Entropie

P. Pansu

20 mai 2012

Conventions

En bleu, les paragraphes qui n'ont pas été traités, ou alors de façon évasive.

1 Motivation

1.1 L'entropie dans les sciences

L'entropie est une invention de physiciens. Elle apparaît en thermodynamique macroscopique, comme une nécessité théorique : le second principe de la thermodynamique exprime seulement son existence (Clausius 1854). Il y a une formule pour sa variation lors de certaines transformations, mais pas d'expression directe.

La physique statistique (Boltzmann) donne de nouvelles fondations, microscopiques, à la thermodynamique, et fournit enfin une expression pour l'entropie d'un système : c'est le logarithme du nombre de configurations microscopiques possibles pour le système.

Après la seconde guerre mondiale, l'essor des communications et de nouveaux problèmes (codage, compression) donnent naissance à la théorie de l'information aux Etats-Unis. En 1948, Claude Shannon introduit l'entropie d'une distribution de probabilité finie, l'information mutuelle de deux distributions de probabilité, et la capacité d'un canal de transmission comme l'information mutuelle maximale entre les signaux d'entrée et de sortie. Il prouve que cette capacité est la bien la borne supérieure des taux de transmission autorisés par le canal. Non seulement ces notions capturent bien l'idée de quantité d'information nécessaire pour décrire une variable aléatoire, mais elles permettent de fonder l'entropie des physiciens sur des bases solides.

En URSS, Andrei Kolmogorov et Yakov Sinai définissent en 1958 l'entropie d'une transformation préservant la mesure. En 1966, Roy Adler, Alan Konheim et M. McAndrew introduisent l'entropie topologique d'une transformation continue d'un espace topologique compact, les deux notions sont très utiles en systèmes dynamiques. Kolmogorov, et, aux Etats-Unis, Ray Solomonov et Gregory Chaitin, définissent vers 1965 l'entropie d'un mot écrit dans un alphabet fini comme la longueur minimale d'un programme permettant à un ordinateur de produire ce mot. Cette notion s'avère étroitement reliée à l'entropie de Shannon.

L'entropie est une notion récente. Elle a déjà envahi beaucoup de domaines des mathématiques (probabilités, statistiques, systèmes dynamiques) et aussi de la physique et de l'informatique théorique. C'est le signe de son intérêt.

1.2 Simulation d'une distribution de probabilité

On lance un dé truqué n fois. Truqué signifie que, pour $k = 1, \dots, 6$, la probabilité $p(k)$ de sortie du chiffre k n'est pas égale à $\frac{1}{6}$. On obtient des suites $(x_1, \dots, x_n)$, $x_i \in \{1, \dots, 6\}$. Quelle est la probabilité d'obtenir une suite "typique". Par suite typique, on veut dire toutes sauf une fraction qui tend vers 0 quand n tend vers l'infini.

La probabilité d'obtenir la suite $(x_1, \dots, x_n)$ est le produit

$$p(x_1)p(x_2)\cdots p(x_n) = \prod_{k=1}^{6} p(k)^{N_k},$$

où N_k est le nombre de fois que le chiffre k apparaît dans la suite $(x_1, \dots, x_n)$. Or, d'après la loi des grands nombres, lorsque n est grand, avec forte probabilité, ce nombre N_k est proche de $np(k)$. Donc, pour la très grande majorité des tirages, la probabilité cherchée est proche de

$$\prod_{k=1}^{6} p(k)^{np(k)} = \left(\prod_{k=1}^{6} p(k)^{p(k)}\right)^n = 2^{-nH},$$

où

$$H = -\sum_{k=1}^{6} p(k)\log_2(p(k)).$$

Pour un dé non pipé, toutes les suites sont obtenues avec la même probabilité $2^{-n\log_2(6)}$. Pour un dé pipé, la répartition est inégale. La suite typique est obtenue avec une probabilité plus forte, et cette probabilité s'exprime au moyen de la quantité H. On appelle H l'entropie de la distribution de probabilité p.

Le fait que $2^{-nH} \geq 2^{-n\log_2(6)}$ ne saute pas aux yeux. C'est l'une des nombreuses identités et inégalités qui constituent les paroles de la théorie de l'information. La musique, ce sont les nombreuses interprétations qu'on peut donner des résultats dans différents champs de la science. On va commencer par établir un certain nombre d'identités et d'inégalités. On passera ensuite aux interprétations.

1.3 Bibliographie

On suit de près, et dans cet ordre, les chapitres 2, 5, 4, 3, 7, 16 du livre

Thomas Cover and Joy Thomas, *Elements of Information Theory*, John Wiley and Sons, Hoboken, NJ (2006). Cote 003.54 COV ele à la BU (rez de jardin).

Puis on passe aux livres

Vladimir Arnold et André Avez, *Problèmes ergodiques de la mécanique classique*, Gauthier-Villars, Paris (1967). Chapitre 12 et appendices 18 et 19. Ne se trouve pas à la BU, mais à la Bibliothèque Jacques Hadamard.

Karl Petersen, *Ergodic Theory*, Cambridge Studies in Advanced Mathematics, 2. Cambridge University Press, Cambridge (1989). Chapitres 5 et 6. En magasin, cote K41366 à la BU.

2 Premières propriétés

2.1 Définition

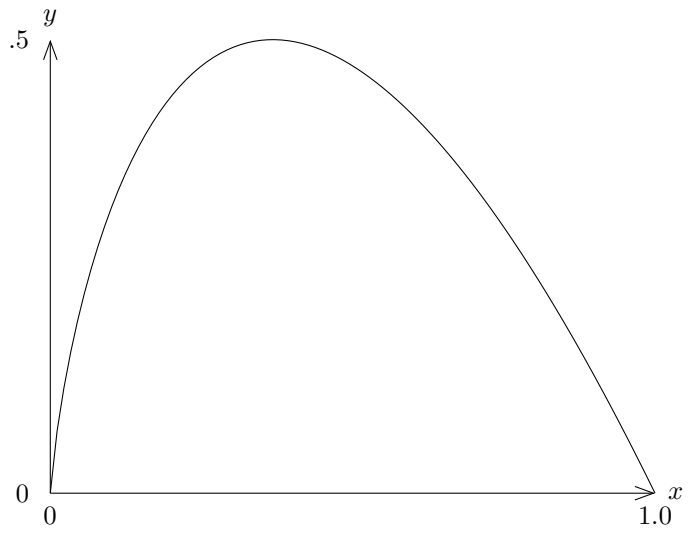
Définition 1 *Soit E un ensemble fini. Une distribution de probabilité sur E, c'est une fonction $p : E \to \mathbb{R}_+$ telle que $\sum_{x\in E} p(x) = 1$.*

*L'*entropie *de la distribution de probabilité p est le nombre*

$$H(p) = -\sum_{x\in E} p(x)\log_2(p(x)).$$

Elle est mesurée en bits.

Par convention, $0\log_2(0) = 0$. Remarquer que $H(p) \geq 0$, avec égalité si et seulement si p est concentrée sur un seul élément de E.



$y = -x \log_2(x)$

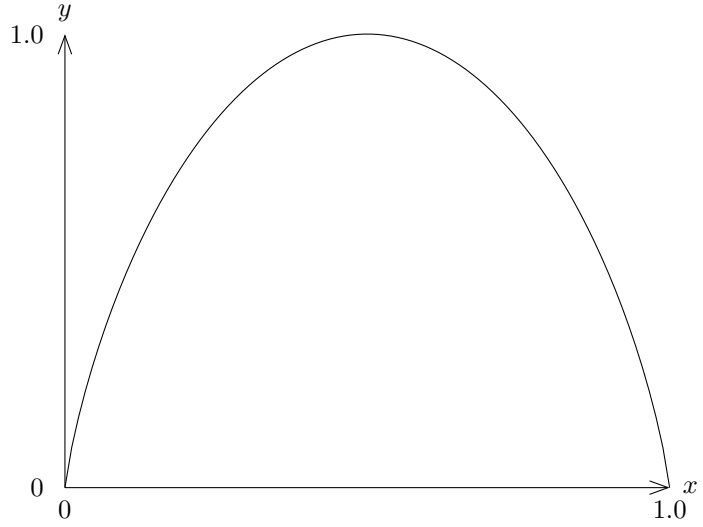
Exemple 2 *L'entropie de la distribution uniforme sur E vaut $\log_2 |E|$ bits.*

Une distribution uniforme sur un ensemble E à 2^n éléments a donc une entropie de n bits. D'ailleurs, on peut numéroter les éléments de E avec n bits seulement. Plus généralement, on verra plus tard que l'entropie est aussi le nombre moyen (au sens de la distribution p) de bits nécessaire pour coder E. C'est pourquoi on prend le logarithme en base 2 et l'unité choisie est appelée bit.

Suggestion d'exercice : n^{0}1, feuille 1.

Exemple 3 *Soit $a \in [0,1]$. Soit $E = \{0,1\}$. Soit $p_a = \mathcal{B}(a)$ la distribution de Bernoulli de paramètre a, i.e $p_a(1) = a$, $p_a(0) = 1 - a$. L'entropie de $\mathcal{B}(a)$ vaut $h(a) := -a \log_2(a) - (1 - a) \log_2(1 - a)$ bits.*

Voir courbe représentative de h.



$y = -x \log_2(x) - (1 - x) \log_2(1 - x)$

On constate que h est une fonction concave de a. On va le démontrer bientôt (Exemple 39).

2.2 Vocabulaire probabiliste

On fixe une fois pour toutes un ensemble Ω muni d'une tribu et d'une mesure de probabilité $\mathbb{P}$. Soit E un ensemble muni d'une tribu (ce sera pratiquement toujours un ensemble fini muni

de la tribu de tous ses sous-ensembles). Une variable aléatoire à valeurs dans E est une fonction mesurable $X : \Omega \to E$. La *loi* ou la *distribution* de X est la distribution de probabilité p_X sur E définie par

$$p_X(x) = \mathbb{P}(X = x).$$

L'espérance de X est le nombre

$$\mathbb{E}(X) = \int_\Omega X \, d\mathbb{P} = \int_E x \, dp_X(x) = \sum_{x \in E} x p_X(x)$$

dans le cas où E est fini.

Notation 4 *Si X est une variable aléatoire de distribution p_X, on note $H(X) = H(p_X)$, et on l'appelle l'*entropie *de X. On peut l'exprimer comme une espérance,*

$$H(X) = \mathbb{E}(-\log_2 p(X)).$$

Si on cherche à mesurer la quantité d'information $I(A)$ contenue dans l'évènement $X \in A$, $A \subset E$, la formule qui s'impose est $I(A) = \log \frac{1}{\mathbb{P}(X \in A)}$. En effet, I doit être fonction décroissante de $\mathbb{P}(X \in A)$, et pour deux évènements A et B indépendants, on souhaite que

$$I(A \cap B) = I(A) + I(B),$$

donc I est positivement proportionnelle à $\log \frac{1}{\mathbb{P}(X \in A)}$. L'entropie est donc la quantité d'information fournie en moyenne par X. Tant qu'on ne connait pas X, il s'agit plutôt d'une quantité d'incertitude.

Si X est constante, $H(X) = 0$. De toutes les variables de Bernoulli, c'est $\mathcal{B}(\frac{1}{2})$ qui est la plus incertaine.

Suggestion d'exercice : n^{0}2, feuille 1.

2.3 Entropie relative et information mutuelle

Définition 5 *Soient p et q deux distributions de probabilité sur le même ensemble E. Leur* entropie relative *est*

$$D(p||q) := \sum_{x \in E_X} p(x) \log_2 (\frac{p(x)}{q(x)}).$$

On verra bientôt qu'on peut penser à l'entropie relative comme à une sorte de distance entre p et q (elle est positive, elle est nulle seulement si $p = q$). Toutefois, elle n'est pas symétrique. Attention, elle prend la valeur $+\infty$ s'il existe x tel que $q(x) = 0$ mais $p(x) \neq 0$.

Exemple 6 *Soit $\mathcal{B}(a)$ la distribution de Bernoulli sur $\{0, 1\}$, qui met le poids a sur 1 et $1 - a$ sur 0. Alors*

$$D(\mathcal{B}(a)||\mathcal{B}(a')) = (1 - a) \log_2 (\frac{1 - a}{1 - a'}) + a \log_2 (\frac{a}{a'})$$

n'est pas symétrique en a, a'.

Attention à la notation $||$.

Définition 7 *Soient X et Y deux variables aléatoires. Leur* information mutuelle *est l'entropie relative de la loi du couple $p_{(X,Y)}$ et du produit des lois marginales $p_X \otimes p_Y$ sur $E_X \times E_Y$,*

$$\begin{aligned} I(X;Y) : &= D(p_{(X,Y)}||p_X \otimes p_Y) \\ &= \sum_{(x,y) \in E_X \times E_Y} p_{(X,Y)}(x,y) \log_2 (\frac{p_{(X,Y)}(x,y)}{p_X(x) p_Y(y)}). \end{aligned}$$

Attention au ;. On verra qu'on peut interpréter $I(X;Y)$ comme la quantité d'information sur X qu'on gagne en apprenant Y. Par définition, $I(Y;X) = I(X;Y)$, donc on apprend autant sur X en découvrant Y qu'on apprend sur Y en découvrant X.

Exemple 8 *Si X et Y sont indépendantes, leur information mutuelle est nulle.*

2.4 Inégalité de Jensen

C'est l'outil utilisé pour prouver les inégalités annoncées plus haut.

Définition 9 *Soit I un intervalle de $\mathbb{R}$. Une fonction $f : I \to \mathbb{R}$ est* convexe *si sa courbe représentative est en-dessous de toutes ses cordes. Autrement dit, si, pour tous $x, y \in I$ et $t \in [0,1]$,*

$$f((1-t)x + ty) \leq (1-t)f(x) + tf(y).$$

f est concave *si $-f$ est convexe.*

On dit que f est strictement convexe *si l'inégalité est stricte dès que $x \neq y$ et $t \in]0,1[$.*

Plus généralement, si $C \subset \mathbb{R}^n$ est un ensemble convexe, on dit qu'une fonction $f : C \to \mathbb{R}$ est convexe si sa restriction à tout segment de droite est convexe.

Proposition 10 *Si $f : I \to \mathbb{R}$ est de classe C^2 et $f'' \geq 0$, alors f est convexe. Si $f'' > 0$, alors f est strictement convexe.*

Preuve On pose $g_\epsilon(t) = f((1-t)x + ty) - (1-t)f(x) - tf(y) - \epsilon t(1-t)$. Alors $g(0) = g(1) = 0$. Par l'absurde. Si g_ϵ prend une valeur > 0, alors elle atteint son maximum en $c \in]0,1[$. En ce point, $g'_\epsilon(c) = 0$ et $g''_\epsilon(c) \leq 0$. Or $g''_\epsilon(c) = f''((1-c)x + cy) + 2\epsilon > 0$, contradiction. On conclut que $g_\epsilon \leq 0$ partout. Comme ceci est vrai pour tout $\epsilon > 0$, on conclut que $f((1-t)x+ty)-(1-t)f(x)-tf(y) \leq 0$ pour tout $t \in [0,1]$.

Lorsque $f'' > 0$ et $x \neq y$, on modifie le raisonnement par l'absurde comme suit. On pose $\epsilon = 0$. On suppose que g_0 prend une valeur ≥ 0 dans l'intervalle $]0,1[$. Elle atteint son maximum en $c \in]0,1[$. En ce point, $g''_0(c) \leq 0$. Or $g''_0(c) = (y-x)^2 f''((1-c)x + cy) > 0$, contradiction. On conclut que $g_0 < 0$ sur $]0,1[$, donc $f((1-t)x + ty) - (1-t)f(x) - tf(y) < 0$ pour tout $t \in]0,1[$. ■

Exemple 11 *$f(x) = |x|$ est convexe, x^2, 2^x sont strictement convexes, $f(x) = 1/x$, $x \log_2 x$ sont strictement convexes sur $]0, +\infty[$, $f(x) = \sqrt{x}$, $\log_2(x)$ sont strictement concaves sur $]0, +\infty[$. Les fonctions affines sont à la fois convexes et concaves sur $\mathbb{R}$.*

Proposition 12 *Soit $f : I \to \mathbb{R}$ une fonction convexe. Alors f est la borne supérieure d'une famille de fonctions affines.*

Preuve Soit $x \in I$. Comme le graphe de f passe en dessous de ses cordes, la pentes des cordes issues de x, i.e. le taux de variation $y \mapsto \Delta(y) = \frac{f(y)-f(x)}{y-x}$, est une fonction croissante. Il possède donc une limite à droite $f'(x+)$ en x. Pour $y \geq x$,

$$f(y) = f(x) + \Delta(y)(y-x) \geq f(x) + f'(x+)(y-x).$$

Pour $y \leq x$, la même inégalité a lieu ($\Delta(y) \leq f'(x+)$ mais $y - x \leq 0$). Dans les deux cas, $f \geq a_x(y)$ où a_x est la fonction affine définie par $a_x(y) = f(x) + f'(x+)(y-x)$. Il y a égalité en x. On conclut que f est la borne supérieure des fonctions affines a_x lorsque x décrit I. ■

Théorème 1 (Inégalité de Jensen) *Soit I un intervalle de $\mathbb{R}$ et $f : I \to \mathbb{R}$ une fonction convexe. Pour toute variable aléatoire X à valeurs dans I,*

$$f(\mathbb{E}(X)) \leq \mathbb{E}(f(X)).$$

Si f est strictement convexe, l'égalité entraîne que X est presque partout constante.

Remarque 13 *Lorsque X ne prend qu'un nombre fini n de valeurs, l'inégalité de Jensen n'est autre que*

$$f(\frac{\sum_{i=1}^n a_i x_i}{\sum_{i=1}^n a_i}) \leq \frac{\sum_{i=1}^n a_i f(x_i)}{\sum_{i=1}^n a_i},$$

lorsque $a_i \geq 0$ pour tout i, avec égalité si et seulement si f prend la même valeur à tous les points x_i dont le poids a_i est non nul, dans le cas où f est strictement convexe.

Preuve Bien que ce ne soit pas nécessaire, on commence par traiter le cas fini. On raisonne par récurrence sur n. Par définition, elle est vraie pour $n = 2$. Supposons l'inégalité vraie pour n points. Etant donnés $a_0, a_1, \ldots, a_n \geq 0$ et $x_0, x_1, \ldots, x_n$, posons $a'_1 = a_0 + a_1$, $x'_1 = \frac{a_0 x_0 + a_1 x_1}{a_0 + a_1}$, et, pour $i \geq 2$, $a'_i = a_i$, $x'_i = x_i$. Par convexité de f,

$$f(x'_1) \leq \frac{a_0 f(x_0) + a_1 f(x_1)}{a_0 + a_1}.$$

Avec l'hypothèse de récurrence,

$$\begin{aligned} f(\frac{\sum_{i=0}^n a_i x_i}{\sum_{i=0}^n a_i}) &= f(\frac{\sum_{i=1}^n a'_i x'_i}{\sum_{i=1}^n a'_i}) \\ &\leq \frac{\sum_{i=1}^n a'_i f(x'_i)}{\sum_{i=1}^n a'_i} \\ &= \frac{(a_0 + a_1) f(x'_1) + \sum_{i=2}^n a_i f(x_i)}{\sum_{i=0}^n a_i} \\ &\leq \frac{\sum_{i=0}^n a_i f(x_i)}{\sum_{i=0}^n a_i}, \end{aligned}$$

ce qu'il fallait démontrer. Par l'hypothèse de récurrence, l'égalité entraîne que f prend la même valeur en tous les x'_i dont les poids sont non nuls, et que cette valeur est égale à $\frac{a_0 f(x_0) + a_1 f(x_1)}{a_0 + a_1}$. Si a_0 et a_1 sont non nuls, cela entraîne que $f(x_0) = f(x_1) = f(x_2) = \cdots = f(x_n)$.

Dans le cas général, on utilise la Proposition 12. Par hypothèse, $f = \sup_{j \in J} a_j$ où chaque fonction a_j est affine. Alors pour tout j, $\mathbb{E}(f(X)) \geq \mathbb{E}(a_j(X)) = a_j(\mathbb{E}(X))$. En prenant le sup sur J, il vient $f(\mathbb{E}(X)) \leq \mathbb{E}(f(X))$.

Cas d'égalité. Supposons que $f(\mathbb{E}(X)) = \mathbb{E}(f(X))$. On pose $x = \mathbb{E}(X)$. On decompose Ω en $\Omega_1 = \{X < x\}$ et $\Omega_2 = \{X \geq x\}$. Soit $t = \mathbb{P}(\Omega_2)$. Supposons que $t > 0$ et $1 - t > 0$. On munit Ω_1 de la mesure de probabilité $\frac{1}{1-t}\mathbb{P}_{|\Omega_1}$ et Ω_2 de la mesure de probabilité $\frac{1}{t}\mathbb{P}_{|\Omega_2}$. Soit X_i la restriction de X à Ω_i. Alors X_1 et X_2 sont des variables aléatoires (ce sont les conditionnements de X aux évènements $\{X < \mathbb{E}(X))\}$ et $\{X \geq \mathbb{E}(X)\}$). On pose $x_i = \mathbb{E}(X_i)$ (on pourrait les noter $\mathbb{E}(X|X < x)$ et $\mathbb{E}(X|X \geq x)$). Comme

$$x_1 = \mathbb{E}(X_1) = \frac{1}{1-t} \int_{\Omega_1} X \, d\mathbb{P}, \quad x_2 = \mathbb{E}(X_2) = \frac{1}{t} \int_{\Omega_2} X \, d\mathbb{P},$$

il vient $x = (1-t)x_1 + t x_2$. De même, $\mathbb{E}(f(X)) = (1-t)\mathbb{E}(f(X_1)) + t\mathbb{E}(f(X_2))$. On a montré que $f(x_i) = f(\mathbb{E}(X_i)) \leq \mathbb{E}(f(X_i))$, d'où

$$\begin{aligned} (1-t)f(x_1) + t f(x_2) &\leq (1-t)\mathbb{E}(f(X_1)) + t\mathbb{E}(f(X_2)) \\ &= E(f(X)) = f(\mathbb{E}(X)) = f(x) \\ &= f((1-t)x_1 + t x_2), \end{aligned}$$

ce qui contredit la stricte convexité de f. On conclut que $t = 0$ ou 1, i.e. $X \leq \mathbb{E}(X)$ ou $X \geq \mathbb{E}(X)$ presque partout. Cela entraîne que $X = \mathbb{E}(X)$ presque partout. ■

2.5 Conséquences de l'inégalité de Jensen

Théorème 2 (Positivité de l'entropie relative) *Soient p et q deux distributions de probabilité sur le même ensemble E. Alors $D(p||q) \geq 0$ avec égalité si et seulement si $p = q$.*

Preuve Soit X une variable aléatoire de distribution p. En utilisant l'inégalité de Jensen, il vient

$$\begin{aligned} D(p||q) &= \mathbb{E}(\log_2(\frac{p(X)}{q(X)})) = \mathbb{E}(-\log_2(\frac{q(X)}{p(X)})) \\ &\geq -\log_2(\mathbb{E}(\frac{q(X)}{p(X)})) = -\log_2(\sum_{x\in E} p(x)\frac{q(x)}{p(x)}) = -\log_2(\sum_{x\in E} q(x)) \\ &= 0. \end{aligned}$$

Par stricte concavité du log, l'égalité entraîne que la variable $\frac{q(X)}{p(X)}$ est constante, donc p et q sont proportionnelles, donc elles sont égales (puisque la somme de leurs valeurs vaut 1). ■

Exemple 14 *Pour toute distribution p sur un ensemble fini E, $H(p) \leq \log_2 |E|$ avec égalité si et seulement si p est la distribution uniforme.*

En effet, soit u la distribution uniforme sur E. Alors

$$0 \leq D(p||u) = \log_2 |E| - H(X),$$

avec égalité si et seulement si $p = u$.

Corollaire 15 *Soient X et Y deux variables aléatoires. Alors $I(X;Y) \geq 0$, avec égalité si et seulement si X et Y sont indépendantes.*

Interprétation : l'information fournie par Y fait diminuer l'incertitude sur X, elle diminue strictement sauf si Y est independante de X.

