



CentraleSupélec

Introduction à la théorie de l'apprentissage statistique

(notes de cours)

Michel BARRET

2020/2021

Introduction à la théorie de l'apprentissage statistique

Michel Barret

2 mars 2020

On s'intéresse aux problèmes rencontrés en apprentissage supervisé.

Pour un ensemble $\mathcal{X}$, et $m \in \mathbb{N}$, $m \neq 0$, on note $|\mathcal{X}|$ son cardinal, $\mathcal{X}^m$ l'ensemble des m -uplets $(x_1, \dots, x_m)$ dont chaque terme est un élément de $\mathcal{X}$, on pose $\mathcal{X}^* = \bigcup_{m=1}^{\infty} \mathcal{X}^m$. Pour $\mathcal{Y}$ un ensemble discret, on note δ le symbole de Kronecker défini sur $\mathcal{Y}^2$ par $\delta_{y,y'} = 1$ si $y = y'$ et 0 si $y \neq y'$. Pour $(\mathcal{X}, \mathcal{M}_{\mathcal{X}})$ et $(\mathcal{Y}, \mathcal{M}_{\mathcal{Y}})$ deux espaces mesurables, on note $\mathcal{M}(\mathcal{X}, \mathcal{Y})$ l'ensemble des fonctions mesurables de $\mathcal{X}$ dans $\mathcal{Y}$. On note $(\Omega, \mathcal{E}, \mathbb{P})$ un espace probabilisé qui modélise l'expérience aléatoire sous-jacente à toutes les variables aléatoires introduites dans la suite.

1 Modèle formel de l'apprentissage statistique supervisé

Un algorithme d'apprentissage supervisé utilise les concepts suivants :

- un espace mesurable $(\mathcal{X}, \mathcal{M}_{\mathcal{X}})$, dont les éléments $x \in \mathcal{X}$ sont des objets à partir desquels une décision est prise ;
- un espace mesurable $(\mathcal{Y}, \mathcal{M}_{\mathcal{Y}})$, dont les éléments $y \in \mathcal{Y}$ sont les résultats possibles de la décision ;
- une séquence finie $S = ((x_1, y_1), \dots, (x_m, y_m)) \in (\mathcal{X} \times \mathcal{Y})^*$ d'objets labélisés, c'est-à-dire d'objet x_i et de l'étiquette (ou décision) associée, appelée *données d'entraînement*.

Un algorithme d'apprentissage donne en sortie une *hypothèse* (ou *prédicteur*) $h \in \mathcal{M}(\mathcal{X}, \mathcal{Y})$, appelée également un *classificateur* quand $\mathcal{Y}$ est discret ¹ ou un *estimateur* quand $\mathcal{Y}$ est continu. Dans le cas supervisé, il admet pour entrée une séquence finie S ci-dessus et s'écrit formellement

$$\begin{aligned} \mathbf{A} : (\mathcal{X} \times \mathcal{Y})^* &\rightarrow \mathcal{M}(\mathcal{X}, \mathcal{Y}) \\ S &\mapsto \mathbf{A}(S). \end{aligned}$$

L'objectif d'un algorithme d'apprentissage supervisé est de fournir un “bon” (voire le meilleur) estimateur ou classificateur de la théorie de la décision dans un cadre statistique, c'est-à-dire sans connaître la ou les vraisemblances, mais à partir des observations $(x_1, y_1), \dots, (x_m, y_m)$ et sous l'hypothèse 1 suivante.

Hypothèse 1 *Les données d'entraînement sont des réalisations de variables aléatoires indépendantes et identiquement distribuées (i.i.d.).*

1.1 Un démarrage en douceur

Considérons un problème de classification binaire, qui consiste à décider si un objet vérifie une propriété particulière ou non, par exemple l'objet est une photo et la propriété “une voiture est visible sur la photo”. De façon formelle, on peut écrire $\mathcal{X} \subset \mathbb{R}^d$, où d est le nombre de pixels de la photo, $\mathcal{Y} = \{0, 1\}$ ($y = 1$ si et seulement si une voiture est visible sur la photo) et on dispose des données d'entraînement $S = ((x_1, y_1), \dots, (x_m, y_m))$, i.e., m réalisations indépendantes d'un couple de variables aléatoires (x, y) à valeurs dans $\mathcal{X} \times \mathcal{Y}$; dans notre exemple : m photos x_i , avec pour chacune d'elle la réponse y_i à la question :

1. C'est-à-dire quand le nombre d'éléments est fini ou infini dénombrable.

« Une voiture est-elle visible sur la photo ? »

On suppose dans cette sous-section qu'il existe une fonction $f \in \mathcal{M}(\mathcal{X}, \mathcal{Y})$ telle que $y = f(x)$.

À toute hypothèse $h \in \mathcal{M}(\mathcal{X}, \mathcal{Y})$, on associe son *risque de généralisation* ou *risque réel* ou simplement *risque*

$$R(h) = \mathbb{P}[h(x) \neq y] = \mathbb{E}[\mathbb{1}_{h(x) \neq y}] \quad (1)$$

et on cherche une hypothèse h de risque réel minimum. Comme on ne connaît pas les vraisemblances (i.e., les lois de probabilité), on ne peut pas calculer ce risque. On introduit alors le *risque empirique* de $h \in \mathcal{M}(\mathcal{X}, \mathcal{Y})$ défini par

$$R_S(h) = \frac{1}{m} \sum_{i=1}^m \mathbb{1}_{h(x_i) \neq y_i} \quad (2)$$

qui, d'après la loi forte des grands nombres (les échantillons (x_i, y_i) étant i.i.d.), converge presque sûrement (p.s.) et en moyenne quadratique (m.q.) vers le risque de généralisation quand m tend vers l'infini. On cherche l'hypothèse h_S qui minimise le risque empirique. C'est la règle du *risque empirique minimum* (ERM).

Mais si on n'applique pas de contrainte à l'hypothèse qui minimise le risque empirique, on obtient la plupart du temps une hypothèse h_S qui admet un fort risque de généralisation, c'est le problème de *sur-apprentissage*, i.e., $R_S(h_S) \simeq 0$ et $R(h_S) \gg 0$. Pour l'éviter on impose des contraintes aux hypothèses retournées par l'algorithme d'apprentissage. Formellement on introduit un sous-ensemble $\mathcal{H} \subset \mathcal{M}(\mathcal{X}, \mathcal{Y})$ et on cherche une hypothèse dans $\mathcal{H}$ qui minimise le risque empirique²

$$h_S = \text{ERM}_{\mathcal{H}}(S) \in \arg \min_{h \in \mathcal{H}} R_S(h). \quad (3)$$

Il est important de noter que h_S est aléatoire, car c'est une fonction des données d'entraînement qui sont aléatoires. Introduisons l'*hypothèse de prédiction singulière*³ (*realizability assumption*)

Hypothèse 2 *Il existe $h^* \in \mathcal{H}$ tel que $R(h^*) = 0$.*

Elle entraîne que $h^*(x) = f(x)$ p.s., donc que $R_S(h_S) = 0$ p.s. En général h^* n'est pas la seule hypothèse qui minimise le risque empirique et un algorithme d'apprentissage retourne une hypothèse $h_S \neq h^*$, qui peut avoir un risque de généralisation important.

Pour étudier précisément cette situation, considérons un écart $\varepsilon \in]0; 1[$ et calculons la probabilité de l'événement $R(h_S) > \varepsilon$. Introduisons le sous-ensemble des hypothèses défavorables $\mathcal{H}_d = \{h \in \mathcal{H}, R(h) > \varepsilon\}$ et l'ensemble des données d'entraînement défavorables $\{S : h_S \in \mathcal{H}_d\} \subset \{S : \exists h \in \mathcal{H}_d, R_S(h) = 0 \text{ p.s.}\} = \bigcup_{h \in \mathcal{H}_d} \{S : R_S(h) = 0 \text{ p.s.}\}$; pour $h \in \mathcal{H}_d$, on a $\mathbb{P}[R_S(h) = 0] = (\mathbb{P}[h(x) = y])^m = [1 - R(h)]^m \leq (1 - \varepsilon)^m \leq e^{-\varepsilon m}$. Donc, si l'espace des hypothèses $\mathcal{H}$ est fini, il vient

$$\mathbb{P}[R(h_S) > \varepsilon] \leq |\mathcal{H}_d| e^{-m\varepsilon} \leq |\mathcal{H}| e^{-m\varepsilon}. \quad (4)$$

Nous avons montré le résultat suivant.

Proposition 1.1 *Soit $\mathcal{H}$ un ensemble fini d'hypothèses. Pour tous $\delta, \varepsilon \in]0; 1[$ et tout entier m qui satisfait*

$$m \geq \frac{\log(|\mathcal{H}|/\delta)}{\varepsilon},$$

pour toute loi de probabilité de x , si $\mathcal{H}$ vérifie l'hypothèse 2 de prédiction singulière, alors avec une probabilité d'au moins $1 - \delta$, $R(h_S) \leq \varepsilon$.

Autrement dit, pour m assez grand, la règle qui consiste à minimiser le risque empirique est probablement (avec une confiance de $1 - \delta$) approximativement (à une erreur ε près) correcte (PAC).

2. Dans ce chapitre, nous utilisons la notation $\min$ dans les expressions mathématiques sans distinguer les notions de minimum et de borne inférieure.

3. En effet, dans ce cas, l'estimation ou la détection est singulière car le risque moyen s'annule.

1.2 Modèle général de l'apprentissage supervisé

Considérons un problème d'estimation ou de classification général en introduisant une fonction coût⁴ $L : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_+$ (voire à valeurs dans $\mathbb{R}$) qui, pour une hypothèse h et un couple $(x, y) \in \mathcal{X} \times \mathcal{Y}$, modélise par $L[h(x), y]$ le coût de la décision $h(x)$ quand la vraie valeur est y . Pour une hypothèse h , les risques, de généralisation et empirique (1-2), deviennent

$$R(h) = \mathbb{E}(L[h(x), y]) \quad (5)$$

$$R_S(h) = \frac{1}{m} \sum_{i=1}^m L[h(x_i), y_i]. \quad (6)$$

L'algorithme d'apprentissage $\text{ERM}_{\mathcal{H}}$ qui minimise le risque empirique reste défini par l'équation (3).

1.3 Capacité d'apprentissage PAC et convergence uniforme

Définition 1.1 *Un ensemble d'hypothèses $\mathcal{H}$ a la capacité d'apprentissage PAC quand il existe une application $m_{\mathcal{H}} :]0; 1]^2 \rightarrow \mathbb{N}$ et un algorithme d'apprentissage ayant la propriété : pour tout $(\varepsilon, \delta) \in]0; 1]^2$ et toute loi de probabilité jointe de (x, y) , la sortie h_S de l'algorithme d'apprentissage appliqué à $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ réalisations i.i.d. de (x, y) vérifie avec une probabilité d'au moins $1 - \delta$ que*

$$R(h_S) \leq \min_{h \in \mathcal{H}} R(h) + \varepsilon.$$

Autrement dit, la probabilité pour que l'excès de risque de h_S (qui vaut $R(h_S) - \min_{h \in \mathcal{H}} R(h)$ par définition) soit supérieur à ε est inférieure à δ .

Pour un ensemble d'hypothèses qui a la capacité d'apprentissage PAC, la fonction $m_{\mathcal{H}}$ minimale, c'est-à-dire telle que pour tout (ε, δ) $m_{\mathcal{H}}(\varepsilon, \delta)$ soit le plus petit entier qui vérifie la propriété de la définition 1.1, détermine la *complexité* (*mémoire* ou *d'échantillon*) de $\mathcal{H}$.

Les auteurs de [2] introduisent une deuxième notion de capacité d'apprentissage PAC plus faible (apparemment), en restreignant la définition 1.1 à toute loi de probabilité vérifiant l'hypothèse 2 d'une prédiction singulière. Nous verrons avec le premier théorème fondamental de l'apprentissage, page 11, que ces deux définitions sont équivalentes, et, sans toujours le justifier, nous utiliserons la notion plus restrictive dans certaines preuves.

Définition 1.2 *Soit $\varepsilon > 0$. Les données d'entraînement S forment un échantillon ε -représentatif (pour un ensemble d'hypothèses $\mathcal{H}$, une fonction coût L et une loi de probabilité) quand*

$$\text{pour tout } h \in \mathcal{H}, \quad |R_S(h) - R(h)| \leq \varepsilon. \quad (7)$$

Lemme 1.2 *Soit S un échantillon $\varepsilon/2$ -représentatif (pour un ensemble d'hypothèses $\mathcal{H}$, une fonction coût L et une loi de probabilité). Alors toute hypothèse $h_S \in \arg \min_{h \in \mathcal{H}} R_S(h)$ vérifie*

$$R(h_S) \leq \min_{h \in \mathcal{H}} R(h) + \varepsilon.$$

Preuve. Pour tout $h \in \mathcal{H}$,

$$R(h_S) \leq R_S(h_S) + \frac{\varepsilon}{2} \leq R_S(h) + \frac{\varepsilon}{2} \leq R(h) + \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = R(h) + \varepsilon.$$

Les première et troisième inégalités proviennent de la $\varepsilon/2$ -représentativité de S , et la deuxième de la définition de h_S . $\diamond$

4. En général, dans les cours d'apprentissage statistique, la fonction coût est définie par

$$\begin{aligned} \ell : \mathcal{H} \times \mathcal{Z} &\rightarrow \mathbb{R}_+ \quad (\text{voire à valeurs dans } \mathbb{R}) \\ (h, z) &\mapsto \ell(h, z) = L[h(x), y] \quad \text{où } z = (x, y), \end{aligned}$$

$\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ et $L : \mathcal{Y}^2 \rightarrow \mathbb{R}_+$ (voire à valeurs dans $\mathbb{R}$) est une *fonction coût* selon la définition du cours de Traitement statistique du signal.

Définition 1.3 Un ensemble d'hypothèses $\mathcal{H}$ a la propriété de convergence uniforme (par rapport à $\mathcal{X} \times \mathcal{Y}$ et à une fonction coût) quand il existe une fonction $m_{\mathcal{H}}^{\text{CU}} :]0; 1]^2 \rightarrow \mathbb{N}$ telle que pour tous $\varepsilon, \delta \in]0, 1[$ et pour toute loi de probabilité sur $\mathcal{X} \times \mathcal{Y}$, si S est un échantillon de $m \geq m_{\mathcal{H}}^{\text{CU}}(\varepsilon, \delta)$ tirages i.i.d. suivant cette loi, alors avec une probabilité d'au moins $1 - \delta$, S est ε -représentatif.

Comme pour la définition de la complexité de $\mathcal{H}$, la fonction $m_{\mathcal{H}}^{\text{CU}}$ mesure la complexité minimale de l'échantillon assurant la propriété de convergence uniforme. Le terme *uniforme* indique ici une taille d'échantillon minimale, valable pour toute hypothèse $h \in \mathcal{H}$, indépendamment de la loi de probabilité.

Nous déduisons du lemme 1.2 la proposition suivante.

Proposition 1.3 Si un ensemble d'hypothèses $\mathcal{H}$ a la propriété de convergence uniforme avec une fonction $m_{\mathcal{H}}^{\text{CU}}$, alors $\mathcal{H}$ a la capacité PAC d'apprendre avec une complexité $m_{\mathcal{H}}(\varepsilon, \delta) \leq m_{\mathcal{H}}^{\text{CU}}(\varepsilon/2, \delta)$. De plus, dans ce cas la règle de minimisation du risque empirique donne un algorithme d'apprentissage PAC pour $\mathcal{H}$.

1.4 Capacité d'apprentissage PAC des ensembles d'hypothèses finis

Nous allons montrer que les ensembles d'hypothèses finis ont la propriété de convergence uniforme. Nous commençons par rappeler l'inégalité de Hoeffding, dont une preuve est donnée en exercice.

Lemme 1.4 Soient $a < b$ deux nombres réels et $x_1, \dots, x_m$ une séquence i.i.d. de VA telles que $\mathbb{E}[x_i] = \mu \in \mathbb{R}$ et $\mathbb{P}[a \leq x_i \leq b] = 1$. Alors pour tout $\varepsilon > 0$,

$$\mathbb{P} \left[\left| \frac{1}{m} \sum_{i=1}^m x_i - \mu \right| > \varepsilon \right] \leq 2 \exp(-2m\varepsilon^2/(b-a)^2).$$

Fixons $\varepsilon, \delta \in]0, 1[$. Nous recherchons une taille m d'échantillon qui garantit que pour toute loi de probabilité de (x, y) , avec une probabilité d'au moins $1 - \delta$, l'échantillon $S = ((x_1, y_1), \dots, (x_m, y_m))$ est ε -représentatif, i.e., pour tout $h \in \mathcal{H}$, $|R_S(h) - R(h)| \leq \varepsilon$. Considérons l'événement complémentaire

$$\{S : \exists h \in \mathcal{H}, |R_S(h) - R(h)| > \varepsilon\} = \bigcup_{h \in \mathcal{H}} \{S : |R_S(h) - R(h)| > \varepsilon\}.$$

Remarquons maintenant que pour $h \in \mathcal{H}$, $\mathbb{E}[R_S(h)] = R(h)$ et en supposant que la fonction coût est à valeurs dans $[0; 1]$, il résulte de l'inégalité de Hoeffding que

$$\mathbb{P}[\{S : \exists h \in \mathcal{H}, |R_S(h) - R(h)| > \varepsilon\}] \leq \sum_{h \in \mathcal{H}} 2 \exp(-2m\varepsilon^2) = 2|\mathcal{H}| \exp(-2m\varepsilon^2),$$

ainsi si $m \geq \frac{\log(2|\mathcal{H}|/\delta)}{2\varepsilon^2}$, alors S est ε -représentatif. Nous avons démontré la proposition suivante.

Proposition 1.5 Soit $\mathcal{H}$ un ensemble fini d'hypothèses et L une fonction coût à valeurs dans $[0; 1]$. Alors $\mathcal{H}$ a la propriété de convergence uniforme avec la complexité

$$m_{\mathcal{H}}^{\text{CU}}(\varepsilon, \delta) \leq \left\lceil \frac{\log(2|\mathcal{H}|/\delta)}{2\varepsilon^2} \right\rceil,$$

de plus $\mathcal{H}$ a la capacité d'apprentissage PAC en utilisant l'algorithme $\text{ERM}_{\mathcal{H}}$ avec la complexité

$$m_{\mathcal{H}}(\varepsilon, \delta) \leq m_{\mathcal{H}}^{\text{CU}}(\varepsilon/2, \delta) \leq \left\lceil \frac{2 \log(2|\mathcal{H}|/\delta)}{\varepsilon^2} \right\rceil.$$

Remarque 1.1 Quand l'espace des hypothèses a la puissance du continu, c'est-à-dire quand chaque hypothèse h_θ dépend de k variables réelles regroupées dans un vecteur $\theta \in \Theta \subset \mathbb{R}^k$, où Θ est borné, si l'on recherche une hypothèse qui minimise le risque empirique avec un ordinateur, les valeurs réelles que l'on peut coder étant en nombre N fini, par exemple si elles sont codées sur 64 bits $N = 2^{64}$, alors $|\mathcal{H}| \simeq N^k$ et la complexité $m_{\mathcal{H}}(\varepsilon, \delta)$ est majorée par $\left\lceil \frac{2k \log(N) + 2 \log(2/\delta)}{\varepsilon^2} \right\rceil$. Une autre façon d'obtenir le même résultat est de considérer que $\Theta = [0, 1]^k$ (ou plus généralement $\Theta = [a, b]^k$ avec $a, b \in \mathbb{R}$) et que chaque composante est uniformément échantillonnée sur N valeurs, alors $|\mathcal{H}| \simeq N^k$.

Exemple 1.1 Dans un problème de classification, l'algorithme des k plus proches voisins, avec $k = 1$ et m données d'entraînement $(\mathbf{x}_i, y_i)$ ($1 \leq i \leq m$), où $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^d$, donne un espace d'hypothèses qui dépend de md paramètres réels : les points $\mathbf{x}_1, \dots, \mathbf{x}_m$. Il est clair qu'avec la discrétisation de la remarque 1.1, on a $\frac{\log(|\mathcal{H}|)}{m} \geq 1$, donc l'excès de risque est majoré par une valeur supérieure à 1 et nous verrons avec le 2ème théorème fondamental de l'apprentissage page 12 que pour le problème de classification binaire, il n'est pas possible d'améliorer cette majoration de l'erreur d'estimation. Autrement dit, l'algorithme du plus proche voisin a trop de degrés de liberté pour contrôler l'excès de risque.

1.5 Le dilemme biais-complexité

Nous pouvons décomposer l'erreur du prédicteur $h_S = \text{ERM}_{\mathcal{H}}(S)$ en la somme de deux termes :

$$R(h_S) = \epsilon_{\text{app}} + \epsilon_{\text{est}} \quad \text{avec} \quad \epsilon_{\text{app}} = \min_{h \in \mathcal{H}} R(h) \text{ et } \epsilon_{\text{est}} = R(h_S) - \min_{h \in \mathcal{H}} R(h). \quad (8)$$

Le premier terme est l'erreur d'approximation ou le biais dû aux contraintes de $\mathcal{H}$ et le deuxième terme est l'excès de risque ou l'erreur de fluctuation ou encore l'erreur d'estimation due au fait que le risque empirique n'est qu'une estimation du risque réel. L'erreur d'estimation dépend

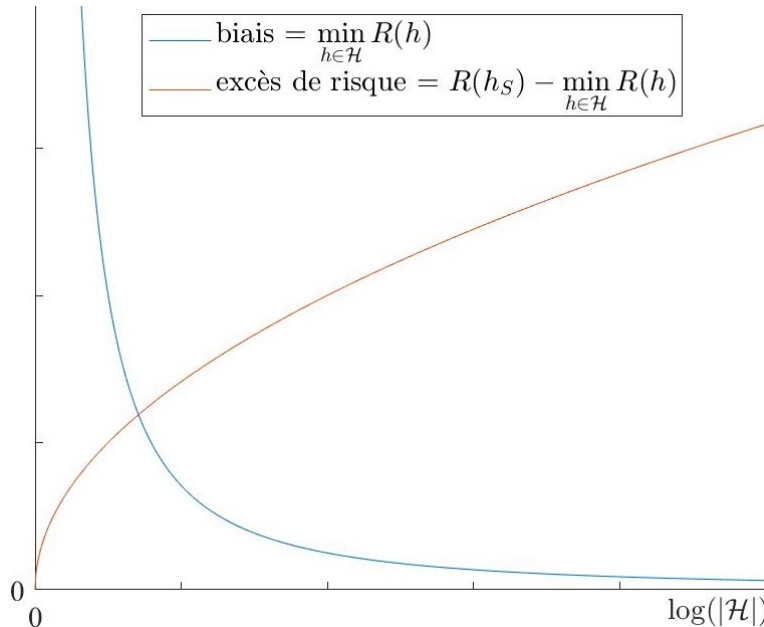


FIGURE 1 – Allure des courbes de biais et d'excès de risque en fonction de la complexité de $\mathcal{H}$, pour m fixé.

de la taille m de l'échantillon et de la complexité de $\mathcal{H}$. Nous avons vu que pour un ensemble d'hypothèses fini, ϵ_{est} croît (logarithmiquement) avec $|\mathcal{H}|$ et décroît avec m . Nous pouvons considérer l'entropie de $\mathcal{H}$ ($\log(|\mathcal{H}|)$) comme une mesure de la complexité de $\mathcal{H}$.

Puisque notre objectif est de minimiser le risque de généralisation $R(h_S)$, il faut trouver le bon compromis entre le biais et la complexité de $\mathcal{H}$. Pour réduire le biais, il faut augmenter la complexité, mais attention au sur-apprentissage. De même, si l'on réduit trop la complexité de $\mathcal{H}$, le terme de biais devient prépondérant, c'est le problème du *sous-apprentissage*. Tout l'art du *data scientist* est d'exploiter l'information *a priori* qu'il a du problème de décision pour choisir un bon ensemble d'hypothèses, bien adapté à la taille des données d'entraînement dont il dispose.

D'après la proposition 1.5, l'excès de risque $R(h_S) - \min_{h \in \mathcal{H}} R(h)$ est majoré par 2ε dès que $2m\varepsilon^2 \geq \log(|\mathcal{H}|) + \log(2/\delta)$; ainsi en écrivant $\log(2/\delta) = m\varepsilon^2$ et $\log(|\mathcal{H}|) = m\varepsilon^2$ et en supposant qu'il existe des constantes $C > 0$ et $\alpha > 0$ telles que l'erreur d'approximation soit majorée par $C \log(|\mathcal{H}|)^{-\alpha}$, on trouve que le risque réel $R(h_S)$ est majoré par 3ε avec une probabilité d'au moins $1 - \delta \geq 1 - 2e^{-(C/\varepsilon)^{1/\alpha}}$ si $C(m\varepsilon^2)^{-\alpha} \leq \varepsilon$, i.e., si $m \geq C^{1/\alpha} \varepsilon^{-2-1/\alpha}$, autrement dit nous avons démontré la proposition suivante de Stéphane Mallat [1].

