Probabilités

Chapitre 2 : Théorie de l'information

Lucie Le Briquer

Sommaire

1 Introduction et motivation

Pour l'instant on a toujours considéré une somme de v.a. réelles indépendantes. On aimerait faire un peu pareil pour les mesures. Reprenons $X_1, ..., X_n$ v.a. indépendantes à valeurs dans $\{1, ..., r\}$ de loi μ.

À chaque X_i, on peut associer une mesure aléatoire qui est S_{X_i}.

S_{X_i} est une v.a. à valeurs dans l'espace des mesures sur $\{1, ..., r\}$.

$$\begin{aligned} S_{X_i} : \Omega &\longrightarrow \mathcal{P}(\{1, ..., r\}) \\ \omega &\longrightarrow S_{X_i(\omega)} \end{aligned}$$

$$\begin{aligned} \mathbb{E}(S_{X_i}) &= \sum (\text{réalisation de } S_{X_i}) \times \mathbb{P}(\text{réalisation}) \\ &= S_1 \mathbb{P}(X_i = 1) + ... + S_r \mathbb{P}(X_i = r) \\ &= \mu \end{aligned}$$

Prenons $\mu_n = \frac{1}{n} \sum_{k=1}^{n} S_{X_k}$ somme de v.a. indépendantes de même loi.

Si on voulait étudier la déviation de μ_n autour de μ, il nous faudrait une sorte de distance qui nous permettrait de comparer les mesures. Mais dans un premier temps, comment comprendre une mesure ?

Exemple.
X de loi uniforme sur $\{1, ..., 32\}$. Si on veut deviner le numéro choisi par X, on a besoin de 5 questions binaires (5 bits), en faisant une dichotomie.

$$H(X) = -\sum_{i=1}^{32} p(i) \log_2(p(i)) = -\log_2\left(\frac{1}{32}\right) = 5$$

Et si on n'avait pas la loi uniforme ?
Par exemple, on a une course à 8 chevaux, ayant une proba de gagner chacun $\{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\}$.
Si on procède comme avant on a besoin de 3 questions. Ici on a intérêt à demander si le premier cheval a gagné d'abord.
Le nombre de bits moyen est $1 \times \frac{1}{2} + 2 \times \frac{1}{4} + 3 \times \frac{1}{8} + 4 \times \frac{1}{16} + 6 \times 4 \times \frac{1}{64} = 2$
Or :

$$-\sum_{i=1}^{8} p(i) \log_2(p(i)) = 2$$

Est-ce une coïncidence ? D'où vient le log ?

Shannon a voulu modéliser l'information donnée par un évènement ; ceci est lié à sa probabilité. Ainsi, Shannon a voulu associer à chaque évènement E une fonction $h(E)$ qui dépend de $\mathbb{P}(E)$ et qui donne l'information découlant de la réalisation de cet évènement.

- $h(E)$ doit être décroissante en $\mathbb{P}(E)$; plus un évènement est récurrent moins sa réalisation ramène de l'information
- $h(E) = 0$ lorsque $\mathbb{P}(E) = 1$; puisque si on sait que E est vrai alors sa réalisation ne nous rapporte aucune information
- si E et F indépendants on doit avoir $h(E \cap F) = h(E) + h(F)$

La fonction $h(E) = \log\left(\frac{1}{\mathbb{P}(E)}\right)$ vérifie ces propriétés. h est l'information d'un évènement E.

2 Entropie dans le cas discret

Définition 1 (entropie de Shannon)

X v.a. à valeurs dans un ensemble dénombrable K de loi $\mathbb{P}(X = x) = p_x \ \forall x \in K$.
L'entropie de Shannon (ou juste entropie) est définie par :

$$H(X) = \mathbb{E}(-\ln p(X)) = \sum_{x \in K} p_x \ln\left(\frac{1}{p_x}\right)$$

avec la convention $0 \times \ln(0) = 0$

Remarques.

- On a vu que l'entropie approche le nombre de bits pour décrire la v.a.
- Plus l'entropie est grande, plus il y a de l'incertitude ; ainsi $\mathcal{B}\left(\frac{1}{2}\right)$ est la Bernouilli ayant la plus grande entropie
- On a choisi le log en base e au lieu du log en base 2. Cela ne change rien, toutes les entropies sont proportionnelles.

$$H(X) = \ln(2) H_2(X) \quad \text{où } H_2 \text{ est l'entropie en base 2}$$

- On a toujours $H(X) \geq 0$ car $0 \leq p_x \leq 1$

Définition 2 (entropie relative)

Soient P et Q deux probabilités sur un ensemble dénombrable K et soient p, q leurs fonctions de masse. On définit l'*entropie relative* entre P et Q (ou *distance de Kullback-Leibler*) par :

$$\mathcal{D}(P||Q) = \sum_{x \in K} p(x) \ln\left(\frac{p(x)}{q(x)}\right)$$

si P est absolument continue par rapport à Q et $+\infty$ sinon.
(avec la convention $0 \times \ln\left(\frac{0}{q}\right) = 0$ et $p \ln\left(\frac{p}{0}\right) = +\infty$)

Remarques.

- $\mathcal{D}$ n'est pas vraiment une distance, elle n'est pas symétrique et ne vérifie pas l'inégalité triangulaire
- $\mathcal{D}$ mesure l'erreur de supposer que la loi d'une v.a. est q alors qu'en réalité elle est p

Théorème 1 (inégalité de l'information)

On a toujours $\mathcal{D}(P||Q) \geq 0$ avec égalité ssi $p(x) = q(x)\ \forall x \in K$

Preuve.