Preuve Par définition, $I(X;Y)$ est l'entropie relative de la loi du couple et de la loi produit. Elle est donc positive, et nulle seulement si $p_{(X,Y)} = p_X p_Y$, i.e. X et Y sont indépendantes. ■

Suggestion d'exercice : n^{0}3, feuille 1.

2.6 Entropie conditionnelle

Rappel 16 *Soient X et Y deux variables aléatoires, à valeurs dans E_X et E_Y respectivement. Le couple (X,Y) est une variable aléatoire à valeurs dans $E_X \times E_Y$, on note $p_{(X,Y)}$ sa distribution, i.e. pour $x \in E_X$ et $y \in E_Y$,*

$$p_{(X,Y)}(x,y) = \mathbb{P}(X = x \text{ et } Y = y).$$

Les distributions p_X et p_Y s'en déduisent,

$$p_X(x) = \sum_{y\in E_Y} p_{(X,Y)}(x,y), \quad p_Y(y) = \sum_{x\in E_X} p_{(X,Y)}(x,y).$$

Notation 17 *Soient X et Y deux variables aléatoires. L'entropie du couple (X,Y) est notée $H(X,Y)$ plutôt que $H((X,Y))$. Autrement dit,*

$$H(X,Y) = -\sum_{(x,y)\in E_X\times E_Y} p_{(X,Y)}(x,y)\log_2(p_{(X,Y)}(x,y)).$$

Rappel 18 *La* loi conditionnelle de Y sachant que $X = x$ *est la distribution de probabilité notée* $p_{Y|X=x}$ *définie par*

$$p_{Y|X=x}(y) = \frac{p_{(X,Y)}(x,y)}{p_X(x)}.$$

On dit que X et Y sont indépendantes *si* $p_{Y|X=x} = p_Y$ *pour tout* x, *i.e. si* $p_{(X,Y)}(x,y) = p_X(x)p_Y(y)$ *pour tout* (x,y).

Définition 19 *Soient* X *et* Y *deux variables aléatoires, à valeurs dans* E_X *et* E_Y *respectivement. On appelle* entropie conditionnelle de Y sachant X, *et on note* $H(Y|X)$ *le nombre*

$$H(Y|X) = \sum_{x \in E_X} p_X(x) H(p_{Y|X=x}).$$

Autres expressions :

$$\begin{aligned} H(Y|X) &= -\sum_{x \in E_X} p_X(x) \sum_{y \in E_Y} p_{Y|X=x}(y) \log_2(p_{Y|X=x}(y)) \\ &= -\sum_{(x,y) \in E_X \times E_Y} p(x,y) \log_2(\frac{p_{(X,Y)}(x,y)}{p_X(x)}). \end{aligned}$$

Exemple 20 *Lorsque* $Y = X$, $H(X|X) = 0$.

En effet, la loi conditionnelle de X sachant que $X = x$ est concentrée en x, son entropie est nulle. Informellement, si on connait X, il n'y a plus aucune incertitude concernant X, d'où une entropie conditionnelle nulle.

Proposition 21 *Soient* X *et* Y *deux variables aléatoires. Alors*

$$H(X,Y) = H(X) + H(Y|X).$$

Preuve

$$\begin{aligned} H(Y|X) &= -\sum_{(x,y) \in E_X \times E_Y} p(x,y) \log_2(\frac{p_{(X,Y)}(x,y)}{p_X(x)}) \\ &= -\sum_{(x,y) \in E_X \times E_Y} p(x,y) \log_2(p_{(X,Y)}(x,y)) + \sum_{(x,y) \in E_X \times E_Y} p(x,y) \log_2(p_X(x)) \\ &= H(X,Y) + \sum_{x \in E_X} p_X(x) \log_2(p_X(x)) \\ &= H(X,Y) - H(X). \end{aligned}$$

■

Fin du cours n^{0}1

Suggestion d'exercice : n^{0}4, feuille 1.

Remarque 22 *En général,* $H(Y|X) \neq H(X|Y)$.

Car $H(X,Y) = H(Y,X)$ et $H(X) \neq H(Y)$ en général.

Voici une version conditionnelle de la Proposition 21.

Corollaire 23 *Soient X, Y et Z trois variables aléatoires. Alors*

$$H(X,Y|Z) = H(X|Z) + H(Y|X,Z).$$

Preuve

$$\begin{aligned} H(X,Y|Z) &= H(X,Y,Z) - H(Z) \\ &= H(X,Y,Z) - H(X,Z) + H(X,Z) - H(Z) \\ &= H(Y|X,Z) + H(X|Z). \end{aligned}$$

■

En utilisant de façon répétée la Proposition 21, on obtient

Corollaire 24 *Soient $(X_1, \ldots, X_n)$ des variables aléatoires. Alors*

$$H(X_1, \ldots, X_n) = H(X_1) + \sum_{i=2}^{n} H(X_i|X_{i-1}, \ldots, X_1).$$

Preuve

$$\begin{aligned} H(X_1, X_2) &= H(X_1) + H(X_2|X_1), \\ H(X_1, X_2, X_3) &= H(X_1, X_2) + H(X_3|(X_1, X_2)) \\ &= H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2) \end{aligned}$$

etc... ■

2.7 Information mutuelle et entropie conditionnelle

Proposition 25 *Soient X et Y deux variables aléatoires. Alors*

$$I(X;Y) = H(X) - H(X|Y).$$

Interprétation : $I(X;Y)$ mesure la diminution de l'incertitude contenue dans X lorsqu'on apprend Y.

Preuve

$$\begin{aligned} I(X;Y) &= \sum_{(x,y)\in E_X \times E_Y} p_{(X,Y)}(x,y) \log_2\left(\frac{p_{(X,Y)}(x,y)}{p_X(x)p_Y(y)}\right) \\ &= \sum_{(x,y)\in E_X \times E_Y} p(x,y) \log_2\left(\frac{p_{(X,Y)}(x,y)}{p_Y(y)}\right) - \sum_{(x,y)\in E_X \times E_Y} p_{(X,Y)}(x,y) \log_2(p_X(x)) \\ &= -H(X|Y) - \sum_{x\in E_X} p_X(x) \log_2(p_X(x)) \\ &= -H(X|Y) + H(X). \end{aligned}$$

■

Exemple 26 $I(X;X) = H(X)$

En effet, $H(X|X) = 0$. Informellement, en apprenant X, on sait tout sur X, donc l'incertitude passe de $H(X)$ à 0, la diminution d'incertitude est $H(X)$.

Corollaire 27 $I(X;Y) = H(X) + H(Y) - H(X,Y)$.

Preuve Combinaison des propositions 21 et 25. ■

Suggestion d'exercice : n^{0}4, feuille 1.

2.8 Conditionner diminue l'entropie

Proposition 28 *Soient X et Y deux variables aléatoires. Alors $H(X|Y) \leq H(X)$, avec égalité si et seulement si X et Y sont indépendantes.*

Preuve $0 \leq I(X;Y) = H(X) - H(X|Y)$. ■

Interprétation : conditionnellement à Y, l'incertitude sur X diminue toujours, et diminue strictement sauf si Y est independante de X.

Attention, c'est seulement vrai en moyenne : pour une valeur particulière y, $H(X|Y=y)$ peut être plus grand que $H(X)$ (certaines informations supplémentaires peuvent ajouter à la confusion, alors que d'autres renseignent sur X). Voir un exemple en Exercice 5, TD1.

Suggestion d'exercice : n^{0}5, feuille 1.

Corollaire 29 *Soient X et Y deux variables aléatoires. Alors $H(X,Y) \leq H(X) + H(Y)$, avec égalité si et seulement si X et Y sont indépendantes.*

Preuve $H(X,Y) = H(X) + H(Y|X) \leq H(X) + H(Y)$. ■

2.9 Information mutuelle conditionnelle

On va voir que le principe "conditionner diminue l'entropie" a une portée plus générale. Il s'applique aussi à l'entropie conditionnelle (Corollaire 32). Pour le montrer, on a recours à une notion supplémentaire, celle d'information mutuelle conditionnelle.

Définition 30 *Soient X, Y et Z des variables aléatoires. L'*information mutuelle conditionnelle *de X et Y sachant Z est la moyenne, pondérée par la loi p_Z de Z, des informations mutuelles des variables conditionnées $X|Z=z$ et $Y|Z=z$,*

$$I(X;Y|Z) = \sum_{z \in E_Z} p_Z(z) I(X|Z=z; Y|Z=z).$$

Autre expression :

$$I(X;Y|Z) = \sum_{x,y,z} p_{(X,Y,Z)}(x,y,z) \log_2 \frac{\frac{p_{(X,Y,Z)}(x,y,z)}{p_Z(z)}}{\frac{p_{(X,Z)}(x,z)}{p_Z(z)} \frac{p_{(Y,Z)}(y,z)}{p_Z(z)}}.$$

Comme c'est une moyenne d'informations mutuelles, $I(X;Y|Z) \geq 0$. Comme dans l'information mutuelle, X et Y jouent des rôles symétriques : $I(X;Y|Z) = I(Y;X|Z)$.

Il y a une version conditionnelle de la formule $I(X;Y) = H(X) - H(X|Y)$ (Proposition 25) :

Proposition 31 *Soient X, Y et Z des variables aléatoires. Alors*

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z).$$

Preuve Pour tout $z \in E_Z$, $I(X|Z=z; Y|Z=z) = H(X|Z=z) - H((X|Z=z)|(Y|Z=z))$, d'où

$$\begin{aligned}
I(X;Y|Z) &= \sum_{z \in E_Z} p_Z(z) I(X|Z=z; Y|Z=z) \\
&= \sum_{z \in E_Z} p_Z(z) H(X|Z=z) - \sum_{z \in E_Z} p_Z(z) H((X|Z=z)|(Y|Z=z)) \\
&= H(X|Z) - \sum_{(y,z) \in E_Y \times E_Z} p_Z(z) p_{Y|Z=z}(y) H((X|Z=z)|(Y|Z=z)=y) \\
&= H(X|Z) - \sum_{(y,z) \in E_Y \times E_Z} p_{(Y,Z)}(y,z) H(X|Y=y, Z=z) \\
&= H(X|Z) - H(X|Y,Z).
\end{aligned}$$

Pour calculer $H((X|Z=z)|(Y|Z=z)=y)$, on a calculé la loi conditionnelle de $X|Y=y$ sous la probabilité conditionnelle sachant $Z=z$, notée $\mathbb{P}_z$,

$$\begin{aligned}
\mathbb{P}_{(X|Z=z)|(Y|Z=z)=y}(x) &= (\mathbb{P}_z)_{X|Y=y}(x) \\
&= \frac{\mathbb{P}_z(X=x,Y=y)}{\mathbb{P}_z(Y)} \\
&= \frac{\mathbb{P}(X=x,Y=y|Z=z)}{\mathbb{P}(Y|Z=z)} \\
&= \frac{\mathbb{P}_{(X,Y,Z)}(x,y,z)}{\mathbb{P}_{(Y,Z)}(y,z)} \\
&= \mathbb{P}_{X|(Y=y,Z=z)}(x).
\end{aligned}$$

■

Corollaire 32 *Soient X, Y et Z des variables aléatoires. Alors $H(X|Y,Z) \leq H(X|Y)$.*

De même, avec la Proposition 23,

Corollaire 33 *Soient X, Y et Z des variables aléatoires. Alors $H(X,Y|Z) \leq H(X|Z)+H(Y|Z)$.*

(Cela peut aussi se montrer directement : pour tout z, $H(X,Y|Z=z) \leq H(X|Z=z)+H(Y|Z=z)$, donc c'est vrai en moyenne).

Enfin, l'information mutuelle satisfait une règle d'addition analogue à la Proposition 21 et au Corollaire 24.

Proposition 34 *Soient X, Y, Z des variables aléatoires. Alors*

$$I(X,Y;Z) = I(X;Z) + I(Y;Z|X), \quad I(X;Y,Z) = I(X;Z) + I(X;Y|Z).$$

Plus généralement, soient $X_1,\ldots,X_n$ et Z des variables aléatoires. Alors

$$I(X_1,\ldots,X_n;Z) = I(X_1;Z) + \sum_{i=2}^{n} I(X_i;Z|X_{i-1},\ldots,X_1).$$

Preuve On exprime l'information mutuelle en fonction d'entropies conditionnelles, on applique la Proposition 21 et on conclut avec la Proposition 31.

$$\begin{aligned}
I(X,Y;Z) &= H(X,Y) - H(X,Y|Z) \\
&= H(X) + H(Y|X) - (H(X|Z) + H(Y|X,Z)) \\
&= I(X;Z) + I(Y;Z|X).
\end{aligned}$$

La seconde formule s'obtient en échangeant X et Z et en observant que $I(Y;Z|X) = I(Z;Y|X)$.

Quand il y a davantage de variables, le Corollaire 24 remplace la Proposition 21.

$$\begin{aligned}
I(X_1,\ldots,X_n;Z) &= H(X_1,\ldots,X_n) - H(X_1,\ldots,X_n|Z) \\
&= H(X_1) + \sum_{i=2}^{n} H(X_i|X_{i-1},\ldots,X_1) \\
&\quad - \left(H(X_1|Z) + \sum_{i=2}^{n} H(X_i|X_{i-1},\ldots,X_1,Z) \right) \\
&= I(X_1|Z) + \sum_{i=2}^{n} I(X_i;Z|X_{i-1},\ldots,X_1).
\end{aligned}$$

■

Suggestion d'exercice : n^{0}6, feuille 1.

Suggestion d'exercice : n^{0}7, feuille 1.

2.10 Entropie relative conditionnelle

Définition 35 *Soient E_X, E_Y des ensembles finis, soient p et q des distributions de probabilité sur $E_X \times E_Y$. L'*entropie relative conditionnelle $D(p_{Y|X}||q_{Y|X})$ *est la moyenne selon la marginale p_X des entropies relatives $D(p_{Y|X=x}||q_{Y|X=x})$, i.e.*

$$D(p_{Y|X}||q_{Y|X}) = \sum_{x \in E_X} p_X(x) D(p_{Y|X=x}||q_{Y|X=x}).$$

Remarquer que $D(p_{Y|X}||q_{Y|X})$ est une moyenne de quantités positives ou nulles, donc $D(p_{Y|X}||q_{Y|X}) \geq 0$. Autres expressions :

$$\begin{aligned} D(p_{Y|X}||q_{Y|X}) &= \sum_{x \in E_X} p_X(x) \sum_{y \in E_Y} p_{Y|X=x}(y) \log \frac{p_{Y|X=x}(y)}{q_{Y|X=x}(y)} \\ &= \sum_{(x,y) \in E_X \times E_Y} p(x,y) \log \frac{\frac{p(x,y)}{\sum_z p(x,z)}}{\frac{q(x,y)}{\sum_z q(x,z)}}. \end{aligned}$$

Proposition 36 *Soient E_X, E_Y des ensembles finis, soient p et q des distributions de probabilité sur $E_X \times E_Y$. On peut exprimer l'entropie relative p et q comme suit.*

$$D(p||q) = D(p_X||q_X) + D(p_{Y|X}||q_{Y|X}).$$

Preuve

$$\begin{aligned} D(p_{Y|X}||q_{Y|X}) &= \sum_{(x,y) \in E_X \times E_Y} p(x,y) \log \frac{p(x,y)}{q(x,y)} + \sum_{(x,y) \in E_X \times E_Y} p(x,y) \log \frac{\sum_z q(x,z)}{\sum_z p(x,z)} \\ &= D(p||q) - D(p_X||q_X). \end{aligned}$$

■

Remarque 37 *En particulier, $D(p||q) \geq D(p_{Y|X}||q_{Y|X})$, encore une manifestation du principe selon lequel conditionner diminue l'entropie.*

On utilisera l'entropie relative conditionnelle dans la preuve du Théorème 9.

2.11 Convexité de l'entropie relative

Théorème 3 *Soit E un ensemble fini. L'entropie relative est une fonction convexe sur l'ensemble $\mathcal{P}(E) \times \mathcal{P}(E)$, où $\mathcal{P}(E)$ est l'ensemble des distributions de probabilité sur E.*

Remarquer que $\mathcal{P}(E)$ est un polyèdre convexe de $\mathbb{R}^{|E|}$, et $\mathcal{P}(E) \times \mathcal{P}(E)$ un polyèdre convexe de $\mathbb{R}^{2|E|}$. On peut donc parler de convexité d'une fonction définie sur $\mathcal{P}(E) \times \mathcal{P}(E)$.

Fin du cours n^{0}2

Preuve Soient p_1, p_2, q_1 et q_2 des distributions de probabilité sur E. Soit $t \in]0,1[$. Soit $x \in E$. On écrit

$$\frac{(1-t)p_1(x) + tp_2(x)}{(1-t)q_1(x) + tq_2(x)} = (1-\lambda)a + \lambda b, \quad \text{où } a = \frac{p_1(x)}{q_1(x)},\ b = \frac{p_2(x)}{q_2(x)},\ \lambda = \frac{tq_2(x)}{(1-t)q_1(x) + tq_2(x)}.$$

Comme $x \mapsto f(x) = x \log_2 x$ est convexe,

$$\begin{aligned} f\left(\frac{(1-t)p_1(x) + tp_2(x)}{(1-t)q_1(x) + tq_2(x)}\right) &= f((1-\lambda)a + \lambda b) \\ &\leq (1-\lambda)f(a) + \lambda f(b) \\ &= \frac{(1-t)q_1(x)}{(1-t)q_1(x) + tq_2(x)} f\left(\frac{p_1(x)}{q_1(x)}\right) + \frac{tq_2(x)}{(1-t)q_1(x) + tq_2(x)} f\left(\frac{p_2(x)}{q_2(x)}\right) \\ &= \frac{(1-t)p_1(x)}{(1-t)q_1(x) + tq_2(x)} \log_2\left(\frac{p_1(x)}{q_1(x)}\right) + \frac{tp_2(x)}{(1-t)q_1(x) + tq_2(x)} \log_2\left(\frac{p_2(x)}{q_2(x)}\right), \end{aligned}$$

d'où

$$((1-t)p_1(x)+tp_2(x))\log_2(\frac{(1-t)p_1(x)+tp_2(x)}{(1-t)q_1(x)+tq_2(x)}) \leq (1-t)p_1(x)\log_2(\frac{p_1(x)}{q_1(x)})+tp_2(x)\log_2(\frac{p_2(x)}{q_2(x)}).$$

En sommant sur $x \in E$, il vient

$$D((1-t)p_1+tp_2||(1-t)q_1+tpq_2) \leq (1-t)D(p_1||q_1)+tD(p_2||q_2).$$

■

Corollaire 38 *L'entropie est une fonction concave sur l'ensemble des distributions de probabilité sur E. Elle atteint son maximum à la distribution uniforme.*

Preuve Soit $u : x \mapsto \frac{1}{|E|}$ la distribution uniforme sur E. Alors pour toute distribution de probabilité p,

$$D(p||u) = \sum_{x\in E} p(x)\log_2(\frac{p(x)}{u(x)}) = \log_2(|E|) - H(p),$$

est une fonction convexe de p, positive ou nulle, et nulle exactement en u. ■

Exemple 39 *L'entropie de la distribution de Bernoulli $q \mapsto H(\mathcal{B}(q))$ est une fonction concave sur $[0,1]$, elle atteint son maximum en $\frac{1}{2}$.*

Suggestion d'exercice : n^{0}8, feuille 1.

2.12 A retenir

- Les notions absolues : entropie, entropie relative, information mutuelle.
- Leurs versions conditionnelles : entropie conditionnelle, entropie relative conditionnelle, information mutuelle conditionnelle. Les formules d'addition correspondantes.
- Le lien entre information mutuelle et entropie conditionnelle (et sa version conditionnelle).
- Les inégalités : toutes les entropies/informations sont positives ou nulles, conditionner diminue l'entropie, concavité de l'entropie (resp. convexité de l'entropie relative) comme fonction de la distribution (resp. de deux distributions).

Il s'agit d'un chapitre théorique, avec plein de définitions et d'énoncés qui seront utilisés ensuite.

3 Compression de données

3.1 Motivation

Il est notoire que l'anglais est plus concis que le français : la traduction anglaise d'un texte français est presque toujours nettement plus courte que le texte original. Jusqu'où peut on aller dans cette direction ? Peut on fabriquer une langue artificielle qui soit encore plus économe ?

Ignorons la grammaire pour ne travailler que sur le vocabulaire. La langue artificielle - appelons la le codage - est définie par un dictionnaire : à chaque mot français x correspond son équivalent $C(x)$, une chaîne de caractères pris dans l'alphabet latin $\mathcal{D}$, qui possède $D = 26$ lettres. On note $\ell(x)$ la longueur de $C(x)$. Etant donné un texte français T, i.e. une suite de mots, la longueur totale de sa traduction est $\sum_{x\in E} N(x,T)\ell(x)$ où E est l'ensemble des mots français et $N(x,T)$ est le nombre d'apparitions du mot x dans le texte T. La performance du codage est

$$\sum_{x\in E} \frac{N(x,T)}{N(T)}\ell(x) = \sum_{x\in E} p_T(x)\ell(x),$$

où $N(T)$ est le nombre total de mots dans T, et $p_T(x)$ la fréquence d'apparition du mot x dans T. Pour de grands textes, l'expérience montre que les fréquences p_T convergent vers une distribution de probabilité sur E, la fréquence d'utilisation de chaque mot en français.

3.2 Modélisation

La langue française est donc modélisée par une variable aléatoire X à valeurs dans un ensemble fini E, le codage par une application $C : E \to \mathcal{D}^*$ (ensemble des suites finies de lettres prises dans $\mathcal{D}$), et la performance du codage par l'espérance $\mathbb{E}(\ell(X))$.

Le codage des textes est l'application $C^* : E^* \to \mathcal{D}^*$ qui envoie le texte (i.e. la suite de mots) $(x_1, x_2, \ldots, x_k)$ sur sa traduction $C(x_1)C(x_2)\cdots C(x_k)$. Par souci d'économie, les mots codés sont concaténés sans séparateurs (pas d'espace ni de ponctuation).

Définition 40 *Un codage C est* uniquement décodable *si l'application $C^* : E^* \to \mathcal{D}^*$ est injective.*

C'est la moindre des choses.

Question 41 (Problème du codage de source) *Quelle est la meilleure performance $\mathbb{E}(\ell(X))$ accomplie par les codages uniquement décodables ?*

Ce problème s'intitule problème du *codage de source*, par opposition au codage des transmissions, qu'on étudiera plus loin. On voit la variable aléatoire X comme une source de mots. Dans notre exemple initial, la source, c'est la littérature française.

Suggestion d'exercice : n^{0}1, feuille 2.

3.3 Inégalité de Kraft

Dans un codage uniquement décodable, les mots codés ne peuvent pas tous être courts. Le théorème suivant donne une borne inférieure optimale sur les longueurs.

Théorème 4 (L. Kraft (1949), B. McMillan (1953)) *Soit $C : E \to \mathcal{D}^*$ un codage uniquement décodable dans un alphabet à $D = |\mathcal{D}|$ lettres. Alors*

$$\sum_{x \in E} D^{-\ell(x)} \leq 1.$$

Inversement, toute fonction $\ell : E \to \mathbb{N} \setminus \{0\}$ qui satisfait l'inégalité précédente est la fonction longueur d'un codage uniquement décodable.

Preuve Soit $k \geq 1$ un entier. On développe

$$\begin{aligned} (\sum_{x \in E} D^{-\ell(x)})^k &= \sum_{(x_1,\ldots,x_k) \in E^k} D^{-\sum_i \ell(x_i)} \\ &= \sum_{T \in E^k} D^{-|C^*(T)|}. \end{aligned}$$

Ici E^k désigne l'ensemble des textes à k mots et $|C^*(T)|$ est la longueur de la traduction de T. Soit

$$L = \max_{x \in E} \ell(x)$$

la longueur maximale d'un mot codé. On note $N(m)$ le nombre de textes à k mots codés par des suites de m lettres. Alors

$$\sum_{T \in E^k} D^{-|C^*(T)|} = \sum_{m=1}^{kL} N(m) D^{-m}$$

Comme C^* est injective, et comme il y a seulement D^m suites de m lettres distinctes, $N(m) \leq D^m$. Il vient

$$(\sum_{x \in E} D^{-\ell(x)})^k \leq \sum_{m=1}^{kL} D^m D^{-m} = kL,$$

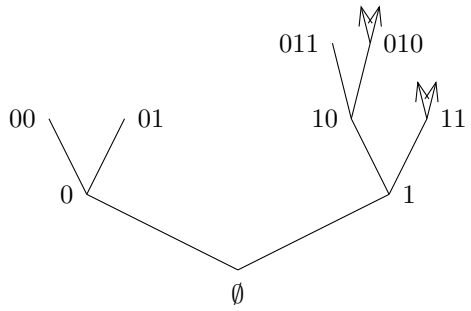
d'où $\sum_{x\in E} D^{-\ell(x)} \leq (kL)^{1/k}$ pour tout $k \geq 1$. En faisant tendre k vers $+\infty$, on trouve que $\sum_{x\in E} D^{-\ell(x)} \leq 1$.

