Chapitre 2

Entropie

2.1 Information, entropie

2.1.1 Définitions

Si les $(A_i)_{i\in I}$ forment une partition mesurable dénombrable de $(\Omega, \mathcal{F}, \mathbb{P})$ par des ensembles de mesure non nulle, alors la fonction d'information I_A est la fonction qui vaut $-\log \mathbb{P}(A_i)$ sur les éléments qui sont dans A_i :

$$I_A(x) = \sum_{i\in I} -\log \mathbb{P}(A_i) \mathbb{1}_{\mathbb{P}(A_i)}$$

Si on a des partitions dénombrables $A = (A_i)_{i\in I}$ et $B = (B_i)_{i\in I}$, on peut encore considérer la partition engendrée par A et B : $A \vee B = (A_i \cap B_j)_{(i,j)\in(I\times}$. Si les deux partitions sont indépendantes, on a

$$I_{A\vee B}(x) = I_A(x) + I_B(x) \tag{2.1}$$

Dans le cas où $(X_n)_{n\geq 1}$ est une suite de variables aléatoires indépendantes de même loi μ à support dans I dénombrable et que A_n est la partition engendrée par les valeurs prises par la variable X_n, on a alors

$$I_{A_1\vee\dots\vee A_n}(x) = \sum_{k=1}^{n} I_{A_i}(x).$$

C'est une somme de variables aléatoires indépendantes : si $h(\mu) = -\sum_i \mu(i) \log \mu(i) < +\infty$, d'après la loi forte des grands nombres, on a presque sûrement et dans $L^1(\mathbb{P})$

$$\lim_{n\to+\infty} \frac{1}{n} I_{A_1\vee\dots\vee A_n}(x) = \mathbb{E}(I_{A_1}) = h(\mu) \tag{2.2}$$

Ceci nous amène à définir (cette fois dans le cas général), l'entropie de la partition A

$$\mathcal{H}(A) = \mathbb{E}(I_A) = \sum_{i\in A} -\log(\mathbb{P}(A_i))\mathbb{P}(A_i) = \sum_{i\in A} h(\mathbb{P}(A_i)),$$

avec $h(x) = -x\log x$. Cette fonction reviendra souvent. Il est important de noter que h est concave.

Maintenant, si $\mathcal{G}$ est une sous-tribu de $\mathcal{F}$, on définit l'information conditionnelle de la partition en remplaçant $\mathbb{P}$ par $\mathbb{P}(\cdot|\mathcal{G})$ dans la définition de l'entropie :

$$I_{A|\mathcal{G}}(x) = \sum_{i\in I} -\log \mathbb{P}(A_i|\mathcal{G})\mathbb{1}_{A_i}$$

Quant à l'entropie conditionnelle, c'est la valeur moyenne de l'entropie conditionnelle :

$$\begin{aligned} H(A|\mathcal{G}) &= \mathbb{E}(I_{A|\mathcal{G}}) = \mathbb{E}(\mathbb{E}(I_{A|\mathcal{G}}|\mathcal{G})) \\ &= \mathbb{E}\left(\sum_{i\in A} -\log(\mathbb{P}(A_i|\mathcal{G})\mathbb{P}(A_i|\mathcal{G}))\right) \\ &= \sum_{i\in A} \mathbb{E}\left(\varphi(\mathbb{P}(A_i|\mathcal{G}))\right). \end{aligned}$$

Bien noter que bien que ce soit une quantité qualifiée de conditionnelle, $H(A|\mathcal{G})$ n'est pas une variable aléatoire ; c'est une constante.

