Fondamentaux de l'apprentissage statistique
Lecture Notes

M. Lerasle

2021-2022

Table des matières

Chapitre 1

Minimiseurs du risque empirique

1.1 Classification binaire

Les méthodes présentées dans ce cours seront illustrées dans le problème de la classification binaire supervisée, un des problèmes les plus simples de l'apprentissage. Le but est de vous fournir les outils et méthodes pour attaquer ce type de problèmes d'un point de vue *théorique*. Vous ne verrez pas dans ce cours les derniers algorithmes, mais des outils méthodologiques et mathématiques fondamentaux, communs à différents problèmes.

Le but de la classification binaire est d'attribuer une étiquette $y \in \{0, 1\}$ (retrouver la classe) à un vecteur $x \in \mathcal{X}$. Le vecteur x peut encoder un mail par exemple et l'étiquette peut être 0 si le mail n'est pas un spam et 1 s'il en est un.

Pour procéder, les algorithmes d'apprentissage *supervisés* disposent d'un dataset d'entraînement $\mathcal{D}_n = \{(x_i, y_i)_{i=1,\dots,n}\}$ de vecteurs déjà étiquetés, possiblement avec des erreurs (certains mails ont pu être mal classés dans ce dataset). Le but de l'algorithme est de *généraliser* à partir de cette base. Il va chercher à construire une fonction $\hat{f} : \mathcal{X} \to \{0, 1\}$, de façon à pouvoir ensuite, pour tout nouveau vecteur X, lui attribuer l'étiquette $\hat{f}(X)$. Cette fonction dépend typiquement de $\mathcal{D}_n$, une information résumée par le chapeau sur la fonction $\hat{f}$ plutôt que par la notation plus lourde $f(\mathcal{D}_n, \cdot)$.

Pour comparer différents algorithmes, on utilise une fonction de perte ℓ (comme loss en anglais) qui, à toute fonction $f : \mathcal{X} \to \{0, 1\}$ et à tout couple (x, y) associe un réel $\ell_f(x, y)$ mesurant l'erreur commise par le *classifieur* f pour prédire l'étiquette y à partir du vecteur x. On utilisera dans tout le début du cours la perte $0 - 1$ définie par $\ell_f(x, y) = \mathbf{1}_{\{f(x) \neq y\}}$ qui vaut donc 1 si le classifieur f a mal prédit l'étiquette y de x et 0 sinon.

En toute généralité, il est possible d'espérer faire des prédictions pour un vecteur x qu'on aurait déjà observé dans la base d'entraînement (i.e.

si $x \in \{x_1, \dots, x_n\}$), par exemple en faisant un vote majoritaire parmi les y_i tels que $x_i = x$, mais dans la plupart des applications, il est illusoire d'espérer avoir déjà observé (plusieurs fois) tous les scénarios envisageables (un détecteur de spams doit ainsi pouvoir faire une prédiction sans être entraîné sur tous les textes possibles).

Pour généraliser à partir d'un dataset nécessairement partiel, on utilise un formalisme probabiliste inspiré des statistiques. On voit les données d'entraînement comme des réalisations de variables aléatoires $(X_i, Y_i)_{i=1,\dots,n}$ de loi inconnue P et la nouvelle donnée (X, Y) comme une nouvelle variable aléatoire indépendante du dataset d'entraînement, encore de loi P. Muni de ce nouveau formalisme, on peut définir le *risque* de tout classifieur f comme $R(f) = \mathbb{E}[\ell_f(X, Y)]$, c'est à dire la perte moyenne (sous la loi P) du classifieur f. En cherchant à minimiser le risque plutôt que la perte, on ne va plus chercher un classifieur se comportant bien *pour toute valeur de* x, mais se comportant bien en moyenne, c'est à dire typiquement très bien sur des x “typiques", qui ressemblent à des x_i déjà observés, et peut-être moins bien sur des x_i plus atypiques. Cette nouvelle tâche a toutes les chances d'être plus raisonnable. Formellement, un classifieur $\hat{f}$ construit à partir des données d'entraînement est alors un classifieur *aléatoire*, de sorte que la valeur de la fonction de risque en un tel classifieur aléatoire $R(\hat{f})$ est elle-même une variable aléatoire. Pour le voir, on peut formellement l'écrire comme

$$R(\hat{f}) = \mathbb{E}[\ell_{\hat{f}}(X, Y) | \mathcal{D}_n] \ ,$$

c'est à dire qu'on intègre la perte *par rapport à l'aléa de la nouvelle donnée uniquement.*

Le but de ce premier chapitre est essentiellement de donner les outils principaux permettant de construire des bornes de généralisation, c'est à dire des bornes de la forme

$$\mathbb{P}(R(\hat{f}) > \Delta) \leqslant \epsilon \ ,$$

pour quelques estimateurs classiques $\hat{f}$.

1.2 Estimateur de Bayes

Au vu du formalisme introduit à la section précédente, un classifieur idéal minimise le risque, et est donc défini comme n'importe quel élément

$$f^* \in \operatorname{argmin}_{f \in \mathcal{F}_0} R(f) \ ,$$

où $\mathcal{F}_0$ désigne l'ensemble de toutes les fonctions $f : \mathcal{X} \to \{0, 1\}$. Lorsque ℓ est la perte $0-1$, la valeur de ce classifieur idéal (appelé *classifieur de Bayes*) peut être explicitée à partir de la fonction de régression définie informellement par

$$\eta(x) = \mathbb{P}(Y = 1 | X = x) = \mathbb{E}[Y | X = x] \ .$$

Rappelons que cette fonction de régression vérifie en effet que, pour toute fonction $\psi : \mathcal{X} \to \mathbb{R}$ intégrable

$$\mathbb{E}[Y\psi(X)] = \mathbb{E}[\eta(X)\psi(X)] \ .$$

Pour exprimer la valeur de l'estimateur de Bayes à partir de η, on va alors utiliser le fait que y et $f(x)$ valent 0 ou 1 pour écrire

$$\ell_f(x,y) = \mathbf{1}_{\{f(x)\neq y\}} = y(1-f(x)) + (1-y)f(x) \ .$$

En utilisant la propriété de l'espérance conditionnelle, on a alors, pour tout $f \in \mathcal{F}_0$,

$$R(f) = \mathbb{E}[\eta(X)(1-f(X)) + (1-\eta(X))f(X)] \ . \tag{1.1}$$

Or, pour tout $x \in \mathcal{X}$, la fonction intégrée

$$\eta(x)(1-f(x)) + (1-\eta(x))f(x) = \begin{cases} \eta(x) & \text{si } f(x) = 0 \ , \\ 1-\eta(x) & \text{si } f(x) = 1 \ . \end{cases}$$

Pour toute valeur de x, cette fonction est donc minimale pour toute fonction f^* telle que

$$\begin{aligned} f^*(x) &= \begin{cases} 1 & \text{si } \eta(x) > 1-\eta(x) \\ 0 & \text{si } \eta(x) < 1-\eta(x) \end{cases} \\ &= \begin{cases} 1 & \text{si } \eta(x) > 1/2 \ , \\ 0 & \text{si } \eta(x) < 1/2 \ . \end{cases} \end{aligned}$$

La valeur de f^* sur l'ensemble $\{\eta(x) = 1/2\}$ peut être choisie arbitrairement. Autrement dit, tout estimateur idéal f^* (aussi appelé estimateur de Bayes) attribue l'étiquette 1 à x lorsque la probabilité que $Y = 1$ sachant x est supérieure à $1/2$ et 0 sinon (tout ça pour ça !!).

Notons qu'on déduit alors immédiatement de l'équation (1.1) que le risque de l'estimateur de Bayes vaut alors

$$R(f^*) = \mathbb{E}[\min(\eta(X), 1-\eta(X))] \ .$$

De plus, $R(f^*)$ étant clairement une borne inférieure sur le risque $R(f)$ de tout classifieur, on va chercher à borner, plutôt que le risque lui même, l'excès de risque défini par

$$\mathcal{E}(f) = R(f) - R(f^*) \ .$$

Toujours d'après l'équation (1.1), cet excès de risque vaut alors

$$\begin{aligned} \mathcal{E}(f) &= \mathbb{E}[\eta(X)(f^*(X) - f(X)) + (1-\eta(X))(f(X) - f^*(X))] \\ &= \mathbb{E}[(1-2\eta(X))(f(X) - f^*(X))] \ . \end{aligned}$$

Par définition de f^*, $(1-2\eta(X))$ et $(f(X)-f^*(X))$ sont de même signe, donc

$$\mathcal{E}(f) = \mathbb{E}[|1-2\eta(X)||f(X)-f^*(X)|] \ .$$

Rassemblons les informations importantes de cette section dans le théorème suivant.

Théorème 1. *L'estimateur de Bayes minimisant le risque* $0-1$ *est égal à*

$$f^*(x) = \mathbf{1}_{\{\eta(x)>1/2\}} \ ,$$

où η *est la fonction de régression* $\eta(x) = \mathbb{E}[Y|X=x]$.

Son risque (pour la perte $0-1$*) vaut*

$$R(f^*) = \mathbb{E}[\min(\eta(X), 1-\eta(X))] \ .$$

L'excès de risque de tout classifieur f *vaut alors*

$$\mathcal{E}(f) = R(f) - R(f^*) = \mathbb{E}[|1-2\eta(X)||f(X)-f^*(X)|] \ .$$

1.3 Minimiseur du risque empirique

Le classifieur de Bayes f^* est un classifieur idéal mais inaccessible en pratique car la loi P est inconnue, donc la fonction de risque R aussi. Pour l'approcher, l'idée est de définir plutôt que le risque de tout classifieur f $R(f)$, le risque empirique

$$R_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell_f(X_i, Y_i) \ .$$

La loi des grands nombres garantit que $R_n(f) \to R(f)$, pour tout $f \in \mathcal{F}_0$, en probabilité, cet estimateur est donc raisonnable a priori. Le risque empirique est calculable facilement pour tout classifieur f et il est (au moins en théorie) possible de minimiser la fonction R_n sur tout ensemble $\mathcal{F} \subset \mathcal{F}_0$ de classifieurs, définissant ainsi le *minimiseur du risque empirique*

$$\hat{f}_{\mathcal{F}} \in \operatorname{argmin}_{f\in\mathcal{F}} R_n(f) \ .$$

Remarquons qu'on devrait plutôt parler des minimiseurs du risque empirique puisqu'on a le choix de la classe $\mathcal{F}$ et rien ne garantit en général l'unicité d'un minimiseur (quand déjà il en existe). Nous conviendrons dans la suite qu'un tel minimiseur existe et que $\hat{f}_{\mathcal{F}}$ en désigne un particulier s'il y en a plusieurs. Ceci n'est qu'une convention pratique et le lecteur (suspicieux et) intéressé pourra vérifier que les propriétés que nous allons donner sur $\hat{f}_{\mathcal{F}}$ sont partagées (à de mineures modifications près dans les notations et quelques

lourdeurs techniques supplémentaires) par tous les "quasi-minimiseurs" $\hat{f}_{\mathcal{F},n}$ tels que

$$\forall f \in \mathcal{F}, \qquad R_n(\hat{f}_{\mathcal{F},n}) \leqslant R_n(f) + \frac{1}{n} \ .$$

Notons que $\mathcal{F}$ est un sous-ensemble de $\mathcal{F}_0$ en général, et dans l'essentiel des cas particuliers également. Ceci n'est pas étonnant : pour la perte $0-1$ par exemple, pour peu que tous les x_i soient distincts (ce qui est le cas p.s. dès que $\mathcal{X}$ et P sont continus), la perte empirique est minimisée sur $\mathcal{F}_0$ par tout classifieur $\hat{f}$ classant parfaitement les données d'entraînement. Il est aisé de se convaincre que cette stratégie de surapprentissage revenant à accorder une confiance aveugle à toutes les étiquettes de l'ensemble d'apprentissage alors que celles-ci ont pu être entachées d'erreur est nocive en général et peut conduire à de mauvaises capacités de généralisation de l'estimateur $\hat{f}_{\mathcal{F}}$.

Cela dit, dès lors que $\hat{f}_{\mathcal{F}}$ est défini comme un élément de $\mathcal{F}$, son excès de risque se décompose de la façon suivante :

$$\mathcal{E}(\hat{f}_{\mathcal{F}}) = R(\hat{f}_{\mathcal{F}}) - \min_{f\in\mathcal{F}} R(f) + \min_{f\in\mathcal{F}} R(f) - R(f^*) \ . \tag{1.2}$$

Là encore, par commodité, nous allons admettre l'existence dans la suite d'un classifieur $f^*_{\mathcal{F}}$ idéal dans la classe $\mathcal{F}$ défini par

$$f^*_{\mathcal{F}} \in \operatorname{argmin}_{f\in\mathcal{F}} R(f) \ .$$

Ainsi, on a $\min_{f\in\mathcal{F}} R(f) = R(f^*_{\mathcal{F}})$ et $\min_{f\in\mathcal{F}} R(f) - R(f^*) = \mathcal{E}(f^*_{\mathcal{F}})$. Dans la décomposition (1.2) de l'excès de risque de l'estimateur $\hat{f}_{\mathcal{F}}$ les deux termes

$$R(\hat{f}_{\mathcal{F}}) - R(f^*_{\mathcal{F}}) \quad \text{et} \quad \mathcal{E}(f^*_{\mathcal{F}}) \ ,$$

sont positifs presque sûrement. Le premier décrit l'erreur commise en minimisant sur la classe $\mathcal{F}$ le risque empirique plutôt que le vrai risque, tandis que le second mesure l'erreur commise en utilisant le sous-ensemble $\mathcal{F}$ plutôt que la classe complète $\mathcal{F}_0$.

Le choix d'un bon ensemble $\mathcal{F}$ est donc le fruit d'un arbitrage entre ces deux types d'erreur : d'un côté, il doit être suffisamment simple pour qu'on puisse garantir un bon contrôle du terme aléatoire $R(\hat{f}_{\mathcal{F}}) - R(f^*_{\mathcal{F}})$. Cette perte se réduisant avec le nombre de données, "simple" est relatif au nombre de données disponibles. De l'autre, il doit être suffisamment complexe pour espérer approcher une grande variété de classifieurs et contrôler $\mathcal{E}(f^*_{\mathcal{F}})$. Typiquement, on voit en particulier que le choix d'un modèle optimal va dépendre du nombre n de données à disposition et de la complexité du classifieur optimal f^*. Dans ce chapitre, nous nous concentrons principalement sur la compréhension du terme aléatoire $R(\hat{f}_{\mathcal{F}}) - R(f^*_{\mathcal{F}})$ dans cette décomposition, même si nous décrivons rapidement dans la prochaine section les grandes étapes et types de résultats permettant de contrôler le terme déterministe.

1.4 Contrôle du biais de modèle $\mathcal{E}(f^*_{\mathcal{F}})$

Ce terme est indépendant des données *une fois le modèle $\mathcal{F}$ choisi*. Pour qu'il diminue asymptotiquement avec le nombre de données, il sera donc nécessaire qu'il grandisse avec n. Le choix d'un modèle optimal dépend typiquement d'a priori qu'on a sur la fonction f^*. Plus celle-ci est complexe (ce qui est le cas, par exemple si la frontière entre les ensembles $\mathcal{X}_i = \{x \in \mathcal{X} : f^*(x) = i\}$ est très irrégulière), plus on va devoir choisir un modèle $\mathcal{F}$ complexe pour que cette erreur soit rendue plus petite qu'une erreur ϵ. En l'absence d'informations fortes sur cette fonction inconnue f^*, on utiliser des modèles $\mathcal{F}$ d'approximation universelle. Pour procéder, on va suivre généralement les étapes suivantes :

1. On approche f^* par sa restriction $f^*_K = f^*\mathbf{1}_K$ où K est un "grand" ensemble compact, par exemple la boule $B(0, \rho)$ où ρ est grand.
2. On approche $f^*\mathbf{1}_K$ par un ensemble $\mathcal{F}$ de fonctions "simples".