Réciproquement, soit $\ell : E \to \mathbb{N} \setminus \{0\}$ une fonction qui satisfait l'inégalité $\sum_{x\in E} D^{-\ell(x)} \leq 1$. On ordonne les éléments de $E = \{x_1, x_2, \ldots\}$ de sorte que ℓ soit croissante. On ordonne $\mathcal{D}$ aussi. L'ensemble $\mathcal{A}_0 = \mathcal{D}^*$ est ordonné par l'ordre lexicographique. On construit par récurrence des éléments $C(x_1), C(x_2), \ldots$ et des sous-ensembles décroissants $\mathcal{A}_0 \supset \mathcal{A}_1 \supset \mathcal{A}_2 \supset \cdots$ comme suit. Soit $C(x_1)$ le premier élément de $\mathcal{D}^*$ (dans l'ordre lexicographique) de longueur $\ell(x_1)$. Soit $\mathcal{A}_1 = \mathcal{A}_0$ privé de toutes les suites qui commencent par $C(x_1)$. Soit $C(x_2)$ le premier élément de $\mathcal{A}_1$ (dans l'ordre lexicographique) de longueur $\ell(x_2)$. Soit $\mathcal{A}_2 = \mathcal{A}_1$ privé de toutes les suites qui commencent par $C(x_2)$, etc.... La condition $\sum_i D^{-\ell(x_i)} \leq 1$ garantit qu'à chaque étape, $\mathcal{A}_i$ possède bien des éléments de longueur $\ell(x_i)$. En effet, soit de nouveau $L = \max_{x\in E} \ell(x)$. Soit $N(j)$ le nombre de suites de longueur L qui n'appartiennent pas à $\mathcal{A}_j$. Par construction, $N(0) = 0$. En passant de $\mathcal{A}_{j-1}$ à $\mathcal{A}_j$, on supprime toutes les suites de longueur L commençant par $C(x_j)$, il y en a exactement $D^{L-\ell(x_j)}$. Par conséquent,

$$N(j) = N(j-1) + D^{L-\ell(x_i)}, \quad \text{d'où} \quad N(j) = \sum_{i=1}^{j} D^{L-\ell(x_i)} \leq D^L,$$

avec inégalité stricte tant que E n'est pas épuisé. Il y a donc des éléments de longueur L (et donc, de toute longueur $\leq L$) dans $\mathcal{A}_{j-1}$ jusqu'à la dernière étape de la construction.

Il est commode de visualiser chaque ensemble ordonné $\mathcal{A}_j$ sous la forme d'un arbre enraciné, sous-arbre de l'arbre associé à $\mathcal{D}^*$. Ci-dessous, l'arbre final lorsque $D = 2$, $E = \{x_1, x_2, x_3\}$, $\ell(x_1) = \ell(x_2) = 2$, $\ell(x_3) = 3$ (les flèches désignent les branches de l'arbre qui se poursuivent sans changement). Le codage est $C(x_1) = 00$, $C(x_2) = 01$, $C(x_3) = 011$.



Arbre schématisant la construction du codage

Par construction, aucun des $C(x)$, $x \in E$, n'est un préfixe d'un autre $C(x')$. Par conséquent, une concaténation $w = C(x_{i_1})C(x_{i_2})\cdots C(x_{i_k})$ se décode aisément comme suit. On lit les lettres dans l'ordre jusqu'à ce qu'on reconnaisse un élément de $C(E)$. Il y a exactement un élément x de E tel que $C(x)$ commence par les mêmes lettres que w, c'est x_{i_1}, donc w est la traduction d'un mot commençant par x_{i_1}. On recommence avec $w_1 = C(x_{i_2})\cdots C(x_{i_k})$, etc... Cela prouve que $C^* : E^* \to \mathcal{D}^*$ est injective. ∎

3.4 Théorème du codage de source

L'inégalité de Kraft ramène la question de minimiser la longueur moyenne d'un codage uniquement décodable à une simple inégalité sur les fonctions sur E. La réponse fait intervenir l'entropie.

Théorème 5 (C. Shannon (1948)) *Soit X une variable aléatoire à valeurs dans un ensemble fini E. Soit $C : E \to \mathcal{D}^*$ un codage uniquement décodable dans un alphabet à $D = |\mathcal{D}|$ lettres. Soit $\ell : E \to \mathbb{N}$ la longueur des mots-codes. Alors la longueur moyenne satisfait*

$$\mathbb{E}(\ell(X)) \geq \frac{H(X)}{\log_2(D)}.$$

Cette borne est atteinte si et seulement si les valeurs $p(x)$, $x \in E$, de la loi de X sont des puissances de D.

En général, il existe un code C qui réalise

$$\frac{H(X)}{\log_2(D)} \leq \mathbb{E}(\ell(X)) < \frac{H(X)}{\log_2(D)} + 1.$$

Autrement dit, si $L_{ud}(X)$ désigne la borne inférieure des longueurs moyennes des codages uniquement décodables, alors

$$\frac{H(X)}{\log_2(D)} \leq L_{ud}(X) < \frac{H(X)}{\log_2(D)} + 1.$$

Preuve Soit $c = \sum_{x \in E} D^{-\ell(x)}$. On définit une distribution de probabilité q sur E par $q(x) = \frac{1}{c} D^{-\ell(x)}$. On calcule

$$\begin{aligned} D(p||q) &= \sum_{x \in E} p(x) \log_2(\frac{p(x)}{q(x)}) \\ &= -H(X) + \log_2(c) + \log_2(D)\mathbb{E}(\ell(X)). \end{aligned}$$

Comme $D(p||q) \geq 0$ (Théorème 2) et $c \leq 1$ (Théorème 4), il vient $-H(X) + \log_2(D)\mathbb{E}(\ell(X)) \geq 0$. L'égalité entraîne que $D(p||q) = 0$ et $c = 1$, et donc que $p = q$ est à valeurs dans les puissances de D.

Réciproquement, si p est à valeurs dans les puissances de D, on pose $\ell = -\log_2(p)/\log_2(D)$. C'est une fonction à valeurs entières qui satisfait l'inégalité de Kraft. D'après le Théorème 4, il existe un codage C uniquement décodable dont les longueurs sont données par ℓ, il réalise la performance $\mathbb{E}(\ell(X)) = \frac{H(X)}{\log_2(D)}$.

En général, on pose $\ell = \lceil -\log_2(p)/\log_2(D) \rceil$, fonction à valeurs entières qui satisfait l'inégalité de Kraft, donc réalisable par un codage uniquement décodable. On calcule

$$\begin{aligned} \mathbb{E}(\ell(X)) &= \sum_{x \in E} p(x) \lceil -\frac{\log_2(p)}{\log_2(D)} \rceil \\ &< \sum_{x \in E} p(x)(-\frac{\log_2(p)}{\log_2(D)} + 1) \\ &= \frac{H(X)}{\log_2(D)} + 1. \end{aligned}$$

■

Cela confirme l'idée que l'entropie mesure la quantité d'information nécessaire pour décrire une variable aléatoire X : c'est la borne inférieure du nombre de bits nécessaires en moyenne pour coder X.

Remarque 42 *Supposons qu'on est mal informé sur la loi de X. On n'en connait qu'une approximation q. Le codage qui semble presque optimal a pour longueurs de mots $\ell(x) = \lceil -\frac{\log_2(q)}{\log_2(D)} \rceil$, donc pour longueur moyenne*

$$\frac{H(X) + D(p||q)}{\log_2(D)} \leq \mathbb{E}(\ell(X)) < \frac{H(X) + D(p||q)}{\log_2(D)} + 1.$$

Autrement dit, l'entropie relative $D(p||q)$ mesure l'augmentation de la complexité de X due au fait qu'on a une information erronnée sur sa loi.

Remarque 43 *Un code non uniquement décodable ne satisfait pas nécessairement $\mathbb{E}(\ell(X)) \geq H(X)$.*

Fin du cours n^{0}3

Suggestion d'exercice : n^{0}2, feuille 2.

Suggestion d'exercice : n^{0}3, feuille 2.

3.5 Construction de codes optimaux

Le Théorème 5 ne fournit un codage optimal que dans le cas où la distribution de probabilité donnée est à valeurs dans les puissances de la taille D de l'alphabet. Il existe un algorithme, dû à D. Huffman, qui produit pour toute distribution de probabilité un codage optimal, voir le livre de Cover et Thomas, chapitre 5, pages 118 à 127. On le décrit ici uniquement pour $D = 2$.

3.5.1 Jeu dans la cour d'un collège

On commence par relier ce cas particulier du problème du codage optimal au *jeu des 10 questions* : "Choisis un nombre à trois chiffres, mémorise-le sans le dire. Je vais te poser une série de questions auxquelles tu répondras par oui ou par non. En moins de 10 questions, je peux deviner de quel nombre il s'agit !"

Ce qui nous intéresse, ce n'est pas de déterminer le nombre de questions nécessaire dans le pire des cas, mais d'optimiser le nombre de questions en moyenne, lorsque le nombre est tiré suivant une distribution de probabilité connue.

Autrement dit, on se donne une variable aléatoire X à valeurs dans un ensemble fini E. Une stratégie est un procédé déterministe qui, à une suite de réponses $s \in \{0,1\}^*$ associe une fonction $f_s : X \to \{0,1\}$. Si s est de longueur $i-1$, f_s représente la question que je pose à l'étape i au vu de la suite de réponses s. Lorsque $X = x$, le déroulement du jeu produit des suites de réponses $s_1 = f_\emptyset(x)$, $s_2 = s_1 f_{s_1}(x)$,..., $s_i = s_{i-1} f_{s_{i-1}}$,...,$s_Q$, $Q = Q(x)$. Le jeu s'arrête lorsque la suite de réponses $s_Q(x)$ détermine uniquement x, i.e. lorsque pour tout $y \in E$, $s_Q(y) = (x) \Rightarrow y = x$. Autrement dit, $Q(x)$ est la durée du jeu. Il s'agit de trouver la stratégie qui minimise la durée moyenne du jeu, i.e. l'espérance de $Q(X)$.

Question 44 *Quel est le nombre minimum de questions nécessaire, en moyenne, pour déterminer X ?*

On va utiliser une représentation graphique commode, sous forme d'arbre binaire. Il y a une terminologie pour ces arbres.

Définition 45 *Un* arbre binaire *est un arbre muni d'un sommet particulier, la* racine*. Les arêtes sont orientées dans la direction opposée à la racine. Chaque sommet autre que la racine reçoit une arête, qui provient de son* ascendant*. Chaque sommet a ou bien 2* descendants *(on parle de* noeud*), ou bien 0 (on parle de* feuille*). La* profondeur *d'un sommet est le nombre d'arêtes qui le séparent de la racine, on note cette fonction ℓ (ou ℓ_A quand il faut spécifier qu'il s'agit de l'arbre A). On note ∂A l'ensemble des feuilles d'un arbre binaire A.*

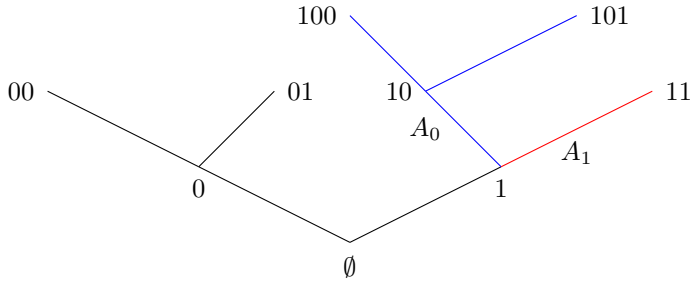
Proposition 46 *Etant donnée une distribution de probabilité p sur un ensemble fini E, les trois problèmes suivants sont équivalents.*

1. *Construire une stratégie au jeu des questions qui minimise la durée moyenne du jeu.*
2. *Construire un arbre binaire A et une bijection de E sur l'ensemble des feuilles de A, qui minimise la profondeur moyenne des feuilles.*
3. *Construire un codage de longueur moyenne minimale parmi ceux qui ont la propriété suivante : aucun mot-code n'est un préfixe d'un autre mot-code.*

Preuve La suite des réponses aux questions jusqu'à terminaison associe une suite $C(x) \in \{0,1\}^*$ à chaque élément de E. L'application C est un codage sans préfixe. En effet, supposons qu'il existe $x \neq y \in E$ tels que $C(x)$ est un préfixe de $C(y)$. Les tirages x et y donnent la même suite de réponses de longueur $Q(x)$. Si le jeu s'arrête là dans le cas du tirage x, c'est que cette suite détermine uniquement x. En particulier, $y = x$, contradiction. La fonction longueur ℓ du codage C coïncide avec Q, donc la durée moyenne du jeu est égale à la longueur moyenne du codage.

On peut représenter graphiquement C par un arbre binaire : la racine a deux descendants numérotés 0 et 1, qui ont chacun deux descendants numérotés respectivement 00 et 01, 10 et 11, etc... On arrête la construction récursive de l'arbre dès que la suite de questions s'arrête, i.e. quand elle caractérise uniquement un élément de E. Les feuilles correspondent bijectivement aux éléments x de E, elles héritent d'un poids $p(x)$. La profondeur de la feuille associée à x est égale à la longueur $\ell(x)$ du mot-code $C(x)$, la profondeur moyenne est égale à la longueur moyenne du codage. Si un noeud ne possède qu'un seul descendant, on peut le supprimer en contractant l'unique arête qui en part. Cela retire une lettre aux mots-codes dont il était un préfixe, donc cela diminue strictement la longueur moyenne. Cela ne peut pas se produire pour un codage minimal. Les codages minimaux correspondent donc à des arbres binaires.

Enfin, un arbre binaire fournit une stratégie pour le jeu des 10 questions. Une suite $s \in \{0,1\}^*$ amène à un noeud $n(s)$ de l'arbre. De ce noeud émanent deux sous-arbres A_0 et A_1. Soit f_s la fonction qui vaut 1 sur les feuilles de A_1, et 0 sur toutes les autres feuilles. La réponse à cette question permet de de faire un pas de plus, dans A_0 ou A_1 suivant que la réponse est 0 ou 1.



Stratégie associée à un arbre : sous arbres lorsque $s = 1$

Muni de cette stratégie, le jeu s'arrête lorsqu'on atteint une feuille, marquée par un élément x de E. Il permet d'identifier uniquement x. La durée du jeu en cas de tirage x est égale à la profondeur de x, donc profondeur moyenne et durée moyenne du jeu sont égales. ■

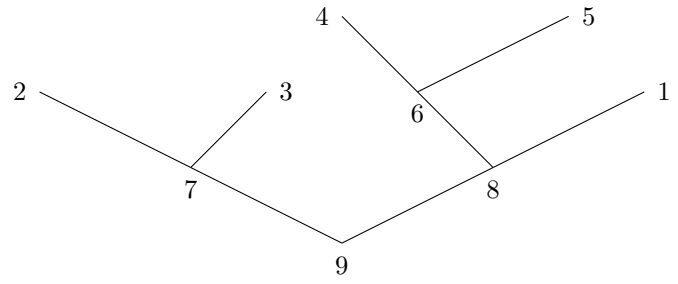
3.5.2 Algorithme de D. Huffman

En entrée : un ensemble fini ordonné E muni d'une distribution de probabilité p. En sortie : un arbre binaire A et une bijection de E sur ∂A.

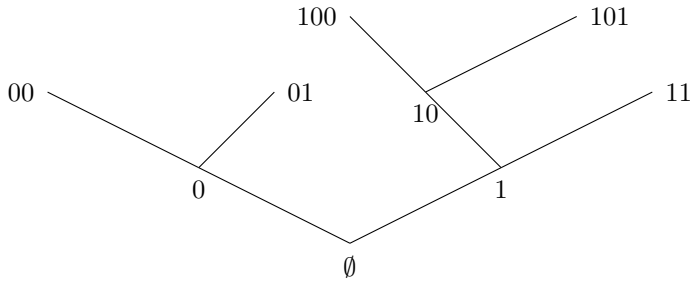
On construit une suite d'ensembles ordonnés $E = E_0, \ldots, E_{|E|-1}$ et de distributions de probabilité $p = p_0, \ldots, p_{|E|-1}$. L'algorithme est récursif. On répète l'opération suivante, qui fait diminuer $|E|$ d'une unité, jusqu'à ce que E n'ait plus qu'un élément.

Soient x et x' les éléments de E dont les probabilités sont les plus basses (s'il y a plusieurs choix, on choisit les derniers dans l'ordre fixé sur E). On les fusionne, i.e. on leur donne un ascendant commun y auquel on affecte la probabilité $p'(y) = p(x) + p(x')$. On pose $E' = E/(x \sim x') = (E \setminus \{x, x'\}) \cup \{y\}$.

Exemple. Soit $E = \{1,2,3,4,5\}$ muni de la distribution $(\frac{1}{4}, \frac{1}{4}, \frac{1}{5}, \frac{3}{20}, \frac{3}{20})$.
- Au premier tour, on fusionne 4 et 5 en un noeud baptisé 6, de probabilité $\frac{3}{20} + \frac{3}{20} = \frac{3}{10}$, d'où $E_1 = \{1,2,3,6\}$.
- Au second tour, on fusionne 2 et 3 en un noeud baptisé 7, de probabilité $\frac{1}{5} + \frac{1}{4} = \frac{9}{20}$, d'où $E_2 = \{1,6,7\}$.
- Au troisième tour, on fusionne 1 et 6 en un noeud baptisé 8, de probabilité $\frac{1}{4} + \frac{3}{10} = \frac{11}{20}$, d'où $E_3 = \{7,8\}$.
- Au dernier tour, on fusionne 7 et 8 en un noeud baptisé 9, de probabilité $\frac{9}{20} + \frac{11}{20} = 1$.



Arbre de Huffman



Codage de Huffman

Cela donne le codage suivant.

i	1	2	3	4	5
$C(i)$	11	00	01	100	101

Le codage minimal n'est pas unique, loin de là.

Proposition 47 *Soit E un ensemble fini muni d'une distribution de probabilité p. L'algorithme qui vient d'être décrit produit un arbre dont les feuilles sont numérotées par E, de profondeur moyenne minimale.*

Preuve Par récurrence sur $|E|$, on montre que pour tout arbre A et toute bijection $\phi : E \to \partial A$, la profondeur moyenne de A, notée $\mathbb{E}(\ell_A \circ \phi)$ est supérieure ou égale à celle de l'arbre $(A_0, \phi_0 : E \to \partial A_0)$ construit par l'algorithme.

Initialisation. Lorsque $|E| = 1$, il n'y a qu'un arbre possible, $A = A_0$ ont même profondeur moyenne.

Supposons que l'algorithme construise un arbre optimal pour tous les ensembles E à n éléments. Soit E un ensemble ordonné à $n+1$ éléments, muni d'une distribution de probabilité p. Soit A un arbre et $\phi : E \to \partial A$ une bijection. Soit $(A_0, \phi_0 : E \to \partial A_0)$ l'arbre construit par l'algorithme.

Opération préliminaire. Soit $\{a, a'\}$ une paire de feuilles de profondeur maximale dans A, ayant le même ascendant. Soit $\{x, x'\}$ la paire d'éléments de E fusionnée au premier tour. En composant ϕ avec une permutation de E, remplaçons ϕ par $\tilde{\phi} = \phi \circ \sigma : E \to \partial A$, de sorte que $\tilde{\phi}(a) = x$ et $\tilde{\phi}(a') = x'$. La permutation σ répartit différemment les probabilités sans changer les profondeurs. Comme x et x' sont de probabilités minimales et a, a' de profondeurs maximales, on peut choisir σ de sorte qu'elle n'augmente pas la longueur moyenne, $\mathbb{E}(\ell_A \circ \phi) \geq \mathbb{E}(\ell_A \circ \tilde{\phi})$.

Supprimons les feuilles a et a'. On obtient un arbre A' et une bijection $\phi' : E' \to \partial A'$ qu'il faut comparer à l'arbre A_1, muni de sa bijection évidente $\phi_1 : E' \to \partial A_1$, produit par l'algorithme. La profondeur moyenne satisfait

$$\begin{aligned} \mathbb{E}(\ell_{A'} \circ \phi') &= \mathbb{E}(\ell_A \circ \tilde{\phi}) - p'(y), \\ \mathbb{E}(\ell_{A_1} \circ \phi_1) &= \mathbb{E}(\ell_{A_0} \circ \phi_0) - p'(y). \end{aligned}$$

Par l'hypothèse de récurrence, $\mathbb{E}(\ell_{A'} \circ \phi') \geq \mathbb{E}(\ell_{A_1} \circ \phi_1)$. Par conséquent, $\mathbb{E}(\ell_A \circ \tilde{\phi}) \geq \mathbb{E}(\ell_{A_0} \circ \phi_0)$, et donc $\mathbb{E}(\ell_A \circ \phi) \geq \mathbb{E}(\ell_{A_0} \circ \phi_0)$. ■

Suggestion d'exercice : n^{0}4, feuille 2.

Fin du cours n^{0}4

Corollaire 48 *Au jeu des 10 questions, la durée moyenne du jeu est au moins égale à l'entropie.*

Preuve D'après la Proposition 46, la question se traduit en termes de codage sans préfixes. Ce codage est en particulier uniquement décodable (on s'en est servi dans la preuve du théorème de codage de source), donc d'après le théorème de codage de source, sa longueur moyenne satisfait $\mathbb{E}(\ell(X)) \geq H(X)$. ■

Exemple 49 *Soit X une variable qui suit une loi géométrique de paramètre $\frac{1}{2}$ (par exemple, le temps d'obtention de face au jeu de pile ou face). Alors il existe une stratégie dont la durée moyenne est $H(X) = 2$ et on ne peut pas faire mieux.*

Voir l'Exercice 2 de la feuille 1, pour la stratégie de durée moyenne 2.

Suggestion d'exercice : n^{0}5, feuille 2.

Suggestion d'exercice : n^{0}6, feuille 2.

3.6 A retenir

- L'interprétation de l'entropie comme longueur moyenne minimale d'un codage d'une source (asymptotiquement).
- L'algorithme de Huffman.

Il s'agit d'un chapitre pratique, où on a utilisé les outils du chapitre précédent pour résoudre un problème concret.

4 Entropie des processus stationnaires

Dans ce chapitre, on cherche à établir le second principe de la thermodynamique. On étudie l'entropie de processus sensés modéliser des systèmes physiques. Il s'avère que leur entropie elle-même n'augmente pas, c'est l'entropie relative à la condition initiale qui augmente. On généralise ensuite la notion d'entropie aux processus aléatoires stationnaires. Pour un processus, c'est l'entropie "par symbole"qui remplace l'entropie d'une variable isolée. Ce taux de croissance se calcule particulièrement bien dans le cas des chaînes de Markov stationnaires.

4.1 Chaînes de Markov

Définition 50 *Une* chaîne de Markov *est une suite $X_0, X_1, \ldots$ de variables aléatoires à valeurs dans un ensemble E (l'ensemble des* états *de la chaîne) dont la dépendance au passé se résume à la dépendance à la dernière position. Autrement dit, pour tous états $x_0, \ldots, x_{n+1} \in E$,*

$$\mathbb{P}(X_{n+1} = x_{n+1} | X_n = x_n,\, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0) = \mathbb{P}(X_{n+1} = x_{n+1} | X_n = x_n).$$

On fera, sans le dire, l'hypothèse supplémentaire que la chaîne est indépendante du temps*, i.e. que les probabilités conditionnelles ne dépendent pas de n,*

$$\mathbb{P}(X_{n+1} = j | X_n = i) = P_{ij}.$$

La matrice $P = (P_{ij})_{i,\, j \in E}$ s'appelle la matrice des probabilités de transition.