Si $\mathcal{H} \subset \mathcal{G}$, on a

$$\mathbb{E}\left(\varphi(\mathbb{P}(A_i|\mathcal{G}))\right) = \mathbb{E}\left(\mathbb{E}\left(\varphi(\mathbb{P}(A_i|\mathcal{G}))\,|\mathcal{H}\right),\right.$$

cependant, l'inégalité de Jensen conditionnelle nous dit que

$$\mathbb{E}\left(\varphi(\mathbb{P}(A_i|\mathcal{G}))|\mathcal{H}\right) \le \varphi(\mathbb{E}(\mathbb{P}(A_i|\mathcal{G})|\mathcal{H})) = \varphi(\mathbb{P}(A_i|\mathcal{H})).$$

En intégrant et en sommant sur i, on a

$$\mathcal{H} \subset \mathcal{G} \Longrightarrow H(A|\mathcal{H}) \ge H(A|\mathcal{G}).$$

En particulier, prenant pour $\mathcal{H}$ la tribu triviale, on obtient $H(A) \ge H(A|\mathcal{G})$.

2.1.2 Continuité croissante de l'information, de l'entropie

Si $(\mathcal{F}_n)_{n\ge 0}$ est une filtration croissante, avec $\mathcal{F}_\infty = \sigma(\cup_{n\ge 1}\mathcal{F}_n)$, la suite $H(A|\mathcal{F}_n)$ est décroissante. On peut conjecturer que $H(A|\mathcal{F}_n)$ converge vers $H(A|\mathcal{F}_n)$.

En réalité, sous l'hypothèse que $H(A) < +\infty$, on peut l'obtenir comme corollaire d'un résultat plus général.

Lemme 4. *On a presque sûrement*

$$\lim_{n\to+\infty} I_{A|\mathcal{F}_n} = I_{A|\mathcal{F}_\infty}.$$

Démonstration. Soit A_i un élément de la partition A : sur A_i, on a pour tout n : $I_{A|\mathcal{F}_n} = -\log \mathbb{P}(A_i|\mathcal{F}_n)$. D'après le théorème de convergence des martingales $\mathbb{P}(A_i|\mathcal{F}_n)$ converge presque sûrement vers $\mathbb{P}(A_i|\mathcal{F}_\infty)$; par continuité du log, $I_{A|\mathcal{F}_n}$ converge vers $-\log \mathbb{P}(A_i|\mathcal{F}_\infty) = I_{A|\mathcal{F}_\infty}$. □

Lemme 5. *Posons $g = \sup_{n\ge 0} I_{A|\mathcal{F}_n}$. Si la partitition A est finie, alors g a des moments exponentiels. Si $H(A) < +\infty$, alors $\mathbb{E}(g) < +\infty$.*

Démonstration. Fixons $i \in I$ et posons $g_i = \sup_{n\ge 0} -\log \mathbb{P}(A_i|\mathcal{F}_n)$. Soit $N_t^i = \inf\{n \ge 0; -\log \mathbb{P}(A_i|\mathcal{F}_n) > t\}$. On a

$$\begin{aligned}
\mathbb{P}(A_i, N_t^i = n, g > t) = \mathbb{P}(A_i, N_t^i = n, g_i > t) &= \mathbb{E}(\mathbb{P}(A_i, N_t^i = n, g_i > t|\mathcal{F}_n)) \\
&= \mathbb{E}(\mathbb{1}_{\{N_t^i=n, g_i>t\}}\mathbb{P}(A_i|\mathcal{F}_n)) \le \mathbb{E}(\mathbb{1}_{\{N_t^i=n, g_i>t\}}e^{-t}) \\
&\le \mathbb{E}(\mathbb{1}_{\{N_t^i=n\}}e^{-t})
\end{aligned}$$

En sommant sur n on obtient $\mathbb{P}(A_i, g > t) \le e^{-t}$. Si la partition A est finie, en sommant sur i, on obtient tout de suite l'intégrabilité de g (g a même des moments exponentiels).

