E}_{\mathbb{P}}[\exp f]=1} \mathbb{E}_{\mathbb{P}}\left[gf\right].$$

Soient $f : \Omega \to \mathbb{R}$ mesurable telle que $\mathbb{E}_{\mathbb{P}}[\exp f] = 1$ et $x \in \Omega$, définissons

$$\begin{cases} h_{1,x}(y) := f_{1,x}(y) - \log \mathbb{E}_{\mu^1}[\exp f_x^1] \text{ pour tout } y \in \Omega_1; \\ \forall i \geq 2,\ h_{i,x}(y) := \log \mathbb{E}_{\mu^{i-1}}[\exp f^{i-1}_{(x_1,..,x_{i-1},y,x_{i+1},..,x_n)}] - \log \mathbb{E}_{\mu^i}[\exp f_x^i]; \end{cases}$$

où pour tout $i \in \{1, ..., n\}$:

— $\mu^i = \bigotimes_{j=1}^i \mu_j$

— $f_x^i : (y_1, ..., y_i) \in \prod_{j=1}^i \Omega_i \mapsto f(y_1, ..., y_i, x_{i+1}, .., x_n)$

— $f_{i,x} : y \in \Omega_i \mapsto f(x_1, .., x_{i-1}, y, x_{i+1}, .., x_n)$

Alors, pour tout $i \in \{1, \dots, n\}$, $\mathbb{E}_{\mu_i}[\exp h_{i,x}] = \mathbb{E}_{\mu^i}[\exp f_x^i]/\mathbb{E}_{\mu^i}[\exp f_x^i] = 1$ et $f(x) = \sum_{i=1}^n h_{i,x}(x_i) = \sum_{i=1}^n h_{i,x} \circ \pi_i(x)$ où $\pi_i : (x_1, ..., x_n) \in \Omega \mapsto x_i$ par téléscopage et parce que $\mathbb{E}_{\mathbb{P}}[\exp f] = 1$.

Par la formule variationnelle on a donc pour tout $i \in \{1, ..., n\}$ et tout $x \in \Omega$:

$$\mathbb{E}_{\mu_i}[g_{i,x} h_{i,x}] \leq \mathbb{H}_{\mu_i}(g_{i,x}).$$

Or $\mathbb{E}_{\mathbb{P}}[gf] = \sum_{i=1}^n \mathbb{E}_{X\sim\mathbb{P}}[g(X) h_{i,X} \circ \pi_i(X)] = \sum_{i=1}^n \mathbb{E}_{X\sim\mathbb{P}}\left[\mathbb{E}_{\mu_i}[g_{i,X} h_{i,X}]\right]$, d'où :

$$\mathbb{E}_{\mathbb{P}}[gf] \leq \sum_{i=1}^n \mathbb{E}_{X\sim\mathbb{P}}\left[\mathbb{H}_{\mu_i}(g_{i,X})\right].$$

Ceci étant vrai pour tout f, on obtient une majoration du supremum de $\mathbb{E}_{\mathbb{P}}[gf]$ et donc de l'entropie par la formule variationnelle. □

Lemme 16.
Soient Z une variable aléatoire $\mathcal{A}$-mesurable et $(Z_k)_{1\leq k\leq n}$ des variables aléatoire respectivement $\mathcal{A}_k$-mesurable définies par $Z_k = F(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$, alors pour tout $\lambda \in \mathbb{R}$:

$$\lambda\mathbb{E}[Z\exp(\lambda Z)] - \mathbb{E}[\exp(\lambda Z)]\log\mathbb{E}[\exp(\lambda Z)] \leq \mathbb{E}\left[\sum_{k=1}^n \psi(\lambda(Z - Z_k))\exp(\lambda Z)\right].$$

Démonstration. Par l'inégalité de tensorisation, en considérant les espaces probabilisés $\{(\Omega, A, \mathbb{P})\}_{1\leq k\leq n}$ et les variables aléatoires $g = Z = F(X_1, \dots, X_n)$, $g_{i,x} = Z_i$ pour tout $x \in \Omega$:

$$\mathbb{H}(Z) \leq \sum_{i=1}^n \mathbb{E}\left[\mathbb{E}_i[\Phi(Z)] - \Phi(\mathbb{E}_i[Z])\right].$$

Puis, par la formule variationnelle de l'entropie, pour tout $i \in \{1, \dots, n\}$:

$$\mathbb{E}_i[\Phi(Z)] - \Phi(\mathbb{E}_i[Z]) = \inf_{U\geq 0,\ \mathcal{A}_k\text{-mesu.}} \mathbb{E}_i[Z(\log Z - \log U) - (Z - U)].$$

