Fragment 7

Quelques mots sur l'entropie

Le mot *entropie* vint du grec *entropê*, qui signifie « retour ». Le concept d'entropie a été introduit en thermodynamique par Clausius il y a cent cinquante ans. La croissance au cours du temps pour un système isolé de cette variable thermodynamique extensive est sensée exprimer l'irréversibilité, le « non-retour », certains parlent même de « flèche du temps » à ce propos. Le concept général et désormais classique d'entropie a eu un succès sans précédent bien au delà de la physique et le mot « entropie » est aujourd'hui assez galvaudé. Notre objectif ici est de montrer comment l'entropie utilisée par Boltzmann en théorie cinétique des gaz et par Shannon en théorie de l'information apparaît naturellement. Nous en donnons les aspects les plus importants, qui reposent sur des propriétés élémentaires de combinatoire, de convexité, et des fonctions $x \mapsto \log x$ et $x \mapsto e^x$.

7.1 L'entropie selon Boltzmann

Soit Σ_r un système « macroscopique » constitué de r particules « microscopiques » *indiscernables* pouvant être chacune dans l'un des n états possibles. L'état macroscopique du système est donnée par le nombre de particules dans chaque état, autrement dit, par le vecteur $(r_1, \ldots, r_n)$ où r_i est le nombre de particules dans l'état i. Pour un état macroscopique donné, le nombre de « degrés de liberté » du système Σ_r est donné naturellement par le nombre d'états microscopiques compatibles avec l'état macroscopique spécifié. Malheureusement, cette grandeur n'est pas *extensive* car le nombre de degrès de liberté de la juxtaposition de deux systèmes est le produit des degrès de liberté de chacun des deux systèmes, et pas leur somme. Il est donc plus commode de considérer le logarithme du nombre de degrés de liberté. Ainsi, par définition, l'entropie $\mathbf{S}(\Sigma_r)$ du système Σ_r est le logarithme du nombre d'états microscopiques compatibles avec l'état macroscopique donné. Or il y a exactement

$$C_r^{r_1,\ldots,r_n} := \frac{r!}{r_1! \cdots r_n!}$$

états microscopiques possibles pour le système Σ_r lorsque l'état macroscopique $(r_1, \ldots, r_n)$ est fixé. En vertu de de qui précède, l'entropie moyenne du système par particule est alors donné par

$$\mathbf{S}_{\text{moy}}(\Sigma_r) := \frac{1}{r} \log C_r^{r_1,\ldots,r_n}.$$

Le vecteur $r^{-1}(r_1, \ldots, r_n)$ est une loi de probabilité discrète puisque par définition $r = r_1 + \cdots + r_n$. Supposons que les fréquences $r^{-1}(r_1, \ldots, r_n)$ convergent vers la loi de probabilité $(p_1, \ldots, p_n)$ lorsque le

nombre de particules r tend vers l'infini. La formule de Striling[1] indique alors que l'entropie moyenne par particule pour le système infini Σ_∞ est

$$\mathbf{S}_{\text{moy}}(\Sigma_\infty) := \lim_{r\to+\infty} \frac{1}{r} \log C_r^{r_1,\dots,r_n} = -\sum_{i=1}^{n} p_i \log p_i.$$

Ce raisonnement est à peu de chose près celui qu'a fait Boltzmann. Nous désignerons la quantité

$$\mathbf{H}(p_1,\dots,p_n) := \sum_{i=1}^{n} p_i \log_2 p_i \tag{7.1}$$

par le terme « entropie de Boltzmann » de la loi de probabilité discrète $(p_1,\dots,p_n)$. En particulier, lorsque les n états microscopiques possibles sont équiprobables, on a $p_1 = \cdots = p_n = n^{-1}$ et $\mathbf{S} = \log n$, qui n'est qu'une forme de la célèbre formule[2] $\mathbf{S} = \kappa \log W$. Boltzmann a utilisé cette notion en théorie cinétique des gaz. Considérons un gaz constitué d'un très grand nombre de particules identiques possédant chacune une position et une vitesse repérées par un vecteur (x,v) dans $\mathbb{R}^6$. Le grand nombre de particules fait qu'il n'est pas envisageable d'écrire les équations du mouvement pour chacune d'entre elles car le système obtenu serait gigantesque et les conditions initiales inconnues. Boltzmann adopte alors une approche probabiliste, on disait « statistique » à l'époque, d'où le nom de « mécanique statistique ». Il considère la densité de probabilité $(x,v) \in \mathbb{R}^6 \mapsto f_t(x,v)$ qui représente la répartition des positions-vitesses des particules du système à l'instant t. Par homogénéité, il suppose pour simplifier que f_t ne dépend pas de la position x. Il écrit ensuite une équation aux dérivées partielles qui exprime l'évolution de cette densité au cours du temps en tenant compte des chocs entre particules. Il montre enfin en substance dans son célèbre « théorème-**H** » que sous certaines hypothèses simplificatrices, la quantité $\int_{\mathbb{R}^3} f_t(v) \log f_t(v)\, dv$, analogue continu de l'entropie discrète **H**, décroit au cours du temps vers une valeur minimale qui est atteinte lorsque f est gaussienne[3] de la forme $\mathcal{N}(u, T\,Id_3)$ où T est la température absolue du gaz.

Dans toute la suite, l'entropie de Boltzmann $\mathbf{H}(f)$ d'une densité de probabilité $f : \mathbb{R}^d \to \mathbb{R}$ par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$ est définie par

$$\mathbf{H}(f) := \int_{\mathbb{R}^d} f(u) \log f(u)\, du. \tag{7.2}$$

Boltzmann note **H** « son entropie » – d'où le nom de son fameux théorème – sans doute pour éviter les confusions avec l'entropie en thermodynamique qui est notée traditionnellement **S**. Comme nous allons le voir dans la section 7.3 page 113, l'entropie de Shannon est l'opposée en signe de l'entropie de Boltzmann. Elle est malheureusement notée **H** par Shannon lui même. Pour clarifier les choses, nous avons choisit de noter **H** l'entropie de Boltzmann et **I** l'entropie de Shannon. L'opposition de signe entre **H** et **I** fait que l'entropie de Shannon est parfois appelée négentropie, on parle aussi d'incertitude ou d'information (d'où notre notation).

7.2 Le logarithme et la notion intuitive d'information

Nous avons tous appris tôt ou tard que le logarithme népérien log est la primitive de la fonction $x \to x^{-1}$ sur $\mathbb{R}_+^*$ et que le logarithme $\log_b$ de base $b > 0$ est la fonction réciproque de la fonction $x \to b^x$ sur $\mathbb{R}_+^*$. Le logarithme a cependant une signification plus intuitive : pour tout $b \in \mathbf{N}^*$ et tout $x \geqslant 1$, la

[1] $n! \sim \sqrt{2\pi n}\, e^{-n} n^n$, formule qui peut s'établir à partir de la fonction $\Gamma(x) := \int_{\mathbb{R}_+} t^{x-1}\, e^{-t}\, dt$ d'Euler.

[2] Où κ est une constante universelle appelée « constante de Boltzmann », et où W est le nombre d'états d'énergie microscopiques possibles. Cette formule est gravée sur la tombe de Boltzmann à Vienne. On peut réécrire la formule sous la forme $\log_{e^{1/\kappa}} W$.

[3] On parle de distribution de Maxwell en physique dans ce contexte.

partie entière de $\log_b(x)$ représente le nombre de symboles nécessaires à l'écriture de x en base b. Cette propriété élémentaire découle de la monotonie du logarithme et du fait que pour tout $n \in \mathbb{N}^*$

$$\log_b(b^n) = n.$$

Le cas $0 < x < 1$ mène à une interprétation similaire relative aux symboles « après la virgule ». Ainsi, le logarithme interpole entre $n = \log_b(b^n)$ et $n+1 = \log_b(b^{n+1})$ de la même manière que la fonction gamma d'Euler interpole entre $n!$ et $(n+1)!$. La propriété de multiplicativité du logarithme $\log_b(xy) = \log_b(x) + \log_b(y)$ exprime alors le fait qu'à peu de chose près, en base b, le nombre de symboles nécessaires à l'écriture de xy est la somme du nombre de symboles nécessaires à l'écriture de x et de y.

Connaître x en base b revient à donner les $\log_b(x)$ symboles nécessaires à son écriture en base b. Ainsi, $\log_b(x)$ représente la *quantité d'information* nécessaire à la connaissance de x en base b. En informatique, on utilise en général la base $b = 2$ pour des raisons évidentes, et l'on parle de *bit* pour désigner une unité. Ainsi, pour $x \geqslant 1$, $\log_2 x$ represente le nombre de *bits d'information* nécessaires à la connaissance de x.

