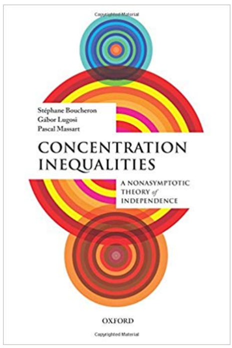
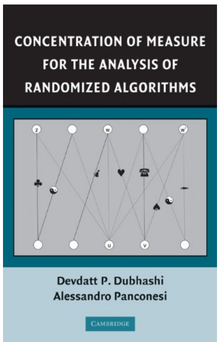
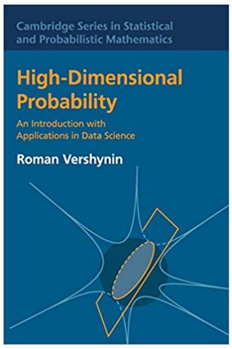
Master 2 Probabilités + M2A Algorithmes et Apprentissage
Sorbonne Université
2019-2020

Inégalités de concentration







Références :
Boucheron, Lugosi, Massart, *Concentration Inequalities* [4].
Dubhashi, Panconesi, *Concentration of Measure for the Analysis of Randomised Algorithms* [7].
Vershynin, *High-dimensional Probability* [11].

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

Table des matières

Chapitre 1

Variance, entropie, influences

Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes définies sur un espace mesurable $(\Omega, \mathcal{F}, \mathbf{P})$, à valeurs dans un espace mesurable $\mathcal{X}$, et soit $Z = f(X_1, \ldots, X_n)$ 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 $\mathbf{E}_i = \mathbf{E}[\cdot \mid X_1, \ldots, X_i]$ 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 $X_1, \ldots, X_n$. Celle-ci intervient maintenant pour écrire

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

où $\mathbf{E}^{(i)} = \mathbf{E}[\cdot \mid X^{(i)}]$ avec $X^{(i)} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$. 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 $X_1, \ldots, X_n$ des variables indépendantes à valeurs dans un espace mesurable $\mathcal{X}$, et soit $Z = f(X_1, \ldots, X_n)$ une fonction mesurable. Alors*

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

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

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

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

Dans le reste de ce chapitre, on s'intéresse au cas (simple mais déjà très riche) où $\mathcal{X} = \{0,1\}$ et où $X = (X_1, \ldots, X_n) \sim \mathcal{B}(p)^{\otimes n}$ 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

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

où $\bar{X}^{(i)} = (X_1, \ldots, X_{i-1}, 1-X_i, X_{i+1}, \ldots, X_n)$ est le vecteur obtenu en flippant la $i^{\text{ième}}$ coordonnée. 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.

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

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

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

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 $\mathbf{E}_\mu h = 1$, on a

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

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 1.2 (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 $\mathbf{E}^{(i)} = \mathbf{E}[\cdot \mid X^{(i)}]$ et $X^{(i)} = (X_1,\dots,X_{i-1},X_{i+1},\dots,X_n)$.

Preuve de la Proposition 1.2. 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

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

$\blacksquare$

Proposition 1.3 (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

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

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

Preuve de la Proposition 1.3. 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 1.4. *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 1.3.

Preuve de la Proposition 1.4. Remarquons d'abord que

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

En effet, pour $X \sim \mathcal{B}(p)^{\otimes n}$, on a

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

où pour la première égalité on a utilisé

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

En appliquant la Proposition 1.3, 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 1.4 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. [8] 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 1.5 (Kahn et al. [8])**.** *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 1.5. 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 1.4 implique que

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

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 1.6 (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 1.6. 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 $X = (X_1, \ldots, X_n) \sim \mu_p$, et si $X_i' = \mathbb{1}_{U_i \leq q}$, alors le vecteur $\widetilde{X}^{(i)} = (X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_n)$ 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 1.7. *Soit $f : \{0,1\}^n \to \{0,1\}$ une fonction monotone symétrique. Alors pour tout $\varepsilon \in]0,1/2[$,*

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

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

Preuve de la Proposition 1.7. Soit $p \leq p_{1/2}$. Par la Proposition 1.5 et le Lemme 1.6, et comme $\mu_p(A) \leq 1/2$, on a

