french

1 TD : rappels

Exercice 1 On considère le modèle de régression de dimension N

$$x_k = \theta_k + \epsilon_k, \quad k = 1, \dots, N,$$

où $\mathbf{x} = (x_k)_{k=1,\dots,N}$ est l'observation dans $\mathbb{R}^N$, $\boldsymbol{\theta} = (\theta_k)_{k=1,\dots,N}$ le paramètre de $\mathbb{R}^N$ et les ϵ_k sont des variables i.i.d. centrées de variance connue σ^2.

1. On cherche un estimateur de $\boldsymbol{\theta}$ de la forme $\tilde{\boldsymbol{\theta}}(\lambda) = \lambda \mathbf{x}$ avec $\lambda \in [0, 1]$. Calculer son biais $\mathbb{E}[\tilde{\boldsymbol{\theta}}(\lambda)] - \boldsymbol{\theta}$ et sa variance sommée $\sum_{k=1}^N \mathrm{Var}(\tilde{\theta}_k(\lambda))$. Quel est l'effet de λ sur le biais et la variance ?
2. Calculer $\lambda^* \in [0, 1]$ qui minimise le risque d'oracle de cet estimateur :
$$R(\lambda) = \mathbb{E}\left[\sum_{k=1}^N (\theta_k - \lambda x_k)^2\right] .$$
3. Proposer un estimateur sans biais de θ_k^2 à partir de x_k et σ^2.
4. En déduire un estimateur sans biais $\mathcal{C}(\lambda)$ de $R(\lambda) - \sum_{k=1}^N \theta_k^2$.
5. Quel est l'estimateur $\widehat{\boldsymbol{\theta}} = \lambda(\mathbf{x})\, \mathbf{x}$, avec $\lambda : \mathbb{R}^N \to [0, 1]$, obtenu par minimisation du critère $\mathcal{C}$ sur $[0, 1]$?

Exercice 2 Let $f : \mathbb{R} \to \mathbb{R}$ be an unknown function. Suppose we observe noisy values of f :

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

where $(X, \epsilon), (X_i, \epsilon_i)_{i=1,\dots,n} \subset \mathbb{R} \times \mathbb{R}$ is an i.i.d. collection of random variables such that $\epsilon \perp X$, $\mathbb{E}[\epsilon] = 0$, $\mathbb{E}[\epsilon^2] = \sigma^2$ and $\mathbb{E}[f(X)^2] < \infty$. To estimate the function f, we shall use a sequence of functions $(h_k)_{k \geq 1}$ where each $h_k : \mathbb{R} \to \mathbb{R}$ and $G = \mathbb{E}[h(X)h(X)^T]$ is invertible. . The procedure is as follows. We choose K, we set $h(x) = (h_1(x), \dots h_K(x))^T$ and then compute

$$\theta_{K,n} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^K} \sum_{i=1}^n (Y_i - h(X_i)^T \theta)^2.$$

The estimate of f is given by $f_{K,n}(x) = h(x)^T \theta_{K,n}$. Define the risk $R(\theta) = \mathbb{E}[(f(X) - h(X)^T \theta)^2]$. The aim is to study the $R(\theta_{n,K})$.

1. Let $\theta_K^* \in \operatorname{argmin}_{\theta \in \mathbb{R}^K} R(\theta)$. Give the normal equations satisfied by θ_K^* and deduce an expression for θ_K^*.
2. Give the expression for $\theta_{K,n}$.
3. Show that the estimated function $f_{K,n}$ is invariant by any linear invertible transform on the set of functions h, i.e. h is replaced by Ah where $A \in \mathbb{R}^{K \times K}$ is invertible. In the following we shall assume that $G = I_K$.
4. Show that for any $\theta \in \mathbb{R}^K$, $R(\theta) = R(\theta_K^*) + \|\theta - \theta_K^*\|^2$ where the norm $\| \cdot \|$ should be specified.

5. From now on, we suppose that the smallest eigenvalue of $n^{-1}\sum_{i=1}^{n} h(X_i)h(X_i)^T$ is lower bounded by $\lambda > 0$. Show that $\|\theta_{K,n} - \theta_K^*\| \leq \lambda^{-1/2}\|n^{-1}\sum_{i=1}^{n} \xi_i h(X_i)\|$ where $\xi_i \neq \epsilon_i$ should be specified.
6. Show that $n^{-1}\sum_{i=1}^{n} \xi_i h(X_i) \to 0$ almost surely (hint : use the Cauchy-Schwarz inequality).
7. Suppose that $f \in \text{span}((h_k)_{k\geq 1})$. Conclude that choosing K large enough, $\limsup_n R(\theta_{K,n})$ can be made arbitrarily small.

