

# THÉORIE STATISTIQUE DE L'APPRENTISSAGE

OLIVIER CATONI

*DEA Probabilités et Applications  
Corrigé de l'examen du 7 juin 2002.*

## PREMIÈRE PARTIE

**I.1)** On range les éléments à coder par probabilités décroissantes et on leur fait correspondre des codes rangés par longueurs croissantes, par exemple 0; 1; 00; 01; 10; 11; 000; 001 etc. On obtient ainsi le code

$$(c_i)_{i=1}^8 = (10; 0; 1; 000; 11; 001; 01; 00),$$

de longueur moyenne 1,58.

**I.2)** Pour obtenir un code binaire préfixe optimal, il convient d'utiliser l'algorithme de Huffman. Il consiste à agréger de façon récursive les deux éléments de probabilités les plus faibles, et à ramener ainsi (d'après la démonstration du cours) le calcul d'un code optimal pour un vecteur de probabilités de taille  $n$  à celui d'un code optimal pour un vecteur de taille  $n - 1$ , un code optimal pour un vecteur de probabilités de taille 2 étant de façon évidente égal à (0;1). On peut présenter les agrégations successives dans un tableau :

$i$	vecteurs de probabilités								$c(i)$
1	0,1	0,1	0,1						010
2	0,3	0,3	0,3	0,3	0,3	0,3	4▷ 0,58	7▷ 1	11
3	0,2	0,2	0,2	0,2	0,2				00
4	0,05	6▷ 0,08	5▷ 0,15	0,15	8▷ 0,28	0,28			10111
5	0,07	0,07						1010	
6	0,03							10110	
7	0,12	0,12	0,12	1▷ 0,22	0,22	3▷ 0,42	0,42		011
8	0,13	0,13	0,13	0,13					100

Dans la dernière colonne du tableau, on a calculé les codes d'après la règle suivante: pour calculer  $c(i)$ , on lit le tableau des vecteurs de probabilités, de la gauche vers la droite, en partant de la ligne  $i$ , et en suivant chaque ligne jusqu'à son extrémité. Lorsqu'on atteint la fin d'une ligne, on saute à la ligne à laquelle elle est agrégée, jusqu'à parvenir à la huitième et dernière colonne du tableau. On écrit le code  $c(i)$  de l'élément  $i$  de la droite vers la gauche, en ajoutant un 0 chaque fois que l'on change de ligne et un 1 chaque fois que la ligne que l'on parcourt reçoit la contribution d'une autre ligne.

La longueur moyenne du code de Huffman ainsi construit est égale à 2,73. (NB: la construction d'un code de Huffman n'est pas unique, on peut en effet à chaque branchement inverser le rôle de 0 et de 1, mais la longueur moyenne de tous les codes de Huffman est la même, puisque ces codes sont tous optimaux.)

**I.3)** On peut considérer un code binaire préfixe n'utilisant que des mots de longueur paire comme un code préfixe sur l'alphabet à quatre lettres  $\{00; 01; 10; 11\}$ . La question posée concerne donc la généralisation de l'algorithme de Huffman à des codes utilisant un alphabet  $A$  à  $k$  symboles (ici  $k = 4$ ). Un dictionnaire préfixe complet  $\mathcal{D}$  sur un alphabet à  $k$  lettres vérifie  $|\mathcal{D}| = 1 + |\mathring{\mathcal{D}}|(k-1)$ . Ainsi, lorsque l'ensemble  $E$  à coder ne vérifie pas la congruence  $(|E|-1)/(k-1) \in \mathbb{N}$ , les codes préfixes optimaux ne peuvent être complets. On peut néanmoins adapter la construction de Huffman et sa preuve de la façon suivante: soit  $c$  un code optimal vérifiant les propriétés requises et  $c(E) = \mathcal{D}$  l'ensemble de ses valeurs. Soit  $\bar{\mathcal{D}}$  le complété de  $\mathcal{D}$  (le plus petit dictionnaire préfixe complet contenant  $\mathcal{D}$ ). Dans la représentation des dictionnaires préfixes par des arbres,  $\bar{\mathcal{D}}$  s'obtient à partir de  $\mathcal{D}$  en « complétant les fratries incomplètes », ce qui peut aussi s'exprimer par la relation :

$$\bar{\mathcal{D}} = (\mathring{\mathcal{D}} \times A) \setminus \mathring{\mathcal{D}}.$$

