M2 Mathématiques fondamentales
Concentration de la mesure et statistique non-asymptotique
Examen final

La durée de l'examen est de 1h30. Tous les documents sont autorisés. L'utilisation de résultats existants est permise, à condition de les mentionner clairement (indiquer la référence dans le polycopié ou rappeler le résultat s'il provient d'une autre source).

La clarté ainsi que la présentation des réponses seront prises en compte dans la notation.

L'exercice 2 est, dans une certain sens, plus élémentaire que l'exercice 1. L'exercice 3 est plus compliqué, notamment les questions 4 et 5.

Exercice 1 Soient f et q deux densités sur $\mathbb{R}^d$, c'est-à-dire $f, q : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ et

$$\int_{\mathbb{R}^d} f(x)\,dx = \int_{\mathbb{R}^d} q(x)\,dx = 1.$$

On suppose la condition de domination suivante $\forall x \in \mathbb{R}^d$, $f(x) \leq c\,q(x)$, où $c > 0$ est une constante. Soient $Z, Z_1, \ldots, Z_n$ des variables aléatoires i.i.d. de densité q. On définit les poids d'importance

$$w_i = \frac{f(Z_i)}{q(Z_i)}, \qquad i = 1, \ldots, n.$$

On souhaite estimer la fonction de répartition F de f, définie par

$$F(t) = \int_{(-\infty, t_1] \times \cdots \times (-\infty, t_d]} f(z)\,dz, \qquad t = (t_1, \ldots, t_d) \in \mathbb{R}^d.$$

La relation $Z_i \leq t$ est entendue composante par composante. On considère l'estimateur pondéré

$$\hat{F}_w(t) = \frac{1}{n} \sum_{i=1}^{n} w_i \mathbf{1}_{\{Z_i \leq t\}}, \qquad t \in \mathbb{R}^d.$$

Soient $\eta_1, \ldots, \eta_n$ des variables aléatoires i.i.d., indépendantes des Z_i, telles que $\mathbb{P}(\eta_i = 1) = \mathbb{P}(\eta_i = -1) = \frac{1}{2}$ (variables de Rademacher).

1) Soit l'estimateur non-pondéré:

$$\hat{F}(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{Z_i \leq t\}}, \qquad t \in \mathbb{R}^d.$$

L'estimateur $\hat{F}$ permet-il d'estimer F? Démontrer que

$$\mathbb{E}\left[\sup_{t \in \mathbb{R}^d} |\hat{F}(t) - \mathbb{E}[\hat{F}(t)]|\right] \leq 2\sqrt{2n^{-1}\log(2(n+1)^d)}$$

On montrera d'abord une inégalité de symetrisation impliquant les variables de Rademacher:

$$R_t = \sum_{i=1}^{n} \eta_i \mathbf{1}_{\{Z_i \leq t\}}.$$

Puis on obtiendra la borne donnée. On pourra admettre que la dimension de Vapnik des rectangles $]-\infty, t]$ est d. On admettra aussi que si $W_1, \ldots, W_M$ sont sous-Gaussiennes de facteur commun v, on a que $\mathbb{E}[\max_{k=1,\ldots,M} |W_k|] \leq \sqrt{2v \log(2M)}$.

2) Montrer que $\hat{F}_w(t)$ est un estimateur sans biais de $F(t)$ et calculer $\mathrm{Var}(\hat{F}_w(t))$. On montrera que

$$\mathrm{Var}\Big(\hat{F}_w(t)\Big) \leq \frac{c}{n} F(t).$$

3) On définit

$$R_t^w = \sum_{i=1}^{n} \eta_i \, w_i \mathbf{1}_{\{Z_i \leq t\}}.$$

Montrer que, conditionnellement à $Z_1, \ldots, Z_n$, la variable aléatoire R_t^w est sous-Gaussienne avec facteur de sous-Gaussianité majoré par $c^2 n$.

4) Montrer un résultat de type Glivenko-Cantelli (en espérance) pour $\hat{F}_w$. Discuter de l'influence de la constante c sur la qualité de l'estimation. L'approche par entropy-metrique de Dudley, permet-elle d'améliorer l'analyse de $\hat{F}_w$?

Exercice 2 Soit $(B_i)_{i=1,\ldots,n}$ est une collection de variables indépendantes chacune de loi de Bernoulli de parametre p. On rappelle l'inégalité suivante pour tout $\eta \in (0,1)$ et pour $\lambda = -\log(1-\eta)$,

$$\exp(-\lambda) - 1 + \lambda(1-\eta) \leq -\eta^2/2.$$

1) Montrer que pour tout $\eta \in (0,1)$, et pour un certain $\lambda > 0$, on a

$$\mathbb{E}\left[\exp\left(\lambda\left(\left(np - \sum_{i=1}^{n} B_i\right) - np\eta\right)\right)\right] \leq \exp(-\eta^2 np/2).$$

2) En déduire que, pour tout $t \in (0, \sqrt{np})$, et pour un certain $\lambda > 0$,

$$\mathbb{E}\left[\exp\left(\lambda\left(\left(np - \sum_{i=1}^{n} B_i\right) - t\sqrt{np}\right)\right)\right] \leq \exp(-t^2/2).$$

3) En déduire une inégalité avec probabilité $1-\delta$, sous laquelle $\sum_{i=1}^{n} B_i$ est bornée inférieurement.

