Master 2 Probabilités + M2A Algorithmes et Apprentissage
Sorbonne Université

Inégalités de concentration

Références :
Boucheron, Lugosi, Massart, *Concentration inequalities* [3].
Devroye, Györfi, Lugosi, *A probabilistic theory of pattern recognition* [7].
Dubhashi, Panconesi, *Concentration of measure for the analysis of randomised algorithms* [8].
Ledoux, *The concentration of measure phenomenon* [15].
Tropp, *An introduction to matrix concentration inequalities* [24].
Vershynin, *High-dimensional probability* [25].

Anna Ben-Hamou
anna.ben-hamou@upmc.fr

Table des matières

Chapitre 1

Variance, entropie, influences

Soient $X_1, \dots, X_n$ des variables aléatoires indépendantes définies sur un espace mesurable $(\Omega, \mathcal{F}, \mathbf{P})$, à valeurs dans un espace mesurable $\mathcal{X}$, et soit $Z = f(X_1, \dots, X_n)$ avec $f : \mathcal{X}^n \to \mathbb{R}$ une fonction mesurable. Si l'on s'intéresse à la façon dont Z se concentre autour de son espérance $\mathbf{E}Z$, une première quantité que l'on peut étudier est la variance. Si la fonction f correspond à une somme, alors le problème est simplement celui des variances individuelles des variables :

$$\mathbf{Var}(Z) = \sum_{i=1}^{n} \mathbf{Var}(X_i)\,.$$

Mais que peut-on dire de la variance d'une fonction éventuellement bien plus complexe que la somme ? Notons que l'on peut toujours décomposer $Z - \mathbf{E}Z$ comme une somme d'incréments de martingale pour la filtration de Doob et utiliser l'orthogonalité de ces incréments. Plus précisément, notons $\mathbf{E}_i = \mathbf{E}[\cdot \mid X_1, \dots, X_i]$ et $\mathbf{E}_0 = \mathbf{E}$. Alors

$$Z - \mathbf{E}Z = \sum_{i=1}^{n} \mathbf{E}_i Z - \mathbf{E}_{i-1} Z\,,$$

et

$$\mathbf{Var}(Z) = \sum_{i=1}^{n} \mathbf{E}\left[(\mathbf{E}_i Z - \mathbf{E}_{i-1} Z)^2\right] + 2 \sum_{i<j} \mathbf{E}\left[(\mathbf{E}_j Z - \mathbf{E}_{j-1} Z)(\mathbf{E}_i Z - \mathbf{E}_{i-1} Z)\right]\,.$$

En remarquant que pour $j > i$, $\mathbf{E}_i[\mathbf{E}_j Z - \mathbf{E}_{j-1} Z] = 0$, on voit que les covariances sont nulles et l'on obtient

$$\mathbf{Var}(Z) = \sum_{i=1}^{n} \mathbf{E}\left[(\mathbf{E}_i Z - \mathbf{E}_{i-1} Z)^2\right]\,.$$

Jusqu'ici, on n'a pas utilisé l'hypothèse d'indépendance sur les $X_1, \dots, X_n$. Celle-ci intervient maintenant pour pouvoir écrire

$$\mathbf{E}_{i-1} Z = \mathbf{E}_i \mathbf{E}^{(i)} Z\,,$$

où $\mathbf{E}^{(i)} = \mathbf{E}[\cdot \mid X^{(i)}]$ avec $X^{(i)} = (X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$. C'est l'observation-clé dans la preuve du résultat principal de ce chapitre, l'inégalité d'Efron–Stein.

1. L'inégalité d'Efron–Stein

Proposition 1.1 (Inégalité d'Efron–Stein)**.** *Soient $X_1, \dots, X_n$ des variables indépendantes à valeurs dans un espace mesurable $\mathcal{X}$, et soit $Z = f(X_1, \dots, X_n)$ une fonction mesurable. Alors*

$$\mathbf{Var}(Z) \leq \sum_{i=1}^{n} \mathbf{E}\left[\left(Z - \mathbf{E}^{(i)} Z\right)^2\right]\,.$$

Preuve de la Proposition 1.1. Par le théorème de Fubini, si P_i est la loi de X_i, on a

$$\begin{aligned}\mathbf{E}_i\mathbf{E}^{(i)}Z &= \mathbf{E}_i\left[\int_{\mathcal{X}} f(X_1,\dots,X_{i-1},x_i,X_{i+1},\dots,X_n)\mathrm{d}P_i(x_i)\right]\\ &= \int_{\mathcal{X}^{n-i+1}} f(X_1,\dots,X_{i-1},x_i,x_{i+1},\dots,x_n)\mathrm{d}P_i(x_i)\dots\mathrm{d}P_n(x_n)\\ &= \mathbf{E}_{i-1}Z\,.\end{aligned}$$

Ainsi, en utilisant l'inégalité de Jensen conditionnellement à $X_1,\dots,X_i$,

$$(\mathbf{E}_iZ-\mathbf{E}_{i-1}Z)^2=\left(\mathbf{E}_i[Z-\mathbf{E}^{(i)}Z]\right)^2\le\mathbf{E}_i\left[(Z-\mathbf{E}^{(i)}Z)^2\right]\,,$$

et

$$\mathbf{Var}(Z)=\sum_{i=1}^n\mathbf{E}\left[(\mathbf{E}_iZ-\mathbf{E}_{i-1}Z)^2\right]\le\sum_{i=1}^n\mathbf{E}\left[\mathbf{E}_i\left[(Z-\mathbf{E}^{(i)}Z)^2\right]\right]=\sum_{i=1}^n\mathbf{E}\left[\left(Z-\mathbf{E}^{(i)}Z\right)^2\right]\,.$$

■

Remarque 1.1. La borne $v=\sum_{i=1}^n\mathbf{E}\left[\left(Z-\mathbf{E}^{(i)}Z\right)^2\right]$ de l'inégalité d'Efron–Stein peut se ré-écrire de plusieurs façons. Rappelons que si X est une variable aléatoire réelle et Y une copie indépendante de X, on peut écrire $\mathbf{Var}(X)=\frac{1}{2}\mathbf{E}[(X-Y)^2]$. Si X_i' est une copie indépendante de X_i, alors, conditionnellement à $X^{(i)}$, la variable

$$Z_i'=f(X_1,\dots,X_{i-1},X_i',X_{i+1},\dots,X_n)$$

est une copie indépendante de Z, et l'on a

$$\mathbf{Var}^{(i)}(Z)=\mathbf{E}^{(i)}\left[\left(Z-\mathbf{E}^{(i)}Z\right)^2\right]=\frac{1}{2}\mathbf{E}^{(i)}\left[(Z-Z_i')^2\right]=\mathbf{E}^{(i)}\left[(Z-Z_i')_+^2\right]\,.$$

Ainsi

$$v=\frac{1}{2}\sum_{i=1}^n\mathbf{E}\left[(Z-Z_i')^2\right]=\sum_{i=1}^n\mathbf{E}\left[(Z-Z_i')_+^2\right]\,.$$

De plus, en utilisant que pour toute variable aléatoire réelle X, $\mathbf{Var}(X)=\inf_{a\in\mathbb{R}}\mathbf{E}[(X-a)^2]$, on a

$$\mathbf{Var}^{(i)}(Z)=\inf_{Z_i}\mathbf{E}^{(i)}\left[(Z-Z_i)^2\right]\,,$$

où l'infimum est pris sur les fonctions mesurables de $X^{(i)}$ de carré intégrable. Ainsi

$$v=\sum_{i=1}^n\mathbf{E}\left[\inf_{Z_i}\mathbf{E}^{(i)}\left[(Z-Z_i)^2\right]\right]\,.$$

Exemple 1.2 (Bins and balls)**.** Soit $X_1,\dots,X_n$ des variables i.i.d. à valeurs dans $\mathbb{N}^*$, de loi $(p_j)_{j\ge1}$. Pour $r\ge1$, on note $K_{n,r}$ le nombre d'entiers représentés exactement r fois dans l'échantillon $(X_1,\dots,X_n)$, soit

$$K_{n,r}=\sum_{j\ge1}\mathbb{1}_{\{\sum_{i=1}^n\mathbb{1}_{X_i=j}=r\}}\,.$$

On définit aussi $\bar{K}_{n,r} = \sum_{s \geq r} K_{n,s}$ le nombre d'entiers représentés au moins r fois, et $K_n = \bar{K}_{n,1}$ le nombre d'entiers distincts présents dans l'échantillon. On a

$$\mathbf{E}K_{n,r} = \sum_{j \geq 1} \binom{n}{r} p_j^r (1-p_j) n - r \,,$$

et

$$\mathbf{E}K_n = \sum_{j \geq 1} \left(1 - (1-p_j)^n\right) .$$

Que peut-on dire de la variance de ces variables ? Soit $K_n^{(i)}$ le nombre de symboles distincts dans l'échantillon lorsque l'on omet la $i^{\text{ième}}$ variable. Alors

$$K_n^{(i)} = \begin{cases} K_n - 1 & \text{si } X_i \text{ n'est présent qu'une seule fois,} \\ K_n & \text{sinon.} \end{cases}$$

Ainsi l'inégalité d'Efron–Stein donne

$$\mathbf{Var}(K_n) \leq \mathbf{E}K_{n,1} \,.$$

De façon plus générale, on a

$$\mathbf{Var}(\bar{K}_{n,r}) \leq r\mathbf{E}K_{n,r} \,.$$

De plus, $\mathbf{Var}(K_{n,r}) \leq r\mathbf{E}K_{n,r} + (r+1)\mathbf{E}K_{n,r+1}$.

Définition 1.1 (Fonction à différences bornées)**.** On dit que $f : \mathcal{X}^n \to \mathbb{R}$ est à différences bornées s'il existe des constantes $c_1, \ldots, c_n \geq 0$ telles que pour tout $i \in [\![1, n]\!]$, on a

$$\sup_{\substack{x_1,\ldots,x_n \in \mathcal{X} \\ x_i' \in \mathcal{X}}} \left| f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n) \right| \leq c_i \,.$$

L'inégalité d'Efron–Stein donne la borne suivante sur la variance des fonctions à différences bornées.

Proposition 1.2 (Variance des fonctions à différences bornées)**.** *Si $f : \mathcal{X}^n \to \mathbb{R}$ est à différences bornées avec constantes $c_1, \ldots, c_n \geq 0$ et si $Z = f(X_1, \ldots, X_n)$ avec $X_1, \ldots, X_n$ indépendantes, alors*

$$\mathbf{Var}(Z) \leq \frac{1}{4} \sum_{i=1}^n c_i^2 \,.$$

Preuve de la Proposition 1.2. On définit

$$Z_i = \frac{1}{2} \left(Z_i^- + Z_i^+ \right) ,$$

avec

$$Z_i^- = \inf_{x_i \in \mathcal{X}} f(X_1, \ldots, X_{i-1}, x_i, X_{i+1}, \ldots, X_n) \quad \text{et} \quad Z_i^+ = \sup_{x_i \in \mathcal{X}} f(X_1, \ldots, X_{i-1}, x_i, X_{i+1}, \ldots, X_n) \,.$$

Alors $|Z - Z_i|$ correspond à la distance entre Z et le milieu de l'intervalle $[Z_i^-, Z_i^+]$. Comme cette intervalle est de longueur inférieure à c_i par hypothèse, on a $|Z - Z_i| \leq c_i/2$. Ainsi, par l'inégalité d'Efron–Stein, on a

$$\mathbf{Var}(Z) \leq \sum_{i=1}^n \mathbf{E}\left[(Z - Z_i)^2 \right] \leq \frac{1}{4} \sum_{i=1}^n c_i^2 \,.$$

■

Dans le reste de ce chapitre, on s'intéresse au cas (simple mais déjà très riche) où $\mathcal{X} = \{0,1\}$ et où $X = (X_1, \dots, X_n) \sim \mathcal{B}(p)^{\otimes n}$ avec $\mathcal{B}(p) = p\delta_1 + (1-p)\delta_0$ (notons que quand $p = 1/2$, la loi de X est uniforme sur $\{0,1\}^n$). Dans ce cas, l'inégalité d'Efron–Stein correspond exactement à une inégalité de Poincaré. L'énergie d'une fonction $f : \{0,1\}^n \to \mathbb{R}$ est définie par

$$\mathcal{E}(f) = \frac{1}{2} \sum_{i=1}^{n} \mathbf{E}\left[\left(f(X) - f(\widetilde{X}_i) \right)^2 \right] ,$$

où $\widetilde{X}_i = (X_1, \dots, X_{i-1}, X_i', X_{i+1}, \dots, X_n)$ est le vecteur obtenu en rejouant la $i^{\text{ième}}$ coordonnée de X indépendamment des autres. L'inégalité d'Efron–Stein donne

$$\mathbf{Var}(f) \leq \mathcal{E}(f) \tag{1.1}$$

avec égalité pour $f(x) = \sum_{i=1}^n x_i$. On dit que la mesure produit $\mathcal{B}(p)^{\otimes n}$ vérifie une inégalité de Poincaré, avec constante de Poincaré égale à 1, i.e.

$$\sup_{\substack{f:\{0,1\}^n \to \mathbb{R} \\ f \text{ non-constante}}} \frac{\mathbf{Var}(f)}{\mathcal{E}(f)} = 1 .$$

Nous allons voir que l'on peut aussi montrer que

$$\mathbf{Var}(f) \log \left(\frac{\mathbf{Var}(f)}{\sum_{j=1}^n (\mathbf{E}|\Delta_j|)^2} \right) \leq c(p) \mathcal{E}(f) , \tag{1.2}$$

où $\Delta_j = \mathbf{E}_j f - \mathbf{E}_{j-1} f$, et où $c(p)$ est une constante qui ne dépend que de p. Dès que

$$\mathbf{Var}(f) = \sum_{j=1}^{n} \mathbf{E}\left[(\mathbf{E}_j f - \mathbf{E}_{j-1} f)^2 \right] \gg \sum_{j=1}^{n} \mathbf{E}\left[|\mathbf{E}_j f - \mathbf{E}_{j-1} f| \right]^2 ,$$

l'inégalité (1.2) constitue une significative amélioration par rapport à (1.1). Avant d'établir (1.2), montrons que la mesure $\mathcal{B}(p)^{\otimes n}$ vérifie aussi une inégalité de Sobolev logarithmique.

2. Inégalité de Sobolev logarithmique

Soit μ une mesure de probabilité sur $\{0,1\}^n$. L'entropie sous μ d'une fonction positive $g : \{0,1\}^n \to \mathbb{R}_+$ est définie comme

$$\mathbf{Ent}_\mu(g) = \mathbf{E}_\mu[g \log g] - \mathbf{E}_\mu[g] \log \mathbf{E}_\mu[g] ,$$

avec la convention $0 . \log 0 = 0$. Si $(X_1, \dots, X_n) \sim \mu$ et $Z = g(X_1, \dots, X_n)$ on écrira indifféremment $\mathbf{Ent}(Z)$ ou $\mathbf{Ent}_\mu\, g$. On a

$$\mathbf{Ent}_\mu(g) = \sup_{\substack{h:\{0,1\}^n \to \mathbb{R}_+, \\ \mathrm{Supp}(g) \subset \mathrm{Supp}(h) \\ \mathbf{E}_\mu h = 1}} \mathbf{E}_\mu\left[g \log h \right] ,$$

où $\mathrm{Supp}(g) = \{x \in \{0,1\}^n,\, g(x) > 0\}$. En effet, d'une part il y a égalité pour $h = \frac{g}{\mathbf{E}_\mu g}$. D'autre part, pour toute fonction $h : \{0,1\}^n \to \mathbb{R}_+$ avec $\mathrm{Supp}(g) \subset \mathrm{Supp}(h)$ et $\mathbf{E}_\mu h = 1$, on a

$$\mathbf{E}_\mu[g \log g] - \mathbf{E}_\mu[g \log h] = \mathbf{E}_\mu\left[g \log \frac{g}{h} \right] = \mathbf{E}_\nu\left[\frac{g}{h} \log \frac{g}{h} \right] ,$$

où ν est la loi de probabilité sur $\{0,1\}^n$ donnée par $\nu(x) = h(x)\mu(x)$. Par l'inégalité de Jensen, on a

$$\mathbf{E}_\nu\left[\frac{g}{h} \log \frac{g}{h} \right] \geq \mathbf{E}_\nu\left[\frac{g}{h} \right] \log \mathbf{E}_\nu\left[\frac{g}{h} \right] = \mathbf{E}_\mu[g] \log \mathbf{E}_\mu[g] ,$$

soit $\mathbf{Ent}_\mu(g) \geq \mathbf{E}_\mu[g \log h]$. Cette caractérisation variationnelle de l'entropie (que nous retrouverons en plus grande généralité au Chapitre 4) a de nombreuses implications, la plus importante d'entre elles étant probablement la sous-additivité de l'entropie.

Proposition 1.3 (Sous-additivité de l'entropie)**.** *Soit $X = (X_1, \dots, X_n) \sim \mathcal{B}(p)^{\otimes n}$ et $Z = g(X)$ pour $g : \{0,1\}^n \to \mathbb{R}_+^*$ une fonction positive. Alors*

$$\mathbf{Ent}(Z) \leq \mathbf{E}\left[\sum_{i=1}^n \mathbf{Ent}^{(i)}(Z)\right],$$

où $\mathbf{Ent}^{(i)}(Z) = \mathbf{E}^{(i)}[Z \log Z] - \mathbf{E}^{(i)}[Z] \log \mathbf{E}^{(i)}[Z]$ avec $\mathbf{E}^{(i)} = \mathbf{E}[\cdot \mid X^{(i)}]$.

Preuve de la Proposition 1.3. On rappelle la notation $\mathbf{E}_i = \mathbf{E}[\cdot \mid X_1, \dots, X_i]$ avec $\mathbf{E}_0 = \mathbf{E}$, et le fait que $\mathbf{E}_{i-1} Z = \mathbf{E}^{(i)} \mathbf{E}_i Z$ par indépendance des X_i. On peut alors écrire

$$Z\left(\log Z - \log \mathbf{E} Z\right) = \sum_{i=1}^n Z\left(\log \mathbf{E}_i Z - \log \mathbf{E}_{i-1} Z\right) = \sum_{i=1}^n Z\left(\log \mathbf{E}_i Z - \log \mathbf{E}^{(i)} \mathbf{E}_i Z\right).$$

En appliquant l'inégalité $\mathbf{Ent}_\mu\, g \geq \mathbf{E}_\mu[g \log h - \log \mathbf{E}_\mu h]$ avec μ la loi de X sachant $X^{(i)}$ et $h(X) = \mathbf{E}_i g(X)$, on obtient

$$\mathbf{E}^{(i)}\left[Z\left(\log \mathbf{E}_i Z - \log \mathbf{E}^{(i)} \mathbf{E}_i Z\right)\right] \leq \mathbf{Ent}^{(i)}(Z),$$

et en prenant l'espérance dans la somme, on obtient bien

$$\mathbf{Ent}(Z) = \mathbf{E}\left[\sum_{i=1}^n \mathbf{E}^{(i)}\left[Z\left(\log \mathbf{E}_i Z - \log \mathbf{E}_{i-1} Z\right)\right]\right] \leq \mathbf{E}\left[\sum_{i=1}^n \mathbf{Ent}^{(i)}(Z)\right].$$

■

Proposition 1.4 (Inégalité de Sobolev logarithmique sur le cube)**.** *Soit $\mu = \mathcal{B}(p)^{\otimes n}$. Pour toute fonction $f : \{0,1\}^n \to \mathbb{R}$,*

$$\mathbf{Ent}_\mu(f^2) \leq c(p) \mathcal{E}(f),$$

avec

$$c(p) = \begin{cases} 2 & \text{si } p = \frac{1}{2}, \\ \frac{1}{1-2p} \log\left(\frac{1-p}{p}\right) & \text{sinon.} \end{cases}$$

On dit que la mesure $\mathcal{B}(p)^{\otimes n}$ vérifie une inégalité de Sobolev logarithmique avec constante $c(p)$.

Preuve de la Proposition 1.4. Soit $X \sim \mathcal{B}(p)^{\otimes n}$. Par sous-additivité de l'entropie, on a

$$\mathbf{Ent}\left(f(X)^2\right) \leq \sum_{i=1}^n \mathbf{E}\left[\mathbf{Ent}^{(i)}\left(f(X)^2\right)\right].$$

Il suffit donc de montrer que

$$\mathbf{Ent}^{(i)}\left(f(X)^2\right) \leq c(p) p(1-p) \mathbf{E}^{(i)}\left[\left(f(X) - f(\bar{X}^{(i)})\right)^2\right].$$

Pour toute réalisation de $X^{(i)}$, la fonction $f(X)$ ne peut prendre que deux valeurs selon que $X_i = 1$ ou $X_i = 0$. En notant a et b ces deux valeurs possibles, il s'agit de montrer

$$pa^2 \log(a^2) + (1-p) b^2 \log(b^2) - \left(pa^2 + (1-p)b^2\right) \log\left(pa^2 + (1-p)b^2\right) \leq c(p) p(1-p)(a-b)^2.$$

On laisse en exercice la démonstration de cette inégalité. ■

Montrons maintenant comment l'inégalité de Sobolev logarithmique peut être utilisée pour montrer l'inégalité 1.2.

Proposition 1.5. *Sous la loi $\mu = \mathcal{B}(p)^{\otimes n}$, pour toute fonction $f : \{0,1\}^n \to \mathbb{R}$,*

$$\mathbf{Var}(f) \log\left(\frac{\mathbf{Var}(f)}{\sum_{j=1}^n (\mathbf{E}|\Delta_j|)^2}\right) \leq c(p)\mathcal{E}(f)\,,$$

où $\Delta_j = \mathbf{E}_j f - \mathbf{E}_{j-1} f$, et où $c(p)$ est la constante de Sobolev logarithmique de la Proposition 1.4.

Preuve de la Proposition 1.5. Remarquons d'abord que

$$\mathcal{E}(f) = \sum_{j=1}^n \mathcal{E}(\Delta_j)\,.$$

En effet, pour $X \sim \mathcal{B}(p)^{\otimes n}$ et $\widetilde{X}_i = (X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_n)$, on a

$$\begin{aligned}
2\sum_{j=1}^n \mathcal{E}(\Delta_j) &= \sum_{j=1}^n \sum_{i=1}^n \mathbf{E}\left[\left(\Delta_j(X) - \Delta_j(\widetilde{X}_i)\right)^2\right] \\
&= \sum_{i,j=1}^n \left\{\mathbf{E}\left[\left(\mathbf{E}_j f(X) - \mathbf{E}_j f(\widetilde{X}_i)\right)^2\right] - \mathbf{E}\left[\left(\mathbf{E}_{j-1} f(X) - \mathbf{E}_{j-1} f(\widetilde{X}_i)\right)^2\right]\right\} \\
&= \sum_{i=1}^n \left\{\mathbf{E}\left[\left(\mathbf{E}_n f(X) - \mathbf{E}_n f(\widetilde{X}_i)\right)^2\right] - \mathbf{E}\left[\left(\mathbf{E}_0 f(X) - \mathbf{E}_0 f(\widetilde{X}_i)\right)^2\right]\right\} \\
&= \sum_{i=1}^n \mathbf{E}\left[\left(f(X) - f(\widetilde{X}_i)\right)^2\right] \\
&= 2\mathcal{E}(f)\,,
\end{aligned}$$

où pour la première égalité on a utilisé

$$\mathbf{E}\left[\left(\mathbf{E}_j f(X) - \mathbf{E}_j f(\widetilde{X}_i)\right)\left(\mathbf{E}_{j-1} f(X) - \mathbf{E}_{j-1} f(\widetilde{X}_i)\right)\right] = \mathbf{E}\left[\left(\mathbf{E}_{j-1} f(X) - \mathbf{E}_{j-1} f(\widetilde{X}_i)\right)^2\right].$$

En appliquant la Proposition 1.4, on a

$$\mathcal{E}(f) = \sum_{j=1}^n \mathcal{E}(\Delta_j) \geq \frac{1}{c(p)} \sum_{j=1}^n \mathbf{Ent}\left(\Delta_j^2\right).$$

Pour toute fonction positive g, on a $\mathbf{Ent}(g^2) \geq \mathbf{E}[g^2] \log \frac{\mathbf{E}[g^2]}{(\mathbf{E}[g])^2}$. En effet, en utilisant que pour tout $x > 0$, $\log(x) \leq x - 1$, on a

$$\mathbf{E}\left[g^2 \log \frac{\mathbf{E}g^2}{g\mathbf{E}g}\right] \leq \mathbf{E}\left[g^2\left(\frac{\mathbf{E}g^2}{g\mathbf{E}g} - 1\right)\right] = 0\,,$$

ce qui équivaut à l'inégalité voulue. Ainsi

$$\begin{aligned}\mathcal{E}(f) &\geq \frac{1}{c(p)} \sum_{j=1}^{n} \mathbf{E}[\Delta_j^2] \log \frac{\mathbf{E}[\Delta_j^2]}{\mathbf{E}[|\Delta_j|]^2} \\ &= -\frac{\mathbf{Var}(f)}{c(p)} \sum_{j=1}^{n} \frac{\mathbf{E}[\Delta_j^2]}{\mathbf{Var}(f)} \log \frac{\mathbf{E}[|\Delta_j|]^2}{\mathbf{E}[\Delta_j^2]} \\ &\geq -\frac{\mathbf{Var}(f)}{c(p)} \log \left(\frac{\sum_{j=1}^{n} \mathbf{E}[|\Delta_j|]^2}{\mathbf{Var}(f)} \right) ,\end{aligned}$$

où l'on a utilisé l'inégalité de Jensen et le fait que $\mathbf{Var}(f) = \sum_{j=1}^{n} \mathbf{E}[\Delta_j^2]$. ■

3. Influences

Nous allons voir les conséquences surprenantes de la Proposition 1.5 quant à l'influence des fonctions booléennes. Soit $f : \{0,1\}^n \to \{0,1\}$ une fonction booléenne et notons $A = \{x \in \{0,1\}^n,\ f(x) = 1\}$. Pour $X \sim \mathcal{B}(p)^{\otimes n}$, l'influence de i sur f est définie comme

$$\mathbf{I}_i(f) = \mathbf{P}\left(f(X) \neq f(\bar{X}^{(i)}) \right) ,$$

soit la probabilité qu'un flip de la coordonnée implique un changement de la valeur de f. Quand $f(X) \neq f(\bar{X}^{(i)})$, on dit que i est un pivot pour X. L'influence totale est définie comme la somme des influences individuelles :

$$\mathbf{I}(f) = \sum_{i=1}^{n} \mathbf{I}_i(f) .$$

Exemple 1.3 (Fonction parité)**.** Pour f la fonction qui vaut 1 si le nombre de coordonnées égales à 1 est impair, 0 sinon, appelée fonction parité, on a toujours $f(X) \neq f(\bar{X}^{(i)})$. Ainsi $\mathbf{I}_i(f) = 1$ et $\mathbf{I}(f) = n$. Clairement, c'est la plus grande influence possible.

Exemple 1.4 (Fonction majorité)**.** Pour $f(x) = \mathbb{1}_{\sum x_i > n/2}$ la fonction majorité, la coordonnée i est pivot uniquement quand $\sum_{j \neq i} x_j = \lceil \frac{n-1}{2} \rceil$. Ainsi, pour $p = \frac{1}{2}$, par la formule de Stirling,

$$\mathbf{I}_i(f) = \mathbf{P}\left(\mathrm{Bin}\left(n-1, \frac{1}{2} \right) = \left\lceil \frac{n-1}{2} \right\rceil \right) \sim \sqrt{\frac{2}{\pi n}} ,$$

et $\mathbf{I}(f) \sim \sqrt{\frac{2n}{\pi}}$.

Exemple 1.5 (Fonction dictature)**.** Pour $f(x) = x_1$ la fonction dictature qui ne retient que la valeur de la première coordonnée, l'influence de toutes les coordonnées est nulle, sauf celle de la première qui vaut 1. Ainsi $\mathbf{I}(f) = \mathbf{I}_1(f) = 1$.

Peut-on obtenir des bornes générales pour l'influence d'une fonction f ? Par l'inégalité d'Efron–Stein,

$$\mathbf{Var}(f) = \mathbf{P}(A)(1 - \mathbf{P}(A)) \leq p(1-p) \sum_{i=1}^{n} \mathbf{I}_i(f) = p(1-p)\mathbf{I}(f) .$$

En particulier, si $\mathbf{P}(A) = p$, l'influence doit être au moins égale à 1, et cette borne inférieure est atteinte par la fonction dictature. Pour cette fonction, il n'y a qu'une seule coordonnée qui a une influence non-nulle. Plus généralement, si seulement k coordonnées ont une influence non-nulle

sur f (pour k fixé ne dépendant pas de n, on dit que f est un *junta*), alors clairement $\mathbf{I}(f) \leq k$. Une question que l'on peut se poser est la suivante : si f est symétrique au sens où toutes les coordonnées ont la même influence sur f, $\mathbf{I}_1(f) = \cdots = \mathbf{I}_n(f) = \frac{\mathbf{I}(f)}{n}$, jusqu'à quel point l'influence totale peut-elle être petite ? Un résultat fondamental de Kahn et al. [13] implique que l'influence d'une fonction symétrique est au moins égale à $\mathbf{Var}(f)\log n$, ce qui contraste fortement avec le cas de la fonction dictature ou plus généralement des juntas qui ont une influence bornée.

Proposition 1.6 (Kahn et al. [13])**.** *Soit $f : \{0,1\}^n \to \{0,1\}$ une fonction booléenne. Alors, sous la loi $\mu = \mathcal{B}(p)^{\otimes n}$,*

$$\max_{1\leq i\leq n} \mathbf{I}_i(f) \geq \frac{\mathbf{Var}(f)\log n}{n}\,,$$

En particulier, si f est symétrique, alors $\mathbf{I}(f) \geq \mathbf{Var}(f)\log n$.

Preuve de la Proposition 1.6. On a

$$\begin{aligned}\mathbf{E}|\Delta_j| &= \mathbf{E}\left[\left|\mathbf{E}_i f(X) - \mathbf{E}_i\mathbf{E}^{(i)} f(X)\right|\right]\\ &\leq \mathbf{E}\left[\left|f(X) - \mathbf{E}^{(i)} f(X)\right|\right]\\ &= 2p(1-p)\mathbf{I}_i(f)\,.\end{aligned}$$

Ainsi, la Proposition 1.5 implique que

$$\mathbf{Var}(f)\log\left(\frac{\mathbf{Var}(f)}{4p^2(1-p)^2\sum_{j=1}^n \mathbf{I}_j(f)^2}\right) \leq c(p)p(1-p)\mathbf{I}(f)\,,$$

soit

$$\sum_{j=1}^n \mathbf{I}_j(f)^2 \geq \frac{\mathbf{Var}(f)}{4p^2(1-p)^2}\exp\left(-\frac{c(p)p(1-p)\mathbf{I}(f)}{\mathbf{Var}(f)}\right)\,,$$

On distingue deux cas. Soit $\mathbf{I}(f) \geq \frac{\alpha\,\mathbf{Var}(f)}{c(p)p(1-p)}\log n$, pour $\alpha = 1 - \frac{\log\left(\mathbf{Var}(f)\log^2 n\right)}{\log n}$, auquel cas

$$\sum_{j=1}^n \mathbf{I}_j(f)^2 \geq \frac{\mathbf{I}(f)^2}{n} \geq \frac{\alpha^2}{c(p)^2p^2(1-p)^2}\cdot\frac{\mathbf{Var}(f)^2\log^2 n}{n}\,,$$

et

$$\max_{1\leq i\leq n} \mathbf{I}_i(f) \geq \frac{\alpha}{c(p)p(1-p)}\cdot\frac{\mathbf{Var}(f)\log n}{n}\,.$$

Soit $\mathbf{I}(f) \leq \frac{\alpha\,\mathbf{Var}(f)}{c(p)p(1-p)}\log n$, auquel cas

$$\sum_{j=1}^n \mathbf{I}_j(f)^2 \geq \frac{\mathbf{Var}(f)}{4p^2(1-p)^2}\cdot n^{-\alpha} = \frac{1}{4p^2(1-p)^2}\cdot\frac{\mathbf{Var}(f)^2\log^2 n}{n}\,,$$

et

$$\max_{1\leq i\leq n} \mathbf{I}_i(f) \geq \frac{1}{2p(1-p)}\cdot\frac{\mathbf{Var}(f)\log n}{n}\,.$$

Dans les deux cas, en utilisant que $\alpha \geq \frac{1}{2}$, que $c(p)p(1-p) \leq \frac{1}{2}$, et que $p(1-p) \leq \frac{1}{4}$, on obtient

$$\max_{1\leq i\leq n} \mathbf{I}_i(f) \geq \frac{\mathbf{Var}(f)\log n}{n}\,.$$

■

4. Phénomènes de transition de phase

Soit $f : \{0,1\}^n \to \{0,1\}$ une fonction booléenne monotone, au sens où elle est croissante en chacune de ses coordonnées (par exemple la fonction majorité et la fonction dictature sont toutes les deux monotones). Pour $A = \{x \in \{0,1\}^n,\ f(x) = 1\}$, on s'intéresse à la fonction

$$p \mapsto \mu_p(A) = \mathbf{P}(X \in A) = \sum_{x \in A} p^{\|x\|}(1-p)^{n-\|x\|}\,,$$

où $\|x\| = \sum_{i=1}^n x_i$ et $X \sim \mu_p = \mathcal{B}(p)^{\otimes n}$ (on rend maintenant la dépendance en p explicite). La monotonicité de f implique que $\mu_0(A) = 0$, $\mu_1(A) = 1$, et $p \mapsto \mu_p(A)$ est une fonction strictement croissante et différentiable. Le message principal de cette section est que si la fonction f ne dépend pas trop de chaque coordonnée individuellement, alors on observe une transition abrupte de 0 à 1. Plus précisément, si l'on note p_ε la valeur de p pour laquelle $\mu_p(A) = \varepsilon$, alors la différence $p_{1-\varepsilon} - p_\varepsilon$ est très petite.