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

soit

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

Ainsi pour $\varepsilon < 1/2$,

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

soit

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

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 $(u_1,\ldots,u_k)$ telle que pour tout $j \leq k-1$, $|u_j^1 - u_{j+1}^1| + |u_j^2 - u_{j+1}^2| = 1$). 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 1.7 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 $t \geq 0$,

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

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 $\lambda \in \mathbb{R}$,

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

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 $\lambda \in \mathbb{R}$,

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

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 $c/3$.*

3. Sommes de variables indépendantes

Soit $(X_1, \dots, X_n)$ 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 $\log(1+x) \leq x$. En particulier, Z est sous-poisson avec facteur de variance np et facteur d'échelle 1. Ainsi, pour $t \geq 0$,

$$\mathbf{P}\left(Z - \mathbf{E}Z \geq t\right) \leq e^{-\sup_{\lambda \geq 0}\left\{\lambda t - np(e^\lambda - \lambda - 1)\right\}} = e^{-nph\left(\frac{t}{np}\right)}\,,$$

où $h(x) = (1+x)\log(1+x) - x$. En utilisant l'inégalité $h(u) \geq \frac{u^2}{2(1+u/3)}$ (laissée en exercice), on obtient que pour $t \geq 0$,

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

Pour les déviations vers la gauche, on a pour $t \geq 0$ et $\lambda \geq 0$,

$$\mathbf{P}\left(Z - \mathbf{E}Z \leq -t\right) \leq e^{-\sup_{\lambda \geq 0}\{\lambda t - \psi_Z(-\lambda)\}} \leq e^{-\sup_{\lambda \geq 0}\left\{\lambda t - np(e^{-\lambda} + \lambda - 1)\right\}}\,.$$

En utilisant le fait que pour $\lambda \geq 0$, $e^{-\lambda} + \lambda - 1 \leq \frac{\lambda^2}{2}$, et en optimisant en $\lambda \geq 0$, on obtient

$$\mathbf{P}\left(Z - \mathbf{E}Z \leq -t\right) \leq \exp\left\{-\frac{t^2}{2np}\right\}\,, \tag{2.2}$$

En combinant (2.1) et (2.2), et en supposant $t \in [0, np]$, on obtient finalement

$$\mathbf{P}\left(|Z - \mathbf{E}Z| \geq t\right) \leq 2\exp\left\{-\frac{3t^2}{8np}\right\}\,. \tag{2.3}$$

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 2.2 (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 2.2. Montrons d'abord le résultat suivant : si X est une v.a.r. telle que $a \leq X \leq b$, alors pour tout $\lambda \in \mathbb{R}$,

$$\log \mathbf{E}\left[e^{\lambda(X - \mathbf{E}X)}\right] \leq \frac{\lambda^2 (b-a)^2}{8}. \tag{2.4}$$

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 $e^{\lambda Y} \leq Y e^{\lambda} + 1 - Y$, 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 $X_1, \dots, X_n$ 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. Dans ce cas, le comportement de la somme n'est plus gaussien, mais poissonien et appliquer l'inégalité de Hoeffding n'est pas judicieux.

3.2. Inégalité de Bennett.

Proposition 2.3 (Inégalité de Bennett)**.** *Soient $X_1, \dots, X_n$ 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 $t \geq 0$,

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

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

$$\mathbf{P}(Z - \mathbf{E}Z \geq t) \leq \exp\left\{-\frac{t^2}{2(v + ct/3)}\right\}. \tag{2.5}$$

Preuve de la Proposition 2.3. 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 $\log(1+x) \le x$. Maintenant, pour $t \ge 0$, 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 (2.5)) avec $c = 1$ et $v = 1$, on a

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

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 (2.3), pour $\varepsilon \in]0,1[$,

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

Et en utilisant une borne union,

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

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 repose 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 2.4 (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 $\lambda \in [0, 1/c[$,

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

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

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

Preuve de la Proposition 2.4. En notant $\phi(u) = e^u - u - 1$ et en utilisant l'inégalité $\log(1+x) \leq x$, on a

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

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

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

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

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

■

Exemple 2.10 (Norme d'un vecteur sous-gaussien)**.** Soit $X = (X_1, \ldots, X_n)$ 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

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

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

$$\sqrt{n \pm O(\sqrt{n})} = \sqrt{n} \pm O(1) \quad !!!$$

Chapitre 3

L'approche par martingales

1. L'inégalité d'Azuma-Hoeffding

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

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

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 $a_i \le Z_i - Z_{i-1} \le b_i$, alors, pour tout $t > 0$,*

