Exercices pour le cours d'Apprentissage et Grande Dimension en statistique

Laëtitia Comminges et Marc Hoffmann

Janvier 2015

Table des matières

1 Modèles statistiques en grande dimension.

1.1 Comparaison d'information pour des vecteurs gaussiens

Soit $M \geq 1$ et soient $0 < \sigma_1 < \sigma_2$. On considère

$$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\theta}, \sigma_1^2 \, \mathrm{Id}_{\mathbb{R}^M}) \quad \text{et} \quad \boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\theta}, \sigma_2^2 \, \mathrm{Id}_{\mathbb{R}^M})$$

deux vecteurs gaussiens sur $\mathbb{R}^M$ de même moyenne $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_M)$ et de matrice de variance-covariance $\sigma_1^2 \mathrm{Id}_{\mathbb{R}^M}$ et $\sigma_2^2 \mathrm{Id}_{\mathbb{R}^M}$ respectivement. On note $\mathcal{E}$ l'expérience statistique engendrée par $\boldsymbol{X}$ et $\mathcal{F}$ l'expérience statistique engendrée par $\boldsymbol{Y}$. Les paramètres σ_1 et σ_2 sont connus.

1. Calculer la matrice d'information de Fisher des expériences $\mathcal{E}$ et $\mathcal{F}$.
2. Montrer que pour le risque quadratique, il existe un estimateur sans biais dans $\mathcal{E}$ plus performant que tous les estimateurs sans biais dans $\mathcal{F}$.
3. Montrer que ce résultat est valable pour une perte quelconque *.

On dit que l'expérience gaussienne $\mathcal{F}$ est moins informative au sens de la variance que l'expérience $\mathcal{E}$. En particulier, toute précision pour une perte donnée dans $\mathcal{F}$ est aussi atteignable pour cette perte dans $\mathcal{E}$.

1.2 Approximation gaussienne du modèle de densité

Soit $\boldsymbol{X}^n = (X_1, \ldots, X_n)$ un n-échantillon de variables aléatoires réelles, de densité f par rapport à la mesure de Lebesgue, continue, bornée et de carré intégrable. Soit $\varphi : \mathbb{R} \to \mathbb{R}$ une fonction de carré intégrable telle que $\int_{\mathbb{R}} \varphi(x)^2 dx = 1$. On pose

$$\theta(\varphi)_n = n^{-1} \sum_{i=1}^{n} \varphi(X_i) \;\text{ et }\; \theta(\varphi) = \int_{\mathbb{R}} f(x)\varphi(x)dx.$$

1. Montrer que
$$\theta_n(\varphi) = \theta(\varphi) + n^{-1/2}\varepsilon_n(\varphi, f),$$
où $\varepsilon_n(\varphi) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$ pour une variance σ^2 que l'on déterminera.
2. Montrer que $\sigma^2 \leq \|f\|_{L^\infty}$.

*. Autrement dit, si $\ell : \mathbb{R}^M \times \mathbb{R}^M \to [0, \infty)$ est une fonction de perte et si $\widehat{\boldsymbol{\theta}}(\boldsymbol{Y})$ est un estimateur dans $\mathcal{F}$ tel que
$$\mathbb{E}\big[\ell(\widehat{\boldsymbol{\theta}}(\boldsymbol{Y}), \boldsymbol{\theta})\big] \leq \delta$$
pour un $\delta > 0$, alors on peut construire un estimateur dans $\mathcal{E}$ ayant la même propriété, quitte à se placer sur un espace de probabilité élargi.

3. Soit $M \geq 1$ et $\{\varphi_1, \ldots, \varphi_M\}$ des fonctions intégrables et orthonormales dans $L^2(\mathbb{R})$. On considère l'ensemble des paramètres

$$\mathcal{P}_M = \Big\{ f : \mathbb{R} \to [0,\infty), f(x) = \sum_{k=1}^{M} \theta_k \varphi_k(x),\ |\theta_k| \leq C, \int_{\mathbb{R}} f(x)dx = 1 \Big\}.$$

Montrer que la suite de vecteurs

$$\boldsymbol{Y}^n = \sqrt{n}\big(\theta(\varphi_1) - \theta_n(\varphi_1), \ldots, (\theta(\varphi_M) - \theta_n(\varphi_M)\big)$$

converge en loi vers un vecteur $\boldsymbol{Y}$ gaussien de moyenne 0 et de matrice de variance-covariance que l'on déterminera. En déduire que l'expérience $\mathcal{E}$ engendrée par le vecteur limite est dominée au sens de la variance par une expérience de suite gaussienne $\mathcal{F}$ pour l'espace des paramètres $\mathcal{P}_M$.

En conclusion, à partir d'un n-échantillon $\boldsymbol{X}^n$ de densité $f \in \mathcal{P}_M$, on construit une (suite de) sous-expériences engendrées par $\boldsymbol{Y}^n$. Les $\boldsymbol{Y}^n$ convergent en loi vers $\boldsymbol{Y}$ qui engendre une expérience plus informative qu'une expérience de suite gaussienne. On peut donc s'attendre à ce que les performances d'un estimateur dans le modèle de densité soient atteignables dans le modèle de suite gaussienne †.

1.3 Approximation gaussienne du modèle de régression : design aléatoire

Soit $\mu : [0,1] \to [0,\infty)$ une densité continue. On définit

$$L^2(\mu) = \{g : [0,1] \to \mathbb{R}, \|g\|^2_{L^2(\mu)} = \int_0^1 g(x)^2 \mu(x)dx < \infty\}.$$

On observe un n-échantillon (X_i, Y_i), avec

$$Y_i = f(X_i) + \xi_i,$$

où $f : [0,1] \to \mathbb{R}$ est une fonction de $L^2(\mu)$ bornée et les X_i sont distribuées suivant la densité μ et indépendantes des variables aléatoires gaussiennes standard ξ_i. On pose

$$\theta(\varphi)_n = n^{-1} \sum_{i=1}^{n} Y_i \varphi(X_i) \ \text{ et } \ \theta(\varphi) = \int_0^1 f(x)\mu(x)\varphi(x)dx.$$

†. Pour rendre cette heuristique rigoureuse, il faudrait montrer la convergence de vecteur $\boldsymbol{Y}^n$ dans un sens plus fort que la convergence en loi (en variation totale) et se restreindre à des fonctions de pertes particulières. Il faudrait aussi retravailler l'Exercice 1.1 en considérant le cas où σ_1 peut dépendre de $\boldsymbol{\theta}$, ce qui rend plus difficile en particulier la question 3 de l'Exercice 1.1.