Cette stratégie peut être raffinée de diverses façons, mais elle conduit à deux types de bornes :

1. Des résultats de consistance universelle, sans hypothèse sur f^*, car pour K suffisamment grand l'erreur commise à l'étape 1 sera plus petite que ϵ et, si $\mathcal{F}$ est suffisamment complexe, l'erreur commise à l'étape 2 sera elle aussi plus petite que ϵ.
2. Des vitesses de convergence sous hypothèses fortes sur f^*, typiquement f^* est à support compact et à la régularité contrôlée, de sorte qu'on puisse avoir une version quantitative des erreurs commises lors des deux étapes d'approximation dans l'approche précédente.

Vous étudierez plus en détail certaines de ces approches dans la partie du cours d'Erwan, on va se concentrer dans le reste de ce chapitre sur le contrôle du terme aléatoire $R(\hat{f}_{\mathcal{F}}) - R(f^*_{\mathcal{F}})$ de l'excès de risque.

1.5 Concentration de l'excès de risque

Pour borner le terme $R(\hat{f}_{\mathcal{F}}) - R(f^*_{\mathcal{F}})$, on procède généralement en deux étapes :

1. On utilise la *concentration de la mesure* pour majorer les déviations de $R(\hat{f}_{\mathcal{F}}) - R(f^*_{\mathcal{F}})$ autour de son espérance $\mathbb{E}[R(\hat{f}_{\mathcal{F}}) - R(f^*_{\mathcal{F}})]$.
2. On utilise ensuite diverses méthodes pour contrôler l'espérance $\mathbb{E}[R(\hat{f}_{\mathcal{F}}) - R(f^*_{\mathcal{F}})]$.

Le but de cette section est de donner des outils de concentration de la mesure suffisants pour la première étape de ce programme. Le gros avantage de cette approche est son coté *non-asymptotique* qui permet d'accéder aux ordre de grandeurs importants, sans avoir à calibrer a priori le modèle $\mathcal{F}$.

1.5.1 Concentration de la mesure

Le but de cette section est de démontrer l'inégalité de Mc-Diarmid. Ce résultat est un résultat de concentration de la mesure au sens suivant : Selon cette théorie, toute fonction *régulière* h de variables aléatoires $X_1, \dots, X_n$ *indépendantes* est pratiquement constante (et donc à peu près égale à son espérance). La théorie formalise cette heuristique.

Pour l'inégalité de Mc-Diarmid, on s'intéresse aux fonctions régulières au sens suivant :

Définition 2. *Une fonction $h : \prod_{i=1}^n \mathcal{X}_i \to \mathbb{R}$ est dite à variations bornées par $c = (c_1, \dots, c_n)$ si elle vérifie*

$$\forall x = (x_1, \dots, x_n), y = (y_1, \dots, y_n), \qquad |h(x) - h(y)| \leqslant \sum_{i=1}^n c_i \mathbf{1}_{\{x_i \neq y_i\}} \ .$$

Dit autrement, si x et y ne diffèrent que sur la i-ème coordonnées, h ne dévie que de c_i au maximum. L'hypothèse d'être à variations bornées revient donc à être Lipschitzienne par rapport à une distance de Hamming à poids.

Pour les fonctions à variations bornées, l'inégalité de Mc-Diarmid quantifie le "presque constant" en majorant, pour tout $z > 0$, la probabilité que la variable $h(X_1, \dots, X_n)$ dévie de son espérance de plus de z : $\mathbb{P}(h(X_1, \dots, X_n) - \mathbb{E}[h(X_1, \dots, X_n)] > z)$.

Théorème 3. *Soit $h : \prod_{i=1}^n \mathcal{X}_i \to \mathbb{R}$ une fonction à variations bornées par un vecteur $c \in \mathbb{R}^n$, de norme Euclidienne $\|c\|$ et $X_1, \dots, X_n$ des variables aléatoires indépendantes. Alors, pour tout $s > 0$, on a*

$$\mathbb{E}[\exp(s(h(X_1, \dots, X_n) - \mathbb{E}[h(X_1, \dots, X_n)]))] \leqslant \exp\left(\frac{\|c\|^2 s^2}{8}\right) \ ,$$

En particulier, pour tout $z > 0$, on a

$$\mathbb{P}(h(X_1, \dots, X_n) - \mathbb{E}[h(X_1, \dots, X_n)] > z) \leqslant \exp\left(-\frac{2z^2}{\|c\|^2}\right) \ .$$

Démonstration. On procède en trois temps.

1ère étape : Décomposition en incréments de martingale.

On note $\mathcal{G}_0 = \{\emptyset, \Omega\}$ la tribu triviale et, pour tout $i \geqslant 1$, $\mathcal{G}_i$ la tribu engendrée par $X_1, \dots, X_i$, de sorte que $h(X_1, \dots, X_n) = \mathbb{E}[h(X_1, \dots, X_n)|\mathcal{G}_n]$, $\mathbb{E}[h(X_1, \dots, X_n)] = \mathbb{E}[h(X_1, \dots, X_n)|\mathcal{G}_0]$ et donc

$$h(X_1, \dots, X_n) - \mathbb{E}[h(X_1, \dots, X_n)] = \sum_{i=1}^n \Delta_i \ ,$$
$$\Delta_i = \mathbb{E}[h(X_1, \dots, X_n)|\mathcal{G}_i] - \mathbb{E}[h(X_1, \dots, X_n)|\mathcal{G}_{i-1}] \ .$$

2ème étape : Contrôle de la transformée de Laplace via le lemme d'Hoeffding.

Le lemme d'Hoeffding donne une borne supérieure sur la transformée de Laplace d'une variable aléatoire bornée.

Lemma 1 (Hoeffding)**.** *Soit X une variable aléatoire à valeurs dans un intervalle $[a,b]$ presque-sûrement, alors, pour tout $s \in \mathbb{R}$,*

$$\mathbb{E}[\exp(s(X - \mathbb{E}[X]))] \leqslant \exp\left(\frac{s^2(b-a)^2}{8}\right) .$$

Ce résultat est la pierre angulaire de l'inégalité de Mc-Diarmid. Le lecteur pressé peut l'admettre dans un premier temps. On en donne deux preuves, la première complètement élémentaire mais fournissant un résultat légèrement sous-optimal, la seconde plus astucieuse donnant le résultat avec les meilleures constantes.

Preuve élémentaire. Notons qu'on a toujours $|X - \mathbb{E}[X]| \leqslant b - a$, donc

$$\mathbb{E}[\exp(s(X - \mathbb{E}[X]))] \leqslant \mathbb{E}[\exp(|s|(b-a))] ,$$

et par conséquent, si $|s|(b-a) \geqslant 1$, on a $|s|(b-a) \leqslant s^2(b-a)^2$ et donc

$$\mathbb{E}[\exp(s(X - \mathbb{E}[X]))] \leqslant \mathbb{E}[\exp(s^2(b-a)^2)] .$$

Il suffit donc de montrer le résultat pour les s tels que $|s|(b-a) \leqslant 1$. Or, sous cette condition, on a

$$\begin{aligned}
\exp(s(X - \mathbb{E}[X])) &= 1 + s(X - \mathbb{E}[X]) + \sum_{k=2}^{+\infty} \frac{s^k (X - \mathbb{E}[X])^k}{k!} \quad \text{dvpt en série entière} \\
&\leqslant 1 + s(X - \mathbb{E}[X]) + \sum_{k=2}^{+\infty} \frac{|s|^k (b-a)^k}{k!} \quad \text{car } |X - \mathbb{E}[X]| \leqslant b - a \\
&\leqslant 1 + s(X - \mathbb{E}[X]) + \sum_{k=2}^{+\infty} \frac{|s|^k (b-a)^k}{2 * 3^{k-2}} \quad \text{car } k! \geqslant 2 * 3^{k-2} \\
&= 1 + s(X - \mathbb{E}[X]) + \frac{s^2(b-a)^2}{2(1 - |s|(b-a)/3)} \\
&\leqslant 1 + s(X - \mathbb{E}[X]) + \frac{s^2(b-a)^2}{2(1-1/3)} \quad \text{car } |s|(b-a) \leqslant 1 \\
&\leqslant 1 + s(X - \mathbb{E}[X]) + s^2(b-a)^2 \\
&\leqslant s(X - \mathbb{E}[X]) + \exp(s^2(b-a)^2) \quad \text{car } 1 + u \leqslant \exp(u) .
\end{aligned}$$

En prenant l'espérance des deux cotés de la dernière inégalité, on obtient donc

$$\mathbb{E}[\exp(s(X - \mathbb{E}[X]))] \leqslant \exp(s^2(b-a)^2) .$$

On obtient donc le résultat avec une constante 1 au lieu de 1/8. □

Donnons maintenant la preuve fournissant les constantes optimales.

Preuve optimale. Supposons X centrée et posons $\psi : s \to \log \mathbb{E}[\exp(sX)]$. La fonction ψ est de classe $\mathcal{C}^\infty$ et vérifie $\psi(0) = 0$, $\psi'(0) = \mathbb{E}[X]/1 = 0$ et enfin

$$\psi''(s) = \frac{\mathbb{E}[X^2 \exp(sX)]}{\mathbb{E}[\exp(sX)]} - \left(\frac{\mathbb{E}[X \exp(sX)]}{\mathbb{E}[\exp(sX)]}\right)^2 = \mathrm{Var}_P(U) \ ,$$

où P est la loi sur $[a, b]$ telle que, pour toute fonction g,

$$\mathbb{E}_P[g(U)] = \mathbb{E}\left[g(X)\frac{\exp(sX)}{\mathbb{E}[\exp(sX)]}\right] \ .$$

Cette dernière remarque est la grande astuce de la preuve du lemme d'Hoeffding. En effet, on en déduit

$$\psi''(s) = \mathrm{Var}_P\left(U - \frac{b-a}{2}\right) \leqslant \mathbb{E}_P\left[\left(U - \frac{b-a}{2}\right)^2\right] \leqslant \frac{(b-a)^2}{4} \ .$$

Finalement, on peut majorer

$$\psi(s) = \int_0^s \psi'(u)\mathrm{d}u = \int_0^s \int_0^u \psi''(y)\mathrm{d}y \leqslant \frac{(b-a)^2 s^2}{8} \ .$$

□

On peut maintenant conclure la preuve de l'inégalité de Mc-Diarmid. On procède pour la première partie récursivement en conditionnant d'abord par $\mathcal{G}_{n-1}$, de sorte que

$$\mathbb{E}[\exp(s\sum_{i=1}^n \Delta_i)] = \mathbb{E}[\exp(s\sum_{i=1}^{n-1} \Delta_i)\mathbb{E}[\exp(s\Delta_n)|\mathcal{G}_{n-1}]] \ .$$

Or conditionnellement à $X_1, \ldots, X_{n-1}$, Δ_n est une variable aléatoire centrée à valeur dans l'intervalle $[a, b]$, où

$$\begin{aligned} a &= \inf_{x_n \in \mathcal{X}_n} f(X_1, \ldots, X_{n-1}, x_n) \\ b &= \sup_{x_n \in \mathcal{X}_n} f(X_1, \ldots, X_{n-1}, x_n) \ . \end{aligned}$$

Comme f est à variations bornées par c, on a $b - a \leqslant c_n$ p.s. On a donc, d'après le lemme d'Hoeffding

$$\mathbb{E}[\exp(s\Delta_n)|\mathcal{G}_{n-1}] \leqslant \exp\left(\frac{c_n^2 s^2}{8}\right) \ .$$

En procédant récursivement, on obtient bien la première conclusion que

$$\mathbb{E}[\exp(s\sum_{i=1}^{n}\Delta_i)] \leqslant \exp\left(\frac{\|c\|^2 s^2}{8}\right) .$$

Pour la seconde partie du théorème, on utilise la méthode suivante.

3ème étape : Inégalité de Markov exponentielle.

Soit $Z = h(X_1,\dots,X_n) - \mathbb{E}[h(X_1,\dots,X_n)] = \sum_{i=1}^{n}\Delta_i$. On a, pour tout $s > 0$,

$$\begin{aligned}
\mathbb{P}(Z > z) &= \mathbb{P}(\exp(sZ) > \exp(st)) \qquad \text{croissance de } x \mapsto \exp(sx) \\
&\leqslant \frac{\mathbb{E}[\exp(sZ)]}{\exp(st)} \qquad \text{Markov + } \exp(sZ) \geqslant 0 \text{ p.s.} \\
&= \frac{\mathbb{E}[\exp(s\sum_{i=1}^{n}\Delta_i)]}{\exp(st)} .
\end{aligned}$$

Fin de la preuve : On injecte le résultat de la première partie du théorème dans l'inégalité de Markov exponentielle, on a

$$\mathbb{P}(Z > z) \leqslant \exp\left(-sz + \frac{\|c\|^2 s^2}{8}\right) .$$

Pour $s = 4z/\|c\|^2$, on obtient le résultat. □

1.5.2 Application aux maxima de processus empiriques

Dans cette section, on donne une application fort utile de l'inégalité de Mc-Diarmid. Soient $Z_1,\dots,Z_n$ des variables aléatoires indépendantes à valeurs dans des espaces mesurables $\mathcal{Z}_1,\dots,\mathcal{Z}_n$. Soit $\mathcal{G}_1,\dots,\mathcal{G}_n$ des ensembles de fonctions $g_i : \mathcal{Z}_i \to [a_i, b_i]$. On définit la fonction

$$f(z_1,\dots,z_n) = \sup_{g_1\in\mathcal{G}_1,\dots,g_n\in\mathcal{G}_n}\sum_{i=1}^{n} g_i(z_i) .$$

On vérifie que f est à variations bornées par le vecteur $c \in \mathbb{R}^n$ de coordonnées $c_i = b_i - a_i$. En effet, si $z = (z_1,\dots,z_n)$ et $z' = (z'_1,\dots,z'_n)$ sont deux vecteurs tels que $z_i = z'_i$, pour tout $i \neq j$, alors

$$\begin{aligned}
f(z) - f(z') &\leqslant \sup_{g_1\in\mathcal{G}_1,\dots,g_n\in\mathcal{G}_n}\sum_{i=1}^{n}(g_i(z_i) - g_i(z'_i)) \\
&= \sup_{g_j\in\mathcal{G}_j} g_j(z_j) - g_j(z'_j) \leqslant b_j - a_j .
\end{aligned}$$

On déduit alors directement de l'inégalité de Mc-Diarmid le résultat suivant.