Cette relation montre que

$$\max\{\ell(m); m \in \mathcal{D}\} = \max\{\ell(m); m \in \mathring{\mathcal{D}}\} + 1 = \max\{\ell(m); m \in \bar{\mathcal{D}}\} \stackrel{\text{def}}{=} M.$$

Soit  $\mathcal{D}_{\max} = \{m \in \mathcal{D}; \ell(m) = M\}$  les mots de  $\mathcal{D}$  de longueur maximum,

$\bar{\mathcal{D}}_{\max} = \{m \in \bar{\mathcal{D}}; \ell(m) = M\}$  les mots de  $\bar{\mathcal{D}}$  de longueur maximum

et  $\mathring{\mathcal{D}}_{\max} = \{m \in \mathring{\mathcal{D}}; \ell(m) = M-1\}$  les mots de  $\mathring{\mathcal{D}}$  de longueur maximum. On voit que

$$\bar{\mathcal{D}}_{\max} = \mathring{\mathcal{D}}_{\max} \times A$$

(car cet ensemble ne rencontre pas  $\mathring{\mathcal{D}}$ ). On voit aussi que  $\bar{\mathcal{D}} \setminus \mathcal{D} \subset \bar{\mathcal{D}}_{\max}$ , en effet si ce n'était pas le cas, on obtiendrait un code de longueur moyenne strictement inférieure à celle de  $c$  en remplaçant un mot codé de  $\mathcal{D}_{\max}$  par un mot codé plus court choisi dans  $\bar{\mathcal{D}} \setminus \mathcal{D}$ . Quitte à modifier  $c$  sans changer la longueur des mots codés, on peut supposer que  $\mathcal{D}_{\max}$  est constitué des premiers éléments de  $\bar{\mathcal{D}}_{\max}$  pour l'ordre lexicographique. Montrons que

$$\{m \in \mathring{\mathcal{D}}_{\max}; (m \times A) \not\subset \mathcal{D}_{\max}\}$$

contient au plus un élément (qui est donc dans ce cas le dernier de  $\mathring{\mathcal{D}}_{\max}$  pour l'ordre lexicographique). En effet s'il contenait deux mots distincts  $m_1 < m_2$  (pour l'ordre lexicographique), on aurait  $(m_2 \times A) \cap \mathcal{D}_{\max} = (m_2 \times A) \cap \mathcal{D} = \emptyset$ , en contradiction avec le fait que  $m_2 \in \mathring{\mathcal{D}}$ . Quitte à modifier  $c$  comme indiqué tout en lui gardant son caractère optimal, on a donc montré que le dernier élément  $m_{\max}$  de  $\mathring{\mathcal{D}}_{\max}$  pour l'ordre lexicographique vérifiait

$$(m_{\max} \times A) \setminus \mathcal{D} = \bar{\mathcal{D}} \setminus \mathcal{D}.$$

Il en découle

$$\begin{aligned}
 |(m_{\max} \times A) \cap \mathcal{D}| &= k - |(m_{\max} \times A) \setminus \mathcal{D}| \\
 &= k - |\overset{\circ}{\mathcal{D}} \setminus \mathcal{D}| \\
 &= k - (k-1)|\overset{\circ}{\mathcal{D}}| - 1 + |E| \\
 &= |E| - (k-1)(|\overset{\circ}{\mathcal{D}}| - 1) \\
 &= (|E| - 2) \pmod{(k-1)} + 2 \\
 &\stackrel{\text{def}}{=} r.
 \end{aligned}$$

(En effet,  $2 \leq |(m_{\max} \times A) \cap \mathcal{D}| \leq k$ : le cas  $(m_{\max} \times A) \cap \mathcal{D} = \{(m_{\max}, a)\}$  est exclu, car on pourrait alors remplacer le mot codé  $(m_{\max}, a)$  par le mot plus court  $m_{\max}$ .)  
 Quitte à échanger des codes sans augmenter la longueur moyenne de  $c$ , on peut de plus supposer que les mots de  $(m_{\max} \times A) \cap \mathcal{D}$  codent un sous-ensemble arbitraire de  $r$  éléments de  $E$  ayant les probabilités les plus faibles.