Dans le cas général, il faut raffiner un peu : on a

$$\begin{aligned}
\mathbb{E}[g\mathbb{1}_{A_i}] &= \int_0^{+\infty} \mathbb{P}(g\mathbb{1}_{A_i} > t)\ d\lambda(t) \\
&\le (-\log \mathbb{P}(A_i))\mathbb{P}(A_i) + \int_{-\log \mathbb{P}(A_i)}^{+\infty} \mathbb{P}(A_i, g > t)\ dt \\
&\le (-\log \mathbb{P}(A_i))\mathbb{P}(A_i) + \int_{-\log \mathbb{P}(A_i)}^{+\infty} e^{-t}\ dt \\
&= (-\log \mathbb{P}(A_i))\mathbb{P}(A_i) + \mathbb{P}(A_i)
\end{aligned}$$

En sommant sur i, on obtient $\mathbb{E}(g) \le H(A) + 1$. □

Corollaire 2. *Si $H(A) < +\infty$, alors $I_{A|\mathcal{F}_n}$ converge presque sûrement et dans L^1 vers $I_{A|F_\infty}$. En particulier* $\lim_{n\to+\infty} H(A|\mathcal{F}_n) = H(A|\mathcal{F}_\infty)$.

Démonstration. Cela découle du théorème de convergence dominée, avec domination par g. □

Revenons à l'équation (2.2). Si la suite de variables aléatoires indépendantes est réalisée sur l'espace canonique $(\mathbb{R}, \mathcal{B}(\mathbb{R}^{\mathbb{N}}), T, \mathbb{P})$ avec $\mathbb{P} = \mu^{\otimes n}$, on a $X_n = \pi_0 \circ T^n$, et on a $A_n = T^{-n}A$, où A est la partition engendrée par la variable aléatoire Π_0. Ainsi, l'équation peut se réécrire comme la convergence, presque sûre et dans L^1, de la quantité

$$\frac{1}{n} I_{A\vee\cdots\vee T^{-n}(A)} \tag{2.3}$$

C'est à ce problème, pour un système dynamique général, que nous allons nous attaquer maintenant.

2.2 Addition de l'information

Pour étudier $I_{A\vee\cdots\vee T^{-n}(A)}$, on a besoin de comprendre comment s'additionnent les informations portées en raffinant les partitions. En fait, il s'agit de comprendre ce que devient l'équation (2.1) dans le cas non-indépendant.

Théorème 20. *Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace probabilisé dont A et B sont des partitions dénombrables mesurables. Alors pour toute sous-tribu $\mathcal{G}$, on a :*

$$I_{A\vee B|\mathcal{G}} = I_{A|\mathcal{G}} + I_{B|\mathcal{G}\vee\sigma(A)}$$

Démonstration. On suppose que $A = (A_i)_{i\in I}$ et $B = (B_j)_{j\in J}$. Soit $\ell \in J$. Posons

$$\Psi = \sum_{i\in I} \frac{\mathbb{P}(B_\ell \cap A_i|\mathcal{G})}{\mathbb{P}(A_i|\mathcal{G})} \mathbb{1}_{A_i}$$

On va d'abord montrer que Ψ est la probabilité conditionnelle de B_ℓ sachant $\mathcal{G} \vee \sigma(A)$. Comme Ψ est $\mathcal{G} \vee \sigma(A)$-mesurable, il suffit de montrer que pour tout j et pour tout $B' \in \mathcal{G}$, on a

$$\mathbb{E}(\Psi \mathbb{1}_{A_j} \mathbb{1}_{B'}) = \mathbb{E}(\mathbb{1}_{B_\ell} \mathbb{1}_{A_j} \mathbb{1}_{B'})$$

Comme $B' \in \mathcal{G}$, il vient $\Psi \mathbb{1}_{A_j} \mathbb{1}_{B'} = \frac{\mathbb{P}(B_\ell\cap B'\cap A_j|\mathcal{G})}{\mathbb{P}(A_j|\mathcal{G})} \mathbb{1}_{A_j}$ et le résultat s'ensuit en conditionnant par $\mathcal{G}$.