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

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 $\lambda \ge 0$,

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

En utilisant (2.4), on a $\mathbf{E}\left[e^{\lambda(Z_n - Z_{n-1})} \mid \mathcal{F}_{n-1}\right] \le e^{\frac{\lambda^2(b_n - a_n)^2}{8}}$. On peut alors procéder de la même manière en conditionnant successivement par $\mathcal{F}_{n-1}, \ldots, \mathcal{F}_0$, et l'on obtient

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

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 $0 \le X_{n-1} - X_n \le 1$ et que $0 \le X_{n-1} \le N - n + 1$, on a

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

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

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

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

2. L'inégalité des différences bornées

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 $f : \mathcal{X}^n \to \mathbb{R}$. Supposons que pour tout $i \in [\![1, n]\!]$, il existe $c_i > 0$ tel que pour tous $x_1, \ldots, x_n, x_i'$,*

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

(On dit que f satisfait la condition de différences bornées). Alors pour tout $t > 0$,

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

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

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

Ainsi

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

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

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

Nous verrons en Section 6 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

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

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

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

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 est inconnue (on conjecture que γ vaut $\frac{2}{1+\sqrt{2}}$). 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 :

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

Exemple 3.5 (Le nombre chromatique d'un graphe aléatoire)**.** Soit $\chi(G_{n,p})$ le nombre chromatique d'un graphe aléatoire d'Erdös–Renyi, i.e. le nombre minimal de couleurs nécessaires pour pouvoir colorier les sommets de telle sorte que deux sommets voisins aient toujours des couleurs différentes. Par un célèbre résultat de Bollobás [3],

$$\frac{2\log_q(n)\chi(G_{n,p})}{n}\xrightarrow{\mathbf{P}}1\,,$$

où $q=1-p$.

Exemple 3.6 (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

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

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

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

Cherchons maintenant à appliquer plus finement l'inégalité d'Azuma–Hoeffding pour obtenir une inégalité exponentielle. En notant $\mathcal{F}_i = \sigma(X_1,\dots,X_i)$ 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. 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. *Supposons que $\langle Z\rangle_n \leq V$ 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

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

■

Chapitre 4

La méthode entropique

1. Entropie de Shannon

Soit $(\mathcal{X}, \mathcal{E}, \mu)$ un espace mesuré, avec μ une mesure σ-finie sur $\mathcal{X}$, et soit X une variable aléatoire à valeurs dans $\mathcal{X}$, de loi $P \ll \mu$. On pose $p = \frac{\mathrm{d}P}{\mathrm{d}\mu}$ la densité de P par rapport à μ. L'entropie de P (ou indifféremment l'entropie de X) est définie comme

$$H(P) = H(X) = \mathbf{E}[-\log p(X)] = -\int_{\mathcal{X}} \log p(x) \mathrm{d}P(x) = -\int_{\mathcal{X}} p(x) \log p(x) \mathrm{d}\mu(x)\,,$$

avec $0 \log 0 = 0$. On a toujours $H(X) \geq 0$. Dans le cas où l'ensemble $\mathcal{X}$ est dénombrable avec μ la mesure de comptage, on a

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

avec $p(x) = \mathbf{P}(X = x)$. Notons que 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

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

avec $X \sim P$. Si Q n'est pas absolument continue par rapport à P, on pose $D(Q \mid P) = +\infty$. Par l'inégalité de Jensen, on voit facilement que $D(Q \mid P) \geq 0$ et que $D(Q \mid P) = 0$ si et seulement si $P = Q$. Remarquons que $D(Q \mid P)$ ne dépend pas du choix de la mesure dominante μ.

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 $P_{(X,Y)} \ll \mu \otimes \mu$, 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.

