Inégalités de concentration
Sorbonne Université
2019-2020

Examen

1. Échauffement

Les questions de cet exercice sont complètement indépendantes.

(1) Montrer que si X est une variable aléatoire réelle positive avec $\mathbf{E}X^2 < \infty$, alors pour tout $\lambda \geq 0$,

$$\log \mathbf{E}e^{-\lambda(X-\mathbf{E}X)} \leq \frac{\lambda^2 \mathbf{E}X^2}{2},$$

et en déduire que pour tout $t \geq 0$, $\mathbf{P}(X - \mathbf{E}X \leq -t) \leq e^{-\frac{t^2}{2\mathbf{E}X^2}}$.

(2) Soient $X_1, \ldots, X_n$ des variables aléatoires réelles indépendantes symétriques (i.e. X_i et $-X_i$ ont la même loi, ou encore $X_i = \varepsilon_i |X_i|$ avec $\varepsilon_i = \mathrm{signe}(X_i)$ valant $+1$ ou -1 avec probabilité $1/2$ et ε_i indépendante de $|X_i|$). On suppose de plus que pour tout $i = 1, \ldots, n$, $\mathbf{P}(X_i = 0) = 0$. Montrer que pour tout $t \geq 0$,

$$\mathbf{P}\left(\frac{X_1 + \cdots + X_n}{\sqrt{X_1^2 + \cdots + X_n^2}} \geq t\right) \leq \exp\left\{-\frac{t^2}{2}\right\}.$$

(3) On effectue des tirages indépendants $X_1, X_2, \ldots$ uniformément distribués dans $\{1, \ldots, n\}$. On s'intéresse à la variable T_n égale au premier temps où tous les éléments ont été tirés au moins une fois, soit

$$T_n = \inf\{t \geq 1, \{X_1, \ldots, X_t\} = \{1, \ldots, n\}\}.$$

(a) Écrire T_n comme une somme de variables aléatoires géométriques indépendantes dont on précisera les paramètres.

(b) Donner un équivalent asymptotique (quand $n \to +\infty$) de l'espérance et de la variance de T_n.

(c) Soit $X \sim$ Geom(p), pour $p \in]0,1[$. Montrer que pour tout $\lambda < \log\left(\frac{1}{1-p}\right)$,

$$\log \mathbf{E}e^{\lambda(X-\mathbf{E}X)} = \log\left(\frac{p}{e^{-\lambda} - 1 + p}\right) - \frac{\lambda}{p},$$

et en déduire que X est sous-Gamma à droite avec facteur variance $\frac{1}{p^2}$ et facteur d'échelle $\frac{1}{p}$ (on pourra utiliser $e^{-\lambda} \geq 1 - \lambda$, et que pour tout $u \in [0,1[$, $-u - \log(1-u) \leq \frac{u^2}{2(1-u)}$), et sous-gaussienne à gauche avec facteur d'échelle $\frac{1}{p^2}$.

(d) Montrer que pour tout $\lambda \in [0, 1/n[$,

$$\log \mathbf{E}e^{\lambda(T_n - \mathbf{E}T_n)} \leq \frac{\pi^2 n^2 \lambda^2}{12(1-n\lambda)},$$

et que pour tout $\lambda \leq 0$,

$$\log \mathbf{E}e^{\lambda(T_n - \mathbf{E}T_n)} \leq \frac{\pi^2 n^2 \lambda^2}{12}.$$

2. Estimateur à noyau

Soient $X_1, \ldots, X_n$ des variables aléatoires réelles indépendantes, de densité φ par rapport à la mesure de Lebesgue. Pour $x \in \mathbb{R}$, l'estimateur à noyau de $\varphi(x)$ est défini comme

$$\varphi_n(x, X_1, \ldots, X_n) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right),$$

avec $h_n > 0$ et K une fonction positive telle que $\int_{\mathbb{R}} K(u)\mathrm{d}u = 1$. On s'intéresse à l'erreur L^1 :

$$Z = f(X_1,\dots,X_n) = \int_{\mathbb{R}} |\varphi_n(x, X_1,\dots,X_n) - \varphi(x)|\,\mathrm{d}x\,.$$

(1) Montrer que pour tous $x_1,\dots,x_i,x_i',\dots,x_n$, on a

$$\left|f(x_1,\dots,x_{i-1},x_i,x_{i+1},\dots,x_n) - f(x_1,\dots,x_{i-1},x_i',x_{i+1},\dots,x_n)\right| \le \frac{2}{n}\,.$$