Proposition 51 *Soit $X_0, X_1, \ldots$ une chaîne de Markov. La loi de X_n est entièrement déterminée par celle de X_0 et par la matrice P des probabilités de transition. Si on représente les distributions de probabilité par des vecteurs lignes, alors*

$$p_{X_n} = p_{X_{n-1}} P = p_{X_0} P^n.$$

Plus généralement, la loi jointe de $(X_0, X_1, \ldots, X_n)$ est déterminée par P et par la loi de X_0.

Preuve La loi de X_{n+1}

$$\begin{aligned}
\mathbb{P}(X_{n+1} = y) &= \sum_{x \in E} \mathbb{P}(X_{n+1} = y \text{ et } X_n = x) \\
&= \sum_{x \in E} \mathbb{P}(X_{n+1} = y | X_n = x) \mathbb{P}(X_n = x) \\
&= \sum_{x \in E} P_{xy} \mathbb{P}(X_n = x)
\end{aligned}$$

s'exprime en fonction de P et de la loi de X_n.

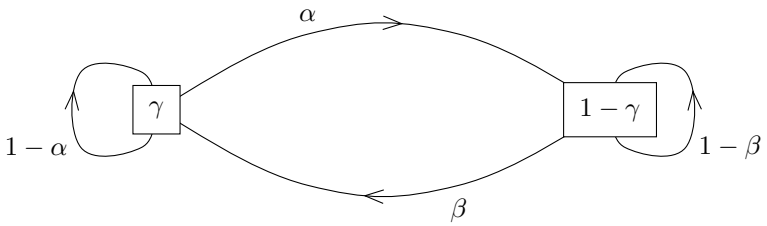
Plus généralement, par la définition d'une chaîne de Markov, la loi jointe

$$\begin{aligned}
p_{(X_0,\ldots,X_n)}(x_0, \ldots, x_n) &= \mathbb{P}(X_0 = x_0, \ldots, X_n = x_n) \\
&= \mathbb{P}(X_n = x_n | X_0 = x_0, \ldots, X_{n-1} = x_{n-1}) \mathbb{P}(X_0 = x_0, \ldots, X_{n-1} = x_{n-1}) \\
&= \mathbb{P}(X_n = x_n | X_{n-1} = x_{n-1}) \mathbb{P}(X_0 = x_0, \ldots, X_{n-1} = x_{n-1}) \\
&= P_{x_{n-1} x_n} \mathbb{P}(X_0 = x_0, \ldots, X_{n-1} = x_{n-1}) \\
&\vdots \\
&= P_{x_{n-1} x_n} P_{x_{n-2} x_{n-1}} \cdots P_{x_1 x_2} p_{X_0}(x_0)
\end{aligned}$$

s'exprime directement en fonction de P et p_{X_0}. ■

Remarque 52 *Représentation graphique. Une chaîne de Markov sur E peut être schématisée par un graphe orienté dont l'ensemble des sommets est E. Chaque arête orientée (x, y) porte la probabilité de transition P_{xy}. Chaque sommet x porte sa probabilité $p(x)$.*

Voici le schéma correspondant à la chaîne à 2 états de probabilités respectives γ et $1 - \gamma$ et de matrice de probabilités de transition $\begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$.



Graphe orienté schématisant une chaîne de Markov

On admettra le théorème suivant (on reviendra sur ce point au paragraphe 5.6).

Théorème 6 *Soit P une matrice* stochastique *(les coefficients sont positifs ou nuls, les sommes par ligne valent 1) de taille d. Soit μ une distribution de probabilité sur $\{1, \ldots, d\}$. Il existe une chaîne de Markov $X_0, X_1, \ldots$ à valeurs dans $\{1, \ldots, d\}$ dont la matrice des probabilités de transition est P et telle que la loi de X_0 est μ.*

Exemple 53 (Marches aléatoires) *Soit $G = (E, w)$ un graphe pondéré fini : il y a un ensemble fini E de sommets, et pour chaque paire de sommets distincts $\{x, y\}$, un réel positif ou nul $w(x, y) = w(y, x)$ ($w(x, y) = 0$ signifie que le graphe G ne contient pas l'arête xy). La marche aléatoire d'origine x_0 est la chaîne de Markov à valeurs dans E telle que $X_0 = x_0$ et dont la matrice des probabilités de transitions est définie par*

$$P_{xy} = \frac{w(x, y)}{\sum_{z \neq x} w(x, z)}.$$

si $x \neq y$, et $P_{xx} = 0$.

Autrement dit, on saute d'un sommet à l'un de ses voisins le long d'une arête choisie au hasard, avec probabilité proportionnelle à son poids.

4.2 Processus stationnaires

Définition 54 *Un processus aléatoire $\Xi = (X_n)_{n \in \mathbb{N} \text{ ou } \mathbb{Z}}$ est* stationnaire *si pour tout $k \in \mathbb{N}$, la loi jointe de $(X_{n+1}, \ldots, X_{n+k})$ ne dépend pas de n.*

Autrement dit, une translation dans le temps n'a pas d'effet sur la loi du processus.

Exemple 55 *Une suite de variables aléatoires indépendantes est un processus stationnaire si et seulement si les variables ont toutes même loi.*

Proposition 56 *Une chaîne de Markov $X_0, X_1, \ldots$ de loi initiale p et de matrice de probabilités de transition P est stationnaire si et seulement si $pP = p$.*

Preuve Si la chaîne est stationnaire, alors X_0 et X_1 ont même loi, $p = p_{X_0} = p_{X_1} = pP$. Réciproquement, si $p = pP$, alors $p_{X_1} = p$. Le processus $Y_n = X_{n+1}$ a même loi initiale et même matrice de probabilités de transition que X_n, donc mêmes lois jointes (Proposition 51). Cela entraîne que pour tout k et tout n, la loi jointe de $(X_n, \ldots, X_{n+k})$ coïncide avec la loi de $(X_0, \ldots, X_k)$. ■

Exemple 57 *Pour une chaîne de Markov à deux états, la matrice des probabilités de transition est de la forme $\begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$. Il y a une unique distribution stationnaire, le vecteur ligne $(\frac{\beta}{\alpha+\beta} \quad \frac{\alpha}{\alpha+\beta})$.*

On admettra le théorème élémentaire suivant (algèbre linéaire).

Théorème 7 *Soit P une matrice* stochastique *(les coefficients sont positifs ou nuls, les sommes par ligne valent 1). On suppose P* irréductible *(il existe une puissance de P dont tous les coefficients sont strictement positifs) et* apériodique *(pour tout i, le pgcd des entiers n tels que $(P^n)_{ii} > 0$ vaut 1). Alors il existe une unique distribution de probabilité stationnaire μ. Pour toute chaîne de Markov de matrice de probabilités de transitions P, la loi de X_n tend vers μ quand n tend vers l'infini.*

Exemple 58 *Pour la marche aléatoire sur le graphe pondéré $G = (E, w)$, la distribution de probabilité suivante*

$$\mu(x) = \frac{\sum_{z \neq x} w(x, z)}{\sum_{z \in E} \sum_{z' \neq z} w(z, z')}$$

est stationnaire.

Voir Exercice 1, feuille 3.

Suggestion d'exercice : n^{0}1, feuille 3.

4.3 Second principe de la thermodynamique

On considère qu'une chaîne de Markov constitue un bon modèle d'un système physique isolé. En effet, le caractère aléatoire résulte de la simplification d'un modèle déterministe (on substitue aux variables microscopiques des observables macroscopiques obtenues en faisant des moyennes), conformément au paradigme de la physique statistique. La condition sur la dépendance par rapport au passé traduit le fait qu'au niveau microscopique, le système suit une évolution déterministe gouvernée par des équations différentielles (une position et une vitesse initiales, i.e. une position dans un espace des phases, déterminent l'évolution future).

L'entropie des physiciens est le logarithme du nombre de configurations microscopiques correspondant à l'état macroscopique du système. On donne fréquemment l'exemple suivant. Soit un système isolé comportant N molécules identiques, qui peuvent occuper différents états numérotés de 1 à ℓ. On considère que l'état macroscopique du système est décrit par les effectifs $N_1, \ldots, N_\ell$ de chaque état, où, de façon équivalente, par les proportions $p_i = \frac{N_i}{N}$. Un même état macrocopique est réalisé à l'échelle microscopique de multiples manières. Le nombre de façons de répartir les N molécules entre les états dans les proportions prescrites p_i est le coefficient multinômial

$$\nu(p_1, \ldots, p_\ell) = \frac{N!}{N_1! \ldots N_\ell}.$$

D'après Boltzmann, l'entropie du système dans l'état macroscopique $p = (p_1, \ldots, p_\ell)$ est

$$S = k \log \nu(p),$$

où k est une constante physique, la *constante de Boltzmann*. En utilisant la formule de Stirling, on voit que, lorsque les N_i sont grands,

$$\begin{aligned} S &\sim N \log_2 N - \sum_{i=1}^{\ell} N_i \log_2 N_i \\ &= N(-\sum_{i=1}^{\ell} \frac{N_i}{N} \log_2 \frac{N_i}{N}) \\ &= NH(p), \end{aligned}$$

donc l'entropie par molécule vaut $H(p)$.

Il semble que seules les chaînes de Markov dont la distribution stationnaire est uniforme aient une réalité physique.

Théorème 8 *Soit $X_0, X_1, \ldots$ une chaîne de Markov. On suppose que la distribution uniforme est stationnaire. Alors l'entropie augmente : la suite $H(X_n)$ est croissante.*

Fin du cours n^{0}5

Remarque 59 *Il est facile de caractériser les chaînes de Markov pour lesquelles la distribution uniforme est stationnaire. Voir Exercice 2, feuille 3.*

Suggestion d'exercice : n^{0}2, feuille 3.

Le Théorème 8 résulte du fait plus général suivant.

Théorème 9 *Soient μ et μ' deux distributions de probabilité sur un ensemble fini E. Soient $\mu_n = \mu P^n$ et $\mu'_n = \mu' P^n$ les évolutions de ces distributions sous une même matrice stochastique P. Alors l'entropie relative $D(\mu_n, \mu'_n)$ décroît.*

Preuve du Théorème 9.

Pour $(x, y) \in E \times E$, on pose $p_n(x, y) = \mu_n(x)P_{xy}$ et $q_n(x, y) = \mu'_n(x)P_{xy}$. Par construction,

$$p_{n,X} = \mu_n, \quad q_{n,X} = \mu'_n, \quad p_{n,Y} = \mu_{n+1}, \quad q_{n,Y} = \mu'_{n+1},$$

et, pour $(x, y) \in E \times E$,

$$p_{n,Y|X=x}(y) = \frac{\mu_n(x)P_{xy}}{p_{n,X}(x)} = P_{xy} = q_{n,Y|X=x}(y),$$

les $D(p_{n,Y|X=x}||q_{n,Y|X=x})$ sont toutes nulles, donc $D(p_{n,Y|X}||q_{n,Y|X}) = 0$. En utilisant deux fois la proposition 36, on obtient

$$\begin{aligned} D(\mu_n||\mu'_n) &= D(p_{n,X}||q_{n,X}) \\ &= D(p_{n,X}||q_{n,X}) + D(p_{n,Y|X}||q_{n,Y|X}) \\ &= D(p_n||q_n) \\ &= D(p_{n,Y}||q_{n,Y}) + D(p_{n,X|Y}||q_{n,X|Y}) \\ &\geq D(\mu_{n+1}||\mu'_{n+1}), \end{aligned}$$

car $D(p_{n,X|Y}||q_{n,X|Y}) \geq 0$. ■

Preuve du Théorème 8.

On applique le Théorème 9 à la loi μ de X_0 et à la distribution uniforme μ'. Par hypothèse, μ' est stationnaire, donc $\mu'_n = \mu'$. En revanche, μ_n est la loi de X_n, qui dépend de n. On sait que $D(\mu_n, \mu') = \log|E| - H(\mu_n)$, donc $H(\mu_n) = H(X_n)$ décroît. ■

Suggestion d'exercice : n^{0}3, feuille 3.

Suggestion d'exercice : n^{0}4, feuille 3.

Pour une autre manifestation de la croissance de l'entropie, voir l'Exercice 5, feuille 3.

Suggestion d'exercice : n^{0}5, feuille 3.

Suggestion d'exercice : n^{0}6, feuille 3.

4.4 Les langues naturelles comme processus stochastiques

Au chapitre précédent, on a vu une langue naturelle comme une source, de lettres ou de mots. Autrement dit, comme une distribution de probabilité sur l'ensemble des 27 lettres (inclure l'espace), ou sur l'ensemble des mots du dictionnaire.

Dans l'hypothèse où les textes seraient des tirages aléatoires indépendants (de lettres ou de mots), l'entropie de cette distribution mesure la probabilité des textes typiques. Ce n'est réaliste que pour les textes produits par un singe tapant à la machine ou tirant des mots dans un dictionnaire.

Une modélisation plus fidèle du langage consiste à voir un texte comme une réalisation d'un processus stochastique $\Xi = (X_1, X_2, \ldots)$ non nécessairement indépendant. En première approximation, on peut penser que ce processus est stationnaire : les règles de production de phrases varient suffisamment lentement dans le temps pour qu'on puisse négliger cette variation. Ce qui nous renseigne sur la probabilité de phrases typiques, c'est l'entropie jointe $H(X_1, \ldots, X_n)$. Pour les phrases écrites par le singe, les X_i sont indépendantes, donc $H(X_1, \ldots, X_n) = nH(X_1)$ est proportionnelle aux nombre de mots. Pour avoir une quantité qui ne tend pas vers l'infini, on considère l'entropie "par symbole" (lettre ou mot) $\frac{1}{n}H(X_1, \ldots, X_n)$ même dans le cas général.

4.5 Entropie par symbole

Définition 60 *Soit $\Xi = (X_n)_{n\in\mathbb{N}}$ un processus aléatoire. L'entropie par symbole du processus est la limite*

$$H(\Xi) = \lim_{n\to\infty} \frac{1}{n} H(X_1, \ldots, X_n),$$

lorsqu'elle existe.

Remarque 61 *Cette limite n'existe pas toujours. Par exemple, pour une suite de variables indépendantes $\frac{1}{n}H(X_1,\ldots,X_n) = \frac{1}{n}\sum_{i=1}^n H(X_i)$ peut faire à peu près n'importe quoi.*

Théorème 10 *Soit $\Xi = (X_n)_{n\in\mathbb{N}}$ un processus stationnaire. Alors les limites suivantes existent et sont égales.*

$$H(\Xi) = \lim_{n\to\infty} \frac{1}{n} H(X_1, \ldots, X_n) = \lim_{n\to\infty} H(X_n | X_{n-1}, \ldots, X_1).$$

Preuve D'après le principe "conditionner diminue l'entropie",

$$H(X_{n+1}|X_n, \ldots, X_1) \leq H(X_{n+1}|X_{n-1}, \ldots, X_2) = H(X_n|X_{n-1}, \ldots, X_1)$$

par stationnarité. Une suite décroissante positive ayant une limite, la suite $u_n = H(X_n|X_{n-1}, \ldots, X_1)$ converge. D'autre part, d'après la Proposition 24,

$$H(X_1, \ldots, X_n) = \sum_{i=1}^n H(X_i|X_{i-1}, \ldots, X_1) = \sum_{i=1}^n u_i.$$

Pour tout suite convergente u_n, $v_n = \frac{1}{n}\sum_{i=1}^n u_i$ converge vers la même limite. ∎

Exemple 62 *Si Ξ est une suite de variables indépendantes de même loi p, alors $H(\Xi) = p$.*

Il faut donc penser à l'entropie par symbole $H(\Xi)$ comme une généralisation aux processus de l'entropie d'une variable isolée.

Proposition 63 *Soit $\Xi = (X_n)_{n\in\mathbb{N}}$ une chaîne de Markov stationnaire. Alors*

$$H(\Xi) = H(X_1|X_0) = -\sum_{x,\, y\in E} \mu(x) P_{xy} \log_2(P_{xy}),$$

où μ désigne la distribution de X_0, qui est stationnaire.

Preuve Par définition d'une chaîne de Markov, la loi conditionnelle de $X_n|X_{n-1} = x_{n-1}, \ldots, X_0 = x_0$ coïncide avec la loi conditionnelle $X_n|X_{n-1} = x_{n-1}$, donc $H(X_n|X_{n-1}, \ldots, X_1) = H(X_n|X_{n-1}) = H(X_1, X_0)$ par stationnarité.

Par définition de la matrice des probabilités de transition, la loi conditionnelle $p_{X_1|X_0=x}(y) = P_{xy}$, donc $H(p_{X_1|X_0=x}) = -\sum_{y\in E}$ et $H(X_1, X_0) = -\sum_{x,\, y\in E} \mu(x) P_{xy} \log_2(P_{xy})$. ∎

Corollaire 64 *Soit $\Xi = (X_n)_{n\in\mathbb{N}}$ une chaîne de Markov irréductible et apériodique. Alors l'entropie par symbole est bien définie et vaut $H(\Xi) = -\sum_{x,\, y\in E} \mu(x) P_{xy} \log_2(P_{xy})$, où μ est la distribution stationnaire et P la matrice des probabilités de transition.*

Exemple 65 *Soit $G = (E, w)$ un graphe pondéré fini. On pose, pour $x \in E$, $W(x) = \sum_{y\neq x} w(x, y)$ et $W = \sum_{x\in E} W(x)$. On introduit deux distributions de probabilité, $p(x, y) = \frac{w(x,y)}{W}$ sur $E \times E$ et $\mu(x) = \frac{W(x)}{W}$ sur E. Soit Ξ une marche aléatoire sur G. On suppose qu'elle est irréductible et apériodique. Alors $H(\Xi) = H(p) - H(\mu)$.*

Voir Exercice 7, feuille 3.

Suggestion d'exercice : n^{0}7, feuille 3.

Suggestion d'exercice : n^{0}8, feuille 3.

4.6 Entropie par symbole d'une langue naturelle

La mesure directe de l'entropie par lettre (ou par mot) d'une langue naturelle est malaisée. Par exemple, si X_n désigne la n-ème lettre d'un texte tiré au hasard, les lois conditionnelles de $X_5|X_4 = x_4, X_3 = x_3, X_2 = x_2, X_1 = x_1$ sont au nombre de $27^4 = 531441$. Evaluer directement ces probabilités nécessite de traiter des millions de bouts de textes. Aller loin au delà n'est pas envisageable, sans parler d'évaluer une limite.

En 1951, Shannon a suggéré de procéder autrement, au moyen d'un protocole expérimental faisant intervenir des volontaires (étudiants) à qui on présente des textes tronqués, de longueur 75. Pour chaque texte, chaque volontaire doit deviner la lettre suivante. Il doit fournir une liste de 27 lettres ordonnée par probabilité décroissante. Dans chaque liste, l'expérimentateur note le rang de la lettre qui était effectivement la suivante dans le texte examiné. Par exemple, confronté au texte "*allons enfants de la patrie le jour de gloire est arrive contre nous de la* ", le volontaire répond la liste ordonnée : *m, t, e, c, a, b,...,* et l'expérimentateur note : 2. Il fait de même avec 11 autres volontaires, confrontés à 10 textes chacun, soigneusement tirés au hasard dans un gros livre. Il obtient 120 notes comprises entre 1 et 27, dont il fait un histogramme, i.e. le tableau des effectifs $N_1, \ldots, N_{27}$. Il calcule une *entropie empirique*

$$h = -\sum_{i=1}^{27} \frac{N_i}{120} \log_2(\frac{N_i}{120}).$$

L'expérience, conduite par Shannon en 1950, a donné une valeur de 1.3 bits par lettre pour la langue anglaise.

Pourquoi cette valeur expérimentale constitue t'elle une valeur approchée de l'entropie de la langue anglaise ? On fait l'hypothèse que les volontaires connaissent bien leur langue, et qu'ils produisent tous leurs listes en suivant le même raisonnement déterministe. La note ne dépend alors que du texte. Si le texte est tiré au hasard, cette note devient une variable aléatoire Y à valeurs dans $E = \{1, \ldots, 27\}$. Cette variable contient la même information que la variable X_{76} : étant donné un texte $t = t_1 \ldots t_{76}$, la 76ème lettre $t_{76} = X_{76}(t)$ de t est uniquement déterminée par $t_1 \ldots t_{75}$ et $Y(t_1 \ldots t_{75})$. Donc les lois conditionnelles de Y et de X_{76} sachant les 75 premières lettres sont les mêmes, à une bijection de E sur l'alphabet près. En particulier, l'entropie $H(Y) = H(X_{76}|X_1, \ldots, X_{75})$. L'expérience constitue une simulation de la variable Y. D'après la loi des grands nombres, quand le nombre de volontaires tend vers l'infini, l'histogramme des notes converge vers la loi de Y, donc l'entropie empirique calculée à partir de cet histogramme converge vers $H(X_{76}|X_1, \ldots, X_{75})$. Celle-ci constitue un majorant de l'entropie par lettre de la langue anglaise. Les hésitations et erreurs introduites par les volontaires tendent à augmenter l'entropie de Y, cela va dans le même sens : la valeur réelle de l'entropie par lettre de la langue anglaise est inférieure à la valeur empirique.

Comment interpréter la valeur numérique 1.3 bits par lettre ? Il faut d'abord la comparer au maximum possible, $\log_2(27) = 4.75$, atteinte lorsque les lettres sont tirées indépendamment selon la distribution uniforme. Lorsqu'on tire les lettres indépendamment selon la distribution qu'elles ont en anglais, donnée par le tableau suivant,

lettre	e	t	a	o	i	n	s	h	r	d	l	c	u
%	13	9	8	7,5	7	6,8	6,5	6,2	6	4	3,8	2,8	2,7
lettre	m	w	f	g	y	p	b	v	k	j	x	q	z
%	2,5	2,4	2,3	2,1	2	1,9	1,5	1,1	0,9	0,2	0,2	0,1	0,1

l'entropie vaut 4.2. Une meilleure approximation est obtenue par une chaîne de Markov stationnaire dont la matrice des probabilités de transition est tirée du tableau des fréquences des couples de lettres en anglais. L'entropie par lettre de cette chaîne vaut 4.03 bits par lettre. On peut raffiner en tirant chaque lettre supplémentaire en respectant la loi jointe de 3 ou de 4 lettres, dont l'entropie vaut 2.3 bits par lettre. Une entropie nettement plus basse, de 1.3 bits par lettre, signifie que l'anglais est bien plus déterministe, du fait des règles de grammaire et de construction des mots.

4.7 A retenir

- La notion de chaîne de Markov.
- La notion d'entropie par symbole.
- Le cas des chaînes de Markov.

Il s'agit d'un chapitre mi-théorique, mi-pratique, où on donne une définition précise à une notion qui provient de problèmes concrets, tout en posant les bases de développements ultérieurs.

5 Equipartition asymptotique

Dans cette section, on donne une nouvelle preuve du théorème du codage de source, basée sur l'idée que la loi jointe de variables iid est concentrée.

5.1 Comportement typique pour une suite de variables iid

Notation 66 *Un* processus aléatoire *à valeurs dans un ensemble E, c'est simplement une suite de variables aléatoires $\Xi = (X_n)_{n\in I}$ à valeurs dans E, indexée par $I = \mathbb{N}$, $\mathbb{N}\setminus\{0\}$ ou $\mathbb{Z}$. Pour $m \leq n$, on note $X_m^n = (X_m, X_{m+1}, \ldots, X_n)$. On note p_m^n sa loi, une distribution de probabilité sur E^n.*

On s'intéresse à la probabilité de sortie d'une suite $(x_1, \ldots, x_n) \in E^n$, $p_1^n(x_1, \ldots, x_n)$, vue comme variable aléatoire. Autrement dit, p_1^n est une fonction sur E^n, et on considère la variable aléatoire $p_1^n(X_1, \ldots, X_n)$, qu'on peut noter $p_1^n(X_1^n)$.

Question 67 *Etant donné un processus aléatoire $\Xi = (X_n)_{n\in\mathbb{N}}$, quel est le comportement asymptotique de la variable aléatoire $p_1^n(X_1^n)$ lorsque n tend vers l'infini ?*

On va formaliser le raisonnement indiqué au début du cours (paragraphe 1.2).

