1 Complexité de Rademacher

Exercice 1. Inégalité de Ledoux Le but de cet exercice est de prouver le théorème suivant :

Theorème 1 (Inégalité de Ledoux). Si φ est B-Lipshitz et $\varepsilon_1^n = \{\varepsilon_i\}_{i=1}^n$ est une suite i.i.d. de variables Rad(1/2), nous avons

$$\mathbb{E}\left[\sup_{\theta\in\Theta}\frac{1}{n}\sum_{i=1}^n \varepsilon_i\varphi\left(\theta^\top x_i\right)\right] \leq B\mathbb{E}\left[\sup_{\theta\in\Theta}\frac{1}{n}\sum_{i=1}^n \varepsilon_i\theta^\top x_i\right].$$

Nous allons montrer par récurrence que pour tout $k \in \{1,\dots,n\}$, **toutes fonctions** $b:\Theta\to\mathbb{R}$, $a_i:\Theta\to\mathbb{R}$, $i\in\{1,\dots,k\}$ et toutes fonctions 1-Lipschitz $\varphi_i:\mathbb{R}\to\mathbb{R}$, $i=1,\dots,k$,

$$\mathbb{E}\left[\sup_{\theta\in\Theta} b(\theta) + \sum_{i=1}^k \varepsilon_i\varphi_i\left(a_i(\theta)\right)\right] \leqslant \mathbb{E}\left[\sup_{\theta\in\Theta} b(\theta) + \sum_{i=1}^k \varepsilon_i a_i(\theta)\right] \tag{1}$$

1. Pour toutes fonctions $\varphi,\psi:\Theta\to\mathbb{R}$ et $\varepsilon\sim\mathrm{Rad}(1/2)$, montrer que :

$$\begin{aligned}\mathbb{E}\left[\sup_{\theta\in\Theta}\{\varphi(\theta)+\varepsilon_i\psi(\theta)\}\right] &= \frac{1}{2}\left[\sup_{\theta,\theta'\in\Theta^2}\{\varphi(\theta)+\varphi(\theta')+\psi(\theta)-\psi(\theta')\}\right]\\ &= \frac{1}{2}\left[\sup_{\theta,\theta'\in\Theta^2}\{\varphi(\theta)+\varphi(\theta')+|\psi(\theta)-\psi(\theta')|\}\right]\end{aligned} \tag{2}$$

2. Supposons que (1) est vérifiée pour $k\in\mathbb{N}$. Montrer que

$$\begin{aligned}&\mathbb{E}\left[\sup_{\theta\in\Theta} b(\theta) + \sum_{i=1}^{k+1}\varepsilon_i\varphi_i\left(a_i(\theta)\right)\right]\\ &\leq \mathbb{E}\left[\mathbb{E}\left[\sup_{\theta,\theta'\in\Theta}\frac{b(\theta)+b\left(\theta'\right)}{2} + \sum_{i=1}^k\varepsilon_i\frac{\varphi_i\left(a_i(\theta)\right)+\varphi_i\left(a_i\left(\theta'\right)\right)}{2} + \frac{\left|a_{k+1}(\theta)-a_{k+1}\left(\theta'\right)\right|}{2}\,\middle|\,\varepsilon_1^k\right]\right]\\ &= \mathbb{E}\left[\mathbb{E}\left[\sup_{\theta\in\Theta} b(\theta) + \sum_{i=1}^k\varepsilon_i\varphi_i\left(a_i(\theta)\right) + \varepsilon_{k+1}a_{k+1}(\theta)\,\middle|\,\varepsilon_1^k\right]\right]\\ &= \mathbb{E}\left[\sup_{\theta\in\Theta} b(\theta) + \sum_{i=1}^{k+1}\varepsilon_i a_i(\theta)\right]\end{aligned}$$

3. Conclure.

Exercice 2 (Dernière étape pour calculer la borne supérieure de la complexité de Rademacher). Nous considérons $X_1^n = \{X_i\}_{i=1}^n$ un n-échantillon de distribution $\mathbb{P}$. Nous supposons qu'il existe $R>0$, tel que $\mathbb{E}[\|X\|^2]\leq R^2$. On considère $\varepsilon_1^n = \{\varepsilon_i\}_{i=1}^n$ une suite de variables aléatoires indépendantes de Rademacher $\mathcal{R}(1/2)$ indépendante de X_1^n. Montrer que :

$$\mathbb{E}\left[\sup_{\|\theta\|\leq D}\frac{1}{n}\sum_{i=1}^n\varepsilon_i\theta^\top X_i\right] \leq \frac{RD}{\sqrt{n}}. \tag{3}$$