Considérons à présent un ensemble fini Ω de cardinal $|\Omega| \in \mathbb{N}^*$. L'axiome du choix le plus simple fait que l'on peut donc numéroter les éléments de Ω, du n°1 au n°$|\Omega|$. Ainsi, il faut $\log_2 |\Omega|$ bits d'information pour désigner un élément particulier de Ω. Comme l'ensemble des parties de Ω est de cardinal $2^{|\Omega|}$, il faut $|\Omega|$ bits d'information pour désigner une partie de Ω. On voit bien que la base 2 est la plus naturelle ici. De même, il faut $\log_2(C^k_{|\Omega|})$ bit d'information pour désigner une partie de cardinal k de Ω.

La propriété de multiplicativité du logarithme fait qu'il faut $\log_2 |\Omega_1| + \log_2 |\Omega_2|$ bits d'information pour désigner un élément particulier du produit cartésien $\Omega_1 \times \Omega_2$. Si $A \subset \Omega$, la quantité $\log_2 |\Omega| - \log_2 |A|$ représente l'information résiduelle nécessaire pour décrire Ω après description de A. Si l'on pose $p_A := |\Omega|/|A| \in [0,1]$, on a alors

$$\log_2 |\Omega| - \log_2 |A| = \log_2 \frac{|\Omega|}{|A|} = -\log_2 p_A.$$

Ainsi, si $p_A = 0$ (i.e. ici $A = \emptyset$), la description de A n'a rien changé au problème de la description de Ω, tandisque lorsque $p_A = 1$ (i.e. $A = \Omega$), la description de A suffit entièrement à décrire Ω. Soit à présent $A_1, \ldots, A_n$ une partition de Ω. On définit la loi de probabilité discrète $(p_1, \ldots, p_n)$ par $p_i := p_{A_i}$. On peut définir la quantité moyenne en bits d'information nécessaire à la description de Ω après description de l'un des A_i par :

$$-\sum_{i=1}^{n} p_i \log_2 p_i.$$

Cette quantité est exactement – au facteur $\log 2$ et au signe près – l'entropie de Boltzmann (7.1). Le modèle d'équiprobabilité ici est caché dans Ω, et la formule $p_A = |A|/|\Omega|$ correspond à la sacro-sainte formule « cas favorables sur cas totaux ».

7.3 L'entropie selon Shannon

Considérons un ensemble $\mathcal{A} := \{a_1, \ldots, a_n\}$ de cardinal n, que l'on appelera *alphabet*, et dont les éléments seront appelés *symboles*. Un message de longueur $r \in \mathbb{N}$ écrit avec cet alphabet ne sera qu'une suite finie de longueur r d'éléments de $\mathcal{A}$. En d'autres termes, ces messages sont exactement les éléments de $\mathcal{A}^r$. Soit $x = x_r \cdots x_r$ un message de longueur r écrit avec l'alphabet $\mathcal{A}$. Si r_i désigne le nombre d'occurrence du symbole a_i dans le message x, on a forcément $r = r_1 + \cdots + r_n$ et la fréquence d'apparition $f_{r,i}$ du symbole a_i est donnée par $f_{r,i} := r^{-1} r_i$. Lorsque $\mathcal{A}, r, r_1, \ldots, r_n$ sont fixés, le nombre de messages possibles est donné par le coefficient multinomial suivant :

$$C_r^{r_1,\ldots,r_n} := \frac{r!}{r_1! \cdots r_n!}.$$

Pour transmettre un message de longueur quelconque r écrit avec l'alphabet $\mathcal{A}$, il suffit donc par simple numérotation de transmettre d'abord $r_1, \ldots, r_n$, ce qui requière d'après la section 7.2 moins de $n \log_2 r$ bits, puis de transmettre un nombre plus petit ou égal à $C_r^{r_1,\ldots,r_n}$. Le nombre $\mathrm{I}(r)$ de bits requis vérifie donc

$$\log_2 C_r^{r_1,\ldots,r_n} \leqslant \mathrm{I}(r) \leqslant n \log_2 r + \log_2 C_r^{r_1,\ldots,r_n}.$$

Supposons que lorsque la longueur r du message est grande, les fréquences $(f_{r,1}, \ldots, f_{r,n})$ d'apparitions des symboles de l'alphabet convergent vers une loi de probabilité discrète $(p_1, \ldots, p_n)$. La formule de Stirling donne alors

$$\mathrm{I}(r) \underset{r=+\infty}{\sim} r \left(- \sum_{i=1}^{n} p_i \log_2 p_i \right).$$

Ainsi, la quantité moyenne d'information en bits par symbole nécessaire à la transmission d'un message écrit avec l'alphabet $\mathcal{A}$ est asymptotiquement donnée par l'entropie (7.1) où p_i représente la probabilité d'apparition du symbole a_i dans le message.

Ce raissonnement peut alors être inversé de la façon suivante : si une source émet de façon i.i.d. des symboles de l'alphabet $\mathcal{A} := \{a_1, \ldots, a_n\}$ avec probabilités $(p_1, \ldots, p_n)$, alors la quantité d'information moyenne en bit par symbole nécessaire à la transmission de la sortie de cette source est donnée par l'entropie (7.1).

Sur le plan combinatoire, le raisonnement est exactement celui qu'a fait Boltzmann : asymptotique de la loi multinômiale lorsque les fréquences convergent, et utilisation de la formule de Stirling pour faire apparaître la formule de l'entropie discrète via la propriété clé $\log(m!) \sim m \log m$.

Nous définissons l'*entropie de Shannon discrète* de la loi discrète $(p_1, \ldots, p_n)$ par la formule

$$\mathbf{I}(p_1, \ldots, p_n) := - \sum_{i=1}^{n} p_i \log_2 p_i, \tag{7.3}$$

qui – au signe et au facteur $\log 2$ près – est exactement l'entropie de Boltzmann définie par (7.1). De même, nous définissons l'*entropie de Shannon continue* de la densité de probabilité $f : \mathbb{R}^d \to \mathbb{R}$ par rapport à la mesure de Lebesgue par la formule

$$\mathbf{I}(f) := - \int_{\mathbb{R}^d} f(x) \log_2 f(x)\, dx, \tag{7.4}$$

qui – au signe et au facteur $\log 2$ près – est exactement l'entropie de Boltzmann continue définie par (7.2). Comme les entropies de Boltzmann $\mathbf{H}$ et de Shannon $\mathbf{I}$ sont semblables, nous ne parleront que de celle de Shannon, et ce choix et tout à fait arbitraire. D'autre part, pour une variable aléatoire X, on notera $\mathbf{I}(X)$ l'entropie $\mathbf{I}(\mathcal{L}(X))$ de la loi de X. Remarquons que $\mathbf{I}(f)$ est bien définie seulement lorsque $f \log_2 f$ est Lebesgue intégrable.

Remarque 7.3.1 *(Entropie relative de Kullback-Leibler).* Soient μ et ν deux mesures positives définies sur le même espace mesurable $(\Omega, \mathcal{F})$. L'entropie relative de Kullback-Leibler $\mathbf{Ent}(\nu \,|\, \mu)$ est définie par la formule (2.6) page 51. Lorsque μ est une loi de probabilité, l'inégalité de Jensen entraîne que $\mathbf{Ent}(\nu \,|\, \mu) \in \mathbb{R}_+ \cup \{+\infty\}$. De plus $\mathbf{Ent}(\nu \,|\, \mu) = 0$ si et seulement si $\mu = \nu$ et $\mathbf{Ent}(\nu \,|\, \mu) < +\infty$ si et seulement si $\frac{d\nu}{d\mu} \log \frac{d\nu}{d\mu} f \in \mathrm{L}^1(\Omega, \mathcal{F}, \mu; \mathbb{R})$.

7.4 Propriétés de l'entropie discrète

L'entropie de Shannon $\mathbf{I}(p_1, \ldots, p_n)$ d'une loi de probabilité discrète $(p_1, \ldots, p_n)$ est définie par (7.3). On remarquera qu'elle ne dépend que des coefficients p_i, et pas du support de la loi discrète considérée.

Elle est *continue* sur le simplexe Λ_n des lois de probabilités discrètes de taille au plus n :

$$\Lambda_n := \{(p_1, \dots, p_n) \in (\mathbb{R}_+)^n,\ p_1 + \cdots + p_n = 1\}. \tag{7.5}$$

La convention $0 \log 0 = 0$ fait qu'il n'y a pas de problèmes aux bord de Λ_n. La fonction $x \in \mathbb{R}_+^* \to x \log x$ étant strictement convexe, on a :

$$\begin{aligned}\mathbf{I}(p_1, \dots, p_n) &= -n \sum_{i=1}^{n} \frac{1}{n} (p_i \log p_i) \\ &\leqslant -n \left(\sum_{i=1}^{n} \frac{p_i}{n} \right) \log \left(\sum_{i=1}^{n} \frac{p_i}{n} \right) \\ &= \log n = \mathbf{I}\left(\frac{1}{n}, \dots, \frac{1}{n} \right).\end{aligned}$$

La convexité stricte entraîne que le cas d'égalité n'a lieu que pour la mesure uniforme. D'autre part, la fonction $\mathbf{I}$ est strictement concave sur Λ_n en tant que fonction de n variables car $x \log x$ est strictement convexe. Ainsi, $\mathbf{I} : \Lambda_n \to [0, \log n]$ atteint sa valeur minimale 0 aux points extrémaux du convexe compact Λ_n de $\mathbb{R}^n$, qui sont exactement les masses de Dirac, et sa valeur maximale $\log n$ en un unique point qui est la loi uniforme[4].