On vient donc de montrer le résultat suivant :

**Lemme 0.1.** *Étant donné  $r$  éléments de  $E$  arbitraires dont la somme des probabilités est minimale, il existe un code optimal  $c$  et un mot codé  $m_{\max}$  tel que  $(m_{\max} \times A) \cap c(E)$  code ces  $r$  éléments.*

On continue ensuite comme dans la preuve de Huffman. On considère l'ensemble réduit  $E'$  de cardinal  $|E| - r + 1$  obtenu en agrégeant  $r$  éléments dont la somme des probabilités est minimale et le vecteur de probabilité agrégé  $p'$  correspondant. Soit  $c'$  un code préfixe quelconque sur  $E'$  et  $c$  le code sur  $E$  obtenu en posant

$$\begin{aligned}
 c(e) &= c'(e) \text{ quand } e \in E' \cap E, \\
 c(e) &= (c'(e_*), a_e) \text{ quand } e \in E \setminus E',
 \end{aligned}$$

où  $e_*$  désigne l'élément agrégé, et  $\{a_e; e \in E \setminus E'\}$  une famille quelconque de  $r$  lettres distinctes. Le code  $c$  est préfixe et vérifie

$$\sum_{e \in E} p(e) \ell[c(e)] = \sum_{e \in E'} p'(e) \ell[c'(e)] + p'(e_*).$$

Le code  $c$  est donc optimal parmi les codes construits à partir d'un code préfixe de  $E'$  si et seulement si  $c'$  l'est. Comme il existe un code optimal parmi tous les codes préfixe qui est construit à partir d'un code de  $E'$  (d'après le lemme précédemment démontré), le code  $c$  est optimal parmi tous les codes préfixe, dès que  $c'$  l'est. On construit alors un code optimal par agrégations successives. Remarquons que  $r$  est nécessairement égal à  $k$  dès la deuxième étape de la construction.

**Application à l'exemple:**  $r = (6 \pmod 3) + 2 = 2$ . On peut consigner les agrégations successives dans le tableau suivant :

$i$	vecteurs de probabilités				$c(i)$
1	0,1	0,1			1100
2	0,3	0,3	0,3		00
3	0,2	0,2	0,2		01
4	0,05	6> 0,08			110101
5	0,07	0,07			1110
6	0,03				110100
7	0,12	0,12	1,4,5> 0,37	2,3,8> 1	1111
8	0,13	0,13	0,13		10

La longueur moyenne du code optimal est égale à 2,9.

**I.4)** Construction d'un code arithmétique : on choisit un ordre quelconque sur  $E$ , par exemple ici l'ordre croissant sur les entiers. On code l'élément  $e$  par le code correspondant à un intervalle dyadique de longueur  $2^{-\lceil -\log_2(p(e)) \rceil - 1}$  inclus dans  $\left[ \sum_{e' < e} p(e'), \sum_{e' \leq e} p(e') \right]$ . On obtient par exemple la borne droite d'un tel intervalle en tronquant la représentation binaire de  $\sum_{e' \leq e} p(e')$  à la  $\ell^{\text{e}}$  décimale où  $\ell = \lceil -\log_2(p(e)) \rceil + 1$ .