Proposition 1.6 *S'il existe des constantes $C > 0$ et $\alpha > 0$ telles que $\min_{h \in \mathcal{H}} R(h) \leq C(\log |\mathcal{H}|)^{-\alpha}$, alors $\mathbb{P}[R(h_S) \leq 3\varepsilon] \geq 1 - 2e^{-(C/\varepsilon)^{1/\alpha}}$ si $m \geq C^{1/\alpha} \varepsilon^{-2-1/\alpha}$.*

Ainsi, si α est petit, alors le nombre d'échantillons doit être très grand, de l'ordre de $\varepsilon^{-1/\alpha}$, pour garantir un risque réel probablement de l'ordre de ε .

Si $\mathcal{Y} = \mathbb{R}$ et $L(y, y') = |y - y'|^p$ (où $p \geq 1$), et s'il existe une fonction $f \in \mathcal{M}(\mathcal{X}, \mathcal{Y})$ telle que $y = f(x)$, alors $R(h) = \mathbb{E}[|h(X) - f(X)|^p] \leq \sup_{x \in \mathcal{X}} |h(x) - f(x)|^p = \|h - f\|_\infty^p$ et l'erreur d'approximation satisfait la condition

$$(\epsilon_{\text{app}})^{\frac{1}{p}} = \min_{h \in \mathcal{H}} \|h - f\|_\infty. \quad (9)$$

On tombe sur le problème d'approximation de la fonction f par une fonction $h \in \mathcal{H}$, à la différence que f est inconnue ici. Supposons que nous ayons quelque *information a priori* sur le problème, à savoir que $f \in \mathcal{C}$, où $\mathcal{C} \subset \mathcal{M}(\mathcal{X}, \mathcal{Y})$ est une classe de fonctions connue, on peut alors majorer l'inégalité (9) par une borne qui ne dépend plus de f :

$$\min_{h \in \mathcal{H}} \|h - f\|_\infty \leq \max_{f \in \mathcal{C}} \min_{h \in \mathcal{H}} \|h - f\|_\infty. \quad (10)$$

Supposons enfin que le membre de droite de l'inégalité (10) soit majoré par $C(\log |\mathcal{H}|)^{-\alpha}$, avec α grand, alors on peut appliquer la proposition 1.6 ci-dessus et affirmer que le risque réel est probablement approximativement petit pour un nombre d'échantillons m pas trop grand.

Comment traduire l'information *a priori* que l'on a sur f dans des classes $\mathcal{C}$ de fonctions? Stéphane Mallat propose d'utiliser la notion de régularité des fonctions. Si $\mathcal{X} \subset \mathbb{R}^d$ et $\mathcal{Y} \subset \mathbb{R}$, la notion de régularité la plus simple est associée à celle de dérivabilité, ou plus généralement à la notion de régularité lipschitzienne.

Définition 1.4 *Soit $\mathcal{D} \subset \mathbb{R}^d$ un domaine ouvert. Une application $f : \mathcal{D} \rightarrow \mathbb{R}$ est localement lipschitzienne en $\mathbf{x} \in \mathcal{D}$ quand il existe $C_x > 0$ tel que pour tout $\mathbf{x}' \in \mathcal{D}$, $|f(\mathbf{x}) - f(\mathbf{x}')| \leq C_x \|\mathbf{x} - \mathbf{x}'\|$, où $\|\mathbf{x}\|$ désigne la norme euclidienne. L'application f est C -lipschitzienne dans $\mathcal{D}$ quand pour tous $\mathbf{x}, \mathbf{x}' \in \mathcal{D}$, $|f(\mathbf{x}) - f(\mathbf{x}')| \leq C \|\mathbf{x} - \mathbf{x}'\|$. L'application f est lipschitzienne quand il existe $C > 0$ tel que f est C -lipschitzienne.*

Il est facile de vérifier que si $f : \mathcal{D} \rightarrow \mathbb{R}$ est continûment différentiable dans $\mathcal{D}$, alors elle est localement lipschitzienne. Nous admettrons le théorème de Rademacher, qui dit que si $f : \mathcal{D} \rightarrow \mathbb{R}$ est lipschitzienne, alors elle est différentiable presque partout dans $\mathcal{D}$ pour la mesure de Lebesgue.

Revenons au problème de classification et supposons f C -lipschitzienne, i.e., $\mathcal{X}$ est un domaine compact $\mathcal{D} \subset \mathbb{R}^d$ et $\mathcal{C}$ est l'ensemble des fonctions C -lipschitziennes de $\mathcal{D}$ dans $\mathbb{R}$. L'algorithme du plus proche voisin donne une approximation de f par une fonction constante sur les cellules de Voronoï définies par les données d'entraînement : $\forall \mathbf{x} \in \mathcal{D}$, $h_S(\mathbf{x}) = f(\mathbf{x}_i)$ pour

$i \in \arg \min_{j \in \llbracket 1; m \rrbracket} \|\mathbf{x} - \mathbf{x}_j\|$. Ainsi, $\forall \mathbf{x} \in \mathcal{D}$, $|f(\mathbf{x}) - h_S(\mathbf{x})| \leq C \min_i \|\mathbf{x} - \mathbf{x}_i\|$ et le problème revient à avoir $\max_{\mathbf{x}} \min_i \|\mathbf{x} - \mathbf{x}_i\| = \varepsilon$ petit. La question alors est de savoir combien il faut de données d'entraînement pour être sûr que ce soit vrai.

Nous allons voir que cette question conduit à la malédiction de la dimensionnalité, c'est-à-dire que le nombre de données d'entraînement nécessaire croît plus vite qu'exponentiellement avec la dimension d . Choisissons pour $\mathcal{X}$ l'hypercube $\mathcal{D} = [0; 1]^d$ et supposons que les $\mathbf{x}_i$ soient régulièrement répartis dans le cube sur le réseau obtenu en divisant chaque côté de $\mathcal{D}$ en $\Delta^{-1} - 1$ segments de longueur Δ et 2 segments de longueur $\Delta/2$, un sur chacun des bords

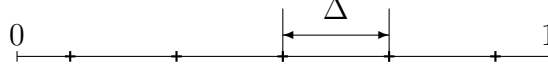


FIGURE 2 – Découpage d'un côté de l'hypercube, ici $\Delta = 1/5$.

(voir la figure 2), le nombre de points du réseau vaut $m = (\Delta^{-1})^d$ et pour un point $\mathbf{x} \in \mathcal{X}$, si $\mathbf{x}_i$ désigne le (ou un) point du réseau le plus proche de $\mathbf{x}$, on a

$$\|\mathbf{x} - \mathbf{x}_i\|^2 = \sum_{u=1}^d (\mathbf{x}(u) - \mathbf{x}_i(u))^2 \leq \frac{d\Delta^2}{4},$$

car chaque terme dans la somme est majoré par $(\frac{\Delta}{2})^2$. On obtient donc

$$\varepsilon = \sup_{\mathbf{x} \in \mathcal{D}} \min_i \|\mathbf{x} - \mathbf{x}_i\| \leq \frac{\sqrt{d}\Delta}{2} \leq \frac{\sqrt{d}m^{-1/d}}{2}. \quad (11)$$

Par ailleurs, si l'on veut recouvrir l'hypercube $\mathcal{D}$ par des boules de rayon ε , en notant $B(\mathbf{x}, \varepsilon)$ la boule de centre $\mathbf{x}$ et de rayon ε , on a $\mathcal{D} \subset \bigcup_{i=1}^m B(\mathbf{x}_i, \varepsilon)$, qui entraîne que

$$1 = \text{vol}(\mathcal{D}) \leq \text{vol}\left(\bigcup_{i=1}^m B(\mathbf{x}_i, \varepsilon)\right) \leq m \text{vol}(B(\mathbf{0}, \varepsilon)) = \frac{m\varepsilon^d \pi^{d/2}}{\Gamma(\frac{d}{2} + 1)},$$

où vol désigne le volume. En utilisant la formule de Stirling⁵ on obtient l'inégalité

$$1 \leq \left(\frac{2e\pi}{d}\right)^{\frac{d}{2}} \frac{m\varepsilon^d}{\sqrt{\pi d}} \quad \text{ou encore} \quad \sqrt{\pi d} \left(\frac{d}{2e\pi}\right)^{\frac{d}{2}} \varepsilon^{-d} \leq m, \quad (12)$$

qui montre que le nombre de données d'entraînement minimal, pour une erreur d'approximation petite avec l'algorithme des k plus proches voisins, croît beaucoup plus vite que ε^{-d} . Ce phénomène est connu sous le nom de *la malédiction de la dimensionnalité*, et correspond au fait que $\pi^{d/2}/\Gamma(\frac{d}{2} + 1)$, le rapport du volume de l'hypersphère inscrite dans l'hypercube sur le volume de l'hypercube, tend très vite vers 0 quand la dimension augmente.

Nous concluons ce paragraphe en montrant qu'en grande dimension, la régularité lipschitzienne ne suffit pas comme information *a priori* pour assurer un risque réel petit avec l'algorithme des k plus proches voisins. En effet, si f est C -lipschitzienne et le coût quadratique, nous avons vu avec les inégalités (10) et (11) que $\epsilon_{\text{app}} \leq C^2 \varepsilon^2 \leq \frac{C^2 d m^{-2/d}}{4}$, or pour $\mathcal{H}$ l'ensemble des fonctions constantes sur les cellules de Voronoï construites à partir de m points, nous avons vu à la remarque 1.1 que $\log(|\mathcal{H}|) = md \log(N)$ (où $N = 2^{64}$ si les flottants sont codés sur 64 bits) donc pour tout d , il existe $C'(d) > 0$ tel que $\min_{h \in \mathcal{H}} R(h) \leq C'(d) (\log(|\mathcal{H}|))^{-2/d}$ et si d n'est pas un petit entier (i.e., s'il est supérieur à 5 ou 6), d'après la proposition 1.6, le nombre m de données assurant un risque réel petit (e.g., 10^{-2}) est astronomique (supérieur à $100^{2+\frac{d}{2}}$).

Remarquons qu'en revanche, quand d est petit (inférieur à 2 ou 3), l'algorithme des k plus proches voisins donne un estimateur non paramétrique de la ddp de $\mathbf{x}$ (en supposant qu'elle

5. $\Gamma(\frac{d}{2} + 1) = (\frac{d}{2})! \simeq \sqrt{\pi d} (\frac{d}{2e})^{\frac{d}{2}}$ quand d tend vers l'infini, et $(\frac{d}{2})! \leq \sqrt{\pi d} (\frac{d}{2e})^{\frac{d}{2}}$ pour tout $d \geq 0$.

existe)⁶ et que dans un problème de détection (i.e., de classification binaire) le rapport de vraisemblance (i.e., le rapport de la ddp sous l'une des hypothèses sur la ddp de l'hypothèse alternative) est une statistique suffisante. Autrement dit, si l'on connaît les deux vraisemblances, le problème de détection est parfaitement résolu.

En grande dimension, par exemple si $\mathbf{x}$ est une image de $d = 10^6$ pixels, la régularité lipschitzienne de f ou des vraisemblances ne suffit pas pour trouver une approximation de f sans que le nombre de données d'entraînement nécessaires explose. Dans un problème de classification, les objets d'une même classe présentent une grande variabilité et, pour reprendre les termes de Stéphane Mallat :

« Tout le problème est de contrôler cette variabilité, la fonction f est régulière par rapport à une source de variabilité qui est beaucoup plus compliquée [à modéliser]. C'est le centre de tous les problèmes d'approximation que de comprendre cette régularité. »

En grande dimension, les objets $x \in \mathcal{X}$ à partir desquels un algorithme d'apprentissage supervisé prend une décision, sont très souvent des signaux numériques, c'est-à-dire le résultat d'un échantillonnage appliqué à une fonction $x(u)$ où la variable $u \in \mathbb{R}^\ell$ est de petite dimension $1 \leq \ell \leq 4$, et les signaux que l'on rencontre dans les applications (par exemple les images ou les sons) sont des fonctions $u \mapsto x(u)$ qui ont une certaine régularité. La régularité de la fonction $u \mapsto x(u)$ nous donne de l'information *a priori* sur la fonction f , elle implique que la donnée $\mathbf{x}$ n'occupe pas tout le domaine $\mathcal{X} = [0; 1]^d$, mais seulement un domaine de volume négligeable et la démonstration ci-dessus de la malédiction de la dimensionnalité ne s'applique plus.

En petite dimension, la notion de régularité d'une fonction est bien comprise d'un point de vue mathématique, elle repose sur les notions de transformée de Fourier, de décompositions en ondelettes et de représentation parcimonieuse d'un signal, qui sont enseignées dans les cours de "Traitement du signal".

1.6 Il n'existe pas d'algorithme d'apprentissage universel

Nous allons montrer qu'il n'existe pas d'algorithme d'apprentissage universel. C'est l'objet du théorème suivant (dit *no-free-lunch*).

Théorème 1.7 *Soit $\mathbf{A}$ un algorithme d'apprentissage pour un problème de classification binaire. Soit une taille d'échantillon m strictement inférieure à $|\mathcal{X}|/2$. Alors il existe une loi de probabilité sur $\mathcal{X} \times \{0, 1\}$ telle que :*

1. *il existe une fonction $f : \mathcal{X} \rightarrow \{0, 1\}$ avec $R(f) = 0$;*
2. *avec une probabilité d'au moins $1/7$ sur le choix de S , on a $R[\mathbf{A}(S)] \geq 1/8$.*

Preuve. Soit C un sous-ensemble de $\mathcal{X}$ de taille $2m$. Il y a $T = 2^{2m}$ fonctions de C dans $\{0, 1\}$, numérotons-les $f_1, \dots, f_T$. Pour chacune de ces fonctions, considérons la loi de probabilité

$$\mathcal{D}_i(\{(x, y)\}) = \begin{cases} 1/|C| & \text{si } y = f_i(x) \\ 0 & \text{sinon.} \end{cases}$$

En notant $R_{\mathcal{D}_i}(h)$ le risque réel de h pour la loi $\mathcal{D}_i$, il est clair que $R_{\mathcal{D}_i}(f_i) = 0$. Nous allons montrer que tout algorithme d'apprentissage admettant pour entrée un échantillon S de m couples dans $C \times \{0, 1\}$ et retournant une fonction $\mathbf{A}(S) : C \rightarrow \{0, 1\}$ vérifie

$$\max_{i \in [1; T]} \mathbb{E}_{\mathcal{D}_i^m} [R_{\mathcal{D}_i}[\mathbf{A}(S)]] \geq 1/4. \quad (13)$$

Il y a $k = (2m)^m$ séquences S possibles, notons-les $S_j = (x_1^j, \dots, x_m^j)$ pour $j \in [1; k]$, et pour $i \in [1; T]$ notons $S_j^i = ((x_1^j, f_i(x_1^j)), \dots, (x_m^j, f_i(x_m^j)))$. Si la loi de probabilité est $\mathcal{D}_i$, alors les

6. Voir Y. P. Mack et M. Rosenblatt, "Multivariate k -nearest neighbor density estimates", *Journal of Multivariate Analysis*, vol. 9, pp. 1–15, 1979 et E. Parzen, "On estimation of a density probability function and mode", *Ann. Inst. Statist. Math.*, vol. 33, pp. 1065–1076, 1962.

données d'entraînement que peut recevoir $\mathbf{A}$ sont les S_j^i pour $1 \leq j \leq k$ et tous ces échantillons sont équiprobables. Par conséquent, nous avons

$$\mathbb{E}_{\mathcal{D}_i^m} [R_{\mathcal{D}_i}[\mathbf{A}(S)]] = \frac{1}{k} \sum_{j=1}^k R_{\mathcal{D}_i}[\mathbf{A}(S_j^i)] \quad (14)$$

et, le maximum étant supérieur à la moyenne qui est supérieure au minimum,

$$\begin{aligned} \max_{i \in \llbracket 1; T \rrbracket} \frac{1}{k} \sum_{j=1}^k R_{\mathcal{D}_i}[\mathbf{A}(S_j^i)] &\geq \frac{1}{T} \sum_{i=1}^T \frac{1}{k} \sum_{j=1}^k R_{\mathcal{D}_i}[\mathbf{A}(S_j^i)] = \frac{1}{k} \sum_{j=1}^k \frac{1}{T} \sum_{i=1}^T R_{\mathcal{D}_i}[\mathbf{A}(S_j^i)] \\ &\geq \min_{j \in \llbracket 1; k \rrbracket} \frac{1}{T} \sum_{i=1}^T R_{\mathcal{D}_i}[\mathbf{A}(S_j^i)]. \end{aligned} \quad (15)$$

Fixons maintenant $j \in \llbracket 1; k \rrbracket$, notons $S_j = (x_1, \dots, x_m)$ et $v_1, \dots, v_p$ les éléments de C qui n'apparaissent pas dans S_j . Les tirages étant avec remise, on a $p \geq m$ et pour toute hypothèse $h : C \rightarrow \{0, 1\}$ et tout i , nous avons

$$R_{\mathcal{D}_i}(h) = \frac{1}{2m} \sum_{x \in C} \mathbb{1}_{[h(x) \neq f_i(x)]} \geq \frac{1}{2m} \sum_{r=1}^p \mathbb{1}_{[h(v_r) \neq f_i(v_r)]} \geq \frac{1}{2p} \sum_{r=1}^p \mathbb{1}_{[h(v_r) \neq f_i(v_r)]}, \quad (16)$$

donc

$$\frac{1}{T} \sum_{i=1}^T R_{\mathcal{D}_i}[\mathbf{A}(S_j^i)] \geq \frac{1}{2p} \sum_{r=1}^p \frac{1}{T} \sum_{i=1}^T \mathbb{1}_{[\mathbf{A}(S_j^i)(v_r) \neq f_i(v_r)]} \geq \frac{1}{2} \min_{r \in \llbracket 1; p \rrbracket} \frac{1}{T} \sum_{i=1}^T \mathbb{1}_{[\mathbf{A}(S_j^i)(v_r) \neq f_i(v_r)]}. \quad (17)$$

Enfin, fixons $r \in \llbracket 1; p \rrbracket$. Nous pouvons partitionner l'ensemble de toutes les fonctions f_i en $T/2$ paires disjointes, où pour toute paire $\{f_i, f_{i'}\}$, nous avons quel que soit $c \in C$, $f_i(c) \neq f_{i'}(c)$ si et seulement si $c = v_r$. Enfin, puisque que pour chaque paire nous avons $S_j^i = S_j^{i'}$, il vient $\mathbb{1}_{[\mathbf{A}(S_j^i)(v_r) \neq f_i(v_r)]} + \mathbb{1}_{[\mathbf{A}(S_j^{i'})(v_r) \neq f_{i'}(v_r)]} = 1$, qui entraîne $\frac{1}{T} \sum_{i=1}^T \mathbb{1}_{[\mathbf{A}(S_j^i)(v_r) \neq f_i(v_r)]} = \frac{1}{2}$. La démonstration de (13) en résulte, d'après les inégalités (17), (15) et (14).

Finalement, $\mathbb{P}(R_{\mathcal{D}_i}[\mathbf{A}(S)] \geq 1 - a) \geq \frac{\mathbb{E}(R_{\mathcal{D}_i}[\mathbf{A}(S)]) - (1 - a)}{a}$, d'après l'inégalité de Markov appliquée à $1 - R_{\mathcal{D}_i}[\mathbf{A}(S)]$, et $a = 7/8$ donne $\mathbb{P}(R_{\mathcal{D}_i}[\mathbf{A}(S)] \geq \frac{1}{8}) \geq \frac{\mathbb{E}(R_{\mathcal{D}_i}[\mathbf{A}(S)]) - (1/8)}{7/8} \geq \frac{1}{7}$. $\diamond$

Corollaire 1.8 Soient $\mathcal{X}$ un ensemble infini et $\mathcal{Y} = \{0, 1\}$. Alors l'ensemble $\mathcal{H} = \mathcal{M}(\mathcal{X}, \mathcal{Y})$ n'a pas la capacité d'apprentissage PAC.

Preuve. Supposons que $\mathcal{H}$ ait la capacité PAC d'apprendre. Soient $\varepsilon < 1/8$ et $\delta < 1/7$, il résulte de la capacité PAC d'apprendre qu'il existe un algorithme d'apprentissage $\mathbf{A}$ et un entier $m = m(\varepsilon, \delta)$ tels que pour toute loi de probabilité sur $\mathcal{X} \times \mathcal{Y}$, s'il existe une fonction $f \in \mathcal{M}(\mathcal{X}, \mathcal{Y})$ telle que $R(f) = 0$, alors avec une probabilité d'au moins $1 - \delta$, quand $\mathbf{A}$ est appliqué à des données d'entraînements S de taille m , on a $R[\mathbf{A}(S)] \leq \varepsilon$. Mais cela est en contradiction avec le théorème 1.7 puisque $|\mathcal{X}| > 2m$. $\diamond$

Nous voyons ainsi clairement apparaître le phénomène de sur-apprentissage, si l'on n'impose pas de contrainte à l'ensemble d'hypothèses $\mathcal{H}$.

1.7 La dimension VC (Vapnik-Chervonenkis) d'un ensemble d'hypothèses

Nous commençons par donner l'exemple d'un espace d'hypothèses continu ayant la capacité PAC d'apprendre. Pour $a \in \mathbb{R}$, on définit le détecteur à seuil $h_a : \mathbb{R} \rightarrow \{0, 1\}$ par $h_a(x) = 1$ si $x \geq a$ et $h_a(x) = 0$ si $x < a$.

Lemme 1.9 Pour $\mathcal{X} \times \mathcal{Y} = \mathbb{R} \times \{0, 1\}$, l'ensemble des détecteurs à seuil a la capacité PAC d'apprendre pour l'algorithme d'apprentissage $\text{ERM}_{\mathcal{H}}$ avec une complexité $m_{\mathcal{H}}(\varepsilon, \delta) \leq \left\lceil \frac{\log(2/\delta)}{\varepsilon} \right\rceil$.

Preuve. Nous donnons une démonstration dans le cas où l'ensemble des détecteurs à seuil vérifie l'hypothèse de prédiction singulière, le cas général se déduira du premier théorème fondamental de l'apprentissage (théorème 1.14). Soit $a^* \in \mathbb{R}$ tel que $R(h_{a^*}) = 0$. Fixons $\varepsilon, \delta \in]0, 1[$. Avec la convention $\inf \emptyset = +\infty$ et $\sup \emptyset = -\infty$, introduisons

$$a_0 = \sup\{a < a^* : \mathbb{P}(a \leq x < a^*) \geq \varepsilon\} \quad \text{et} \quad a_1 = \inf\{a > a^* : \mathbb{P}(a^* \leq x < a) \geq \varepsilon\}.$$

Pour un échantillon $S = ((x_i, y_i))_{1 \leq i \leq m}$, avec $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$, posons $b_1(S) = \inf\{x_i : y_i = 1\}$ et $b_0(S) = \sup\{x_i : y_i = 0\}$. Soit b_S un seuil correspondant à $\text{ERM}_{\mathcal{H}}(S)$. On a $b_0(S) < b_S \leq b_1(S)$ et $R(h_S) \leq \varepsilon$ si et seulement si $a_0 \leq b_0(S)$ et $b_1(S) \leq a_1$. Donc $\mathbb{P}[R(h_S) > \varepsilon] \leq \mathbb{P}[a_0 > b_0(S)] + \mathbb{P}[a_1 < b_1(S)]$, mais $\mathbb{P}[a_1 < b_1(S)] = \mathbb{P}(x \notin [a^*, a_1])^m \leq (1 - \varepsilon)^m \leq e^{-m\varepsilon}$ et de même $\mathbb{P}[a_0 > b_0(S)] = \mathbb{P}(x \notin [a_0, a^*])^m \leq (1 - \varepsilon)^m \leq e^{-m\varepsilon}$, d'où $\mathbb{P}[R(h_S) > \varepsilon] \leq 2e^{-m\varepsilon} \leq \delta$ puisque $m \geq \left\lceil \frac{\log(2/\delta)}{\varepsilon} \right\rceil$. $\diamond$

Définition 1.5 Soient $\mathcal{C} \subset \mathcal{X}$. La restriction de $\mathcal{H}$ à $\mathcal{C}$, notée $\mathcal{H}_{\mathcal{C}}$, est $\{h|_{\mathcal{C}} : \mathcal{C} \rightarrow \mathcal{Y}, h \in \mathcal{H}\}$.

Quand $\mathcal{C} = \{c_1, \dots, c_m\}$ avec $m < \infty$, on identifie la restriction $h|_{\mathcal{C}}$ d'une hypothèse h au m -uplet $(h(c_1), \dots, h(c_m)) \in \mathcal{Y}^m$.

