THÉORIE DE L'INFORMATION

Antoine Chambert-Loir

Antoine Chambert-Loir
Université Paris-Diderot.
E-mail : Antoine.Chambert-Loir@math.univ-paris-diderot.fr

Version du 5 décembre 2018, 15h55
La version la plus à jour de ce texte devrait être accessible en ligne, à l'adresse
http://webusers.imj-prg.fr/~antoine.chambert-loir/enseignement/2018-19/shannon/it.pdf

BIBLIOGRAPHIE

T. M. Cover & J. A. Thomas (2006), *Elements of information theory*, Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second édition.

Hearing range (2005), « Hearing range — Wikipedia, the free encyclopedia ». URL *https://en.wikipedia.org/w/index.php?title=Hearing_range*, consulté le 20 novembre 2018.

C. E. Shannon (1948), « A mathematical theory of communication ». *Bell System Tech. J.*, **27**, p. 379–423, 623–656.

C. E. Shannon (1949), « Communication in the presence of noise ». *Proc. I.R.E.*, **37**, p. 10–21.

C. E. Shannon & W. Weaver (2018), *La théorie mathématique de la communication*, Cassini.

Démonstration. — Considérons la série

$$\mathrm{F}(x) = \sum_{m\in\mathbf{Z}} f(x+am).$$

La décroissance de f et de $\widehat{f}$ entraînent que cette série converge uniformément, de même que sa dérivée terme à terme, lorsque x parcourt un intervalle borné. La fonction F ainsi définie est ainsi de classe $\mathscr{C}^1$ sur $\mathbf{R}$; elle est aussi de période a. Calculons ses coefficients de Fourier : pour tout $n \in \mathbf{Z}$, on a

$$c_n(\mathrm{F}) = \frac{1}{a}\int_0^a \mathrm{F}(x)e^{-2i\pi nx/a}\,dx = \frac{1}{a}\int_0^a \sum_{m\in\mathbf{Z}} f(x+ma)e^{-2i\pi nx/a}\,dx.$$

Par convergence uniforme de la série pour $x \in [0;a]$, on peut intervertir intégration et sommation, d'où

$$\begin{aligned}
c_n(\mathrm{F}) &= \frac{1}{a}\sum_{m\in\mathbf{Z}}\int_0^a f(x+ma)e^{-2i\pi nx/a}\,dx\\
&= \frac{1}{a}\sum_{m\in\mathbf{Z}}\int_{ma}^{(m+1)a} f(x)e^{-2i\pi n(x-ma)/a}\,dx\\
&= \frac{1}{a}\sum_{m\in\mathbf{Z}}\int_{ma}^{(m+1)a} f(x)e^{-2i\pi nx/a}\,dx\\
&= \frac{1}{a}\int_{-\infty}^{\infty} f(x)e^{-2i\pi nx/a}\,dx\\
&= \widehat{f}(n/a).
\end{aligned}$$

Comme F est de classe $\mathscr{C}^1$, elle est somme de sa série de Fourier, d'où la relation :

$$\sum_{m\in\mathbf{Z}} f(x+am) = \frac{1}{a}\sum_{n\in\mathbf{Z}} \widehat{f}(n/a)e^{2i\pi nx/a},$$

pour tout $x \in \mathbf{R}$. En prenant $x = 0$, on obtient l'autre relation. □

TABLE DES MATIÈRES

C'est une formule analogue à la formule de reconstruction (3.6.1.1) que l'on retrouve formellement en prenant $W = W_0$ et pour fonction ψ la fonction indicatrice de $[-W_0; W_0]$. Toutefois, lorsque $W > W_0$ (il y a alors *suréchantillonnage* par rapport à la fréquence de Nyquist $2W_0$), on peut prendre pour ψ une fonction de classe $\mathscr{C}^\infty$. Sa transformée de Fourier $\widehat{\psi}$ décroît alors rapidement à l'infini si bien que la série donnée dans la proposition converge très vite.

Démonstration. — On note encore φ la fonction de période W qui coïncide avec $\widehat{f}$ sur $[-W; W]$; elle est $\mathscr{C}^1$ par morceaux et ses coefficients de Fourier sont donnés par

$$c_n(\varphi) = f(-n/2W)/2W,$$

comme dans la démonstration du théorème 3.6.1. Pour tout $y \in \mathbf{R}$, on peut alors écrire

$$\widehat{f}(y) = \varphi(y)\psi(y),$$

puisque cette égalité devient $\widehat{f}(y) = \widehat{f}(y)\psi(y)$ si $y \in [-W; W]$, égalité vraie car $\psi(y) = 1$ si $\widehat{f}(y) \neq 0$, et qu'elle se réduit à $0 = 0$ si $y \notin [-W; W]$. Ainsi, en écrivant φ comme la somme de sa série de Fourier, on a

$$\widehat{f}(y) = \sum_{n\in\mathbf{Z}} \frac{1}{2W} f(n/2W) e^{-2i\pi ny} \psi(y).$$

Appliquons maintenant la formule d'inversion de Fourier à cette égalité; on trouve

$$f(x) = \sum_{n\in\mathbf{Z}} \frac{1}{2W} f(-n/2W) \int_{-\infty}^{\infty} e^{-2i\pi ny/2W} \psi(y) e^{2i\pi xy}\, dy.$$

Cette dernière intégrale vaut $\widehat{\psi}(-x + n/2W)$, d'où la proposition. □

La démontration du théorème d'échantillonnage est en fait très proche de celle d'une formule importante en mathématiques, la *formule sommatoire de Poisson*.

Théorème **(3.6.5)** (Formule de Poisson). — *Soit $f : \mathbf{R} \to \mathbf{C}$ une fonction décroissant assez vite, de même que sa transformée de Fourier $\widehat{f}$. Alors, pour tout nombre réel $a > 0$, on a*

(3.6.5.1)
$$\sum_{m\in\mathbf{Z}} f(am) = \frac{1}{a} \sum_{n\in\mathbf{Z}} \widehat{f}(n/a).$$

Plus généralement, pour tout nombre réel $a > 0$ et tout nombre réel x, on a

(3.6.5.2)
$$\sum_{m\in\mathbf{Z}} f(x + am) = \frac{1}{a} \sum_{n\in\mathbf{Z}} \widehat{f}(n/a) e^{2i\pi nx/a}.$$

3.6.3. Démonstration du théorème d'échantillonnage. — Puisque f est continue, intégrable, et que $\widehat{f}$ est nulle en dehors de l'intervalle borné $[-W; W]$, on a la formule d'inversion de Fourier

$$f(x) = \int_{-W}^{W} \widehat{f}(y) e^{2\pi i x y}\, dy.$$

En particulier, f est indéfiniment dérivable.

Soit φ la fonction de période W qui coïncide avec $\widehat{f}$ sur $[-W; W]$. Comme $\widehat{f}(W) = \widehat{f}(-W) = 0$, elle est continue. Calculons ses coefficients de Fourier : pour $n \in \mathbf{Z}$, on a

$$\begin{aligned} c_n(\varphi) &= \frac{1}{2W} \int_{-W}^{W} \varphi(y) e^{-2\pi i n y/2W}\, dy \\ &= \frac{1}{2W} \int_{-\infty}^{\infty} \widehat{f}(y) e^{-2\pi i n y/2W}\, dy \\ &= \frac{1}{2W} f(-n/2W). \end{aligned}$$

Le développement en série de Fourier de φ est ainsi

$$\varphi(y) = \sum_{n\in\mathbf{Z}} c_n(\varphi) e^{2\pi i n y/2W} = \frac{1}{2W} \sum_{n\in\mathbf{Z}} f(n/2W) e^{-2\pi i n y/2W}.$$

Comme $n^2 f(n/2W)$ est borné, cette série converge uniformément. On peut donc la reporter dans la formule d'inversion de Fourier et intégrer terme à terme, ce qui fournit, puisque $\widehat{f}(y) = \varphi(y)$ pour $y \in [-W; W]$ et $\widehat{f}(y) = 0$ sinon,

$$\begin{aligned} f(x) &= \frac{1}{2W} \sum_{n\in\mathbf{Z}} f(n/2W) \int_{-W}^{W} e^{2\pi i (x - n/2W) y}\, dy \\ &= \sum_{n\in\mathbf{Z}} f(n/2W)\, \mathrm{sinc}(x - n/2W), \end{aligned}$$

comme il fallait démontrer.

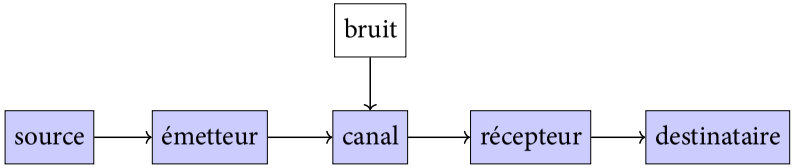
Proposition **(3.6.4).** — *Soit $f : \mathbf{R} \to \mathbf{C}$ une fonction intégrable ; on suppose que $\widehat{f}$ est nulle hors d'un intervalle $[-W_0; W_0]$. Soit W un nombre réel tel que $W \geqslant W_0$ et soit $\psi : \mathbf{R} \to \mathbf{C}$ une fonction nulle hors de $[-W; W]$ et identiquement égale à 1 sur l'intervalle $[-W_0; W_0]$. Alors, pour tout $x \in \mathbf{R}$, on a*

$$f(x) = \frac{1}{2W} \sum_{n\in\mathbf{Z}} f(n/2W) \widehat{\psi}(n/2W - x).$$

INTRODUCTION

La *théorie mathématique de la communication* vise à étudier de façon mathématiques dans quelles conditions on peut transmettre des données, en particulier à quelle vitesse, et avec quelle fiabilité.

Dans l'article fondateur de Shannon (1948), un système de communication est modélisé par le diagramme suivant :



– La *source* est l'entité qui possède l'information à transmettre à son *destinataire* ; ce peut-être une station de radio ou de télévision, un journal, un site web, vous ou moi désirant annoncer une mauvaise nouvelle au téléphone ou par courrier électronique, une sonde spatiale prenant des photos des planètes qu'elle survole, etc. L'information peut-être un texte, une photographie, un signal sonore, une combinaison de ceux-ci. Dans le cas du téléphone ou de la radio, ce sera un signal sonore dont l'amplitude sera représentée par une fonction du temps, ou par deux telles fonctions pour un signal stéréo ; dans le cas de la télévision couleur, il s'agira de transmettre les trois amplitudes (rouge/vert/bleu) en chaque point de l'écran, et les deux composantes du son, le tout dépendant du temps.

– L'*émetteur* est l'appareil physique par lequel nous allons créer cette information, l'émetteur de radio ou de télévision. À l'époque de Shannon, la transmission était souvent analogique ; dans le cas du téléphone, par exemple, l'amplitude de la pression sonore était transformée en un signal électrique proportionnel.

De nos jours, la transmission est aujourd'hui essentiellement numérique[1] : le signal est transformé en une suite de nombres qu'il s'agit de transmettre.

– Le *récepteur* est l'appareil par lequel le destinataire reçoit cette information, un poste de radio ou de télévision, éventuellement associé à un « décodeur » dans le cas de la télévision numérique terrestre ou de la télévision par Internet, un téléphone, un ordinateur relié au réseau Internet, etc.

– Le *canal* est le medium physique par lequel l'information est transmise de l'émetteur au récepteur : l'air pour la transmission de la radio/télévision par voie hertzienne, la fibre optique du fournisseur Internet, les câbles en cuivre du réseau de téléphone, etc. Comme tout objet physique, ce canal est sujet à des perturbations — du *bruit* — par lesquelles le signal qui parvient au récepteur diffère de celui envoyé par l'émetteur.

La théorie mathématique de la communication vise à analyser dans quelles conditions un canal donné, soumis à un certain bruit, peut, ou pas, transmettre l'information voulue. Deux théorèmes de Shannon (1948) répondent ainsi aux questions suivantes :

a) À quelle vitesse est-il possible de transmettre cette information ?

b) En présence de bruit, est-il possible de transmettre cette information de manière fiable ?

Si le canal permet de diffuser c symboles par unité de temps, il semble évident qu'on peut transmettre un message de N symboles pendant un temps N/c, mais peut-on faire mieux ? Ensuite, comment détecter une mauvaise transmission de certains symboles et, éventuellement, les corriger ?

Le présupposé de base de la théorie est que les messages à transmettre, du fait-même de leur origine, ne sont pas arbitraires. Si c'est un texte, certaines lettres seront plus fréquentes que d'autres ; si c'est l'enregistrement d'une voix ou d'un morceau de musique certaines fréquences seront absentes du signal, sans même tenir compte du fait que l'oreille humaine ne les perçevra pas.

Dans son article, Shannon (1948) propose une mesure de la « quantité d'information » contenue dans un signal, qu'il appelle *entropie*. Plus exactement, il s'agit de la quantité d'information contenue dans l'ensemble des signaux susceptibles d'être transmis. Sa définition est de nature probabiliste.

[1] À l'exception notable de la radio FM, la radio numérique terrestre (DAB) peinant à décoller.

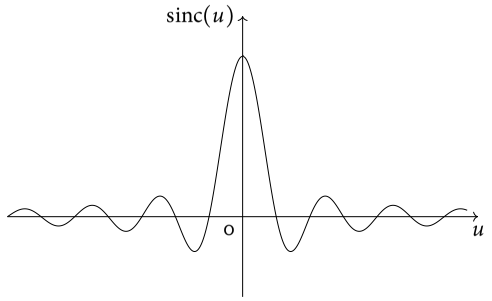


FIGURE 3.5.2.2. Graphe du sinus cardinal

3.6. Théorème d'échantillonnage

Théorème **(3.6.1)**. — *Soit* $f : \mathbf{R} \to \mathbf{C}$ *une fonction continue telle que* $x^2 f(x)$ *soit borné, lorsque* x *varie. Alors* f *est intégrable sur* $\mathbf{R}$*, et on suppose qu'il existe un nombre réel* W *tel que* $\widehat{f}(y) = 0$ *pour tout* $y \in \mathbf{R}$ *tel que* $|y| \geqslant W$*. Alors, pour tout* $x \in \mathbf{R}$*, on a la formule,*

$$\textbf{(3.6.1.1)} \qquad f(x) = \sum_{n=-\infty}^{\infty} f\left(\frac{n}{2W}\right) \operatorname{sinc}\left(x - \frac{n}{2W}\right).$$

Autrement dit, en échantillonnant le signal donné par f à la fréquence 2W, on peut le reconstituer exactement !

3.6.2. — Selon Shannon (1949), l'idée que l'on puisse reconstruire un signal en l'échantillonnant à une fréquence au moins double de la plus grande fréquence intervenant dans un signal était bien connue des spécialistes de la théorie de la communication — « *common knowledge in the communication art* ». L'apport de cet article est ainsi de donnner une formule explicite pour cette reconstruction. Shannon (1949) admet aussi que cette formule avait déjà été démontrée, sous une forme ou sous une autre, par divers mathématiciens ; citons par exemple E. Whittaker (1915), Oguro (1920), Kotel'nikov (1933). Il insiste cependant que c'est la première fois qu'elle est explicitée dans le contexte de la théorie de la communication.

Cette histoire un peu balbutiante explique peut-être pourquoi ce théorème d'échantillonnage est parfois dénommé *théorème de Nyquist–Shannon*.

Posons

$$\mathrm{F}_n(t) = \frac{1}{n+1}\sum_{k=0}^{n} \mathrm{D}_k(t)\,;$$

cette fonction F_n est appelée *noyau de Fejér* et on déduit de l'équation (3.3.3.3) que l'on a

$$\mathrm{T}_n(f)(x) = \int_0^{\mathrm{T}} f(t)\mathrm{F}_n(x-t)\,dt.$$

Le noyau de Fejér est pair, T-périodique et d'intégrale 1, et un calcul explicite prouve qu'il est positif et, plus précisément, vérifie les conditions de l'argument de convolution. Ainsi, si f a des limites à droite et à gauche en x, $\mathrm{T}_n(f)(x)$ converge vers sa régularisée $f^*(x)$.

3.4. Formule de Parseval

3.5. Transformation de Fourier

3.5.1. — Soit $f : \mathbf{R} \to \mathbf{C}$ une fonction localement intégrable telle que $\int_{-\infty}^{\infty}|f| < +\infty$. On définit alors sa *transformée de Fourier* $\widehat{f}$ par la formule

(3.5.1.1)
$$\widehat{f}(y) = \int_{-\infty}^{\infty} f(x)e^{-2\pi i x y}\,dx.$$

Exemple **(3.5.2).** — Soit W un nombre réel $\geqslant 0$ et soit $f : \mathbf{R} \to \mathbf{C}$ la fonction définie par $f(x) = 1$ pour $x \in [-\mathrm{W};\mathrm{W}]$, et par $f(x) = 0$ sinon. On a, pour $y \neq 0$,

$$\begin{aligned}\widehat{f}(y) &= \int_{-\infty}^{\infty} f(x)e^{-2\pi i x y}\,dx = \int_{-\mathrm{W}}^{\mathrm{W}} e^{-2\pi i x y}\,dx \\ &= \left[\frac{1}{-2\pi i y}e^{-2\pi i x y}\right]_{-\mathrm{W}}^{\mathrm{W}} = \frac{e^{-2\pi i \mathrm{W} y} - e^{2\pi i \mathrm{W} y}}{-2\pi i y} = \frac{\sin(2\pi y \mathrm{W})}{\pi y} \\ &= 2\mathrm{W}\,\mathrm{sinc}(2\pi y \mathrm{W}),\end{aligned}$$

où la fonction sinc, usuellement appelée *sinus cardinal*, est définie par

(3.5.2.1)
$$\mathrm{sinc}(y) = \sin(y)/y$$

pour $y \neq 0$, et $\mathrm{sinc}(0) = 1$. Pour $y = 0$, on trouve $\widehat{f}(y) = 2\mathrm{W}$, de sorte que cette formule est encore valable.

CHAPITRE 1

ENTROPIE ET INFORMATION MUTUELLE

1.1. Entropie d'une variable aléatoire

1.1.1. — On désigne ici par $\log : \mathbf{R}_{>0} \to \mathbf{R}$ la fonction logarithme usuelle, fonction réciproque de la fonction exponentielle, autrement dit le logarithme néperien. C'est une fonction de classe $\mathscr{C}^\infty$, concave, stictement croissante. On a aussi les limites, pour tout nombre réel $\alpha > 0$:

(1.1.1.1)
$$\lim_{x\to 0^+} x^\alpha \log(x) = 0, \qquad \lim_{x\to+\infty} x^{-\alpha}\log(x) = 0.$$

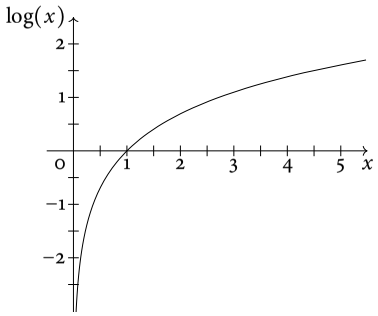


FIGURE 1.1.1.2. Graphe de la fonction « logarithme néperien »

La fonction $x \mapsto -x\log(x)$ de $]0;1]$ dans $\mathbf{R}$ est à valeurs positives ou nulles. Elle a pour limite 0 en 0, ce qui permet de la prolonger par continuité en 0, de valeur 0. Cela justifie aussi la convention d'écriture $0 \times \log(0) = 0$.

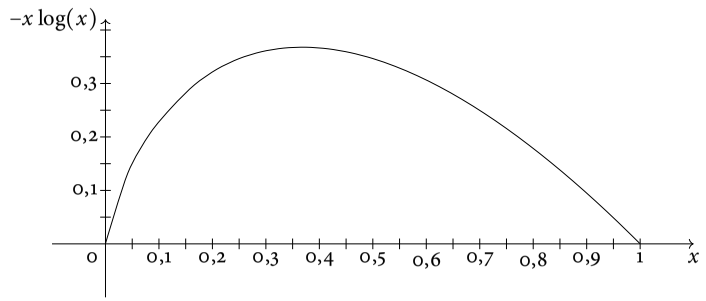


FIGURE 1.1.1.3. Graphe de la fonction $x \mapsto -x\log(x)$

1.1.2. — Si a est un nombre réel > 0, le « logarithme en base a » est la fonction donnée par

$$\log_a(x) = \frac{\log(x)}{\log(a)}.$$

Elle vérifie des propriétés similaires au logarithme néperien, qui est le cas où $a = e = 2{,}718\ldots$. Le cas $a = 10$ est courant en physique ; en théorie de l'information, nous verrons qu'il est naturel de prendre $a = 2$.

Définition **(1.1.3)**. — *L'entropie d'une variable aléatoire discrète* X *est définie par*

(1.1.3.1) $$\mathrm{H}(\mathrm{X}) = \sum_x \left(-\mathbf{P}(\mathrm{X} = x)\log(\mathbf{P}(\mathrm{X} = x))\right).$$

L'entropie est donc définie comme la somme d'une série, indexée par les valeurs possibles x de la variable aléatoire X. Dans cette définition, on utilie la convention $0\log(0) = 0$; on peut donc ne considérer, si l'on veut, que les valeurs x pour lesquelles $\mathbf{P}(\mathrm{X} = x)$ est strictement positive. Cette série est à termes positifs car une probabilité appartient à $[0;1]$, donc son logarithme est négatif. Il en résulte donc que la somme de cette série est bien définie, en tant qu'élément de $[0;+\infty]$. Elle est finie si X ne prend qu'un nombre fini de valeurs.

On peut bien sûr la définir dans toute base $a > 0$.

(1.1.3.2) $$\mathrm{H}_a(\mathrm{X}) = \sum_x \left(-\mathbf{P}(\mathrm{X} = x)\log_a(\mathbf{P}(\mathrm{X} = x))\right) = \frac{\mathrm{H}(\mathrm{X})}{\log(a)}.$$

On verra que l'entropie en base 2 d'une variable aléatoire est le nombre moyen de questions binaires qu'il faut poser pour espérer connaître son résultat.

Pour tout nombre réel r tel que $0 \leqslant r < 1$, posons

(3.3.7.1) $$\mathrm{P}_r(f)(x) = \sum_{k=-\infty}^{\infty} c_k(f) r^{|k|} e^{2\pi i k x/\mathrm{T}}.$$

Comme les coefficients de Fourier de f sont bornés, le facteur supplémentaire $r^{|k|}$ entraîne une majoration des termes de cette série (infinie dans les deux sens) par les termes d'une série géométrique, si bien que cette série converge normalement et sa limite est une fonction continue de x. On peut exprimer $\mathrm{P}_r(f)$ de façon analogue à la formule (3.3.3.3) :

$$\begin{aligned}\mathrm{P}_r(f)(x) &= \sum_{k=-\infty}^{\infty} \frac{1}{\mathrm{T}} \int_0^{\mathrm{T}} f(t) e^{-2\pi i k t/\mathrm{T}}\, dt r^{|k|} e^{2\pi k x/\mathrm{T}} \\ &= \int_0^{\mathrm{T}} f(t) \frac{1}{\mathrm{T}} \sum_{k=-\infty}^{\infty} r^{|k|} e^{2\pi i k (x-t)/\mathrm{T}}\, dt \\ &= \int_0^{\mathrm{T}} f(t) \mathrm{K}_r(x-t)\, dt,\end{aligned}$$

où la fonction K_r, *noyau de Poisson*, est défini par

$$\mathrm{K}_r(t) = \frac{1}{\mathrm{T}} \sum_{k=-\infty}^{\infty} r^{|k|} e^{2\pi i k t/\mathrm{T}}.$$

On peut, en fait, calculer précisément K_r :

$$\begin{aligned}\mathrm{K}_r(t) &= 1 + \sum_{k=1}^{\infty} r^k e^{2\pi i k t/\mathrm{T}} + \sum_{k=1}^{\infty} r^k e^{-2\pi i k t/\mathrm{T}} \\ &= 1 + \frac{r e^{2\pi i t/\mathrm{T}}}{1 - r e^{2\pi i t/\mathrm{T}}} + \frac{r e^{-2\pi i t/\mathrm{T}}}{1 - r e^{-2\pi i t/\mathrm{T}}} \\ &= \frac{1 - r^2}{1 - 2r\cos(2\pi t/\mathrm{T}) + r^2}\end{aligned}$$

Cette formule permet de constater que le noyau de Poisson est positif et converge uniformément vers 0 sur tout intervalle $[\delta; \mathrm{T} - \delta]$; sa définition montre aussi qu'il est pair et d'intégrale 1. Alors, un argument de convolution démontre que $\mathrm{K}_r(t)(x)$ converge vers la régularisée $f^*(x)$ de f en tout point x tel que f ait des limites à droite et à gauche.

Un autre procédé consiste à introduire les moyennes, au sens de Cesáro, des sommes partielles $\mathrm{S}_n(f)(x)$, et de poser

(3.3.7.2) $$\mathrm{T}_n(f)(x) = \frac{1}{n+1} \sum_{k=0}^{n} \mathrm{S}_n(f)(x).$$

soient inférieures à ε.

Choisissons ensuite une suite finie croissante $(a_0, \dots, a_n)$ telle que $a = a_0$ et $a_n = b$, et telle que pour tout $k \in \{1, \dots, n\}$, la fonction h soit continue sur $]a_{k-1}, a_k[$ et ait une limite à droite en a_{k-1}, et à gauche en a_k. Notons h_k la fonction continue sur $[a_{k-1}, a_k]$ qui coïncide avec h sur l'intervalle ouvert $]a_{k-1}, a_k[$. Elle est uniformément continue : il existe donc un nombre réel $\delta > 0$ tel que pour tous $x, y \in [a_{k-1}, a_k]$ tels que $|x - y| \leqslant \delta$, on ait $|h_k(x) - h_k(y)| < \varepsilon$. Subdivisons alors l'intervalle $[a_{k-1}, a_k]$ en sous-intervalles de longueur $< \delta$ et notons g_k la fonction constante sur l'intérieur de chacun de ces intervalles et qui coïncide avec h_k en leur milieu $b_k = (a_k + a_{k-1})/2$. Sur chacun de ces intervalles I, on a $|g_k(x) - h_k(x)| < \varepsilon$, donc l'intégrale sur I de $|g_k - h_k|$ est majorée par $\varepsilon\ell(\mathrm{I})$. Finalement, $\int_{a_{k-1}}^{a_k} |g_k - h_k| < \varepsilon(a_k - a_{k-1})$. Soit g une fonction sur $\mathbf{R}$ qui, pour tout k, coïncide avec g_k sur l'intervalle $]a_{k-1}, a_k[$, et est nulle hors de l'intervalle $[a; b]$. C'est une fonction en escalier et l'on a

$$\int_{-\infty}^{\infty} |g - h| = \int_{-\infty}^{a} + \int_{a}^{b} + \int_{b}^{\infty} |g - h| \leqslant \varepsilon(2 + b - a).$$

De cette inégalité, on déduit la majoration

$$|\widehat{g}(\omega) - \widehat{h}(\omega)| = \left|\int_{-\infty}^{\infty} (g(t) - h(t)) e^{-i\omega t}\, dt\right| \leqslant \int_{-\infty}^{\infty} |g(t) - h(t)|\, dt \leqslant \varepsilon(2 + b - a),$$

pour tout $\omega \in \mathbf{R}$. Par ailleurs, on a

$$\begin{aligned}\widehat{g}(\omega) &= \sum_{k=1}^{n} h(b_k) \int_{a_{k-1}}^{a_k} e^{-i\omega t}\, dt \\ &= \frac{1}{\omega} \sum_{k=1}^{n} ih(b_k)(e^{-i\omega a_k} - e^{-i\omega a_{k-1}}),\end{aligned}$$

formule qui prouve que

$$\lim_{\omega \to \pm\infty} \widehat{g}(\omega) = 0.$$

En particulier, il existe $\mathrm{W} \in \mathbf{R}$ tel que $|\widehat{g}(\omega)| \leqslant \varepsilon$ dès que ω vérifie $|\omega| \geqslant \mathrm{W}$. Pour un tel ω, on a donc $|\widehat{h}(\omega)| \leqslant \varepsilon(3 + b - a)$. La démonstration est ainsi terminée. □

Remarque **(3.3.7)**. — Dans les cours de Licence, on apprend à « sommer » une série en considérant ses sommes partielles, mais il y a de nombreuses autres procédés. Deux d'entre eux, au moins, sont particulièrement adaptés à la théorie des séries de Fourier.