Exercice 3 Soit $\mathcal{A}$ une classe d'ensembles de S. Soit $(Z_1, \ldots, Z_n)$ une collection de variables aléatoires indépendantes et identiquement distribués de loi P sur S. Soit $\sigma^*, (\sigma_i^*)_{i=1,\ldots,n}$ une collection de variables aléatoires indépendantes et identiquement distribuées de loi uniforme sur $\{1, \ldots, n\}$. Les variables aléatoires $\sigma^*, (\sigma_i^*)_{i=1,\ldots,n}$ sont indépendentes des $(Z_1, \ldots, Z_n)$. On définit, pour tout $A \in \mathcal{A}$,

$$Z_A^* = \sum_{i=1}^{n} (1_A(Z_i) - 1_A(Z_{\sigma_i^*})).$$

On souhaite obtenir une inégalité pour Z_A^* normalisée par $\sqrt{P_n(A)}$ et uniforme sur $A \in \mathcal{A}$. On notera dans la suite $\mathbb{E}^*$ l'espérance sur $(\sigma_i^*)_{i=1,\dots,n}$, conditionnellement à $(Z_1, \dots, Z_n)$, et $\mathbb{E}$ l'espérance portant sur $(Z_1, \dots, Z_n)$. On introduit aussi la notation $Z^* = Z_{\sigma^*}$ ainsi que $Z_i^* = Z_{\sigma_i^*}$.

1) Donner la loi de $1_A(Z^*)$ conditionnellement à $(Z_1, \dots, Z_n)$. On pourra introduire la notation $P_n(A) = n^{-1} \sum_{i=1}^n 1_A(Z_i)$.

2) Montrer que, pour tout $A \in \mathcal{A}$, Z_A^* est une variable aléatoire centrée, conditionnellement à $(Z_1, \dots, Z_n)$. Calculer sa variance, conditionnellement à $(Z_1, \dots, Z_n)$. Montrer que $\mathbb{E}^*[\sup_{A\in\mathcal{A}} Z_A^*] \geq 0$.

3) Montrer que

$$\sum_{i=1}^n 1_A(Z_i^*) = \sum_{i=1}^n w_i^* 1_A(Z_i)$$

avec (w_i^*) une collection de variable aléatoire indépendante des (Z_i). En déduire que

$$\sup_{A\in\mathcal{A}} Z_A^* \leq \max_{v\in\mathcal{V}} \sum_{i=1}^n (1 - w_i^*) v_i$$

où $\mathcal{V}$ est un ensemble fini de $\mathbb{R}^n$ à préciser.

4) Pour $t > 0$ notons $\mathcal{A}_t = \{A \in \mathcal{A} : nP_n(A) > t^2\}$ et remarquons que cet ensemble est vide quand $t \geq \sqrt{n}$. Montrer que, pour $\lambda_A = -\log(1 - t/\sqrt{nP_n(A)})$,

$$\mathbb{E}^* \left[\exp\left(\sup_{A\in\mathcal{A}_t} \lambda_A \left(\{Z_A^* - t\sqrt{nP_n(A)}\} \right) \right) \right] \leq \mathbb{S}_{\mathcal{A}_t}(n) \exp(-t^2/2).$$

(on pourra s'assurer du passage au maximum puis appliquer un des résultats de l'exercice précédent en précisant chacune des quantités impliquées).

5) En déduire que pour tout $t > 0$,

$$\mathbb{P}^* \left[\sup_{A\in\mathcal{A}} \{Z_A^* - t\sqrt{nP_n(A)}\} > 0 \right] \leq \mathbb{S}_{\mathcal{A}}(n) \exp(-t^2/2).$$

(on pourra commencer par montrer que l'événement $\{\sup_{A\in\mathcal{A}_t}(Z_A^* - t\sqrt{nP_n(A)}) > 0\}$ est identique à l'évènement $\{\sup_{A\in\mathcal{A}}(Z_A^* - t\sqrt{nP_n(A)}) > 0\}$).

6) Montrer que avec probabilité $1 - \delta$: pour tout $A \in \mathcal{A}$,

$$Z_A^* \leq \sqrt{2nP_n(A) \log(\mathbb{S}_{\mathcal{A}}(n)/\delta)}.$$