

Section 3

SOURCES SYMBOLIQUES CLASSIQUES**3.1. Introduction : définitions**

Nous utiliserons les notations de la section précédente, proches des notations utilisées par Rényi [Rényi, 1962, 1992].

On considère une source symbolique Σ de messages σ constitués de chaînes de m signes ξ pris dans une bibliothèque Ξ de N symboles χ_1, \dots, χ_N :

$$\xi \in \Xi = \{\chi_1, \dots, \chi_i, \dots, \chi_N\}$$

$$\sigma = \xi_1 \dots \xi_k \dots \xi_m \in \Sigma \subset \Xi^*$$

ξ est considéré comme une variable aléatoire à valeur dans l'ensemble Ξ . On suppose que la répartition statistique observée des symboles χ_i dans le message σ est la suivante :

χ_1 présent k_1 fois dans σ : fréquence k_1 / m

.....

χ_N présent k_N fois dans σ : fréquence k_N / m

avec :

$$\sum_{i=1}^N k_i = m$$

On note $\Theta(\sigma)$ la quantité d'information ou *masse informationnelle* du message σ . On pose :

$$\Theta = V(\sigma) \cdot H(\xi)$$

où :

- $V(\sigma)$ est une grandeur extensive désignant le *volume d'information* contenue dans le message σ . Nous adopterons dans la suite la définition de la fonction de volume caractérisant le message par la quantité :

$$V = l(\sigma) = m$$

où $l(\sigma)$ est la longueur du message exprimée en nombre de signes.

- $H(\xi)$ est une grandeur intensive désignant la *densité d'information diacritique* [Oswald 1986], information contenue dans un volume élémentaire de référence. Si on choisit ce volume

élémentaire comme étant égal à un signe ($V = 1$), alors $H(\xi)$ désigne la densité diacritique moyenne par symbole, ou *entropie*, contenue dans le signe ξ .

Remarque :

En toute rigueur, la répartition statistique des symboles est normalement calculée sur l'ensemble des messages lus ou écrits par Σ , qui peuvent être considérés comme la réunion (par concaténation) en un seul message σ' de longueur m' . Mais, quelle que soit la manière de considérer ce qui est lu ou écrit par la source, ce qui évolue est la variable ξ , à qui sont affectées successivement des valeurs prises dans la bibliothèque Ξ .

3.2. Entropie statistique d'une source unique

3.2.1. Approche intuitive

i) Soit un message de volume unité, émis par une source dont la bibliothèque comprend $N = 2^\Theta$ symboles χ_i distincts ($i = 1, \dots, N$). En l'absence de toute autre information que posséderait éventuellement l'observateur sur cette source, la masse informationnelle est égale au nombre *d'étapes* (ou au nombre de *questions*, ou, à une constante près, au nombre de *pas de calcul*) nécessaires pour déterminer le symbole contenu dans le message en exécutant un algorithme dichotomique arborescent :

$$\Theta = \log N + O(1)$$

Remarques :

- on prendra arbitrairement : $O(1) = 0$ (ce qui suppose une machine construite pour exécuter un pas de calcul par alternative binaire).

- ce résultat se généralise au cas où N n'est pas une puissance de deux.

ii) Soit un message de volume m . La fréquence du symbole χ_i est notée k_i . On a :

$$m = \sum_{i=1}^N k_i$$

Si l'observateur constate que tous les symboles ont, dans le message, la même fréquence, on peut écrire :

$$k_i = k \forall i \Rightarrow k = m/N \Leftrightarrow N = m/k$$

La masse informationnelle totale vaut :

$$\Theta = m \cdot \log N = m \cdot \log \frac{m}{k}$$

Remarque : la densité diacritique est, dans ce cas particulier, égale à : $H(\xi) = \log(m/k)$. Elle s'exprime en unité de masse par unité de volume élémentaire, soit en *shannon/symbole* (abréviation : "sh/symb").

iii) Si l'observateur constate que les fréquences sont différentes, la masse globale du message est la somme des contributions en masse θ_i due à chaque symbole, qui contribue pour un volume k_i au message, avec une densité $\log m/k_i$ sh/symb. On remarque que cette dernière quantité est d'autant plus grande que le caractère est plus rare dans le message.

$$\theta_i = k_i \cdot \log \frac{m}{k_i} \Rightarrow \Theta = \sum_{i=1}^N \theta_i = \sum_{i=1}^N k_i \log \frac{m}{k_i}$$

La densité diacritique moyenne par symbole, exprimée en sh/symb, est donc :

$$H(\xi) = \frac{\Theta}{V} = \frac{\Theta}{m} = \sum_{i=1}^N \frac{k_i}{m} \log \frac{m}{k_i}$$

Si on définit les probabilités comme limites des fréquences lorsque $m \rightarrow \infty$ on aboutit à la formule classique :

$$\frac{k_i}{m} \xrightarrow{m \rightarrow \infty} p(i) \Rightarrow H(\xi) = - \sum_{i=1}^N p(i) \log p(i)$$

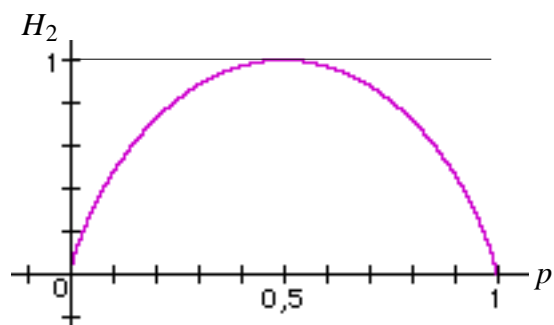
Plus précisément, si l'on considère la quantité H comme l'espérance mathématique des densités $\log(1/p_i)$ des différents états possibles :

$$H = \mathbb{E} \left(\log \frac{1}{p_i} \right) = - \sum_i p_i \log p_i$$

Exemple :

Soit une source binaire telle que $\Xi = \{\emptyset, 1\}$. Soit $p(\emptyset) = p$, $p(1) = 1 - p$. On note $H_2(p)$ l'entropie de cette source. Il vient :

$$H_2 = -p \log p - (1-p) \log(1-p)$$



• Figure 3.1

On montre [voir par exemple Shannon 1948, Giasu et Theodorescu 1971, Oswald 1986] que la convergence de H vers $-\sum_i p_i \log p_i$ est une conséquence de la loi des grands nombres :

$$\forall \delta, \varepsilon, p \left(\left| \sum_i \frac{k_i}{m} \log \frac{m}{k_i} - H \right| \leq \delta \right) > 1 - \varepsilon$$

et qu'en conséquence il est légitime de remplacer les fréquences par des probabilités.

3.2.2. Approche axiomatique

On peut aussi appliquer à la quantité d'information une définition purement axiomatique. Nombre d'auteurs, après Shannon, en partant de certaines des propriétés exposées ci-dessus prises comme axiomes, aboutissent à une formulation commune de l'entropie [voir par ex. Khintchine 1957, Faddeev 1956, Kullback 1958, Rényi 1962], et établissent l'axiomatique de l'information à partir des probabilités. Il est intéressant de constater que des axiomatiques fort différentes aboutissent à une même formulation de l'information.

Réciproquement, d'autres travaux [voir par ex. Ingarden et Urbanik 1962, Kampé de Fériet et Forte 1967, Giascu et Theodorescu 1971] proposent de déduire la probabilité de l'information. Ces auteurs définissent l'information sur une classe d'algèbres booléennes d'où il est possible de construire une axiomatique des événements élémentaires en attachant à chaque élément une classe de probabilité (mais en réalité cette probabilité est remplaçable par toute mesure mathématique ayant les mêmes propriétés). Cette démarche s'est avérée féconde, puisqu'elle conduit au principe du maximum d'entropie, important théorème en théorie de la décision [Réfrégier 1993].

Schématiquement, l'axiomatisation de l'information à partir des probabilités se présente ainsi :

i) On considère un système Σ dont les N états sont supposés : *identifiés* (on en connaît la liste), *de probabilités* $p_1, p_2, \dots, p_i, \dots, p_N$ *connues*.. Soit $H(p_1, p_2, \dots, p_N)$ la fonction attachée à ces probabilités, qui possède les propriétés axiomatiques suivantes :

□ Axiomes

- A1) H est une fonction continue des p_i

- A2) H est maximale lorsque les événements sont équiprobables (cela correspond à la plus grande incertitude) :

$$H_N(p_1, \dots, p_N) \leq H_N\left(\frac{1}{N}, \frac{1}{N}, \dots, \frac{1}{N}\right)$$

- A3) Si H est une fonction composée de fonctions de probabilités élémentaires, alors cette composition est une somme pondérée par les probabilités de ces fonctions (suivant la loi des probabilités composées). Par exemple, Khintchine pose :

$$p_i = \sum_{j=1}^M r_{ij} \quad \text{avec : } r_{ij} > 0, \quad 1 \leq i \leq N, \quad 1 \leq j \leq M, \quad \sum_{i=1}^N \sum_{j=1}^M r_{ij} = 1$$

$$H_{NM}(r_{11}, r_{12}, \dots, r_{NM}) = H_N(p_1, \dots, p_N) + \sum_{i=1}^N p_i \cdot H_M\left(\frac{r_{i1}}{p_i}, \dots, \frac{r_{iM}}{p_i}\right)$$

▣

ii)

□ Démonstration. Sans détailler complètement la démonstration, on peut en donner succinctement les grandes lignes. Notons : $h(N) = H_N\left(\frac{1}{N}, \frac{1}{N}, \dots, \frac{1}{N}\right)$. L'application de

A3 pour $N=S$ et $M=S$ donne dans le cas de l'équiprobabilité : $p_i = \frac{1}{S}$; $r_{ij} = \frac{1}{S^2}$; $h(S^2) =$

$2 h(S)$. Par récurrence on aboutit à l'équation fonctionnelle :

$$h(S^\alpha) = \alpha \cdot h(S)$$

dont on déduit que h doit être de la forme (avec $\lambda > 0$) :

$$h(x) = \lambda \cdot \log_S x$$

Supposons que l'on puisse décomposer les probabilités p_i si celles-ci sont des rationnels quelconques, ayant même dénominateur :

$$p_i = \frac{q_i}{q} \quad \text{avec } 1 \leq i \leq N ; \quad \sum_{i=1}^N p_i = 1 \Rightarrow \sum_{i=1}^N q_i = q$$

En utilisant à nouveau l'axiome A3, il y a deux façons de calculer H :

$$\bullet H(r_{11}, \dots, r_{NM}) = H_q \left(\underbrace{\frac{1}{q}, \dots, \frac{1}{q}}_{q \text{ fois}} \right) = h(q)$$

$$\bullet H(r_{11}, \dots, r_{NM}) = H(p_1, \dots, p_N) + \sum_{i=1}^N p_i H_{q_i} \left(\frac{1}{q_i}, \dots, \frac{1}{q_i} \right)$$

~~1 2 3~~
q_i fois

$$\Leftrightarrow H(r_{11}, \dots, r_{NM}) = H(p_1, \dots, p_N) + \sum_{i=1}^N p_i h(q_i)$$

D'où :

$$h(q) = H(p_1, \dots, p_N) + \sum_{i=1}^N p_i h(q_i)$$

Or : $h(q) = \lambda \cdot \log q$

et $h(q_i) = \lambda \cdot \log q_i = \lambda \cdot \log p_i + \lambda \cdot \log q$

En utilisant ces relations dans l'expression précédente, il vient :

$$H(p_1, \dots, p_N) = -\lambda \sum_{i=1}^N p_i \log p_i \quad \text{avec } \lambda > 0.$$

Cette fonction étant supposée continue, cette relation s'étend aux nombres irrationnels.



La formule de Shannon, qui résulte d'une évaluation statistique de la probabilité d'apparition de chaque symbole pris séparément, trouve son origine dans une démarche intuitive basée sur l'évaluation fondamentale en $\log N$ de la quantité d'information. Ce choix est ainsi confirmé par une démarche axiomatique rigoureuse : c'est bien le seul choix possible.

