INTRODUCTION A LA THEORIE DE L'INFORMATION

I Définition de l'information

Transmission de l'information : l'employé du télégraphe, pour calculer la somme à payer, ne fait entrer en ligne de compte que le nombre de mots du télégramme.
On code les lettres en binaire : 32 symboles 0 ou 1.
Une information = un signe distinctif d'un élément d'un ensemble E.
Si $\text{card}\, E = 2^n$, on peut numéroter les éléments de E en binaire à n chiffres.
$N = 2^n$, $n = \log_2 N$ bits.

Définition de Hartley : $I(E_N) = \log_2 N$ même si N n'est pas une puissance de 2 ($N = \text{card}\, E_N$).

Proposition : $\begin{cases} I(E_{MN}) = I(E_M) + I(E_N) \\ I(E_N) \leq I(E_{N+1}) \\ I(E_2) = 1 \end{cases}$

démonstration : $E_{MN} = E_M^{(1)} \cup E_M^{(2)} \cup \dots \cup E_M^{(N)}$. Pour caractériser un élément de E_{MN}, on doit d'abord savoir à quel $E_M^{(i)}$ il appartient (ce qui nécessite l'information $I(E_N)$ puisqu'il y a N parties), puis on doit identifier l'élément dans $E_M^{(i)}$ (ce qui nécessite l'information $I(E_M^{(i)}) = I(E_M)$). D'où le besoin de $I(E_N) + I(E_M)$ informations.

La démonstration vient d'être faite dans le cas où les $E_M^{(i)}$ ont même information.
Cas général : $E_N = E_{N_1}^{(1)} \cup \dots \cup E_{N_n}^{(n)}$ avec les $E_{N_i}^{(i)}$ deux à deux disjoints, $N = \sum_{i=1}^{n} N_i$.

Pour caractériser un élément de E_N, il faut savoir à quel $E_{N_i}^{(i)}$ il appartient (ce qui demande une quantité d'information I_1), puis sachant qu'il est dans $E_{N_i}^{(i)}$, il faut une quantité d'information $I(E_{N_i}^{(i)})$ supplémentaire pour l'identifier. Or les N_i ne sont pas nécessairement tous identiques. Il faut donc une information moyenne $\sum_{i=1}^{n} p_i I(E_{N_i}^{(i)})$ où $p_i = \frac{N_i}{N}$.

Au total : $I(E_N) = I_1 + \sum_{i=1}^{n} p_i I(E_{N_i}^{(i)})$

$$\Rightarrow I_1 = \log_2 N - \sum_{i=1}^{n} p_i \log_2 N_i = \log_2 N - \sum_{i=1}^{n} p_i \log_2 N - \sum_{i=1}^{n} p_i \log_2 p_i$$

Formule de Shannon : $\boxed{I_1 = \sum_{i=1}^{n} p_i \log_2 \frac{1}{p_i}}$
Claude Elwood Shannon, 1916-...

Si $p_1 = \dots = p_n = \frac{1}{n}$ on trouve $I_1 = \log_2 n$ ($N = n N_i, \forall i$)

Définition : Entropie : soit $\mathcal{P} = (p_1, \dots, p_n)$ une distribution de probabilité discrète finie (i.e. $p_i \in [0,1]$ et $\sum p_i = 1$). L'entropie de $\mathcal{P}$ est

$$H(\mathcal{P}) = \sum_{i=1}^{n} p_i \log_2 \frac{1}{p_i}$$

On parle aussi de quantité d'information, à comprendre plutôt comme quantité d'indécision (ou d'incertitude).
Plus l'entropie est grande, plus l'énergie du système est petite : tendance vers l'état d'énergie minimale (chaos) ... Lien avec la thermodynamique avec Boltzmann.

Si X est une v.a. suivant la loi $\mathcal{P}$: $P(X = x_i) = p_i$, on note aussi

$H(X) = H(\mathcal{P})$, définition indépendante des valeurs prises par X.

Si f est une injection sur $\{x_1, \dots, x_n\}$, $H(X) = H(f(X))$.