Exemple 1.6 (Fonction dictature)**.** Soit $f(x) = x_1$ la fonction dictature. Dans ce cas, on a $\mu_p(A) = p$. La fonction $p \mapsto \mu_p(A)$ croit linéairement de 0 à 1, il n'y a pas de transition abrupte.

Exemple 1.7 (Fonction majorité)**.** Soit $f(x) = \mathbb{1}_{\sum x_i > n/2}$ la fonction majorité. On a $p_{1/2} = 1/2$, et, pour $p < 1/2$, par l'inégalité de Hoeffding,

$$\begin{aligned}\mu_p(A) &= \mathbf{P}_p\left(\sum_{i=1}^n X_i > \frac{n}{2}\right)\\ &= \mathbf{P}_p\left(\sum_{i=1}^n X_i - np > \left(\frac{1}{2} - p\right)n\right)\\ &\leq \exp\left\{-2\left(\frac{1}{2}-p\right)^2 n\right\}.\end{aligned}$$

Ainsi $\mu_p(A) \leq \varepsilon$ dès que $p \leq \frac{1}{2} - \sqrt{\frac{\log(1/\varepsilon)}{2n}}$. De même, $\mu_p(A) \geq 1-\varepsilon$ dès que $p \geq \frac{1}{2} + \sqrt{\frac{\log(1/\varepsilon)}{2n}}$. Ainsi, la valeur de $\mu_p(A)$ saute de ε à $1-\varepsilon$ dans un intervalle de longueur $\sqrt{\frac{2\log(1/\varepsilon)}{n}}$, il y a une transition abrupte autour de $p = \frac{1}{2}$.

Ce phénomène de transition de phase s'étend à une large classe de fonctions monotones, grosso modo celles qui dépendent un peu mais pas trop de chaque variable individuellement. Le Lemme de Russo ci-dessous relie la dérivée de la fonction $p \mapsto \mu_p(A)$ à l'influence de f.

Lemme 1.7 (Lemme de Russo)**.** *Soit $f : \{0,1\}^n \to \{0,1\}$ une fonction booléenne monotone et $A = \{x \in \{0,1\}^n,\ f(x) = 1\}$. Alors pour tout $p \in]0,1[$,*

$$\frac{\mathrm{d}\mu_p(A)}{\mathrm{d}p} = \mathbf{I}^p(f)\,.$$

Preuve du Lemme 1.7. Soit $\mu_p = \mathcal{B}(p)^{\otimes n}$ et pour $i \in [\![1,n]\!]$ et $q \in [0,1]$, soit

$$\mu_p^{(i)} = \mathcal{B}(p)^{\otimes i-1} \otimes \mathcal{B}(q) \otimes \mathcal{B}(p)^{\otimes n-i}\,.$$

En considérant des variables $U_1, \ldots, U_n$ indépendantes uniformes sur $[0,1]$, on a

$$X_i = \mathbb{1}_{U_i \leq p} \sim \mathcal{B}(p)\,,$$

et $X = (X_1, \dots, X_n) \sim \mu_p$, et si $X_i' = \mathbb{1}_{U_i \le q}$, alors le vecteur $\widetilde{X}^{(i)} = (X_1, \dots, X_{i-1}, X_i', X_{i+1}, \dots, X_n)$ est de loi $\mu_p^{(i)}$. Supposons $q > p$. Par monotonicité de f, on a

$$\begin{aligned}\mu_p^{(i)}(A) - \mu_p(A) &= \mathbf{P}\left(\widetilde{X}^{(i)} \in A, X \notin A\right) \\ &= \mathbf{P}\left(U_i \in]p,q], \, f(\widetilde{X}^{(i)}) \neq f(X)\right) \\ &= (q-p)\mathbf{I}_i^p(f)\,.\end{aligned}$$

Par un argument similaire, si $q < p$, $\mu_p^{(i)}(A) - \mu_p(A) = (q-p)\mathbf{I}_i^p(f)$. Ainsi, en divisant par $q-p$ et en faisant tendre q vers p, on a

$$\frac{\partial \mu_p(A)}{\partial p_i} = \mathbf{I}_i^p(f)\,.$$

Et ainsi

$$\frac{\mathrm{d}\mu_p(A)}{\mathrm{d}p} = \sum_{i=1}^n \frac{\partial \mu_p(A)}{\partial p_i} = \mathbf{I}^p(f)\,.$$

■

Proposition 1.8. *Soit $f : \{0,1\}^n \to \{0,1\}$ une fonction monotone symétrique. Alors pour tout $\varepsilon \in]0, 1/2[$,*

$$p_{1-\varepsilon} - p_\varepsilon \le \frac{4\log\left(\frac{1}{2\varepsilon}\right)}{\log n}\,,$$

où p_ε est la valeur de p telle que $\mu_p(f=1) = \varepsilon$.

Preuve de la Proposition 1.8. Soit $p \le p_{1/2}$. Par la Proposition 1.6 et le Lemme 1.7, et comme $\mu_p(A) \le 1/2$, on a

$$\frac{\mathrm{d}\mu_p(A)}{\mathrm{d}p} \ge \mu_p(A)(1-\mu_p(A))\log n \ge \frac{\mu_p(A)\log n}{2}\,,$$

soit

$$\frac{\mathrm{d}\log\mu_p(A)}{\mathrm{d}p} \ge \frac{\log n}{2}\,.$$

Ainsi pour $\varepsilon < 1/2$,

$$\log(1/2) - \log(\varepsilon) \ge (p_{1/2} - p_\varepsilon)\frac{\log n}{2}\,,$$

soit

$$p_{1/2} - p_\varepsilon \le \frac{2\log\left(\frac{1}{2\varepsilon}\right)}{\log n}\,.$$

Comme la même borne supérieure est valable pour $p_{1-\varepsilon} - p_{1/2}$, on obtient bien l'inégalité voulue. ■

Exemple 1.8 (Percolation sur $[\![1, \sqrt{n}]\!]^2$)**.** Soit $\mathcal{C} = [\![1, \sqrt{n}]\!]^2$ la grille carrée de côté $\sqrt{n}$. Indépendamment pour chaque sommet (i,j), on tire une variable de loi $\mathcal{B}(p)$. On dit qu'un sommet est ouvert si la variable en ce sommet est égale à 1 (fermé sinon), et l'on note A l'ensemble des configurations dans lesquelles il existe un chemin de sommets ouverts allant de gauche à droite (ici, un chemin est une suite de sommets $(u_1, \dots, u_k)$ telle que pour tout $j \le k-1$, $|u_j^1 - u_{j+1}^1| + |u_j^2 - u_{j+1}^2| = 1$). La fonction $f = \mathbb{1}_A$ est clairement monotone. Par symétrie, on voit que $p_{1/2} = 1/2$. En effet, si l'on note B l'ensemble des configurations dans lesquelles il existe un chemin de sommets fermés allant de bas en haut, alors $A = B^c$, et pour $p = 1/2$, on a clairement

$\mu_{1/2}(A) = \mu_{1/2}(B) = 1 - \mu_{1/2}(A)$, d'où $\mu_{1/2}(A) = 1/2$. En admettant que toutes les variables ont à peu près la même influence sur f, la Proposition 1.8 implique qu'il y a une transition abrupte autour de $p = 1/2$.

Chapitre 2

Méthode de Cramér-Chernoff et inégalités classiques

1. Méthode de Cramér-Chernoff

Soit Z une variable aléatoire réelle d'espérance $\mathbf{E}Z$ finie. La fonction génératrice des moments de Z, ou transformée de Laplace, est la fonction qui à $\lambda \in \mathbb{R}$ associe $\mathbf{E}e^{\lambda Z} \in \mathbb{R}_+ \cup \{+\infty\}$. On note

$$\psi_Z(\lambda) = \log \mathbf{E}e^{\lambda(Z-\mathbf{E}Z)}\,.$$

En passant à l'exponentielle et en appliquant l'inégalité de Markov, on a, pour tout $t \geq 0$ et $\lambda \geq 0$,

$$\mathbf{P}(Z - \mathbf{E}Z \geq t) \leq \mathbf{P}\left(e^{\lambda(Z-\mathbf{E}Z)} \geq e^{\lambda t}\right) \leq e^{-\lambda t}\mathbf{E}e^{\lambda(Z-\mathbf{E}Z)} = e^{-\{\lambda t - \psi_Z(\lambda)\}}\,.$$

Comme cela est vrai pour tout $\lambda \geq 0$, on peut choisir celui qui minimise la quantité ci-dessus et l'on a

$$\mathbf{P}(Z - \mathbf{E}Z \geq t) \leq e^{-\sup_{\lambda\geq 0}\{\lambda t - \psi_Z(\lambda)\}}\,.$$

Si l'on s'intéresse aux déviations de Z « vers la gauche », on peut écrire, pour tout $t \geq 0$ et $\lambda \geq 0$,

$$\mathbf{P}(Z - \mathbf{E}Z \leq -t) = \mathbf{P}(-Z + \mathbf{E}Z \geq t) \leq e^{-\lambda t}\mathbf{E}e^{-\lambda(Z-\mathbf{E}Z)}\,,$$

et ainsi

$$\mathbf{P}(Z - \mathbf{E}Z \leq -t) \leq e^{-\sup_{\lambda\geq 0}\{\lambda t - \psi_Z(-\lambda)\}} = e^{-\sup_{\lambda\geq 0}\{\lambda t - \psi_{-Z}(\lambda)\}}\,.$$

La fonction $\psi_Z^* : t \mapsto \sup_{\lambda\geq 0}\{\lambda t - \psi_Z(\lambda)\}$ s'appelle la transformée de Cramér de $Z - \mathbf{E}Z$. Comme $\psi_Z(0) = 0$, on a $\psi_Z^* \geq 0$. De plus, par l'inégalité de Jensen, $\lambda t - \psi_Z(\lambda) \leq \lambda t$, qui est négatif pour $t \geq 0$ et $\lambda \leq 0$. Pour $t \geq 0$, on peut donc écrire

$$\psi_Z^*(t) = \sup_{\lambda\in\mathbb{R}}\{\lambda t - \psi_Z(\lambda)\}\,.$$

Ainsi, sur $\mathbb{R}_+$, la transformée de Cramér ψ^* correspond à la transformée de Legendre de ψ. Remarquons aussi que, si $\mathbf{E}e^{\lambda Z} = +\infty$ pour tout $\lambda > 0$, alors la fonction ψ_Z^* est identiquement nulle, ce qui n'a pas beaucoup d'intérêt (la borne de Cramér–Chernoff est triviale dans ce cas). Maintenant, si l'ensemble $I = \left\{\lambda \geq 0,\, \mathbf{E}e^{\lambda Z} < +\infty\right\}$ n'est pas réduit à $\{0\}$ (I est alors de la forme $[0, b[$ avec $0 < b \leq +\infty$), la fonction ψ_Z est convexe et continuement différentiable sur I avec $\psi_Z(0) = \psi_Z'(0) = 0$, et on peut écrire

$$\psi_Z^*(t) = \sup_{\lambda\in I}\{\lambda t - \psi_Z(\lambda)\} = \lambda_t t - \psi_Z(\lambda_t)\,,$$

où λ_t vérifie $\psi_Z'(\lambda_t) = t$.

Exemple 2.1 (Variable gaussienne)**.** Si $Z \sim \mathcal{N}(0, \sigma^2)$, on a $\psi_Z(\lambda) = \frac{\lambda^2\sigma^2}{2}$ et $t = \frac{t}{\sigma^2}$. La méthode de Cramér-Chernoff donne alors, pour tout $t \geq 0$,

$$\mathbf{P}(Z \geq t) \leq e^{-\frac{t^2}{2\sigma^2}}\,.$$

Exemple 2.2 (Variable de Poisson)**.** Si $Z \sim \mathcal{P}(\theta)$, avec $\theta > 0$, on a pour tout $\lambda \in \mathbb{R}$,

$$\mathbf{E}e^{\lambda Z} = \sum_{k\geq 0} e^{\lambda k} \frac{e^{-\theta}\theta^k}{k!} = e^{\theta(e^\lambda - 1)}\,,$$

et ainsi

$$\psi_Z(\lambda) = \theta(e^\lambda - \lambda - 1)\,.$$

On obtient alors, pour $t \geq 0$,

$$\lambda_t = \log\left(1 + \frac{t}{\theta}\right) \quad \text{et} \quad \psi_Z^*(t) = \theta h\left(\frac{t}{\theta}\right)\,,$$

avec h définie pour $x \geq -1$ par $h(x) = (1+x)\log(1+x) - x$. On obtient de même, pour $0 \leq t \leq \theta$, $\psi_{-Z}^*(t) = \theta h\left(-\frac{t}{\theta}\right)$.

Exemple 2.3 (Variable Gamma)**.** Soit $Z \sim \Gamma(p, \theta)$ une variable de loi Gamma de paramètres $p, \theta > 0$, de densité donnée par

$$x \mapsto \frac{\theta^p}{\Gamma(p)} x^{p-1} e^{-\theta x} \mathbb{1}_{\{x\geq 0\}}\,.$$

On peut facilement vérifier que $\mathbf{E}Z = \frac{p}{\theta}$ et $\mathbf{Var}\, Z = \frac{p}{\theta^2}$. On a pour tout $\lambda < \theta$,

$$\psi_Z(\lambda) = -\frac{\lambda p}{\theta} - p\log\left(1 - \frac{\lambda}{\theta}\right)\,.$$

En utilisant l'inégalité (laissée en exercice)

$$\forall u \in [0,1[,\ -\log(1-u) - u \leq \frac{u^2}{2(1-u)}\,,$$

on obtient que pour tout $\lambda \in [0, \theta[$,

$$\psi_Z(\lambda) \leq \frac{p\lambda^2}{2\theta^2(1 - \lambda/\theta)}\,.$$

Pour $\lambda \leq 0$, on peut utiliser l'inégalité $-\log(1-u) - u \leq \frac{u^2}{2}$ pour tout $u \leq 0$ et obtenir que pour tout $\lambda \leq 0$,

$$\psi_Z(\lambda) \leq \frac{p\lambda^2}{2\theta^2}\,.$$

2. Variables sous-gaussiennes, sous-Poisson, sous-gamma

On dit qu'une variable Z est sous-gaussienne avec facteur de variance $\sigma^2 > 0$ si pour tout $\lambda \in \mathbb{R}$,

$$\psi_Z(\lambda) \leq \frac{\lambda^2\sigma^2}{2}\,.$$

Si cette égalité est vérifiée pour $\lambda \geq 0$ (resp. $\lambda \leq 0$), on dit qu'elle est sous-gaussienne à droite (resp. à gauche).

On dit qu'une variable Z est sous-Poisson avec facteur de variance $v > 0$ et facteur d'échelle $c > 0$ si pour tout $\lambda \in \mathbb{R}$,

$$\psi_Z(\lambda) \leq \frac{v}{c^2}(e^{c\lambda} - c\lambda - 1)\,.$$

Si cette égalité est vérifiée pour $\lambda \geq 0$ (resp. $\lambda \leq 0$), on dit qu'elle est sous-Poisson à droite (resp. à gauche).

On dit qu'une variable Z est sous-gamma à droite avec facteur de variance $v > 0$ et facteur d'échelle $c > 0$ si pour tout $\lambda \in [0, 1/c[$,

$$\psi_Z(\lambda) \leq \frac{v\lambda^2}{2(1-c\lambda)}\,.$$

Ainsi par example, une variable de loi $\Gamma(p, \theta)$ est sous-gamma à droite avec facteur variance $v = p/\theta^2$ et facteur d'échelle $c = 1/\theta$, et sous-gaussienne à gauche avec facteur de variance v.

Proposition 2.1. *Une variable sous-Poisson avec facteur de variance $v > 0$ et facteur d'échelle $c > 0$ est sous-gaussienne à gauche avec facteur de variance v, et sous-gamma à droite avec facteur de variance v et facteur d'échelle $c/3$.*

Preuve de la Proposition 2.1. Supposons que pour tout $\lambda \in \mathbb{R}$,

$$\psi_Z(\lambda) \leq \frac{v}{c^2}(e^{c\lambda} - c\lambda - 1)\,.$$

Pour $\lambda \leq 0$, on a $e^{c\lambda} - c\lambda - 1 \leq \frac{(c\lambda)^2}{2}$. Ainsi, pour tout $\lambda \leq 0$, on a bien $\psi_Z(\lambda) \leq \frac{v\lambda^2}{2}$.

Pour $\lambda \geq 0$, on a

$$e^{c\lambda} - c\lambda - 1 = \sum_{k=2}^{+\infty} \frac{(c\lambda)^k}{k!} = (c\lambda)^2 \sum_{k=0}^{+\infty} \frac{(c\lambda)^k}{(k+2)!}\,.$$

En utilisant que pour tout $k \geq 0$, $(k+2)! \geq 2 \cdot 3^k$, on obtient

$$\psi_Z(\lambda) \leq \frac{v\lambda^2}{2} \sum_{k=0}^{+\infty} \left(\frac{c\lambda}{3}\right)^k\,.$$

Ainsi, pour tout $\lambda \in [0, 3/c[$, on a

$$\psi_Z(\lambda) \leq \frac{v\lambda^2}{2\,(1 - c\lambda/3)}\,.$$

■

Proposition 2.2. *Si Z est sous-gamma à droite avec facteur de variance $v > 0$ et facteur d'échelle $c > 0$, alors, pour tout $t \geq 0$,*

$$\mathbf{P}(Z - \mathbf{E}Z \geq t) \leq \exp\left\{-\frac{t^2}{2(v + ct)}\right\}\,.$$

*Démonstration.*Voir feuilles d'exercices. ■

3. Sommes de variables indépendantes

Soit $(X_1, \ldots, X_n)$ une suite de v.a.r. indépendantes et $Z = \sum_{i=1}^n X_i$. La transformée de Laplace de Z s'exprime facilement en fonction de celle des X_i : $\mathbf{E}e^{\lambda Z} = \prod_{i=1}^n \mathbf{E}e^{\lambda X_i}$, et ainsi

$$\psi_Z(\lambda) = \sum_{i=1}^n \psi_{X_i}(\lambda)\,.$$

Exemple 2.4 (Loi binomiale)**.** Soient $X_1, \dots, X_n$ des variables i.i.d. de loi de Bernoulli $\mathcal{B}(p)$, pour $p \in [0,1]$, et $Z = \sum_{i=1}^n X_i$. On a, pour tout $\lambda \in \mathbb{R}$,

$$\psi_Z(\lambda) = n\left\{\log\left(pe^\lambda + 1 - p\right) - \lambda p\right\} \leq np(e^\lambda - \lambda - 1)\,,$$

où l'on a utilisé $\log(1+x) \leq x$.Ainsi Z est sous-poisson avec facteur de variance np et facteur d'échelle 1. En combinant les Propositions 2.1 et 2.2, on a, pour $t \geq 0$,

$$\mathbf{P}\left(Z - \mathbf{E}Z \geq t\right) \leq \exp\left\{-\frac{t^2}{2(np + t/3)}\right\}\,,$$

et

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

Remarquons que si $t \in [0, np]$, on obtient

$$\mathbf{P}\left(|Z - \mathbf{E}Z| \geq t\right) \leq 2\exp\left\{-\frac{3t^2}{8np}\right\}\,. \tag{2.1}$$

Dans l'exemple ci-dessus, on sait calculer explicitement la transformée de Laplace de chaque variable. Ce que nous allons voir maintenant, c'est que l'on peut obtenir des bornes parfois très bonnes sur la transformée de Laplace avec seulement très peu d'informations sur la loi des variables.

3.1. Inégalité d'Hoeffding.

Proposition 2.3 (Inégalité d'Hoeffding)**.** *Soit $(X_1, \dots, X_n)$ une suite de v.a.r. indépendantes et $Z = \sum_{i=1}^n X_i$. Si pour tout $i \in [\![1, n]\!]$, il existe $a_i, b_i \in \mathbb{R}$ tels que $a_i \leq X_i \leq b_i$, alors, pour tout $t \geq 0$,*

$$\mathbf{P}\left(Z - \mathbf{E}Z \geq t\right) \leq \exp\left\{-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right\}\,.$$

Preuve de la Proposition 2.3. Montrons d'abord le résultat suivant : si X est une v.a.r. telle que $a \leq X \leq b$, alors pour tout $\lambda \in \mathbb{R}$,

$$\log \mathbf{E}\left[e^{\lambda(X - \mathbf{E}X)}\right] \leq \frac{\lambda^2 (b-a)^2}{8}\,. \tag{2.2}$$

En effet, posons $Y = \frac{X-a}{b-a}$ et $p = \frac{\mathbf{E}X - a}{b-a}$, de telle sorte que $0 \leq Y \leq 1$ et $\mathbf{E}Y = p$. Par convexité de $\lambda \mapsto e^{\lambda Y}$, on a $e^{\lambda Y} \leq Ye^\lambda + 1 - Y$, si bien que

$$\log \mathbf{E}e^{\lambda(Y - \mathbf{E}Y)} \leq \log(pe^\lambda + 1 - p) - \lambda p := \varphi(\lambda)\,.$$

Par le théorème de Taylor, il existe θ entre 0 et λ tel que

$$\varphi(\lambda) = \varphi(0) + \lambda\varphi'(0) + \frac{\lambda^2}{2}\varphi''(\theta)\,.$$

Il suffit maintenant de remarquer que $\varphi(0) = \varphi'(0) = 0$ et que

$$\varphi''(\theta) = \frac{p(1-p)e^\theta}{(pe^\theta + 1 - p)^2} \leq \frac{1}{4}\,.$$

On obtient alors

$$\log \mathbf{E}e^{\lambda(X - \mathbf{E}X)} = \log \mathbf{E}e^{(b-a)\lambda(Y - \mathbf{E}Y)} \leq \frac{\lambda^2 (b-a)^2}{8}\,.$$

Maintenant, par indépendance des X_i, on a

$$\log \mathbf{E} e^{\lambda(Z-\mathbf{E}Z)} = \sum_{i=1}^{n} \log \mathbf{E} e^{\lambda(X_i - \mathbf{E}X_i)} \leq \frac{\lambda^2}{8} \sum_{i=1}^{n} (b_i - a_i)^2 \,,$$

et la méthode de Cramér-Chernoff donne alors, pour tout $t > 0$,

$$\mathbf{P}\left(Z - \mathbf{E}Z \geq t\right) \leq e^{-\sup_{\lambda \geq 0}\left\{\lambda t - \frac{\lambda^2 v}{8}\right\}} = e^{-\frac{2t^2}{v}} \,,$$

où l'on a noté $v = \sum_{i=1}^{n} (b_i - a_i)^2$. ■

Remarque 2.5. En appliquant le résultat à $-Z$ et en utilisant une borne union, on a

$$\mathbf{P}\left(|Z - \mathbf{E}Z| \geq t\right) \leq 2 \exp\left\{-\frac{2t^2}{\sum_{i=1}^{n} (b_i - a_i)^2}\right\} .$$

Exemple 2.6 (Loi binomiale)**.** Si $X_1, \ldots, X_n$ sont des variables i.i.d. de loi de Bernoulli $\mathcal{B}(p)$, $p \in [0,1]$, l'inégalité de Hoeffding donne, pour tout $t \geq 0$,

$$\mathbf{P}\left(\sum_{i=1}^{n} X_i - np \geq t\right) \leq \exp\left\{-\frac{2t^2}{n}\right\} .$$

On obtient ainsi une inégalité sous-gaussienne avec un facteur de variance de l'ordre de n. Pour p fixé dans $]0,1[$, c'est bien le bon ordre pour la variance d'une variable binomiale. Mais si $p \ll 1$, par exemple pour $p = \frac{1}{n}$, cela devient une très mauvaise borne, la vraie variance étant d'ordre 1. Le comportement de la somme n'est plus gaussien, mais poissonien (ce que l'on avait remarqué dans l'exemple 2.4) et appliquer l'inégalité de Hoeffding n'est pas judicieux.

3.2. Inégalité de Bennett.

Proposition 2.4 (Inégalité de Bennett)**.** *Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes de variance finie et telles que $X_i \leq c$ avec $c > 0$. On pose $Z = \sum_{i=1}^{n} X_i$ et $v = \sum_{i=1}^{n} \mathbf{E}[X_i^2]$. Alors, pour tout $\lambda \geq 0$,*

$$\log \mathbf{E} e^{\lambda(Z-\mathbf{E}Z)} \leq \frac{v}{c^2} \phi(c\lambda) \,,$$

avec $\phi(\lambda) = e^{\lambda} - \lambda - 1$. De plus, pour tout $t \geq 0$,

$$\mathbf{P}\left(Z - \mathbf{E}Z \geq t\right) \leq \exp\left\{-\frac{v}{c^2} h\left(\frac{ct}{v}\right)\right\} ,$$

avec $h(x) = (1+x)\log(1+x) - x$.

Remarque 2.7. L'inégalité de Bennett affirme que si chaque variable d'une suite indépendante est majorée par c, alors la somme est sous-Poisson à droite avec facteur de variance v, la somme des moments d'ordre 2, et facteur d'échelle c, donc sous-gamma à droite avec facteurs v et $c/3$. Ainsi, on a la borne plus manipulable

$$\mathbf{P}\left(Z - \mathbf{E}Z \geq t\right) \leq \exp\left\{-\frac{t^2}{2(v + ct/3)}\right\} . \tag{2.3}$$

Preuve de la Proposition 2.4. Par homogénéité, on peut supposer $c = 1$. Remarquons d'abord que la fonction $u \mapsto \frac{\phi(u)}{u^2}$ (prolongée par continuité en 0) est croissante sur $\mathbb{R}$. Comme $X_i \leq 1$, on a alors, pour $\lambda \geq 0$,

$$e^{\lambda X_i} = 1 + \lambda X_i + \phi(\lambda X_i) \leq 1 + \lambda X_i + X_i^2 \phi(\lambda) \,.$$

Ainsi,

$$\begin{aligned}\log \mathbf{E} e^{\lambda(Z-\mathbf{E}Z)} &= \sum_{i=1}^{n} \log \mathbf{E}\left[e^{\lambda(X_i-\mathbf{E}X_i)}\right] \\ &\leq \sum_{i=1}^{n} \left\{\log\left(1+\lambda \mathbf{E}X_i + \mathbf{E}[X_i^2]\phi(\lambda)\right) - \lambda \mathbf{E}X_i\right\} \\ &\leq \sum_{i=1}^{n} \mathbf{E}[X_i^2]\phi(\lambda)\,,\end{aligned}$$

où pour la dernière inégalité on a utilisé $\log(1+x) \leq x$. Maintenant, pour $t \geq 0$, la fonction $\lambda \mapsto \lambda t - v\phi(\lambda)$ est maximale en $\lambda = \log\left(1+\frac{t}{v}\right)$ et vaut en ce point

$$t\log\left(1+\frac{t}{v}\right) - v\left(\frac{t}{v} - \log\left(1+\frac{t}{v}\right)\right) = vh\left(\frac{t}{v}\right)\,.$$

■

Exemple 2.8 (Loi binomiale)**.** Reprenons l'exemple de la loi binomiale de paramètres n et $1/n$. En appliquant l'inégalité de Bennett (ou plutôt la version (2.3)) avec $c=1$ et $v=1$, on a

$$\mathbf{P}\left(\sum_{i=1}^{n} X_i - 1 \geq t\right) \leq \exp\left\{-\frac{t^2}{2(1+t/3)}\right\}\,.$$

Pour t fixé (i.e. ne dépendant pas de n), cela donne une bien meilleure concentration que celle qui provenait de l'inégalité de Hoeffding. Morale de l'histoire : ne pas forcer une variable poissonnienne à être sous-gaussienne !

Exemple 2.9 (Degrés dans un graphe aléatoire dense)**.** Soit $G=(V,E) \sim \mathcal{G}(n,p_n)$ un graphe aléatoire d'Erdös–Renyi, c'est-à-dire un graphe dont l'ensemble de sommets V est de cardinal n et dont l'ensemble d'arêtes E est formé en connectant chaque paire de sommets distincts indépendamment avec probabilité p_n. On suppose $np_n \gg \log n$ (régime dit dense). Soit D_u le degré du sommet u, i.e.

$$D_u = \sum_{v \neq u} \mathbb{1}_{\{\{u,v\}\in E\}}\,.$$

Comme les indicatrices sont i.i.d. de loi $\mathcal{B}(p_n)$, on a $D_u \sim \mathrm{Bin}(n-1,p_n)$. En particulier $\mathbf{E}D_u = (n-1)p_n \gg \log n$. Par l'inégalité (2.1), pour $\varepsilon \in]0,1[$,

$$\mathbf{P}\left(\left|\frac{D_u}{(n-1)p_n} - 1\right| \geq \varepsilon\right) \leq 2\exp\left\{-\frac{3\varepsilon^2(n-1)p_n}{8}\right\}\,.$$

Et en utilisant une borne union,

$$\mathbf{P}\left(\bigcup_{u\in V}\left\{\left|\frac{D_u}{(n-1)p_n} - 1\right| \geq \varepsilon\right\}\right) \leq 2n\exp\left\{-\frac{3\varepsilon^2(n-1)p_n}{8}\right\} = o(1)\,.$$

Ainsi, $\sup_{u\in V}\left|\frac{D_u}{(n-1)p_n} - 1\right| \xrightarrow{\mathbf{P}} 0$. Dans le régime dense, le graphe d'Erdös–Renyi est presque régulier : tous les degrés sont concentrés autour de $(n-1)p_n$.

3.3. Inégalité de Bernstein. À la fois l'inégalité de Hoeffding et l'inégalité de Bennett reposent sur le fait que les variables sont bornées (soit des deux soit d'un seul côté). L'inégalité de Bernstein montre que l'on peut établir le comportement sous-gamma d'une somme de variables indépendantes en faisant seulement une hypothèse sur la croissance des moments.