2 TD : histogramme

Exercice 3 On souhaite estimer "globalement" une densité de probabilité inconnue sur un intervalle donné, disons $[0,1]$ pour simplifier, à partir de l'observation de la réalisation d'un n-échantillon $(X_1,\dots,X_n)$. Cela signifie que les variables aléatoires réelles $X_1,X_2,\dots,X_n$ sont indépendantes et identiquement distribuées. Dans toute la suite, on notera $x\mapsto f(x)$ leur densité de probabilité commune définie sur $[0,1]$. Le but est d'estimer les valeurs $\big(f(x), x\in[0,1]\big)$ simultanément.

Soit $m\geq 1$ un entier. On définit les *boites* $B_1,B_2,\dots,B_m$ en posant :

$$B_1=\big[0,\tfrac{1}{m}\big)\,,\quad B_2=\big[\tfrac{1}{m},\tfrac{2}{m}\big)\,,\dots,\quad B_m=\Big[\tfrac{(m-1)}{m},1\Big]\ .$$

On appelle *largeur de bande* associée aux boites B_j le nombre $h=1/m$. Pour $j=1,\dots,m$, on définit

$$\widehat{p_j}=\frac{1}{n}\#\left\{X_i\in B_j,\ i=1,\dots,n\right\}=\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{B_j}(X_i)\ .$$

1. Montrer que $\widehat{p_j}$ est un estimateur sans biais de $p_j=\int_{B_j}f(u)\mathrm{d}u$. Quelle est sa variance ?

L'estimateur par histogramme de la densité est alors défini par la formule

$$\widehat{f_n}(x)=\sum_{j=1}^{m}\frac{\widehat{p_j}}{h}\,\mathbb{1}_{B_j}(x)\qquad\text{pour}\ \ x\in[0,1]\ .$$

2. Soit j l'indice de la boite contenant x. Montrer que

$$\mathbb{E}[\widehat{f_n}(x)]=\frac{p_j}{h}\quad\text{et}\quad\mathrm{Var}\Big[\widehat{f_n}(x)\Big]=\frac{p_j(1-p_j)}{nh^2}.$$

3. Supposons que f est continue en x. Que dire du biais de $\widehat{f_n}(x)$ pour estimer $f(x)$ quand h tend vers 0 ?

On définit *l'erreur quadratique moyenne intégrée* de l'estimateur $\widehat{f_n}$ de f en posant

$$\mathcal{R}(\widehat{f_n},f)=\mathbb{E}\left[\int_0^1\big(\widehat{f_n}(u)-f(u)\big)^2\mathrm{d}u\right]\ .$$

On suppose désormais que f est 2 fois continûment dérivable sur $[0,1]$.

On note $b(x)=\mathbb{E}[\widehat{f_n}(x)]-f(x)$ le biais de l'estimateur $\widehat{f_n}(x)$ et $v(x)=\mathrm{Var}\Big[\widehat{f_n}(x)\Big]$ sa variance.

4. Montrer que
$$b(x)=f'(x)\big(h(j-\tfrac{1}{2})-x\big)+O(h^2)\ ,$$
où $x\in B_j$.
5. Montrer que
$$\int_0^1 b(x)^2\mathrm{d}x=\frac{h^2}{12}\int_0^1\big(f'(x)\big)^2\mathrm{d}x+o(h^2)\ .$$
On pourra utiliser que $f'(x)=f'(x_j)+O(h)$ pour $x\in B_j$, où x_j désigne le centre de la boite B_j.

6. Comment varie le biais en fonction de h ?
7. En reproduisant les arguments précédents, montrer que

$$v(x) = \frac{f(x) + O(h)}{nh} \ ,$$

puis que

$$\int_0^1 v(x)\mathrm{d}x = \frac{1}{nh} + O(1/n) \ .$$

8. Comment varie la variance en fonction de h ?
9. Déduire des questions précédentes que

$$\mathcal{R}(\widehat{f}_n, f) = \frac{h^2}{12}\int_0^1 f'(u)^2\mathrm{d}u + \frac{1}{nh} + o(h^2) + o(1/(nh)) \ . \qquad (1)$$

10. On note $\widehat{f}_n = \widehat{f}_{n,h}$ pour mettre en évidence la dépendance en h de l'estimateur. Montrer que

$$\lim_{n\to\infty} n^{2/3} \inf_h \mathcal{R}^0(\widehat{f}_{n,h}, f) = (3/4)^{2/3}\left(\int_0^1 f'(u)^2\mathrm{d}u\right)^{1/3} \ ,$$

où $\mathcal{R}^0$ est l'approximation du risque obtenue en négligeant les termes en $o(\dots)$ dans (1).