3.2.3. Calcul approché

Une autre approche de la définition de la densité diacritique moyenne consiste à calculer la masse informationnelle d'après la $m^{\text{ème}}$ extension Ξ^m de la bibliothèque Ξ , où chaque chaîne ou "super-signe" est lui-même un message σ . Soit N_m le nombre d'états de cette bibliothèque étendue. Le message étant formé d'un unique "super-signe", on a, en admettant toujours que l'observateur ne possède aucune autre information :

$$\Theta = \log N_m$$

Pour calculer N_m , il vient, par un calcul classique :

i) nombre de façons de disposer k_1 caractères dans une chaîne de longueur m , ou nombre de combinaisons de m objets pris k_1 à k_1 :

$$\binom{k_1}{m} = \frac{m!}{k_1!(m-k_1)!}$$

Il reste $m-k_1$ espaces disponibles. Il y a $\binom{k_2}{m-k_1}$ façons de disposer k_2 caractères

dans $m-k_1$ espaces. Etc. Il vient :

$$N_m = \binom{k_1}{m} \binom{k_2}{m-k_1} \binom{k_3}{m-k_1-k_2} \cdots \binom{k_{N-1}}{m-\sum_{i=1}^{N-2} k_i}$$

$$N_m = \frac{m!}{k_1!(m-k_1)!} \cdot \frac{(m-k_1)!}{k_2!(m-k_1-k_2)!} \cdots \frac{\left(m-\sum_{i=1}^{N-2} k_i\right)!}{k_{N-1}!k_N!} \quad \text{avec} \quad m-\sum_{i=1}^{N-1} k_i = k_N$$

$$\Rightarrow N_m = \frac{m!}{\prod_{i=1}^N k_i!}$$

$$\Rightarrow \Theta = \log m! - \sum_{i=1}^N \log k_i!$$

ii) Pour calculer cette dernière expression, il faut employer l'approximation de Stirling :

$$x! = x^x \cdot e^{-x} \cdot \sqrt{2\pi x} \cdot \left(1 + \frac{1}{12x} + \frac{a}{x^2} + \dots\right)$$

Dans cette expression, la parenthèse tend vers 1 si x augmente indéfiniment. Donc, à une constante $O(1)$ près, qui tend vers zéro lorsque m augmente, il vient :

$$\Theta = \log N_m \approx 1,44 \cdot \ln N_m$$

$$\Rightarrow \Theta / 1,44 = m \cdot \ln m - m + \frac{1}{2} \ln 2\pi m - \sum_{i=1}^N \left(k_i \cdot \ln k_i - k_i + \frac{1}{2} \ln 2\pi k_i \right) + O(1)$$

Soit, en réarrangeant les termes :

$$\Rightarrow \Theta = -m \cdot \sum_{i=1}^N \frac{k_i}{m} \cdot \log \frac{k_i}{m} - \frac{N-1}{2} \log 2\pi m - \frac{1}{2} \sum_{i=1}^N \log 2\pi \frac{k_i}{m} + O(1)$$

En restreignant à ses deux premiers termes l'approximation de Stirling, Brillouin aboutit par ce calcul à la formule définie précédemment. Mais la quantité d'information réelle, calculée sur la totalité du message, est en réalité *plus faible* que l'estimation que fournit la formule de

Shannon, d'un terme en $O\left(\frac{N}{2} \log 2\pi m\right)$ [voir par ex. Collot *et al.*, 1977a et 1977b]. Ce calcul

est conduit globalement, sur la $m^{\text{ème}}$ extension de Ξ , *en connaissant d'avance* la structure statistique du message, c'est-à-dire les fréquences k_i de chaque symbole. Opérant sur des chaînes et non des symboles, ce calcul fait apparaître une distinction fondamentale dans l'évaluation de la quantité d'information portée par un message selon que l'on s'intéresse aux symboles ou aux chaînes de symboles, en tenant compte ou non de la structure de celles-ci. C'est précisément cette différence qui rend le codage et la compression d'information possibles. Mais cette différence concerne une information implicite (i.e. qui n'est pas transmise) sur la connaissance du fait qu'un message constitué d'un ensemble de signes est structuré, ou non, en mots indépendants et successifs selon une syntaxe donnée.

On pourrait imaginer par exemple une source de mots de longueur $n = 10$, avec un alphabet de 10 symboles équiprobables, où chaque symbole serait employé une fois et une seule dans chaque mot. *Si l'on ignore cette structure particulière des mots*, la formule de Shannon conduit à estimer l'information moyenne portée par chaque mot à $H = \log 10^{10} = 33,2$ sh/mot. Alors que dans le calcul de Brillouin, il y a seulement $10!$ façons d'écrire un mot long de dix symboles tous différents, ce qui conduit à $H = \log 10! = 21,8$ sh/mot. En portant la longueur des mots à $n = 100$, on trouve respectivement 332 et 306 sh/mot. Et ceci alors que, si le message est suffisamment long, les fréquences des symboles sont les mêmes dans les deux calculs.

Clairement, une connaissance des caractéristiques structurelles du message plus fine que la simple connaissance de la répartition statistique des symboles influe de façon implicite sur la masse informationnelle que l'observateur peut en extraire. Nous voyons dans cette remarque un premier indice de la présence d'un contenu d'identification, dont tient compte implicitement l'observateur lorsqu'il connaît préalablement certaines caractéristiques de la source.

3.2.4. Propriétés

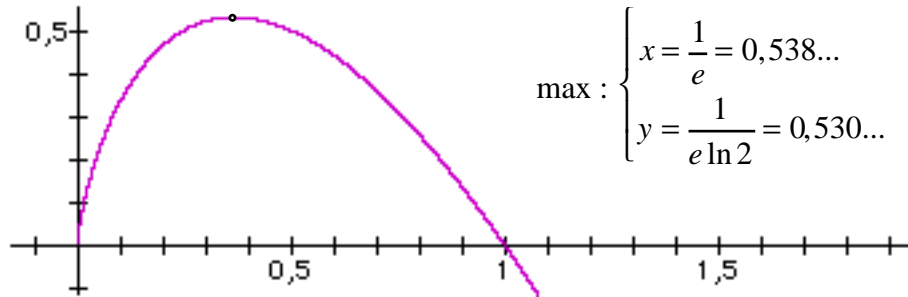
Théorème 3.1 : positivité de l'entropie :

$$H(p_1, p_2, \dots, p_N) \geq 0$$

H est nulle si l'une des probabilités est égale à 1 : l'un des événements est certain (toutes les autres probabilités sont nulles), et sa réalisation n'apporte aucune information.

Théorème 3.2 : concavité de l'entropie.

□ La fonction $-x \log x$, définie pour tout $x \geq 0$, est concave :



• Figure 3.2 : fonction $y = -x \log_2 x$

Cette propriété est liée à l'inégalité de Jensen pour les fonctions concaves : si $f(x)$ est une fonction concave dans un intervalle $[a, b]$, si $x_1, \dots, x_j, \dots, x_N$ sont des réels arbitraires, $a < x_j <$

b , et $\lambda_1, \dots, \lambda_j, \dots, \lambda_N$ sont des nombres positifs, avec $\sum_{j=1}^N \lambda_j = 1$, alors :

$$f\left(\sum_{j=1}^N \lambda_j x_j\right) \geq \sum_{j=1}^N \lambda_j f(x_j)$$

avec ici : $f(x) = -x \log_2 x$.

Notons une distribution de probabilités $(p_1, \dots, p_k, \dots, p_N)$ sous forme vectorielle : $\dot{p} =$

$(p_1, \dots, p_k, \dots, p_N)$. Avec cette notation, $\sum_{k=1}^N f(p_k) = H(\dot{p})$.

Considérons une suite $\dot{p}_1, \dots, \dot{p}_j, \dots, \dot{p}_N$ de distributions de probabilités telles que $\dot{p}_j = (p_{1j}, \dots, p_{kj}, \dots, p_{Nj})$. $\mathbf{P} = (p_{kj})$ est une matrice à N lignes et N colonnes dont tous les éléments sont positifs ou nuls et la somme des termes de chaque colonne égale à 1.

Construisons la distribution de probabilité $\dot{q} = (q_1, \dots, q_k, \dots, q_N)$ telle que :

$$\dot{q} = \sum_{j=1}^N \lambda_j \dot{p}_j \Leftrightarrow q_k = \sum_{j=1}^N \lambda_j p_{kj} \quad ; \quad \text{avec} \quad \sum_{j=1}^N \lambda_j = 1$$

\dot{q} est bien une distribution de probabilité. En effet :

$$\sum_{k=1}^N q_k = \sum_{k=1}^N \sum_{j=1}^N \lambda_j p_{kj} = \sum_{j=1}^N \sum_{k=1}^N \lambda_j p_{kj} = \sum_{j=1}^N \lambda_j \sum_{k=1}^N p_{kj} = \sum_{j=1}^N \lambda_j \cdot 1 = 1$$

Alors :

$$H(\mathbf{q}) = \sum_{k=1}^N f(q_k) = \sum_{k=1}^N f\left(\sum_{j=1}^N \lambda_j p_{kj}\right)$$

et :

$$\sum_{j=1}^N \lambda_j H(\mathbf{p}_j) = \sum_{j=1}^N \lambda_j \sum_{k=1}^N f(p_{kj}) = \sum_{j=1}^N \sum_{k=1}^N \lambda_j f(p_{kj}) = \sum_{k=1}^N \sum_{j=1}^N \lambda_j f(p_{kj})$$

Or d'après l'inégalité de Jensen appliquée à l'ensemble $(p_{k1}, \dots, p_{kj}, \dots, p_{kN})$:

$$f\left(\sum_{j=1}^N \lambda_j p_{kj}\right) \geq \sum_{j=1}^N \lambda_j f(p_{kj}) \Rightarrow \sum_{k=1}^N f\left(\sum_{j=1}^N \lambda_j p_{kj}\right) \geq \sum_{k=1}^N \sum_{j=1}^N \lambda_j f(p_{kj})$$

donc :

$$H\left(\sum_{j=1}^N \lambda_j \mathbf{p}_j\right) \geq \sum_{j=1}^N \lambda_j H(\mathbf{p}_j)$$

L'indétermination sur un ensemble de distributions de probabilités calculée à partir d'une superposition linéaire de ces distributions est supérieure à la somme pondérée des indéterminations calculées séparément sur chacune d'elles. Ce principe, que nous retrouverons dans le cas quantique, est fondamental : savoir que des distributions de probabilité sont réparties et comment elles le sont est déjà en soi une information, qui diminue notre incertitude par rapport à une estimation faite globalement à partir d'un "mélange" de distributions de probabilités.



Théorème 3.3 : majoration de l'entropie :

$$H(\xi) \leq \log N$$

l'égalité étant réalisée dans le cas d'une loi uniforme.

□ Dans l'inégalité de Jensen, il suffit de remplacer x_j par p_k et de choisir $\lambda_j = 1/N$

$\forall j \in [1, N]$:

$$-\sum_{k=1}^N \frac{1}{N} p_k \log \sum_{k=1}^N \frac{1}{N} p_k \geq -\sum_{k=1}^N \frac{1}{N} p_k \log p_k$$

Sachant que $\sum_k p_k = 1$:

$$-\log \frac{1}{N} \geq -\sum_{k=1}^N p_k \log p_k = H(\xi)$$



Définition 3.1 : on appelle *entropie relative* la quantité $h = H / \log N$.

Définition 3.2 : on appelle *redondance* la quantité $R = 1 - h$, exprimée en % (NB : parfois la quantité $\log N - H$ est aussi appelée redondance).

3.3. Entropies statistiques de sources conjointes

3.3.1. Information composée

On considère une deuxième source symbolique Γ de messages γ constitués de chaînes de m signes η pris dans une bibliothèque Π de M symboles π_1, \dots, π_M :

$$\eta \in \Pi = \{\pi_1, \dots, \pi_j, \dots, \pi_M\}$$

$$\gamma = \eta_1 \dots \eta_k \dots \eta_m \in \Gamma \subset \Pi^*$$

On remarque que les deux sources Σ et Γ émettent des messages de même longueur m . On associe ces deux sources en une source unique $\{\Sigma, \Gamma\}$ dont la bibliothèque est formée de $N \times M$ symboles composites constitués par tous les couples (χ_i, π_j) , $1 \leq i \leq N$, $1 \leq j \leq M$. Cette source lit ou écrit des messages $\sigma\gamma = (\xi_1\eta_1) \dots (\xi_k\eta_k) \dots (\xi_m\eta_m)$.

Sur un total de m signes distincts, le couple (χ_i, π_j) apparaît k_{ij} fois. En généralisant le raisonnement précédent, la masse informationnelle sur m couples de signes successifs est :

$$\Theta = \sum_{i=1}^N \sum_{j=1}^M k_{ij} \log \frac{m}{k_{ij}} = m.H(\xi, \eta) \Rightarrow H(\xi, \eta) = \sum_{i=1}^N \sum_{j=1}^M \frac{k_{ij}}{m} \log \frac{m}{k_{ij}}$$

et :

$$\frac{k_{ij}}{m} \xrightarrow{m \rightarrow \infty} p(i, j) \Rightarrow H(\xi, \eta) = - \sum_{i=1}^N \sum_{j=1}^M p(i, j) \log p(i, j)$$

$H(\xi, \eta)$ est la densité diacritique moyenne par couple de symboles. Cette relation est décomposable selon les lois du calcul des probabilités conditionnelles (la notation " a, b " se lisant a et b ; la notation " $a | b$ " se lisant a si b ou a connaissant b) :

$$p(i, j) = \text{Prob} (\xi = \chi_i, \eta = \pi_j)$$

$$p(i) = \text{Prob} (\xi = \chi_i)$$

$$p(j|i) = \text{Prob} (\eta = \pi_j \mid \xi = \chi_i)$$

avec :

$$p(i, j) = p(i) \cdot p(j|i) \quad p(i, j) = p(j) \cdot p(i|j)$$

$$(i = 1 \dots N) \quad p(i) = \sum_j p(i, j) \quad p(j) = \sum_i p(i, j)$$

$$\begin{aligned}
(j = 1 \dots M) \quad & \sum_i p(i) = 1 & \sum_j p(j) = 1 \\
& \sum_{ij} p(i,j) = 1 & \sum_{ij} p(i)p(j) = 1 \\
& \sum_i p(i|j) = 1 & \sum_j p(j|i) = 1
\end{aligned}$$

Il vient :

$$\begin{aligned}
H(\xi, \eta) &= - \sum_{i=1}^N \sum_{j=1}^M p(i, j) \log p(i) p(j|i) \\
H(\xi, \eta) &= - \sum_{i=1}^N p(i) \log p(i) - \sum_{i=1}^N p(i) \sum_{j=1}^M p(j|i) \log p(j|i)
\end{aligned}$$

On reconnaît dans le premier terme la quantité $H(\xi)$, quantité d'information moyenne par symbole de type Ξ .