1.1.4. Exemple : lancer d'un dé. — Considérons un dé à 6 faces, équilibré. La probabilité d'apparition de chacune des faces est donc 1/6 ; l'entropie de la variable aléatoire correspondante est ainsi égale à $6 \cdot (-\frac{1}{6}\log(\frac{1}{6})) = \log(6)$.

Plus généralement, une variable aléatoire X prenant N valeurs, chacune avec probabilité 1/N, a pour entropie $\log(\mathrm{N})$. Imaginons que N soit une puissance de 2, $\mathrm{N} = 2^n$, et que X prenne ses valeurs parmi $\{0, \dots, \mathrm{N} - 1\}$. Alors, on peut connaître le résultat de X en posant successivement n questions « binaires », à savoir quels sont les chiffres du développement binaire de X. Dans ce cas, on a $\mathrm{H}_2(\mathrm{X}) = \log_2(\mathrm{N}) = n$. Plus généralement, on verra que l'entropie en base 2 d'une variable aléatoire est (à une unité près) le nombre moyen de questions binaires qu'il faut poser pour espérer connaître son résultat.

Considérons maintenant deux dés à 6 faces, équilibrés, et prenons considérons la variable aléatoire Y la somme des valeurs des deux faces. Elle peut prendre les valeurs $2, 3, \dots, 12$; la valeur 2 n'est possible que pour le tirage $(1, 1)$, la valeur 3 apparaît pour deux tirages $(1, 2)$ et $(2, 1)$, etc. Les probabilités des événements $\mathrm{X} = x$ sont ainsi résumées par le tableau :

2	3	4	5	6	7	8	9	10	11	12
1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

et son entropie en base 2 est égale à

$$\mathrm{H}_2(\mathrm{Y}) = -\frac{1}{36}\log_2\left(\frac{1}{36}\right) - \cdots - \frac{1}{36}\log_2\left(\frac{1}{36}\right) \approx 3{,}27440191928877$$

alors que l'entropie d'une variable aléatoire identiquement distribuée parmi $\{2, \dots, 12\}$ est égale à

$$\log_2(11) \approx 3{,}45943161863730.$$

Il y a un peu moins de hasard dans le résultat de la somme de deux dés que dans le tirage d'un dé équilibré dont les faces indiqueraient les entiers de 2 à 12.

1.1.5. Exemple : variable de Bernoulli. — Soit p un élément de $[0; 1]$. Rappelons qu'une variable aléatoire X suit une *loi de Bernoulli* de paramètre p si elle prend la valeur 1 avec probabilité p et la valeur 0 avec la probabilité $1 - p$. L'entropie d'une telle variable aléatoire est donc égale à

$$\textbf{(1.1.5.1)} \qquad h(p) = \begin{cases} -p\log(p) - (1-p)\log(1-p) & \text{si } 0 < p < 1, \\ 0 & \text{si } p = 0 \text{ ou } p = 1. \end{cases}$$

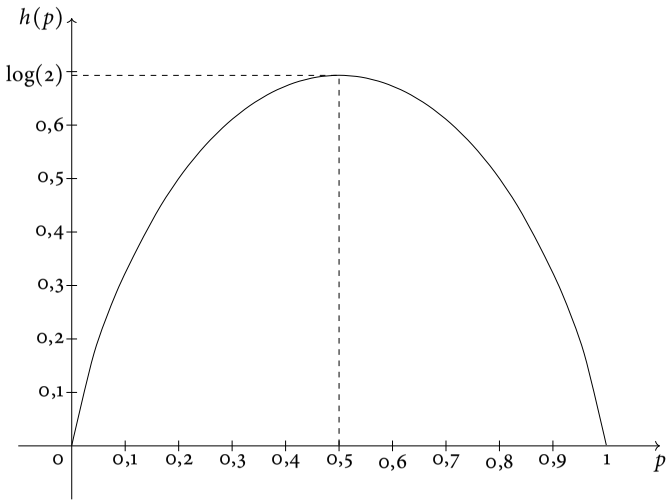


FIGURE 1.1.5.2. Graphe de la fonction « entropie »

On a vu que la fonction $x \mapsto -x\log(x)$ est continue sur $[0;1]$; elle est aussi indéfiniment dérivable sur $]0;1[$, de dérivée $x \mapsto -\log(x) - 1 = -\log(ex)$. Par suite, la fonction h est donc continue, indéfiniment dérivable sur $]0;1[$, de dérivée

$$h'(p) = -\log(ep) + \log(e(1-p)) = \log(1-p) - \log(p).$$

Sa dérivée seconde, donnée par

$$h''(p) = -\frac{1}{1-p} - \frac{1}{p}$$

est strictement négative sur $]0;1[$, si bien que la fonction h est strictement concave. On a $h'(1/2) = 0$, ce qui entraîne que $h'(p) > 0$ pour $p \in]0;1/2[$ et $h'(p) < 0$ pour $p \in]1/2;1[$. La fonction h est donc strictement croissante sur $[0;1/2]$ et strictement décroissante sur $[1/2;1]$. Elle atteint son maximum en le point $p = 1/2$, de valeur $h(1/2) = \log(2)$.

On voit là l'intérêt de la base 2 : la fonction h_2 définie par

$$h_2(p) = h(p)/\log(2) = -p\log_2(p) - (1-p)\log_2(1-p)$$

a pour image $[0;1]$.

Soit A une partie de **R** formée de points de continuité de f ; en particulier, $f(x) = f^*(x)$ pour tout $x \in A$, et on peut écrire, comme précédemment,

$$S_n(f)(x) - f(x) = \int_{-T/2}^{T/2} (f(x-t) - f(x))D_n(t)\,dt.$$

De plus, on peut appliquer l'inégalité des accroissements finis, d'où une majoration $|f(x-t) - f(x)| \leqslant M|t|$, avec $M = \sup|f'|$, de sorte que la fonction h_x introduite plus haut est majorée par M, pour tout $x \in A$. (...)

Lemme **(3.3.6)** (« Lemme de Riemann–Lebesgue »). — *Soit* $h : \mathbf{R} \to \mathbf{C}$ *une fonction localement intégrable telle que* $\int_{-\infty}^{\infty}|h(t)|\,dt < +\infty$. *On a :*

$$\lim_{\omega\to\pm\infty} \int_{-\infty}^{+\infty} h(t)e^{-i\omega t}\,dt = 0.$$

Ce lemme vaut en particulier pour toute fonction continue par morceaux qui est nulle hors d'un intervalle compact de **R**. Nous allons d'ailleurs essentiellement nous contenter de le prouver dans ce cas particulier.

Démonstration. — Pour alléger l'écriture, posons, pour toute fonction h qui est localement intégrable et telle que $\int_{-\infty}^{\infty}|h(t)|\,dt < +\infty$,

$$\widehat{h}(\omega) = \int_{-\infty}^{\infty} h(t)e^{-i\omega t}\,dt.$$

Supposons tout d'abord que h soit de classe $\mathscr{C}^1$ sur un intervalle compact $[a;b]$, et nulle en dehors. Alors, on peut intégrer par parties dans la définition de $\widehat{h}(\omega)$, d'où :

$$\widehat{h}(\omega) = \left[h(t)\frac{i}{\omega}e^{-i\omega t}\right]_a^b - \frac{i}{\omega}\int_a^b h'(t)e^{-i\omega t}\,dt.$$

Cette expression montre que l'on a

$$|\omega\widehat{h}(\omega)| \leqslant |h(a)| + |h(b)| + \int_a^b |h'(t)|\,dt,$$

si bien que $|\widehat{h}(\omega)| = O(1/\omega)$. En particulier, $\lim_{\pm\infty}\widehat{h}(\omega) = 0$.

Si h est seulement intégrable, cet argument ne suffit pas car la formule d'intégration par parties ne sera pas valable. On peut, en revanche, l'approcher « en moyenne » par une fonction en escalier, nulle hors d'un intervalle bornée. Expliquons comment faire lorsque h est continue par morceaux. Soit $\varepsilon > 0$; choisissons des nombres réels a et b tels que $a < b$ et tels que les deux intégrales

$$\int_{-\infty}^{a}|h(t)|\,dt, \qquad \int_b^{+\infty}|h(t)|\,dt$$

de sorte que

$$\begin{aligned}
S_n(f)(x) - f^*(x) &= \int_{-T/2}^{T/2} f(x-t)D_n(t)\,dt \\
&\qquad - \int_{-T/2}^{0} f(x^-)D_n(t)\,dt - \int_0^{T/2} f(x^+)D_n(t)\,dt \\
&= \int_0^{T/2} (f(x-t) - f(x^-))D_n(t)\,dt \\
&\qquad + \int_{-T/2}^{0} (f(x-t) - f(x^+))D_n(t)\,dt \\
&= \int_0^{T/2} \big((f(x-t) - f(x^-)) + (f(x+t) - f(x^+))\big)D_n(t)\,dt.
\end{aligned}$$

Comme f est de classe $\mathscr{C}^1$ par morceaux, $\frac{f(x-t)-f(x^-)}{t}$ et $\frac{f(x+t)-f(x^+)}{t}$ ont une limite quand t tend vers 0. Autrement dit, il existe une fonction continue g_x sur $[0;T/2]$ telle que

$$(f(x-t) - f(x^-)) + (f(x+t) - f(x^+)) = tg_x(t)$$

pour tout $t \in [0;T/2]$. On écrit alors

$$S_n(f)(x) - f^*(x) = \int_0^{T/2} g_x(t)\frac{t}{T\sin(\pi t/T)}\sin(\pi(2n+1)t/T)\,dt.$$

La fonction définie par $t \mapsto t/\sin(\pi t/T)$ pour $t \in]0;T/2]$ est continue, et a une limite (T/π) en 0 ; elle se prolonge par continuité en une fonction continue sur $[0;T/2]$. Alors, la fonction h_x sur $[0;T/2]$ définie par

$$h_x(t) = g_x(t)\frac{t}{T\sin(\pi t/T)}$$

pour $t \in]0;T/2]$ se prolonge par continuité en 0, et l'on a

(3.3.5.1) $$S_n(f)(x) - f^*(x) = \int_0^{T/2} h_x(t)\sin(\pi(2n+1)t/T)\,dt.$$

D'après le lemme ci-dessous (lemme 3.3.6), le membre de droite de cette égalité tend vers 0, on a donc

$$\lim_{x\to\infty} S_n(f)(x) = f^*(x).$$

Cela conclut la démonstration de la première partie du théorème de Dirichlet.

1.2. Entropie conditionnelle

1.2.1. — Soit A un événement de probabilité non nulle, c'est-à-dire une partie de l'univers probabiliste Ω telle que $\mathbf{P}(A) > 0$. Alors, A lui-même peut être vu comme un univers probabiliste, lorsqu'on pose, pour toute partie mesurable B de A,

$$\mathbf{P}(B \mid A) = \frac{\mathbf{P}(B)}{\mathbf{P}(A)}.$$

Si X est une variable aléatoire discrète, on peut alors la *conditionner* à A en considérant sa restriction à A, ici notée $X \mid A$.

Définition **(1.2.2)**. — *Soit* X *et* Y *des variables aléatoires discrètes. On appelle* entropie conditionnelle *de* X *sachant* Y *l'expression*

(1.2.2.1) $$H(X \mid Y) = \sum_y \mathbf{P}(Y = y)H(X \mid Y = y).$$

C'est un élément de $[0;+\infty]$. A priori, la variable aléatoire $X \mid Y = y$ n'est définie que si $\mathbf{P}(Y = y) \neq 0$; dans le cas contraire, on enlève le terme correspondant de la somme. Cette somme est infinie s'il existe y tel que $\mathbf{P}(Y = y) > 0$ et $H(X \mid Y = y) = +\infty$; s'il n'existe pas de tel y, il est aussi possible que la série diverge. Quoi qu'il en soit, si X et Y ne prennent qu'un nombre fini de valeurs, cette expression est finie.

Si la variable aléatoire Y est certaine, c'est-à-dire s'il existe y tel que $\mathbf{P}(Y = y) = 1$, alors $H(X \mid Y) = H(X \mid Y = y) = H(X)$.

Proposition **(1.2.3)**. — *Soit* X *et* Y *des variables aléatoires discrètes. On a*

(1.2.3.1) $$H(X,Y) = H(Y) + H(X \mid Y).$$

Démonstration. — Partons de la définition de l'entropie conditionnelle :

$$H(X \mid Y) = \sum_y \mathbf{P}(Y = y)H(X \mid Y = y).$$

Pour tout y tel que $\mathbf{P}(\mathrm{Y}=y)>0$, on a alors

$$\begin{aligned}\mathrm{H}(\mathrm{X}\mid\mathrm{Y}=y) &= -\sum_x \mathbf{P}(\mathrm{X}=x\mid\mathrm{Y}=y)\log\mathbf{P}(\mathrm{X}=x\mid\mathrm{Y}=y)\\ &= -\sum_x \frac{\mathbf{P}(\mathrm{X}=x,\mathrm{Y}=y)}{\mathbf{P}(\mathrm{Y}=y)}\log\frac{\mathbf{P}(\mathrm{X}=x,\mathrm{Y}=y)}{\mathbf{P}(\mathrm{Y}=y)}\\ &= -\sum_x \frac{\mathbf{P}(\mathrm{X}=x,\mathrm{Y}=y)}{\mathbf{P}(\mathrm{Y}=y)}\log(\mathbf{P}(\mathrm{X}=x,\mathrm{Y}=y))\\ &\quad+\sum_x \frac{\mathbf{P}(\mathrm{X}=x,\mathrm{Y}=y)}{\mathbf{P}(\mathrm{Y}=y)}\log(\mathbf{P}(\mathrm{Y}=y))\\ &= -\sum_x \frac{\mathbf{P}(\mathrm{X}=x,\mathrm{Y}=y)}{\mathbf{P}(\mathrm{Y}=y)}\log(\mathbf{P}(\mathrm{X}=x,\mathrm{Y}=y))+\log(\mathbf{P}(\mathrm{Y}=y))\end{aligned}$$

puisque

$$\sum_x \mathbf{P}(\mathrm{X}=x,\mathrm{Y}=y)=\mathbf{P}(\mathrm{Y}=y).$$

Alors,

$$\begin{aligned}\mathrm{H}(\mathrm{X}\mid\mathrm{Y}) &= -\sum_y\sum_x \mathbf{P}(\mathrm{X}=x,\mathrm{Y}=y)\log(\mathbf{P}(\mathrm{X}=x,\mathrm{Y}=y))\\ &\quad+\sum_y \mathbf{P}(\mathrm{Y}=y)\log(\mathbf{P}(\mathrm{Y}=y))\\ &= \mathrm{H}(\mathrm{X},\mathrm{Y})-\mathrm{H}(\mathrm{Y}).\end{aligned}$$

Ces calculs ont un sens lorsque l'entropie $\mathrm{H}(\mathrm{Y})$ est finie, avec la convention $+\infty-h=+\infty$ pour tout nombre réel h. Lorsque $\mathrm{H}(\mathrm{Y})=+\infty$, on rappelle qu'une entropie est positive ou nulle, de sorte que $\mathrm{H}(\mathrm{X}\mid\mathrm{Y}=y)\geqslant 0$ et

$$-\sum_x \mathbf{P}(\mathrm{X}=x,\mathrm{Y}=y)\log(\mathbf{P}(\mathrm{X}=x,\mathrm{Y}=y))\geqslant -\mathbf{P}(\mathrm{Y}=y)\log(\mathbf{P}(\mathrm{Y}=y)).$$

Lorsqu'on somme sur les valeurs possibles de Y, on obtient que l'entropie de $\mathrm{H}(\mathrm{X},\mathrm{Y})$ est infinie. □

Corollaire **(1.2.4)**. — *Soit* X, Y, Z *des variables aléatoires discrètes. On a*

$$\mathrm{H}(\mathrm{X},\mathrm{Y}\mid\mathrm{Z})=\mathrm{H}(\mathrm{Y}\mid\mathrm{Z})+\mathrm{H}(\mathrm{X}\mid\mathrm{Y},\mathrm{Z}).$$

Démonstration. — Soit z une valeur de Z telle que $\mathbf{P}(\mathrm{Z}=z)>0$. Appliquons la proposition aux variables aléatoires $\mathrm{X}\mid\mathrm{Z}=z$ et $\mathrm{Y}\mid\mathrm{Z}=z$: on obtient

$$\mathrm{H}(\mathrm{X},\mathrm{Y}\mid\mathrm{Z}=z)=\mathrm{H}(\mathrm{Y}\mid\mathrm{Z}=z)+\mathrm{H}(\mathrm{X}\mid\mathrm{Y},\mathrm{Z}=z).$$

La somme qui apparaît est la somme des $(2n+1)$ premiers termes d'une suite géométrique de raison $e^{2\pi iu/\mathrm{T}}$; si $u\in\mathbf{R}-\mathrm{T}\mathbf{Z}$, $e^{2\pi iu/\mathrm{T}}\neq 1$, de sorte que

$$\begin{aligned}\mathrm{D}_n(u) &= \frac{1}{\mathrm{T}}e^{-2\pi inu/\mathrm{T}}\frac{1-e^{2\pi i(2n+1)u/\mathrm{T}}}{1-e^{2\pi iu/\mathrm{T}}}\\ &= \frac{1}{\mathrm{T}}e^{-2\pi inu/\mathrm{T}}\frac{e^{\pi i(2n+1)u/\mathrm{T}}(-2i\sin(\pi(2n+1)u/\mathrm{T}))}{e^{\pi iu/\mathrm{T}}(-2i\sin(piu/\mathrm{T})}\\ &= \frac{1}{\mathrm{T}}\exp((-2n+(2n+1)-1)\pi iu/\mathrm{T})\frac{\sin(\pi(2n+1)u/\mathrm{T})}{\sin(\pi u/\mathrm{T})}\\ &= \frac{1}{\mathrm{T}}\frac{\sin(\pi(2n+1)u/\mathrm{T})}{\sin(\pi u/\mathrm{T})}.\end{aligned}$$

Le lemme est ainsi démontré. □

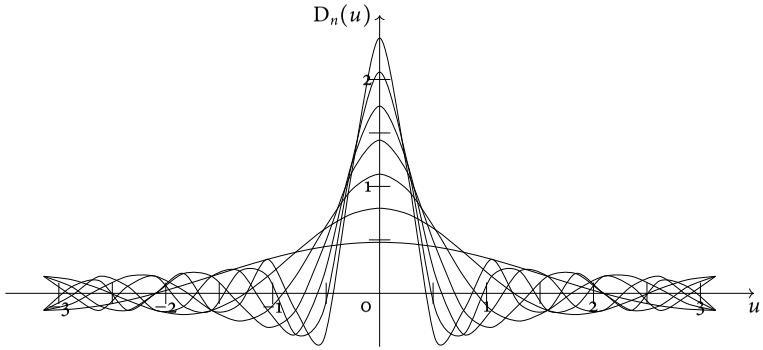


FIGURE 3.3.4.1. Graphe du noyau de Dirichlet ($\mathrm{T}=2\pi$), pour $1\leqslant n\leqslant 7$

3.3.5. — La représentation graphique du noyau de Dirichlet sur $[-\mathrm{T}/2;\mathrm{T}/2]$ montre que lorsque n grandit, il se concentre autour de 0 (modulo T). Jointe à la première relation du lemme précédent, cette observation aurait pu être la clé de la démonstration du théorème de Dirichlet si les changements de signe de D_n n'avaient pas eu comme conséquence que $\frac{1}{\mathrm{T}}\int_0^{\mathrm{T}}|\mathrm{D}_n(u)|\,du$ tende vers $+\infty$, bien que $\frac{1}{\mathrm{T}}\int_0^{\mathrm{T}}\mathrm{D}_n(u)\,du$ soit constante, égale à 1.

Reprenons donc plutôt la relation (3.3.3.3). Comme D_n est paire, on a

$$\int_0^{\mathrm{T}/2}\mathrm{D}_n(u)\,du=\int_{-\mathrm{T}/2}^{0}\mathrm{D}_n(u)\,du=\frac{1}{2}\int_{-\mathrm{T}/2}^{\mathrm{T}/2}\mathrm{D}_n(u)\,du=\frac{1}{2},$$

3.3.3. — La preuve commence par récrire $S_n(f)(x)$ sous la forme d'une intégrale. En effet, on a

$$\begin{aligned} S_n(f)(x) &= \sum_{p=-n}^{n} c_p(f) e^{2\pi i p x/T} \\ &= \sum_{p=-n}^{n} \frac{1}{T} \int_0^T f(t) e^{-2\pi i p t/T}\, dt e^{2\pi i p x/T} \\ &= \frac{1}{T} \int_0^T f(t) \sum_{p=-n}^{n} e^{2\pi i p(x-t)/T}\, dt. \end{aligned}$$

Posons ainsi, pour tout $u \in \mathbf{R}$,

$$\textbf{(3.3.3.1)} \qquad D_n(u) = \frac{1}{T} \sum_{p=-n}^{n} e^{2\pi i p u/T} ;$$

la fonction D_n est un polynôme trigonométrique qu'on appelle *noyau de Dirichlet*. Par définition, on a

$$\textbf{(3.3.3.2)} \qquad S_n(f)(x) = \int_0^T f(t) D_n(x-t)\, dt.$$

Le changement de variables $t' = x - t$ et la périodicité de f et de D_n permettent de récrire

$$\textbf{(3.3.3.3)} \qquad S_n(f)(x) = \int_{x-T}^{x} f(x-u) D_n(u)\, du = \int_{-T/2}^{T/2} f(x-u) D_n(u)\, du.$$

Lemme **(3.3.4)**. — a) *On a* $\int_0^T D_n(u)\, du = 1$.

b) *Pour tout entier* $n \geqslant 1$ *et tout* $u \in \mathbf{R} - T\mathbf{Z}$, *on a*

$$D_n(u) = \frac{1}{T} \frac{\sin(\pi(2n+1)u/T)}{\sin(\pi u/T)}.$$

Démonstration. — La première relation est immédiate. Démontrons la seconde. On part de la définition de $D_n(u)$:

$$D_n(u) = \frac{1}{T} \sum_{p=-n}^{n} e^{2\pi i p u/T} = \frac{1}{T} e^{-2\pi i n u/T} \sum_{p=0}^{2n} e^{2\pi i p u/T}.$$

Multiplions cette égalité par $\mathbf{P}(Z = z)$ et ajoutons-les ; par définition des entropies $H(X, Y \mid Z)$ et $H(Y \mid Z)$, il vient :

$$H(X, Y \mid Z) = H(Y \mid Z) + \sum_z \mathbf{P}(Z = z) H(X \mid Y, Z = z).$$

Pour calculer ce dernier terme, revenons à la définition de l'entropie conditionnelle $H(X \mid Y, Z = z)$; on a

$$H(X \mid Y, Z = z) = \sum_y \mathbf{P}(Y = y \mid Z = z) H(X \mid Y = y, Z = z),$$

de sorte que

$$\begin{aligned} &\sum_{\mathbf{P}(Z=z)>0} \mathbf{P}(Z = z) H(X \mid Y, Z = z) \\ &\qquad = \sum_y \sum_{\mathbf{P}(Z=z)>0} \mathbf{P}(Z = z) \mathbf{P}(Y = y \mid Z = z) H(X \mid Y = y, Z = z) \\ &\qquad = \sum_y \sum_{\mathbf{P}(Z=z)>0} \mathbf{P}(Y = y, Z = z) H(X \mid Y = y, Z = z) \\ &\qquad = \sum_{\mathbf{P}(Y=y, Z=z)>0} \mathbf{P}(Y = y, Z = z) H(X \mid Y = y, Z = z) \\ &\qquad = H(X \mid Y, Z). \end{aligned}$$

Le corollaire est ainsi démontré. □

Remarque **(1.2.5)**. — On peut aussi démontrer simplement cette égalité lorsque l'entropie $H(Y, Z)$ est finie. Dans ce cas, $H(Z)$ est également finie et l'on a

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

Corollaire **(1.2.6)**. — *Soit* $X_1, \ldots, X_n$ *des variables aléatoires discrètes. On a*

$$H(X_1, \ldots, X_n) = \sum_{k=1}^{n} H(X_k \mid X_1, \ldots, X_{k-1}).$$

Démonstration. — En effet, on a

$$\begin{aligned}
\mathrm{H}(\mathrm{X}_1,\dots,\mathrm{X}_n) &= \mathrm{H}(\mathrm{X}_1) + \mathrm{H}(\mathrm{X}_2,\dots,\mathrm{X}_n \mid \mathrm{X}_1)\\
&= \mathrm{H}(\mathrm{X}_1) + \mathrm{H}(\mathrm{X}_2 \mid \mathrm{X}_1) + \mathrm{H}(\mathrm{X}_3,\dots,\mathrm{X}_n \mid \mathrm{X}_1,\mathrm{X}_2)\\
&= \dots\\
&= \mathrm{H}(\mathrm{X}_1) + \mathrm{H}(\mathrm{X}_2 \mid \mathrm{X}_1) + \mathrm{H}(\mathrm{X}_3 \mid \mathrm{X}_1,\mathrm{X}_2) + \dots\\
&\qquad + \mathrm{H}(\mathrm{X}_n \mid \mathrm{X}_1,\mathrm{X}_2,\dots,\mathrm{X}_{n-1}),
\end{aligned}$$

ce qu'il fallait démontrer □

1.3. Information mutuelle

Définition **(1.3.1)**. — *Soit p, q des lois discrètes sur un ensemble* A. *On appelle* divergence *de p par rapport à q l'expression*

$$\mathrm{D}(p \mid q) = \sum_{\substack{a\in\mathrm{A}\\ p(a)>0}} p(a)\log\left(\frac{p(a)}{q(a)}\right).$$

Rien ne garantit, a priori, que cette famille soit sommable ; d'ailleurs, s'il existe un élément a tel que $p(a) > 0$ et $q(a) = 0$, on a $\mathrm{D}(p \mid q) = +\infty$. On va en fait vérifier que la famille $(p(a)\inf\log(p(a)/q(a), 0))$ est sommable, ce qui entraîne que $\mathrm{D}(p \mid q)$ est un élément bien défini de $[0;+\infty]$.