Nous avons vu que la taille de fenêtre optimale $h_n^\star = h_n^\star(f)$ dépend de f, qui est inconnu. On s'intéresse maintenant au problème du choix automatique de la fenêtre pour l'estimation de la densité par histogramme. Nous allons donc chercher un choix de h dicté par l'observation $X_1, \dots, X_n$ uniquement, et dont l'erreur imite le mieux possible l'erreur idéale fournie par le choix de $h_n^\star$. Nous considérons la méthode de **validation croisée** de type *leave one out*. Nous écrivons désormais $\widehat{f_n(x)} = \widehat{f_{n,h}(x)}$ et définissions

$$L_n(h) = \int_0^1 \left(\widehat{f_{n,h}(u)} - f(u)\right)^2 du = \int_0^1 (\widehat{f_{n,h}(u)})^2 du - 2\int_0^1 \widehat{f_{n,h}(u)} f(u) du + \int_0^1 f(u)^2 du.$$

Définition 1
L'estimateur du risque par validation croisée est

$$\widehat{J_n}(h) = \int_0^1 (\widehat{f_{n,h}(u)})^2 du - \frac{2}{n}\sum_{i=1}^n \widehat{f_{n,h,i}}(X_i) \ ,$$

où $\widehat{f_{n,h,i}}(x)$ est l'estimateur de f au point x obtenu en ignorant la donnée X_i.

Exercice 4 1. Montrer que minimiser $\mathcal{R}(\widehat{f_{n,h}}, f)$ est équivalent à minimiser l'espérance de

$$J_n(h) = \int_0^1 (\widehat{f_{n,h}(u)})^2 du - 2\int_0^1 \widehat{f_{n,h}(u)} f(u) du.$$

2. Comparer $\mathbb{E}[\widehat{J_n}(h)]$ et $\mathbb{E}[J_n(h)]$.

En principe, pour minimiser $h \mapsto \widehat{J_n}(h)$, on doit reconstruire n histogrammes pour chaque valeur de h. Heureusement, on dispose du raccourci suivant.

3. Montrer que :

$$\widehat{J_n}(h) = \frac{2}{(n-1)h} - \frac{1}{h}\frac{n+1}{n-1}\sum_{j=1}^m \widehat{p_j}^2.$$

Solution de l'exercice 1

1. For all $1 \le k \le N$, $\quad \mathbb{E}[x_k] = \mathbb{E}[\theta_k + \epsilon_k] = \theta_k$ so $\mathbb{E}[\mathbf{x}] = \boldsymbol{\theta}$ and the bias is equal to

$$Bias(\lambda) = \mathbb{E}[\tilde{\boldsymbol{\theta}}(\lambda)] - \boldsymbol{\theta} = \lambda \mathbb{E}[\mathbf{x}] - \boldsymbol{\theta} = (\lambda - 1)\boldsymbol{\theta}.$$

For the variance, we have $\mathrm{Var}(\tilde{\theta}_k(\lambda)) = \lambda^2 \mathrm{Var}(x_k) = \lambda^2\sigma^2$ and $\sum_{k=1}^N \mathrm{Var}(\tilde{\theta}_k(\lambda)) = N\lambda^2\sigma^2$. We recover the well-known bias/variance trade-off
- $Bias(\lambda) \xrightarrow{\lambda\to 1} 0, \quad \mathrm{Var}(\lambda) \xrightarrow{\lambda\to 1} N\sigma^2.$
- $Bias(\lambda) \xrightarrow{\lambda\to 0} -\boldsymbol{\theta}, \quad \mathrm{Var}(\lambda) \xrightarrow{\lambda\to 0} 0.$

2. Define $\mathcal{L}(\lambda) = \sum_{k=1}^N (\theta_k - \lambda x_k)^2$ and $R(\lambda) = \mathbb{E}\left[\mathcal{L}(\lambda)\right]$. For all $1 \le k \le N$,

$$\begin{aligned}(\theta_k - \lambda x_k) &= \theta_k - \lambda\left(\theta_k + \epsilon_k\right) = (1-\lambda)\theta_k - \lambda\epsilon_k,\\ (\theta_k - \lambda x_k)^2 &= (1-\lambda)^2\theta_k^2 - 2\lambda(1-\lambda)\theta_k\epsilon_k + \lambda^2\epsilon_k^2.\end{aligned}$$