Soient (j,ℓ) tels que $\mathbb{P}(A_j \cap B_\ell) > 0$. Sur $A_j \cap B_\ell$, $I_{A \vee B|\mathcal{G}}$ vaut

$$\begin{aligned}\log \mathbb{P}(A_j \cap B_\ell|\mathcal{G}) &= \log \mathbb{P}(A_j|\mathcal{G}) + \log \frac{\mathbb{P}(B_\ell \cap A_j|\mathcal{G})}{\mathbb{P}(A_j|\mathcal{G})} \\ &= \log \mathbb{P}(A_j|\mathcal{G}) + \log \mathbb{P}(B_\ell|\mathcal{G} \vee \sigma(A)) \quad \mathbb{P}\text{-p.s.},\end{aligned}$$

ce qui donne p.s. l'identité voulue sur $A_j \cap B_\ell$. En considérant toutes les parties de mesure $\mathbb{P}(A_j \cap B_\ell)$ non nulle, cela donne p.s. l'égalité voulue. □

Corollaire 3. *Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace probabilisé dont A et B sont des partitions dénombrables mesurables. Alors pour toute sous-tribu $\mathcal{G}$, on a :*

1. $I_{A \vee B} = I_A + I_{B|\sigma(A)}$
2. $H(A \vee B|\mathcal{G}) = H(A|\mathcal{G}) + H(B|\sigma(A) \vee \mathcal{G})$
3. $H(A \vee B) = H(A) + H(B|\sigma(A))$.
4. $H(A \vee B|\mathcal{G}) \leq H(A|\mathcal{G}) + H(B|\mathcal{G})$ *et* $H(A \vee B|\mathcal{G}) \leq H(A|\mathcal{G}) + H(B|\sigma(A))$.

2.3 Entropie dynamique relative à une partition

Lemme 6. *Soit $(\Omega, \mathcal{F}, \mathbb{P}, T)$ un système dynamique mesuré dont A est une partition dénombrable mesurable. Alors pour toute sous-tribu $\mathcal{G}$ et tout entier naturel n, on a :*

$$I_{T^{-n}A|T^{-n}\mathcal{G}} = I_{A|\mathcal{G}} \circ T^n.$$

En particulier, prenant pour $\mathcal{G}$ la tribu triviale, on a $I_{T^{-n}A} = I_A \circ T^n$.

Démonstration. On suppose que $A = (A_i)_{i \in I}$. Il suffit de montrer que

$$\exp(I_{T^{-n}A|T^{-n}\mathcal{G}}) = \exp(I_{A|\mathcal{G}} \circ T^n). \tag{2.4}$$

On va les identifier sur chaque élément de la partition $T^{-n}A$: soit $i \in I$ avec $\mathbb{P}(T^{-n}(A_i)) = \mathbb{P}(A_i) \neq 0$. Par définition de l'information conditionnelle, le membre de gauche de (2.4) y vaut $\mathbb{P}(T^{-n}(A_i)|T^{-n}\mathcal{G})$. Je dis qu'il coïncide avec $r \circ T^n$, où r est une version de $\mathbb{P}(A_i|\circ G)$. Par définition de la tribu $T^{-n}\mathcal{G}$, la mesurabilité de r par rapport à $\mathcal{G}$ entraîne celle de $r \circ T^n$ par rapport à $T^{-n}\mathcal{G}$. Si B est $\sigma(T^{-n}\mathcal{G})$-mesurable, il s'écrit $B = T^{-n}(C)$ avec $C \in \mathcal{G}$. On a ainsi

$$\begin{aligned}\mathbb{E}(\mathbb{1}_B r \circ T^n) &= \mathbb{E}((\mathbb{1}_C r) \circ T^n) \\ &= \mathbb{E}(\mathbb{1}_C r) \\ &= \mathbb{E}(\mathbb{1}_C \mathbb{1}_{A_i}) \\ &= \mathbb{E}((\mathbb{1}_C \mathbb{1}_{A_i}) \circ T^n) \\ &= \mathbb{E}(\mathbb{1}_B \mathbb{1}_{T^{-n}(A_i)})\end{aligned}$$

ce qui montre qu'on a bien

$$\mathbb{P}(T^{-n}(A_i)|T^{-n}\mathcal{G}) = r \circ T^n.$$

Or pour x dans $T^{-n}(A_i)$ $T^n(x) \in A_i$; or sur A_i, r coïncide avec $\exp(I_{A|\mathcal{G}})$, ce qui donne le résultat recherché.