En particulier, soit $\lambda \in \mathbb{R}$, pour $U = \exp(\lambda Z_i)$ et en considérant $\exp(\lambda Z)$, ce qui ne change pas les résultats établis précédemment :

$$\mathbb{E}_i[\Phi(\exp(\lambda Z))] - \Phi(\mathbb{E}_i[\exp(\lambda Z)]) \leq \mathbb{E}_i[\exp(\lambda Z)(\lambda Z - \lambda Z_i) - (\exp(\lambda Z) - \exp(\lambda Z_i))].$$

Enfin, en intégrant et en utilisant Jensen parce que $\Phi : x \mapsto x\log x$ est convexe, on obtient pour tout $i \in \{1, \dots, n\}$:

$$\mathbb{E}[\Phi(\exp\lambda Z)] - \Phi\left(\mathbb{E}[\exp\lambda Z]\right) \leq \mathbb{E}\left[\exp\lambda Z\psi(-\lambda(Z - Z_i))\right].$$

On conclut en développant le terme de gauche, en sommant et en combinant les deux inégalités démontrées. □

Lemme 17. (Découplage)
Soient Z, V deux variables aléatoires $\mathcal{A}$-mesurables, alors pour tout $\lambda \in \mathbb{R}$ et $\theta > 0$:

$$\lambda \mathbb{E}[V \exp(\lambda Z)] \leq \theta \mathbb{H}[\exp(\lambda Z)] + \theta \mathbb{E}[\exp(\lambda Z)] \log \mathbb{E}[\exp(\lambda V/\theta)].$$

Démonstration. Soient V une variable aléatoire $\mathcal{A}$-mesurable, $\lambda \in \mathbb{R}$ et $\theta > 0$. On a la formule variationnelle suivante pour l'entropie :

$$\mathbb{H}[\exp(\lambda Z)] = \sup_T \mathbb{E}\left[\exp(\lambda Z) \log \frac{T}{\mathbb{E}[T]}\right],$$

où le supremum est pris sur les variables aléatoires T positives telles que $\log \frac{T}{\mathbb{E}[T]}$ est définit. Alors, $\exp(\lambda V/\theta)$ vérifie cette condition et donc :
$\mathbb{H}[\exp(\lambda Z)] \geq \mathbb{E}[\exp(\lambda Z)\lambda V/\theta] - \mathbb{E}[\exp(\lambda Z)] \log \mathbb{E}[\exp(\lambda V/\theta)]$, d'où le résultat. □

Lemme 18.
Soient Z et $(Z_1, \ldots, Z_n)$ les variables aléatoires définies au début de la section. Si V est une variable aléatoire $\mathcal{A}$-mesurable telle que $\sum_{k=1}^n Z - Z_k \leq V$, alors pour tout $\lambda > 0$:

$$\sum_{k=1}^n \mathbb{E}[\exp(\lambda Z) - \exp(\lambda Z_k)] \leq \mathbb{E}[\exp(\lambda Z)] \log \mathbb{E}[\exp(\lambda V)].$$

Démonstration. Par le lemme de découplage appliqué à Z, V et $\theta = 1$, puis par le premier lemme :

$$\begin{aligned}
&\mathbb{E}[\exp(\lambda Z)] \log \mathbb{E}[\exp(\lambda V)] - \mathbb{E}[\lambda V \exp(\lambda Z)] \geq \mathbb{E}[\exp(\lambda Z)] \log \mathbb{E}[\exp(\lambda Z)] - \mathbb{E}[\lambda Z \exp(\lambda Z)] \\
&\geq -\mathbb{E}\left[\sum_{k=1}^n \psi(\lambda(Z - Z_k)) \exp(\lambda Z)\right] \\
&= \mathbb{E}\left[\sum_{k=1}^n \exp(\lambda Z) - \exp(\lambda Z_k) - \lambda(Z - Z_k) \exp(\lambda Z)\right].
\end{aligned}$$

Or par hypothèse,

$$\mathbb{E}[\lambda V \exp(\lambda Z)] \geq \mathbb{E}\left[\sum_{k=1}^n \lambda(Z - Z_k) \exp(\lambda Z)\right],$$

d'où le résultat. □

Lemme 19.
Pour tout $\lambda \geq 0$, $k \in \{1, \ldots, n\}$:

$$\psi(\lambda(Z - Z_k)) \exp(\lambda Z) \leq \frac{\phi(-\lambda)}{\psi(-\lambda) + \frac{\lambda}{1+u}} \left(\exp(\lambda Z) - \exp(\lambda Z_k) + \lambda \exp(\lambda Z_k) \left(\frac{(Z_k')^2}{1+u} - Z_k'\right)\right).$$