Soit $S = \text{Supp}(P) = \{x \in K | p(x) > 0\}$

$$-\mathcal{D}(P||Q) = -\sum_{x \in S} p(x) \ln\left(\frac{p(x)}{q(x)}\right) = \sum_{x \in S} p(x) \ln\left(\frac{q(x)}{p(x)}\right) \leq \ln(\sum_{x \in S} p(x) \frac{q(x)}{p(x)}) \leq \ln 1 = 0$$

Si $p(x) = q(x) \forall x \in K$ alors $\mathcal{D}(P||Q) = 0$

Pour la réciproque, on sait que log est strictement concave, ainsi on a égalité si $\frac{q(x)}{p(x)} = \text{cste} \forall x \in S$ et $S = K$ donc $\Rightarrow \frac{q(x)}{p(x)} = \text{cste} \forall x \in K \Rightarrow q(x) = p(x) \forall x \in K$

□

Corollaire 2

X à valeurs dans un ensemble dénombrable K. Alors :

$$H(X) \leq \ln |K| \quad \text{avec égalité ssi } X \text{ suit la loi uniforme}$$

Preuve.

Soit Q loi uniforme et P la loi de X. On sait que $\mathcal{D}(P||Q) \geq 0$

$$\mathcal{D}(P||Q) = \sum_{x \in K} p(x) \ln \frac{p(x)}{q(x)} = \sum_{x \in K} p(x) \ln(p(x)) - \sum_{x \in K} p(x) \ln(\underbrace{q(x)}_{=\frac{1}{|K|}})$$

Donc $\mathcal{D}(P||Q) = -H(X) + \ln(|K|) \geq 0$ donc $H(X) \leq \ln |K|$
avec égalité ssi X suit la loi uniforme par le théorème précédent.

□

Définition 3 (entropie jointée)

X, Y v.a. discrètes à valeurs dans K et L respectivement. L'*entropie jointée* $H(X,Y)$ est l'entropie du couple (X,Y).

L'*information mutuelle* entre X et Y est l'entropie relative du couple $\mathbb{P}_{(X,Y)}$ et du produit des lois marginales $\mathbb{P}_X \otimes \mathbb{P}_Y$ sur $K \times L$

$$I(X,Y) = \mathcal{D}(\mathbb{P}_{(X,Y)} || \mathbb{P}_X \otimes \mathbb{P}_Y) = \sum_{x \in K \ y \in L} \mathbb{P}_{(X,Y)}(x,y) \ln \left(\frac{\mathbb{P}_{(X,Y)}(x,y)}{\mathbb{P}_X(x) \mathbb{P}_Y(y)} \right)$$

Remarques.

- $I(X,Y) = H(X) + H(Y) - H(X,Y)$
- $I(X,Y) \geq 0 \Rightarrow H(X,Y) \leq H(X) + H(Y)$ avec égalité ssi X et Y sont indépendantes ; c'est la propriété de sous-additivité de l'entropie de Shannon
- $I(X,Y)$ représente l'information qu'on gagne sur X connaissant Y
- $I(X,Y) = I(Y,X)$

Plus tard, on définira les entropies conditionnelles $H(X|Y) = H(X,Y) - H(Y)$

3 Entropie : cas général

On note $\phi(x) = x \ln x$ définie sur $[0; +\infty[$ avec la convention $\phi(0) = 0$

Définition 4 (entropie fonctionnelle)

Étant donné f une fonction positive et μ une mesure sur $\mathbb{R}$. Alors :

$$\mathrm{Ent}_\mu(f) = \int \phi(f) d\mu - \phi\left(\int f d\mu\right) = \int f \ln f d\mu - \left(\int f d\mu\right) \ln\left(\int f d\mu\right)$$

Remarques.

- Si on ne spécifie pas la mesure, cela sous-entend que c'est par rapport à la mesure de Lebesgue
- $\mathrm{Ent}_\mu(f) \geq 0$ car ϕ convexe + Jensen

Définition 5 (entropie d'une v.a. ≥ 0)

$(\Omega, \mathcal{A}, P)$ un espace de probabilité et Y une v.a. positive telle que $\mathbb{E}(Y) < +\infty$. On définit :

$$\mathrm{Ent}(Y) = \mathbb{E}(\phi(Y)) - \phi(\mathbb{E}(Y)) = \mathbb{E}(Y \ln Y) - \mathbb{E}(Y) \ln(\mathbb{E}(Y))$$

Remarques.

- $\mathrm{Ent}(Y) \geq 0$ par Jensen
- $\mathrm{Ent}(Y) < +\infty$ ssi $\mathbb{E}(\phi(Y)) < +\infty$
- $\mathrm{Ent}(\mathrm{cste}) = 0$
- $\mathrm{Ent}(\lambda Y) = \lambda \mathrm{Ent}(Y) \quad \forall \lambda \geq 0$
- Dans la littérature, on définit l'entropie d'une v.a. comme étant l'entropie fonctionnelle de sa densité

Définition 6 (entropie relative)

L'entropie relative de Q par rapport à P est donnée par $\mathcal{D}(Q||P) = \mathrm{Ent}(Y)$ si Q absoluement continue par rapport à P et $+\infty$ sinon.

Où Y est obtenue comme suit :

- Y de loi P et $\mathbb{E}(Y) = 1$
- $\forall\, A \in \mathcal{A}, \quad Q(A) = \mathbb{E}(Y 1_A)$, on écrit alors $Q = YP$

Remarque.

En fait, ce qu'on est en train de dire est que si $Q \ll P$ alors par Radon-Nikodym, $\exists f \geq 0$ tel que $\forall A \in \mathcal{A} \quad Q(A) = \int_A f dP$ et on définit alors $\mathcal{D}(Q||P)$ comme l'entropie fonction de f par rapport à P i.e. $\text{Ent}_P(f)$

Remarque.