Exercice 3 (Inégalité de McDiarmid). Le but de cet exercice est de prouver l'inégalité de McDiarmid, énoncée ci-dessous :

Dfinition 1 (Fonction de différence bornée). $g : \mathsf{Z}^n \to \mathbb{R}$ est une fonction de différence bornée s'il existe des constantes $\{c_i\}_{i=1}^n$ telles que, pour tout $z_1^n = \{z_i\}_{i=1}^n$ et $\tilde{z}_1^n = \{\tilde{z}_i\}_{i=1}^n$,

$$|g(z_1^n) - g(\tilde{z}_1^n)| \leq \sum_{i=1}^n c_i \mathbb{1}_{\{z_i \neq \tilde{z}_i\}} .$$

Theorème 2 (Inégalité de McDiarmid). Si g est une fonction de différence bornée et Z_i sont des variables aléatoires indépendantes alors

$$\mathbb{P}\left(g(Z_1, \ldots, Z_n) - \mathbb{E}\left[g(Z_1, \ldots, Z_n)\right] \geq \varepsilon\right) \leq \mathrm{e}^{\frac{-2\varepsilon^2}{\sum_{i=1}^n c_i^2}}$$
$$\mathbb{P}\left(\mathbb{E}\left[g(Z_1, \ldots, Z_n)\right] - g(Z_1, \ldots, Z_n) \geq \varepsilon\right) \leq \mathrm{e}^{\frac{-2\varepsilon^2}{\sum_{i=1}^n c_i^2}}$$

Pour $i \in \{1, \ldots, n\}$ et $z_1^i = (z_1, \ldots, z_i) \in \mathsf{Z}^i$, nous posons $g_i(z_1^i) = \mathbb{E}[g(z_1^i, Z_{i+1}^n)]$. Par convention, nous posons $g_0 = \mathbb{E}[g(Z_1^n)]$. Notons que $g_n(z_1^n) = g(z_1^n)$, pour tout $i \in \{1, \ldots, n\}$, $\mathbb{E}[g_i(Z_1^i)] = \mathbb{E}[g(Z_1^n)]$ et pour $n \geq 1$

$$g(z_1^n) = g_n(z_1^n) = g_n(z_1^n) - g_{n-1}(z_1^{n-1}) + g_{n-1}(z_1^{n-1}) = \sum_{j=1}^n \{g_j(z_1^j) - g_{j-1}(z_1^{j-1})\} ,$$

avec la convention que $g_0(z_1^0) = g_0$.

1. Montrer que pour tout $\lambda \in \mathbb{R}$,
$$\mathbb{E}\left[\mathrm{e}^{\lambda\{g_n(Z_1^n) - g_0\}}\right] = \mathbb{E}\left[\mathbb{E}\left[\mathrm{e}^{\lambda\{g_n(Z_1^n) - g_{n-1}(Z_1^{n-1})\}} \,\middle|\, Z_1^{n-1}\right] \mathrm{e}^{\lambda\{g_{n-1}(Z_1^{n-1}) - g_0\}}\right]$$
2. Montrer que pour tout $i \in \{1, \ldots, n\}$, $\inf_{z \in \mathsf{Z}} g_i(z_1^{i-1}, z) \leq g_i(z_1^i) \leq \inf_{z \in \mathsf{Z}} g_i(z_1^{i-1}, z) + c_i$ et que
$$g_{i-1}(Z_1^{i-1}) = \mathbb{E}\left[g_i(Z_1^i) \,\middle|\, Z_1^{i-1}\right] , \quad \mathbb{P} - \text{p.s.}$$
3. En déduire que
$$\mathbb{E}\left[\mathrm{e}^{\lambda\{g_n(Z_1^n) - g_{n-1}(Z_1^{n-1})\}} \,\middle|\, Z_1^{n-1}\right] \leq \mathrm{e}^{\lambda^2 c_n^2/8} .$$
4. Montrer que :
$$\mathbb{E}\left[\mathrm{e}^{\lambda(g(Z_1, \ldots, Z_n) - \mathbb{E}[g(Z_1, \ldots, Z_n)])}\right] \leq \mathrm{e}^{\frac{\lambda^2 \sum_{i=1}^n c_i^2}{8}} .$$
5. Prouver le théorème

Nous considérons

$$\Delta_{n,01}(\mathcal{C}) = \sup_{f \in \mathcal{C}} R(f) - \widehat{\mathrm{R}}_{n,01}(f) = \sup_{f \in \mathcal{C}} \left(\mathbb{E}\left[\ell_{01}(Y, f(X))\right] - \frac{1}{n} \sum_{i=1}^n \ell_{01}(Y_i, f(X_i))\right) .$$