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

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

$$H(X \mid Y) = H(X, Y) - H(Y)\,.$$

On a

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

Cela montre que $H(X) \geq H(X \mid Y)$: 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 $Q \ll \mu^{\otimes n}$. 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

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

■

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

$$D(Q \mid P) \leq \sum_{i=1}^{n} \left(D(Q \mid P) - D(Q^{(i)} \mid P^{(i)}) \right)\,.$$

Preuve du Corollaire 4.2. Notons $p = p_1 \dots p_n$, q, $p^{(i)}$ et $q^{(i)}$ les densités 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

$$\begin{aligned} \int_{\mathcal{X}^n} \log p(x) \mathrm{d}Q(x) &= \frac{1}{n} \sum_{i=1}^{n} \int_{\mathcal{X}^n} \left(\log p_i(x_i) + \log p^{(i)}(x^{(i)}) \right) \mathrm{d}Q(x) \\ &= \frac{1}{n} \int_{\mathcal{X}^n} \log p(x) \mathrm{d}Q(x) + \frac{1}{n} \sum_{i=1}^{n} \int_{\mathcal{X}^{n-1}} \log p^{(i)}(x^{(i)}) \mathrm{d}Q^{(i)}(x^{(i)})\,. \end{aligned}$$

Ainsi

$$\int_{\mathcal{X}^n} \log p(x) \mathrm{d}Q(x) = \frac{1}{n-1} \sum_{i=1}^n \int_{\mathcal{X}^{n-1}} \log p^{(i)}(x^{(i)}) \mathrm{d}Q^{(i)}(x^{(i)}) .$$

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

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

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

1.5. Entropie fonctionnelle. Soit $f : \mathcal{X}^n \to \mathbb{R}_+$ une fonction positive et $Z = f(X_1, \ldots, X_n)$ avec $(X_1, \ldots, X_n)$ de loi P. Si $\mathbf{E}[Z \log Z] < +\infty$, on définit l'entropie fonctionnelle de f sous P (ou indifféremment de Z) par

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

Si $\mathbf{E}Z = 1$, on peut définir une autre mesure de probabilité Q par

$$\mathrm{d}Q = f \mathrm{d}P .$$

On a alors $Q \ll P$ et

$$\mathbf{Ent}[Z] = \mathbf{E}[Z \log Z] = \int_{\mathcal{X}^n} \log \left(\frac{\mathrm{d}Q}{\mathrm{d}P} \right) \mathrm{d}Q = D(Q \mid P) .$$

Si $P \ll \mu^{\otimes n}$ est une loi produit de densité $p = p_1 \ldots p_n$, alors Q est de densité $q = fp$ et l'on a

$$q^{(i)}(X^{(i)}) = \int_{\mathcal{X}} q(X_1, \ldots, X_{i-1}, y, X_{i+1}, \ldots, X_n) \mathrm{d}\mu(y) = p^{(i)}(X^{(i)}) \mathbf{E}^{(i)}[Z] ,$$

où $\mathbf{E}^{(i)} = \mathbf{E}[\cdot \mid X^{(i)}]$. Ainsi

$$D(Q^{(i)} \mid P^{(i)}) = \mathbf{E} \left[\mathbf{E}^{(i)}[Z] \log \mathbf{E}^{(i)}[Z] \right] ,$$

et le Corollaire 4.2 se traduit alors par

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

où $\mathbf{Ent}^{(i)}[Z] = \mathbf{E}^{(i)} \left[Z \log Z - \mathbf{E}^{(i)}[Z] \log \mathbf{E}^{(i)}[Z] \right]$ est l'entropie conditionnelle de Z sachant $X^{(i)}$. Quitte à considérer la variable $Z/\mathbf{E}Z$, on voit que l'inégalité ci-dessus reste valable même quand $\mathbf{E}Z \neq 1$. Cette inégalité correspond à ce qu'on appelle la sous-additivité de l'entropie. Nous allons voir dans la section suivante que cette propriété est valable en général, et pas seulement dans le cadre à densité considéré jusqu'ici.

2. Sous-additivité de l'entropie

Soit Z une variable aléatoire positive telle que $\mathbf{E}[Z \log Z] < +\infty$. On définit comme ci-dessus l'entropie (fonctionnelle) de Z comme

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

On note que si $Q \ll P$, alors $D(Q \mid P) = \mathbf{Ent}\left[\frac{\mathrm{d}Q}{\mathrm{d}P}\right]$.

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.