Ceci généralise le cas discret. Si Ω dénombrable, P et Q deux probabilités sur Ω :

$$Y(\omega) = \begin{cases} \frac{q(\omega)}{p(\omega)} & \text{si } p(\omega) > 0 \\ 0 & \text{sinon} \end{cases}$$

On retrouve donc la définition qu'on avait avant.

Théorème 2 (formule de dualité de l'entropie)

1. Y v.a. positive sur $(\Omega, \mathcal{A}, P)$ telle que $\mathbb{E}(\phi(Y)) < +\infty$. Alors :

$$\text{Ent}(Y) = \sup_{u \in \mathcal{U}} \mathbb{E}(uY)$$

où $\mathcal{U} = \{u \text{ v.a. sur } \Omega \text{ tq } \mathbb{E}(e^u) = 1\}$

2. D'autre part, si u est telle que $\mathbb{E}(uY) \leq \text{Ent}(Y)$ alors $\mathbb{E}(e^u) \leq 1$

Remarque.

On peut tout énoncer pour l'entropie fonctionnelle :

1. $\forall f \geq 0$ et $\forall \mu$ mesure sur $\mathbb{R}$

$$\text{Ent}_\mu(f) = \sup_{g \in \mathcal{U}} \int f g d\mu$$

où $\mathcal{U} = \{g \text{ fonction sur } \mathbb{R} \text{ tq } \int e^g d\mu = 1\}$

2. pareil

Preuve.

1. Ici on a pu faire ce changement car $e^u P$ et aussi une probabilité puisque $\mathbb{E}_P(e^u) = 1$.

   Pour l'égalité, on prend u telle que $\frac{Y}{\mathbb{E}(Y)} = e^u$ comme $\mathbb{E}(e^u) = \frac{\mathbb{E}(Y)}{\mathbb{E}(Y)} = 1$ alors $u \in \mathcal{U}$ et $\frac{e^{-u}Y}{\mathbb{E}(Y)} = 1$.

   Donc $\text{Ent}_{e^{-u}P}(e^{-u}\frac{Y}{\mathbb{E}(Y)}) = 0$ et du coup $\mathbb{E}(uY) = \text{Ent}(Y)$.

   Si $\mathbb{E}(uY) \leq \text{Ent}(Y) \quad \forall Y \geq 0, \mathbb{E}(Y) = 1$ alors $\mathbb{E}(e^u) \leq 1$

2. Soit u telle que $\mathbb{E}(uY) \leq \underset{=\mathbb{E}(Y\ln Y)}{\mathrm{Ent}(Y)} \quad \forall Y \geq 0, \mathbb{E}(Y) = 1$

$$Y = \frac{e^u}{\mathbb{E}(e^u)} \sim \frac{\mathbb{E}(ue^u)}{\mathbb{E}(e^u)} \leq \mathbb{E}(\frac{e^u}{\mathbb{E}(e^u)} \ln \frac{e^u}{\mathbb{E}(e^u)}) \sim \mathbb{E}(ue^u) \leq \mathbb{E}(e^u \ln e^u) - \mathbb{E}(e^u) \ln \mathbb{E}(e^u)$$

$$\Rightarrow \mathbb{E}(e^u) \ln(\mathbb{E}(e^u)) \leq 0 \sim \mathbb{E}(e^u) \leq 1$$

Ici il manquerait que $\phi(Y)$ est intégrable mais rien ne garantit ça du coup on rend $Y_n = \frac{e^{\min(u,n)}}{\mathbb{E}(e^{\min(u,n)})}$ pour lequel $\mathbb{E}(\phi(Y_n)) < +\infty$. On utilise le même raisonnement pour déduire $\forall n, \mathbb{E}(e^{\min(u,n)}) \leq 1$, par le théorème de convergence monotone $\Rightarrow \mathbb{E}(e^u) \leq 1$

□

Remarque.

On peut reformuler la conclusion :

$$\mathrm{Ent}(Y) = \sup_{T \in \{\text{v.a.} \geq 0\}} \mathbb{E}(Y[\log T - \log(\mathbb{E}(T))])$$

en écrivant $e^u = \frac{T}{\mathbb{E}(T)}$

Corollaire 3

P et Q deux probabilités sur Ω. Alors :

$$\mathcal{D}(Q||P) = \sup_Z \left[\mathbb{E}_Q(Z) - \log(\mathbb{E}_P e^Z)\right]$$

où le sup est pris sur les Z tels que $\mathbb{E}_P(e^Z) < +\infty$

Preuve.

Si Q n'est pas absolument continue par rapport à $P \quad \Rightarrow \exists A \in \mathcal{A}$ tel que $P(A) = 0$ mais $Q(A) > 0$

prenons $\forall n \; Z_n = n1_A, \quad \mathbb{E}_Q(Z_n) = Q(A) > 0, \quad \mathbb{E}_P(e^{Z_n}) = 1$

$$e^{Z_n} = e^{1_A n} \sim \mathbb{E}_Q(Z_n) - \log_P(\mathbb{E}(e^{Z_n})) = nQ(A) \underset{n\to+\infty}{\longrightarrow} +\infty$$

si $Q \ll P$, $\mathcal{D}(Q||P) = \mathrm{Ent}\left(\frac{dQ}{dP}\right) = \sup_T \mathbb{E}\left(\frac{dQ}{dP}[\underbrace{\log T}_{=Z} - \underbrace{log(\mathbb{E}(T))}_{=\mathbb{E}_e(Z)}]\right)$

□

Définition 7 (transformée de Legendre)

Si f est une fonction, la transformée de Legendre est :

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

Pour X v.a., on note $\Lambda_X(\lambda) = \log M_X(\lambda)$

Corollaire 8

Z v.a. centrée de loi P tq $M_Z(\lambda) < +\infty \quad \forall \lambda$. Alors :

$$\forall t > 0, \quad \Lambda_Z^*(t) = \inf_Q \{\mathcal{D}(Q||P), \mathbb{E}_Q(Z) \geq t\}$$