Taking the expectation leads to

$$\begin{aligned}\mathbb{E}\left[(\theta_k - \lambda x_k)^2\right] &= (1-\lambda)^2\theta_k^2 + \lambda^2\sigma^2\\ \mathbb{E}\left[(\theta_k - \lambda x_k)^2\right] &= \lambda^2(\theta_k^2 + \sigma^2) - 2\lambda\theta_k^2 + \theta_k^2,\end{aligned}$$

and by sum we finally get a 2nd ordre polynomial in λ,

$$R(\lambda) = \left(\sum_{k=1}^N (\theta_k^2 + \sigma^2)\right)\lambda^2 - \left(2\sum_{k=1}^N \theta_k^2\right)\lambda + \left(\sum_{k=1}^N \theta_k^2\right).$$

The minimizer λ^* is given by

$$\lambda^* = \frac{\sum_{k=1}^N \theta_k^2}{\sum_{k=1}^N (\theta_k^2 + \sigma^2)}.$$

3. For all $1 \le k \le N$, $\quad x_k^2 = \theta_k^2 + 2\theta_k\epsilon_k + \epsilon_k^2$ so $\mathbb{E}[x_k^2 - \sigma^2] = \theta_k^2$, and by sum

$$\mathbb{E}\left[\sum_{k=1}^N (x_k^2 - \sigma^2)\right] = \sum_{k=1}^N \theta_k^2.$$

4. Define $S_N = \sum_{k=1}^N \theta_k^2$, we want an estimate $\mathcal{C}(\lambda)$ such that $\mathbb{E}\left[\mathcal{C}(\lambda)\right] = R(\lambda) - S_N$. By the previous question, we have an estimate of S_N so that

$$\begin{aligned}\mathcal{C}(\lambda) &= \mathcal{L}(\lambda) - \sum_{k=1}^N (x_k^2 - \sigma^2) = \sum_{k=1}^N (\theta_k - \lambda x_k)^2 - \sum_{k=1}^N (x_k^2 - \sigma^2)\\ \mathcal{C}(\lambda) &= \left(\sum_{k=1}^N x_k^2\right)\lambda^2 - \left(2\sum_{k=1}^N \theta_k x_k\right)\lambda + \left(\sum_{k=1}^N \theta_k^2 - x_k^2 + \sigma^2\right).\end{aligned}$$

5. Minimizing $\mathcal{C}(\lambda)$ w.r.t. λ yields an estimate $\widehat{\boldsymbol{\theta}} = \lambda(\mathbf{x})\,\mathbf{x}$, with $\lambda : \mathbb{R}^N \to [0,1]$,

$$\lambda(x_1, \ldots, x_N) = \frac{\left(\sum_{k=1}^N \theta_k x_k\right)_+}{\sum_{k=1}^N x_k^2}.$$

Solution de l'exercice 2

1. $\theta^*_K \in \operatorname{argmin}_{\theta\in\mathbb{R}^K} R(\theta)$ satisfies $\nabla R(\theta^*_K) = 0$ with $\nabla R(\theta) = -2\mathbb{E}\left[h(X)(f(X) - h(X)^T\theta)\right]$. Therefore we have the normal equation

$$\underbrace{\mathbb{E}\left[h(X)h(X)^T\right]}_{G=I_K}\theta^*_K = \mathbb{E}\left[h(X)f(X)\right].$$

We can recover this expression using Hilbert projection theorem since $\mathrm{Span}((h_k)_{k\geq 1})$ is a closed linear subspace of L^2. Define $\widehat{f}(X) = h(X)^T\theta^*_K$, it is unique and characterized by $f - \widehat{f} \perp \mathrm{Span}((h_k)_{k\geq 1})$, i.e.,

$$\begin{aligned}
\mathbb{E}\left[(f(X) - \widehat{f}(X))h(X)\right] &= 0\\
\mathbb{E}\left[(f(X) - h(X)^T\theta^*_K)h(X)\right] &= 0.
\end{aligned}$$

2. Consider the empirical risk $R_n(\theta) = \sum_{i=1}^n (Y_i - h(X_i)^T\theta)^2$ along with its minimizer $\theta_{K,n} \in \operatorname{argmin}_{\theta\in\mathbb{R}^K} R_n(\theta)$ which is a stationnary point : $\nabla R_n(\theta_{K,n}) = 0$.