Comme expliqué dans les sections 7.2 et 7.3, l'entropie de Shannon $\mathbf{I}(p_1, \dots, p_n)$ mesure l'*information* (en bits) nécessaire au codage d'une source de loi $(p_1, \dots, p_n)$. Cependant, selon le point de vue adopté, elle mesure également l'*incertitude*. Ainsi, l'entropie $\mathbf{I}(p_1, \dots, p_n)$ est d'autant plus grande que son argument est « aléatoire ». Une masse de Dirac représente la certitude tandis que la loi uniforme représente l'incertitude totale. La certitude se code avec très peu de bits tandisque que l'incertitude totale se code avec autant de bits que la taille du support. Ce qui précède est parfaitement cohérent avec ce qui est expliqué dans la section 7.2.

7.4.1 Extensivité de l'entropie

La fonction logarithme intervenant dans la définition de $\mathbf{I}$ n'est pas vraiment nécessaire à l'obtention des propriétés de monotonie, et l'on peut remplacer $x \log x$ par une fonction strictement convexe quelconque de $\mathbb{R}_+$ dans $\mathbb{R}$, s'anullant en 0 et en 1. En revanche, la fonction $x \log x$ fait de $\mathbf{I}$ une fonction *extensive*, comme l'exprime le théorème suivant.

Théorème 7.4.1 (Extensivité de l'entropie). *Si X et Y sont deux v.a. discrètes à valeurs dans $\{0, \dots, n\}$, alors*

$$\mathbf{I}((X, Y)) \leqslant \mathbf{I}(X) + \mathbf{I}(Y),$$

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

Démonstration. Notons $\mathcal{L}(X) = p_1 \delta_{x_1} + \cdots + p_n \delta_{x_n}$ la loi de X et $\mathcal{L}(Y) = q_1 \delta_{y_1} + \cdots + q_m \delta_{y_m}$ la loi de Y. La propriété fondamentale du logarithme $\log(ab) = \log a + \log b$ permet d'écrire :

$$\sum_{\substack{1 \leqslant i \leqslant n \\ 1 \leqslant j \leqslant m}} p_i q_j \log(p_i q_j) = \sum_{1 \leqslant j \leqslant m} q_j \log(q_j) \underbrace{\sum_{1 \leqslant i \leqslant n} p_i}_{=1} + \sum_{1 \leqslant i \leqslant n} p_j \log(p_j) \underbrace{\sum_{1 \leqslant j \leqslant m} q_j}_{=1}.$$

[4] Le fait que l'ensemble des minima et l'ensemble des maxima soient invariants par permutation des coordonnées est dû à la symétrie de $\mathbf{I}$.

Ainsi, nous avons $\mathbf{I}(\mathcal{L}(X) \otimes \mathcal{L}(Y)) = \mathbf{I}(X) + \mathbf{I}(Y)$. Posons $r_{i,j} := \mathbb{P}(X = x_i\,,\, Y = y_j)$. La continuité de l'entropie $\mathbf{I}$ permet de se ramener au cas où les p_i et les q_j sont tous strictement positifs. On a alors

$$\mathbf{I}((X,Y)) - \mathbf{I}(\mathcal{L}(X) \otimes \mathcal{L}(Y)) \quad = \quad \sum_{\substack{1 \leqslant i \leqslant n \\ 1 \leqslant j \leqslant m}} \left(\frac{r_{i,j}}{p_i q_j} \log \frac{r_{i,j}}{p_i q_j} \right) p_i q_j .$$

Pour terminer, il suffit de remarquer que l'inégalité de Jensen pour la variable aléatoire $(i,j) \mapsto r_{i,j}/p_i q_j$, la fonction convexe $x \mapsto x \log x$ et la loi de probabilité $\mathcal{L}(X) \otimes \mathcal{L}(Y)$ donne :

$$\sum_{\substack{1 \leqslant i \leqslant n \\ 1 \leqslant j \leqslant m}} \left(\frac{r_{i,j}}{p_i q_j} \log \frac{r_{i,j}}{p_i q_j} \right) p_i q_j \quad \geqslant \quad 0.$$

La convexité stricte entraîne que l'égalité n'a lieu que lorsque $r_{ij} = p_i q_j$ pour tous i et j, c'est-à-dire lorsque X et Y sont indépendantes, ce qui termine la preuve. On remarquera que l'inégalité précédente découle de la remarque 7.3.1 car $\mathbf{Ent}(\mathcal{L}(X,Y) \,|\, \mathcal{L}(X) \otimes \mathcal{L}(Y)) \geqslant 0$. □

Cette propriété « extensive » est tout à fait naturelle : l'incertitude d'un couple de v.a. est toujours plus petite que la somme des incertitudes, avec égalité si et seulement si elles sont indépendantes. On vérifie immédiatement que $\mathbf{I}((X,X)) = \mathbf{I}(X)$, en d'autres termes, la duplication d'une v.a. n'ajoute aucune incertitude. La notion d'entropie de Shannon mène à celles d'entropie conditionnelle et de capacité d'un canal de communication, cf. [App96, chap. 6], [Rom92, Rom97] et [YY59].

7.4.2 Le problème du codage optimal sans bruit

Alphabets, mots et messages

Un *n-alphabet* $\mathcal{A} := \{a_1, \ldots, a_n\}$ est la donnée de n symboles $a_1, \ldots, a_n$ tous différents. Un *message* ou un *mot* x écrit dans cet alphabet n'est qu'une suite finie $x_1, x_2, x_2, \ldots, x_m$ où les x_i sont tous dans $\mathcal{A}$. On note $\mathcal{A}^*$ l'ensemble des mots écrits avec l'alphabet $\mathcal{A}$. On note $|x|$ la longueur m du mot x. Les mots peuvent être *concaténés* pour obtenir des mots plus longs. Par exemple, pour tout entier k, le produit cartésien $\mathcal{A}^k$ s'injecte naturellement dans $\mathcal{A}^*$ par la concaténation

$$(a_{i_1}, \ldots, a_{i_k}) \in \mathcal{A}^k \longrightarrow a_{i_1} \cdots a_{i_k} \in \mathcal{A}^*.$$

On parlera indistinctement de *mots* ou de *messages* car la notion d'espace ne nous préoccupe pas ici. On pourra tout naturellement penser à l'exemple donné par des messages en français pour lesquels $\mathcal{A}$ est l'alphabet usuel enrichit de quelques symboles additionnels comme les signes de ponctuation par exemple.

Canal et signaux

Comment transmettre de tels messages à travers un canal de communication ? Et qu'est-ce qu'un « canal de communication » au juste ? Un *canal de communication sans bruit* fonctionne à la manière du télégraphe : seuls un nombre finis de *signaux* sont transmissibles. Pour le télégraphe, on dispose de trois signaux qui sont : le point, le tiret, et le temps de pause. De manière abstraite, un r-canal de communication sans bruit ne sait transmettre que r signaux différents $s_1, \ldots, s_r$, qui forment un alphabet $\mathcal{S} := \{s_1, \ldots, s_r\}$. Notre problème est alors d'utiliser ces r signaux – et donc l'alphabet $\mathcal{S}$ – pour transmettre des messages écrits dans l'alphabet $\mathcal{A}$, qui n'est pas $\mathcal{S}$. Lorsque $r \leqslant n$, la solution au problème est triviale car alors $\mathcal{A} \subset \mathcal{S}$. En général, et tout comme dans l'exemple du télégraphe, n est bien plus grand que r, et c'est là tout le problème. En informatique, r vaut la plupart du temps 2. En somme, nous cherchons en particulier un moyen pour communiquer rapidement du Shakespeare en utilisant 2 grognements différents... Comme nous allons le voir, cela est algorithmiquement possible et qui plus est de manière optimale !