Théorème **(1.3.2)**. — *Soit p, q des lois discrètes sur un ensemble* A. *La famille* $(p(a)\inf\log(p(a)/q(a), 0))$ *est sommable ; on a* $\mathrm{D}(p \mid q) \geqslant 0$, *avec égalité si et seulement si* $p = q$.

Démonstration. — La fonction logarithme est strictement concave ; son graphe donc en-dessous de sa tangente en tout point, et ne coupe cette tangente qu'un le point. En particulier, pour tout $x \in \mathbf{R}_{>0}$, on a $\log(x) \leqslant x - 1$ (inégalité que l'on peut aussi vérifier par analyse de fonction, ou bien par la formule de Taylor), et l'inégalité est stricte si $x \neq 1$. On écrit plutôt

$$\log\frac{1}{x} = -\log x \geqslant 1 - x,$$

tout entier n, on a

$$\begin{aligned}
c_n(f') &= \frac{1}{\mathrm{T}}\int_0^{\mathrm{T}} f'(t)e^{-2\pi int/\mathrm{T}}\,dt\\
&= \frac{1}{\mathrm{T}}\left[f(t)e^{-2\pi int/\mathrm{T}}\right]_0^{\mathrm{T}} - \frac{1}{\mathrm{T}}\int_0^{\mathrm{T}} f(t)(-2\pi in/\mathrm{T})e^{-2\pi int/\mathrm{T}}\,dt\\
&= \frac{2\pi in}{\mathrm{T}}\,\frac{1}{\mathrm{T}}\int_0^{\mathrm{T}} f(t)e^{-2\pi int/\mathrm{T}}\,dt\\
&= \frac{2\pi in}{\mathrm{T}}c_n(f). \qquad \textbf{(3.2.7.1)}
\end{aligned}$$

Pour les coefficients de Fourier a_n et b_n, on obtient alors

$$a_n(f') = -\frac{2\pi n}{\mathrm{T}}b_n(f) \qquad \text{et} \qquad b_n(f') = \frac{2\pi n}{\mathrm{T}}a_n(f). \qquad \textbf{(3.2.7.2)}$$

Ces formules valent, en fait, sous l'hypothèse un peu plus générale que f est de classe $\mathscr{C}^1$ par morceaux et continue, et se démontrent par le même raisonnement, grâce à la formule d'intégration par parties pour les fonctions de classe $\mathscr{C}^1$ par morceaux et continues.

3.3. Théorème de Dirichlet

3.3.1. — Soit f une fonction continue par morceaux sur $\mathbf{R}$. On note f^* la fonction définie sur $\mathbf{R}$ par la formule

$$f^*(t) = \frac{1}{2}\big(f(t^-) + f(t^+)\big),$$

où $f(t^+)$ et $f(t^-)$ désignent les limites à droite et à gauche de f en t. On dit que c'est la *régularisée*de f au sens de Dirichlet.

Si I est un intervalle ouvert tel que $f|_{\mathrm{I}}$ est continue, alors $f^* = f$ sur I. Par suite, la trace de l'ensemble des points où f et f^* diffèrent rencontre tout intervalle compact en un ensemble fini.

Théorème **(3.3.2)**. — *Soit* $f : \mathbf{R} \to \mathbf{C}$ *une fonction de classe* $\mathscr{C}^1$ *par morceaux, de période* T. *Pour tout* $x \in \mathbf{R}$, *on a*

$$\mathrm{S}_n(f)(x) \to f^*(x).$$

De plus, si A *est un intervalle compact formé de points de continuité de* f, *alors la convergence est uniforme sur* A.

Exemple **(3.2.6)**. — On dit que f est un *polynôme trigonométrique* si c'est une combinaison linéaire (finie) de fonctions de la forme $t \mapsto e^{2i\pi pt/\mathrm{T}}$, autrement dit s'il existe un entier $\mathrm{N} \geqslant 0$ et une famille $(c_p)_{-\mathrm{N}\leqslant p\leqslant \mathrm{N}}$ de nombres complexes telle que

$$f(t) = \sum_{p=-\mathrm{N}}^{\mathrm{N}} c_p e^{2i\pi pt/\mathrm{T}}.$$

En décomposant l'exponentielle en cosinus et sinus, cela revient aussi à l'existence de deux familles $(a_p)_{0\leqslant p\leqslant \mathrm{N}}$ et $(b_p)_{1\leqslant p\leqslant \mathrm{N}}$ de nombres complexes tels que

$$f(t) = \frac{1}{2}a_0 + \sum_{p=1}^{n} a_p \cos(2\pi pt/\mathrm{T}) + b_p \sin(2\pi pt/\mathrm{T}),$$

les a_p, b_p et c_p étant reliés par les formules $a_p = c_p + c_{-p}$ et $b_p = i(c_p - c_{-p})$.

Ses coefficients de Fourier vérifient précisément $c_n(f) = c_n$, si $|n| \leqslant \mathrm{N}$, $a_n(f) = a_n$ si $0 \leqslant n \leqslant \mathrm{N}$, $b_n(f) = b_n$ si $1 \leqslant n \leqslant \mathrm{N}$, et tous ses autres coefficients sont nuls. Pour le démontrer, il suffit, par linéarité, de traiter le cas où $f(t) = e^{2i\pi pt/\mathrm{T}}$. Dans ce cas, on a

$$c_n(f) = \frac{1}{\mathrm{T}}\int_0^{\mathrm{T}} e^{2i\pi pt/\mathrm{T}} e^{-2i\pi nt/\mathrm{T}}\, dt = \frac{1}{\mathrm{T}}\int_0^{\mathrm{T}} e^{2i\pi(p-n)t/\mathrm{T}}\, dt.$$

Lorsque $n = p$, on obtient

$$c_p(f) = \frac{1}{\mathrm{T}}\int_0^{\mathrm{T}} 1\, dt = 1,$$

tandis que si $n \neq p$, on a

$$c_n(f) = \frac{1}{\mathrm{T}}\left[\frac{\mathrm{T}}{2i\pi(p-n)} e^{2i\pi(p-n)t/\mathrm{T}}\right]_0^{\mathrm{T}} = 0.$$

Les formules pour $a_n(f)$ et $b_n(f)$ s'en déduisent.

On observe alors que pour tout entier n tel que $n \geqslant \mathrm{N}$, on a $\mathrm{S}_n(f)(t) = f(t)$.

Exemple **(3.2.7)**. — Supposons que f soit une fonction de classe $\mathscr{C}^1$ sur $\mathbf{R}$, de période T. Dans ce cas, sa dérivée f' est une fonction continue, également de période T, donc dispose de coefficients de Fourier. Montrons comment on peut, par intégration par parties, les calculer en fonction de ceux de f. En effet, pour

avec égalité si et seulement si $x = 1$. Appliquons cette inégalité à $x = q(a)/p(a)$, pour $a \in \mathrm{A}$ tel que $p(a) > 0$. Il vient

$$\log\frac{p(a)}{q(a)} \geqslant 1 - \frac{q(a)}{p(a)} = \frac{p(a) - q(a)}{p(a)},$$

d'où

$$p(a)\log\frac{p(a)}{q(a)} \geqslant p(a) - q(a),$$

avec égalité si et seulement si $q(a) = p(a) > 0$. En sommant sur l'ensemble des valeurs de a telles que $p(a) > 0$, on obtient

$$\mathrm{D}(p \mid q) \geqslant 1 - \sum_{\substack{a\in\mathrm{A}\\ p(a)>0}} q(a) \geqslant 0.$$

Il y a égalité si et seulement si $q(a) = p(a)$ pour tout a tel que $p(a) > 0$; comme $\sum_{a\in\mathrm{A}} q(a) = 1$, cela signifie $p = q$. □

Remarque **(1.3.3)**. — Dans la littérature, la quantité $\mathrm{D}(p \mid q)$ s'appelle *divergence de Kullback-Leibler*, ou aussi distance de Kullback-Leibler. De fait, elle mesure la différence entre les deux lois de probabilité p et q : elle est positive, et ne s'annule que lorsque $p = q$. Mais ce n'est pas tout à fait une distance, car elle n'est pas symétrique et ne vérifie pas l'inégalité triangulaire. D'ailleurs, d'autres expressions ont les mêmes propriétés et peuvent rendre des services similaires ; comme nous ne les utiliserons pas dans ce cours, nous avons préféré l'expression *divergence*.

1.3.4. — Soit X et Y des variables aléatoires discrètes. Sur l'ensemble des valeurs possibles du couple (X, Y), on dispose alors de deux lois discrètes :

a) La loi du couple (X, Y), c'est-à-dire $(x, y) \mapsto \mathbf{P}(\mathrm{X} = x, \mathrm{Y} = y)$;

b) Le produit des deux lois marginales de ce couple, c'est-à-dire $(x, y) \mapsto \mathbf{P}(\mathrm{X} = x)\mathbf{P}(\mathrm{Y} = y)$.

Définition **(1.3.5)**. — *Soit* X *et* Y *des variables aléatoires discrètes. On appelle* information mutuelle *de* X *et* Y *la divergence de la loi du couple* (X, Y) *par rapport à la loi* $(x, y) \mapsto \mathbf{P}(\mathrm{X} = x)\mathbf{P}(\mathrm{Y} = y)$, *produit des deux lois marginales du couple* (X, Y).

Corollaire **(1.3.6)**. — *Soit* X *et* Y *des variables aléatoires discrètes. L'information mutuelle* $\mathrm{I}(\mathrm{X},\mathrm{Y})$ *est un élément de* $[0;+\infty]$ *; il est nul si et seulement si* X *et* Y *sont indépendantes.*

Démonstration. — L'inégalité $\mathrm{I}(\mathrm{X},\mathrm{Y}) \geqslant 0$ est un cas particulier du théorème. De plus, il y a égalité $\mathrm{I}(\mathrm{X},\mathrm{Y}) = 0$ si et seulement si

$$\mathbf{P}(\mathrm{X}=x,\mathrm{Y}=y) = \mathbf{P}(\mathrm{X}=x)\mathbf{P}(\mathrm{Y}=y)$$

pour tout couple (x,y), ce qui signifie exactement que X et Y sont indépendantes. □

Corollaire **(1.3.7)**. — *Soit* X *et* Y *des variables aléatoires discrètes. On a les égalités*

(1.3.7.1) $$\mathrm{H}(\mathrm{X}) = \mathrm{I}(\mathrm{X},\mathrm{Y}) + \mathrm{H}(\mathrm{X}\mid\mathrm{Y})$$

(1.3.7.2) $$\mathrm{H}(\mathrm{X},\mathrm{Y}) + \mathrm{I}(\mathrm{X},\mathrm{Y}) = \mathrm{H}(\mathrm{X}) + \mathrm{H}(\mathrm{Y}).$$

En particulier, on a l'inégalité

(1.3.7.3) $$\mathrm{H}(\mathrm{X}\mid\mathrm{Y}) \leqslant \mathrm{H}(\mathrm{X}),$$

avec égalité si et seulement si X *et* Y *sont indépendantes.*

Si l'entropie d'une variable aléatoire est une mesure d'incertitude, la conditionner à une seconde variable aléatoire diminue cette incertitude.

Démonstration. — Il n'y a rien à démontrer si $\mathrm{H}(\mathrm{X})$ est infinie ; supposons donc $\mathrm{H}(\mathrm{X})$ finie. De la définition des entropies $\mathrm{H}(\mathrm{X})$ et $\mathrm{H}(\mathrm{X}\mid\mathrm{Y})$, on tire

$$\begin{aligned}\mathrm{H}(\mathrm{X}) - \mathrm{H}(\mathrm{X}\mid\mathrm{Y}) &= \mathrm{H}(\mathrm{X}) + \mathrm{H}(\mathrm{Y}) - \mathrm{H}(\mathrm{X},\mathrm{Y})\\ &= \sum_{x,y}\sum_{x}\mathbf{P}(\mathrm{X}=x,\mathrm{Y}=y)\log\frac{\mathbf{P}(\mathrm{X}=x)\mathbf{P}(\mathrm{Y}=y)}{\mathbf{P}(\mathrm{X}=x,\mathrm{Y}=y)}\\ &= \mathrm{I}(\mathrm{X},\mathrm{Y}).\end{aligned}$$

La première relation découle aussitôt, de même que l'inégalité finale et son cas d'égalité. Quant à la seconde, elle resulte alors des égalités $\mathrm{H}(\mathrm{X},\mathrm{Y}) + \mathrm{I}(\mathrm{X},\mathrm{Y}) = \mathrm{H}(\mathrm{Y}) + \mathrm{H}(\mathrm{X}\mid\mathrm{Y}) + \mathrm{I}(\mathrm{X},\mathrm{Y}) = \mathrm{H}(\mathrm{Y}) + \mathrm{H}(\mathrm{X})$. □

Définition **(1.3.8)**. — *Soit* X, Y, Z *des variables aléatoires discrètes. On dit que* X *et* Z *sont indépendantes conditionnellement à* Y*, et l'on note* $\mathrm{X} \perp_{\mathrm{Y}} \mathrm{Z}$ *si l'on a*

$$\mathbf{P}(\mathrm{X}=x,\mathrm{Z}=z\mid\mathrm{Y}=y) = \mathbf{P}(\mathrm{X}=x\mid\mathrm{Y}=y)\mathbf{P}(\mathrm{Z}=z\mid\mathrm{Y}=y)$$

pour tous x, y, z.

Fourier, on trouve

$$a_n(\check{f}) = a_n(f), \qquad b_n(\check{f}) = -b_n(f), \qquad c_n(\check{f}) = c_{-n}(f).$$

En particulier, si f est paire ($\check{f} = f$), ses coefficients b_n sont nuls, tandis que si f est impaire ($\check{f} = -f$), ses coefficients a_n sont nuls.

3.2.5. — Soit encore une fonction f, définie sur $\mathbf{R}$, localement intégrable et de période T. Sa *série de Fourier* est la série (illimitée dans les deux directions)

(3.2.5.1) $$\sum_{n\in\mathbf{Z}} c_n(f)e^{2\pi int/\mathrm{T}}.$$

Sa série de Fourier « réelle » est la série

(3.2.5.2) $$\frac{1}{2}a_0(f) + \sum_{n=1}^{\infty} a_n(f)\cos(2\pi int/\mathrm{T}) + b_n(f)\sin(2\pi int/\mathrm{T}).$$

Notons qu'elles dépendent de t : ce ce sont des séries de fonctions. À ce stade là du cours, on n'affirme pas encore que ces séries convergent, ce n'est d'ailleurs pas toujours le cas, et encore moins qu'elles convergent vers $f(t)$.

Pour tout entier $n \geqslant 0$, on notera

(3.2.5.3) $$\mathrm{S}_n(f)(t) = \sum_{p=-n}^{n} c_p(f)e^{2\pi ipt/\mathrm{T}}.$$

Si l'on exprime les coefficients c_p en fonction de a_p et b_p, on obtient l'égalité

(3.2.5.4) $$\mathrm{S}_n(f) = \frac{1}{2}a_0(f) + \sum_{p=1}^{n}\big(a_p(f)\cos(2\pi ipt/\mathrm{T}) + b_p(f)\sin(2\pi ipt/\mathrm{T})\big),$$

qui justifie la définition (3.2.5.2) de la série de Fourier réelle de la fonction f. En effet, on a :

$$\begin{aligned}\mathrm{S}_n(f)(t) &= c_0(f) + \sum_{p=1}^{n}\big(c_p(f)(\cos(2\pi ipt/\mathrm{T}) + i\sin(2\pi ipt/\mathrm{T}))\\ &\qquad + c_{-p}(f)(\cos(2\pi ipt/\mathrm{T}) - i\sin(2\pi ipt/\mathrm{T}))\big)\\ &= \frac{1}{2}a_0(f) + \sum_{p=1}^{n}\big((c_p(f)+c_{-p}(f))\cos(2\pi ipt/\mathrm{T})\\ &\qquad + i(c_p(f)-c_{-p}(f))\sin(2\pi ipt/\mathrm{T})\big)\\ &= \frac{1}{2}a_0(f) + \sum_{p=1}^{n}\big(a_p(f)\cos(2\pi ipt/\mathrm{T}) + b_p(f)\sin(2\pi it/\mathrm{T})\big).\end{aligned}$$

Définition **(3.2.4).** — *Soit* $f : \mathbf{R} \to \mathbf{C}$ *une fonction localement intégrable, de période* $\mathrm{T} > 0$. *Ses* coefficients de Fourier *sont les nombres complexes :*

$$a_n(f) = \frac{2}{\mathrm{T}} \int_0^{\mathrm{T}} f(t) \cos(2\pi n t/\mathrm{T})\, dt, \tag{3.2.4.1}$$

$$b_n(f) = \frac{2}{\mathrm{T}} \int_0^{\mathrm{T}} f(t) \sin(2i\pi n t/\mathrm{T})\, dt, \tag{3.2.4.2}$$

$$c_n(f) = \frac{1}{\mathrm{T}} \int_0^{\mathrm{T}} f(t) e^{-2i\pi n t/\mathrm{T}}\, dt, \tag{3.2.4.3}$$

pour $n \in \mathbf{Z}$.

Comme les fonctions intégrées sur de période T, on peut en fait intégrer sur n'importe quel intervalle de longueur T ; en particulier, il est parfois utile de considérer l'intervalle $[-\mathrm{T}/2; \mathrm{T}/2]$.

Ces coefficients ne sont pas indépendants. La parité de la fonction cosinus et l'imparité de la fonction sinus entraînent que l'on a $a_{-n}(f) = a_n(f)$ et $b_{-n}(f) = b_n(f)$ pour tout entier n ; en particulier, $b_0(f) = 0$. De même, en décomposant l'exponentielle complexe

$$e^{-2i\pi n t/\mathrm{T}} = \cos(2\pi n t/\mathrm{T}) - i \sin(2\pi n t/\mathrm{T}),$$

on obtient les formules

$$c_n(f) = \frac{1}{2}(a_n(f) - i b_n(f)) \qquad \text{et} \qquad c_{-n}(f) = \frac{1}{2}(a_n(f) + i b_n(f)),$$

que l'on peut inverser en

$$a_n(f) = c_n(f) + c_{-n}(f) \qquad \text{et} \qquad b_n(f) = i(c_n(f) - c_{-n}(f)).$$

Ainsi, dans la pratique, il suffira de travailler avec les coefficients c_n ; Ils sont linéaires en f :

$$c_n(\lambda f) = \lambda c_n(f), \qquad c_n(f + g) = c_n(f) + c_n(g),$$

et de même pour les coefficients a_n et b_n.

Les fonctions cosinus et sinus sont à valeurs réelles, tandis que la conjugaison complexe change $e^{-2i\pi n t/\mathrm{T}}$ en $e^{2i\pi n t/\mathrm{T}}$. Par suite, on a aussi

$$a_n(\overline{f}) = \overline{a_n(f)}, \qquad b_n(\overline{f}) = \overline{b_n(f)}, \qquad c_n(\overline{f}) = \overline{c_{-n}(f)}.$$

Notons $\check{f}$ la fonction $t \mapsto f(-t)$; elle est également de période T. En faisant le changement de variables $t' = -t$ dans l'intégrale qui définit ses coefficients de

Exemple **(1.3.9).** — S'il existe une fonction f telle que $\mathrm{Z} = f(\mathrm{Y})$, alors $\mathrm{X} \perp_{\mathrm{Y}} \mathrm{Z}$. En particulier, X et Y sont indépendantes conditionnellement à Y.

Soit en effet x, y, z tels que $\mathbf{P}(\mathrm{Y} = y) > 0$. Si $z \neq f(y)$, on a $\mathbf{P}(\mathrm{X} = x, \mathrm{Z} = z \mid \mathrm{Y} = y) = 0$ et $\mathbf{P}(\mathrm{Z} = z \mid \mathrm{Y} = y) = 0$. En revanche, si $z = f(y)$, on a $\mathbf{P}(\mathrm{X} = x, \mathrm{Z} = z \mid \mathrm{Y} = y) = \mathbf{P}(\mathrm{X} = x \mid \mathrm{Y} = y)$ et $\mathbf{P}(\mathrm{Z} = z \mid \mathrm{Y} = y) = 1$. $\mathbf{P}(\mathrm{X} = x, \mathrm{Z} = z \mid \mathrm{Y} = y) = 0$.

1.3.10. — Pour étudier cette notion d'indépendance conditionnelle de X, Z relativement à la variable aléatoire Y, il est utile d'introduire la notion d'information mutuelle de X, Z relativement à Y, définie par

$$\mathrm{I}(\mathrm{X}, \mathrm{Z} \mid \mathrm{Y}) = \sum_y \mathbf{P}(\mathrm{Y} = y) \mathrm{I}(\mathrm{X} \mid \mathrm{Y} = y, \mathrm{Z} \mid \mathrm{Y} = y).$$

C'est un élément de $[0; +\infty]$, nul si et seulement si $\mathrm{X} \perp_{\mathrm{Y}} \mathrm{Z}$.

Par ailleurs,

$$\begin{aligned}
\mathrm{I}(\mathrm{X}, (\mathrm{Y}, \mathrm{Z})) &= \sum_{x,y,z} \mathbf{P}(\mathrm{X} = x, \mathrm{Y} = y, \mathrm{Z} = z) \log \frac{\mathbf{P}(\mathrm{X} = x, \mathrm{Y}, y, \mathrm{Z} = z)}{\mathbf{P}(\mathrm{X} = x)\mathbf{P}(\mathrm{Y} = y, \mathrm{Z} = z)} \\
&= \sum_{x,y,z} \mathbf{P}(\mathrm{X} = x, \mathrm{Y} = y, \mathrm{Z} = z) \log \frac{\mathbf{P}(\mathrm{X} = x, \mathrm{Y} = y, \mathrm{Z} = z)\mathbf{P}(\mathrm{Y} = y)}{\mathbf{P}(\mathrm{X} = x, \mathrm{Y} = y)\mathbf{P}(\mathrm{Y} = y, \mathrm{Z} = z)} + \\
&\qquad + \sum_{x,y,z} \mathbf{P}(\mathrm{X} = x, \mathrm{Y} = y, \mathrm{Z} = z) \log \frac{\mathbf{P}(\mathrm{X} = x, \mathrm{Y} = y)}{\mathbf{P}(\mathrm{X} = x)\mathbf{P}(\mathrm{Y} = y))} \\
&= \sum_y \mathbf{P}(\mathrm{Y} = y) \sum_{x,z} \mathbf{P}(\mathrm{X} = x, \mathrm{Z} = z \mid \mathrm{Y} = y) \log \frac{\mathbf{P}(\mathrm{X} = x, \mathrm{Z} = z \mid \mathrm{Y} = y)}{\mathbf{P}(\mathrm{X} = x \mid \mathrm{Y} = y)\mathbf{P}(\mathrm{Z} = z \mid \mathrm{Y} = y)} \\
&\qquad + \sum_{x,y} \mathbf{P}(\mathrm{X} = x, \mathrm{Y} = y) \log \frac{\mathbf{P}(\mathrm{X} = x, \mathrm{Y} = y)}{\mathbf{P}(\mathrm{X} = x)\mathbf{P}(\mathrm{Y} = y))} \\
&= \mathrm{I}(\mathrm{X}, \mathrm{Z} \mid \mathrm{Y}) + \mathrm{I}(\mathrm{X}, \mathrm{Y}).
\end{aligned}$$

Par symétrie, on a également

$$\mathrm{I}(\mathrm{X}, (\mathrm{Y}, \mathrm{Z})) = \mathrm{I}(\mathrm{X}, \mathrm{Y} \mid \mathrm{Z}) + \mathrm{I}(\mathrm{X}, \mathrm{Z}),$$

de sorte que $\mathrm{I}(\mathrm{X}, \mathrm{Y} \mid \mathrm{Z}) \geqslant 0$, et $\mathrm{I}(\mathrm{X}, (\mathrm{Y}, \mathrm{Z})) \geqslant \mathrm{I}(\mathrm{X}, \mathrm{Z})$.

Dans l'hypothèse où X et Z sont conditionnellement indépendantes relativement à Y, on a $\mathrm{I}(\mathrm{X}, \mathrm{Z} \mid \mathrm{Y}) = 0$. Ces deux expressions pour $\mathrm{I}(\mathrm{X}, (\mathrm{Y}, \mathrm{Z}))$ entraînent alors le théorème suivant.

Théorème **(1.3.11)**. — *Soit* X, Y, Z *des variables aléatoires discrètes. Si* $X \perp_Y Z$, *alors* $I(X,Y) \geqslant I(X,Z)$.

Corollaire **(1.3.12)**. — *Soit* X, Y *des variables aléatoires discrètes et soit* f *une fonction. On a* $I(X, f(Y)) \leqslant I(X,Y)$.

1.4. Taux d'entropie

Une suite (X_n) de variables aléatoires est aussi appelée *processus stochastique*.

Définition **(1.4.1)**. — *On appelle* taux d'entropie *d'un processus stochastique* $X = (X_n)$ *l'expression*

$$H(X) = \lim_{n\to\infty} \frac{1}{n+1} H(X_0, \dots, X_n),$$

pourvu que la limite existe.

En remplaçant la limite par des limites supérieure et inférieure, on définit les taux d'entropie supérieur, $\overline{H}(X)$, et inférieur, $\underline{H}(X)$. On a l'inégalité $\underline{H}(X) = \overline{H}(X)$; le taux d'entropie existe si et seulement si ces deux expressions coïncident, et il leur est alors égal.