Théorème 4. *Soient $Z_1, \dots, Z_n$ des variables aléatoires indépendantes à valeurs dans des espaces mesurables $\mathcal{Z}_1, \dots, \mathcal{Z}_n$. Soit $\mathcal{G}_1, \dots, \mathcal{G}_n$ des ensembles de fonctions $g_i : \mathcal{Z}_i \to [a_i, b_i]$. Alors, la variable aléatoire $S = \sup_{g_1 \in \mathcal{G}_1, \dots, g_n \in \mathcal{G}_n} \sum_{i=1}^n g_i(Z_i)$ vérifie, pour tout $s > 0$,*

$$\mathbb{E}[\exp(s(S - \mathbb{E}[S])) \leqslant \exp\left(\frac{s^2 \sum_{i=1}^n (b_i - a_i)^2}{8}\right) .$$

En particulier, pour tout $z > 0$,

$$\mathbb{P}(S - \mathbb{E}[S] > z) \leqslant \exp\left(-\frac{2z^2}{\sum_{i=1}^n (b_i - a_i)^2}\right) .$$

1.5.3 Concentration de l'excès de risque

Rappelons que le but du jeu est de majorer le terme $R(\hat{f}_{\mathcal{F}}) - R(f^*_{\mathcal{F}})$. On commence par la remarque suivante

$$R(\hat{f}_{\mathcal{F}}) - R(f^*_{\mathcal{F}}) = R(\hat{f}_{\mathcal{F}}) - R_n(\hat{f}_{\mathcal{F}}) + R_n(f^*_{\mathcal{F}}) - R(f^*_{\mathcal{F}}) + R_n(\hat{f}_{\mathcal{F}}) - R_n(f^*_{\mathcal{F}}) .$$

La définition de l'estimateur $\hat{f}_{\mathcal{F}}$ implique que $R_n(\hat{f}_{\mathcal{F}}) - R_n(f^*_{\mathcal{F}}) \leqslant 0$, donc

$$R(\hat{f}_{\mathcal{F}}) - R(f^*_{\mathcal{F}}) \leqslant R(\hat{f}_{\mathcal{F}}) - R_n(\hat{f}_{\mathcal{F}}) + R_n(f^*_{\mathcal{F}}) - R(f^*_{\mathcal{F}}) .$$

On peut alors majorer "brutalement"

$$R(\hat{f}_{\mathcal{F}}) - R(f^*_{\mathcal{F}}) \leqslant \sup_{f \in \mathcal{F}} \{R(f) - R_n(f) + R_n(f^*_{\mathcal{F}}) - R(f^*_{\mathcal{F}})\} .$$

On pose alors $z_i = (x_i, y_i)$ et

$$\forall f \in \mathcal{F}, \qquad g_f(z_i) = \frac{1}{n}(\ell_{f^*}(x_i, y_i) - R(f^*) - (\ell_f(x_i, y_i) - R(f))) ,$$

de sorte que $S = \sup_{f \in \mathcal{F}} \sum_{i=1}^n g_f(Z_i) = \sup_{f \in \mathcal{F}} \{R(f) - R_n(f) + R_n(f^*_{\mathcal{F}}) - R(f^*_{\mathcal{F}})\}$.

Clairement, $g_f(z_i) \in [-2/n, 2/n]$ donc, d'après le théorème 4 pour tout $z > 0$, on a

$$\mathbb{P}(S - \mathbb{E}[S] > z) \leqslant \exp\left(-\frac{nz^2}{2}\right) .$$

Cette inégalité peut être réécrite, pour tout $\delta \in (0, 1)$,

$$\mathbb{P}\left(S - \mathbb{E}[S] \leqslant \sqrt{\frac{2\log(1/\delta)}{n}}\right) \geqslant 1 - \delta .$$

De plus, on a

$$\mathbb{E}[S] = \mathbb{E}[\sup_{f \in \mathcal{F}} \{R(f) - R_n(f) + R_n(f^*_{\mathcal{F}}) - R(f^*_{\mathcal{F}})\}] = \mathbb{E}[\sup_{f \in \mathcal{F}} \{R(f) - R_n(f)\}] .$$

Résumons les résultats obtenus dans cette section dans un théorème.

Théorème 5. *Le minimiseur du risque empirique sur le modèle $\mathcal{F}$ vérifie*

$$\begin{aligned} R(\hat{f}_{\mathcal{F}}) - R(f^*_{\mathcal{F}}) &\leqslant R(\hat{f}_{\mathcal{F}}) - R_n(\hat{f}_{\mathcal{F}}) + R_n(f^*_{\mathcal{F}}) - R(f^*_{\mathcal{F}}) \\ &\leqslant \sup_{f\in\mathcal{F}}\{R(f) - R_n(f) + R_n(f^*_{\mathcal{F}}) - R(f^*_{\mathcal{F}})\} \ . \end{aligned}$$

De plus, pour tout $\delta \in (0,1)$, on a en outre

$$\mathbb{P}\left(R(\hat{f}_{\mathcal{F}}) - R(f^*_{\mathcal{F}}) \leqslant \mathbb{E}[\sup_{f\in\mathcal{F}}\{R(f) - R_n(f)\}] + \sqrt{\frac{2\log(1/\delta)}{n}} \right) \geqslant 1-\delta \ .$$

Notons que la première partie du résultat est vraie pour toute perte ℓ. Pour la seconde partie, nous n'avons utilisé que le fait que $\ell_f(x_i, y_i) \in [0,1]$, ce résultat est donc également valide pour toute perte bornée à de mineures modifications près, et pas seulement pour la perte $0-1$.

Pour majorer le risque du minimiseur du risque empirique, il reste donc à majorer le terme

$$\mathbb{E}[\sup_{f\in\mathcal{F}}\{R(f) - R_n(f)\}] \ ,$$

souvent appelé la *complexité statistique* du modèle $\mathcal{F}$.

1.6 Lemme de symétrisation

Nous allons procéder en deux temps pour cela en établissant d'abord un résultat général appelé lemme de symétrisation, qui nous permettra ensuite d'utiliser la théorie de Vapnik-Chervonenkis pour arriver à nos fins.

Dans toute la suite de ce chapitre, on va noter $\epsilon_1, \dots, \epsilon_n$ des variables aléatoires i.i.d. de Rademacher (i.e. uniformes sur $\{-1,1\}$, indépendantes de $\mathcal{D}_n$). Le lemme de symétrisation énonce que la complexité statistique est majorée par la complexité de Rademacher de la classe $\mathcal{F}$.

Définition 6. *La complexité de Rademacher de la classe $\mathcal{F}$ est définie par*

$$\mathrm{Rad}(\mathcal{F}) = \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i \ell_f(X_i, Y_i) \right] \ .$$

Théorème 7 (Symétrisation)**.** *On a*

$$\mathbb{E}[\sup_{f\in\mathcal{F}}\{R(f) - R_n(f)\}] \leqslant 2\,\mathrm{Rad}(\mathcal{F}) \ .$$

Démonstration. Soit $Z'_1, \dots, Z'_n$, $Z'_i = (X'_i, Y'_i)$ des données indépendantes de $\mathcal{D}_n$ et de $\epsilon_1, \dots, \epsilon_n$, de même loi que $\mathcal{D}_n$, de sorte que, pour tout $f \in \mathcal{F}$,

$$R(f) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \ell_f(X_i, Y_i)\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \ell_f(X'_i, Y'_i)\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \ell_f(X'_i, Y'_i)|\mathcal{D}_n\right] \ .$$

Comme on a aussi

$$R_n(f) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \ell_f(X_i, Y_i)|\mathcal{D}_n\right] \ ,$$

on en déduit

$$\mathbb{E}[\sup_{f\in\mathcal{F}}\{R(f) - R_n(f)\}] = \mathbb{E}\left[\sup_{f\in\mathcal{F}} \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\{\ell_f(X_i', Y_i') - \ell_f(X_i, Y_i)\}|\mathcal{D}_n\right]\right] \ .$$

Ensuite, on a, pour tout $f \in \mathcal{F}$,

$$\frac{1}{n}\sum_{i=1}^{n}\{\ell_f(X_i', Y_i') - \ell_f(X_i, Y_i)\} \leqslant \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\{\ell_f(X_i', Y_i') - \ell_f(X_i, Y_i)\} \ ,$$

donc

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\{\ell_f(X_i', Y_i') - \ell_f(X_i, Y_i)\}|\mathcal{D}_n\right] \leqslant \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\{\ell_f(X_i', Y_i') - \ell_f(X_i, Y_i)\}|\mathcal{D}_n\right] \ .$$

Ainsi

$$\sup_{f\in\mathcal{F}} \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\{\ell_f(X_i', Y_i') - \ell_f(X_i, Y_i)\}|\mathcal{D}_n\right] \leqslant \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\{\ell_f(X_i', Y_i') - \ell_f(X_i, Y_i)\}|\mathcal{D}_n\right] \ .$$

Finalement,

$$\mathbb{E}[\sup_{f\in\mathcal{F}}\{R(f) - R_n(f)\}] \leqslant \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\{\ell_f(X_i', Y_i') - \ell_f(X_i, Y_i)\}\right] \ .$$

On a ensuite, par symétrie de $\ell_f(X_i', Y_i') - \ell_f(X_i, Y_i)$, que ces variables aléatoires ont même loi que $\epsilon_i(\ell_f(X_i', Y_i') - \ell_f(X_i, Y_i))$ (le vérifier en calculant leur transformée de Fourier par exemple). Par indépendance, on en déduit alors que

$$\begin{aligned}\mathbb{E}[\sup_{f\in\mathcal{F}}\{R(f) - R_n(f)\}] &\leqslant \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\epsilon_i\{\ell_f(X_i', Y_i') - \ell_f(X_i, Y_i)\}\right] \\ &\leqslant \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\epsilon_i\ell_f(X_i', Y_i')\right] + \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}(-\epsilon_i)\ell_f(X_i, Y_i)\right] \ .\end{aligned}$$

On conclut en vérifiant que les deux termes dans cette dernière majoration sont tous deux égaux à $\mathrm{Rad}(\mathcal{F})$.

□

Les théorèmes 5 et 7 permettent d'assurer que, avec une probabilité supérieure à $1-\delta$,

$$R(\hat{f}_{\mathcal{F}}) - R(f^*_{\mathcal{F}}) \leqslant 2\,\mathrm{Rad}(\mathcal{F}) + \sqrt{\frac{2\log(1/\delta)}{n}} \ . \tag{1.3}$$

La majoration de la complexité statistique par la complexité de Rademacher est là encore valable pour toute fonction de perte ℓ comme vous pourrez le vérifier. Le but de la fin du chapitre est de décrire comment majorer $\mathrm{Rad}(\mathcal{F})$ en utilisant la théorie de Vapnik et Chervonenkis qui elle, est propre aux fonctions booléennes, donc à la perte $\{0,1\}$.

1.7 Théorie de Vapnik et Chervonenkis

Soit $\mathcal{G}$ une collection de fonctions booléennes, définies sur un ensemble $\mathcal{Z}$, c'est à dire un ensemble de fonctions $g : \mathcal{Z} \to \{0,1\}$.

1.7.1 Dimension de Vapnik-Chervonenkis (VC)

Définition 8. *On dit qu'un ensemble fini de points $z_1, \ldots, z_k$ de $\mathcal{Z}$ est éclaté par $\mathcal{G}$ si toutes les fonctions $\{z_1, \ldots, z_k\} \to \{0,1\}$ peuvent être obtenues par restriction de fonctions de $\mathcal{G}$, i.e. si*

$$|\{(g(z_1), \ldots, g(z_k)), g \in \mathcal{G}\}| = 2^k \ .$$

Exemple 9. *Sur la droite réelle $\mathbb{R}$, tout ensemble de deux points distincts est éclaté par l'ensemble $\mathcal{G}$ des indicatrices de demi-droites. Il n'existe pas d'exemples de 3 points distincts qui soit éclaté par cet ensemble $\mathcal{G}$.*

Dans le plan Euclidien $\mathbb{R}^2$, tout ensemble de 3 points non alignés est éclaté par la collection $\mathcal{G}$ des indicatrices de demi-plans. Il n'existe pas d'ensemble de 4 points du plan qui soit éclaté par cette même collection $\mathcal{G}$.

Plus généralement, on peut montrer que, dans l'espace Euclidien $\mathbb{R}^d$, il existe un sous-ensemble de $d+1$ points éclaté par l'ensemble $\mathcal{G}$ des indicatrices de demi-espaces, mais qu'il n'existe pas de sous-ensembles de $d+2$ points éclaté par cette collection $\mathcal{G}$.

Définition 10 (Dimension de Vapnik)**.** *La dimension de Vapnik $VC(\mathcal{G})$ de l'ensemble $\mathcal{G}$ est le plus grand entier k tel qu'il existe un sous ensemble de k points de $\mathcal{Z}$ éclaté par $\mathcal{G}$ mais tout sous-ensemble de $k+1$ points de $\mathcal{Z}$ n'est pas éclaté par $\mathcal{G}$.*

Ainsi, dans l'espace Euclidien $\mathbb{R}^k$ la dimension de Vapnik de l'ensemble des indicatrices de demi-espaces est $k+1$.

1.7.2 Le lemme de Pajor

Comme $\mathcal{G}$ éclate un ensemble $\Lambda \subset \mathcal{Z}$ de cardinal $\mathrm{VC}(\mathcal{G})$, on a toujours $|\mathcal{G}| \geqslant 2^{\mathrm{VC}(\mathcal{G})}$.

Le lemme de Pajor majore $|\mathcal{G}|$ par le nombre de sous-ensembles de $\mathcal{Z}$ éclatés par $\mathcal{G}$ lorsque $\mathcal{Z}$ est fini.

Lemma 2 (Lemme de Pajor)**.** *Soit $\mathcal{G}$ un ensemble de fonctions booléennes définies sur un ensemble fini $\mathcal{Z}_0$. Alors,*

$$|\mathcal{G}| \leqslant |\{\Lambda \subset \mathcal{Z}_0 : \Lambda \textit{ est éclaté par } \mathcal{G}\}| \ .$$

Démonstration. On procède par récurrence sur le cardinal de $\mathcal{Z}_0$. Le résultat est trivial lorsque $|\mathcal{Z}_0| = 0$ (l'ensemble vide étant toujours éclaté par $\mathcal{G}$), on suppose que le lemme de Pajor est vrai pour tout ensemble de cardinal n et on considère un sous-ensemble $\mathcal{Z}_0$ de cardinal $n+1$. Soit $z_0 \in \mathcal{Z}_0$ et

$$\mathcal{G}_0 = \{g \in \mathcal{G} : g(z_0) = 0\}, \qquad \mathcal{G}_1 = \{g \in \mathcal{G} : g(z_0) = 1\} \ .$$

Ainsi, on a clairement $|\mathcal{G}| = |\mathcal{G}_0| + |\mathcal{G}_1|$.

Définissons $\mathcal{S} = \{\Lambda \subset \mathcal{Z}_0 : \Lambda \text{ est éclaté par } \mathcal{G}\}$. Notre hypothèse de récurrence implique immédiatement que, pour tout $i \in \{0,1\}$,

$$|\mathcal{G}_i| \leqslant |\mathcal{S}_i|, \qquad \text{où} \qquad \mathcal{S}_i = \{\Lambda \subset \mathcal{Z}_0 \setminus \{z_0\} : \Lambda \text{ est éclaté par } \mathcal{G}_i\} \ .$$

On a $\mathcal{S}_0 \cup \mathcal{S}_1 \subset \mathcal{S}$ et

$$|\mathcal{S}_0| + |\mathcal{S}_1| = |\mathcal{S}_0 \cup \mathcal{S}_1| + |\mathcal{S}_0 \cap \mathcal{S}_1| \ .$$

On construit alors deux injections, l'une de $\mathcal{S}_0 \cup \mathcal{S}_1$ dans $\mathcal{S}$ et l'autre de $\mathcal{S}_0 \cap \mathcal{S}_1$ dans $\mathcal{S}$, d'images disjointes, de sorte qu'on aura

$$|\mathcal{G}| = |\mathcal{G}_0| + |\mathcal{G}_1| \leqslant |\mathcal{S}_0| + |\mathcal{S}_1| = |\mathcal{S}_0 \cup \mathcal{S}_1| + |\mathcal{S}_0 \cap \mathcal{S}_1| \leqslant |\mathcal{S}| \ .$$

On va noter $\varphi_\cup$ la première injection et $\varphi_\cap$ la seconde.