Rappel 68 (Loi des grands nombres) *Soit $\Xi = (X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires indépendantes, de même loi, intégrables (i.e. $\mathbb{E}(|X_0|) < +\infty$). Alors la suite $S_n = \frac{1}{n}\sum_{i=1}^n X_i$ converge presque sûrement vers la constante $\mathbb{E}(X_0)$. En particulier, elle converge en probabilité, i.e. pour tout $\epsilon > 0$, la probabilité $\mathbb{P}(|S_n - \mathbb{E}(X_0)| > \epsilon)$ tend vers 0 quand n tend vers $+\infty$.*

A toutes fins utiles, on donne au paragraphe 5.11 une preuve de la loi des grands nombres sous l'hypothèse (suffisante pour l'application ci-dessous) que X_0 est bornée. Et on explique pourquoi convergence presque sûre implique convergence en probabilité.

Théorème 11 (Equipartition asymptotique) *Soit $\Xi = (X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires indépendantes, de même loi p. Alors $-\frac{1}{n}\log_2 p_1^n(X_1^n)$ converge presque partout et en probabilité vers la constante $H(p)$. En particulier, pour tout $\epsilon > 0$, la probabilité $\mathbb{P}(|-\frac{1}{n}\log_2 p_1^n(X_1^n) - H(p)| > \epsilon)$ tend vers 0 quand n tend vers $+\infty$.*

Preuve Par indépendance, $p_1^n(x_1, \ldots, x_n) = p(x_1)\cdots p(x_n)$. Les variables $Z_i = -\log_2 p(X_i)$ sont indépendantes, donc, d'après la loi des grands nombres

$$-\frac{1}{n}\log_2 p_1^n(X_1^n) = \frac{1}{n}\sum_{i=1}^n Z_i \to \mathbb{E}(-\log_2 p(X_0)) = H(p),$$

en probabilité. ∎

Interprétation : Etant donné $\epsilon > 0$ et $n \in \mathbb{N}$, soit T le sous-ensemble de E^n défini par

$$T = \{t \in E^n\,;\, 2^{-n(H(p)+\epsilon)} \leq p_1^n(t) \leq 2^{-n(H(p)-\epsilon)}\},$$

et soit S son complémentaire. Alors $p_1^n(S)$ tend vers 0 quand n tend vers $+\infty$ à ϵ fixé. De plus, comme p est presque constante sur T, le nombre d'éléments de T est de l'ordre de $2^{nH(p)}$. Précisément, pour n assez grand,

$$(1-\epsilon)2^{n(H(p)-\epsilon)} \leq |T| \leq 2^{n(H(p)+\epsilon)}.$$

On pense à T comme à l'ensemble des "suites typiques", ce sont celles qui ont le plus de chances de sortir.

Suggestion d'exercice : n^{0}1, feuille 4.

5.2 Suites conjointement typiques

Il s'agit d'une variante de l'équipartition qui permet de donner une interprétation à l'information mutuelle. De plus, cette notion va être à l'origine d'un procédé de décodage utilisé au chapitre 6.

Définition 69 *Considérons deux variables aléatoires X et Y de loi jointe $p_{(X,Y)}$. Soient $n \in \mathbb{N}$ et $\epsilon > 0$. L'ensemble $T = T(n,\epsilon) \subset (E_X \times E_Y)^n$ des suites* conjointement typiques *est l'ensemble des suites $(x,y)_1^n$ telles que*

1. $|-\frac{1}{n}\log_2(p_X(x_1)\cdots p_X(x_n)) - H(X)| < \epsilon$.
2. $|-\frac{1}{n}\log_2(p_Y(y_1)\cdots p_Y(y_n)) - H(Y)| < \epsilon$.
3. $|-\frac{1}{n}\log_2(p_{(X,Y)}(x_1,y_1)\cdots p_{(X,Y)}(x_n,y_n)) - H(X,Y)| < \epsilon$.

Autrement dit, x_1^n est typique pour p_X, y_1^n est typique pour p_Y et $((x_1,y_1),\dots(x_n,y_n))$ est typique pour $p_{(X,Y)}$.

Proposition 70 *Soit $(X,Y)_1^n$ une suite de variables indépendantes et de même loi $p_{(X,Y)}$. Alors, pour n assez grand,*

1. $\mathbb{P}((X,Y)_1^n \in T)$ *tend vers* 1 *quand n tend vers $+\infty$.*
2. $(1-\epsilon)2^{n(H(X,Y)+\epsilon)} \leq |T| \leq 2^{n(H(X,Y)+\epsilon)}$.
3. *Soient $\tilde{X}_1^n$ et $\tilde{Y}_1^n$ des variables indépendantes, de lois respectives p_X et p_Y. Alors, pour n assez grand,*

$$(1-\epsilon)2^{-n(I(X;Y)+3\epsilon)} \leq \mathbb{P}((\tilde{X},\tilde{Y})_1^n \in T) \leq 2^{-n(I(X;Y)-3\epsilon)}.$$

Preuve 1. On applique la loi des grands nombres aux trois variables

$$-\log_2 p_X(X_n), \quad -\log_2 p_Y(Y_n), \quad -\log_2 p_{(X,Y)}(X_n,Y_n).$$

Presque sûrement, les trois convergent respectivement vers $H(X)$, $H(Y)$ et $H(X,Y)$. En particulier, il y a convergence en probabilité.

2. Pour chaque $t \in T$, pour n assez grand,

$$2^{-n(H(X,Y)+\epsilon)} \leq \mathbb{P}((X,Y)_1^n = t) \leq 2^{-n(H(X,Y)-\epsilon)}.$$

On écrit

$$1 \geq \mathbb{P}((X,Y)_1^n \in T) = \sum_{t\in T} \mathbb{P}((X,Y)_1^n = t) \geq |T| 2^{-n(H(X,Y)+\epsilon)},$$

D'où $|T| \leq 2^{n(H(X,Y)+\epsilon}$. Inversement, pour n assez grand, $\mathbb{P}((X,Y)_1^n \in T) \geq 1-\epsilon$, ce qui s'écrit

$$1-\epsilon \leq \sum_{t\in T} \mathbb{P}((X,Y)_1^n = t) \leq |T| 2^{-n(H(X,Y)-\epsilon)},$$

soit $|T| \geq (1-\epsilon)2^{n(H(X,Y)-\epsilon)}$.

3. Par hypothèse, pour $t = ((x_1,y_1),\dots,(x_n,y_n)) \in T$,

$$\begin{aligned} P((\tilde{X},\tilde{Y})_1^n = t) &= p_X(x_1)\cdots p_X(x_n)p_Y(y_1)\cdots p_Y(y_n) \\ &\leq 2^{-n(H(X)-\epsilon)}2^{-n(H(Y)-\epsilon)}, \end{aligned}$$

d'où

$$\mathbb{P}((\tilde{X},\tilde{Y})_1^n \in T) \leq 2^{-n(H(X)-\epsilon)}2^{-n(H(Y)-\epsilon)}|T| \leq 2^{-n(H(X,Y)-H(X)-H(Y)-3\epsilon)}.$$

De même, pour $t \in T$,

$$P((\tilde{X},\tilde{Y})_1^n = t) \geq 2^{-n(H(X)+\epsilon)}2^{-n(H(Y)+\epsilon)},$$

d'où

$$\mathbb{P}((\tilde{X},\tilde{Y})_1^n \in T) \geq 2^{-n(H(X)+\epsilon)}2^{-n(H(Y)+\epsilon)}|T| \geq (1-\epsilon)2^{-n(H(X)+H(Y)-H(X,Y)+3\epsilon)}.$$

■

Fin du cours n^{0}6

5.3 Application au codage de source

Proposition 71 *Soit $\Xi = (X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires indépendantes, de même loi p. Soit $\mathcal{D}$ un alphabet à $D = |\mathcal{D}|$ lettres. Soit $\epsilon > 0$. Il existe $n_0 \in \mathbb{N}$ tel que, pour tout $n \geq n_0$, il se produit la chose suivante.*

1. *Pour toute application injective $C^* : E^n \to \mathcal{D}^*$, la longueur moyenne*

$$\mathbb{E}(\frac{1}{n}\ell(X_1^n)) \geq \frac{1}{\log_2 D}H(p) - \epsilon.$$

2. *Il existe une application injective $C^* : E^n \to \mathcal{D}^*$ dont la longueur moyenne*

$$\mathbb{E}(\frac{1}{n}\ell(X_1^n)) \leq \frac{1}{\log_2 D}H(p) + \epsilon.$$

Le premier énoncé est un peu plus fort que celui du Théorème 5. En effet, si $C : E \to \mathcal{D}^*$ est un codage uniquement décodable, son extension aux textes $C^* : E^* \to \mathcal{D}^*$ est injective, et en particulier sa restriction à l'ensemble E^n des textes à n mots est injective. De plus, dans le contexte du Théorème 5, $C^*(x_1^n) = C(x_1)\cdots C(x_n)$, donc $\ell(x_1^n) = \sum_{i=1}^n \ell(x_i)$, $\mathbb{E}(\ell(X_1^n)) = \sum_{i=1}^n \mathbb{E}(\ell(X_i)) = n\mathbb{E}(\ell(X_0))$. En faisant varier ϵ et en faisant tendre n vers l'infini, l'inégalité $\mathbb{E}(\ell(X_0)) = \mathbb{E}(\frac{1}{n}\ell(X_1^n)) \geq \frac{1}{\log_2 D}H(p) - \epsilon$ entraîne que $\mathbb{E}(\ell(X_0)) \geq \frac{1}{\log_2 D}H(X_0) = \frac{H(p)}{\log_2 D}$. Donc la proposition 71 implique bien la borne inférieure pour la longueur moyenne des codages uniquement décodables donnée par le Théorème 5. C'était d'ailleurs l'argument que Shannon avait en tête.

Le second énoncé ne se compare pas directement à celui de Shannon. Néanmoins, il indique bien que la borne par l'entropie est essentiellement optimale.

Preuve On étend à D quelconque le résultat de l'exercice 3 de la feuille 2, limité au cas où $D = 2$. Par hypothèse, C^* est un codage non singulier de E^n. Dans un codage non singulier optimal, les D premiers éléments sont codés par des suites de longueur 1, les D^2 suivant par des suites de longueur 2, etc... La longueur $\ell(x_i)$ du i-ème élément de E satisfait

$$\sum_{k=1}^{\ell(x_i)-1} D^k < i \leq \sum_{k=1}^{\ell(x_i)} D^k,$$

i.e.

$$\frac{D^{\ell(x_i)}-1}{D-1} - 1 < i \leq \frac{D^{\ell(x_i)+1}-1}{D-1} - 1,$$

ce qui s'écrit aussi

$$\ell(x_i) < \log_D((D-1)(i+1)+1) \leq \ell(x_i)+1,$$

soit $\ell(x_i) = \lceil \log_D((D-1)(i+1)+1)\rceil = \lceil \log_D(\frac{(D-1)i}{D}+1)\rceil$. D'où l'inégalité

$$\mathbb{E}(\ell(X)) \geq \sum_{i=1}^{|E|} p(x_i)\lceil \log_D(\frac{(D-1)i}{D}+1)\rceil \geq M := \sum_{i=1}^{|E|} p(x_i)\log_D(\frac{(D-1)i}{D}+1).$$

On pose

$$c = \sum_{i=1}^{|E^n|} \frac{1}{\frac{(D-1)i}{D}+1}, \quad q(x_i) = \frac{1}{c}\frac{1}{\frac{(D-1)i}{D}+1}.$$

Comme q est une distribution de probabilité, on peut appliquer le Théorème 2,

$$\begin{aligned} 0 &\leq \frac{D(p||q)}{\log_2(D)} = \sum_{i=1}^{|E|^n} p(x_i)\log_D p(x_i) - \sum_{i=1}^{|E|^n} p(x_i)\log_D q(x_i) \\ &= -\frac{H(p)}{\log_2 D} + \sum_{i=1}^{|E|^n} p(x_i)\log_D(c(\frac{(D-1)i}{D}+1)) \\ &= -\frac{H(p)}{\log_2 D} + \log_D(c) + M, \end{aligned}$$

d'où $\frac{H(p)}{\log_2 D} - M \leq \log_D(c)$.

Il reste à majorer c. On utilise l'encadrement de la série harmonique par le logarithme népérien : si $h(k) = \sum_{j=1}^{n} \frac{1}{j}$, alors $h(k) \leq \ell n(k)+1$. Ici,

$$c = \sum_{i=1}^{|E|^n} \frac{1}{\frac{(D-1)i}{D}+1} = \frac{D}{D-1}\sum_{i=1}^{|E|^n} \frac{1}{i+\frac{D}{D-1}} \leq 2(h(|E|^n+1)-1) \leq 2\ell n(|E|^n+1)-2 \leq 2n\ell n(|E|).$$

On voit que $\log_D(c) \leq O(\log(n))$. D'où la minoration $\mathbb{E}(\frac{1}{n}\ell(X)) \geq \frac{H(p)}{\log_2 D} - O(\frac{\log n}{n}) \geq \frac{H(p)}{\log_2 D} - \epsilon$ pour n assez grand.

Inversement, soit $T = T(n,\epsilon)$ l'ensemble des suites typiques et S son complémentaire. Soit L le plus petit entier tel que $|T| \leq D^L$. Alors $L \leq n\frac{(H(p)+\epsilon)}{\log_2 D}+1$. Il existe une application injective $C : T \to \mathcal{D}^L$, on choisit n'importe laquelle. Soit L' le plus petit entier $> L$ tel que $|E^n| \leq D^{L'}$. Alors $L' \leq n\frac{\log_2|E|}{\log_2 D}+1$. Sur S, on choisit une application injective à valeurs dans $\mathcal{D}^{L'}$. On calcule

$$\begin{aligned} \mathbb{E}(\frac{1}{n}\ell(X_1^n)) &= \sum_{t\in T} p_1^n(t)\frac{L}{n} + \sum_{s\in S} p_1^n(s)\frac{L'}{n} \\ &\leq \mathbb{P}(X_1^n \in T)\frac{H(p)+\epsilon}{\log_2 D} + \mathbb{P}(X_1^n \in S)\frac{\log_2|E|}{\log_2 D} + o(1), \end{aligned}$$

qui tend vers $\frac{H(p)+\epsilon}{\log_2 D}$ quand n tend vers l'infini. ∎

5.4 Equirépartition asymptotique pour les processus stationnaires

On va étendre aux processus aléatoire stationnaires le théorème d'équirépartition asymptotique. On n'ira pas jusqu'au bout de la preuve, assez difficile, mais on établira la traduction du problème dans le langage des systèmes dynamiques. Cela nous conduira naturellement à deux pierres angulaires de cette théorie, la notion d'ergodicité et le Théorème Ergodique de Birkhoff.

Comme au paragraphe 5.1, étant donné un processus aléatoire Ξ, on s'intéresse à la probabilité de sortie d'une suite $(x_1, \ldots, x_n) \in E^n$, $p_1^n(x_1, \ldots, x_n)$, vue comme variable aléatoire. Autrement dit, p_1^n est une fonction sur E^n, et on considère la variable aléatoire $p_1^n(X_1, \ldots, X_n)$, qu'on peut noter $p_1^n(X_1^n)$. On vise le théorème suivant.

Théorème 12 (Shannon, McMillan, Breiman) *Soit $\Xi = (X_n)_{n\in\mathbb{Z}}$ un processus stationnaire d'entropie par symbole $H(\Xi)$. Alors*

$$\lim_{n\to\infty} -\frac{1}{n}\log_2 p_1^n(X_1^n) = H(\Xi) \quad \textit{presque sûrement.}$$

On renvoie au Chapitre 16.8 du livre de Cover et Thomas. On va présenter quelques ingrédients de la preuve.

5.5 Schéma de la preuve

L'espérance de la variable $-\log_2 p_1^n(X_1^n)$ est l'entropie jointe $H(X_1^n)$. Il s'agit donc de montrer qu'une variable aléatoire se concentre autour de son espérance. Comme dans la preuve de l'équirépartition asymptotique pour les suites iid, on voudrait utiliser la loi des grands nombres. Pour faire apparaître une moyenne de variables, on introduit les lois conditionnelles

$$p(x_i|x_1^{i-1}) = \mathbb{P}(X_i = x_i | X_{i-1} = x_{i-1}, \ldots, X_1 = x_1),$$

(vue ici comme une simple fonction sur E^{i+1}) et on écrit que

$$p_1^n(x_1^n) = p(x_i)\prod_{i=2}^{n} p(x_i|x_1^{i-1}),$$

d'où

$$p_1^n(X_1^n) = p(X_i)\prod_{i=2}^{n} p(X_i|X_1^{i-1}),$$

Les variables $-\log_2 p(X_i|X_1^{i-1})$ n'étant pas indépendantes, la loi des grands nombres ordinaire (Proposition 78) ne s'applique pas. La forme la plus générale de la loi des grands nombres, c'est le Théorème Ergodique de Birkhoff. Il se formule en termes de transformation préservant une mesure de probabilité. On va maintenant expliquer le lien entre processus stationnaires et transformations préservant une mesure de probabilité.

Fin du cours n^{0}7

5.6 Processus stationnaires et transformations préservant une mesure de probabilité

Proposition 72 *Soit $\Xi = ((X_n)_{n\in\mathbb{Z}})$ un processus aléatoire stationnaire à valeurs dans un ensemble fini E. Sur l'ensemble $E^{\mathbb{Z}}$ des suites bi-infinies d'éléments de E, muni de la tribu $\mathcal{B}$ engendrée par les cylindres, il existe une unique mesure de probabilité μ possédant les propriétés suivantes.*

- *μ est invariante par le* décalage *$T : (x_n)_{n\in\mathbb{Z}} \mapsto (y_n)_{n\in\mathbb{Z}}$, où $y_n = x_{n+1}$.*
- *Le processus Υ formé par les fonctions coordonnées $Y_m((x_n)_{n\in\mathbb{Z}}) = x_m$ a les mêmes lois jointes que Ξ.*

Réciproquement,

Autrement dit, on peut supposer que X_n est la n-ème fonction coordonnée sur $E^{\mathbb{Z}}$, ce qu'on fera dans la suite.

Preuve Etant donnés $m \leq n \in \mathbb{Z}$ et $z_m^n \in E^{n-m+1}$, on appelle cylindre l'ensemble

$$C_{z_m^n} = \{(x_n)_{n\in\mathbb{Z}}\,;\, x_m = z_m, \ldots, x_n = z_n\}.$$

La mesure μ doit satisfaire

$$\mu(C_{z_m^n}) = \mathbb{P}(X_m = z_m, \ldots, X_n = z_n).$$

Comme les cylindres engendrent la tribu $\mathcal{B}$, μ est uniquement déterminée par ces conditions. Réciproquement, ces conditions définissent une fonction T-invariante sur l'ensemble des cylindres. Le fait que cette fonction se prolonge en une mesure résulte d'un principe général dû à Kolmogorov. ■

Exemple 73 *Soient X_n, $n \in \mathbb{Z}$, des variables indépendantes et de même loi p. Alors $\mu = p_{\mathbb{Z}}$ est la mesure produit.*

Exemple 74 *Soit X_n, $n \in \mathbb{Z}$, une chaîne de Markov stationnaire, de loi p et de matrice de probabilités de transition. Alors μ est caractérisée par*

$$\mu(C_{z_m^n}) = p(z_m)P_{z_m z_{m+1}} \cdots P_{z_{n-1} z_n}.$$

Suggestion d'exercice : n^{0}2, feuille 4.

5.7 Le Théorème Ergodique de Birkhoff

Voici le substitut annoncé de la loi des grands nombres.

Théorème 13 (Birkhoff 1931) *Soit $(\Omega, \mathcal{B}, \mu)$ un espace probabilisé. Soit $T : \Omega \to \Omega$ une bijection qui préserve la mesure. Soit $f : \Omega \to \mathbb{R}$ une fonction intégrable. Alors la limite*

$$\bar{f}(x) = \lim_{n\to+\infty} \frac{1}{n}\left(f(x) + f(T(x)) + \cdots + f(T^{n-1}(x))\right)$$

existe μ-presque partout. La fonction $\bar{f}$ est intégrable, son espérance est égale à celle de f. Enfin, $\bar{f} \circ T = \bar{f}$ presque partout.

Preuve Voir le livre de Petersen, pages 27 à 33. ■

Lorsque $\Omega = E^{\mathbb{Z}}$, $\mathcal{B}$ est la tribu produit et $\mu = p^{\mathbb{Z}}$ est la mesure produit, les coordonnées X_n sont des variables aléatoires indépendantes et de même loi, le Théorème 13 affirme que les sommes $\frac{1}{n}\sum_{i=1}^n f(X_i)$ convergent presque sûrement vers une fonction T-invariante $\bar{f}$. Il faudrait encore montrer que $\bar{f}$ est presque partout constante. C'est vrai pour les suites de variables indépendantes, mais pas pour tous les processus stationnaires, comme on va le voir.

5.8 Ergodicité

Définition 75 *On dit qu'une transformation préservant la mesure $T : \Omega \to \Omega$ est* ergodique *si, pour tout ensemble mesurable $A \in \mathcal{B}$,*

$$T(A) = A \quad \Rightarrow \quad \mu(A)(1 - \mu(A)) = 0.$$

En particulier, une fonction invariante coïncide avec son espérance presque partout. Cela entraîne que les sommes de Birkhoff convergent presque partout vers l'espérance de f.