Formalisation du problème du codage

Le problème du codage consiste à associer à chaque symbole $a_i \in \mathcal{A}$ un mot $c_i := s_{i,1} \cdots s_{i,r_i} \in \mathcal{S}^*$ écrit dans l'alphabet $\mathcal{S}$. On comprend aisément que la nature exacte des signaux s_i n'a absolument pas d'importance et que seul leur nombre r est pertinent, et il en est de même pour $\mathcal{A}$ et n. La suite $c := (c_1, \ldots, c_n) \in (\mathcal{S}^*)^n$ est appelée un (r,n)-*code*, et chaque c_i est le *code* de a_i. Un message $x := a_{i_1} a_{i_1} \cdots a_{i_m} \in \mathcal{A}^*$ sera donc codé en un message $c_{i_1} c_{i_2} \ldots c_{i_m} \in \mathcal{S}^*$ en remplaçant chaque occurrence du symbole a_i dans x par le mot $c_i \in \mathcal{S}^*$. Les c_i sont donc *concaténés* dans $\mathcal{S}^*$. Ce procédé fournit au final un *message codé* :

$$\underbrace{a_{i_1} a_{i_1} \cdots a_{i_m}}_{\text{message d'origine}} \in \mathcal{A}^* \overset{\text{codage}}{\longrightarrow} \underbrace{c_{i_1} c_{i_2} \ldots c_{i_m}}_{\text{message codé}} \in \mathcal{S}^*.$$

Le message codé n'est qu'une suite de signaux, qu'il faut ensuite *décoder* pour retrouver le message d'origine. La longueur du message d'origine $a_{i_1} a_{i_1} \cdots a_{i_m}$ dans $\mathcal{A}^*$ est m. En revanche, la longueur du message codé $c_{i_1} c_{i_2} \ldots c_{i_m}$ dans $\mathcal{S}^*$ n'est pas m mais $|c_{i_1}| + |c_{i_2}| + \cdots + |c_{i_m}|$.

Solution naïve au problème du codage

Une solution naïve au problème du codage consiste à choisir des codes c_i qui sont tous de même longueur, et il est alors facile de voir que cette longueur vaut $\log_r n$ à une unité près, en vertu des explications de la section 7.2. Choisir des codes qui sont tous de la même longueur n'est pas très satisfaisant car les symboles fréquents de $\mathcal{A}$ seront codés de la même manière que les symboles rarement utilisés. Ce problème a été résolu pour le télégraphe en utilisant des codes à base de tirets et de points de longueur inversement proportionnelle à la fréquence du symbole qu'ils codent dans une langue de référence (l'anglais, en l'occurrence). C'est le fameux code Morse inventé vers 1835, et pour lequel $r = 3$. Le décodage est garantit par l'utilisation du temps de pause pour séparer les codes. En effet, lorsque les codes ne sont pas de même longueur, il faut être en mesure de déterminer le début où la fin de chaque code lors de la réception des signaux.

Codes instantanés et à décodage unique

Poursuivons notre démarche quelque peu abstraite en donnant deux définitions liées à la possibilité de décoder un message codé en signaux par un (r,n)-code.

1. *Codes à déchiffrage unique.* Un (r,n)-code est dit *à déchiffrage unique* lorsque le décodage permet toujours de retrouver le message d'origine dans $\mathcal{A}^*$ de façon unique après réception de la totalité de la version codée du message dans $\mathcal{S}^*$;
2. *Codes instantanés.* Un (r,n)-code est dit *instantané* lorsque le décodage d'un message peut être fait au fur et à mesure lors de la lecture des signaux constituant la version codée du message.

Il est clair qu'un code instantané est à déchiffrage unique, et les codes instantanés sont de loin les plus pratiques. Kraft a montré en 1949 que si un (r,n)-code $(c_1, \ldots, c_n)$ est instantané, il satisfait à l'inégalité suivante :

$$\sum_{i=1}^{n} r^{-l_i} \leqslant 1,$$

où $l_i = |c_i|$. Réciproquement, si l'on se donne des entiers naturels $r, n, l_1, \ldots, l_n$ qui satisfont à la condition de Kraft, alors il existe un (r,n)-code instantané dont les longueurs de codes sont précisément ces $l_1, \ldots, l_n$. Enfin, Mac Millan a montré en 1956 que tout code à déchiffrage unique satisfait à la condition de Kraft.

L'utilisation du temps de pause dans le code Morse pour séparer les codes garantit le décodage instantané. En réalité un (r,n)-code $(c_1, \ldots, c_n)$ est instantané si et seulement si il a la *propriété de préfixe* : aucun c_i n'est le préfixe d'un c_j avec $j \neq i$. En d'autres termes, si $c_k = s_{i_1} \cdots s_{i_k}$, alors pour tout $j < k$, $s_{i_1} \cdots s_{i_j} \notin \{c_1, \ldots, c_n\}$.

L'entropie de Shannon et le problème du codage optimal

Supposons que les messages à coder dans $\mathcal{A}^*$ sont tels que la probabilité d'apparition du symbole a_i est p_i. On a alors une loi de probabilité discrète $(p_1, \ldots, p_n)$ sur $\mathcal{A}$ qui décrit les fréquences d'apparitions des symboles. Si $c := (c_1, \ldots, c_n)$ est un (r, n)-code, la longueur moyenne dans $\mathcal{S}^*$ du code d'un symbole de $\mathcal{A}$ par le code c est donnée par

$$\mathcal{L}(c) := \sum_{i=1}^{n} p_i \, |c_i|.$$

Soit $\mathcal{C}(r, n)$ l'ensemble des (r, n)-code instantanés. Le problème du *codage optimal* consiste à construire un (r, n)-code instantané $c_{\text{opt.}}$ tel que

$$\mathcal{L}(c_{\text{opt.}}) \simeq \inf_{c \in \mathcal{C}(r,n)} \mathcal{L}(c).$$

Il est facile de voir par compacité (exercice !) que cet infimum est atteint et qu'un tel code optimal $c_{\text{opt.}}$ existe. Naturellement, un tel code n'a aucune raison d'être unique (penser à la symétrie) et il dépendra essentiellement de r et de la loi de probabilité discrète $(p_1, \ldots, p_n)$ qui décrit la *source* des messages. Reste à trouver des méthode de constructions de codes optimaux, et cela fait l'objet des sections suivantes. Lorsque $\mathcal{A}$ est l'alphabet usuel, on pourra penser par exemple que $(p_1, \ldots, p_n)$ représente les fréquences d'utilisation des lettres dans les textes en français. Lorsque l'on a affaire à un message suffisamment long dans $\mathcal{A}^*$, on peut également imaginer que $(p_1, \ldots, p_n)$ est obtenue en calculant les fréquences empiriques des symboles de l'alphabet $\mathcal{A}$ dans le message en question. Le *théorème de codage non-bruité de Shannon* affirme que l'on a

$$\mathbf{I}(p_1, \ldots, p_n) \leqslant \inf_{c \in \mathcal{C}(2,n)} \mathcal{L}(c) \leqslant \mathbf{I}(p_1, \ldots, p_n) + 1.$$

La version pour un $r \neq 2$ s'obtient en remplaçant $\log_2$ par $\log_r$ dans la définition (7.3) de l'entropie de Shannon $\mathbf{I}$. En codant des blocs de k symboles de $\mathcal{A}$ plutôt que les symboles eux-mêmes, on obtient immédiatement en appliquant ce qui précède à $\mathcal{A}^k$ et en vertu du théorème 7.4.1 que pour tout $k \in \mathbb{N}^*$

$$\mathbf{I}(p_1, \ldots, p_n) \leqslant \inf_{c \in \mathcal{C}(2,n^k)} \frac{1}{k} \mathcal{L}(c) \leqslant \mathbf{I}(p_1, \ldots, p_n) + \frac{1}{k}.$$

Notons que $\frac{1}{k}\mathcal{L}(c)$ est toujours une longueur moyenne par symbole de $\mathcal{A}$ car $c \in \mathcal{C}(2, n^k)$ est un code qui concerne les symboles de $\mathcal{A}^k$ qui sont de longueur k dans $\mathcal{A}^*$. Ainsi, l'entropie de Shannon $\mathbf{I}(p_1, \ldots, p_n)$ mesure la longueur moyenne du meilleur $(2, n^k)$-code, et cette mesure est d'autant plus fine que k est grand. Les preuves de toutes les affirmations qui précèdent sont élémentaires et figurent par exemple dans le premier chapitre de [Rom97]. On pourra consulter également le chapitre 7 du très accessible [App96].

