Exercice 1 Le support d'une distribution μ pour $\mathbb{R}^d$ est défini par l'ensemble

$$\operatorname{supp}\mu := \{x \in \mathbb{R}^d,\ \forall \epsilon > 0,\ \mu(B(x,\epsilon)) > 0\}$$

où on note $B(x,\epsilon)$ la boule fermée centrée en x de rayon ϵ. On suppose que μ admet une densité par rapport à la mesure de Lebesgue. Soit $(X_i, Y_i)_{1\leq i\leq n}$ i.i.d. observations, on définit pour tout $k < n$ le point $X_{(k)}(x)$ vérifiant

$$\#\{1 \leq i \leq n,\ d(x, X_i(x)) \leq d(x, X_{(k)}(x))\} = k.$$

où $d : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ est la distance Euclidienne.

1. Montrer que pour tout $x \in \operatorname{supp}\mu$, pour un entier k fixé, la convergence p.s.

   $$\lim_{n\to\infty} d(X_{(k)}(x), x) = 0$$

   a lieu. Déduire, pour une v.a. X indépendante des observations, la convergence p.s.

   $$\lim_{n\to\infty} d(X_{(k)}(X), X) = 0.$$

2. Admettons le lemme technique suivant

   Lemma 1 (Stone)**.** *Soit une fonction $f \in \mathbb{L}^1(\mathbb{R}^d, \mu)$. Pour tout $(k,n) \in \mathbb{N}^2$ satisfaisant $k \leq n$, il existe une constant γ_d qui ne dépend que de la dimension d vérifiant*

   $$\frac{1}{k}\sum_{i=1}^{k} \|f(X_{(i)}(X))\|_1 \leq \gamma_d \|f(X)\|_1.$$

   Montrer pour toute fonction $f \in \mathbb{L}^1(\mathbb{R}^d, \mu)$

   $$\lim_{n\to\infty} \frac{1}{k}\sum_{i=1}^{k} \|f(X) - f(X_{(i)}(X))\|_1 = 0.$$

3. On peut représenter le couple (X, Y) à l'aide d'une v.a. uniforme U indépendante de X

   $$Y = 1_{\eta(X)>U}$$

   et la règle de knn (k-nearest neighbor) par

   $$g_n(x) = \begin{cases} 1, & \phi(x, Y_{(1)}(x), \cdots, Y_{(k)}(x)) > 0 \\ 0, & sinon \end{cases}.$$

   Pour un x fixé, on remplace $Y_{(i)}(x) = 1_{\eta(X_{(i)}(x))>U}$ par $Y'_{(i)}(x) = 1_{\eta(x)>U}$ pour tout $1 \leq i \leq k$ et définit $g'_n(x)$ de la même manière. Montrer

   $$\mathbb{P}(g_n(X) \neq g'_n(X)) = \sum_{i=1}^{k} \|\eta(X) - \eta(X_{(k)}(X))\|_1.$$

4. Déduire, pour un k impair, la convergence

$$L_{knn} = \lim_{n\to\infty} \mathbb{E}L_n = L^* + \mathbb{E}\big[(1 - 2\min(\eta(X), 1-\eta(X)))\mathbb{P}(B(k, \min(\eta(X), 1-\eta(X))) > \frac{k}{2})|X)\big] \geq L^*$$

où on pose $B(k, \alpha)$ une v.a. de la loi binomiale de paramètre α.

5. Déduire, sous l'hypothèse précédente,

$$L_{knn} \leq L^* + \frac{1}{\sqrt{ke}}.$$

6. Soit une marche aléatoire à pas $(Z_i)_{i\geq 1}$ i.i.d. vérifiant $\mathbb{P}(Z_i = 1) = 1 - \mathbb{P}(Z_i = -1) < 1/2$. Montrer

$$\forall m > 1,\ \mathbb{P}(\sum_{i=1}^{2m+1} Z_i > 0) \leq \mathbb{P}(\sum_{i=1}^{2m-1} Z_i > 0).$$

7. Conclure avec l'inégalité de Cover-Hart $L^* \leq L_{(2m+1)nn} \leq L_{1nn} \leq 2L^*$. Si $L^* = 0$, on établit la consistance universelle de knn.

8. Supposons maintenant que k peut croître avec le nombre d'échantillon n et qu'il vérifie $\lim_{n\to\infty} k/n = 0$ et $\lim_{n\to\infty} k = +\infty$. Montrer en utilisant le théorème de Stone que la méthode est universellement consistante.