Proposition 76 (Ergodicité des chaînes de Markov) *Soit Ξ une chaîne de Markov stationnaire et irréductible. Alors le décalage $T : E^{\mathbb{Z}} \to E^{\mathbb{Z}}$ est ergodique pour la mesure μ. La réciproque est vraie : l'ergodicité de T entraîne que la chaîne est irréductible, pourvu que la mesure stationnaire charge tous les points de E.*

Preuve Soit, pour $x \in E$, f_x la fonction caractéristique de l'ensemble C_x des suites $(x_n)_{n\in\mathbb{Z}}$ telle $x_0 = x$. Le théorème ergodique de Birkhoff affirme que la suite de fonctions $\frac{1}{n}\sum_{i=1}^n f_x \circ T^i$ converge presque partout. Par convergence dominée, pour tous x, $y \in E$, la limite

$$q_{yx} = \frac{1}{\mathbb{E}(f_y)} \lim_{n\to+\infty} \mathbb{E}(\frac{1}{n}\sum_{i=1}^n (f_x \circ T^i) f_y)$$

existe. Or

$$\begin{aligned}\frac{1}{\mathbb{E}(f_y)}\mathbb{E}((f_x \circ T^i)f_y) &= \frac{1}{p(y)}\mathbb{P}(X_0 = y \text{ et } X_i = x) \\ &= \mathbb{P}(X_i = x | X_0 = y) = P^i_{yx},\end{aligned}$$

(ici, P^i désigne la puissance i-ème de la matrice des probabilités de transition P), donc

$$\begin{aligned}q_{xy} &= \lim_{n\to+\infty} \frac{1}{n}\sum_{i=1}^n \mathbb{E}((f_x \circ T^i)f_y) \\ &= \lim_{n\to+\infty} \frac{1}{n}\sum_{i=1}^n P^i_{yx}.\end{aligned}$$

Autrement dit, la limite

$$Q = \lim_{n\to+\infty} \frac{1}{n}\sum_{i=1}^n P^i$$

existe (on pourrait aussi le démontrer directement par l'algèbre linéaire). Chaque ligne q de Q est une distribution de probabilité et satisfait $qP = q$. Par unicité de la distribution stationnaire, $q = p$, donc

$$\lim_{n\to+\infty} \frac{1}{n}\sum_{i=1}^n \mathbb{E}((f_x \circ T^i)f_y) = \mathbb{E}(f_x)\mathbb{E}(f_y).$$

Plus généralement, soit $f_{z_r^{r+\ell}}$ la fonction caractéristique de l'ensemble $C_{z_r^{r+\ell}}$ des suites $(x_n)_{n\in\mathbb{Z}}$ telles $x_r = z_r, \ldots, x_{r+\ell} = z_{r+\ell}$. Alors, dès que $i + r > s + m$,

$$\begin{aligned}\mathbb{E}((f_{z_r^{r+\ell}} \circ T^i) f_{z_s^{s+m}}) &= \mathbb{P}(X_s^{s+m} = z_s^{s+m} \text{ et } X_{r+i}^{r+i+\ell} = z_{r+i}^{r+i+\ell}) \\ &= p(z_s) P_{z_s z_{s+1}} \cdots P_{z_{m-1} z_m} P^{i-m}_{z_m z_{r+i}} P_{z_{r+i} z_{r+i+1}} \cdots P_{z_{r+i+\ell-1} z_{r+i+\ell}}.\end{aligned}$$

donc

$$\begin{aligned}\lim_{n\to+\infty} \frac{1}{n}\sum_{i=s+m+1}^n \mathbb{E}((f_{z_r^{r+\ell}} \circ T^i) f_{z_s^{s+m}}) &= p(z_s) P_{z_s z_{s+1}} \cdots P_{z_{z+m-1} z_m} \\ &\quad Q_{z_m z_{r+i}} P_{z_{r+i} z_{r+i+1}} \cdots P_{z_{r+i+\ell-1} z_{r+i+\ell}} \\ &= \mathbb{P}(X_s^{s+m} = z_s^{s+m}) p(z_r) P_{z_{r+i} z_{r+i+1}} \cdots P_{z_{r+i+\ell-1} z_{r+i+\ell}} \\ &= \mathbb{P}(X_s^{s+m} = z_s^{s+m}) \mathbb{P}(X_r^{r+\ell} = z_r^{r+\ell}) \\ &= \mathbb{E}(f_{z_r^{r+\ell}}) \mathbb{E}(f_{z_s^{s+m}}).\end{aligned}$$

Ajouter le terme $\sum_{i=1}^{s+m} \mathbb{E}((f_{z_r^{r+\ell}} \circ T^i) f_{z_s^{s+m}})$ ne change pas la limite. L'espace vectoriel engendré par la famille de fonctions $f_{z_s^{s+m}}$ étant dense dans $L^2(E^{\mathbb{Z}}, \mu)$, on en déduit que pour toutes les fonctions f et $g \in L^2(E^{\mathbb{Z}}, \mu)$,

$$\lim_{n\to+\infty} \frac{1}{n}\sum_{i=1}^n \mathbb{E}((f \circ T^i) g) = \mathbb{E}(f)\mathbb{E}(g).$$

On peut appliquer ceci à la fonction caractéristique $f = g = 1_A$ d'un ensemble invariant A. Il vient $\mu(A) = \mathbb{E}(1_A^2) = \mathbb{E}(1_A)\mathbb{E}(1_A) = \mu(A)^2$, d'où $\mu(A) = 0$ ou 1. On conclut que T est ergodique. ∎

Exemple 77 *Soit Ξ un processus indépendant du temps, i.e. pour tout $n \in \mathbb{Z}$, $X_n = X_1$. Alors Ξ est une chaîne de Markov dont la matrice des probabilités de transition est l'identité. Si X_1 n'est pas presque partout constante, la transformation correspondante sur $E^{\mathbb{Z}}$ n'est pas ergodique.*

En effet, la mesure μ ne charge que les suites constantes. Si X_1 n'est pas presque partout constante, il existe $x \in E$ tel que $0 < \mathbb{P}(X_1 = x) < 1$. La suite constante $(\ldots, x, x, \ldots)$ est un ensemble invariant qui a pour mesure $\mathbb{P}(X_1 = x)$.

Suggestion d'exercice : n^{0}3, feuille 4.

Suggestion d'exercice : n^{0}4, feuille 4.

5.9 Preuve du Théorème 12, cas des chaînes de Markov stationnaires irréductibles

Dans ce cas, $-\log_2 p(x_i|x_1^{i-1}) = -\log_2 p(x_i|x_{i-1}) = f \circ T^{i-1}((x_n)_{n\in\mathbb{Z}})$ où la fonction $f : E^{\mathbb{Z}} \to \mathbb{R}$ est définie par

$$f((x_n)_{n\in\mathbb{Z}}) = -\log_2 p(x_1|x_0).$$

Par conséquent, la quantité qui nous intéresse est une somme de Birkhoff de f, à un terme tendant vers 0 près,

$$-\frac{1}{n}\log_2 p_1^n(x_1^n) = \frac{1}{n}\sum_{i=1}^{n-1} f \circ T^i((x_n)_{n\in\mathbb{Z}}) + \frac{1}{n}\log_2 p_1(x_1).$$

Le Théorème ergodique de Birkhoff affirme que $-\frac{1}{n}\log_2 p_1^n(x_1^n)$ converge presque sûrement. Si la transformation T est ergodique, la limite est l'espérance $H(\Xi)$. Ceci achève la démonstration, dans le cas particulier des chaînes de Markov stationnaires ergodiques. Or, d'après la Proposition 76, elles le sont si elles sont irréductibles.

Fin du cours n^{0}8

5.10 Cas général

Dans le cas d'un processus stationnaire général, la quantité qui nous intéresse n'est pas exactement une somme de Birkhoff. On l'encadre par deux variantes qui sont des sommes de Birkhoff. On pose

$$f_k((x_n)_{n\in\mathbb{Z}}) = -\log_2 p(x_0|x_{-1}, x_{-2}, \ldots, x_{-k}), \quad f_\infty((x_n)_{n\in\mathbb{Z}}) = -\log_2 p(x_0|x_{-1}, x_{-2}, \ldots, x_{-k}, \ldots).$$

Alors

$$\begin{aligned}\frac{1}{n}\sum_{i=0}^{n-1} f_\infty \circ T^i &= -\frac{1}{n}\log_2 \prod_{i=0}^{n-1} p(x_i|x_{i-1}, x_{i-2}, \ldots) \\ &= -\frac{1}{n}\log_2 p(x_0^{n-1}|x_{-\infty}^{-1}).\end{aligned}$$

Le Théorème ergodique de Birkhoff affirme donc que $-\frac{1}{n}\log_2 p(X_0^{n-1}|X_{-\infty}^{-1})$ converge presque sûrement vers $\mathbb{E}(f_\infty)$. Par analogie, on note

$$p^k(x_0^{n-1}) = p(x_0^{k-1}) \prod_{i=k}^{n-1} p(x_i|x_{i-k}^{i-1}),$$

de sorte que $-\frac{1}{n}\log_2 p^k(X_0^{n-1})$ converge presque sûrement vers $\mathbb{E}(f_k)$ quand n tend vers $+\infty$.

D'après le Théorème 10,

$$\begin{aligned}\mathbb{E}(f_k) &= \mathbb{E}(-\log_2 p(X_0|X_{-1},\dots,X_{-k})) \\ &= \mathbb{E}(-\log_2 p(X_k|X_{k-1},\dots,X_0)) \\ &= H(X_k|X_{k-1},\dots,X_0)\end{aligned}$$

tend vers $H(\Xi)$ quand k tend vers $+\infty$. Comme $f_\infty = \lim_{k\to+\infty} f_k$ (cela résulte du théorème de convergence des martingales), le théorème de convergence dominée entraîne que $\mathbb{E}(f_\infty) = H(\Xi)$. Donc les deux quantités ci-dessus ont la même limite presque sûre.

On montre enfin que pour tout k,

$$p^k(x_0^{n-1}) \le p(x_0^{n-1}) \le p(x_0^{n-1}|x_{-\infty}^{-1}),$$

à des termes tendant vers 0 près. En faisant tendre en plus k vers $+\infty$, on conclut que $-\frac{1}{n}\log_2 p(X_0^{n-1})$ converge presque sûrement vers $H(\Xi)$.

5.11 Appendice : preuve d'une forme faible de la loi des grands nombres

Proposition 78 *Soit $\Xi = (X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires indépendantes, de même loi, bornées. Alors la suite $S_n = \frac{1}{n}\sum_{i=1}^n X_i$ converge presque sûrement vers la constante $\mathbb{E}(X_0)$.*

Preuve Quitte à retrancher à X_n son espérance, on peut supposer que $\mathbb{E}(X_n) = 0$. Comme les X_i sont bornées, $|X_i| \le M$, on peut étudier l'espérance de produits comme $\mathbb{E}(X_{i_1}X_{i_2}X_{i_3}X_{i_4})$. Si l'un des indices, disons i_1, n'apparaît qu'une fois, alors, par indépendance,

$$\mathbb{E}(X_{i_1}X_{i_2}X_{i_3}X_{i_4}) = \mathbb{E}(X_{i_1})\mathbb{E}(X_{i_2}X_{i_3}X_{i_4}) = 0.$$

On développe

$$\begin{aligned}\mathbb{E}(|S_n|^4) &= \frac{1}{n^4}\sum_{i_1,i_2,i_3,i_4} \mathbb{E}(X_{i_1}X_{i_2}X_{i_3}X_{i_4}) \\ &= \frac{1}{n^4}(\sum_i \mathbb{E}(X_i^2) + 3\sum_{i_1<i_2} \mathbb{E}(X_{i_1}^2X_{i_2}^2) \\ &\le \frac{1}{n^4}(3n^2-2n)M^4 \sim \frac{3M^4}{n^2}.\end{aligned}$$

La série d'intégrales $\sum_{n=1}^\infty \mathbb{E}(|S_n|^4)$ converge. Par conséquent, la somme de la série $\sum_{n=1}^\infty |S_n|^4$ est intégrable (Théorème de convergence monotone), et donc presque sûrement finie. Cela entraîne que S_n tend vers 0 presque sûrement. ■

Proposition 79 *La convergence presque sûre implique la convergence en probabilité. Autrement dit, si une suite variables aléatoires X_n tend vers 0 presque sûrement, alors pour tout $\epsilon > 0$, la probabilité $\mathbb{P}(|X_n| > \epsilon)$ tend vers 0 quand n tend vers $+\infty$.*

Preuve Soit $Y_n = \min\{|X_n|, 1\}$. Cette suite de fonctions converge presque partout vers 0. Elle est majorée par la constante 1, qui est intégrable. Par convergence dominée, $\mathbb{E}(Y_n)$ tend vers 0. Par l'inégalité de Bienaymé-Tchebycheff, $\mathbb{P}(|X_n| > \epsilon) = \mathbb{P}(|Y_n| > \epsilon) \le \frac{1}{\epsilon}\mathbb{E}(Y_n)$ tend vers 0 pour tout $\epsilon > 0$. ■

5.12 A retenir

- L'équipartition asymptotique, et l'interprétation de l'entropie et de l'information mutuelle en termes de suites typiques.
- Le lien entre processus stationnaires et systèmes dynamiques.
- Le Théorème ergodique de Birkhoff, substitut à la loi des grands nombres.
- La notion d'ergodicité.

Il s'agit d'un chapitre théorique, où on passe progressivement de considérations élémentaires sur des probabilités finies à un niveau d'abstraction plus élevé, avec des espaces non discrets, des tribus, et des théorèmes de convergence délicats.

6 Capacité d'un canal de transmission

On cherche à faire passer des textes au travers d'un canal de transmission imparfait : pour chaque lettre envoyée, il y a une probabilité non nulle que la lettre reçue soit différente. Par un codage judicieux, on peut diminuer cette probabilité d'erreur. Par exemple, en répétant chaque lettre 3 fois, la probabilité d'erreur est élevée au carré, elle devient négligeable. Le prix à payer est un allongement du message, ce que l'on souhaite éviter. Quel est le meilleur compromis entre risque d'erreur et longueur de la transmission ?

6.1 Modélisation d'un canal de transmission

Le canal prend des caractères pris dans un ensemble fini E_X et retourne des caractères pris dans un autre ensemble E_Y. On se limite aux canaux *sans mémoire*, i.e. tels que la probabilité d'observer un caractère y en sortie ne dépend que du caractère x arrivé en entrée. Le canal est donc entièrement décrit par la matrice de probabilités de transition $P_{xy} = p(y|x)$. Si X est une variable aléatoire à valeurs dans E_X, le canal produit à partir de X mis en entrée une variable aléatoire Y à valeurs dans E_Y. La loi du couple (X, Y) est donnée par

$$\mathbb{P}(Y = y \text{ et } X = x) = \mathbb{P}(Y = y|X = x)P(X = x) = P_{xy}p_X(x).$$

Définition 80 *La* capacité κ *d'un canal est la borne supérieure des informations mutuelles $I(X;Y)$ sur toutes les variables d'entrée X.*

Exemple 81 *Canal binaire symétrique : il reçoit et renvoie des bits, chaque bit est renversé avec probabilité α. Alors $\kappa = 1 - h(\alpha)$.*

En effet, la matrice des probabilités de transition est $\begin{pmatrix} 1-\alpha & \alpha \\ \alpha & 1-\alpha \end{pmatrix}$. Pour toute variable X,

$$H(Y|X) = \sum_{x \in E_X} p_X(x)H(Y|X = x) = h(\alpha),$$

d'où

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) - h(\alpha) \leq 1 - h(\alpha),$$

car Y ne prend que deux valeurs, et l'égalité a lieu si la loi de X est uniforme.

Suggestion d'exercice : n^{0}1, feuille 5.

Fin du cours n^{0}9

Exemple 82 *Canal effaceur : il reçoit des bits et renvoie des bits ou bien la mention "illisible". Chaque bit est rendu illisible avec probabilité α. Alors $\kappa = 1 - \alpha$.*

En effet, la matrice des probabilités de transition est $\begin{pmatrix} 1-\alpha & 0 & \alpha \\ 0 & 1-\alpha & \alpha \end{pmatrix}$. A nouveau, pour toute variable X, $H(Y|X) = h(\alpha)$, d'où $I(X;Y) \leq \log_2 3 - h(\alpha)$, mais le cas d'égalité n'est jamais réalisé. Il faut trouver une meilleure majoration de $H(Y)$. Soit Z la variable aléatoire qui vaut 1 quand Y est illisible, 0 sinon. Alors Z est une fonction de Y donc

$$\begin{aligned} H(Y) &= H(Y) + H(Z|Y) = H(Y,Z) = H(Z) + H(Y|Z) \\ &= h(\alpha) + \mathbb{P}(Z = 1)H(Y|Z = 1) + \mathbb{P}(Z = 0)H(Y|Z = 0) \\ &= h(\alpha) + (1 - \alpha)H(X) \leq h(\alpha) + 1 - \alpha, \end{aligned}$$

avec égalité si la loi de X est uniforme. Il vient

$$I(X;Y) = H(Y) - H(Y|X) \leq 1 - \alpha,$$

atteint pour X uniformément distribué, et $\kappa = 1 - \alpha$.

Remarque 83 *Dans cet exemple, il est intuitivement clair qu'une fraction α de l'information est irrémédiablement perdue.*

Exemple 84 *Machine à écrire défectueuse : quand on frappe sur une touche, la lettre imprimée est ou bien celle voulue ou bien la suivante dans l'ordre alphabétique (cyclique), avec égale probabilité. Alors $\kappa = \log_2 13$.*

La matrice des probabilités de transition a des $1/2$ sur la diagonale et sur la diagonale qui se trouve juste en dessous. L'entropie relative $H(Y|X) = h(\frac{1}{2}) = 1$. Par conséquent, $I(X;Y) = H(Y) - H(Y|X) \leq \log_2(26) - 1 = \log_2(13)$, atteint lorsque la loi de X est uniforme.

Suggestion d'exercice : n^{0}2, feuille 5.

Suggestion d'exercice : n^{0}3, feuille 5.

6.2 Modélisation du codage et du décodage

En amont du canal, il y a un dispositif déterministe qui code le texte original, écrit dans un alphabet A, de sorte qu'il puisse être digéré par le canal. Chaque lettre $a \in A$ est codée par une chaîne de n caractères $C(a) \in E_X^n$. En aval, il y a un dispositif déterministe qui reconstitue des lettres $g(y_1, \ldots, y_n) \in A$ à partir de ce qui sort du canal. L'entier n est appelé la *longueur des blocs.* Ce qui nous intéresse, c'est la *probabilité maximale d'erreur* $\max_{a \in A} \lambda_a$, où λ_a est la probabilité conditionnelle d'erreur sur la lettre a,

$$\lambda_a = \mathbb{P}(g(\text{sortie}) \neq a | \text{entrée} = C(a)).$$

Elle dépend à la fois du canal, du codeur et du décodeur. L'efficacité d'un codeur/décodeur se mesure à son *taux* $\frac{\log_2 |A|}{n}$. L'unité est le bit par transmission.

Exemple 85 *On transmet des bits au travers d'un canal binaire symétrique de probabilité d'erreur α. On considère le dispositif qui, en amont du canal, répète chaque bit 3 fois et en aval, choisit dans chaque suite de 3 bits celui qui a la majorité. Son taux vaut $\frac{1}{3}$ bits par transmission. Sa probabilité maximale d'erreur vaut $3\alpha^2 - 2\alpha^3$.*

En effet, il y a erreur lorsque le canal a changé au moins deux des trois bits identiques qui constituent le codage d'une lettre. Ceci se produit avec probabilité $3\alpha^2(1-\alpha) + \alpha^3 = 3\alpha^2 - 2\alpha^3$.

Définition 86 *On dit qu'un canal autorise un taux de transmission τ s'il existe une suite de codeurs/décodeurs de taux tendant vers τ et dont la probabilité maximale d'erreur tend vers 0.*

Exemple 87 *Machine à écrire défectueuse : Il est aisé de voir qu'on peut transmettre 13 lettres sans aucune erreur (il suffit de ne transmettre que les lettres de numéros impairs : A, C, E,...), ce qui donne un taux de transmission au moins égal à $\log_2(13)$ bits par transmission.*

Ici, A est l'ensemble des 13 lettres de numéros impairs, et la longueur des bloc vaut 1. Le codage est l'injection de A dans l'alphabet à 26 lettres. Le décodage consiste à remplacer toute lettre lue en sortie par la lettre de numéro impair immédiatement inférieure. La probabilité maximale d'erreur vaut 0. Le taux vaut $\log_2(13)$.

6.3 Le théorème du codage de canal

Dans l'exemple précédent, le dispositif de codage/décodage est optimal : il réalise le taux de transmission maximal, en vertu du théorème suivant.

Théorème 14 (Shannon, 1948) *Le taux de transmission maximal autorisé par un canal sans mémoire est égal à sa capacité κ. Autrement dit,*

1. *Pour toute suite de codeurs/décodeurs dont la probabilité maximale d'erreur tend vers 0, la limite supérieure des taux est* $\leq \kappa$.
2. *Il existe une suite de codeurs/décodeurs dont le taux tend vers* κ *et dont la probabilité maximale d'erreur tend vers 0.*

Idée de la preuve. Soit un dispositif codeur/décodeur utilisant des blocs de longueur n assez grande. Notons X_1^n la variable mise en entrée du canal, après codage d'une variable uniforme, et Y_1^n la variable lue en sortie. Supposons, pour simplifier, que la probabilité maximale d'erreur est nulle. Alors $\log_2 |A| = H(X_1^n) = H(X_1^n|Y_1^n) + I(X_1^n; Y_1^n) = I(X_1^n; Y_1^n) \leq nI(X;Y) \leq n\kappa$, car dans ce cas, X_1^n est une fonction de Y_1^n. L'argument s'étend au cas général (prise en compte d'une faible probabilité d'erreur).

Réciproquement, on montre que par un codage judicieux, on peut forcer le canal, ou plutôt, son extension aux blocs de longueur n, à se comporter comme la machine à écrire défectueuse, où il y a un grand sous-ensemble de blocs de longueur n dont les images possibles en sortie constituent des sous-ensembles presque disjoints de E_Y^n. Ces blocs, ce sont les suites typiques.

D'après la propriété d'équirépartition asymptotique, pour chaque bloc de E_X^n, il y a environ $2^{nH(Y|X)}$ suites typiques dans E_Y^n, d'égale probabilité. Or le nombre total de suites typiques dans E_Y^n est $2^{nH(Y)}$. Donc le nombre de blocs de E_X^n qui donnent des ensembles disjoints en sortie est au plus $2^{n(H(Y)-H(Y|X))} = 2^{nI(X;Y)}$. Cette borne est atteinte par un codage tiré au hasard. Autrement dit, on peut transmettre $2^{nI(X;Y)}$ blocs distincts, ce qui correspond à un taux de transmission $\geq I(X;Y)$ bits par transmission, lorsqu'à l'entrée arrive un texte tiré selon X.

Fin du cours n^{0}10

6.4 Deux inégalités

D'abord, quand un canal doit transmettre des blocs de longueur n, sa capacité est au pire multipliée par n, i.e., sa capacité "par symbole" n'augmente pas.

Lemme 88 *Soient* E_X *et* E_Y *les alphabets d'entrée et de sortie d'un canal sans mémoire de capacité* κ. *Soit* X_1^n *une variable aléatoire à valeurs dans* E_X^n. *On fait passer successivement les composantes de* X_1^n *à travers le canal et on obtient une variable* Y_1^n *à valeurs dans* E_Y^n. *Alors*

$$I(X_1^n; Y_1^n) \leq \sum_{i=1}^n I(X_i; Y_i) \leq n\kappa.$$

Preuve

$$\begin{aligned} I(X_1^n; Y_1^n) &= H(Y_1^n) - H(Y_1^n|X_1^n) \\ &= H(Y_1^n) - H(Y_1|X_1^n) - \sum_{i=2}^n H(Y_i|Y_1, \ldots, Y_{i-1}, X_1^n) \\ &= H(Y_1^n) - \sum_{i=1}^n H(Y_i|X_i), \end{aligned}$$

car Y_i ne dépend que de X_i et, conditionnellement à X_1^n, Y_i est indépendant des Y_j, $j \neq i$. Il vient

$$\begin{aligned} I(X_1^n; Y_1^n) &\leq \sum_{i=1}^n H(Y_i) - \sum_{i=1}^n H(Y_i|X_i) \\ &= \sum_{i=1}^n H(Y_i) - H(Y_i|X_i) = \sum_{i=1}^n I(Y_i; X_i) \leq n\kappa \end{aligned}$$

par définition de la capacité. ∎

Ensuite, une propriété des chaînes de Markov.

Lemme 89 *Soient X, Y et Z des variables aléatoires. On suppose que le processus (X, Y, Z) est une chaîne de Markov. Alors*

$$I(X;Z) \leq I(Y;Z) \quad et \quad I(X;Z) \leq I(X;Y).$$

Preuve D'après la propriété de Markov, X est conditionnellement indépendant de Z sachant Y, i.e.

$$\mathbb{P}(X = x|Y = y, Z = z) = \mathbb{P}(X = x|Y = y).$$

Autrement dit,

$$\frac{p_{(X,Y,Z)}(x,y,z)}{p_{(Y,Z)}(y,z)} = \frac{p_{(X,Y)(x,y)}}{p_Y(y)},$$

ce qui, par définition de l'information mutuelle conditionnelle, entraîne que $I(X;Z|Y) = 0$. D'après la Proposition 34,

$$I(X;Y,Z) = I(X;Y) + I(X;Z|Y) = I(X;Y),$$

d'où, en appliquant une seconde fois la Proposition 34,

$$I(X;Y) = I(X;Y,Z) = I(X;Z) + I(X;Y|Z)) \geq I(X;Z).$$

Comme $I(Z;X|Y) = I(X;Z|Y) = 0$, on peut échanger X et Z, et conclure que $I(Z;Y) \geq I(X;Z)$. ■

Suggestion d'exercice : n^{0}4, feuille 5.

6.5 Inégalité de Fano

Pour prendre en compte une probabilité d'erreur non nulle mais faible, on utilise une majoration de l'entropie conditionnelle entre une variable et une valeur approchée de cette variable en fonction de la probabilité d'erreur. Il s'agit de l'inégalité de Fano.