(2) En déduire que $\mathbf{Var}(Z) \le \frac{1}{n}$ et que pour tout $t \ge 0$,

$$\mathbf{P}\left(Z - \mathbf{E}Z \ge t\right) \le \exp\left\{-\frac{nt^2}{2}\right\}\,.$$

3. Moyenne de Rademacher conditionnelle

Soient $X_1,\dots,X_n$ des variables aléatoires indépendantes à valeurs dans la sphère $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d,\ \|x\| \le 1\}$ (où $\|\cdot\|$ est la norme euclidienne), et $\varepsilon_1,\dots,\varepsilon_n$ des variables aléatoires i.i.d. de loi de Rademacher (i.e. égales à -1 ou 1 avec probabilité $1/2$), indépendantes de $(X_1,\dots,X_n)$. On s'intéresse à la variable

$$Z = \mathbf{E}\left[\left\|\sum_{i=1}^n \varepsilon_i X_i\right\| \,\Big|\, X_1,\dots,X_n\right]\,.$$

On écrit $Z = f(X_1,\dots,X_n)$ avec, pour $x_1,\dots,x_n \in \mathbb{S}^{d-1}$, $f(x_1,\dots,x_n) = \mathbf{E}\left[\|\sum_{i=1}^n \varepsilon_i x_i\|\right]$.

(1) Montrer que la fonction f est auto-bornée. On pourra considérer les fonctions f_i données par

$$f_i(x_1,\dots,x_{i-1},x_{i+1},\dots,x_n) = \mathbf{E}\left[\Big\|\sum_{\substack{j=1\\ j\neq i}}^n \varepsilon_j x_j\Big\|\right]\,,$$

et utiliser le fait que pour $y \in \mathbb{R}^d$, $\|y\| = \sup_{\alpha\in\mathbb{R}^d,\|\alpha\|\le 1}\langle \alpha, y\rangle$.

(2) En déduire que pour tout $t \ge 0$,

$$\mathbf{P}\left(Z - \mathbf{E}Z \ge t\right) \le \exp\left\{-\frac{t^2}{2\left(\mathbf{E}Z + \frac{t}{3}\right)}\right\} \quad \text{et} \quad \mathbf{P}\left(Z - \mathbf{E}Z \le -t\right) \le \exp\left\{-\frac{t^2}{2\mathbf{E}Z}\right\}\,.$$

4. Une amélioration de l'inégalité de McDiarmid

Soit $f : \mathcal{X}^n \to \mathbb{R}$ une fonction mesurable et $Z = f(X_1,\dots,X_n)$ avec $X_1,\dots,X_n$ des variables aléatoires indépendantes à valeurs dans $\mathcal{X}$.

(1) En utilisant la sous-additivité de l'entropie, montrer que pour tout $\lambda \in \mathbb{R}$,

$$\mathbf{Ent}\left[e^{\lambda Z}\right] \le \sum_{i=1}^n \mathbf{E}\left[e^{\lambda Z}\phi\left(-\lambda(Z - Z_i')\right)\right]\,,$$

avec $Z_i' = f(X_1,\dots,X_{i-1},X_i',X_{i+1},\dots,X_n)$ où X_i' est une copie indépendante de X_i, et $\phi(z) = e^z - z - 1$.

(2) En écrivant $Z - Z_i' = (Z - Z_i')_+ - (Z_i' - Z)_+$ et le fait que Z et Z_i' sont échangeables, montrer que pour tout $\lambda \in \mathbb{R}$,

$$\mathbf{E}\left[e^{\lambda Z}\phi\left(-\lambda(Z - Z_i')\right)\right] = \mathbf{E}\left[e^{\lambda Z}\tau\left(-\lambda(Z - Z_i')_+\right)\right]\,,$$

où $\tau(z) = z(e^z - 1)$.

(3) En déduire que pour tout $\lambda \ge 0$, $\mathbf{Ent}\left[e^{\lambda Z}\right] \le \sum_{i=1}^n \mathbf{E}\left[\lambda^2 e^{\lambda Z}(Z - Z_i')_+^2\right]$.

(4) Montrer que s'il existe $v > 0$ tel que

$$\mathbf{E}\left[\sum_{i=1}^n (Z - Z_i')_+^2 \mid X_1,\dots,X_n\right] \le v\,,$$

alors pour tout $t \ge 0$, $\mathbf{P}\left(Z - \mathbf{E}Z \ge t\right) \le \exp\left\{-\frac{t^2}{4v}\right\}$.