Proposition 2.5 (Inégalité de Bernstein)**.** *Soient $X_1, \dots, X_n$ des variables indépendantes et soit $Z = \sum_{i=1}^n X_i$. On suppose qu'il existe v et c tels que $\sum_{i=1}^n \mathbf{E}X_i^2 \le v$ et*

$$\forall k \ge 3, \qquad \sum_{i=1}^n \mathbf{E}(X_i)_+^k \le \frac{vk!c^{k-2}}{2}\,.$$

Alors pour tout $\lambda \in [0, 1/c[$,

$$\log \mathbf{E}e^{\lambda(Z-\mathbf{E}Z)} \le \frac{v\lambda^2}{2(1-c\lambda)}\,,$$

ce qui implique que pour tout $t \ge 0$,

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

Preuve de la Proposition 2.5. En notant $\phi(u) = e^u - u - 1$ et en utilisant l'inégalité $\log(1+x) \le x$, on a

$$\begin{aligned}\log \mathbf{E}e^{\lambda(Z-\mathbf{E}Z)} &= \sum_{i=1}^n \left\{\log\left(1 + \lambda\mathbf{E}X_i + \mathbf{E}\phi(\lambda X_i)\right) - \lambda\mathbf{E}X_i\right\} \\ &\le \sum_{i=1}^n \mathbf{E}\phi(\lambda X_i)\,.\end{aligned}$$

Comme pour tout $u \le 0$, $\phi(u) \le \frac{u^2}{2}$ (et que $\phi(0) = 0$), on a

$$\phi(\lambda X_i) = \phi(\lambda(X_i)_-) + \phi(\lambda(X_i)_+) \le \frac{\lambda^2(X_i)_-^2}{2} + \sum_{k\ge 2}\frac{\lambda^k(X_i)_+^k}{k!} = \frac{\lambda^2 X_i^2}{2} + \sum_{k\ge 3}\frac{\lambda^k(X_i)_+^k}{k!}\,.$$

Ainsi pour $\lambda \in [0, 1/c[$,

$$\begin{aligned}\sum_{i=1}^n \mathbf{E}\phi(\lambda X_i) &\le \frac{\lambda^2}{2}\sum_{i=1}^n \mathbf{E}[X_i^2] + \sum_{k\ge 3}\frac{\lambda^k}{k!}\sum_{i=1}^n \mathbf{E}[(X_i)_+^k] \\ &\le \frac{\lambda^2 v}{2} + \sum_{k\ge 3}\frac{\lambda^k v c^{k-2}}{2} = \frac{\lambda^2 v}{2}\sum_{k\ge 0}(\lambda c)^k = \frac{v\lambda^2}{2(1-c\lambda)}\,.\end{aligned}$$

■

Exemple 2.10 (Norme d'un vecteur sous-gaussien)**.** Soit $X = (X_1, \dots, X_n)$ un vecteur aléatoire de $\mathbb{R}^n$ dont les coordonnées sont indépendantes. On note

$$\|X\|_2 = \sqrt{\sum_{i=1}^n X_i^2}$$

la norme euclidienne de X. Supposons que chaque coordonnée est d'espérance nulle et de variance 1. En particulier, $\mathbf{E}\|X\|_2^2 = n$. À quel point la norme est-elle concentrée par rapport à son

espérance ? Faisons l'hypothèse supplémentaire que chaque entrée est sous-gaussienne (avec facteur de variance 1) :

$$\forall i \in [\![1, n]\!],\ \forall \lambda \in \mathbb{R},\ \psi_{X_i}(\lambda) \leq \frac{\lambda^2}{2}\,.$$

Montrons que si X est sous-gaussienne avec facteur de variance 1, alors $X^2 - 1$ est sous-gamma à droite avec facteur de variance 16 et facteur d'échelle 2, et sous-gaussienne à gauche avec facteur de variance 16. En effet,

$$\begin{aligned}\mathbf{E}X^{2k} &= \int_0^{+\infty} \mathbf{P}\left(X^{2k} > t\right) \mathrm{d}t = \int_0^{+\infty} \mathbf{P}\left(|X| > t^{\frac{1}{2k}}\right) \mathrm{d}t \\ &\leq \int_0^{+\infty} 2\exp\left\{-\frac{t^{1/k}}{2}\right\} \mathrm{d}t = 2^{k+1}k \int_0^{+\infty} u^{k-1}e^{-u}\mathrm{d}u = 2^{k+1}k!\,,\end{aligned}$$

où l'on a utilisé la majoration de Cramér–Chernoff, puis le changement de variable $u = \frac{t^{1/k}}{2}$. Ainsi, pour tout $\lambda \geq 0$,

$$\log \mathbf{E}e^{-\lambda(X^2-1)} = \lambda + \log \mathbf{E}e^{-\lambda X^2} \leq \lambda + \log\left(1 - \lambda + \frac{\lambda^2}{2}\mathbf{E}X^4\right) \leq 8\lambda^2\,.$$

Par indépendance des X_i, on a donc

$$\forall \lambda \geq 0, \quad \log \mathbf{E}e^{-\lambda(\|X\|_2^2 - n)} \leq 8n\lambda^2\,,$$

ce qui implique

$$\forall t \geq 0, \quad \mathbf{P}\left(\|X\|_2^2 - n \leq -t\right) \leq \exp\left\{-\frac{t^2}{32n}\right\}\,.$$

D'autre part, pour tout $\lambda \in [0, 1/2[$,

$$\log \mathbf{E}\left[e^{\lambda(X^2-1)}\right] = -\lambda + \log\left(1 + \lambda + \mathbf{E}\left[\sum_{k\geq 2} \frac{(\lambda X^2)^k}{k!}\right]\right) \leq \sum_{k\geq 2} \lambda^k 2^{k+1} = \frac{8\lambda^2}{1-2\lambda}\,.$$

et ainsi

$$\forall \lambda \in [0, 1/2[, \quad \log \mathbf{E}e^{\lambda(\|X\|_2^2 - n)} \leq \frac{8n\lambda^2}{1-2\lambda}\,,$$

ce qui implique

$$\forall t \geq 0, \quad \mathbf{P}\left(\|X\|_2^2 \geq n + t\right) \leq \exp\left\{-\frac{t^2}{4(8n+t)}\right\}\,.$$

En combinant les deux inégalités, on obtient

$$\forall t \geq 0, \quad \mathbf{P}\left(\left|\|X\|_2^2 - n\right| \geq t\right) \leq 2\exp\left\{-\frac{t^2}{4(8n+t)}\right\}\,. \tag{2.4}$$

Maintenant utilisons le fait que pour tout $z, \delta \geq 0$, $|z-1| \geq \delta$ implique $|z^2 - 1| \geq \max\{\delta, \delta^2\}$ pour obtenir que

$$\mathbf{P}\left(\left|\frac{\|X\|_2}{\sqrt{n}} - 1\right| \geq \delta\right) \leq \mathbf{P}\left(\left|\frac{\|X\|_2^2}{n} - 1\right| \geq \max\{\delta, \delta^2\}\right) \leq 2\exp\left\{-\frac{n\delta^2}{36}\right\}\,.$$

En posant $t = \delta\sqrt{n}$, on a finalement

$$\mathbf{P}\left(\left|\|X\|_2 - \sqrt{n}\right| \geq t\right) \leq 2\exp\left\{-\frac{t^2}{36}\right\}\,.$$

On a ainsi montré que X était extrêmement proche de la sphère de rayon $\sqrt{n}$: avec grande probabilité, X est à distance *constante* de cette sphère. Le fait que les déviations soient si petites

peut paraître surprenant mais tentons d'en donner l'intuition. On a d'abord montré que la norme au carré $\|X\|_2^2$ était concentrée autour de n avec des fluctuations d'ordre $\sqrt{n}$ (cela est naturel, $\|X\|_2^2$ étant une somme de n variables aléatoires possédant un moment d'ordre 2, cf. TCL). De façon non-rigoureuse, $\|X\|_2^2 = n \pm O(\sqrt{n})$. Mais

$$\sqrt{n \pm O(\sqrt{n})} = \sqrt{n} \pm O(1)\,.$$

Exemple 2.11 (Le lemme de Johnson-Lindenstrauss)**.** Une application surprenante de l'inégalité de Bernstein est le lemme de Johnson-Lindenstrauss, qui énonce qu'étant donnés n points dans un espace euclidien de dimension arbitrairement grande, on peut les plonger dans un espace euclidien de dimension d de telle sorte que toutes les distances entre deux points soient préservées à un facteur $1 \pm \varepsilon$ près, pourvu que d soit plus grand qu'une constante fois $\frac{\log n}{\varepsilon^2}$. La preuve repose sur la méthode probabiliste. Nous allons définir une notion naturelle de projection aléatoire et montrer qu'une projection tirée aléatoirement selon cette loi vérifie la propriété de préservation des distances avec grande probabilité. Pour simplifier, supposons que l'espace de départ est $\mathbb{R}^D$ avec $D \geq 1$. Mais insistons sur le fait que le résultat est complètement indépendant de la dimension de départ D, et que l'on pourrait en fait remplacer $\mathbb{R}^D$ par un espace de Hilbert séparable de dimension infinie.

On construit une application linéaire aléatoire $W : \mathbb{R}^D \to \mathbb{R}^d$ de la façon suivante. Soient $(X_{i,j})_{1\leq i\leq d, 1\leq j\leq D}$ une suite i.i.d. de variables aléatoires sous-gaussiennes avec facteur de variance 1. Pour $\alpha = (\alpha_1, \ldots, \alpha_D) \in \mathbb{R}^D$, on pose

$$W(\alpha) = \frac{1}{\sqrt{d}} \left(W_1(\alpha), \ldots, W_d(\alpha)\right)\,,$$

avec, pour $1 \leq i \leq d$,

$$W_i(\alpha) = \sum_{j=1}^{D} \alpha_j X_{i,j}\,.$$

Soient $x_1, \ldots, x_n$ des points distincts de $\mathbb{R}^D$ et notons $\mathcal{S}$ le sous-ensemble de la boule unité défini par

$$\mathcal{S} = \left\{ \frac{x_i - x_j}{\|x_i - x_j\|},\, 1 \leq i < j \leq n \right\}\,.$$

Nous allons montrer que pour tout $\varepsilon, \delta \in]0,1[$, si $d \geq \frac{36}{\varepsilon^2} \log\left(\frac{n^2}{\delta}\right)$, alors

$$\mathbf{P}\left(\sup_{\alpha\in\mathcal{S}} \left| \|W(\alpha)\|^2 - 1 \right| \leq \varepsilon \right) \geq 1 - \delta\,.$$

Autrement dit, avec probabilité $1-\delta$, l'application W vérifie que pour tout $i, j \in [\![1, n]\!]$,

$$(1-\varepsilon)\|x_i - x_j\|^2 \leq \|W(x_i) - W(x_j)\|^2 \leq (1+\varepsilon)\|x_i - x_j\|^2\,.$$

Soit $\alpha \in \mathbb{R}^D$ tel que $\|\alpha\| = 1$. Remarquons que pour tout $i \in [\![1, d]\!]$, et pour tout $\lambda \in \mathbb{R}$,

$$\mathbf{E}\left[e^{\lambda W_i(\alpha)}\right] = \prod_{j=1}^{D} \mathbf{E}\left[e^{\lambda \alpha_j X_{i,j}}\right] \leq \prod_{j=1}^{D} e^{\frac{\alpha_j^2 \lambda^2}{2}} = e^{\frac{\lambda^2}{2}}\,.$$

Ainsi, les variables $W_i(\alpha)$ sont sous-gaussiennes avec facteur de variance 1. Et comme elles sont indépendantes, l'exemple précédent 2.10 s'applique et l'inégalité (2.4), appliquée avec d au lieu

de n, et $d\varepsilon$ au lieu de t, donne

$$\mathbf{P}\left(\left|\|W(\alpha)\|^2 - 1\right| \geq \varepsilon\right) \leq 2\exp\left\{-\frac{d\varepsilon^2}{4(8+\varepsilon)}\right\}.$$

En utilisant que $\varepsilon \leq 1$ et que $|\mathcal{S}| \leq \frac{n(n-1)}{2}$, on obtient

$$\mathbf{P}\left(\sup_{\alpha\in\mathcal{S}}\left|\|W(\alpha)\|^2 - 1\right| \geq \varepsilon\right) \leq n^2\exp\left\{-\frac{d\varepsilon^2}{36}\right\} \leq \delta\,,$$

pour $d \geq \frac{36}{\varepsilon^2}\log\left(\frac{n^2}{\delta}\right)$.

Chapitre 3

L'approche par martingales

Soient $(\Omega, \mathcal{F}, \mathbf{P})$ un espace de probabilité et $Z : \Omega \to \mathbb{R}$ une variable aléatoire intégrable. Soit $(\mathcal{F}_i)_{i=0}^n$ une filtration, i.e. une suite croissante de tribus sur Ω avec $\{\emptyset, \Omega\} = \mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}_n = \mathcal{F}$. La martingale de Doob associée à Z et $(\mathcal{F}_i)$ est la suite de variables aléatoires $(Z_i)_{i=0}^n$ définies par

$$Z_i = \mathbf{E}[Z \mid \mathcal{F}_i] .$$

La suite (Z_i) est adaptée à la filtration $(\mathcal{F}_i)$ (Z_i est $\mathcal{F}_i$-mesurable) et $\mathbf{E}[Z_i \mid \mathcal{F}_{i-1}] = Z_{i-1}$. En remarquant que $Z_n = Z$ et que $Z_0 = \mathbf{E}Z$, on peut écrire

$$Z - \mathbf{E}Z = \sum_{i=1}^{n} (Z_i - Z_{i-1}) .$$

1. L'inégalité d'Azuma-Hoeffding

Proposition 3.1 (Inégalité d'Azuma–Hoeffding)**.** *Si pour tout $i \in [\![1, n]\!]$, il existe $a_i, b_i \in \mathbb{R}$ tels que $a_i \leq Z_i - Z_{i-1} \leq b_i$, alors, pour tout $t > 0$,*

$$\mathbf{P}\left(Z - \mathbf{E}Z \geq t\right) \leq \exp\left\{ -\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2} \right\} .$$

Preuve de la Proposition 3.1. Comme les variables $Z_0, \ldots, Z_{n-1}$ sont $\mathcal{F}_{n-1}$-mesurables, on a , pour tout $\lambda \geq 0$,

$$\mathbf{E}e^{\lambda(Z - \mathbf{E}Z)} = \mathbf{E}\left[e^{\lambda \sum_{i=1}^{n-1} (Z_i - Z_{i-1})} \mathbf{E}\left[e^{\lambda(Z_n - Z_{n-1})} \mid \mathcal{F}_{n-1} \right] \right] .$$

En utilisant (2.2), on a $\mathbf{E}\left[e^{\lambda(Z_n - Z_{n-1})} \mid \mathcal{F}_{n-1} \right] \leq e^{\frac{\lambda^2 (b_n - a_n)^2}{8}}$. On peut alors procéder de la même manière en conditionnant successivement par $\mathcal{F}_{n-1}, \ldots, \mathcal{F}_0$, et l'on obtient

$$\log \mathbf{E}e^{\lambda(Z - \mathbf{E}Z)} \leq \frac{\lambda^2}{8} \sum_{i=1}^{n} (b_i - a_i)^2 .$$

On conclut en appliquant la méthode de Cramér-Chernoff. ■

Exemple 3.1 (Tirage sans remise)**.** Initialement, une urne contient K boules noires et $N - K$ boules blanches. À chaque temps, on tire uniformément au hasard une boule dans l'urne, sans la remettre. Pour $n \in [\![0, N-1]\!]$, on note X_n le nombre de boules noires dans l'urne après n tirages et $M_n = \frac{X_n}{N-n}$ la proportion correspondante. Remarquons que la suite $(M_n)_{n=0}^{N-1}$ est une martingale adaptée à la filtration $\mathcal{F}_n = \sigma(X_0, \ldots, X_n)$. En effet, comme au temps n on tire une boule noire avec probabilité M_{n-1}, on a

$$\mathbf{E}[M_n \mid \mathcal{F}_{n-1}] = \frac{X_{n-1} - M_{n-1}}{N - n} = \frac{X_{n-1}}{N - n + 1} = M_{n-1} .$$

En particulier, $\mathbf{E}M_n = M_0 = \frac{K}{N}$. De plus, en utilisant que $0 \leq X_{n-1} - X_n \leq 1$ et que $0 \leq X_{n-1} \leq N - n + 1$, on a

$$-\frac{1}{N-n} \leq M_n - M_{n-1} \leq \frac{1}{N-n}\cdot$$

Ainsi $\sum_{i=1}^n (b_i - a_i)^2 \leq 4\sum_{i=1}^n \frac{1}{(N-i)^2} \leq \frac{4n}{(N-n)^2}$ et l'inégalité d'Azuma–Hoeffding donne

$$\mathbf{P}\left(M_n - \mathbf{E}M_n \geq \varepsilon\right) \leq \exp\left\{-\frac{\varepsilon^2(N-n)^2}{2n}\right\}.$$

Une conséquence majeure de l'inégalité d'Azuma–Hoeffding est l'inégalité des différences bornées.

2. L'inégalité des différences bornées

Dans le cas où $Z = f(X_1, \ldots, X_n)$ avec $X_1, \ldots, X_n$ indépendantes, l'inégalité d'Azuma-Hoeffding a pour conséquence un résultat essentiel : l'inégalité des différences bornées.

Corollaire 3.2 (L'inégalité des différences bornées)**.** *Soit $(X_1, \ldots, X_n)$ une suite de variables aléatoires indépendantes à valeurs dans un espace mesurable $\mathcal{X}$ et $Z = f(X_1, \ldots, X_n)$, avec $f : \mathcal{X}^n \to \mathbb{R}$ à différences bornées avec constantes $c_1, \ldots, c_n \geq 0$ (voir Définition 1.1). Alors pour tout $t > 0$,*

$$\mathbf{P}\left(Z - \mathbf{E}Z \geq t\right) \leq \exp\left\{-\frac{t^2}{2\sum_{i=1}^n c_i^2}\right\}.$$

Preuve du Corollaire 3.2. Posons $\mathcal{F}_k = \sigma(X_1, \ldots, X_k)$ et soit $(X_1', \ldots, X_n')$ une copie indépendante de $(X_1, \ldots, X_n)$. Par indépendance, on a

$$\mathbf{E}\left[Z \,\middle|\, \mathcal{F}_{k-1}\right] = \mathbf{E}\left[f(X_1, \ldots, X_{k-1}, X_k', X_{k+1}, \ldots, X_n) \,\middle|\, X_1, \ldots, X_k\right].$$

Ainsi

$$\mathbf{E}\left[Z \,\middle|\, \mathcal{F}_k\right] - \mathbf{E}\left[Z \,\middle|\, \mathcal{F}_{k-1}\right] = \mathbf{E}\left[f(X_1, \ldots, X_k, \ldots, X_n) - f(X_1, \ldots, X_k', \ldots, X_n) \,\middle|\, \mathcal{F}_k\right],$$

qui est contenu dans $[-c_k, c_k]$ par hypothèse. On conclut en appliquant l'inégalité d'Azuma–Hoeffding. ■

Exemple 3.2 (Bins and balls)**.** Reprenons l'exemple 1.2. Si l'on change le résultat du $i^{\text{ième}}$ lancer, la variable K_n soit reste la même, soit est modifiée de -1 ou 1. L'inégalité des différences bornées donne

$$\mathbf{P}\left(K_n - \mathbf{E}K_n \geq t\right) \leq \exp\left\{-\frac{t^2}{2n}\right\}.$$

Nous verrons en Section 7 que cette inégalité peut être significativement améliorées.

Exemple 3.3 (Bin packing)**.** Étant donnés $x_1, \ldots, x_n \in [0,1]$, quel est le nombre minimum de cases de taille unitaire nécessaires pour contenir ces éléments, de telle sorte que la somme des éléments dans chaque case n'excède pas 1 ? Notons $M_n = f(X_1, \ldots, X_n)$ ce nombre minimum, où $X_1, \ldots, X_n$ sont i.i.d. à support dans $[0,1]$. Comme changer un des X_i ne peut pas changer la valeur de M_n de plus que -1 ou 1, l'inégalité des différences bornées donne

$$\mathbf{P}\left(M_n - \mathbf{E}M_n \geq t\right) \leq \exp\left\{-\frac{t^2}{2n}\right\}.$$

Exemple 3.4 (Plus longue sous-suite commune)**.** Dans sa version la plus simple, le problème de la plus longue sous-suite commune peut s'énoncer ainsi : soit $(X_1, \dots, X_n, Y_1, \dots, Y_n)$ une suite i.i.d. de loi de Bernoulli $\mathcal{B}(1/2)$. Quelle est la longueur de la plus longue sous-suite commune à $(X_1, \dots, X_n)$ et $(Y_1, \dots, Y_n)$? Formellement, on s'intéresse à

$$L_n = \max\{k,\, X_{i_1} = Y_{j_1}, \dots, X_{i_k} = Y_{j_k},\ \text{ avec } i_1 < \cdots < i_k \text{ et } j_1 < \cdots < j_k\}\,.$$

On sait qu'il existe $\gamma \in [0,1]$ tel que $\frac{\mathbf{E}L_n}{n} \underset{n\to\infty}{\longrightarrow} \gamma$ mais la valeur de γ, appelée constante de Chvátal-Sankoff, est inconnue (on sait que $0,78807 \leq \gamma \leq 0,82628$). Même sans connaître la valeur précise de l'espérance, on peut s'intéresser aux propriétés de concentration de L_n. Là encore, changer une des variables ne peut pas perturber L_n de plus que -1 ou 1, et l'inégalité des différences bornées (appliquée à droite et à gauche) donne

$$\mathbf{P}\left(|L_n - \mathbf{E}L_n| \geq t\right) \leq 2\exp\left\{-\frac{t^2}{4n}\right\}\,.$$

En particulier, pour tout $\varepsilon > 0$,

$$\mathbf{P}\left(\left|\frac{L_n}{\mathbf{E}L_n} - 1\right| \geq \varepsilon\right) \leq 2\exp\left\{-\frac{\varepsilon^2(\mathbf{E}L_n)^2}{4n}\right\}\,.$$

Comme on sait que $\mathbf{E}L_n \approx n$, la borne ci-dessus est en $e^{-c_\varepsilon n}$. En particulier, elle correspond donc au terme général d'une série convergente et le lemme de Borel-Cantelli assure alors que L_n vérifie la loi forte des grands nombres :

$$\frac{L_n}{\mathbf{E}L_n} \overset{\text{p.s.}}{\longrightarrow} 1\,.$$

Exemple 3.5 (Le voyageur de commerce)**.** Le problème du voyageur de commerce est un problème classique (et très difficile) d'optimisation combinatoire. Un voyageur de commerce doit visiter n villes en revenant à son point de départ et en empruntant le chemin le plus court. Considérons ici une version aléatoire de ce problème. Supposons que les positions des n villes sont données par des variables i.i.d. $X_1, \dots, X_n$ de loi uniforme sur le carré $[0,1]^2$. On s'intéresse à la variable

$$L_n = \min_{\sigma \in \mathcal{S}_n} \sum_{i=1}^{n} \|X_{\sigma(i)} - X_{\sigma(i+1)}\|\,,$$

où $\sigma(n+1) = \sigma(1)$. Le théorème de Beardwood–Halton–Hammersley affirme qu'il existe $\beta > 0$ tel que

$$\frac{L_n}{\sqrt{n}} \overset{\text{p.s.}}{\longrightarrow} \beta\,.$$

Que peut-on dire de la concentration de L_n par rapport à son espérance ? Voyons ce que donne l'inégalité des différences bornées. Si l'on rejoue la position de la ville i, on modifie L_n d'au plus $2\sqrt{2}$ (et cette borne est atteinte dans le cas extrême où toutes les villes sont d'abord placées en $(0,0)$ et où l'on déplace la ville i en $(1,1)$). On obtient alors

$$\mathbf{P}\left(|L_n - \mathbf{E}L_n| \geq t\right) \leq 2\exp\left(-\frac{t^2}{16n}\right)\,.$$

Cette inégalité n'est pas très satisfaisante : le facteur de variance est de l'ordre de n, alors que l'inégalité d'Efron–Stein nous dit que $\mathbf{Var}\, L_n = O(1)$. En effet, si l'on note $L_n(i)$ la longueur du plus petit parcours lorsque l'on ne prend pas en compte la ville i, on peut observer que

$$L_n(i) \leq L_n \leq L_n(i) + 2\xi_i\,,$$

où $\xi_i = \min_{j\neq i} \|X_i - X_j\|$. Ainsi

$$\mathbf{Var}\, L_n \leq \sum_{i=1}^{n} \mathbf{E}[(L_n - L_n(i))^2] \leq 4\sum_{i=1}^{n} \mathbf{E}\xi_i^2 = 4n\mathbf{E}\xi_1^2 = O(1)\,.$$

Cherchons maintenant à appliquer plus finement l'inégalité d'Azuma–Hoeffding pour obtenir une inégalité exponentielle. En notant $\mathcal{F}_i = \sigma(X_1, \ldots, X_i)$ et en observant que $\mathbf{E}[L_n(i) \mid \mathcal{F}_i] = \mathbf{E}[L_n(i) \mid \mathcal{F}_{i-1}]$, on obtient

$$-2\mathbf{E}[\xi_i \mid \mathcal{F}_{i-1}] \leq \mathbf{E}[L_n \mid \mathcal{F}_i] - \mathbf{E}[L_n \mid \mathcal{F}_{i-1}] \leq 2\mathbf{E}[\xi_i \mid \mathcal{F}_i]\,.$$

Or

$$\max\left\{\mathbf{E}[\xi_i \mid \mathcal{F}_i], \mathbf{E}[\xi_i \mid \mathcal{F}_{i-1}]\right\} \leq \max_{x\in[0,1]^2} \mathbf{E}\left[\min_{i+1\leq k\leq n} \|X_k - x\|\right] \leq \frac{c}{\sqrt{n-i+1}}\,.$$

L'inégalité d'Azuma–Hoeffding donne alors

$$\mathbf{P}\left(|L_n - \mathbf{E}L_n| \geq t\right) \leq 2\exp\left(-\frac{t^2}{8c^2\sum_{i=1}^{n}\frac{1}{n-i+1}}\right) \leq 2\exp\left(-\frac{t^2}{8c^2(\log n + 1)}\right)\,.$$

On verra au Chapitre 5 que l'on peut encore améliorer cette inégalité en supprimant le facteur $\log n$.

3. L'inégalité de Grable

Dans de nombreuses situations, la borne $\sum_{i=1}^{n}(b_i - a_i)^2$ s'avère trop grande par rapport à la vraie variance (cf. l'exemple de la loi $\mathrm{Bin}(n, 1/n)$). On peut néanmoins obtenir une inégalité de type Bernstein faisant intervenir une estimée plus fine de la variance.

Revenons au cadre général du début de chapitre, avec $(Z_i)_{i=0}^n$ la martingale de Doob associée à Z pour la filtration $(\mathcal{F}_i)_{i=0}^n$. Notons $V_i = \mathbf{E}\left[(Z_i - Z_{i-1})^2 \mid \mathcal{F}_{i-1}\right]$. On définit le processus de variation quadratique associé à (Z_i) par

$$\langle Z\rangle_i = \sum_{j=1}^{i} V_j\,.$$

Proposition 3.3 ([11]). *Supposons que $\langle Z\rangle_n \leq v$ avec $v > 0$, et que pour tout $i \in [\![1, n]\!]$, $|Z_i - Z_{i-1}| \leq c$. Alors pour tout $\lambda \in [0, 1/c[$,*

$$\log \mathbf{E}e^{\lambda(Z-\mathbf{E}Z)} \leq \frac{\lambda^2 v}{2(1-\lambda c)}\,.$$

Preuve de la Proposition 3.3. En développant l'exponentielle, on a, pour $\lambda \in [0, 1/c[$,

$$\begin{aligned}
\mathbf{E}\left[e^{\lambda(Z_i - Z_{i-1})} \mid \mathcal{F}_{i-1}\right] &= 1 + \mathbf{E}\left[\sum_{k=2}^{+\infty} \frac{(\lambda(Z_i - Z_{i-1}))^k}{k!} \mid \mathcal{F}_{i-1}\right] \\
&\leq 1 + \lambda^2 V_i \sum_{k=0}^{+\infty} \frac{(\lambda c)^k}{(k+2)!} \\
&\leq 1 + \frac{\lambda^2 V_i}{2(1-\lambda c)} \\
&\leq \exp\left(\frac{\lambda^2 V_i}{2(1-\lambda c)}\right)\,.
\end{aligned}$$

Ainsi, le processus

$$\left(\exp\left(\lambda Z_i - \frac{\lambda^2\langle Z\rangle_i}{2(1-\lambda c)}\right)\right)_{i=0}^{n}$$

est une supermartingale. En particulier

$$\mathbf{E}\left[\exp\left(\lambda Z_n - \frac{\lambda^2\langle Z\rangle_n}{2(1-\lambda c)}\right)\right] \leq \exp\left(\lambda Z_0 - \frac{\lambda^2\langle Z\rangle_0}{2(1-\lambda c)}\right) = \exp\left(\lambda \mathbf{E}Z\right) .$$

Comme $\langle Z\rangle_n \leq v$, on obtient bien

$$\log \mathbf{E}e^{\lambda(Z-\mathbf{E}Z)} \leq \frac{\lambda^2 v}{2(1-\lambda c)} .$$

■

4. L'inégalité de Freedman

L'inégalité de Grable requiert une borne sur la variation quadratique. Or cette borne peut ne pas être valable presque sûrement, mais seulement avec grande probabilité. L'astuce de Freedman pour s'affranchir de l'hypothèse $\langle Z\rangle_n \leq v$ est de passer par un temps d'arrêt bien choisi.

Soit $(Z_n)_{n\geq 0}$ une martingale adaptée à la filtration $(\mathcal{F}_n)_{n\geq 0}$, avec $Z_0 = 0$.

Proposition 3.4 ([10])**.** *Supposons que pour tout $i \in \mathbb{N}^*$, $|Z_i - Z_{i-1}| \leq 1$. Alors pour tout $t \geq 0$, et pour tout $v > 0$, on a*

$$\mathbf{P}\left(\exists n \in \mathbb{N}\,,\, Z_n \geq t\,,\, \langle Z\rangle_n \leq v\right) \leq \left(\frac{v}{v+t}\right)^{v+t} e^t \leq \exp\left\{-\frac{t^2}{2(v+t/3)}\right\} .$$

Preuve de la Proposition 3.4. Soit $\lambda \geq 0$ et $\phi(u) = e^u - u - 1$. En utilisant le fait que $u \mapsto \frac{\phi(u)}{u^2}$ est croissante sur $\mathbb{R}$, on a

$$\begin{aligned}\mathbf{E}\left[e^{\lambda(Z_n - Z_{n-1})} \mid \mathcal{F}_{n-1}\right] &= 1 + \mathbf{E}\left[\phi\left(\lambda(Z_n - Z_{n-1})\right) \mid \mathcal{F}_{n-1}\right] \\ &\leq 1 + \phi(\lambda)V_n \\ &\leq \exp\left(\phi(\lambda)V_n\right) .\end{aligned}$$

Ainsi le processus

$$\left(\exp\left(\lambda Z_n - \phi(\lambda)\langle Z\rangle_n\right)\right)_{n\geq 0}$$

est une supermartingale. En particulier, pour tout temps d'arrêt borné τ, on a

$$\mathbf{E}\left[\exp\left(\lambda Z_\tau - \phi(\lambda)\langle Z\rangle_\tau\right)\right] \leq 1 .$$

Soit $\tau = \inf\{n \geq 0,\, Z_n \geq t\} \cup \{\infty\}$ et soit $\mathcal{E}$ l'événement

$$\mathcal{E} = \left\{\exists n \in \mathbb{N}\,,\, Z_n \geq t\,,\, \langle Z\rangle_n \leq v\right\} .$$

Sur $\mathcal{E}$, on a $\tau < \infty$, $Z_\tau \geq t$, et $\langle Z\rangle_\tau \leq v$ (puisque le processus $(\langle Z\rangle_n)_{n\geq 0}$ est croissant). Ainsi

$$\mathbf{P}(\mathcal{E}) \exp\left(\lambda t - \phi(\lambda)v\right) \leq \mathbf{E}\left[\exp\left(\lambda Z_\tau - \phi(\lambda)\langle Z\rangle_\tau\right)\mathbb{1}_{\mathcal{E}}\right] \leq 1 .$$

Donc pour tout $\lambda \geq 0$, on a $\mathbf{P}(\mathcal{E}) \leq \exp\left(-\{\lambda t - v\phi(\lambda)\}\right)$. Pour $\lambda = \log\left(1 + \frac{t}{v}\right)$, on obtient

$$\mathbf{P}(\mathcal{E}) \leq \exp\left\{-vh\left(\frac{t}{v}\right)\right\} = \left(\frac{v}{v+t}\right)^{v+t} e^t ,$$

avec $h(x) = (1+x)\log(1+x) - x$. La dernière borne de la proposition s'obtient en utilisant $h(u) \geq \frac{u^2}{2(1+u/3)}$. ■

Chapitre 4

La méthode entropique

1. Entropie de Shannon

Soit $\mathcal{X}$ un ensemble dénombrable et X une variable aléatoire à valeurs dans $\mathcal{X}$, de loi P. Pour $x \in \mathcal{X}$, on note $p(x) = \mathbf{P}(X = x)$. L'entropie de P (ou indifféremment l'entropie de X) est définie comme

$$H(P) = H(X) = \mathbf{E}[-\log p(X)] = -\sum_{x\in\mathcal{X}} p(x)\log p(x)\,,$$

avec $0\log 0 = 0$. On a $H(P) \geq 0$, et, si $\mathcal{X}$ est un ensemble fini, alors $H(P) \leq \log|\mathcal{X}|$, la borne étant atteinte par la loi uniforme sur $\mathcal{X}$.

1.1. Un peu de théorie de l'information. D'un point de vue théorie de l'information, il est souvent plus naturel de définir l'entropie en base 2 :

$$H_2(X) = -\sum_{x\in\mathcal{X}} p(x)\log_2 p(x)\,.$$

En effet, $H_2(X)$ représente alors le nombre minimal de bits (0 ou 1) nécessaires pour coder un mot de $\mathcal{X}$ de loi P. Plus précisément, on appelle *code uniquement décodable* une fonction φ de $\mathcal{X}$ dans $\cup_{n\geq 1}\{0,1\}^n$, l'ensemble des suites finies de 0 et de 1, telle que si $(x_1,\ldots,x_n) \in \mathcal{X}^n$ et $(y_1,\ldots,y_m) \in \mathcal{X}^m$ sont deux suites d'éléments de $\mathcal{X}$,

$$\varphi(x_1)\ldots\varphi(x_n) = \varphi(y_1)\ldots\varphi(y_m) \quad \Rightarrow \quad n = m \ \text{ et } \ x_1 = y_1, \ldots, x_n = y_n\,.$$

Autrement dit, on n'a pas besoin de séparer les mots de code pour décoder, la ponctuation est inclue dans le code. Pour $x \in \mathcal{X}$, on note $|\varphi(x)|$ la longueur du mot de code associé à x. Le théorème de Kraft–McMillan affirme que pour tout code uniquement décodable φ,

$$\sum_{x\in\mathcal{X}} 2^{-|\varphi(x)|} \leq 1\,,$$

et qu'inversement, si ℓ est une fonction de $\mathcal{X}$ dans $\mathbb{N}^*$ telle que

$$\sum_{x\in\mathcal{X}} 2^{-\ell(x)} \leq 1\,,$$

alors il existe un code uniquement décodable φ tel que $|\varphi| = \ell$. Par la première assertion du théorème, à tout code uniquement décodable φ est associée une sous-probabilité Q sur $\mathcal{X}$ donnée par $q(x) = 2^{-|\varphi(x)|}$ (on dit que l'on code *selon la loi* Q), et la longueur moyenne d'un mot de code satisfait

$$\mathbf{E}\left[|\varphi(X)|\right] = \sum_{x\in\mathcal{X}} p(x)|\varphi(x)| = -\sum_{x\in\mathcal{X}} p(x)\log_2 q(x) \geq H_2(X)\,.$$

En effet, par l'inégalité de Jensen,

$$H_2(X) + \sum_{x\in\mathcal{X}} p(x)\log_2 q(x) = \sum_{x\in\mathcal{X}} p(x)\log_2 \frac{q(x)}{p(x)} \leq \log_2\left(\sum_{x\in\mathcal{X}} q(x)\right) \leq 0\,.$$

Ainsi, la longueur moyenne de tout mot de code uniquement décodable est au moins égale à l'entropie de la source. Inversement, si l'on pose $\ell(x) = \lceil -\log_2 p(x)\rceil$, alors

$$\sum_{x\in\mathcal{X}} 2^{-\ell(x)} \leq 1\,,$$

et, par la deuxième assertion du théorème, il existe un code uniquement décodable φ tel que $|\varphi| = \ell$. Pour ce code-là, on a alors

$$\mathbf{E}\left[|\varphi(X)|\right] = \sum_{x\in\mathcal{X}} p(x)\lceil -\log_2 p(x)\rceil \leq H_2(X) + 1\,.$$

Si l'on connaît la loi de la source P, on peut donc coder de façon optimale et atteindre la borne inférieure de l'entropie (éventuellement $+1$). En pratique cependant, on ne connaît pas la loi de la source. On ne peut donc pas coder selon P. La longueur moyenne additionnelle due au fait de coder selon Q et non pas selon P est alors donnée par

$$-\sum_{x\in\mathcal{X}} p(x)\log_2 q(x) - H(P) = \sum_{x\in\mathcal{X}} p(x)\log_2 \frac{p(x)}{q(x)}\,.$$

Cette quantité s'appelle la divergence de Kullback–Leibler (ou entropie relative) de P par rapport à Q. C'est le nombre moyen de bits additionnels lorsque l'on code selon Q alors que la source est de loi P.