Définition 1.6 Un ensemble d'hypothèses $\mathcal{H} \subset \mathcal{M}(\mathcal{X}, \{0, 1\})$ brise⁷ $\mathcal{C}$ quand $\mathcal{H}_{\mathcal{C}}$ est l'ensemble des fonctions de $\mathcal{C}$ dans $\{0, 1\}$ (autrement dit $|\mathcal{H}_{\mathcal{C}}| = 2^{|\mathcal{C}|}$).

Autrement dit, $\mathcal{H}$ brise $\mathcal{C}$, quand pour tout sous-ensemble $\mathcal{B} \subset \mathcal{C}$, il existe $h \in \mathcal{H}$ telle que $h|_{\mathcal{C}}$ coïncide sur $\mathcal{C}$ avec la fonction indicatrice de $\mathcal{B}$. Ou encore, si $\mathcal{C}$ correspond aux objets $x \in \mathcal{X}$ des données d'entraînement, alors quelles que soient les labels (y) associés, il existe $h \in \mathcal{H}$ qui annule le risque empirique.

Par exemple, l'ensemble des détecteurs à seuil brise tout singleton $\{c\} \subset \mathbb{R}$, mais il ne brise aucune paire $\{c_1, c_2\} \subset \mathbb{R}$ (en effet, si $c_1 < c_2$, aucun détecteur à seuil h_a ci-dessus ne donnera $h_a(c_1) = 1$ et $h_a(c_2) = 0$).

Corollaire 1.10 Soient $\mathcal{H} \subset \mathcal{M}(\mathcal{X}, \{0, 1\})$ et m la taille de données d'entraînement. Supposons qu'il existe un sous-ensemble $\mathcal{C}$ de $\mathcal{X}$ de cardinal $2m$ qui est brisé par $\mathcal{H}$. Alors, pour tout algorithme d'apprentissage, $\mathbf{A}$, il existe une loi de probabilité sur $\mathcal{X} \times \{0, 1\}$ et un prédicteur $h \in \mathcal{H}$ tels que $R(h) = 0$ et $\mathbb{P}(R[\mathbf{A}(S)] \geq 1/8) \geq 1/7$.

Preuve. La démonstration du corollaire 1.8 s'applique ici. $\diamond$

Définition 1.7 La dimension-VC d'un ensemble d'hypothèses $\mathcal{H}$, notée $\text{dimVC}(\mathcal{H})$, est la taille maximale d'un sous-ensemble fini $\mathcal{C} \subset \mathcal{X}$ qui peut être brisé par $\mathcal{H}$. Si $\mathcal{H}$ peut briser des sous-ensembles finis de n'importe quelle taille, sa dimension-VC est infinie.

Le théorème suivant résulte directement du corollaire 1.10.

Théorème 1.11 Aucun ensemble d'hypothèses de dimension-VC infinie n'a la capacité d'apprentissage PAC.

7. *shatter* en anglais.

1.8 Le théorème fondamental de l'apprentissage PAC

Définition 1.8 Soit $\mathcal{H}$ un ensemble d'hypothèses. L'application $\tau_{\mathcal{H}} : \mathbb{N} \rightarrow \mathbb{N}$ définie par

$$\tau_{\mathcal{H}}(m) = \max_{\mathcal{C} \subset \mathcal{X} : |\mathcal{C}|=m} |\mathcal{H}_{\mathcal{C}}|$$

est la fonction de croissance de $\mathcal{H}$.

Il est clair que si $\dim \text{VC}(\mathcal{H}) = d < \infty$, alors pour tout $m \leq d$, $\tau_{\mathcal{H}}(m) = 2^m$. Le lemme suivant montre que la fonction de croissance est polynomiale quand $m > d$.

Lemme 1.12 (Sauer-Shelah-Perles) Si $\dim \text{VC}(\mathcal{H}) = d < \infty$ pour un ensemble d'hypothèses $\mathcal{H}$, alors pour tout m , $\tau_{\mathcal{H}}(m) \leq \sum_{i=0}^d \binom{m}{i}$. En particulier pour $m > d+1$ on a $\tau_{\mathcal{H}}(m) \leq (em/d)^d$.

Preuve. Nous allons démontrer une assertion plus forte : pour tout $\mathcal{C} \subset \mathcal{X}$ de taille m

$$\forall \mathcal{H} \subset \mathcal{M}(\mathcal{X}, \{0, 1\}), \quad |\mathcal{H}_{\mathcal{C}}| \leq |\{\mathcal{B} \subset \mathcal{C} : \mathcal{H} \text{ brise } \mathcal{B}\}|. \quad (18)$$

En effet, puisqu'aucun sous-ensemble de taille strictement supérieure à d ne peut être brisé par $\mathcal{H}$, on a $|\{\mathcal{B} \subset \mathcal{C} : \mathcal{H} \text{ brise } \mathcal{B}\}| \leq \sum_{i=0}^d \binom{m}{i}$ et l'inégalité (18) implique alors la première assertion du lemme. La preuve de la deuxième assertion du lemme est laissée à titre d'exercice.

La démonstration de (18) se fait par récurrence sur m . Pour $m = 1$, les deux membres de l'inégalité (18) ont tous les deux la même valeur, qui est 2 ou 1 (par convention l'ensemble vide est toujours brisé par $\mathcal{H}$). Supposons que l'inégalité (18) soit satisfaite pour tout sous-ensemble de taille $k < m$ et prouvons-la pour m . Fixons $\mathcal{C} = \{c_1, \dots, c_m\}$ et $\mathcal{H}$, posons $\mathcal{C}' = \{c_2, \dots, c_m\}$ et définissons les ensembles

$$\begin{aligned} \mathcal{Y}_0 &= \{(y_2, \dots, y_m) \in \{0, 1\}^{m-1} : (0, y_2, \dots, y_m) \in \mathcal{H}_{\mathcal{C}} \text{ ou } (1, y_2, \dots, y_m) \in \mathcal{H}_{\mathcal{C}}\} \\ \mathcal{Y}_1 &= \{(y_2, \dots, y_m) \in \{0, 1\}^{m-1} : (0, y_2, \dots, y_m) \in \mathcal{H}_{\mathcal{C}} \text{ et } (1, y_2, \dots, y_m) \in \mathcal{H}_{\mathcal{C}}\}. \end{aligned}$$

Il est clair que $|\mathcal{H}_{\mathcal{C}}| = |\mathcal{Y}_0| + |\mathcal{Y}_1|$ et, puisque $\mathcal{Y}_0 = \mathcal{H}_{\mathcal{C}'}$, l'hypothèse de récurrence appliquée à $\mathcal{H}$ et $\mathcal{C}'$ donne

$$|\mathcal{Y}_0| \leq |\{\mathcal{B} \subset \mathcal{C}' : \mathcal{H} \text{ brise } \mathcal{B}\}| = |\{\mathcal{B} \subset \mathcal{C} : c_1 \notin \mathcal{B} \text{ et } \mathcal{H} \text{ brise } \mathcal{B}\}|.$$

Maintenant, introduisons les paires d'hypothèses qui coïncident sur $\mathcal{C}'$ et qui diffèrent sur c_1 :

$$\mathcal{H}' = \{h \in \mathcal{H} : \exists h' \in \mathcal{H}, (1 - h'(c_1), h'(c_2), \dots, h'(c_m)) = (h(c_1), h(c_2), \dots, h(c_m))\}.$$

Il est clair que si $\mathcal{H}'$ brise un sous-ensemble $\mathcal{B} \subset \mathcal{C}'$, alors il brise $\mathcal{B} \cup \{c_1\}$ et réciproquement. En remarquant que $\mathcal{Y}_1 = \mathcal{H}'_{\mathcal{C}'}$ et en appliquant l'hypothèse de récurrence à $\mathcal{H}'$ et $\mathcal{C}'$ on obtient

$$\begin{aligned} |\mathcal{Y}_1| \leq |\{\mathcal{B} \subset \mathcal{C}' : \mathcal{H}' \text{ brise } \mathcal{B}\}| &= |\{\mathcal{B} \subset \mathcal{C}' : \mathcal{H}' \text{ brise } \mathcal{B} \cup \{c_1\}\}| \\ &= |\{\mathcal{B} \subset \mathcal{C} : c_1 \in \mathcal{B} \text{ et } \mathcal{H}' \text{ brise } \mathcal{B}\}| \\ &\leq |\{\mathcal{B} \subset \mathcal{C} : c_1 \in \mathcal{B} \text{ et } \mathcal{H} \text{ brise } \mathcal{B}\}|. \end{aligned}$$

Finalement, nous avons $|\mathcal{H}_{\mathcal{C}}| = |\mathcal{Y}_0| + |\mathcal{Y}_1|$ et l'inégalité

$$|\mathcal{Y}_0| + |\mathcal{Y}_1| \leq |\{\mathcal{B} \subset \mathcal{C} : c_1 \notin \mathcal{B} \text{ et } \mathcal{H} \text{ brise } \mathcal{B}\}| + |\{\mathcal{B} \subset \mathcal{C} : c_1 \in \mathcal{B} \text{ et } \mathcal{H} \text{ brise } \mathcal{B}\}|,$$

dont le membre de droite vaut $|\{\mathcal{B} \subset \mathcal{C} : \mathcal{H} \text{ brise } \mathcal{B}\}|$, ce qui termine la démonstration. $\diamond$

Théorème 1.13 (1^{er} théorème fondamental d'apprentissage) Soit un ensemble d'hypothèses $\mathcal{H} \subset \mathcal{M}(\mathcal{X}, \{0, 1\})$ et un coût d'erreur constant ($L(y, y') = 1 - \delta_{y, y'}$). Les assertions suivantes sont équivalentes.

- (i) $\mathcal{H}$ a la propriété de convergence uniforme ;

- (ii) $\mathcal{H}$ a la capacité d'apprentissage PAC avec l'algorithme $\text{ERM}_{\mathcal{H}}$;
- (iii) la dimension-VC de $\mathcal{H}$ est finie.

Nous avons déjà démontré que (i) $\Rightarrow$ (ii) (lemme 1.2 et proposition 1.1). Remarquons que l'assertion (ii) entraîne trivialement que $\mathcal{H}$ a la capacité d'apprentissage PAC “faible”, où ne sont considérées que les lois de probabilité pour lesquelles $\mathcal{H}$ vérifie l'hypothèse 2 de prédiction singulière, et que la capacité d'apprentissage “faible” entraîne l'assertion (iii) (théorème 1.11). La démonstration de l'implication (iii) $\Rightarrow$ (i) et celle du théorème suivant sont reportées aux prochaines sous-sections.

Remarque 1.2 *Il résulte du premier théorème fondamental d'apprentissage que si $\mathcal{H}$ a la capacité d'apprentissage PAC pour toute loi de probabilité satisfaisant l'hypothèse de prédiction singulière, alors il a la capacité d'apprentissage PAC pour toute loi de probabilité (même quand l'hypothèse de prédiction singulière n'est pas satisfaite).*

Théorème 1.14 (2^{ème} théorème fondamental d'apprentissage) *Considérons un problème de classification binaire $\mathcal{X} \times \{0, 1\}$ et un coût d'erreur constant ($L(y, y') = 1 - \delta_{y, y'}$). Alors il existe des constantes C_1 et C_2 telles que pour tout ensemble d'hypothèses $\mathcal{H} \subset \mathcal{M}(\mathcal{X}, \{0, 1\})$ de dimension-VC finie égale à d ,*

- (i) $\mathcal{H}$ a la propriété de convergence uniforme avec la complexité

$$C_1 \frac{d + \log(1/\delta)}{\varepsilon^2} \leq m_{\mathcal{H}}^{\text{CU}}(\varepsilon, \delta) \leq C_2 \frac{d + \log(1/\delta)}{\varepsilon^2}$$

- (ii) $\mathcal{H}$ a la capacité d'apprentissage PAC avec la complexité

$$C_1 \frac{d + \log(1/\delta)}{\varepsilon^2} \leq m_{\mathcal{H}}(\varepsilon, \delta) \leq C_2 \frac{d + \log(1/\delta)}{\varepsilon^2}$$

- (iii) si, de plus, $\mathcal{H}$ satisfait l'hypothèse 2 de prédiction singulière, alors $\mathcal{H}$ a la la capacité d'apprentissage PAC avec la complexité

$$C_1 \frac{d + \log(1/\delta)}{\varepsilon} \leq m_{\mathcal{H}}(\varepsilon, \delta) \leq C_2 \frac{d + \log(1/\delta)}{\varepsilon}.$$

Les constantes C_1 et C_2 des assertions (i), (ii) et (iii) ne sont pas les mêmes.

Lemme 1.15 *Soit X une VA réelle. Supposons qu'il existe $a > 0$ et $b \geq e$ tels que pour tout $t \geq 0$, $\mathbb{P}(|X| > t) \leq 2be^{-t^2/a^2}$. Alors, $\mathbb{E}[|X|] \leq a(2 + \sqrt{\log(b)})$.*

Preuve. Pour $i \in \mathbb{N}$, posons $t_i = a(i + \sqrt{\log(b)})$. On a

$$\begin{aligned} \mathbb{E}[|X|] &= \int_0^{t_0} u dP_{|X|}(u) + \sum_{i=1}^{\infty} \int_{t_{i-1}}^{t_i} u dP_{|X|}(u) \leq t_0 + \sum_{i=1}^{\infty} t_i \mathbb{P}(t_{i-1} < |X| \leq t_i) \\ &\leq t_0 + \sum_{i=1}^{\infty} t_i \mathbb{P}(|X| > t_{i-1}) \leq a\sqrt{\log(b)} + 2ab \sum_{i=1}^{\infty} (i + \sqrt{\log(b)}) e^{-(i-1+\sqrt{\log(b)})^2} \\ &\leq a\sqrt{\log(b)} + 2ab \int_{1+\sqrt{\log(b)}}^{+\infty} x e^{-(x-1)^2} dx \leq a\sqrt{\log(b)} + 2ab \int_{\sqrt{\log(b)}}^{+\infty} (u+1) e^{-u^2} du, \end{aligned}$$

car, puisque $\log(b) \geq 1$, $x \mapsto x e^{-(x-1)^2}$ est décroissante sur $[1 + \sqrt{\log(b)}; +\infty[$. Enfin,

$$2ab \int_{\sqrt{\log(b)}}^{+\infty} (u+1) e^{-u^2} du \leq 4ab \int_{\sqrt{\log(b)}}^{+\infty} u e^{-u^2} du = 2a$$

◇

Théorème 1.16 Soient $\mathcal{H}$ un ensemble d'hypothèses et $\tau_{\mathcal{H}}$ sa fonction de croissance. Alors, pour toute loi de probabilité et tout $\delta \in]0; 1[$, avec une probabilité d'au moins $1 - \delta$ sur les données d'entraînement S de taille m , nous avons

$$\forall h \in \mathcal{H}, \quad |R(h) - R_S(h)| \leq \sqrt{\frac{2}{m}} \left(\frac{2 + \sqrt{\log[\tau_{\mathcal{H}}(2m)]}}{\delta} \right).$$

Preuve. Il suffit de montrer

$$\mathbb{E} \left[\sup_{h \in \mathcal{H}} |R(h) - R_S(h)| \right] \leq \sqrt{\frac{2}{m}} \left(2 + \sqrt{\log[\tau_{\mathcal{H}}(2m)]} \right), \quad (19)$$

car le théorème en découle directement d'après l'inégalité de Markov. Pour majorer le membre de gauche de (19), introduisons un nouvel échantillon $S' = ((x'_i, y'_i))_{1 \leq i \leq m}$ i.i.d. et indépendant de S . Pour toute hypothèse h , on a $R(h) = \mathbb{E}[R_{S'}(h)]$ et

$$|\mathbb{E}[R_{S'}(h) - R_S(h) \mid S]| \leq \mathbb{E}[|R_{S'}(h) - R_S(h)| \mid S] \leq \mathbb{E} \left[\sup_{g \in \mathcal{H}} |R_{S'}(g) - R_S(g)| \mid S \right],$$

donc
$$\sup_{h \in \mathcal{H}} |R(h) - R_S(h)| = \sup_{h \in \mathcal{H}} |\mathbb{E}[R_{S'}(h) - R_S(h) \mid S]| \leq \mathbb{E} \left[\sup_{h \in \mathcal{H}} |R_{S'}(h) - R_S(h)| \mid S \right] \text{ et}$$

$$\mathbb{E} \left[\sup_{h \in \mathcal{H}} |R(h) - R_S(h)| \right] \leq \mathbb{E} \left[\sup_{h \in \mathcal{H}} |R_{S'}(h) - R_S(h)| \right]. \quad (20)$$

Maintenant, on a

$$|R_{S'}(h) - R_S(h)| = \frac{1}{m} \left| \sum_{i=1}^m (L[h(x'_i), y'_i] - L[h(x_i), y_i]) \right|$$

et, les $2m$ données S et S' étant i.i.d., l'égalité précédente est inchangée en moyenne si on permute un (x_i, y_i) avec un (x'_i, y'_i) , autrement dit si on remplace un terme $L[h(x'_i), y'_i] - L[h(x_i), y_i]$ par son opposé $-(L[h(x'_i), y'_i] - L[h(x_i), y_i])$.

Ainsi pour tout vecteur $\sigma = (\sigma_i)_{1 \leq i \leq m} \in \{-1, 1\}^m$, le membre de droite de l'inégalité (20) est égal à

$$\mathbb{E} \left[\sup_{h \in \mathcal{H}} \frac{1}{m} \left| \sum_{i=1}^m \sigma_i (L[h(x'_i), y'_i] - L[h(x_i), y_i]) \right| \right], \quad (21)$$

qui reste inchangé si l'on considère le vecteur σ aléatoire avec une loi uniforme. Fixons alors S et S' et posons $\mathcal{C} = \{x_1, \dots, x_m, x'_1, \dots, x'_m\}$; la borne supérieure de l'expression (21) coïncide avec le maximum obtenu sur $\mathcal{H}_{\mathcal{C}}$. Pour $h \in \mathcal{H}_{\mathcal{C}}$, introduisons la VA

$$\theta_h = \frac{1}{m} \sum_{i=1}^m \sigma_i (L[h(x_i), y_i] - L[h(x'_i), y'_i]),$$

égale à la moyenne statistique de m VA i.i.d. centrées et à valeurs dans $[-1; 1]$ (qui sont les $\sigma_i (L[h(x_i), y_i] - L[h(x'_i), y'_i])$ sachant S et S'). Il résulte alors de l'inégalité de Hoeffding que pour tout $\eta > 0$,

$$\mathbb{P}[|\theta_h| > \eta \mid S, S'] \leq 2 \exp(-m\eta^2/2),$$

donc
$$\mathbb{P} \left[\max_{h \in \mathcal{H}_{\mathcal{C}}} |\theta_h| > \eta \mid S, S' \right] \leq 2|\mathcal{H}_{\mathcal{C}}| \exp(-m\eta^2/2), \text{ qui entraîne, d'après le lemme 1.15,}$$

$$\mathbb{E} \left[\max_{h \in \mathcal{H}_{\mathcal{C}}} |\theta_h| \mid S, S' \right] \leq \sqrt{\frac{2}{m}} \left(2 + \sqrt{\log(|\mathcal{H}_{\mathcal{C}}|)} \right) \leq \sqrt{\frac{2}{m}} \left(2 + \sqrt{\log[\tau_{\mathcal{H}}(2m)]} \right),$$

ce qui termine la démonstration avec l'inégalité (20). $\diamond$

1.9 Preuve du premier théorème fondamental

D'après le lemme de Sauer, nous avons $\tau_{\mathcal{H}}(2m) \leq (2me/d)^d$, qui combiné avec le théorème 1.16 montre qu'avec une probabilité d'au moins $1 - \delta$,

$$|R_S(h) - R(h)| \leq \sqrt{\frac{2}{m} \frac{\left(2 + \sqrt{d \log(2em/d)}\right)}{\delta}} \leq \frac{2\sqrt{2d \log(2em/d)}}{\delta\sqrt{m}},$$

où pour simplifier nous avons supposé que $2 \leq \sqrt{d \log(2em/d)}$.

Maintenant $\frac{2\sqrt{2d \log(2em/d)}}{\delta\sqrt{m}} \leq \varepsilon$, si et seulement si,

$$\frac{m\varepsilon^2\delta^2}{8d} \geq \log\left(\frac{2em}{d}\right) = 1 + \log\left(\frac{m\varepsilon^2\delta^2}{8d}\right) + 2\log\left(\frac{4}{\varepsilon\delta}\right),$$

ou encore, si et seulement si, $g\left(\frac{m\varepsilon^2\delta^2}{8d}\right) \geq 2\log\left(\frac{4}{\varepsilon\delta}\right)$ avec $g(u) = u - \log(u) - 1$. La fonction g , définie pour $u > 0$, est convexe positive, elle atteint son minimum (qui est nul) pour $u = 1$. Donc $g(u/2) \geq 0$, ce qui entraîne $g(u) \geq u/2 - \log(2)$. Ainsi, si

$$m \geq \frac{32d}{\varepsilon^2\delta^2} \log\left(\frac{4\sqrt{2}}{\varepsilon\delta}\right), \quad (22)$$

alors avec une probabilité d'au moins $1 - \delta$, $|R_S(h) - R(h)| \leq \varepsilon$, et donc $\mathcal{H}$ a la propriété de convergence uniforme avec une complexité $m_{\mathcal{H}}^{\text{CU}}(\varepsilon, \delta)$ majorée par le membre de droite de l'inégalité (22). $\diamond$

1.10 La capacité d'apprentissage PAC non uniforme

Définition 1.9 *Un ensemble d'hypothèses $\mathcal{H} \subset \mathcal{M}(\mathcal{X}, \mathcal{Y})$ a la capacité PAC d'apprentissage non uniforme quand il existe un algorithme d'apprentissage $\mathbf{A}$ et $m_{\mathcal{H}}^{\text{ANU}} :]0; 1]^2 \times \mathcal{H} \rightarrow \mathbb{N}$ une fonction telle que pour tous $\varepsilon, \delta \in]0, 1[$, pour tout $h \in \mathcal{H}$, si $m \geq m_{\mathcal{H}}^{\text{ANU}}(\varepsilon, \delta, h)$, alors pour toute loi de probabilité sur $\mathcal{X} \times \mathcal{Y}$, avec une probabilité d'au moins $1 - \delta$ sur le choix des données d'entraînement S de taille m , on a*

$$R[\mathbf{A}(S)] \leq R(h) + \varepsilon.$$

Remarquons que si $\mathcal{H}$ a la capacité PAC d'apprentissage, alors il a la capacité d'apprentissage non uniforme.

Théorème 1.17 *Soit un ensemble d'hypothèses $\mathcal{H} = \bigcup_{n=1}^{\infty} \mathcal{H}_n$, où, pour tout $n \geq 1$, $\mathcal{H}_n$ a la propriété de convergence uniforme. Alors $\mathcal{H}$ a la capacité d'apprentissage non uniforme.*

Un énoncé plus précis de ce théorème est donné à la section suivante (théorème 1.20), avec sa démonstration.

Théorème 1.18 *Pour $\mathcal{Y} = \{0, 1\}$ (i.e., un problème de classification binaire), un ensemble d'hypothèses a la capacité d'apprentissage non uniforme si et seulement si c'est la réunion dénombrable d'ensembles d'hypothèses ayant la capacité d'apprentissage PAC.*

Preuve. Supposons $\mathcal{H} = \bigcup_{n=1}^{\infty} \mathcal{H}_n$, où, pour tout $n \geq 1$, $\mathcal{H}_n$ a la capacité d'apprentissage PAC. Alors, d'après le théorème fondamental d'apprentissage, chaque $\mathcal{H}_n$ a la propriété de convergence uniforme et donc $\mathcal{H}$ a la capacité d'apprentissage non uniforme, d'après le théorème 1.17. Réciproquement, pour $n \geq 1$, posons $\mathcal{H}_n = \{h \in \mathcal{H} : m_{\mathcal{H}}^{\text{ANU}}(1/8, 1/7, h) \leq n\}$. Il est clair que $\mathcal{H} = \bigcup_{n=1}^{\infty} \mathcal{H}_n$. De plus, d'après la définition de $m_{\mathcal{H}}^{\text{ANU}}$, pour toute loi de probabilité qui vérifie l'hypothèse de prédiction singulière dans $\mathcal{H}_n$, on a une probabilité d'au moins $6/7$ de tirer un échantillon S de taille n tel que $R[\mathbf{A}(S)] \leq 1/8$. Ceci entraîne, d'après le corollaire 1.10 et le théorème 1.11, que la dimension VC de $\mathcal{H}_n$ est finie et donc, d'après le théorème fondamental d'apprentissage, que $\mathcal{H}_n$ a la capacité PAC d'apprentissage. $\diamond$

1.11 Minimisation du risque structurel

Soit un ensemble d'hypothèses $\mathcal{H} = \bigcup_{n=1}^{\infty} \mathcal{H}_n$, où pour tout $n \geq 1$, $\mathcal{H}_n$ a la propriété de convergence uniforme avec une complexité d'échantillon $m_{\mathcal{H}_n}^{\text{CU}}(\varepsilon, \delta)$. Définissons pour $n \geq 1$

$$\begin{aligned} \epsilon_n : \mathbb{N} \times]0; 1[&\rightarrow]0; 1[\\ (m, \delta) &\mapsto \inf\{\varepsilon \in]0; 1[: m_{\mathcal{H}_n}^{\text{CU}}(\varepsilon, \delta) \leq m\}, \end{aligned} \quad (23)$$

autrement dit, pour une taille m d'échantillon donnée, on cherche la borne inférieure de l'écart atteignable entre les risques, empirique et réel, pour tout $h \in \mathcal{H}_n$. Ainsi, avec une probabilité d'au moins $1 - \delta$ sur l'échantillon S de taille m , on a

$$\forall h \in \mathcal{H}_n, |R(h) - R_S(h)| \leq \epsilon_n(m, \delta). \quad (24)$$

Soit $w : \mathbb{N} \setminus \{0\} \rightarrow [0; 1]$ telle que $\sum_{n \geq 1} w(n) \leq 1$. Nous disons que c'est une *fonction de pondération* sur les classes d'hypothèses $\mathcal{H}_1, \mathcal{H}_2, \dots$. Le poids $w(n)$ correspond à l'importance que l'on porte *a priori* sur l'ensemble d'hypothèses $\mathcal{H}_n$. La règle de minimisation du risque structurel consiste à minimiser une certaine borne supérieure du risque réel.