La sommation sur j est une quantité d'information moyenne calculée à partir de la probabilité $p(j|i) = \text{Prob}(\eta = \pi_j \text{ si } \xi = \chi_i \text{ donnée})$. C'est la quantité d'information moyenne par symbole de type Π , quand une valeur symbolique χ_i de la variable aléatoire ξ est donnée :

$$H(\eta|\chi_i) = - \sum_j p(j|i) \log p(j|i).$$

La deuxième sommation sur i représente la moyenne pondérée de cette quantité, pour l'ensemble des symboles de type Ξ . C'est donc la moyenne sur l'ensemble des symboles de type Ξ de la quantité d'information moyenne par symbole de type Π , *sous condition* que la valeur χ_i de ξ soit connue :

$$H(\eta|\xi) = \sum_i p(i) H(\eta|\chi_i)$$

Soit :

$$H(\eta|\xi) = - \sum_{i=1}^N p(i) \sum_{j=1}^M p(j|i) \log p(j|i)$$

Cette dernière quantité est appelée "entropie conditionnelle".

Par conséquent l'entropie composée est égale à :

$$H(\xi, \eta) = H(\xi) + H(\eta|\xi)$$

Cette quantité est additive, car dans le cas où ξ et η sont indépendants, $H(\eta|\xi) = H(\eta)$ et $H(\xi, \eta) = H(\xi) + H(\eta)$.

Dans le cas général, on a :

$$H(\xi, \eta) \leq H(\xi) + H(\eta)$$

3.3.2. Information mutuelle

(i) On pose :

$$H(\eta) = H(\eta|\xi) + H(\xi;\eta) \Leftrightarrow H(\xi;\eta) = H(\eta) - H(\eta|\xi)$$

L'information sur η est égale à l'information sur η connaissant ξ , c'est-à-dire à la nouveauté apportée par η , ξ étant connu, plus une quantité qui mesure ce qui n'est pas nouveau. Cette quantité représente l'information "mutuelle", notée $H(\xi;\eta)$, contenue à la fois dans ξ et η . Il vient :

$$H(\xi;\eta) = H(\xi) + H(\eta) - H(\xi,\eta)$$

Théorème 3.4 : l'information mutuelle est une quantité symétrique :

$$H(\xi;\eta) = H(\eta;\xi)$$

La structure des calculs des quantités d'information dans la théorie statistique des sources symboliques est donc une structure symétrique.

□ **Démonstration** : on connaît les relations :

$$H(\xi,\eta) = H(\xi) + H(\eta|\xi) = H(\eta) + H(\xi|\eta)$$

$$\Rightarrow H(\eta) - H(\eta|\xi) = H(\xi) - H(\xi|\eta)$$

$$\Rightarrow H(\xi;\eta) = H(\eta;\xi)$$

□

D'après $H(\xi;\eta) = H(\xi) + H(\eta) - H(\xi,\eta)$, on peut donner à cette quantité une expression symétrique :

$$H(\xi;\eta) = \sum_{i=1}^N \sum_{j=1}^M p(i,j) \log \frac{p(i,j)}{p(i) \cdot p(j)}$$

Soit, en résumé :

$$H(\xi) = - \sum_i p(i) \log p(i)$$

$$H(\xi,\eta) = - \sum_{i,j} p(i,j) \log p(i,j)$$

$$H(\xi|\eta) = - \sum_{i,j} p(i,j) \log p(i|j) \quad \text{avec } p(i|j) = p(i,j) / p(j)$$

$$H(\xi;\eta) = - \sum_i p(i,j) \log p(i;j) \quad \text{avec } p(i;j) = p(i) \cdot p(j) / p(i,j)$$

(ii) $H(\xi:\eta) = H(\xi) - H(\xi|\eta)$ s'exprime encore par :

$$H(\xi:\eta) = H(\xi) + \sum_{j=1}^M p(j) \sum_{i=1}^N p(i|j) \log p(i|j)$$

Comme $\sum_j p(j) = 1$, on peut remplacer dans cette expression $H(\xi)$ par $\sum_j p(j).H(\xi)$. Il vient :

$$H(\xi:\eta) = \sum_{j=1}^M p(j) \left(H(\xi) + \sum_{i=1}^N p(i|j) \log p(i|j) \right)$$

$$\Rightarrow H(\xi:\eta) = \sum_{j=1}^M p(j) \left(H(\xi) - H(\xi|\pi_j) \right)$$

si l'on appelle $H(\xi|\pi_j)$ la quantité $-\sum_i p(i|j) \cdot \log p(i|j)$, quantité d'information moyenne par symbole de type ξ , quand une valeur symbolique π_j de la variable aléatoire η est donnée.

Cette formulation montre que l'information mutuelle entre ξ et η peut s'interpréter comme la moyenne, sur tous les signes de type Π , du gain d'information ou *diminution de l'indétermination* sur ξ , lorsqu'on remplace une estimation grossière $H(\xi)$ de l'information concernant ξ par une estimation plus fine $H(\xi|\pi_j)$ tenant compte de l'information qu'apporte la connaissance de la valeur de la variable associée η .

La réciproque est vraie, en échangeant respectivement ξ et η . On a symétriquement :

$$H(\xi:\eta) = \sum_{i=1}^N p(i) \left(H(\eta) - H(\eta|\chi_i) \right)$$

Cas particulier : supposons que ξ et η représentent deux distributions de probabilité évaluées sur deux messages σ et γ de même longueur $N = M$, toutes les deux uniformes de sorte que $p(i) = p(j) = 1/N$. Ces distributions de probabilité représentent l'état des connaissances *a priori* que l'on a *indépendamment* sur ξ et η , en l'absence de toute autre information. Supposons qu'une analyse des relations entre ξ et η conduise *a posteriori* à produire une connaissance sur η , la distribution ξ étant connue, sous la forme d'une certaine quantité d'information moyenne $H(\zeta)$ (avec $\zeta = \eta|\chi$ – il n'y a pas lieu de préciser l'indice sur χ puisque ces symboles ont une distribution uniforme). De la formule précédente, on déduit :

$$I = \log N - H(\zeta)$$

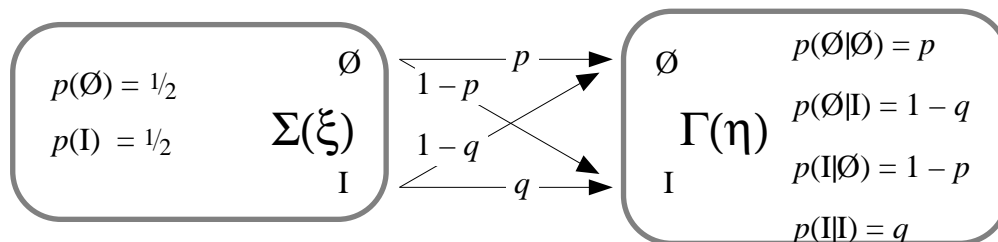
représente le gain d'information (ou diminution d'indétermination) obtenu en remplaçant

des distributions uniformes de probabilités a priori par une distribution ζ de probabilité a posteriori.

ξ et η pourraient en fait représenter le même objet, vu avant et après une certaine série d'expériences. Avant, les symboles sont équiprobables... faute de mieux. Après, une analyse d'une partie du message fournit par exemple la valeur des fréquences de chaque symbole, et permet en conséquence de prédire la quantité d'information moyenne contenue dans les prochains signes lus ou écrits par la source. Cette distinction éclaire la discussion faite au paragraphe 3.4 : toute connaissance apportée a posteriori sur le message diminue l'incertitude et représente un gain d'information qui vaut respectivement, dans l'exemple cité, $33,2-21,8 = 11,4$ sh/mot ou $332-306 = 26$ sh/mot.

Exemple : calcul de l'information mutuelle dans un canal binaire unidirectionnel.

On considère deux sources binaires (voir exemple 3.x) liées par les probabilités suivantes :



- Figure 3.3 : canal binaire unidirectionnel. Soient respectivement $p'(\emptyset)$ et $p'(I)$ les probabilités des symboles lus par Γ . Il vient :
 - $p'(\emptyset) = p(\emptyset) \cdot p(\emptyset|\emptyset) + p(I) \cdot p(\emptyset|I) = \frac{1}{2}(p + 1 - q)$
 - $p'(I) = p(I) \cdot p(I|I) + p(\emptyset) \cdot p(I|\emptyset) = \frac{1}{2}(q + 1 - p)$
 - On vérifie que : $p'(\emptyset) + p'(I) = 1$

Ce schéma, nommé "canal binaire", symbolise usuellement une transmission binaire bruitée : les deux sources sont l'émetteur et le récepteur, les probabilités de transition $\emptyset \rightarrow I$ et $I \rightarrow \emptyset$ sont dues au "bruit" qui affecte le canal de communication. L'information mutuelle $H(\xi;\eta)$ représente l'information transmise de Ξ à Γ . Mais cette interprétation n'est pas la seule possible. On en verra plus loin un exemple différent à travers la théorie quantique de l'information. Plus généralement, tout couple (Σ, Γ) de sources binaires liées de la sorte est un modèle pour ce calcul.

L'information mutuelle se calcule par exemple par : $H(\xi;\eta) = H(\eta) - H(\eta|\xi)$

(i) Calcul de $H(\eta)$:

$$H(\eta) = - \sum_j p(j) \log p(j)$$

$$H(\eta) = -p'(\emptyset) \log p'(\emptyset) - p'(\text{I}) \log p'(\text{I})$$

$$H(\eta) = -\frac{1}{2} (p + 1 - q) \log \frac{1}{2} (p + 1 - q) - \frac{1}{2} (q + 1 - p) \log \frac{1}{2} (q + 1 - p)$$

$$H(\eta) = H_2\left(\frac{1}{2} (1 + p - q)\right)$$

(ii) Calcul de $H(\eta|\xi)$

$$H(\eta|\xi) = -\sum_{i,j} p(i,j) \log p(j|i)$$

$$H(\eta|\xi) = -\sum_i p(i) \sum_j p(j|i) \log p(j|i)$$

$$H(\eta|\xi) = -p(\emptyset) \left[p(\emptyset|\emptyset) \log p(\emptyset|\emptyset) + p(\text{I}|\emptyset) \log p(\text{I}|\emptyset) \right] \\ - p(\text{I}) \left[p(\text{I}|\text{I}) \log p(\text{I}|\text{I}) + p(\emptyset|\text{I}) \log p(\emptyset|\text{I}) \right]$$

$$H(\eta|\xi) = -\frac{1}{2} \left[p \log p + q \log q + (1-p) \log (1-p) + (1-q) \log (1-q) \right]$$

$$H(\eta|\xi) = \frac{1}{2} \left[H_2(p) + H_2(q) \right]$$

D'où :

$$H(\xi:\eta) = H_2\left(\frac{1}{2} (1 + p - q)\right) - \frac{1}{2} \left[H_2(p) + H_2(q) \right]$$

(iii) A.N.1 : "Canal Binaire Symétrique" (CBS) : c'est un système tel que $p = q$:

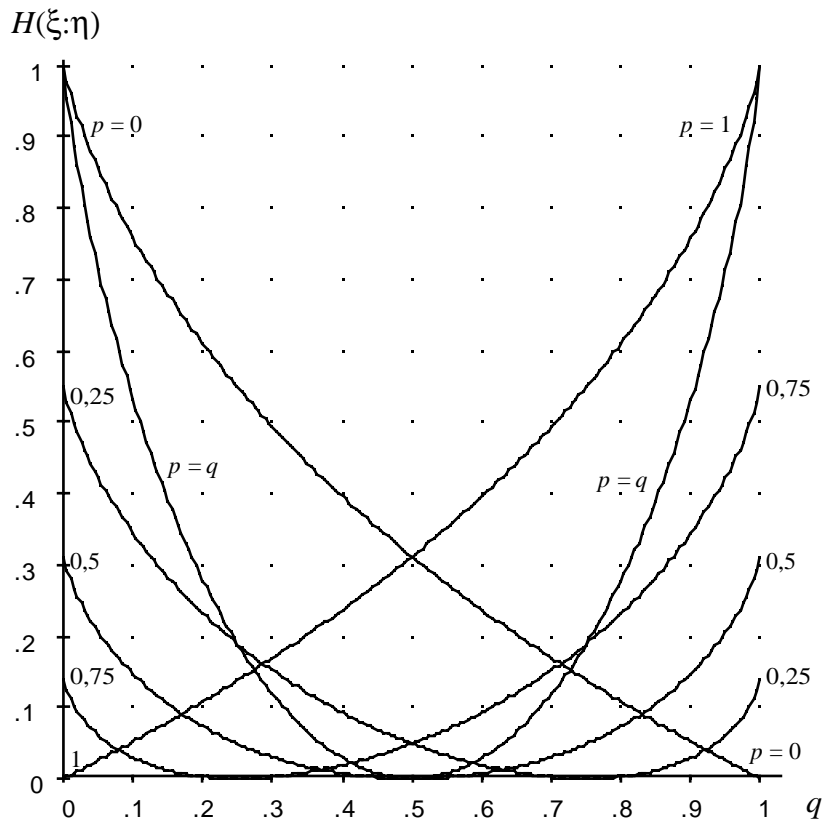
$$H(\xi:\eta) = 1 - H_2(p)$$

L'information est nulle pour $p = q = \frac{1}{2}$, ce cas correspondant à l'indétermination la plus grande; elle est maximale (= 1) pour $p = q = 1$.