Codes de Shannon-Fano et de Huffman

FIXME: **Partie à intégrer**

```
function [code,arbre,entropie,longmoy]=huffman2(P)
%
% Determine le code de Huffman binaire (2-ary en anglais)
% qui correspond a la distribution de probabilite P.
%
%
%% Exemple 7.11 page 125 du Applebaum, cf. aussi figure 7.6.
%
% [code,arbre,entrop,longmoy]=huffman2([.14,.24,.33,.06,.11,.12])
%
```

```
% code =
%    [0.1400]    '101'
%    [0.2400]    '01'
%    [0.3300]    '11'
%    [0.0600]    '000'
%    [0.1100]    '001'
%    [0.1200]    '100'
%
% arbre =
%    0.1700    4.0000    5.0000
%    0.2600    6.0000    1.0000
%    0.4100    7.0000    2.0000
%    0.5900    8.0000    3.0000
%    1.0000    9.0000   10.0000
%
% entrop =
%   2.3800
%
% longmoy =
%    2.4300
%
%
%
%% Exemple 7.10 page 124 du Applebaum, cf. aussi figure 7.5.
%
% [code,arbre,entrop,longmoy]=huffman2([1/2,1/4,1/8,1/16,1/16])
%
% code =
%   [0.5000]    '0'
%   [0.2500]    '10'
%   [0.1250]    '110'
%   [0.0625]    '1110'
%   [0.0625]    '1111'
%
% arbre =
%   0.1250    4.0000    5.0000
%   0.2500    3.0000    6.0000
%   0.5000    2.0000    7.0000
%   1.0000    1.0000    8.0000
%
% entrop =
%   1.8750
%
% longmoy =
%    1.8750
%%

entropie = -sum(P.*log2(P));

n = length(P);
PP = [P 2*ones(1,n-1)];
arbre = zeros(n-1, 3);
for i = 1:n-1,
```

```
    [index1, index2] = findmins(PP);
    arbre(i, 1) = PP(index1) + PP(index2);
    PP(n + i) = arbre(i, 1);
    arbre(i, 2) = index1;
    arbre(i, 3) = index2;
    PP(index1) = 2;
    PP(index2) = 2;
end
%
codes = cell(length(PP),1);
codes{arbre(n-1,2)} = '0';
codes{arbre(n-1,3)} = '1';
for i = n-2:-1:1,
    codes{arbre(i,2)} = [ codes{i+n} '0' ];
    codes{arbre(i,3)} = [ codes{i+n} '1' ];
end

code = cell(n,2);
longmoy = 0.;
for i=1:n,
     code{i,1} = P(i);
     code{i,2} = codes{i};
     longmoy = longmoy + length(codes{i})*P(i);
end

return

%
function [index1, index2] = findmins(P)
%
min1 = 3;
min2 = 3;
n = length(P);
for i = 1:n,
    if P(i) < min1,
       index1 = i;
       min1 = P(i);
    end
end
P(index1) = 2;
for i = 1:n,
    if P(i) < min2,
       index2 = i;
       min2 = P(i);
    end
end
```

7.4.3 Entropie exponentielle de Shannon

L'entropie exponentielle de Shannon de la loi discrère $(p_1, \ldots, p_n)$ est définie par

$$2^{\mathbf{I}(p_1,\ldots,p_n)}. \tag{7.6}$$

Elle correspond au « nombre de degrès de liberté » évoqué au début de la section 7.1 sur l'entropie de Botlzmann en physique, ou tout simplement à $|\Omega|$ dans l'esprit de la section 7.2. Ainsi, l'entropie exponentielle d'une masse de Dirac vaut 1, celle d'une loi de Bernoulli symétrique vaut 2 et plus généralement, celle de la loi uniforme de taille n vaut n. Rappelons que parmis les lois discrètes de support de taille au plus n, la loi uniforme maximise l'entropie de Shannon. Ainsi, l'entropie exponentielle de Shannon $2^{\mathbf{I}(p_1,\dots,p_n)}$ donne la taille de la loi uniforme qui a la même entropie que la loi $(p_1,\dots,p_n)$. L'entropie exponentielle d'une loi de Bernoulli asymétrique est entre 1 et 2 car une telle loi favorise une des deux valeurs possibles et est donc « moins incertaine » qu'une loi de Bernoulli symétrique.

Un générateur pseudo-aléatoire idéal pour la loi uniforme sur $\{0,\dots,n-1\}$ devrait avoir une entropie exponentielle de n. Bien entendu, l'entropie des générateurs existants est toujours plus petite que cette valeur maximale.

7.4.4 Information mutuelle et capacité d'un canal de communication

La preuve du théorème 7.4.1 sur l'extensivité de l'entropie de Shannon est essentiellement basée sur le fait que pour deux variables aléatoires discrètes à valeurs dans $\{0,\dots,n\}$

$$\mathbf{I}(\mathcal{L}(X)\otimes\mathcal{L}(Y)) = \mathbf{I}(X)+\mathbf{I}(Y)-\mathbf{I}((X,Y)).$$

Cette propriété suggère de définir l'*information mutuelle* $\mathbf{M}(X,Y)$ de X et Y par la formule symétrique

$$\mathbf{M}(X,Y) := \mathbf{I}(X)+\mathbf{I}(Y)-\mathbf{I}((X,Y)) = \mathbf{I}(\mathcal{L}(X)\otimes\mathcal{L}(Y)).$$

Elle représente l'incertitude maximale que l'on peut obtenir pour un couple de v.a. dont les incertitudes marginales sont celles de X et de Y. Dans la même veine, on définit l'*entropie conditionnelle* $\mathbf{I}(X\,|\,Y)$ par la formule

$$\mathbf{I}(X\,|\,Y) := \mathbf{I}((X,Y))-\mathbf{I}(Y).$$

Elle représente en quelque sorte l'incertitude résiduelle dans X lorsque l'on connaît Y. On a alors en vertu de ce qui précède,

$$\mathbf{M}(X,Y) = \mathbf{I}((X,Y))-\mathbf{I}(X\,|\,Y)-\mathbf{I}(Y\,|\,X).$$

Dans l'esprit de la section 7.2, les relations entre information mutuelle et entropies conditionnelles se comprennent très facilement en terme d'ensembles, comme expliqué sur la figure 7.1.

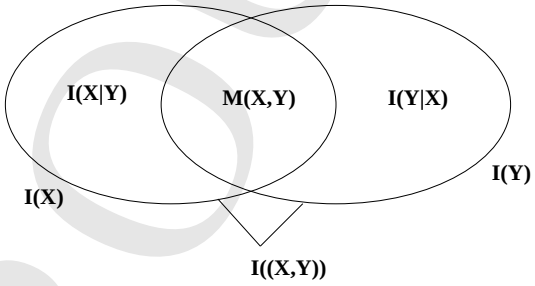

FIG. 7.1 – Relations entre entropie de Shannon, entropies conditionnelles, et information mutuelle.

FIXME: **Partie à intégrer**
définir la capacité

Théorème de codage bruité

FIXME: **Partie à intégrer**

7.4.5 Caractérisations axiomatiques de l'entropie discrète

L'entropie de Shannon définie en (7.3) possède de très belles caractérisations axiomatiques. Rappelons que Λ_n désigne le simplexe définit en (7.5). La seule famille de fonction $F_n : \Lambda_n \to \mathbb{R}$, $n \in \mathbb{N}^*$, qui vérifient

1. Pour tout $n \in \mathbb{N}^*$, F_n est symétrique, positive et continue;
2. $F_2(1/2, 1/2) = 1$;
3. Pour tout $n \in \mathbb{N}^*$ et $p \in \Lambda_n$, $F_n(p) = F_{n-1}(q, p_3, \ldots, p_n) + q\, F_2(p_1/q, p_2/q)$ où $q := p_1 + p_2$;

est l'entropie de Shannon $\mathbf{I}$ définie en (7.3). Il existe d'autres caractérisations du même type. Ainsi, les seules familles de fonctions $F_n : \Lambda_n \to \mathbb{R}$, $n \in \mathbb{N}^*$, qui vérifient pour tout $n \in \mathbb{N}^*$

1. F_n est positive et continue;
2. $F_n(1/n, \ldots, 1/n)) < F_{n+1}(1/(n+1), \ldots, 1/(n+1))$;
3. Pour tout $(n_1, \ldots, n_k) \in \mathbb{N}^{*k}$ avec $n_1 + \cdots + n_k = n$,

$$F_n\Big(\frac{1}{n}, \ldots, \frac{1}{n}\Big) = F_k\Big(\frac{n_1}{n}, \ldots, \frac{n_k}{n}\Big) + \sum_{i=1}^{k} \frac{n_i}{n} F_{n_i}\Big(\frac{1}{n_i}, \ldots, \frac{1}{n_i}\Big);$$

sont de la forme

$$F_n(p_1, \ldots, p_n) \;=\; -\sum_{i=1}^{n} p_i \log_b p_i,$$

avec $b \in \mathbb{R}_+^*$. Ces axiomes sont tout à fait naturels[5] et rendent d'une certaine façon l'entropie de Boltzmann-Shannon « canonique ». La démonstration de ce qui précède est élémentaire et a été donnée par Shannon lui même. On la trouvera par exemple dans les premières pages de [Rom97].