1. Si $\Lambda \in \mathcal{S}_0 \setminus \mathcal{S}_1$ ou $\Lambda \in \mathcal{S}_1 \setminus \mathcal{S}_0$, on note $\varphi_\cup(\Lambda) = \Lambda$.
2. Si $\Lambda \in \mathcal{S}_0 \cap \mathcal{S}_1$, alors, pour tout $g : \Lambda \to \{0,1\}$, il existe $f_0 \in \mathcal{S}_0$ et $f_1 \in \mathcal{S}_1$ telles que g est la restriction simultanément de f_0 et f_1. Ainsi, les deux ensembles Λ et $\Lambda \cup \{x_0\}$ sont éclatés par $\mathcal{G}$, et donc appartiennent à $\mathcal{S}$. On peut donc définir $\varphi_\cup(\Lambda) = \Lambda$ et $\varphi_\cap(\Lambda) = \Lambda \cup \{x_0\}$.

Par construction, les deux injections $\varphi_\cup$ et $\varphi_\cap$ sont donc d'images disjointes et donc, par hypothèse de récurrence, comme annoncé

$$|\mathcal{S}| \geqslant |\mathcal{S}_0 \cup \mathcal{S}_1| + |\mathcal{S}_0 \cap \mathcal{S}_1| = |\mathcal{S}_0| + |\mathcal{S}_1| \geqslant |\mathcal{G}_0| + |\mathcal{G}_1| = |\mathcal{G}| \ .$$

□

1.7.3 Lemme de Sauer-Shelah

Un corollaire important du lemme de Pajor est le résultat suivant, connu sous le nom de Lemme de Sauer-Shelah, qui fournit une borne sur la croissance des ensembles de fonctions Booléennes définies sur un ensemble fini, en fonction de la dimension de Vapnik.

Lemma 3 (Lemme de Sauer-Shelah)**.** *Soit $\mathcal{G}$ un ensemble de fonctions Booléennes définies sur un ensemble $\mathcal{Z}_0$ de n-points, tel que $VC(\mathcal{G}) = d$. Alors,*

$$|\mathcal{G}| \leqslant \sum_{k=0}^{d} \binom{n}{k} \leqslant \left(\frac{en}{d}\right)^d \ .$$

Le lemme de Sauer montre un résultat intéressant de combinatoire : Etant donné un ensemble $\mathcal{F}$ de fonctions Booléennes définies sur un ensemble $\mathcal{X}$, alors le cardinal de l'ensemble des vecteurs de l'hypercube $\{(f(x_1), \ldots, f(x_n)) \in \{0,1\}^n\}$ vaut 2^n ou est au plus polynomial $\leqslant n^d$.

Démonstration. Par le lemme de Pajor

$$\begin{aligned} |\mathcal{G}| &\leqslant |\{\Lambda \subset \mathcal{Z}_0 : \Lambda \text{ est éclaté par } \mathcal{G}\}| \\ &= \sum_{k=0}^{n} |\{\Lambda \subset \mathcal{Z}_0 : \Lambda \text{ est éclaté par } \mathcal{G} \text{ et } |\Lambda| = k\}| \ . \end{aligned}$$

Par définition de la dimension VC, les cardinaux de cette dernière borne sont nuls pour tous les $k > d$, ainsi,

$$\begin{aligned} |\mathcal{G}| &= \sum_{k=0}^{d} |\{\Lambda \subset \mathcal{Z}_0 : \Lambda \text{ est éclaté par } \mathcal{G} \text{ et } |\Lambda| = k\}| \\ &\leqslant \sum_{k=0}^{d} |\{\Lambda \subset \mathcal{Z}_0 : |\Lambda| = k\}| \\ &= \sum_{k=0}^{d} \binom{n}{k} \ . \end{aligned}$$

La seconde inégalité est très classique : Elle est triviale si $d > n$ et si $d/n \leqslant 1$, on a

$$\sum_{k=0}^{d} \binom{n}{k} \left(\frac{d}{n}\right)^d \leqslant \sum_{k=0}^{d} \binom{n}{k} \left(\frac{d}{n}\right)^k \leqslant \sum_{k=0}^{n} \binom{n}{k} \left(\frac{d}{n}\right)^k = \left(1 + \frac{d}{n}\right)^n \leqslant \mathrm{e}^d \ .$$

□

1.8 Majoration de la complexité statistique

Le but de cette section est d'expliquer comment majorer la complexité de Rademacher à partir de la dimension de Vapnik. Dans toute cette section, on note $\mathcal{G} = \{\ell_f(\cdot), f \in \mathcal{F}\}$, ensemble de fonctions Booléennes définies sur $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, et

$$\mathrm{VC}(\mathcal{F}) = \mathrm{VC}(\mathcal{G}) \ . \tag{1.4}$$

On suppose dans la suite cette quantité finie.

Le but est de majorer

$$\mathrm{Rad}(\mathcal{F}) = \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \ell_f(X_i, Y_i)\right] \ .$$

On veut appliquer le lemme de Sauer-Shelah et, pour se ramener à des fonctions Booléennes définies sur un ensemble fini, on procède d'abord brutalement en majorant

$$\mathrm{Rad}(\mathcal{F}) \leqslant \sup_{z_1,\dots,z_n\in\mathcal{Z}} \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \ell_f(z_i)\right] \ .$$

Fixons alors $z_1, \dots, z_n$ n points de $\mathcal{Z}$. L'ensemble $\mathcal{G}_n$ des restrictions des fonctions de $\mathcal{G}$ à $z_1, \dots, z_n$ est alors clairement de dimension de Vapnik inférieure à $\mathrm{VC}(\mathcal{F})$ donc, d'après le lemme de Sauer-Shelah, on a

$$|\mathcal{G}_n| \leqslant \left(\frac{en}{\mathrm{VC}(\mathcal{F})}\right)^{\mathrm{VC}(\mathcal{F})} \ . \tag{1.5}$$

Pour conclure à la majoration de la complexité de Rademacher, on utilise alors le lemme suivant attribué à Pisier et Massart.

Lemma 4 (Lemme de Pisier-Massart)**.** *Soient $X_1, \dots, X_N$ des variables aléatoires telles que*

$$\forall i \in \{1, \dots, N\}, \forall s > 0, \qquad \mathbb{E}[\exp(sX_i)] \leqslant \exp(s^2K^2) \ .$$

Alors,

$$\mathbb{E}[\max_{i\in\{1,\dots,N\}} X_i] \leqslant 2K\sqrt{\log(N)} \ .$$

Démonstration. On fixe $s > 0$ et on utilise les majorations suivantes

$$\begin{aligned}
\mathbb{E}[\max_{i\in\{1,\dots,N} X_i] &= \frac{1}{s}\mathbb{E}[\max_{i\in\{1,\dots,N\}} \log\exp(sX_i)] \\
&= \frac{1}{s}\mathbb{E}[\log \max_{i\in\{1,\dots,N\}} \exp(sX_i)] \quad \text{croissance de log} \\
&\leqslant \frac{1}{s}\log\mathbb{E}[\max_{i\in\{1,\dots,N\}} \exp(sX_i)] \quad \text{Jensen} \\
&\leqslant \frac{1}{s}\log\mathbb{E}[\sum_{i\in\{1,\dots,N\}} \exp(sX_i)] \quad \exp(sX_i) \geqslant 0 \\
&\leqslant \frac{1}{s}\log(N\exp(s^2K^2)) \quad \text{par hypothèse} \\
&= \frac{\log N}{s} + sK^2 \ .
\end{aligned}$$

On conclut en appliquant ce résultat avec $s = \sqrt{\log N}/K$. □

Le but est d'appliquer le lemme de Pisier-Massart aux variables aléatoires $X_g = n^{-1}\sum_{i=1}^n \epsilon_i g(z_i)$, pour tous les $g \in \mathcal{G}_n$. On doit vérifier l'hypothèse sur ces variables aléatoires. Or, les variables aléatoires $\epsilon_i g(z_i)/n \in [-1/n, 1/n]$ et sont centrées donc, d'après la première conclusion du Théorème 4,

$$\forall g \in \mathcal{G}_n, \forall s > 0, \qquad \mathbb{E}[\exp(sX_g)] \leqslant \exp\left(\frac{s^2\sum_{i=1}^n (2/n)^2}{8}\right) = \exp\left(\frac{s^2}{2n}\right) \ .$$

Il vient donc du Lemme de Pisier-Massart que

$$\forall z_1,\dots,z_n \in \mathcal{Z}, \qquad \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i \ell_f(z_i)\right] = \mathbb{E}\left[\max_{g\in\mathcal{G}_n} X_g\right] \leqslant \sqrt{\frac{2\log(|\mathcal{G}_n|)}{n}} \ .$$

En injectant la borne de Sauer-Shelah (1.5) dans ce résultat, on en déduit

$$\mathrm{Rad}(\mathcal{F}) \leqslant \sqrt{\frac{2\,\mathrm{VC}(\mathcal{F})}{n}\log\left(\frac{en}{\mathrm{VC}(\mathcal{F})}\right)} \tag{1.6}$$

Finalement, on peut injecter ce résultat dans la borne d'excès de risque du minimiseur du risque empirique (1.3) pour obtenir le résultat suivant.

Théorème 11. *Soit $\hat{f}_{\mathcal{F}}$ un minimiseur du risque empirique sur un ensemble de classifieurs $\mathcal{F}$ de dimension de Vapnik $\mathrm{VC}(\mathcal{F})$ finie. Alors, on a, pour tout $\delta \in (0,1)$,*

$$\mathbb{P}\left(R(\hat{f}_{\mathcal{F}}) - R(f^*_{\mathcal{F}}) \leqslant 2\sqrt{\frac{2\,\mathrm{VC}(\mathcal{F})}{n}\log\left(\frac{en}{\mathrm{VC}(\mathcal{F})}\right)} + \sqrt{\frac{2\log(1/\delta)}{n}}\right) \geqslant 1-\delta \ .$$

Remark 5. *On peut montrer qu'il est impossible en général* pour n'importe quel estimateur *de converger à une vitesse meilleure que* $\sqrt{\mathrm{VC}(\mathcal{F})/n}$.

Une première conséquence est donc que l'hypothèse $\mathrm{VC}(\mathcal{F})$ *finie ne peut être évitée en général, puisqu'aucun estimateur ne pourrait converger sinon.*

Une seconde conséquence est que les minimiseurs du risque empirique sont "optimaux" au sens où, en l'absence d'hypothèses sur P, *aucun estimateur ne peut converger plus rapidement qu'eux, éventuellement à un log près.*

Attention toutefois à cette dernière conclusion, puisqu'on peut démontrer de meilleures vitesses sous des hypothèses sur P.

1.9 Meilleures vitesses

Pour illustrer la dernière remarque, nous allons étudier le problème jouet suivant et illustrer comment on peut espérer la vitesse $1/\sqrt{n}$ sous de bonnes hypothèses sur P.

On suppose dans toute cette section que $\mathcal{F} = \{f_1, \dots, f_M\}$ est un ensemble fini de classifieurs *et* que $f^* \in \mathcal{F}$. Rappelons que, pour tout $f \in \mathcal{F}$, on a établi dans le Théorème 1 que l'excès de risque de tout classifieur $f \in \mathcal{F}$ vaut alors

$$\mathcal{E}(f) = R(f) - R(f^*) = \mathbb{E}[|1 - 2\eta(X)||f(X) - f^*(X)|] \ .$$

D'autre part, le Théorème 5 nous garantit que le minimiseur du risque empirique sur le modèle $\mathcal{F}$ vérifie

$$R(\hat{f}_{\mathcal{F}}) - R(f^*_{\mathcal{F}}) \leqslant R(\hat{f}_{\mathcal{F}}) - R_n(\hat{f}_{\mathcal{F}}) + R_n(f^*_{\mathcal{F}}) - R(f^*_{\mathcal{F}}) \ .$$

1.9.1 Inégalité de Bernstein

Soient $X_1, \dots, X_n$ des variables aléatoires indépendantes, X_i étant à valeurs dans $[a_i, b_i]$. Le Théorème 4 montre que, dans ce cas, avec probabilité $1 - \delta$, on a

$$\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) \leqslant K\sqrt{\frac{2\log(1/\delta)}{n}}, \qquad K^2 = \frac{1}{4}\sum_{i=1}^{n}(b_i - a_i)^2 \ .$$

Cette inégalité peut s'avérer très imprécise pour les niveaux de confiance δ assez grands, car l'inégalité de Chebyshev assure que, avec la même probabilité,

$$\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) \leqslant \sigma\sqrt{\frac{1}{n\delta}}, \qquad \sigma^2 = \sum_{i=1}^{n}\mathrm{Var}(X_i) \ .$$

Comme $X_i \in [a_i, b_i]$, on a toujours $\mathrm{Var}(X_i) \leqslant (b_i - a_i)^2/4$, mais cette majoration peut être pessimiste. Ainsi, si $X_i \sim \mathcal{B}(p)$ est une variable aléatoire

de Bernoulli de paramètre p, on a $\mathrm{Var}(X_i) = p(1-p) \leqslant 1/4$, cette inégalité étant très imprécise si p est proche de 0. L'inégalité de Bernstein va donner un résultat plus précis que ces deux inégalités pour les "grandes" valeurs de δ.

Commençons par un lemme donnant un contrôle de la transformée de Laplace de X_i faisant intervenir sa variance.

Lemma 6. *Soit X une variable aléatoire, de variance σ^2, à valeurs dans $[a,b]$. Alors, pour tout $s \in (0, 3/(b-a))$, on a*

$$\mathbb{E}[\exp(s(X - \mathbb{E}[X]))] \leqslant \exp\left(\frac{s^2\sigma^2}{2(1 - s(b-a)/3)}\right) .$$

Démonstration. La preuve reprend essentiellement les mêmes étapes que la première preuve du lemme d'Hoeffding. On suppose sans perte de généralité que X est centrée et on a donc $|X| \leqslant (b-a)$ p.s. On en déduit

$$\begin{aligned}
\mathbb{E}[\exp(sX)] &= 1 + \sum_{k \geqslant 2} \frac{s^k \mathbb{E}[X^k]}{k!} \quad \text{dvpt en série entière} \\
&\leqslant 1 + \sum_{k \geqslant 2} \frac{s^k \sigma^2 (b-a)^{k-2}}{k!} \quad \text{car } X^k \leqslant X^2 (b-a)^{k-2} \\
&\leqslant 1 + \sum_{k \geqslant 2} \frac{s^k \sigma^2 (b-a)^{k-2}}{2 * 3^{k-2}} \quad \text{car } k! \geqslant 2 * 3^{k-2} \\
&= 1 + \frac{s^2\sigma^2}{2(1 - s(b-a)/3)} \quad \text{somme géométrique .}
\end{aligned}$$

On conclut la preuve par l'inégalité $1 + u \leqslant \exp(u)$ valable pour tout $u \in \mathbb{R}$. □

Pour démontrer l'inégalité de Bernstein, on va utiliser la méthode de Chernoff (Markov exponentiel). Soit $K = \max_i (b_i - a_i)$. Pour tout $s \in$

$(0, 1/K)$ et tout $z > 0$, on a

$$\begin{aligned}
\mathbb{P}(\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) > z) &= \mathbb{P}(\exp(s\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i])) > \exp(sz)) \\
&\leqslant \frac{\mathbb{E}[\exp(s\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]))]}{\exp(sz)} \\
&= \frac{\prod_{i=1}^{n}\mathbb{E}[\exp(s(X_i - \mathbb{E}[X_i]))]}{\exp(sz)} \\
&= \exp(-sz)\prod_{i=1}^{n}\exp\left(\frac{s^2\operatorname{Var}(X_i)}{2(1 - s(b_i - a_i)/3)}\right) \\
&\leqslant \exp(-sz)\prod_{i=1}^{n}\exp\left(\frac{s^2\operatorname{Var}(X_i)}{2(1 - sK/3)}\right) \\
&\leqslant \exp\left(-sz + \frac{s^2\sum_{i=1}^{n}\operatorname{Var}(X_i)}{2(1 - sK/3)}\right) .
\end{aligned}$$

On utilise alors le lemme d'analyse 7 pour conclure la preuve de l'inégalité de Bernstein donnée au Théorème 12.