(iv) A.N.2 : considérons le cas $p = 1$ et q quelconque ($q \in [0,1]$) :

$$H(\xi:\eta) = H_2\left(\frac{1}{2} q\right) - \frac{1}{2} H_2(q)$$

(v) Cas général : on donne ci-dessous l'évolution de $H(\xi:\eta)$ en fonction de q pour différentes valeurs de p ($p = 0 ; 0,25 ; 0,5 ; 0,75 ; 1$) ainsi que pour $p = q$.



- *Figure 3.4* : information mutuelle dans un système binaire en fonction des probabilités p et q (voir texte).

3.3.3 Ecart entropique

Définition 3.3 : soient ξ et η deux variables aléatoires telles que $\sum_{i=1}^N p(i) = 1$ et

$\sum_{i=1}^N q(i) = 1$. On suppose en outre que $q(i) \neq 0 \forall i$. On appelle *écart entropique* [Battail, 1997]

ou *cross-entropy* [Cover et Thomas, 1991] ou *gain d'information* [Rényi, 1962] ou *quantité d'information de Kullback-Leibler* [Réfrégier, 1993] la quantité :

$$H(\xi \parallel \eta) = \sum_{i=0}^N p(i) \log \frac{p(i)}{q(i)} \geq 0$$

- L'inégalité $H(\xi \parallel \eta) \geq 0$, appelée aussi inégalité de Gibbs, se déduit de l'inégalité de Jensen, mais peut aussi se démontrer directement : sachant que $\ln x \leq x-1$, en remplaçant x par $p(i)/q(i)$ et en multipliant par $p(i)$, il vient :

$$\sum_i p(i) \ln \frac{q(i)}{p(i)} \leq \sum_i p(i) \left(\frac{q(i)}{p(i)} - 1 \right) = \sum_i q(i) - \sum_i p(i) = 1 - 1 = 0$$

La multiplication de l'inégalité par $\log_2 e$ renvoie aux logarithmes en base 2.

□

Si $p(i) \neq 0 \forall i$, on définit de même $H(\eta||\xi)$. Cette grandeur permet d'évaluer la "distance" entre deux distributions de probabilité pour une source donnée, bien qu'elle n'ait pas les caractéristiques d'une "distance" au sens mathématique du terme car en général $H(\xi||\eta) \neq H(\eta||\xi)$.

Si $q(i) = \frac{1}{N} \forall i$, il vient :

$$H(\xi||\eta) = \log N - H(\xi)$$

L'écart entropique mesure donc la différence entre l'entropie maximum (d'une distribution uniforme) et l'entropie d'une distribution p , c'est-à-dire la *redondance* (déf. 3.2).

Soient deux variables ξ et η , avec $r(k) = p(i,j)$ la distribution des probabilités conjointes où $k \in [1, N.M]$, et $q(k) = p(i).p(j)$ le produit direct de leurs vraies distributions (qui serait égal à $r(k)$ si les deux variables étaient indépendantes), alors :

$$H(\xi;\eta) = \sum_{i=1}^N \sum_{j=1}^M p(i,j) \log \frac{p(i,j)}{p(i).p(j)} = \sum_{k=1}^{N.M} r(k) \log \frac{r(k)}{q(k)} = H((\xi, \eta) || \xi, \eta)$$

L'information mutuelle est donc l'écart entropique entre la source (Σ, Γ) de couples (ξ, η) et le produit des sources Σ et Γ . On en déduit le

Théorème 3.5 : l'information mutuelle est une grandeur positive ou nulle.

$$H(\xi;\eta) \geq 0$$

D'où l'on tire l'inégalité :

$$H(\xi) \geq H(\xi|\eta)$$

3.4. Exemple. Structure ternaire et théories de l'information

La structure ternaire de l'information est implicite dans la théorie de l'information. Pour s'en convaincre, nous examinerons ici deux formes complémentaires de celle-ci, d'une part statistique, d'autre part algorithmique, à travers l'exemple d'une horloge. La théorie statistique est une théorie de la mesure de l'information par le récepteur du message, alors que la théorie algorithmique est une théorie de la mesure de l'information par l'émetteur qui construit le message. C'est donc à deux points de vue complémentaires que nous conduisent ces calculs :

message reçu par un récepteur, vs complexité intrinsèque d'une source faisant l'objet d'un tel message.

3.4.1. Point de vue du récepteur : analyse statistique du message

Considérons une horloge indiquant l'heure et les minutes, de 00h00mn à 23h59mn. Supposons qu'un observateur prend connaissance de l'indication portée par l'horloge à des instants aléatoires, avec les hypothèses suivantes :

- l'observateur connaît l'alphabet $\{0, \dots, 9\}$ et le standard convenu préalablement entre le constructeur de l'horloge et lui-même : le premier digit (à gauche du mot) est le chiffre des dizaines de l'heure, etc. Soit un format, ordonné de gauche à droite, " $h_d h_u m_d m_u$ ", avec : $0 \leq h_d \leq 2$; $0 \leq h_u \leq 9$; $0 \leq m_d \leq 5$; $0 \leq m_u \leq 9$. Remarquons que cette connaissance n'est pas anodine : ainsi une date s'écrit sous forme condensée, en français, jour/mois/année, alors qu'en anglais on écrit mois/jour/année ; la méconnaissance de cette convention peut être source d'erreurs importantes.

- l'observateur est sans mémoire, au sens où la lecture de l'heure à l'instant t_1 ne lui apporte aucune indication sur la lecture de l'heure à l'instant $t_2 > t_1$.

- il est possible de distinguer le mot codant l'heure par rapport au reste de l'univers : on ne tient pas compte de l'existence de séparateurs (par exemple un blanc), du rapport signal sur bruit (le cadran est lisible), etc.

Quelle est la quantité d'information apportée par une lecture de l'horloge? Plusieurs calculs sont possibles :

1°) Il est clair que si l'horloge codait l'heure comme une source symbolique comprenant $24 \times 60 = 1440$ symboles distincts (à la manière des idéogrammes chinois), chaque symbole porterait, les états de l'horloge étant bien sûr équiprobables, une information $\Theta_1 = \log_2 1440 = 10,49$ sh (10,491853 plus exactement!). C'est l'information contenue dans la connaissance *de l'heure* en soi, et non dans la connaissance *du message* représentant l'heure (comme indiqué ci-dessous, cette représentation de l'heure est plus complexe et plus riche que la simple énumération d'une bibliothèque de 1440 symboles distincts).

2°) Si l'on considère le message comme constitué globalement de 4 digits, le calcul a trait aux messages de 4 symboles émis par une source symbolique à 10 états $\{0, \dots, 9\}$. Le calcul des probabilités, effectué à partir des fréquences calculées sur $4 \times 1440 = 5760$ digits émis par jour, donne $p \approx 0,2$ pour $\{0\}$ et $\{1\}$; $0,14$ pour $\{2\}$; etc. On trouve :

$$H = \sum_{i=0}^9 p_i \cdot \log p_i = 3,087 \text{ sh/digit}$$

Soit, pour des messages de longueur 4, $\Theta_2 \approx 12,35$ sh/message (cf tableau 3.1) :

	Hd	Hu	Md	Mu		
0	600	180	240	144	1164	0,466201754337
1	600	180	240	144	1164	0,466201754337
2	240	180	240	144	804	0,396528529469
3	0	180	240	144	564	0,328246212427
4	0	120	240	144	504	0,307525152623
5	0	120	240	144	504	0,307525152623
6	0	120	0	144	264	0,203841869778
7	0	120	0	144	264	0,203841869778
8	0	120	0	144	264	0,203841869778
9	0	120	0	144	264	0,203841869778
					Σ	Q (HdHuMdMu)
					5760	12,3503841397

- Tableau 3.1 : calcul des fréquences journalières de chaque chiffre dans un mot " $h_d h_u m_d m_u$ ".

Où est la différence ? Dans ce dernier calcul, on a ignoré la structure du message, pour ne s'intéresser qu'à la fréquence des symboles qui le composent. La différence $\Theta_2 - \Theta_1$ est proche de deux shannons. Or cette quantité représente la quantité d'information nécessaire pour spécifier la position, c'est-à-dire *l'adresse* de chaque digit. On pourrait par exemple affecter un message complémentaire écrit dans un alphabet binaire pour déterminer l'identité de chaque digit : 00 pour h_d ; 01 pour h_u ; 10 pour m_d ; 11 pour m_u par exemple.

3°) En réalité, l'observateur connaît cette structure, qui apparaît dans la convention préalable faite sur le standard. On peut calculer la quantité d'information portée par chaque digit pris séparément : l'affichage de l'heure associe quatre sources symboliques distinctes produisant chacune des messages de longueur unité. Il vient :

$$H(h_d) = \sum_{i=0..2} p(i) \log p(i) \approx 1,483 \text{ sh}$$

$$H(h_u) = \sum_{i=0..9} p(i) \log p(i) \approx 3,292 \text{ sh}$$

$$H(m_d) = \log 6 \approx 2,585 \text{ sh}$$

$$H(m_u) = \log 10 \approx 3,322 \text{ sh}$$

Soit un total, par mot, $\Theta_3 \approx 10,68$ sh (cf tableau 3.2).

	Hd		Hu		Md		Mu	
0	10	0,526264335764	3	0,375	10	0,430827083454	6	0,332192809489
1	10	0,526264335764	3	0,375	10	0,430827083454	6	0,332192809489
2	4	0,430827083454	3	0,375	10	0,430827083454	6	0,332192809489
3	0	0	3	0,375	10	0,430827083454	6	0,332192809489
4	0	0	2	0,29874687506	10	0,430827083454	6	0,332192809489
5	0	0	2	0,29874687506	10	0,430827083454	6	0,332192809489
6	0	0	2	0,29874687506	0	0	6	0,332192809489
7	0	0	2	0,29874687506	0	0	6	0,332192809489
8	0	0	2	0,29874687506	0	0	6	0,332192809489
9	0	0	2	0,29874687506	0	0	6	0,332192809489
	Σ	Q(Hd)	Σ	Q(Hu)	Σ	Q(Md)	Σ	Q(Mu)
	24	1,48335575498	24	3,29248125036	60	2,58496250072	60	3,32192809489
								Q(HdHuMdMu)
								10,682727601

- **Tableau 3.2** : calcul des fréquences journalières de chaque chiffre dans les mots " h_d ", " h_u ", " m_d ", " m_u " pris séparément.

En tenant ainsi compte de la structure du mot (calcul effectué séparément sur chaque digit), le résultat obtenu s'approche de Θ_1 , l'information contenue dans la lecture de l'heure proprement dite. Il subsiste toutefois une légère différence entre Θ_3 et Θ_1 .

4°) Un autre facteur n'a pas été pris en compte : l'heure étant comprise entre 00 et 23, la lecture du chiffre des dizaines apporte quelque information sur le chiffre des unités. Si $h_d = 2$, on sait que $0 \leq h_u \leq 3$ au lieu de $0 \leq h_u \leq 9$. La réciproque est vraie : si $0 \leq h_u \leq 3$, alors $0 \leq h_d \leq 2$, sinon $0 \leq h_d \leq 1$. Cette remarque ne concerne pas les minutes : pour tous les m_d , $0 \leq m_u \leq 9$.

Il est donc nécessaire d'effectuer un calcul de probabilités conditionnelles pour connaître la quantité d'information contenue dans la partie du message $h_u h_d$: ces deux sources symboliques (variables h_u et h_d) sont liées.

On trouve, avec les notations usuelles :

-Quantité d'information apportée par la lecture de h_u connaissant h_d :

$$H(h_u|h_d) = -\sum_{i=0..2} \sum_{j=0..9} p(i,j) \log p(j|i) \approx 3,102 \text{ sh}$$

-Quantité d'information apportée par la lecture de $h_d h_u$:

$$H(h_u, h_d) = H(h_d) + H(h_u|h_d) = 4,585 \text{ sh}$$

-Quantité d'information mutuelle :

$$H(h_u \cdot h_d) = H(h_u) - H(h_u|h_d) = H(h_u) + H(h_d) - H(h_u, h_d) = 0,19 \text{ sh}$$

Autrement dit, la lecture de l'un des deux chiffres de l'heure apporte une quantité d'information de 0,19 sh sur la valeur de l'autre chiffre.