Proposition 90 (Fano, 1952) *Soient W et $\hat{W}$ deux variables aléatoires à valeurs dans un ensemble fini E_W. Soit $p_e = \mathbb{P}(W \neq \hat{W})$. Alors*

$$H(W|\hat{W}) \leq h(p_e) + p_e \log_2 |E_W|.$$

Preuve Soit Z la variable aléatoire qui vaut 1 si $W \neq \hat{W}$ et 0 sinon. Z est une fonction de $(W, \hat{W})$, donc $H(Z|W, \hat{W}) = 0$. En appliquant de deux façons différentes le Corollaire 23, il vient

$$\begin{aligned} H(W|\hat{W}) &= H(W|\hat{W}) + H(Z|W,\hat{W}) \\ &= H(Z,W|\hat{W}) \\ &= H(Z|\hat{W}) + H(W|Z,\hat{W}). \end{aligned}$$

D'après le principe "conditionner diminue l'entropie" (Proposition 28), $H(Z|\hat{W}) \leq H(Z) = h(p_e)$. Sachant que $Z = 0$, $W = \hat{W}$, donc $H(W|\hat{W}, Z = 0) = 0$. Sachant que $Z = 1$, $H(W|\hat{W}, Z = 1) \leq H(W) \leq \log_2 |E_W|$. Par définition de l'entropie conditionnelle,

$$H(W|Z,\hat{W}) = \mathbb{P}(Z = 0)H(W|\hat{W}, Z = 0) + \mathbb{P}(Z = 1)H(W|\hat{W}, Z = 1) \leq p_e \log_2 |E_W|.$$

On conclut que $H(W|\hat{W}) \leq h(p_e) + p_e \log_2 |E_W|$. ■

6.6 Preuve du Théorème 14, sens direct

Supposons donné un canal de capacité κ et un dispositif de codage $C : A \to E_X^n$ et de décodage $g : E_Y^n \to A$, de taux τ, utilisant des blocs de taille n, pour ce canal. Soit p_e sa probabilité maximale d'erreur. Soit W une variable aléatoire de loi uniforme dans A. Soit $X_1^n = C(W)$ la variable mise en entrée du canal, soit Y_1^n la variable observée en sortie, soit $\hat{W} = g(W)$ le résultat du décodage. Alors

$$\begin{aligned}
\mathbb{P}(W \neq \hat{W}) &= \sum_{a \in A} \mathbb{P}(W = a)\mathbb{P}(\hat{W} \neq a | W = a) \\
&= \sum_{a \in A} \mathbb{P}(W = a)\mathbb{P}(g(Y_1^n) \neq a | X_1^n = C(a)) \\
&= \sum_{a \in A} \mathbb{P}(W = a)\lambda_a \\
&\leq \max_{a \in A} \lambda_a = p_e.
\end{aligned}$$

Le processus $(W, X_1^n, Y_1^n, \hat{W})$ est une chaîne de Markov. En effet, X_1^n dépend de façon déterministe de W, Y_1^n dépend de façon markovienne de X_1^n (car le canal est sans mémoire), et $\hat{W}$ dépend de façon déterministe de Y_1^n. D'après le Lemme 89,

$$I(W; \hat{W}) \leq I(W; Y_1^n) \leq I(X_1^n; Y_1^n).$$

D'après le Lemme 88,

$$I(X_1^n; Y_1^n) \leq n\kappa.$$

L'inégalité de Fano donne

$$H(W|\hat{W}) \leq h(p_e) + p_e \log_2 |A|.$$

$$\begin{aligned}
H(W) &= H(W|\hat{W}) + I(W; \hat{W}) \\
&\leq h(p_e) + p_e \log_2 |A| + I(X_1^n; Y_1^n) \\
&\leq 1 + p_e \log_2 |A| + n\kappa.
\end{aligned}$$

En choisissant W uniformément réparti dans A, on trouve

$$\tau \leq \kappa + \frac{1}{n} + p_e\tau.$$

Lorsque la taille des blocs tend vers l'infini et la probabilité maximale d'erreur tend vers 0, on trouve asymptotiquement $\tau \leq \kappa$.

Fin du cours n^{0}11

6.7 Schéma de la preuve du Théorème 14, sens réciproque

On montre que si $\tau < \kappa$, il existe un codeur/décodeur de taux de transmission $> \tau$.

Le décodage est fondé sur la notion de suite conjointement typique. Soit $C : A \to E_X^n$ le codeur. Avec forte probabilité, le dispositif codeur/canal produit des suites conjointement typiques pour le couple (X, Y). Si on observe y_1^n en sortie, l'entrée x_1^n doit être telle que le couple (x_1^n, y_1^n) est conjointement typique. Le décodage $g(y_1^n)$ doit être un $a \in A$ tel que $(C(a), x_1^n)$ est une suite conjointement typique. Si $|A| < 2^{n\tau}$, avec forte probabilité, cet élément a existe et est unique.

Le codeur est tiré au hasard. On montre que l'espérance, parmi les choix aléatoires de C, de la probabilité maximale d'erreur du dispositif codeur/décodeur, est faible. Il existe donc un codeur dont la probabilité maximale d'erreur est faible. C'est une illustration de ce qu'on appelle parfois la méthode probabiliste en combinatoire.

6.8 Majoration de la probabilité d'erreur

Par hypothèse, il existe une distribution p_X sur E_X (et donc une variable aléatoire X et la variable de sortie Y correspondante) telle que $I(X;Y) > \tau$. Soit A un ensemble à $2^{n\tau}$ éléments. On interprète le fait de tirer indépendamment $2^{n\tau}$ éléments de E_X^n selon la loi produit $p_X^{\otimes n}$ comme une application aléatoire $C : A \to E_X^n$. Autrement dit, chaque lettre de $C(a)$, $a \in A$, est tirée indépendamment selon p_X, ce qu'on note

$$\mathbb{P}_C(C(a) = x_1^n) = p_X(x_1) \cdots p_X(x_n),$$

car l'aléa provient du choix de l'application C. De plus les différentes variables $C(a)$, $a \in A$, sont indépendantes.

Le procédé de décodage retenu repose sur la notion de suite conjointement typique pour le couple (X, Y). On fixe $\epsilon < \frac{1}{4}(I(X;Y) - \tau)$. Soit W une variable uniformément répartie dans A, soit $\tilde{X}_1^n = C(W)$ le signal mis en entrée du canal, et $\tilde{Y}_1^n$ la variable observée en sortie. Le décodage $a = g(\tilde{Y}_1^n)$ est l'élément $a \in A$ tel que $(C(a), \tilde{Y}_1^n) \in T$. Il y a deux sources d'erreurs de décodage.

1. Il n'existe aucun $a \in A$ tel que $(C(a), \tilde{Y}_1^n)$ est conjointement typique. Même $a = W$ ne convient pas. Donc la probabilité de cet évènement est majorée par celle de $\mathcal{E} = \{(\tilde{X}_1^n, \tilde{Y}_1^n) \notin T)\}$.
2. Il existe plusieurs $a \in A$ tels que $(C(a), \tilde{Y}_1^n) \in T$. Notons cet évènement $\mathcal{F}$.

Fixons $a_0 \in A$ et raisonnons conditionnellement à l'évènement $\{W = a_0\}$. L'aléa restant provient du tirage au hasard du codeur C. Alors les composantes de $\tilde{X}_1^n$ sont (conditionnellement) indépendantes, de loi p_X, et, comme le canal est sans mémoire, les composantes de $\tilde{Y}_1^n$ sont (conditionnellement) indépendantes, de loi p_Y. Chacun des couples $(\tilde{X}_i, \tilde{Y}_i)$ a même loi que (X, Y). D'après la Proposition 70, point 1,

$$\mathbb{P}_C(\mathcal{E}|W = a_0) = \mathbb{P}((X,Y)_1^n \notin T)$$

tend vers 0, donc est $< \epsilon$ pour n assez grand.

Pour $a \neq a_0$, les variables conditionnées $\tilde{X}_1^n|W = a$ et $\tilde{X}_1^n|W = a_0$ sont indépendantes. Comme le canal est sans mémoire, il en est de même de $\tilde{Y}_1^n|W = a$ et $\tilde{X}_1^n|W = a_0$. Donc pour tout $a \neq a_0$, la Proposition 70, point 3, donne

$$\mathbb{P}_C((\tilde{X}|W = a, \tilde{Y}|W = a_0)_1^n \in T) \leq 2^{-n(I(X;Y)-3\epsilon)} \leq 2^{-n(\tau+\epsilon)}.$$

En sommant sur les $a \neq a_0$,

$$\mathbb{P}(\mathcal{F}|W = a_0) \leq 2^{n\tau}\mathbb{P}((\tilde{X}, \tilde{Y})_1^n \in T) \leq 2^{-n(I(X;Y)-\tau-3\epsilon)} \leq 2^{-n\epsilon} < \epsilon$$

pour n assez grand. En faisant la moyenne sur les tirages $a_0 \in A$ de W, il vient

$$\mathbb{E}_W(\mathbb{P}_C(\mathcal{E} \cup \mathcal{F})) < 2\epsilon$$

pour n assez grand. On écrit cela $\mathbb{E}_W(\mathbb{E}_C(1_{\mathcal{E}\cup\mathcal{F}})) < 2\epsilon$, ou, avec Fubini, $\mathbb{E}_C(\mathbb{E}_W(1_{\mathcal{E}\cup\mathcal{F}})) < 2\epsilon$

Il existe donc un codage C qui rend la probabilité moyenne d'erreur $\mathbb{E}_W(1_{\mathcal{E}\cup\mathcal{F})}) < 2\epsilon$. On fixe un tel C désormais. Remarquer que

$$\mathbb{E}(\lambda_W) = 2^{-n\tau} \sum_{a \in A} \lambda_a = \mathbb{E}_W(\mathbb{P}(\mathcal{E} \cup \mathcal{F})) < 2\epsilon.$$

On ordonne les éléments de A par probabilité d'erreur λ_a croissante, et on note $A' \subset A$ la première moitié de A pour cette ordre. Alors

$$p_e = \max_{a \in A'} \lambda_a < 4\epsilon.$$

En effet, $p_e \leq \min_{a \notin A'} \lambda_a$, d'où

$$|A|\mathbb{E}(\lambda_W) = \sum_{a \in A'} \lambda_a + \sum_{a \notin A'} \lambda_a \geq |A \setminus A'|p_e = \frac{1}{2}|A|p_e.$$

On modifie le codage C choisi en remplaçant A par A' et C par sa restriction C' à A'. Pour le codage C', la probabilité maximale d'erreur est $p_e < 4\epsilon$. D'autre part, le taux de transmission de C' est $\frac{\log|A'|}{n} = \tau - \frac{1}{n}$. On faisant tendre n vers l'infini, on conclut que τ constitue un taux de transmission autorisé par le canal. Ceci achève la preuve du Théorème 14.

6.9 Postérité

Le Théorème du codage de canal contient tous les ingrédients pour attirer l'attention :
- Un problème concret, d'intérêt industriel.
- Une modélisation simple, un résultat d'impossibilité théorique qui, pour une fois, est réaliste.
- Une preuve lumineuse, basée sur des idées heuristiques simples, mais un peu délicate à mettre au point.
- Une preuve probabiliste d'existence, qui ne fournit aucune piste pour construire des codeurs/décodeurs réels, et constitue donc un défi.

Le théorème est à l'origine d'innombrables travaux qui ont produit des codeurs/décodeurs de plus en plus performants. Il ne suffit pas que le taux de transmission soit élevé. Il faut aussi des algorithmes peu coûteux de codage et de décodage. L'algèbre sur les corps finis a été mis à contribution. Les turbo-codes, inventés par des ingénieurs de l'Ecole Nationale Supérieure des Télécommunications, à Lannion, se sont imposés comme solution technologique. Ils équipent aussi bien les missions spatiales de la NASA et de l'Agence Spatiale Européenne que les lecteurs de DVD. Ils constituent aussi une excellente solution théorique : ils réalisent presque la borne du Théorème du codage de canal sur le taux de transmission autorisé pour de vastes classes de canaux.

6.10 A retenir

- L'interprétation de l'information mutuelle comme borne sur le taux de transmission autorisé par un canal.
- Le calcul de la capacité dans quelques exemples.

Il s'agit d'un chapitre pratique, où on utilise les outils du chapitre précédent dans un cas simple (indépendance) pour résoudre un problème concret.

7 Entropie métrique

Un processus stationnaire à valeurs dans un ensemble fini n'est qu'un exemple de transformation préservant une mesure de probabilité. L'entropie par symbole, définie pour l'instant seulement pour les processus stationnaires à valeurs dans des ensembles finis, s'étend à cette situation plus vaste.

Reprenons nos deux exemples favoris de processus stationnaires, et traduisons les en termes de dynamique sur un espace probabilisé.

Exemple 91 (Décalage de Bernoulli) *Soit p une distribution de probabilité sur un ensemble fini E. Le décalage sur $E^{\mathbb{Z}}$ muni de la tribu produit et de la mesure produit $\mu = p^{\otimes\mathbb{Z}}$ est appelé* décalage de Bernoulli $\mathcal{B}(p)$ *(`Bernoulli shift` en anglais).*

La tribu produit est engendrée par les cylindres de la forme

$$C_{z_m^n} = \{\text{suites } (x_n)_{n\in\mathbb{N}} \text{ telles que } x_m = z_m, \ldots, x_n = z_n\}.$$

La mesure produit donne au cylindre $C_{z_m^n}$ la mesure $\mu(C_{z_m^n}) = p(z_m)\cdots p(z_n)$.

Exemple 92 (Décalages de Markov) *Soit p une distribution de probabilité sur un ensemble fini E. Soit P une matrice stochastique pour laquelle p est stationnaire, i.e. $pP = p$. Le* décalage de Markov $\mathcal{M}(p, P)$ *(`Markov shift` en anglais) est le décalage sur $E^{\mathbb{Z}}$ muni de la tribu produit et de la mesure μ telle que*

$$\mu(C_{z_m^n}) = p(z_m)P_{z_m z_{m+1}} \cdots P_{z_{n-1} z_n}.$$

Evidemment, le décalage de Bernoulli est un cas particulier de décalage de Markov.

7.1 Entropie d'une transformation préservant la mesure

Voici la démarche qui conduit à la généralisation. A une transformation T préservant une mesure de probabilité μ sur un espace Ω sont associés des processus stationnaires à valeurs dans des ensembles finis. Il suffit pour cela de choisir une partition finie $\alpha = (A_x)_{x\in E_\alpha}$ de Ω, et de poser

$$X_n(\omega) = x \text{ si } T^n(\omega) \in A_x.$$

C'est bien un processus stationnaire : pour tous $k \in \mathbb{N}$ et $n \in \mathbb{Z}$, la loi jointe satisfait

$$\begin{aligned}
\mathbb{P}(X_n^{n+k} = x_0^k) &= \mathbb{P}(T^n(\omega) \in A_{x_0}, \dots, T^{n+k}(\omega) \in A_{x_k}) \\
&= \mu(T^{-n}(A_{x_0}) \cap \dots \cap T^{-n-k}A_{x_k}) \\
&= \mu(T^{-n}\left(A_{x_0} \cap \dots \cap T^{-k}A_{x_k}\right)) \\
&= \mu\left(A_{x_0} \cap \dots \cap T^{-k}A_{x_k}\right) \\
&= \mathbb{P}(\omega \in A_{x_0}, \dots, T^k(\omega) \in A_{x_k}) \\
&= \mathbb{P}(X_0^k = x_0^k).
\end{aligned}$$

On note Ξ_α ce processus. Il correspond à ce qu'on peut extraire de T à l'aide d'un appareil de mesure, qui ne possède forcément qu'un nombre fini de graduations. Pour affiner la mesure, il faut des appareils de plus en plus précis, donc des partitions de plus en plus fines. Cela conduit à la définition suivante.

Définition 93 *Soit $(\Omega, \mathcal{B}, \mu)$ un espace probabilisé. Soit $T : \Omega \to \Omega$ une application qui préserve la mesure. Etant donné une partition mesurable finie $\alpha = (A_x)_{x\in E_\alpha}$ de Ω, on note Ξ_α le processus stationnaire correspondant. On appelle* entropie métrique *de T la borne supérieure*

$$h_\mu(T) = \sup_\alpha H(\Xi_\alpha),$$

prise sur toutes les partitions mesurables finies de Ω.

On abrégera souvent entropie métrique en entropie, quand le contexte exclut toute confusion.

7.2 Entropie des partitions

On peut s'affranchir de la notion de processus stationnaire dans la définition de l'entropie métrique. En effet, les lois jointes ne sont autres que les probabilités d'évènements engendrés par la transformation T à partir de la partition α.

Définition 94 *Soient $\alpha = (A_x)_{x\in E_\alpha}$ et $\beta = (B_y)_{y\in E_\beta}$ deux partitions finies d'un espace probabilisé Ω, et $T : \Omega \to \Omega$ une application qui préserve la mesure.*

- *On note $T\alpha$ la partition dont les pièces sont les $(T^{-1}A_x)_{x\in E_\alpha}$.*
- *On dit que β* raffine *α, et on note $\alpha \preceq \beta$, si chaque pièce de β est contenue dans une pièce de α.*
- *On note $\alpha \vee \beta$ la partition dont les pièces sont les $(A_x \cap B_y)_{(x,y)\in E_\alpha \times E_\beta}$. Cette opération est commutative et associative.*
- *On appelle entropie de la partition α le nombre $H(\alpha) = -\sum_{x\in E} \mu(A_x)\log_2(\mu(A_x))$.*
- *On appelle entropie conditionnelle des partitions α et β le nombre*

$$H(\alpha|\beta) = -\sum_{(x,y)\in E_\beta} \mu(A_x \cap B_y)\log_2 \frac{\mu(A_x \cap B_y)}{\mu(B_y)}.$$

Lemme 95 *L'entropie des partitions a les propriétés suivantes.*

- $0 \le H(\alpha|\beta) \le H(\alpha)$.
- $H(\alpha \vee \beta) = H(\beta) + H(\alpha|\beta)$.
- $H(\alpha \vee \beta) \le H(\alpha) + H(\beta)$.

– $\alpha \preceq \alpha' \Rightarrow H(\alpha) \leq H(\alpha')$ *et* $H(\alpha|\beta) \leq H(\alpha'|\beta)$.
– *Si* $T : \Omega \to \Omega$ *préserve la mesure,* $T(\alpha \vee \beta) = T(\alpha) \vee T(\beta)$, $H(\alpha|\beta) = H(T\alpha|T\beta)$.

Preuve $H(\alpha) = H(f_\alpha)$ où la variable aléatoire $f_\alpha : \Omega \to E_\alpha$ vaut x sur A_x. ■

Suggestion d'exercice : n^{0}1, feuille 6.

Proposition 96 *Soit* $T : \Omega \to \Omega$ *une application qui préserve la mesure. Pour toute partition mesurable finie* α,

$$H(\Xi_\alpha) = \lim_{n\to\infty} \frac{1}{n} H(\alpha \vee T\alpha \vee \cdots \vee T^{n-1}\alpha).$$

Preuve Par construction, la variable $X_{\alpha,n}$ du processus stationnaire Ξ_α est $f_{T^n\alpha}$. Par conséquent, l'entropie de la n-ème partition engendrée par T est égale à l'entropie jointe,

$$H(\alpha \vee T\alpha \vee \cdots \vee T^{n-1}\alpha) = H(X_{\alpha,1}, \ldots, X_{\alpha,n}).$$

■

Exemple 97 (Décalage de Bernoulli) *Soit* $T = \mathcal{B}(p)$ *un décalage de Bernoulli sur* $E^{\mathbb{Z}}$. *Soit* α *la partition telle que* $A_x = \{$*suites* $(x_n)_{n\in\mathbb{Z}}$ *telles que* $x_0 = x\}$, *i.e. la partition induite par la projection sur un facteur. Le processus* Ξ_α *est une suite de variables indépendantes de loi p, donc* $H(\Xi_\alpha) = H(p)$.

Exemple 98 (Décalage de Markov) *Soit* $T = \mathcal{M}(p, P)$ *un décalage de Markov sur* $E^{\mathbb{Z}}$. *Soit* α *la partition induite par la projection sur un facteur. Le processus* Ξ_α *est une chaîne de Markov stationnaire de matrice de probabilités de transition* P *et de distribution stationnaire p, donc* $H(\Xi_\alpha) = -\sum_{x,\,y\in E} p(x)P_{xy}\log_2(P_{xy})$.

7.3 Partitions génératrices

La définition de l'entropie métrique par une borne supérieure est une belle construction théorique, qui paraît parfaitement incalculable. Il n'en est rien. Il se trouve que les "bonnes" partitions α donnent toutes la même valeur de $H(\Xi_\alpha)$.

Notation 99 *Soit* α *une partition finie d'un ensemble* Ω. *On note* $\alpha_m^n = T^m\alpha \vee \cdots \vee T^n\alpha$. *On note* $\alpha_{-\infty}^n$ *la tribu engendrée par les pièces des partitions* α_m^n, $m \leq n$. *Et de même,* $\alpha_{-\infty}^\infty$ *est la tribu engendrée par les pièces de toutes les partitions* α_m^n, m, $n \in \mathbb{Z}$.

Suggestion d'exercice : n^{0}2, feuille 6.

Suggestion d'exercice : n^{0}3, feuille 6.

Exemple 100 (Tribu produit) *Soit* E *un ensemble fini,* $\Omega = E^{\mathbb{Z}}$, $T =$ *décalage,* α *la partition* $A_x = \{$*suites* $(x_n)_{n\in\mathbb{Z}}$ *telles que* $x_0 = x\}$ *induite par la projection sur un facteur. La tribu produit sur* Ω *coïncide avec* $\alpha_{-\infty}^\infty$.

En effet, pour tout $(z_m, \ldots, z_n) \in E^{n-m+1}$,

$$C_{z_1^n} = T^{-m}A_{z_m} \cap \cdots \cap T^{-n}A_{z_n},$$

donc α_m^n coïncide avec l'ensemble des cylindres sur les coordonnées m à n. On conclut par définition de la tribu produit.

Définition 101 *Soit $(\Omega, \mathcal{B}, \mu)$ un espace probabilisé. Soit $T : \Omega \to \Omega$ une application qui préserve la mesure. On dit qu'une partition mesurable finie α de Ω est* génératrice *pour T si $\mathcal{B} = \alpha_{-\infty}^{\infty}$ aux ensembles de mesure nulle près.*

Exemple 102 *Pour un décalage de Bernoulli ou, plus généralement, de Markov, la partition induite par la projection sur un facteur est génératrice.*

Fin du cours n^{0}12

Le lemme suivant aide à comprendre ce que signifie $\alpha_{-\infty}^{\infty}$, et par conséquent, la notion de partition génératrice.

Lemme 103 *Soit $(\Omega, \mathcal{B}, \mu)$ un espace probabilisé. Soit $\beta_0 \preceq \cdots \preceq \beta_n \preceq$ une suite de partitions mesurables finies de Ω, de plus en plus fines. Soit β_∞, la tribu engendrée par toutes les β_n, $n \in \mathbb{N}$. Si $C \in \beta_\infty$, alors pour tout $\delta > 0$, il existe n et un élément B de la tribu engendrée par β_n tel que $\mu(B \Delta C) < \delta$.*

Preuve Notons $\beta_\rightarrow$ la famille de toutes les intersections $B_1 \cap \cdots B_n \cap \cdots$ où B_i appartient à la tribu engendrée par β_i. Soit $\mathcal{T}$ l'ensemble des réunions dénombrables d'éléments de $\beta_\rightarrow$. On montre d'abord que $\beta_\infty = \mathcal{T}$.

Par construction, T est stable par réunion dénombrable. Si $A \in \beta_\rightarrow$, le complémentaire A^c est une réunion dénombrable de complémentaires B_i^c. Chaque B_i^c appartient à la tribu engendrée par β_i, donc à $\beta_\rightarrow$. Par conséquent, $A^c \in \mathcal{T}$, i.e. $\mathcal{T}$ est stable par complémentaire. Cela prouve que $\mathcal{T}$ est une tribu. Pour tout i, elle contient la tribu engendrée par β_i, donc $\mathcal{T}$ contient β_∞. Réciproquement, comme β_∞ est stable par intersection dénombrable, $\beta_\rightarrow \subset \beta_\infty$, et comme β_∞ est stable par réunion dénombrable, $\mathcal{T} \subset \beta_\infty$. On conclut que $\mathcal{T} = \beta_\infty$.