7.4.6 Maximum d'entropie discrète

Nous avons déjà vu dans la section 7.4 que l'entropie de Shannon est maximisée, à taille du support fixée, par la loi uniforme. Que se passe-t-il lorsque l'on ajoute une contrainte sur sa moyenne par exemple ? La réponse est donnée par le théorème suivant.

Théorème 7.4.2 (Maximum d'entropie). *Soit $\eta \in \mathbb{R}^n$. Pour tout $\eta_0 \in \mathbb{R}$ vérifiant $\min(\eta_1, \ldots, \eta_n) < \eta_0 < \max(\eta_1, \ldots, \eta_n)$, il existe $\beta \in \mathbb{R}$ tel que*

$$\max_{\substack{(p_1,\ldots,p_n)\in\Lambda_n \\ \eta_1 p_1 + \cdots + \eta_n p_n = \eta_0}} \mathbf{I}(p_1, \ldots, p_n) \;=\; \mathbf{I}\Big((q_1^\beta, \ldots, q_n^\beta)\Big),$$

où $(q_1^\beta, \ldots, q_n^\beta) \in \Lambda_n$ est définie par

$$q_k^\beta := (Z_\beta)^{-1}\, e^{-\beta\, \eta_k}$$

où $Z_\beta := \sum_{i=1}^n e^{-\beta\, \eta_i}$. De plus, β est unique lorsque les $\eta_1, \ldots, \eta_n$ ne sont pas tous égaux. Enfin, β à le signe de $(\eta_1 + \cdots + \eta_n)/n - \eta_0$, et vaut 0 en cas d'égalité.

[5] Pourquoi ? Exercice !

Démonstration. Pour alléger les notations, on notera $q^\beta := (q_1^\beta, \ldots, q_n^\beta)$ et $p := (p_1, \ldots, p_n)$ et enfin $\mathbb{E}_q(\eta) := \eta_1 q_1 + \cdots + \eta_n q_n$ pour tout $q \in \Lambda_n$. Il est clair que pour tout $\beta \in \mathbb{R}$, $q^\beta \in \Lambda_n$. Montrons que β peut être choisit tel que $\mathbb{E}_{q^\beta}(\eta) = \eta_0$. Pour cela, on considère la fonction $\varphi : \mathbb{R} \to \mathbb{R}$ définie par

$$\varphi(\beta) := \mathbb{E}_{q^\beta}(\eta) := \frac{\sum_{i=1}^n \eta_i \, e^{-\beta \eta_i}}{\sum_{i=1}^n e^{-\beta \eta_i}}.$$

Cette fonction est clairement de classe $\mathcal{C}^\infty$ et $\varphi(0) = (\eta_1 + \cdots + \eta_n)/n$. D'autre part, il existe toujours i dans $\{1, \ldots, n\}$ tel que $\eta_i = \min(\eta_1, \ldots, \eta_n)$, et l'on a alors

$$\varphi(\beta) := \frac{\eta_i + \sum_{j=1}^n \eta_j \, e^{-\beta(\eta_j - \eta_i)}}{1 + \sum_{j=1}^n e^{-\beta(\eta_j - \eta_i)}},$$

qui converge vers $\min(\eta_1, \ldots, \eta_n)$ lorsque β tend vers $+\infty$. De la même manière, on montre que $\varphi(\beta)$ converge vers $\max(\eta_1, \ldots, \eta_n)$ lorsque β tend vers $-\infty$. Le théorème des valeurs intermédiaires assure alors l'existence d'un $\beta \in \mathbb{R}$ tel que $\varphi(\beta) = \eta_0$. Pour montrer que β est unique, il suffit de montrer que φ est strictement décroissante, or nous avons par un simple calcul

$$\varphi'(\beta) := -\mathbf{Var}_{q^\beta}(\eta) \leqslant 0,$$

avec égalité si et seulement si les $\eta_1, \ldots, \eta_n$ sont tous égaux. Montrons à présent que si $p \in \Lambda_n$ vérifie $\mathbb{E}_p(\eta) = \eta_0$, alors $\mathbf{I}(p) \leqslant \mathbf{I}(q^\beta)$. On a

$$\mathbf{I}(q^\beta) - \mathbf{I}(p) = \sum_{i=1}^n \left(\frac{p_i}{q_i^\beta} \log \frac{p_i}{q_i^\beta} \right) q_i^\beta,$$

qui est exactement l'entropie relative de Kullback-Leibler $\mathbf{Ent}\left(p \,|\, q^\beta\right)$ définie en (2.6) page 51. Elle est donc positive et nulle si et seulement si $p = q^\beta$. □

On remarquera que lorsque les $\eta_1, \ldots, \eta_n$ sont tous égaux, la condition sur η_0 dans l'énoncé du théorème précédent n'est jamais satisfaite, et de toute manière, le problème revient alors à constater que l'entropie de Shannon est maximisée par la loi uniforme, ce que nous savons déjà. De plus, la contrainte $\mathbf{E}_p(\eta) = \eta_0$ imposée à l'entropie entraîne immédiatement que $\min(\eta_1, \ldots, \eta_n) \leqslant \eta_0 \leqslant \max(\eta_1, \ldots, \eta_n)$, ce qui montre donc que la condition imposée sur η n'est pas restrictive. Le cas où $\eta_0 = (\eta_1 + \ldots + \eta_n)/n$ est intéressant. Il entraîne qu'à taille du support fixée et à η-moyenne fixée, l'entropie de Shannon est maximisée par la loi uniforme. En langage géométrique, la contrainte imposée à l'entropie revient à la considérer comme une fonction définie sur l'intersection du simplexe Λ_n avec l'hyperplan passant par η_0 et orthogonal au vecteur η. On voit alors qu'il suffit de considérer les vecteurs η appartenant à la sphère de $\mathbb{R}^n$ pour la norme $\|\cdot\|_1$, et l'on a donc $\eta_0 \in]-1, +1[$.

Les lois de probabilités de la forme q_β sont appelées « lois de Boltzmann-Gibbs ». On peut énoncer un théorème similaire pour les lois de probabilités sur $\mathbb{N}$ ou $\mathbb{Z}$. En particulier, la loi géométrique de paramètre $p = 1/(m+1)$ sur $\mathbb{N}$ maximise l'entropie parmi toutes les lois sur $\mathbf{N}$ d'entropie finie et de moyenne m. Le théorème 122 possède une version semblable 7.5.1 pour l'entropie continue (7.4).

7.4.7 Entropie de Shannon et complexité de Kolmogorov

Revenons à présent à la notion de complexité de Kolmogorov introduire dans la section 1.1 page 20 et faisons le lien avec l'entropie de Boltzmann-Shannon. Soit $m \in \mathbb{N}^*$ un entier non nul, U une machine universelle et $\mathcal{P}_m$ l'ensemble des programmes écrits pour cette machine dont la longueur de la sortie est

de m. On note $\mathbf{K}(y|m)$ la complexité de Kolmogorov de taille m d'une suite $y = (y_1, \ldots, y_m)$ de longueur m, définie par :

$$\mathbf{K}(y \,|\, m) := \min_{p \in \mathcal{P}_m,\, s(p)=y} l(p).$$

Soit à présent une suite de variables aléatoires i.i.d. $(X_i)_{i\in\mathbb{N}}$ de loi discrète $p_1\delta_{a_1} + \cdots + p_n\delta_{a_n}$ où les $a_1, \ldots, a_n$ sont tous différents, et l'on note $\mathcal{A} := \{a_1, \ldots, a_n\}$ l'alphabet correspondant. Alors, on montre[6] qu'il existe une constante $c > 0$ telle que pour tout $m \in \mathbb{N}^*$:

$$\mathbf{I}(p_1, \ldots, p_n) \;\leqslant\; \frac{1}{m} \sum_{y \in \mathcal{A}^m} p_{y_1} \cdots p_{y_m}\, \mathbf{K}(y \,|\, m) \;\leqslant\; \mathbf{I}(p_1, \ldots, p_n) + \frac{n \log m}{m} + \frac{c}{m},$$

où $p_z := \mathbb{P}(X_1 = z_1, \ldots, X_m = z_m)$ pour tout $z \in \mathcal{A}^m$. Ainsi, on a :

$$\frac{1}{m}\, \mathbb{E}(\mathbf{K}((X_1, \ldots, X_m) \,|\, m)) \underset{m\to+\infty}{\longrightarrow} \mathbf{I}(p_1, \ldots, p_n).$$

L'entropie de Boltzmann-Shannon apparaît donc « asymptotiquement » comme une « complexité de Kolmogorov moyenne ». De ce point de vue, la complexité de Kolmogorov est est une notion plus fondamentale que l'entropie de Boltzmann-Shannon.