$$\begin{aligned}
&\nabla R_n(\theta_{K,n}) = -2\sum_{i=1}^n h(X_i)(Y_i - h(X_i)^T\theta_{K,n}) = 0,\\
&\left(\sum_{i=1}^n h(X_i)h(X_i)^T\right)\theta_{K,n} = \left(\sum_{i=1}^n h(X_i)Y_i\right) = \left(\sum_{i=1}^n h(X_i)f(X_i)\right) + \left(\sum_{i=1}^n h(X_i)\epsilon_i\right).
\end{aligned}$$

Define $\widehat{G}_n = n^{-1}\left(\sum_{i=1}^n h(X_i)h(X_i)^T\right)$ the empirical Gram matrix, we have

$$\begin{aligned}
\widehat{G}_n\theta_{K,n} &= \frac{1}{n}\sum_{i=1}^n h(X_i)Y_i\\
&= \frac{1}{n}\sum_{i=1}^n h(X_i)(Y_i - h(X_i)^T\theta^*_K) + \frac{1}{n}\sum_{i=1}^n h(X_i)h(X_i)^T\theta^*_K\\
\widehat{G}_n\theta_{K,n} &= \frac{1}{n}\sum_{i=1}^n h(X_i)(Y_i - h(X_i)^T\theta^*_K) + \widehat{G}_n\theta^*_K\\
\widehat{G}_n\left(\theta_{K,n} - \theta^*_K\right) &= \frac{1}{n}\sum_{i=1}^n h(X_i)(Y_i - h(X_i)^T\theta^*_K).
\end{aligned}$$

3. Any linear invertible transform $A \in \mathbb{R}^{K\times K}$ on the set of functions changes $\widehat{G}_n$ into $\widehat{G}_n A^{-1}$ and these two matrices both share the same column space. Indeed, consider $\tilde{h} = Ah$ so that $h = A^{-1}\tilde{h}$, we have

$$A^{-1}\left(\sum_{i=1}^n \tilde{h}(X_i)\tilde{h}(X_i)^T\right)\left(A^{-T}\theta_{K,n}\right) = A^{-1}\left(\sum_{i=1}^n \tilde{h}(X_i)Y_i\right),$$

meaning that $\tilde{\theta}_{K,n} = A^{-T}\theta_{K,n}$ and the invariance of the estimate

$$\tilde{f}_{K,n}(x) = \tilde{h}(x)^T\tilde{\theta}_{K,n} = h(x)^T A^T A^{-T}\theta_{K,n} = h(x)^T\theta_{K,n} = f_{K,n}(x).$$

4. For any $\theta \in \mathbb{R}^K$,

$$
\begin{aligned}
R(\theta) &= \mathbb{E}\left[(f(X) - h(X)^T\theta)^2\right] \\
&= \mathbb{E}\left[(f(X) - h(X)^T\theta_K^* + h(X)^T(\theta_K^* - \theta))^2\right] \\
&= \mathbb{E}\left[(f(X) - h(X)^T\theta_K^*)^2 - 2(f(X) - h(X)^T\theta_K^*)h(X)^T(\theta_K^* - \theta) + (h(X)^T(\theta_K^* - \theta))^2\right] \\
&= \mathbb{E}\left[(f(X) - h(X)^T\theta_K^*)^2\right] - 2\underbrace{\mathbb{E}\left[(f(X) - \widehat{f}(X))h(X)^T\right]}_{=0}(\theta_K^* - \theta) + \mathbb{E}\left[(h(X)^T(\theta_K^* - \theta))^2\right]
\end{aligned}
$$

$$
\begin{aligned}
R(\theta) &= R(\theta_K^*) + (\theta_K^* - \theta)^T\mathbb{E}\left[h(X)h(X)^T\right](\theta_K^* - \theta) \\
R(\theta) &= R(\theta_K^*) + (\theta_K^* - \theta)^T G(\theta_K^* - \theta).
\end{aligned}
$$

In the general case, we have the norm associated to the matrix G and in the particular case $G = I$ we have the euclidian norm,

$$
R(\theta) = R(\theta_K^*) + \|\theta_K^* - \theta\|_2^2.
$$

5. Define $\xi_i = Y_i - h(X_i)^T\theta_K^*$ and use the expression of question 2,

$$
\widehat{G}_n\left(\theta_{K,n} - \theta_K^*\right) = \frac{1}{n}\sum_{i=1}^n h(X_i)\xi_i.
$$

Therefore

$$
\|\theta_{K,n} - \theta_K^*\|^2 = \|\widehat{G}_n^{-1}\left(\frac{1}{n}\sum_{i=1}^n h(X_i)\xi_i\right)\|^2.
$$