Application au codage de l'élément 4 :  $\sum_{e' \leq e} p(e') = 0,65$ ,  $\lceil -\log_2(p(4)) \rceil + 1 = \lceil -\log_2(0,05) \rceil + 1 = \ll \text{le nombre de fois où il faut multiplier } 0,05 \text{ par } 2 \text{ pour obtenir un résultat supérieur ou égal à } 1 \gg + 1 = 6$ . De plus  $0,65 = 0,101001\dots$  (base 2), d'où  $c(4) = 101000$ . NB: il y a deux intervalles dyadiques de même longueur inclus dans  $\left[ \sum_{e' < e} p(e'), \sum_{e' \leq e} p(e') \right]$ , si bien que la réponse  $c(4) = 100111$  convenait aussi. On l'obtient en suivant la construction du cours qui consiste à prendre comme borne gauche de l'intervalle dyadique  $2^{-\ell} \lceil \sum_{e' < e} p(e') 2^\ell \rceil$ .

## DEUXIÈME PARTIE

D'après la proposition 2.4 du chapitre 1 du cours, il suffit de trouver une loi  $\rho$  sur  $[\epsilon, 1]$  pour laquelle  $\mathcal{R}(P_\theta, P_\rho)$  est  $\rho$  presque partout égale à son maximum. Comme  $P_\rho = P_{\int \theta \rho(d\theta)}$ , on voit immédiatement que le maximum de  $\mathcal{R}(P_\theta, P_\rho)$  est atteint sur l'ensemble  $\{\epsilon, 1\}$ . S'il était atteint en un seul de ces deux points, on devrait avoir  $\rho = \delta_\epsilon$  où  $\rho = \delta_1$ , qui ne convient pas, par contre, si  $\theta(\epsilon)$  est tel que

$$(1) \quad \mathcal{R}(P_\epsilon, P_{\theta(\epsilon)}) = \mathcal{R}(P_1, P_{\theta(\epsilon)}),$$

alors il existe  $\rho_\epsilon \in \mathcal{M}_+^1(\{\epsilon, 1\})$  telle que  $\int \theta \rho_\epsilon(d\theta) = \rho_\epsilon(\epsilon)\epsilon + \rho_\epsilon(1) = \theta(\epsilon)$ , et qui vérifie bien que  $\mathcal{R}(P_\theta, P_{\rho_\epsilon})$  est  $\rho_\epsilon$  presque sûrement égale à son maximum. Ceci montre que la loi de codage minimax sur  $[\epsilon, 1]$  est l'unique loi  $P_{\theta(\epsilon)}$  vérifiant (1). Le paramètre  $\theta(\epsilon)$  est donc solution de l'équation

$$\epsilon \log\left(\frac{\epsilon}{\theta}\right) + (1 - \epsilon) \log\left(\frac{1 - \epsilon}{1 - \theta}\right) = \log\left(\frac{1}{\theta}\right),$$

ce qui prouve que

$$\theta(\epsilon) = \left(1 + (1 - \epsilon)\epsilon^{\frac{\epsilon}{1 - \epsilon}}\right)^{-1}.$$

## TROISIÈME PARTIE

Il suffit de reprendre la démonstration du théorème de Rissanen en posant  $K = \frac{N^\beta}{\log(N)}$ .

## QUATRIÈME PARTIE

**IV.1.1)** Considérons sur les simplexes  $\mathcal{M}_+^1(\mathcal{X})$  et  $\mathcal{M}_+^1(\mathcal{Y})$  les lois a priori de Krichevski – Trofimov :

$$\rho_{\mathcal{X}}(d\theta) = |\mathcal{X}|^{-1/2} \frac{\Gamma\left(\frac{|\mathcal{X}|}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^{|\mathcal{X}|}} \prod_{x \in \mathcal{X}} \theta(x)^{-1/2} \lambda_{\mathcal{X}}(d\theta),$$

$$\rho_{\mathcal{Y}}(d\theta) = |\mathcal{Y}|^{-1/2} \frac{\Gamma\left(\frac{|\mathcal{Y}|}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^{|\mathcal{Y}|}} \prod_{y \in \mathcal{Y}} \theta(y)^{-1/2} \lambda_{\mathcal{Y}}(d\theta),$$

où  $\lambda_{\mathcal{X}}$  (respectivement  $\lambda_{\mathcal{Y}}$ ) désigne la mesure de Lebesgue sur  $\mathcal{M}_+^1(\mathcal{X})$ .