□

Notons $\mathcal{F}_0$ la tribu triviale et, pour $n \geq 1$, $\mathcal{F}_n$ la tribu engendrée par $A, T^{-1}(A), \dots T^{-(n-1)}(A)$.

On a

$$\begin{aligned} I_{\mathcal{F}_n} &= I_{A \vee T^{-1}\mathcal{F}_{n-1}} \\ &= I_{T^{-1}\mathcal{F}_{n-1}} + I_{A|T^{-1}\mathcal{F}_{n-1}} \quad \text{(1. du corollaire 3)} \\ &= I_{A|T^{-1}\mathcal{F}_{n-1}} + I_{\mathcal{F}_{n-1}} \circ T \quad \text{(2}^{\text{e}}\text{ point du lemme 6)} \end{aligned}$$

Ainsi, on obtient successivement

$$\begin{aligned} I_{\mathcal{F}_1} &= I_A \\ I_{\mathcal{F}_2} &= I_{A|T^{-1}\mathcal{F}_1} + I_{\mathcal{F}_1} \circ T = I_{A|T^{-1}\mathcal{F}_1} + I_A \circ T \\ I_{\mathcal{F}_3} &= I_{A|T^{-1}\mathcal{F}_2} + I_{\mathcal{F}_2} \circ T = I_{A|T^{-1}\mathcal{F}_2} + I_{A|T^{-1}\mathcal{F}_1} \circ T + I_A \circ T^2 \end{aligned}$$

et par récurrence

$$I_{\mathcal{F}_n} = \sum_{i=0}^{n-1} I_{A|T^{-1}(\mathcal{F}_i)} \circ T^{n-1-i} \tag{2.5}$$

En prenant l'espérance, on a

$$H(\mathcal{F}_n) = \sum_{i=0}^{n-1} H(A|T^{-1}(\mathcal{F}_i)).$$

Or, on sait que $H(A|T^{-1}(\mathcal{F}_i))$ converge vers $H(A|T^{-1}(\mathcal{F}_\infty))$:avec le théorème de Cesaro, on obtient :

Théorème 21. *Soit A une partition dénombrable d'entropie finie :*

$$\lim_{n\to+\infty} \frac{1}{n} H(\mathcal{F}_n) = h(A,T) = H(A|T^{-1}(\mathcal{F}_\infty)).$$

On peut maintenant démontrer de théorème de Shannon–Mc Millan–Breiman

Théorème 22. *Soit A une partition dénombrable d'entropie finie : on a presque sûrement et dans $L^1(\mathbb{P})$*

$$\lim_{n\to+\infty} \frac{1}{n} I_{\mathcal{F}_n} = \mathbb{E}(I_{A|T^{-1}(\mathcal{F}_\infty)}|\tau),$$

où τ est la tribu des invariants.

Démonstration. On revient à l'équation (2.5) : posant $g_i = I_{A|T^{-1}(\mathcal{F}_i)}$, on a $I_{\mathcal{F}_n} = \sum_{i=0}^{n-1} g_k \circ T^{n-1-i}$. Grâce au lemme 4, la suite (g_i) converge presque sûrement vers $I_{A|T^{-1}(\mathcal{F}_\infty)}$.

Vu le lemme 5, il suffit de montrer que si une suite g_k converge presque partout vers g_∞ et que $\sup_{n\ge 0} |g_n|$ est intégrable, alors $\frac{1}{n}\sum_{k=0}^{n-1} g_k \circ T^{n-1-k}$ converge presque sûrement et dans L^1 vers $\mathbb{E}(g|\tau)$.