Démonstration. Une analyse de ϕ et ψ permet de montrer le résultat. Nous ne la ferons pas dans ce mémoire, pour le lecteur intéressé, des détails se trouvent dans [Bousquet, 2003] □

Lemme 20.
Pour toute variable aléatoire U $\mathcal{A}$-mesurable positive, tout $\lambda \geq 0$ et tout $k \in \{1, \ldots, n\}$:

$$\mathbb{E}[\exp(\lambda Z_k)U] \leq \mathbb{E}[\exp(\lambda Z)\mathbb{E}_k[U]].$$

Démonstration. Soient $\lambda \geq 0$ et $k \in \{1, \ldots, n\}$, alors par hypothèse $Z'_k \geq Z - Z_k \geq 1$. Donc $\mathbb{E}_k[Z'_k] \leq \mathbb{E}_k[Z] - Z_k$ car Z_k est $\mathcal{A}_k$-mesurable. Donc par croissance de $x \mapsto \exp(\lambda x)$,

$$\begin{aligned}
\mathbb{E}[\exp(\lambda Z_k)U] &\leq \mathbb{E}[\exp(\lambda \mathbb{E}_k[Z]) \exp(-\lambda \mathbb{E}_k[Z'_k])U] \\
&\leq \mathbb{E}[\exp(\lambda \mathbb{E}_k[Z])U] \text{ car } \mathbb{E}_k[Z'_k] \geq 0 \\
&= \mathbb{E}\ \mathbb{E}_k[\exp(\lambda \mathbb{E}_k[Z])U] \\
&= \mathbb{E}[\exp(\lambda \mathbb{E}_k[Z])\mathbb{E}_k[U]] \\
&\leq \mathbb{E}[\mathbb{E}_k[\exp(\lambda Z)]\mathbb{E}_k[U]] \text{ par Jensen} \\
&= \mathbb{E}[\mathbb{E}_k[\exp(\lambda Z)\mathbb{E}_k[U]]] \\
&= \mathbb{E}[\exp(\lambda Z)\mathbb{E}_k[U]]
\end{aligned}$$

□

Démonstration du théorème. Grâce aux lemmes établis, on va construire une inégalité différentielle vérifiée par $F(\lambda) := \mathbb{E}[\exp(\lambda Z)]$, la transformée de Laplace de Z.

D'après les lemme 3, 4 et 5 :

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^{n} \psi(\lambda(Z - Z_k)) \exp(\lambda Z)\right] \\
&\leq \frac{\phi(-\lambda)}{\psi(-\lambda) + \frac{\lambda}{1+u}} \sum_{k=1}^{n} \mathbb{E}\left[\exp(\lambda Z) - \exp(\lambda Z_k) + \lambda \exp(\lambda Z_k)\left(\frac{(Z'_k)^2}{1+u} - Z'_k\right)\right] \\
&\leq \frac{\phi(-\lambda)}{\psi(-\lambda) + \frac{\lambda}{1+u}} \sum_{k=1}^{n} \mathbb{E}\left[\exp(\lambda Z) - \exp(\lambda Z_k) + \lambda \exp(\lambda Z_k)\frac{(Z'_k)^2}{1+u}\right] \\
&\leq \frac{\phi(-\lambda)}{\psi(-\lambda) + \frac{\lambda}{1+u}} \left(\mathbb{E}[\exp(\lambda Z)] \log \mathbb{E}[\exp(\lambda Z)] + \frac{\lambda}{1+u} \sum_{k=1}^{n} \mathbb{E}\left[\exp(\lambda Z_k)(Z'_k)^2\right]\right) \\
&\leq \frac{\phi(-\lambda)}{\psi(-\lambda) + \frac{\lambda}{1+u}} \left(\mathbb{E}[\exp(\lambda Z)] \log \mathbb{E}[\exp(\lambda Z)] + \frac{\lambda}{1+u} \sum_{k=1}^{n} \mathbb{E}\left[\exp(\lambda Z_k)\mathbb{E}_k[(Z'_k)^2]\right]\right) \\
&\leq \frac{\phi(-\lambda)}{\psi(-\lambda) + \frac{\lambda}{1+u}} \left(\mathbb{E}[\exp(\lambda Z)] \log \mathbb{E}[\exp(\lambda Z)] + \frac{n\sigma^2\lambda}{1+u}\mathbb{E}\left[\exp(\lambda Z)\right]\right)
\end{aligned}$$