Autre forme de $H(X)$: soit $f_X(x) = \mathbb{P}\{X = x\}$. Alors $H(X) = \mathbb{E}\left[\log_2 \frac{1}{f_X(X)}\right]$

<u>Proposition</u> : $H(\mathcal{P}) \leq \log_2 n$ et l'égalité a lieu ssi $p_1 = \dots = p_n = \frac{1}{n}$ (i.e. $\mathcal{P}$ uniforme).

<u>Démonstration</u> : $\varphi(x) = x \log_2 x$ est convexe donc

$$H(\mathcal{P}) = -\sum_{i=1}^{n} p_i \log_2 p_i = -\sum_{i=1}^{n} \varphi(p_i) \leq -n\,\varphi\Big(\frac{1}{n}\underbrace{\sum_{i=1}^{n} p_i}_{1}\Big) = \log_2 n .$$

Il y a égalité ssi les p_i sont tous égaux de somme 1 i.e. $p_i = \frac{1}{n}$. $\square$

<u>Interprétation</u> : s'il y a n possibilités pour le résultat d'une épreuve, l'indétermination est maximale quand toutes les éventualités sont équiprobables (pas de site privilégié, tendance vers le chaos).
S'il y a un site privilégié e.g. $p_1 = 1$, $p_2 = \dots = p_n = 0$, $H(\mathcal{P}) = 0 \to$ aucune indétermination.

<u>Accroissement de l'indétermination.</u>

Soit $\mathcal{P} = (p_1, \dots, p_n)$ et $W = (w_{ij})_{\substack{1 \leq i \leq n \\ 1 \leq j \leq m}}$ une matrice doublement stochastique $\begin{cases} \sum_{i=1}^{n} w_{ij_0} = 1, \ \forall j_0 \\ \sum_{j=1}^{m} w_{i_0 j} = 1, \ \forall i_0 \\ w_{ij} \geq 0 \end{cases}$

$Q = \mathcal{P}W = (q_1, \dots, q_m)$ est une distribution de probabilité $\left(\sum_{i=1}^{m} q_i = 1\right)$

et $H(\mathcal{P}) \leq H(Q)$ (l'indétermination s'est accrue).

En effet : $q_j \log_2 q_j = \varphi\left(\sum_{i=1}^{n} p_i w_{ij}\right) \leq \sum_{i=1}^{n} w_{ij}\,\varphi(p_i) =$

$$-H(Q) = \sum_{j=1}^{m} q_j \log_2 q_j \leq \sum_{i=1}^{n} \varphi(p_i) \underbrace{\sum_{j=1}^{m} w_{ij}}_{1} = \sum p_i \log_2 p_i = -H(\mathcal{P}).$$

<u>II Information conditionnelle - Transinformation.</u>

Soit X, Y deux v.a. discrètes finies. On pose

$$\begin{cases} P(X = x_i) = p_i & \mathcal{P} = (p_1, \dots, p_n) \\ P(Y = y_j) = q_j & Q = (q_1, \dots, q_m) \\ P(X = x_i, Y = y_j) = r_{ij} & \mathcal{R} = (r_{ij})_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}} \end{cases}$$

$$\begin{cases} P(X = x_i \mid Y = y_j) = p_{i|j} & \mathcal{P}_j = (p_{1|j}, \dots, p_{m|j}) \\ P(Y = y_j \mid X = x_i) = q_{j|i} & Q_i = (q_{1|i}, \dots, q_{n|i}) \end{cases}$$

On a $r_{ij} = p_{i|j}\, q_j = q_{j|i}\, p_i$, $\sum_{j=1}^{m} r_{ij} = p_i$, $\sum_{i=1}^{m} r_{ij} = q_j$

Définition : $\ast$ $H(\mathcal{R}) = \boxed{H(X,Y) = \sum_{i,j} r_{ij} \log_2 \frac{1}{r_{ij}}} = E\left[\log_2 \frac{1}{f_{(X,Y)}(X,Y)}\right]$ avec $f_{(X,Y)}(x,y) = \mathbb{P}(X=x, Y=y)$