Exemple **(1.4.2)**. — Soit (X_n) un processus stochastique. On suppose que les variables aléatoires X_n sont indépendantes. Alors,

$$\frac{1}{n+1} H(X_0, \dots, X_n) = \frac{1}{n+1} \sum_{k=0}^{n} H(X_k);$$

le taux d'entropie est alors la limite au sens de Cesáro de la suite $(H(X_n))$.

Supposons de plus que les variables aléatoires sont identiquement distribuées. Alors, $H(X_k) = H(X_0)$ pour tout k, et l'on a $H(X) = H(X_0)$.

Lemme **(1.4.3)** (Cesáro). — *Soit* (a_n) *une suite de nombres réels ; pour tout entier* $n \geqslant 0$, *posons* $A_n = (a_0 + \cdots + a_n)/(n+1)$. *On a les inégalités*

$$\underline{\lim}\, a_n \leqslant \underline{\lim}\, A_n \leqslant \overline{\lim}\, A_n \leqslant \overline{\lim}\, a_n.$$

En particulier, si la suite (a_n) *a une limite* ℓ *dans* $[-\infty; +\infty]$, *la suite* (A_n) *converge également vers* ℓ.

Si f est une fonction de classe $\mathscr{C}^k$ par morceaux, on notera abusivement $f^{(k)}$ sa dérivée k-ième ; elle est définie seulement presque partout, plus précisément sauf sur un ensemble dont la trace sur tout intervalle compact est finie.

Lemme **(3.2.3)**. — a) *Soit* $f : [a;b] \to \mathbf{C}$ *une fonction de classe* $\mathscr{C}^1$ *par morceaux et continue. On a*

$$\int_a^b f'(t)\,dt = f(b) - f(a).$$

b) *Soit* $u, v : [a;b] \to \mathbf{C}$ *des fonctions de classe* $\mathscr{C}^1$ *par morceaux et continues. On a la formule d'intégration par parties :*

$$\int_a^b u'(t)v(t)\,dt = \big[u(t)v(t)\big]_a^b - \int_a^b u(t)v'(t)\,dt.$$

Démonstration. — *a*) Soit $(a_0, \dots, a_n)$ une suite finie, croissante, telle que $a = a_0$, $b = a_n$, et telle que f soit de classe $\mathscr{C}^1$ sur chaque intervalle $]a_{k-1}, a_k[$. Si x et y sont des éléments de $]a_{k-1}, a_k[$ tels que $a_{k-1} < x \leqslant y < a_k$, on a

$$\int_x^y f'(t)\,dt = f(y) - f(x),$$

par la formule fondamentale du calcul différentiel et intégral. Lorsqu'on fait tendre x vers a_{k-1} par valeurs supérieures et y vers a_k par valeurs inférieures, $f(x)$ tend vers $f(a_{k-1})$ et $f(y)$ tend vers $f(a_k)$, parce que f est continue. On obtient alors

$$\int_{a_{k-1}}^{a_k} f'(t)\,dt = f(a_k) - f(a_{k-1}).$$

Finalement, on a

$$\int_a^b f'(t)\,dt = \sum_{k=1}^{n} \int_{a_{k-1}}^{a_k} f'(t)\,dt = \sum_{k=1}^{n} \big(f(a_k) - f(a_{k-1})\big) = f(b) - f(a).$$

b) La preuve de la seconde formule est analogue. Plus simplement, on peut observer que la fonction f définie par $f(t) = u(t)v(t)$ est également de classe $\mathscr{C}^1$ par morceaux et continue, et que l'on a $f'(t) = u'(t)v(t) + u(t)v'(t)$ pour tout $t \in [a;b]$, sauf pour un nombre fini d'exceptions. En appliquant la formule *a*) à cette fonction f, on retrouve la formule *b*). □

peut bien sûr échantillonner à une fréquence plus basse, par exemple 8 000 Hz pour des téléphones basiques ou le protocole VoIP (*Voice over IP*).

Pour terminer cette introduction technologique, rappelons qu'en plus d'être échantillonnés, les signaux doivent être quantifiés. La norme Audio-CD, par exemple, les code sur 16 bits (2 octets), d'où, en principe, pour un signal stéréo, une quantité d'information de 176 kO par seconde. En fait, selon cette norme, 6 échantillons sont regroupés en un *frame* de 192 bits (24 octets), auquel s'ajoutent 8 octets de code correcteur d'erreur et un octet de contrôle (*subcode*) ; finalement, ce sont 33 octets pour 6 échantillons, d'où une quantité d'information de 242 kO par seconde. La durée d'un CD musical est ainsi de l'ordre d'une heure. L'utilisation d'algorithmes de compression permet d'y stocker un signal plus long ; c'est ainsi que certains CDs contiennent, non pas, un signal comme décrit ci-dessus, mais des fichiers compressés selon la norme MP3.

3.2. Série de Fourier d'une fonction périodique

3.2.1. — La théorie des *séries de Fourier* permet l'analyse fréquentielle des fonctions périodiques. Elle repose sur l'observation qu'une fonction trigonométrique $t \mapsto \exp(\omega t)$ est de période T si et seulement si ω est un multiple entier de $2\pi/\mathrm{T}$, et sur l'idée que « toute » fonction de période T est, en un sens qu'il faudra préciser, *somme* de telles fonctions trigonométriques.

La seule hypothèse que doit vérifier une fonction f sur $\mathbf{R}$ pour que l'on puisse envisager sa série de Fourier est d'être *localement intégrable* au sens de la théorie de Lebesgue, ce qui permettra de calculer son intégrale sur tout intervalle borné. Un cadre plus restrictif, mais souvent suffisant, est celui des *fonctions continues par morceaux*.

3.2.2. — Soit k un entier $\geqslant 0$. On dit qu'une fonction f définie sur un intervalle compact $[a, b]$ de $\mathbf{R}$ est de *classe $\mathscr{C}^k$ par morceaux* s'il existe une suite finie croissante $(a_0, a_1, \ldots, a_n)$ de nombres réels telle que $a_0 = a$, $a_n = b$, et telle que pour tout entier $p \in \{1, \ldots, n\}$, la fonction f soit de classe $\mathscr{C}^k$ sur $]a_{p-1}, a_p[$ et ait, ainsi que toutes ses dérivées d'ordres $\leqslant k$, une limite à droite en a_{p-1} et une limite à gauche en a_p.

Si f est définie sur un intervalle arbitraire de $\mathbf{R}$, on dit qu'elle est de classe $\mathscr{C}^k$ par morceaux si c'est le cas de sa restriction à tout intervalle compact contenu dans son intervalle de définition.

Démonstration. — Démontrons l'inégalité $\overline{\lim}\, \mathrm{A}_n \leqslant \overline{\lim}\, a_n$. Il n'y a rien à démontrer lorsque $\overline{\lim}\, a_n = +\infty$; supposons donc que $\overline{\lim}\, a_n < \infty$ et soit λ un nombre réel tel que $\overline{\lim}\, a_n < \lambda$. Alors, par définition de la limite supérieure, il existe un entier N tel que, pour tout entier $n \geqslant \mathrm{N}$, on ait $a_n \leqslant \lambda$. Pour $n \geqslant \mathrm{N}$, on a alors

$$\mathrm{A}_n = \frac{1}{n+1}\sum_{k=0}^{n} a_k = \frac{1}{n+1}\sum_{k=0}^{\mathrm{N}-1} a_k + \frac{1}{n+1}\sum_{k=\mathrm{N}}^{n} a_k \leqslant \frac{1}{n+1}\sum_{k=0}^{\mathrm{N}-1} a_k + \frac{n+1-\mathrm{N}}{n+1}\lambda.$$

Lorsque n tend vers l'infini, le membre de droite tend vers λ ; par suite, $\overline{\lim}\, \mathrm{A}_n \leqslant \lambda$. Comme λ est arbitraire, on a $\overline{\lim}\, \mathrm{A}_n \leqslant \overline{\lim}\, a_n$.

En remplaçant la suite (a_n) par la suite (b_n) définie par $b_n = -a_n$, la suite (A_n) est remplacée par la suite (B_n) définie par $\mathrm{B}_n = -\mathrm{A}_n$, et l'on a $\underline{\lim}\, a_n = -\overline{\lim}\, b_n$ et $\underline{\lim}\, \mathrm{A}_n = -\overline{\lim}\, \mathrm{B}_n$. L'inégalité de limites supérieures appliquée à la suite (b_n) entraîne alors que $\underline{\lim}\, a_n \leqslant \underline{\lim}\, \mathrm{A}_n$.

Lorsque la suite (a_n) converge vers un élément ℓ de $[-\infty; +\infty]$, on a $\underline{\lim}\, a_n = \ell = \overline{\lim}\, a_n$, et les inégalités précédentes entraînent que $\underline{\lim}\, \mathrm{A}_n = \overline{\lim}\, \mathrm{A}_n = \ell$, de sorte que la suite (A_n) converge vers ℓ. □

Définition **(1.4.4)**. — *On dit qu'un processus stochastique* (X_n) *est* stationnaire *si pour tout entier n et toute suite* $(x_0, \ldots, x_m)$*, on a*

$$\mathbf{P}(\mathrm{X}_n = x_0, \mathrm{X}_{n+1} = x_1, \ldots, \mathrm{X}_{n+m} = x_m) = \mathbf{P}(\mathrm{X}_0 = x_0, \mathrm{X}_1 = x_1, \ldots, \mathrm{X}_m = x_m).$$

Proposition **(1.4.5)**. — *Soit* $\mathrm{X} = (\mathrm{X}_n)$ *un processus stochastique stationnaire. Alors, le taux d'entropie* $\mathrm{H}(\mathrm{X})$ *existe, et est donné par*

$$\mathrm{H}(\mathrm{X}) = \lim_{n\to\infty} \mathrm{H}(\mathrm{X}_n \mid \mathrm{X}_{n-1}, \ldots, \mathrm{X}_0).$$

Démonstration. — Pour tout entier n, posons

$$\mathrm{H}'(\mathrm{X})_n = \mathrm{H}(\mathrm{X}_n \mid \mathrm{X}_{n-1}, \ldots, \mathrm{X}_0).$$

Puisque l'entropie diminue par conditionnement, on a, pour tout entier n, l'inégalité

$$\mathrm{H}'(\mathrm{X})_{n+1} = \mathrm{H}(\mathrm{X}_{n+1} \mid \mathrm{X}_n, \ldots, \mathrm{X}_0) \leqslant \mathrm{H}(\mathrm{X}_{n+1} \mid \mathrm{X}_n, \ldots, \mathrm{X}_1).$$

Puisque le processus X est stationnaire,

$$\mathrm{H}(\mathrm{X}_{n+1} \mid \mathrm{X}_n, \ldots, \mathrm{X}_1) = \mathrm{H}(\mathrm{X}_n \mid \mathrm{X}_{n-1}, \ldots, \mathrm{X}_0) = \mathrm{H}'(\mathrm{X})_n.$$

Ainsi, la suite $(H'(X))_n$ est croissante. Elle converge donc vers un élément de $[0;+\infty]$ que nous notons $H'(X)$.

Alors, pour tout entier n, on a

$$\begin{aligned} H(X_0,\dots,X_n) &= H(X_0) + H(X_1 \mid X_0) + \cdots + H(X_n \mid X_{n-1},\dots,X_0) \\ &= \sum_{k=0}^{n} H'(X)_k, \end{aligned}$$

de sorte que la suite $(\frac{1}{n+1}H(X_0,\dots,X_n))$ est la moyenne au sens de Cesáro de la suite $(H'(X)_n)$. Elle converge donc vers sa limite, ce qu'il fallait démontrer. □

1.5. Taux d'entropie des processus markoviens

Définition **(1.5.1)**. — *On dit qu'un processus stochastique* (X_n) *est* markovien *(ou est un processus de Markov, ou est une* chaîne de Markov*) si pour tout entier* n, $(X_0,\dots,X_{n-1})$ *et* X_{n+1} *sont conditionnellement indépendantes relativement à* X_n.

Cela signifie que pour tout entier n et toute suite $(x_0,\dots,x_{n+1})$, on a

$$\mathbf{P}(X_{n+1} = x_{n+1} \mid X_n = x_n,\dots,X_0 = x_0) = \mathbf{P}(X_{n+1} = x_{n+1} \mid X_n = x_n).$$

1.5.2. — Soit $X = (X_n)$ un processus markovien. On fait l'hypothèse supplémentaire qu'il est *homogène* c'est-à-dire que pour tout couple (a,b), on a

$$\mathbf{P}(X_{n+1} = b \mid X_n = a) = \mathbf{P}(X_1 = b \mid X_0 = a).$$

Supposons que l'ensemble A des valeurs possibles de (X_n) soit fini ; pour tout couple (a,b) d'éléments de A, posons $p_{a,b} = \mathbf{P}(X_1 = b \mid X_0 = a)$ et notons P la matrice $(p_{a,b})$.

C'est une matrice carrée à indices dans l'ensemble A ; même si A n'est pas forcément de la forme $\{1,\dots,m\}$, la théorie est identique. Les coefficients de la matrice P sont des probabilités conditionnelles ; ils sont donc positifs ou nuls. Pour tout a, on a

$$\sum_{b\in A} p_{a,b} = \sum_{b\in A} \mathbf{P}(X_1 = b \mid X_0 = a) = 1.$$

Autrement dit la somme des coefficients de chaque ligne de P est égale à 1. On dit que P est une *matrice stochastique.*

La matrice P est appelée la *matrice de transition du processus markovien* X.

existent également, mais dans une proportion bien plus faible. (Dans certains instruments, tels les toms d'une batterie, les fréquences anharmoniques sont très présentes, si bien qu'on peut difficilement leur attribuer une note, mais la façon dont l'instrument est joué, la baguette utilisée par exemple, influe beaucoup sur le son et même sur sa hauteur.)

piano	27,5 Hz	4 186 Hz
saxophone soprano	233 Hz	1480 Hz
alto	139 Hz	831 Hz
ténor	104 Hz	659 Hz
baryton	65 Hz	440 Hz
batterie – cymbales	200 Hz	10 000 Hz
caisse claire	240 Hz	6 000 Hz
toms	120 Hz	5 000 Hz
grosse caisse	60 Hz	4 000 Hz
voix – soprano	261 Hz	1 047 Hz
ténor	123 Hz	440 Hz
basse	82 Hz	300 Hz
violon	196 Hz	2 794 Hz

FIGURE 3.1.0.1. Domaines de fréquences de quelques instruments de musique

La perception des sons dépend ensuite du fonctionnement de notre oreille, du tympan qui transforme les oscillations de la pression de l'air en vibrations mécaniques qu'il transmet à la cochlée, un petit os en forme de limaçon rempli d'un liquide où des milliers de cellules ciliées réagissent aux diverses fréquences du son et les transforment en signal nerveux. Ainsi, l'oreille humaine est sensible aux signaux de fréquences variant de 20 Hz à 20 000 Hz ; selon HEARING RANGE (2005), des conditions idéales permettent d'observer une sensibilité plus large, de 12 Hz à 28 000 Hz, et la partie où l'audition est le plus efficace est entre 2 000 et 5 000 Hz.

En conclusion, pour les applications aux signaux sonores, on peut prendre pour fréquence de Nyquist toute fréquence supérieure à 40 000 Hz. Celle choisie par la norme Audio-CD est 44,1 kHz permet donc, en théorie, de recréer tout le spectre audible d'un signal. Si l'on se contente d'un signal de moindre qualité, on

fois dans Shannon (1949). Cependant, Shannon y insiste que ce résultat était *common knowledge in the communication art;* l'idée de l'échantillonnage à une fréquence double était également bien connue de H. Nyquist. C'est pourquoi on trouve aussi l'appellation *théorème de Nyquist–Shannon*.

La preuve de ce théorème repose sur la théorie des séries et de la transformation de Fourier qui permet de décomposer tout signal en une combinaison de signaux trigonométriques « purs ». C'est aussi cette théorie qui fournira une définition précise de l'ensemble des *fréquences* qui apparaissent dans un signal donné — ce sera le support de sa transformée de Fourier.

La considération d'un tel signal trigonométrique pur explique déjà la nécessité de l'échantillonnage à la fréquence de Nyquist. Si la fonction F définie par $F(t) = \sin(\omega t)$ (de période $2\pi/\omega$, de fréquence $f = \omega/2\pi$) est échantillonnée à la fréquence double ω/π, on obtient les données $F(n\pi/\omega) = \sin(n\pi) = 0$. On ne peut donc distinguer le signal F du signal nul !

Nous pouvons aussi expliquer tout de suite la façon dont ce théorème, appliqué aux signaux sonores, intervient dans la vie de tous les jours.

Les sons simples que nous percevons, ceux d'une voix chantée, d'un instrument de musique, etc., ont une *fréquence fondamentale* f_1, qui détermine la note (do, ré,…) que nous attribuerons à ce son. Lorsque cette fréquence fondamentale est doublée, nous entendons la « même » note, à l'octave supérieur, et ce qui fait la richesse de ces sons est qu'ils « contiennent » des *harmoniques*, c'est-à-dire des fréquences multiples $f_2 = 2f_1$, $f_3 = 3f_1$,…, de la fréquence fondamentale. C'est notre perception du son qui unifie toutes ces signaux en un son unique. Au subtilités (fondamentales) des gammes près, lorsque la fréquence fondamentale f_1 correspond à un do, les harmoniques suivantes correspondent, $f_2 = 2f_1$ au do de l'octave supérieur, $f_3 = 3f_1$ au sol de cet octable, $f_4 = 4f_1$ au do de l'octave encore supérieur, puis $f_5 = 5f_1$ au mi de cet octave et $f_6 = 6f_1$ au sol de cet octave, un octave plus haut que la troisième harmonique $f_3 = 3f_1$. On devine là d'où provient l'impression de solidité que procure l'accord « parfait » (do-mi-sol, par exemple) en harmonie classique.

Nous avons indiqué figure 3.1.0.1 le spectre des fréquences fondamentales (arrondies au Hz le plus proche) de quelques instruments de musique. Elle ne tient pas compte des harmoniques qui, nous l'avons dit, sont responsables de la nature du son ; disons que les 10 premières harmoniques ont une importance. Cette figure tient encore moins compte des fréquences *anharmoniques* qui

Dans le vocabulaire des chaînes de Markov, les éléments de A sont appelés *états*, et $p_{a,b}$ est la probabilité de passage de l'état a à l'état b. On représente souvent une telle chaîne par un carquois (graphe orienté) dont les sommets sont les états de la chaîne, muni pour chaque couple d'états (a, b), d'une flèche de l'état a à l'état b étiquetée de la probabilité $p_{a,b}$.

Ainsi, le graphe de la figure 1.5.2.1 représente une chaîne de Markov à deux états $\{a, b\}$. La probabilité de passer de a à b est égale à p, celle de passer de b à a est égale à q. Sa matrice de transition est ainsi donnée par

$$\mathrm{P} = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}.$$

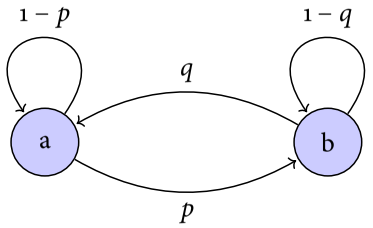


Figure 1.5.2.1. Une chaîne de Markov à deux états.

Si $\mathrm{M} = (\mu_a)$ est la loi de X_n, considérée comme un vecteur-ligne, la loi $\mathrm{M}' = (\mu'_a)$ de X_{n+1} est donnée par

$$\mu'_a = \mathbf{P}(\mathrm{X}_{n+1} = a) = \sum_{b \in \mathrm{A}} \mathbf{P}(\mathrm{X}_n = b) \cdot \mathbf{P}(\mathrm{X}_{n+1} = a \mid \mathrm{X}_n = b) = \sum_{b \in \mathrm{A}} \mu_b p_{b,a}.$$

Autrement dit, on a $\mathrm{M}' = \mathrm{MP}$.

Par récurrence, si le vecteur M représente la loi de X_0, la loi de X_n est représentée par le vecteur MP^n.

Proposition **(1.5.3)**. — *Soit* $\mathrm{X} = (\mathrm{X}_n)$ *un processus markovien homogène. Soit* A *l'ensemble des valeurs de* X_0, *soit* $\mathrm{P} = (p_{a,b})$ *la matrice de transition de* X *et soit* $\mathrm{M} = (\mu_a)$ *la loi de* X_0. *Pour que* X *soit stationnaire, il faut et il suffit que l'on ait* $\mathrm{M} = \mathrm{MP}$. *Dans ce cas, on a*

$$\mathrm{H}(\mathrm{X}) = -\sum_{a,b} \mu_a p_{a,b} \log(p_{a,b}).$$

Démonstration. — Si X est stationnaire, alors X_0 et X_1 ont même loi, donc M = MP. Supposons inversement que M = MP et prouvons que X est un processus stationnaire. On sait déjà que pour tout entier n, la loi de X_n est donnée par M. Démontrons par récurrence sur m que $\mathbf{P}(X_n = x_0, \dots, X_{n+m} = x_m) = \mathbf{P}(X_0 = x_0, \dots, X_m = x_m)$ pour tout entier m, tous $x_0, \dots, x_m \in A$ et tout $n \in \mathbf{N}$. Par définition d'un processus markovien homogène, on a

$$\begin{aligned}
&\mathbf{P}(X_n = x_0, \dots, X_{n+m} = x_m) \\
&\quad = \mathbf{P}(X_{n+m} = x_m \mid X_n = x_0, \dots, X_{n+m-1} = x_{m-1}) \cdot \mathbf{P}(X_n = x_0, \dots, X_{n+m-1} = x_{m-1}) \\
&\quad = \mathbf{P}(X_{n+m} = x_m \mid X_{n+m-1} = x_{m-1}) \cdot \mathbf{P}(X_n = x_0, \dots, X_{n+m-1} = x_{m-1}) \\
&\quad = \mathbf{P}(X_1 = x_m \mid X_0 = x_{m-1}) \cdot \mathbf{P}(X_n = x_0, \dots, X_{n+m-1} = x_{m-1}).
\end{aligned}$$

Par l'hypothèse de récurrence, on a

$$\mathbf{P}(X_n = x_0, \dots, X_{n+m-1} = x_{m-1}) = \mathbf{P}(X_0 = x_0, \dots, X_{m-1} = x_{m-1}).$$

Ainsi

$$\begin{aligned}
&\mathbf{P}(X_n = x_0, \dots, X_{n+m} = x_m) \\
&\qquad = \mathbf{P}(X_1 = x_m \mid X_0 = x_{m-1}) \cdot \mathbf{P}(X_0 = x_0, \dots, X_{m-1} = x_{m-1}) \\
&\qquad = \mathbf{P}(X_0 = x_0, \dots, X_{m-1} = x_{m-1}, X_m = x_m)
\end{aligned}$$

en utilisant une fois de plus l'hypothèse que le processus X est markovien. Cela prouve que X est un processus stationnaire.

Supposons maintenant que X est stationnaire. Pour tout entier n, on a $H(X_n \mid X_{n-1}, \dots, X_0) = H(X_n \mid X_{n-1})$, par la propriété markovienne, puis $H(X_n \mid X_{n-1}) = H(X_1 \mid X_0)$ par stationarité. On a ainsi $H(X_n \mid X_{n-1}, \dots, X_0) = H(X_1 \mid X_0)$, d'où la formule $H(X) = H(X_1 \mid X_0)$ en vertu de la proposition 1.4.5.

Par ailleurs, la définition de l'entropie conditionnelle entraîne

$$H(X_1 \mid X_0) = \sum_a \mathbf{P}(X_0 = a) H(X_1 \mid X_0 = a) = -\sum_a \mu_a \sum_b p_{a,b} \log(p_{a,b}),$$

ainsi qu'il fallait démontrer. □

Exemple **(1.5.4)**. — Reprenons l'exemple d'une chaîne de Markov X à deux états $\{a, b\}$ représentée par le graphe de la figure 1.5.2.1.

CHAPITRE 3

ÉCHANTILLONAGE

3.1. Signaux continus et signaux discrets

Considérons un signal; ce peut être le chant d'une artiste, représenté par exemple par la pression d'air exercée au cours du temps sur un microphone, ou un signal visuel, tel une image à photographier, alors représentée par la luminosité émise par chaque point de la scène. *Échantillonner* ce signal, c'est le mesurer à divers instants, ou à divers lieux, souvent régulièrement espacés; cela transforme ainsi un signal continu (une fonction du temps) en un signal discret (une suite de valeurs). La question de l'échantillonnage est apparue très tôt en théorie de la communication. Les ingénieurs l'ont par exemple utilisée dès le milieu du XIX^{e} siècle pour faire passer plusieurs signaux sur un même canal. Elle est aujourd'hui fondamentale pour le traitement numérique du signal puisque les ordinateurs ne manipulent qu'une quantité finie d'information.

D'ailleurs, les ordinateurs doivent faire plus qu'échantillonner : ils doivent également *quantifier* le signal c'est-à-dire transformer une valeur continue (la pression, une tension électrique) en une valeur discrète, que l'on peut représenter sur 8 bits ou 16 bits, par exemple...). C'est cette combinaison échantillonnage/quantification, et la reconstruction ultérieure du signal, qui est au cœur de l'ingéniérie du traitement du signal.

Comme l'indique son titre, nous nous contenterons dans ce chapitre de la question de l'échantillonnage en démontrant que l'*on peut reconstruire un signal échantillonné s'il ne contenait pas de fréquences supérieures à la moitié de la fréquence d'échantillonnage.* Dit autrement, si l'on échantillonne un signal à une fréquence au moins deux fois supérieure à celles qu'il contient, on pourra le reconstruire exactement, au moins théoriquement. Ce théorème mathématique est le plus souvent attribué à Shannon, car il apparaît semble-t-il, pour la première

Les lois de X_o pour lesquelles cette chaîne est stationnaire sont données par un vecteur (μ, ν) tel que

$$\begin{cases} \mu(1-p) + \nu q &= \mu \\ \mu p + \nu(1-q) &= \nu \end{cases}$$

On obtient l'égalité $p\mu = q\nu$. Joint à la condition $\mu + \nu = 1$, on voit qu'il existe une unique telle loi, donnée par

$$(\mu, \nu) = \left(\frac{q}{p+q}, \frac{p}{p+q}\right).$$

Alors, le taux d'entropie de X est égal à

$$\begin{aligned} \mathrm{H}(\mathrm{X}) &= \frac{q}{p+q}\left(-(1-p)\log(1-p) - p\log(p)\right) \\ &\qquad + \frac{p}{p+q}\left(-q\log(q) - (1-q)\log(1-q)\right) \\ &= \frac{q}{p+q}h(p) + \frac{p}{p+q}h(q), \end{aligned}$$

où h est la fonction entropie représentée dans la figure 1.1.5.2.