1. Montrer que
$$\theta_n(\varphi) = \theta(\varphi) + n^{-1/2}\varepsilon_n(\varphi),$$
où $\mathcal{L}\big(\varepsilon_n(\varphi)\,\big|X_1,\ldots,X_n\big) \overset{d}{=} \mathcal{N}(M_n(\varphi),\Sigma_n(\varphi)^2)$ où $M_n(\varphi)$ et $\Sigma_n^2(\varphi)$ sont deux variables aléatoires que l'on déterminera.
2. On suppose que $\|\varphi\|_{L^2(\mu)} = 1$. Montrer que $\varepsilon_n(\varphi) \overset{d}{\to} \mathcal{N}(0,\sigma(f,\varphi)^2)$, où $\sigma(f,\varphi)^2 \le 1 + \|f\|^2_{L^\infty(\mu)}$.

1.4 Approximation gaussienne du modèle de régression : design déterministe

Pour $n \ge 1$, on définit la subdivision
$$0 = x_{0,n} < x_{1,n} < x_{2,n} < \ldots < x_{n,n} = 1$$
On suppose qu'il existe une fonction $g : [0,1] \to [0,\infty)$ continûment différentiable et non-nulle de sorte que pour tout $n \ge 1$
$$\int_{x_{i-1,n}}^{x_{i,n}} g(s)ds = \frac{1}{n}, \;\; i = 1,\ldots,n.$$
On observe
$$y_{i,n} = f(x_{i,n}) + \xi_{i,n}, \;\; i = 1,\ldots,n$$
où $f : [0,1] \to \mathbb{R}$ est une fonction continue et les $\xi_{i,n}$ des gaussiennes standard indépendantes. Pour $\varphi : [0,1] \to \mathbb{R}$ continue, on pose
$$\theta_n(\varphi) = \sum_{i=1}^n y_{i,n}\varphi(x_{i,n})\Delta x_{i,n} \;\text{ et }\; \theta(\varphi) = \int_0^1 \varphi(x)f(x)dx.$$
Soit $\alpha > 0$. On dit qu'une fonction $f : [0,1] \to \mathbb{R}$ est α-hölderienne si $|f(x) - f(y)| \le C|x-y|^\alpha$ pour tous $x, y \in [0,1]$ et une constante $C \ge 0$.

1. Montrer que si $\alpha > 1$, alors f est constante.
2. Montrer que si f et g sont α-hölderiennes, alors le produit fg est encore α-hölderien.
3. On suppose φ et f α-hölderiennes et que $\|\frac{\varphi}{\sqrt{g}}\|_{L^2} = 1$. Montrer que
$$\theta_n(\varphi) = \theta(\varphi) + n^{-1/2}\varepsilon_n(\varphi),$$
où $\varepsilon_n(\varphi)$ est une variable gaussienne dont on précisera la moyenne et la variance. Montrer que si $\alpha > 1/2$, alors $\varepsilon_n(\varphi) \overset{d}{\to} \mathcal{N}(0,1)$.
4. On suppose désormais cette condition vérifiée et $g(x) = 1$ pour tout $x \in [0,1]$. Pour un entier $M \ge 1$, on pose
$$\varphi_k(x) = M^{1/2}\mathbf{1}_{[(k-1)M^{-1},kM^{-1}[}(x), \;\; k = 1,\ldots,M.$$
Pour $M \ge 1$, montrer que la suite de vecteurs $\big(\varepsilon_n(\varphi_1),\ldots,\varepsilon_n(\varphi_M)\big)$ converge en loi vers un vecteur gaussien que l'on caractérisera.

5. Montrer que si f est α-hölderienne, alors

$$\|f - \sum_{k=1}^{M_n} \langle f, \varphi_k \rangle \varphi_k \|_{L^2([0,1])} \leq C(f) M_n^{-\alpha}$$

où $\langle f, g \rangle = \int_0^1 f(x)g(x)dx$ désigne le produit scalaire de $L^2([0,1])$.

6. On considère l'estimateur

$$\widehat{f}_{n,M}(x) = \sum_{k=1}^{M_n} \theta_n(\varphi_k)\varphi_k(x).$$

Etudier

$$\mathbb{E}[\|\widehat{f}_{n,M} - f\|_{L^2}^2].$$

Comment choisir M de façon optimale ?

2 Estimateurs linéaires et régression

2.1 EMC non-paramétrique et design uniforme

Soit $\{\varphi_k\}_{k\geq 1}$ la base trigonométrique de $L^2([0,1])$ définie par $\varphi_1(x)=1$, $\varphi_{2k}(x)=\sqrt{2}\cos(2k\pi x)$ et $\varphi_{2k+1}(x)=\sqrt{2}\sin(2k\pi x)$. Soit f dans la boule de Sobolev $W(\beta,L)$, $\beta\geq 1$, $L>0$ (périodique).

1. Montrer que
$$n^{-1}\sum_{i=1}^{n}\varphi_j(i/n)\varphi_k(i/n)=\mathbf{1}_{\{j=k\}},\ \ 1\leq j,k\leq n-1.$$
2. Montrer que la série
$$\sum_{j\geq 1}\theta_j\varphi_j(x)$$
converge ponctuellement vers $f(x)$, où l'on a posé $\theta_j=\langle f,\varphi_j\rangle$.
3. Soit $M\geq 1$. Montrer que si $\|g\|_n^2=n^{-1}\sum_{i=1}^n g(i/n)^2$, alors ‡
$$\|\sum_{j=M+1}^{\infty}\theta_j\varphi_j\|_n^2\leq 2\sum_{j=M+1}^{n-1}\theta_j^2+4\big(\sum_{j\geq n}|\theta_j|\big)^2.$$
4. En déduire
$$\inf_{\theta\in\mathbb{R}^M}\|f_\theta-f\|_n^2\leq C(\beta,L)\big(M^{-2\beta}\mathbf{1}_{\{M\leq n-1\}}+n^{-1}\big)$$
où $f_\theta=\sum_{j=1}^M\theta_j\varphi_j$ si $\theta=(\theta_1,\dots,\theta_M)\in\mathbb{R}^M$.
5. Montrer alors rigoureusement que pour un choix de M judicieux, l'estimateur des moindres carrés non-paramétrique atteint la vitesse $n^{-\beta/(2\beta+1)}$ sur la classe $W(\beta,L)$ pour la perte $\|\cdot\|_n$ lorsque les points du design X_i sont donnés par $X_i=i/n$.

