EXAMEN

Durée : 2h. Tous les documents sont autorisés. Les trois exercices sont complètement indépendants.

Exercice 1 [7 points]

On considère le problème de classification binaire avec $\mathcal{Y} = \{0;1\}$ et $\mathcal{X} \subset \mathbb{R}^2$. On suppose que $Y_i = \mathbb{1}(U_i > 1)$ où $(X_1, U_1), \ldots, (X_n, U_n) \in \mathbb{R}^2 \times \mathbb{R}$ sont iid de densité (par rapport à la mesure de Lebesgue) :

$$f_{(X,U)}(x_1, x_2, u) = C|x_1|\mathbb{1}(x_1^2 + x_2^2 \leq 4)e^{-u|x_1|}\mathbb{1}(u \geq 0), \qquad (x_1, x_2, u) \in \mathbb{R}^3.$$

Ici, C est une constante de normalisation.

1. Quelle est la densité marginale du vecteur X_i ? Est-ce la densité d'une loi usuelle ?
2. Quelle est la densité conditionnelle $f_{U|X}(u|x)$ de U_i sachant X_i ? Reconnaissez-vous la densité d'une loi usuelle ?
3. Exprimer $\mathbf{E}[Y_i|X_i = x]$ à l'aide de la densité $f_{U|X}(u|x)$. En déduire la forme explicite du classifieur de Bayes g^*. A t-on besoin de la deuxième coordonnée des X_i pour prévoir Y_i ?
4. Montrer que $C = 1/4\pi$ et que le risque de classification de g^* est donnée par la formule :

$$\pi^{-1} \int_0^{\ln 2} (1 - e^{-x_1})\sqrt{4 - x_1^2}\, dx_1 + \pi^{-1} \int_{\ln 2}^{2} e^{-x_1}\sqrt{4 - x_1^2}\, dx_1.$$

A titre d'information, un calcul numérique montre que cette expression est égale à 0.2943.

Exercice 2 [7 points]

On considère le modèle de régression à design fixe :

$$Y_i = f(x_i) + \sigma\xi_i, \qquad i = 1, \ldots, n.$$

Soit $\{\phi_1, \ldots, \phi_M\}$ un dictionnaire dans lequel f admet une représentation parcimonieuse :

$$f(x) = \sum_{m=1}^{M} \vartheta_m^* \phi_m(x), \qquad \text{avec} \qquad \|\boldsymbol{\vartheta}^*\|_0 \ll M.$$

Dans ce qui suit, on utilise la notation $f_{\boldsymbol{\vartheta}}(x) = \sum_{m=1}^{M} \vartheta_m \phi_m(x)$. Soit $\widehat{\boldsymbol{\vartheta}}$ l'estimateur Lasso défini par

$$\widehat{\boldsymbol{\vartheta}} \in \arg\min_{\boldsymbol{\vartheta} \in \mathbb{R}^M} \bigg\{ \frac{1}{n} \sum_{i=1}^{n} \Big(Y_i - f_{\boldsymbol{\vartheta}}(x_i)\Big)^2 + \lambda \|\boldsymbol{\vartheta}\|_1 \bigg\}, \tag{1}$$

où λ est un paramètre positif.

1. Posons $Z_m = \frac{1}{n} \sum_{i=1}^{n} Y_i \phi_m(x_i)$. Montrer que si la matrice

$$\Phi = \frac{1}{\sqrt{n}} \begin{pmatrix} \phi_1(x_1) & \ldots & \phi_M(x_1) \\ \vdots & \ddots & \vdots \\ \phi_1(x_n) & \ldots & \phi_M(x_n) \end{pmatrix}$$

est orthogonale (c'est-à-dire $\Phi^\top \Phi = I_M$), alors l'estimateur Lasso vérifie

$$\widehat{\boldsymbol{\vartheta}} \in \arg\min_{\boldsymbol{\vartheta} \in \mathbb{R}^M} \sum_{m=1}^{M} \big\{ (Z_m - \vartheta_m)^2 + \lambda|\vartheta_m| \big\}. \tag{2}$$

2. On suppose que les ξ_i sont iid d'espérance 0 et de variance 1. Calculer l'espérance de Z_m et sa variance (sous la condition $\Phi^\top \Phi = I_M$).
3. Prouver que si $Z \geq 0$, alors

$$\arg\min_{u \in \mathbb{R}} \Big[(Z-u)^2 + \lambda|u|\Big] = \begin{cases} Z - \frac{\lambda}{2}, & \text{si } Z \geq \lambda/2, \\ 0, & \text{sinon.} \end{cases}$$