1.2. Entropie relative. Revenons en base e. Si $Q \ll P$, on définit l'entropie relative de Q par rapport à P par

$$\mathrm{D}(Q \,\big|\, P) = \sum_{x\in\mathcal{X}} q(x)\log\frac{q(x)}{p(x)} = \int_{\mathcal{X}} \log\left(\frac{\mathrm{d}Q}{\mathrm{d}P}\right)\mathrm{d}Q = \mathbf{E}\left[\frac{q}{p}(X)\log\left(\frac{q}{p}(X)\right)\right]\,,$$

avec $X \sim P$. Si Q n'est pas absolument continue par rapport à P, on pose $\mathrm{D}(Q \,\big|\, P) = +\infty$. Par l'inégalité de Jensen, on voit facilement que $\mathrm{D}(Q \,\big|\, P) \geq 0$ et que $\mathrm{D}(Q \,\big|\, P) = 0$ si et seulement si $P = Q$.

1.3. Entropie conditionnelle et *chain rule*. Si (X, Y) est un couple de variables aléatoires à valeurs dans $\mathcal{X}\times\mathcal{X}$, de loi $P_{(X,Y)}$, et si l'on note P_X (resp. P_Y) la loi marginale de X (resp. de Y), alors l'*information mutuelle* de X et Y, notée $I(X, Y)$, est l'entropie relative de la loi $P_{(X,Y)}$ par rapport à la loi produit $P_X \otimes P_Y$, i.e.

$$I(X,Y) = \mathrm{D}(P_{(X,Y)} \,\big|\, P_X \otimes P_Y) = H(X) + H(Y) - H(X,Y)\,.$$

En particulier, cela montre que $H(X, Y) \leq H(X) + H(Y)$, avec égalité si et seulement X et Y sont indépendantes.

L'entropie conditionnelle de X sachant Y est définie par

$$H(X \,\big|\, Y) = H(X,Y) - H(Y)\,.$$

On a

$$I(X,Y) = H(X) - H(X \,\big|\, Y) = H(Y) - H(Y \,\big|\, X)\,.$$

Cela montre que $H(X) \geq H(X \,\big|\, Y)$: ajouter de l'information réduit l'entropie.

En itérant la définition de l'entropie conditionnelle, on obtient que si $X_1, \dots, X_n$ sont des v.a. sur $\mathcal{X}$,

$$H(X_1, \dots, X_n) = H(X_1) + H(X_2 \mid X_1) + \cdots + H(X_n \mid X_1, \dots, X_{n-1}) .$$

C'est ce qu'on appelle la règle de la chaîne (*chain rule* en anglais).

1.4. Inégalité de Han.

Proposition 4.1 (Inégalité de Han)**.** *Soit $(X_1, \dots, X_n)$ une variable aléatoire sur $\mathcal{X}^n$ de loi Q. Alors*

$$H(X_1, \dots, X_n) \leq \frac{1}{n-1} \sum_{i=1}^{n} H(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n) .$$

Autrement dit,

$$H(Q) \leq \frac{1}{n-1} \sum_{i=1}^{n} H(Q^{(i)}) ,$$

où $Q^{(i)}$ est la loi de $(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$.

Preuve de la Proposition 4.1. Par la définition de l'entropie conditionnelle et le fait que le conditionnement réduit l'entropie,

$$\begin{aligned} H(X_1, \dots, X_n) &= H(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n) + H(X_i \mid X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n) \\ &\leq H(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n) + H(X_i \mid X_1, \dots, X_{i-1}) . \end{aligned}$$

Et sommant ces n inégalités et en utilisant la règle de la chaîne, on obtient

$$n H(X_1, \dots, X_n) \leq \sum_{i=1}^{n} H(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n) + H(X_1, \dots, X_n) .$$

■

Corollaire 4.2 (Inégalité de Han pour l'entropie relative)**.** *Soient P et Q deux probabilités sur $\mathcal{X}^n$, avec $P = P_1 \otimes \cdots \otimes P_n$ une mesure produit. Alors*

$$\mathrm{D}(Q \mid P) \leq \sum_{i=1}^{n} \Big(\mathrm{D}(Q \mid P) - \mathrm{D}(Q^{(i)} \mid P^{(i)}) \Big) .$$

Preuve du Corollaire 4.2. Notons $p = p_1 \dots p_n$, q, $p^{(i)}$ et $q^{(i)}$ les densités par rapport à la mesure de comptage de P, Q, $P^{(i)}$ et $Q^{(i)}$ respectivement, et $x^{(i)} = (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$. Remarquons d'abord que, comme P est produit, on a

$$\begin{aligned} \sum_{x \in \mathcal{X}^n} q(x) \log p(x) &= \frac{1}{n} \sum_{i=1}^{n} \sum_{x \in \mathcal{X}^n} q(x) \Big(\log p_i(x_i) + \log p^{(i)}(x^{(i)}) \Big) \\ &= \frac{1}{n} \sum_{x \in \mathcal{X}^n} q(x) \log p(x) + \frac{1}{n} \sum_{i=1}^{n} \sum_{x^{(i)} \in \mathcal{X}^{n-1}} q^{(i)}(x^{(i)}) \log p^{(i)}(x^{(i)}), . \end{aligned}$$

Ainsi

$$\sum_{x \in \mathcal{X}^n} q(x) \log p(x) = \frac{1}{n-1} \sum_{i=1}^{n} \sum_{x^{(i)} \in \mathcal{X}^{n-1}} q^{(i)}(x^{(i)}) \log p^{(i)}(x^{(i)}) .$$

D'autre part, par l'inégalité de Han, $H(Q) \leq \frac{1}{n-1}\sum_{i=1}^n H(Q^{(i)})$, et l'on obtient

$$\begin{aligned}\mathrm{D}(Q \mid P) &= -\sum_{x\in\mathcal{X}^n} q(x)\log p(x) - H(Q)\\ &\geq \frac{1}{n-1}\sum_{i=1}^n \sum_{x^{(i)}\in\mathcal{X}^{n-1}} q^{(i)}(x^{(i)})\log\frac{q^{(i)}(x^{(i)})}{p^{(i)}(x^{(i)})}\\ &= \frac{1}{n-1}\sum_{i=1}^n \mathrm{D}(Q^{(i)} \mid P^{(i)})\,.\end{aligned}$$

En réarrangeant, on obtient bien l'inégalité voulue. ■

2. Sous-additivité de l'entropie

On considère désormais un espace mesurable $(\mathcal{X},\mathcal{E})$ quelconque. Soit $f : \mathcal{X}^n \to \mathbb{R}_+$ une fonction positive et $Z = f(X_1,\dots,X_n)$ avec $(X_1,\dots,X_n)$ à valeurs dans $\mathcal{X}^n$, de loi P. Si $\mathbf{E}[Z\log Z] < +\infty$, on définit l'entropie fonctionnelle de f sous P (ou indifféremment de Z) par

$$\mathbf{Ent}_P f = \mathbf{Ent}[Z] = \mathbf{E}[Z\log Z] - \mathbf{E}Z\log\mathbf{E}Z\,.$$

On peut faire le lien avec la section suivante en remarquant que si $Q \ll P$, alors $\mathrm{D}(Q \mid P) = \mathbf{Ent}\left[\frac{\mathrm{d}Q}{\mathrm{d}P}\right]$.

Proposition 4.3 (Formule de dualité pour l'entropie)**.** *Soit Z une variable aléatoire positive avec $\mathbf{E}[Z\log Z] < +\infty$. On a*

$$\mathbf{Ent}[Z] = \sup_T \mathbf{E}\left[Z(\log T - \log\mathbf{E}T)\right]\,,$$

où le supremum est pris sur les variables aléatoires positives intégrables et telles que $\mathrm{Supp}(Z) \subset \mathrm{Supp}(T)$.

Preuve de la Proposition 4.3. Quitte à considérer la variable $Z/\mathbf{E}Z$, il suffit de montrer que pour toute variable positive avec $\mathbf{E}Z = 1$,

$$\mathbf{Ent}[Z] = \sup_{T\geq 0,\,\mathbf{E}T=1,\mathrm{Supp}(Z)\subset\mathrm{Supp}(T)} \mathbf{E}\left[Z\log T\right]\,.$$

Soit T une v.a. positive avec $\mathbf{E}T = 1$ et $\mathrm{Supp}(Z) \subset \mathrm{Supp}(T)$. En posant $\mathrm{d}\mathbf{Q}(\omega) = T(\omega)\mathrm{d}\mathbf{P}(\omega)$, on a

$$\begin{aligned}\mathbf{Ent}[Z] - \mathbf{E}\left[Z\log T\right] &= \int_{\mathrm{Supp}(T)} Z(\omega)\log\left(\frac{Z(\omega)}{T(\omega)}\right)\mathrm{d}\mathbf{P}(\omega)\\ &= \int_{\mathrm{Supp}(T)} \frac{Z(\omega)}{T(\omega)}\log\left(\frac{Z(\omega)}{T(\omega)}\right)\mathrm{d}\mathbf{Q}(\omega)\,.\end{aligned}$$

Par l'inégalité de Jensen,

$$\int_{\mathrm{Supp}(T)} \frac{Z(\omega)}{T(\omega)}\log\left(\frac{Z(\omega)}{T(\omega)}\right)\mathrm{d}\mathbf{Q}(\omega) \geq \mathbf{E}_{\mathbf{Q}}\left[\frac{Z}{T}\right]\log\mathbf{E}_{\mathbf{Q}}\left[\frac{Z}{T}\right] = \mathbf{E}[Z]\log\mathbf{E}[Z] = 0\,.$$

On voit de plus que le supremum est atteint pour $T = Z$. ■

On peut maintenant établir la sous-additivité de l'entropie dans le cas général.

Proposition 4.4 (Sous-additivité de l'entropie)**.** *Soit $f : \mathcal{X}^n \to \mathbb{R}_+$ et $Z = f(X_1, \dots, X_n)$ avec $X_1, \dots, X_n$ indépendantes. On suppose $\mathbf{E}[Z \log Z] < +\infty$. Alors*

$$\mathbf{Ent}[Z] \leq \mathbf{E}\left[\sum_{i=1}^{n} \mathbf{Ent}^{(i)}[Z]\right] .$$

Preuve de la Proposition 4.4. Introduisons la notation $\mathbf{E}_i = \mathbf{E}[\cdot \mid X_1, \dots, X_i]$ avec $\mathbf{E}_0 = \mathbf{E}$. On peut alors écrire

$$Z (\log Z - \log \mathbf{E}Z) = \sum_{i=1}^{n} Z (\log \mathbf{E}_i Z - \log \mathbf{E}_{i-1} Z) . \tag{4.1}$$

En remarquant que $\mathbf{E}_{i-1} Z = \mathbf{E}^{(i)} \mathbf{E}_i Z$ et en utilisant la Proposition 4.3, on a

$$\mathbf{E}^{(i)} [Z (\log \mathbf{E}_i Z - \log \mathbf{E}_{i-1} Z)] = \mathbf{E}^{(i)} \left[Z \left(\log \mathbf{E}_i Z - \log \mathbf{E}^{(i)} \mathbf{E}_i Z\right)\right] \leq \mathbf{Ent}^{(i)}[Z] ,$$

et en prenant l'espérance dans (4.1), on obtient bien

$$\mathbf{Ent}[Z] = \mathbf{E}\left[\sum_{i=1}^{n} \mathbf{E}^{(i)} [Z (\log \mathbf{E}_i Z - \log \mathbf{E}_{i-1} Z)]\right] \leq \mathbf{E}\left[\sum_{i=1}^{n} \mathbf{Ent}^{(i)}[Z]\right] .$$

■

Donnons une autre caractérisation de l'entropie qui nous sera utile par la suite.

Proposition 4.5. *Soit Z une variable positive avec $\mathbf{E}[Z \log Z] < +\infty$. On a*

$$\mathbf{Ent}[Z] = \inf_{u>0} \mathbf{E} [Z(\log Z - \log u) - (Z - u)] .$$

Preuve de la Proposition 4.5. Cela découle du résultat plus général suivant : soit $\Phi : I \to \mathbb{R}$ une fonction convexe et dérivable définie sur une intervalle ouvert $I \subset \mathbb{R}$ et soit X une variable aléatoire à valeurs dans I. Alors

$$\mathbf{E}\Phi(X) - \Phi(\mathbf{E}X) = \inf_{u \in I} \mathbf{E}\left[\Phi(X) - \Phi(u) - \Phi'(u)(X - u)\right] .$$

En effet, soit $u \in I$. On a

$$\mathbf{E}\left[\Phi(X) - \Phi(u) - \Phi'(u)(X - u)\right] - (\mathbf{E}\Phi(X) - \Phi(\mathbf{E}X)) = \Phi(\mathbf{E}X) - \Phi(u) - \Phi'(u)(\mathbf{E}X - u) .$$

Comme Φ est convexe, la quantité ci-dessus est positive. D'autre part, on voit que le supremum est atteint en $u = \mathbf{E}X$. La Proposition 4.5 vient alors en prenant $I = \mathbb{R}_+$ et $\Phi(x) = x \log x$ (prolongée par continuité en 0). ■

3. Lien avec la transformée de Laplace

Quel est le lien de tout cela avec la concentration ? Nous avons vu qu'une façon d'obtenir une inégalité de concentration pour une variable $Z = f(X_1, \dots, X_n)$ par rapport à son espérance était de majorer la transformée de Laplace $\lambda \mapsto \mathbf{E}[e^{\lambda(Z - \mathbf{E}Z)}]$. Or pour des fonctions f autres que la somme, la transformée de Laplace est généralement difficile à manier. Nous allons voir qu'une majoration de $\mathbf{E}[e^{\lambda(Z-\mathbf{E}Z)}]$ peut être déduite d'une majoration de $\mathbf{Ent}[e^{\lambda Z}]$. En particuler, l'argument de Herbst permet d'obtenir une majoration sous-gaussienne. Comme il est plus facile de majorer l'entropie, notamment grâce à la sous-additivité, cela permet alors d'obtenir des inégalités de concentration exponentielle pour des fonctions de variables indépendantes bien plus complexes que la somme. C'est la méthode entropique.

Proposition 4.6. *Soit Z une variable aléatoire intégrable. Pour tout $\lambda \geq 0$, on a*

$$\psi(\lambda) = \lambda \int_0^\lambda \frac{1}{\gamma^2} \frac{\mathbf{Ent}\left[e^{\gamma Z}\right]}{\mathbf{E}\left[e^{\gamma Z}\right]} \mathrm{d}\gamma \,,$$

où $\psi(\lambda) = \log \mathbf{E} e^{\lambda(Z - \mathbf{E}Z)}$.

Preuve de la Proposition 4.6. On vérifie facilement que

$$\frac{\mathbf{Ent}[e^{\lambda Z}]}{\mathbf{E}e^{\lambda Z}} = \lambda \psi'(\lambda) - \psi(\lambda) \,. \tag{4.2}$$

Et l'on a

$$\lambda \int_0^\lambda \frac{\gamma\psi'(\gamma) - \psi(\gamma)}{\gamma^2} \mathrm{d}\gamma = \lambda \left[\frac{\psi(\gamma)}{\gamma)}\right]_0^\lambda = \psi(\lambda) \,,$$

où l'on a utilisé que $\frac{\psi(\gamma)}{\gamma} \underset{\gamma \to 0^+}{\to} \psi'(0)$. ■

La Proposition 4.6 fournit immédiatement une condition entropique suffisante pour obtenir une inégalité sous-gaussienne, connue sous le nom d'argument de Herbst.

Proposition 4.7 (Argument de Herbst)**.** *Soit Z une variable aléatoire intégrable. S'il existe $v > 0$ tel que pour tout $\lambda \geq 0$,*

$$\mathbf{Ent}\left[e^{\lambda Z}\right] \leq \frac{\lambda^2 v}{2} \mathbf{E}\left[e^{\lambda Z}\right] \,,$$

alors pour tout $\lambda \geq 0$,

$$\psi(\lambda) \leq \frac{\lambda^2 v}{2} \,.$$

Preuve de la Proposition 4.7. Pour tout $\lambda \geq 0$, on a

$$\psi(\lambda) = \lambda \int_0^\lambda \frac{1}{\gamma^2} \frac{\mathbf{Ent}\left[e^{\gamma Z}\right]}{\mathbf{E}\left[e^{\gamma Z}\right]} \mathrm{d}\gamma \leq \lambda \int_0^\lambda \frac{v}{2} \mathrm{d}\gamma = \frac{\lambda^2 v}{2} \,.$$

■

4. Inégalité de Mc Diarmid

Comme première application de la méthode entropique, présentons une amélioration de l'inégalité des différences bornées due à McDiarmid [18].

Proposition 4.8 (Inégalité de McDiarmid)**.** *Soit $f : \mathcal{X}^n \to \mathbb{R}$ et notons*

$$c_i\left(x^{(i)}\right) = \sup_{x_i, x_i'} \left| f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n) \right| \,.$$

S'il existe $v > 0$ tel que pour tout $x \in \mathcal{X}^n$, $\sum_{i=1}^n c_i^2\left(x^{(i)}\right) \leq v$, et si $Z = f(X_1, \ldots, X_n)$ avec $X_1, \ldots, X_n$ indépendantes, alors pour tout $t \geq 0$,

$$\mathbf{P}\left(Z - \mathbf{E}Z \geq t\right) \leq \exp\left\{-\frac{2t^2}{v}\right\} \,.$$

Preuve de la Proposition 4.8. Montrons d'abord le résultat suivant (qui correspond en fait à une version plus forte l'inégalité de Hoeffding) : si X est une variable aléatoire prenant ses valeurs dans l'intervalle $[a,b]$, alors pour tout $\lambda \in \mathbb{R}$,

$$\frac{\mathbf{Ent}[e^{\lambda X}]}{\mathbf{E}e^{\lambda X}} \leq \frac{(b-a)^2\lambda^2}{8}\,. \tag{4.3}$$

En effet, par (4.2), on a

$$\frac{\mathbf{Ent}[e^{\lambda X}]}{\mathbf{E}e^{\lambda X}} = \lambda\psi'(\lambda) - \psi(\lambda) = \int_0^\lambda u\psi''(u)\mathrm{d}u\,,$$

où $\psi(\lambda) = \log \mathbf{E}e^{\lambda(X-\mathbf{E}X)}$. Or

$$\begin{aligned}\psi''(u) &= \frac{\mathbf{E}[X^2e^{uX}]\mathbf{E}[e^{uX}] - \mathbf{E}[Xe^{uX}]^2}{\mathbf{E}[e^{uX}]^2}\\ &= \mathbf{E}_{\mathbf{Q}_u}[X^2] - \mathbf{E}_{\mathbf{Q}_u}[X]^2\,,\end{aligned}$$

où $\mathbf{Q}_u$ est la mesure de probabilité donnée par

$$\mathrm{d}\mathbf{Q}_u(\omega) = \frac{e^{uX(\omega)}}{\mathbf{E}[e^{uX}]}\mathrm{d}\mathbf{P}(\omega)\,.$$

Ainsi

$$\psi''(u) = \mathbf{Var}_{\mathbf{Q}_u}(X) \leq \mathbf{E}_{\mathbf{Q}_u}\left[\left(X - \frac{a+b}{2}\right)^2\right] \leq \frac{(b-a)^2}{4}\,.$$

On a donc

$$\frac{\mathbf{Ent}[e^{\lambda X}]}{\mathbf{E}e^{\lambda X}} = \int_0^\lambda u\psi''(u)\mathrm{d}u \leq \frac{(b-a)^2\lambda^2}{8}\,.$$

Maintenant par la sous-additivité de l'entropie, on a

$$\mathbf{Ent}[e^{\lambda Z}] \leq \sum_{i=1}^n \mathbf{E}\left[\mathbf{Ent}^{(i)}[e^{\lambda Z}]\right]\,.$$

Notons que conditionnellement à $X^{(i)}$, la variable Z prend ses valeurs dans un intervalle de taille $c_i\left(X^{(i)}\right)$ par hypothèse. Ainsi en utilisant (4.3), on obtient

$$\mathbf{Ent}[e^{\lambda Z}] \leq \sum_{i=1}^n \mathbf{E}\left[\frac{c_i^2\left(X^{(i)}\right)\lambda^2}{8}\mathbf{E}^{(i)}[e^{\lambda Z}]\right] = \sum_{i=1}^n \mathbf{E}\left[\frac{c_i^2\left(X^{(i)}\right)\lambda^2}{8}e^{\lambda Z}\right] \leq \frac{v\lambda^2}{8}\mathbf{E}\left[e^{\lambda Z}\right]\,.$$

L'argument de Herbst nous dit alors que pour tout $\lambda \geq 0$,

$$\log \mathbf{E}\left[e^{\lambda(Z-\mathbf{E}Z)}\right] \leq \frac{v\lambda^2}{8}\,.$$

On conclut par la méthode de Chernoff. ■

5. Une inégalité de Sobolev logarithmique modifiée

De façon plus générale, la sous-additivité de l'entropie implique la majoration suivante sur $\mathbf{Ent}[e^{\lambda Z}]$.

Proposition 4.9. *Soit $f : \mathcal{X}^n \to \mathbb{R}$ et $Z = f(X_1, \dots, X_n)$ avec $X_1, \dots, X_n$ indépendantes. On note $Z_i = f_i(X^{(i)}) = f_i(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$ où $f_i : \mathcal{X}^{n-1} \to \mathbb{R}$ est une fonction arbitraire. Alors pour tout $\lambda \in \mathbb{R}$,*

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

avec $\phi(x) = e^x - x - 1$.

Preuve de la Proposition 4.9. Commençons par utiliser la sous-additivité de l'entropie :

$$\mathbf{Ent}[e^{\lambda Z}] \leq \sum_{i=1}^{n} \mathbf{E}\left[\mathbf{Ent}^{(i)}[e^{\lambda Z}]\right] .$$

En appliquant la Proposition 4.5 avec $u = e^{\lambda Z_i}$, on a

$$\begin{aligned}
\mathbf{Ent}^{(i)}[e^{\lambda Z}] &\leq \mathbf{E}^{(i)}\left[e^{\lambda Z}(\lambda Z - \lambda Z_i) - (e^{\lambda Z} - e^{\lambda Z_i})\right] \\
&= \mathbf{E}^{(i)}\left[e^{\lambda Z}\left(e^{-\lambda(Z - Z_i)} + \lambda(Z - Z_i) - 1\right)\right] \\
&= \mathbf{E}^{(i)}\left[e^{\lambda Z}\phi(-\lambda(Z - Z_i))\right] .
\end{aligned}$$

■

En utilisant la Proposition 4.9, on peut alors améliorer la Proposition 4.8 comme suit.

Proposition 4.10. *Soit $f : \mathcal{X}^n \to \mathbb{R}$ et $Z = f(X_1, \dots, X_n)$ avec $X_1, \dots, X_n$ indépendantes. Notons*

$$Z_i = \inf_{x_i \in \mathcal{X}} f(X_1, \dots, X_{i-1}, x_i, X_{i+1}, \dots, X_n) .$$

S'il existe $v > 0$ tel que $\sum_{i=1}^{n}(Z - Z_i)^2 \leq v$, alors pour tout $t \geq 0$,

$$\mathbf{P}\left(Z - \mathbf{E}Z \geq t\right) \leq \exp\left\{-\frac{t^2}{2v}\right\} .$$

Remarque 4.1. En remplaçant Z par $-Z$, on voit que si $\sum_{i=1}^{n}(Z_i - Z)^2 \leq v$ avec $Z_i = \sup_{x_i \in \mathcal{X}} f(X_1, \dots, X_{i-1}, x_i, X_{i+1}, \dots, X_n)$, alors pour tout $t \geq 0$,

$$\mathbf{P}\left(Z - \mathbf{E}Z \leq -t\right) \leq \exp\left\{-\frac{t^2}{2v}\right\} .$$

Preuve de la Proposition 4.10. Par définition de Z_i et pour $\lambda \geq 0$, on a $-\lambda(Z - Z_i) \leq 0$. Or pour tout $x \leq 0$, $\phi(x) \leq \frac{x^2}{2}$. Ainsi, la Proposition 4.9 donne

$$\mathbf{Ent}[e^{\lambda Z}] \leq \frac{\lambda^2}{2}\sum_{i=1}^{n} \mathbf{E}\left[e^{\lambda Z}(Z - Z_i)^2\right] \leq \frac{v\lambda^2}{2}\mathbf{E}\left[e^{\lambda Z}\right] .$$

On conclut avec l'argument de Herbst et la méthode de Chernoff. ■

Exemple 4.2 (Plus grande valeur propre de matrices aléatoires symétriques)**.** Soit $X = (X_{i,j})_{1\le i,j\le n}$ une matrice aléatoire symétrique dont les entrées $(X_{i,j})_{i\le j}$ sont indépendantes avec $|X_{i,j}| \le 1$, et soit $Z = \lambda_1(A)$ la plus grande valeur propre de A. Soit $v \in \mathbb{R}^n$ tel que $\|v\| = 1$ et

$$Z = {}^t vXv = \sup_{u\in\mathbb{R}^n, \|u\|=1} {}^t uXu\,.$$

Notons $Z_{i,j}$ la plus grande valeur propre de la matrice $X^{i,j}$ où l'on a rejoué l'entrée $X_{i,j}$ par $X'_{i,j}$. On a

$$(Z - Z_{i,j})_+ \le {}^t v(X - X^{i,j})v \mathbb{1}_{Z\ge Z_{i,j}} \le 2|(X_{i,j} - X'_{i,j})v_i v_j| \le 4|v_i v_j|\,.$$

Ainsi,

$$\sum_{i\le j}(Z - Z_{i,j})^2 \le 16\sum_{i,j}|v_i v_j|^2 = 16\|v\|^4 = 16\,,$$

et par la Proposition 4.10,

$$\mathbf{P}(Z - \mathbf{E}Z \ge t) \le \exp\left\{-\frac{t^2}{32}\right\}\,.$$

Remarquons que cet argument ne marche pas pour les déviations à gauche (on verra cependant au Chapitre 5 qu'une borne similaire peut être obtenue).

Une autre conséquence importante de la Proposition 4.10 est la sous-gaussianité des fonctions convexes lipschitziennes.

Proposition 4.11. *Soient $X_1, \dots, X_n$ des variables aléatoires indépendantes à valeurs dans $[0,1]$ et soit $f : [0,1]^n \to \mathbb{R}$ une fonction séparément convexe (i.e. convexe en chaque coordonnée lorsque les autres sont fixées), 1-lipschitzienne (i.e. pour tous $x, y \in [0,1]^n$, $|f(x)-f(y)| \le \|x-y\|$), et continument différentiable. Alors si $Z = f(X_1, \dots, X_n)$, pour tout $t \ge 0$,*

$$\mathbf{P}(Z - \mathbf{E}Z \ge t) \le e^{-\frac{t^2}{2}}\,.$$

Preuve de la Proposition 4.11. Nous allons montrer que $\sum_{i=1}^n (Z - Z_i)^2 \le v$ avec $Z_i = \inf_{x_i} f(X_1, \dots, X_{i-1}, x_i, X_{i+1}, \dots, X_n)$ et utiliser la Proposition 4.10. Soit X'_i la valeur de x_i en laquelle l'infimum dans la définition de Z_i est atteint. Par convexité en chaque coordonnée, on a

$$\sum_{i=1}^n (Z - Z_i)^2 \le \sum_{i=1}^n \left(\frac{\partial f}{\partial x_i}(X)\right)^2 (X_i - X'_i)^2 \le \sum_{i=1}^n \left(\frac{\partial f}{\partial x_i}(X)\right)^2 = \|\nabla f(X)\|^2\,.$$

Et comme f est 1-lipschitzienne, $\|\nabla f(X)\|^2 \le 1$. Notons que là aussi, l'argument ne marche plus pour les déviations à gauche. ■

6. Une autre inégalité de Mc Diarmid

Proposition 4.12. *Soit $Z = f(X_1, \dots, X_n)$, avec $X_1, \dots, X_n$ indépendantes et $f : \mathcal{X}^n \to \mathbb{R}$. On suppose qu'il existe $v, c > 0$ tels que pour tout $i \in [\![1, n]\!]$, on a $Z - \mathbf{E}^{(i)}Z \le c$ et $\sum_{i=1}^n \mathbf{Var}^{(i)} Z \le v$. Alors pour tout $\lambda \ge 0$, on a*

$$\log \mathbf{E}e^{\lambda(Z-\mathbf{E}Z)} \le \frac{v}{c^2}\phi(c\lambda)\,,$$

où $\phi(u) = e^u - u - 1$.

Preuve de la Proposition 4.12. Sans perdre en généralité, on suppose $c = 1$. Montrons d'abord que si X est une variable aléatoire réelle telle que $X - \mathbf{E}X \leq 1$, alors pour tout $\lambda \geq 0$, on a

$$\mathbf{Ent}\left[e^{\lambda X}\right] \leq (\lambda, e^{\lambda} - e^{\lambda} + 1)\,\mathbf{Var}(X)\mathbf{E}\left[e^{\lambda X}\right] .$$

En notant $\psi(\lambda) = \log \mathbf{E}e^{\lambda(X-\mathbf{E}X)}$, on a

$$\frac{\mathbf{Ent}\left[e^{\lambda X}\right]}{\mathbf{E}\left[e^{\lambda X}\right]} = \lambda\psi'(\lambda) - \psi(\lambda) = \int_0^{\lambda} u\psi''(u)\mathrm{d}u\,,$$

et l'on a vu dans la preuve de l'inégalité de Mc Diarmid que $\psi''(u) = \mathbf{Var}_{\mathbf{Q}_u}(X)$, où $\mathbf{Q}_u$ est la mesure de probabilité donnée par $\mathrm{d}\mathbf{Q}_u(\omega) = \frac{e^{uX(\omega)}}{\mathbf{E}[e^{uX}]}\mathrm{d}\mathbf{P}(\omega)$. Or,

$$\mathbf{Var}_{\mathbf{Q}_u}(X) = \mathbf{Var}_{\mathbf{Q}_u}(X - \mathbf{E}X) \leq \mathbf{E}_{\mathbf{Q}_u}[(X - \mathbf{E}X)^2] = \frac{\mathbf{E}\left[e^{uX}(X - \mathbf{E}X)^2\right]}{\mathbf{E}\left[e^{uX}\right]} .$$

Par l'inégalité de Jensen, le dénominateur $\mathbf{E}\left[e^{uX}\right]$ est plus grand que $e^{u\mathbf{E}X}$. On obtient

$$\mathbf{Var}_{\mathbf{Q}_u}(X) \leq \mathbf{E}\left[e^{u(X-\mathbf{E}X)}(X - \mathbf{E}X)^2\right] \leq e^u\mathbf{E}\left[(X - \mathbf{E}X)^2\right] = e^u\,\mathbf{Var}(X)\,,$$

où l'on a utilisé $X - \mathbf{E}X \leq 1$. Ainsi

$$\frac{\mathbf{Ent}\left[e^{\lambda X}\right]}{\mathbf{E}\left[e^{\lambda X}\right]} \leq \mathbf{Var}(X)\int_0^{\lambda} ue^u\mathrm{d}u = \mathbf{Var}(X)(\lambda e^{\lambda} - e^{\lambda} + 1)\,.$$

Maintenant, par la sous-additivité de l'entropie et en appliquant l'inégalité ci-dessus à la loi conditionnelle de Z sachant $X^{(i)}$, on a, pour tout $\lambda \geq 0$,

$$\begin{aligned}
\mathbf{Ent}\left[e^{\lambda Z}\right] &\leq \mathbf{E}\left[\sum_{i=1}^{n}\mathbf{Ent}^{(i)}\left[e^{\lambda Z}\right]\right] \\
&\leq (\lambda e^{\lambda} - e^{\lambda} + 1)\mathbf{E}\left[\sum_{i=1}^{n}\mathbf{E}^{(i)}\left[e^{\lambda Z}\right]\mathbf{Var}^{(i)}(Z)\right] \\
&= (\lambda e^{\lambda} - e^{\lambda} + 1)\mathbf{E}\left[e^{\lambda Z}\sum_{i=1}^{n}\mathbf{Var}^{(i)}(Z)\right] .
\end{aligned}$$

Si $\sum_{i=1}^{n}\mathbf{Var}^{(i)}(Z) \leq v$, alors on a

$$\frac{\mathbf{Ent}\left[e^{\lambda Z}\right]}{\mathbf{E}\left[e^{\lambda Z}\right]} \leq v(\lambda e^{\lambda} - e^{\lambda} + 1)\,.$$

Par la Proposition 4.6, on obtient alors

$$\log \mathbf{E}e^{\lambda(Z-\mathbf{E}Z)} \leq \lambda\int_0^{\lambda}\frac{v(\gamma e^{\gamma} - e^{\gamma} + 1)}{\gamma^2}\mathrm{d}\gamma = v\lambda\left[\frac{e^{\gamma} - 1}{\gamma}\right]_0^{\lambda} = v(e^{\lambda} - \lambda - 1)\,,$$

ce qu'il fallait démontrer. ■

7. Concentration des fonctions auto-bornées

Une fonction $f : \mathcal{X}^n \to \mathbb{R}$ est dite auto-bornée s'il existe des fonctions $f_i : \mathcal{X}^{n-1} \to \mathbb{R}$ telles que pour tout $x \in \mathcal{X}^n$ et $i \in [\![1, n]\!]$,

$$0 \leq f(x) - f_i(x^{(i)}) \leq 1\,,$$

et

$$\sum_{i=1}^{n} \Big(f(x) - f_i(x^{(i)}) \Big) \leq f(x)\,.$$

Proposition 4.13. *Soit $f : \mathcal{X}^n \to \mathbb{R}$ une fonction auto-bornée et $Z = f(X_1, \ldots, X_n)$ avec $X_1, \ldots, X_n$ indépendantes. Alors pour tout $\lambda \in \mathbb{R}$,*

$$\log \mathbf{E} e^{\lambda(Z - \mathbf{E}Z)} \leq \mathbf{E}[Z]\phi(\lambda)\,,$$

où $\phi(\lambda) = e^\lambda - \lambda - 1$.