2.2 Estimation sans biais du risque

Avec les notations et les hypothèses du cours, on considère le modèle de régression donné par l'observation d'un n-échantillon
$$Y_i=f(X_i)+\xi_i,\dots,i=1,\dots,n$$
où les ξ_i sont centrés et de variance σ^2 que l'on réécrit sous la forme vectorielle et les X_i sont considérés comme déterministes §
$$\boldsymbol{y}=\boldsymbol{f}+\boldsymbol{\xi}.$$

‡. avec la convention $\sum_p^q a_i=0$ si $p\geq q$.

§. ou encore, dans la suite, les espérances sont des espérances conditionellement à $(X_1,\dots,X_n)$.

Soit $\widehat{f}_M = S_M \boldsymbol{y}$ l'estimateur des moindres carrés non-paramétrique associé à un dictionnaire $\{\varphi_1, \ldots, \varphi_M\}$, de matrice de lissage S_M, où M est la taille du dictionnaire employé pour reconstruire f via la perte empirique $\|g\|_n^2 = n^{-1}\sum_{i=1}^n g(X_i)^2$.

1. Montrer que la valeur M qui minimise $\mathbb{E}\big[\|\widehat{f}_M - f\|_n^2\big]$ est aussi celle qui minimise
$$M \rightsquigarrow \mathbb{E}[\|S_M\boldsymbol{y}\|_n^2] - \frac{2}{n}\boldsymbol{f}^T\mathbb{E}[S_M\boldsymbol{y}].$$
2. Montrer que si S est une matrice de lissage quelconque, alors
$$\widehat{J}(S) = \|S\boldsymbol{y}\|_n^2 - \frac{2}{n}\big(\boldsymbol{y}^T S\boldsymbol{y} - \sigma^2\mathrm{Tr}(S)\big)$$
est un estimateur sans biais de
$$\mathbb{E}[\|S\boldsymbol{y}\|_n^2] - \frac{2}{n}\boldsymbol{f}^T\mathbb{E}[S\boldsymbol{y}].$$
3. Montrer que pour l'estimateur des moindres carrés non-paramétrique, minimiser $M \rightsquigarrow \widehat{J}(S_M)$ est équivalent à minimiser
$$M \rightsquigarrow \|\boldsymbol{y} - S_M\boldsymbol{y}\|_n^2 + \frac{2\sigma^2\mathrm{Tr}(S)}{n}$$
C'est le critère C_p de Mallows.

On suppose désormais que le dictionnaire $\{\varphi_1, \ldots, \varphi_M\}$ est constitué par les premiers éléments de la base trigonométrique définie dans l'Exercice 2.1 et que le design est donné ¶ par $X_i = i/n$.

4. Montrer que si $M \leq n-1$, l'estimateur des moindres carrés non-paramétrique $\widehat{f}_M^{MC}(x) = \sum_{j=1}^M \widehat{\theta}_j^{MC}\varphi_j(x)$ est donné ‖ par
$$\widehat{\theta}_j^{MC} = n^{-1}\sum_{i=1}^n Y_i\varphi_j(i/n).$$

On admettra le théorème suivant du à Kneipp (1994), et appelé *Inégalité Oracle* : Si les ξ_i sont des gaussiennes centrées de variance σ^2 et si $f^\star$ est l'estimateur obtenu en minimisant le critère C_p de Mallows, alors, il existe $K > 0$ telle que pour tout f et tout $\varepsilon > 0$, on a
$$\mathbb{E}[\|f^\star - f\|_n^2] \leq (1+\varepsilon)\min_{M=1,\ldots,n-1}\mathbb{E}[\|\widehat{f}_M^{MC} - f\|_n^2] + \frac{K}{\varepsilon n}.$$

5. En déduire que si $f \in W(\beta, L)$, alors $\mathbb{E}[\|f^\star - f\|_n^2] \leq C(\beta, L)n^{-2\beta/(2\beta+1)}$.
6. Quelle propriété remarquable cet estimateur possède-t-il par rapport à l'estimateur du cours ?

¶. comprendre que l'on travaille conditionnellement aux variables explicatives $X_i = i/n$, $i = 1, \ldots, n$.

‖. Indication : utiliser la Question 1 de l'Exercice 2.1 et la représentation de l'EMC non-paramétrique comme estimateur linéaire.

3 Parcimonie et régression non-paramétrique

3.1 Lemme de concentration gaussienne

Soit $M \geq 2$ un entier et soient $\eta_1, \ldots, \eta_M$ des variables aléatoires gaussiennes standard.

1. Montrer que pour tout $t > 0$, on a
$$\mathbb{P}\big(\max_{1\leq j\leq M} |\eta_j| \geq t\sqrt{\log M}\big) \leq \frac{2M}{\sqrt{2\pi}} \int_{t\sqrt{\log M}}^{\infty} e^{-u^2/2} du.$$
2. Montrer que pour tout $x > 0$, on a
$$\int_x^{\infty} e^{-u^2/2} du \leq \frac{1}{x} e^{-x^2/2}.$$
3. En appliquant cette dernière inégalité au point $x = t\sqrt{\log M}$, en déduire
$$\mathbb{P}\big(\max_{1\leq j\leq M} |\eta_j| \geq t\sqrt{\log M}\big) \leq M^{1-t^2/2}.$$
4. Montrer que pour tout $x > 0$, on a
$$\mathbb{E}\big(\eta^2 \mathbf{1}_{|\eta|>x}\big) \leq \sqrt{\frac{2}{\pi}}\big(x + \frac{2}{x}\big) e^{-x^2/2}.$$

3.2 Seuillage dans le modèle de suite gaussienne

On se place dans le modèle de suite gaussienne
$$y_j = \theta_j^\star + \frac{\sigma}{\sqrt{n}}\eta_j, \quad j = 1, \ldots, M,$$
où les η_j sont des gaussiennes standard. On note
$$\widehat{\boldsymbol{\theta}}^H = (y_j \mathbf{1}_{|y_j|\geq\tau}, j = 1, \ldots, M)$$
l'estimateur par "hard-thresholding" de $\boldsymbol{\theta}^\star$ et
$$\widehat{\boldsymbol{\theta}}^S = \Big(y_j\big(1 - \frac{\tau}{|y_j|}\big)_+, j = 1, \ldots, M\Big)$$
l'estimateur par "soft-thresholding".