Lemma 7. *Soit $v > 0$, $c > 0$, $z > 0$, alors*

$$\sup_{s\in(0,1/c)}\left\{sz - \frac{s^2 v}{2(1 - cs)}\right\} = \frac{v}{c^2}h\left(\frac{cz}{v}\right) ,$$

où la fonction h est définie par

$$h(x) = 1 + x - \sqrt{1 + 2x} = \frac{1}{2}(\sqrt{1 + 2x} - 1)^2 .$$

Démonstration. Notons

$$\psi(s) = sz - \frac{s^2 v}{2(1 - cs)} .$$

On vérifie successivement que

$$\begin{aligned}
\psi'(s) &= z - \frac{v}{2}\frac{2s - s^2 c}{(1 - cs)^2} \\
&= \frac{z - 2s(cz + v/2) + s^2(c^2 z + vc/2)}{(1 - cs)^2} \\
&= \frac{c(cz + v/2)\big[(s - 1/c)^2 - 1/c^2 + z/[c(cz + v/2)]\big]}{(1 - cs)^2} .
\end{aligned}$$

Il s'en suit que ψ atteint son maximum en

$$s^* = \frac{1}{c}\left(1 - \sqrt{1 - \frac{z}{(z + v/2c)}}\right) = \frac{1}{c}(1 - (1 + 2x)^{-1/2}) ,$$

où on a posé $x = cz/v$. Ce s^* vérifie donc

$$(1 - cs^*)^{-1} = \sqrt{1+2x}, \qquad s^*(1-cs^*)^{-1} = \frac{1}{c}(\sqrt{1+2x} - 1) \ .$$

Il s'en suit que

$$\begin{aligned}
\psi(s^*) &= s^*\bigg(z - \frac{v}{2c}(\sqrt{1+2x} - 1)\bigg) \\
&= \frac{v}{2c} s^*\big(2x + 1 - \sqrt{1+2x}\big) \\
&= \frac{v}{2c^2}(1 - (1+2x)^{-1/2})((1+2x) - \sqrt{1+2x}) \\
&= \frac{v}{c^2} h(x) \ .
\end{aligned}$$

□

La première conclusion de l'inégalité de Bernstein donnée au Théorème 12 s'en déduit directement. Pour la seconde partie, on vérifie que

$$h^{-1}(x) = \frac{1}{2}\big[(1+\sqrt{2x})^2 - 1\big] = \sqrt{2x} + x \ .$$

et donc, si

$$u = \frac{9\sigma^2}{K^2} h\bigg(\frac{Kz}{3\sigma^2}\bigg), \qquad z = \frac{3\sigma^2}{K} h^{-1}\bigg(\frac{K^2 u}{9\sigma^2}\bigg) = \sigma\sqrt{2u} + \frac{Ku}{3} \ .$$

Théorème 12. *Soient $X_1, \dots, X_n$ des variables aléatoires, X_i étant à valeurs dans $[a_i, b_i]$ p.s. On note*

$$\sigma^2 = \sum_{i=1}^n \mathrm{Var}(X_i), \qquad K = \max_i (b_i - a_i) \ .$$

On a pour tout $z > 0$,

$$\mathbb{P}\Big(\sum_{i=1}^n (X_i - \mathbb{E}[X_i]) > z\Big) \leqslant \exp\bigg(-\frac{9\sigma^2}{K^2} h\bigg(\frac{Kz}{3\sigma^2}\bigg)\bigg) \ .$$

De manière équivalente, on a aussi,pour tout $z > 0$,

$$\mathbb{P}\bigg(\sum_{i=1}^n (X_i - \mathbb{E}[X_i]) > \sigma\sqrt{2z} + \frac{Kz}{3}\bigg) \leqslant \exp\big(-z\big) \ .$$

Pour terminer cette section, on va appliquer l'inégalité de Bernstein pour contrôler l'excès de risque du minimiseur du risque empirique. Rappelons qu'on avait démontré

$$\begin{aligned}
R(\hat{f}_{\mathcal{F}}) - R(f^*_{\mathcal{F}}) &\leqslant R(\hat{f}_{\mathcal{F}}) - R_n(\hat{f}_{\mathcal{F}}) + R_n(f^*_{\mathcal{F}}) - R(f^*_{\mathcal{F}}) \\
&= \sum_{i=1}^n \frac{\ell_{f^*}(Z_i) - \ell_{\hat{f}}(Z_i) - \mathbb{E}[\ell_{f^*}(Z_i) - \ell_{\hat{f}}(Z_i)]}{n} \ .
\end{aligned}$$

Pour tout $f \in \mathcal{F}$, les variables aléatoires $X_i = \frac{\ell_{f^*}(Z_i)-\ell_f(Z_i)}{n}$ sont indépendantes et à valeurs dans $[-1/n, 1/n]$ p.s., donc, d'après l'inégalité de Bernstein, on a, pour tout $f \in \mathcal{F}$ et $z > 0$,

$$\mathbb{P}\bigg(\sum_{i=1}^n \frac{\ell_{f^*}(Z_i) - \ell_f(Z_i) - \mathbb{E}[\ell_{f^*}(Z_i) - \ell_f(Z_i)]}{n} > \sigma_f \sqrt{\frac{2z}{n}} + \frac{2z}{n} \bigg) \leqslant \exp(-z) \ ,$$

où $\sigma_f^2 = n^{-1} \sum_{i=1}^n \mathrm{Var}(\ell_{f^*}(Z) - \ell_f(Z)) = \mathrm{Var}(\ell_{f^*}(Z) - \ell_f(Z))$.

Par une borne d'union, on en déduit donc que, pour tout $z > 0$,

$$\mathbb{P}\bigg(\exists f \in \mathcal{F} : \sum_{i=1}^n \frac{\ell_{f^*}(Z_i) - \ell_f(Z_i) - \mathbb{E}[\ell_{f^*}(Z_i) - \ell_f(Z_i)]}{n} > \sigma_f \sqrt{\frac{2z}{n}} + \frac{2z}{n} \bigg) \leqslant |\mathcal{F}| \exp(-z) \ ,$$

ce qui implique en particulier que, pour tout $z > 0$, avec probabilité $1 - \exp(-z)$,

$$R(\hat{f}) - R(f^*) \leqslant \sigma_{\hat{f}} \sqrt{\frac{2(\log(|\mathcal{F}|) + z)}{n}} + \frac{2(\log(|\mathcal{F}|) + z)}{n} \ . \tag{1.7}$$

1.9.2 Condition de Bernstein

On va pouvoir déduire de la borne précédente une meilleure borne de risque pour $\hat{f}$ si on peut relier $\sigma_f = \sqrt{\mathrm{Var}(\ell_{f^*}(Z) - \ell_f(Z))}$ à $\mathcal{E}(f) = R(f) - R(f^*) = \mathbb{E}[\ell_f(Z) - \ell_{f^*}(Z)]$. L'inégalité de Cauchy-Schwarz assure que

$$R(f) - R(f^*) \leqslant \sigma_f \ .$$

S'il s'avère que cette inégalité est précise ou presque, alors on va pouvoir en déduire de bonnes bornes. Plus précisément, on dit que $(P, \mathcal{F}, \ell)$ satisfait l'hypothèse de Bernstein (C, α), où $\alpha \in [0, 1]$, si

$$\sigma_f \leqslant C \mathcal{E}(f)^\alpha \ .$$

Sous l'hypothèse de Bernstein, on déduit de (1.7) que, pour tout $z > 0$, avec probabilité $1 - \exp(-z)$,

$$\mathcal{E}(\hat{f}) \leqslant C \mathcal{E}(\hat{f})^\alpha \sqrt{r_n(z)} + r_n(z), \qquad r_n(z) = \frac{2(\log(|\mathcal{F}|) + z)}{n} \ .$$

On applique alors l'inégalité de Minkowsky $ab \leqslant p^{-1}a^p + q^{-1}b^q$ avec $p = 1/\alpha$, $q = 1/(1-\alpha)$ pour en déduire que

$$\mathcal{E}(\hat{f}) \leqslant \frac{1}{2}\mathcal{E}(\hat{f}) + (1-\alpha)[(2\alpha)^\alpha C \sqrt{r_n(z)}]^{1/(1-\alpha)} + r_n(z) \ .$$

De manière équivalente, on a donc

$$\mathcal{E}(\hat{f}) \leqslant 2(1-\alpha)[(2\alpha)^\alpha C \sqrt{r_n(z)}]^{1/(1-\alpha)} + 2r_n(z) \ . \tag{1.8}$$

Ainsi, on voit que le minimiseur du risque empirique converge à la vitesse $n^{-1 \wedge 1/(2(1-\alpha))}$ qui est toujours meilleure que $1/\sqrt{n}$, si $(P, \mathcal{F}, \ell)$ satisfait l'hypothèse de Bernstein (C, α). On va maintenant s'intéresser aux hypothèses sur P permettant de vérifier cette hypothèse.

1.9.3 Hypothèses de marge

Il s'agit de majorer la variance par l'excès de risque. On a d'une part

$$\begin{aligned}\sigma_f &= \sqrt{\mathrm{Var}(\ell_{f^*}(Z) - \ell_f(Z))} \leqslant \sqrt{\mathbb{E}[(\ell_{f^*}(Z) - \ell_f(Z))^2]} \\ &= \sqrt{\mathbb{E}[|\ell_{f^*}(Z) - \ell_f(Z)|]} = \sqrt{\mathbb{E}[|f(X) - f^*(X)|]} \ ,\end{aligned}$$

et d'autre part

$$\mathcal{E}(f) = R(f) - R(f^*) = \mathbb{E}[|1 - 2\eta(X)||f(X) - f^*(X)|] \ .$$

Il n'est pas surprenant que l'hypothèse de Bernstein puisse être vérifiée sous des conditions de marge, sur la loi P, i.e. sous des hypothèses sur la probabilité que $\eta(X)$ soit proche de $1/2$. Notons qu'il est raisonnable d'espérer pouvoir construire de meilleurs classifieurs sous de telles hypothèses : si $\eta(X)$ a peu de chance d'être proche de $1/2$, alors X apporte une information plus pertinente sur l'étiquette Y, donc on doit pouvoir utiliser cette information pour proposer un meilleur choix de Y.

Définition 13 (Marge dure)**.** *On dit que P vérifie la condition de marge dure s'il existe $h \in [0, 1/2]$ tel que*

$$|\eta(X) - 1/2| > h, \qquad p.s. \ .$$

Sous l'hypothèse de marge dure, on a

$$\mathcal{E}(f) = \mathbb{E}[|1 - 2\eta(X)||f(X) - f^*(X)|] \geqslant 2h\mathbb{E}[|f(X) - f^*(X)|] \geqslant 2h\sigma_f^2 \ .$$

On en déduit que $(P, \mathcal{F}, \ell)$ satisfait l'hypothèse de Bernstein $((2h)^{-1/2}, 1/2)$, donc que

$$\mathcal{E}(\hat{f}) \leqslant \left(\frac{1}{2h} + 2\right) r_n(z) \ .$$

Il existe une version plus douce de l'hypothèse de marge.

Définition 14 (Marge douce)**.** *On dit que P vérifie la condition de marge douce s'il existe $C > 0$ et β tel que*

$$\forall h \in [0, 1/2], \qquad \mathbb{P}(|\eta(X) - 1/2| \leqslant h) \leqslant Ch^\beta \ .$$

Si P vérifie l'hypothèse de marge douce, on a

$$\begin{aligned}\mathcal{E}(f) &= \mathbb{E}[|1 - 2\eta(X)||f(X) - f^*(X)|] \\ &\geqslant \mathbb{E}[|1 - 2\eta(X)|\mathbf{1}_{\{|\eta(X)-1/2|>h\}}|f(X) - f^*(X)|] \\ &\geqslant h(\mathbb{E}[|f(X) - f^*(X)|] - \mathbb{E}[\mathbf{1}_{\{|\eta(X)-1/2|\leqslant h\}}|f(X) - f^*(X)|]) \\ &\geqslant h(\mathbb{E}[|f(X) - f^*(X)|] - \sqrt{\mathbb{P}(|\eta(X) - 1/2| \leqslant h)\mathbb{E}[|f(X) - f^*(X)|]}) \\ &\geqslant h(\mathbb{E}[|f(X) - f^*(X)|] - \sqrt{Ch^\beta\mathbb{E}[|f(X) - f^*(X)|]}) \ .\end{aligned}$$

Cette inégalité étant vraie pour tout h, on peut le choisir tel que

$$\mathbb{E}[|f(X)-f^*(X)|] = 2\sqrt{Ch^\beta \mathbb{E}[|f(X)-f^*(X)|]} \quad \Leftrightarrow h = \left(\frac{\mathbb{E}[|f(X)-f^*(X)|]}{4C}\right)^{1/\beta} .$$

On a alors

$$\begin{aligned}\mathcal{E}(f) &\geqslant \frac{1}{2}h\mathbb{E}[|f(X)-f^*(X)|] \\ &\geqslant \frac{(\mathbb{E}[|f(X)-f^*(X)|])^{1+1/\beta}}{2(4C)^{1/\beta}} \\ &\geqslant \frac{\sigma_f^{2(1+1/\beta)}}{2(4C)^{1/\beta}} .\end{aligned}$$

Ainsi, l'hypothèse de Bernstein est satisfaite avec

$$C = (2(4C)^{1/\beta})^{\beta/(2(1+\beta))}, \qquad \alpha = \frac{\beta}{2(\beta+1)} .$$

En injectant cette information dans la borne générique d'excès de risque sous condition de Bernstein (1.8), on déduit que le minimiseur du risque empirique converge, sous hypothèse de marge douce à la vitesse $n^{(\beta+1)/(\beta+2)}$, vitesse qui interpole entre la vitesse $1/\sqrt{n}$ qu'on retrouve pour $\beta = 0$ (i.e. sans hypothèse de marge) et la vitesse rapide $1/n$ qu'on retrouve pour $\beta \to \infty$ (i.e. sous hypothèse de marge dure).

Chapitre 2

Apprentissage avec pertes convexes

Le but de ce chapitre est de se rapprocher d'un cadre plus pratique. Le gros défaut de l'approche avec perte $0-1$ vue au chapitre précédent est que les minimiseurs de risque empirique

$$\hat{f} \in \operatorname{argmin}_{f \in \mathcal{F}} R_n(f) \ ,$$

sont solutions d'un problème NP-hard en général, et donc difficilement calculable, ou même approchable en général. L'idée de ce chapitre est de remplacer la perte $0-1$ par une perte convexe de manière à définir un minimiseur de risque empirique solution d'un problème convexe, donc beaucoup plus facile à approcher en général. Nous en profitons pour présenter la méthode de localisation qui est utile pour déduire des vitesses rapides.