7.4.8 Quelques mots sur l'entropie en cryptographie

La méthode cryptographique la plus élémentaire – celle que tout le monde connait – est sans doute le code « César ». Elle consiste à chiffrer le message avec une permutation de l'alphabet, c'est-à-dire avec un élément du groupe symétrique $\mathbb{S}_{26}$, dont le cardinal est de l'ordre de $10^{25} = 10$ millions de milliards de milliards. Cette méthode aurait été utilisée par Jules César. Or on sait depuis le IX$^{\text{e}}$ siècle[7] que ce procédé est beaucoup trop simple pour être efficace. En effet, permuter les lettres de l'alaphabet ne fait que permuter leur fréquence d'apparition. Cette observation permet de retrouver facilement le message originel, s'il n'est pas trop court, en utilisant les fréquences d'apparition des lettres dans la langue du message. Ainsi, la permutation n'a pas modifié l'entropie de Shannon du message, qui dépend des $p_1, \ldots, p_{26}$ mais pas des $x_1 =$ 'a', $\ldots, x_{26} =$ 'z'. Claude Shannon a montré dans [Sha49] que l'entropie du message codé doit être plus élevée que celle du message originel pour éviter ce genre d'attaques (notion de diffusion/confusion). C'est sur ce principe que sont basés les algorithmes cryptographiques symétriques standards comme le déjà ancien DES (Data Encryption Standard) et le tout nouveau AES (Advanced Encryption Standard), cf. par exemple [PlS02].

Augmenter l'entropie d'un message correspond à le bruiter. Le principe de nombreuses méthodes cryptographiques dites *symétriques* est d'utiliser une fonction de bruitage qui dépend d'une clé que seuls l'émetteur et le récepteur possèdent. Ainsi, seule la bonne personne sera à même d'effectuer le débruitage. Le qualificatif *symétrique* rappelle que la même clé sert au bruitage (cryptage) et au débruitage (décryptage), et la fonction de bruitage utilisée est souvent son propre inverse, comme dans le DES. Les méthodes de cryptographie symétrique souffrent du problème de la communication de la clé entre l'émetteur et le récepteur. C'est ce qui a motivé l'introduction de méthodes *asymétriques*, comme l'algorithme RSA. Cet algorithme est basé sur l'arithmétique modulo[8] n, i.e. dans $\mathbb{Z}/n\mathbb{Z}$. Des nombres premiers de plusieurs milliers de chiffres sont alors nécessaires, ce qui nécessite l'utilisation d'algorithme probabilistes de test de primalité, bien plus rapides que les algorithmes déterministes.

Pour la petite histoire, le code César a été amélioré au 16$^{\text{e}}$ siècle par Vigenère. La nouvelle méthode consistait à utiliser une permutation de l'alphabet différente à chaque lettre du message à coder, le choix

[6] Cf. [CT91, théorème 7.3.1 page 154] ou [LV97] pour une preuve.

[7] Le fameux « Manuscrit sur le déchiffrement des messages cryptographiques » d'Al-Kindi.

[8] Cela constitue d'ailleurs un développement classique de certaines leçons orales d'algèbre... Pour rester dans l'informatique et les corps finis, vous pouvez aussi penser aux codes correcteurs d'erreur, cf. [Rom92] et [Rom97] par exemple.

de la permutation se faisant de façon cyclique selon une clé préétablie de quelques lettres qui donne la translation de l'alphabet à utiliser. Ce code a résisté longtemps à la cryptanalyse mais a été cassé par Charles Babbage vers 1850 en utilisant la périodicité engendrée par l'utilisation répétée de la clé.

Les cryptalgorithmes symétriques incassables existent. L'un des plus simples est sans doute l'algorithme probabiliste *symétrique* inventé par Vernam, qui a été utilisé en particulier pour le « téléphone rouge » entre Washington et Moscou pendant la guere froide. Il consiste à reprendre la méthode de Vigenère, mais en utilisant pour chaque lettre du message à coder un alphabet de substitution aléatoirement choisit. La clé est alors la suite aléatoire utilisée, qui est aussi longue que le message, et qui ne doit bien entendu servir qu'une fois. On peut facilement imaginer d'autres variantes : si le message à crypter est converti en base 2, il devient une séquence $(x_n)_n$ de 0 et de 1. La méthode consiste alors à combiner cette suite à une suite aléatoire $(a_n)_n$ constituée de réalisations i.i.d. de loi de Bernoulli symétrique $(\delta_0+\delta_1)/2$. Le message crypté $(y_n)_n$ est alors donné par $y_n = x_n \text{XOR } a_n$, où XOR est la notation standard du « OU exclusif ». Pour décrypter, il suffit de répéter l'opération avec la même suite $(a_n)_n$, qui constitue la clé. Cette clé est malheureusement aussi longue que le message... Ces méthodes probabilistes sont très sûres à condition de les pratiquer avec rigueur : clé à usage unique, générateur aléatoire de bonne qualité avec une bonne « entropie », etc. Notre société de l'information numérique fait que la cryptographie est en plein essor, et il se trouve même que nous l'utilisons parfois à notre insu. Son champ d'application est très vaste : sécurisation des échanges commerciaux sur Internet, authentification des cartes bancaires, sécurisation et signature des transferts de données et des messages électroniques, verouillage des portes et des moteurs de voitures, ...

7.5 Quelques propriétés de l'entropie continue

L'entropie de Shannon continue est définie en (7.4). On notera $\mathbf{I}(X)$ l'entropie $\mathbf{I}(f)$ d'un vecteur aléatoire de densité f par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$. Certaines propriétés de l'entropie discrète (7.3) sont perdues. Ainsi par exemple, l'entropie continue peut prendre toutes les valeurs de $\mathbb{R}$. En effet, si X est un vecteur aléatoire de $\mathbb{R}^d$ dont la loi a une densité par rapport à la mesure de Lebesgue, on a pour tout $\alpha \in \mathbb{R}$:

$$\mathbf{I}(\alpha\, X) = \mathbf{I}(X) + d \log_2 |\alpha|,$$

et l'entropie de la loi gaussienne standard n'est pas nulle et vaut $\frac{d}{2}\log_2(4\pi)$. L'entropie continue est invariante par translations : $\mathbf{I}(X+\alpha) = \mathbf{I}(X)$, et cette propriété est le pendant de l'invariance de l'entropie discrète par rapport au support de son argument. Tout comme l'entropie discrète, l'entropie continue est extensive :

$$\mathbf{I}((X,Y)) \leqslant \mathbf{I}(X) + \mathbf{I}(Y),$$

avec égalité si et seulement si les deux vecteurs aléatoires à densité X et Y sont indépendants. La preuve est identique à celle du théorème 7.4.1 pour le cas discret.

Comme nous l'avons montré précédemment, l'entropie continue, contrairement à l'entropie discrète, n'a pas de maximum et peut prendre toutes les valeurs dans $\mathbb{R}$. Comme $0 \log 0 = 0$, l'entropie d'une densité f à support inclus dans le compact $K := [a_1, b_1] \times \cdots \times [a_d, b_d] \subset \mathbb{R}^d$ a un sens lorsque $f \log f$ est Lebesgue intégrable sur K. Il se trouve que parmi toutes les lois de probabilités sur $\mathbb{R}^d$ dont le support est dans K, l'entropie continue est maximisée par la loi uniforme $\mathcal{U}_K$ sur K. En effet, si μ est une loi de probabilité à support dans K et de densité f par rapport à $\mathcal{U}_K$, on a en notant $|K|$ la mesure de Lebesgue de K

$$\mathbf{I}(\mathcal{U}_K) - \mathbf{I}(\mu) = \log|K| + \int_K f(x) \log f(x)\, dx = \int_K f(x) \log(|K|\, f)\, dx = \mathbf{Ent}(\mu \,|\, \mathcal{U}_K)\,,$$

où le membre de droite est l'entropie relative de Kullback-Leibler, cf. remarque 7.3.1. On a donc

$$\mathbf{Ent}(\mu \,|\, \mathcal{U}_K) \geqslant 0,$$

avec égalité si et seulement si $\mu = \mathcal{U}_K$. Ainsi, la prescription d'un support compact pour l'entropie continue est l'analogue de la prescription de la taille du support pour l'entropie discrète, ce qui n'est pas surprenant. De manière générale, l'entropie continue a un maximum sous contrainte de « moment » sur tout $\mathbb{R}^d$, comme l'exprime le théorème suivant. On notera que cette contrainte de moment est *linéaire en la loi*, exactement comme l'est la contrainte sur l'entropie discrète dans le théorème 7.4.2.