Par ailleurs, la loi de X_n est, pour tout entier n, donnée par (μ, ν), de sorte que l'entropie de X_n est égale à

$$\mathrm{H}(\mathrm{X}_n) = -\frac{q}{p+q}\log\left(\frac{q}{p+q}\right) - \frac{p}{p+q}\log\left(\frac{p}{p+q}\right) = h\Big(\frac{p}{p+q}\Big).$$

1.5.5. — Soit $\mathrm{X} = (\mathrm{X}_n)$ un chaîne de Markov homogène à ensemble d'états A fini ; soit $\mathrm{P} = (p_{a,b})$ sa matrice de transition. On dit que la chaîne X, ou que la matrice stochastique P, est *primitive* s'il existe un entier $m \geqslant 1$ tel que tous les coefficients de la matrice P^m soient strictement positifs.

Théorème **(1.5.6)** (O. Perron, 1907). — *Soit* P *une matrice stochastique primitive. La suite de matrices* (P^n) *converge ; sa limite* Q *est une matrice stochastique de rang* 1.

Démonstration. — On munit $\mathbf{R}^{\mathrm{A}}$ de la norme définie par $\|\mathrm{X}\| = \sum_{a\in\mathrm{A}}|x_a|$, pour $\mathrm{X} = (x_a)$; on pose aussi $f(\mathrm{X}) = \sum_{a\in\mathrm{A}} x_a$.

La démonstration du théorème requiert plusieurs étapes. On note Σ l'ensemble des $\mathrm{X} \in \mathbf{R}_+^{\mathrm{A}}$ tels que $f(\mathrm{X}) = 1$; c'est une partie convexe et compacte de $\mathbf{R}^{\mathrm{A}}$. Soit m

un entier $\geqslant 1$ tel que tous les coefficients $(p'_{a,b})$ de la matrice $P' = P^m$ soient strictement positifs ; notons c leur borne inférieure.

a) *Pour tout* $X \in \mathbf{R}^A$, *on a* $\|XP\| \leqslant \|X\|$ *et* $f(XP) = f(X)$.

Soit $X \in \mathbf{R}^A$. Posons $Y = XP$; on a $Y = (y_b)$, avec $y_b = \sum_a x_a p_{a,b}$. Grâce à l'inégalité triangulaire, on a

$$\|Y\| = \sum_{b\in A} |y_b| \leqslant \sum_{a,b} |x_a| p_{a,b} = \sum_a |x_a| \sum_b p_{a,b} = \sum_a |x_a| = \|X\|,$$

ce qui prouve l'inégalité voulue.

De même,

$$f(Y) = \sum_b y_b = \sum_{a,b} x_a p_{a,b} = \sum_a x_a = f(X).$$

b) *Pour tout* $X \in \mathbf{R}^A_+$, *les coefficients de* XP *sont positifs, et ceux de* XP' *sont supérieurs à* $c\,\|X\|$.

Soit $X \in \mathbf{R}^A_+$. Posons $Y = XP$ et notons $Y = (y_b)$. On a $y_b = \sum_a x_a p_{a,b}$; on voit donc que $y_b \geqslant 0$ pour tout b.

De même, posons $Z = XP' = (z_b)$. Puisque $p'_{a,b} \geqslant c$ et $x_a \geqslant 0$ pour tous a, b, on a

$$z_b = \sum_a x_a p'_{a,b} \geqslant \sum_a x_a c = c.$$

c) *Soit* $X \in \mathbf{R}^A$ *tel que* $f(X) = 0$. *Alors,* $\|XP'\| \leqslant (1 - c\,\mathrm{Card}(A))\,\|X\|$.

Soit $Z = XP'$; notons $Z = (z_b)$, de sorte que $z_b = \sum_a x_a p'_{a,b}$. Comme $\sum x_a = 0$, on a aussi $z_b = \sum_a x_a (p'_{a,b} - c)$, de sorte que

$$|z_b| \leqslant \sum_a |x_a| (p'_{a,b} - c) \leqslant \sum_a |x_a| p'_{a,b} - c\,\|X\|.$$

En sommant sur b, on obtient

$$\|Z\| \leqslant \|X\| - c\,\mathrm{Card}(A)\,\|X\| = (1 - c\,\mathrm{Card}(A))\,\|X\|.$$

d) *L'ensemble* Σ *est stable par l'application* $X \mapsto XP'$, *et cette application est contractante pour la norme* $\|\cdot\|$.

La stabilité de Σ découle de ce qui a été dit plus haut. Par ailleurs, soit X, X' des éléments de Σ ; posons $Y = XP'$ et $Y' = X'P'$. On a $f(X' - X) = 0$ et $Y' - Y = (X' - X)P'$; alors

$$\|Y' - Y\| \leqslant (1 - c\,\mathrm{Card}(A))\,\|X' - X\|,$$

d'où l'assertion pusique $1 - c\,\mathrm{Card}(A) < 1$.

□

2.6.11. Conclusion de la démonstration du théorème 2.6.4. — Rappelons qu'il s'agit de prouver qu'il existe, pour tout nombre réel ρ tel que $\rho < I(C)$, tout nombre réel $\alpha > 0$ et tout entier n assez grand, un code de longueur n adapté au canal C, de taux de transmission au moins ρ et de probabilité d'erreur maximale $< \alpha$. Soit ρ' un nombre réel tel que $\rho < \rho' < I(C)$. D'après le corollaire 2.6.9, il existe, pour tout entier n assez grand, un code Φ de longueur n, adapté au canal C, de taux de transmission $\geqslant \rho'$ et de probabilité d'erreur moyenne $< \alpha/2$. Soit Φ' un code tel que construit dans le lemme 2.6.10. Son taux de transmission est au moins égal à $\rho' - \frac{\log(2)}{n}$, donc $\tau(\Phi') \geqslant \rho$ si n est assez grand, précisément, si $n \geqslant \log(2)/(\rho' - \rho)$, et sa probabilité d'erreur maximale est au plus α. Le théorème est ainsi démontré.

Par suite, il existe Φ tel que $\mathbf{P}(U = 1 \mid f = f_\Phi) < \alpha$. Par ailleurs, comme la variable aléatoire W est indépendante de Φ est uniforme dans M, on a

$$\mathbf{P}(U = 1 \mid f = f_\Phi) = \frac{1}{\mathrm{Card}(M)} \sum_{m \in M} \mathbf{P}(U = 1 \mid f = f_\Phi, W = m).$$

Par définition, $\mathbf{P}(U = 1 \mid f = f_\Phi, W = m) = \lambda_m(\Phi)$, la probabilité d'erreur de transmission lorsque le mot m est transmis dans le canal C au moyen du code Φ. Ainsi,

$$\lambda_{\mathrm{moy}}(\Phi) = \mathbf{P}(U = 1 \mid f = f_\Phi) < \alpha.$$

Enfin, le taux de transmission du code Φ vérifie

$$\tau(\Phi) = \frac{\log(\mathrm{Card}(\Phi))}{n} \geqslant \rho,$$

ce qui conclut la démonstration du corollaire. □

Lemme **(2.6.10)**. — *Soit Φ un code de longueur n adapté à un canal sans mémoire C. Il existe un code Φ' de même longueur tel que*

$$\tau(\Phi') \geqslant \tau(\Phi) - \frac{\log(2)}{n} \quad \textit{et} \quad \lambda_{\max}(\Phi') \leqslant 2\lambda_{\mathrm{moy}}(\Phi).$$

Démonstration. — Soit M le domaine du code Φ, soit f sa fonction de codage et g sa fonction de décodage.

Soit M′ l'ensemble des éléments $m \in M$ tels que $\lambda_m(\Phi) \leqslant 2\lambda_{\mathrm{moy}}(\Phi)$. Appliquons l'inégalité de Markov à la fonction $m \mapsto \lambda_m(\Phi)$ sur l'univers M muni de la mesure de probabilité uniforme ; son espérance est $\lambda_{\mathrm{moy}}(\Phi)$. On a donc

$$\mathbf{P}(\lambda_m(\Phi) > 2\lambda_{\mathrm{moy}}(\Phi)) \leqslant \frac{1}{2},$$

c'est-à-dire $\mathrm{Card}(M - M') \leqslant \frac{1}{2}\mathrm{Card}(M)$, soit encore $\mathrm{Card}(M') \geqslant \frac{1}{2}\mathrm{Card}(M)$.

Soit $f' : M' \to A^n$ la restriction à M′ de la fonction de codage f. On choisit une fonction $g' : B^n \to M'$ telle que $g'(b) = g(b)$ si $g(b) \in M'$. Alors, (f', g') est un code Φ' sur l'ensemble M′ adapté au canal C. Pour $m \in M'$, on a

$$\begin{aligned}\lambda_m(\Phi') &= \mathbf{P}(g'(Y) \neq m \mid X = f'(m)) \\ &\leqslant \mathbf{P}(g(Y) \neq m \mid X = f(m)) \\ &= \lambda_m(\Phi) \leqslant 2\lambda_{\mathrm{moy}}(\Phi),\end{aligned}$$

donc $\lambda_{\max}(\Phi') \leqslant 2\lambda_{\mathrm{moy}}(\Phi)$. Enfin, le taux de transmission de ce code Φ' vérifie

$$\tau(\Phi') = \frac{\log(\mathrm{Card}(M'))}{n} \geqslant \frac{\log(\mathrm{Card}(M)) - \log(2)}{n} \geqslant \tau(\Phi) - \frac{\log(2)}{n}.$$

e) *La suite de matrices $((P')^n)$ converge ; sa limite est une matrice stochastique de rang* 1.

D'après le théorème du point fixe de Picard, l'application $X \mapsto XP'$ possède un unique point fixe M dans Σ et, pour tout vecteur $X \in \Sigma$, la suite $(X(P')^n)$ converge vers M.

L'espace Σ contient les vecteurs X_a de la base canonique. Pour chacun d'entre eux, on a donc $X_a(P')^n \to M$. Cela prouve que la ligne a de la suite de matrices $((P')^n)$ converge vers M. La suite $((P')^n)$ converge donc vers la matrice Q dont toutes les lignes sont égales à M ; c'est une matrice stochastique de rang 1.

f) *La suite de matrices (P^n) converge vers* Q.

En écrivant la division euclidienne de n par m, $n = mk + d$, où $0 \leqslant d \leqslant m - 1$, on a $P^n = P^d(P')^k$. Supposons que n tende vers l'infini en restant dans la classe de d modulo m ; on a donc $P^n \to P^dQ$. Or, comme toutes les vecteurs-ligne de P^d appartiennent à Σ, on a $P^dQ = Q$. Par suite, toutes ces sous-suites ont même limite, Q, ce qui entraîne que P^n converge vers Q. □

Remarque **(1.5.7)**. — Toutes les lignes de la matrice Q sont égales à un même vecteur M à coefficients positifs, de somme égale à 1. On a $MP = M$, ce qui prouve que M est un « vecteur propre » à gauche de P, pour la valeur propre 1.

Soit N un vecteur propre à gauche de P pour une valeur propre λ. On a donc $NP = \lambda N$ puis, par itération, $NP^n = \lambda^n N$ pour tout n. Lorsque n tend vers l'infini, le membre de gauche tend vers NQ. Comme $N \neq 0$, cela entraîne que la suite (λ^n) converge : on a donc $\lambda = 1$ ou $|\lambda| < 1$.

Supposons $\lambda = 1$. Soit $X = N - f(N)M$, de sorte que $f(X) = 0$. D'après le point c) de la preuve, on a donc $\|XQ\| \leqslant (1 - c\,\mathrm{Card}(A))\,\|X\|$. Or, $X = XP = XQ$, ce qui entraîne $X = 0$ et $N = f(N)M$.

Sous les conditions du théorème, la matrice P possède une unique valeur propre de module $\geqslant 1$; cette valeur propre est égale à 1 et l'espace propre (à gauche) correspondant est de dimension 1, engendré par M.

Théorème **(1.5.8)**. — *Soit* X *une chaîne de Markov homogène, à ensemble d'états fini,* primitive. *Soit* $P = (p_{a,b})$ *sa matrice de transitions et soit* $M = (\mu_a)$ *son unique loi stationnaire. Alors, le taux d'entropie existe et est donné par*

$$H(X) = \lim_{n \to \infty} H(X_n \mid X_{n-1}) = -\sum_{a,b} \mu_a p_{a,b} \log(p_{a,b}).$$

Démonstration. — Soit A l'ensemble des états de X et soit $M_0 = (m_a^{(0)})$ le vecteur de $\mathbf{R}^A$ décrivant la loi de X_0. Pour tout entier n, la loi de X_n est représentée par le vecteur $M_n = M_0 P^n = (m_a^{(n)})$. Par suite,

$$H(X_n \mid X_{n-1}) = \sum_a m_a^{(n-1)} H(X_n \mid X_{n-1} = a) = -\sum_{a,b} m_a^{(n-1)} p_{a,b} \log(p_{a,b}).$$

Puisque $m_a^{(n)} \to \mu_a$ pour tout a, on a donc

$$H(X_n \mid X_{n-1}) \to -\sum_{a,b} \mu_a p_{a,b} \log(p_{a,b}).$$

Notons $H'(X)$ cette expression.

Par ailleurs, on a

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

Comme X est un processur markovien, on a l'égalité

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

si bien que

$$H(X_1, \dots, X_n) = H(X_1) + \sum_{k=2}^{n} H(X_k \mid X_{k-1}).$$

D'après le lemme de Cesáro, on a donc

$$\frac{1}{n} H(X_1, \dots, X_n) \to H'(X),$$

ce qui prouve que le taux d'entropie de X existe et est égal à $H'(X)$. □

de sorte que

$$\mathbf{P}(U = 1 \mid W = m) \leqslant \mathbf{P}(E \mid W = m) + \sum_{m' \neq m} \mathbf{P}(E_{m'} \mid W = m).$$

Alors,

$$\mathbf{P}(U = 1) \leqslant \mathbf{P}((X, Y) \notin C_\varepsilon^n) + \frac{1}{\mathrm{Card}(M)} \sum_{m \neq m'} \mathbf{P}((f(m'), Y) \in C_\varepsilon^n \mid X = f(m)).$$

Introduisons une variable aléatoire X′ à valeurs dans A^n, de même loi que X mais indépendante de X; alors X′ et Y sont indépendantes. Soit $m' \in M$ tel que $m' \neq m$; conditionnée à $X = m$, la variable aléatoire $f(m')$ se comporte comme X′, de sorte que que le couple $(f(m'), Y)$ a même loi que (X', Y). Ainsi,

$$\frac{1}{\mathrm{Card}(M)} \sum_{m \neq m'} \mathbf{P}((f(m'), Y) \in C_\varepsilon^n \mid X = f(m)) \leqslant \mathrm{Card}(M) \mathbf{P}((X', Y) \in C_\varepsilon^n \mid X = f(m)),$$

de sorte que

$$\mathbf{P}(U = 1) \leqslant \mathbf{P}((X, Y) \notin C_\varepsilon^n) + \mathbf{P}((X', Y) \in C_\varepsilon^n.$$

Le premier terme est majoré par $(c(X) + c(Y) + c(Z))/2n\varepsilon^2$, et le second est majoré par

$$\mathrm{Card}(M) e^{-n(I(X_1, Y_1)) - 3\varepsilon} = \lceil e^{n\rho} \rceil e^{-n(I(C) - 3\varepsilon)} \sim e^{n(\rho - I(C) + 3\varepsilon)}.$$

Comme $\rho < I(C)$, il existe $\varepsilon > 0$ tel que $\rho - I(C) + 3\varepsilon < 0$. La majoration de $\mathbf{P}(U = 1)$ tend alors vers 0 quand n tend vers $+\infty$; pour n assez grand, on a donc $\mathbf{P}(U = 1) < \alpha$. □

Corollaire **(2.6.9)**. — *Soit* C *un canal sans mémoire. Soit* $\alpha > 0$. *Pour tout nombre réel* $\rho < I(C)$ *et tout entier* n *assez grand, il existe un code* Φ *de longueur* n *adapté au canal* C *dont le taux de transmission est au moins* ρ *et dont la probabilité d'erreur moyenne* $\lambda_{\mathrm{moy}}(\Phi)$ *est inférieure à* α.

Démonstration. — Choisissons n assez grand de sorte que $\mathbf{P}(U = 1) < \alpha$ (proposition 2.6.8). En conditionnant sur tous les codes possibles, on a

$$\mathbf{P}(U = 1) = \frac{1}{\mathrm{Card}(\{\Phi\})} \sum_{\Phi} \mathbf{P}(U = 1 \mid f = f_\Phi).$$

La méthode suivie par Shannon (1948), et peu modifiée depuis, consiste, non pas à construire explicitement un code adapté au canal C, mais à évaluer l'espérance de la probabilité d'erreur lorsque le code Φ est choisi aléatoirement. Ainsi, pour tout $m \in \mathrm{M}$, $f(m)$ est une variable aléatoire sur A^n ; on suppose que ces variables $f(m)$ sont indépendantes et ont toutes pour loi la loi sur A^n qui réalise la capacité du canal C. (Rappelons que $\mathrm{I}(\mathrm{C})$ est la borne supérieure, pour toutes les lois sur A, de l'information mutuelle $\mathrm{I}(\mathrm{X},\mathrm{Y})$, où X et Y sont des variables aléatoires à valeurs dans A et B respectivement, liées par la relation $\mathbf{P}(\mathrm{Y}=b \mid \mathrm{X}=a) = p(b \mid a)$, $p(\cdot \mid \cdot)$ désignant la probabilité de transmission du canal C. Cette borne supérieure est réalisée par une loi ; si elle ne l'avait pas été, on aurait choisi une loi qui l'approche.)

La fonction de décodage g_Φ n'est pas aléatoire, mais est définie par la stratégie de la « correction typique ». Si elle n'est pas explicite, cette stratégie a l'avantage de permettre une évaluation relativement simple de la probabilité d'erreur. Dans $\mathrm{C}^n = \mathrm{A}^n \times \mathrm{B}^n$, soit C_ε^n l'ensemble typique adapté à un paramètre $\varepsilon > 0$; autrement dit, $(a,b) \in \mathrm{C}_\varepsilon^n$ si et seulement si $\mathbf{P}(\mathrm{X}=a, \mathrm{Y}=b)$ est de l'ordre de $e^{-n\mathrm{H}(\mathrm{X},\mathrm{Y})}$, $\mathbf{P}(\mathrm{X}=a)$ est de l'ordre de $e^{-n\mathrm{H}(\mathrm{X})}$ et $\mathbf{P}(\mathrm{Y}=b)$ est de l'ordre de $e^{-n\mathrm{H}(\mathrm{Y})}$. Par définition, la fonction de décodage g applique un élément $b \in \mathrm{B}^n$ sur un élément $m \in \mathrm{M}$ tel que $(f(m), b) \in \mathrm{C}_\varepsilon^n$, s'il existe un tel élément et un seul, et sur un élément non précisé sinon.

On choisit aussi aléatoirement le mot $m \in \mathrm{M}$ qui est transmis, au moyen d'une variable aléatoire W uniforme dans M, indépendante du code Φ. Le mot transmis dans le canal est $\mathrm{X} = f(\mathrm{W})$, celui qui est reçu est Y, et le mot décodé est $\mathrm{W}' = g(\mathrm{Y})$. Notons U la variable aléatoire qui vaut 1 lorsque $\mathrm{W}' \neq \mathrm{W}$ et 0 sinon ; il s'agit pour commencer de montrer que l'espérance de U est petite.

Proposition **(2.6.8)**. — *Soit $\alpha > 0$. Il existe $\varepsilon > 0$ tel que, dès que n est assez grand, on a $\mathbf{P}(\mathrm{U}=1) \leqslant \alpha$.*

Démonstration. — Puisque $\mathbf{P}(\mathrm{W}=m) = 1/\operatorname{Card}(\mathrm{M})$ pour tout $m \in \mathrm{M}$, on a

$$\mathbf{P}(\mathrm{U}=1) = \sum_{m\in\mathrm{M}} \mathbf{P}(\mathrm{W}=m)\mathbf{P}(\mathrm{U}=1 \mid \mathrm{W}=m) = \frac{1}{\operatorname{Card}(\mathrm{M})} \sum_{m\in\mathrm{M}} \mathbf{P}(\mathrm{U}=1 \mid \mathrm{W}=m).$$

Fixons un élément $m \in \mathrm{M}$ et conditionnons la situation à $\mathrm{W}=m$; on a erreur lorsque, soit $(f(m), \mathrm{Y})$ n'appartient pas à C_ε^n (événement E ; rappelons que $\mathrm{X} = f(\mathrm{W})$), soit il existe $m' \neq m$ tel que $(f(m'), \mathrm{Y})$ appartient à C_ε^n (événement $\mathrm{E}_{m'}$),

CHAPITRE 2

CODAGE

Nous abordons maintenant la contribution de Shannon au codage.

2.1. Codes

2.1.1. Alphabets et mots. — Soit A un ensemble. On considère les éléments de A comme les lettres d'un *alphabet* et on considère les mots que l'on peut écrire avec ces symboles. Par définition, un *mot* est une suite finie $(a_1, \ldots, a_n)$ d'éléments de A ; l'entier n est la longueur de ce mot. On notera A^* l'ensemble des mots écrits dans l'alphabet A ; c'est la réunion, lorsque l'entier n varie, des ensembles A^n des mots de longueur n.

Il y a un seul mot de longueur 0, c'est le mot vide, parfois noté ε. L'ensemble A^* est muni d'une loi interne de concaténation. Si $a = (a_1, \ldots, a_n)$ et $b = (b_1, \ldots, b_m)$ sont des mots, le mot ab est donné par la suite $(a_1, \ldots, a_n, b_1, \ldots, b_m)$. Sa longueur est la somme des longueurs des mots a et b. Cette loi interne est associative ; le mot vide est son un élément neutre.

On notera $\ell(a)$ la longueur du mot a.

2.1.2. — Un code C est la donnée d'un second alphabet B et d'une application, également notée C, de A dans B^*, telle que $\ell(\mathrm{C}(a)) > 0$ pour tout a. Par cette application, les symboles de A sont représentés par des mots non vides dans l'alphabet B.

Un exemple représentatif consisterait à prendre pour A un ensemble de symboles assez vaste contenant, par exemple, les lettres de l'alphabet latin et les symboles de ponctuation, et pour C l'application qui à un tel symbole associe son *code* dans le système ASCII.

2.1.3. — Soit C un code sur l'alphabet A. En pratique, on ne code pas uniquement des symboles de l'alphabet A mais des mots dans cet alphabet : si $(a_1, \dots, a_n)$ est un mot, son code est le mot concaténé $C(a_1) \dots C(a_n)$. On notera C^*, ou parfois encore C, l'application de A^* dans B^* ainsi définie. Elle vérifie $C^*(ab) = C(a)C(b)$ pour tous a, b dans A^*.

On dit que le code C est *uniquement décodable* si cette application C^* est injective, c'est-à-dire si deux mots distincts ont des codes distincts.

Exemple **(2.1.4)**. — On dit qu'un code C sur un alphabet A est *préfixe* si, pour tous $a, b \in A$ tels que $a \neq b$, aucun des deux mots $C(a)$, $C(b)$ n'est le début de l'autre.

Démontrons qu'un tel code est uniquement décodable. Soit en effet $a = (a_1, \dots, a_n)$ et $b = (b_1, \dots, b_m)$ des mots dans l'alphabet A tels que $C^*(a) = C^*(b)$, c'est-à-dire $C(a_1) \dots C(a_n) = C(b_1) \dots C(b_m)$; démontrons que $a = b$.

Si $a = \varepsilon$, alors $C(a) = \varepsilon$, donc les mots $C(b_1), \dots, C(b_m)$ sont vides, ce qui impose $m = 0$. Ainsi $a = \varepsilon = b$. De même, si $b = \varepsilon$, alors $a = \varepsilon$.

Supposons maintenant $n \geqslant 1$, d'où $m \geqslant 1$ d'après ce qui précède. Posons $p = \ell(a_1)$ et supposons, quitte à échanger a et b, que $p \leqslant \ell(b_1)$. Par définition des mots $C(a)$ et $C(b)$, le mot de B^* formé des p premiers symboles de $C(a)$ est égal à $C(a_1)$. Comme $C(a_1)C(a_2) \dots C(a_n) = C(a) = C(b) = C(b_1) \dots C(b_m)$ et $\ell(b_1) \geqslant p$, c'est aussi le mot formé des p premiers symboles de $C(b_1)$. Par suite, $C(a_1)$ est le début du mot $C(b_1)$. Puisque C est un code préfixe, on a donc $a_1 = b_1$. Les mots formés à partir de $C(a)$ et $C(b)$ en enlevant les p premiers symboles sont aussi égaux ; on a donc $C(a_2) \dots C(a_n) = C(b_2) \dots C(b_m)$. Par récurrence, cela entraîne $m = n$ et $a_2 = b_2, \dots, a_n = b_n$.

Nous avons ainsi démontré que $a = b$.

Remarque **(2.1.5)**. — On dit également qu'un code préfixe est *instantanément décodable*. En effet, si $a \in A^*$, il n'est pas besoin de parcourir tout le mot $C^*(a)$ pour déterminer le premier symbole de a, il suffit de reconnaître, en tête de $C^*(a)$, l'un des mots $C(x)$, pour $x \in A$. Alors, x est le premier symbole de a, on peut écrire $a = xa'$, pour $a' \in A^*$, et il reste à décoder le mot $C^*(a')$ obtenu en enlevant de la tête de $C^*(a)$ le mot $C(x)$.

on obtient donc

$$\log(\mathrm{Card}(M)) \leqslant h(\pi) + \pi \log(\mathrm{Card}(M)) + I(W, W'),$$

d'où

$$(1 - \pi) \log(\mathrm{Card}(M)) \leqslant h(\pi) + I(W, W').$$

Dans la chaîne de variables aléatoires $W \to X \to Y \to W'$, W et Y sont indépendantes conditionnellement à X (le canal ne connaît pas le mot W d'où est issu X), et W et W′ sont conditionnellement indépendantes conditionnellement à Y (car W′ est certaine conditionnellement à Y). D'après le théorème 1.3.11, on a donc

$$I(W, W') \leqslant I(W, Y) = I(Y, W) \leqslant I(Y, X) = I(X, Y).$$

Par définition de la capacité d'un canal répété, on a enfin $I(X, Y) \leqslant nI(C)$, d'où l'inégalité $I(W, W') \leqslant nI(C)$.