1. Montrer (à l'aide du cours) que pour un choix judicieux de τ (dépendant de n, M et σ), on a
$$\mathbb{E}\big[\|\widehat{\boldsymbol{\theta}}^H - \boldsymbol{\theta}^\star\|^2\big] \leq C\big(1 + M(\boldsymbol{\theta}^\star)\big)\frac{\log M}{n}$$
où $M(\boldsymbol{\theta}^\star)$ désigne le nombre de composantes non-nulles pour $\boldsymbol{\theta}^\star$.

2. Montrer que le même résultat est valable pour l'estimateur par soft-thresholding.
3. Montrer que $\widehat{\boldsymbol{\theta}}^H$ est solution du problème variationnel

$$\widehat{\boldsymbol{\theta}}^H = \mathrm{Argmin}_{\boldsymbol{\theta}\in\mathbb{R}^M}\big(\sum_{i=1}^M (y_j-\theta_j)^2 + \tau^2 M(\boldsymbol{\theta})\big).$$

4. Montrer que $\widehat{\boldsymbol{\theta}}^S$ est solution du problème variationnel

$$\widehat{\boldsymbol{\theta}}^H = \mathrm{Argmin}_{\theta\in\mathbb{R}^M}\big(\sum_{i=1}^M (y_j-\theta_j)^2 + 2\tau\|\boldsymbol{\theta}\|_{\ell^1}\big),$$

où $\|\boldsymbol{\theta}\|_{\ell^1} = \sum_{i=1}^M |\theta_j|$.

3.3 Méthode LASSO

On se place dans le modèle de régression

$$\boldsymbol{y} = X\boldsymbol{\theta}^\star + \sigma\boldsymbol{\xi}$$

où $\boldsymbol{y}\in\mathbb{R}^n$ et $\boldsymbol{\xi}$ est un vecteur gaussien centré de matrice de variance-covariance l'identité (avec les notations du cours). On suppose l'hypothèse ORT sur la matrice de design, c'est-à-dire $n^{-1}X^TX = \mathrm{Id}_M$.

1. Montrer que

$$\sum_{j=1}^M (y_j-\theta_j)^2 = \|\boldsymbol{y}-X\boldsymbol{\theta}\|_n^2 + C$$

où C est une constante qui ne dépend pas de $\boldsymbol{\theta}$. En déduire une autre formulation variationnelle pour les estimateurs de hard-thresholding et de soft-thresholding. L'estimateur associé dans le cas du soft-thresholding est appelé estimateur LASSO.

2. On suppose que la matrice de design $X = (X_{ij})_{1\le i\le,n,1\le j\le M}$ s'écrit $X_{ij} = \varphi_j(X_i)$ pour un dictionnaire $\{\varphi_1,\dots,\varphi_M\}$ vérifiant la propriété dite de cohérence
(i) $\|\varphi_j\|_{L^2} = 1,\ \ j=1,\dots,M.$
(ii)

$$\max_{1\le k,j\le M, k\ne j}\Big|n^{-1}\sum_{i=1}^n \varphi_j(X_i)\varphi_k(X_i)\Big| \le \frac{1}{7s},\ \text{ pour un entier } s>1.$$

Alors on a l'inégalité oracle suivante (admise) : pour $\widehat{\boldsymbol{\theta}}^L$ l'estimateur LASSO

$$\|f_{\widehat{\boldsymbol{\theta}}^L} - f\|_n^2 \le \min_{\boldsymbol{\theta}, M(\boldsymbol{\theta})\le s}\Big\{3\|f_{\boldsymbol{\theta}}-f\|_n^2 + CA^2\sigma^2\frac{M(\boldsymbol{\theta})\log M}{n}\Big\}$$

avec probabilité au moins $1 - M^{1-A^2/8}$, pour le choix $\tau = A\sigma\sqrt{\frac{\log M}{n}}$ avec $A > 2\sqrt{2}$ et $C > 0$ est une constante absolue.
En déduire qu'avec probabilité supérieure ou égale à $1 - M^{1-A^2/8}$, on a

$$\left\|X\left(\widehat{\boldsymbol{\theta}}^L - \boldsymbol{\theta}^\star\right)\right\|_n^2 \leq C\frac{M(\boldsymbol{\theta}^\star)\log M}{n}.$$

3. Si $\{\varphi_j, 1, \ldots, M\}$ sont les premiers éléments de la base trigonométrique et les $X_i = i/n$, et $M = n - 1$, montrer que pour le choix de $\tau = A\sigma\sqrt{\log M/n}$ avec $A > 2\sqrt{2}$, on a

$$\sup_{f \in W(\beta,L)} \mathbb{P}\Big(\|f_{\widehat{\boldsymbol{\theta}}^L} - f\|_n \leq C(\log n/n)^{\beta/(2\beta+1)}\Big) \geq 1 - (n-1)^{1-A^2/8},$$

où $C = C(\sigma, \beta, L)$ et $W(\beta, L)$ désigne la boule de l'espace de Sobolev (périodique) avec $\beta \geq 1$.

4 Simulations numériques

Préliminaires méthodologiques

Dans toute la suite, on trouvera souvent la question "Etudier les performances de l'estimateur en fonction de n" et d'un (ou plusieurs) autre(s) paramètre(s) noté(s) ici μ. Cela signifie que l'on sait simuler une variable aléatoire de la forme $Z_n^{(\mu)}$, où par exemple

$$Z_n^{(\mu)} = \|\widehat{\theta}_{n,\mu} - \theta\|^2$$

et θ_n est un estimateur de θ dépendant de μ (μ peut être égal à la taille du dictionaire, le niveau de variance à travers les observations que l'on a injectées dans l'estimateur, la régularité de la fonction sous-jacente inconnue, etc.). Quand la théorie nous dit que

$$\mathbb{E}[Z_n^{(\mu)}] \leq C n^{-p(\mu)}, \qquad (\star),$$

nous proposons le protocole suivant pour mettre en évidence ce résultat :

- On se fixe un nombre de répliques de Monte-Carlo K (par exemple $K = 30$).
- Pour $\ell = 1, \ldots, K$, on simule la perte de l'estimateur pour μ fixé (choisi), c'est-à-dire $Z_{n,\ell}^{(\mu)}$ et pour différentes valeurs de n (par exemple $n = 10^k$, $k = 1, \ldots 6$ ou 7 voire plus).
- On construit

$$n \rightsquigarrow K^{-1} \sum_{\ell=1}^{K} Z_{n,\ell}^{(\mu)}, \quad \text{pour} \quad n = 10^k, k = 1, 2, \ldots$$

  qui approche $\mathbb{E}[Z_n^{(\mu)}]$ par la loi des grands nombres, c'est la méthode de Monte-Carlo.
- On représente graphiquement $\log K^{-1} \sum_{\ell=1}^{K} Z_{n,\ell}^{(\mu)}$ en fonction de $\log n$ (ce que l'on appelle un log-log plot).
- On trace la droite des moindres carrés pour le nuage de points

$$(\log n, K^{-1} \sum_{\ell=1}^{K} Z_{n,\ell}^{(\mu)}), n = 10^k, k = 1, \ldots, 2...$$

  La pente (négative) obtenue est une estimation de $-p(\mu)$.
- On peut répéter l'expérience pour plusieurs valeurs de μ.