Définissons alors la loi de codage  $Q$  par la formule

$$\begin{aligned} Q((x_i)_{i=1}^N, (y_i)_{i=1}^N | x_0) \\ = \int Q_{p,q}((x_i)_{i=1}^N, (y_i)_{i=1}^N | x_0) \bigotimes_{x \in \mathcal{X}} \rho_{\mathcal{X}}(dp(x, \cdot)) \bigotimes_{y \in \mathcal{Y}} \rho_{\mathcal{Y}}(dq(y, \cdot)). \end{aligned}$$

Introduisons les compteurs

$$\begin{aligned} a(x, x') &= \sum_{i=1}^N \mathbb{1}(x_{i-1} = x, x_i = x') \\ a(x) &= \sum_{x' \in \mathcal{X}} a(x, x'), \\ b(x, y) &= \sum_{i=1}^N \mathbb{1}(x_i = x, y_i = y), \\ b(x) &= \sum_{y \in \mathcal{Y}} b(x, y), \end{aligned}$$

dans le but de calculer  $Q$  plus explicitement :

$$\begin{aligned} Q((x_i)_{i=1}^N, (y_i)_{i=1}^N | x_0) \\ = \prod_{x \in \mathcal{X}} \int \left( \prod_{x' \in \mathcal{X}} p(x, x')^{a(x, x')} \right) \rho_{\mathcal{X}}(dp(x, \cdot)) \int \left( \prod_{y \in \mathcal{Y}} q(x, y)^{b(x, y)} \right) \rho_{\mathcal{Y}}(dq(x, \cdot)) \\ = \prod_{x \in \mathcal{X}} \frac{\Gamma\left(\frac{|\mathcal{X}|}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^{|\mathcal{X}|}} \frac{\prod_{x' \in \mathcal{X}} \Gamma\left(a(x, x') + \frac{1}{2}\right)}{\Gamma\left(a(x) + \frac{|\mathcal{X}|}{2}\right)} \\ \times \frac{\Gamma\left(\frac{|\mathcal{Y}|}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^{|\mathcal{Y}|}} \frac{\prod_{y \in \mathcal{Y}} \Gamma\left(b(x, y) + \frac{1}{2}\right)}{\Gamma\left(b(x) + \frac{|\mathcal{Y}|}{2}\right)} \\ \geq \sup_{(p,q) \in \Theta} Q_{p,q}((x_i)_{i=1}^N, (y_i)_{i=1}^N | x_0) \exp\left(-\sum_{x \in \mathcal{X}} \psi_{|\mathcal{X}|}(a(x)) + \psi_{|\mathcal{Y}|}(b(x))\right). \end{aligned}$$

d'après le lemme 4.1, la proposition 4.1 et le corollaire 4.1 du chapitre 1 du cours. De plus par concavité des fonctions  $\psi_d$ ,

$$\sum_{x \in \mathcal{X}} \psi_{|\mathcal{X}|}(a(x)) = |\mathcal{X}| \sum_{x \in \mathcal{X}} \frac{1}{|\mathcal{X}|} \psi_{|\mathcal{X}|}(a(x)) \leq |\mathcal{X}| \psi_{|\mathcal{X}|}\left(\sum_{x \in \mathcal{X}} \frac{a(x)}{|\mathcal{X}|}\right) = |\mathcal{X}| \psi_{|\mathcal{X}|}\left(\frac{N}{|\mathcal{X}|}\right)$$

De même

$$\sum_{x \in \mathcal{X}} \psi_{|\mathcal{Y}|}(b(x)) \leq |\mathcal{X}| \psi_{|\mathcal{Y}|}\left(\frac{N}{|\mathcal{X}|}\right).$$