Théorème 1.19 *Soit $w : \mathbb{N} \setminus \{0\} \rightarrow [0; 1]$ telle que $\sum_{n \geq 1} w(n) \leq 1$ et un ensemble d'hypothèses $\mathcal{H} = \bigcup_{n=1}^{\infty} \mathcal{H}_n$, où, pour tout $n \geq 1$, $\mathcal{H}_n$ a la propriété de convergence uniforme avec une complexité d'échantillon $m_{\mathcal{H}_n}^{\text{CU}}$. Soit ϵ_n définie à la relation (23). Alors, pour tout $\delta \in]0; 1[$, pour toute loi de probabilité, avec une probabilité d'au moins $1 - \delta$ sur le choix d'un échantillon S de taille m , la borne suivante est valable (simultanément) pour tout $n \geq 1$ et tout $h \in \mathcal{H}_n$:*

$$|R(h) - R_S(h)| \leq \epsilon_n(m, w(n)\delta).$$

Par conséquent, pour tout $\delta \in]0; 1[$ et toute loi de probabilité, avec une probabilité d'au moins $1 - \delta$ on a

$$\forall h \in \mathcal{H}, \quad R(h) \leq R_S(h) + \inf_{n: h \in \mathcal{H}_n} \epsilon_n(m, w(n)\delta). \quad (25)$$

Preuve. Fixons $\delta \in]0; 1[$ et posons $\delta_n = w(n)\delta$ pour $n \geq 1$, alors d'après la propriété de convergence uniforme et la relation (24), pour n fixé, pour toute loi de probabilité et tout échantillon S de taille m , avec une probabilité d'au moins $1 - \delta_n$, on a

$$\forall h \in \mathcal{H}_n, \quad |R(h) - R_S(h)| \leq \epsilon_n(m, \delta_n).$$

Et donc, avec une probabilité d'au moins $1 - \sum_{n \geq 1} \delta_n = 1 - \delta \sum_{n \geq 1} w(n) \geq 1 - \delta$, l'inégalité précédente est valable pour tout $n \geq 1$, ce qui conclut la preuve. $\diamond$

Posons

$$n(h) = \min\{n \geq 1 : h \in \mathcal{H}_n\}, \quad (26)$$

l'équation (25) implique que pour toute loi de probabilité, avec une probabilité d'au moins $1 - \delta$, pour tout échantillon S de taille m , pour tout $h \in \mathcal{H}$, $R(h) \leq R_S(h) + \epsilon_{n(h)}(m, w[n(h)]\delta)$. La règle de minimisation du risque structurel consiste à minimiser le membre de droite de cette inégalité :

$$\text{SRM}_{\mathcal{H}, w}(S) \in \arg \min_{h \in \mathcal{H}} [R_S(h) + \epsilon_{n(h)}(m, w[n(h)]\delta)]. \quad (27)$$

Théorème 1.20 *Soit un ensemble d'hypothèses $\mathcal{H} = \bigcup_{n=1}^{\infty} \mathcal{H}_n$, où, pour tout $n \geq 1$, $\mathcal{H}_n$ a la propriété de convergence uniforme avec une complexité d'échantillon $m_{\mathcal{H}_n}^{\text{CU}}$. Soit $w : \mathbb{N} \setminus \{0\} \rightarrow [0; 1]$, défini par $w(n) = \frac{6}{\pi^2 n^2}$. Alors $\mathcal{H}$ a la capacité d'apprentissage non uniforme en utilisant la règle de minimisation du risque structurel, avec la taille d'échantillon minimale*

$$m_{\mathcal{H}}^{\text{ANU}}(\varepsilon, \delta, h) \leq m_{\mathcal{H}_{n(h)}}^{\text{CU}} \left(\varepsilon/2, \frac{6\delta}{\pi^2 n^2(h)} \right).$$

Preuve. Soit $\mathbf{A}$ l'algorithme d'apprentissage défini par la relation (27). Pour tous $h \in \mathcal{H}$, $\varepsilon, \delta \in]0; 1[$, soit $m \geq m_{\mathcal{H}_n(h)}^{\text{CU}}(\varepsilon/2, w[n(h)]\delta)$; puisque $\sum_{n=1}^{\infty} w(n) = 1$, il résulte du théorème 1.19 qu'avec une probabilité d'au moins $1 - \delta$ sur le choix de l'échantillon S de taille m , nous avons pour tout $h' \in \mathcal{H}$

$$R(h') \leq R_S(h') + \epsilon_{n(h')}(m, w[n(h')]\delta);$$

en particulier, pour $h' = \mathbf{A}(S)$. De plus, par définition de $\text{SRM}_{\mathcal{H},w}$ nous avons

$$R[\mathbf{A}(S)] \leq \min_{h' \in \mathcal{H}} [R_S(h') + \epsilon_{n(h')}(m, w[n(h')]\delta)] \leq R_S(h) + \epsilon_{n(h)}(m, w[n(h)]\delta).$$

Finalement, puisque $m \geq m_{\mathcal{H}_n(h)}^{\text{CU}}(\varepsilon/2, w[n(h)]\delta)$, on a $\epsilon_{n(h)}(m, w[n(h)]\delta) \leq \varepsilon/2$, et la propriété de convergence uniforme de chaque $\mathcal{H}_n$ impliquant que $R_S(h) \leq R(h) + \varepsilon/2$, nous obtenons $R[\mathbf{A}(S)] \leq R(h) + \varepsilon$, ce qui termine la preuve. $\diamond$

1.12 La règle de longueur de code minimale

Supposons que l'ensemble d'hypothèses $\mathcal{H}$ soit dénombrable, que L soit à valeurs dans $[0, 1]$ et écrivons $\mathcal{H} = \bigcup_{n=1}^{\infty} \{h_n\}$. D'après l'inégalité de Hoeffding, chaque singleton $\{h_n\}$ a la propriété de convergence uniforme avec la complexité $m^{\text{CU}}(\varepsilon, \delta) = \frac{\log(2/\delta)}{2\varepsilon^2}$ et la fonction ϵ_n de la relation (23) devient $\epsilon_n(m, \delta) = \sqrt{\frac{\log(2/\delta)}{2m}}$ et la règle de minimisation du risque structural devient

$$\text{SRM}_{\mathcal{H},w}(S) \in \arg \min_{h_n \in \mathcal{H}} \left[R_S(h_n) + \sqrt{\frac{-\log[w(n)] + \log(2/\delta)}{2m}} \right],$$

de façon équivalente, on peut considérer w comme une fonction de $\mathcal{H}$ dans $[0; 1]$ et écrire

$$\text{SRM}_{\mathcal{H},w}(S) \in \arg \min_{h \in \mathcal{H}} \left[R_S(h) + \sqrt{\frac{-\log[w(h)] + \log(2/\delta)}{2m}} \right].$$

Ainsi, notre connaissance *a priori* est uniquement déterminée par le poids que nous assignons à chaque hypothèse. Le poids est d'autant plus important, que nous croyons l'hypothèse correcte et dans l'algorithme d'apprentissage, nous préférons les hypothèses qui ont les plus grands poids. Considérons une fonction d'encodage des hypothèses

$$\begin{aligned} \sigma : \mathcal{H} &\rightarrow \{0, 1\}^* \\ h &\mapsto \sigma(h) \end{aligned}$$

qui à chaque hypothèse $h \in \mathcal{H}$ associe une suite finie de bits, $\sigma(h)$. La taille de $\sigma(h)$ (exprimée en nombre de bits) est notée $|h|$. On suppose que la fonction d'encodage donne des mots de codes dont aucun est le préfixe d'un autre (on dit alors que le code σ est *instantané*). Il vérifie alors l'inégalité de Kraft :

$$\sum_{h \in \mathcal{H}} \frac{1}{2^{|h|}} \leq 1. \quad (28)$$

En posant $w(h) = 2^{-|h|}$, le théorème suivant résulte du théorème 1.19 avec $\epsilon_n(m, \delta) = \sqrt{\frac{\log(2/\delta)}{2m}}$ en remarquant que $\log(2^{|h|}) = |h| \log(2) < |h|$.

Théorème 1.21 *Supposons que la fonction coût soit à valeurs dans $[0, 1]$. Soient $\mathcal{H}$ un ensemble dénombrable d'hypothèses et $\sigma : \mathcal{H} \rightarrow \{0, 1\}^*$ un code instantané pour $\mathcal{H}$. Alors pour toute taille m d'échantillon, pour toute valeur du paramètre de confiance δ et pour toute loi de probabilité, avec une probabilité d'au moins $1 - \delta$ sur le choix de l'échantillon S de taille m , nous avons*

$$\forall h \in \mathcal{H}, \quad R(h) \leq R_S + \sqrt{\frac{|h| + \log(2/\delta)}{2m}},$$

où $|h|$ est la longueur du mot de code $\sigma(h)$.

2 Preuve du 2^{ème} théorème fondamental d'apprentissage

2.1 Définitions et notations

Soient un ensemble d'hypothèses $\mathcal{H} \in \mathcal{M}(\mathcal{X}, \mathcal{Y})$, une fonction coût $L : \mathcal{Y}^2 \rightarrow \mathbb{R}_+$ et un échantillon $S = ((x_1, y_1), \dots, (x_m, y_m))$ de m réalisations i.i.d. d'un vecteur aléatoire (x, y) à valeurs dans $\mathcal{X} \times \mathcal{Y}$, appelées données d'entraînement.

Nous définissons la *représentativité* de S par rapport à $\mathcal{H}$ et L comme étant l'écart maximal entre le risque réel et le risque empirique de $h \in \mathcal{H}$:

$$\text{Rep}(S, \mathcal{H}, L) \stackrel{\text{def}}{=} \sup_{h \in \mathcal{H}} (R(f) - R_S(f)). \quad (29)$$

C'est une variable aléatoire, fonction de S .

L'ensemble des coûts des observations filtrées (avec contrainte) est

$$\mathcal{L}(\mathcal{H} | S) = \left\{ (L(h(x_i), y_i))_{1 \leq i \leq m} : h \in \mathcal{H} \right\}. \quad (30)$$

2.2 Complexité de Rademacher

Soit $(\sigma_i)_{1 \leq i \leq m}$ une séquence de m variables aléatoires i.i.d. uniformes à valeurs dans $\{-1, +1\}$.

La *complexité de Rademacher* d'un sous-ensemble $\mathcal{A} \subset \mathbb{R}^m$ est définie par

$$\rho(\mathcal{A}) \stackrel{\text{def}}{=} \frac{1}{m} \mathbb{E} \left[\sup_{\underline{a} \in \mathcal{A}} \sum_{i=1}^m \sigma_i a_i \right], \quad (31)$$

où $\underline{a} = (a_1, \dots, a_m)$.

Introduisons des données de test $S' = ((x'_i, y'_i))_{1 \leq i \leq m}$ i.i.d., de même loi que (x, y) et indépendantes de S . Le raisonnement de la preuve du théorème 1.16 peut être refait ici pour montrer que

$$\begin{aligned} \mathbb{E} [\text{Rep}(S, \mathcal{H}, L)] &\leq \mathbb{E} \left[\sup_{h \in \mathcal{H}} R_{S'}(h) - R_S(h) \right] \\ &= \frac{1}{m} \mathbb{E} \left[\sup_{h \in \mathcal{H}} \sum_{i=1}^m \sigma_i (L(h(x'_i), y'_i) - L(h(x_i), y_i)) \right] \\ &\leq \frac{2}{m} \mathbb{E} \left\{ \sup_{h \in \mathcal{H}} \sum_{i=1}^m \sigma_i L[(h(x_i), y_i)] \right\}, \end{aligned}$$

d'où le lemme suivant.

Lemme 2.1

$$\mathbb{E} [\text{Rep}(S, \mathcal{H}, L)] \leq 2 \mathbb{E} \{ \rho[\mathcal{L}(\mathcal{H} | S)] \}.$$

Théorème 2.2 Pour tous $\mathcal{H}$, L et S nous avons

$$\mathbb{E} \{ R[\text{ERM}_{\mathcal{H}}(S)] - R_S[\text{ERM}_{\mathcal{H}}(S)] \} \leq 2 \mathbb{E} \{ \rho[\mathcal{L}(\mathcal{H} | S)] \}.$$

De plus, pour tout $h \in \mathcal{H}$

$$\mathbb{E} \{ R[\text{ERM}_{\mathcal{H}}(S)] \} - R(h) \leq 2 \mathbb{E} \{ \rho[\mathcal{L}(\mathcal{H} | S)] \}$$

et pour tout $\delta \in]0; 1[$, avec une probabilité d'au moins $1 - \delta$ sur le choix de S nous avons

$$R[\text{ERM}_{\mathcal{H}}(S)] - \min_{h \in \mathcal{H}} R(h) \leq \frac{2 \mathbb{E} \{ \rho[\mathcal{L}(\mathcal{H} | S)] \}}{\delta}.$$

Preuve. La première inégalité résulte du lemme 2.1, la deuxième de l'optimalité de $\text{ERM}_{\mathcal{H}}(S)$: $R_S(h) \geq R_S[\text{ERM}_{\mathcal{H}}(S)]$, donc $R(h) \geq \mathbb{E}\{R_S[\text{ERM}_{\mathcal{H}}(S)]\}$ et la troisième de l'inégalité de Markov en remarquant que la VA $R[\text{ERM}_{\mathcal{H}}(S)] - \min_{h \in \mathcal{H}} R(h)$ est positive. $\diamond$

Lemme 2.3 (Inégalité de Mc Diarmid) Soient $X_1, \dots, X_m$ m VA indépendantes à valeurs dans $\mathcal{X}$ et $f : \mathcal{X}^m \rightarrow \mathbb{R}$. Si pour tout $i \in \llbracket 1; m \rrbracket$ il existe $c_i > 0$ tel que pour tous $x_1, \dots, x_m, x'_i \in \mathcal{X}$ on ait

$$|f(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_m) - f(x_1, \dots, x_{i-1}, x'_i, x_{i+1}, \dots, x_m)| \leq c_i, \quad (32)$$

alors pour tout $\varepsilon > 0$, on a

$$\begin{aligned} \mathbb{P}\{f(X_1, \dots, X_m) - \mathbb{E}[f(X_1, \dots, X_m)] \geq \varepsilon\} &\leq \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^m c_i^2}\right) \\ \mathbb{P}\{f(X_1, \dots, X_m) - \mathbb{E}[f(X_1, \dots, X_m)] \leq -\varepsilon\} &\leq \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^m c_i^2}\right) \\ \mathbb{P}\{|f(X_1, \dots, X_m) - \mathbb{E}[f(X_1, \dots, X_m)]| \geq \varepsilon\} &\leq 2 \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^m c_i^2}\right). \end{aligned}$$

Une preuve de ce lemme est donnée en exercice.

Théorème 2.4 Supposons qu'il existe $c > 0$ tel que pour tout $h \in \mathcal{H}$ et tout $(x, y) \in \mathcal{X} \times \mathcal{Y}$, $|L(h(x), y)| \leq c$, alors pour tout $\delta \in]0; 1[$

(i) avec une probabilité d'au moins $1 - \delta$ sur l'échantillon S , pour tout $h \in \mathcal{H}$

$$R(h) - R_S(h) \leq 2 \mathbb{E}\{\rho[\mathcal{L}(\mathcal{H} | S)]\} + c \sqrt{\frac{2 \log(1/\delta)}{m}}, \quad (33)$$

en particulier pour $h = \text{ERM}_{\mathcal{H}}(S)$;

(ii) avec une probabilité d'au moins $1 - \delta$ sur l'échantillon S , pour tout $h \in \mathcal{H}$

$$R(h) - R_S(h) \leq 2\rho[\mathcal{L}(\mathcal{H} | S)] + 3c \sqrt{\frac{2 \log(2/\delta)}{m}}, \quad (34)$$

en particulier pour $h = \text{ERM}_{\mathcal{H}}(S)$;

(iii) avec une probabilité d'au moins $1 - \delta$ sur l'échantillon S , pour tout $h \in \mathcal{H}$

$$R(\text{ERM}_{\mathcal{H}}(S)) - R(h) \leq 2\rho[\mathcal{L}(\mathcal{H} | S)] + 4c \sqrt{\frac{2 \log(4/\delta)}{m}},$$

Preuve. Remarquons que pour tout échantillon S de taille m et toute fonction $h \in \mathcal{H}$, la fonction $R_S(h)$ vérifie la propriété (32) avec $c_i = \frac{2c}{m}$, donc la fonction $\text{Rep}(S, \mathcal{H}, L) = \sup_{h \in \mathcal{H}} (R(h) - R_S(h))$ vérifie également la propriété (32) avec la même valeur $c_i = \frac{2c}{m}$. Il résulte alors de l'inégalité de Mc Diarmid appliquée à $\text{Rep}(S, \mathcal{H}, L)$ qu'avec une probabilité d'au moins $1 - \delta$

$$\text{Rep}(S, \mathcal{H}, L) \leq \mathbb{E}[\text{Rep}(S, \mathcal{H}, L)] + c \sqrt{\frac{2 \log(1/\delta)}{m}}$$

et l'inégalité (33) se déduit du lemme 2.1 et de la définition de $\text{Rep}(S, \mathcal{H}, L)$. De même, la fonction $\rho[\mathcal{L}(\mathcal{H} | S)]$ vérifie la propriété (32) avec $c_i = \frac{2c}{m}$, il résulte alors de l'inégalité de Mc Diarmid appliquée à $\rho[\mathcal{L}(\mathcal{H} | S)]$ pour $\delta/2$, de l'inégalité (33) appliquée pour $\delta/2$ que l'inégalité (34) est valide. Enfin, en posant $h_S = \text{ERM}_{\mathcal{H}}(S)$, on a pour tout $h \in \mathcal{H}$

$$\begin{aligned} R(h_S) - R(h) &= R(h_S) - R_S(h_S) + R_S(h_S) - R_S(h) + R_S(h) - R(h) \\ &\leq R(h_S) - R_S(h_S) + R_S(h) - R(h). \end{aligned}$$

Le premier terme du membre de droite de la dernière inégalité est majoré avec une probabilité d'au moins $1 - \delta/2$ grâce à l'inégalité (34) et le deuxième terme, où h^* ne dépend pas de S , est majoré grâce à l'inégalité de Hoeffding avec une probabilité d'au moins $1 - \delta/2$ par

$$R_S(h) - R() \leq c \sqrt{\frac{2 \log(2/\delta)}{m}},$$

ce qui termine la démonstration. $\diamond$

Terminons ce paragraphe par deux lemmes sur la complexité de Rademacher. La preuve du premier est immédiate.

Lemme 2.5 *Pour $\mathcal{A} \subset \mathbb{R}^m$, $c \in \mathbb{R}$ et $\mathbf{a}_0 \in \mathbb{R}^m$, $\rho(c\mathcal{A} + \mathbf{a}_0) = |c|\rho(\mathcal{A})$.*

Lemme 2.6 (Lemme de Massart) *Soient $\mathcal{A} = \{\mathbf{a}_1, \dots, \mathbf{a}_N\} \subset \mathbb{R}^m$ et $\bar{\mathbf{a}} = \frac{1}{N} \sum_{i=1}^N \mathbf{a}_i$, alors*

$$\rho(\mathcal{A}) \leq \max_{\mathbf{a} \in \mathcal{A}} \|\mathbf{a} - \bar{\mathbf{a}}\| \frac{\sqrt{2 \log(N)}}{m}.$$

Preuve. Supposons sans perte de généralité que $\bar{\mathbf{a}} = 0$ (d'après le lemme 2.5). Soit $\lambda > 0$, posons $\mathcal{A}' = \lambda\mathcal{A}$. On a

$$\begin{aligned} m\rho(\mathcal{A}') &= \mathbb{E} \left(\max_{\mathbf{a} \in \mathcal{A}'} \sum_{i=1}^m \sigma_i a_i \right) = \mathbb{E} \left[\log \left(\max_{\mathbf{a} \in \mathcal{A}'} \exp \left(\sum_{i=1}^m \sigma_i a_i \right) \right) \right] \\ &\leq \mathbb{E} \left[\log \left(\sum_{\mathbf{a} \in \mathcal{A}'} \exp \left(\sum_{i=1}^m \sigma_i a_i \right) \right) \right] \leq \log \left[\mathbb{E} \left(\sum_{\mathbf{a} \in \mathcal{A}'} \exp \left(\sum_{i=1}^m \sigma_i a_i \right) \right) \right] \\ &= \log \left[\sum_{\mathbf{a} \in \mathcal{A}'} \prod_{i=1}^m \mathbb{E} (e^{\sigma_i a_i}) \right] \leq \log \left[\sum_{\mathbf{a} \in \mathcal{A}'} \prod_{i=1}^m e^{a_i^2/2} \right] = \log \left[\sum_{\mathbf{a} \in \mathcal{A}'} e^{\|\mathbf{a}\|^2/2} \right] \\ &\leq \log \left(|\mathcal{A}'| \max_{\mathbf{a} \in \mathcal{A}'} e^{\|\mathbf{a}\|^2/2} \right) = \log(|\mathcal{A}'|) + \max_{\mathbf{a} \in \mathcal{A}'} \|\mathbf{a}\|^2/2. \end{aligned}$$

Puisque $\rho(\mathcal{A}) = \frac{1}{\lambda} \rho(\mathcal{A}')$, nous obtenons

$$\rho(\mathcal{A}) \leq \frac{\log(|\mathcal{A}|) + \lambda^2 \max_{\mathbf{a} \in \mathcal{A}} \|\mathbf{a}\|^2/2}{\lambda m},$$

et l'inégalité obtenue pour $\lambda = \sqrt{2 \log(|\mathcal{A}|) / \max_{\mathbf{a} \in \mathcal{A}} \|\mathbf{a}\|^2}$ termine la preuve. $\diamond$

2.3 Preuve de la borne supérieure

Nous allons démontrer une borne supérieure un peu plus faible que celle du deuxième théorème fondamental d'apprentissage, à savoir

$$m_{\mathcal{H}}(\varepsilon, \delta) \leq C \frac{d \log(1/\varepsilon) + \log(1/\delta)}{\varepsilon^2}. \quad (35)$$

Soit $((x_1, y_1), \dots, (x_m, y_m))$ des données d'entraînement pour un problème de classification. Il résulte du lemme 1.12 de Sauer-Shelah-Perles que si $\dim \text{VC}(\mathcal{H}) = d$, alors

$$|\mathcal{L}(\mathcal{H} | S)| \leq \left(\frac{em}{d} \right)^d$$

et donc $\mathcal{A} = \{(\mathbb{1}_{h(x_1) \neq y_1}, \dots, \mathbb{1}_{h(x_m) \neq y_m}) : h \in \mathcal{H}\}$ vérifie $|\mathcal{A}| \leq \left(\frac{em}{d} \right)^d$, qui combiné avec le lemme 2.6 de Massart, donne $\rho(\mathcal{A}) \leq \sqrt{\frac{2d \log(em/d)}{m}}$. Enfin, il résulte du théorème 2.4 (ii) qu'avec une probabilité d'au moins $1 - \delta$, pour tout $h \in \mathcal{H}$

$$R(h) - R_S(h) \leq 2 \sqrt{\frac{2d \log(em/d)}{m}} + 3 \sqrt{\frac{2 \log(2/\delta)}{m}}$$

donc avec une probabilité d'au moins $1 - \delta$, pour tout $h \in \mathcal{H}$

$$|R(h) - R_S(h)| \leq 2\sqrt{\frac{2d \log(em/d)}{m}} + 3\sqrt{\frac{2 \log(4/\delta)}{m}} \leq 2\sqrt{\frac{4d \log(em/d) + 9 \log(4/\delta)}{m}}$$

et cette quantité est majorée par ε si et seulement si

$$\frac{\varepsilon^2 m}{16d} \geq 1 + \log\left(\frac{\varepsilon^2 m}{16d}\right) + \frac{9}{4d} \log\left(\frac{4}{\delta}\right) + \log\left(\frac{16}{\varepsilon^2}\right),$$

i.e., si et seulement si $g\left(\frac{\varepsilon^2 m}{16d}\right) \geq \frac{9}{4d} \log\left(\frac{4}{\delta}\right) + \log\left(\frac{16}{\varepsilon^2}\right)$, où $g(u) = u - \log(u) - 1$, définie pour $u > 0$, est convexe positive et atteint son minimum (qui s'annule) en $u = 1$; donc $g(u/2) \geq 0$, ce qui entraîne que $g(u) \geq u/2 - \log(2)$. Ainsi, si $m \geq \frac{32d}{\varepsilon^2} \left[\frac{9}{4d} \log\left(\frac{4}{\delta}\right) + \log\left(\frac{16}{\varepsilon^2}\right) + \log(2) \right]$, alors avec une probabilité d'au moins $1 - \delta$ pour tout $h \in \mathcal{H}$, $|R(h) - R_S(h)| \leq \varepsilon$ et nous avons montré qu'il existe une constante $C > 0$ pour laquelle l'inégalité (35) est vérifiée.