Preuve.

$$\lambda_Z * (t) = \sup_{\lambda \in \mathbb{R}} \{\lambda t - \Lambda_Z \lambda\}$$

Montrons d'abord que $\sup_{\lambda \in \mathbb{R}} \{\lambda t - \Lambda_Z(\lambda)\} \leq \inf_{Q:\mathbb{E}_Q(Z) \geq t} \{\mathcal{D}(Q||P)\}$

Soit $\lambda \in \mathbb{R}$, et Q tq $\mathbb{E}_Q(Z) \geq t$

$$\mathcal{D}(Q||P) = \sup_{Y;\mathbb{E}_P(e^Y)<+\infty} [\mathbb{E}_Q(Y) - \log(\mathbb{E}_P(e^Y))] \quad \text{corollaire précédent}$$

Donc :

$$\mathcal{D}(Q||P) \geq \lambda \mathbb{E}_Q(Z) - \log(\mathbb{E}_P(e^{\lambda Z})) \geq \lambda t - \Lambda_Z(t)$$

Il reste à montrer l'égalité. Prenons $\frac{dQ}{dP} = \frac{e^{\lambda Z}}{\mathbb{E}(e^{\lambda Z})}$ avec $\lambda = \Lambda_Z^{'-1}(t)$

Il faut que $\mathbb{E}_Q(Z) \geq t$; mais $\mathbb{E}_Q(Z) = \mathbb{E}_P\left(\frac{e^{\lambda Z} Z}{\mathbb{E}(e^{\lambda Z})}\right) = \Lambda_Z'(\lambda) = t$

$$\mathcal{D}(Q||P) = \text{Ent}\left(\frac{dQ}{dP}\right) = \frac{1}{\mathbb{E}(e^{\lambda Z})}[\mathbb{E}(e^{\lambda Z}\lambda Z) - \mathbb{E}(e^{\lambda Z})\ln(\mathbb{E}(e^{\lambda Z}))] = \lambda \Lambda_Z'(\lambda) - \Lambda_Z(\lambda) = \lambda t - \Lambda_Z(\lambda)$$

□

Remarques.

- $\mathbb{E}_Q(Z) \geq t \Leftrightarrow Q \in \Gamma = \{\mu | \mathbb{E}_\mu(Z) \geq t\}$, $Q = \delta_Z$ donc $\delta_Z \in \Gamma$ veut dire $Z \geq t$
- Rappelons nous le phénomène de concentration du premier chapitre. Inégalité de Chernoff :

  $$\mathbb{P}(X \geq t) \leq \inf_{\lambda > 0} \{\exp(-\lambda t) M_X(\lambda)\} = \exp(-\sup_{\lambda > 0} \{\lambda t \Lambda_X(\lambda)\}) = \exp(-\Lambda_X^*(t))$$

  Ainsi $Z \geq t$ correspond à $\delta_Z \in \Gamma$ et $\Lambda_Z^*(t)$ correspond à $\inf\{\mathcal{D}(Q||P); Q \in P\}$ qui correspond en gros à $d(P, \Gamma)$
- Reprenons l'exemple du début de chapitre. $X_1, ..., X_n$ indépendantes de loi P, notons $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_n}$ alors :

  $$\frac{1}{n}\sum_{i=1}^n X_i \geq t \Leftrightarrow P_n \in \Gamma$$

  Avant on avait tiré avantage du fait que la concentration était contrôlée par $M_{S_n}(\lambda) = \prod_{i=1}^n M_{X_i}(\lambda)$ où $S_n = \sum_{i=1}^n X_i$. Ici la concentration dépendra de $d(P_n, \Gamma)$ ou encore de $\mathcal{D}(Q||P_n)$ pour $Q \in \Gamma$. On aimerait que l'entropie satisfasse des propriétés de tensorisation comme M_X.

Cours du 10 mars

Remarque.
Pour toute fonction positive f, et toute mesure μ

$$\mathrm{Ent}_\mu f = \int f \ln f d\mu - \int f d\mu \ln \int f d\mu$$

Si μ est une proba alors $\mathrm{Ent}_\mu f \geq 0$.
Dans la littérature on définit l'entropie d'une densité f ($\int f d\mu = 1$) par $-f \ln f d\mu$.

4 Conditionnement

4.1 Cas discret

Définition 8 (probabilité conditionnelle)

$(\Omega, \mathcal{A}, \mathbb{P})$ un espace de probabilité. Soit $B \in \mathcal{A}$ tq $\mathbb{P}(B) \geq 0$. On définit la probabilité conditionnelle sachant B par :

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \quad \forall A \in \mathcal{A}$$

Remarque.
Si X est intégrable, $\mathbb{E}[X|B] = \frac{\mathbb{E}(X.1_B)}{\mathbb{P}(B)}$. On peut définir l'espérance conditionnelle d'une v.a. X sachant une autre v.a. Y. Si Y est à valeurs dans un espace dénombrable E et $E' = \{y \in E \mid \mathbb{P}(Y = y) > 0\}$, on définit :

$$\mathbb{E}[X|Y = y] = \frac{\mathbb{E}(X.1_{\{Y=y\}})}{\mathbb{P}(Y = y)} \quad y \in E'$$

Définition 9 (espérance conditionnelle)

$X \in L^1(\Omega, \mathcal{A}, \mathbb{P})$. L'espérance conditionnellle de X sachant Y est la v.a. définie par :

$$\mathbb{E}[X|Y] = \varphi(Y)$$

Où :

$$\varphi(y) = \begin{cases} \mathbb{E}[X|Y = y] & \text{si } y \in E' \\ 0 \quad \text{sinon} \end{cases}$$