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 $\mathbf{E}Z = 1$,

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

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

$$\mathbf{Ent}[Z] - \mathbf{E}\left[Z \log T\right] = \int_\Omega Z(\omega) \log\left(\frac{Z(\omega)}{T(\omega)}\right) \mathrm{d}\mathbf{P}(\omega) = \int_\Omega \frac{Z(\omega)}{T(\omega)} \log\left(\frac{Z(\omega)}{T(\omega)}\right) \mathrm{d}\mathbf{Q}(\omega) \,,$$

qui correspond à l'entropie de Z/T sous la loi $\mathbf{Q}$ et est donc positive. 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 $Z = f(X_1, \ldots, X_n)$ avec $X_1, \ldots, X_n$ 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 $\mathbf{E}_i = \mathbf{E}[\cdot \mid X_1, \ldots, X_i]$ avec $\mathbf{E}_0 = \mathbf{E}$. On peut alors écrire

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

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

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

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

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

■

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*

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

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. L'argument de Herbst

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 $Z = f(X_1, \ldots, X_n)$ 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 $\mathbf{Ent}[e^{\lambda Z}]$. C'est l'argument de Herbst. 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 (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 $\lambda \geq 0$,

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

Preuve de la Proposition 4.6. Posons $\psi(\lambda) = \log \mathbf{E}e^{\lambda(Z-\mathbf{E}Z)}$. On vérifie facilement que

$$\frac{\mathbf{Ent}[e^{\lambda Z}]}{\mathbf{E}e^{\lambda Z}} = \lambda\psi'(\lambda) - \psi(\lambda) . \tag{4.1}$$

La condition se traduit alors par $\lambda\psi'(\lambda) - \psi(\lambda) \leq \frac{\lambda^2 v}{2}$, ou, de façon équivalente par $\Psi'(\lambda) \leq \frac{v}{2}$ avec $\Psi(\lambda) = \frac{\psi(\lambda)}{\lambda}$. Comme $\Psi(\lambda) \to 0$ quand $\lambda \to 0$ par la règle de l'Hôpital, on a

$$\Psi(\lambda) = \int_0^\lambda \Psi'(u)\mathrm{d}u \leq \frac{\lambda v}{2} ,$$

et l'on obtient bien $\psi(\lambda) \leq \frac{\lambda^2 v}{2}$. ■

4. Inégalité de Mc Diarmid

Comme première application de la méthode entropique, présentons une amélioration de l'inégalité des différences bornées due à McDiarmid [9].

Proposition 4.7 (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 4.7. Montrons d'abord le résultat suivant (qui correspond en fait à une autre formulation de l'inégalité de Hoeffding) : si X est une variable aléatoire prenant ses valeurs dans l'intervalle $[a, b]$, alors pour tout $\lambda \in \mathbb{R}$,

$$\frac{\mathbf{Ent}[e^{\lambda X}]}{\mathbf{E}e^{\lambda X}} \leq \frac{(b-a)^2\lambda^2}{8} . \tag{4.1}$$

En effet, par (4.1), 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

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

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

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

Ainsi

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

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 (4.1), 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. ■

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

Proposition 4.8. *Soit $f : \mathcal{X}^n \to \mathbb{R}$ et $Z = f(X_1, \ldots, X_n)$ avec $X_1, \ldots, X_n$ indépendantes. On note $Z_i = f_i(X^{(i)}) = f_i(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$ 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 4.8. 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 4.8, on peut alors améliorer la Proposition 4.7 comme suit.

Proposition 4.9. *Soit $f : \mathcal{X}^n \to \mathbb{R}$ et $Z = f(X_1, \ldots, X_n)$ avec $X_1, \ldots, X_n$ indépendantes. Notons*

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

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 $\sum_{i=1}^{n}(Z - Z_i)^2 \leq v$ avec $Z_i = \sup_{x_i \in \mathcal{X}} f(X_1, \ldots, X_{i-1}, x_i, X_{i+1}, \ldots, X_n)$, 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 4.9. 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 4.8 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

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

Ainsi,

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

et par la Proposition 4.9,

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

5. Concentration des fonctions convexes lipschitziennes

Proposition 4.10. *Soient $X_1, \ldots, X_n$ 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) et 1-lipschitzienne (i.e. pour tous $x, y \in [0,1]^n$, $|f(x) - f(y)| \le \|x - y\|$). On suppose de plus que les dérivées partielles de f existent. Alors si $Z = f(X_1, \ldots, X_n)$, pour tout $t \ge 0$,*