2.4 Démonstration de la borne inférieure

Lemme 2.7 Soit X une loi binomiale $\mathcal{B}(m, p)$ avec $p = (1 - \varepsilon)/2$, alors

$$\mathbb{P}(X \geq m/2) \geq \frac{1}{2} \left(1 - \sqrt{1 - \exp(-m\varepsilon^2/(1 - \varepsilon^2))} \right).$$

Preuve. La preuve repose sur l'inégalité de Slud⁸ [3] : si $p \leq 1/2$ et m pair alors

$$\mathbb{P}(X \geq m/2) \geq 1 - \Phi\left(\frac{m/2 - mp}{\sqrt{mp(1 - p)}}\right),$$

où Φ est la fonction de répartition d'une VA réelle gaussienne centrée réduite. Ici,

$$\frac{m/2 - mp}{\sqrt{mp(1 - p)}} = \sqrt{\frac{m\varepsilon^2}{1 - \varepsilon^2}}.$$

En posant pour $u \geq 0$, $g(u) = 2(1 - \Phi(u))$ et $f(u) = 1 - \sqrt{1 - e^{-u^2}}$, nous allons montrer que $g(u) \geq f(u)$ pour $u \geq 0$. On a $g(0) - f(0) = 0 = \lim_{u \rightarrow \infty} g(u) - f(u)$,

$$g'(u) - f'(u) = \frac{ue^{-u^2}}{\sqrt{1 - e^{-u^2}}} - \sqrt{\frac{2}{\pi}} e^{-u^2/2},$$

$\lim_{u \rightarrow 0^+} g'(u) - f'(0) = 1 - \sqrt{\frac{2}{\pi}} > 0$, $\lim_{u \rightarrow \infty} g'(u) - f'(u) = 0$ et $g'(u) - f'(u) = 0$ si et seulement si

$$h(u) \stackrel{\text{def}}{=} \frac{ue^{-u^2/2}}{\sqrt{1 - e^{-u^2}}} = \sqrt{\frac{2}{\pi}}. \quad (36)$$

Enfin, $h'(u) = \frac{-e^{-u^2/2} (e^{-u^2} - (1 - u^2))}{(1 - e^{-u^2})^{3/2}} < 0$, $h(0) = 1 > \sqrt{2/\pi} > 0 = \lim_{u \rightarrow \infty} h(u)$, donc l'équation (36) n'a qu'une racine $u_0 > 0$ et $g'(u) - f'(u)$ ne s'annule qu'une fois, ainsi $g(u) - f(u)$ est croissante sur $[0, u_0]$, puis décroissante sur $[u_0, \infty[$ et $g(u) \geq f(u)$ pour $u \geq 0$. $\diamond$

Pour montrer qu'il existe $C > 0$ tel que $m_{\mathcal{H}}(\varepsilon, \delta) \geq C \frac{d + \log(1/\delta)}{\varepsilon^2}$, nous procédons en deux étapes. Nous commençons par montrer que $m_{\mathcal{H}}(\varepsilon, \delta) \geq (\frac{1}{2}) \log[1/(4\delta)]/\varepsilon^2$, puis nous montrons que pour tout $\delta \leq 1/8$, nous avons $m_{\mathcal{H}}(\varepsilon, \delta) \geq 8d/\varepsilon^2$.

8. Pour une preuve, voir [4].

Montrons que pour tout $\varepsilon < \frac{1}{\sqrt{2}}$ et tout $\delta \in]0; 1[$, nous avons $m_{\mathcal{H}}(\varepsilon, \delta) \geq (\frac{1}{2}) \log[1/(4\delta)]/\varepsilon^2$. Pour cela, remarquons que c'est trivial si $\delta \geq \frac{1}{4}$, nous supposons donc $\delta < \frac{1}{4}$, et montrons que si $m \leq (\frac{1}{2}) \log[1/(4\delta)]/\varepsilon^2$, alors $\mathcal{H}$ n'a pas la capacité PAC d'apprentissage. Considérons un exemple $c \in \mathcal{X}$ qui est brisé par $\mathcal{H}$, i.e., il existe $h_+, h_- \in \mathcal{H}$ tels que $h_+(c) = 1$ et $h_-(c) = -1$. Considérons les deux lois de probabilité $\mathcal{D}_+$ et $\mathcal{D}_-$ définies par

$$\mathcal{D}_b(\{(x, y)\}) = \begin{cases} \frac{1by\varepsilon}{2} & \text{si } x = c \\ 0 & \text{sinon.} \end{cases} \quad (37)$$

avec $b \in \{+, -\}$. Soient un algorithme d'apprentissage quelconque $\mathbf{A}$ et la loi de probabilité $\mathcal{D}_b$, un échantillon de m données d'entraînement est alors donné par $S = ((c, y_1), \dots, (c, y_m))$. Il est donc entièrement caractérisé par le vecteur $\mathbf{y} = (y_1, \dots, y_m) \in \{-1, +1\}^m$ et $\mathbf{A}(S) = h_S$ est caractérisé par $h_S(c) \in \{-1, +1\}$. Ainsi l'algorithme d'apprentissage peut être considéré comme une application de $\{-1, +1\}^m$ dans $\{-1, +1\}$ et l'on note $\mathbf{A}(\mathbf{y}) = h_S(c)$. Remarquons que pour tout $h \in \mathcal{H}$, on a $R(h) = \mathbb{P}[h(c) \neq y] = \frac{1-h(c)b\varepsilon}{2}$, en particulier pour le classificateur optimal de Bayes h_B on a

$$R[\mathbf{A}(S)] - R(h_B) = \frac{1 - \mathbf{A}(\mathbf{y})b\varepsilon}{2} - \frac{1 - \varepsilon}{2} = \begin{cases} \varepsilon & \text{si } \mathbf{A}(\mathbf{y}) \neq b \\ 0 & \text{sinon.} \end{cases}.$$

Fixons l'algorithme d'apprentissage $\mathbf{A}$. On a

$$\mathbb{P}\{R[\mathbf{A}(S)] - R(h_B) = \varepsilon\} = \sum_{\mathbf{y} \in \{-1, +1\}^m} \mathcal{D}_b^m(\mathbf{y}) \mathbb{1}_{\mathbf{A}(\mathbf{y}) \neq b},$$

notons $\mathcal{N}^+ = \{\mathbf{y} \in \{-1, +1\}^m : |\{i \in \llbracket 1; m \rrbracket : y_i = 1\}| \geq m/2\}$ et $\mathcal{N}^- = \{-1, +1\}^m \setminus \mathcal{N}^+$. Remarquons que pour tout $\mathbf{y} \in \mathcal{N}^+$, on a $\mathcal{D}_+^m(\mathbf{y}) \geq \mathcal{D}_-^m(\mathbf{y})$ et pour tout $\mathbf{y} \in \mathcal{N}^-$, on a $\mathcal{D}_-^m(\mathbf{y}) \geq \mathcal{D}_+^m(\mathbf{y})$. Par conséquent,

$$\begin{aligned} \max_{b \in \{-1, +1\}} \mathbb{P}\{R[\mathbf{A}(S)] - R(h_B) = \varepsilon\} &= \max_{b \in \{-1, +1\}} \sum_{\mathbf{y} \in \{-1, +1\}^m} \mathcal{D}_b^m(\mathbf{y}) \mathbb{1}_{\mathbf{A}(\mathbf{y}) \neq b} \\ &\geq \frac{1}{2} \sum_{\mathbf{y} \in \{-1, +1\}^m} \mathcal{D}_+^m(\mathbf{y}) \mathbb{1}_{\mathbf{A}(\mathbf{y}) \neq +1} + \frac{1}{2} \sum_{\mathbf{y} \in \{-1, +1\}^m} \mathcal{D}_-^m(\mathbf{y}) \mathbb{1}_{\mathbf{A}(\mathbf{y}) \neq -1} \\ &= \frac{1}{2} \sum_{\mathbf{y} \in \mathcal{N}^+} [\mathcal{D}_+^m(\mathbf{y}) \mathbb{1}_{\mathbf{A}(\mathbf{y}) \neq +1} + \mathcal{D}_-^m(\mathbf{y}) \mathbb{1}_{\mathbf{A}(\mathbf{y}) \neq -1}] \\ &\quad + \frac{1}{2} \sum_{\mathbf{y} \in \mathcal{N}^-} [\mathcal{D}_+^m(\mathbf{y}) \mathbb{1}_{\mathbf{A}(\mathbf{y}) \neq +1} + \mathcal{D}_-^m(\mathbf{y}) \mathbb{1}_{\mathbf{A}(\mathbf{y}) \neq -1}] \\ &\geq \frac{1}{2} \sum_{\mathbf{y} \in \mathcal{N}^+} [\mathcal{D}_-^m(\mathbf{y}) \mathbb{1}_{\mathbf{A}(\mathbf{y}) \neq +1} + \mathcal{D}_-^m(\mathbf{y}) \mathbb{1}_{\mathbf{A}(\mathbf{y}) \neq -1}] \\ &\quad + \frac{1}{2} \sum_{\mathbf{y} \in \mathcal{N}^-} [\mathcal{D}_+^m(\mathbf{y}) \mathbb{1}_{\mathbf{A}(\mathbf{y}) \neq +1} + \mathcal{D}_+^m(\mathbf{y}) \mathbb{1}_{\mathbf{A}(\mathbf{y}) \neq -1}] = \frac{1}{2} \sum_{\mathbf{y} \in \mathcal{N}^+} \mathcal{D}_-^m(\mathbf{y}) + \frac{1}{2} \sum_{\mathbf{y} \in \mathcal{N}^-} \mathcal{D}_+^m(\mathbf{y}). \end{aligned}$$

Remarquons maintenant que $\sum_{\mathbf{y} \in \mathcal{N}^-} \mathcal{D}_+^m(\mathbf{y}) = \sum_{\mathbf{y} \in \mathcal{N}^+} \mathcal{D}_-^m(\mathbf{y})$ et que chacune de ces probabilité est égale à la probabilité pour qu'une loi binomiale $\mathcal{B}(m, (1-\varepsilon)/2)$ prenne une valeur supérieure ou égale à $m/2$. D'après le lemme 2.7, cette probabilité est minorée par

$$\frac{1}{2} \left(1 - \sqrt{1 - \exp(-m\varepsilon^2/(1-\varepsilon^2))} \right) \geq \frac{1}{2} \left(1 - \sqrt{1 - \exp(-2m\varepsilon^2)} \right),$$

où nous avons utilisé la condition $\varepsilon^2 \leq 1/2$. On en déduit que si $m \leq (\frac{1}{2}) \log(1/(4\delta))/\varepsilon^2$, alors il existe b tel que $\mathcal{D}_b^m \{R[\mathbf{A}(S)] - R(h_B) = \varepsilon\} \geq \frac{1}{2} (1 - \sqrt{1 - 4\delta}) \geq \delta$. Ceci termine la première étape.

Montrons maintenant que $m(\varepsilon, 1/8) \geq 8d/\varepsilon^2$ dès que $\varepsilon < 1/(8\sqrt{2})$. Soit $\rho = 8\varepsilon \in]0; 1/\sqrt{2}[$ et $\mathcal{C} = \{c_1, \dots, c_d\} \subset \mathcal{X}$ qui est brisé par $\mathcal{H}$. Pour chaque vecteur $b = (b_1, \dots, b_d) \in \{-1, +1\}^d$ définissons une loi de probabilité $\mathcal{D}_b$ sur $\mathcal{X} \times \mathcal{Y}$ par

$$\mathcal{D}_b(\{(x, y)\}) = \begin{cases} \frac{1}{d} \cdot \frac{1+y b_i \rho}{2} & \text{si } \exists i \in \llbracket 1; d \rrbracket : x = c_i \\ 0 & \text{sinon.} \end{cases}$$

Le classificateur de Bayes pour $\mathcal{D}_b$ est défini par $h_B(c_i) = b_i$ pour $1 \leq i \leq d$ et son risque réel vaut $R(h_B) = \mathbb{P}[h_B(x) \neq y] = \frac{1-\rho}{2}$. De plus, pour toute hypothèse $h \in \mathcal{H}$, son risque vaut

$$R(h) = \frac{1+\rho}{2} \cdot \frac{|\{i \in \llbracket 1; d \rrbracket : h(c_i) \neq b_i\}|}{d} + \frac{1-\rho}{2} \cdot \frac{|\{i \in \llbracket 1; d \rrbracket : h(c_i) = b_i\}|}{d},$$

donc

$$R(h) - \min_{g \in \mathcal{H}} R(g) = \rho \cdot \frac{|\{i \in \llbracket 1; d \rrbracket : h(c_i) \neq b_i\}|}{d}. \quad (38)$$

Maintenant, fixons un algorithme d'apprentissage $\mathbf{A}$ et, comme dans la preuve du théorème *no free lunch*, en supposant b aléatoire avec une loi uniforme nous avons

$$\begin{aligned} \max_{\mathcal{D}_b : b \in \{-1, +1\}^d} \mathbb{E} \left[R[\mathbf{A}(S)] - \min_{g \in \mathcal{H}} R(g) \mid b \right] \\ \geq \mathbb{E} \left[R[\mathbf{A}(S)] - \min_{g \in \mathcal{H}} R(g) \right] = \mathbb{E} \left[\rho \cdot \frac{|\{i \in \llbracket 1; d \rrbracket : \mathbf{A}(S)(c_i) \neq b_i\}|}{d} \right] \end{aligned} \quad (39)$$

$$= \frac{\rho}{d} \sum_{i=1}^d \mathbb{E} [\mathbb{1}_{[\mathbf{A}(S)(c_i) \neq b_i]}]. \quad (40)$$

Le tirage d'un échantillon S de taille m suivant la loi $\mathcal{D}_b$ se fait de la manière suivante : choisir au hasard m entiers $j = (j_1, \dots, j_m)$ dans $\llbracket 1; d \rrbracket$ (m tirages indépendants avec remise), on a alors $x_\ell = c_{j_\ell}$ ($1 \leq \ell \leq m$) et pour chaque ℓ , choisir $y_\ell = b_{j_\ell}$ avec une probabilité de $\frac{1+\rho}{2}$. Ainsi, on peut remplacer l'échantillon S par le couple (j, y) où j est défini ci-dessus et $y = (y_1, \dots, y_m)$. Fixons $i \in \llbracket 1; d \rrbracket$, notons $b^{(-i)} = (b_1, \dots, b_{i-1}, b_{i+1}, \dots, b_d)$, et pour j fixé, notons $I(j) = \{\ell \in \llbracket 1; m \rrbracket : j_\ell = i\}$ et $I^c(j) = \llbracket 1; m \rrbracket \setminus I(j)$, on a

$$\mathbb{P}(y \mid b, j) = \mathbb{P}((y_\ell)_{\ell \in I^c(j)}, (y_\ell)_{\ell \in I(j)} \mid b^{(-i)}, b_i, j) = \mathbb{P}((y_\ell)_{\ell \in I^c(j)} \mid b^{(-i)}, j) \mathbb{P}((y_\ell)_{\ell \in I(j)} \mid b_i, j),$$

donc

$$\begin{aligned} \mathbb{E} [\mathbb{1}_{[\mathbf{A}(S)(c_i) \neq b_i]} \mid j] &= \frac{1}{2} \sum_{b_i \in \{-1, +1\}} \mathbb{E} [\mathbb{1}_{[\mathbf{A}(j, y)(c_i) \neq b_i]} \mid b_i, j] \\ &= \frac{1}{2^{d-1}} \sum_{b^{(-i)}} \sum_{(y_\ell)_{\ell \in I^c(j)}} \mathbb{P}((y_\ell)_{\ell \in I^c(j)} \mid b^{(-i)}, j) \frac{1}{2} \sum_{(y_\ell)_{\ell \in I(j)}} \left(\sum_{b_i} \mathbb{P}((y_\ell)_{\ell \in I(j)} \mid b_i, j) \mathbb{1}_{[\mathbf{A}(j, y)(c_i) \neq b_i]} \right). \end{aligned}$$

Le terme entre parenthèses est minimisé quand $\mathbf{A}(j, y)(c_i)$ maximise $\mathbb{P}((y_\ell)_{\ell \in I(j)} \mid b_i, j)$, c'est-à-dire quand il correspond au maximum de vraisemblance :

$$\mathbf{A}(j, y)(c_i) = \sum_{\ell \in I(j)} y_\ell, \quad (41)$$

correspondant au vote majoritaire entre les (y_ℓ) pour lesquels $\ell \in I(j)$ (i.e., $j_\ell = i$). Minorons maintenant l'expression (40) quand $\mathbf{A}$ est l'estimateur du maximum de vraisemblance (41). Pour $i \in \llbracket 1; d \rrbracket$ et $j \in \llbracket 1; d \rrbracket^m$ fixés, $\mathbb{E} [\mathbb{1}_{[\mathbf{A}(S)(c_i) \neq b_i]} \mid j]$ est la probabilité pour qu'une loi binomiale $\mathcal{B}(|I(j)|, \frac{1-\rho}{2})$ prenne une valeur supérieure ou égale à $\frac{|I(j)|}{2}$, or quand $2\rho^2 \leq 1$, cette probabilité est minorée par

$$\frac{1}{2} \left(1 - \sqrt{1 - \exp(-2|I(j)|\rho^2)} \right) \geq \frac{1}{2} \left(1 - \sqrt{2|I(j)|\rho^2} \right).$$

Maintenant, en moyennant suivant j , il résulte de l'inégalité de Jensen (la racine carrée étant concave) que

$$\mathbb{E} [\mathbb{1}_{[\mathbf{A}(S)(c_i) \neq b_i]}] \geq \frac{1}{2} \left(1 - \rho \sqrt{2 \mathbb{E}[|I(j)|]} \right) = \frac{1}{2} \left(1 - \rho \sqrt{\frac{m}{d}} \right)$$

donc le minimum de (40) est minoré par $\frac{\rho}{2} (1 - \rho \sqrt{\frac{m}{d}})$ et tant que $m < \frac{d}{8\rho^2}$, ce terme sera minoré par $\rho/4$.

En résumé, nous avons montré que si $m < \frac{d}{8\rho^2}$, alors pour tout algorithme d'apprentissage il existe une loi de probabilité telle que $\mathbb{E}[R[\mathbf{A}(S)] - \min_{h \in \mathcal{H}} R_S(h)] \geq \rho/4$. Finalement, en posant $\Delta = \frac{1}{\rho} (R[\mathbf{A}(S)] - \min_{h \in \mathcal{H}} R_S(h))$ et en remarquant que $0 \leq \Delta \leq 1$ (d'après (39)), il résulte de l'inégalité de Markov que $\mathbb{P}(\Delta > \frac{\varepsilon}{\rho}) \geq \mathbb{E}[\Delta] - \frac{\varepsilon}{\rho} \geq \frac{1}{4} - \frac{\varepsilon}{\rho}$.

En choisissant $\rho = 8\varepsilon$, nous concluons que si $m < \frac{8d}{\varepsilon^2}$, alors avec une probabilité d'au moins $1/8$, nous avons $R[\mathbf{A}(S)] - \min_{h \in \mathcal{H}} R_S(h) \geq \varepsilon$. $\diamond$

3 Bornes bayésiennes PAC pour les déviations des moyennes empiriques par rapport à leurs espérances

Cette section est très largement inspirée d'un cours d'Olivier Catoni du 12-06-2013, disponible sur internet à l'adresse <http://ocatoni.perso.math.cnrs.fr/learning2013-02.pdf>, voir également [5].

Nous commençons par généraliser l'inégalité de Hoeffding à des variables aléatoires réelles du deuxième ordre en suivant un raisonnement que nous reproduirons à la section suivante pour établir des bornes bayésiennes PAC sur l'écart entre des moyennes empiriques et leurs espérances. Cela nous permettra de démontrer la propriété de convergence uniforme d'espaces d'hypothèses mesurables, pour n'importe quelle fonction coût.

3.1 Déviations de sommes de VA indépendantes

Soit $(X_i)_{1 \leq i \leq m}$ un échantillon de m variables aléatoires (VA) réelles indépendantes et

$$M = \frac{1}{m} \sum_{i=1}^m X_i$$

leur moyenne empirique. Soit $\mu = \mathbb{E}[M]$. Considérons les fonctions génératrices des moments $\phi_i(\lambda) = \mathbb{E}[\exp(\lambda X_i)]$ et leurs logarithmes

$$\begin{aligned} \psi_i(\lambda) &= \log \{ \mathbb{E}[\exp(\lambda X_i)] \} \\ \psi(\lambda) &= \frac{1}{m} \sum_{i=1}^m \psi_i(\lambda). \end{aligned}$$

Ces derniers sont des fonctions convexes⁹, nulles en 0, à valeurs dans $\mathbb{R} \cup \{+\infty\}$.

Considérons la fonction duale

$$\psi^*(x) = \sup_{\lambda \in \mathbb{R}_+} [\lambda x - \psi(\lambda)],$$

qui est à valeurs dans $\mathbb{R} \cup \{+\infty\}$.

9. En effet, pour λ, λ' et $\alpha \in]0, 1[$, l'inégalité de Hölder appliquée à $f = e^{\alpha \lambda X_i}$ et $g = e^{(1-\alpha)\lambda' X_i}$ s'écrit

$$\phi_i(\alpha\lambda + (1-\alpha)\lambda') = \mathbb{E}[fg] \leq \left(\mathbb{E}\left[f^{\frac{1}{\alpha}}\right] \right)^\alpha \left(\mathbb{E}\left[g^{\frac{1}{1-\alpha}}\right] \right)^{1-\alpha} = \phi_i(\lambda)^\alpha \phi_i(\lambda')^{1-\alpha},$$

et en prenant le logarithme on obtient : $\psi_i(\alpha\lambda + (1-\alpha)\lambda') \leq \alpha\psi_i(\lambda) + (1-\alpha)\psi_i(\lambda')$ (voir également la proposition 3.3 page 24 et sa démonstration).

Proposition 3.1 *Pour tout $x \in \mathbb{R}$, $\mathbb{P}(M \geq x) \leq \exp[-m\psi^*(x)]$.*

Preuve. Pour tout $x \in \mathbb{R}$ et tout $\lambda \in \mathbb{R}_+^*$ on a

$$\begin{aligned}\mathbb{P}(M \geq x) &= \mathbb{E}\{\mathbb{1}[\exp(m\lambda(M-x)) \geq 1]\} \\ &\leq \mathbb{E}\{\exp[m\lambda(M-x)]\} = \exp\{m[\psi(\lambda) - \lambda x]\}\end{aligned}$$

et par conséquent $\mathbb{P}(M \geq x) \leq \inf_{\lambda \in \mathbb{R}_+} \exp\{-m[\lambda x - \psi(\lambda)]\} = \exp[-m\psi^*(x)]$. $\diamond$

Proposition 3.2 *Posons $\Lambda_i = \sup\{\lambda \in \mathbb{R}_+ : \phi_i(\lambda) < \infty\}$ et $\Lambda = \min_{1 \leq i \leq m} \Lambda_i$. Alors, pour tout $\lambda \in [0, \Lambda_i[$, $\phi_i(\lambda) < \infty$ et ϕ_i est $\mathcal{C}^\infty$ sur $]0, \Lambda_i[$ lorsque $\Lambda_i > 0$. Si de plus $\mathbb{E}(|X_i|^k) < \infty$, alors ϕ_i est $\mathcal{C}^k$ sur $[0, \Lambda_i[$.*

Preuve. Soit $\lambda \in [0, \Lambda_i[$. Par définition de Λ_i , il existe $\beta > \lambda$ tel que $\phi_i(\beta) < \infty$ et d'après l'inégalité de Jensen $\phi_i(\beta) = \mathbb{E}\left[(e^{\lambda X_i})^{\frac{\beta}{\lambda}}\right] \geq \phi_i(\lambda)^{\frac{\beta}{\lambda}}$, donc $\phi_i(\lambda) < \infty$.