Ainsi

$$\begin{aligned} -\log\left[Q((x_i)_{i=1}^N, (y_i)_{i=1}^N | x_0)\right] \\ \leq \inf_{\theta \in \Theta} -\log\left[Q_{\theta}((x_i)_{i=1}^N, (y_i)_{i=1}^N | x_0)\right] + |\mathcal{X}| \left[\psi_{|\mathcal{X}|}\left(\frac{N}{|\mathcal{X}|}\right) + \psi_{|\mathcal{Y}|}\left(\frac{N}{|\mathcal{X}|}\right)\right]. \end{aligned}$$

(N.B.: il y avait malheureusement une coquille dans l'énoncé, où on pouvait lire  $\psi_{|y|} \left( \frac{N}{|y|} \right)$  au lieu de  $\psi_{|y|} \left( \frac{N}{|\mathcal{X}|} \right)$ .)

D'après le théorème 5.1 du premier chapitre du cours, pour  $n$  suffisamment grand,

$$\begin{aligned} \frac{d-1}{2} \log(n) + \min \left\{ \log(d), -\frac{d-1}{2} \log \left( \frac{d-2}{2} \right) + \frac{(d-1)^2}{4n} + \frac{d}{2} \right\} \\ \leq \psi_d(n) \leq \frac{d-1}{2} \log(n) + \log(d). \end{aligned}$$

D'où

$$\lim_{N \rightarrow \infty} \frac{\gamma(N)}{\log(N)} = \left( |\mathcal{X}| + |\mathcal{Y}| - 2 \right) \frac{|\mathcal{X}|}{2}.$$

**IV.1.2)** Pour tout  $\theta \in \Theta$ , tout  $(x_i)_{i=1}^N \in \mathcal{Y}^N$ ,

$$Q((x_i)_{i=1}^N, (y_i)_{i=1}^N | x_0) \geq Q_\theta((x_i)_{i=1}^N, (y_i)_{i=1}^N | x_0) \exp[-\gamma(N)],$$

d'où en sommant ces inégalités :

$$\begin{aligned} Q((y_i)_{i=1}^N | x_0) &= \sum_{(x_i)_{i=1}^N \in \mathcal{X}^N} Q((x_i)_{i=1}^N, (y_i)_{i=1}^N | x_0) \\ &\geq \sum_{(x_i)_{i=1}^N \in \mathcal{X}^N} Q_\theta((x_i)_{i=1}^N, (y_i)_{i=1}^N | x_0) \exp[-\gamma(N)] = \exp[-\gamma(N)] Q_\theta((y_i)_{i=1}^N | x_0). \end{aligned}$$

**IV.2)** Une solution simple (mais légèrement sous-optimale) consiste à poser

$$Q'((x_i)_{i=1}^N, (y_i)_{i=1}^N) = |\mathcal{X}|^{-1} |\mathcal{Y}|^{-1} Q((x_i)_{i=2}^N, (y_i)_{i=2}^N | x_1).$$

On obtient immédiatement

$$\begin{aligned} -\log \left[ Q'((x_i)_{i=1}^N, (y_i)_{i=1}^N) \right] &\leq \inf_{\theta \in \Theta} -\log \left[ Q_\theta((x_i)_{i=2}^N, (y_i)_{i=2}^N | x_1) \right] \\ &\quad + \gamma(N-1) + \log(|\mathcal{X}|) + \log(|\mathcal{Y}|) \\ &\leq \inf_{\theta \in \Theta} -\log \left[ Q_\theta((x_i)_{i=1}^N, (y_i)_{i=1}^N | x_0) \right] \\ &\quad + \gamma(N-1) + \log(|\mathcal{X}|) + \log(|\mathcal{Y}|). \end{aligned}$$

On en déduit en sommant en  $(x_i)_{i=1}^N$  que

$$-\log \left[ Q'((y_i)_{i=1}^N) \right] \leq \inf_{\theta \in \Theta, x_0 \in \mathcal{X}} -\log \left[ Q_\theta((y_i)_{i=1}^N | x_0) \right] + \gamma'(N),$$

où  $\gamma'(N) = \gamma(N-1) + \log(|\mathcal{X}|) + \log(|\mathcal{Y}|)$ .