2.1 Régression logistique

Dans tout ce chapitre, les données $Z, Z_1, \dots, Z_n$ sont des variables aléatoires i.i.d. de même loi P. Chaque $Z_i = (X_i, Y_i)$ est un couple, X_i est un vecteur à valeurs dans $\mathcal{X} = \mathbb{R}^d$ et $Y \in \mathcal{Y} = \{-1, 1\}$ décrit la classe de l'objet X_i. Pour fixer les idées, on va supposer pendant tout le chapitre que X est un vecteur Gaussien standard de $\mathbb{R}^d$. On note $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. L'ensemble mesurable $\Theta = \mathbb{R}^d$ des paramètres $\theta \in \Theta$ indexe les classifieurs linéaires $f_\theta(x) = \mathrm{Sign}(\langle x, \theta \rangle)$. La fonction de perte $\ell : \Theta \times \mathcal{Z} \to \mathbb{R}$, $(\theta, z) \mapsto \ell_\theta(z)$ est la perte logistique qui s'écrit

$$\ell_\theta(x, y) = \varphi(-y \langle \theta, x \rangle), \qquad \varphi(u) = \log(1 + \exp(u)) \ .$$

Elle mesure l'adéquation du paramètre θ à la donnée z, i.e. la qualité de la prédiction de y par $f_\theta(x)$. En effet, si y et $f_\theta(x)$ sont de même signe, la perte

est plus petite car φ est une fonction croissante. Le risque de tout paramètre $\theta \in \Theta$ est mesuré par

$$R(\theta) = P\ell_\theta = \mathbb{E}[\ell_\theta(Z)] \ .$$

Si $\hat{\theta}$ est un paramètre aléatoire dépendant des observations $\mathcal{D}_n = (Z_1, \dots, Z_n)$, alors $R(\hat{\theta}) = P\ell_{\hat{\theta}}$ est une variable aléatoire définie formellement par

$$R(\hat{\theta}) = P\ell_{\hat{\theta}} = \mathbb{E}[\ell_{\hat{\theta}}(Z)|\mathcal{D}_n] \ .$$

Le gros avantage de remplacer la perte $0-1$ par la logistique en pratique est que le problème d'optimisation définissant le minimiseur du risque empirique :

$$\hat{\theta} \in \mathrm{argmin}_{\theta\in\Theta} P_n\ell_\theta, \qquad P_n g = \frac{1}{n}\sum_{i=1}^{n} g(Z_i), \qquad \forall g : \mathcal{Z} \to \mathbb{R} \ .$$

est un problème convexe car $\Theta = \mathbb{R}^d$ est convexe, dont on peut approcher facilement la solution par un algorithme de descente de gradient.

D'un point de vue théorique, ce cadre va permettre de développer la méthode de localisation, qui permet de démontrer des vitesses rapides pour les minimiseurs du risque empirique comme nous le verrons dans la section 2.2.

En revanche, une grosse différence avec la perte $0-1$ est que la logistique n'est pas bornée. En particulier, on ne peut donc pas utiliser a priori l'inégalité de Mc-Diarmid comme on l'a fait au chapitre précédent pour concentrer le supremum du processus empirique. Nous allons donc développer de nouveaux outils de concentration dans ce chapitre pour majorer le risque du minimiseur du risque empirique.

2.2 Localisation

On souhaite tirer partie du fait que l'estimateur du risque empirique

$$\hat{\theta} \in \mathrm{argmin}_{\theta\in\Theta} P_n\ell_\theta \ ,$$

doit être proche que sa cible

$$\theta^* \in \mathrm{argmin}_{\theta\in\Theta} P\ell_\theta \ .$$

Alors, la majoration utilisée au Théorème 5

$$(P - P_n)(\ell_{\hat{\theta}} - \ell_{\theta^*}) \leqslant \sup_{\theta\in\Theta}(P - P_n)(\ell_\theta - \ell_{\theta^*}) \ ,$$

est pessimiste et doit pouvoir être améliorée en une inégalité

$$(P - P_n)(\ell_{\hat{\theta}} - \ell_{\theta^*}) \leqslant \sup_{\theta\in\mathcal{V}(\theta^*)} (P - P_n)(\ell_\theta - \ell_{\theta^*}) \ ,$$

pour un voisinage $\mathcal{V}(\theta^*)$ de θ^* bien choisi.

On va procéder indirectement en montrant qu'il suffit de contrôler le processus empirique $(P-P_n)(\ell_\theta - \ell_{\theta^*})$ uniformément sur un voisinage $\mathcal{V}(\theta^*)$ de θ^* pour en déduire une vitesse de convergence pour $\hat{\theta}$. Pour cela, on fait d'abord la remarque clé suivante :

Remarque clé 1 : *Si, pour tout $\theta \notin \mathcal{V}(\theta^*)$, $P_n\ell_\theta > P_n\ell_{\theta^*}$, alors, comme par définition $P_n\ell_{\hat{\theta}} \leqslant P_n\ell_{\theta^*}$, on a nécessairement $\hat{\theta} \in \mathcal{V}(\theta^*)$.*

Cette première remarque est toujours vraie et implique donc que, si on montre que

$$\inf_{\mathcal{V}(\theta^*)^c} P_n(\ell_\theta - \ell_{\theta^*}) > 0 \ ,$$

alors $\hat{\theta} \in \mathcal{V}(\theta^*)$. Ce n'est pas encore un argument de localisation car il s'agit de contrôler le processus empirique *hors* d'un voisinage. Pour se ramener à un argument de localisation, on va utiliser la convexité.

Remarque clé 2 : *Si, pour tout θ sur la frontière $\partial\mathcal{V}(\theta^*)$ de $\mathcal{V}(\theta^*)$, $P_n\ell_\theta \geqslant P_n\ell_{\theta^*}$, alors, on a pour tout $\theta \notin \mathcal{V}(\theta^*)$, l'existence d'un $t \in (0,1)$ tel que $\theta^* + t(\theta - \theta^*) = t\theta + (1-t)\theta^* \in \partial\mathcal{V}(\theta^*)$, et donc, par stricte convexité de $\theta \mapsto P_n\ell_\theta$,*

$$0 \leqslant P_n\ell_{t\theta+(1-t)\theta^*} - P_n\ell_{\theta^*} < tP_n\ell_\theta + (1-t)P_n\ell_{\theta^*} - P_n\ell_{\theta^*} = tP_n(\ell_\theta - \ell_{\theta^*}) \ .$$

Donc, d'après la première remarque clé, on a $\hat{\theta} \in \mathcal{V}(\theta^*)$.

On a ainsi un vrai argument de localisation :

$$\{P_n(\ell_\theta - \ell_{\theta^*}) \geqslant 0 \quad \forall\theta \in \partial\mathcal{V}(\theta^*)\} \qquad \Longrightarrow \qquad \hat{\theta} \in \mathcal{V}(\theta^*) \ .$$

Pour montrer que $P_n(\ell_\theta - \ell_{\theta^*}) \geqslant 0$ pour tout $\theta \in \partial\mathcal{V}(\theta^*)$, on procède généralement en deux temps en décomposant

$$\begin{aligned} P_n(\ell_\theta - \ell_{\theta^*}) &= P(\ell_\theta - \ell_{\theta^*}) + (P_n - P)(\ell_\theta - \ell_{\theta^*}) \\ &\geqslant \inf_{\theta\in\partial\mathcal{V}(\theta^*)} P(\ell_\theta - \ell_{\theta^*}) - \sup_{\theta\in\mathcal{V}(\theta^*)} (P-P_n)(\ell_\theta - \ell_{\theta^*}) \ . \end{aligned} \tag{2.1}$$

On se ramène ainsi à des choses semblables à ce qu'on a vu au premier chapitre : on va déduire des bornes sur les minimiseurs du risque empirique si on est capable :

- de contrôler uniformément le processus empirique $(P-P_n)(\ell_\theta - \ell_{\theta^*})$ sur un voisinage de θ^*,
- de minorer uniformément l'excès de risque sur ce même voisinage.

Nous allons montrer dans les prochaines sections les résultats suivants. Soit $\mathbf{M}^*$ la matrice de l'application linéaire qui envoie θ^* sur $\theta^*/(1+\|\theta^*\|_2)^3$ et les vecteurs u orthogonaux à θ^* sur $u/(1+\|\theta^*\|_2)$. La matrice $\mathbf{M}^*$ est symétrique et positive, elle définit les ellipsoïdes

$$r\mathcal{E} = \{t \in \mathbb{R}^d : t^T\mathbf{M}^*t \leqslant r^2\} \ .$$

Soit alors r_0 le plus grand r tel que $r\mathcal{E}$ est inclus dans la boule de rayon $\|\theta^*\|_2/2$. On a,

$$\forall t \in r_0\mathcal{E}, \qquad P(\ell_{\theta^*+t} - \ell_{\theta^*}) \asymp t^T \mathbf{M}^* t \ .$$

En particulier, il existe donc une constante numérique telle que, pour tout $r \leqslant r_0$,

$$\inf_{t \in \partial r\mathcal{E}} P(\ell_{\theta^*+t} - \ell_{\theta^*}) \geqslant cr^2 \ .$$

De plus, il existe une autre constante numérique C telle que, pour tout $\delta \in (0,1)$, avec probabilité $1-\delta$, on a

$$\sup_{t \in r\mathcal{E}} (P - P_n)(\ell_{\theta^*+t} - \ell_{\theta^*}) \leqslant \frac{C\mathrm{Comp}(\mathcal{E},\delta)}{\sqrt{n}} r$$
$$\mathrm{Comp}(\mathcal{E},\delta) = \sqrt{\|\theta^*\|_2^3 + d(1+\|\theta^*\|_2)} + \sqrt{(\|\theta^*\|^3 \vee 1)\log(1/\delta)} \ .$$

En injectant ces contrôles dans la borne inférieure (2.1), on en déduit que, avec probabilité $1-\delta$, pour tout $r \leqslant r_0$ et tout θ tel que $\theta - \theta^* \in r\mathcal{E}$,

$$P_n(\ell_\theta - \ell_{\theta^*}) \geqslant cr^2 - \frac{C\mathrm{Comp}(\mathcal{E},\delta)}{\sqrt{n}} r \ .$$

Cette dernière quantité est $\geqslant 0$ si

$$r \geqslant \frac{C}{c} \frac{\mathrm{Comp}(\mathcal{E},\delta)}{\sqrt{n}} \ .$$

Ainsi, si

$$\frac{C}{c} \frac{\mathrm{Comp}(\mathcal{E},\delta)}{\sqrt{n}} \leqslant r_0 \ ,$$

on peut en déduire que

$$\hat{\theta} - \theta^* \in \frac{C}{c} \frac{\mathrm{Comp}(\mathcal{E},\delta)}{\sqrt{n}} \mathcal{E} \ ,$$

ce qui équivaut en particulier à

$$P(\ell_{\hat{\theta}} - \ell_{\theta^*}) \leqslant C' \frac{\mathrm{Comp}(\mathcal{E},\delta)^2}{n} \ .$$

On voit ainsi que la méthode de localisation conduit à des vitesses de convergence en $1/n$ pour l'excès de risque du minimiseur du risque empirique.

Le but de la section 2.3 est maintenant de donner les éléments pour comprendre le comportement de l'excès de risque tandis que la section 2.4 développera les outils pour majorer le processus empirique.

2.3 Excès de risque

Le but de cette section est de comprendre comment gérer le premier terme dans l'inégalité (2.1). La fonction φ étant régulière, on peut utiliser un développement de Taylor pour écrire, pour tout x, h,

$$\varphi(x+h) - \varphi(x) = h\varphi'(x) + h^2 \int_0^1 \varphi''(x+uh)(1-u)\mathrm{d}u \ .$$

On en déduit

$$\begin{aligned}\ell_{\theta^*+t}(x,y) - \ell_{\theta^*}(x,y) &= \varphi(-y\langle x, \theta^*+t\rangle) - \varphi(-y\langle x,\theta^*\rangle) \\ &= -y\varphi'(-y\langle x,\theta^*\rangle)\langle x,t\rangle + \langle x,t\rangle^2 \int_0^1 \varphi''(-y\langle x,\theta^*+ut\rangle)(1-u)\mathrm{d}u \ .\end{aligned}$$

En prenant l'espérance, on obtient

$$P(\ell_{\theta^*+t} - \ell_{\theta^*}) = \langle \mathcal{L}, t\rangle + t^T Q_t t \ ,$$

où $\mathcal{L} = \nabla P\ell_{\theta^*}$ et $Q_t = P[(\int_0^1 \varphi''(-Y\langle X, \theta^*+ut\rangle)(1-u)\mathrm{d}u)XX^T]$.

Une première première remarque élémentaire est que, comme θ^* est le minimiseur de $P\ell_{\theta^*}$, on a $\mathcal{L} = 0$ et donc

$$P(\ell_{\theta^*+t} - \ell_{\theta^*}) = t^T Q_t t \ .$$

D'autre part, en conditionnant par X et en notant $\eta(x) = \mathbb{E}[Y|X=x]$ la fonction de régression, on peut écrire

$$\begin{aligned}&P(\varphi''(-Y\langle X,\theta^*+ut\rangle)XX^T) \\ &\quad = \mathbb{E}[(\eta(X)\varphi''(-\langle X,\theta^*+ut\rangle) + (1-\eta(X))\varphi''(\langle X,\theta^*+ut\rangle))XX^T] \ .\end{aligned}$$

D'autre part, on vérifie que la fonction logistique vérifie

$$\varphi''(x) = \varphi'(x)\varphi'(-x) \ ,$$

donc en particulier $\varphi''(-x) = \varphi''(x)$ et donc

$$P(\varphi''(-Y\langle X,\theta^*+ut\rangle)XX^T) = \mathbb{E}[\varphi''(\langle X,\theta^*+ut\rangle)XX^T] \ .$$

Rappelons que X est un vecteur Gaussien standard de $\mathbb{R}^d$. On peut vérifier que, si $\|t\|_2 \leqslant \|\theta^*\|_2/2$, en notant $u^* = \theta^*/\|\theta^*\|_2$,

$$\mathbb{E}[\varphi''(\langle X,\theta^*+ut\rangle)XX^T] \asymp \frac{\langle t,u^*\rangle^2}{(1+\|\theta^*\|_2)^3} + \frac{\|t\|_2^2 - \langle t,u^*\rangle^2}{1+\|\theta^*\|_2} = t^T\mathbf{M}^*t \ , \quad (2.2)$$

où

$$\mathbf{M}^* = P\begin{bmatrix}(1+\|\theta^*\|_2)^{-3} & 0 & & 0 \\ 0 & (1+\|\theta^*\|_2)^{-1} & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & & 0 & (1+\|\theta^*\|_2)^{-1}\end{bmatrix}P^T \ ,$$

et P est la matrice de passage de la base canonique vers la base de premier vecteur u^* complétée en une base orthonormée de $\mathbb{R}^d$. Sans faire le calcul précis (que les courageux s'y attèlent!!), on peut se convaincre de cette affirmation en rappelant que, comme X est un vecteur Gaussien standard, on a, si $\langle v, u^* \rangle = 0$, $\langle v, X \rangle$ et $\langle u^*, X \rangle$ indépendants. De plus, par régularité de φ, la matrice $Q_t \approx Q = \mathbb{E}[\varphi''(\langle X, \theta^* \rangle) X X^T]$ pour les petits t. Enfin, en notant $G \sim \mathsf{N}(0,1)$ une Gaussienne standard,

$$(u^*)^T Q u^* = \mathbb{E}[\varphi''(\|\theta^*\|_2 G) G^2] \ ,$$

tandis que, pour tout v orthogonal à u^*,

$$v^T Q v = \mathbb{E}[\varphi''(\|\theta^*\|_2 G)] \mathbb{E}[G^2] \ ,$$

On a alors, comme $\varphi''(x)$ est presque nulle hors d'un compact de taille constante,

$$\mathbb{E}[\varphi''(\|\theta^*\|_2 G) G^2] \asymp \int_{-K/\|\theta^*\|_2}^{K/\|\theta^*\|_2} x^2 \mathrm{d}x \asymp \frac{1}{\|\theta^*\|_2^3}, \qquad \mathbb{E}[\varphi''(\|\theta^*\|_2 G)] \asymp \frac{1}{\|\theta^*\|_2} \ .$$

En intégrant l'équation (2.2), on voit que l'excès de risque se comporte de la façon suivante :

$$\forall t : \|t\|_2 \leqslant \|\theta^*\|_2/2, \qquad P(\ell_{\theta^*+t} - \ell_{\theta^*}) \asymp t^T \mathbf{M}^* t \ .$$

En particulier donc, il peut être minoré de la façon suivante : Pour tout $r > 0$, soit $r\mathcal{E} = \{t \in \mathbb{R}^d : t^T \mathbf{M}^* t \leqslant r^2\}$ et soit r_0 tel que $r_0 \mathcal{E}$ soit inclus dans la boule Euclidienne de rayon $\|\theta^*\|_2/2$. Alors, il existe une constante numérique $c > 0$ telle que

$$\forall r \leqslant r_0, \forall t \in \partial r\mathcal{E}, \qquad P(\ell_{\theta^*+t} - \ell_{\theta^*}) \geqslant c r^2 \ .$$

2.4 Majoration de processus stochastiques

Dans cette section, on explique comment majorer le processus empirique $(P - P_n)(\ell_{\theta^*+t} - \ell_{\theta^*})$ uniformément sur l'ellipsoïde $r\mathcal{E}$. Pour cela, on va prendre un peu de recul et noter $(X_t)_{t \in T}$ un processus stochastique, c'est à dire une collection de variables aléatoires à valeurs réelles indexées par un ensemble fini T. On s'intéresse à l'obtention de majoration de $\sup_{t \in T} X_t$.