Exemple.
$\Omega = \{1, ..., 6\}, \mathbb{P}(\{i\}) = \frac{1}{6}$

$$Y(\omega) = \begin{cases} 1 \text{ si } \omega \text{ est impair} \\ 0 \text{ sinon} \end{cases}$$

Pour $X(\omega) = \omega$:

$$\mathbb{E}[X|Y=1] = \frac{\mathbb{E}(X.1_{Y=1})}{\mathbb{P}(Y=1)} = \frac{(1+3+5)\frac{1}{6}}{\frac{1}{2}} = 3$$

$$\mathbb{E}[X|Y=0] = \frac{\mathbb{E}(X.1_{Y=0})}{\mathbb{P}(Y=0)} = \frac{(2+4+6)\frac{1}{6}}{\frac{1}{2}} = 4$$

Remarques.

- $\mathbb{E}[X|Y]$ est donc une v.a. qui est $\sigma(Y)-$mesurable
- on peut changer la définition de φ sur $E \backslash E'$, on obtiendrait des v.a. p.s. égales car $E \backslash E'$ est négligeable
- $Z = \mathbb{E}[X|Y]$; $Z(y) = \mathbb{E}[X|Y=y]$

Proposition 8

- On a toujours $\mathbb{E}|\mathbb{E}[X|Y]| \leq \mathbb{E}|X|$ (donc $\mathbb{E}[X|Y] \in L^1(\Omega, \mathcal{A}, \mathbb{P})$)
- Si Z v.a. $\sigma(Y)-$mesurable (bornée) alors :

$$\mathbb{E}(ZX) = \mathbb{E}[Z\ \mathbb{E}[X|Y]]$$

Preuve.

- Pour le premier point :

$$\begin{aligned} \mathbb{E}|\mathbb{E}[X|Y]| &= \sum_{y \in E'} \mathbb{P}(Y=y).|\mathbb{E}[X|Y=y]| \\ &= \sum_{y \in E'} \mathbb{P}(Y=y).\frac{|\mathbb{E}(X.1_{Y=y})|}{\mathbb{P}(Y=y)} \\ &= \sum_{y \in E'} |\mathbb{E}(X.1_{Y=y})| \\ &\leq \mathbb{E}|X| \end{aligned}$$

– $Z = \psi(Y)$ avec ψ bornée

$$\begin{aligned}
\mathbb{E}[\psi(Y).\mathbb{E}[X|Y]] &= \sum_{y\in E'} \mathbb{P}(Y=y).\psi(y).\mathbb{E}[X|Y=y] \\
&= \sum_{y\in E'} \mathbb{P}(Y=y).\psi(y).\frac{\mathbb{E}(X.1_{Y=y})}{\mathbb{P}(Y=y)} \\
&= \sum_{y\in E'} \psi(y).\mathbb{E}(X.1_{Y=y}) \\
&= \mathbb{E}\left(\sum_{y\in E'} X.\psi(y).1_{Y=y}\right) \\
&= \mathbb{E}(X\underbrace{\psi(Y)}_{Z})
\end{aligned}$$

□

4.2 Plus généralement

Théorème 9 (et définition)

$X \in L^1(\Omega, \mathcal{A}, P)$ et $\mathcal{B}$ sous-tribu de $\mathcal{A}$. Il existe une *unique* v.a. $\in L^1(\Omega, \mathcal{B}, P)$ notée $\mathbb{E}[X|\mathcal{B}]$ telle que :

$$\forall B \in \mathcal{B},\ \mathbb{E}[X.1_B] = \mathbb{E}[\mathbb{E}[X|\mathcal{B}].1_B]$$

Ou aussi $\forall Z$ $\mathcal{B}-$mesurable on a $\mathbb{E}(XZ) = \mathbb{E}[\mathbb{E}[X|\mathcal{B}]Z]$

Si Y v.a., on définit $\mathbb{E}[X|Y] = \mathbb{E}[X|\sigma(Y)]$

Preuve.

1. Existence :
   Supposons que $X \geq 0$ (sinon on écrit $X = X^+ - X^-$ et puis on refait la même chose pour X^+ et X^-). On définit Q une mesure sur $\mathcal{B}$ comme :

   $$\forall B \in \mathcal{B},\ Q(B) = \mathbb{E}[X.1_B]$$

   On a $Q \ll P$ sur $(\Omega, \mathcal{B})$. Alors d'après Radon-Nikodym $\exists \widetilde{X}$ $\mathcal{B}-$mesurable tq :

   $$\forall B \in \mathcal{B},\ Q(B) = \mathbb{E}[\widetilde{X}.1_B]$$

   On a $\widetilde{X} \in L^1(\Omega, \mathcal{B}, P)$ (en prenant $B = \Omega$ et utilisant que $X \in L^1$). On prend $\mathbb{E}[X|\mathcal{B}] = \widetilde{X}$

2. Unicité : soit X' et X'' $\in L^1(\Omega, \mathcal{B}, P)$ telles que $\forall$ $B \in \mathcal{B}$ $\mathbb{E}[X'.1_B] = \mathbb{E}[X''.1_B]$
   Prenons $B = \{X' > X''\}$ $\mathcal{B}-$mesurable car X' et X'' le sont, alors :

   $$\mathbb{E}(X' - X'')1_{\{X'-X''>0\}} = 0 \quad \Rightarrow \quad P(X' - X'' > 0) = 0$$

   et idem pour $X'' - X'$ donc $X' = X''$

□

Propriété 10

1. Si X est $\mathcal{B}-$mesurable, $\mathbb{E}[X|\mathcal{B}] = X$
2. $X \longrightarrow \mathbb{E}[X|\mathcal{B}]$ est linéaire
3. $\mathbb{E}[\mathbb{E}[X|\mathcal{B}]] = \mathbb{E}(X)$
4. Si $X \geq X'$ alors $\mathbb{E}[X|\mathcal{B}] \geq \mathbb{E}[X'|\mathcal{B}]$
5. X et Y sont indépendantes ssi $\forall h,\ \mathbb{E}[h(X)|Y] = \mathbb{E}(h(X))$
6. On a Jensen pour tout fonction f convexe positive : $f(\mathbb{E}[X|Y]) \leq \mathbb{E}[f(X)|Y]$

Pour finir, énonçons le cadre des variables dans L^2 qui donnera une bonne intuition.