Remarque 4.3. Une fonction auto-bornée $Z = f(X_1, \ldots, X_n)$ est donc sous-Poisson avec facteur de variance $\mathbf{E}Z$ et facteur d'échelle 1. Par la Proposition 2.1, on a donc, pour tout $t \geq 0$,

$$\mathbf{P}(Z - \mathbf{E}Z \leq -t) \leq e^{-\frac{t^2}{2\mathbf{E}Z}} \quad \text{et} \quad \mathbf{P}(Z - \mathbf{E}Z \geq t) \leq e^{-\frac{t^2}{2(\mathbf{E}Z + t/3)}}\,.$$

Preuve de la Proposition 4.13. Posons $Z_i = f_i(X^{(i)})$. Par la Proposition 4.9, on a, pour tout $\lambda \in \mathbb{R}$,

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

Comme ϕ est une fonction convexe, que $Z - Z_i \in [0, 1]$ et que $\phi(0) = 0$, on a $\phi(-\lambda(Z - Z_i)) \leq (Z - Z_i)\phi(-\lambda)$. Ainsi, comme $\sum_{i=1}^{n}(Z - Z_i) \leq Z$,

$$\mathbf{Ent}[e^{\lambda Z}] \leq \phi(-\lambda)\mathbf{E}\left[Z e^{\lambda Z} \right]\,. \tag{4.4}$$

Par un argument similaire à l'argument de Herbst, montrons que cette inégalité entraîne que pour tout $\lambda \in \mathbb{R}$, $\psi(\lambda) \leq \mathbf{E}Z\phi(\lambda)$ avec $\psi(\lambda) = \log \mathbf{E} e^{\lambda(Z - \mathbf{E}Z)}$. En effet, l'inégalité (4.4) peut se réécrire

$$(1 - e^{-\lambda})\psi'(\lambda) - \psi(\lambda) \leq \mathbf{E}Z\phi(-\lambda)\,,$$

ou encore

$$\psi_1'(\lambda) \leq \mathbf{E}Z\psi_2'(\lambda)\,,$$

où $\psi_1(\lambda) = \frac{\psi(\lambda)}{e^\lambda - 1}$, prolongée par continuité en 0 par $\psi_1(0) = 0$, et $\psi_2(\lambda) = \frac{-\lambda}{e^\lambda - 1}$, prolongée par continuité en 0 par $\psi_2(0) = -1$. En intégrant entre 0 et λ (séparément pour $\lambda \geq 0$ et $\lambda \leq 0$), on obtient bient $\psi(\lambda) \leq \mathbf{E}Z\phi(\lambda)$. ∎

Exemple 4.4 (Bins and balls)**.** Reprenons l'exemple 3.2 du nombre de symboles distincts dans un échantillon $X_1, \ldots, X_n$ i.i.d. de loi $(p_j)_{j \geq 1}$ sur $\mathbb{N}^*$. On a

$$K_n = f(X_1, \ldots, X_n) = \sum_{j \geq 1} \mathbb{1}_{\{\exists i \in [\![1,n]\!],\, X_i = j\}}\,.$$

La fonction f est auto-bornée. En effet, si l'on considère les fonctions f_i données par

$$f_i(x^{(i)}) = \sum_{j \geq 1} \mathbb{1}_{\{\exists k \in [\![1,n]\!] \setminus \{i\},\, x_k = j\}}\,,$$

autrement dit le nombre de symboles distincts lorsque l'on retire l'observation i, alors on a clairement $0 \leq f(x) - f_i(x^{(i)}) \leq 1$ et

$$\begin{aligned}\sum_{i=1}^{n} \Big(f(x) - f_i(x^{(i)}) \Big) &\leq \sum_{i=1}^{n} \sum_{j \geq 1} \mathbb{1}_{\{x_i = j, \text{ et } \forall k \neq i, x_k \neq j\}} \\ &\leq \sum_{j \geq 1} \mathbb{1}_{\{\exists ! i \in [\![1,n]\!], x_i = j\}} \leq f(x)\,.\end{aligned}$$

Ainsi par la Proposition 4.13, K_n est sous-Poissonienne avec facteur de variance $\mathbf{E}K_n$. Notons que ce facteur de variance est toujours bien plus petit que n (le facteur de variance dans l'inégalité sous-gaussienne des différences bornées). En effet, on peut montrer que, quelle que soit la loi sous-jacente, $\frac{\mathbf{E}K_n}{n} \to 0$. Pour la loi géométrique par exemple, $\mathbf{E}K_n$ est de l'ordre de $\log n$. Cependant, cela ne permet toujours pas d'atteindre le facteur variance $\mathbf{E}K_{n,1}$ (l'espérance du nombre de symboles observés une seule fois), donné par l'inégalité d'Efron–Stein, qui dans certains cas peut être bien encore plus petit que $\mathbf{E}K_n$ (par exemple pour la loi géométrique où $\mathbf{E}K_{n,1}$ est d'ordre constant). On peut montrer que la variable K_n est en fait bien sous-Poissonnienne avec facteur de variance $\mathbf{E}K_{n,1}$ (voir Chapitre 8, Section 2).

Chapitre 5

La méthode de transport

L'idée de la méthode de transport est de relier la concentration d'une variable Z à un coût de transport, c'est-à-dire au « prix » à payer pour calculer l'espérance de Z sous une loi Q plutôt que sous la loi P.

1. Le lemme de transport

On rappelle que si P et Q sont deux mesures de probabilité sur un espace mesurable $(\mathcal{X}, \mathcal{E})$, alors l'entropie relative de Q par rapport à P, dite aussi divergence de Kullback–Leibler, est donnée par

$$\mathrm{D}(Q \mid P) = \begin{cases} \int_{\mathcal{X}} \log\left(\frac{\mathrm{d}Q}{\mathrm{d}P}(x)\right) \mathrm{d}Q(x) & \text{si } Q \ll P, \\ +\infty & \text{sinon.} \end{cases}$$

On remarque que si $Q \ll P$, alors $\mathrm{D}(Q \mid P) = \mathbf{Ent}(U)$, avec $U = \frac{\mathrm{d}Q}{\mathrm{d}P}$.

Dans ce qui suit, pour $f : \mathcal{X} \to \mathbb{R}$, on note $\mathbf{E}f = \mathbf{E}_P f = \mathbf{E}[f(X)]$ avec $X \sim P$, et $\mathbf{E}_Q f = \mathbf{E}[f(Y)]$ avec $Y \sim Q$.

Lemme 5.1 (Lemme de transport)**.** *Soit X une variable aléatoire définie sur $(\Omega, \mathcal{F}, \mathbf{P})$, à valeurs dans $(\mathcal{X}, \mathcal{E})$, et soit $Z = f(X)$ avec $f : \mathcal{X} \to \mathbb{R}$ mesurable. Pour tout $\lambda \in \mathbb{R}$, on a*

$$\log \mathbf{E} e^{\lambda(Z - \mathbf{E}Z)} = \sup_{Q \ll P} \left\{ \lambda(\mathbf{E}_Q f - \mathbf{E}f) - \mathrm{D}(Q \mid P) \right\} .$$

En particulier, pour $v > 0$, il y a équivalence entre

(i) pour toute loi $Q \ll P$,

$$\mathbf{E}_Q f - \mathbf{E}f \leq \sqrt{2v\mathrm{D}(Q \mid P)} \,,$$

et

(ii) pour tout $\lambda \geq 0$,

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

Preuve du Lemme 5.1. Soit $\lambda \in \mathbb{R}$. En posant $T = e^{\lambda Z}$ et en identifiant une loi $Q \ll P$ à la variable aléatoire $U = \frac{\mathrm{d}Q}{\mathrm{d}P}(X)$, cela revient à montrer que

$$\log \mathbf{E}T = \sup_{U \geq 0,\, \mathbf{E}U = 1} \left\{ \mathbf{E}[U \log T] - \mathbf{E}[U \log U] \right\} .$$

Or par la Proposition 4.3, pour toute variable $U \geq 0$ avec $\mathbf{E}U = 1$,

$$\mathbf{E}[U \log U] = \mathbf{Ent}(U) \geq \mathbf{E}[U \log T] - \log \mathbf{E}[T] \,,$$

et il y a égalité pour $U = \frac{T}{\mathbf{E}T}$.

Maintenant, si pour tout loi $Q \ll P$, on a $\mathbf{E}_Q f - \mathbf{E} f \leq \sqrt{2v\mathrm{D}(Q \mid P)}$, alors, pour tout $\lambda \geq 0$, on a

$$\log \mathbf{E} e^{\lambda(Z-\mathbf{E}Z)} \leq \sup_{Q \ll P} \left\{ \lambda \sqrt{2v\mathrm{D}(Q \mid P)} - \mathrm{D}(Q \mid P) \right\} .$$

En remarquant que le supremum est atteint pour Q telle que $\mathrm{D}(Q \mid P) = \frac{\lambda^2 v}{2}$, on obtient bien $\log \mathbf{E} e^{\lambda(Z-\mathbf{E}Z)} \leq \frac{\lambda^2 v}{2}$. Inversement, si pour tout $\lambda \geq 0$, $\log \mathbf{E} e^{\lambda(Z-\mathbf{E}Z)} \leq \frac{v\lambda^2}{2}$, alors, pour tout $\lambda > 0$ et toute loi $Q \ll P$, on a

$$\lambda(\mathbf{E}_Q f - \mathbf{E} f) - \mathrm{D}(Q \mid P) \leq \frac{v\lambda^2}{2} ,$$

soit

$$\mathbf{E}_Q f - \mathbf{E} f \leq \frac{\mathrm{D}(Q \mid P)}{\lambda} + \frac{v\lambda}{2} .$$

En prenant $\lambda = \sqrt{\frac{2\mathrm{D}(Q \mid P)}{v}}$, on obtient bien $\mathbf{E}_Q f - \mathbf{E} f \leq \sqrt{2v\mathrm{D}(Q \mid P)}$. ■

Voyons comment se servir du Lemme 5.1 pour établir l'inégalité des différences bornées (avec un facteur de variance divisé par 4). Soit $(\mathcal{X}, \mathcal{E})$ un espace mesurable, $f : \mathcal{X}^n \to \mathbb{R}$ une fonction mesurable, et $Z = f(X)$ avec $X = (X_1, \ldots, X_n)$ de loi produit $P = P_1 \otimes \cdots \otimes P_n$. Soit maintenant $Q \ll P$ et Y une variable aléatoire (définie sur $(\Omega, \mathcal{F}, \mathbf{P})$ aussi) de loi Q.

Jusqu'ici, on a spécifié uniquement les lois marginales P et Q de X et de Y, mais l'on peut définir la loi du couple (X, Y) comme on le souhaite. Notons $\mathcal{C}(P, Q)$ l'ensemble des *couplages* de P et Q, i.e. l'ensemble des couples (X, Y) tels que X est de loi P et Y de loi Q, et soit $(X, Y) \in \mathcal{C}(P, Q)$. Si l'on suppose que pour $c_1, \ldots, c_n \geq$, la fonction f vérifie pour tous $x, y \in \mathcal{X}^n$,

$$f(y) - f(x) \leq \sum_{i=1}^{n} c_i \mathbb{1}_{\{x_i \neq y_i\}} , \tag{5.1}$$

alors on peut écrire

$$\mathbf{E} f(Y) - \mathbf{E} f(X) \leq \sum_{i=1}^{n} c_i \mathbf{P}\left(X_i \neq Y_i\right) \leq \left(\sum_{i=1}^{n} c_i^2 \right)^{1/2} \left(\sum_{i=1}^{n} \mathbf{P}\left(X_i \neq Y_i\right)^2 \right)^{1/2} ,$$

où l'on a utilisé l'inégalité de Cauchy-Schwarz. Comme cela est vrai pour tout couplage (X, Y), on voit que si l'on peut montrer l'inégalité de transport de Marton :

$$\min_{(X,Y) \in \mathcal{C}(P,Q)} \sum_{i=1}^{n} \mathbf{P}\left(X_i \neq Y_i\right)^2 \leq \frac{1}{2} \mathrm{D}(Q \mid P) , \tag{5.2}$$

alors on obtient une inégalité sous-gaussienne avec facteur de variance $v = \frac{1}{4} \sum_{i=1}^{n} c_i^2$, ce qui correspond à l'inégalité des différences bornées (Corollaire 3.2) avec un facteur de variance divisé par 4.

Pour $n = 1$, il s'agit en fait de l'inégalité de Pinsker, qui relie l'entropie relative de deux lois P, Q sur $(\mathcal{X}, \mathcal{E})$ avec leur distance en variation totale, définie par

$$d_{\mathrm{TV}}(P, Q) = \sup_{A \in \mathcal{E}} \{P(A) - Q(A)\} . \tag{5.3}$$

Si p et q sont les densités respectives de P et Q par rapport à une mesure dominante μ (par exemple $\mu = P + Q$), alors on a

$$d_{\mathrm{TV}}(P,Q) = \int_{\mathcal{X}} (p(x) - q(x))_+ \,\mathrm{d}\mu(x) = \frac{1}{2}\int_{\mathcal{X}} |p(x) - q(x)|\,\mathrm{d}\mu(x) = 1 - \int_{\mathcal{X}} p(x) \wedge q(x)\mathrm{d}\mu(x)\,.$$

La première égalité peut se vérifier en remarquant que le sup dans la définition (5.3) est atteint pour l'ensemble $A = \{x \in \mathcal{X},\, p(x) > q(x)\}$, la deuxième vient du fait que p et q sont des densités, et la troisième du fait que $|p(x) - q(x)| = p(x) + q(x) - 2p(x) \wedge q(x)$. Cette distance peut aussi être caractérisée en termes de couplage :

$$d_{\mathrm{TV}}(P,Q) = \min_{(X,Y)\in\mathcal{C}(P,Q)} \mathbf{P}(X \neq Y)\,.$$

En effet, pour tout couplage (X,Y) et pour tout $A \in \mathcal{E}$, on a

$$\mathbf{P}(X \neq Y) \geq \mathbf{P}(X \in A, Y \notin A) \geq \mathbf{P}(X \in A) - \mathbf{P}(Y \in A) = P(A) - Q(A)\,.$$

Ainsi $\min_{(X,Y)\in\mathcal{C}(P,Q)} \mathbf{P}(X \neq Y) \geq d_{\mathrm{TV}}(P,Q)$. Inversement, si $d_{\mathrm{TV}}(P,Q) = 0$, alors $P = Q$ et l'on peut considérer le couplage $X = Y$ qui vérifie l'égalité. Si $d_{\mathrm{TV}}(P,Q) = 1$, alors P et Q ont des supports disjoints et tout couplage vérifie $\mathbf{P}(X \neq Y) = 1$. Enfin, si $d_{\mathrm{TV}}(P,Q) \in]0,1[$, on considère le couplage (X,Y) donné par

$$R = d_{\mathrm{TV}}(P,Q) R_1 + (1 - d_{\mathrm{TV}}(P,Q))\, R_2\,, \tag{5.4}$$

avec, pour toute fonction mesurable bornéee $\Psi : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$,

$$\int_{\mathcal{X}\times\mathcal{X}} \Psi(x,y)\mathrm{d}R_1(x,y) = \int_{\mathcal{X}\times\mathcal{X}} \frac{(p(x) - q(x))_+ (q(y) - p(y))_+}{d_{\mathrm{TV}}(P,Q)^2} \Psi(x,y)\mathrm{d}(\mu\otimes\mu)(x,y)\,,$$

et

$$\int_{\mathcal{X}\times\mathcal{X}} \Psi(x,y)\mathrm{d}R_2(x,y) = \frac{1}{1 - d_{\mathrm{TV}}(P,Q)} \int_{\mathcal{X}} p(x) \wedge q(x) \Psi(x,x)\mathrm{d}\mu(x)\,.$$

On vérifie qu'il s'agit bien d'un couplage de P et Q. De plus, comme R_2 est concentrée sur $\{(x,y) \in \mathcal{X}\times\mathcal{X},\, x = y\}$, on a $\mathbf{P}(X \neq Y) \leq d_{\mathrm{TV}}(P,Q)$.

Proposition 5.2 (Inégalité de Pinsker)**.** *Soient P et Q deux probabilités sur un ensemble mesurable $(\mathcal{X},\mathcal{E})$ avec $Q \ll P$. Alors*

$$d_{\mathrm{TV}}(P,Q)^2 \leq \frac{1}{2}\mathrm{D}(Q \mid P)\,.$$

Preuve de la Proposition 5.2. On remarque que la distance en variation totale peut s'écrire $\mathbf{E}_Q g - \mathbf{E} g$ avec, pour tout $x \in \mathcal{X}$, $g(x) = \mathbb{1}_{\frac{\mathrm{d}Q}{\mathrm{d}P}(x) \geq 1}$. Comme g est à valeurs dans $[0,1]$, la borne de Hoeffding (2.2) donne pour tout $\lambda \geq 0$,

$$\log \mathbf{E} e^{\lambda(g(X) - \mathbf{E}g(X))} \leq \frac{\lambda^2}{8}\,.$$

Et par le Lemme de transport 5.1, on conclut que

$$d_{\mathrm{TV}}(P,Q) = \mathbf{E}_Q g - \mathbf{E} g \leq \sqrt{\frac{\mathrm{D}(Q \mid P)}{2}}\,.$$

■

Établissons maintenant l'inégalité (5.2). Soient $P = P_1 \otimes \cdots \otimes P_n$ et Q deux lois sur $\mathcal{X}^n$ avec $Q \ll P$, et soient $X = (X_1,\ldots,X_n) \sim P$ et $Y = (Y_1,\ldots,Y_n) \sim Q$. Notons $Q_i^{Y_1,\ldots,Y_{i-1}}$

la loi de Y_i sachant $Y_1, \dots, Y_{i-1}$. Pour coupler X et Y, on procède de la façon suivante : on commence par générer (X_1, Y_1) selon le couplage optimal des lois P_1 et Q_1, i.e. tel que $\mathbf{P}(X_1 \neq Y_1) = d_{\mathrm{TV}}(P_1, Q_1)$. Puis, pour $i = 2, \dots, n$, si $(X_1, \dots, X_{i-1})$ et $(Y_1, \dots, Y_{i-1})$ ont été générées, on génère (X_i, Y_i) selon le couplage optimal des lois P_i et $Q_i^{Y_1,\dots,Y_{i-1}}$, i.e. tel que $\mathbf{P}(X_i \neq Y_i \mid Y_1, \dots, Y_{i-1}) = d_{\mathrm{TV}}(P_i, Q_i^{Y_1,\dots,Y_{i-1}})$. En utilisant l'inégalité de Jensen, les propriétés du couplage, puis l'inégalité de Pinsker, on a

$$\begin{aligned}\sum_{i=1}^n \mathbf{P}(X_i \neq Y_i)^2 &\leq \sum_{i=1}^n \mathbf{E}\left[\mathbf{P}(X_i \neq Y_i \mid Y_1, \dots, Y_{i-1})^2\right] \\ &= \sum_{i=1}^n \mathbf{E}\left[d_{\mathrm{TV}}\left(P_i, Q_i^{Y_1,\dots,Y_{i-1}}\right)^2\right] \\ &\leq \frac{1}{2}\sum_{i=1}^n \mathbf{E}\left[\mathrm{D}\left(Q_i^{Y_1,\dots,Y_{i-1}} \mid P_i\right)\right].\end{aligned}$$

Il n'y a plus qu'à utiliser ce qu'on appelle parfois la chain rule pour l'entropie relative pour remarquer que

$$\sum_{i=1}^n \mathbf{E}\left[\mathrm{D}\left(Q_i^{Y_1,\dots,Y_{i-1}} \mid P_i\right)\right] = \mathrm{D}(Q \mid P).$$

2. L'inégalité de transport conditionnelle de Marton

Relâchons l'hypothèse (5.1) et supposons qu'il existe des fonctions mesurables $c_i : \mathcal{X}^n \to \mathbb{R}_+$ telles que pour touts $x, y \in \mathcal{X}^n$,

$$f(y) - f(x) \leq \sum_{i=1}^n c_i(x)\mathbb{1}_{\{x_i \neq y_i\}}. \tag{5.5}$$

En utilisant deux fois l'inégalité de Cauchy-Schwarz, on a

$$\begin{aligned}\mathbf{E}f(Y) - \mathbf{E}f(X) &\leq \sum_{i=1}^n \mathbf{E}\left[c_i(X)\mathbf{P}(X_i \neq Y_i \mid X)\right] \\ &\leq \sum_{i=1}^n \left(\mathbf{E}\left[c_i(X)^2\right]\mathbf{E}\left[\mathbf{P}(X_i \neq Y_i \mid X)^2\right]\right)^{1/2} \\ &\leq \left(\sum_{i=1}^n \mathbf{E}c_i(X)^2\right)^{1/2}\left(\sum_{i=1}^n \mathbf{E}\left[\mathbf{P}(X_i \neq Y_i \mid X)^2\right]\right)^{1/2}.\end{aligned}$$

On voit ainsi que si l'on peut trouver un couplage (X, Y) tel que

$$\sum_{i=1}^n \mathbf{E}\left[\mathbf{P}(X_i \neq Y_i \mid X)^2\right] \leq 2\mathrm{D}(Q \mid P),$$

alors on obtient une inégalité de concentration sous-gaussienne à droite avec facteur de variance $v = \sum_{i=1}^n \mathbf{E}c_i(X)^2$.

Proposition 5.3 (Inégalité de transport conditionnelle de Marton)**.** *Soit $P = P_1 \otimes \cdots \otimes P_n$ une loi produit sur $\mathcal{X}^n$ et $Q \ll P$. Alors*

$$\min_{(X,Y)\in\mathcal{C}(P,Q)} \mathbf{E}\left[\sum_{i=1}^n \left(\mathbf{P}(X_i \neq Y_i \mid X)^2 + \mathbf{P}(X_i \neq Y_i \mid Y)^2\right)\right] \leq 2\mathrm{D}(Q \mid P),$$

Avant de prouver la Proposition 5.3, énonçons deux lemmes importants. Comme ci-dessus, considérons p et q les densités respectives de P et Q par rapport à une même mesure dominante μ, et définissons

$$d_2^2(Q,P) = \int_{\mathcal{X}} \left(1 - \frac{q(x)}{p(x)}\right)_+^2 p(x)\mathrm{d}\mu(x) = \mathbf{E}\left[\left(1 - \frac{q(X)}{p(X)}\right)_+^2\right].$$

Lemme 5.4. *Soit P et Q deux lois de probabilité sur $\mathcal{X}$. Alors*

$$\min_{(X,Y)\in\mathcal{C}(P,Q)} \mathbf{E}\left[\mathbf{P}(X \neq Y \mid X)^2 + \mathbf{P}(X \neq Y \mid Y)^2\right] = d_2^2(Q,P) + d_2^2(P,Q)\,.$$

Preuve du Lemme 5.4. D'une part, pour tout couplage (X,Y), $\mathbf{P}(X = Y \mid X) \leq \min\left\{1, \frac{q(X)}{p(X)}\right\}$. En effet, pour toute fonction mesurable positive φ,

$$\mathbf{E}\left[\varphi(X)\mathbf{P}(X = Y \mid X)\right] = \mathbf{E}\left[\varphi(X)\mathbb{1}_{\{X=Y\}}\right] \leq \mathbf{E}\left[\varphi(Y)\right] = \mathbf{E}\left[\varphi(X)\frac{q(X)}{p(X)}\right].$$

Ainsi $\min_{(X,Y)\in\mathcal{C}(P,Q)} \mathbf{P}(X \neq Y \mid X) \geq \left(1 - \frac{q(X)}{p(X)}\right)_+$. Inversement, en considérant le couplage défini en (5.4), on a

$$\mathbf{P}(X = Y \mid X) = \frac{p(X) \wedge q(X)}{p(X)} = \min\left\{1, \frac{q(X)}{p(X)}\right\},$$

et

$$\mathbf{P}(X = Y \mid Y) = \frac{p(Y) \wedge q(Y)}{q(Y)} = \min\left\{1, \frac{p(Y)}{q(Y)}\right\}.$$

■

Lemme 5.5. *Soit P et Q deux lois de probabilité sur $\mathcal{X}$ avec $Q \ll P$. Alors*

$$d_2^2(Q,P) + d_2^2(P,Q) \leq 2\mathrm{D}(Q \mid P)\,.$$

Preuve du Lemme 5.5. Soit $X \sim P$ et $U = \frac{q(X)}{p(X)}$. On a

$$d_2^2(Q,P) = \mathbf{E}[(1-U)_+^2]\,, \quad \text{et} \quad d_2^2(P,Q) = \mathbf{E}\left[\left(1 - \frac{1}{U}\right)_+^2 U\right].$$

D'autre part,

$$\mathrm{D}(Q \mid P) = \mathbf{E}\left[U\log U + 1 - U\right] = \mathbf{E}\left[h((1-U)_+)\right] + \mathbf{E}\left[h(-(U-1)_+)\right],$$

avec, pour tout $u \geq 0$, $h(t) = (1-t)\log(1-t) + t$. Le résultat s'obtient en vérifiant que pour tout $t \in [0,1]$, $h(t) \geq \frac{t^2}{2}$, et que pour tout $t \geq 0$, $h(-t) \geq \frac{t^2}{2(1+t)}$. ■

Preuve de la Proposition 5.3. Soit $P = P_1 \otimes \cdots \otimes P_n$ et Q deux lois sur $\mathcal{X}^n$ avec $Q \ll P$. Pour coupler X et Y, on procède comme dans la preuve de (5.2). Pour $i = 1, \ldots, n$, si $(X_1, \ldots, X_{i-1})$ et $(Y_1, \ldots, Y_{i-1})$ ont été générées, on génère X_i de loi P_i et Y_i de loi $Q_i^{Y_1,\ldots,Y_{i-1}}$ de telle sorte que

$$\begin{aligned}&\mathbf{E}\left[\mathbf{P}(X_i \neq Y_i \mid Y_1, \ldots, Y_{i-1}, X_i)^2 + \mathbf{P}(X_i \neq Y_i \mid Y_1, \ldots, Y_i)^2 \mid Y_1, \ldots, Y_{i-1}\right] \\ &\quad = d_2^2(Q_i^{Y_1,\ldots,Y_{i-1}}, P_i) + d_2^2(P_i, Q_i^{Y_1,\ldots,Y_{i-1}})\,.\end{aligned}$$

Cela est possible par le Lemme 5.4. Remarquons que, pour ce couplage, la loi conditionnelle de X_i sachant Y est égale à la loi conditionnelle de X_i sachant Y_i, et la loi conditionnelle de Y_i sachant X est égale à la loi conditionnelle de Y_i sachant X_i. Ainsi

$$\mathbf{E}\left[\sum_{i=1}^{n}\left(\mathbf{P}(X_i \neq Y_i \mid X)^2 + \mathbf{P}(X_i \neq Y_i \mid Y)^2\right)\right] = \mathbf{E}\left[\sum_{i=1}^{n}\left(\mathbf{P}(X_i \neq Y_i \mid X_i)^2 + \mathbf{P}(X_i \neq Y_i \mid Y_i)^2\right)\right].$$

En utilisant successivement l'inégalité de Jensen, les propriétés du couplage, le Lemme 5.5, et la chain rule pour l'entropie relative, on a

$$\begin{aligned}
&\mathbf{E}\left[\sum_{i=1}^{n}\left(\mathbf{P}(X_i \neq Y_i \mid X_i)^2 + \mathbf{P}(X_i \neq Y_i \mid Y_i)^2\right)\right] \\
&\leq \mathbf{E}\left[\sum_{i=1}^{n}\left(\mathbf{P}(X_i \neq Y_i \mid Y_1, \dots, Y_{i-1}, X_i)^2 + \mathbf{P}(X_i \neq Y_i \mid Y_1, \dots, Y_i)^2\right)\right] \\
&\leq \mathbf{E}\left[\sum_{i=1}^{n} d_2^2(Q_i^{Y_1,\dots,Y_{i-1}}, P_i) + d_2^2(P_i, Q_i^{Y_1,\dots,Y_{i-1}})\right] \\
&\leq 2\mathbf{E}\left[\sum_{i=1}^{n} \mathrm{D}\left(Q_i^{Y_1,\dots,Y_{i-1}} \mid P_i\right)\right] \\
&= 2\mathrm{D}(Q \mid P)\,.
\end{aligned}$$

■

Proposition 5.6. *Soit $f : \mathcal{X}^n \to \mathbb{R}$ et $X_1, \dots, X_n$ des variables aléatoires indépendantes à valeurs dans $\mathcal{X}$. Notons $Z = f(X_1, \dots, X_n)$. Supposons qu'il existe des fonctions mesurables $c_i : \mathcal{X}^n \to \mathbb{R}_+$ telles que, pour tous $x, y \in \mathcal{X}^n$,*

$$f(y) - f(x) \leq \sum_{i=1}^{n} c_i(x)\mathbb{1}_{y_i \neq x_i}\,,$$

et notons

$$v = \sum_{i=1}^{n} \mathbf{E}[c_i(X)^2]\,, \qquad \text{et} \quad v_\infty = \sup_{x \in \mathcal{X}^n} \sum_{i=1}^{n} c_i(x)^2\,.$$

Alors pour tout $\lambda \geq 0$,

$$\log \mathbf{E}e^{\lambda(Z - \mathbf{E}Z)} \leq \frac{v\lambda^2}{2}\,, \qquad \text{et} \quad \log \mathbf{E}e^{\lambda(\mathbf{E}Z - Z)} \leq \frac{v_\infty \lambda^2}{2}\,.$$

Preuve de la Proposition 5.6. Soit $P = P_1 \otimes \cdots \otimes P_n$ la loi de X. Comme on l'a vu en début de section, pour toute loi $Q \ll P$, si $Y \sim Q$,

$$\mathbf{E}f(Y) - \mathbf{E}f(X) \leq \sqrt{v}\left(\min_{(X,Y) \in \mathcal{C}(P,Q)} \sum_{i=1}^{n} \mathbf{E}\left[\mathbf{P}(X_i \neq Y_i \mid X)^2\right]\right)^{1/2}.$$

Par la Proposition 5.3, on obtient

$$\mathbf{E}f(Y) - \mathbf{E}f(X) \leq \sqrt{2v\mathrm{D}(Q \mid P)}\,,$$

et le lemme de transport donne, pour tout $\lambda \geq 0$, $\log \mathbf{E} e^{\lambda(Z-\mathbf{E}Z)} \leq \frac{v\lambda^2}{2}$. Pour la deuxième inégalité, en considérant la fonction $g = -f$, on a

$$\begin{aligned}\mathbf{E}g(Y) - \mathbf{E}g(X) &\leq \left(\sum_{i=1}^n \mathbf{E}[c_i(Y)^2]\right)^{1/2} \left(\min_{(X,Y)\in\mathcal{C}(P,Q)} \sum_{i=1}^n \mathbf{E}\left[\mathbf{P}(X_i \neq Y_i \mid Y)^2\right]\right)^{1/2} \\ &\leq \sqrt{2v_\infty \mathrm{D}(Q \mid P)}\,.\end{aligned}$$

■

3. L'inégalité de distance convexe de Talagrand

L'inégalité de concentration ci-dessus a de nombreuses conséquences importantes. Elle permet notamment d'obtenir l'inégalité de distance convexe de Talagrand. Pour $A \subset \mathcal{X}^n$ et $\alpha \in \mathbb{R}_+^n$ un vecteur de réels positifs, on définit la distance pondérée $d_\alpha(x, A)$ de $x \in \mathcal{X}^n$ à A par

$$d_\alpha(x, A) = \inf_{y\in A} d_\alpha(x, y) = \inf_{y\in A} \sum_{i=1}^n \alpha_i \mathbb{1}_{x_i \neq y_i}\ .$$

La distance convexe de x à A est alors définie par

$$d_T(x, A) = \sup_{\alpha\in\mathbb{R}_+^n,\, \|\alpha\|\leq 1} d_\alpha(x, A)\,.$$

Proposition 5.7 (Inégalité de distance convexe de Talagrand)**.** *soit $X = (X_1, \ldots, X_n)$ avec $X_1, \ldots, X_n$ indépendantes à valeurs dans $\mathcal{X}$. Pour tout $A \subset \mathcal{X}^n$ et pour tout $t \geq 0$,*

$$\mathbf{P}(X \in A)\mathbf{P}\left(d_T(X, A) \geq t\right) \leq e^{-\frac{t^2}{4}}\,.$$