L'idée est d'appliquer les résultats généraux que nous allons développer ici au processus $X_t = (P - P_n)(\ell_{\theta^*+t} - \ell_{\theta^*})$. Ce processus n'est pas indexé par un ensemble fini, mais séparable dans les applications, de sorte que

$$\sup_{t \in T} X_t = \sup_{T_0 \subset T, |T_0| < \infty} \sup_{t \in T_0} X_t \ ,$$

et l'extension du cas fini au cas séparable se fait par simple application du théorème de convergence dominée. On laissera au lecteur intéressé le soin de vérifier les détails de cette extension. Par ailleurs, on va toujours faire l'hypothèse $\mathbb{E}[X_t] = 0$ et comme $X_0 = 0$, on va aussi supposer $0 \in T$ et que cette hypothèse est vérifiée en général.

Avant de majorer $\sup_{t\in T} X_t$, on va s'intéresser aux hypothèses sous lesquelles on sait majorer X_t seul. On passera ensuite à un résultat uniforme.

2.4.1 Majoration d'un X_t

Rappelons que l'idée est d'étendre le résultat vu au premier chapitre déduit de l'inégalité de Mc-Diarmid, à des pertes convexes, donc non-bornées.

Le résultat suivant caractérise précisément les variables pour lesquelles on est capable de montrer une inégalité de déviation aussi bonne que l'inégalité de Mc-Diarmid.

Théorème 15. *Soit X une variable aléatoire. Les propriétés suivantes sont équivalentes.*

(i) Pour tout $x > 0$, $\mathbb{P}(|X| > x) \leqslant 2\exp(-x^2/K_1^2)$.

(ii) Pour tout entier $p \geqslant 1$, $\|X\|_p := \mathbb{E}[|X|^p]^{1/p} \leqslant K_2\sqrt{p}$.

(iii) Pour tout $|s| < 1/K_3$, $\mathbb{E}[\exp(s^2X^2)] \leqslant \exp(s^2K_3^2)$.

(iv) $\mathbb{E}[\exp(X^2/K_4^2)] \leqslant 2$.

Si de plus, X est centré, alors ces propriétés sont aussi équivalentes à

(v) Pour tout $s \in \mathbb{R}$, $\mathbb{E}[\exp(sX)] \leqslant \exp(s^2K_5^2)$.

Si on a une de ces propriétés, on dit que X est sous-Gaussienne et les différents K_i ne diffèrent que d'une constante numérique multiplicative. La plus petite constante K_4 telle que (iv) est vérifiée est appelée la norme sous-Gaussienne de X, elle est notée $\|X\|_{\psi_2}$.

On a utilisé déjà à de nombreuses reprises le résultat $(v) \implies (i)$. On a vu que l'inégalité de Markov exponentielle implique que, si (v) est vérifiée pour une constante K_5, alors (i) est vérifiée avec $K_1 = 2K_5$.

En exercice, vous pouvez vérifier qu'on a nécessairement $\mathbb{E}[X] = 0$ si (v) est vérifiée.

Démonstration. On prouve la chaîne d'implications.

(i) $\implies$ (ii) Supposons $K_1 = 1$. Alors, pour tout $p \geqslant 1$, on a

$$\begin{aligned}
\mathbb{E}[|X|^p] &= \int_0^{+\infty} \mathbb{P}(|X|^p > t)\mathrm{d}t \quad \text{car } |X|^p \geqslant 0 \text{ a.s.} \\
&= p\int_0^{+\infty} \mathbb{P}(|X| > u)u^{p-1}\mathrm{d}u \quad \text{en posant } t = u^p \\
&\leqslant p\int_0^{+\infty} (u^2)^{p/2-1}\exp(-u^2)2u\mathrm{d}u \\
&= p\Gamma(p/2) \quad \text{en posant } u^2 = v \ .
\end{aligned}$$

On utilise alors les estimés suivants de Stirling, qui sont vérifiés pour tout $x > 0$,

$$\left(\frac{x}{e}\right)^x \leqslant \Gamma(x+1) \leqslant x^x \ . \tag{2.3}$$

On a alors

$$\|X\|_p \leqslant p^{1/p}\sqrt{\frac{p}{2}} \leqslant 2\sqrt{p} \ .$$

Soit maintenant $K_1 > 0$, alors $\mathbb{P}(|X/K_1| > x/K_1) \leqslant \exp(-(x/K_1)^2)$, donc $|X/K_1|$ vérifie (i) avec $K_1 = 1$, donc $\|X/K_1\|_p \leqslant 2\sqrt{p}$ et donc $\|X\|_p \leqslant 2K_1\sqrt{p}$.

(ii) $\Longrightarrow$ (iii) Supposons $K_2 = 1$. On développe l'exponentielle en série entière pour obtenir

$$\mathbb{E}[\exp(s^2X^2)] \leqslant \sum_{k=0}^{+\infty} \frac{s^{2k}k^k}{\Gamma(k+1)} \leqslant \sum_{k=0}^{+\infty} s^{2k}e^k \ .$$

La dernière inégalité vient alors de (2.3). Cette dernière borne supérieure est finie si $|s| < 1/\sqrt{e}$ et alors

$$\mathbb{E}[\exp(s^2X^2)] \leqslant \frac{1}{1-es^2} \ .$$

De plus, pour tout $|s| < 1/\sqrt{2e}$, on a

$$\mathbb{E}[\exp(s^2X^2)] \leqslant 1 + \frac{es^2}{1-es^2} \leqslant 1 + 2es^2 \leqslant \exp(2es^2) \ .$$

Soit alors $K_2 > 0$, alors $\|X/K_2\|_p \leqslant \sqrt{p}$ pour tout $p \geqslant 1$ et donc pour tout $|s| < 1/\sqrt{2e}$

$$\mathbb{E}[\exp((s/K_2)^2X^2)] \leqslant \exp(2eK_2^2(s/K_2)^2) \ .$$

On a donc, pour tout $|s| < 1/\sqrt{2e}K_2$, $\mathbb{E}[\exp(s^2X^2)] \leqslant \exp(s^2 2eK_2^2)$.

(iii) $\Longrightarrow$ (iv) est triviale.

(iv) $\Longrightarrow$ (i) Soit $x > 0$, on a

$$\mathbb{P}(|X| > x) = \mathbb{P}(\exp(X^2/K_4^2) > \exp(x^2/K_4^2)) \leqslant 2\exp(-x^2/K_4^2) \ ,$$

le dernière inégalité étant une conséquence de l'inégalité de Markov.

(iii) $\Longrightarrow$ (v) Soit $s \in \mathbb{R}$. Si $|s| < 1/K_3$, on utilise l'inégalité $\exp(x) \leqslant x + \exp(x^2)$ qui donne

$$\mathbb{E}[\exp(sX)] \leqslant \mathbb{E}[\exp(s^2X^2)] \leqslant \exp(s^2K_3^2) \ .$$

Si $|s| \geqslant 1/K_3$, on utilise la majoration $sx \leqslant \frac{1}{4}s^2K_3^2 + \frac{x^2}{K_3^2}$ pour obtenir

$$\begin{aligned}\mathbb{E}[\exp(sX)] &\leqslant \exp(s^2K_3^2/4)\mathbb{E}[\exp(X^2/K_3^2)] \\ &\leqslant 2\exp(s^2K_3^2/4) \leqslant \exp(s^2K_3^2) \ .\end{aligned}$$

(v) $\Longrightarrow$ (i) Soit $x > 0$, on a d'après l'inégalité de Markov exponentielle

$$\begin{aligned}\mathbb{P}(X > x) &= \inf_{s>0} \mathbb{P}(\exp(sX) > \exp(sx)) \\ &\leqslant \inf_{s>0} \exp(-sx + \log \mathbb{E}[\exp(sX)]) \\ &= \exp(-\sup_{s>0}\{sx - \log \mathbb{E}[\exp(sX)]\}) \ . \end{aligned} \tag{2.4}$$

Si (v) est vérifiée, on a alors

$$\mathbb{P}(X > x) \leqslant \exp(-\sup_{s>0}\{sx - s^2K_5^2\}) = \exp(-x^2/4K_5^2) \ .$$

De même, on a

$$\mathbb{P}(X < -x) = \mathbb{P}(-X > x) \leqslant \exp(-x^2/4K_5^2) \ .$$

Finalement

$$\mathbb{P}(|X| > x) \leqslant 2\exp(-x^2/4K_5^2) \ .$$

□

Rassemblons avant de terminer cette section quelque remarques qui s'avèreront utiles.

1. La première est que $\|X\|_{\psi_2}$ définit bien norme. (En particulier, vérifiez que l'inégalité triangulaire découle de la convexité de la fonction $x \mapsto \exp(x^2)$).
2. La seconde est que, si $\|X\|_{\psi_2} \leqslant K$, alors

$$\|X - \mathbb{E}[X]\|_{\psi_2} \leqslant \|X\|_{\psi_2} + |\mathbb{E}[X]| \leqslant C\|X\|_{\psi_2} \leqslant CK \ .$$

La première inégalité vient du fait que $\|\cdot\|_{\psi_2}$ est une norme, la seconde du point (ii).

3. Les variables bornées sont sous-Gaussiennes. En effet, si $X \in [a, b]$ p.s., alors $\|X - \mathbb{E}[X]\|_{\psi_2} \leqslant (b-a)/\log 2$ car $\exp(-\log 2(X - \mathbb{E}[X])^2/(b-a)^2) \leqslant 2$ p.s. Ainsi, le point (iv) est vérifié. Le lemme d'Hoeffding montre aussi que (v) est vérifié avec $K_5 = (b-a)/\sqrt{8}$.
4. Une variable aléatoire peut ne pas être bornée mais être sous-Gaussienne. Par exemple, si $X \sim \mathrm{N}(0, \sigma^2)$ et φ est une fonction L-Lipschitzienne, alors $Y = \varphi(X) - \mathbb{E}[\varphi(X)]$ est sous Gaussienne. En effet, si $X' \sim \mathrm{N}(0, \sigma^2)$ est indépendant de X, $\mathbb{E}[\varphi(X)] = \mathbb{E}[\varphi(X')|X]$, donc

$$\begin{aligned}\mathbb{E}\big[\exp\big(s^2Y^2\big)\big] &\leqslant \mathbb{E}\big[\exp\big(s^2(\varphi(X) - \varphi(X'))^2\big)\big] \quad \text{Jensen} \\ &\leqslant \mathbb{E}\big[\exp\big(L^2s^2(X - X')^2\big)\big] \quad \text{Lipshitz} \\ &= \mathbb{E}\big[\exp\big(2L^2s^2X^2\big)\big] \quad X - X' \sim \sqrt{2}X \\ &= \int \exp\left(-\left(\frac{1}{2\sigma^2} - 2L^2s^2\right)x^2\right)\frac{\mathrm{d}x}{\sqrt{2\pi\sigma^2}} \\ &= \frac{1}{\sqrt{1 - 4\sigma^2s^2L^2}} \quad \text{si } 4\sigma^2s^2L^2 < 1 \ . \end{aligned}$$

Il s'en suit que $\|Y\|_{\psi_2} \leqslant 4\sigma L/\sqrt{3}$.

Vérifions que les variables $X_t = (P-P_n)(\ell_{\theta^*+t}-\ell_{\theta^*})$ sont sous-Gaussiennes. Le vecteur X étant un vecteur Gaussien standard, on a pour tout vecteur u de la sphère unité $\langle u, X\rangle \sim \mathsf{N}(0,1)$, donc, pour tout $|s| < 1/\sqrt{2}$,

$$\mathbb{E}\big[\exp\big(s^2\langle u, X\rangle^2\big)\big] = \int_{\mathbb{R}} \exp\left(-\left(\frac{1}{2}-s^2\right)x^2\right)\frac{\mathrm{d}x}{\sqrt{2\pi}} = \frac{1}{\sqrt{1-2s^2}} \leqslant \exp(s^2) \ .$$

Ainsi $\|\langle X, u\rangle\|_{\psi_2} = \sqrt{1/\log 2} =: C_0$. On en déduit

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\frac{(\varphi(-Y\langle\theta^*+t, X\rangle) - \varphi(-Y\langle\theta^*, X\rangle))^2}{C_0^2\|t\|_2^2}\right)\right] \\
&\leqslant \mathbb{E}\left[\exp\left(\frac{(-Y\langle t, X\rangle)^2}{C_0^2\|t\|_2^2}\right)\right] \qquad \varphi\,1-\text{Lipschitz} \\
&= \mathbb{E}\left[\exp\left(\frac{\langle t, X\rangle^2}{C_0^2\|t\|_2^2}\right)\right] \qquad Y \in \{-1,1\} \\
&= \mathbb{E}\left[\exp\left(\frac{\langle t/\|t\|_2, X\rangle^2}{C_0^2}\right)\right] \leqslant 2 \ .
\end{aligned}$$

Ainsi $\|(\ell_{\theta^*+t}-\ell_{\theta^*})(X,Y)\|_{\psi_2} \leqslant C_0$. Par la remarque sur le centrage, il existe donc une constante numérique C_1 telle que

$$\|(\ell_{\theta^*+t}-\ell_{\theta^*})(X,Y) - P(\ell_{\theta^*+t}-\ell_{\theta^*})\|_{\psi_2} \leqslant C_1\|t\|_2 \ .$$

Par la caractérisation (v), on a donc, pour tout $s \in \mathbb{R}$,

$$\mathbb{E}\big[\exp\big(s(\ell_{\theta^*+t}-\ell_{\theta^*})(X,Y) - P(\ell_{\theta^*+t}-\ell_{\theta^*})\big)\big] \leqslant \exp\big(C_2^2\|t\|_2^2 s^2\big) \ .$$

Par indépendance, il s'en suit donc que, pour tout $s \in \mathbb{R}$,

$$\mathbb{E}\big[\exp\big(s(P_n-P)(\ell_{\theta^*+t}-\ell_{\theta^*})\big)\big] \leqslant \exp\left(\frac{C_2^2\|t\|_2^2 s^2}{n}\right) \ .$$

Autrement dit

$$\|X_t\|_{\psi_2} \leqslant \frac{C_3\|t\|_2}{\sqrt{n}}, \qquad X_t = (P_n-P)(\ell_{\theta^*+t}-\ell_{\theta^*}) \ .$$

La concentration de X_t s'obtient alors en utilisant la première caractérisation des variables sous-Gaussiennes. Résumons les résultats obtenus dans cette section dans le résultat suivant.