Théorème 11

Si $X \in L^2(\Omega, \mathcal{A}, P)$ alors $\mathbb{E}[X|\mathcal{B}]$ est la projection orthogonale de X sur $L^2(\Omega, \mathcal{B}, P)$

Preuve.
$\mathbb{E}\left((\mathbb{E}[X|\mathcal{B}])^2\right) \leq \mathbb{E}[\mathbb{E}[X^2|\mathcal{B}]] < +\infty$ ainsi $\mathbb{E}[X|\mathcal{B}] \in L^2(\Omega, \mathcal{B}, P)$
$\mathbb{E}[X|\mathcal{B}]$ est la projection orthogonale de X sur $L^2(\Omega, \mathcal{B}, P)$ veut dire que pour tout $Z \in L^2(\Omega, \mathcal{B}, P)$ on a :

$$Z \perp X - \mathbb{E}[X|\mathcal{B}] \ \Rightarrow\ \mathbb{E}(Z.(X - \mathbb{E}[X|\mathcal{B}])) = 0 \ \Rightarrow\ \mathbb{E}(ZX) = \mathbb{E}[Z.\mathbb{E}[X|\mathcal{B}]]$$

qui est la définition de l'espérance conditionnelle.

□

5 Sous-additivité de l'entropie

Le théorème suivant sera la clé pour établir des inégalités de concentration.

Théorème 12 (sous-additivité de l'entropie)

$X_1, ..., X_n$ v.a. indépendantes, $Y = f(X_1, ..., X_n)$ une fonction mesurable tq $\phi(Y) = Y \ln Y$ est intégrable
Pour tout $i \leq n$, $\mathbb{E}^{(i)}$ est l'espérance conditionnelle par rapport à :

$$X^{(i)} = (X_1, ..., X_{i-1}, X_{i+1}, ..., X_n)$$

($\mathbb{E}^{(i)} = \mathbb{E}[\ .\ |X_1, ..., X_{i-1}, X_{i+1}, ..., X_n]$).

$$\mathrm{Ent}^{(i)} = \mathbb{E}^{(i)}\phi(Y) - \phi(\mathbb{E}^{(i)}Y)$$

on a alors $\mathrm{Ent}(Y) \leq \mathbb{E}(\sum_{i=1}^n \mathrm{Ent}^{(i)}Y)$

Preuve.
Notons $\mathbb{E}_i[.] = \mathbb{E}[.|X_1...X_i]$ donc $\mathbb{E}_0 = \mathbb{E}$ et $\mathbb{E}_n Y = Y$

$$\text{Ent}Y = \mathbb{E}Y \ln Y - \mathbb{E}Y \ln \mathbb{E}Y = \mathbb{E}(Y \ln Y - Y \ln \mathbb{E}Y) = \mathbb{E}(Y(\ln Y - \ln \mathbb{E}Y))$$

$$\ln Y - \ln \mathbb{E}Y = \ln \mathbb{E}_n Y - \ln \mathbb{E}_0 Y = \sum_{i=1}^{n} \ln(\mathbb{E}_i Y) - \ln(\mathbb{E}_{i-1} Y)$$

Donc :

$$Y(\ln Y - \ln \mathbb{E}Y) = \sum_{i=1}^{n} Y(\ln(\mathbb{E}_i Y) - \ln(\mathbb{E}_{i-1} Y))$$

Formule de dualité de l'entropie $\rightarrow$ $\text{Ent}(Y) = \sup_{T \geq 0} \mathbb{E}(Y(\ln T - \ln \mathbb{E}T))$

Prenons $T = \mathbb{E}_i T \rightarrow \text{Ent}^{(i)}(Y) \geq \mathbb{E}^{(i)} Y(\ln(\mathbb{E}_i Y) - \ln(\mathbb{E}^{(i)} \mathbb{E}_i Y))$

$$\begin{aligned} \mathbb{E}^{(i)} \mathbb{E}_i Y &= \mathbb{E}(i)\mathbb{E}[Y|X_1, ..., X_i] \\ &= \mathbb{E}[\mathbb{E}[Y|X_1, ..., X_i]|X_1, ..., X_{i-1}, X_{i+1}, ..., X_n] \end{aligned}$$

Z est une fonction de $X_1, ..., X_i$ et les $(X_j)_j$ sont indépendants donc Z est indépendante de $X_{i+1}, ..., X_n$.

$$\begin{aligned} \rightarrow \mathbb{E}^{(i)} \mathbb{E}_i Y &= \mathbb{E}[\mathbb{E}[Y|X_1, ..., X_i]|X_1, ..., X_{i-1}] \\ &= \mathbb{E}[Y|X_1, ..., X_{i-1}] = \mathbb{E}_{i-1} Y \end{aligned}$$

$$\begin{aligned} \rightarrow \mathbb{E}^{(i)} \mathbb{E}_i Y &\geq \mathbb{E}^{(i)} Y(\ln \mathbb{E}_i Y - \ln \mathbb{E}_{i-1} Y) \\ \mathbb{E}\left(\sum_{i=1}^{n} \text{Ent}^{(i)} Y\right) &\geq \mathbb{E}\left(\sum_{i=1}^{n} \mathbb{E}^{(i)} Y(\ln \mathbb{E}_i Y - \ln \mathbb{E}_{i-1} Y)\right) \\ &\geq \mathbb{E}\left(\sum_{i=1}^{n} Y(\ln \mathbb{E}_i Y - \ln \mathbb{E}_{i-1} Y)\right) \qquad = \text{Ent}Y \end{aligned}$$