## CINQUIÈME PARTIE

Il suffit de montrer que

$$\mathbb{E} \left\{ -\log \left[ \hat{Q}(Y_{N+1} | X_{N+1}) \right] \right\} \leq \inf_{\theta \in \Theta} \mathbb{E} \left\{ -\log \left[ Q_\theta(Y_{N+1} | X_{N+1}) \right] \right\} + \gamma(N),$$

où  $(X_{N+1}, Y_{N+1}) \sim P$  est un couple de variables aléatoires indépendantes de  $(X_i, Y_i)_{i=1}^N$  et où

$$\gamma(N) = \frac{\log(N-K+2)}{\beta(N-K+1)} + \frac{2}{K+1}.$$

Posons

$$\mathcal{E}(\eta) = \log \left( \sum_{j=K+1}^{N+2} \left( \prod_{i=K+1}^N \hat{p}_{\mathbb{1}(X_i \geq X_j)}^{X_j}(Y_i) \right)^\beta \left( \hat{p}_{\mathbb{1}(x \geq X_j)}^{X_j}(Y_{N+1}) \right)^\eta \right)$$

et remarquons que  $\frac{1}{K+2} \leq \hat{p}_{\mathbb{1}(x \geq X_j)}^{X_j}(Y_{N+1}) \leq 1$ . La preuve du théorème 1.1 s'applique sans modification à la fonction  $\mathcal{E}$  et montre que, sous l'hypothèse sur  $\beta$  contenue dans l'énoncé,  $\mathcal{E}(1) - \mathcal{E}(0) \geq \mathcal{E}'(\beta)$ , qui peut encore s'écrire

$$\begin{aligned} & -\log[\hat{Q}(Y_{N+1} | X_{N+1})] \\ & \leq \frac{-\sum_{j=K+1}^{N+2} \left( \left( \prod_{i=K+1}^{N+1} \hat{p}_{\mathbb{1}(X_i \geq X_j)}^{X_j}(Y_i) \right)^\beta \log \left[ \hat{p}_{\mathbb{1}(X_{N+1} \geq X_j)}^{X_j}(Y_{N+1}) \right] \right)}{\sum_{j=K+1}^{N+2} \left( \prod_{i=K+1}^{N+1} \hat{p}_{\mathbb{1}(X_i \geq X_j)}^{X_j}(Y_i) \right)^\beta}. \end{aligned}$$

On en déduit en utilisant l'échangeabilité de  $P^{\otimes(N+1)}$  que

$$\begin{aligned} & \mathbb{E} \left\{ -\log[\hat{Q}(Y_{N+1} | X_{N+1})] \right\} \\ & \leq \mathbb{E} \left\{ \frac{-\sum_{j=K+1}^{N+2} \left( \left( \prod_{i=K+1}^{N+1} \hat{p}_{\mathbb{1}(X_i \geq X_j)}^{X_j}(Y_i) \right)^\beta \log \left( \prod_{i=K+1}^{N+1} \hat{p}_{\mathbb{1}(X_i \geq X_j)}^{X_j}(Y_i) \right) \right)}{(N-K+1) \sum_{j=K+1}^{N+2} \left( \prod_{i=K+1}^{N+1} \hat{p}_{\mathbb{1}(X_i \geq X_j)}^{X_j}(Y_i) \right)^\beta} \right\} \\ & \leq \mathbb{E} \left\{ \inf_{j=K+1, \dots, N+2} \frac{-\log \left[ \prod_{i=K+1}^{N+1} \hat{p}_{\mathbb{1}(X_i \geq X_j)}^{X_j}(Y_i) \right]}{N-K+1} + \frac{\log(N-K+2)}{\beta(N-K+1)} \right\} \end{aligned}$$