On a donc

$$(1 - \pi) \log(\mathrm{Card}(M)) \leqslant h(\pi) + nI(C),$$

d'où l'inégalité

$$\tau(\Phi) = \frac{\log(\mathrm{Card}(M))}{n} \leqslant \frac{I(C) + h(\pi)/n}{1 - \pi}.$$

Appliquons cette inégalité à des codes de longueur arbitrairement grande (n tend vers $+\infty$) et dont la probabilité d'erreur est arbitrairement petite (π tend vers 0, donc $h(\pi)$ tend vers 0) ; le membre de droite de l'inégalité précédente tend vers $I(C)$, donc a limite supérieure des taux de transmission $\tau(\Phi)$ sera au plus égale à $I(C)$.

Cela prouve que tout taux de transmission atteignable par le canal C est inférieur ou égal à $I(C)$.

2.6.7. — Démontrons maintenant la partie « positive » du théorème de Shannon, c'est-à-dire que tout nombre réel ρ tel que $\rho < I(C)$ est atteignable. On fixe un entier $n \geqslant 1$ et un ensemble M de cardinal $\lceil \exp(n\rho) \rceil$; il s'agit de prouver qu'il existe, pourvu que n soit assez grand, un code de longueur n sur M adapté au canal C dont la probabilité maximale d'erreur est petite. On va commencer par prouver qu'il existe un tel code dont la probabilité moyenne d'erreur est petite, on verra ensuite comment en déduire un code de même longueur et de taux de transmission un peu plus faible.

Proposition **(2.6.5)** (Inégalité de Fano). — *Soit* X, Z *des variables aléatoires discrètes à valeurs dans un ensemble fini* A. *Posons* $\pi = \mathbf{P}(X \neq Z)$*; alors,*

$$H(X \mid Z) \leqslant h(\pi) + \pi \log(\mathrm{Card}(A) - 1).$$

Démonstration. — Soit U la variable aléatoire qui vaut 1 si $Z = X$ et 0 sinon. On a

$$H(X, U \mid Z) = H(X \mid Z) + H(U \mid X, Z) = H(U \mid Z) + H(X \mid U, Z).$$

Comme U est certaine conditionnée à (X, Z), on a $H(U \mid X, Z) = 0$. L'entropie décroit par conditionnement, donc

$$H(U \mid Z) \leqslant H(U) = h(\pi),$$

puisque U suit une loi de Bernoulli de paramètre π. Par définition de l'entropie conditionnelle, on a aussi

$$H(X \mid U, Z) = (1 - \pi)H(X \mid Z, U = 1) + \pi H(X \mid Z, U = 0).$$

Le premier terme est nul, car si $U = 1$, $X = Z$ est certaine conditionnellement à Z. Dans le second terme, le facteur $H(X \mid Z, U = 0)$ est majoré par l'entropie d'une variable aléatoire à valeurs dans A, donc est au plus égal à l'entropie $\log(\mathrm{Card}(A))$ d'une loi uniforme sur A. En fait, conditionné à l'événement $U = 0$, c'est-à-dire $X \neq Z$, cette variable aléatoire évite une valeur, donc son entropie est majorée par $\log(\mathrm{Card}(A) - 1)$. La proposition en résulte. □

2.6.6. — Commençons par démontrer que tout taux de transmission atteignable est $\leqslant I(C)$.

Soit donc Φ un code de longueur n sur un ensemble M pour le canal C ; soit f et g les applications de codage et de décodage. Soit W une variable aléatoire à valeurs dans M, de loi uniforme ; son entropie est $\log(\mathrm{Card}(M))$. Alors $X = f(W)$ est une variable aléatoire à valeurs dans A^n, transmise par le canal, et le mot Y reçu à l'autre extrémité du canal est une variable aléatoire à valeurs dans B^n. Le symbole décodé est alors $W' = g(Y)$, qu'il faut comparer à W. Posons $\pi = \mathbf{P}(W \neq W')$. La variable aléatoire W est uniforme dans l'ensemble M, donc $H(W) = \log(\mathrm{Card}(M))$. L'inégalité de Fano appliquée aux variables W, W′ entraîne aussi $H(W \mid W') \leqslant h(\pi) + \pi \log(\mathrm{Card}(M))$. En écrivant l'égalité

$$H(W) = H(W \mid W') + I(W, W'),$$

2.2. L'inégalité de Kraft–McMillan

Proposition **(2.2.1)** (Kraft, McMillan). — *Soit* A *un ensemble et soit* C *un code sur* A *à valeurs dans les mots sur un alphabet fini* B. *Posons* $D = \mathrm{Card}(B)$. *Si le code* C *est uniquement décodable, on a l'inégalité*

$$\sum_{a \in A} D^{-\ell(C(a))} \leqslant 1.$$

Démonstration. — Pour prouver l'inégalité, on peut supposer que l'ensemble A est fini. Soit N un entier tel que $N \geqslant \ell(C(a))$ pour tout $a \in A$.

Soit k un entier $\geqslant 1$. On a

$$\left(\sum_{a \in A} D^{-\ell(C(a))}\right)^k = \sum_{(a_1, \ldots, a_k) \in A^k} D^{-\ell(C(a_1))} \ldots D^{-\ell(C(a_k))} = \sum_{a \in A^k} D^{-\ell(C(a))}.$$

Pour tout entier m, soit c_m le nombre d'éléments $a \in A^k$ tels que $C(a)$ soit de longueur m. Par hypothèse, l'application de A^k dans B^* donnée par $(a_1, \ldots, a_k) \mapsto C(a_1) \ldots C(a_k)$ est injective. Ainsi, c_m est le nombre de mots de B^m de la forme $C(a)$, pour $a \in A^k$, si bien que $c_m \leqslant D^m$. On a aussi $c_m = 0$ si $m > kN$. Alors,

$$\sum_{a \in A^k} D^{-\ell(C(a))} = \sum_m c_m D^{-m} \leqslant \sum_{m=1}^{kN} D^m D^{-m} = kN.$$

Par suite,

$$\sum_{a \in A} D^{-\ell(C(a))} \leqslant (kN)^{1/k}.$$

Lorsqu'on fait tendre k vers $+\infty$, on obtient l'inégalité voulue. □

Théorème **(2.2.2)** (Shannon). — *Soit* X *une variable aléatoire discrète à valeurs dans un ensemble* A. *Soit* C *un code sur un alphabet* A, *à valeurs dans un alphabet fini* B *de cardinal* $D \geqslant 2$. *Si* C *est uniquement décodable, la longueur moyenne de* $C(X)$ *vérifie l'inégalité*

$$\mathbf{E}(\ell(C(X))) \geqslant H_D(X),$$

où $H_D(X)$ *est l'entropie en base* D *de* X. *Il y a égalité si et seulement si* $\ell(C(a)) = -\log_D(\mathbf{P}(X = a))$ *pour tout* $a \in A$.

Démonstration. — Déduisons l'inégalité de Shannon de l'inégalité de Kraft–McMillan. Par définition de la longueur moyenne de $C(X)$ et de l'entropie de X,

on a

$$\begin{aligned}\mathbf{E}(\ell(\mathrm{C}(\mathrm{X}))) - \mathrm{H}_{\mathrm{D}}(\mathrm{X}) &= \sum_{a\in\mathrm{A}} \mathbf{P}(\mathrm{X}=a)\ell(\mathrm{C}(a)) + \sum_{a\in\mathrm{A}} \mathbf{P}(\mathrm{X}=a)\log(\mathbf{P}(\mathrm{X}=a)) \\ &= -\sum_{a\in\mathrm{A}} \mathbf{P}(\mathrm{X}=a)\log_{\mathrm{D}}\left(\frac{\mathrm{D}^{-\ell(\mathrm{C}(a))}}{\mathbf{P}(\mathrm{X}=a)}\right).\end{aligned}$$

Comme la fonction logarithme est concave, on a

$$\begin{aligned}\sum_{a\in\mathrm{A}} \mathbf{P}(\mathrm{X}=a)\log_{\mathrm{D}}\left(\frac{\mathrm{D}^{-\ell(\mathrm{C}(a))}}{\mathbf{P}(\mathrm{X}=a)}\right) &\leqslant \log_{\mathrm{D}}\left(\sum_{a\in\mathrm{A}} \mathbf{P}(\mathrm{X}=a)\frac{\mathrm{D}^{-\ell(\mathrm{C}(a))}}{\mathbf{P}(\mathrm{X}=a)}\right) \\ &= \log_{\mathrm{D}}\left(\sum_{a\in\mathrm{A}} \mathrm{D}^{-\ell(\mathrm{C}(a))}\right).\end{aligned}$$

D'après l'inégalité de Kraft, l'argument du logarithme est $\leqslant 1$; puisque la fonction logarithme est croissante, on a donc

$$\sum_{a\in\mathrm{A}} \mathbf{P}(\mathrm{X}=a)\log_{\mathrm{D}}\left(\frac{\mathrm{D}^{-\ell(\mathrm{C}(a))}}{\mathbf{P}(\mathrm{X}=a)}\right) \leqslant 0.$$

Ainsi, $\mathbf{E}(\ell(\mathrm{C}(\mathrm{X}))) - \mathrm{H}_{\mathrm{D}}(\mathrm{X}) \geqslant 0$, d'où l'inégalité de Shannon. La fonction logarithme est strictement concave et strictement croissante. Pour qu'il y ait égalité, il faut donc, et il suffit, d'une part que $\sum \mathrm{D}^{-\ell(\mathrm{C}(a))} = 1$, et d'autre part que tous les termes $\mathrm{D}^{-\ell(\mathrm{C}(a))}/\mathbf{P}(\mathrm{X}=a)$ soient égaux. Puisque $\sum \mathbf{P}(\mathrm{X}=a) = 1$, cela signifie que $\mathrm{D}^{-\ell(\mathrm{C}(a))} = \mathbf{P}(\mathrm{X}=a)$ pour tout a, autrement dit, $\ell(\mathrm{C}(a)) = -\log_{\mathrm{D}}(\mathbf{P}(\mathrm{X}=a))$ pour tout a. □

2.2.3. Codes efficaces. — Soit A un alphabet et soit p une loi de probabilité sur A. Soit B un alphabet fini de cardinal $\mathrm{D} \geqslant 2$.

D'après l'inégalité de Shannon, on a $\mathbf{E}(\ell(\mathrm{C}(\mathrm{X}))) \geqslant \mathrm{H}_{\mathrm{D}}(\mathrm{X})$, pour toute variable aléatoire X à valeurs dans A : l'entropie fournit une limite irrémédiable à la compression d'un message. Nous allons maintenant voir que cette limite peut essentiellement être atteinte, qui plus est, par un code préfixe ! Nous commençons par une réciproque à l'inégalité de Kraft-McMillan.

Proposition **(2.2.4)**. — *Soit* A *un ensemble, soit* D *un entier* $\geqslant 1$ *et soit* $\ell : \mathrm{A} \to \mathbf{N}^*$ *une application telle que l'inégalité*

$$\sum_{a\in\mathrm{A}} \mathrm{D}^{-\ell(a)} \leqslant 1$$

Ce sont ces mots qui sont transmis par le canal avec bruit, puis décodés en des mots de M au moyen de l'application g.

Le *taux de transmission* d'un tel code est le quotient

$$\tau(\Phi) = \frac{\log(\mathrm{Card}(\mathrm{M}))}{n}.$$

La probabilité d'erreur lorsqu'on transmet un symbole $m \in \mathrm{M}$ est donnée par

$$\lambda_m(\Phi) = \mathbf{P}(g(\mathrm{Y}) \neq m \mid \mathrm{X} = f(m)),$$

où X et Y sont des variables aléatoires à valeurs dans A^n et B^n liées par les probabilités de transmission définies par le canal C. Comme il s'agit d'un canal sans mémoire, on a, si $f(m) = a_1 \dots a_n$,

$$\lambda_m(\Phi) = \sum_{\substack{b=(b_1,\dots,b_n)\in\mathrm{B}^n \\ g(b)\neq m}} \prod_{i=1}^{n} p(b_i \mid a_i).$$

Cela montre que ces probabilités d'erreur ne dépendent que des probabilités de transmission du canal C et pas du choix de variables aléatoires X et Y adaptées au canal.

On définit aussi la probabilité d'erreur *maximale* :

$$\lambda_{\max}(\Phi) = \sup_{m\in\mathrm{M}} \lambda_m(\Phi)$$

et la probabilité d'erreur *moyenne* :

$$\lambda_{\mathrm{moy}}(\Phi) = \frac{1}{\mathrm{Card}(\mathrm{M})} \sum_{m\in\mathrm{M}} \lambda_m(\Phi).$$

Définition **(2.6.3)**. — *Soit* C *un canal de transmission d'un alphabet* A *à un alphabet* B. *On dit qu'un nombre réel* ρ *est un taux de transmission* atteignable *par le canal* C *s'il existe, pour tout nombre réel* $\varepsilon > 0$ *et tout entier* n *assez grand, un code* Φ *de longueur* n *de taux de transmission* $\geqslant \rho$ *et de probabilité d'erreur maximale* $\leqslant \varepsilon$.

Théorème **(2.6.4)** (Shannon). — *Soit* C *un canal de transmission d'un alphabet* A *à un alphabet* B. *Tout taux de transmission atteignable par le canal* C *est inférieur ou égal à* $\mathrm{I}(\mathrm{C})$; *inversement, tout nombre réel* $\rho < \mathrm{I}(\mathrm{C})$ *est un taux de transmission atteignable par le canal* C.

Le premier terme se calcule par récurrence :

$$\begin{aligned}\mathrm{H}(\mathrm{Y}) &= \mathrm{H}(\mathrm{Y}_1) + \mathrm{H}(\mathrm{Y}_2 \mid \mathrm{Y}_1) + \cdots + \mathrm{H}(\mathrm{Y}_n \mid \mathrm{Y}_1, \ldots, \mathrm{Y}_{n-1}) \\ &\leqslant \mathrm{H}(\mathrm{Y}_1) + \mathrm{H}(\mathrm{Y}_2) + \cdots + \mathrm{H}(\mathrm{Y}_n),\end{aligned}$$

puisque l'entropie diminue par conditionnement ; on a même égalité lorsque les X_i sont indépendantes. Encore par récurrence, on a

$$\mathrm{H}(\mathrm{Y} \mid \mathrm{X}) = \mathrm{H}(\mathrm{Y}_1 \mid \mathrm{X}) + \mathrm{H}(\mathrm{Y}_2 \mid \mathrm{Y}_1, \mathrm{X}) + \cdots + \mathrm{H}(\mathrm{Y}_n \mid \mathrm{Y}_1, \ldots, \mathrm{Y}_{n-1}, \mathrm{X}).$$

Soit $p \in \{1, \ldots, n\}$. Par définition du canal C^n, la variable aléatoire Y_p est indépendante des Y_i et des X_i (pour $i \neq p$) conditionnellement à X_p ; ainsi,

$$\mathrm{H}(\mathrm{Y}_p \mid \mathrm{Y}_1, \ldots, \mathrm{Y}_{p-1}, \mathrm{X}) = \mathrm{H}(\mathrm{Y}_p \mid \mathrm{X}_p).$$

Ainsi, on a

$$\mathrm{I}(\mathrm{X}, \mathrm{Y}) \leqslant \sum_{p=1}^{n} \mathrm{H}(\mathrm{Y}_p) - \sum_{p=1}^{n} \mathrm{H}(\mathrm{Y}_p \mid \mathrm{X}_p) = \sum_{p=1}^{n} \mathrm{I}(\mathrm{X}_p, \mathrm{Y}_p),$$

avec égalité si les X_p sont indépendantes.

On obtient donc d'une part l'inégalité $\mathrm{I}(\mathrm{X}, \mathrm{Y}) \leqslant n\mathrm{I}(\mathrm{C})$, d'où $\mathrm{I}(\mathrm{C}^n) \leqslant n\mathrm{I}(\mathrm{C})$. D'autre part, si les X_p sont indépendantes et vérifient $\mathrm{I}(\mathrm{X}_p, \mathrm{Y}_p) = \mathrm{I}(\mathrm{C})$, on obtient $\mathrm{I}(\mathrm{X}, \mathrm{Y}) = n\mathrm{I}(\mathrm{C})$. Finalement, $\mathrm{I}(\mathrm{C}^n) = n\mathrm{I}(\mathrm{C})$.

2.6. Codage adapté à un canal avec bruit

2.6.1. — À moins qu'il ne soit en fait *sans* bruit, il n'est pas possible de transmettre, dans un canal avec bruit, un message avec certitude. Le théorème de Shannon (1948) que nous allons maintenant démontrer affirme que c'est toutefois possible de le transmettre de sorte que la probabilité d'erreur soit aussi petite que désirée, et que la vitesse de transmission n'est alors limitée que par la capacité du canal.

Pour cela, il va être nécessaire de préparer les messages qu'on envoie.

Définition **(2.6.2)**. — *Soit* C *un canal de transmission d'un alphabet* A *à un alphabet* B. *Soit* M *un ensemble fini ; un code* Φ *de longueur* n *sur* M *pour le canal* C *est la donnée de deux applications* $f : \mathrm{M} \to \mathrm{A}^n$ *et* $g : \mathrm{B}^n \to \mathrm{M}$.

Explicitement, les symboles qu'on souhaite transmettre sont ceux de l'ensemble M ; l'application f les code en des mots de longueur n sur l'alphabet A.

soit vérifiée. Il existe un code préfixe C *sur* A *à valeurs dans un alphabet de cardinal* D *tel que* $\ell(\mathrm{C}(a)) = \ell(a)$ *pour tout* $a \in \mathrm{A}$.

Démonstration. — Numérotons les éléments de A en une suite $a_1, a_2, \ldots$, de sorte que $\ell(a_1) \leqslant \ell(a_2) \leqslant \ldots$. C'est évidemment possible lorsque l'ensemble A est fini. Lorsqu'il est infini, on observe que pour tout entier n, l'ensemble des $a \in \mathrm{A}$ tels que $\ell(a) = n$ est fini, car sinon la somme $\sum_a \mathrm{D}^{-\ell(a)}$ serait infinie. Il suffit alors de numéroter d'abord les éléments a de A tels que $\ell(a) = 1$, puis ceux tels que $\ell(a) = 2$, etc.

On définit alors une suite strictement croissante de nombres rationnels en posant

$$z_n = \sum_{m<n} \mathrm{D}^{-\ell(a_m)},$$

pour tout entier n tel que $n \leqslant \mathrm{Card}(\mathrm{A})$. Puisque $\sum_{a \in \mathrm{A}} \mathrm{D}^{-\ell(a)} \leqslant 1$, on a $z_n \leqslant 1 - \mathrm{D}^{-\ell(a_n)} < 1$ pour tout entier $n \leqslant \mathrm{Card}(\mathrm{A})$. Considérons le développement en base D de z_n : il est de la forme $z_n = 0{,}y_1y_2\ldots y_p$, où l'entier p vérifie $p \leqslant \ell(a_{n-1})$. Associons alors au symbole a_n le code $\mathrm{C}(a_n) = y_1 \ldots y_p 0 \ldots 0$ dans l'alphabet $\{0; 1; \ldots; \mathrm{D}-1\}$, complété par $\ell(a_n) - p$ symboles 0 de sorte que $\ell(\mathrm{C}(a_n)) = \ell(a_n)$.

Soit m, n des entiers tels que $m < n \leqslant \mathrm{Card}(\mathrm{A})$. On a

$$z_n - z_m = \sum_{q=m}^{n-1} \mathrm{D}^{-\ell(a_q)} \geqslant \mathrm{D}^{-\ell(a_m)}.$$

Par suite, les développements en base D de z_m et z_n diffèrent au moins par le $\ell(a_m)$-ième chiffre, de sorte que $\mathrm{C}(a_m)$ n'est pas préfixe de $\mathrm{C}(a_n)$. Mais $\mathrm{C}(a_n)$ n'est pas non plus préfixe de $\mathrm{C}(a_m)$: c'est évident si $\ell(a_n) > \ell(a_m)$, et dans le cas où $\ell(a_m) = \ell(a_n)$, cela signifierait que $\mathrm{C}(a_m) = \mathrm{C}(a_n)$.

L'application $\mathrm{C} : \mathrm{A} \to \{0; \ldots; \mathrm{D}-1\}^*$ est un code préfixe tel que $\ell(\mathrm{C}(a)) = \ell(a)$ pour tout a. □

Exemple **(2.2.5)**. — Supposons que $\mathrm{A} = \{a, b, c, d, e\}$ et supposons que l'on ait $\ell(a) = \ell(b) = 1$, $\ell(c) = 2$, $\ell(d) = \ell(e) = 3$; on numérote A par $a_1 = a, a_2 = b, \ldots, a_5 = e$. Prenons $\mathrm{D} = 3$. Les nombres réels $z_1, \ldots, z_5$ construits par la preuve de la proposition sont (en base 3) $z_1 = 0$, $z_2 = 0{,}1$, $z_3 = 0{,}2$, $z_4 = 0{,}21$ et $z_5 = 0{,}211$. On pose alors $\mathrm{C}(a) = 0$, $\mathrm{C}(b) = 1$, $\mathrm{C}(c) = 20$, $\mathrm{C}(d) = 210$ et $\mathrm{C}(e) = 211$.

Théorème **(2.2.6)** (Shannon). — *Soit* X *une variable aléatoire discrète à valeurs dans un alphabet* A. *Soit* B *un ensemble fini de cardinal* $D \geqslant 2$. *Il existe un code préfixe* C *sur* A, *à valeurs dans* B, *tel que*

$$H_D(X) \leqslant \mathbf{E}(\ell(C(X))) < H_D(X) + 1.$$

Pour qu'il existe un tel code C *vérifiant l'égalité* $\mathbf{E}(\ell(C(X))) = H_D(X)$, *il faut et il suffit que pour tout élément* $a \in A$, $\mathbf{P}(X = a)$ *soit de la forme* D^{-m} *pour un entier* $m \geqslant 0$; *on a alors* $\ell(C(a)) = -\log_D(\mathbf{P}(X = a))$.

Démonstration. — Pour tout $a \in A$ tel que $\mathbf{P}(X = a) > 0$, posons $\lambda(a) = \lceil -\log_D(\mathbf{P}(X = a)) \rceil$, le plus petit entier supérieur ou égal à $\log_D(\mathbf{P}(X = a))$, de sorte que $D^{-\lambda(a)} \leqslant \mathbf{P}(X = a)$. On a donc $\sum_{a\in A} D^{-\lambda(a)} \leqslant \sum_{a\in A} \mathbf{P}(X = a) = 1$. D'après la proposition 2.2.4, il existe ainsi un code préfixe $C : A \to \{0; 1; \dots, D-1\}^*$ tel que $\ell(C(a)) = \lambda(a)$ pour tout $a \in A$. Démontrons que ce code vérifie la conclusion du théorème.

L'inégalité $H_D(X) \leqslant \mathbf{E}(\ell(C(X)))$ est un cas particulier du théorème 2.2.2. D'autre part, on a

$$\begin{aligned}\mathbf{E}(\ell(C(X))) &= \sum_{a\in A} \mathbf{P}(X = a)\ell(C(a)) = \sum_{a\in A} \mathbf{P}(X = a)\lceil \log_D(\mathbf{P}(X = a)) \rceil \\ &< \sum_{a\in A} \mathbf{P}(X = a)\,(\log_D(\mathbf{P}(X = a)) + 1) \\ &= H_D(X) + \sum_{a\in A} \mathbf{P}(X = a) = H_D(X) + 1.\end{aligned}$$

Cela conclut la démonstration.

La dernière assertion résulte de cela et du théorème 2.2.2 : d'après ce théorème, le cas d'égalité se produit si et seulement si $\ell(C(a)) = -\log_D(\mathbf{P}(X = a))$ pour tout $a \in A$. D'autre part, si $\mathbf{P}(X = a)$ est de la forme D^{-m}, on a $\lambda(a) = -\log_D(\mathbf{P}(X = a))$ et le code construit vérifie $\mathbf{E}(\ell(C(X)) = H_D(X)$. □

2.3. Codes optimaux

2.3.1. — Bien que sa longueur moyenne soit proche de l'optimum (limité par l'entropie), le code préfixe construit par la méthode de la preuve du théorème 2.2.6 n'est pas toujours optimal. Par exemple, lorsque X suit une loi de Bernoulli de paramètre $p \in]0; 1[$ et que l'alphabet-but a deux symboles, les mots codés ont longueur $\lceil -\log_2(p) \rceil$ et $\lceil -\log_2(1-p) \rceil$, alors qu'on pourrait se contenter de recopier X ! Ce code trivial a longueur moyenne 1, alors que la longueur moyenne

(Si on n'avait pas su faire cette constatation, il restait à calculer u en fonction de t : on trouve $u = (t-p)/(1-2p) = 1/2$ si $t = 1/2$.)

c) Un canal est dit *faiblement symétrique* si les lignes de sa matrice de probabilités de transmission diffèrent uniquement l'une de l'autre par des permutations et si la somme des coefficients de chaque colonne est constante. On parle de canal *symétrique* si, de plus, les colonnes de sa matrice de probabilités de transmission diffèrent par permutations l'une de l'autre. C'est le cas des deux canaux précédents. Une façon d'obtenir un canal symétrique consiste à prendre pour alphabets $A = B = \mathbf{Z}/d\mathbf{Z}$ et à poser $p(b \mid a) = q(b-a)$, où q est une loi de probabilité sur A. On peut bien sûr remplacer $\mathbf{Z}/d\mathbf{Z}$ par un groupe fini arbitraire.

Soit C un tel canal et soit X, Y des variables aléatoires sur A, B respectivement, liées par la condition $\mathbf{P}(Y = b) = p(b \mid a)\mathbf{P}(Y = a)$. Pour tout $a \in A$,

$$H(Y \mid X = a) = -\sum_b p(b \mid a)\log(p(b \mid a)),$$

expression indépendante de a par la condition de symétrie des lignes de la matrice de probabilités de transmission du canal. D'autre part,

$$H(Y) \leqslant \log(\mathrm{Card}(B)).$$

Lorsque la loi de X est uniforme, la condition sur la somme des coefficients de chaque colonne entraîne que la loi de Y est également uniforme : pour tout $b \in B$, on a en effet

$$\mathbf{P}(Y = b) = \sum_{a\in A} \mathbf{P}(Y = b \mid X = a)\mathbf{P}(X = a) = \frac{1}{\mathrm{Card}(A)} \sum_{a\in A} p(b \mid a),$$

expression indépendante de b. Dans ce cas, on a donc $H(Y) = \log(\mathrm{Card}(B))$. Puique $I(X, Y) = H(Y) - H(Y \mid X)$, ce qui précède entraîne ainsi la formule

$$I(C) = \log(\mathrm{Card}(B)) - H(Y \mid X = a),$$

où a est un élément quelconque de A.

d) Considérons un canal C d'un alphabet A à un alphabet B, soit n un entier $\geqslant 2$ et définissons un canal C^n de l'alphabet A^n à l'alphabet B^n de probabilités de transmission d'un canal sans mémoire : $p(b \mid a) = \prod_{i=1}^n p(b_i \mid a_i)$, pour $a = (a_1, \dots, a_n) \in A^n$ et $b = (b_1, \dots, b_n) \in B^n$. Démontrons que $I(C^n) = nI(C)$.