□

6 Inégalités de Sobolev Logarithmique

Définition 10 (inégalité de Sobolev Logarithmique)

Soit μ mesure de probabilité sur $\mathbb{R}^n$. On dit que μ satisfait une inégalité de Sobolev Logarithmique (ISL(c)) avec constante $c > 0$ si :

$$\forall f : \mathbb{R}^n \longrightarrow \mathbb{R} \quad \text{ on a } \quad \text{Ent}_\mu(f^2) \leq c \int |\nabla f|^2 d\mu$$

De même, si X est un vecteur aléatoire de $\mathbb{R}^2$ de loi μ, il satisfait une ISL(c) si:

$$\text{Ent}(f^2(X)) \leq c\mathbb{E}(\|\bigtriangledown f\|_2^2(X))$$

Remarques.

- Plus généralement, on peut définir ISL pour un espace quelconque muni d'une proba μ, quitte à trouver une bonne notion de gradient et $|.|$
- Une inégalité du même esprit est l'inégalité de Poincaré où la conclusion est :

$$\mathrm{Var}_\mu(f) \leq c \int |\nabla f|^2 d\mu$$

  Où $\mathrm{Var}_\mu(f) = \int f^2 d\mu - (\int f d\mu)^2$
- On a log-Sobolev $\Rightarrow$ Poincaré

6.1 Mesure produit Bernoulli

Posons $\Omega_n = \{-1, 1\}^n$ le cube discret dans $\mathbb{R}^n$. Soit σ_n la mesure uniforme sur Ω_n. On a $\forall x \in \Omega_n, \sigma_n(\{x\}) = \frac{1}{2^n}$. On peut voir σ_n comme la mesure produit $\sigma_n = \sigma^{\bigotimes n}$, où $\sigma(\{1\}) = \sigma(\{-1\}) = \frac{1}{2}$ mais aussi comme la loi du vecteur aléatoire $\xi = (\xi_1, ..., \xi_n)$ où les ξ_i sont i.i.d (Rademacher).

On aimerait définir une ISL pour σ_n. Définissons d'abord une métrique:

$$\forall x, y \in \Omega_n, d(x, y) = |\{i \leq n, x_i \neq y_i\}| = \sum_{i=1}^{n} 1_{\{x_i \neq y_i\}}$$

On appelle cette distance la *distance de Hamming*. Elle mesure le nombre d'arêtes à traverser sur le cube pour passer de x à y.
Notons τ_i le *flip* de la i^{ieme} coordonnée : $\tau_i(x) = (x_1, ..., x_{i-1}, -x_i, x_{i+1}, ..., x_n)$. On a :

$d(x, y) = k \Leftrightarrow$ il y a k flips à effectuer pour passer de x à $y \Leftrightarrow \exists i_1, ...i_k$ distincts t.q $y = \tau_{i_k} \circ ... \circ \tau_{i_1}(x)$

On dira que x et y sont *voisins* et on note $x \sim y$ si $\exists i \in \mathbb{N}$ t.q $y = \tau_i(x)$.
Définissons maintenant la norme. Pour $f : \Omega_n \to \mathbb{R}$ et $x \in \Omega_n$, posons:

$$\|\bigtriangledown f\|^2 (x) = \frac{1}{2} \sum_{\text{y où } y \sim x} (f(y) - f(x))^2$$

$$(\bigtriangledown f. \bigtriangledown g)(x) = \frac{1}{2} \sum_{\text{y où } y \sim x} (f(y) - f(x))(g(y) - g(x))$$

($\bigtriangledown$ est une notation)

Soit $x \in \Omega_n$ et $i \leq n$. On note $\hat{x_i} = (x_1, ..., x_{i-1}, x_{i+1}, ..., x_n) \in \Omega_{n-1}$. Fixons $\hat{x_i}$. Soit $f : \Omega_n \to \mathbb{R}$.
On définit $f_{\hat{x_i}} : \Omega \to \mathbb{R}, x \mapsto f_{\hat{x_i}}(x) = f(x_1, ..., x_{i-1}, x, x_{i+1}, ..., x_n)$.

Pour $x \in \Omega_n$, on pose :

$$\|\bigtriangledown_i f_i\|^2(x) = \|\bigtriangledown f_{\hat{x_i}}\|^2(x) = \frac{1}{2}(f_{\hat{x_i}}(1) - f_{\hat{x_i}}(-1))^2 = \frac{1}{2}(f(x) - f(\tau_i(x)))^2$$

Ainsi, on a $\|\bigtriangledown f\|^2(x) = \sum_{i=1}^n \|\bigtriangledown_i f_i\|^2(x)$. (Comme pour le gradient usuel).

Théorème 13

(Ω_n, σ_n) satisfait ISL_1 i.e

$$\forall g : \Omega_n \to \mathbb{R}, \quad \mathrm{Ent}_{\sigma_n}(g^2) \leq \int \|\bigtriangledown g\|^2(x) d\sigma_n(x)$$

i.e si $\xi = (\xi_1, ..., \xi_n)$ où les ξ_i sont des Rademacher, $\mathrm{Ent}(g^2(\xi)) \leq \mathbb{E}(\|\bigtriangledown g\|^2(\xi))$.

Preuve.