4.1 Approximation gaussienne dans le modèle de la densité

1. Simuler un n-échantillon $\varepsilon_1, \ldots, \varepsilon_n$ de loi normale de variance $\sigma^2 = 0.1$.

2. Soit $f(x) = \min\{x, 1-x\}$ pour $x \in [0,1]$. Soit $\{\varphi_1, \ldots, \varphi_M\}$ un dictionnaire (on choisira par exemple la base des fonctions constantes par morceaux définies par

$$\varphi_k(x) = M^{1/2}\mathbf{1}_{[(k-1)M^{-1}, kM^{-1})}(x), \quad k = 1, \ldots, M.$$

Simuler le modèle d'observation (X_i, Y_i), $i = 1, \ldots, n$ où

$$Y_i = f(X_i) + \xi_i,$$

avec $X_i = i/n$. Représenter la fonction f et les données sur un même graphique (on pourra choisir n de l'ordre de 100 ou 1000 par exemple).

3. Avec les notations de l'Exercice 1.4, construire le vecteur

$$U_n = \sqrt{n}\big(\theta_n(\varphi_1) - \theta(\varphi_1), \ldots, \theta_n(\varphi_M) - \theta(\varphi_M)\big)$$

Montrer que U_n converge en loi vers un vecteur gaussien centré.

4. Visualiser l'approximation gaussienne à l'aide d'un histogramme des fréquence et/ou un QQ-plot pour les données $\{\sqrt{n}(\theta_n(\varphi_k) - \theta(\varphi_k)), k = 1, \ldots, M\}$. (Voir pour cela les fonctions `hist()` et `qqplot()` en R. On pourra imposer un nombre de classes pour l'histogramme.)

5. A l'aide de la statistique $\|U_n\|^2$ construire un test d'adéquation à la loi normale centrée (on calculera la p-valeur du test de Wald). On pourra essayer la validité de ce test avec d'autres lois de bruit (uniforme, Laplace, Cauchy).

6. Construire et représenter graphiquement l'estimateur

$$\widehat{f}_{n,M}(x) = \sum_{k=1}^{M} \theta_n(\varphi_k)\varphi_k(x)$$

pour différentes valeurs de M (par exemple M variant entre 10 et 50 pour n de l'ordre de grandeur 100 ou 1000).

7. Représenter (par approximation de Monte-Carlo) la fonction d'erreur $M \mapsto \mathbb{E}[\|\widehat{f}_{n,M} - f\|_n^2]$ et retrouver le phénomène d'équilibre biais-variance en jouant sur les paramètres n et M (avec les spécifications précédentes).

8. Reprendre cette étude pour d'autres choix ** de fonctions f, par exemple $f(x) = \sin(10x)$.

9. Reprendre l'étude pour des X_i indépendants et uniformes sur $[0,1]$.

**. On pourra aussi par exemple en choisir l'un des quatre signaux cibles (benchmark) Blocks, Bumps, Heavisine et Doppler de Donoho et Johnstone téléchargeables à l'adresse

`ftp://ftp.sas.com/pub/neural/data/`dojo_`test.txt 286K, 4097 cases`

4.2 Estimation de la variance dans un modèle de régression

On considère le modèle de régression

$$Y_i = f(X_i) + \xi_i, \quad i = 1, \dots, n$$

où les ξ_i sont i.i.d. de variance commune σ^2 et $X_i = i/n$. On ne suppose pas σ^2 connu. Lorsque la fonction f (le paramètre inconnu) est lisse, l'estimateur de Rice défini par

$$\widehat{\sigma}_n^2 = \frac{1}{2(n-1)} \sum_{i=2}^{n} (Y_i - Y_{i-1})^2$$

est un candidat naturel pour l'estimation de σ^2.

1. Etudier les propriétés asymptotiques de $\widehat{\sigma}_n^2$ lorsque f est constante (mais inconnue). On admettra que $\widehat{\sigma}_n^2$ converge lorsque f satisfait une propriété de type hölderienne.
2. Simuler le modèle d'observation (X_i, Y_i), $i = 1, \leq n$ pour des bruits ξ_i gaussiens standard, avec $\sigma^2 = 1$ et

   $$f(x) = \max\{x, (1-x)\}^\beta$$

   pour différentes valeurs de $\beta > 0$.
3. Etudier en fonction de n et de β la vitesse de convergence de l'estimateur de Rice.

4.3 Seuillage et détection du nombre de variables significatives

On se place dans le modèle séquentiel gaussien :

$$y_j = \theta_j^\star + \frac{\sigma}{\sqrt{n}} \eta_j, \quad j = 1, \dots, M$$

où $\theta_j^\star$ est la j-ième coordonnée de $\boldsymbol{\theta}^\star = (\theta_1, \dots, \theta_M)$. On note

$$\widehat{J}_\tau = \{j, \widehat{\theta}_j^H(\tau) \neq 0\}$$

l'estimateur de l'ensemble révélateur des coefficients non-nuls de $\theta^\star$. où $\widehat{\theta}_j^H(\tau) = y_j \mathbf{1}_{\{|y_j| \geq \tau\}}$ est l'estimateur de hard-thresholding au seuil $\tau > 0$.