La quantité d'information totale, sur le mot, est alors :

$$\Theta_4 = H(h_u, h_d) + H(m_d) + H(m_u) = 10,49 \text{ sh}$$

	Hd		Hu			Md		Mu		
			2	1	0					
0	10	0,52626433576	1	1	1	0,36016067457	10	0,43082708345	6	0,33219280949
1	10	0,52626433576	1	1	1	0,36016067457	10	0,43082708345	6	0,33219280949
2	4	0,43082708345	1	1	1	0,36016067457	10	0,43082708345	6	0,33219280949
3	0	0	1	1	1	0,36016067457	10	0,43082708345	6	0,33219280949
4	0	0	0	1	1	0,27682734124	10	0,43082708345	6	0,33219280949
5	0	0	0	1	1	0,27682734124	10	0,43082708345	6	0,33219280949
6	0	0	0	1	1	0,27682734124	0	0	6	0,33219280949
7	0	0	0	1	1	0,27682734124	0	0	6	0,33219280949
8	0	0	0	1	1	0,27682734124	0	0	6	0,33219280949
9	0	0	0	1	1	0,27682734124	0	0	6	0,33219280949
	Σ	Q(Hd)	Σ	Σ	Σ	Q(Hu Hd)	Σ	Q(Md)	Σ	Q(Mu)
	24	1,483355755	4	10	10	3,1016067457	60	2,5849625007	60	3,3219280949
					Σ					Q(HdHuMdMu)
					24					10,491853096

- Tableau 3.3 : calcul des fréquences journalières de chaque chiffre dans les mots " $h_d h_u$ ", " m_d ", " m_u ".

On trouve : $\Theta_4 = \Theta_1$. La connaissance précise de la structure du message permet à la lecture de porter seulement sur la connaissance de l'heure proprement dite. Cette connaissance de la structure du message est implicite, et porte sur la convention préalable faite entre l'émetteur et le récepteur. C'est donc dans la connaissance de la répartition des sources symboliques dans le message et de leurs relations que git l'information implicite utilisée par le lecteur pour connaître l'heure effective à travers l'heure affichée.

On peut quantifier ces informations de la façon suivante : sur 12,35 h "aveugles", il y a :

- 0,19 sh provenant du code utilisé, le système décimal codant une énumération sexagésimale, ce qui entraîne la dépendance de certaines sources.
- 1,67 sh provenant du format " $h_d h_u m_d m_u$ " - heures/minutes, dizaines/unités, 12/24h, lecture de gauche à droite, etc - le système de numération décimal étant connu, c'est-à-dire la structure de l'affichage en une succession ordonnée de quatre sources symboliques.
- 10,49 sh d'information en "valeur propre", c'est-à-dire portant effectivement sur la connaissance de l'heure qu'il est, le système décimal et le format étant connus.

3.4.2. Point de vue de l'émetteur : construction algorithmique du message

Soit une procédure effective (c'est-à-dire un algorithme) affichant l'heure à l'aide d'un appareil à 1440 états distincts. Cette procédure peut être simulée à l'aide d'une machine de Turing. Pour la clarté de l'exposé, nous utiliserons une machine RAM (*Random Access Memory*) à programme enregistré (machine RASP : *Random Access with Stored Program*)[Autebert, 1992], dont on sait qu'une telle machine peut être simulée par une machine de Turing. Le programme d'une machine RAM est constitué d'une suite d'instructions prises parmi un certain nombre d'opérations élémentaires. Dans la pratique, nous utiliserons un sous-ensemble du jeu d'instructions du processeur 80x86, programmé en assembleur, pour matérialiser cette machine. Nous disposons ainsi d'un certain nombre de registres internes au processeur (registres AX, BX, CX, DX, CS, DS, ES, etc du 80x86) et surtout de la mémoire de l'ordinateur, qui met à la disposition du programmeur un nombre extrêmement grand de "registres" formés d'une ou plusieurs positions mémoire de un octet chacune. Un programme dans la RAM est constitué d'une suite d'instructions prises parmi les opérations élémentaires indiquées ci-dessous, compilées sous la forme d'un programme exécutable simplifié d'extension .COM.

<i>Type</i>	<i>Instruction</i>	<i>Implémentation 80x86</i> (<i>r</i> : registre; <i>m</i> : mémoire)
Instruction d'affectation	a:= b	
CHARGER opérande	r ← m	MOV r,m ou r,r
RANGER opérande	m ← r	MOV m,r ou r,r
Instructions d'entrée/sortie		
LIRE clavier	entrer	MOV r,m (adr. clavier)
ÉCRIRE écran	imprimer	MOV m (adr. écran),r
Instructions arithmétiques	+, -	
INCRÉMENTER	opérande	r/m + 1 INC r/m
DÉCRÉMENTER	opérande	r/m - 1 DEC r/m
Instructions de contrôle		
ARRET	fin	INT
SAUT SI NON ZÉRO	si... alors...	JNZ m

Une dernière opération permet d'exécuter séquentiellement une boucle finie, mais ne nécessite pas d'instruction nouvelle, par décomptage ou comptage dans un alphabet binaire

limité ici à 8 ou 16 bits :

Décomptage:

FAIRE...

pour $i:=n$ à 0

boucle: MOV r,n
... (instructions)
DEC r
JNZ boucle

Comptage (la retenue étant ignorée) :

FAIRE...

pour $i:=0$ à n

boucle: MOV r,256-n
ou 65536-n
... (instructions)
INC r
JNZ boucle

Nous nous interdisons des instructions de type ADD, SUB, CMP (comparer a et b, en soustrayant b de a), car celles-ci supposent l'exécution d'opérations plus compliquées que la simple décrémentation ou incrémentation (leur profondeur logique au sens de Benett [Delahaye, 1994] est supérieure). Voir par exemple l'algorithme qui vérifie l'égalité de deux grandeurs a et b sans utiliser l'instruction CMP :

a = b ?

MOV ax,a
MOV bx,b
boucle: DEC ax
DEC bx
JNZ boucle
INC ax
DEC ax
JNZ suite
... ; a=b
suite ... ; a≠b

Enfin, nous nous interdisons l'indirection, ce qui est toujours possible avec une machine RASP, et permet donc de simuler toute machine de ce type à l'aide d'une machine RAM. Cela interdit aussi l'usage d'une table de transcodage réalisée à l'aide d'un pointeur (par adressage indexé), ce qui reviendrait à exécuter une addition (d'adresses), hypothèse que nous avons d'emblée éliminée en ne conservant que les instructions arithmétiques INC et DEC.

Néanmoins, malgré la simplicité des instructions auxquelles nous nous limitons, on montre [Autebert 1992, Delahaye 1994] que ce langage de programmation suffit pour énumérer l'ensemble (dénombrable) des fonctions primitives récursives.

Un certain nombre de détails restent à régler pour implémenter en assembleur un programme d'horloge sur la machine :

- une temporisation (tempo) devrait fixer à une minute la durée d'exécution d'une itération. Pour des raisons pratiques, elle sera ici volontairement réduite à quelques dixièmes de seconde.

- la machine possède un interface de sortie constituée d'un écran de 25 lignes de 80 caractères, ce qui est amplement suffisant pour afficher le message de l'heure (ici, en haut à droite de l'écran, adresse hexadécimale B800:0090h).

- chaque signe est pris dans une bibliothèque de 256 symboles (code ASCII étendu) avec 256 attributs de couleur (couleur du caractère + couleur du fond) soit au total $256 \times 256 = 65536$ symboles distincts, ce qui là aussi suffit amplement pour représenter les 1440 états de la source.

- la table ASCII étendue est utilisable quelque soit son sens de parcours (sens lexicographique des codes croissants ou sens inverse).

Un programme qui affiche l'heure doit alors obéir au cahier des charges suivant:

- compter ou décompter 1440 états distincts.
- afficher 1440 messages distincts, choisis dans une bibliothèque unique.

Cette dernière opération est importante : il ne suffit pas d'énumérer 1440 messages distincts, encore faut-il les présenter selon les conventions implicites d'écriture de l'heure déjà évoquées. Si l'on se passe de ces conventions, alors le programme (1) fourni en annexe (*voir annexe 1*) suffit. Il s'agit d'un programme minimum, qui affiche l'heure à l'aide de 255 symboles ASCII (le symbole NUL de code \emptyset est évité) avec 6 couleurs différentes (bleu, vert, bleu de cobalt, rouge, marron, gris clair) de sorte qu'il est possible d'afficher 1440 états distincts codés dans une bibliothèque de $255 \times 6 = 1530$ symboles. Ces symboles codent bijectivement l'état du registre (dénommé "compteur"). Si l'on effectue un décomptage de 1440 à 1 (DEC), la table des codes ASCII est parcourue à l'envers ; dans le cas contraire d'un comptage de 64536-1440 à 64535 (INC), elle est parcourue dans le sens usuel des codes croissants. Mais peu importe : il n'y a à ce stade aucune convention concernant l'ordre lexicographique. Au total, nous obtenons un programme de longueur (i.e. de complexité) égale à 29 octets.

Le programme (2) quant à lui exécute une tâche supplémentaire : il affiche l'heure en base soixante, selon l'écriture traditionnelle du temps (qui remonte aux babyloniens !). 60 symboles ASCII monochromes suffisent. La complexité de ce programme est de 48 octets.

Mais le programme (2) est faux ! Étant donné une bibliothèque Ξ comprenant 60 symboles $\{\chi_1, \dots, \chi_{60}\}$, il nous faut énumérer respectivement les symboles χ_1, \dots, χ_{60} pour les minutes et χ_1, \dots, χ_{24} pour les heures. L'affichage des états d'un compteur ou d'un décompteur ne suffit pas. Il faut ajouter un test d'arrêt lorsque $h = 24$, car le comptage des minutes impose

un sens de balayage de la bibliothèque Ξ_{60} , de χ_1 à χ_{60} , si bien qu'il est faux d'énumérer les heures, comme le fait le programme (2), de χ_{36} à χ_{60} (à moins de considérer que le message des heures est écrit dans un alphabet différent de celui des minutes, ce qui est contraire aux hypothèses de départ : nous voulons que h_1 et m_1, \dots , soient codés par des symboles de même type). Donc, si nous voulons éviter l'utilisation d'un test d'égalité par l'instruction CMP, et éviter l'indirection (pour l'utilisation d'une table), nous sommes conduits à compliquer le programme en distinguant la variable de comptage de la variable d'affichage des heures – cf programme (3).

Remarquons que la complexité de ce programme est la même, qu'il exécute un comptage ou un décomptage : comme dans la théorie shannonienne de l'information, la théorie algorithmique de l'information ne prend pas en compte le sens d'énumération des messages, qui est ici le sens de parcours de la bibliothèque Ξ_{60} fixé par le sens d'énumération des minutes. L'information implicite concernant ce sens n'est pas une donnée numérique évaluée par ces théories. *En revanche*, le fait de devoir tenir compte de ce sens dans l'énumération des heures complique le programme (3) – dont la longueur est de 52 octets – par rapport au programme (2) : ce n'est donc pas le sens qui est mesuré par la complexité de l'algorithme, c'est son existence, le fait qu'il y ait un sens ou qu'il n'y en ait pas.

Même avec une machine de Turing travaillant en code unaire, il ne serait pas possible de faire l'économie de cette complication supplémentaire si l'on suppose par hypothèse que le fonctionnement de toute machine énumérant une suite d'états distincts fait appel à l'arithmétique des entiers naturels (faute d'une telle hypothèse, aucune évaluation numérique de complexité n'est plus possible). En effet, de deux choses l'une. Ou bien la machine compte dans l'ordre numérique croissant. La première minute ou la première heure est alors codée par un 1, la vingt-quatrième heure par vingt-quatre 1, la soixantième minute par soixante 1 successifs. Mais alors cela nécessite pour l'arrêt l'usage d'une machine de Turing sachant exécuter le test d'égalité ($h = 24$?), d'où une complication algorithmique et une profondeur logique supplémentaires. Ou bien la machine décompte l'ordre numérique jusqu'à zéro (arrêt si zéro). La première minute ou la première heure sont codées par soixante 1 – si l'on veut préserver la règle de l'alphabet commun aux deux messages (heure, minute) – la vingt-quatrième heure par trente-six 1, la soixantième minute par un 1. Cela nécessite alors de faire la distinction entre le décompteur horaire (initialisé à vingt-quatre 1) et le message affichant l'heure (initialisé à soixante 1). D'où à nouveau une complexité algorithmique accrue.

Enfin le programme (4) affiche l'énumération sexagésimale des heures et des minutes en base 10, ce qui apporte deux contraintes supplémentaires :

- il faut compter séparément les dizaines et les unités des heures et des minutes, ce qui ajoute deux boucles de calcul, une par source symbolique supplémentaire.

- il faut insérer une condition d'arrêt sur la vingt-quatrième heure, ce qui correspond, dans la théorie shannonienne de l'information, à la connaissance que l'on a sur le chiffre des unités d'heure quand on lit celui des dizaines, ou vice versa.

Le programme (4), de longueur 126 octets, affiche les entiers 0 à 9 à l'aide des chiffres "arabes" habituels -ce qui impose encore le sens (lexicographique) de parcours de la bibliothèque Ξ_{10} , et donc la distinction entre variables de comptage et variables d'affichage.

En résumé (cf tableau 3.4) , cette évaluation succincte de la complexité algorithmique de l'affichage de l'heure met bien en évidence les différents contenus d'identification ajoutés à l'information horaire proprement dite. Les conclusions apportées ici au niveau de l'émetteur rejoignent les résultats examinés au paragraphe précédent concernant le récepteur :

- 29 octets sont nécessaires pour fabriquer la source à 1440 états
- il faut ajouter 19 octets supplémentaires pour tenir compte du système de numération (base 60)...
- et 4 octets pour spécifier le sens d'énumération des états.
- un supplément de 74 octets vient enfin compléter l'algorithme pour respecter le format ($h_d h_u m_d m_u$) et le codage décimal.