Recall that for any matrix $A \in \mathbb{R}^{p\times p}$ and $u \in \mathbb{R}^p$, $\|Au\|^2 \le \lambda_{\max}(A)\|u\|^2$. By assumption, we have $\lambda_{\min}(\widehat{G}_n) \ge \lambda$ so the maximum eigenvalue of the inverse is such that $\lambda_{\max}(\widehat{G}_n^{-1}) = 1/\lambda_{\min}(\widehat{G}_n) \le 1/\lambda$ and we get the following bound

$$
\|\theta_{K,n} - \theta_K^*\| \le \frac{1}{\sqrt{\lambda}}\|\frac{1}{n}\sum_{i=1}^n h(X_i)\xi_i\|.
$$

6. We apply the strong law of large numbers to the random variables $h(X_i)\xi_i$. Notice that for all $k \in K, h_k \in L^2$ and since $\mathbb{E}[f(X)^2] < \infty$ and $\mathbb{E}[\epsilon_i^2] = \sigma^2 < \infty$, we have $Y_i \in L^2$ and $\xi_i \in L^2$. Using Cauchy-Schwarz inequality, we have

$$
\mathbb{E}\left[|h_k(X_i)\xi_i|\right]^2 \le \mathbb{E}\left[|h_k(X_i)|^2\right]\mathbb{E}\left[|\xi_i|^2\right] < \infty.
$$

Besides the expectation is given by

$$
\begin{aligned}
\mathbb{E}\left[h(X_i)\xi_i\right] &= \mathbb{E}\left[h(X_i)(Y_i - h(X_i)^T\theta_K^*)\right] \\
&= \mathbb{E}\left[h(X_i)(f(X_i) + \epsilon_i - h(X_i)^T\theta_K^*)\right] \\
&= \mathbb{E}\left[h(X_i)(f(X_i) - h(X_i)^T\theta_K^*)\right] + \mathbb{E}\left[h(X_i)\epsilon_i\right] \\
\mathbb{E}\left[h(X_i)\xi_i\right] &= 0,
\end{aligned}
$$

where we used the normal equation to treat the first term and the fact that the noise ϵ is centered for the second term.

7. Thanks to questions 4 and 5, we have

$$\begin{aligned} R(\theta_{K,n}) &= R(\theta_K^*) + \|\theta_{K,n} - \theta_K^*\| \\ &\leq R(\theta_K^*) + \frac{1}{\sqrt{\lambda}} \|\frac{1}{n}\sum_{i=1}^n h(X_i)\xi_i\|. \end{aligned}$$

Assume that $f \in \mathrm{span}((h_k)_{k\geq 1})$ and set $\varepsilon > 0$. We can choose K large enough to have

$$\|f - \sum_{k=1}^K \alpha_k h_k\|_{L^2} \leq \varepsilon,$$

and the risk is such that $R(\theta_K^*) \leq \varepsilon$. Using the law of large numbers, the second term goes to 0 so we can take the limit sup and write

$$\limsup_{n\to\infty} R(\theta_{K,n}) \leq \varepsilon.$$

Solution de l'exercice 3

1. By linearity of the expectation, we have

$$\mathbb{E}\left[\widehat{p}_j\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\mathbb{1}_{B_j}(X_i)] \; .$$

Using the fact that the sequence $(X_1, \ldots, X_n)$ is independently and identically distributed, we have

$$\mathbb{E}\left[\widehat{p}_j\right] = \mathbb{E}[\mathbb{1}_{B_j}(X_1)] = \int_{B_j} f(u)\mathrm{d}u \; .$$

Because for any random variable Z and any constant a, $\mathrm{Var}[Z-a] = \mathrm{Var}[Z]$, the variance is given by

$$\begin{aligned} \mathrm{Var}[\widehat{p}_j] &= \mathrm{var}[\widehat{p}_j - p_j] \\ &= n^{-1}\mathrm{var}(\mathbb{1}_{B_j}(X_1)) \\ &= n^{-1}p_j(1-p_j) \; . \end{aligned}$$

2. As x is lying in the box with index j, we have $\widehat{f}(x) = \widehat{p}_j/h$. Using again the linearity of the expectation, we get

$$\mathbb{E}[\widehat{f}_n(x)] = h^{-1}\mathbb{E}\left[\widehat{p}_j\right] = h^{-1}\int_{B_j} f(u)\mathrm{d}u \; .$$

The computation of the variance is as follows,

$$\mathrm{Var}\left[\widehat{f}_n(x)\right] = \mathrm{Var}\left[\widehat{p}_j/h\right] = (nh^2)^{-1}p_j(1-p_j) \; .$$