On simule un signal $\theta^\star$ de la manière suivante : on se donne une suite i.i.d. de variables aléatoires uniformes U_j sur $[-1, 1]$ et une suite i.i.d. de variables aléatoires de Bernoulli κ_j de paramètre $1 - p$ indépendantes des U_i. On pose

$$\theta_j^\star = \kappa_j U_j, \quad j = 1, \dots, M.$$

1. Montrer que $\mathbb{E}[M(\boldsymbol{\theta}^\star)] = pM$.

2. Mettre en oeuvre l'estimateur par seuillage dur et étudier ses performances en fonction de M, n et p pour un choix de τ que l'on se donnera en fonction des résultats du cours lorsque l'on se donne un niveau de confiance $1-\alpha$ pour obtenir la bonne estimation.
3. Etudier les performances de l'estimateur de l'ensemble révélateur des coefficients non-nuls de $\boldsymbol{\theta}^\star$ lorsque l'on se donne un niveau de confiance $1-\alpha$ pour obtenir la bonne estimation.

4.4 Reconstruction d'un signal : estimation sans biais du risque

On se place dans le cadre de l'Exercice 2.2. Télécharger les quatre signaux cibles (benchmark) Blocks, Bumps, Heavisine et Doppler de Donoho et Johnstone à l'adresse

`ftp://ftp.sas.com/pub/neural/data/`dojo_`test.txt 286K, 4097 cases`

1. Simuler le modèle d'observation de l'Exercice 2.2 pour $X_i = i/n$ dans le cadre des quatre signaux de Donoho & Johnstone, pour un bruit gaussien de variance σ^2 connu.
2. Calculer numériquement $\widehat{\theta}_j$ (on pourra utiliser l'algorithme FFT si on le connaît, ou bien faire le calcul linéaire direct).
3. Calculer numériquement l'estimateur de $f^\star$ obtenu en minimisant le critère C_p de Mallows pour les quatre signaux.
4. Représenter sur un même graphique le signal, le signal bruité et l'estimateur $f^\star$.
5. Que se passe-t-il si σ^2 est inconnu ? Reprendre les résultats précédents en implémentant l'estimateur de Rice pour la variance. Retrouve-t-on une sensibilité à la régularité de $\theta^\star$?

5 Classification

Notations : pour une fonction g pouvant dépendre de $\mathcal{D}_n$, la notation $L(g)$ correspond à la notation $R_{\mathcal{D}_n}(g)$ du cours et L_n désigne le risque empirique (avec la notation du cours $L_n = R_n$).

5.1 Classifieur bayésien

Soit (X, Y) un couple de loi $\mathbb{P}$ avec $Y \in \{0, 1\}$. Soit f^* le classifieur de Bayes, et L^* son risque de classification. On note également $p = \mathbb{P}(Y = 1)$.

1. Montrer que $L^* \leq \min\{p, 1-p\}$.
2. Montrer que si X et Y sont indépendants, $L^* = \min\{p, 1-p\}$.
3. Fabriquer un exemple où $L^* = \min\{p, 1-p\}$ et où X et Y ne sont pas indépendants.

5.2 Classification et régression

Etant donné l'apprentissage, $\mathcal{D}_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$, soit f une règle de classification définie par

$$f(x, \mathcal{D}_n) = \mathbf{1}_{\{\hat{\eta}(x, \mathcal{D}_n) > \frac{1}{2}\}}$$

où $\hat{\eta}(\cdot, \mathcal{D}_n)$ est un estimateur de la fonction de régression $\eta(\cdot) = \mathbb{E}[Y|X = \cdot]$. Montrer

$$L(f) - L^* \leq 2\mathbb{E}\left[|\hat{\eta}(X, \mathcal{D}_n) - \eta(X)|\right].$$

On suppose de plus que $L^* = 0$. Montrer que pour tout $p \geq 1$,

$$L(f) \leq 2^p \mathbb{E}\left[|\hat{\eta}(X, \mathcal{D}_n) - \eta(X)|^p\right].$$

5.3 Consistance universelle

Soit (X, Y) un couple de loi P avec $Y \in \{0, 1\}$ avec $X \in \{1, ..., p\}$ et $Y \in \{0, 1\}$. On note toujours f^* le classifieur de Bayes et L^* son risque de classification. On se propose d'étudier la règle de classification suivante, étant donné $\mathcal{D}_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$,

$$f(x, \mathcal{D}_n) = \mathbf{1}_{\{\hat{\eta}(x, \mathcal{D}_n) > \frac{1}{2}\}},$$

avec

$$\hat{\eta}(x, \mathcal{D}_n) = \frac{1}{\text{card}\{i, X_i = x\}} \sum_{i, X_i = x} Y_i$$

et $\hat{\eta}(x, \mathcal{D}_n) = 0$ si $\text{card}\{i, X_i = x\} = 0$. Montrer que f est universellement consistante, c'est-à-dire que : $E(R_{\mathcal{D}_n}(f)) \to R^*$ (ce qui est équivalent à la convergence en probabilité de $R_{\mathcal{D}_n}(f)$ vers R^*).

5.4 Classification et maximum de vraisemblance

On suppose que la loi P du couple (X, Y) est donnée de la manière suivante : $Y \sim \mathcal{B}(p)$ pour un paramètre $p \in]0,1[$ puis $X \in \mathbb{R}^d$ est donnée par sa loi conditionnelle sachant Y :

$$X|Y \sim \mathcal{N}(V_Y, \Sigma)$$

où Σ est une matrice définie positive et V_0, V_1 sont deux vecteurs distincts de $\mathbb{R}^d$.

1. Déterminer la fonction de régression

$$\eta(x) = E(Y|X = x).$$

2. En déduire la forme du classifieur de Bayes, f^*, et montrer que son risque de classification s'écrit

$$L^* = pP\big[Z > \delta/2 + \delta^{-1}\log(\frac{p}{1-p})\big] + (1-p)P\big[Z < -\delta/2 + \delta^{-1}\log(\frac{p}{1-p})\big]$$

où $\delta = \|\Sigma^{-1/2}(V_1 - V_0)\|$ et $Z \sim \mathcal{N}(0,1)$.

3. En pratique, on dispose d'un n-échantillon (X_i, Y_i), $1 \leq i \leq n$, de la loi P. On suppose que l'on est dans un cas où l'on connaît p et Σ et on se propose d'estimer V_0 et V_1 par maximum de vraisemblance. Donner la forme des estimateurs $\hat{V}_0$ et $\hat{V}_1$ ainsi obtenus.
4. En déduire un estimateur de la fonction de régression, $\hat{\eta}$ et une règle de classification, f.
5. Montrer que $L(f) \to L^*$ en probabilité, quand $n \to \infty$.
6. Montrer que f n'est pas universellement consistante. Il suffira de fabriquer une autre loi P' telle que si (X_i, Y_i) et (X, Y) sont iid de loi P', alors $L(f)$ ne converge pas vers L^*.