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

Preuve de la Proposition 4.10. Nous allons montrer que $\sum_{i=1}^n (Z - Z_i)^2 \le v$ avec $Z_i = \inf_{x_i} f(X_1, \ldots, X_{i-1}, x_i, X_{i+1}, \ldots, X_n)$ et utiliser la Proposition 4.9. 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, $\|\nabla f(X)\|^2 \le 1$. ■

6. 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 $i \in [\![1, n]\!]$,

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

et

$$\sum_{i=1}^n \left(f(x) - f_i(x^{(i)})\right) \le f(x)\,.$$

Proposition 4.11. *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 $t \geq 0$,

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

Preuve de la Proposition 4.11. Posons $Z_i = f_i(X^{(i)})$. Par la Proposition 4.8, 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 $\sum_{i=1}^{n}(Z - Z_i) \leq Z$,

$$\mathbf{Ent}[e^{\lambda Z}] \leq \phi(-\lambda)\mathbf{E}\left[Ze^{\lambda Z}\right]\,. \tag{4.1}$$

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 (4.1) 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 $\psi_1(0) = 0$ (règle de l'Hôpital), 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

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

Ainsi par la Proposition 4.11, K_n est sous-Poissonienne avec facteur de variance $\mathbf{E}K_n$. Notons que ce facteur de variance est toujours inférieur à n (le facteur de variance dans l'inégalité sous-gaussienne des différences bornées), et que dans certains cas il peut être bien plus petit. 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 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 [2]).

Chapitre 5

La méthode de transport

1. Le lemme de transport

Lemme 5.1 (Lemme de transport). *Soit Z une variable aléatoire réelle définie sur $(\Omega, \mathcal{F}, \mathbf{P})$ et soit $v > 0$. Il y a équivalence entre*

(i) pour toute probabilité $\mathbf{Q} \ll \mathbf{P}$ avec $D(\mathbf{Q} \mid \mathbf{P}) < +\infty$,

$$\mathbf{E}_{\mathbf{Q}} Z - \mathbf{E} Z \leq \sqrt{2vD(\mathbf{Q} \mid \mathbf{P})}\,,$$

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

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

Voyons comment se servir du Lemme 5.1 pour établir l'inégalité des différences bornées. Soit $(\mathcal{X}, \mathcal{E})$ un espace mesurable, $f : \mathcal{X}^n \to \mathbb{R}$ une fonction mesurable, et $Z = f(X)$ avec $X = (X_1, \ldots, X_n)$ un vecteur indépendant défini sur $(\Omega, \mathcal{F}, \mathbf{P})$ à valeurs dans $\mathcal{X}^n$ de loi $P = P_1 \otimes \cdots \otimes P_n$, i.e. $P = \mathbf{P} \circ X^{-1}$. Soit maintenant $\mathbf{Q} \ll \mathbf{P}$ avec $D(\mathbf{Q} \mid \mathbf{P}) < +\infty$ et Y une variable aléatoire définie aussi sur $(\Omega, \mathcal{F}, \mathbf{P})$ de loi $Q = \mathbf{Q} \circ X^{-1} = \mathbf{P} \circ Y^{-1}$. Remarquons que $D(\mathbf{Q} \mid \mathbf{P}) = D(Q \mid P)$ et que

$$\mathbf{E}_{\mathbf{Q}} f(X) = \int_{\Omega} f(X(\omega)) \mathrm{d}\mathbf{Q}(\omega) = \int_{\mathcal{X}^n} f(y) \mathrm{d}Q(y) = \mathbf{E}_{\mathbf{P}} f(Y) = \mathbf{E} f(Y)\,.$$

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 :

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

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.

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

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

Cela peut se vérifier en remarquant que le sup dans la définition (5.3) est atteint pour l'ensemble $A = \{x \in \mathcal{X},\, p(x) > q(x)\}$. 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, considérons le couplage (X, Y) de densité par rapport à $\mu \otimes \mu$

$$r(x,y) = p(x) \wedge q(x) \mathbb{1}_{\{x=y\}} + \frac{(p(x) - q(x))_+ (q(y) - p(y))_+}{d_{\mathrm{TV}}(P,Q)} . \tag{5.4}$$

Alors $\mathbf{P}(X \neq Y) = d_{\mathrm{TV}}(P,Q)$.

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*

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

Preuve de la Proposition 5.2.