Soit $X = (X_1, \dots, X_n)$ et $Y = (Y_1, \dots, Y_n)$ des variables aléatoires à valeurs dans A^n et B^n respectivement vérifiant $\mathbf{P}(Y = b \mid X = a) = p(b \mid a)$ pour $a \in A^n$ et $b \in B^n$. On a

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

Soit E la variable aléatoire qui vaut 1 si $Y = e$ et 0 sinon. Puisque E est conséquence de Y, on a $H(Y) = H(Y, E)$; alors,

$$H(Y) = H(Y, E) = H(E) + H(Y \mid E).$$

D'autre part, $\mathbf{P}(E = 1) = q$ et $\mathbf{P}(E = 0) = 1 - q$, de sorte que $H(E) = h(q)$. De plus, conditionnée à l'événement $Y = e$, la variable aléatoire Y est certaine, donc d'entropie nulle ; conditionnée à l'événement complémentaire, elle a pour loi une loi de Bernoulli de paramètre t, de sorte que

$$H(Y \mid E) = qH(Y \mid Y = e) + (1 - q)H(Y \mid Y \neq e) == (1 - q)h(t),$$

de sorte que

$$H(Y) = h(q) + (1 - q)h(t).$$

On applique un argument similaire pour le terme $H(Y \mid X)$. On a tout d'abord

$$H(Y \mid X) = H(Y, E \mid X) = H(E \mid X) + H(Y \mid E, X).$$

Le premier terme est de nouveau égal à $h(q)$. Conditionnée à l'événement $Y = e$, de probabilité q, la variable aléatoire Y est certaine ; conditionnée à l'événement $(E = 0) \cap (X = 0)$, de probabilité $(1 - q)u$, elle se comporte comme une loi de Bernoulli de paramètre p, de même que conditionnée à l'événement $(E = 0) \cap (X = 1)$ qui est de probabilité $(1 - q)(1 - u)$. Ainsi,

$$\begin{aligned} H(Y \mid E, X) &= qH(Y \mid E = 1) + (1 - q)uH(Y \mid E = 0, X = 0) \\ &\quad + (1 - q)(1 - u)H(Y \mid E = 0, X = 1) \\ &= q \cdot 0 + (1 - q)u \cdot h(p) + (1 - q)(1 - u) \cdot h(p) \\ &= (1 - q)h(p). \end{aligned}$$

On a donc

$$H(Y \mid X) = h(q) + (1 - q)h(p).$$

Finalement,

$$I(X, Y) = (1 - q)(h(t) - h(p)).$$

Cette expression est maximale lorsque $h(t)$ est maximale. Lorsque la base des logarithmes est 2, on a $h(t) \leqslant 1$, de sorte que $I(X, Y) \leqslant (1 - q)(1 - h(p))$. On a aussi $h(t) = 1$ pour $t = 1/2$. Or, rappelons que $t = u(1-p) + (1-u)p$; on constate que pour $u = 1/2$, on a également $t = 1/2$, d'où

$$I(C) = (1 - q)(1 - h(p)) \text{ bits.}$$

du code proposé par Shannon est égale à $S(p) = p\lceil -\log_2(p)\rceil + (1-p)\lceil -\log_2(1-p)\rceil$; il n'y a égalité que si $p = 1/2$, qui est d'ailleurs le seul cas où les probabilités p et $1 - p$ sont toutes deux de la forme 2^{-m}. En fait, lorsque $2^{-m-1} \leqslant p < 2^{-m}$, avec $m \geqslant 1$, on a $\lceil -\log_2(p)\rceil = m + 1$, tandis que $1 - p > 1/2$, de sorte que $\lceil -\log_2(1-p)\rceil = 1$. Alors, $S(p) = p(m + 1) + (1 - p) = mp + 1 < 1 + m/2^m < 1{,}5$. Et lorsque $p \to 1/2$ par valeurs inférieures, $S(p)$ converge vers 1,5.

Définition **(2.3.2)**. — *Soit* C *un code uniquement décodable sur un alphabet* A *à valeurs dans un un alphabet* B. *Soit* p *une loi de probabilité sur* A. *On dit que* C *est optimal (par rapport à* p*) si le code* C *minimise l'expression* $\sum_{a\in A} p(a)\ell(C(a))$ *parmi tous les codes uniquement décodables à valeurs dans l'alphabet* B.

L'expression $\ell_p(C) = \sum_{a\in A} p(a)\ell(C(a))$ est la valeur moyenne du code d'un symbole de A, lorsque ces symboles sont pris avec la loi p.

Lemme **(2.3.3)**. — *Soit* A *un alphabet fini et soit* p *une loi de probabilité sur* A. *Soit* B *un alphabet fini*

a) *Il existe un code préfixe sur* A *à valeurs dans* B *qui est optimal.*

b) *Si* C *est un code optimal, et si* a, b *sont des éléments de* A *tels que* $p(a) < p(b)$, *alors* $\ell(C(a)) \geqslant \ell(C(b))$ *— les symboles moins probables ont des codes plus longs ;*

c) *Si* C *est un code préfixe qui est optimal, alors pour chaque symbole* $a \in A$ *tel que* $C(a)$ *soit de longueur maximale, il existe* $b \in A$ *tel que* $C(b)$ *soit de même longueur que* $C(a)$ *et en diffère uniquement par le dernier symbole.*

Démonstration. — a) Les éléments $a \in A$ tels que $p(a) = 0$ n'interviennent pas dans la définition de la longueur moyenne d'un code ; quitte à remplacer A par l'ensemble des éléments a tels que $p(a) > 0$, on suppose que $p(a) > 0$ pour tout $a \in A$.

Soit C_1 un code uniquement décodable et soit C un autre code uniquement décodable tel que $\ell_p(C) \leqslant \ell_p(C_1)$. On a en particulier $p(a)\ell(C(a)) < \ell_p(C_1)$ pour tout $a \in A$, donc $\ell(C(a)) < \ell_p(C_1)/p(a)$. Ainsi, le code C est une application de A dans l'ensemble fini des mots de longueurs $< \ell_p(C_1)/\inf(p)$. L'ensemble de ces applications étant fini, il n'y a qu'un nombre fini de codes uniquement décodables de longueur moyenne inférieure ou égale à celle de C_1. On trouvera dans cet ensemble fini un code uniquement décodable de longueur moyenne minimale, c'est-à-dire optimal.

Soit alors C un code optimal. Les longueurs $\ell(C(a))$ vérifient l'inégalité de Kraft $\sum D^{-\ell(C(a))} \leqslant 1$. Il existe alors un code *préfixe* C′ tel que $\ell(C'(a)) = \ell(C(a))$ pour tout $a \in A$. Le code C′ a même longueur moyenne que C, donc est un code optimal.

b) Soit C un code optimal et soit a, b des éléments de A tels que $p(a) < p(b)$ et $\ell(C(a)) < \ell(C(b))$. Considérons le code C′ qui coïncide avec C sur $A - \{a, b\}$ mais qui échange les codes de a et b : $C'(a) = C(b)$ et $C'(b) = C(a)$. On a

$$\begin{aligned}
\ell_p(C) - \ell_p(C') &= \sum_{x\in A} p(x)\ell(C(x)) - \sum_{x\in A} \ell(C'(x)) \\
&= p(a)\ell(C(a)) + p(b)\ell(C(b)) - p(a)\ell(C'(a)) - p(b)\ell(C'(b)) \\
&= p(a)\ell(C(a)) + p(b)\ell(C(b)) - p(a)\ell(C(b)) - p(b)\ell(C(a)) \\
&= (p(a) - p(b))(\ell(C(a)) - \ell(C(b))) \\
&> 0.
\end{aligned}$$

Cela contredit l'hypothèse que C est un code optimal.

c) Soit C un code préfixe optimal. Soit a un élément de A dont le code est de longueur maximale ; écrivons $C(a) = mx$, où $m \in B^*$ et $x \in B$. Supposons que m n'est pas préfixe d'un mot de code. Soit alors C′ le code qui coïncide avec C sur $A - \{a\}$ et tel que $C'(a) = m$. C'est encore un code préfixe : $C'(a)$ n'est pas hypothèse pas préfixe d'un autre mot, et un autre mot, disons $C'(b) = C(b)$, ne peut être préfixe de $C'(a)$, puisqu'il serait alors préfixe de $C(a)$. En particulier, le code C′ est uniquement décodable, mais sa longueur moyenne est strictement plus petite que celle de C, ce qui contredit l'hypothèse que C est optimal. Donc m est préfixe d'un autre mot de code, disons $C(b)$; écrivons $C(b) = mp$, avec $p \in B^*$. Comme $C(a) = mx$ est de longueur maximale, égale à $\ell(m) + 1$, on a $\ell(p) = \ell(C(b)) - \ell(m) \leqslant \ell(C(a)) - \ell(m) = 1$, c'est-à-dire que p est soit le mot vide, soit réduit à un seul symbole. Si p est vide, $C(b) = m$ est préfixe de $C(a) = mx$, ce qui contredit l'hypothèse que C est un code préfixe. Il existe donc $y \in B$ tel que $p = (y)$. Les mots $C(a) = mx$ et $C(b) = my$ diffèrent donc uniquement par leur dernier symbole. □

2.3.4. Code de Huffman. — Soit A un ensemble fini et soit $(p(a))_{a\in A}$ une loi de probabilité sur A. Lorsque l'alphabet d'arrivée B est $\{0, 1\}$, le code de Huffman est construit explicitement par récurrence sur le cardinal de A, de la façon suivante.

Sa matrice de probabilités de transmission est donc

$$\begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}.$$

Soit X et Y des variables aléatoires sur A et B respectivement telles que $\mathbf{P}(Y = b \mid X = a) = p(b \mid a)$ pour $a \in A$ et $b \in B$. Notons $u = \mathbf{P}(X = 0)$; on a donc $\mathbf{P}(X = 1) = 1 - u$, puis

$$\begin{aligned}
H(Y \mid X) &= uH(Y \mid X = 0) + (1-u)H(Y \mid X = 1) \\
&= uh(p) + (1-u)h(1-p) = h(p).
\end{aligned}$$

D'autre part, Y étant une variable aléatoire binaire, on a $H(Y) \leqslant \log(2)$. Par suite,

$$I(X, Y) \leqslant \log(2) - h(p).$$

Lorsque X suit une loi uniforme ($u = 1/2$), il en est de même de la loi Y — on peut le justifier par symétrie des probabilités, ou bien faire le calcul. Cela montre que la capacité de ce canal est $\log(2) - h(p)$. À ce stade, il est vraiment plus pratique de fixer à 2 la base des logarithmes, ce qui mesure les entropies en *bits*. Alors, la capacité du canal symétrique à bruit est $1 - h(p)$.

Lorsque $p = 0$, la capacité de ce canal est 1 ; lorsque $p = 1/2$, elle est nulle. Lorsque $p = 1$, on trouve encore $I(C) = 1$: ce canal modifie *systématiquement* le symbole reçu — il n'y a en fait pas de perte d'information !

b) Une variante du canal précédent utilise l'alphabet $A = \{0, 1\}$ mais a pour but l'alphabet $B = \{0, 1, e\}$, où e est un symbole auxiliaire indiquant une erreur de transmission, commise avec probabilité q, tandis que la probabilité de transmettre un symbole erroné est $(1-q)p$. On a $p(1 \mid 0) = p(0 \mid 0) = (1-q)q$ et $p(e \mid 1) = p(e \mid 0) = q$. La matrice de probabilités de transmission de ce canal est

$$\begin{pmatrix} (1-q)(1-p) & (1-q)p & q \\ (1-q)p & (1-q)(1-p) & q \end{pmatrix}.$$

Notons encore $\mathbf{P}(X = 0) = u$. Alors, en posant $t = u(1-p) + (1-u)p$, on a

$$\begin{aligned}
\mathbf{P}(Y = 0) &= u(1-q)(1-p) + (1-u)(1-q)p = (1-q)t, \\
\mathbf{P}(Y = 1) &= (1-u)(1-q)(1-p) + u(1-q)p = (1-q)(1-t) \\
\mathbf{P}(Y = e) &= q.
\end{aligned}$$

du nombre d'erreurs prévues. Cela n'est pas satisfaisant, puisqu'on ne tient pas compte du fait que le destinataire ne sait pas où se trouvent les erreurs. Prenons le cas extrême où le bruit est si grand que les symboles reçus sont entièrement indépendants des symboles transmis. La probabilité de recevoir 1 est 1/2 quel que soit ce qui est transmis, et de même pour 0. Alors, environ la moitié des symboles reçus sont corrects, du seul fait du hasard, et on pourrait de la même façon dire que le système transmet 500 bits par seconde, alors qu'aucune information n'est transmise en réalité. »

Shannon continue : « *La correction adéquate à appliquer à la quantité d'information transmise est évidemment la quantité de cette information qui est manquante dans le signal reçu, ou, ce qui revient au même, l'incertitude lors de la réception du signal sur ce qui a été réellement émis. Vu notre discussion antérieure, il semble raisonnable d'utiliser l'entropie conditionnelle du message connaissant le signal reçu comme mesure de cette information manquante.* » C'est ce que dit la troisième égalité dans la formule précédente : on obtient l'information mutuelle $\mathrm{I}(\mathrm{X},\mathrm{Y})$ en partant de la quantité d'information dans le message envoyé X, mesurée par son entropie $\mathrm{H}(\mathrm{X})$, et en lui soustrayant l'incertitude $\mathrm{H}(\mathrm{X} \mid \mathrm{Y})$ que mesure l'entropie de X conditionnellement à Y. La seconde égalité présente cette quantité comme la quantité d'information $\mathrm{H}(\mathrm{Y})$ du message reçu minorée du bruit que mesure l'entropie conditionnelle $\mathrm{H}(\mathrm{Y} \mid \mathrm{X})$.

Le fait de prendre la borne supérieure sur toutes les lois possibles sur les symboles X reflète la possibilité pour l'émetteur d'*adapter* la façon dont il écrit le message pour tenir compte des problèmes du canal. Par exemple, si tous les symboles étaient envoyés sans erreur, sauf un qui était systématiquement corrompu, il serait malin d'utiliser un codage qui permette de ne jamais l'utiliser.

Continuons avec quelques exemples de capacité de transmission.

Exemple **(2.5.3)**. — a) Prenons pour alphabets $\mathrm{A} = \mathrm{B} = \{0,1\}$; soit p un élément de $[0;1]$. Le canal avec bruit de paramètre p transmet le mauvais symbole avec probabilité p. Lorsque $p = 1$, ce canal retransmet exactement le symbol émis : on parle de canal *sans bruit*. Lorsque $p = 1/2$, ce canal envoie chaque symbole avec probabilité $1/2$, indépendamment du symbole émis ; on comprend bien, et on verra, qu'un tel canal n'est pas très utile.

Si $\mathrm{Card}(\mathrm{A}) = 2$, le code H_p associe aux deux mots de A les mots 0 et 1, tous deux de longueur 1. Supposons $\mathrm{Card}(\mathrm{A}) > 2$ et soit a, b deux éléments de A minimisant p : explicitement, on a $p(a), p(b) \leqslant \inf_{c \neq a,b} p(c)$. Soit A' la réunion de l'ensemble $\mathrm{A} \smallsetminus \{a,b\}$ et d'un élément auxiliaire noté ab ; on définit une loi de probabilité sur A' par $p'(c) = p(c)$ si $c \in \mathrm{A} \smallsetminus \{a,b\}$, et $p(ab) = p(a) + p(b)$. Soit $\mathrm{H}_{p'}$ le code de Huffman associé par récurrence à la loi p' sur A'. Le code H_p associe à un symbole $c \in \mathrm{A} \smallsetminus \{a,b\}$ le mot $\mathrm{H}_{p'}(c)$; si m est le code $\mathrm{H}_{p'}(ab)$, on pose $\mathrm{H}_p(a) = m0$ et $\mathrm{H}_p(b) = m1$.

Exemple **(2.3.5)**. — Supposons $\mathrm{A} = a, b, c, d, e$, avec les probabilités données par le tableau

a	b	c	d	e
0,25	0,25	0,20	0,15	0,15

La méthode commence par combiner d et e et leur associer la probabilité $p'(de) = 0,30$, les autres symboles étant a, b, c, avec leurs probabilités initiales, d'où le tableau

a	b	c	de
0,25	0,25	0,20	0,30

Puis elle combine, disons a et c et leur associe la probabilité $p''(ac) = 0,45$, les autres symboles étant b, de de probabilités 0,25 et 0,30 :

ac	b	de
0,45	0,25	0,30

Ensuite, elle combine b et de, d'où les deux symboles bde et ac, avec probabilités 0,55 et 0,45.

ac	bde
0,45	0,55

On parcourt maintenant le chemin en sens inverse. À la dernière étape, on code ac par 0, bde par 1. À l'avant dernière, on code ac par 0, b par 10 et de par 11. Puis on code c par 00, a par 01, b par 10 et de par 11. Finalement, le code

obtenu est

a	b	c	d	e
01	10	00	110	111

L'entropie (en base 2) d'une variable aléatoire X de loi p est égale à

$$\mathrm{H}_2(\mathrm{X}) = -2 \cdot 0{,}25 \log_2(0{,}25) - 0{,}20 \log_2(0{,}20) - 2 \cdot 0{,}15 \log_2(0{,}15) \approx 2{,}285.$$

La longueur moyenne du code ci-dessus est alors

$$\mathbf{E}(\ell_{\mathrm{H}}(\mathrm{X})) = (2 \cdot 0{,}25 + 0{,}20) \cdot 2 + (2 \cdot 0{,}15) \cdot 3 = 2{,}3.$$

Pour comparaison, les longueurs du code construit par la méthode de Shannon sont 2, 2, 3, 3, 3, de sorte que sa longueur moyenne est

$$\mathbf{E}(\ell_{\mathrm{S}}(\mathrm{X})) = (2 \cdot 0{,}25) \cdot 2 + (0{,}20) + 2 \cdot 0{,}15) \cdot 3 = 2{,}5.$$

Voici d'ailleurs un tel code :

a	b	c	d	e
00	01	100	110	101

Proposition **(2.3.6)** (D. Huffman, 1952). — *Soit* A *un ensemble fini et soit* p *une loi de probabilité sur* A. *Le code de Huffman* H_p *est un code préfixe; il est optimal (relativement à la loi* p*).*

Démonstration. — Prouvons par récurrence sur le cardinal de A que le code H_p est un code préfixe. C'est évident si $\mathrm{Card}(\mathrm{A}) \leqslant 2$; supposons maintenant $\mathrm{Card}(\mathrm{A}) \geqslant 3$. Reprenons les notations de la construction : a, b sont deux éléments de A de probabilités minimales, l'ensemble A' est la réunion de $\mathrm{A} - \{a, b\}$ et d'un symbole ab, et la loi de probabilité p' sur A' attache à tout $c \neq a, b$ la probabilité qu'il avait pour p, et à ab la somme des probabilités de a et b.

Par récurrence, le code H' associé à A' et à la loi p' est un code préfixe. Les mots de H sont les $\mathrm{H}(c) = \mathrm{H}'(c)$, pour $c \neq a, b$, et les deux mots $\mathrm{H}(a) = \mathrm{H}'(ab)0$ et $\mathrm{H}(b) = \mathrm{H}'(ab)1$. Par récurrence, $\mathrm{H}(c)$ et $\mathrm{H}(c')$ ne sont pas préfixes l'un de l'autre si c, c' sont des éléments distincts, distincts de a, b. Les mots $\mathrm{H}(a) = \mathrm{H}'(ab)0$ et $\mathrm{H}(b) = \mathrm{H}'(ab)1$ ne sont pas préfixes l'un de l'autre, puisqu'ils ont même longueur et sont distincts. Pour $c \neq a, b$, le mot $\mathrm{H}(c) = \mathrm{H}'(c)$ n'est pas préfixe du mot $\mathrm{H}'(ab)$, par l'hypothèse de récurrence. S'il est préfixe de l'un des mots $\mathrm{H}'(ab)0$ ou $\mathrm{H}'(ab)1$, c'est qu'il leur est égal, mais alors $\mathrm{H}'(ab)$ est préfixe de $\mathrm{H}'(c)$, une

Définition **(2.5.1)**. — *Soit* A *et* B *des alphabets. Un* canal de transmission sans mémoire *de l'alphabet* A *à l'alphabet* B *est donné par une famille* $(p(\cdot \mid a))$ *de lois de probabilités sur* B, *indexée par l'ensemble* A.

Pour $a \in \mathrm{A}$ et $b \in \mathrm{B}$, $p(b \mid a)$ est la probabilité que le canal transmette le symbole b sachant que le symbole émis était a. La matrice $(p(b \mid a))$ de type $\mathrm{A} \times \mathrm{B}$ est la *matrice de probabilités de transmission* du canal.

Le fait que le canal soit *sans mémoire* signifie que la transmission des symboles d'un mot $(a_1, \ldots, a_n)$ est faite symbole par symbole, de façon indépendante. Autrement dit, la probabilité que le mot $(a_1, \ldots, a_n)$ soit transmis en $(b_1, \ldots, b_n)$ est donnée par

$$\mathbf{P}(b_1 \ldots b_n \mid a_1 \ldots a_n) = p(b_1 \mid a_1) \ldots p(b_n \mid a_n).$$

Définition **(2.5.2)**. — *Soit* C *un canal de matrice de probabilités de transmission* $(p(b \mid a))$. *On appelle* capacité de transmission *de ce canal l'expression*

$$\mathrm{I}(\mathrm{C}) = \sup \mathrm{I}(\mathrm{X}, \mathrm{Y})$$

où X *parcourt l'ensemble des variables aléatoires sur* A *et* Y *parcourt l'ensemble des variables aléatoires sur* B *telles que* $\mathbf{P}(\mathrm{Y} = b \mid \mathrm{X} = a) = p(b \mid a)$ *pour tout* $a \in \mathrm{A}$ *et tout* $b \in \mathrm{B}$.

Rappelons que l'on a

$$\mathrm{I}(\mathrm{X}, \mathrm{Y}) = \mathrm{H}(\mathrm{X}) + \mathrm{H}(\mathrm{Y}) - \mathrm{H}(\mathrm{X}, \mathrm{Y}) = \mathrm{H}(\mathrm{Y}) - \mathrm{H}(\mathrm{Y} \mid \mathrm{X}) = \mathrm{H}(\mathrm{X}) - \mathrm{H}(\mathrm{X} \mid \mathrm{Y}).$$

Pour tout couple (X, Y), on a $0 \leqslant \mathrm{I}(\mathrm{X}, \mathrm{Y}) \leqslant \max(\log(\mathrm{Card}(\mathrm{A})), \log(\mathrm{Card}(\mathrm{B})))$, de sorte que

$$0 \leqslant \mathrm{I}(\mathrm{C}) \leqslant \max(\log(\mathrm{Card}(\mathrm{A})), \log(\mathrm{Card}(\mathrm{B}))).$$

Expliquons tout d'abord cette définition en reprenant, comme Shannon (1948, §12), le cas où il n'y a que deux symboles 0 et 1 à transmettre « *à un débit de 1000 symboles par seconde avec les probabilités* $p_0 = p_1 = 1/2$. *Ainsi,* continue Shannon, *notre source produit de l'information avec un débit de 1000 bits par seconde. Lors de la transmission, le bruit introduit des erreurs, de sorte que, en moyenne, un symbole sur 100 est reçu incorrectement(0 pour 1, ou 1 pour 0). Quel est le débit d'information transmis? Certainement moins de 1000 bits par seconde puisque 1% environ des symboles reçus sont incorrects. Notre première réaction pourrait être de dire que ce débit est de 990 bits par seconde, par simple soustraction*

d'où la minoration voulue pour $\mathbf{P}(\mathrm{C}_\varepsilon^n)$. On prouve aussi, par le même argument que

$$\left(1 - \frac{c(\mathrm{X}) + c(\mathrm{Y}) + c(\mathrm{Z})}{2n\varepsilon^2}\right) e^{n(\mathrm{H}(\mathrm{X}_1,\mathrm{Y}_1)-\varepsilon)} \leqslant \mathrm{Card}(\mathrm{C}_\varepsilon^n) \leqslant e^{n(\mathrm{H}(\mathrm{X}_1,\mathrm{Y}_1)+\varepsilon)}.$$

Considérons alors des variables aléatoires X', Y', indépendantes et de mêmes lois que X, Y, et posons $\mathrm{Z}' = (\mathrm{X}', \mathrm{Y}')$. On a

$$\begin{aligned}\mathbf{P}(\mathrm{Z}' \in \mathrm{C}_\varepsilon^n) &= \sum_{(a,b)\in \mathrm{C}_\varepsilon^n} \mathbf{P}(\mathrm{Z}' = (a,b)) \\ &= \sum_{(a,b)\in \mathrm{C}_\varepsilon^n} \mathbf{P}(\mathrm{X}' = a)\mathbf{P}(\mathrm{Y}' = b) \\ &= \sum_{(a,b)\in \mathrm{C}_\varepsilon^n} \mathbf{P}(\mathrm{X} = a)\mathbf{P}(\mathrm{Y} = b).\end{aligned}$$

Par définition de C_ε^n et la majoration de $\mathrm{Card}(\mathrm{C}_\varepsilon^n)$, on a donc

$$\begin{aligned}\mathbf{P}(\mathrm{Z}' \in \mathrm{C}_\varepsilon^n) &\leqslant \mathrm{Card}(\mathrm{C}_\varepsilon^n) e^{-n(\mathrm{H}(\mathrm{X}_1)-\varepsilon)} e^{-n(\mathrm{H}(\mathrm{Y}_1)-\varepsilon)} \\ &\leqslant e^{-n(\mathrm{H}(\mathrm{X}_1)+\mathrm{H}(\mathrm{Y}_1)-\mathrm{H}(\mathrm{X}_1,\mathrm{Y}_1)-3\varepsilon)} \\ &= e^{-n(\mathrm{I}(\mathrm{X}_1,\mathrm{Y}_1)-3\varepsilon)}.\end{aligned}$$

La preuve de la minoration est analogue :

$$\begin{aligned}\mathbf{P}(\mathrm{Z}' \in \mathrm{C}_\varepsilon^n) &\geqslant \mathrm{Card}(\mathrm{C}_\varepsilon^n) e^{-n(\mathrm{H}(\mathrm{X}_1)+\varepsilon)} e^{-n(\mathrm{H}(\mathrm{Y}_1)+\varepsilon)} \\ &\geqslant \left(1 - \frac{c(\mathrm{X}) + c(\mathrm{Y}) + c(\mathrm{Z})}{2n\varepsilon^2}\right) e^{-n(\mathrm{H}(\mathrm{X}_1)+\mathrm{H}(\mathrm{Y}_1)-\mathrm{H}(\mathrm{X}_1,\mathrm{Y}_1)+3\varepsilon)} \\ &= \left(1 - \frac{c(\mathrm{X}) + c(\mathrm{Y}) + c(\mathrm{Z})}{2n\varepsilon^2}\right) e^{-n(\mathrm{I}(\mathrm{X}_1,\mathrm{Y}_1)+3\varepsilon)}.\end{aligned}$$

Ceci conclut la preuve du théorème. □

2.5. Capacité de transmission d'un canal

Les deux thèmes étudiés jusqu'à présent — entropie d'une variable aléatoire et codage — concernaient uniquement les deux premières étapes du diagramme de communication présenté dans l'introduction. Nous allons maintenant faire intervenir la troisième : le canal de transmission et, en particulier, le problème du *bruit* à cause duquel un symbole reçu ne coïncide pas forcément avec le symbole émis.

contradiction. Enfin, si $\mathrm{H}(a) = \mathrm{H}'(ab)$o est préfixe de $\mathrm{H}(c)$, pour $c \neq a, b$, alors $\mathrm{H}'(ab)$ est préfixe de $\mathrm{H}'(c) = \mathrm{H}(c)$, une contradiction également. De même, $\mathrm{H}(b)$ n'est pas préfixe de $\mathrm{H}(c)$, pour aucun $c \neq a, b$. Cela prouve que le code H est un code préfixe.