6. Montrer que
$$\mathbb{P}\left(\Delta_n(\mathcal{C}) - \mathbb{E}\left[\Delta_n(\mathcal{C})\right] \leq \varepsilon\right) \geq 1 - \mathrm{e}^{-2n\varepsilon^2}$$

2 Convexification du risque

Exercice 4 (Lemme de Zhang, (cours)). On considère le problème de classification binaire, où $y \in \{\pm 1\}$. Nous considérons également la règle de prédiction suivante : prédire $y = 1$ si $g(x) \geq 0$, et prédire $y = -1$ sinon. La perte 0-1 de classification de $g(\cdot)$ en un point (x, y) est donnée par

$$\ell_{01}(g(x), y) = \begin{cases} 1, & \text{si } -g(x)y > 0, \\ 1, & \text{si } g(x) = 0 \text{ et } y = -1, \\ 0, & \text{sinon.} \end{cases}$$

Notre objectif est de trouver un prédicteur $g(x)$ de sorte à minimiser le risque 0-1 : $R_{01}(g) = \mathbb{E}[\ell_{01}(g(X), Y)]$. Étant donné un ensemble de données d'apprentissage $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$ i.i.d. la minimisation du risque empirique 0-1 consiste à trouver g dans une classe de fonctions $\mathcal{C}$ qui minimise

$$\widehat{\mathrm{R}}_{n,01}(g) = \frac{1}{n} \sum_{i=1}^{n} \ell_{01}\left(g\left(X_i\right), Y_i\right) \tag{4}$$

1. Pourquoi la minimisation du risque empirique $\widehat{\mathrm{R}}_{n,01}$ est elle problématique ?

Soit $\phi : \mathbb{R} \to \mathbb{R}_+$ une fonction convexe. On considère la minimisation dans une classe de fonctions $\mathcal{C}$ du risque empirique suivant

$$\widehat{\mathrm{R}}_{n,\phi}(g) = \frac{1}{n} \sum_{i=1}^{n} \phi\left(-g\left(X_i\right) Y_i\right),$$

qui est une approximation stochastique du risque $R_\phi(g) = \mathbb{E}[\phi(-g(X)Y)]$.

2. Montrer que
$$R_\phi(g) = \mathbb{E}[\eta(X)\phi(-g(X)) + (1 - \eta(X))\phi(g(X))]$$
où $\eta(X) = \mathbb{P}\left(Y = +1 \,|\, X\right)$.

On rappelle les exemples suivants :

Nom de la perte	Perte ϕ	Nom de la méthode
Moindres carrés	$\phi_2(v) = (1 + v)^2$	Moindres carrés.
Hinge	$\phi_H(v) = \max(1 + v, 0)$	SVM
Exponentielle	$\phi_E(v) = \exp(v)$	AdaBoost
Logistique	$\phi_L(v) = \log_2(1 + \exp(v))$	Régression Logistique

3. Quelles sont les propriétés cruciales de ces fonctions qui permettront de les optimiser ? Montrer que pour tous les exemples ci dessus (sauf ϕ_2),
$$\phi \text{ est croissante, convexe, et } \phi(v) \geq \mathbb{1}_{\{v \geq 0\}}. \tag{H1}$$

On s'intéresse dans la suite à une fonction ϕ générique satisfaisant (H1). Au regard de la question 2, nous considérons la fonction $H_\phi : [0, 1] \times \mathbb{R} \to \mathbb{R}$

$$H_\phi(\eta, p) = \eta\phi(-p) + (1 - \eta)\phi(p).$$

4. Montrer que $p \mapsto H_\phi(\eta, p)$ est convexe.
5. Montrer que pour $\eta \in]0;1[$, $p \mapsto H_\phi(\eta, p)$ admet un minimum unique dans $\mathbb{R}$, pour $\phi \in \{\phi_2, \phi_E, \phi_L\}$. Qu'en est il pour ϕ_H ? Qu'en est il pour $\eta \in \{0;1\}$?