Dans cet exemple, l'heure est le contenu d'information effectivement transmis, étant entendu que le récepteur et l'émetteur du message sont implicitement d'accord pour compter l'heure en base 60 et l'écrire en base 10 dans un format pondéré sur 4 digits justifié à droite (le chiffre de poids faible, situé à droite, jouant le rôle de référence).

Un codage élémentaire à partir d'une bibliothèque complexe conduit à un programme de courte longueur, alors qu'un codage sophistiqué à l'aide d'une bibliothèque réduite conduit à un programme beaucoup plus long. Cette constatation élémentaire est une clé des techniques de compression d'information, qui consistent à rejeter dans le codage (non transmis) un maximum d'information, pour que le message transmis soit le plus court possible.

Mais inversement, peut-on expliciter complètement, dans un processus de transfert informationnel, le contenu d'identification implicite qui lui est inhérent ?

Tableau 3.4 : affichage de l'heure (résumé) :

□ A l'aide d'une bibliothèque de 1440 symboles :

• *Quantité d'information* :

$$\Theta_1 = \log(24 \times 60) = \log 1440 = 10,491853 \text{ sh}$$

• *Programme* :

```
B8B8008EC0A1011EFEC426A30090BAFFFF4A75FDFF0E011E75
EBB44CCD2105A0
```

□ A l'aide d'une bibliothèque de 60 symboles :

• *Quantité d'information apparente* :

$$\Theta_2 = 2 \cdot \sum_{i=0}^{59} p_i \cdot \log p_i = 11,28 \text{ sh}$$

• *Quantité d'information réellement transmise* :

$$\Theta_3 = \log 24 + \log 60 = 10,49 \text{ sh}$$

• *Programme* :

```
B8B8008EC0B41FA0013726A30090C60601363C90A0013626A3
0098BAFFFF4A75FDFF0E013675EDFE0E0137FE0E013575D6B4
4CCD21183C3C
```

□ A l'aide d'une bibliothèque de 10 symboles :

• *Quantité d'information apparente globale* :

$$\Theta_2 = 4 \times \sum_{i=0}^9 p_i \log p_i = 12,32$$

• *Quantité d'information apparente, la structure du message ($h_d h_u m_d m_u$) étant connue* :

$$\Theta_3 = \sum_{d=0}^2 h_d \log h_d + \sum_{u=0}^9 h_u \log h_u + \sum_{d=0}^6 m_d \log m_d + \sum_{u=0}^9 m_u \log m_u$$

$$= 1,483 + 3,292 + 2,585 + 3,322 = 10,68 \text{ sh}$$

• *Quantité d'information réellement transmise* :

$$\Theta_4 = \sum_{d=0}^2 h_d \log h_d + \sum_{d=0}^2 \sum_{u=0}^9 h_{d,u} \log h_{d,u} + \sum_{d=0}^6 m_d \log m_d + \sum_{u=0}^9 m_u \log m_u$$

$$= 1,483 + 3,102 + 2,585 + 3,322 = 10,491853 \text{ sh}$$

• *Programme* :

```
B8B8008EC0B41FA0018426A30090C60601853090C60601800A
90A0018526A30092C60601863090C60601810690A0018626A3
0098C60601873090C60601820A90A0018726A3009ABAFF4A
75FDFF0E0187FE0E018275E9FE060186FE0E018175CCFE0E01
837504B44CCD21FE060185FE0E018075A5FE060184FE0E017F
39017D7588030A060A1830303030
```

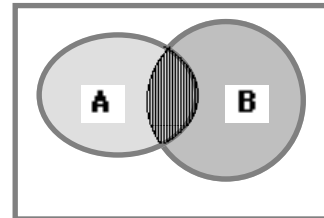
3.5. Conclusions

3.5.1. Interprétation ensembliste

Considérons les sources Σ et Γ comme deux ensembles A et B. Ecrivons $\mu(A)$ et $\mu(B)$ quelque mesure (par exemple l'aire) associée à A et B. Dans le cas discret, les entropies sont positives ou nulles, et additives. On peut formuler un parallèle entre les trois formalismes, ensembliste, probabiliste et informationnel :

° mathématiques ensemblistes :	$\mu(A \cup B) = \mu(A) + \mu(B) - \mu(A \cap B)$
° probabilités	$p(A \cup B) = p(A) + p(B) - p(A \cap B)$
° théorie de l'information	$H(A, B) = H(A) + H(B) - H(A : B)$
avec :	

$\mu(A)$: mesure de l'ensemble A
 $p(A)$: probabilité de l'événement A
 $H(A)$: entropie de la source A

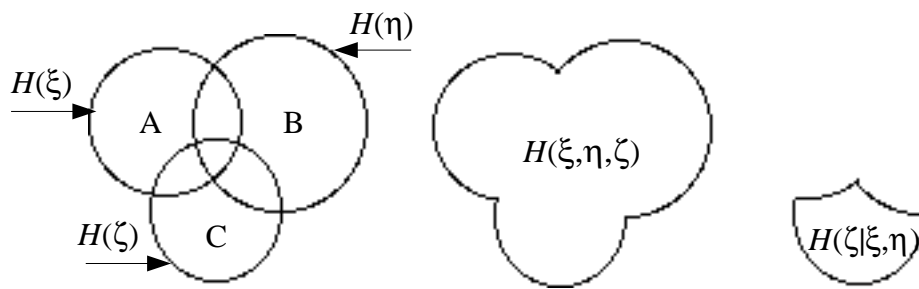


Le bilan de l'interprétation ensembliste de la théorie de l'information est le suivant :

μ	p	H
$\mu(A)$	$p(i)$	$H(\xi)$
$\mu(A \cup B)$	$p(i, j)$	$H(\xi, \eta)$
$\mu(A \cap B)$		$H(\xi : \eta)$
$\mu(A \cap \neg B)$	$p(i/j) = p(i, j) / p(j)$	$H(\xi \eta)$
$\mu(A \cup B) \leq \mu(A) + \mu(B)$		$H(\xi, \eta) \leq H(\xi) + H(\eta)$
$\mu(A \cap \neg B) \leq \mu(A)$		$H(\xi \eta) \leq H(\xi)$

L'intérêt de cette représentation est qu'elle s'étend à un nombre quelconque de sources. Avec trois sources, citons en exemple quelques relations (démontrées formellement par Khinchin) :

$\mu(A \cup B \cup C) \leq \mu(A) + \mu(B) + \mu(C)$	$H(\xi, \eta, \zeta) \leq H(\xi) + H(\eta) + H(\zeta)$
$\mu[C \cap \neg(A \cup B)] \leq \mu(C \cap \neg B)$	$H(\zeta \xi, \eta) \leq H(\zeta \eta)$

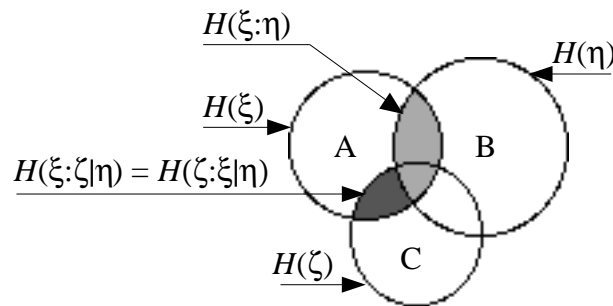


- Figure 3.5 : présentation ensembliste associée à trois sources de variables ξ , η , ζ .

De même :

$$\mu[A \cap (B \cup C)] = \mu(A \cap B) + \mu[A \cap (C \cap \neg B)] \quad H(\xi : \eta, \zeta) = H(\xi : \eta) + H(\xi : \zeta | \eta)$$

$$\mu[(B \cup C) \cap A] = \mu(B \cap A) + \mu[C \cap (A \cap \neg B)] \quad H(\eta, \zeta : \xi) = H(\eta : \xi) + H(\zeta : \xi | \eta)$$



- Figure 3.6 : présentation ensembliste des informations mutuelles $H(\xi : \eta, \zeta)$, $H(\xi : \eta)$ et $H(\xi : \zeta | \eta)$

Enfin, dans le même ordre d'idée, on montre que :

Théorème 3.6 : l'entropie classique possède la propriété appelée *sous-additivité "forte"* :

$$H(\xi, \eta, \zeta) + H(\eta) \leq H(\xi, \eta) + H(\eta, \zeta)$$

Remarque sur les notations

On constate que la notation ensembliste est plus claire que la notation adoptée généralement en théorie de l'information, dès que celle-ci traite de plus de deux sources. Il convient d'avoir à l'esprit les conventions de lecture suivante, sachant qu'on lit les expressions de gauche à droite :

$$H(\xi : (\text{expression})) \quad \text{lire} \quad H(\xi \cap (\text{expression}))$$

$$H(\xi | (\text{expression})) \quad \text{lire} \quad H(\xi \cap \neg (\text{expression}))$$

$$H(\xi, (\text{expression})) \quad \text{lire} \quad H(\xi \cup (\text{expression}))$$

Aussi serait-il préférable d'utiliser, en cas d'ambiguïté, la notation ensembliste, par exemple $H(\xi \cap (\zeta \cap \neg \eta))$ pour $H(\xi; \zeta | \eta)$.

3.5.2. Résumé des propriétés de l'entropie discrète

Nous donnons dans le tableau 3.5 un bilan de certaines propriétés de l'entropie H qui ont été exposées dans les paragraphes précédents. Ces propriétés ont été choisies pour faciliter l'étude comparative des autres entropies (différentielle, quantique, algorithmique,...) avec H . Les propriétés de H sont en quelque sorte "canoniques", mais non nécessairement justifiées par rapport au concept d'information proprement dit. Elles sont issues d'un calcul statistique expérimentalement et axiomatiquement solide fondé sur des flux de données, mais ne sauraient caractériser que ce calcul lui-même, et non le concept de quantité d'information dont il n'est qu'une modélisation élémentaire.

<i>Tableau 3.5</i>	Entropie statistique classique	(discrète)
<i>Minimum</i>	$H(0) = H(1) = 0$ $H(\xi \xi) = 0$	L'entropie d'une variable certaine est nulle
<i>Majoration</i>	$H(\xi) \leq \log N$	L'entropie d'une variable à distribution uniforme $p = 1/N$ est maximale
<i>Concavité</i>	$H\left(\sum_{j=1}^N \lambda_j \mathbf{p}_j^{\mathbf{r}}\right) \geq \sum_{j=1}^N \lambda_j H(\mathbf{p}_j^{\mathbf{r}})$	L'entropie d'une moyenne est supérieure à la moyenne des entropies élémentaires.
<i>Monotonie</i>	$H(\xi, \eta) \geq H(\xi)$	L'entropie du tout est supérieure à l'entropie des parties.
<i>Additivité</i>	$H(\xi \eta) \leq H(\xi)$ $H(\xi, \eta) = H(\xi) + H(\eta \xi)$	La connaissance d'une information sur A ne peut que diminuer l'incertitude sur B.
<i>Sous-additivité</i>	$H(\xi, \eta) \leq H(\xi) + H(\eta)$	Les corrélations entre deux parties composant un système ne peuvent que diminuer l'entropie totale
<i>Sous-additivité "forte"</i>	$H(\xi, \eta, \zeta) + H(\eta) \leq H(\xi, \eta) + H(\eta, \zeta)$	Si 2 systèmes AB et BC (union ABC) se recouvrent partiellement en leur intersection B, alors $H\{ABC\} + H\{B\} \leq H\{AB\} + H\{BC\}$
<i>Signe</i>	$H(\xi) \geq 0$ $H(\xi, \eta) \geq 0$ $H(\xi \eta) \geq 0$ $H(\xi:\eta) \geq 0$	L'entropie est positive ou nulle
<i>Symétrie</i>	$H(\xi:\eta) = H(\eta:\xi)$ $= H(\xi) + H(\eta) - H(\xi, \eta)$	L'information contenue dans A à propos de B est la même que l'information contenue dans B à propos de A
<i>Changement de coordonnées</i>	$\xi \rightarrow \zeta \Rightarrow H(\xi) = H(\zeta)$	L'entropie est invariante dans un changement de coordonnées
<i>Aléatoire</i>	oui	Par définition, il existe une loi de probabilité.



3.5.3. Structure heuristique de la théorie statistique classique

Pour conclure, nous considérerons la structure heuristique de la théorie shannonienne de l'information : à quelles opérations cognitives et à quels calculs doit-on se livrer pour évaluer des quantités d'information ? Nous représenterons cette activité de modélisation à l'aide de quelques symboles graphiques. Le schéma général de calcul est :

$$i \rightarrow p(i) \rightarrow H(p)$$

index \rightarrow *probabilité* \rightarrow *entropie*

(i) Remarquons tout d'abord que le calcul booléen est une représentation sous forme algébrique de processus cognitifs discursifs relevant de la logique des propositions et de la logique des classes, qu'illustrent les diagrammes de Venn. D'un point de vue cognitif, il est significatif de remarquer qu'un même diagramme de Venn peut être lu dans différents langages (ensembliste, probabiliste, informationnel) : cela va dans le sens de l'hypothèse selon laquelle il s'agit bien d'une primitive cognitive.