Pour montrer que ϕ_i est $\mathcal{C}^\infty$, fixons $0 < \alpha < \beta < \Lambda_i$, remarquons que $\forall j \geq 1$, $X_i^{j-1}e^{\beta X_i} = X_i^{j-1}e^{\alpha X_i} + \int_\alpha^\beta X_i^j e^{\lambda X_i} d\lambda$. De plus, $\mathbb{E}[\sup_{\lambda \in [\alpha, \beta]} |X_i^j e^{\lambda X_i}|] < \infty$, car pour $\gamma \in]\beta, \Lambda_i[$, en posant $c_1 = \sup_{x \in \mathbb{R}_+} x^j e^{-(\gamma-\beta)x}$ et $c_2 = \sup_{x \in \mathbb{R}_+} x^j e^{-\alpha x}$, on a

$$|X_i^j e^{\lambda X_i}| \leq \begin{cases} c_1 e^{\gamma X_i} & \text{si } X_i \geq 0 \\ c_2 & \text{si } X_i \leq 0 \end{cases},$$

donc $\mathbb{E}[\sup_{\lambda \in [\alpha, \beta]} |X_i^j e^{\lambda X_i}|] \leq c_1 \mathbb{E}[e^{\gamma X_i}] + c_2 < \infty$. On peut ainsi appliquer le théorème de Fubini pour obtenir

$$\mathbb{E}[X_i^{j-1}e^{\beta X_i}] = \mathbb{E}[X_i^{j-1}e^{\alpha X_i}] + \int_\alpha^\beta \mathbb{E}[X_i^j e^{\lambda X_i}] d\lambda \quad (42)$$

et déduire du théorème de convergence dominée que l'application de $[\alpha, \beta]$ dans $\mathbb{R}$ définie par $\lambda \mapsto \mathbb{E}[X_i^j e^{\lambda X_i}]$ est continue, donc $\beta \mapsto \mathbb{E}[X_i^{j-1}e^{\beta X_i}]$ est de classe $\mathcal{C}^1$, de dérivée $\mathbb{E}[X_i^j e^{\beta X_i}]$ et ϕ_i est $\mathcal{C}^\infty$ sur $]0, \Lambda_i[$, enfin il en est de même pour $\psi_i = \log(\phi_i)$.

Supposons maintenant que $\mathbb{E}[|X_i|^k] < \infty$, alors le raisonnement ci-dessus s'applique pour $\alpha = 0$ tant que $j \leq k$ et montre que pour tout $\beta \in]0, \Lambda_i[$ et pour tout $j \in \llbracket 0, k \rrbracket$, l'égalité (42) est valable pour $\alpha = 0$, ce qui entraîne que ϕ_i et ψ_i sont $\mathcal{C}^k$ sur $[0, \Lambda_i[$. $\diamond$

Proposition 3.3 *Si $\mathbb{E}[X_i^2] < \infty$ et $\Lambda_i > 0$, alors $\forall \lambda \in [0, \Lambda_i[$, $\psi_i''(\lambda)$ est la variance d'une VA :*

$$\begin{aligned}\psi_i''(\lambda) &= \frac{\mathbb{E}[X_i^2 e^{\lambda X_i}]}{\mathbb{E}[e^{\lambda X_i}]} - \left(\frac{\mathbb{E}[X_i e^{\lambda X_i}]}{\mathbb{E}[e^{\lambda X_i}]} \right)^2 \quad \text{et} \\ \psi_i(\lambda) &= \lambda \mathbb{E}[X_i] + \int_0^\lambda (\lambda - \alpha) \psi_i''(\alpha) d\alpha.\end{aligned}$$

Preuve. D'après la proposition 3.2, ψ_i est $\mathcal{C}^2$ sur $[0, \Lambda_i[$, avec $\psi_i'(\lambda) = \frac{\mathbb{E}[X_i e^{\lambda X_i}]}{\mathbb{E}[e^{\lambda X_i}]}$ et $\psi_i''(\lambda) = \frac{\mathbb{E}[X_i^2 e^{\lambda X_i}]}{\mathbb{E}[e^{\lambda X_i}]} - \left(\frac{\mathbb{E}[X_i e^{\lambda X_i}]}{\mathbb{E}[e^{\lambda X_i}]} \right)^2$. Considérons la VA Y_i de loi $\mathbb{P}(Y_i \in \mathcal{A}) = \frac{\mathbb{E}[\mathbb{1}(X_i \in \mathcal{A}) e^{\lambda X_i}]}{\mathbb{E}[e^{\lambda X_i}]}$ ($\forall \mathcal{A}$ borélien), alors pour toute fonction borélienne f telle que $\mathbb{E}[|f(X_i)| e^{\lambda X_i}] < \infty$, on a $\mathbb{E}[f(Y_i)] = \frac{\mathbb{E}[f(X_i) e^{\lambda X_i}]}{\mathbb{E}[e^{\lambda X_i}]}$, ce qui montre que $\psi_i''(\lambda) = \mathbb{E}[Y_i^2] - \mathbb{E}[Y_i]^2$ est bien une variance et la dernière égalité vient du développement de Taylor avec reste intégral $\psi_i(\lambda) = \psi_i(0) + \lambda \psi_i'(0) + \int_0^\lambda (\lambda - \alpha) \psi_i''(\alpha) d\alpha$. $\diamond$

Proposition 3.4 *Supposons que $\Lambda > 0$ et que $\mathbb{E}[X_i^2] < \infty$ ($1 \leq i \leq m$). Posons pour $\lambda \in [0, \Lambda[$*

$$\begin{aligned}\bar{V}(\lambda) &= \frac{2}{\lambda^2}[\psi(\lambda) - \lambda\mu] = \frac{2}{\lambda^2} \int_0^\lambda (\lambda - \alpha)\psi''(\alpha) d\alpha, \quad V(\lambda) = \sup_{\beta \in [0, \lambda]} \bar{V}(\beta) \quad \text{et} \\ v &= V(0) = \frac{1}{m} \sum_{i=1}^m \text{var}(X_i).\end{aligned}$$

Alors la fonction V est croissante, continue, à valeurs dans $\mathbb{R}$ et pour tout ¹⁰ x et tout $\delta \in]0; 1[$

$$\mathbb{P}(M \geq \mu + x) \leq \exp\left(\frac{-mx^2}{2V(x/v)}\right) \quad (43)$$

$$\mathbb{P}\left(M \geq \mu + \sqrt{\frac{2\log(\delta^{-1})}{m} V\left(\sqrt{\frac{2\log(\delta^{-1})}{mv}}\right)}\right) \leq \delta. \quad (44)$$

Preuve. On a $v = \lim_{\lambda \rightarrow 0} V(\lambda) = \psi''(0) = \frac{1}{m} \sum_{i=1}^m \text{var}(X_i)$. De plus, V est continue, car étant croissante, pour tout $\lambda \in]0, \Lambda[$ elle admet une limite à droite et une limite à gauche en λ : $V(\lambda^-) \leq V(\lambda^+)$ et, comme $\bar{V}$ est continue d'après la proposition 3.2, on a $V(\lambda^-) = \sup_{\beta \in [0, \lambda]} \bar{V}(\beta) = V(\lambda^+) (= \bar{V}(\beta_0))$ pour un $\beta_0 \in [0, \lambda]$. Pour tout $\lambda \in]0, \Lambda[$ et tout $\beta \in [0, \lambda]$,

$$-\psi^*(\mu + x) = \inf_{\gamma \in \mathbb{R}_+} \psi(\gamma) - \gamma(\mu + x) \leq \psi(\beta) - \beta(\mu + x) = \frac{\beta^2}{2} \bar{V}(\beta) - \beta x \leq -\left[\beta x - \frac{\beta^2}{2} V(\lambda)\right],$$

donc d'après la proposition 3.1 $\mathbb{P}(M \geq \mu + x) \leq \exp[-m\psi^*(\mu + x)] \leq \exp\left[-m\left(\beta x - \frac{\beta^2}{2} V(\lambda)\right)\right]$ et pour $\lambda = x/v$ et $\beta = x/V(\lambda) \leq \lambda$, on obtient l'inégalité (43). Enfin, posons

$$\delta = \exp\left[-m\left(\beta x - \frac{\beta^2}{2} V(\lambda)\right)\right],$$

donc $\frac{\log(\delta^{-1})}{m\beta} + \frac{\beta}{2} V(\lambda) = x$ et $\mathbb{P}\left(M \geq \mu + \frac{\beta}{2} V(\lambda) + \frac{\log(\delta^{-1})}{m\beta}\right) \leq \delta$, et choisissons $\lambda = \sqrt{\frac{2\log(\delta^{-1})}{mv}} \geq \beta = \sqrt{\frac{2\log(\delta^{-1})}{mV(\lambda)}}$ pour conclure. $\diamond$

Proposition 3.5 *[Inégalité de Hoeffding]. Supposons que pour tout $i \in \llbracket 1, m \rrbracket$, il existe des constantes a_i et b_i telles que $a_i \leq X_i \leq b_i$, alors pour tout $x \in \mathbb{R}$ et tout $\delta \in]0; 1[$*

$$\begin{aligned}P(M \geq \mu + x) &\leq \exp\left(\frac{-2m^2x^2}{\sum_{i=1}^m (b_i - a_i)^2}\right) \quad \text{et} \\ P\left(M \geq \mu + \sqrt{\frac{\sum_{i=1}^m (b_i - a_i)^2 \log(\delta^{-1})}{2m^2}}\right) &\leq \delta.\end{aligned}$$

Preuve. ψ_i'' est la variance d'une VA à valeurs dans $[a_i, b_i]$ et ne peut donc pas dépasser $\frac{(b_i - a_i)^2}{4}$, donc $\psi(\lambda) \leq \lambda\mu + \frac{\lambda^2}{8m} \sum_{i=1}^m (b_i - a_i)^2$ et

$$\psi^*(\mu + x) = \sup_{\lambda \in \mathbb{R}_+} [\lambda\mu + \lambda x - \psi(\lambda)] \geq \sup_{\lambda \in \mathbb{R}_+} \left(\lambda x - \frac{\lambda^2}{8m} \sum_{i=1}^m (b_i - a_i)^2\right) = \frac{2mx^2}{\sum_{i=1}^m (b_i - a_i)^2}.$$

L'inégalité de Hoeffding résulte donc de la proposition 3.1 et de l'égalité $\frac{2m^2x^2}{A} = \log(\delta^{-1})$, qui implique $x = \sqrt{\frac{A \log(\delta^{-1})}{2m^2}}$. $\diamond$

10. Sous réserve que $0 \leq x < \Lambda v$ et que $\delta > e^{-mv\Lambda^2/2}$.

3.2 Divergence de Kullback-Leibler

Pour $(\Theta, \mathcal{M}_\Theta)$ un espace mesurable, nous noterons $\mathcal{P}_+^1(\Theta, \mathcal{M}_\Theta)$ (ou $\mathcal{P}_+^1(\Theta)$ quand il n'y a pas d'ambiguïté sur la tribu associée) l'ensemble des mesures de probabilité sur $(\Theta, \mathcal{M}_\Theta)$.

Pour ρ et ν dans $\mathcal{P}_+^1(\Theta)$, la *divergence de Kullback-Leibler*, appelée aussi *entropie relative*, des probabilités ρ et ν est définie par

$$\mathcal{K}(\rho, \nu) = \begin{cases} \int_{\Theta} \log \left(\frac{d\rho}{d\nu}(\theta) \right) d\rho(\theta) & \text{si } \rho \ll \nu \\ +\infty & \text{sinon,} \end{cases} \quad (45)$$

où $\frac{d\rho}{d\nu}(\theta)$ désigne une dérivée de Radon-Nikodym de ρ par rapport à ν . Quand $\rho \ll \nu$, on a

$$\mathcal{K}(\rho, \nu) = \int_{\Theta} \left[1 - \frac{d\rho}{d\nu}(\theta) + \frac{d\rho}{d\nu}(\theta) \log \left(\frac{d\rho}{d\nu}(\theta) \right) \right] d\nu(\theta)$$

et la fonction définie sur $\mathbb{R}_+^*$ par $x \mapsto 1 - x + x \log(x)$ étant positive, strictement convexe et ne s'annulant que pour $x = 1$, on en déduit que la divergence de Kullback-Leibler est toujours positive et ne s'annule que si $\rho = \nu$, ν -presque sûrement.

3.3 Bornes bayésiennes PAC des déviations entre moyennes empirique et réelle

Soient m VA indépendantes $Z_1, \dots, Z_m$ à valeurs dans $\mathcal{Z}$, un espace mesurable (on notera $\mathcal{M}_{\mathcal{Z}}$ la tribu associée). Soient Θ un espace de paramètres, également mesurable (avec $\mathcal{M}_\Theta$ la tribu associée), et $f : \mathcal{Z} \times \Theta \rightarrow \mathbb{R}$ une fonction mesurable.

Exemple 3.1 *Un problème d'apprentissage supervisé¹¹, où $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, $Z_i = (x_i, y_i)$ —la base d'entraînement—, avec un espace d'hypothèses h_θ dépendant d'un paramètre $\theta \in \Theta$, autrement dit il existe une application $h : \Theta \rightarrow \mathcal{M}(\mathcal{X}, \mathcal{Y})$ telle que $\mathcal{H} = h(\Theta)$ —on écrit h_θ l'hypothèse $h(\theta)$ —, et pour une fonction coût $L : \mathcal{Y}^2 \rightarrow \mathbb{R}_+$ (ou à valeurs dans $\mathbb{R}$), $f(Z_i, \theta) = L[h_\theta(x_i), y_i]$.*

Supposons que pour tout $\theta \in \Theta$ et tout $i \in \llbracket 1, m \rrbracket$, $\mathbb{E}[f(Z_i, \theta)^2] < \infty$. Posons

$$M(\theta) = \frac{1}{m} \sum_{i=1}^n f(Z_i, \theta), \quad \mu(\theta) = \mathbb{E}[M(\theta)],$$

$$\psi_i(\lambda, \theta) = \log \{ \mathbb{E}(\exp[\lambda f(Z_i, \theta)]) \}, \quad \psi(\lambda, \theta) = \frac{1}{m} \sum_{i=1}^m \psi_i(\lambda, \theta)$$

et $\Lambda = \sup \{ \lambda : \forall \theta \in \Theta, \psi(\lambda, \theta) < \infty \}$.

Avec l'exemple 3.1, $M(\theta)$ et $\mu(\theta)$ correspondent aux risques empirique $R_S(h_\theta)$ et réel $R(h_\theta)$.

Proposition 3.6 *Supposons $\Lambda > 0$. Soit $\nu \in \mathcal{P}_+^1(\Theta)$ une mesure de référence sur l'espace des paramètres Θ . Alors $\forall \lambda \in [0, \Lambda[$*

$$\mathbb{E} \left[\exp \left(\sup \left\{ \int_{\Theta} m[\lambda M(\theta) - \psi(\lambda, \theta)] d\rho(\theta) - \mathcal{K}(\rho, \nu) : \rho \in \mathcal{P}_+^1(\Theta) \text{ et } \theta \mapsto \lambda M(\theta) - \psi(\lambda, \theta) \in L^1(\rho) \text{ et } \mathcal{K}(\rho, \nu) < \infty \right\} \right) \right] \leq 1.$$

Par conséquent, pour tout $\delta > 0$, avec une probabilité d'au moins $1 - \delta$, pour toute distribution de probabilité $\rho \in \mathcal{M}_+^1(\Theta)$ telle que $\theta \mapsto \lambda M(\theta) - \psi(\lambda, \theta) \in L^1(\rho)$ et $\mathcal{K}(\rho, \nu) < \infty$, pour tout $\lambda \in [0, \Lambda[$

$$\int_{\Theta} M(\theta) d\rho(\theta) \leq \frac{1}{\lambda} \int_{\Theta} \psi(\lambda, \theta) d\rho(\theta) + \frac{\mathcal{K}(\rho, \nu) + \log(\delta^{-1})}{m\lambda}.$$

11. Voir la section 1.

Preuve. D'après l'inégalité de Jensen

$$\begin{aligned}
& \exp \left\{ \int_{\Theta} m[\lambda M(\theta) - \psi(\lambda, \theta)] d\rho(\theta) - \mathcal{K}(\rho, \nu) \right\} \\
& \leq \int_{\Theta} \exp \{ m[\lambda M(\theta) - \psi(\lambda, \theta)] \} \left(\frac{d\rho}{d\nu}(\theta) \right)^{-1} \mathbf{1} \left[\frac{d\rho}{d\nu}(\theta) > 0 \right] d\rho(\theta) \\
& = \int_{\Theta} \exp \{ m[\lambda M(\theta) - \psi(\lambda, \theta)] \} \mathbf{1} \left[\frac{d\rho}{d\nu}(\theta) > 0 \right] d\nu(\theta) \\
& \leq \int_{\Theta} \exp \{ m[\lambda M(\theta) - \psi(\lambda, \theta)] \} d\nu(\theta).
\end{aligned}$$

On peut ensuite appliquer le théorème de Fubini pour les fonctions positives en ayant remarqué que $\mathbb{E} [\exp \{ m[\lambda M(\theta) - \psi(\lambda, \theta)] \} d\nu(\theta)] = \prod_{i=1}^m \frac{\mathbb{E}[e^{\lambda f(Z_i, \theta)}]}{e^{\psi_i(\lambda, \theta)}} = 1$ pour conclure que

$$\begin{aligned}
& \mathbb{E} \left[\exp \left\{ \int_{\Theta} m[\lambda M(\theta) - \psi(\lambda, \theta)] d\rho(\theta) - \mathcal{K}(\rho, \nu) \right\} \right] \\
& \leq \mathbb{E} \left[\int_{\Theta} \exp \{ m[\lambda M(\theta) - \psi(\lambda, \theta)] \} d\nu(\theta) \right] = 1.
\end{aligned}$$

L'espérance indiquée dans la proposition ne porte pas forcément sur une fonction mesurable, mais cette fonction est majorée par une fonction mesurable d'espérance inférieure ou égale à 1. C'est ce que montre la preuve et c'est le sens à donner à la proposition. La deuxième partie de la proposition s'obtient en appliquant l'inégalité de Markov¹² : considérons $\rho \in \mathcal{P}_+^1(\Theta)$ satisfaisant aux conditions de la proposition. On a

$$\begin{aligned}
& \mathbb{P} \left[\exp \left(\int_{\Theta} m[\lambda M(\theta) - \psi(\lambda, \theta)] d\rho(\theta) - \mathcal{K}(\rho, \nu) \right) \geq \delta^{-1} \right] \\
& \leq \delta \mathbb{E} \left[\exp \left(\int_{\Theta} m[\lambda M(\theta) - \psi(\lambda, \theta)] d\rho(\theta) - \mathcal{K}(\rho, \nu) \right) \right] \leq \delta,
\end{aligned}$$

donc $\mathbb{P} \left[\int_{\Theta} M(\theta) d\rho(\theta) \leq \frac{1}{\lambda} \int_{\Theta} \psi(\lambda, \theta) d\rho(\theta) + \frac{\mathcal{K}(\rho, \nu) + \log(\delta^{-1})}{m\lambda} \right] \geq 1 - \delta$. Ici aussi, l'événement n'est pas forcément mesurable, il faut comprendre qu'il contient un événement mesurable de probabilité supérieure ou égale à $1 - \delta$. $\diamond$

Posons pour $\theta \in \Theta$ et $\lambda \in [0, \Lambda[$

$$\begin{aligned}
\mu_i(\theta) &= \mathbb{E}[f(Z_i, \theta)], & v(\theta) &= \frac{1}{m} \sum_{i=1}^m \mathbb{E} \{ [f(Z_i, \theta) - \mu_i(\theta)]^2 \}, \\
\bar{V}(\lambda, \theta) &= \frac{2}{\lambda^2} [\psi(\lambda, \theta) - \lambda \mu(\theta)], & V(\lambda, \theta) &= \sup_{\beta \in [0, \lambda]} \bar{V}(\beta, \theta), \\
\bar{V}(\lambda) &= \sup_{\theta \in \Theta} \bar{V}(\lambda, \theta), & V(\lambda) &= \sup_{\theta \in \Theta} V(\lambda, \theta)
\end{aligned}$$

et supposons que

$$v = \sup_{\theta \in \Theta} v(\theta) < \infty \text{ et } \exists \Lambda' > 0, \forall \lambda \in [0, \Lambda'[, V(\lambda) < \infty. \quad (46)$$

Proposition 3.7 *Sous les hypothèses ci-dessus, on a pour tout¹³ $c > 0$*

$$\begin{aligned}
& \mathbb{E} \left[\sup \left\{ \int_{\Theta} [M(\theta) - \mu(\theta)] d\rho(\theta) : \rho \in \mathcal{P}_+^1(\Theta) \text{ et } \theta \mapsto M(\theta) - \mu(\theta) \in L^1(\rho) \text{ et } \mathcal{K}(\rho, \nu) \leq c \right\} \right] \\
& \leq \inf_{\lambda \in [0, \Lambda'[} \left(\frac{\lambda \bar{V}(\lambda)}{2} + \frac{c}{\lambda m} \right) \leq \sqrt{\frac{2c}{m} V \left(\sqrt{\frac{2c}{mv}} \right)}.
\end{aligned}$$

12. Pour Z une VA positive, $\mathbb{P}(Z \geq a) \leq \frac{\mathbb{E}[Z]}{a}$.

13. En fait, pour tout c tel que $0 < c < \frac{\Lambda'^2 m v}{2}$, qui assure que $\sqrt{\frac{2c}{mv}} < \Lambda'$.

En particulier quand $|\Theta| < \infty$ avec $c = \log(|\Theta|)$, $\rho = \delta_\theta$ et $\nu(\theta) = \frac{1}{|\Theta|}$ on obtient

$$\mathbb{E} \left\{ \sup_{\theta \in \Theta} [M(\theta) - \mu(\theta)] \right\} \leq \sqrt{\frac{2 \log(|\Theta|)}{m} V \left(\sqrt{\frac{2 \log(|\Theta|)}{mv}} \right)}.$$

Preuve. D'après la preuve de la proposition précédente, pour $\rho \in \mathcal{P}_+^1(\Theta)$ telle que $\mathcal{K}(\rho, \nu) \leq c$ et $\theta \mapsto M(\theta) - \mu(\theta) \in L^1(\rho)$, on a¹⁴ pour tout $\lambda \in [0, \Lambda']$ et grâce à l'inégalité de Jensen

$$\begin{aligned} \int_{\Theta} [M(\theta) - \mu(\theta)] d\rho(\theta) &= \frac{1}{m\lambda} \left\{ \log \exp \left[\int_{\Theta} m[\lambda M(\theta) - \psi(\lambda, \theta)] d\rho(\theta) - \mathcal{K}(\rho, \nu) \right] + \right. \\ &\quad \left. + \int_{\Theta} m[\psi(\lambda, \theta) - \lambda \mu(\theta)] d\rho(\theta) + \mathcal{K}(\rho, \nu) \right\} \\ &\leq \frac{1}{\lambda m} \left\{ \log \left[\int_{\Theta} \exp \left[m[\lambda M(\theta) - \psi(\lambda, \theta)] \right] d\nu(\theta) \right] + \right. \\ &\quad \left. + c + \frac{m\lambda^2}{2} \bar{V}(\lambda) \right\} \end{aligned}$$

et en appliquant encore une fois l'inégalité de Jensen ($\mathbb{E}[\log(Y)] \leq \log[\mathbb{E}(Y)]$) on déduit de la proposition 3.6 la première inégalité. Le majorant $\inf_{\lambda \in [0, \Lambda']} \frac{\lambda \bar{V}(\lambda)}{2} + \frac{c}{\lambda m}$ peut être affaibli en $\inf_{\lambda \in [0, \beta]} \frac{\lambda V(\beta)}{2} + \frac{c}{\lambda m}$, avec $0 \leq \beta < \Lambda'$. Enfin, on obtient la deuxième inégalité en choisissant $\beta = \sqrt{\frac{2c}{v}}$ (voir la note 13 au bas de la page 27) et $\lambda = \sqrt{\frac{2c}{mV(\beta)}} \leq \beta$. $\diamond$

Proposition 3.8 *Sous les hypothèses précédentes, pour toute constante $c > 0$ et pour tout¹⁵ $\delta > 0$, avec probabilité d'au moins $1 - \delta$*

$$\begin{aligned} \sup \left\{ \int_{\Theta} [M(\theta) - \mu(\theta)] d\rho(\theta) : \rho \in \mathcal{P}_+^1(\Theta) \text{ et } \theta \mapsto M(\theta) - \mu(\theta) \in L^1(\rho) \text{ et } \mathcal{K}(\rho, \nu) \leq c \right\} \\ \leq \inf_{\lambda \in [0, \Lambda']} \frac{\lambda \bar{V}(\lambda)}{2} + \frac{c + \log(\delta^{-1})}{\lambda m} \\ \leq \sqrt{\frac{2[c + \log(\delta^{-1})]}{m} V \left(\sqrt{\frac{2[c + \log(\delta^{-1})]}{mv}} \right)}. \end{aligned}$$

En particulier quand Θ est fini, avec probabilité d'au moins $1 - \delta$

$$\sup_{\theta \in \Theta} [M(\theta) - \mu(\theta)] \leq \sqrt{\frac{2 \log(|\Theta|/\delta)}{m} V \left(\sqrt{\frac{2 \log(|\Theta|/\delta)}{mv}} \right)}.$$

Preuve. C'est une conséquence directe de la deuxième partie de la proposition 3.6, de la majoration $\psi(\lambda, \theta) - \lambda \mu(\theta) \leq \frac{\lambda^2 \bar{V}(\lambda)}{2}$ et de la fin de la démonstration de la proposition 3.7. $\diamond$

Remarque 3.1 *Remarquons que la proposition 3.8 est encore vraie en remplaçant f par $-f$, c'est-à-dire M par $-M$ et μ par $-\mu$.*

14. Remarquons que l'hypothèse (46) entraîne que $\theta \mapsto \psi(\lambda, \theta) - \lambda \mu(\theta) \in L^1(\rho)$ pour tout $\lambda \in [0, \Lambda']$ et donc $\theta \mapsto \psi(\lambda, \theta) - \lambda M(\theta)$ étant la somme de deux fonctions de $L^1(\rho)$, elle est aussi dans $L^1(\rho)$.