Preuve de la Proposition 5.7. Notons $\alpha(x) = (\alpha_1(x), \ldots, \alpha_n(x))$ l'élément $\alpha \in \mathbb{R}_+^n$ avec $\|\alpha\| \leq 1$ où le supremum dans la définition de la distance convexe $d_T(x, A)$ est atteint. On a

$$d_T(x, A) - d_T(y, A) \leq \inf_{x'\in A} \sum_{i=1}^n \alpha_i(x)\mathbb{1}_{x_i\neq x'_i} - \inf_{y'\in A} \sum_{i=1}^n \alpha_i(x)\mathbb{1}_{y_i\neq y'_i} \leq \sum_{i=1}^n \alpha_i(x)\mathbb{1}_{x_i\neq y_i}\,.$$

Ainsi la fonction $f : x \mapsto -d_T(x, A)$ vérifie la condition de la Proposition 5.6 avec $c_i = \alpha_i$. Et comme pour tout $x \in \mathcal{X}^n$, on a $\sum_{i=1}^n \alpha_i(x)^2 \leq 1$, la Proposition 5.6 implique que si $X = (X_1, \ldots, X_n)$ est un vecteur de variables indépendantes à valeurs dans $\mathcal{X}$, alors la variable $Z = d_T(X, A)$ est sous-gaussienne avec facteur de variance 1 (à gauche et à droite). Les déviations à droite donnent que tout $t \geq 0$,

$$\mathbf{P}(Z \geq t) \leq e^{\frac{(\mathbf{E}Z)^2}{2} - \frac{t^2}{4}}\,.$$

En effet, si $t < \mathbf{E}Z$, l'inégalité est trivialement vraie et si $t \geq \mathbf{E}Z$, on a $\mathbf{P}(Z \geq t) \leq e^{-\frac{(t-\mathbf{E}Z)^2}{2}} \leq e^{\frac{(\mathbf{E}Z)^2}{2} - \frac{t^2}{4}}$, où l'on a utilisé que $t\mathbf{E}Z \leq (\mathbf{E}Z)^2 + \frac{t^2}{4}$. D'autre part, en utilisant les déviations à gauche, on a

$$\mathbf{P}(X \in A) = \mathbf{P}(Z = 0) \leq \mathbf{P}(Z - \mathbf{E}Z \leq -\mathbf{E}Z) \leq e^{-\frac{(\mathbf{E}Z)^2}{2}}\,.$$

Ainsi $\mathbf{P}(X \in A)\mathbf{P}\left(d_T(X, A) \geq t\right) \leq e^{-\frac{t^2}{4}}$. ■

Exemple 5.1 (Le voyageur de commerce)**.** Reprenons l'exemple 3.5 du voyageur de commerce. Tout d'abord, montrons par récurrence que pour tout $n \geq 1$, pour tout $h > 0$, et pour tous points $x_1, \ldots, x_n$ dans un triangle rectangle d'hypoténuse h, il existe un parcours qui part d'un bout de l'hypoténuse, arrive à l'autre, passe par tous les points, et est tel que la somme des carrés des longueurs de chaque arête est inférieure à h^2. Notons que par le théorème de Pythagore, il suffit de montrer le résultat pour un plus petit triangle rectangle contenant ces n points. Pour $n = 1$, c'est bon. Supposons le résultat vrai jusqu'au rang $n \geq 1$, prenons $n + 1$ points dans le plan et notons h la longueur de l'hypoténuse d'un plus petit triangle rectangle contenant ces points. Si l'on divise ce triangle en deux selon la hauteur issue du sommet opposé à l'hypoténuse, alors il y a au plus n points dans chacun des deux triangles et l'hypothèse de récurrence et le théorème de Pythagore permettent de conclure. Ainsi, pour $x = (x_1, \ldots, x_n)$ avec $x_i \in [0, 1]^2$, on peut trouver un parcours cyclique σ_x passant par $x_1, \ldots, x_n$ tel que la somme des carrés des longueurs de chaque arête est inférieure à 4. Notons $\alpha_i(x)$ deux fois la longueur de l'arête précédant x_i dans ce parcours. On a, pour tous $x, y \in [0, 1]^{2n}$,

$$L_n(x) \leq L_n(y) + \sum_{i=1}^{n} \alpha_i(x) \mathbb{1}_{y_i \neq x_i} .$$

En effet, si x et y n'ont pas de points en communs, alors c'est clair. Sinon, soit σ_y^* un parcours cyclique de longueur minimale passant par $y_1, \ldots, y_n$. On peut alors parcourir les points $x_1, \ldots, x_n$ de la façon suivante : partant d'un point commun à x et y, on parcourt σ_x tant que les points visités ne sont pas communs à y. Si le point suivant sur σ_x, disons u, est commun à y, alors on revient en arrière jusqu'au point commun précédant et on va en u en empruntant σ_y^*. Et ainsi de suite jusqu'à revenir au point de départ. On voit que la longueur de ce parcours est bien plus petite que $L_n(y) + \sum_{i=1}^{n} \alpha_i(x) \mathbb{1}_{y_i \neq x_i}$. Comme

$$\sum_{i=1}^{n} \alpha_i(x)^2 \leq 4 \sum_{i=1}^{n} \|x_{\sigma_x(i)} - x_{\sigma_x(i+1)}\|^2 \leq 16 ,$$

la Proposition 5.6 donne, pour tout $t \geq 0$,

$$\mathbf{P}(|L_n - \mathbf{E}L_n| \geq t) \leq 2e^{-\frac{t^2}{32}} .$$

Chapitre 6

Classification et théorie de Vapnik-Chervonenkis

1. Un problème d'apprentissage statistique

Soit (X, Y) un couple de variables aléatoires avec X à valeurs dans un espace $\mathcal{X}$ et Y à valeurs dans $\{0, 1\}$. Un classifieur est une fonction mesurable $f : \mathcal{X} \to \{0, 1\}$, dont l'objectif est de prédire, à partir de X, la valeur de Y. Le risque d'un classifieur f est donné par

$$\mathbf{R}(f) = \mathbf{P}(Y \neq f(X)) .$$

Si l'on pose $\eta(X) = \mathbf{P}(Y = 1 \mid X)$, il est facile de voir que le classifieur

$$f^\bullet(X) = \mathbb{1}_{\{\eta(X) \geq \frac{1}{2}\}} ,$$

appelé classifieur de Bayes, atteint le plus petit risque possible :

$$\mathbf{R}(f^\bullet) = \min_f \mathbf{R}(f) = \mathbf{E}\left[\min\{\eta(X), 1 - \eta(X)\}\right] .$$

En pratique, le classifieur de Bayes n'est pas d'une grande utilité. Pour pouvoir le calculer, il faut connaître la loi du couple (X, Y) qui est généralement inconnue. Il faut *apprendre* à classifier à partir d'observations issues de cette loi. Plus précisément, on observe un échantillon i.i.d. de même loi que (X, Y) :

$$\mathcal{D}_n = ((X_1, Y_1), \ldots, (X_n, Y_n)) .$$

L'objectif est de construire, à partir de $\mathcal{D}_n$, un classifieur $\widehat{f}_n$ dont le risque soit le plus petit possible. On cherche en fait à minimiser la quantité aléatoire

$$\mathbf{R}(\widehat{f}_n) = \mathbf{P}\left(Y \neq \widehat{f}_n(X) \mid \mathcal{D}_n\right) ,$$

où

Exemple 6.1. Dans le cas où l'ensemble $\mathcal{X}$ est un ensemble discret, un classifieur naturel, appelé classifieur par majorité, est construit de la façon suivante : pour tout $x \in \mathcal{X}$, on calcule

$$N_0(x) = |\{i \in [\![1, n]\!],\, X_i = x,\, Y_i = 0\}| ,$$

et

$$N_1(x) = |\{i \in [\![1, n]\!],\, X_i = x,\, Y_i = 1\}| ,$$

et on pose

$$\widehat{f}_n^{\mathsf{maj}}(x) = \begin{cases} 1 & \text{si } N_1(x) \geq N_0(x), \\ 0 & \text{si } N_0(x) > N_1(x). \end{cases}$$

Autrement dit, on attribue à x le label majoritaire parmi les observations de $\mathcal{D}_n$ pour lesquelles $X_i = x$.

Une méthode souvent utilisée pour construire un classifieur $\widehat{f}_n$ consiste à minimiser le risque empirique. L'idée est d'approcher le risque $\mathbf{R}(f)$ d'un classifieur f par sont équivalent empirique

$$\mathbf{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{Y_i \neq f(X_i)\}}\,.$$

Par la loi des grands nombres, $\mathbf{R}_n(f) \xrightarrow{\mathbf{P}} \mathbf{R}(f)$. Étant donné un ensemble $\mathcal{F}$ de classifieurs (ici supposé dénombrable), souvent appelé dictionnaire, la méthode de minimisation du risque empirique consiste à choisir

$$f_n^* \in \arg\min_{f\in\mathcal{F}} \mathbf{R}_n(f)\,.$$

Remarque 6.2. Le choix de $\mathcal{F}$ est crucial. Prendre $\mathcal{F}$ égal à l'ensemble de tous les classifieurs est souvent un très mauvais choix et conduit au sur-apprentissage (le classifieur colle parfaitement à l'aléa des données mais n'est pas capable de prendre en compte des nouvelles observations). En effet, si $\mathcal{X}$ est assez grand pour que, presque sûrement, toutes les observations X_i soient distinctes, alors le risque empirique est toujours minimisé par le classifieur qui s'ajuste parfaitement aux données, i.e.

$$\widehat{f}_n(x) = \sum_{i=1}^{n} \mathbb{1}_{x=X_i} Y_i\,.$$

Autrement dit, si $x = X_i$, le classifieur répond Y_i et si $x \notin \{X_1, \dots, X_n\}$, il répond, de façon arbitraire, 0. On a alors $\mathbf{R}_n(\widehat{f}_n) = 0$ mais $\mathbf{R}(\widehat{f}_n)$ peut être bien plus grand (plus $\mathcal{F}$ est grand, plus $\sup_{f\in\mathcal{F}} |\mathbf{R}_n(f) - \mathbf{R}(f)|$ est grand). En fait, il faut choisir $\mathcal{F}$ assez grand pour pouvoir approcher le classifieur de Bayes par des éléments de $\mathcal{F}$ mais assez petit pour que $\mathbf{R}_n(f)$ reste une bonne approximation de $\mathbf{R}(f)$, uniformément sur $\mathcal{F}$.

À supposer que le minimum est atteint, une solution idéale au problème d'apprentissage sur $\mathcal{F}$ est donnée par

$$f^* \in \arg\min_{f\in\mathcal{F}} \mathbf{R}(f)\,.$$

La principale question est alors de savoir quelle est l'amplitude de l'excès de risque $\mathbf{R}(f_n^*) - \mathbf{R}(f^*)$. On peut commencer par observer que

$$\mathbf{R}(f_n^*) - \mathbf{R}(f^*) \leq 2\sup_{f\in\mathcal{F}} |\mathbf{R}_n(f) - \mathbf{R}(f)|\,. \tag{6.1}$$

En effet,

$$\begin{aligned}
\mathbf{R}(f_n^*) &\leq \mathbf{R}_n(f_n^*) + \sup_{f\in\mathcal{F}} |\mathbf{R}_n(f) - \mathbf{R}(f)| && \text{puisque } f_n^* \in \mathcal{F} \text{ par construction}\ ,\\
&\leq \mathbf{R}_n(f^*) + \sup_{f\in\mathcal{F}} |\mathbf{R}_n(f) - \mathbf{R}(f)| && \text{puisque } f_n^* \text{ minimise } \mathbf{R}_n \text{ sur } \mathcal{F}\ ,\\
&\leq \mathbf{R}(f^*) + 2\sup_{f\in\mathcal{F}} |\mathbf{R}_n(f) - \mathbf{R}(f)| && \text{puisque } f^* \in \mathcal{F} \text{ par construction}\ .
\end{aligned}$$

Le but de ce chapitre est de contrôler des quantités de la forme $\sup_{f\in\mathcal{F}} |\mathbf{R}_n(f) - \mathbf{R}(f)|$.

2. Inégalités de Vapnik–Chervonenkis

Soit $(\mathcal{X}, \mathcal{E})$ un ensemble mesurable et P une mesure de probabilité sur $\mathcal{X}$.

Définition 6.1. Soit $\mathcal{A}$ un ensemble d'éléments de la tribu $\mathcal{E}$ et $x = (x_1, \dots, x_n)$ un vecteur de n points de $\mathcal{X}$. La trace de $\mathcal{A}$ sur x est définie comme

$$\mathsf{tr}(\mathcal{A}, x) = \{A \cap \{x_1, \dots, x_n\}, A \in \mathcal{A}\} .$$

Le coefficient d'éclatement (*shatter coefficient*) d'ordre n est donné par

$$\mathsf{s}(\mathcal{A}, n) = \max_{x \in \mathcal{X}^n} |\mathsf{tr}(\mathcal{A}, x)| .$$

La dimension de Vapnik-Chervonenkis de $\mathcal{A}$ est définie par

$$\mathsf{V}(\mathcal{A}) = \sup \{n \in \mathbb{N}, \mathsf{s}(\mathcal{A}, n) = 2^n\} .$$

Exemple 6.3 (Dimension de certaines classes)**.**

- si $\mathcal{X} = \mathbb{R}$ et $\mathcal{A} = \{]-\infty, x], x \in \mathbb{R}\}$, alors $\mathsf{V}(\mathcal{A}) = 1$.
- si $\mathcal{X} = \mathbb{R}^d$ et $\mathcal{A} = \left\{\prod_{j=1}^d]-\infty, x_j], x_j \in \mathbb{R}\right\}$, alors $\mathsf{V}(\mathcal{A}) = d$.
- si $\mathcal{X} = \mathbb{R}^d$ et si $\mathcal{A}$ est l'ensemble de tous les demi-espaces, i.e. de tous les sous-ensembles de la forme $\{x \in \mathbb{R}^d, \langle a, x\rangle \geq b\}$, pour $a \in \mathbb{R}^d$ et $b \in \mathbb{R}$, alors $\mathsf{V}(\mathcal{A}) = d + 1$.

Lemme 6.1 (Lemme de Sauer–Shelah)**.** *Pour toute classe $\mathcal{A}$, et pour tout $n \in \mathbb{N}$, on a*

$$\mathsf{s}(\mathcal{A}, n) \leq \sum_{k=0}^{\mathsf{V}(\mathcal{A})} \binom{n}{k} \leq (n+1)^{\mathsf{V}(\mathcal{A})} .$$

Preuve du Lemme 6.1. Voir Devroye et al. [7], Chapitre 13. ■

Soit $X = (X_1, \dots, X_n)$ un vecteur i.i.d. de loi P. Pour $A \in \mathcal{E}$, on note

$$P_n(A) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{X_i \in A\}} ,$$

la probabilité empirique de A. Pour une partie donnée $\mathcal{A}$ de $\mathcal{E}$, on s'intéresse dans cette section à la variable

$$\sup_{A \in \mathcal{A}} |P_n(A) - P(A)| ,$$

Proposition 6.2. *Soit $A \subset \mathcal{E}$ et $X = (X_1, \dots, X_n)$ un échantillon i.i.d. de loi P. Alors, pour tout $\varepsilon > 0$, on a*

$$\mathbf{P}\left(\sup_{A \in \mathcal{A}} |P_n(A) - P(A)| > \varepsilon\right) \leq 8\, \mathsf{s}(\mathcal{A}, n) e^{-\frac{n\varepsilon^2}{32}} .$$

Preuve de la Proposition 6.2. Remarquons déjà que l'on peut supposer que $n\varepsilon^2 \geq 2$ (sinon, le résultat est immédiat). Soit $X' = (X'_1, \dots, X'_n)$ une copie indépendante de X et notons $P'_n(A) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{X'_i \in A\}}$. Montrons que, pour $n\varepsilon^2 \geq 2$, on a

$$\mathbf{P}\left(\sup_{A \in \mathcal{A}} |P_n(A) - P(A)| > \varepsilon\right) \leq 2\mathbf{P}\left(\sup_{A \in \mathcal{A}} \left|P_n(A) - P'_n(A)\right| > \frac{\varepsilon}{2}\right) .$$

Soit $A^\star$ un élément de $\mathcal{A}$ tel que $|P_n(A) - P(A)| > \varepsilon$ s'il en existe un, ou bien un élément quelconque de $\mathcal{A}$ s'il n'en existe pas. On a

$$\begin{aligned}\mathbf{P}\left(\sup_{A\in\mathcal{A}}\left|P_n(A)-P'_n(A)\right|>\frac{\varepsilon}{2}\right) &\geq \mathbf{P}\left(\left|P_n(A^\star)-P'_n(A^\star)\right|>\frac{\varepsilon}{2}\right)\\ &\geq \mathbf{P}\left(|P_n(A^\star)-P(A^\star)|>\varepsilon\,,\,\left|P'_n(A^\star)-P(A^\star)\right|<\frac{\varepsilon}{2}\right)\\ &= \mathbf{E}\left[\mathbb{1}_{\{|P_n(A^\star)-P(A^\star)|>\varepsilon\}}\mathbf{P}\left(\left|P'_n(A^\star)-P(A^\star)\right|<\frac{\varepsilon}{2}\mid X\right)\right].\end{aligned}$$

Par l'inégalité de Chebyshev, on a

$$\mathbf{P}\left(\left|P'_n(A^\star)-P(A^\star)\right|\geq\frac{\varepsilon}{2}\mid X\right)\leq\frac{4P(A^\star)(1-P(A^\star))}{n\varepsilon^2}\leq\frac{1}{n\varepsilon^2}\leq\frac{1}{2}\,,$$

pour $n\varepsilon^2\geq 2$. Ainsi

$$\mathbf{P}\left(\sup_{A\in\mathcal{A}}\left|P_n(A)-P'_n(A)\right|>\frac{\varepsilon}{2}\right)\geq\frac{1}{2}\mathbf{P}\left(|P_n(A^\star)-P(A^\star)|>\varepsilon\right)=\frac{1}{2}\mathbf{P}\left(\sup_{A\in\mathcal{A}}|P_n(A)-P(A)|>\varepsilon\right).$$

Montrons maintenant que

$$\mathbf{P}\left(\sup_{A\in\mathcal{A}}\left|P_n(A)-P'_n(A)\right|>\frac{\varepsilon}{2}\right)\leq 4\,\mathsf{s}(\mathcal{A},n)e^{-\frac{n\varepsilon^2}{32}}$$

Soit $(\varepsilon_i)_{i=1}^n$ une suite indépendante de variables de Rademacher (uniformes sur $\{-1,1\}$), indépendante de (X,X'). Par symétrie, on a

$$\left(\mathbb{1}_{X_i\in A}-\mathbb{1}_{X'_i\in A}\right)_{i=1}^n\sim\left(\varepsilon_i\left(\mathbb{1}_{X_i\in A}-\mathbb{1}_{X'_i\in A}\right)\right)_{i=1}^n.$$

Ainsi

$$\begin{aligned}\mathbf{P}\left(\sup_{A\in\mathcal{A}}\left|P_n(A)-P'_n(A)\right|>\frac{\varepsilon}{2}\right) &= \mathbf{P}\left(\sup_{A\in\mathcal{A}}\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_i\left(\mathbb{1}_{X_i\in A}-\mathbb{1}_{X'_i\in A}\right)\right|>\frac{\varepsilon}{2}\right)\\ &\leq 2\mathbf{P}\left(\sup_{A\in\mathcal{A}}\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_i\mathbb{1}_{X_i\in A}\right|>\frac{\varepsilon}{4}\right).\end{aligned}$$

Remarquons maintenant que si A et A' sont deux éléments de $\mathcal{A}$ qui ont la même intersection avec $\{X_1,\ldots,X_n\}$, alors $\frac{1}{n}\sum_{i=1}^n\varepsilon_i\mathbb{1}_{X_i\in A}=\frac{1}{n}\sum_{i=1}^n\varepsilon_i\mathbb{1}_{X_i\in A'}$. On peut donc prendre le supremum uniquement sur $\mathsf{tr}(\mathcal{A},X)$:

$$\begin{aligned}\mathbf{P}\left(\sup_{A\in\mathcal{A}}\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_i\mathbb{1}_{X_i\in A}\right|>\frac{\varepsilon}{4}\right) &= \mathbf{P}\left(\sup_{A\in\mathsf{tr}(\mathcal{A},X)}\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_i\mathbb{1}_{X_i\in A}\right|>\frac{\varepsilon}{4}\right)\\ &\leq \mathbf{E}\left[\sum_{A\in\mathsf{tr}(\mathcal{A},X)}\mathbf{P}\left(\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_i\mathbb{1}_{X_i\in A}\right|\geq\frac{\varepsilon}{4}\mid X\right)\right].\end{aligned}$$

Or l'inégalité de Hoeffding donne

$$\mathbf{P}\left(\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_i\mathbb{1}_{X_i\in A}\right|>\frac{\varepsilon}{4}\mid X\right)\leq 2e^{-\frac{n\varepsilon^2}{32}}.$$

Ainsi

$$\mathbf{P}\left(\sup_{A\in\mathcal{A}}\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_i\mathbb{1}_{X_i\in A}\right|>\frac{\varepsilon}{4}\right)\leq 2\mathbf{E}\left[|\mathsf{tr}(\mathcal{A},X)|\right]e^{-\frac{n\varepsilon^2}{32}}\leq 2\,\mathsf{s}(\mathcal{A},n)e^{-\frac{n\varepsilon^2}{32}}.$$

■

Exemple 6.4. Soient $X_1, \dots, X_n$ des variables aléatoires réelles, i.i.d. de fonction de répartition F que l'on suppose continue. La fonction de répartition empirique est donnée par

$$\forall x \in \mathbb{R}\,,\ F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \le x\}}\,.$$

Pour quantifier de façon uniforme la distance entre F_n et F, on introduit

$$K_n = \sup_{x\in\mathbb{R}} |F_n(x) - F(x)|\ .$$

Notons que la loi de K_n ne dépend pas de F. En effet, le vecteur $(F(X_1), \dots, F(X_n))$ a la même loi qu'un vecteur $(U_1, \dots, U_n)$ i.i.d. de loi uniforme sur $[0,1]$. Ainsi

$$K_n \sim \sup_{x\in\mathbb{R}} \left| \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{U_i \le F(x)\}} - F(x) \right| = \sup_{u\in[0,1]} \left| \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{U_i \le u\}} - u \right|\ ,$$

où pour la deuxième égalité, on a utilisé la continuité de F. La loi de K_n s'appelle loi de Kolmogorov–Smirnov, et l'on peut montrer que $\sqrt{n}K_n$ converge en loi vers le supremum d'un pont brownien entre 0 et 1. En notant $\mathcal{A} = \{]-\infty, x],\ x \in \mathbb{R}\}$, on a toujours

$$|\,\mathsf{tr}(x_1, \dots, x_n)| \le n+1\,,$$

avec égalité quand les x_i sont tous distincts. Ainsi la Proposition 6.2 donne

$$\mathbf{P}\left(\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| > \varepsilon \right) \le 8(n+1)e^{-\frac{n\varepsilon^2}{32}}\ . \tag{6.2}$$

On en déduit en particulier le théorème de Glivenko-Cantelli :

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \xrightarrow{\text{p.s.}} 0\,.$$

La borne (6.2) est loin d'être optimale et Massart [17] a montré que

$$\mathbf{P}\left(\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| > \varepsilon \right) \le 2e^{-\frac{n\varepsilon^2}{2}}\ .$$

Exemple 6.5. Reprenons le cadre de l'apprentissage statistique, avec $\mathcal{D}_n = ((X_1, Y_1), \dots, (X_n, Y_n))$ un échantillon i.i.d. à valeurs dans $\mathcal{X} \times \{0,1\}$ et $\mathcal{F}$ un ensemble de classifieurs $f : \mathcal{X} \to \{0,1\}$. En associant un classifieur f à l'événement $A = \{f(X) \ne Y\}$, on définit

$$\mathcal{A} = \{\{(x,y) \in \mathcal{X} \times \{0,1\}\,,\ f(x) \ne y\}\,,\ f \in \mathcal{F}\}\ ,$$

En notant

$$\mathcal{A}' = \{\{x \in \mathcal{X}\,,\ f(x) = 1\}\,,\ f \in \mathcal{F}\}\ ,$$

on peut montrer que $\mathsf{s}(\mathcal{A}, n) = \mathsf{s}(\mathcal{A}', n)$, et donc que $\mathsf{V}(\mathcal{A}) = \mathsf{V}(\mathcal{A}')$ (voir Devroye et al. [7][Chapitre 13]). Dans le cadre de la classification, on écrira en fait $\mathsf{s}(\mathcal{F}, n)$ et $\mathsf{V}(\mathcal{F})$ pour désigner $\mathsf{s}(\mathcal{A}, n)$ et $\mathsf{V}(\mathcal{A})$. Avec ces identifications et en posant $Z_i = (X_i, Y_i) \sim P$, on a

$$\mathbf{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{Z_i \in A\}} = P_n(A) \qquad \text{et } \mathbf{R}(f) = P(A)\,.$$

La Proposition 6.2 donne alors

$$\mathbf{P}\left(\sup_{f\in\mathcal{F}} |\mathbf{R}_n(f) - \mathbf{R}(f)| > \varepsilon \right) \le 8\,\mathsf{s}(\mathcal{F}, n)e^{-\frac{n\varepsilon^2}{32}}\ .$$

En utilisant la borne (6.1), on obtient

$$\mathbf{P}\left(\mathbf{R}(f_n^*) - \mathbf{R}(f^*) > \varepsilon\right) \leq 8\,\mathsf{s}(\mathcal{F}, n) e^{-\frac{n\varepsilon^2}{128}}\,.$$

Autrement dit, pour tout $\delta \in]0,1[$, on a, avec probabilité au moins $1-\delta$,

$$\begin{aligned}\mathbf{R}(f_n^*) - \mathbf{R}(f^*) &\leq \sqrt{\frac{128}{n}\log\left(\frac{8\,\mathsf{s}(\mathcal{F},n)}{\delta}\right)} \\ &\leq \sqrt{\frac{128}{n}\left(\log\left(\frac{8}{\delta}\right) + \mathsf{V}(\mathcal{F})\log(n+1)\right)}\,, \qquad (6.3)\end{aligned}$$

où la deuxième inégalité vient du Lemme 6.1. Dans la section suivante, nous allons voir que l'on peut remplacer $\log(n+1)$ par un terme d'ordre constant.

3. Chaînage et inégalité de Dudley

Soit $X = (X_1, \ldots, X_n)$ i.i.d. de loi P, et $\varepsilon_1, \ldots, \varepsilon_n$ des variables i.i.d. de loi de Rademacher, indépendantes de X. Pour $A \in \mathcal{A}$, on définit

$$P_n^\varepsilon(A) = \frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathbb{1}_{\{X_i \in A\}}\,,$$

la version « rademacherisée » de $P_n(A)$. Dans le cadre de la classification, on écrira $\mathbf{R}_n^\varepsilon(f) = \frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathbb{1}_{\{f(X_i)\neq Y_i\}}$.

Proposition 6.3. *On a*

$$\mathbf{E}\left[\sup_{A\in\mathcal{A}} |P_n(A) - P(A)|\right] \leq 2\mathbf{E}\left[\sup_{A\in\mathcal{A}} |P_n^\varepsilon(A)|\right]$$

Preuve de la Proposition 6.3. Soit $X' = (X_1', \ldots, X_n')$ une copie indépendante de X. En écrivant

$$P(A) = \mathbf{E}\left[\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{X_i'\in A\}} \mid X\right]\,,$$

et en utilisant la convexité de la valeur absolue et du supremum, on a

$$\mathbf{E}\left[\sup_{A\in\mathcal{A}} |P_n(A) - P(A)|\right] \leq \mathbf{E}\left[\sup_{A\in\mathcal{A}} \left|P_n(A) - P_n'(A)\right|\right]\,.$$

Maintenant, si $(\varepsilon_i)_{i=1}^n$ est une suite de Rademacher indépendantes de (X, X'), on a

$$\begin{aligned}\mathbf{E}\left[\sup_{A\in\mathcal{A}} \left|P_n(A) - P_n'(A)\right|\right] &= \mathbf{E}\left[\sup_{A\in\mathcal{A}} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\left(\mathbb{1}_{\{X_i\in A\}} - \mathbb{1}_{\{X_i'\in A\}}\right)\right|\right] \\ &\leq 2\mathbf{E}\left[\sup_{A\in\mathcal{A}} |P_n^\varepsilon(A)|\right]\,.\end{aligned}$$

■

La Proposition 6.3, combinée avec l'inégalité de McDiarmid, donne la borne suivante.

Proposition 6.4. *Pour tout $\delta \in]0,1[$, avec probabilité au moins $1-\delta$, on a*

$$\sup_{A\in\mathcal{A}} |P_n(A) - P(A)| \leq 2\mathbf{E}\left[\sup_{A\in\mathcal{A}} |P_n^\varepsilon(A)|\right] + \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2n}}\,.$$

Preuve de la Proposition 6.4. Par la Proposition 6.3, on a

$$\sup_{A\in\mathcal{A}}|P_n(A)-P(A)| \le 2\mathbf{E}\left[\sup_{A\in\mathcal{A}}|P_n^\varepsilon(A)|\right] + \sup_{A\in\mathcal{A}}|P_n(A)-P(A)| - \mathbf{E}\left[\sup_{A\in\mathcal{A}}|P_n(A)-P(A)|\right].$$

On remarque maintenant que la variable $\sup_{A\in\mathcal{A}}|P_n(A)-P(A)|$ vérifie l'hypothèse de l'inégalité de McDiarmid avec $c_i(x^{(i)}) \le \frac{1}{n}$. Ainsi pour tout $\varepsilon>0$, on a

$$\mathbf{P}\left(\sup_{A\in\mathcal{A}}|P_n(A)-P(A)| - \mathbf{E}\left[\sup_{A\in\mathcal{A}}|P_n(A)-P(A)|\right] > \varepsilon\right) \le e^{-2\varepsilon^2 n},$$

ce qui donne bien le résultat voulu. ■

Dans la suite de cette section, on cherche à majorer $\mathbf{E}\left[\sup_{A\in\mathcal{A}}|P_n^\varepsilon(A)|\right]$. Voyons déjà comment obtenir une borne similaire à (6.3). On a

$$\mathbf{E}\left[\sup_{A\in\mathcal{A}}|P_n^\varepsilon(A)|\right] = \mathbf{E}\left[\sup_{A\in\mathcal{A}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathbb{1}_{X_i\in A}\right|\right] = \mathbf{E}\left[\sup_{A\in\mathsf{tr}(\mathcal{A},X)}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathbb{1}_{X_i\in A}\right|\right].$$

Conditionnellement à $X=(X_1,\dots,X_n)$, les variables $\frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathbb{1}_{X_i\in A}$ et $-\frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathbb{1}_{X_i\in A}$ sont sous-gaussiennes avec facteur variance $\frac{1}{n}$. Or, si $Y_1,\dots,Y_m$ sont sous-gaussiennes avec facteur variance v, on a

$$\mathbf{E}\left[\max_{1\le j\le m} Y_j\right] \le \sqrt{2v\log(m)}, \tag{6.4}$$

ce qui donne (pour $m = 2|\mathsf{tr}(\mathcal{A},X)|$),

$$\mathbf{E}\left[\sup_{A\in\mathsf{tr}(\mathcal{A},X)}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathbb{1}_{X_i\in A}\right| \;\middle|\; X\right] \le \sqrt{\frac{2\log\left(2|\mathsf{tr}(\mathcal{A},X)|\right)}{n}}.$$

En utilisant le Lemme 6.1, on obtient

$$\mathbf{E}\left[\sup_{A\in\mathcal{A}}|P_n^\varepsilon(A)|\right] \le \sqrt{\frac{2\log\left(2\,\mathsf{s}(\mathcal{A},n)\right)}{n}} \le \sqrt{\frac{2}{n}\left(\log(2)+\mathsf{V}(\mathcal{A})\log(n+1)\right)}.$$

Par une méthode dite de chaînage, on peut en fait se passer du terme logarithmique. Pour cela, commençons par fixer un vecteur $x=(x_1,\dots,x_n)\in\mathcal{X}^n$ fixé, et, pour $a\in\mathcal{A}$, notons

$$Y_a = \frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathbb{1}_{\{x_i\in a\}}.$$

On introduit une pseudo-distance sur $\mathcal{A}$ donnée par

$$\forall a,b\in\mathcal{A},\ d_x(a,b) = \sqrt{\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{x_i\in a\Delta b\}}},$$

où $a\Delta b = a\cup b - a\cap b$ est la différence symétrique entre a et b. Cela fait de $\mathcal{A}$ un espace pseudo-métrique totalement borné. Pour $\delta>0$, un δ-net pour d_x est un ensemble fini $\mathcal{A}_\delta\subset\mathcal{A}$ de cardinal maximal et tel que pour tous $a,b\in\mathcal{A}_\delta$ avec $a\neq b$, on a $d_x(a,b)>\delta$. On note

$$H_x(\delta,\mathcal{A}) = \log|\mathcal{A}_\delta|.$$

La fonction $\delta\mapsto H_x(\delta,\mathcal{A})$ s'appelle l'entropie métrique de $\mathcal{A}$ pour la pseudo-distance d_x. On définit l'entropie métrique universelle de $\mathcal{A}$ par

$$H(\delta,\mathcal{A}) = \sup_Q H_Q(\delta,\mathcal{A}),$$

où le supremum est pris sur toutes les lois de probabilité concentrées sur un sous-ensemble fini de $\mathcal{X}$, et où $H_Q(\delta, \mathcal{A})$ correspond à l'entropie métrique pour la pseudo-distance

$$d_Q(a,b) = \sqrt{Q(a \Delta b)}\,.$$

(En fait $d_x(a,b) = d_Q(a,b)$ pour Q la mesure empirique associée au vecteur x.) Pour simplifier, on suppose ici que $\mathcal{A}$ est fini mais le résultat suivant se généralise au cas infini.