Démontrons maintenant qu'il est optimal relativement à la loi p. Soit C un code binaire uniquement décodable optimal ; par définition, la longueur moyenne de C est inférieure à celle de H, c'est-à-dire que l'on a

$$\sum_{a\in \mathrm{A}} p(a)\ell(\mathrm{C}(a)) \leqslant \sum_{a\in \mathrm{A}} p(a)\ell(\mathrm{H}(a)).$$

Démontrons que l'on a en fait égalité. Puisque C est optimal, il existe deux symboles x, y tels que $\mathrm{C}(x)$ et $\mathrm{C}(y)$ soit de longueur maximale, disons n, et différent l'un de l'autre par leur dernier symbole uniquement. De plus, les longueur des mots de code des deux symboles a, b les moins probables choisis pour le codage de Huffman sont également de longueur n. Quitte à modifier le code C en échangeant les valeurs de $\mathrm{C}(a)$ et $\mathrm{C}(x)$ d'une part, et de $\mathrm{C}(b)$ et $\mathrm{C}(y)$ d'autre part, on peut supposer que $\mathrm{C}(a)$ et $\mathrm{C}(b)$ diffèrent uniquement de leur dernier symbole. Quitte à échanger $\mathrm{C}(a)$ et $\mathrm{C}(b)$, on suppose aussi que $\mathrm{C}(a)$ se termine par 0 et $\mathrm{C}(b)$ se termine par 1.

Le codage de Huffman a introduit l'alphabet $\mathrm{A}' = \mathrm{A} - \{a, b\} \cup \{ab\}$, muni de la loi de probabilité p' qui coïncide avec p sur $\mathrm{A} - \{a, b\}$et telle que $p'(ab) = p(a) + p(b)$. Définissons un code C' sur cet alphabet en posant $\mathrm{C}'(x) = \mathrm{C}(x)$ si $x \in \mathrm{A} - \{a, b\}$, et en prenant pour $\mathrm{C}'(ab)$ le mot déduit de $\mathrm{C}(a)$ (ou de $\mathrm{C}(b)$) en enlevant le dernier symbole. Sa longueur moyenne est

$$\begin{aligned}\ell_{p'}(\mathrm{C}') &= \sum_{x\neq a,b} p'(x)\ell(\mathrm{C}'(x)) + p'(ab)\ell(\mathrm{C}'(ab)) \\ &= \sum_{x\neq a,b} p(x)\ell(\mathrm{C}(x)) + (p(a) + p(b))(\ell(\mathrm{C}(a) - 1) \\ &= \ell_p(\mathrm{C}) - (p(a) + p(b)),\end{aligned}$$

car $\ell(\mathrm{C}(a)) = \ell(\mathrm{C}(b))$. Le même calcul pour le code de Huffman $\mathrm{H}_{p'}$ montre qu'il est de longueur moyenne

$$\ell_p(\mathrm{H}) - (p(a) + p(b)).$$

Par récurrence, le code de Huffman H' est optimal, de sorte que $\ell_{p'}(\mathrm{C}') \geqslant \ell_{p'}(\mathrm{H}')$. On a donc $\ell_p(\mathrm{C}) \geqslant \ell_p(\mathrm{H})$, d'où l'égalité, comme il fallait démontrer. □

2.4. Loi des grands nombres et compression

Dans son article originel, SHANNON (1948) utilisait une autre description de l'entropie, liée à la loi des grands nombres en théorie des probabilités.

Commençons par rappeler deux inégalités importantes.

Proposition **(2.4.1)**. — *Soit* X *une variable aléatoire possédant une espérance.*

a) *Pour tout nombre réel* $t > 0$, *on a l'*inégalité de Markov :

$$\mathbf{P}(|X| > t) \leqslant \mathbf{E}(|X|)/t.$$

b) *Supposons, de plus, que* X *possède une variance* $V(X)$. *Alors, pour tout nombre réel* $t > 0$, *on a l'*inégalité de Bienaymé–Tchebitcheff :

$$\mathbf{P}(|X - \mathbf{E}(X)| > t) \leqslant V(X)/t^2.$$

Démonstration. — a) Soit A l'ensemble des éléments ω de l'univers tels que $|X(\omega)| > t$; il s'agit de majorer $\mathbf{P}(A)$. Observons que sur A, la variable aléatoire $|X|/t$ est supérieure à 1 ; sur son complémentaire $\complement A$, elle est positive ou nulle. On a donc $\mathbf{E}(|X|/t) \geqslant 1 \cdot \mathbf{P}(A) + 0 \cdot \mathbf{P}(\complement A)$. Comme $\mathbf{E}(|X|/t) = \mathbf{E}(|X|)/t$, l'inégalité de Markov en résulte.

b) On considère maintenant la variable aléatoire $Y = X - \mathbf{E}(X)$. Comme X a une variance, Y^2 possède une espérance, égale à $V(X)$ par définition. Appliquons donc à Y^2 l'inégalité de Markov : on trouve $\mathbf{P}(Y^2 > t^2) \leqslant \mathbf{E}(Y^2)/t^2 = V(X)/t^2$. Puisque $Y^2 > t^2$ équivaut à $|X - \mathbf{E}(X)| > t$, cela prouve l'inégalité de Bienaymé–Tchebitcheff. □

Nous utiliserons directement ces inégalités, mais il est intéressant d'en déduire tout de suite la *loi des grands nombres.*

Théorème **(2.4.2)** (Loi faible des grands nombres). — *Soit* (X_n) *une suite de variables indépendantes et identiquement distribuées, d'espérance finie. Pour tout* $n \geqslant 1$, *on pose* $S_n = (X_1 + \cdots + X_n)/n$. *Pour tout nombre réel* $t > 0$, *on a*

$$\mathbf{P}(|S_n - \mathbf{E}(X_1)| > t) \to 0$$

quand n tend vers $+\infty$.

Dans le langage de la théorie des probabilités, on dit que S_n *converge en probabilité vers* $\mathbf{E}(X_1)$.

Lorsqu'on la divise par n, on obtient

$$\frac{1}{n}\lceil n(H_2(X_1) + \varepsilon)\rceil + \frac{c}{2n\varepsilon^2}\log_2(\mathrm{Card}(A)),$$

quantité majorée par $H_2(X_1) + 1$ lorsque n est assez grand.

2.4.5. — L'information mutuelle, comme l'entropie, dispose d'une interprétation statistique. Soit A et B des ensembles finis, soit $C = A \times B$ l'ensemble produit et soit $Z_1 = (X_1, Y_1), \ldots, Z_n = (X_n, Y_n)$ des variables aléatoires indépendantes et de même loi à valeurs dans C. Posons enfin $X = (X_1, \ldots, X_n)$, $Y = (Y_1, \ldots, Y_n)$ et $Z = (Z_1, \ldots, Z_n) = (X, Y)$ si l'on identifie l'élément (a, b) de $A^n \times B^n$ avec l'élément $((a_1, b_1), \ldots, (a_n, b_n))$ de C^n. On pose $c(Z) = \sum_{a,b\in C} \mathbf{P}(Z_1 = a)\mathbf{P}(Z_1 = b)\log(\mathbf{P}(Z_1 = a)/\mathbf{P}(Z_1 = b))^2$ et on définit $c(X)$ et $c(Y)$ de manière analogue.

Soit ε un nombre réel > 0 ; définissons une partie C_ε^n de C^n par : C_ε^n est l'ensemble des $(a, b) \in C_\varepsilon^n$ tels que

$$e^{-n(H(X_1)+\varepsilon)} \leqslant \mathbf{P}(X = a) \leqslant e^{-n(H(X_1)-\varepsilon)}$$
$$e^{-n(H(Y_1)+\varepsilon)} \leqslant \mathbf{P}(Y = b) \leqslant e^{-n(H(Y_1)-\varepsilon)}$$
$$e^{-n(H(X_1,Y_1)+\varepsilon)} \leqslant \mathbf{P}(X = a, Y = b) \leqslant e^{-n(H(X_1,Y_1)-\varepsilon)}.$$

Théorème **(2.4.6)**. — *On a*

$$\mathbf{P}(Z \in C_\varepsilon^n) \geqslant 1 - \frac{c(X) + c(Y) + c(Z)}{2n\varepsilon^2}$$

et

$$\mathrm{Card}(C_\varepsilon^n) \leqslant e^{n(H(X,Y)+\varepsilon)}.$$

Enfin, si $X'_1, \ldots, X'_n$ *d'une part,* $Y'_1, \ldots, Y'_n$ *d'autre part, sont des variables aléatoires de mêmes lois que* $X_1, \ldots, X_n$ *et* $Y_1, \ldots, Y_n$, *mais indépendantes, alors*

$$\Big(1 - \frac{c(X) + c(Y) + c(Z)}{2n\varepsilon^2}\Big)e^{-n(H(X,Y)+3\varepsilon)} \leqslant \mathbf{P}((X', Y')) \in C_\varepsilon^n) \leqslant e^{-n(I(X,Y)-3\varepsilon)}.$$

Démonstration. — On munit l'ensemble C^n de la loi de Z. Soit $\varphi_Z : C \to \mathbf{R}_+$ la variable aléatoire définie par $\varphi_Z(c) = -\log(\mathbf{P}(Z = c))$ si $\mathbf{P}(Z = c) > 0$, et $\varphi_Z(c) = 0$ sinon. Définissons $\varphi_X : A \to \mathbf{R}_+$ et $\varphi_Y : B \to \mathbf{R}_+$ de façon analogue. Comme dans la démonstration du théorème ..., on prouve que

$$\mathbf{P}(\complement C_\varepsilon^n) < \frac{c(X)}{2n\varepsilon^2} + \frac{c(Y)}{2n\varepsilon^2} + \frac{c(Z)}{2n\varepsilon^2},$$

de sorte que

$$\begin{aligned}\mathrm{V}(\mathrm{U}_1) &= \sum_{a\in\mathrm{A}} \mathbf{P}(\mathrm{X}_1 = a)\log(\mathbf{P}(\mathrm{X}_1 = a))^2 \\ &\quad - \sum_{a,b\in\mathrm{A}} \mathbf{P}(\mathrm{X}_1 = a)\mathbf{P}(\mathrm{X}_1 = b)\log(\mathbf{P}(\mathrm{X}_1) = a)\log(\mathbf{P}(\mathrm{X}_1 = b) \\ &= \sum_{a,b\in\mathrm{A}} p_a p_b\,(\log(p_a)^2 - \log(p_a)\log(p_b)) \\ &= \sum_{a,b\in\mathrm{A}} p_a p_b \log(p_a)\log(p_a/p_b).\end{aligned}$$

Par symétrie, on a aussi

$$\mathrm{V}(\mathrm{U}_1) = \sum_{a,b\in\mathrm{A}} p_a p_b \log(p_b)\log(p_b/p_a),$$

d'où, en additionant ces deux formules, l'égalité

$$2\mathrm{V}(\mathrm{U}_1) = \sum_{a,b\in\mathrm{A}} p_a p_b \log(p_b/p_a)^2 = c.$$

Cela conclut la preuve de la première inégalité.

Ensuite, en utilisant la majoration de $\mathbf{P}(\mathrm{X}_1 = a_1, \dots, \mathrm{X}_n = a_n)$ pour $(a_1, \dots, a_n) \in \mathrm{A}_\varepsilon^n$,

$$1 \geqslant \mathbf{P}(\mathrm{A}_\varepsilon^n) \geqslant \mathrm{Card}(\mathrm{A}_\varepsilon^n) e^{-n(\mathrm{H}(\mathrm{X}_1)+\varepsilon)},$$

d'où la majoration donnée pour $\mathrm{Card}(\mathrm{A}_\varepsilon^n)$. En utilisant la minoration analogue, on obtient aussi

$$1 - \frac{c}{2n\varepsilon^2} \leqslant \mathbf{P}(\mathrm{A}_\varepsilon^n) \leqslant \mathrm{Card}(\mathrm{A}_\varepsilon^n) e^{-n(\mathrm{H}(\mathrm{X}_1)-\varepsilon)},$$

ce qui donne la minoration de $\mathrm{Card}(\mathrm{A}_\varepsilon^n)$. □

2.4.4. — Comment Shannon en déduit-il la possibilité de comprimer un signal (dans la limite permise par l'entropie) ? Fixons un paramètre ε et considérons l'« ensemble typique » A_ε^n de A^n défini dans le théorème 2.4.3. Il est de cardinal au plus $2^{n(\mathrm{H}_2(\mathrm{X}_1)+\varepsilon)}$; donc on numérotant ses éléments, chacun ne requiert que $\lceil n(\mathrm{H}_2(\mathrm{X}_1) + \varepsilon)\rceil$ bits. Les autres vont chacun requérir $\mathrm{Card}(\mathrm{A})^n$ symboles, soit $n\log_2(\mathrm{Card}(\mathrm{A}))$ bits, mais n'apparaissent qu'avec une probabilité faible, majorée par $c/2n\varepsilon^2$. Ainsi, la longueur moyenne du code d'une suite de n symboles est majorée par

$$\lceil n(\mathrm{H}_2(\mathrm{X}_1) + \varepsilon)\rceil + \frac{c}{2n\varepsilon^2} n\log_2(\mathrm{Card}(\mathrm{A})).$$

Démonstration. — Comme toutes les X_n ont même loi, elles ont même espérance, en remplaçant X_n par $\mathrm{X}_n - \mathbf{E}(\mathrm{X}_1)$, on remplace S_n par $\mathrm{S}_n - \mathbf{E}(\mathrm{X}_1)$. Il suffit alors de démontrer que $\mathbf{P}(|\mathrm{S}_n| > t)$ tend vers 0 sous l'hypothèse $\mathbf{E}(\mathrm{X}_1) = 0$.

Commençons par démontrer une majoration précise de cette probabilité sous l'hypothèse supplémentaire que les variables aléatoires X_n sont de variance finie. Dans ce cas, S_n est de variance

$$\mathrm{V}(\mathrm{S}_n) = \mathbf{E}(\mathrm{S}_n^2) = \frac{1}{n^2}\mathbf{E}((\mathrm{X}_1 + \cdots + \mathrm{X}_n)^2) = \frac{1}{n^2}\sum_{i,j}\mathbf{E}(\mathrm{X}_i\mathrm{X}_j).$$

Pour $i \neq j$, l'indépendance de X_i et X_j entraîne $\mathbf{E}(\mathrm{X}_i\mathrm{X}_j) = \mathbf{E}(\mathrm{X}_i)\mathbf{E}(\mathrm{X}_j) = 0$; pour $i = j$, on a $\mathbf{E}(\mathrm{X}_i^2) = \mathrm{V}(\mathrm{X}_i) = \mathrm{V}(\mathrm{X}_1)$ puisque les X_i ont même loi, donc même variance. Ainsi, $\mathrm{V}(\mathrm{S}_n) = \mathrm{V}(\mathrm{X}_1)/n$. (Plus généralement, la variance d'une somme de variables aléatoires indépendantes est la somme de leurs variances.) Appliquons maintenant l'inégalité de Bienaymé–Tchebitcheff : pour tout nombre réel $t > 0$, on a

$$\mathbf{P}(|\mathrm{S}_n| > t) \leqslant \mathbf{E}((\mathrm{S}_n/t)^2) \leqslant \mathrm{V}(\mathrm{S}_n)/t^2.$$

Par suite, $\mathbf{P}(|\mathrm{S}_n| > t) \leqslant \mathrm{V}(\mathrm{X}_1)/nt^2$, d'où le résultat voulu.

Le cas général est plus difficile et se démontre par une méthode classique de troncation. Introduisons un paramètre $\delta > 0$ et posons $\mathrm{X}'_k = \mathrm{X}_k$ si $|\mathrm{X}_k| \leqslant \delta n$, et $\mathrm{X}'_k = 0$ sinon ; posons aussi $\mathrm{X}''_k = \mathrm{X}_k - \mathrm{X}'_k$. On va majorer les probabilités

$$\mathbf{P}(|\mathrm{X}'_1 + \cdots + \mathrm{X}'_n| > nt/2) \qquad \text{et} \qquad \mathbf{P}(|\mathrm{X}''_1 + \cdots + \mathrm{X}''_n| > nt/2).$$

La somme de ces probabilités fournira une majoration de $\mathbf{P}(|\mathrm{X}_1 + \cdots + \mathrm{X}_n| > nt)$ qui sera arbitrairement petite pour tout n assez grand. L'idée sous-jacente à cette méthode est que les X'_k sont majorées, donc de variance finie, et seront justiciables du premier cas, tandis que les X''_k seront rares, car X''_k est nulle lorsque X_k n'est pas trop grande.

Tout d'abord, les $|\mathrm{X}'_j|$ sont de même loi, mutuellement indépendantes, et majorées par δn. Elles ont donc une variance commune, laquelle vérifie

$$\mathrm{V}(\mathrm{X}'_j) = \mathrm{V}(\mathrm{X}'_1) \leqslant \mathbf{E}((\mathrm{X}'_1)^2) \leqslant \delta\mathbf{E}(|\mathrm{X}'_1|)n \leqslant \delta\mathbf{E}(|\mathrm{X}_1|)n.$$

Par suite,

$$\mathrm{V}(\mathrm{X}'_1 + \cdots + \mathrm{X}'_n) = n\mathrm{V}(\mathrm{X}'_1) \leqslant \delta\mathbf{E}(|\mathrm{X}_1|)n^2.$$

On a aussi

$$\mathbf{E}(\mathrm{X}'_1 + \cdots + \mathrm{X}'_n)^2 = \mathbf{E}(\mathrm{X}'_1)^2 n^2.$$

Lorsque n tend vers l'infini, X'_1 tend vers X_1 presque sûrement, tout en étant majorée en valeur absolue par $|\mathrm{X}_1|$; le théorème de convergence dominée entraîne donc que $\mathbf{E}(\mathrm{X}'_1)$ tend vers $\mathbf{E}(\mathrm{X}_1) = 0$. En particulier, pour n assez grand, $\mathbf{E}(\mathrm{X}'_1 + \cdots + \mathrm{X}'_n)^2 \leqslant \delta n^2$, et

$$\mathbf{E}((\mathrm{X}'_1 + \cdots + \mathrm{X}'_n)^2) \leqslant 2\delta\mathbf{E}(|\mathrm{X}_1|)n^2.$$

L'inégalité de Bienaymé–Tchebitcheff entraîne alors que

(2.4.2.1) $$\mathbf{P}(|X'_1 + \cdots + X'_n| > nt/2) \leqslant 8\delta\mathbf{E}(|X_1|)t^{-2}$$

pour tout entier n assez grand. Fixons un nombre réel $\varepsilon > 0$ et choisissons δ de sorte que

$$8\delta\mathbf{E}(|X_1|)t^{-2} < \varepsilon/2.$$

On retient alors de l'inégalité (2.4.2.1) que pour tout n assez grand, on a $\mathbf{P}(|X'_1 + \cdots + X'_n|) \leqslant \varepsilon/2$.

Par ailleurs,

$$\mathbf{P}(|X''_1 + \cdots + X''_n| > nt/2) \leqslant \mathbf{P}(X''_1 + \cdots + X''_n \neq 0) \leqslant n\mathbf{P}(X''_1 \neq 0),$$

puisque les X''_j sont de même loi. Or,

$$\mathbf{P}(X''_1 \neq 0) = \mathbf{P}(|X''_1| > \delta n) \leqslant \mathbf{E}(|X''_1|/\delta n) = \mathbf{E}(|X''_1|)\delta^{-1}n^{-1},$$

d'où l'inégalité

(2.4.2.2) $$\mathbf{P}(|X''_1 + \cdots + X''_n| > nt/2) \leqslant \mathbf{E}(|X''_1|)\delta^{-1}.$$

Quand n tend vers $+\infty$, X''_1 tend presque partout vers 0, et l'on a $|X''_1| \leqslant |X_1|$, de sorte que $\mathbf{E}(|X''_1|)$ tend vers 0. Alors, pour n assez grand, l'inégalité (2.4.2.2) assure que $\mathbf{P}(|X''_1 + \cdots + X''_n| > nt/2) < \varepsilon/2$.

En combinant ces deux majorations, on en déduit que pour tout n assez grand, l'événement $\{|X_1 + \cdots + X_n| > nt\}$ est de probabilité $< \varepsilon$, ce qui conclut la démonstration. □

Théorème **(2.4.3)** (Shannon). — *Soit (X_n) une suite de variables aléatoires prenant leurs valeurs dans un ensemble fini A, indépendantes et de même loi p; posons $c = \sum_{a,b\in A} p_a p_b(\log(p_a/p_b))^2$. Pour tout entier $n \geqslant 1$, munissons l'ensemble A^n de la loi produit. Soit ε un nombre réel > 0 et soit A^n_ε l'ensemble des $(a_1, \ldots, a_n)$ tels que*

$$e^{-n(H(X_1)+\varepsilon)} \leqslant \mathbf{P}(X_1 = a_1, \ldots, X_n = a_n) \leqslant e^{-n(H(X_1)-\varepsilon)}.$$

On a

$$\mathbf{P}(A^n_\varepsilon) > 1 - \frac{c}{2n\varepsilon^2},$$

et

$$\left(1 - \frac{c}{2n\varepsilon^2}\right)e^{n(H(X_1)-\varepsilon)} \leqslant \mathrm{Card}(A^n_\varepsilon) \leqslant e^{n(H(X_1)+\varepsilon)}.$$

Reformulons un peu cet énoncé : $1 - \mathbf{P}(A^n_\varepsilon)$ est la probabilité du complémentaire de A^n_ε, et est majorée par $c/2n\varepsilon^2$; lorsque n est grand, elle est arbitrairement petite. Autrement dit, lorsque n est grand, la plupart des tirages $(a_1, \ldots, a_n)$ ont une probabilité voisine de $e^{-nH(X_1)}$, et il y a environ $e^{nH(X_1)}$ tels tirages. Autrement dit encore, lorsqu'on effectue un grand nombre n de tirages, tout se passe comme si l'on avait effectué un tirage au sort équitable parmi $e^{nH(X_1)}$ valeurs — c'est l'*interprétation statistique* de l'entropie.

Démonstration. — Quitte à modifier l'univers en lui enlevant un ensemble de probabilité 0, puis l'ensemble A par l'ensemble des valeurs des X_k, on suppose que $\mathbf{P}(X_k = a) > 0$ pour tout $a \in A$ et tout $k \in \{1, \ldots, n\}$. Soit alors $\varphi : A \to \mathbf{R}$ l'application définie par $\varphi(a) = -\log(\mathbf{P}(X_1 = a))$.

Comme les X_k sont indépendantes et de même loi, on a

$$\begin{aligned}\mathbf{P}(X_1 = a_1, \ldots, X_n = a_n) &= \mathbf{P}(X_1 = a_1)\ldots\mathbf{P}(X_n = a_n)\\ &= \mathbf{P}(X_1 = a_1)\ldots\mathbf{P}(X_1 = a_n),\end{aligned}$$

soit encore

$$-\frac{1}{n}\log(\mathbf{P}(X_1 = a_1, \ldots, X_n = a_n)) = -\frac{1}{n}\sum_{k=1}^{n}\log(\mathbf{P}(X_1 = a_k)) = \frac{1}{n}\sum_{k=1}^{n}\varphi(a_k).$$

On munit l'ensemble A de la loi de X_1; quitte à le remplacer par l'ensemble des $a \in A$ tels que $\mathbf{P}(X_1 = a) > 0$, on suppose que $\mathbf{P}(X_1 = a) > 0$ pour tout $a \in A$. On munit alors l'ensemble A^n de la loi produit. On a $\mathbf{P}(a_1, \ldots, a_n) = \prod_{k=1}^n \mathbf{P}(X_1 = a_k)$. Sur cet espace probabilisé A^n, on pose $U_k(a_1, \ldots, a_n) = \varphi(a_k)$.

Les variables aléatoires $U_1, \ldots, U_n$ sont indépendantes. Elles ont même loi : pour tout $a \in A$, on a

$$\mathbf{P}(U_k = \varphi(a)) = \mathbf{P}(X_1 = a),$$

et $\mathbf{P}(U_k = t) = 0$ si $t \notin \varphi(A)$. Elles sont d'espérance et de variance finies car elles ne prennent qu'un nombre fini de valeurs. De plus, on a

$$H(X_1) = -\sum_{a\in A}\mathbf{P}(X_1 = a)\log(\mathbf{P}(X_1) = a)) = \mathbf{E}(U_1).$$

D'après l'inégalité de Bienaymé–Tchebitcheff, on a

$$\mathbf{P}\left(|H(X_1) - \frac{1}{n}\sum_{k=1}^{n}U_k| > \varepsilon\right) < V(U_1)/n\varepsilon^2.$$

On a

$$\frac{1}{n}\sum_{k=1}^{n}U_k(a_1, \ldots, a_n) = -\frac{1}{n}\log(\mathbf{P}(X_1 = a_1, X_2 = a_2, \ldots, X_n = a_n)).$$

Il reste à calculer la variance $V(U_1)$ de U_1. Par définition, on a

$$V(U_1) = \mathbf{E}((U_1 - \mathbf{E}(U_1))^2) = \mathbf{E}(U_1^2) - \mathbf{E}(U_1)^2,$$