Théorème 7.5.1 (Maximum d'entropie). *Soit $\eta : \mathbb{R}^d \to \mathbb{R}$ une fonction mesurable telle que $\exp(-\eta)$ soit Lebesgue intégrable sur $\mathbb{R}^d$. Soit μ_η la loi de probabilités sur $\mathbb{R}^d$ de densité $(Z_\eta)^{-1} \exp(-\eta)$ où Z_η est la constante de normalisation. Supposons que η soit μ_η intégrable, et notons η_0 sa moyenne. Soit $\mathcal{V}_d$ l'ensemble des vecteurs aléatoires de $\mathbb{R}^d$ dont l'entropie de Shannon (7.4) est bien définie et est est finie. On a alors*

$$\max_{\substack{X \in \mathcal{V}_d \\ \mathbb{E}(\eta(X)) = \eta_0}} \mathbf{I}(X) \;\leqslant\; \mathbf{I}(\mu_\eta) \;:=\; \eta_0 + \log_2 Z_\eta,$$

et le maximum n'est atteint que pour les vecteurs aléatoires de loi μ_η.

Démonstration. La preuve est similaire à celle du théorème 7.4.2. On a

$$\mathbf{I}(\mu_\eta) - \mathbf{I}(X) = \mathbf{Ent}(\mathcal{L}(X) \,|\, \mu_\eta)\,,$$

où le membre de droite est l'entropie relative de Kullback-Leibler, cf. remarque 7.3.1. □

Les lois de probabilités de la forme μ_η sont appelées « lois de Boltzmann-Gibbs », et la fonction η est parfois appelée « potentiel » ou encore « hamiltonien »[9]. Contrairement au cas discret du théorème 7.4.2, il n'est pas toujours possible de paramétrer la contrainte par un paramètre β. C'est cependant possible dans un certain nombres de cas. En voici quelques exemples importants.

1. **Lois exponentielles.** Lorsque $\eta(x) = \lambda \|x\|_1$ avec $\lambda \in \mathbb{R}_+^*$, la loi μ_η qui est une loi de Laplace $\mathcal{E}(\lambda)$ (ou double-exponentielle) de paramètre λ. On a alors $\eta_0 = \lambda$, et le théorème exprime que les lois de Laplace maximisent l'entropie à moyenne fixée. Notons que dans ce cas, $\mathbf{I}(\eta) = \lambda + d \log_2(2\lambda^{-1})$;
2. **Lois gaussiennes.** Lorsque $\eta(x) = \frac{1}{2}\langle \Sigma^{-1} x\,,\, x\rangle$ où Σ est une matrice symétrique $d \times d$ définie positive, la loi η est une loi gaussienne $\mathcal{N}(0, \Sigma)$. On a alors $\eta_0 = \frac{d}{2}$, et le théorème exprime que les lois gaussiennes de covariance Σ maximisent l'entropie à trace de la matrice de covariance fixée. Par matrice de covariance du vecteur X, on entend $(\mathbf{Cov}(X_i, X_j))_{1 \leqslant i,j \leqslant d}$. Notons que dans ce cas, $\mathbf{I}(\eta) = \frac{d}{2} \log_2 \left(4\pi \mathrm{Det}(\Sigma)^{1/d}\right)$.
3. **Lois gamma.** Tout est explicitable en termes de la fonction gamma. Exercice ! On devrait retrouver la loi exponentielle comme cas particulier;
4. **Lois de Weibull.** Idem.

Remarque 7.5.2 *(Principe de maximum d'entropie).* Le *principe général de maximum d'entropie*, énoncé par exemple par Jaynes dans les années 1950, consiste à dire que la loi de probabilité à choisir parmi un ensemble de lois possibles lors d'une modélisation doit être celle qui maximise l'entropie. Les contraintes proviennent alors de l'information dont on dispose. Par exemple, sous contrainte de support, on considérera une loi uniforme, sous contrainte de moyenne, on choisira une loi exponentielle, sous contrainte de variance, on considèrera une loi gaussienne, etc.

Les lois gaussiennes sont très importantes dans les applications, et remplacent sur tout $\mathbb{R}^d$ les lois uniformes. Par analogie avec le cas discret (7.6), on définit l'entropie exponentielle de Shannon $\mathbf{N}(X)$ du vecteur aléatoire X de la manière suivante :

$$\mathbf{N}(X) := \frac{1}{4\pi} 2^{\frac{2}{d}\mathbf{I}(X)}. \tag{7.7}$$

[9] Et pourquoi donc ?

De cette manière $\mathbf{N}(\mathcal{N}(m,\Sigma)) = \mathrm{Det}(\Sigma)^{1/d}$. L'entropie exponentielle de X donne donc l'écart type de la loi gaussienne standard de même entropie que X. Comme le déterminant est un d-volume, $|\Sigma|^{1/d}$ s'interprette comme une sorte de « rayon d'incertitude ». Le maximum d'entropie pour la loi gaussienne s'exprime alors de la façon suivante : pour tout vecteur aléatoire X de covariance Σ,

$$\mathbf{N}(X) \leqslant \mathrm{Det}(\Sigma)^{1/d}.$$

De manière générale, on peut définir l'équivalent de l'entropie exponentielle pour chaque loi η dans le théorème 7.5.1. Lorsque la loi de X a un support compact K dans $\mathbb{R}^d$, il serait par exemple préférable d'utiliser comme référence la loi uniforme sur K et de redéfinir $\mathbf{N}(X)$ en conséquence. L'entropie $\mathbf{I}(\mathcal{U}_K)$ de la loi uniforme $\mathcal{U}_K$ sur K valant $\log|K|$, on pourrait donc définir $\mathbf{N}_{\mathcal{U}_K}(X) := \exp(\mathbf{I}(X))$. Cependant, dans la pratique, les lois continues à support compact sont beaucoup moins utilisées que les lois gaussiennes.

L'entropie exponentielle de Shannon définie en (7.7) possède une propriété de sous-additivité très importante appelée « inégalité de l'entropie exponentielle de Shannon », qui s'exprime comme suit : pour tous vecteurs aléatoires *indépendants* X et Y de $\mathbb{R}^d$, d'entropies bien définies et finies, on a

$$\mathbf{N}(X+Y) \geqslant \mathbf{N}(X) + \mathbf{N}(Y),$$

avec égalité si et seulement si leur lois sont gaussiennes et de covariances proportionnelles. Cette inégalité joue un rôle clé dans l'établissement du théorème de codage bruité de Shannon, qui sort du cadre de notre propos. Nous renvoyons à [CT91] et [Rom97] pour en savoir plus.

FIXME: **Partie à intégrer**

Donner la preuve du TCL via max d'entropie, cf. Barron.

Un peu d'Histoire. La notion d'entropie a été introduite officiellement en thermodynamique au milieux du XIX$^{\mathrm{e}}$ siècle par Clausius pour compléter le principe de conservation de l'énergie. La théorie cinétique des gaz et plus généralement la mécanique statistique on été étudiées en particulier par Maxwell, Gibbs, Kelvin et Boltzmann pendant la seconde moitiée du XIX$^{\mathrm{e}}$ siècle. Boltzmann a été le premier à obtenir la formulation « probabiliste » de l'entropie que nous avons donnée ici, qui découle d'une hypothèse de quantification des états microscopiques. Cette idée fructueuse de quantification microscopique a sans doute été source d'inspiration pour Planck, qui en introduisant l'hypothèse quantique vers 1900 pour expliquer le rayonnement du corps noir, a jetté les bases de ce qui deviendra plus tard la mécanique quantique. D'une certaine manière, la physique a devancé les mathématiques en utilisant des idées probabilistes bien avant que les probabilités ne soient formalisées – par Kolmogorov, dans les années 1930 – et la « mécanique statistique » serait sans doute appelée aujourd'hui « mécanique stochastique ». L'entropie de Boltzmann a été utilisée par Shannon en 1948 dans un article devenu célèbre[10], pour les besoins de ce qui sera appelé plus tard « théorie de l'information » ou encore « théorie de la communication ». Cette théorie va accompagner la naissance de l'informatique pendant la seconde moitiée du XX$^{\mathrm{e}}$ siècle. Le lecteur en trouvera une présentation générale dans l'ouvrage de référence [CT91] par exemple. Shannon préférait à l'origine le terme « incertitude », mais a été convaincu par John Von Neumann d'utiliser le terme « entropie » car son « incertitude » n'est rien d'autre au signe près que l'entropie de Boltzmann. Le lecteur pourra trouver par exemple dans [Zin96] et [Oll02] quelques pages accessibles sur certains aspects mathématiques de l'entropie. Le principe de maximum d'entropie joue un rôle considérable aussi bien en physique qu'en statistique, et peut être utilisé en particulier pour justifier l'usage des lois gaussiennes et exponentielles. L'entropie de Boltzmann-Shannon apparaît aujourd'hui comme un cas particulier (important !) de l'entropie relative de Kullback-Leibler (2.6), très utile en théorie des probabilités, et qui intervient en particulier dans le principe de grandes déviations de Sanov, cf. théorème 2.5.1 page 51.

[10] Cf. [Sha48] et `http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html`.