Soit $g : \Omega_n \to \mathbb{R}$. Par sous-additivité de l'entropie, on a :

$$\mathrm{Ent}_{\sigma_n}(g^2) \leq \sum_{i=1}^n \int \mathrm{Ent}_{\sigma_1}(g^2_{\hat{x_i}}) d\sigma_n$$

Notons $g_{\hat{x_i}} = g_i$ pour simplifier les notations.
D'autre part, on a $\|\bigtriangledown g\|^2(x) = \sum_{i=1}^n \|\bigtriangledown_i g_i\|^2(x)$.
Ainsi, il suffit de montrer que:

$$\sum_{i=1}^n \int \mathrm{Ent}_{\sigma_1}(g_i^2) d\sigma_n \leq \int \sum_{i=1}^n \|\bigtriangledown_i g_i\|^2(x) d\sigma_n$$

Montrons que : $\forall i \in \{1, .., n\}, \int \mathrm{Ent}_{\sigma_1}(g_i^2) d\sigma_n \leq \int \|\bigtriangledown_i g_i\|^2(x) d\sigma_n$.
Le problème se réduit ainsi à montrer l'ISL_1 en dimension 1 (pour (Ω, σ_1)).
Soit $g : \Omega \to \mathbb{R}$. On a :

$$\int \|\bigtriangledown g\|^2(x) d\sigma_1 = \frac{1}{2}(g(1) - g(-1))^2 = \frac{1}{2}(a-b)^2$$

$$\mathrm{Ent}_\sigma(g^2) = \int \Phi(g^2) d\sigma - \Phi\left(\int g^2 d\sigma\right) = \frac{1}{2}a^2 \ln a^2 + \frac{1}{2}b^2 \ln b^2 - \frac{a^2+b^2}{2} \ln\left(\frac{a^2+b^2}{2}\right)$$

Ainsi, il faut montrer que :

$$\forall a, b \in \mathbb{R}, (a-b)^2 \geq a^2 \ln a^2 + b^2 \ln b^2 - (a^2+b^2) \ln\left(\frac{a^2+b^2}{2}\right)$$

L'inégalité étant symétrique en a et b, supposons que $a \geq b$. Or, on a $||a| - |b|| \leq |a-b|$. Il nous suffit donc de montrer que :

$$\text{pour } a \geq b \geq 0, (|a| - |b|)^2 \geq a^2 \ln a^2 + b^2 \ln b^2 - (a^2+b^2) \ln\left(\frac{a^2+b^2}{2}\right)$$

Fixons $b \geq 0$. Posons $h(a) = a^2 \ln a^2 + b^2 \ln b^2 - (a^2+b^2) \ln(\frac{a^2+b^2}{2}) - (a-b)^2$. On a :

- $h(b) = 0$
- $h'(b) = 0$
- h est concave

Donc h est négative. D'où le résultat.

□

6.2 ISL-gaussienne

Théorème 14

Soit γ_n la mesure gaussienne sur $\mathbb{R}^n$(de densité $\frac{1}{(2\pi)^{\frac{n}{2}}} \exp(-\frac{|x|^2}{2})$). Alors, γ_n (ou $(\mathbb{R}^n, \gamma_n,$ norme euclidienne)) vérifie(nt) ISL_2
i.e $\forall f : \mathbb{R}^n \to \mathbb{R}$ continuement différentiable, on a :

$$\mathrm{Ent}_{\gamma_n}(f^2) \leq 2 \int \|\bigtriangledown f\|_2^2 \, d\gamma_n$$

i.e si g est un vecteur gaussien ($g \sim \mathcal{N}(0, I_{d\mathbb{R}^n})$), alors:

$$\mathrm{Ent}(f^2(g)) \leq 2\mathbb{E}\left(\int \|\bigtriangledown f\|^2 (g) \right)$$

Preuve.
Comme dans la preuve précédente, il suffit de montrer l'inégalité en dimension 1. Dans l'exercice 2 du TD 5, on a montré que : $\forall f : \mathbb{R} \to \mathbb{R}$ uniformément bornée et pour $\xi_1, ..., \xi_n$ des Rademacher i.i.d, on a :

$$\limsup_n \sum_{i=1}^n \mathbb{E}\left[\left(f\left(\tilde{S_n} + \frac{1-\xi_i}{\sqrt{n}} \right) - f\left(\tilde{S_n} - \frac{1+\xi_i}{\sqrt{n}} \right) \right)^2 \right] = 4\mathbb{E}(f'(X)^2)$$

où $\tilde{S_n} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i$ et $X \sim \mathcal{N}(0,1)$.
Ainsi, d'après l'ISL discret, on a, en posant $g(\xi_1, .., \xi_n) = f(\tilde{S_n})$:

$$\begin{aligned}
\mathrm{Ent}(g^2(\xi)) &= \mathrm{Ent}(f^2(\tilde{S_n})) \\
&\leq \mathbb{E}(\|\bigtriangledown g\|^2 (\xi)) \\
&\leq \frac{1}{2}\mathbb{E}(\sum_{i=1}^n \left(\left(g(\xi_1, ..., \xi_{i-1}, 1, \xi_{i+1}, ..., \xi_n) - g(\xi_1, ..., -1, ..., \xi_n) \right)^2 \right) \\
&\leq \frac{1}{2}\mathbb{E}\left(\sum_{i=1}^n \left(f\left(\tilde{S_n} + \frac{1-\xi_i}{\sqrt{n}} \right) - f\left(\tilde{S_n} - \frac{1+\xi_i}{\sqrt{n}} \right) \right)^2 \right)
\end{aligned}$$

Enfin, d'après le TCL, on a :

$$\mathrm{Ent}(f^2(\tilde{S_n})) \to \mathrm{Ent}(f^2(X))$$

où $X \sim \mathcal{N}(0,1)$. D'où le résultat.

□