15. Sous réserve que $c + \log \delta^{-1} < \frac{\Lambda'^2 mv}{2}$.

Remarque 3.2 Si f est à valeurs dans $[0; 1]$ ou dans $[-1; 0]$, alors $V(\lambda) \leq \frac{1}{4}$ pour tout λ , et si $|\Theta|$ est fini, alors avec une probabilité d'au moins $1 - \delta$, $\sup_{\theta \in \Theta} |M(\theta) - \mu(\theta)| \leq \sqrt{\frac{\log(2|\Theta|/\delta)}{2m}}$ et l'on retrouve la propriété de convergence uniforme de Θ (que l'on identifie à l'espace d'hypothèse) donnée à la proposition 1.5.

Proposition 3.9 Supposons que $\Theta = \mathbb{B}_d = \{\theta \in \mathbb{R}^d : \|\theta\| \leq 1\}$ et qu'il existe deux constantes strictement positives B et g telles que

$$\begin{aligned} \sup_{z \in \mathcal{Z}} f(z, \theta) - \inf_{z \in \mathcal{Z}} f(z, \theta) &\leq B & (\forall \theta \in \mathbb{B}_d) \\ |f(z, \theta) - f(z, \theta')| &\leq g \|\theta - \theta'\| & (\forall z \in \mathcal{Z}, \forall \theta, \theta' \in \mathbb{B}_d). \end{aligned}$$

Considérons un estimateur qui minimise le risque empirique

$$\hat{\theta} \in \arg \min_{\theta \in \mathbb{B}_d} M(\theta).$$

Alors, avec une probabilité d'au moins $1 - \delta$

$$\mu(\hat{\theta}) \leq \inf_{\theta \in \mathbb{B}_d} \mu(\theta) + B \left\{ \sqrt{\frac{d}{2m} \log \left(1 + \frac{4g}{B} \sqrt{\frac{2m}{d}} \right) + \frac{\log(2/\delta)}{2m}} + \sqrt{\frac{d}{8m}} + \sqrt{\frac{\log(2/\delta)}{2m}} \right\}.$$

Sous ces hypothèses très simples, il apparaît que la qualité de l'estimation du risque minimum

$$\inf_{\theta \in \mathbb{B}_d} \mu(\theta)$$

par le risque de généralisation $\mu(\hat{\theta})$ dépend de la dimension d de l'espace des paramètres, et plus précisément du rapport d/m entre cette dimension et la taille de l'échantillon.

Preuve. Commençons par prolonger la fonction f à $\mathbb{R}^d$ en posant

$$f(z, \theta) = f(z, \theta/\|\theta\|) \quad (\theta \in \mathbb{R}^d \setminus \mathbb{B}_d).$$

Soit $\eta > 0$ dont la valeur sera précisée ultérieurement et ν la mesure de probabilité uniforme sur la boule $(1 + \eta)\mathbb{B}_d$ de rayon $1 + \eta$. Pour tout $\theta \in \mathbb{B}_d$, considérons ρ_θ la mesure uniforme sur la boule $\theta + \eta\mathbb{B}_d$ de rayon η centrée en θ . Le volume d'une boule de $\mathbb{R}^d$ étant proportionnel à son rayon élevé à la puissance d , on a

$$\mathcal{K}(\rho_\theta, \nu) = \int_{\mathbb{B}_d} d \log \left(\frac{1 + \eta}{\eta} \right) d\rho_\theta(u) = d \log(1 + \eta^{-1}) \quad (\theta \in \mathbb{B}_d).$$

Il résulte de la proposition 3.3 en y remplaçant X_i par $f(Z_i, \theta)$ que pour tout $\theta \in \mathbb{R}^d$, la dérivée seconde $\frac{\partial^2 \psi_i(\lambda, \theta)}{\partial \lambda^2}$ est la variance d'une VA à valeurs dans $\left[\inf_{z \in \mathcal{Z}} f(z, \theta), \sup_{z \in \mathcal{Z}} f(z, \theta) \right]$, ce qui entraîne, d'après les hypothèses de la proposition 3.9, que

$$\overline{V}(\lambda, \theta) \leq \frac{B^2}{4} \quad (\forall \lambda \in [0, \Lambda], \forall \theta \in \mathbb{R}^d).$$

Ainsi, la proposition 3.8 et la remarque 3.1 impliquent (avec $c = \mathcal{K}(\rho_\theta, \nu) = d \log(1 + \eta^{-1})$) qu'avec une probabilité d'au moins $1 - \delta$, on a

$$\int_{\mathbb{R}^d} \mu(\theta') d\rho_\theta(\theta') \leq \int_{\mathbb{R}^d} M(\theta') d\rho_\theta(\theta') + B \sqrt{\frac{d \log(1 + \eta^{-1}) + \log(\delta^{-1})}{2m}}$$

et pour $\theta = \hat{\theta}$ on obtient qu'avec une probabilité d'au moins $1 - \delta$

$$\mu(\hat{\theta}) \leq M(\hat{\theta}) + 2g\eta + B \sqrt{\frac{d \log(1 + \eta^{-1}) + \log(\delta^{-1})}{2m}}$$

car

$$\begin{aligned} \int_{\mathbb{R}^d} \mu(\theta') d\rho_{\hat{\theta}}(\theta') &\geq \mu(\hat{\theta}) - \frac{1}{m} \left| \sum_{i=1}^m \int_{\mathbb{R}^d} \mathbb{E}[f(Z_i, \theta') - f(Z_i, \hat{\theta})] d\rho_{\hat{\theta}}(\theta') \right| \\ &\geq \mu(\hat{\theta}) - g \int_{\mathbb{R}^d} \|\hat{\theta} - \theta'\| d\rho_{\hat{\theta}}(\theta') \geq \mu(\hat{\theta}) - g\eta \end{aligned}$$

et de même

$$\int_{\mathbb{R}^d} M(\theta') d\rho_{\hat{\theta}}(\theta') \leq M(\hat{\theta}) + \frac{1}{m} \left| \sum_{i=1}^m \int_{\mathbb{R}^d} [f(Z_i, \theta') - f(Z_i, \hat{\theta})] d\rho_{\hat{\theta}}(\theta') \right| \leq M(\hat{\theta}) + g\eta.$$

Soit $\theta_\star \in \mathbb{B}_d$ tel que

$$\mu(\theta_\star) = \inf_{\theta \in \mathbb{B}_d} \mu(\theta)$$

(qui existe car $\theta \mapsto \mu(\theta)$ est continue sur le compact $\mathbb{B}_d$). Il r  sulte de l'in  galit   de Hoeffding qu'avec une probabilit   d'au moins $1 - \delta$

$$M(\theta_\star) \leq \mu(\theta_\star) + B\sqrt{\frac{\log(\delta^{-1})}{2m}}.$$

Par d  finition de $\hat{\theta}$, on a $M(\hat{\theta}) \leq M(\theta_\star)$, donc avec une probabilit   d'au moins $1 - 2\delta$

$$\mu(\hat{\theta}) \leq \mu(\theta_\star) + B \left\{ \sqrt{\frac{d \log(1 + \eta^{-1}) + \log(\delta^{-1})}{2m}} + \sqrt{\frac{\log(\delta^{-1})}{2m}} \right\} + 2g\eta.$$

Enfin, le choix $\eta = \frac{B}{4g} \sqrt{\frac{d}{2m}}$ permet de conclure, en rempla  ant δ par $\delta/2$. $\diamond$

Proposition 3.10 *Supposons que $\Theta = \mathbb{R}^d$, qu'il existe une fonction $(z, \theta) \mapsto \nabla f(z, \theta) \in \mathbb{R}^d$ mesurable et des constantes strictement positives g et H telles que $\forall z \in \mathcal{Z}, \forall \theta, \theta' \in \mathbb{R}^d$,*

$$\begin{aligned} |f(z, \theta) - f(z, \theta')| &\leq g\|\theta - \theta'\| \\ |f(z, \theta) - f(z, \theta') - \langle \nabla f(z, \theta), \theta' - \theta \rangle| &\leq \frac{H}{2}\|\theta' - \theta\|^2. \end{aligned}$$

Soient $\theta_\star \in \arg \min_{\theta \in \mathbb{B}_d} \mu(\theta)$ et $\hat{\theta} \in \arg \min_{\theta \in \mathbb{B}_d} M(\theta)$. Introduisons la fonction

$$\mathbb{R}_+^\star \ni h \mapsto \chi(h) = \sup_{\theta \in \mathbb{B}_d} \left\{ \frac{h}{2} \|\theta - \theta_\star\|^2 - \mu(\theta) + \mu(\theta_\star) \right\}.$$

Alors, pour tout $h > 0$, avec une probabilit   d'au moins $1 - \delta$

$$\begin{aligned} \|\hat{\theta} - \theta_\star\|^2 &\leq \frac{8g^2}{mh^2} \left[\left(\frac{8H}{h} + 1 \right) d + 2 \log(\delta^{-1}) \right] + \frac{4\chi(h)}{h} \\ \mu(\hat{\theta}) - \mu(\theta_\star) &\leq \frac{4g^2}{mh} \left[\left(\frac{8H}{h} + 1 \right) d + 2 \log(\delta^{-1}) \right] + \chi(h). \end{aligned}$$

Dans le cas o   il existe $h > 0$ tel que $\chi(h) = 0$, on obtient ainsi une vitesse de convergence en d/m au lieu d'une vitesse en $\sqrt{d/m}$ sous les hypoth  ses plus faibles de la proposition 3.9.

Preuve. Soit $\beta > 0$, choisissons des densit  s gaussiennes $\rho_\theta = \mathcal{N}(\theta, \beta^{-1} \mathbf{I})$ et $\nu = \rho_{\theta_\star}$. On a ¹⁶

$$\mathcal{K}(\rho_\theta, \nu) = \frac{\beta \|\theta - \theta_\star\|^2}{2}.$$

16. En effet,

$$\begin{aligned} \mathcal{K}(\rho_{\theta_0}, \nu) &= \mathbb{E}_{\rho_{\theta_0}} \left\{ \log \left[\exp \left(\frac{-\beta}{2} [(\theta - \theta_0)^T (\theta - \theta_0) - (\theta - \theta_\star)^T (\theta - \theta_\star)] \right) \right] \right\} \\ &= \frac{\beta}{2} \mathbb{E}_{\rho_{\theta_0}} \{ 2\theta^T \theta_0 - \theta_0^T \theta_0 - 2\theta^T \theta_\star + \theta_\star^T \theta_\star \} = \frac{\beta}{2} (\theta_0^T \theta_0 - 2\theta_0^T \theta_\star + \theta_\star^T \theta_\star) = \frac{\beta \|\theta_0 - \theta_\star\|^2}{2}. \end{aligned}$$

Nous allons appliquer la proposition 3.6 à la fonction $(z, \theta) \mapsto f(z, \theta_*) - f(z, \theta) = \tilde{f}(z, \theta)$. Notons respectivement $\tilde{\psi}_i, \tilde{\psi}, \tilde{\mu}$ et $\tilde{M}$ les fonctions ψ_i, ψ, μ et M associées à $\tilde{f}$. On a $\tilde{\mu}(\theta) = \mu(\theta_*) - \mu(\theta)$, $\tilde{M}(\theta) = M(\theta_*) - M(\theta)$ et (d'après la proposition 3.2 en remplaçant X_i par $\tilde{f}(Z_i, \theta)$)

$$\begin{aligned}\tilde{\psi}_i''(\lambda, \theta) &= \text{var}(Y_i) \leq \mathbb{E}[Y_i^2] = \frac{\mathbb{E}\left\{[f(Z_i, \theta_*) - f(Z_i, \theta)]^2 e^{\lambda \tilde{f}(Z_i, \theta)}\right\}}{\mathbb{E}\left[e^{\lambda \tilde{f}(Z_i, \theta)}\right]} \\ &\leq \frac{\mathbb{E}\left\{[g\|\theta - \theta_*\|]^2 e^{\lambda \tilde{f}(Z_i, \theta)}\right\}}{\mathbb{E}\left[e^{\lambda \tilde{f}(Z_i, \theta)}\right]} = g^2 \|\theta - \theta_*\|^2,\end{aligned}$$

donc $\tilde{\psi}''(\lambda, \theta) \leq g^2 \|\theta - \theta_*\|^2$ et $\tilde{\psi}(\lambda, \theta) \leq \lambda \tilde{\mu}(\theta) + \frac{\lambda^2 g^2 \|\theta - \theta_*\|^2}{2} = \lambda [\mu(\theta_*) - \mu(\theta)] + \frac{\lambda^2 g^2 \|\theta - \theta_*\|^2}{2}$. Ainsi, il résulte de la deuxième assertion de la proposition 3.6 qu'avec une probabilité d'au moins $1 - \delta$

$$\begin{aligned}\int_{\mathbb{R}^d} \tilde{M}(\theta') d\rho_\theta(\theta') &\leq \frac{1}{\lambda} \int_{\mathbb{R}^d} \tilde{\psi}(\lambda, \theta') d\rho_\theta(\theta') + \frac{\beta \|\theta - \theta_*\|^2}{2m\lambda} + \frac{\log(\delta^{-1})}{m\lambda} \\ &\leq \int_{\mathbb{R}^d} \left\{ \tilde{\mu}(\theta') + \frac{\lambda g^2 \|\theta' - \theta_*\|^2}{2} \right\} d\rho_\theta(\theta') + \frac{\beta \|\theta - \theta_*\|^2}{2m\lambda} + \frac{\log(\delta^{-1})}{m\lambda}\end{aligned}$$

ce qui s'écrit

$$\begin{aligned}\int_{\mathbb{R}^d} \mu(\theta') d\rho_\theta(\theta') - \mu(\theta_*) &\leq \int_{\mathbb{R}^d} M(\theta') d\rho_\theta(\theta') - M(\theta_*) + \frac{\lambda g^2}{2} \int_{\mathbb{R}^d} \|\theta' - \theta\|^2 d\rho_\theta(\theta') + \\ &\quad + \frac{\beta \|\theta - \theta_*\|^2}{2m\lambda} + \frac{\log(\delta^{-1})}{m\lambda}.\end{aligned}$$

En remarquant que

$$f(Z_i, \theta') = f(Z_i, \theta) + \langle \nabla f(Z_i, \theta), \theta' - \theta \rangle + f(Z_i, \theta') - f(Z_i, \theta) - \langle \nabla f(Z_i, \theta), \theta' - \theta \rangle$$

et $\int_{\mathbb{R}^d} (\theta' - \theta) d\rho_\theta(\theta') = 0$, on obtient après double intégration et moyenne statistique

$$\begin{aligned}\int_{\mathbb{R}^d} \mu(\theta') d\rho_\theta(\theta') &= \mu(\theta) + \frac{1}{m} \sum_{i=1}^m \mathbb{E} \left\{ \int_{\mathbb{R}^d} [f(Z_i, \theta') - f(Z_i, \theta) - \langle \nabla f(Z_i, \theta), \theta' - \theta \rangle] d\rho_\theta(\theta') \right\} \\ &\geq \mu(\theta) - \frac{H}{2} \int_{\mathbb{R}^d} \|\theta' - \theta\|^2 d\rho_\theta(\theta') = \mu(\theta) - \frac{Hd}{2\beta};\end{aligned}$$

de même $\int_{\mathbb{R}^d} M(\theta') d\rho_\theta(\theta') \leq M(\theta) + \frac{Hd}{2\beta}$. Ainsi, avec une probabilité d'au moins $1 - \delta$, pour tout $\theta \in \mathbb{B}_d$, on a

$$\mu(\theta) - \mu(\theta_*) \leq M(\theta) - M(\theta_*) + \frac{Hd}{\beta} + \frac{\lambda g^2 d}{2\beta} + \frac{\lambda g^2 \|\theta - \theta_*\|^2}{2} + \frac{\beta \|\theta - \theta_*\|^2}{2m\lambda} + \frac{\log(\delta^{-1})}{m\lambda}.$$

De plus, par définition de χ , pour tout $\theta \in \mathbb{B}_d$, on a $\mu(\theta) - \mu(\theta_*) \geq \frac{h}{2} \|\theta - \theta_*\|^2 - \chi(h)$ et par construction $M(\hat{\theta}) \leq M(\theta_*)$ donc avec une probabilité d'au moins $1 - \delta$

$$\frac{h}{2} \|\hat{\theta} - \theta_*\|^2 \leq \chi(h) + \frac{d}{\beta} \left(H + \frac{\lambda g^2}{2} \right) + \left(\frac{\lambda g^2}{2} + \frac{\beta}{2m\lambda} \right) \|\hat{\theta} - \theta_*\|^2 + \frac{\log(\delta^{-1})}{m\lambda}$$

ou encore

$$\|\hat{\theta} - \theta_*\|^2 \left(1 - \frac{\lambda g^2}{h} - \frac{\beta}{m\lambda h} \right) \leq \frac{2\chi(h)}{h} + \frac{2d}{\beta h} \left(H + \frac{\lambda g^2}{2} \right) + \frac{2\log(\delta^{-1})}{mh\lambda}.$$

Choisissons alors $\lambda = \frac{h}{4g^2}$ et $\beta = \frac{mh\lambda}{4} = \frac{mh^2}{16g^2}$. On obtient

$$\frac{1}{2} \|\hat{\theta} - \theta_*\|^2 \leq \frac{2\chi(h)}{h} + \frac{32g^2 d}{mh^3} \left(H + \frac{h}{8} \right) + \frac{8g^2 \log(\delta^{-1})}{mh^2}$$

qui donne la première majoration de la proposition. Pour prouver la seconde, on utilise

$$\|\hat{\theta} - \theta_\star\|^2 \leq \frac{2}{h} [\mu(\hat{\theta}) - \mu(\theta_\star) + \chi(h)]$$

pour obtenir

$$\mu(\hat{\theta}) - \mu(\theta_\star) \leq \frac{d}{\beta} \left(H + \frac{\lambda g^2}{2} \right) + \left(\frac{\lambda g^2}{2} + \frac{\beta}{2m\lambda} \right) \frac{2}{h} [\mu(\hat{\theta}) - \mu(\theta_\star) + \chi(h)] + \frac{\log(\delta^{-1})}{m\lambda}$$

et l'on conclut en remplaçant λ et β par leurs valeurs ci-dessus. $\diamond$

4 Bornes bayésiennes PAC pour le problème de classification supervisée

Soient $Z_1, \dots, Z_m$ m VA i.i.d. à valeurs dans $\mathcal{Z}$ un espace mesurable, soit Θ un espace mesurable de paramètres et $f : \mathcal{Z} \times \Theta \rightarrow \{0, 1\}$ une fonction mesurable binaire, que nous appellerons fonction coût¹⁷ par abus de langage. Notre objectif est de minimiser par rapport à θ le risque

$$R(\theta) = \int_{\mathcal{Z}} f(z, \theta) dP_Z(z) = \mathbb{E}[f(Z, \theta) | \theta],$$

où P_Z est la loi marginale des Z_i . De façon plus précise, ne connaissant pas la loi P_Z , l'objectif est de trouver un estimateur $\hat{\theta}(Z_1, \dots, Z_m)$ dépendant de l'échantillon observé $S = Z_1, \dots, Z_m$ ayant un excès de risque

$$R(\hat{\theta}) - \inf_{\theta \in \Theta} R(\theta) = \int_{\mathcal{Z}} f(z, \hat{\theta}) dP_Z(z) - \inf_{\theta \in \Theta} \int_{\mathcal{Z}} f(z, \theta) dP_Z(z)$$

petit. Cette quantité est aléatoire car $\hat{\theta}$ dépend de l'échantillon aléatoire S . Par conséquent, dire quelle est petite peut être compris de différentes manières. Nous nous focaliserons sur les déviations de l'excès de risque, c'est-à-dire que nous rechercherons des estimateurs qui donnent un risque petit avec une probabilité proche de 1.

Un exemple clé d'un tel problème est celui de la classification supervisée, où $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, avec $\mathcal{Y}$ un ensemble fini, les $Z_i = (X_i, Y_i)$ sont les couples (entrée, sortie) de la classification, $\{h_\theta : \mathcal{X} \rightarrow \mathcal{Y}, \theta \in \Theta\}$ est une famille indexée par Θ de règles de classification considérées et la fonction perte est l'erreur de classification

$$L[(x, y), \theta] = \mathbb{1}[h_\theta(x) \neq y].$$

L'objectif est donc de minimiser l'erreur de classification moyenne $P_{X,Y}[h_\theta(X) \neq Y]$ à partir de la connaissance d'un échantillon $(X_i, Y_i)_{1 \leq i \leq m}$ observé.

4.1 Bornes de déviation pour des sommes de VA de Bernoulli

Pour $\lambda \in \mathbb{R}$, considérons la transformation log-Laplace (normalisée) d'une loi de Bernoulli de probabilité p :

$$\Phi_\lambda(p) \stackrel{\text{def}}{=} -\frac{1}{\lambda} \log[1 - p + p \exp(-\lambda)],$$

la divergence de Kullback-Leibler de deux lois de Bernoulli :

$$K(q, p) \stackrel{\text{def}}{=} q \log \left(\frac{q}{p} \right) + (1 - q) \log \left(\frac{1 - q}{1 - p} \right)$$

17. Voir l'exemple 3.1 page 26, où $Z_i = (X_i, Y_i)$ et $f(Z, \theta) = L[h_\theta(X), Y]$.

et notons P_S la mesure de probabilité empirique

$$P_S = \frac{1}{m} \sum_{i=1}^m \delta_{Z_i}$$

d'un échantillon i.i.d. $(Z_i)_{1 \leq i \leq m}$ tiré suivant la loi $P_Z^{\otimes m} \in \mathcal{P}_+^1(\mathcal{Z}^m)$. Nous introduisons des notations courtes pour désigner des intégrales : pour $\rho, \pi \in \mathcal{P}_+^1(\Theta)$ et $g \in L^1(\mathcal{Z} \times \Theta^2, P_Z \otimes \rho \otimes \pi)$,

$$g(P_Z, \rho, \pi) = \int g(z, \theta, \theta') dP_Z(z) d\rho(\theta) d\pi(\theta'),$$

de sorte que $f(P_Z, \rho) = \int f(z, \theta) dP_Z(z) d\rho(\theta)$ (par exemple). Commençons par rappeler la borne de Chernoff.

Proposition 4.1 *Pour toute valeur fixée de $\theta \in \Theta$, l'identité*

$$\int \exp[-m\lambda f(P_S, \theta)] dP_Z^{\otimes m} = \exp\{-m\lambda \Phi_\lambda[f(P_Z, \theta)]\}$$

montre qu'avec une probabilité d'au moins $1 - \delta$,

$$f(P_Z, \theta) \leq B_+[f(P_S, \theta), \log(\delta^{-1})/m],$$

où pour $q \in [0; 1]$ et $\eta \geq 0$, $B_+(q, \eta) = \inf_{\lambda \in \mathbb{R}_+} \Phi_\lambda^{-1}\left(q + \frac{\eta}{\lambda}\right) = \sup\{p \in [0; 1] : K(q, p) \leq \eta\}$.