Établir des tableaux de probabilités (voir par exemple tableau 3.1 à 3.3) résulte de ces opérations cognitives minimales. C'est une activité "multicruciale" par excellence, de classifications horizontale et verticale dans l'ordre d'énumération de chaque bibliothèque symbolique. Le décompte des objets permet d'établir les fréquences, d'où sont tirées les probabilités (on se place dans le cas pratique le plus courant où les probabilités sont calculées à partir des fréquences, mais dans le cas où celles-ci seraient posées axiomatiquement, on se trouverait aussi face à une activité cognitive de nature ensembliste). Convenons de représenter cette activité de classification et de décompte par les symboles  (cas à une dimension) et  (cas à deux dimensions). Notons que cette étape constitue toutefois le "maillon faible" de la théorie statistique classique des sources symboliques : ces pavés à rayures ou à petits carreaux cachent mal la pauvreté de nos connaissances sur l'origine cognitive des nombres $p(i)$, c'est-à-dire des probabilités.

(iii) Les probabilités ayant été évaluées, il reste à faire les calculs $H = \sum_i p \log p$. Nous représenterons ceux-ci d'après les techniques mises en jeu. La toute première d'entre elles est un calcul de moyenne pondérée. Anticipant sur les notations utilisées dans la prochaine section, nous dirons qu'une moyenne pondérée est, à une normalisation près, le résultat d'un produit scalaire de deux vecteurs duaux :

$$\frac{a_1 x_1 + a_2 x_2 + \dots}{N} = \frac{1}{N} (a_1 \quad a_2 \quad \dots) \begin{pmatrix} x_1 \\ x_2 \\ \dots \end{pmatrix} = \frac{1}{N} \langle \mathbf{a} | \mathbf{x} \rangle$$

D'où les symboles :

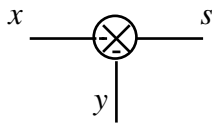
$\langle \cdot | \cdot \rangle$

opérateur produit scalaire

\log

opérateur logarithmique

Il reste enfin à calculer l'information mutuelle par :



opérateur $s = x - y$

Nous noterons également :

----- (*trait discontinu*)

grandeur nécessairement dénombrable

_____ (*trait continu*)

grandeur pouvant être non dénombrable

(iii) Le schéma heuristique de la théorie est symétrique. La connaissance des caractéristiques d'une source n'est qu'une connaissance statistique, qui ne fait pas intervenir la constitution interne de celle-ci, vue seulement comme une "boîte noire". On remarque que la connaissance des entropies conditionnelles nécessite d'associer (de connaître) deux fois les sources Σ et Γ , au niveau des probabilités d'une part (évaluation ensembliste des probabilités conditionnelles), au niveau du calcul de l'entropie proprement dite d'autre part (évaluation de l'espérance mathématique des entropies de chaque symbole).

3.6. Extension au continu : entropie différentielle

3.6.1 Source infinie discrète ou source dénombrable

L'extension de la formule de l'entropie statistique H au cas continu est délicat. Une première généralisation de la notion d'entropie d'une source symbolique finie discrète consiste à étendre cette définition à une source Σ de bibliothèque infinie discrète, c'est-à-dire dénombrable :

$$\Xi = \{\chi_1, \dots, \chi_i, \dots\}$$

qui émet des messages de longueur finie :

$$\sigma = \xi_1 \xi_2 \dots \xi_k \dots \xi_m \quad \text{avec } \xi_k \in \Xi \text{ et } \sigma \in \Xi^m$$

On définit la probabilité :

$$P(\xi = \chi_i) = p(i) \quad \text{qui a un sens si : } \sum_{i=1}^{\infty} p(i) = 1 \quad ; \quad p(i) \geq 0$$

L'entropie est définie par analogie avec le cas fini :

$$H(\xi) = E[-\log p(i)] = -\sum_{i=1}^{\infty} p(i) \log p(i)$$

Contrairement au cas fini, il n'y a pas de majoration de l'entropie par une constante. Mais la principale difficulté est que la convergence de cette série dépend de la loi de probabilité. Pour certaines lois, l'entropie n'est pas définie car la série ne converge pas.

Exemple : soit une loi définie par : $p(i) = \frac{q_i}{\sum_{i=1}^{\infty} q_i}$

La série $p(i) \log p(i)$: - converge pour $q_i = \frac{1}{(i+1)\log(i+1)}$

- diverge pour $q_i = \frac{1}{(i+1)(\log(i+1))^2}$

3.6.2. Entropie différentielle d'une source échantillonnée

On considère une source de messages formés de suites finies dénombrables de signes pris dans une bibliothèque symbolique infinie non dénombrable [Kolmogorov, 1956].

Soit une variable aléatoire ξ absolument continue obéissant à une loi de densité $p(x)$:

$$P(x \leq \xi \leq x + dx) = p(x) dx \quad \text{qui a un sens si : } \int_{-\infty}^{+\infty} p(x) dx = 1$$

En première approximation, l'évaluation de l'entropie résulte, par analogie avec le cas discret, du passage à la limite dans la définition de l'entropie d'une source discrète :

$$H(\xi) = -\sum_{i=1}^n p(i) \log p(i) \rightarrow H_d(\xi) = E[-\log p(x)] = -\int_{-\infty}^{+\infty} p(x) \log p(x) dx$$

Définition 3.4 : on appelle *entropie différentielle* la quantité $H_d(\xi)$.

Comme dans le cas discret, l'entropie différentielle est une fonction concave. Shannon envisage par exemple la moyenne généralisée $\int_a^b \lambda(x, y) p(x) dx$, si $\lambda(x)$ et $p(x)$ sont des fonctions mesurables, avec $\lambda(x) \geq 0$ et $\int_a^b \lambda(x, y) dx = \int_a^b \lambda(x, y) dy = 1$. Il vient :

$$H\left(\int_a^b \lambda(x, y) p(x) dx\right) \geq H(p(x))$$

On peut aussi considérer cette propriété sous l'angle de moyennes sur un ensemble discret de distributions continues :

$$H\left(\sum_{j=1}^N \lambda_j p_j(x)\right) \geq \sum_{j=1}^N \lambda_j H(p_j(x))$$

Exemple : entropie différentielle d'une source gaussienne (loi normale).

On suppose la loi de probabilité centrée :

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{x^2}{2\sigma^2}}$$

Il vient :

$$H_d(\xi) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{x^2}{2\sigma^2}} \left[\log(\sigma\sqrt{2\pi}) + \frac{x^2}{2\sigma^2} \log e \right] dx$$

A l'aide du changement de variable $\frac{x^2}{2\sigma^2} \rightarrow t$ et sachant que $\int_{-\infty}^{+\infty} e^{-t^2} dt = \sqrt{\pi}$, il vient :

$$H_d(\xi) = \log \sigma\sqrt{2\pi e}$$

Théorème 3.7 :

Pour une puissance donnée, la quantité $\log \sigma \sqrt{2\pi e}$ maximise l'entropie H_d .

□ L'inégalité de Gibbs (cf §3.3.3) peut être étendue aux densités $p(x)$ et $q(x)$ telles que

$$\int p(x) = 1 \text{ et } \int q(x) = 1 :$$

$$\int_{-\infty}^{+\infty} p(x) \ln \frac{q(x)}{p(x)} \leq \int_{-\infty}^{+\infty} p(x) \left(\frac{q(x)}{p(x)} - 1 \right) = \int_{-\infty}^{+\infty} q(x) - \int_{-\infty}^{+\infty} p(x) = 1 - 1 = 0$$

De $\int_{-\infty}^{+\infty} p(x) \log \frac{p(x)}{q(x)} \geq 0$, on déduit $H_d(\xi) \leq -\int_{-\infty}^{+\infty} p(x) \log q(x) dx$. Supposons que ξ

soit une v.a. centrée d'ordre 2, c'est-à-dire de "puissance" finie $\int p(x)x^2 dx = \sigma^2$. Il vient :

$$H_d(\xi) \leq \log \sigma \sqrt{2\pi} \int_{-\infty}^{+\infty} p(x) dx + \frac{\log e}{2\sigma^2} \int_{-\infty}^{+\infty} p(x)x^2 dx$$

$$\text{ou : } H_d(\xi) \leq \log \sigma \sqrt{2\pi e}$$

□

L'extension de l'entropie du cas discret au continu, qui concerne des quantités de nature différente ($p(i)$ est un probabilité, alors que $p(x)$ est une densité de probabilité, qui peut être supérieure à 1), soulève cependant un certain nombre de difficultés [voir par ex. Reza, 1961, Rényi, 1962, Réfrégier, 1993], comme le remarquait également Shannon lui-même.

(i) Théorème 3.8 : l'entropie différentielle d'une v.a. à distribution continue peut être négative.

Exemple : Si $p(x)$ est une loi de probabilité de densité constante et égale à $1/A$, alors la relation précédente conduit à $H_d = \log A$. Non seulement un tel résultat, selon la valeur de A , peut être négatif, mais encore H_d dépend de cette amplitude A . Or l'information portée par un signal devrait être invariante dans toute homothétie : l'information portée par un texte devrait être la même quelle que soit la taille des lettres, l'information portée par la parole ne devrait pas dépendre du niveau de la voix – tant qu'elle reste intelligible, etc.

De même, une loi normale centrée conduit à un résultat, positif, négatif ou nul, qui dépend d'une valeur particulière de l'écart-type.

(ii) Théorème 3.9 : l'entropie différentielle d'une v.a. à distribution continue peut être non bornée.

□ Considérons par exemple une v.a. ξ continue définie sur $[a, b]$, dont la valeur x est échantillonnée avec un pas Δx . Sur $[a, b]$ il y a un nombre fini d'intervalles $1, \dots, i, \dots, N$. La

densité de probabilité est définie par :

$$P(a = x_0 \leq x_{i-1} \leq \xi \leq x_i \leq x_N = b) = \int_{\Delta x = x_i - x_{i-1}} p(x) dx = P_i \Delta x$$

Si la loi de probabilité est correctement définie, on a $\int_a^b p(x) dx = 1$, donc $\sum_{i=1}^N P_i \Delta x = 1$,

ce qui permet de définir la loi de probabilité discrète $p(i) = P_i \Delta x$. Aussi l'entropie de la variable échantillonnée ξ_e est :

$$H(\xi_e) = -\sum_{i=1}^N P_i \Delta x \log P_i \Delta x = -\sum_{i=1}^N P_i \Delta x \log P_i - \sum_{i=1}^N P_i \Delta x \log \Delta x$$

Par définition de $H_d(\xi)$, il vient :

$$H_d(\xi) = \lim_{\Delta x \rightarrow 0} H(\xi_e) = -\int_a^b p(x) \log p(x) dx - \lim_{\Delta x \rightarrow 0} \sum_{i=1}^N P_i \Delta x \log \Delta x$$

Le terme $\lim_{\Delta x \rightarrow 0} \sum_{i=1}^N P_i \Delta x \log \Delta x$ ne converge pas, car $\sum_{i=1}^N P_i \Delta x = 1$, mais $\lim_{\Delta x \rightarrow 0} \log \Delta x = -\infty$.

▣

Si l'on diminue le pas d'échantillonnage, l'information devient potentiellement infiniment grande : tout dépend en réalité du pouvoir séparateur du récepteur qui détermine, comme nous le verrons plus loin, l'échantillonnage et sa résolution.

(iii) Théorème 3.10 : l'entropie différentielle d'une v.a. à distribution continue n'est pas nécessairement invariante dans un changement de coordonnées.

□ Dans le cas d'une loi discrète, cette invariance est respectée. Par exemple, aux faces d'un dé $\xi = \{1,2,3,4,5,6\}$ correspondent les probabilités $(1/6, 1/6, 1/6, 1/6, 1/6, 1/6)$, ce qui donne $H(\xi) = \log 6$. Un changement de variable, par exemple $\xi' = \xi^2 = \{1,4,9,16,25,36\}$, ne modifie pas la loi de probabilité, et donc $H(\xi^2) = \log 6$.

Dans le cas continu, un changement $x' = f(x)$ supposé bien formé conduit à une densité :

$$p'(x') = p(x) \left| \frac{dx}{dx'} \right|$$

et à l'entropie associée :

$$H'_d(\xi) = -\int_{-\infty}^{+\infty} p'(x') \log p'(x') dx'$$

$$H'_d(\xi) = -\int_{-\infty}^{+\infty} p(x) \left| \frac{dx}{dx'} \right| \log p(x) \left| \frac{dx}{dx'} \right| dx'$$

$$H'_d(\xi) = -\int_{-\infty}^{+\infty} p(x) \log p(x) dx - \int_{-\infty}^{+\infty} p(x) \log \left| \frac{dx}{dx'} \right| dx$$

$$H'_d(\xi) = H_d(\xi) + \int_{-\infty}^{+\infty} p(x) \log \left| \frac{dx'}{dx} \right| dx$$

▣

Dans le cas continu, l'entropie dans le nouveau système de coordonnées dépend de la loi $\log |dx'/dx|$.

Remarque : dans le cas particulier où $x' = ax + b$, il vient : $H'_d(\xi) = H_d(\xi) + \log |a|$. Pour une transformation linéaire des coordonnées, l'entropie d'une loi continue reste invariante à une constante $\log |a|$ près.