Indication : on pourra commencer par vérifier que la fonction $u \mapsto (Z-u)^2 + \lambda|u|$ est décroissante sur $]-\infty; 0]$ et en déduire que son minimum est forcément attaint sur $[0, +\infty[$.
4. Prouver que si $\widehat{\vartheta}$ est une solution de (2) alors on a

$$\forall m \in \{1, \ldots, M\} \qquad \widehat{\vartheta}_m = \left(|Z_m| - \frac{\lambda}{2}\right)_+ \operatorname{sign}(Z_m),$$

c'est-à-dire que Lasso coïncide avec le seuillage doux.

Exercice 3 [10 points]

Soient $X_i, i = 1, \ldots, n$ des variables iid de densité f sur $[0,1]$. On suppose que $f \in L_2([0,1])$ et désigne par $\langle g, g' \rangle$ le produit scalaire usuel de deux fonctions de L_2. Soit $\{\varphi_k : k \in \mathbb{N}\}$ une base orthonormée de $L_2([0,1])$.

On note $\ell_2(\mathbb{N})$ l'ensemble des suites $(u_k)_{k\in\mathbb{N}}$ de carré sommables. Pour deux suites $u, u' \in \ell_2(\mathbb{N})$, on utilise le produit scalaire usuel $\langle u, u' \rangle = \sum_{k\in\mathbb{N}} u_k u'_k$.

1. On pose $\vartheta_k = \langle f, \varphi_k \rangle$ pour tout $k \in \mathbb{N}$. Montrer que l'estimation de f équivaut à l'estimation de la suite $\vartheta = (\vartheta_k)_{k\in\mathbb{N}}$ pourvu que l'erreur d'estimation soit mesurée dans les deux cas par l'espérance du carré de la distance entre l'estimateur et la vraie valeur.
2. Montrer que

$$\widehat{\vartheta}_k = \frac{1}{n}\sum_{i=1}^{n} \varphi_k(X_i)$$

est un estimateur sans biais de ϑ_k. Vérifier que la variance de $\widehat{\vartheta}_k$ est donnée par

$$\frac{1}{n}\left(\int_0^1 \varphi_k^2(x) f(x)\,dx - \vartheta_k^2\right).$$

En supposant que $\sup_{x\in[0,1]} f(x) \leq L$, montrer que la variance de $\widehat{\vartheta}_k$ est bornée par L/n.
3. On suppose maintenant que (φ_k) vérifie $\varphi_{2p}(x)^2 + \varphi_{2p+1}(x)^2 = 2$ pour tout x et pour tout $p \in \mathbb{N}$. Montrer que le risque de l'estimateur $\widehat{\vartheta} = (\widehat{\vartheta}_k)_{k\in\mathbb{N}}$ est égal à $+\infty$. Quelle est la raison de cette explosion du risque : le sur-lissage (underfitting) ou le sous-lissage (overfitting) ?
4. Pour pallier ce défaut, on utilise un estimateur tronqué (K est un entier positif) :

$$\widehat{\vartheta}^K = \begin{cases} \widehat{\vartheta}_k; & \text{si } k \leq K, \\ 0, & \text{sinon.} \end{cases}$$

Montrer que le biais de cet estimateur décroît avec K alors que sa variance est bornée par LK/n.
5. On suppose maintenant que les coefficients ϑ_k de la vraie densité, en plus d'être de carré sommables, vérifient la propriété d'ellipsoïde suivante :

$$\sum_{k\in\mathbb{N}} k^{2\beta}\vartheta_k^2 \leq L$$

où β et L sont deux constantes positives. Montrer que le biais de l'estimateur tronqué vérifie :

$$\sum_{k\in\mathbb{N}} \left|\mathbf{E}[\widehat{\vartheta}_k^K] - \vartheta_k\right|^2 \leq \frac{L}{K^{2\beta}}.$$

6. En supposant β et L connus, proposer un choix de K pour lequel le risque quadratique $\mathbf{E}[\|f_{\widehat{\vartheta}^K} - f\|_2^2]$ est majoré par $Cn^{-2\beta/(2\beta+1)}$. Comment cela se compare avec les vitesses des convergences vues en cours ?