6 Minimisation du risque empirique

6.1 Dictionnaire fini et inégalité de Hoeffding

Soit $\mathcal{F} = \{f_1, ..., f_p\}$ une famille finie de classifieurs. Soit, avec la notation habituelle, $\mathcal{D}_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$ un famille de couples aléatoires i.i.d. de loi P. Soit L_n le risque empirique. Soit g la règle de minimisation du risque empirique.

1. En utilisant l'inégalité de Markov, montrer que pour tout $t > 0$, pour tout $j \in \{1, ..., p\}$,
$$P\left[L(f_j) - L_n(f_j) > t\right] \leq \frac{L(f_j)(1 - L(f_j))}{nt^2}.$$
2. En déduire que
$$P\left[L(g) - \inf_{1 \leq i \leq p} L(f_p) \leq \sqrt{\frac{p}{n\varepsilon}}\right] \geq 1 - \varepsilon$$
et
$$E\left[L(g) - \inf_{1 \leq i \leq p} L(f_i)\right] \leq 2\sqrt{p/n}.$$
3. Comparer avec le résultat obtenu en cours où l'on utilisait l'inégalité de Hoeffding.
4. Montrer que s'il existe un j tel que $L(f_j) = 0$, on a
$$P\left[L(g) - \inf_{1 \leq i \leq p} L(f_i) \leq \frac{1}{n}\log(p/\epsilon)\right] \geq 1 - \varepsilon.$$
Donner la majoration qui en découle pour $E\left[L(g) - \inf_{1 \leq i \leq p} L(f_i)\right]$.

6.2 Dictionnaire dénombrable et inégalité de Hoeffding

Soit $\mathcal{F} = \{f_j : j \in \mathbb{N}\}$ une famille dénombrable de classifieurs. Soit $\mathcal{D}_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$ un famille de couples aléatoires i.i.d. de loi P. Soit L_n le risque empirique. Enfin, soient $p_j > 0$ des réels tels que
$$\sum_{j=1}^{\infty} p_j = 1.$$

1. En utilisant l'inégalité de Hoeffding, montrer que :
$$P\left[\forall j, L(f_j) - L_n(f_j) \leq \sqrt{\frac{\log\left(\frac{2}{p_j\varepsilon}\right)}{2n}}\right] \geq 1 - \frac{\varepsilon}{2}.$$

2. On suppose qu'il existe $\tilde{j} \in \mathbb{N}^*$ tel que

$$\inf_{j\geq 1} L(f_j) = L(f_{\tilde{j}}).$$

Montrer que la règle de classification

$$g(\mathcal{D}_n, \cdot) = f_j(\cdot) \text{ où } j \in \arg\min_{j\in\mathbb{N}} \left(L_n(f_j) + \sqrt{\frac{\log\left(\frac{2}{p_j \varepsilon}\right)}{2n}} \right)$$

satisfait pour tout $\epsilon > 0$:

$$P\left[L(g) - L(f_{\tilde{j}}) \leq \sqrt{\frac{\log\left(\frac{2}{p_{\tilde{j}}\varepsilon}\right)}{2n}} + \sqrt{\frac{\log\left(\frac{2}{\varepsilon}\right)}{2n}} \right] \geq 1 - \epsilon.$$

6.3 Vitesse de convergence sous hypothèse de régularité de η

On suppose que la variable X est à valeurs dans $[0,1]^d$ pour $d \geq 3$. L'hypothèse sur la loi de probabilité P est que $\eta(\cdot)$ est L-Lipschitz, pour une certaine constante $L > 0$. On note μ la première marginale de P.

Soit f_k le classifieur des k-plus proches voisins : $f_k(x) = \mathbf{1}\hat{\eta}(x) \geq 1/2$ où $\hat{\eta}(x) = \frac{1}{k}\sum_{m=1}^{k} Y_{i_m(x)}$ et $m \mapsto i_m(x)$ est une permutation $\sigma(X_1, \ldots, X_n)$-mesurable telle que $\|X_{i_1(x)} - x\| \leq \cdots \leq \|X_{i_n(x)} - x\|$.

1. Pour $0 < \varepsilon < 1$, montrer qu'il existe une partition de $[0,1]^d$ formée de moins de $(\lfloor \frac{1}{\varepsilon} + 1 \rfloor^d)$ éléments de diamètre maximal au plus $\sqrt{d}\varepsilon$.
   On note $A_1, \ldots, A_{m_\varepsilon}$ les éléments de cette partition.
2. On rappelle que $X_{i_1(x)}$ est le plus proche voisin de x parmi $X_1, \ldots, X_n$. Montrer que

$$\forall x \in A_j, \quad P\Big(\Big\| X_{i_1(x)} - x \Big\| > \varepsilon\sqrt{d} \Big) \leq [1 - \mu(A_j)]^n.$$

3. En déduire que (X est maintenant aléatoire)

$$P\Big(\Big\| X_{i_1(X)} - X \Big\| > \varepsilon\sqrt{d} \Big) \leq m_\varepsilon \sup_{\alpha>0} \alpha \exp(-\alpha n) \leq \left(\frac{2}{\varepsilon}\right)^d \frac{1}{n}$$

4. En déduire que pour $c = \frac{4d^2}{d-2}$ on obtient

$$E(\|X_{i_1(X)} - X\|^2) \leq 2d \int_0^1 \min\left\{ 1, \left(\frac{2}{\varepsilon}\right)^d \frac{1}{n} \right\} \varepsilon d\varepsilon \leq \frac{c}{n^{2/d}}$$

5. Vérifier que

$$\begin{aligned}
&E[(\hat{\eta}(x)-\eta(x))^2]\\
=&E\left[\left(\frac{1}{k}\sum_{m=1}^{k} Y_{i_m(x)}-\eta(X_{i_m(x)})\right)^2\right]+E\left[\left(\frac{1}{k}\sum_{m=1}^{k}\eta(x)-\eta(X_{i_m(x)})\right)^2\right]\\
=&A(x)+B(x)
\end{aligned}$$

6. Montrer que pour $m \neq m'$ (attention $i_m(x)$ et $i_{m'}(x)$ sont aléatoires) :

$$E\left[(Y_{i_m(x)}-\eta(X_{i_m(x)}))(Y_{i_{m'}(x)}-\eta(X_{i_{m'}(x)}))|X_1,\ldots,X_n\right]=0$$

7. En déduire que $A(x) \leq \frac{1}{k}$
8. Soient $S_1,\ldots,S_k$ des sous ensembles disjoints de $\{X_1,\ldots,X_n\}$ tous de taille $\lfloor n/k \rfloor$ et $\hat{X}_j(x)$ le plus proche voisin de x dans S_j. Montrer que

$$B(x) \leq \frac{L^2}{k}\sum_{j=1}^{k} E\|\hat{X}_j(x)-x\|^2$$

9. En déduire que $B(x) \leq cL^2\left(\frac{2k}{n}\right)^{2/d}$ et en déduire la vitesse de convergence des k-plus proches voisins.
10. Qu'obtient-on pour $d=1$ et $d=2$?