■

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 $Y_1, \ldots, Y_{i-1}$. 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 $i = 2, \ldots, n$, si $(X_1, \ldots, X_{i-1})$ et $(Y_1, \ldots, Y_{i-1})$ ont été générées, on génère (X_i, Y_i) selon le couplage optimal des lois P_i et $Q_i^{Y_1,\ldots,Y_{i-1}}$, i.e. tel que $\mathbf{P}(X_i \neq Y_i \mid Y_1, \ldots, Y_{i-1}) = d_{\mathrm{TV}}(P_i, Q_i^{Y_1,\ldots,Y_{i-1}})$. En utilisant l'inégalité de Jensen, les propriétés du couplage, puis l'inégalité de Pinsker, on a

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

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

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

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 $x, y \in \mathcal{X}^n$,

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

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

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

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*

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

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), $\mathbf{P}(X=Y\mid X)\le\min\left\{1,\frac{q(X)}{p(X)}\right\}$. En effet, pour toute fonction mesurable positive φ,

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

Ainsi $\min_{(X,Y)\in\mathcal{C}(P,Q)}\mathbf{P}(X\neq Y\mid X)\ge\left(1-\frac{q(X)}{p(X)}\right)_+$. 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*

$$d_2^2(Q,P)+d_2^2(P,Q)\le 2D(Q\mid P).$$

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

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

Or une simple analyse de fonction montre que pour tout $z\ge 0$,

$$(1-z)_+^2+\left(1-\frac{1}{z}\right)_+^2 z\le 2(z\log z+1-z).$$

■

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 ci-dessus. 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

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

■

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

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

Alors pour tout $\lambda \geq 0$,

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

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

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

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

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

■

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

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

Proposition 5.7 (Inégalité de distance convexe de Talagrand)**.** *Pour tout $A \subset \mathcal{X}^n$ et pour tout $t \ge 0$,*

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

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

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

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 $\sum_{i=1}^n \alpha_i(x)^2 \le 1$, 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 $t \ge 0$,

$$\mathbf{P}(Z \ge t) \le e^{\frac{(\mathbf{E}Z)^2}{2} - \frac{t^2}{4}} .$$

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

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

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

Exemple 5.1 (Le voyageur de commerce)**.** Reprenons l'exemple 3.6 du voyageur de commerce. Tout d'abord, montrons par récurrence que pour tout $n \ge 1$, 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 $n \ge 1$, 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

Concentration sans indépendance

Dans ce chapitre 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 [10], et Chatterjee [6].

1. Paires de Stein

Une paire de variables aléatoires réelles (W, W') est dite échangeable si $(W, W') \sim (W', W)$. Si de plus il existe $a \in]0, 1[$ tel que $\mathbf{E}[W' - W \mid W] = -aW$, alors la paire (W, W') est une a-paire de Stein. Le théorème ci-dessous est dû à Chatterjee [5].

Proposition 6.1. *Soit (W, W') une a-paire de Stein avec $\mathbf{Var}(W) = \sigma^2 < \infty$. Supposons qu'il existe $b, c \geq 0$ tels que*

$$\mathbf{E}\left[(W' - W)^2 \mid W\right] \leq 2a\,(bW + c)\ .$$

Alors, pour tout $t > 0$,

$$\mathbf{P}\,(W > t) \leq \exp\left\{-\frac{t^2}{2(c + bt)}\right\} \qquad et \qquad \mathbf{P}\,(W < -t) \leq \exp\left\{-\frac{t^2}{2c}\right\}\ .$$

Preuve de la Proposition 6.1. Soit $m(\lambda) = \mathbf{E}\left[e^{\lambda W}\right]$. Montrons que

$$m'(\lambda) = \frac{\mathbf{E}\left[(W' - W)(e^{\lambda W'} - e^{\lambda W})\right]}{2a}\ .$$

Par échangeabilité, on a

$$\begin{aligned}
\mathbf{E}\left[(W' - W)(e^{\lambda W'} - e^{\lambda W})\right] &= 2\left(\mathbf{E}\left[We^{\lambda W}\right] - \mathbf{E}\left[W'e^{\lambda W}\right]\right) \\
&= 2\left(\mathbf{E}\left[We^{\lambda W}\right] - \mathbf{E}\left[(1-a)We^{\lambda W}\right]\right) \\
&= 2a\mathbf{E}\left[We^{\lambda W}\right] \\
&= 2am'(\lambda)\,.
\end{aligned}$$