3. By definition, the bias is given by $b(x) = \mathbb{E}[\widehat{f}_n(x)] - f(x)$. Because $\int_{B_j} \mathrm{d}u = h$, we have

$$b(x) = h^{-1}\int_{B_j} f(u)\mathrm{d}u - f(x) = h^{-1}\int_{B_j}(f(u) - f(x))\mathrm{d}u \; . \tag{2}$$

It follows that

$$
\begin{aligned}
|b(x)| &\leq h^{-1} \int_{B_j} |f(u) - f(x)| \mathrm{d}u \\
&\leq \sup_{u \in B_j} |f(u) - f(x)|\, h^{-1} \int_{B_j} \mathrm{d}u \\
&= \sup_{u \in B_j} |f(u) - f(x)| \\
&\leq \sup_{|u-x|<h} |f(u) - f(x)| \ .
\end{aligned}
$$

The latter bound is independant on the block index j. Let $\epsilon > 0$. From the continuity of the function f, there exists $\tilde{h} > 0$ such that whenever $|u - x| < \tilde{h}$, it holds that $|f(u) - f(x)| < \epsilon$. As a consequence, for h sufficently small ($h \leq \tilde{h}$), one has

$$
|b(x)| \leq \epsilon \ .
$$

Since ϵ is arbitrary, this means that $b(x) \to 0$ as $h \to 0$.

4. Using the identity $a^2 - b^2 = (a - b)(a + b)$, note that

$$
\begin{aligned}
\int_{B_j} (u - x) \mathrm{d}u &= \big[(u - x)^2/2\big]_{(j-1)/m}^{j/m} \\
&= (1/m) \left(\frac{j - 1/2}{m} - x \right) \\
&= h\big(h(j - \tfrac{1}{2}) - x\big) \ . \qquad (3)
\end{aligned}
$$

Denote by f' and f'' the first and second derivatives of f. Using the fact that f is 2 times continuously differentiable on $[0, 1]$, we have that (Taylor formula with integral remainder)

$$
f(u) - f(x) - f'(x)(u - x) = \int_x^u f''(v)(u - v) \mathrm{d}v \ .
$$

It follows that, for every $(x, u) \in [0, 1]^2$,

$$
|f(u) - f(x) - f'(x)(u - x)| \leq \frac{1}{2}(u - x)^2 \sup_{v \in [0,1]} |f''(v)| \ .
$$

As a consequence, using (2) and (3), we get

$$
\begin{aligned}
b(x) - f'(x)\big(h(j - \tfrac{1}{2}) - x\big) &= h^{-1} \int_{B_j} (f(u) - f(x)) \mathrm{d}u - h^{-1} f'(x) \int_{B_j} (u - x) \mathrm{d}u \\
&= h^{-1} \int_{B_j} (f(u) - f(x)) - f'(x)(u - x) \mathrm{d}u \ .
\end{aligned}
$$

It follows that

$$
\begin{aligned}
|b(x) - f'(x)\big(h(j - \tfrac{1}{2}) - x\big)| &\leq \frac{1}{2} \sup_{v \in [0,1]} |f''(v)| h^{-1} \int_{B_j} (u - x)^2 \mathrm{d}u \\
&\leq \frac{1}{2} \sup_{v \in [0,1]} |f''(v)| h^{-1} h^2 \int_{B_j} \mathrm{d}u \\
&= \frac{1}{2} \sup_{v \in [0,1]} |f''(v)| h^2 \ .
\end{aligned}
$$

5. First remark that

$$\int_{B_j} \big(h(j-\tfrac{1}{2})-x\big)^2 dx = -\frac{1}{3}\big[\big(h(j-\tfrac{1}{2})-x\big)^3\big]_{h(j-1)}^{hj} = \frac{h^3}{12} \ . \tag{4}$$

Using that f' is continuously differentiable on $[0,1]$, we have that

$$\sup_{x\in B_j} |f'(x)-f'(x_j)| \le h \sup_{x\in[0,1]} |f''(x)| \ ,$$

and that

$$\sup_{x\in B_j} |f'(x)^2-f'(x_j)^2| \le 2h \sup_{x\in[0,1]} |f'(x)| \sup_{x\in[0,1]} |f''(x)| \ .$$