De plus

$$-\eta q \leq B_+(q, \eta) - q - \sqrt{2\eta q(1-q)} \leq 2\eta(1-q).$$

De même, l'identité $\int \exp[m\lambda f(P_S, \theta)] dP_Z^{\otimes m} = \exp\{m\lambda \Phi_{-\lambda}[f(P_Z, \theta)]\}$ montre qu'avec une probabilité d'au moins $1 - \delta$, $f(P_S, \theta) \leq B_-[f(P_S, \theta), \log(\delta^{-1})/m]$, où pour $q \in [0; 1]$ et $\eta \geq 0$, $B_-(q, \eta) = \inf_{\lambda \in \mathbb{R}_+} \left\{ \Phi_{-\lambda}(q) + \frac{\eta}{\lambda} \right\} = \sup\{p \in [0; 1] : K(p, q) \leq \eta\}$ et

$$-\eta q \leq B_-(q, \eta) - q - \sqrt{2\eta q(1-q)} \leq 2\eta(1-q).$$

Présentons maintenant une identité importante.

Proposition 4.2 *Pour toutes mesures de probabilité ρ et π sur un espace mesurable telles que $\mathcal{K}(\rho, \pi) < \infty$, et pour toute fonction mesurable bornée g , définissons la mesure de probabilité transformée $\pi_{\exp(g)} \ll \pi$ par sa densité*

$$\frac{d\pi_{\exp(g)}}{d\pi} = \frac{\exp(g)}{Z}, \quad \text{avec } Z = \int \exp(g) d\pi$$

et

$$\text{var}(g d\pi) = \int \left(g - \int g d\pi \right)^2 d\pi.$$

Les moyennes par rapport à ρ et π et la transformée log-Laplace de g sont reliées par les identités

$$\int g d\rho - \mathcal{K}(\rho, \pi) + \mathcal{K}(\rho, \pi_{\exp(g)}) = \log \left[\int \exp(g) d\pi \right] \quad (47)$$

$$= \int g d\pi + \int_0^1 (1 - \alpha) \text{var}[g d\pi_{\exp(\alpha g)}] d\alpha. \quad (48)$$

Preuve. La première identité est une application directe des définitions et la deuxième est le développement de Taylor à l'ordre 2 avec reste intégral de¹⁸ $f(\alpha) = \log \left[\int \exp(\alpha g) d\pi \right] : f(1) = f(0) + f'(0) + \int_0^1 (1-\alpha) f''(\alpha) d\alpha$, car $f'(\alpha) = \frac{\int g \exp(\alpha g) d\pi}{\int \exp(\alpha g) d\pi}$ et $f''(\alpha) = \text{var}(g d\pi_{\exp(\alpha g)})$. $\square$

Prouvons maintenant la proposition 4.1. Remarquons que

$$f(P_S, \theta) = \int_{\mathcal{Z}} f(z, \theta) dP_S(z) = \frac{1}{m} \sum_{i=1}^m f(Z_i, \theta)$$

et

$$\begin{aligned} \mathbb{E} \{ \exp[-m\lambda f(P_S, \theta)] \mid \theta \} &= \mathbb{E} \left[\prod_{i=1}^n e^{-\lambda f(Z_i, \theta)} \mid \theta \right] = (\mathbb{E}[e^{-\lambda f(Z, \theta)} \mid \theta])^m \\ &= \exp \left[m \log \left(\int_{\mathcal{Z}} e^{-\lambda f(z, \theta)} dP_Z(z) \right) \right] \\ &= \exp \{ -m\lambda \Phi_\lambda[P_Z\{f(Z, \theta) = 1\}] \} = \exp \{ -m\lambda \Phi_\lambda[f(P_Z, \theta)] \}. \end{aligned}$$

Il résulte alors de l'inégalité de Markov, qu'avec une probabilité d'au moins $1 - \delta$:

$$\exp[-m\lambda f(P_S, \theta)] \leq \delta^{-1} \exp \{ -m\lambda \Phi_\lambda[f(P_Z, \theta)] \},$$

soit

$$\Phi_\lambda[f(P_Z, \theta)] \leq f(P_S, \theta) + \frac{\log(\delta^{-1})}{m\lambda}.$$

L'expression $\Phi_\lambda[f(P_Z, \theta)] - \frac{\log(\delta^{-1})}{m\lambda}$ n'étant pas aléatoire, nous pouvons chercher à optimiser l'inégalité précédente en fonction de λ .

Or, $\forall \lambda \neq 0$, la fonction $p \mapsto \Phi_\lambda(p)$ de dérivée $\Phi'_\lambda(p) = (-\frac{1}{\lambda}) \frac{e^{-\lambda} - 1}{1 - p + pe^{-\lambda}}$ est strictement croissante sur $[0; 1]$ et vérifie $\Phi_\lambda(0) = 0$ et $\Phi_\lambda(1) = 1$.

Nous trouvons donc que

$$f(P_Z, \theta) \leq B_+[f(P_S, \theta), \log(\delta^{-1})/m].$$

De plus, pour tout $q \in [0; 1]$, et pour tout $\eta > 0$, il existe $\lambda > 0$, tel que $q + \frac{\eta}{\lambda} < 1$, donc $B_+(q, \eta) < 1$. L'application de l'identité (47) à la fonction $g(x) = -\lambda x$ donne pour $p, q \in [0; 1]$ en notant $\text{Ber}(p)$ la loi de Bernoulli de probabilité p

$$\int g(x) dq(x) - K(q, p) + \mathcal{K}[\text{Ber}(q), \text{Ber}(p)_{\exp(g)}] = \log \left[\int \exp[g(x)] d\text{Ber}(p)(x) \right] = -\lambda \Phi_\lambda(p),$$

soit

$$\lambda \Phi_\lambda(p) = \lambda q + K(q, p) - K(q, p_\lambda) \quad \text{avec} \quad p_\lambda = \frac{p}{p + (1-p)e^\lambda} \in [0; 1],$$

car $\frac{d\text{Ber}(p)_{\exp(g)}}{d\text{Ber}(p)}(x) = \frac{e^{-\lambda x}}{1 - p + pe^{-\lambda}}$ entraîne que $\text{Ber}(p)_{\exp(g)} = \text{Ber}(p_\lambda)$. Ceci montre que

$$\begin{aligned} B_+(q, \eta) &= \sup \{ p \in [0; 1[: \forall \lambda > 0, \lambda \Phi_\lambda(p) \leq \lambda q + \eta \} \\ &= \sup \{ p \in [q, 1[: \forall \lambda > 0, K(q, p) \leq \eta + K(q, p_\lambda) \} \\ &= \sup \{ p \in [q, 1[: K(q, p) \leq \eta \} = \sup \{ p \in [0; 1[: K(q, p) \leq \eta \}, \end{aligned}$$

18. Pour justifier que f est $\mathcal{C}^2$, remarquons que pour tout $\alpha > 0$ et tout $k \in \mathbb{N}$

$$g^k \exp(\alpha g) = g^k + \int_0^\alpha g^{k+1} \exp(\gamma g) d\gamma \quad \text{et} \quad \int \int_0^\alpha |g^{k+1}(x)| \exp[\gamma g(x)] d\gamma d\pi(x) < +\infty,$$

car la fonction g est bornée. On peut donc appliquer le théorème de Fubini :

$$\int g^k(x) \exp[\alpha g(x)] d\pi(x) = \int g^k(x) d\pi(x) + \int_0^\alpha \left(\int g^{k+1}(x) \exp[\gamma g(x)] d\pi(x) \right) d\gamma$$

et l'application $\alpha \mapsto \int g^{k+1}(x) \exp[\gamma g(x)] d\pi(x)$ étant continue d'après le théorème de convergence dominée, on déduit que l'application $\alpha \mapsto \int g^k(x) \exp[\alpha g(x)] d\pi(x)$ est $\mathcal{C}^1$ de dérivée $\alpha \mapsto \int g^{k+1}(x) \exp[\alpha g(x)] d\pi(x)$, donc $\alpha \mapsto \int \exp[\alpha g(x)] d\pi(x)$ est $\mathcal{C}^\infty$, à valeurs dans $\mathbb{R}_+^*$, donc f est infiniment dérivable.

car pour $q \leq p < 1$ et $\lambda_0 = \log\left(\frac{q^{-1}-1}{p^{-1}-1}\right)$, on a $\lambda_0 \geq 0$ et $p_{\lambda_0} = q$.

Remarquons maintenant que $\frac{\partial K(x,p)}{\partial x} = \log\left(\frac{x}{p}\right) - \log\left(\frac{1-x}{1-p}\right)$ et $\frac{\partial^2 K(x,p)}{\partial x^2} = x^{-1}(1-x)^{-1}$ de sorte que la formule de Taylor à l'ordre 2 appliqué $K(x,p)$ en $x = p$ donne $K(q,p) = \frac{(p-q)^2}{2} K''_{xx}(y,p)$ avec $q < y < p$. Ainsi, si $\frac{1}{2} \leq q$, alors $K''_{xx}(y,p) \geq K''_{xx}(q,p)$, donc

$$K(q,p) \geq \frac{(p-q)^2}{2q(1-q)}$$

et l'inégalité $K(q,p) \leq \eta$ entraîne $p \leq q + \sqrt{2\eta q(1-q)}$. Et si $q \leq \frac{1}{2}$, alors $K''_{xx}(y,p) \geq \frac{1}{p(1-q)}$, donc

$$K(q,p) \geq \frac{(p-q)^2}{2p(1-q)}$$

et l'inégalité $K(q,p) \leq \eta$ entraîne $(p-q)^2 \leq 2\eta p(1-q)$, qui implique

$$(p-q)^2 - 2(p-q)\eta(1-q) \leq 2\eta q(1-q), \text{ i.e., } [p-q-\eta(1-q)]^2 \leq 2\eta q(1-q) + \eta^2(1-q)^2$$

et donc

$$p-q \leq \eta(1-q) + \sqrt{2\eta q(1-q) + \eta^2(1-q)^2} \leq \sqrt{2\eta q(1-q)} + 2\eta(1-q).$$

D'autre part, nous avons $K''_{xx}(y,p) < \frac{1}{q(1-p)}$ puisque $q < y < p$, donc

$$K(q,p) < \frac{(p-q)^2}{2q(1-p)}$$

et ainsi si $K(q,p) = \eta$, alors $(p-q)^2 \geq 2\eta q(1-p)$, impliquant

$$p-q \geq -\eta q + \sqrt{2\eta q(1-q) + \eta^2 q^2} \geq \sqrt{2\eta q(1-q)} - \eta q.$$

La preuve de la deuxième partie de la proposition 4.1 se fait en suivant la même démarche.

4.2 Bornes bayésiennes PAC

Nous allons rendre la proposition 4.1 uniforme par rapport à θ . L'approche bayésienne PAC pour répondre à cette question consiste à rendre θ aléatoire, nous considérerons donc des distributions de probabilité jointes de $Z_1, \dots, Z_m, \theta$ qui admettent toujours $P_Z^{\otimes m}$ comme distribution de Z_1^m et la loi conditionnelle de θ sachant l'échantillon Z_1^m est donnée par un noyau de probabilités de transition $\rho : \mathcal{Z}^m \rightarrow \mathcal{P}_+^1(\Theta)$, appelé dans ce contexte¹⁹ "loi *a posteriori*". Cette loi *a posteriori* ρ sera comparée à une loi *a priori* (i.e., non aléatoire) $\pi \in \mathcal{P}_+^1(\Theta)$.

Proposition 4.3 *Introduisons la notation*

$$B_\Lambda(q, \eta) = \inf_{\lambda \in \Lambda} \Phi_\lambda^{-1}\left(q + \frac{\eta}{\lambda}\right).$$

Pour toute loi *a priori* $\pi \in \mathcal{P}_+^1(\Theta)$ et tout $\lambda \in \mathbb{R}_+$,

$$\int \exp \left\{ \sup_{\rho \in \mathcal{P}_+^1(\Theta)} [m\lambda \{ \Phi_\lambda[f(P_Z, \rho)] - f(P_S, \rho) \} - \mathcal{K}(\rho, \pi)] \right\} dP_Z^{\otimes m} \leq 1, \quad (49)$$

et donc pour tout ensemble fini $\Lambda \subset \mathbb{R}_+$, avec une probabilité d'au moins $1 - \delta$, pour tout $\rho \in \mathcal{P}_+^1(\Theta)$,

$$f(P_Z, \rho) \leq B_\Lambda \left(f(P_S, \rho), \frac{\mathcal{K}(\rho, \pi) + \log(|\Lambda|/\delta)}{m} \right).$$

¹⁹. Nous supposons que ρ est un noyau de probabilités de transition régulier, c'est-à-dire que pour tout ensemble mesurable A , l'application $(z_1, \dots, z_m) \mapsto \rho(z_1, \dots, z_m | A)$ est mesurable. Nous supposons également que la tribu sur Θ est générée à partir d'un nombre fini de sous-ensembles de Θ .

Preuve L'identité (47) appliquée à $g(\theta) = m\lambda\{\Phi_\lambda[f(P, \theta)] - f(P_S, \theta)\}$ donne

$$\int m\lambda\{\Phi_\lambda[f(P, \theta)] - f(P_S, \theta)\}d\rho(\theta) - \mathcal{K}(\rho, \pi) \leq \log \left[\int \exp [m\lambda\{\Phi_\lambda[f(P, \theta)] - f(P_S, \theta)\}] d\pi(\theta) \right]$$

et, puisque $mf(P_S, \theta) = \sum_{i=1}^m f(Z_i, \theta)$, en posant $p = f(P, \theta)$ on a

$$\frac{\exp \{m\lambda\Phi_\lambda[f(P, \theta)]\}}{\exp [m\lambda f(P_S, \theta)]} = \frac{\exp [-m \log(1 - p + pe^{-\lambda})]}{\exp [\lambda \sum_i f(Z_i, \theta)]} = \prod_{i=1}^m \left(\frac{1 - p + pe^{-\lambda}}{e^{\lambda f(Z_i, \theta)}} \right) \leq 1,$$

car $\lambda \geq 0$. Enfin, la fonction Φ_λ étant convexe²⁰, l'inégalité de Jensen entraîne

$$\Phi_\lambda[f(P, \rho)] = \Phi_\lambda \left(\int f(P, \theta) d\rho(\theta) \right) \leq \int \Phi_\lambda[f(P, \theta)] d\rho(\theta).$$

Remarquons que P_S et ρ dépendent de S et sont donc aléatoires. Soient ρ une loi *a posteriori* et $\delta > 0$. Pour $\lambda \in \Lambda$, posons

$$A_\lambda = \{z_1^m \in \mathcal{Z}^m : \exp [m\lambda(\Phi_\lambda[f(P, \rho)] - f(P_S, \rho)) - \mathcal{K}(\rho, \pi)] \geq \delta^{-1}\}.$$

Il résulte de l'identité de Markov que $P_Z^{\otimes m}(A_\lambda) \leq \delta$ et donc $P_Z^{\otimes m}(\cup_{\lambda \in \Lambda} A_\lambda) \leq |\Lambda|\delta$. Ainsi, avec une probabilité d'au moins $1 - \delta$, $\forall \lambda \in \Lambda$,

$$\Phi_\lambda[f(P, \rho)] \leq f(P_S, \rho) + \frac{\log(|\Lambda|/\delta) + \mathcal{K}(\rho, \pi)}{m\lambda},$$

ce qui termine la démonstration. ◇

Pour $p \in [0; 1]$, posons

$$\bar{v}(p) = \begin{cases} p(1-p) & \text{si } p \leq 1/2 \\ 1/4 & \text{si } p \geq 1/2. \end{cases}$$

Proposition 4.4 Soient $m \in \mathbb{N}^*$ et

$$t = \frac{1}{4} \log \left(\frac{m}{8 \log[(m+1)/\delta]} \right).$$

Avec une probabilité d'au moins $1 - \delta$, pour toute loi *a posteriori* $\rho \in \mathcal{P}_+^1(\Theta)$,

$$f(P_Z, \rho) \leq f(P_S, \rho) + B_m[f(P_S, \rho), \mathcal{K}(\rho, \pi), \delta],$$

où

$$\begin{aligned} B_m(q, e, \delta) &= \max \left\{ \sqrt{\frac{2\bar{v}(q)\{e + \log[(m+1)/\delta]\}}{m}} \cosh(t/m) \right. \\ &\quad \left. + \frac{2(1-q)\{e + \log[(m+1)/\delta]\}}{m} \cosh^2(t/m), \frac{2\{e + \log[(m+1)/\delta]\}}{m} \right\} \\ &\leq \sqrt{\frac{2\bar{v}(q)\{e + \log[(m+1)/\delta]\}}{m}} \cosh(t/m) + \frac{2\{e + \log[(m+1)/\delta]\}}{m} \cosh^2(t/m). \end{aligned}$$

De plus, dès que $m \geq 5$, $B_{\lfloor \log(m)^2 \rfloor - 1}(q, e, \delta) \leq B(q, e, \delta)$, où

$$\begin{aligned} B(q, e, \delta) &= \sqrt{\frac{2\bar{v}(q)\{e + \log[\log(m)^2/\delta]\}}{m}} \cosh[\log(m)^{-1}] + \\ &\quad + \frac{2\{e + \log[\log(m)^2/\delta]\}}{m} \cosh^2[\log(m)^{-1}], \end{aligned} \tag{50}$$

20. En effet, $\Phi'_\lambda(p) = \frac{-1}{\lambda} \left(\frac{-1+e^{-\lambda}}{1-pe^{-\lambda}} \right)$ et $\Phi''_\lambda(p) = \frac{(-1+e^{-\lambda})^2}{\lambda(1-pe^{-\lambda})^2}$.

ainsi avec une probabilité d'au moins $1 - \delta$, pour toute loi a posteriori ρ ,

$$\begin{aligned} f(P_Z, \rho) &\leq f(P_S, \rho) + \sqrt{\frac{2\bar{v}[f(P_S, \rho)]\{\mathcal{K}(\rho, \pi) + \log[\log(m)^2/\delta]\}}{m}} \cosh[\log(m)^{-1}] \\ &\quad + \frac{2\{\mathcal{K}(\rho, \pi) + \log[\log(m)^2/\delta]\}}{m} \cosh^2[\log(m)^{-1}]. \end{aligned}$$

Preuve. Posons

$$\begin{aligned} q &= f(P_S, \rho), \\ \eta &= \frac{\mathcal{K}(\rho, \pi) + \log[(m+1)/\delta]}{m}, \\ \lambda_{\min} &= \sqrt{\frac{8 \log[(m+1)/\delta]}{m}}, \\ \Lambda &= \{\lambda_{\min}^{1-k/m}, k = 0, \dots, m\}, \\ p &= B_\Lambda(q, \eta), \\ \hat{\lambda} &= \sqrt{\frac{2\eta}{\bar{v}(p)}}. \end{aligned}$$

En appliquant l'identité (48) aux distributions de Bernoulli on obtient pour $g(x) = -\lambda x$:

$$-\lambda \Phi_\lambda(p) = -\lambda p + \int_0^1 (1-\alpha) \lambda^2 p_{\alpha\lambda} (1-p_{\alpha\lambda}) d\alpha = -\lambda p + \int_0^\lambda \frac{\lambda-\alpha}{\lambda} \lambda^2 p_\alpha (1-p_\alpha) \frac{d\alpha}{\lambda}$$

soit

$$\Phi_\lambda(p) = p - \frac{1}{\lambda} \int_0^\lambda (\lambda - \alpha) p_\alpha (1-p_\alpha) d\alpha \leq q + \frac{\eta}{\lambda}.$$

De plus²¹ $p_\alpha \leq p$, donc

$$p - q \leq \inf_{\lambda \in \Lambda} \left[\frac{\lambda \bar{v}(p)}{2} + \frac{\eta}{\lambda} \right] = \inf_{\lambda \in \Lambda} \sqrt{2\eta \bar{v}(p)} \cosh \left[\log \left(\frac{\hat{\lambda}}{\lambda} \right) \right].$$

Puisque $\bar{v}(p) \leq 1/4$ et $\eta \geq \frac{\log[(m+1)/\delta]}{m}$,

$$\sqrt{\frac{2\eta}{\bar{v}(p)}} = \hat{\lambda} \geq \lambda_{\min} = \sqrt{\frac{8 \log[(m+1)/\delta]}{m}}.$$

Si $\lambda_{\min} \leq \hat{\lambda} \leq \max \Lambda = 1$, alors $\log(\hat{\lambda})$ est à une distance d'au plus t/m d'un $\lambda \in \Lambda$, car $\log(\Lambda)$ est une grille à pas constant de taille $\frac{-\log(\lambda_{\min})}{m} = 2t/m$. Donc

$$p - q \leq \sqrt{2\eta \bar{v}(p)} \cosh(t/m).$$

Remarquons que pour tout λ , $\Phi_\lambda(0) = 0$, $\Phi_\lambda(1) = 1$ et $p \mapsto \Phi_\lambda(p)$ étant strictement croissante et convexe sur $[0; 1]$, $q \mapsto \Phi_\lambda^{-1}(q)$ est strictement croissante et concave sur $[0; 1]$. Donc pour tout $q \in [0; 1]$, $q \leq \Phi_\lambda^{-1}(q)$ et $q \leq \inf_{\lambda \in \Lambda} \Phi_\lambda^{-1}(q) \leq \inf_{\lambda \in \Lambda} \Phi_\lambda^{-1}(q + \frac{\eta}{\lambda}) = p$.

Ainsi, $p(1-p) \leq p(1-q)$ et si $q \leq 1/2$, alors $\bar{v}(p) \leq p(1-q)$, et en posant $c = \cosh(t/m)$, nous obtenons l'inégalité quadratique en p : $p^2 - 2p[q + c^2\eta(1-q)] + q^2$, de discriminant $c^2\eta(1-q)[2q + c^2\eta(1-q)]$, donc en notant p_2 la plus grande racine

$$p \leq p_2 = q + c^2\eta(1-q) + c\sqrt{2\eta q(1-q) + c^2\eta^2(1-q)^2} \leq q + c\sqrt{2\eta q(1-q)} + 2c^2\eta(1-q).$$

21. En effet, $(1-p) \leq (1-p)e^\alpha$ entraîne $[p + (1-p)e^\alpha]^{-1} \leq 1$, i.e., $\frac{p_\alpha}{p} \leq 1$.

Maintenant si $q \geq 1/2$, alors $\bar{v}(q) = \bar{v}(p) = \frac{1}{4}$ et $p - q \leq \sqrt{2\bar{v}(q)}c$, donc dans tous les cas

$$p - q \leq \sqrt{2\eta\bar{v}(q)} \cosh(t/m) + 2\eta(1 - q) \cosh^2(t/m). \quad (51)$$

Considérons maintenant le cas $\hat{\lambda} > 1$, donc $\bar{v}(p) < 2\eta$, dans ce cas $\inf_{\lambda \in \Lambda} \cosh \left[\log \left(\frac{\hat{\lambda}}{\lambda} \right) \right]$ est obtenu pour $\lambda = 1$ donc

$$p - q \leq \frac{\bar{v}(p)}{2} + \eta < 2\eta.$$

Pour conclure, appliquons la proposition 4.3 : avec une probabilité d'au moins $1 - \delta$ et pour toute loi *a posteriori* ρ ,

$$f(P_Z, \rho) \leq p \leq q + \max\{2\eta, \sqrt{2\eta\bar{v}(q)} \cosh(t/m) + 2\eta(1 - q) \cosh^2(t/m)\},$$

qui correspond à ce que l'on doit démontrer. Dans le cas particulier où $m = \lfloor \log(m)^2 \rfloor - 1 \geq \log(m)^2 - 2$,

$$\frac{t}{m} \leq \frac{1}{4[\log(m)^2 - 2]} \log \left(\frac{m}{8 \log[\log(m)^2 - 1]} \right) \leq \frac{1}{\log(m)}.$$

La dernière inégalité est valide dès que $m \geq \exp(\sqrt{2}) \simeq 4,11$ pour que $\log(m)^2 - 2 > 0$ et $3 \log(m)^2 - 8 + \log(m) \log\{8 \log[\log(m)^2 - 1]\} \geq 0$ (on peut le vérifier numériquement). $\diamond$

Références

- [1] S. Mallat, *L'apprentissage face à la malédiction de la grande dimension*, Cours du Collège de France, 2018.
- [2] S. Shalev-Shwartz et S. Ben-David, *Understanding Machine Learning, from theory to algorithms*, Cambridge University Press, 2014.
- [3] E. Slud, "Distribution inequalities for the binomial law", *The Annals of Probability*, vol. 5, no. 3, pp. 404–412, 1977.
- [4] M. Telgarsky, "Central binomial tail bounds", preprint, <https://arxiv.org/abs/0911.2077>, 2010.
- [5] O. Catoni, *Measures of complexity, Festschrift for Alexey Chervonenkis*, Vladimir Vovk, Harris Papadopoulos, Alexander Gammerman Editors, chap. 20, Springer, 2015.