[d'après le lemme 1.1 du chapitre 4 du cours]

$$\begin{aligned} & = \mathbb{E} \left\{ \inf_{\tau \in [0,1]} \frac{-\log \left[ \prod_{i=K+1}^{N+1} \hat{p}_{\mathbb{1}(X_i \geq \tau)}^\tau(Y_i) \right]}{N-K+1} \right\} + \frac{\log(N-K+2)}{\beta(N-K+1)} \\ & \leq \inf_{\tau \in [0,1]} -\mathbb{E} \left\{ \frac{\log \left[ \prod_{i=K+1}^{N+1} \hat{p}_{\mathbb{1}(X_i \geq \tau)}^\tau(Y_i) \right]}{N-K+1} \right\} + \frac{\log(N-K+2)}{\beta(N-K+1)} \\ & = \inf_{\tau \in [0,1]} -\mathbb{E} \left\{ \log \left[ \hat{p}_{\mathbb{1}(X_{K+1} \geq \tau)}^\tau(Y_{K+1}) \right] \right\} + \frac{\log(N-K+2)}{\beta(N-K+1)}. \end{aligned}$$

Pour tout  $(\sigma, y) \in \{0; 1\}^2$ , posons

$$\begin{aligned} c(\sigma, y) &= \sum_{i=1}^{K+1} \mathbb{1}[\mathbb{1}(X_i \geq \tau) = \sigma] \mathbb{1}(Y_i = y) \\ c(\sigma) &= c(\sigma, 0) + c(\sigma, 1). \end{aligned}$$

Avec ces notations, en utilisant l'échangeabilité de  $P^{\otimes(K+1)}$ , on peut continuer comme dans la preuve de la proposition 2.1 du chapitre 3 du cours :

$$\begin{aligned} & - \mathbb{E} \left\{ \log \left[ \hat{p}_{\mathbb{1}(X_{K+1} \geq \tau)}^{\tau}(Y_{K+1}) \right] \right\} \\ &= -\frac{1}{K+1} \mathbb{E} \left\{ \log \left[ \prod_{(\sigma, y) \in \{0; 1\}^2} \left( \frac{c(\sigma, y)}{c(\sigma) + 1} \right)^{c(\sigma, y)} \right] \right\} \\ &= \frac{1}{K+1} \mathbb{E} \left\{ -\log \left[ \prod_{(\sigma, y) \in \{0; 1\}^2} \left( \frac{c(\sigma, y)}{c(\sigma)} \right)^{c(\sigma, y)} \right] \right\} \\ &\quad + \frac{1}{K+1} \sum_{\sigma \in \{0; 1\}} c(\sigma) \log \left( 1 + \frac{1}{c(\sigma)} \right) \\ &\leq \frac{1}{K+1} \mathbb{E} \left\{ \inf_{p_0, p_1} -\log \left[ \prod_{i=1}^{K+1} p_{\mathbb{1}(X_i \geq \tau)}^{\tau}(Y_i) \right] \right\} + \frac{2}{K+1} \\ &\leq \inf_{p_0, p_1} \frac{1}{K+1} \mathbb{E} \left\{ -\log \left[ \prod_{i=1}^{K+1} p_{\mathbb{1}(X_i \geq \tau)}^{\tau}(Y_i) \right] \right\} + \frac{2}{K+1} \\ &= \inf_{p_0, p_1} \mathbb{E} \left\{ -\log \left[ p_{\mathbb{1}(X_{K+1} \geq \tau)}^{\tau}(Y_{K+1}) \right] \right\} + \frac{2}{K+1}. \end{aligned}$$

En combinant les deux inégalités démontrées, on obtient

$$\begin{aligned} \mathbb{E} \left\{ -\log [\hat{Q}(Y_{N+1} | X_{N+1})] \right\} &\leq \inf_{(\tau, p_0, p_1) \in \Theta} \mathbb{E} \left\{ -\log \left[ p_{\mathbb{1}(X_{K+1} \geq \tau)}^{\tau}(Y_{K+1}) \right] \right\} \\ &\quad + \frac{2}{K+1} + \frac{\log(N - K + 2)}{\beta(N - K + 1)}, \end{aligned}$$

ce qui termine la preuve, puisque

$$p_{\mathbb{1}(X_{K+1} \geq \tau)}^{\tau}(Y_{K+1}) = Q_{\tau, p_0, p_1}(Y_{K+1} | X_{K+1}).$$