Et par le lemme 1, on peut minorer $\mathbb{E}\left[\sum_{k=1}^{n} \psi(\lambda(Z - Z_k)) \exp(\lambda Z)\right]$ pour finalement obtenir l'inéquation :

$$\lambda F'(\lambda) - F(\lambda) \log F(\lambda) \leq \frac{1 - \exp(\lambda) + \lambda \exp(\lambda)}{\exp(\lambda) - 1 - \lambda + \frac{\lambda}{1+u}} \left(F(\lambda) \log F(\lambda) + \frac{n\sigma^2\lambda}{1+u} F(\lambda)\right).$$

On peut facilement vérifier que $\lambda \mapsto n\sigma^2(\exp(\lambda) - 1 - \lambda)$ est solution du problème d'inconnue L :

$$\begin{cases} \lambda \exp(L)' - \exp(L)L = \frac{1-\exp(\lambda)+\lambda\exp(\lambda)}{\exp(\lambda)-1-\lambda+\frac{\lambda}{1+u}} \left(\exp(L)L + \frac{n\sigma^2\lambda}{1+u}\exp(L)\right) \\ \ell(0) = 0, L'(0) = 0 \end{cases}$$

et donc cette solution majore $\log F$.

Donc,

$$\forall \lambda \geq 0 \log F(\lambda) \leq n\sigma^2\psi(-\lambda).$$

Ensuite,

$$\begin{aligned} \log \mathbb{E}\left[\exp(\lambda(Z - \mathbb{E}[Z]))\right] &= \log \mathbb{E}\left[\exp(\lambda Z)\right] - \lambda\mathbb{E}[Z] \\ &\leq (n\sigma^2 + (1+u)\mathbb{E}[Z])\psi(-\lambda), \end{aligned}$$

car $\mathbb{E}[Z] \geq \sum_{k=1}^n \mathbb{E}[Z - Z_k] \geq \sum_{k=1}^n \mathbb{E}[Z'_k] \geq \sum_{k=1}^n \mathbb{E}[\mathbb{E}_k[Z'_k]] \geq 0$ et ψ est positive. □

Références

[Bartlett et al., 2005] Bartlett, P. L., Bousquet, O., Mendelson, S., et al. (2005). Local rademacher complexities. *The Annals of Statistics*, 33(4) :1497–1537.

[Bartlett and Mendelson, 2002] Bartlett, P. L. and Mendelson, S. (2002). Rademacher and gaussian complexities : Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov) :463–482.

[Bousquet, 2001] Bousquet, O. (2001). A bennett concentration inequality and its application to suprema of empirical processes. *C.R. Acad. Sci. Paris*, 332.

[Bousquet, 2003] Bousquet, O. (2003). Concentration inequalities for sub-additive functions using the entropy method. *Progress in Probability*, 56 :213–248.

[Bousquet and Elisseeff, 2002] Bousquet, O. and Elisseeff, A. (2002). Stability and generalization. *The Journal of Machine Learning Research*, 2 :499–526.

[Chafai, 2005] Chafai, D. (2005). Inégalités de poincaré et de gross pour les mesures de bernoulli, de poisson, et de gauss.

[Hardt et al., 2016] Hardt, M., Recht, B., and Singer, Y. (2016). Train faster, generalize better : Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234. PMLR.

[Liu et al., 2017] Liu, T., Lugosi, G., Neu, G., and Tao, D. (2017). Algorithmic stability and hypothesis complexity. In *International Conference on Machine Learning*, pages 2159–2167. PMLR.

[Massart, 2000] Massart, P. (2000). About the constants in talagrand's concentration inequalities for empirical processes. *The Annals of Probability*, 28(2) :863–884.

[Mikusiński, 1978] Mikusiński, J. (1978). The bochner integral. In *The Bochner Integral*, pages 15–22. Springer.

[Mukherjee et al., 2006] Mukherjee, S., Niyogi, P., Poggio, T., and Rifkin, R. (2006). Learning theory : stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, 25(1) :161–193.

[Pinelis, 1994] Pinelis, I. (1994). Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, 22(4) :1679–1706.

[Shalev-Shwartz and Ben-David, 2014] Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning : From theory to algorithms*. Cambridge university press.

[Shalev-Shwartz et al., 2010] Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. (2010). Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11 :2635–2670.

[Wibisono et al., 2009] Wibisono, A., Rosasco, L., and Poggio, T. (2009). Sufficient conditions for uniform stability of regularization algorithms. *Computer Science and Artificial Intelligence Laboratory Technical Report, MIT-CSAIL-TR-2009-060*.

[Zhang, 2003] Zhang, T. (2003). Leave-one-out bounds for kernel methods. *Neural Computation*, 15(6) :1397–1437.