Théorème 16. *Il existe des constantes numériques C, C' telles que*

$$\|X_t\|_{\psi_2} \leqslant \frac{C\|t\|_2}{\sqrt{n}}, \qquad X_t = (P_n-P)(\ell_{\theta^*+t}-\ell_{\theta^*}) \ .$$

En particulier, on a donc, pour tout $z > 0$,

$$\mathbb{P}(X_t > z) \leqslant \exp\left(-\frac{C'nz^2}{\|t\|_2^2}\right) \ .$$

2.4.2 Bornes uniformes par la méthode de chaînage

La méthode de chaînage permet de déduire de la concentration X_t des bornes sur $\sup_{t\in T} X_t$. On va présenter cette méthode en général sous l'hypothèse que le processus X_t est à accroissements sous-Gaussiens, c'est à dire qu'il existe une distance d telle que, pour tout $z > 0$,

$$\forall t, t' \in T, \qquad \mathbb{P}\big(X_t - X_{t'} > z\big) \leqslant \exp\left(- \frac{z^2}{d(t,t')^2} \right) \ .$$

Il est assez immédiat (faites le si vous n'êtes pas convaincu), d'étendre le Théorème 16 pour vérifier que le processus $X_t = (P_n - P)(\ell_{\theta^*+t} - \ell_{\theta^*})$ est à accroissements sous-Gaussien pour la distance

$$d(t,t') = \frac{C}{\sqrt{n}} \|t - t'\|_2 \ .$$

Le contrôle qu'on va obtenir se base alors sur l'idée de *chaînage* :

1. On construit une suite croissante de sous-ensembles $T_k \subset T$ finis telle que $T_0 = \{0\}$ et $T_N = T$ pour un certain N.
2. Pour tout $k \in \{0, \dots, N\}$ et tout $t \in T$, on note $\pi_k(t)$ un plus proche voisin de t dans T_k.

On peut alors construire une chaîne d'approximations successives de X_t : Pour tout $t \in T$,

$$X_t = X_t - X_0 = X_{\pi_N(t)} - X_{\pi_0(t)} = \sum_{k=1}^{N} X_{\pi_k(t)} - X_{\pi_{k-1}(t)} \ .$$

Remarquons tout de suite (nous y reviendrons) qu'on aurait pu décomposer cette différence seulement le long d'une sous-suite $X_{\pi_{i_k}(t)}$.

On obtient alors

$$\sup_{t\in T} X_t \leqslant \sup_{t\in T} \sum_{k=1}^{N} X_{\pi_k(t)} - X_{\pi_{k-1}(t)} \ .$$

Par hypothèse, on a, pour tout $k \in \{1, \dots, N\}$, tout $t \in T$ et tout $z_k > 0$,

$$\mathbb{P}(X_{\pi_k(t)} - X_{\pi_{k-1}(t)} > z_k) \leqslant \exp\left(- \frac{z_k^2}{d(\pi_{k-1}(t), \pi_k(t))^2} \right) \ ,$$

Par une borne d'union, en notant $r_k(t) = d(\pi_{k-1}(t), \pi_k(t))$, on en déduit donc

$$\mathbb{P}(\exists t \in T,\ X_{\pi_k(t)} - X_{\pi_{k-1}(t)} > r_k(t) z_k) \leqslant |T_k||T_{k-1}| \exp\big(- z_k^2 \big) \ .$$

Finalement, en appliquant encore une borne d'union, on conclut que

$$\mathbb{P}\Big(\sup_{t\in T} X_t > \sup_{t\in T} \sum_{k=1}^N r_k(t) z_k \Big) \leqslant \sum_{k=1}^N |T_k||T_{k-1}| \exp\big(-z_k^2\big) \ . \tag{2.5}$$

Pour n'importe quelle suite ℓ_k telle que $\sum_{k=1}^{+\infty} \exp(-\ell_k^2) = C$, on a donc, pour tout $z > 0$,

$$\mathbb{P}\Big(\sup_{t\in T} X_t > \sup_{t\in T} \sum_{k=1}^N r_k(t)(\sqrt{\log(|T_k||T_{k-1}|)} + z + \ell_k) \Big) \leqslant C\exp(-z^2) \ . \tag{2.6}$$

Il y a naturellement deux façons d'optimiser cette borne : soit en fixant les distances maximales $\sup_{t\in T} r_k(t)$ et en optimisant sous cette contrainte le cardinal $|T_k|$, soit en fixant les cardinaux $|T_k|$ et en optimisant ensuite les distance $r_k(t)$.

Il s'avère que le premier point de vue est légèrement sous-optimal dans certains cas et nous allons donc développer le second point ici. En guise d'exercice, je vous encourage toutefois à développer le premier !

Définition 17. *Une suite $(T_k)_k$ de sous ensembles croissants $T_k \subset T$ est dite admissible si $T_0 = \{t_0\}$ et $|T_k| \leqslant 2^{2^k}$, pour tout $k \geqslant 1$.*

Etant donnée une suite admissible (T_k), on note $d(t, T_k) = \inf_{t'\in T_k} d(t, t')$. On a donc, pour tout $t \in T$,

$$r_k(t) = d(\pi_{k-1}(t), \pi_k(t)) \leqslant d(t, T_{k-1}) + d(t, T_k) \leqslant 2d(t, T_{k-1}) \ ,$$

la dernière égalité se déduisant du fait que la suite T_k est croissante. De plus, on a bien sûr

$$\sqrt{\log(|T_k||T_{k-1}|)} \leqslant \sqrt{2^k + 2^{k-1}} = \sqrt{3 * 2^{k-1}} \ .$$

Cette suite satisfaisant clairement $\sum_{k=0}^{+\infty} \exp(-3*2^k) \leqslant 1$, on déduit de (2.6), en posant $\ell_k = \sqrt{3 * 2^{k-1}}$ que

$$\mathbb{P}\Big(\sup_{t\in T} X_t > 2\sup_{t\in T} \sum_{k=0}^{+\infty} d(t, T_k)(2\sqrt{3 * 2^k} + z) \Big) \leqslant \exp(-z^2) \ .$$

Ainsi, les déviations de $\sup_{t\in T} X_t$ sont contrôlées grâce à deux quantités géométriques

$$\gamma_2(T) = \inf_{(T_k)} \sup_{t\in T} \sum_{k=0}^{+\infty} \sqrt{2^k} d(t, T_k), \qquad \Delta(T) = \sup_{t\in T} \sum_{k=0}^{+\infty} d(t, T_k) \ .$$

Rappelons qu'on aurait pu se limiter dans le chaînage à une sous-suite de k pour lesquels $d(t, T_{i_k}) \leqslant d(t, T_{i_{k-1}})/2$ et donc qu'on peut majorer $\Delta(T)$

par 2diam(T), où diam$(T) = \sup_{t,t'} d(t,t')$ désigne le diamètre de T pour la distance d. Je vous encourage en guise d'exercice qu'on peut encore majorer le premier terme par $\gamma_2(T)$.

D'autre part, un résultat profond, connu sous le nom de théorème des mesures majorantes de Talagrand prouve que, si G_t est un processus Gaussien centré tel que $\mathbb{E}[(G_t - G_{t'})^2] = d(t,t')^2$, alors il existe une constante C numérique telle que

$$\gamma_2(T) \leqslant C\mathbb{E}[\sup_{t\in T} G_t] \ .$$

La quantité $\mathbb{E}[\sup_{t\in T} G_t]$ est appelée la largeur Gaussienne de l'ensemble T et est notée $w_d(T) = \mathbb{E}[\sup_{t\in T} G_t]$. Résumons les résultats obtenus dans cette section dans un théorème.

Théorème 18. *Soit $(X_t)_{t\in T}$ un processus centré indexé par un espace métrique T muni d'une distance d telle que, pour tout $z > 0$,*

$$\forall t, t' \in T, \qquad \mathbb{P}\big(X_t - X_{t'} > z\big) \leqslant \exp\left(-\frac{z^2}{d(t,t')^2}\right) \ .$$

Alors on a, pour tout $z > 0$,

$$\mathbb{P}\left(\sup_{t\in T} X_t > 4\sqrt{3}\gamma_d(T) + 4diam_d(T)z)\right) \leqslant \exp(-z^2) \ ,$$

où diam(T) désigne le diamètre de T pour la métrique d et

$$\gamma_2(T) = \inf_{(T_k)} \sup_{t\in T} \sum_{k=0}^{+\infty} \sqrt{2^k} d(t, T_k) \ ,$$

est la fonctionnelle de Talagrand, l'infimum étant pris sur les suites admissibles. De plus, $\gamma_2(T)$ est majorée par $Cw_d(T)$, où $w_d(T)$ est la largeur Gaussienne de T pour la distance d, i.e. $\mathbb{E}[\sup_{t\in T} G_t]$, où G_t est un processus Gaussien centré tel que $\mathbb{E}[(G_t - G_{t'})^2] = d(t,t')^2$.

On peut maintenant appliquer ce résultat général au processus

$$X_t = (P - P_n)(\ell_{\theta^*+t} - \ell_{\theta^*}) \ .$$

On a vu que ce processus est à accroissements sous-Gaussiens pour la distance $d(t,t') = C\|t - t'\|_2/\sqrt{n}$. On déduit donc du résultat général que, pour tout $z > 0$,

$$\mathbb{P}\left(\sup_{t\in T} X_t > 4\sqrt{3}\gamma_d(T) + 4\mathrm{diam}_d(T)z)\right) \leqslant \exp(-z^2) \ ,$$

On a clairement

$$w_d(T) = \frac{C}{\sqrt{n}} w_2(T), \qquad \mathrm{diam}_d(T) = \frac{C}{\sqrt{n}} \mathrm{diam}_2(T) \ ,$$

où $w_2(T)$ et $\mathrm{diam}_2(T)$ désignent la largeur Gaussienne et le diamètre de T pour la distance Euclidienne. De plus, si G est un vecteur Gaussien standard de $\mathbb{R}^d$, on a, pour tout $t \in \mathbb{R}^d$,

$$\mathrm{Var}(\langle X, t\rangle) = t^T \mathbb{E}[XX^T]t = \|t\|_2^2 \ ,$$

donc le processus Gaussien $G_t = \langle G, t\rangle$ vérifie, pour tout $t, t' \in T$,

$$\mathbb{E}[(G_t - G_{t'})^2] = \|t - t'\|_2^2 \ .$$

On peut alors rassembler les informations obtenues pour le contrôle du processus empirique dans le résultat suivant.

Théorème 19. *Il existe une constante numérique C telle que, pour tout $z > 0$,*

$$\mathbb{P}\left(\sup_{t\in T} (P - P_n)(\ell_{\theta^* + t} - \ell_{\theta^*}) > \frac{C}{\sqrt{n}} (\mathbb{E}[\sup_{t\in T} \langle G, t\rangle] + \mathit{diam}_2(T) z) \right) \leqslant \exp(-z^2) \ ,$$

où G est un vecteur Gaussien standard de $\mathbb{R}^d$.

On peut spécifier maintenant ce résultat lorsque T est un ellipsoïde. Soit S une matrice symétrique symétrique définie positive et soit

$$\mathcal{E} = \{v \in \mathbb{R}^d : v^T S v \leqslant 1\} \ .$$

On écrit la décomposition de S sur une bon de vecteurs propres

$$S = \sum_{i=1}^d s_i u_i u_i^T \ .$$

L'hypothèse S définie positive implique que tous les $s_i > 0$. On a donc $v = \sum_{i=1}^d v_i u_i \in \mathcal{E}$ si

$$\sum_{i=1}^d s_i v_i^2 \leqslant 1 \ .$$

Soit $v \in \mathcal{E}$, on a

$$\|v\|_2^2 = \sum_{i=1}^d v_i^2 \leqslant \frac{1}{\min s_i} \sum_{i=1}^d s_i v_i^2 \leqslant \|S^{-1}\| \ .$$

Il s'en suit que

$$\mathrm{diam}_2(\mathcal{E}) \leqslant 2\sqrt{\|S^{-1}\|} \ .$$

De plus, par Cauchy-Schwarz,

$$\langle v, G\rangle = \sum_{i=1}^d v_i \langle G, u_i\rangle \leqslant \sqrt{\sum_{i=1}^d s_i v_i^2 \sum_{i=1}^d \frac{\langle G, u_i\rangle^2}{s_i}} \leqslant \sqrt{\sum_{i=1}^d \frac{\langle G, u_i\rangle^2}{s_i}} \ .$$

Donc, par Cauchy-Schwarz,

$$\mathbb{E}[\sup_{v\in\mathcal{E}}\langle v, G\rangle] \leqslant \sqrt{\mathbb{E}\Big[\sum_{i=1}^{d}\frac{\langle G, u_i\rangle^2}{s_i}\Big]} = \sqrt{\mathrm{Tr}(S^{-1})}\ .$$

Ainsi, on a pour tout ellipsoïde $\mathcal{E} = \{v \in \mathbb{R}^d : v^T S v \leqslant 1\}$,

$$w_2(\mathcal{E}) \leqslant \sqrt{\mathrm{Tr}(S^{-1})}, \qquad \mathrm{diam}_2(\mathcal{E}) \leqslant 2\sqrt{\|S^{-1}\|}\ .$$

En particulier, pour l'ellipsoïde $r\mathcal{E} = \{v \in \mathbb{R}^d : v^T \mathbf{M}^* v \leqslant r^2\}$, qui correspond à la matrice $S = \mathbf{M}^*/r^2$, on a donc

$$w_2(r\mathcal{E}) \leqslant r\sqrt{\mathrm{Tr}((\mathbf{M}^*)^{-1})} \leqslant Cr\sqrt{\|\theta^*\|_2^3 + d(1+\|\theta^*\|_2)}\ ,$$
$$\mathrm{diam}_2(r\mathcal{E}) \leqslant 2r\sqrt{\|(\mathbf{M}^*)^{-1}\|} \leqslant Cr\sqrt{(\|\theta^*\|_2^3 \vee 1)}\ .$$

En injectant ces bornes dans le résultat général Théorème 19, on en déduit que, pour tout $\delta \in (0,1)$, avec probabilité $1-\delta$,

$$\sup_{t\in r\mathcal{E}}(P - P_n)(\ell_{\theta^*+t} - \ell_{\theta^*}) \leqslant \frac{C\mathrm{Comp}(\mathcal{E},\delta)}{\sqrt{n}} r$$
$$\mathrm{Comp}(\mathcal{E},\delta) = \sqrt{\|\theta^*\|_2^3 + d(1+\|\theta^*\|_2)} + \sqrt{(\|\theta^*\|_2^3 \vee 1)\log(1/\delta)}\ .$$