Soit $C \in \beta_\infty$. On vient de montrer que $C = \bigcup_{k\in\mathbb{N}} I_k$ où I_k est une intersection $I_k = \bigcap_{i\in\mathbb{N}} B_{k,i}$ d'ensembles $B_{k,i}$ appartenant à la tribu engendrée par β_i. Il existe ℓ tel que $\mu(\bigcup_{k\geq\ell} I_k) < \frac{\delta}{2}$. Pour chaque $k < \ell$, il existe $i(k)$ tel que $\mu(\bigcap_{i\leq i(k)} B_{k,i} \setminus I_k) < \frac{\delta}{2\ell}$. Soit $B = \bigcup_{k<\ell} \bigcap_{i\leq i(k)} B_{k,i}$. Alors $\mu(B\Delta C) < \delta$ et B appartient à la tribu engendrée par β_n, pour $n = \max_{k<\ell} i(k)$. ∎

7.4 Calcul de l'entropie métrique

Théorème 15 (Kolmogorov, Sinai 1958) *Soit $(\Omega, \mathcal{B}, \mu)$ un espace probabilisé. Soit $T : \Omega \to \Omega$ une application qui préserve la mesure. Soit α une partition mesurable finie qui est génératrice pour T. Alors*

$$h_\mu(T) = H(\Xi_\alpha).$$

Preuve La preuve que voici, dûe à Yakov Sinai, procède en trois étapes.

1. Soit Φ l'ensemble des partitions mesurables finies de Ω modulo ensembles de mesure nulle. Muni de la distance de l'entropie conditionnelle, Φ est un espace métrique sur lequel la fonction $\beta \mapsto H(\Xi_\beta)$ est continue.
2. Soit $\Psi \subset \Phi$ le sous-ensemble défini par

$$\Psi = \{\beta \in \Phi \,;\, \exists k \text{ tel que } \beta \preceq \alpha_{-k}^k\}.$$

   On montre que pour tout $\beta \in \Psi$, $H(\Xi_\beta) \leq H(\Xi_\alpha)$.
3. Sous l'hypothèse que α est génératrice, on montre que Ψ est dense dans Φ.

1. Pour l'étape 1, on renvoie à l'Exercice 3 du TD6.

2. Soit β une partition mesurable finie. Supposons là moins fine que la partition α_{-k}^k, i.e. $\beta \preceq \alpha_{-k}^k$. Alors

$$\beta_{-n}^n \preceq (\alpha_{-k}^k)_{-n}^n = \alpha_{-n-k}^{n+k},$$

(comparer à l'exercice 2 du TD6), d'où

$$\begin{aligned}\frac{1}{2n+1}H(\beta_1^{2n+1}) &= \frac{1}{2n+1}H(\beta_{-n}^{n}) \\ &\leq \frac{1}{2n+1}H(\alpha_{-n-k}^{n+k}) = \frac{2n+2k+1}{2n+1}\frac{1}{2n+2k+1}H(\alpha_1^{2n+2k+1}),\end{aligned}$$

et, en faisant tendre n vers l'infini, $H(\Xi_\beta) \leq H(\Xi_\alpha)$.

3. Soit $\gamma \in \Phi$ une partition finie dont toutes les pièces ont une mesure non nulle. Par hypothèse, chaque pièce C_z, $z \in E_\gamma$ de γ appartient à la tribu engendrée par les pièces de tous les α_{-n}^n, à un ensemble de mesure nulle près. D'après le Lemme 103, cela entraîne que pour tout $\delta > 0$, il existe n et pour tout z, il existe un élément C'_z de la tribu engendrée par α_{-n}^n tel que $\mu(C_z \Delta C'_z) < \delta$. Soit β la partition engendrée par les C'_z, $z \in E_\gamma$. Les pièces sont toutes les intersections de C'_z. Pour chaque z, il y a une grosse pièce, $B_z = C'_z \setminus \bigcup_{z' \neq z'} C'_{z'}$, et de petites pièces de mesure $< \delta$, donc $E_\gamma \subset E_\beta$. Les variables aléatoires f_γ et f_β ne diffèrent que sur un ensemble de probabilité $p_e < |E_\gamma|\delta$. L'inégalité de Fano 90 entraîne que

$$\rho(\beta, \gamma) = H(\gamma|\beta) + H(\beta|\gamma) \leq 2h(p_e) + 2p_e \log_2 |E_\gamma|.$$

On conclut qu'en choisissant δ assez petit, on peut rendre $\rho(\beta, \gamma)$ arbitrairement petit. Par construction, $\beta \preceq \alpha_{-n}^n$ donc $\beta \in \Psi$. Cela prouve la densité de Ψ dans Φ, et permet d'étendre par continuité l'inégalité $H(\Xi_\beta) \leq H(\Xi_\alpha)$ prouvée pour $\beta \in \Psi$ à Φ tout entier.

On conclut que $h_\mu(T) = \max_{\beta \in \Phi} H(\Xi_\beta) = H(\Xi_\alpha)$. ∎

Corollaire 104 *L'entropie du décalage de Bernoulli $\mathcal{B}(p)$ vaut $H(p)$.*

L'entropie du décalage de Markov $\mathcal{M}(p, P)$ vaut $-\sum_{x,\, y \in E} p(x)P_{xy} \log_2(P_{xy})$.

7.5 Rotations du cercle

Il existe des transformations d'entropie nulle.

Proposition 105 *Soit $T : x \mapsto x + \theta$ une rotation d'angle irrationnel de $\mathbb{R}/\mathbb{Z}$, et λ la mesure de Lebesgue. Alors $h_\lambda(T) = 0$.*

Preuve On montre que la partition en 2 morceaux $\alpha = \{[0, \frac{1}{2}[, [\frac{1}{2}, 1[\}$ est génératrice. On utilise la densité des orbites (Exercice 6 du TD 4). Pour tout intervalle $[a, b]$ et tout $\epsilon > 0$, il existe des entiers n et m tels que $-n\theta \in]a - \epsilon, a[$, $-m\theta \in]b, b + \epsilon[$. Alors l'intersection des intervalles $T^{-n}[0, \frac{1}{2}[\cap T^{-m}[\frac{1}{2}, 1[= [-n\theta, -m\theta[$ est $[a, b]$. Ceci prouve que $[a, b] \in \alpha_{-\infty}^{\infty}$. Comme la tribu engendrée par les intervalles est la tribu borélienne $\mathcal{B}$, $\alpha_{-\infty}^{\infty} = \mathcal{B}$, et α est génératrice.

Le nombre de pièces dans la partition α_1^n est $2n$. En effet, couper une partition selon deux points opposés ne divise que deux pièces de la partition, donc $|\alpha_1^{n+1}| = |\alpha_1^n| + 2$. Par conséquent, $H(\alpha_1^n) \leq \log_2 |\alpha_1^n| = \log_2(2n) = o(n)$, donc $H(\Xi_\alpha) = 0$, et, avec le théorème 15, $h_\lambda(T) = 0$. ∎

7.6 Invariance de l'entropie

Définition 106 *Soient $(\Omega, \mathcal{B}, \mu)$ et $(\Omega', \mathcal{B}', \mu')$ des espaces probabilisés. On dit que deux transformations préservant la mesure $T : \Omega \to \Omega$ et $\Omega' \to \Omega'$ sont* isomorphes *ou* conjuguées *s'il existe une transformation préservant la mesure $\phi : \Omega \to \Omega'$ telle que $T' \circ \phi = \phi \circ T$ presque partout.*

Exemple 107 *Soit $\phi : \{0, 1\}^{\mathbb{Z}} \to [0, 2] \times [0, 1]$, définie pour $z = (z_n)_{n \in \mathbb{Z}}$ par*

$$\phi(x) = (\sum_{n=0}^{\infty} z_n 2^{-n}, \sum_{n=-\infty}^{-1} z_n 2^n).$$

Alors ϕ conjugue le décalage à la transformation $T' : [0, 2] \times [0, 1] \to [0, 2] \times [0, 1]$, définie pour $x \in [0, 2]$ et $y \in [0, 1]$, par

$$T'(x, y) = \begin{cases} (2x, \frac{1}{2}y) & \text{si } x \leq 1, \\ (2x - 2, \frac{1}{2}(1 + y)) & \text{si } x > 1. \end{cases}$$

Soit $w = T(z)$, i.e. $w_n = z_{n+1}$. Soit $(x, y) = \phi(z)$. Si $z_0 = 0$, alors $x \leq 1$. Si $z_0 = 1$, alors $x \geq 1$.

$$\begin{aligned}
\phi(T(z)) &= (\sum_{n=0}^{\infty} z_{n+1} 2^{-n}, \sum_{n=-\infty}^{-1} z_{n+1} 2^n) \\
&= (\sum_{m=1}^{\infty} z_m 2^{-(m-1)}, \sum_{m=-\infty}^{0} z_m 2^{m-1}) \\
&= (-2z_0 + 2\sum_{m=0}^{\infty} z_m 2^{-m}, \frac{z_0}{2} + \frac{1}{2} \sum_{m=-\infty}^{-1} z_m 2^m) \\
&= (-2z_0 + 2x, \frac{z_0}{2} + \frac{1}{2} y) \\
&= T'(x, y) = T'(\phi(z)),
\end{aligned}$$

sauf peut-être lorsque $x = 1$. Enfin, soit μ la mesure produit $p^{\otimes \mathbb{Z}}$ où p est la distribution uniforme sur $\{0, 1\}$. Alors $\phi_* \mu = \frac{1}{2}\lambda$ où λ désigne la mesure de Lebesgue sur $[0, 2] \times [0, 1]$. En effet, étant donnés $m \leq 0 \leq n \in \mathbb{Z}$ et $z_m, \ldots, z_n \in \{0, 1\}$, prolongeons z à $\mathbb{Z}$ par 0 et posons $(x_m, y_n) = \phi(z)$. Le cylindre $C_{z_m, \ldots, z_n}$, de mesure 2^{-n+m-1}, est envoyé sur le rectangle $[x_m, x_m + 2^m] \times [y_n, y_n + 2^{-n}]$, de mesure de Lebesgue 2^{-n+m}.

T' est parfois désignée sous le nom de *transformation du boulanger*, car elle rappelle le geste de l'artisan qui partage une boule de pâte en deux parts, en pose une sur l'autre et presse pour obtenir à nouveau une boule.

Proposition 108 *Si deux transformations sont isomorphes, et si l'une est ergodique, il en est de même de l'autre.*

Proposition 109 *Deux transformations isomorphes ont même entropie.*

En effet, un isomorphisme envoie classes de partitions modulo ensembles de mesure nulle sur classes de partitions modulo ensembles de mesure nulle, et, par naturalité, $H(\Xi_{\phi(\alpha)}) = H(\Xi_\alpha)$, donc $h_{\mu'}(T') = h_\mu(T)$.

Corollaire 110 *Deux décalages de Bernoulli, ou, plus généralement, de Markov, d'entropies différentes ne sont pas isomorphes.*

Exemple 111 *La transformation du boulanger, définie dans l'exemple 107, est ergodique et son entropie vaut* 1.

Nous ne démontrerons pas le théorème, assez difficile, suivant.

Théorème 16 (D. Ornstein 1970) *Deux décalages de Bernoulli qui ont la même entropie sont isomorphes.*

On dit que l'entropie est un *invariant d'isomorphisme complet* pour les décalages de Bernoulli. L'intérêt de cette classe est que de nombreux systèmes dynamiques d'intérêt courant sont isomorphes, du point de vue mesurable, à des décalages de Bernoulli.

Théorème 17 (Y. Katznelson, 1971) *Soit $A \in Gl(n, \mathbb{Z})$ une matrice entière inversible. On suppose que les valeurs propres de A ne sont pas des racines de l'unité. Alors la transformation T_A induite sur le tore est conjuguée à un décalage de Bernoulli.*

En revanche, une rotation du cercle n'est pas conjuguée à un décalage de Bernoulli, puisque son entropie est nulle.

7.7 Automorphismes du tore

Le tore, c'est l'espace quotient $\mathbb{R}^n/\mathbb{Z}^n$. C'est un groupe. On admet que ses automorphismes continus proviennent d'homomorphismes continus du groupe $\mathbb{R}^n$ dans lui-même. Ceux-ci coïncident avec les applications linéaires de $\mathbb{R}^n$ (cette étape n'est pas difficile). Une bijection linéaire $A : \mathbb{R}^n \to \mathbb{R}^n$ passe au quotient si et seulement si $A\mathbb{Z}^n \subset \mathbb{Z}^n$, i.e. si la matrice de A est à coefficients entiers. C'est une bijection du quotient si et seulement si A^{-1} est aussi à coefficients entiers, ce qui signifie que le déterminant de A vaut 1 ou -1. Le groupe des matrices $n \times n$ entières de déterminant ± 1 est noté $Gl(n, \mathbb{Z})$. La formule de changement de variable montre que les bijections linéaires associées à ces matrices préservent la mesure de Lebesgue de $\mathbb{R}^n$. Elles induisent des bijections de $\mathbb{R}^n/\mathbb{Z}^n$ qui préserve la mesure de probabilité naturelle sur le quotient, identifié au cube unité de $\mathbb{R}^n$.

Proposition 112 *Soit T l'automorphisme de $X = \mathbb{R}^n/\mathbb{Z}^n$ associé à la matrice $A \in Gl(n, \mathbb{Z})$. Alors T est ergodique si et seulement si aucune des valeurs propres de A n'est une racine de l'unité.*

Preuve On utilise le développement en série de Fourier des fonctions sur $\mathbb{R}^n/\mathbb{Z}^n$. Si $f \in L^2(\mathbb{R}^n/\mathbb{Z}^n)$, alors f est la somme de la série

$$f = \sum_{\xi \in \mathbb{Z}^n} \hat{f}_\xi e_\xi$$

où la famille de fonctions $e_\xi(x) = e^{2i\pi\xi\cdot x}$ est orthonormée pour le produit scalaire (hermitien) L^2. On remarque que $e_\xi \circ T = e_{A^\top \xi}$.

Si l'une des valeurs propres de A est une racine k-ème de l'unité, alors 1 est valeur propre de A^k. Le système linéaire à coefficients entiers $(A^k)^\top \xi = \xi$ possède une solution non nulle rationnelle, donc une solution non nulle entière ξ. Alors la fonction e_ξ est T^k-invariante. La fonction

$$f : X \to \mathbb{C}, \quad f = \sum_{j=1}^{k} e_\xi \circ T^j,$$

est T-invariante mais n'est pas presque partout constante (elle a des coefficients de Fourier non nuls). Cela prouve que T n'est pas ergodique.

Réciproquement, supposons qu'aucune valeur propre de A n'est une racine de l'unité. Alors pour tous vecteurs entiers non nuls ξ et ξ', et tout k assez grand, $(A^\top)^k \xi \neq \xi'$. En effet, sinon, il existe $k \neq k'$ tels que $(A^\top)^k \xi = \xi' = (A^\top)^{k'} \xi$. Alors $(A^\top)^{k'-k} \xi = \xi$, 1 est valeur propre d'une puissance de $A^\top$, donc A a une valeur propre qui est une racine de l'unité, contradiction. Par conséquent, pour tout ξ, ξ' et k assez grand, le produit scalaire $\langle e_{(A^\top)^k \xi}, e_{\xi'} \rangle = 0$. En particulier,

$$\lim_{k \to \infty} \langle e_\xi \circ T^k, e_{\xi'} \rangle = 0.$$

C'est évidemment encore vrai pour des combinaisons linéaires finies de fonctions e_ξ. D'après la décomposition de Fourier, ce sous-espace vectoriel est dense dans $L^2(\mathbb{R}^n/\mathbb{Z}^n)$. Comme dans la preuve de la Proposition 76, on en déduit que pour tous f, $g \in L^2(\mathbb{R}^n/\mathbb{Z}^n)$,

$$\lim_{k \to \infty} \langle f \circ T^k, g \rangle = 0,$$

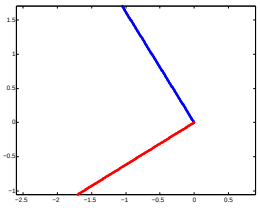
et l'ergodicité de T en résulte. ∎

Théorème 18 (Ya. Sinai, 1959) *Soit $A \in Gl(n, \mathbb{Z})$ une matrice entière inversible. On suppose que les valeurs propres de A ne sont pas des racines de l'unité. Soient $|\lambda_1| \geq \cdots \geq |\lambda_n|$ les modules des valeurs propres de A. Soient $|\lambda_1| \geq \cdots \geq |\lambda_k|$ ceux de ces nombres qui sont > 1. L'entropie métrique de la transformation T_A correspondante vaut*

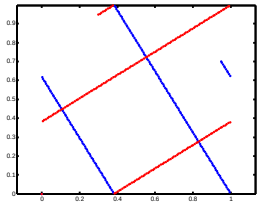
$$h(T_A) = \sum_{i=1}^{k} \log_2 |\lambda_k|.$$

Preuve On donne la preuve lorsque $n = 2$. Dans ce cas, T_A est ergodique si et seulement si ses valeurs propres sont réelles et distinctes (en effet, si elles sont non réelles, les valeurs propres sont conjuguées, de module 1 et de partie réelle entière ou demi-entière, donc des racines cubiques ou quatrièmes de l'unité). Les valeurs propres $\lambda_1 > \lambda_2 = 1/\lambda_1$ sont irrationnelles.

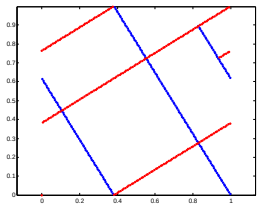
On construit une partition finie α. Chaque pièce est l'image dans $\mathbb{R}^2/\mathbb{Z}^2$ d'un parallélogramme dont les côtés sont parallèles aux deux espaces propres de A. Pour la construire on part de deux longs segments de droites propres issus de l'origine dans $\mathbb{R}^2$.



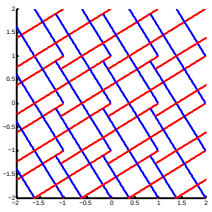
Leurs images γ_1 et γ_2 dans le tore ont une extrémité commune, à l'origine.



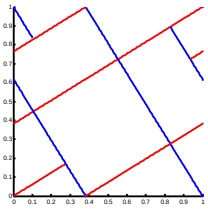
Si l'autre extrémité de γ_1 n'est pas sur γ_2, on prolonge γ_1 jusqu'à ce qu'elle le soit. On fait de même avec γ_2.



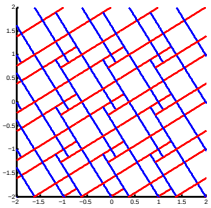
L'image réciproque du réseau de lignes obtenu divise le plan en polygones qui ont un sommet rentrant aux points entiers.



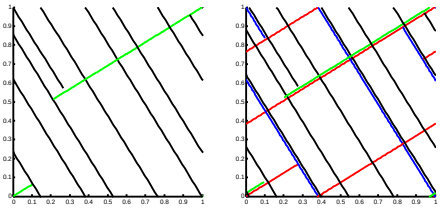
Pour y remédier, on prolonge γ_1 dans la direction opposée, au-delà de l'origine, jusqu'à ce que son extrémité soit sur γ_2, et on fait de même pour γ_2. De la sorte, on obtient un réseau de lignes sur le tore, qui définit une partition α (à ensembles de mesure nulle près).



Cette fois, l'image réciproque du réseau dans le plan divise le plan en polygones localement convexes (donc convexes) à côtés parallèles aux droites propres, ce sont des parallélogrammes.



L'image réciproque par A^n d'un parallélogramme à côtés parallèles aux droites propres est un parallélogramme à côtés parallèles aux droites propres, mais λ_1^{-n} fois plus court dans la direction de γ_1 et λ_2^{-n} fois plus long dans la direction de γ_2. Soit L_i (resp. ℓ_i) la plus grande (resp. plus petite) longueur des côtés parallèles à γ_i des pièces de α. Alors les longueurs correspondantes dans la partition $T_{A^n}\alpha$ sont comprises entre $\ell_i\lambda_i^{-n}$ et $L_i\lambda_i^{-n}$.



Les partitions $T_A\alpha$ et $\alpha \vee T_A\alpha$.

Notons β_i la réunion des lignes parallèles à γ_i (son image réciproque dans le plan est la réunion des images de γ_1 par les translations entières). Comme $\gamma_1 \subset A^{-1}(\gamma_1)$ et $A^{-1}(\gamma_2) \subset \gamma_2$, $\beta_1 \subset A^{-1}(\beta_1)$ et $A^{-1}(\beta_2) \subset \beta_2$. Par conséquent, dans la partition $\alpha \vee T_A\alpha$, chaque parallélogramme est divisé en parallélogrammes de même longueur (dans la direction de γ_1), mais de largeur (dans la direction de γ_2) divisée par un facteur de l'ordre de λ_1. De même, dans la partition $\alpha \vee T_A\alpha \vee \cdots \vee (T_A)^{n-1}\alpha$, les pièces sont des parallélogrammes de longueurs (dans la direction de γ_1) comprises entre ℓ_1 et L_1 et de largeurs (dans la direction de γ_2) comprises entre $\ell_2\lambda_2^{-n}$ et $L_2\lambda_2^{-n}$. Il en résulte que leurs aires sont comprises entre $\ell_1\ell_2\lambda_1^{-n}$ et $L_1L_2\lambda_1^{-n}$. L'entropie satisfait

$$\begin{aligned} H(\alpha \vee T_A\alpha \vee \cdots \vee (T_A)^{n-1}\alpha) &= \sum_B -\mu(B)\log 2\mu(B) \\ &\geq \sum_B -\mu(B)\log 2(L_1L_2\lambda_1^{-n}) \\ &\geq n\log_2(\lambda_1) - \log_2(L_1L_2). \end{aligned}$$

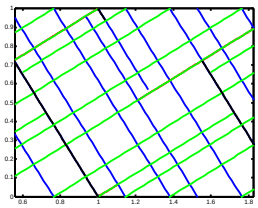
De même,

$$\begin{aligned} H(\alpha \vee T_A\alpha \vee \cdots \vee (T_A)^{n-1}\alpha) &= \sum_B -\mu(B)\log 2\mu(B) \\ &\leq \sum_B -\mu(B)\log 2(\ell_1\ell_2\lambda_1^{-n}) \\ &\leq n\log_2(\lambda_1) - \log_2(\ell_1\ell_2). \end{aligned}$$

On conclut que

$$H(\Xi_{\alpha,T_A}) = \lim_{n\to\infty} \frac{1}{n} H(\alpha \vee T_A\alpha \vee \cdots \vee (T_A)^{n-1}\alpha) = \log_2(\lambda_1).$$

Les pièces de la partition $T^{-n}\alpha$ sont des parallélogrammes dont les côtés parallèles à γ_1 ont des longueurs de l'ordre de λ_2^n. Par conséquent, les pièces de la partition $T^{-n}\alpha \vee T^n\alpha$ ont un diamètre qui tend vers 0.



La partition $T_A^{-1}\alpha \vee T_A\alpha$.

Soit $\epsilon > 0$, soit R un rectangle du tore. Pour n assez grand, il existe une réunion U de pièces de $T^{-n}\alpha \vee T^n\alpha$ telle que $\mu(R\Delta U) < \epsilon$. Cela prouve que la tribu $\alpha_{-\infty}^{\infty}$ contient R. Comme les rectangles engendrent la tribu borélienne du tore, α est génératrice. D'après le Théorème 15,

$$h_\mu(T_A) = H(\Xi_{\alpha,T_A}) = \log_2(\lambda_1).$$

■

7.8 Chaos

Ce terme n'a pas de définition mathématique précise. Il évoque la sensibilité aux conditions initiales : *Un système dynamique est chaotique si deux points très voisins peuvent néanmoins avoir des trajectoires qui s'éloignent.* Les automorphismes ergodiques du tore ont cette propriété. D'une certaine façon, l'entropie quantifie ce phénomène : étant donnée une partition α arbitrairement fine, la transformation en fait des partitions qui recoupent α suivant une multitude de pièces. Entropie h positive signifie qu'il y a en gros 2^{nh} pièces de mesures à peu près égales.

7.9 A retenir

- Le lien entre processus stationnaires et systèmes dynamiques.
- La notion de conjugaison dans la catégorie des espaces probabilisés.
- La notion d'entropie métrique, invariant de conjugaison dans cete catégorie.
- Le lien entre entropie et chaos.

Il s'agit d'un chapitre théorique, où une nouvelle notion a été introduite, dans un autre champ des mathématiques, les systèmes dynamiques. Elle s'applique à certains des exemples les plus fascinants de systèmes dynamiques. Il y a un théorème dont la preuve est difficile, car elle s'appuie sur des manipulations de tribus auxquelles on est peu familier en M1.