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 a alors, pour tout $\lambda \in \mathbb{R}$,

$$\begin{aligned}|m'(\lambda)| &\leq \frac{|\lambda|}{4a}\mathbf{E}\left[(W'-W)^2(e^{\lambda W'}+e^{\lambda W})\right]\\ &= \frac{|\lambda|}{2a}\mathbf{E}\left[(W'-W)^2 e^{\lambda W}\right]\\ &\leq |\lambda|\mathbf{E}\left[(bW+c)e^{\lambda W}\right]\\ &= |\lambda|\left(bm'(\lambda)+cm(\lambda)\right) .\end{aligned}$$

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 $t > 0$, $\mathbf{P}(W > t) \leq \exp\left(-\frac{t^2}{2(c+bt)}\right)$. 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. ∎

2. 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

$$X = \sum_{i=1}^n a_{i,\pi(i)} .$$

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 6.2. *Supposons que les poids $a_{i,j}$ appartiennent tous à $[0,1]$, et soit $X = \sum_{i=1}^n a_{i,\pi(i)}$ où π est une permutation uniforme. Alors pour tout $t \geq 0$,*

$$\mathbf{P}\left(X - \mathbf{E}X \geq t\right) \leq \exp\left\{-\frac{t^2}{4\mathbf{E}X + 2t}\right\} \quad \text{et} \quad \mathbf{P}\left(X - \mathbf{E}X \leq -t\right) \leq \exp\left\{-\frac{t^2}{4\mathbf{E}X}\right\} .$$

Preuve de la Proposition 6.2. Soit π une permutation uniforme de $\{1,\ldots,n\}$ et I, J deux entiers tirés uniformément et indépendamment dans $\{1,\ldots,n\}$. 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 $0 \leq a_{i,j} \leq 1$,

$$\begin{aligned}
\mathbf{E}\left[(W' - W)^2 \mid \pi\right] &= \mathbf{E}\left[\left(a_{I,\pi(J)} + a_{J,\pi(I)} - a_{I,\pi(I)} - a_{J,\pi(J)}\right)^2 \mid \pi\right] \\
&= \frac{1}{n^2}\sum_{i,j}\left(a_{i,\pi(j)} + a_{j,\pi(i)} - a_{i,\pi(i)} - a_{j,\pi(j)}\right)^2 \\
&\leq \frac{2}{n^2}\sum_{i,j}\left(a_{i,\pi(j)} + a_{j,\pi(i)} + a_{i,\pi(i)} + a_{j,\pi(j)}\right)^2 \\
&= \frac{4(X + \mathbf{E}X)}{n} = \frac{4W}{n} + \frac{8\mathbf{E}X}{n}\,.
\end{aligned}$$

Ainsi, en appliquant la Proposition 6.1 avec $b = 1$ et $c = 2\mathbf{E}X$, on obtient bien le résultat voulu.
■

Bibliographie

[1] A. D. Barbour and L. H. Chen. Steins (magic) method. *arXiv preprint arXiv :1411.1179*, 2014.

[2] A. Ben-Hamou, S. Boucheron, M. I. Ohannessian, et al. Concentration inequalities in the infinite urn scheme for occupancy counts and the missing mass, with applications. *Bernoulli*, 23(1) :249–287, 2017.

[3] B. Bollobás. The chromatic number of random graphs. *Combinatorica*, 8(1) :49–55, 1988.

[4] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities*. Oxford University Press, Oxford, 2013. ISBN 978-0-19-953525-5. doi : 10.1093/acprof:oso/9780199535255.001.0001. URL http://dx.doi.org/10.1093/acprof:oso/9780199535255.001.0001.

[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] D. P. Dubhashi and A. Panconesi. *Concentration of measure for the analysis of randomized algorithms*. Cambridge University Press, 2009.

[8] J. Kahn, G. Kalai, and N. Linial. *The influence of variables on Boolean functions*. Citeseer, 1989.

[9] C. McDiarmid. Concentration. In *Probabilistic methods for algorithmic discrete mathematics*, pages 195–248. Springer, 1998.

[10] N. Ross. Fundamentals of Stein's method. *Probab. Surv*, 8 :210–293, 2011.

[11] R. Vershynin. *High-dimensional probability : An introduction with applications in data science*, volume 47. Cambridge university press, 2018.