Proposition 6.5 (Inégalité de Dudley)**.** *Supposons que $\mathcal{A}$ est fini. On a*

$$\mathbf{E}\left[\sup_{a\in\mathcal{A}} Y_a\right] \leq \frac{12}{\sqrt{n}} \int_0^{1/2} \sqrt{H(u,\mathcal{A})}du\,.$$

Preuve de la Proposition 6.5. Remarquons déjà que pour tous $a, b \in \mathcal{A}$ et tout $\lambda > 0$, on a, par l'inégalité de Hoeffding,

$$\begin{aligned}
\log \mathbf{E} e^{\lambda(Y_a - Y_b)} &= \sum_{i=1}^n \log \mathbf{E} e^{\frac{\lambda}{n}\left(\mathbb{1}_{\{x_i\in a\}} - \mathbb{1}_{\{x_i\in b\}}\right)\varepsilon_i} \\
&\leq \sum_{i=1}^n \frac{\lambda^2}{2n^2}\left(\mathbb{1}_{\{x_i\in a\}} - \mathbb{1}_{\{x_i\in b\}}\right)^2 \\
&= \frac{\lambda^2}{2n} d_x^2(a,b)\,.
\end{aligned}$$

Ainsi, $Y_a - Y_b$ est sous-gaussienne avec facteur de variance $\frac{d_x^2(a,b)}{n}$. Pour $j \in \mathbb{N}$, posons $\delta_j = 2^{-j}$ et considérons un δ_j-net $\mathcal{A}_j$ de $\mathcal{A}$ pour d_x. Pour tout $j \in \mathbb{N}$, on peut alors trouver une application $\pi_j : \mathcal{A} \to \mathcal{A}_j$ telle que

$$\forall a \in \mathcal{A}\,,\ d_x(a, \pi_j(a)) \leq \delta_j\,.$$

Comme $\mathcal{A}$ est fini, il existe un entier $J \in \mathbb{N}$ tel que, pour tout $a \in \mathcal{A}$,

$$Y_a = Y_{\pi_0(a)} + \sum_{j=0}^J \left(Y_{\pi_{j+1}(a)} - Y_{\pi_j(a)}\right)\,.$$

D'autre part, comme on a toujours $d_x(a,b) \leq 1$, on peut prendre $\mathcal{A}_0 = \{a_0\}$ pour un élément a_0 quelconque, auquel cas $\pi_0(a) = a_0$ pour tout $a \in \mathcal{A}$. Comme $\mathbf{E}[Y_{a_0}] = 0$, on obtient

$$\mathbf{E}\left[\sup_{a\in\mathcal{A}} Y_a\right] \leq \sum_{j=0}^J \mathbf{E}\left[\sup_{a\in\mathcal{A}} \left(Y_{\pi_{j+1}(a)} - Y_{\pi_j(a)}\right)\right]\,.$$

Maintenant, pour tout $j \in \mathbb{N}$, on a

$$\left|\left\{\left(Y_{\pi_j(a)}, Y_{\pi_{j+1}(a)}\right)\,,\ a \in \mathcal{A}\right\}\right| \leq |\mathcal{A}_j|.|\mathcal{A}_{j+1}| \leq |\mathcal{A}_{j+1}|^2\,,$$

et, par l'inégalité triangulaire,

$$d\left(Y_{\pi_j(a)}, Y_{\pi_{j+1}(a)}\right) \leq \delta_j + \delta_{j+1} = 3\delta_{j+1}\,.$$

Ainsi, en utilisant l'inégalité (6.4), on a

$$\mathbf{E}\left[\sup_{a\in\mathcal{A}} \left(Y_{\pi_{j+1}(a)} - Y_{\pi_j(a)}\right)\right] \leq \sqrt{\frac{18\delta_{j+1}^2}{n} \log\left(|\mathcal{A}_{j+1}|^2\right)} = \frac{6\delta_{j+1}}{\sqrt{n}} \sqrt{H(\delta_{j+1}, \mathcal{A})}\,,$$

où $H(\delta_{j+1}, \mathcal{A})$ correspond à l'entropie universelle. En sommant sur j, et en utilisant la décroissance de $\delta \mapsto H(\delta, \mathcal{A})$ et le fait que $\delta_j = 2(\delta_j - \delta_{j+1})$, on obtient

$$\mathbf{E}\left[\sup_{a\in\mathcal{A}} Y_a\right] \leq \frac{12}{\sqrt{n}} \sum_{j=1}^{J+1} (\delta_j - \delta_{j+1}) \sqrt{H(\delta_j, \mathcal{A})} \leq \frac{12}{\sqrt{n}} \int_0^{1/2} \sqrt{H(u, \mathcal{A})} du \,.$$

■

On obtient donc

$$\mathbf{E}\left[\sup_{A\in\mathcal{A}} |P_n^{\varepsilon}(A)|\right] \leq \frac{24}{\sqrt{n}} \int_0^{1/2} \sqrt{H(u, \mathcal{A})} du$$

Or on peut montrer que

$$H(u, \mathcal{A}) \leq 2\, \mathsf{V}(\mathcal{A}) \log(e^2/u)\,,$$

voir Haussler [12]. Ainsi, en utilisant l'inégalité de Jensen avec $x \mapsto \sqrt{x}$, on a

$$\mathbf{E}\left[\sup_{A\in\mathcal{A}} |P_n^{\varepsilon}(A)|\right] \leq 12 \sqrt{\frac{2\,\mathsf{V}(\mathcal{A})}{n}} \sqrt{2 \int_0^{1/2} \log(e^2/u) du} = 24 \sqrt{\frac{(3 + \log 2)\,\mathsf{V}(\mathcal{A})}{2n}}\,.$$

Finalement, en revenant au résultat de la Proposition 6.4 et en utilisant la borne (6.1), on a, avec probabilité au moins $1 - \delta$,

$$\mathbf{R}(f_n^*) - \mathbf{R}(f^*) \leq 96 \sqrt{\frac{(3 + \log 2)\,\mathsf{V}(\mathcal{F})}{2n}} + \sqrt{\frac{2 \log\left(\frac{1}{\delta}\right)}{n}}\,.$$

Chapitre 7

Concentration de matrices

Dans ce chapitre, nous allons voir comment majorer $\mathbf{P}(\|\mathbf{Z}\| \geq t)$, où $\mathbf{Z}$ est une matrice symétrique réelle et où $\|\cdot\|$ correspond à la norme spectrale (dite aussi norme d'opérateur ℓ_2).

Rappelons d'abord quelques propriétés importantes de $\mathcal{S}_n$ l'ensemble des matrices symétriques de $\mathcal{M}_n(\mathbb{R})$. Tout d'abord, si $A \in \mathcal{S}_n$, alors A est diagonalisable dans une base orthogonale de vecteurs propres. On notera

$$\lambda_1(A) \geq \cdots \geq \lambda_n(A)$$

les n valeurs propres (réelles) de A, rangées par ordre décroissant. La norme spectrale de A est alors donnée par

$$\|A\| = \max\{\lambda_1(A), -\lambda_n(A)\}\,.$$

Une matrice $A \in \mathcal{S}_n$ est dite semi-définie positive si, pour tout $u \in \mathbb{R}^n$, on a ${}^tuAu \geq 0$. De façon équivalente, une matrice de $\mathcal{S}_n$ est semi-définie positive si toutes ses valeurs propres sont positives ou nulles. De façon similaire, on dit qu'une matrice $A \in \mathcal{S}_n$ est dite définie positive si, pour tout $u \in \mathbb{R}^n \setminus \{0\}$, on a ${}^tuAu > 0$, ce qui équivaut à dire que toutes les valeurs propres de A sont strictement positives. On définit l'ordre partiel $\preccurlyeq$ sur $\mathcal{S}_n$ par $A \preccurlyeq B$ ssi $B - A$ est semi-définie positive. De même, on notera $A \prec B$ ssi $B - A$ est définie positive. Une propriété importante de l'ordre partiel $\preccurlyeq$ est la stabilité par conjugaison : si $A \preccurlyeq B$, et si C est une matrice à n lignes, alors ${}^tCAC \preccurlyeq {}^tCBC$.

Remarquons aussi que si $A \succcurlyeq 0$, alors $\lambda_1(A) \leq \mathsf{tr}(A)$. De plus, si $A, B \in \mathcal{S}_n$ avec $A \preccurlyeq B$, alors

$$\forall i \llbracket 1, n \rrbracket\,,\ \lambda_i(A) \leq \lambda_i(B)\,.$$

Ce dernier résultat est appelé le principe de monotonicité de Weyl et découle directement du Théorème de Courant -Fisher :

$$\lambda_i(A) = \max_{E,\,\dim(E)=i} \min_{u \in E} \frac{{}^tuAu}{{}^tuu} \leq \max_{E,\,\dim(E)=i} \min_{u \in E} \frac{{}^tuBu}{{}^tuu} = \lambda_i(B)\,. \tag{7.1}$$

Une façon naturelle d'étendre une fonction sur $\mathbb{R}$ à une fonction sur $\mathcal{S}_n$ est de l'appliquer aux valeurs propres.

Définition 7.1. Soit $f : I \to \mathbb{R}$ où I est un intervalle de $\mathbb{R}$, et soit $A \in \mathcal{S}_n$ une matrice symétrique dont toutes les valeurs propres λ_i appartiennent à I. Si

$$A = Q \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} Q^{-1}\,,$$

alors on définit la matrice $f(A)$ par

$$f(A) = Q \begin{pmatrix} f(\lambda_1) & & \\ & \ddots & \\ & & f(\lambda_n) \end{pmatrix} Q^{-1} .$$

(On peut vérifier que la définition de $f(A)$ ne dépend pas de la décomposition spectrale $A = Q\Lambda Q^{-1}$ choisie.)

On peut ainsi définir l'exponentielle e^A. De façon équivalente, e^A est donnée par

$$e^A = I + \sum_{q=1}^{+\infty} \frac{A^q}{q!} .$$

Le logarithme $\log A$ peut être défini par la définition 7.1, ou bien de façon équivalente comme l'inverse de l'exponentielle : pour tout $A \in \mathcal{S}_n$, $\log e^A = A$.

Une conséquence de l'inégalité (7.1) est que si f est croissante sur I, alors

$$A \preccurlyeq B \quad \Rightarrow \quad \mathsf{tr}\, f(A) \leq \mathsf{tr}\, f(B) . \tag{7.2}$$

Une fonction f est dite opérateur-monotone si elle vérifie la propriété plus forte :

$$A \preccurlyeq B \quad \Rightarrow \quad f(A) \preccurlyeq f(B) . \tag{7.3}$$

Proposition 7.1. *Le logarithme est opérateur-monotone.*

Insistons sur le fait que l'exponentielle n'est pas opérateur-monotone.

Preuve de la Proposition 7.1. Remarquons déjà que pour tout $a > 0$, on a

$$\log a = \int_0^{+\infty} \left(\frac{1}{1+u} - \frac{1}{a+u} \right) \mathrm{d}u .$$

En effet, pour tout $x > 0$,

$$\begin{aligned} \int_0^x \left(\frac{1}{1+u} - \frac{1}{a+u} \right) \mathrm{d}u &= \left[\log(1+u) - \log(a+u) \right]_0^x \\ &= \log a + \log \left(\frac{1+x}{a+x} \right) \\ &\underset{t \to +\infty}{\longrightarrow} \log a . \end{aligned}$$

Soit $A \in \mathcal{S}_n$ définie positive. En appliquant cette identité à toutes les valeurs propres de A, on a

$$\log A = \int_0^{+\infty} \left((1+u)^{-1} I - (A+uI)^{-1} \right) \mathrm{d}u .$$

Montrons maintenant que si $0 \prec A \preccurlyeq B$, et $u \geq 0$, alors $-(A+uI)^{-1} \preccurlyeq -(B+uI)^{-1}$. Notons $A_u = A + uI$ et $B_u = B + uI$. On a $A_u \preccurlyeq B_u$. Par la stabilité par conjugaison, on obtient

$$0 \prec B_u^{-1/2} A_u B_u^{-1/2} \preccurlyeq I .$$

Or, lorsqu'une matrice définie positive a toutes ses valeurs propres inférieures à 1, son inverse a toutes ses valeurs propres supérieures à 1. Ainsi,

$$\left(B_u^{-1/2} A_u B_u^{-1/2} \right)^{-1} = B_u^{1/2} A_u^{-1} B_u^{1/2} \succcurlyeq I .$$

En appliquant à nouveau la stabilité par conjugaison, on a $B_u^{-1} \preccurlyeq A_u^{-1}$, soit $-A_u^{-1} \preccurlyeq -B_u^{-1}$. En appliquant cette inégalité dans la représentation intégrale du logarithme, on obtient $\log A \preccurlyeq \log B$. ■

Énonçons enfin un dernier résultat, le théorème de Lieb. Pour la preuve, nous renvoyons au chapitre 8 de Tropp [23].

Proposition 7.2. *Soit H une matrice symétrique. L'application*

$$A \mapsto \mathsf{tr} \exp\left(H + \log A\right)$$

est concave sur l'ensemble des matrices symétriques définies positives.

Nous sommes maintenant en mesure de montrer un équivalent de l'inégalité de Bernstein pour les sommes de matrices symétriques indépendantes.

1. Une inégalité de Bernstein pour les sommes de matrices

Proposition 7.3. *Soient $X_1, \dots, X_N$ des matrices $n \times n$ symétriques indépendantes telles que $\mathbf{E}X_i = 0$ et $\|X_i\| \leq K$. Alors pour tout $t \geq 0$,*

$$\mathbf{P}\left(\left\|\sum_{i=1}^N X_i\right\| \geq t\right) \leq 2n \exp\left\{-\frac{t^2}{2\left(\sigma^2 + \frac{Kt}{3}\right)}\right\},$$

où $\sigma^2 = \left\|\sum_{i=1}^N \mathbf{E}X_i^2\right\|$.

Preuve de la Proposition 7.3. Notons $S = \sum_{i=1}^N X_i$ et $\lambda_1(S) \geq \cdots \geq \lambda_n(S)$ les valeurs propres de S rangées par ordre décroissant. On a alors $\|S\| = \max\{\lambda_1(S), -\lambda_n(S)\}$. Comme $-\lambda_n(S) = \lambda_1(-S)$, il suffit de montrer que

$$\mathbf{P}\left(\lambda_1(S) \geq t\right) \leq n \exp\left\{-\frac{t^2}{2\left(\sigma^2 + \frac{Kt}{3}\right)}\right\}.$$

On a, pour tout $u \geq 0$,

$$\mathbf{P}\left(\lambda_1(S) \geq t\right) \leq e^{-ut}\mathbf{E}e^{u\lambda_1(S)} = e^{-ut}\mathbf{E}\lambda_1(e^{uS}),$$

où la dernière inégalité vient du fait que $e^{u\lambda_1(S)} = \lambda_1(e^{uS})$. Comme e^{uS} est définie positive, toutes ses valeurs propres sont positives et l'on a $\lambda_1(e^{uS}) \leq \mathsf{tr}(e^{uS})$. À ce stade, on aimerait pouvoir dire $e^{uS} = \prod_{i=1}^N e^{uX_i}$, mais cette identité n'est pas vraie : une exponentielle de matrices ne transforme pas une somme en un produit. En revanche, on a l'inégalité suivante :

$$\mathbf{E}\,\mathsf{tr}(e^{uS}) \leq \mathsf{tr} \exp\left\{\sum_{i=1}^N \log \mathbf{E}e^{uX_i}\right\}. \tag{7.4}$$

Cette inégalité découle du Théorème de Lieb 7.2 appliqué avec $H = \sum_{i=1}^{N-1} uX_i$ et $A = e^{uX_N}$ et de l'inégalité de Jensen appliquée conditionnellement à $X_1, \dots, X_{N-1}$:

$$\mathbf{E}\left[\mathsf{tr}(e^{uS}) \;\middle|\; X_1, \dots, X_{N-1}\right] \leq \mathsf{tr} \exp\left\{u\sum_{i=1}^{N-1} X_i + \log \mathbf{E}e^{uX_N}\right\}.$$

En prenant l'espérance conditionnelle sachant $X_1, \dots, X_{N-2}$, et en répétant le même argument, et ainsi de suite, on obtient bien l'inégalité (7.4). On a ainsi réussi à passer de la fonction génératrice

des moments de S à celle des matrices X_i. Montrons maintenant que si X est une matrice $n \times n$ symétrique avec $\mathbf{E}X = 0$ et $\|X\| \leq K$, alors pour tout $0 \leq u < 3/K$,

$$\log \mathbf{E}e^{uX} \preccurlyeq \frac{u^2 \mathbf{E}X^2}{2\left(1 - \frac{uK}{3}\right)}.$$

Tout d'abord, pour $0 \leq u < 3/K$ et pour x tel que $|x| \leq K$, on a

$$e^{ux} = 1 + ux + \sum_{k \geq 2} \frac{(ux)^k}{k!} \leq 1 + ux + \frac{u^2x^2}{2} \sum_{k \geq 2} \left(\frac{uK}{3}\right)^{k-2} \leq 1 + ux + \frac{u^2x^2}{2\left(1 - \frac{uK}{3}\right)}.$$

Comme toutes les valeurs propres de X sont contenues dans $[-K, K]$, cela implique l'inégalité matricielle

$$e^{uX} \preccurlyeq I + uX + \frac{u^2X^2}{2\left(1 - \frac{uK}{3}\right)}, \tag{7.5}$$

En prenant l'espérance dans (7.5), on a

$$\mathbf{E}e^{uX} \preccurlyeq I + \frac{u^2 \mathbf{E}X^2}{2\left(1 - \frac{uK}{3}\right)}.$$

En utilisant le fait que le logarithme est opérateur-monotone, puis l'inégalité $\log(1+z) \leq z$ pour tout $z \geq 0$ (appliquée à la matrice semi-définie positive $\frac{u^2\mathbf{E}X^2}{2\left(1-\frac{uK}{3}\right)}$), on obtient bien

$$\log \mathbf{E}e^{uX} \preccurlyeq \log\left(I + \frac{u^2 \mathbf{E}X^2}{2\left(1 - \frac{uK}{3}\right)}\right) \preccurlyeq \frac{u^2 \mathbf{E}X^2}{2\left(1 - \frac{uK}{3}\right)}.$$

En revenant à (7.4), on a donc

$$\mathbf{E}\,\mathsf{tr}(e^{uS}) \leq \mathsf{tr} \exp\left\{\frac{u^2 \sum_{i=1}^N \mathbf{E}X_i^2}{2\left(1 - \frac{uK}{3}\right)}\right\} \leq n \exp\left\{\frac{\sigma^2 u^2}{2\left(1 - \frac{uK}{3}\right)}\right\},$$

et

$$\mathbf{P}\left(\lambda_1(S) \geq t\right) \leq ne^{-ut} \exp\left\{\frac{\sigma^2 u^2}{2\left(1 - \frac{uK}{3}\right)}\right\} = n \exp\left\{-ut + \frac{\sigma^2 u^2}{2\left(1 - \frac{uK}{3}\right)}\right\}.$$

En optimisant sur $0 \leq u < 3/K$, on voit que le membre de droit est minimal pour $u = \frac{t}{\sigma^2 + Kt/3}$, ce qui donne

$$\mathbf{P}\left(\lambda_1(S) \geq t\right) \leq n \exp\left\{-\frac{t^2}{2\left(\sigma^2 + \frac{Kt}{3}\right)}\right\}.$$

■

2. Application : connexité du graphe d'Erdös-Renyi

Soit $G \sim \mathcal{G}(n,p)$ un graphe aléatoire d'Erdös-Renyi sur n sommets et soit $\mathbf{A}$ sa matrice d'adjacence, qui peut s'écrire

$$\mathbf{A} = \sum_{1 \leq i < j \leq n} \xi_{ij}(E_{ij} + E_{ji}),$$

où $(\xi_{ij})_{i<j}$ est une suite i.i.d. de loi de Bernoulli $\mathcal{B}(p)$, et où $(E_{ij})_{i,j}$ est la base canonique de $\mathcal{M}_n(\mathbb{R})$. Le Laplacien de G est la matrice $\mathbf{\Delta} = \mathbf{D} - \mathbf{A}$, où $\mathbf{D}$ est la matrice diagonale des degrés ($\mathbf{D}_{i,i} = \deg(i) = \sum_j \mathbf{A}_{i,j}$). Cette matrice peut s'écrire

$$\mathbf{\Delta} = \sum_{1\leq i<j\leq n} \xi_{ij}(E_{ii} + E_{jj} - E_{ij} - E_{ji})\,.$$

La matrice $\mathbf{\Delta}$ est semi-définie positive et le vecteur $\mathbf{1}$ (toutes les coordonnées égales à 1) est vecteur propre pour la valeur propre 0. Le spectre du Laplacien est intimement relié aux propriétés géométriques de G. En particulier, le graphe G est connexe si et seulement si la deuxième plus petite valeur de $\mathbf{\Delta}$ est strictement positive.

Pour simplifier le problème, nous allons d'abord former une matrice $\mathbf{Z}$ dont la plus petite valeur propre correspond à la deuxième plus petite valeur propre de $\mathbf{\Delta}$. Pour cela, considérons la matrice $\mathbf{R} \in \mathcal{M}_{n-1,n}(\mathbb{R})$ d'une isométrie partielle de noyau Vect($\mathbf{1}$), i.e.

$$\mathbf{R}\,{}^t\mathbf{R} = I_{n-1} \qquad \text{et} \qquad \mathbf{R}\mathbf{1} = 0\,.$$

On définit alors la matrice $\mathbf{Z} \in \mathcal{M}_{n-1}(\mathbb{R})$ par

$$\mathbf{Z} = \mathbf{R}\mathbf{\Delta}\,{}^t\mathbf{R} = \sum_{1\leq i<j\leq n} \xi_{ij}\mathbf{R}(E_{ii} + E_{jj} - E_{ij} - E_{ji})\,{}^t\mathbf{R}\,.$$

L'espérance de $\mathbf{Z}$ se calcule facilement :

$$\begin{aligned}\mathbf{E}\mathbf{Z} &= p\mathbf{R}\left(\sum_{1\leq i<j\leq n}(E_{ii} + E_{jj} - E_{ij} - E_{ji})\right){}^t\mathbf{R}\\ &= p\mathbf{R}\left((n-1)I_n - (\mathbf{1}\,{}^t\mathbf{1} - I_n)\right){}^t\mathbf{R}\\ &= pnI_{n-1}\,.\end{aligned}$$

En particulier, $\lambda_{\min}(\mathbf{EZ}) = pn$. Pour $1 \leq i < j \leq n$, notons $X_{ij} = \xi_{ij}\mathbf{R}(E_{ii} + E_{jj} - E_{ij} - E_{ji})\,{}^t\mathbf{R}$, de telle sorte que $\mathbf{Z} = \sum_{i<j} X_{ij}$. Les matrices X_{ij} sont symétriques indépendantes, et, par stabilité par conjugaison, elles restent semi-définies positives. De plus, le théorème de Courant-Fisher implique que la plus petite valeur propre de $\mathbf{Z}$, notée $\lambda_{\min}(\mathbf{Z})$, correspond à la deuxième plus petite valeur propre de $\mathbf{\Delta}$. Remarquons aussi que la norme spectrale de chaque matrice X_{ij} est inférieure à 2. En effet

$$\|X_{ij}\| \leq |\xi_{ij}|\|\mathbf{R}\|\|E_{ii} + E_{jj} - E_{ij} - E_{ji}\|\|\,{}^t\mathbf{R}\|\,.$$

Or $|\xi_{ij}| \leq 1$ (car ξ_{ij} vaut 0 ou 1), $\|\mathbf{R}\| = \|\,{}^t\mathbf{R}\| = 1$ (car $\mathbf{R}$ est une isométrie partielle), et on peut facilement voir que la norme de $E_{ii} + E_{jj} - E_{ij} - E_{ji}$ est inférieure à 2. Soit maintenant $t > 0$. Pour tout $u > 0$, on a

$$\begin{aligned}\mathbf{P}\left(\lambda_{\min}(\mathbf{Z}) \leq t\right) &= \mathbf{P}\left(e^{-u\lambda_{\min}(\mathbf{Z})} \geq e^{-ut}\right)\\ &\leq e^{ut}\mathbf{E}\left[e^{-u\lambda_{\min}(\mathbf{Z})}\right]\\ &= e^{ut}\mathbf{E}\left[e^{\lambda_{\max}(-u\mathbf{Z})}\right]\\ &= e^{ut}\mathbf{E}\left[\lambda_{\max}\left(e^{-u\mathbf{Z}}\right)\right]\\ &\leq e^{ut}\mathbf{E}\left[\mathsf{tr}\left(e^{-u\mathbf{Z}}\right)\right]\,,\end{aligned}$$

où l'on a utilisé la croissance de l'exponentielle et le fait que $e^{-u\mathbf{Z}}$ est définie positive. En utilisant à plusieurs reprises le théorème de Lieb et l'inégalité de Jensen conditionnelle comme dans la preuve de la Proposition 7.3, on obtient

$$\mathbf{E}\left[\mathsf{tr}\left(e^{-u\mathbf{Z}}\right)\right] \leq \mathsf{tr}\exp\left\{\sum_{1\leq i<j\leq n}\log\mathbf{E}\left[e^{-uX_{ij}}\right]\right\}.$$

Par convexité de $x\mapsto e^{-ux}$, on a, pour tout $x\in[0,2]$,

$$e^{-ux}\leq 1+\frac{e^{-2u}-1}{2}x\,.$$

Comme les valeurs propres de X_{ij} sont comprises dans $[0,2]$, on obtient l'inégalité matricielle :

$$e^{-uX_{ij}}\preccurlyeq I+\frac{e^{-2u}-1}{2}X_{ij}\,.$$

En prenant l'espérance, en utilisant que le logarithme est opérateur-monotone puis en appliquant matriciellement l'inégalité $\log(1+z)\leq z$, on obtient

$$\log\mathbf{E}\left[e^{-uX_{ij}}\right]\preccurlyeq\log\left(I+\frac{e^{-2u}-1}{2}\mathbf{E}[X_{ij}]\right)\preccurlyeq\frac{e^{-2u}-1}{2}\mathbf{E}[X_{ij}]\,.$$

Et en sommant sur $i<j$ et en prenant la trace de l'exponentielle, on obtient

$$\mathbf{E}\left[\mathsf{tr}\left(e^{-u\mathbf{Z}}\right)\right]\leq\mathsf{tr}\exp\left\{\frac{e^{-2u}-1}{2}\mathbf{E}[\mathbf{Z}]\right\}=(n-1)\exp\left(\frac{e^{-2u}-1}{2}np\right).$$

Ainsi,

$$\mathbf{P}\left(\lambda_{\min}(\mathbf{Z})\leq t\right)\leq n\exp\left(\inf_{u>0}\left\{ut+\frac{e^{-2u}-1}{2}np\right\}\right).$$

L'infimum est atteint pour $u=\frac{1}{2}\log\left(\frac{np}{t}\right)$. Pour $t=\varepsilon np$ avec $\varepsilon>0$, on obtient

$$\mathbf{P}\left(\lambda_{\min}(\mathbf{Z})\leq\varepsilon np\right)\leq ne^{-\frac{np}{2}(1-\varepsilon+\varepsilon\log(\varepsilon))}\,.$$

On voit que pour $p>\frac{2(1+\delta)\log n}{n}$ avec $\delta>0$, en prenant ε assez petit, la probabilité de G ne soit pas connexe tend vers 0. Cette borne est presque optimale, à un facteur 2 : on peut montrer que si $p>\frac{(1+\delta)\log n}{n}$, alors G est connexe avec grande probabilité, et inversement, que si $p<\frac{(1-\delta)\log n}{n}$, alors G contient des sommets isolées avec grande probabilité.

Chapitre 8

Concentration sans indépendance

1. Concentration pour les chaînes de Markov

Soit $(X_t)_{t\in\mathbb{N}}$ une chaîne de Markov sur un espace d'état Ω fini, de matrice de transition P. On sait que si la chaîne est ergodique au sens où

$$\exists t \in \mathbb{N}\,,\ \forall x, y \in \Omega\,,\ P^t(x,y) > 0\,,$$

alors il existe une unique probabilité stationnaire π et la chaîne converge vers π en variation totale :

$$\max_{x\in\Omega} d_{\mathrm{TV}}\left(P^t(x,\cdot),\pi\right) \underset{t\to+\infty}{\longrightarrow} 0\,.$$

Notons $\mathcal{D}(t) = \max_{x\in\Omega} d_{\mathrm{TV}}\left(P^t(x,\cdot),\pi\right)$. Le temps de mélange est défini pour $\varepsilon \in]0,1[$, par

$$t_{\mathsf{mix}}(\varepsilon) = \min\left\{t \in \mathbb{N}\,,\ \mathcal{D}(t) \le \varepsilon\right\}\,.$$

Définissons aussi

$$\overline{\mathcal{D}}(t) = \max_{x,y\in\Omega} d_{\mathrm{TV}}\left(P^t(x,\cdot),P^t(y,\cdot)\right)\,,$$

et $\tau(\varepsilon) = \min\left\{t \in \mathbb{N}\,,\ \overline{\mathcal{D}}(t) \le \varepsilon\right\}$. On montre facilement que $\mathcal{D}(t) \le \overline{\mathcal{D}}(t) \le 2\mathcal{D}(t)$, ce qui implique $t_{\mathsf{mix}}(\varepsilon) \le \tau(\varepsilon) \le t_{\mathsf{mix}}(\varepsilon/2)$.

Soit $f : \Omega^n \to \mathbb{R}$ une fonction telle que pour tout $x, y \in \Omega^n$,

$$f(y) - f(x) \le \sum_{i=1}^n c_i \mathbb{1}_{x_i\neq y_i}\,,$$

avec $c_1,\dots,c_n \ge 0$, et soit $Z = f(X_1,\dots,X_n)$. Nous allons montrer que pour tout $t \ge 0$, et pour tout $\varepsilon \in]0,1[$,

$$\mathbf{P}\left(Z - \mathbf{E}Z \ge t\right) \le \exp\left\{-\frac{t^2}{2\left(\frac{2-\varepsilon}{1-\varepsilon}\right)^2 \tau(\varepsilon)\sum_{i=1}^n c_i^2}\right\}\,.$$

(Pour des raffinements de cette inégalité et des extensions à des suites dépendantes plus générales que des chaînes de Markov, voir par exemple Samson et al. [22], Kontorovich [14], Paulin et al. [19].)

L'idée va être de décomposer la suite $(X_1,\dots,X_n)$ en blocs de longueur $\tau(\varepsilon)$ et de considérer la martingale de Doob associée à la filtration engendrée pas ces blocs. Pour cela, écrivons

$n = p.\tau(\varepsilon) + s$, avec $p \geq 0$ et $s \in [\![0, \tau(\varepsilon) - 1]\!]$, et posons

$$\begin{aligned} Y_1 &= \left(X_1, \ldots, X_{\tau(\varepsilon)}\right) , \\ &\vdots \\ Y_p &= \left(X_{(p-1)\tau(\varepsilon)+1}, \ldots, X_{p\tau(\varepsilon)}\right) , \\ Y_{p+1} &= \left(X_{p\tau(\varepsilon)+1}, \ldots, X_n\right) . \end{aligned}$$

Pour $k \in [\![1, p]\!]$, on pose $C_k = c_{(k-1)\tau(\varepsilon)+1} + \cdots + c_{k\tau(\varepsilon)}$ et $C_{p+1} = c_{p\tau(\varepsilon)+1} + \cdots + c_n$. Remarquons que pour tout $k \in [\![1, p+1]\!]$, et pour tout $y_1^k = (y_1, \ldots, y_k)$, on a

$$\left|\mathbf{E}\left[Z \mid Y_1^k = y_1^k\right] - \mathbf{E}\left[Z \mid Y_1^{k-1} = y_1^{k-1}\right]\right| \leq \max_{z,z'} \left|\mathbf{E}\left[Z \mid Y_1^k = y_1^{k-1}z\right] - \mathbf{E}\left[Z \mid Y_1^k = y_1^{k-1}z'\right]\right| .$$

En utilisant la définition de $\tau(\varepsilon)$ et la caractérisation de la distance en variation totale par couplage, ainsi que l'hypothèse sur la fonction f, on obtient, pour $k \in [\![1, p]\!]$,

$$\left|\mathbf{E}\left[Z \mid Y_1^k = y_1^{k-1}z\right] - \mathbf{E}\left[Z \mid Y_1^k = y_1^{k-1}z'\right]\right| \leq C_k + C_{k+1} + \varepsilon C_{k+2} + \cdots + \varepsilon^{p-k} C_{p+1} ,$$

et

$$\left|\mathbf{E}\left[Z \mid Y_1^{p+1} = y_1^p z\right] - \mathbf{E}\left[Z \mid Y_1^{p+1} = y_1^p z'\right]\right| \leq C_{p+1} .$$

En notant Δ la matrice triangulaire supérieure donnée par

$$\Delta = \begin{pmatrix} 1 & 1 & \varepsilon & \ldots & \ldots & \varepsilon^{p-1} \\ & 1 & 1 & \varepsilon & \ldots & \varepsilon^{p-2} \\ & & \ddots & \ddots & & \\ & & & \ddots & \ddots & \\ & & & & 1 & 1 \\ & & & & & 1 \end{pmatrix}$$

et C le vecteur $C = (C_1, \ldots, C_{p+1})$, on a, pour tout $k \in [\![1, p+1]\!]$,

$$\left|\mathbf{E}[Z \mid \mathcal{F}_k] - \mathbf{E}[Z \mid \mathcal{F}_{k-1}]\right| \leq (\Delta C)_k ,$$

où $\mathcal{F}_k = \sigma(Y_1, \ldots, Y_k)$. Par l'inégalité d'Azuma-Hoeffding, on obtient

$$\mathbf{P}\left(Z - \mathbf{E}Z \geq t\right) \leq \exp\left\{-\frac{t^2}{2\sum_{k=1}^{p+1}(\Delta C)_k^2}\right\} .$$

Pour conclure, notons que

$$\sum_{k=1}^{p+1}(\Delta C)_k^2 = \|\Delta C\|^2 \leq \|\Delta\|^2 \|C\|^2 .$$

D'une part

$$\|\Delta\| \leq 1 + 1 + \varepsilon + \cdots + \varepsilon^{p-1} \leq 1 + \frac{1}{1-\varepsilon} = \frac{2-\varepsilon}{1-\varepsilon} .$$

D'autre part, par l'inégalité de Cauchy-Schwarz,

$$\|C\|^2 \leq \tau(\varepsilon) \sum_{i=1}^{n} c_i^2 .$$

2. Concentration avec dépendance négative

2.1. Association négative.

Définition 8.1. Une suite $(X_n)_{n\geq 1}$ de variables aléatoires réelles est dite négativement associée (NA) si pour tous sous ensembles finis disjoints $A, B \subset \mathbb{N}^*$, et pour toutes fonctions $f : \mathbb{R}^{|A|} \to \mathbb{R}$ et $g : \mathbb{R}^{|B|} \to \mathbb{R}$, croissantes coordonnée par coordonnée, on a

$$\mathbf{E}\left[f(X_A)g(X_B)\right] \leq \mathbf{E}\left[f(X_A)\right]\mathbf{E}\left[g(X_B)\right]$$

Une conséquence importante de l'association négative (NA) est que toutes les bornes de concentration issues de la méthode de Chernoff pour les sommes de variables indépendantes s'appliquent automatiquement aux sommes de variables négativement associées. En effet, si $X_1, \ldots, X_n$ sont négativement associées, alors, pour tout $\lambda \in \mathbb{R}$, on a

$$\mathbf{E}\left[e^{\lambda \sum_{i=1}^n X_i}\right] \leq \prod_{i=1}^n \mathbf{E}\left[e^{\lambda X_i}\right] .$$

Énonçons deux propriétés simples mais très utiles de l'association négative :

(1) Si $X = (X_1, \ldots, X_n)$ est NA, si $Y = (Y_1, \ldots, Y_m)$ est NA, et si X et Y sont indépendantes, alors (X, Y) est NA.