Grâce aux théorèmes ergodiques p.s et L^1), on se ramène tout de suite au cas où $g_\infty = 0$. Posons $r_N = \sup_{k\ge N} |g_k|$. D'après le théorème de convergence dominée, r_N converge dans L^1 vers 0. Prenons N tel que $\|g_N\|_1 \le \varepsilon$. Posons $M_n = \frac{1}{n}\sum_{k=0}^{n-1} g_k \circ T^{n-1-k}$. On a

$$|M_n| \le \frac{1}{n}\sum_{k=0}^{N-1} g_k \circ T^{n-1-k} + \frac{1}{n}\sum_{k=N}^{n} r_N \circ T^{n-1-k}$$

Le lemme de Borel–Cantelli et l'intégrabité des g_k assure la convergence presque sûre vers 0 du premier terme de la somme. En appliquant le théorème ergodique, il vient

$$\overline{\lim}\ |M_n| \le \mathbb{E}(r_N|\tau),\ \text{d'où}\ \mathbb{E}(\overline{\lim}|M_n|) \le \mathbb{E}(r_N) \le \varepsilon.$$

comme ε peut être pris arbitrairement petit, on a $\mathbb{E}(\overline{\lim}|M_n|) = 0$ ce qui montre que M_n tend presque sûrement vers 0.

De même, en prenant l'espérance,

$$\mathbb{E}(|M_n|) \le \frac{1}{n}\sum_{k=0}^{N-1} \mathbb{E}(g_k) + \frac{1}{n}\sum_{k=N}^{n} \mathbb{E}(r_N),$$

d'où

$$\overline{\lim}\ \mathbb{E}(|M_n|) \le \mathbb{E}(r_N) \le \varepsilon,$$

ce qui achève la preuve.

□

2.4 Théorème de Kolmogorov-Sinaï

On note $h(T)$ le supremum de $h(A,T)$ sur toutes les partitions A d'entropie finie.

Lemme 7.

$$h(A,T) \leq h(B,T) + H(A|B)$$

Démonstration. On a

$$\begin{aligned} H(\vee_{i=0}^{n-1} T^{-i} A) &\leq H(\vee_{i=0}^{n-1} T^{-i} B) + H(\vee_{i=0}^{n-1} T^{-i} A | \vee_{i=0}^{n-1} T^{-i} B) \\ &\leq H(\vee_{i=0}^{n-1} T^{-i} B) + \sum_{j=0}^{n-1} H(T^{-j} A | \vee_{i=0}^{n-1} T^{-i} B) \\ &\leq H(\vee_{i=0}^{n-1} T^{-i} B) + \sum_{j=0}^{n-1} H(T^{-j} A | T^{-j} B) \\ &\leq H(\vee_{i=0}^{n-1} T^{-i} B) + nH(A|B) \end{aligned}$$

En divisant par n et en faisant tendre n vers l'infini, on obtient $h(A,T) \leq h(B,T) + H(A|B)$. □

Théorème 23. *Soit A une partition d'entropie finie et telle que $\mathcal{F}_\infty = \mathcal{F}$. Alors $h(A,T) = h(T)$.*

Démonstration. Soit B une autre partition d'entropie finie. Avec le lemme, on a

$$h(B,T) \leq h(\mathcal{F}_n, T) + H(B|\mathcal{F}_n)$$

Mais en réalité, $h(\mathcal{F}_n, T)$ ne dépend pas de n et coïncide avec $h(A,T)$. En effet $H(\vee_{i=0}^{k-1} T^{-i} \mathcal{F}_n) = H(\mathcal{F}_{n+k-1})$: en divisant par k et en faisant tendre n vers l'infini, on obtient $h(\mathcal{F}_n, T) = h(A,T)$, d'où

$$h(B,T) \leq h(A,T) + H(B|\mathcal{F}_n)$$

En faisant tendre n vers l'infini on obtient

$$h(B,T) \leq h(A,T) + H(B|\mathcal{F}_\infty) = h(A,T) + H(B|\mathcal{F}) = h(A,T),$$

ce qui montre bien que A réalise le supremum des entropies. □