3.6.3. Entropies différentielles de sources échantillonnées conjointes

En étendant le cas discret à deux v.a. absolument continues, on définit les différentes densités de probabilité correspondantes :

$$p(x,y) = p(x).p(y/x) = p(y).p(x/y)$$

$$\text{où : } P(x \leq \xi \leq x + dx) = p(x) dx$$

$$P(y \leq \eta \leq y + dy \text{ si } x \leq \xi \leq x + dx) = p(y/x)$$

$$\text{avec : } p(x) = \int_{\mathbb{R}} p(x,y) dy$$

$$p(y) = \int_{\mathbb{R}} p(x,y) dx$$

$$\int_{\mathbb{R}} p(x) dx = 1$$

$$\int_{\mathbb{R}} p(y) dy = 1$$

$$\iint_{\mathbb{R}^2} p(x,y) dx dy = 1$$

$$\iint_{\mathbb{R}^2} p(x)p(y) dx dy = 1$$

Définition 3.5 : les différentes entropies $H_d(\xi)$, $H_d(\eta)$, $H_d(\xi|\eta)$, $H_d(\eta|\xi)$, $H_d(\xi,\eta)$ en découlent :

$$H_d(\xi,\eta) = -\iint_{\mathbb{R}^2} p(x,y) \log p(x,y) dx dy$$

$$H_d(\xi|\eta) = -\iint_{\mathbb{R}^2} p(x/y) \log p(x/y) dx dy = -\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} p(x,y) \log \frac{p(x,y)}{p(y)} dx dy$$

ainsi que l'information mutuelle :

$$H_d(\xi;\eta) = H_d(\xi) - H_d(\xi|\eta) = H_d(\eta) - H_d(\eta|\xi) = H_d(\xi) + H_d(\eta) - H_d(\xi,\eta)$$

$$H_d(\xi;\eta) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dx dy$$

Dans le passage au continu, les différentes entropies $H_d(\xi|\eta)$, $H_d(\eta|\xi)$, $H_d(\xi,\eta)$ soulèvent les mêmes difficultés que précédemment.

Par contre, celles-ci disparaissent dans le cas de l'information mutuelle, qui, en outre, conserve ses propriétés de symétrie :

(i) *Théorème 3.11* : l'information mutuelle entre deux v.a. continues n'est jamais négative.

□ De l'équation précédente et de la convexité de la fonction logarithme on déduit :

$$H_d(\xi;\eta) \geq \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} p(x,y) \left[\frac{p(x,y)}{p(x)p(y)} - 1 \right] \log e dx dy$$

$$H_d(\xi;\eta) \geq \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} p(x,y) \log e dx dy + \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} p(x), p(y) \log e dx dy$$

$$H_d(\xi;\eta) \geq 1 \cdot \log e - 1 \cdot 1 \cdot \log e$$

$$H_d(\xi;\eta) \geq 0$$

▣

Corollaires

• L'information différentielle est sous-additive :

$$H_d(\xi;\eta) = H_d(\xi) + H_d(\eta) - H_d(\xi,\eta) \geq 0 \Rightarrow H_d(\xi,\eta) \leq H_d(\xi) + H_d(\eta)$$

Mais cette inégalité (égalité dans le cas de variables indépendantes) est maintenant une inégalité algébrique. On ne peut donc plus affirmer, par exemple, que $H_d(\xi,\eta) \leq H_d(\xi)$: l'entropie différentielle n'est pas monotone.

• Par contre, de :

$$H_d(\xi;\eta) = H_d(\xi) - H_d(\xi|\eta) \geq 0$$

on déduit que :

$$H_d(\xi|\eta) \leq H_d(\xi)$$

Comme dans le cas discret, la connaissance d'une information auxiliaire n'augmente pas l'incertitude.

(ii) *Théorème 3.12* : l'information mutuelle est finie.

□ Le passage à la limite des densités absolument continues est défini par :

$$P(x_{i-1} \leq \xi \leq x_i) = \int_{\Delta x} p(x) dx = P_x \Delta x$$

$$P(y_{i-1} \leq \eta \leq y_i) = \int_{\Delta y} p(y) dy = P_y \Delta y$$

$$P(x_{i-1} \leq \xi \leq x_i, y_{i-1} \leq \eta \leq y_i) = \int_{\Delta x} \int_{\Delta y} p(x, y) dx dy = P_{xy} \Delta x \Delta y$$

D'où :

$$H_d(\xi_e; \eta_e) = \sum_{i=1}^N \sum_{j=1}^M P_{xy} \Delta x \Delta y \log \frac{P_{xy} \Delta x \Delta y}{P_x \Delta x P_y \Delta y} = \sum_{i=1}^N \sum_{j=1}^M P_{xy} \Delta x \Delta y \log \frac{P_{xy}}{P_x P_y}$$

Les quantités Δx et Δy se simplifient dans le logarithme, si bien que le passage à la limite dans cette expression ne fait plus apparaître de termes en $\log \Delta x$ ou $\log \Delta y$ qui tendraient vers l'infini :

$$H_d(\xi; \eta) = \lim_{\Delta x, \Delta y \rightarrow 0} H_d(\xi_e; \eta_e) = \int_a^b p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dx dy$$

□

Donc l'information mutuelle reste finie, même si les quantités dont elle est le rapport tendent vers l'infini. Remarquons que ce résultat est établi en considérant l'information mutuelle comme une forme particulière d'écart entropique.

(iii) Théorème 3.13 : l'information mutuelle est strictement invariante sous une transformation linéaire des coordonnées.

□ D'après les résultats précédents, il vient :

$$x' = a x + b$$

$$y' = c y + d$$

$$H'_d(\xi) = H_d(\xi) + \log |a|$$

$$H'_d(\eta) = H_d(\eta) + \log |c|$$

$$H'_d(\xi, \eta) = H_d(\xi, \eta) + \log |ac|$$

D'où :

$$H'_d(\xi; \eta) = H'_d(\xi) + H'_d(\eta) - H'_d(\xi, \eta)$$

$$= H_d(\xi) + H_d(\eta) - H_d(\xi, \eta) + \log |a| + \log |c| - \log |ac|$$

$$H'_d(\xi; \eta) = H_d(\xi; \eta)$$

□

3.6.4. Conclusions

3.6.4.1. Quantification

Le principal problème rencontré lors du calcul de l'entropie différentielle est un problème de divergence : si $Q = 2$, $\Xi = [0,1]$ et ξ uniformément distribuée dans Ξ , alors la transcription binaire de ξ est une suite infinie de variables aléatoires mutuellement indépendantes (caractères $\in \{\emptyset, 1\}$ de probabilité 1/2) fournissant l'information $1+1+1+\dots = \infty$. L'information est infinie si ξ est connue avec une précision infinie. Dans la réalité cela est impossible : il faut donc se contenter d'approximer la distribution continue par une distribution discrète, et observer dans le passage à la limite le comportement de la différence entre ces deux distributions.

Supposons que Ξ soit l'ensemble \mathbb{R} , ou une partie de celui-ci. Soit $\xi_{1:N}$ une quantification de ξ à la N ème décimale binaire :

$$\xi_{1:N} = \frac{\lfloor N \cdot \xi \rfloor}{N}$$

On pose :

$$p_N(i) = p\left(\xi_{1:N} = \frac{i}{N}\right) = p\left(\frac{i}{N} \leq \xi \leq \frac{i+1}{N}\right), \quad i = 0, \pm 1, \pm 2, \dots \quad N = 1, 2, \dots$$

Rényi établit le théorème suivant :

Théorème 3.14 : Soit ξ une variable aléatoire absolument continue et de densité $p(x)$. Soit $\xi_{1:N} = \frac{\lfloor N \cdot \xi \rfloor}{N}$. Si $H(\xi_{1:1})$ est finie et si l'intégrale $-\int_{-\infty}^{+\infty} p(x) \cdot \log p(x) \cdot dx$ existe, alors :

$$\lim_{N \rightarrow \infty} (H(\xi_{1:N}) - \log_2 N) = -\int_{-\infty}^{+\infty} p(x) \cdot \log p(x) \cdot dx$$

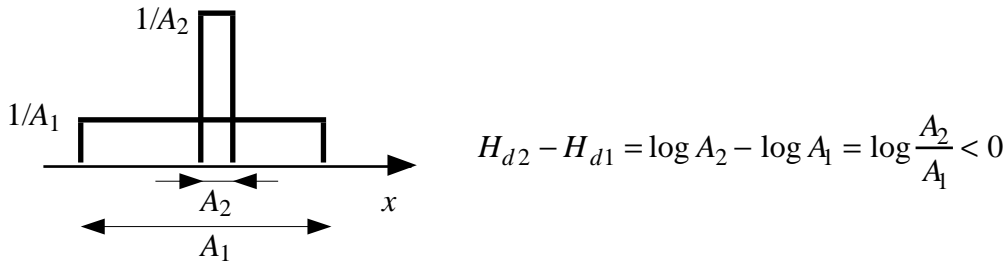
avec :

$$H(\xi_{1:N}) = -\sum_{i=-\infty}^{+\infty} p_N(i) \cdot \log p_N(i)$$

Nous renvoyons à [Rényi 1962, 1992] pour la démonstration de ce théorème.

$H_d(\xi)$ est la limite, pour $N \rightarrow \infty$, d'un écart entropique entre $H(\xi_{1:N})$ et $\log N$, mais au signe près, puisque, comme on vient de le voir, le signe de $H_d(\xi)$ n'est pas défini. Le passage au continu fait donc perdre à l'entropie une caractéristique essentielle, à savoir le signe de la redondance, par rapport à l'entropie maximale à laquelle on peut s'attendre pour un niveau de quantification donné.

En revanche, l'information mutuelle reste positive ou nulle et, en tant qu'écart entropique, représente bien un gain d'information. Cela est conforme au raisonnement simple suivant : si on considère deux segments de longueur A_1 et $A_2 < A_1$, le deuxième segment étant inclu dans le premier, associés respectivement aux distributions $p_1(x) = 1/A_1$ et $p_2(x) = 1/A_2$, il vient :



ce qui correspond à une diminution d'indétermination, donc à un gain d'information.

En résumé, l'extension au continu change la nature des quantités d'information, qui, d'absolues, deviennent relatives ou différentielles. Excepté l'information mutuelle, non seulement les signes des entropies sont indéterminés, mais l'échelle des phénomènes intervient : même si l'on se restreint au cas de changements d'échelle linéaires, les entropies différentielles ne sont définies qu'à une constante près.

Il en découle qu'il est nécessaire de considérer une quantité d'information comme une grandeur quantifiée, irréductiblement liée à la notion de discernabilité. L'extension à des grandeurs continues ne peut être définie que de façon relative.

3.6.4.2. Résumé des propriétés de l'information différentielle

Dans le tableau 3.6, on marque en grisé les propriétés de H_d qui diffèrent de l'entropie discontinue H .

<i>Tableau 3.6</i>	Entropie différentielle	(continue)
<i>Minimum</i>	$H_d(0) = H_d(1) = 0$ mais $\exists p(x) : H_d(p(x)) = 0$ $H_d(\xi \xi) = 0$	L'entropie n'a pas de minimum défini (celui-ci dépend de la forme de la fonction de densité de probabilité)
<i>Majoration</i>	$H_d(\xi) \leq \log(\sigma\sqrt{2\pi e})$ mais $\exists p(x) : H_d(p(x)) \rightarrow \infty$	L'entropie d'une v.a. est maximum lorsque la v.a. est normale. Mais dans le cas général, l'entropie peut être infinie.
<i>Concavité</i>	$H\left(\sum_{j=1}^N \lambda_j p_j(x)\right) \geq \sum_{j=1}^N \lambda_j H(p_j(x))$	L'entropie d'une moyenne est supérieure à la moyenne des entropies élémentaires.
<i>Monotonie</i>	$H_d(\xi, \eta) \not\leq H_d(\xi)$	L'entropie du tout peut être <, > ou = à l'entropie des parties.
<i>Additivité</i>	$H_d(\xi \eta) \leq H_d(\xi)$ $H_d(\xi, \eta) = H_d(\xi) + H_d(\eta \xi)$	La connaissance d'une information sur A ne peut que diminuer l'incertitude sur B.
<i>Sous-additivité</i>	$H_d(\xi, \eta) \leq H_d(\xi) + H_d(\eta)$	Les corrélations entre deux sources composant un système ne peuvent que diminuer l'entropie totale
<i>Sous-additivité "forte"</i>		
<i>Signe</i>	$H_d(\xi) \not\geq 0$ $H_d(\xi \eta) \not\geq 0$ $H_d(\xi, \eta) \not\geq 0$ $H_d(\xi; \eta) \geq 0$	L'entropie différentielle peut être négative. L'information mutuelle est toujours positive.
<i>Symétrie</i>	$H_d(\xi; \eta) = H_d(\eta; \xi)$ $= H_d(\xi) + H_d(\eta) - H_d(\xi, \eta)$	L'information contenue dans A à propos de B est la même que l'information contenue dans B à propos de A
<i>Changement de coordonnées</i>	$H_d'(\xi) = H_d(\xi) + \int_{-\infty}^{+\infty} p(x) \log \left \frac{dx'}{dx} \right dx$	L'entropie n'est pas invariante dans un changement de coordonnées
<i>Aléatoire</i>	oui, sous réserve.	Par définition, il existe une loi de densité de probabilité. Mais les intégrales ne convergent pas nécessairement.