Nous supposons dans la suite qu'il existe une fonction $p^*_\phi(\eta) : [0,1] \to \bar{\mathbb{R}}$ défini par [1]

$$p^*_\phi(\eta) = \arg\min_{p \in \bar{\mathbb{R}}} H_\phi(\eta, p) \ , \quad g^*_\phi(x) = p^*_\phi \circ \eta(x)$$

et nous notons

$$H^*_\phi(\eta) = \inf_{p \in \mathbb{R}} H_\phi(\eta, p) = H_\phi\left(\eta, p^*_\phi(\eta)\right) \ ,$$
$$\Delta H_\phi(\eta, p) = H_\phi(\eta, p) - H_\phi\left(\eta, p^*_\phi(\eta)\right) = H_\phi(\eta, p) - H^*_\phi(\eta).$$

Par symétrie, nous avons $H_\phi(\eta, p) = H_\phi(1-\eta, -p)$. Lorsque que $p^*_\phi(\eta)$ n'est pas déterminé de manière unique, on choisit un minimiseur quelconque mais on suppose dans la suite que p^*_ϕ est choisi de telle sorte que $p^*_\phi(1-\eta) = -p^*_\phi(\eta)$. En particulier, elle implique que $p^*_\phi(1/2) = 0$.

Le prédicteur bayésien pour la perte ϕ est le prédicteur bayésien pour la perte 01

6. Montrer que : $g^*_\phi \in \arg\min_{g \in \mathbb{R}^\mathcal{X}} R_\phi(g)$ et pour tout g :

$$\Delta R_\phi(g) = R_\phi(g) - R_\phi\left(g^*_\phi\right) = \mathbb{E}[\Delta H_\phi(\eta(X), g(X))]$$

7. Montrer que :

Nom de la perte	p^*_ϕ	H^*_ϕ
Moindres carrés	$p^*_\phi(\eta) = 2\eta - 1$	$H^*_\phi(\eta) = 4\eta(1-\eta)$.
Hinge	$p^*_\phi(\eta) = \mathrm{sign}(2\eta - 1)$	$H^*_\phi(\eta) = 1 - \lvert 2\eta - 1 \rvert$.
Logistique	$p^*_\phi(\eta) = \log_2 \frac{\eta}{1-\eta}$	$H^*_\phi(\eta) = -\eta \log_2 \eta - (1-\eta)\log_2(1-\eta)$.

8. Montrer que pour $\phi \in \{\phi_2, \phi_H, \phi_E, \phi_L\}$, la règle de décision signe(g^*_ϕ) est un classifieur de Bayes pour la perte $0-1$.
9. Supposons que ϕ est croissante, convexe et différentiable. Montrer que signe(g^*_ϕ) est la règle de décision bayésienne pour la perte 0-1.

Controle de l'excès de risque pour la perte 01 par l'excès de risque pour la perte ϕ.

Supposons que $p^*_\phi(\eta) > 0$ lorsque $\eta > 1/2$ et qu'il existe $c > 0$ et $s \geq 1$ tels que pour toute $\eta \in [0,1]$,

$$|1/2 - \eta|^s \leq c^s \Delta H_\phi(\eta, 0). \tag{H2}$$

L'objectif de cet partie est d'établir que pour toute fonction mesurable $g(x)$.

$$R_{01}(g) - R^*_{01} \leq 2c \Delta R_\phi(g)^{1/s},$$

où R^* est l'erreur de Bayes optimale pour la perte $0-1$.

1. Soit $\bar{\mathbb{R}}$ la droite réelle étendue $\left(\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}\right)$. Nous étendons une fonction convexe $\psi : \bar{\mathbb{R}} \to \bar{\mathbb{R}}$ en définissant $\psi(\infty) = \lim_{x\to\infty} \psi(x)$ et $\psi(-\infty) = \lim_{x\to-\infty} \psi(x)$. Elle garantit que le minimiseur optimal $p^*_\phi(\eta)$ donné ci-dessous est bien défini à $\eta = 0$ ou 1 pour certaines fonctions de perte.

10. Montrer que

$$R_{01}(g) - R_{01}^* \leq 2c \left(\mathbb{E}[\mathbb{1}_{\{(2\eta(X)-1)g(X)\leq 0\}} \Delta H_\phi(\eta(X), 0)] \right)^{1/s}.$$

11. Montrer que

$$(2\eta(x) - 1)g(x) < 0 \Rightarrow \Delta H_\phi(\eta(x), 0) \leq \Delta H_\phi(\eta(x), g(x))$$

12. Montrer que la condition (H2) est vérifiée avec $c = \frac{1}{2}, s = 2$. pour le critère quadratique.
13. Montrer que la condition (H2) est vérifiée avec $c = \frac{1}{2}, s = 1$ pour la perte hinge.
14. Montrer que la condition (H2) est vérifiée avec $c = \sqrt{\ln 2/2}, s = 2$ pour la perte logistique.