Consequently, defining $\tilde{b}(x) = f'(x)^2 \sum_{j=1}^m \big(h(j-\tfrac{1}{2})-x\big)^2 1_{\{x\in B_j\}}$,

$$\begin{aligned}\int \tilde{r}(x)dx &= \sum_{j=1}^m \int_{B_j} f'(x)^2\big(h(j-\tfrac{1}{2})-x\big)^2 dx \\ &= \sum_{j=1}^m f'(x_j)^2 \int_{B_j} \big(h(j-\tfrac{1}{2})-x\big)^2 dx + r_1(h),\end{aligned}$$

where, using (4),

$$\begin{aligned}|r_1(h)| &= \left|\sum_{j=1}^m \int_{B_j} (f'(x)^2 - f'(x)^2)\big(h(j-\tfrac{1}{2})-x\big)^2 dx\right| \\ &\le 2h \sup_{x\in[0,1]} |f'(x)| \sup_{x\in[0,1]} |f''(x)| \sum_{j=1}^m \int_{B_j} \big(h(j-\tfrac{1}{2})-x\big)^2 dx \\ &= O(h^3) \ .\end{aligned}$$

Finally, using (4) again,

$$\begin{aligned}\int f'(x)^2\big(h(j-\tfrac{1}{2})-x\big)^2 dx &= \frac{h^3}{12}\sum_{j=1}^m f'(x_j)^2 + r_1(h) \\ &= \frac{h^2}{12}\sum_{j=1}^m \int_{B_j} f'(x_j)^2 dx + r_1(h) \\ &= \frac{h^2}{12}\int_0^1 f'(x)^2 dx + r_1(h) + r_2(h) \ ,\end{aligned}$$

with

$$r_2(h) = \frac{h^2}{12}\sum_{j=1}^m \int_{B_j} (f'(x_j)^2 - f'(x)^2) dx \ .$$

We conclude remarking that $|r_2(h)| \le \frac{h^3}{6} \sup_{x\in[0,1]} |f'(x)| \sup_{x\in[0,1]} |f''(x)|$.

6. Let x be a point in the box B_j. From question 4., we have

$$\big|f'(x)\big(h(j-\tfrac{1}{2})-x\big)\big| \le |f'(x)| \sup_{x\in B_j} |h(j-\tfrac{1}{2})-x| = |f'(x)|h/2 \ .$$

As a result, we have that $b(x) = O(h)$. For the previous question, we have also proved that the square integrated bias is of order h^2.

7. For the first point, it is enough to show that there exists $C > 0$ such that for every $j \in \{1, \dots, m\}$,

$$\sup_{x\in B_j} |p_j(1-p_j) - hf(x)| \le C\, h^2 \ .$$

As we have, in virtue of the triangle inequality, that, for every $x \in B_j$,

$$|p_j(1-p_j) - hf(x)| \le |p_j(1-p_j) - h^{-1}\int_{B_j} f(u)du| + |\int_{B_j}(f(u)-f(x))du| \ ,$$

we can proceed in the two following steps. First,

$$|p_j(1-p_j) - h^{-1}\int_{B_j} f(u)du| = p_j^2 \le h^2 \sup_{x\in[0,1]} |f(x)|^2 \ .$$

Second,

$$|\int_{B_j}(f(u)-f(x))du| \le h^2 \sup_{x\in[0,1]} |f'(x)| \ .$$

8. For each $x \in [0,1]$, the variance $v(x)$ goes to infinity as h goes to 0 and n remains fixed. We have shown in question 6. that the bias is going to 0 as h goes to 0. This two facts imply that we should define h as a sequence $h := h_n$ which satisfies

$$h_n \to 0 \ , \qquad nh_n \to +\infty \ .$$

9. Start writing that

$$\mathcal{R}(\widehat{f}_n, f) = \mathbb{E}\left[\int_0^1 (\widehat{f}_n(u) - \mathbb{E}[\widehat{f}_n(u)])^2 \mathrm{d}u\right] + \int_0^1 b(u)^2\mathrm{d}u \ .$$

Then use Tonelli's theorem to obtain that

$$\mathcal{R}(\widehat{f}_n, f) = \int_0^1 v(u)\mathrm{d}u + \int_0^1 b(u)^2\mathrm{d}u \ .$$

Concude by using question 7.

10. Neglecting the terms in the $o(\dots)$, we compute the infinum over h by minimizing the function $h \mapsto \frac{Ih^2}{12} + \frac{1}{nh}$. We find

$$h^{*3} = \frac{6}{I^2 n} \qquad \text{with } I = \int_0^1 f'(u)^2\mathrm{d}u \ .$$

Injecting the previous value gives that

$$\inf_h \mathcal{R}(\widehat{f}_{n,h}, f) = \frac{I^{1/3}}{n^{2/3}}\left(\frac{3}{4}\right)^{2/3} + o\left(\frac{1}{n^{2/3}}\right) \ .$$