(2) Si $X = (X_1, \ldots, X_n)$ est NA, si $A_1, \ldots, A_k$ sont des sous-ensembles disjoints de $[\![1, n]\!]$, et si $h_1, \ldots, h_k$ sont des fonctions réelles croissantes définies respectivement sur $\mathbb{R}^{|A_1|}, \ldots, \mathbb{R}^{|A_k|}$, alors la suite $(h_1(X_{A_1}), \ldots, h_k(X_{A_k}))$ est NA.

Exemple 8.1 (Bins and balls)**.** Reprenons l'exemple de variables aléatoires $X_1, \ldots, X_n$ i.i.d. de loi $(p_j)_{j\geq 1}$ sur $\mathbb{N}^*$. Remarquons que pour tout $i \in [\![1, n]\!]$, la suite $(\mathbb{1}_{X_i=j})_{j\geq 1}$ est NA. Cela provient d'un résultat plus général : si $(Z_j)_{j\geq 1}$ est une suite de variables à valeurs dans $\{0, 1\}$ telle que $\sum_{j\geq 1} Z_j = 1$, alors $(Z_j)_{j\geq 1}$ est NA. En effet, soient $A, B \subset \mathbb{N}^*$ finis disjoints, et soient f et g des fonctions réelles croissantes définies respectivement sur $\{0, 1\}^{|A|}$ et $\{0, 1\}^{|B|}$. Sans perte de généralité, on peut supposer que $f(0, \ldots, 0) = 0$ et que $g(0, \ldots, 0) = 0$. Dans ce cas, $\mathbf{E}[f(Z_A)] \geq 0$ et $\mathbf{E}[g(Z_B)] \geq 0$. Mais comme au plus une variable Z_j, pour $j \in A \cup B$, vaut 1, on a nécessairement $\mathbf{E}[f(Z_A)g(Z_B)] = 0$. Maintenant, par la propriété (1), la suite $(\mathbb{1}_{X_i=j})_{j\geq 1, 1\leq i\leq n}$ est NA. Et par la propriété (2), la suite $(N_j)_{j\geq 1}$ avec

$$N_j = \sum_{i=1}^n \mathbb{1}_{X_i=j} ,$$

est NA. Rappelons que $K_n = \sum_{j\geq 1} \mathbb{1}_{N_j>0}$ correspond au nombre de symboles distincts. Encore par la propriété (2), la suite $(\mathbb{1}_{N_j>0})_{j\geq 1}$ est NA. Ainsi, en utilisant l'inégalité de Bennett, on obtient, pour tout $\lambda \in \mathbb{R}$,

$$\begin{aligned}\log \mathbf{E}\left[e^{\lambda(K_n - \mathbf{E}K_n)}\right] &\leq \sum_{j\geq 1} \log \mathbf{E}\left[e^{\lambda(\mathbb{1}_{N_j>0} - \mathbf{E}\mathbb{1}_{N_j>0})}\right] \\ &\leq \sum_{j\geq 1} \mathbf{Var}(\mathbb{1}_{N_j>0})\phi(\lambda) .\end{aligned}$$

Par l'inégalité d'Efron-Stein, on obtient $\mathbf{Var}(\mathbb{1}_{N_j>0}) \leq \mathbf{E}[\mathbb{1}_{N_j=1}]$. Ainsi

$$\log \mathbf{E}\left[e^{\lambda(K_n - \mathbf{E}K_n)}\right] \leq \mathbf{E}[K_{n,1}]\phi(\lambda) ,$$

où $K_{n,1} = \sum_{j \geq 1} \mathbb{1}_{N_j=1}$ est le nombre de symboles apparaissant une seule fois dans l'échantillon.

2.2. Propriété de recouvrement stochastique.

Définition 8.2. Une mesure μ sur $\{0,1\}^n$ est dite k-homogène ($k \in [\![1,n]\!]$) si son support est inclus dans

$$\left\{ x \in \{0,1\}^n \, , \, \sum_{i=1}^{n} x_i = k \right\} .$$

Pour $x, y \in \{0,1\}^n$, on note $x \geq y$ si pour tout $i \in [\![1,n]\!]$, $x_i \geq y_i$. De plus, on dit que x recouvre y, et l'on note $x \succ y$, si x et y coïncident sur toutes les coordonnées sauf sur au plus une pour laquelle $x_i = 1$ et $y_i = 0$.

Définition 8.3. Soient μ et ν deux mesures sur $\{0,1\}^n$. On dit que μ domine stochastiquement ν si pour tout sous-ensemble $A \subset \{0,1\}^n$ croissant (i.e. fermé supérieurement), on a $\mu(A) \geq \nu(A)$. De façon équivalente, on peut coupler μ et ν de telle sorte que le support soit inclus dans $\{(x,y),\, x \geq y\}$.

Définition 8.4. Soient μ et ν deux mesures sur $\{0,1\}^n$. On dit que μ recouvre stochastiquement ν (et l'on note $\mu \succ \nu$) si l'on peut coupler μ et ν de telle sorte que le support soit inclus dans $\{(x,y),\, x \succ y\}$.

Définition 8.5. Soit μ une mesure de probabilité sur $\{0,1\}^n$. On dit que μ possède la propriété de recouvrement stochastique (SCP) si pour tout $S \subset [\![1,n]\!]$, et pour tous $x, y \in \{0,1\}^{|S|}$, avec $x \succ y$, on a

$$\mu\left(\cdot \,\middle|\, X_S = y\right) \succ \mu\left(\cdot \,\middle|\, X_S = x\right) ,$$

où $\mu\left(\cdot \,\middle|\, X_S = x\right)$ correspond à la loi conditionnelle de X_{S^c} sachant $\{X_S = x\}$.

Donnons quelques exemples de mesures possédant la SCP :

- *mesures déterminantales* : une mesure de probabilité μ sur $\{0,1\}^n$ est dite déterminantale s'il existe une matrice hermitienne $K \in \mathcal{M}_n(\mathbb{C})$ telle que, pour tout $S \subset [\![1,n]\!]$,

  $$\mathbf{E} \prod_{j \in S} X_j = \det K_S \, ,$$

  où $X \sim \mu$ et où K_S correspond à la matrice obtenue en ne retenant que les lignes et les colonnes d'indice appartenant à S. Borcea et al. [2] ont montré que de telles mesures possédaient la SCP (en fait, ils montrent qu'elles vérifient une propriété plus forte appelée propriété de Rayleigh forte).Un exemple de mesure déterminantale est celle des arbres couvrants aléatoires. Soit $G = (V, E)$ un graphe fini connexe, dont les arêtes sont numérotées de 1 à n. Un arbre couvrant est un sous-ensemble de E, sans cycle et qui connecte tous les sommets. On peut tirer un arbre couvrant aléatoirement de la façon suivante : soit $\omega(e) \geq 0$ le poids de l'arête e, et μ la probabilité sur les arbres couvrants telles que $\mu(T) \propto \prod_{e \in T} \omega(e)$. Vue comme une mesure sur $\{0,1\}^n$ ($X_e = 1$ ssi $e \in T$), c'est une mesure déterminantale (voir Burton and Pemantle [4] et Lyons [16]).
- *mesures sur l'ensemble des bases d'un matroïde* : soit E un ensemble fini non-vide et $\mathcal{B}$ une collection non-vide de parties de E de même cardinal. La paire $(E, \mathcal{B})$ est appelée un matroïde si elle vérifie la propriété :

  $$\forall A, B \in \mathcal{B} \, , \; \forall \in A \setminus B \, , \; \exists b \in B \setminus A \, , \; A \cup \{b\} \setminus \{a\} \in \mathcal{B} \, .$$

L'ensemble $\mathcal{B}$ alors appelé l'ensemble des bases du matroïde et le cardinal des bases est appelé le rang. Si E est l'ensemble d'arêtes d'un graphe G fini connexe, et $\mathcal{B}$ l'ensemble des arbres couvrants de G, alors la paire $(E, \mathcal{B})$ est un matroïde. Il est naturel de munir $\mathcal{B}$ de la mesure uniforme. Plus généralement, pour une suite $(\omega(e))_{e\in E}$ de poids positifs, on peut définir la mesure de probabilité pondérée μ_ω telle que pour tout $A \in \mathcal{B}$, $\mu_\omega(A) \propto \prod_{e\in A} \omega(e)$. Si $|E| = n$, ces mesures peuvent être vues comme des mesures sur $\{0,1\}^n$ (en identifiant E avec $[\![1, n]\!]$ et une partie $A \sim \mu_\omega$ au vecteur $X = (X_e)_{e\in E}$ avec $X_e = \mathbb{1}_{e\in A}$). Pour des matroïdes généraux, il est possible d'avoir $\mathbf{E}[X_e X_f] > \mathbf{E}[X_e]\mathbf{E}[X_f]$. Si pour tous $e, f \in E$, on a $\mathbf{E}[X_e X_f] \leq \mathbf{E}[X_e]\mathbf{E}[X_f]$, on dit que la matroïde a la propriété de corrélation négative. Une notion plus forte est celle de matroïde équilibré. Les mineurs d'un matroïde sont tous les matroïdes qui peuvent être obtenus en répétant l'opération de choisir un élément e et de ne garder soit que les bases qui contiennent e, soit celles qui ne le contiennent pas. On dit qu'un matroïde est équilibré si tous ses mineurs (dont lui-même) ont la propriété de corrélation négative. Feder and Mihail [9] ont montré que les mesures μ_ω sur l'ensemble des bases d'un matroïde équilibré possèdent la SCP.

- *Bernoulli indépendantes conditionnées à leur somme* : soit $k \in [\![0, n]\!]$ et $\lambda_1, \ldots, \lambda_n > 0$. La mesure sur $\{0,1\}^n$ donnée par

$$\forall x \in \{0,1\}^n\,,\ \mu(x) = \frac{\prod_{i=1}^n \lambda_i^{x_i} \mathbb{1}_{\|x\|=k}}{\sum_{y,\,\|y\|=k} \prod_{j=1}^n \lambda_j^{y_j}}$$

possède la SCP. En fait, il s'agit d'un cas particulier de mesure pondérée sur l'ensemble des bases d'un matroïde ($E = [\![1, n]\!]$ et $\mathcal{B}$ l'ensemble des parties de E de cardinal k, et $\omega(i) = \lambda_i$). Remarquons aussi qu'il s'agit de la loi de $(X_1, \ldots, X_n)$ où les variables X_i sont indépendantes, et $X_i \sim \mathcal{B}\left(\frac{\lambda_i}{1+\lambda_i}\right)$.

Le résultat suivant est dû à Pemantle and Peres [20].

Proposition 8.1. *Soit μ une mesure de probabilité sur $\{0,1\}^n$, k-homogène et possédant la SCP, et soit $(X_1, \ldots, X_n) \sim \mu$. Soit $f : \{0,1\}^n \to \mathbb{R}$ une fonction 1-lipschitzienne et $Z = f(X_1, \ldots, X_n)$. Alors, pour tout $t \geq 0$,*

$$\mathbf{P}\left(Z - \mathbf{E}Z \geq t\right) \leq \exp\left\{-\frac{t^2}{8k}\right\}\,.$$

Preuve de la Proposition 8.1. À $X = (X_1, \ldots, X_n)$, on associe le vecteur $Y = (Y_1, \ldots, Y_k)$ donc la loi est donnée séquentiellement par : pour tout $j \in [\![1, k]\!]$, pour tout $i \in [\![1, n]\!]$,

$$\mathbf{P}\left(Y_j = i \;\middle|\; Y_1, \ldots, Y_{j-1}\right) = \mathbf{P}\left(X_i = 1 \;\middle|\; X_{Y_1} = \cdots = X_{Y_{j-1}} = 1\right) \mathbb{1}_{i\neq Y_1,\ldots,Y_{j-1}}\,.$$

Le vecteur Y donne l'emplacement des 1 dans le vecteur X, dans un ordre échangeable. Soit g telle que $f(X) = g(Y)$. Remarquons que la fonction g est 2-lipschitzienne. Considérons la martingale de Doob

$$M_j = \mathbf{E}[g(Y) \;|\; \mathcal{F}_j] - \mathbf{E}[g(Y)]\,,$$

associée à la filtration $\mathcal{F}_j$ engendrée par $Y_1, \ldots, Y_j$. Pour $1 \leq j \leq k-1$, et $y_1, \ldots, y_{j+1} \in [\![1, n]\!]$ tels que $\mathbf{P}(Y_1 = y_1, \ldots, Y_{j+1} = y_{j+1}) > 0$, on a, en notant E_j l'événement $\{Y_1 = y_1, \ldots, Y_j = y_j\}$,

$$\begin{aligned}&\mathbf{E}\left[g(Y) \;\middle|\; E_j, Y_{j+1} = y_{j+1}\right] - \mathbf{E}\left[g(Y) \;\middle|\; E_j\right]\\ &\quad = \mathbf{P}\left(Y_{j+1} \neq y_{j+1} \;\middle|\; E_j\right)\left\{\mathbf{E}\left[g(Y) \;\middle|\; E_j, Y_{j+1} = y_{j+1}\right] - \mathbf{E}\left[g(Y) \;\middle|\; E_j, Y_{j+1} \neq y_{j+1}\right]\right\}\,.\end{aligned}$$

Par la SCP, la différence entre les deux espérances conditionnelles ci-dessus est comprise entre -2 et 2. Ainsi, en appliquant l'inégalité d'Azuma–Hoeffding 3.1, on a

$$\mathbf{P}\left(Z - \mathbf{E}Z \geq t\right) \leq \exp\left\{-\frac{2t^2}{4^2 k}\right\} = \exp\left\{-\frac{t^2}{8k}\right\} .$$

■

3. Paires échangeables

Dans cette section, nous introduisons une méthode développée par Chatterjee pour obtenir de la concentration en l'absence d'indépendance. Cette méthode repose sur la notion de paires échangeables, et correspond à une variante de la méthode de Stein. La méthode de Stein est une technique élégante et puissante pour démontrer des approximations distributionnelles. Pour des introductions à cette méthode, voir Barbour and Chen [1], Ross [21], et Chatterjee [6].

Soit (X, X') une paire échangeable de variables aléatoires à valeurs dans un ensemble $\mathcal{X}$, i.e. telle que $(X, X') \sim (X', X)$. Soit $f : \mathcal{X} \to \mathbb{R}$ une fonction mesurable telle que $\mathbf{E}f(X) = 0$, et soit $F : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ une fonction mesurable antisymétrique (i.e. $f(x, x') = -f(x', x)$) et telle que

$$\mathbf{E}\left[F(X, X') \;\middle|\; X\right] = f(X) .$$

On suppose pour commencer que F est donnée, mais on verra comment on peut trouver une telle fonction F à partir de f. Notons déjà un cas particulier où l'on peut facilement trouver une telle fonction : pour $0 < a \leq 1$, une a-paire de Stein est une paire échangeable (X, X') vérifiant $\mathbf{E}[X' \mid X] = (1 - a)X$. Si $(f(X), f(X'))$ est une a-paire de Stein, alors on vérifie facilement que la fonction antisymétrique F donnée par $F(x, x') = \frac{f(x) - f(x')}{a}$ convient.

Remarquons que pour toute fonction mesurable $h : \mathcal{X} \to \mathbb{R}$ telle que $\mathbf{E}\left|h(X)F(X, X')\right| < \infty$, on a

$$\mathbf{E}\left[f(X)h(X)\right] = \frac{1}{2}\mathbf{E}\left[\big(h(X) - h(X')\big)\, F(X, X')\right] . \tag{8.1}$$

En effet, on a $\mathbf{E}\left[f(X)h(X)\right] = \mathbf{E}\left[h(X)F(X, X')\right]$, et, par échangeabilité de (X, X') et antisymétrie de F, on a

$$\mathbf{E}\left[h(X)F(X, X')\right] = \mathbf{E}\left[h(X')F(X', X)\right] = -\mathbf{E}\left[h(X')F(X, X')\right] .$$

En particulier, en prenant $h = f$, on obtient

$$\mathbf{Var}\left(f(X)\right) = \mathbf{E}\left[f(X)^2\right] = \frac{1}{2}\mathbf{E}\left[\big(f(X) - f(X')\big)\, F(X, X')\right] .$$

Pour $x \in \mathcal{X}$, on définit

$$v(x) = \frac{1}{2}\mathbf{E}\left[\left|\big(f(X) - f(X')\big)\, F(X, X')\right| \;\middle|\; X = x\right] .$$

Le théorème ci-dessous est dû à Chatterjee [5].

Proposition 8.2. *Soit (X, X') une paire échangeable, et soient f, F, v les fonctions définies ci-dessus. On suppose que pour tout $\lambda \in \mathbb{R}$, on a $\mathbf{E}\left[e^{\lambda f(X)}|F(X, X')|\right] < \infty$. Supposons qu'il existe $b, c \geq 0$ tels que pour tout $x \in \mathcal{X}$,*

$$v(x) \leq bf(x) + c .$$

Alors, pour tout $t > 0$,

$$\mathbf{P}\left(f(X) > t\right) \leq \exp\left\{-\frac{t^2}{2(c+bt)}\right\} \quad \textit{et} \quad \mathbf{P}\left(f(X) < -t\right) \leq \exp\left\{-\frac{t^2}{2c}\right\}.$$

Preuve de la Proposition 8.2. Notons $m(\lambda) = \mathbf{E}\left[e^{\lambda f(X)}\right]$. En utilisant (8.1), on a, pour tout $\lambda \in \mathbb{R}$,

$$m'(\lambda) = \mathbf{E}\left[f(X)e^{\lambda f(X)}\right] = \frac{1}{2}\mathbf{E}\left[\left(e^{\lambda f(X)} - e^{\lambda f(X')}\right) F(X, X')\right]$$

Par convexité de l'exponentielle, on a pour tous $x \neq y$, $\frac{e^x - e^y}{x-y} \leq \frac{e^x + e^y}{2}$. On obtient ainsi, en utilisant à nouveau l'échangeabilité de (X, X'), puis l'hypothèse sur v,

$$\begin{aligned}
|m'(\lambda)| &\leq \frac{|\lambda|}{4}\mathbf{E}\left[\left(e^{\lambda f(X)} + e^{\lambda f(X')}\right)\left|\left(f(X) - f(X')\right) F(X, X')\right|\right] \\
&= |\lambda|\mathbf{E}\left[e^{\lambda f(X)} v(X)\right] \\
&\leq |\lambda|\mathbf{E}\left[e^{\lambda f(X)}(bf(X) + c)\right] \\
&= |\lambda|\left(bm'(\lambda) + cm(\lambda)\right).
\end{aligned}$$

Comme $\lambda \mapsto m'(\lambda)$ est convexe avec $m'(0) = 0$, on a $\frac{m'(\lambda)}{\lambda} \geq 0$. Pour $0 < \lambda < 1/b$, on obtient

$$\frac{m'(\lambda)}{m(\lambda)} \leq \frac{\lambda c}{1 - \lambda b},$$

et

$$\log m(\theta) \leq \int_0^\theta \frac{cu}{1 - bu}\mathrm{d}u \leq \frac{c\theta^2}{2(1 - b\theta)}.$$

La méthode de Chernoff donne alors que pour tout $t > 0$, $\mathbf{P}(f(X) > t) \leq \exp\left(-\frac{t^2}{2(c+bt)}\right)$. Pour $\lambda < 0$, on a $\frac{m'(\lambda)}{m(\lambda)} \geq \lambda c$, donc par intégration $\log m(\lambda) \leq \frac{c\lambda^2}{2}$ et la méthode de Chernoff permet de conclure. ■

Application : poids d'une permutation. Soit $A = (a_{i,j})_{1 \leq i,j \leq n}$ une matrice réelle et soit π une permutation aléatoire de $\{1, \ldots, n\}$, uniformément distribuée. On s'intéresse au poids de la permutation π défini comme

$$Z = \sum_{i=1}^n a_{i,\pi(i)}.$$

Par exemple, si A est la matrice identité, la variable X correspond au nombre de points fixes. Ou bien si $a_{i,j} = v_j \mathbb{1}_{\{i \leq k\}}$, X correspond à la somme des valeurs d'un échantillon de taille k tiré sans remise dans une population de taille n. Ou encore si $a_{i,j} = |i - j|$, il s'agit de la distance de Spearman entre π et l'identité.

Proposition 8.3. *Supposons que les poids $a_{i,j}$ appartiennent tous à $[0, 1]$, et soit $Z = \sum_{i=1}^n a_{i,\pi(i)}$ où π est une permutation uniforme. Alors pour tout $t \geq 0$,*

$$\mathbf{P}\left(Z - \mathbf{E}Z \geq t\right) \leq \exp\left\{-\frac{t^2}{4\mathbf{E}Z + 2t}\right\} \quad \textit{et} \quad \mathbf{P}\left(Z - \mathbf{E}Z \leq -t\right) \leq \exp\left\{-\frac{t^2}{4\mathbf{E}Z}\right\}.$$

Preuve de la Proposition 8.3. Soit π une permutation uniforme de $\{1,\dots,n\}$ et I, J deux entiers tirés uniformément et indépendamment dans $\{1,\dots,n\}$. Soit $\pi' = \pi \circ (I,J)$ la permutation obtenue à partir de π en transposant I et J. Notons

$$W = \sum_{i=1}^n a_{i,\pi(i)} - \frac{1}{n}\sum_{i,j=1}^n a_{i,j} \quad \text{et } W' = \sum_{i=1}^n a_{i,\pi'(i)} - \frac{1}{n}\sum_{i,j=1}^n a_{i,j}\,.$$

La paire (W,W') est clairement échangeable et l'on a

$$\begin{aligned}\mathbf{E}\left[W'-W \mid \pi\right] &= \mathbf{E}\left[a_{I,\pi(J)} + a_{J,\pi(I)} - a_{I,\pi(I)} - a_{J,\pi(J)} \mid \pi\right] \\ &= 2\left(\frac{1}{n^2}\sum_{i,j} a_{i,\pi(j)} - \frac{1}{n}\sum_{i=1}^n a_{i,\pi(i)}\right) \\ &= -\frac{2}{n}W\,.\end{aligned}$$

Ainsi (W,W') est une $\frac{2}{n}$-paire de Stein. De plus, en utilisant que $0 \le a_{i,j} \le 1$,

$$\begin{aligned}\mathbf{E}\left[(W'-W)^2 \mid \pi\right] &= \mathbf{E}\left[\left(a_{I,\pi(J)} + a_{J,\pi(I)} - a_{I,\pi(I)} - a_{J,\pi(J)}\right)^2 \mid \pi\right] \\ &= \frac{1}{n^2}\sum_{i,j}\left(a_{i,\pi(j)} + a_{j,\pi(i)} - a_{i,\pi(i)} - a_{j,\pi(j)}\right)^2 \\ &\le \frac{2}{n^2}\sum_{i,j}\left(a_{i,\pi(j)} + a_{j,\pi(i)} + a_{i,\pi(i)} + a_{j,\pi(j)}\right) \\ &= \frac{4(Z+\mathbf{E}Z)}{n} = \frac{4W}{n} + \frac{8\mathbf{E}Z}{n}\,.\end{aligned}$$

Ainsi, en appliquant la Proposition 8.2 avec $F(W,W') = \frac{W-W'}{2/n}$, $b=1$ et $c = 2\mathbf{E}Z$, on obtient bien le résultat voulu. ∎

Magnétisation dans le modèle de Curie–Weiss. Soit $\beta \in \mathbb{R}_+$ et $h \in \mathbb{R}$. Le modèle de Curie–Weiss avec interactions ferromagnétiques à température inverse β et champ externe h est donné par la mesure de probabilité μ sur $\{-1,1\}^n$ définie par

$$\mu(\sigma) = \frac{1}{Z}\exp\left\{\frac{\beta}{n}\sum_{i<j}\sigma_i\sigma_j + \beta h\sum_{i=1}^n \sigma_i\right\}\,,$$

où Z est la constante de normalisation. La magnétisation d'une configuration σ est définie comme

$$m(\sigma) = \frac{1}{n}\sum_{i=1}^n \sigma_i\,.$$

Pour n grand et σ distribuée selon μ, la magnétisation satisfait

$$m(\sigma) \approx \tanh\left(\beta m(\sigma) + \beta h\right)\,.$$

Cette équation a une seule solution pour β sous une certaine valeur critique, et plusieurs solutions pour β au-dessus.

Proposition 8.4. *Pour tout $\beta \ge 0$, pour tout $h \in \mathbb{R}$, et pour tout $t \ge 0$,*

$$\mathbf{P}\left(|m(\sigma) - \tanh\left(\beta m(\sigma) + \beta h\right)| \ge \frac{\beta}{n} + \frac{t}{\sqrt{n}}\right) \le 2\exp\left\{-\frac{t^2}{4(1+\beta)}\right\}\,.$$

Preuve de la Proposition 8.4. Soit $\sigma \sim \mu$ et σ' obtenue en choisissant une coordonnée $I \in [\![1, n]\!]$ uniformément au hasard, et en remplaçant σ_I par un élément distribué selon la loi conditionnelle de la $I^{\text{ième}}$ coordonnée sachant toutes les autres. On vérifie facilement que la paire (σ, σ') est échangeable. Soit

$$F(\sigma, \sigma') = \sum_{i=1}^{n} (\sigma_i - \sigma'_i) \,,$$

et

$$m_i(\sigma) = \frac{1}{n} \sum_{j \neq i} \sigma_j \,.$$

On a

$$\mathbf{P}\left(\sigma_i = 1 \,\middle|\, \sigma_j,\, j \neq i\right) = \frac{\exp\left(\beta m_i(\sigma) + \beta h\right)}{\exp\left(\beta m_i(\sigma) + \beta h\right) + \exp\left(-\beta m_i(\sigma) - \beta h\right)} \,.$$

Ainsi

$$\mathbf{E}\left[\sigma_i \,\middle|\, \sigma_j,\, j \neq i\right] = \tanh\left(\beta m_i(\sigma) + \beta h\right) \,,$$

et

$$\begin{aligned} f(\sigma) &= \mathbf{E}\left[F(\sigma, \sigma') \,\middle|\, \sigma\right] \\ &= \frac{1}{n} \sum_{i=1}^{n} \left(\sigma_i - \mathbf{E}\left[\sigma_i \,\middle|\, \sigma_j,\, j \neq i\right]\right) \\ &= m(\sigma) - \frac{1}{n} \sum_{i=1}^{n} \tanh\left(\beta m_i(\sigma) + \beta h\right) \,. \end{aligned}$$

Comme σ et σ' ne diffèrent qu'en au plus une coordonnée, on a $|F(\sigma, \sigma')| \leq 2$, et $|m(\sigma) - m(\sigma')| \leq \frac{2}{n}$. En utilisant le fait que la fonction $x \mapsto \tanh(x)$ est 1-lipschitzienne, on a

$$\left|f(\sigma) - f(\sigma')\right| \leq \left|m(\sigma) - m(\sigma')\right| + \frac{\beta}{n} \sum_{i=1}^{n} \left|m_i(\sigma) - m_i(\sigma')\right| \leq \frac{2(1+\beta)}{n} \,.$$

Ainsi $v(\sigma) \leq \frac{2(1+\beta)}{n}$ et la Proposition 8.2 appliquée avec $b = 0$ et $c = \frac{2(1+\beta)}{n}$ donne

$$\mathbf{P}\left(\left|m(\sigma) - \frac{1}{n} \sum_{i=1}^{n} \tanh\left(\beta m_i(\sigma) + \beta h\right)\right| \geq \frac{t}{\sqrt{n}}\right) \leq 2 \exp\left\{-\frac{t^2}{4(1+\beta)}\right\} \,.$$

Pour conclure, il suffit d'utiliser à nouveau le fait que $x \mapsto \tanh(x)$ est 1-lipschitzienne :

$$\left|\frac{1}{n} \sum_{i=1}^{n} \tanh\left(\beta m_i(\sigma) + \beta h\right) - \tanh\left(\beta m(\sigma) + \beta h\right)\right| \leq \frac{\beta}{n} \sum_{i=1}^{n} |m(\sigma) - m_i(\sigma)| \leq \frac{\beta}{n} \cdot$$

■

Bibliographie

[1] A. D. Barbour and L. H. Chen. Steins (magic) method. *arXiv preprint arXiv :1411.1179*, 2014.

[2] J. Borcea, P. Brändén, and T. Liggett. Negative dependence and the geometry of polynomials. *Journal of the American Mathematical Society*, 22(2) :521–567, 2009.

[3] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities*. Oxford University Press, Oxford, 2013.

[4] R. Burton and R. Pemantle. Local characteristics, entropy and limit theorems for spanning trees and domino tilings via transfer-impedances. *The Annals of Probability*, pages 1329–1371, 1993.

[5] S. Chatterjee. Stein's method for concentration inequalities. *Probability theory and related fields*, 138(1) :305–321, 2007.

[6] S. Chatterjee. A short survey of stein's method. *arXiv preprint arXiv :1404.1392*, 2014.

[7] L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

[8] D. P. Dubhashi and A. Panconesi. *Concentration of measure for the analysis of randomized algorithms*. Cambridge University Press, 2009.

[9] T. Feder and M. Mihail. Balanced matroids. In *Proceedings of the twenty-fourth annual ACM symposium on Theory of computing*, pages 26–38, 1992.

[10] D. A. Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

[11] D. A. Grable. A large deviation inequality for functions of independent, multi-way choices. *Combinatorics, probability and Computing*, 7(1) :57–63, 1998.

[12] D. Haussler. Sphere packing numbers for subsets of the boolean n-cube with bounded vapnik-chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69(2) :217–232, 1995.

[13] J. Kahn, G. Kalai, and N. Linial. *The influence of variables on Boolean functions*. Citeseer, 1989.

[14] L. Kontorovich. *Measure concentration of strongly mixing processes with applications*. PhD thesis, Carnegie Mellon University, School of Computer Science, Machine Learning ?, 2007.

[15] M. Ledoux. *The concentration of measure phenomenon*. Number 89. American Mathematical Soc., 2001.

[16] R. Lyons. Determinantal probability measures. *Publications Mathématiques de l'IHÉS*, 98 : 167–212, 2003.

[17] P. Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The annals of Probability*, pages 1269–1283, 1990.

[18] C. McDiarmid. Concentration. In *Probabilistic methods for algorithmic discrete mathematics*, pages 195–248. Springer, 1998.

[19] D. Paulin et al. Concentration inequalities for markov chains by marton couplings and spectral methods. *Electronic Journal of Probability*, 20, 2015.

[20] R. Pemantle and Y. Peres. Concentration of Lipschitz functionals of determinantal and other strong Rayleigh measures. *Combinatorics, Probability and Computing*, 23(1) :140–160, 2014.

[21] N. Ross. Fundamentals of Stein's method. *Probab. Surv*, 8 :210–293, 2011.

[22] P.-M. Samson et al. Concentration of measure inequalities for markov chains and *phi*-mixing processes. *The Annals of Probability*, 28(1) :416–461, 2000.

[23] J. A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2) :1–230, 2015.

[24] J. A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2) :1–230, 2015.

[25] R. Vershynin. *High-dimensional probability : An introduction with applications in data science*, volume 47. Cambridge university press, 2018.