7 Théorie de Vapnik-Cervonenkis et convexification du risque

7.1 Estimation du coefficient d'éclatement

Soit $B = \{b^{(1)}, \dots, b^{(M)}\} \subset \mathbb{R}^n$ un ensemble fini.

1. Montrer que la complexité de Rademacher de B vérifie

$$\mathcal{R}_n(B) \leq \max_{j=1,\dots,M} \|b^{(j)}\|_{\ell^2} \frac{\sqrt{2\log(2M)}}{n}.$$

2. Montrer que si

$$\mathcal{C}(B) = \Big\{ \sum_{j=1}^{M} c_j b^{(j)}, c_j \in \mathbb{R}, \sum_{j=1}^{M} |c_j| \leq 1 \Big\}$$

désigne le simplexe engendré par B, alors

$$\mathcal{R}_n\big(\mathcal{C}(B)\big) = \mathcal{R}_n(B).$$

7.2 VC-dimension

Vapnik et Chervonenkis ont démontré que, pour une classe de classifieurs donnée $\mathcal{F}$, le classifieur f_n minimisant le risque empirique satisfait

$$E\Big[L(f_n) - \inf_{g\in\mathcal{F}} L(g)\Big] \leq 2\sqrt{\frac{d\left(\log\left(\frac{2n}{d}\right)+1\right)+\log(4)}{n}} \qquad (\star)$$

où d est la VC-dimension de $\mathcal{F}$ qu'il reste à définir.

On dit que $\mathcal{F}$ peut pulvériser l'ensemble de points $X_1, \dots, X_n$ si pour toutes les 2^n valeurs possibles de $Y_1, \dots, Y_n \in \{0,1\}^n$, il existe un $f \in \mathcal{F}$ tel que $f(X_i) = Y_i$ pour tout $i \in \{1, \dots, n\}$. La VC-dimension de $\mathcal{F}$ est simplement le cardinal du plus grand ensemble de points que $\mathcal{F}$ peut pulvériser.

1. Comparer l'inégalité $(\star)$ avec celle vue en cours.
2. Supposons que X est à valeurs dans $\mathbb{R}$ et

$$\mathcal{F} = \big\{1_{]-\infty,x]}; x \in \mathbb{R}\big\} \cup \{1_{[x,\infty[}; x \in \mathbb{R}\}.$$

Calculez la VC dimension de $\mathcal{F}$ ainsi que le risque de l'estimateur minimisant le risque empirique.

3. Généraliser quand $X \in \mathbb{R}^d$ et $\mathcal{F}$ est la classe des classifieurs linéaires, i.e., $\mathcal{F} = \{\mathbf{1}_{\langle \cdot, b\rangle \leq c}; b \in \mathbb{R}^d, c \in \mathbb{R}\}$. Indication : On pourra montrer d'une part que $\mathcal{F}$ pulvérise l'ensemble $\{X_1, \ldots, X_{d+1}\}$ si $X_1 = 0$ et $(X_2, \ldots, X_d)$ est la base canonique de $\mathbb{R}^d$ et d'autre part que si $X_1, \ldots, X_{d+2}$ sont $d+2$ vecteurs de $\mathbb{R}^d$, il existe une partition (I, J) de $\{1, \ldots, d+2\}$ et deux familles $(\alpha_i)_{i\in I}$ et $(\beta_i)_{i\in J}$ de réels positifs tels que
$$\sum_{i\in I} \alpha_i = \sum_{i\in I} \beta_i = 1 \quad \text{et} \quad \sum_{i\in I} \alpha_i X_i = \sum_{i\in J} \beta_i X_i.$$
4. Quelle est la VC-dimension de $\mathcal{F} = \{\mathbf{1}_{\|\cdot - x\|_\infty \leq c}; x \in \mathbb{R}^2, c \in \mathbb{R}\}$?

7.3 Inégalité de Zhang

Soit φ un majorant convexe de $\varphi_0(x) = \mathbf{1}_{\{x\geq 0\}}$, c'est-à-dire φ majore φ_0, est convexe, $\varphi(0) = 1$ et $\varphi(-\infty) = 0$. On suppose de plus $\varphi(z) \geq \varphi(-z)$ pour tout $z \geq 0$. On pose $H_\eta(\alpha) = \eta\varphi(-\alpha) + (1-\eta)\varphi(\alpha)$ et on définit $\tau(\eta) = \inf_\alpha H_\eta(\alpha)$.

On suppose qu'il existe $\gamma \in [0, 1]$ et $c > 0$ de sorte que pour tout $\eta \in [0, 1]$;
$$\left|\tfrac{1}{2} - \eta\right| \leq c(1 - \tau(\eta))^\gamma.$$

1. Montrer que $\tau(\eta) \leq 1$.
2. Montrer que pour toute fonction mesurable de $\mathcal{X}$ à valeurs dans $\mathbb{R}$,
$$R(\mathrm{sign}(f)) - R(h^\star) \leq 2c\big(\mathbb{E}\big[(1 - \tau(\eta(X)))\mathbf{1}_{\{f(X)(\eta(X)-1/2)\leq 0\}}\big]\big)^\gamma.$$
3. Montrer que
$$(1 - \tau(\eta(X)))\mathbf{1}_{\{f(X)(\eta(X)-1/2)\leq 0\}} \leq H_{\eta(X)}(f) - \tau(\eta(X)).$$
4. En déduire l'inégalité de Zhang :
$$R(\mathrm{sign}(f)) - R^\star \leq 2c\big(R_\varphi(f) - R_\varphi^\star\big)^\gamma.$$
5. Montrer que l'inégalité de Zhang est vérifiée pour les fonctions hinge $\varphi(x) = (1+x)_+$, exponentielle et logistique.