$\ast$ $H(\mathcal{P}|\mathcal{Q}) = H(X|Y) = \mathbb{E}\left[\log_2 \frac{1}{f_{X|Y}(X,Y)}\right]$ avec $f_{X|Y}(x,y) = \mathbb{P}(X=x|Y=y) = \frac{f_{(X,Y)}(x,y)}{f_Y(y)}$

$$= \sum_{i,j} r_{ij} \log_2 \frac{1}{p_{i|j}} = \boxed{\sum_{i,j} r_{ij} \log_2 \frac{q_j}{r_{ij}}} = \sum_{j=1}^{m} q_j \underbrace{H(\mathcal{P}_j)}_{\sum_{i=1}^{m} p_{i|j} \log_2 \frac{1}{p_{i|j}}}$$

Proposition : $H(X,Y) = H(Y) + H(X|Y) = H(X) + H(Y|X)$, ou : l'information contenue dans (X,Y) = l'information contenue dans X sachant Y + l'information contenue dans Y.

Démonstration : $H(X,Y) - H(X|Y) = \sum_{i,j} r_{ij} \log \frac{1}{q_j} = \sum_{j=1}^{m} q_j \log \frac{1}{q_j} = H(Y)$. $\square$

Si X, Y sont indépendantes, $H(X|Y) = H(X)$ et $H(X,Y) = H(X) + H(Y)$.
$H(X|X) = 0$ car $r_{ij} = p_i \delta_{ij}$.

Proposition : $H(X,Y) \leq H(X) + H(Y)$ ou encore $H(X|Y) \leq H(X)$: l'information contenue dans X sachant ne peut dépasser l'information totale de X.

Démonstration : $H(X|Y) = \sum_{i,j} r_{ij} \log_2 \frac{q_j}{r_{ij}} = -\sum_{i=1}^{m}\left[\sum_{j=1}^{m} q_j \underbrace{\left(p_{i|j} \log_2 p_{i|j}\right)}_{\varphi(p_{i|j})}\right] \leq -\sum_{i=1}^{m} \varphi\left(\sum_{j=1}^{m} q_j p_{i|j}\right) = -\sum_{i=1}^{m} \varphi(p_i)$
$\leq H(X)$.

égalité ssi les $p_{i|j}$ sont tous égaux, i.e. X, Y indépendantes.

Transinformation : $I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y) = \boxed{\sum_{i,j} r_{ij} \log_2 \frac{r_{ij}}{p_i q_j}}$

C'est l'information relative donnée par Y sur X. $= E\left[\log_2 \frac{f_{(X,Y)}(X,Y)}{f_X(X) f_Y(Y)}\right]$

Proposition : $I(X,Y) = I(Y,X)$ et $0 \leq I(X,Y) \leq H(X) \wedge H(Y)$. Si X, Y sont indépendantes, $I(X,Y) = 0$.

Interprétation : $I(X,Y)$ = diminution d'indétermination résultant de la connaissance de Y
= information sur X que l'on peut extraire de la valeur de Y
$\rightarrow$ mesure de dépendance stochastique entre X et Y.

La propriété de symétrie signifie que Y donne autant d'information sur X que X sur Y.

exemple : Soit Y symétrique par rapport à 0 et $\mathbb{P}(Y=0) = 0$, $X = Y^2$.
Valeurs de Y : $0, \pm y_1, \dots, \pm y_n$. Valeurs de X : $0, y_1^2, \dots, y_n^2$.
$\forall i$, $\mathbb{P}(X = y_i^2) = 2\,\mathbb{P}(Y = y_i)$ $\Rightarrow$ $\log_2 \mathbb{P}(X = y_i^2) = 1 + \log_2 \mathbb{P}(Y = y_i)$

D'où $H(Y) = H(X) + 1$: si on connaît $|Y|$, Y prend les valeurs $\pm |Y|$ avec probabilité $\frac{1}{2}$, il reste donc une unité d'indétermination sur le signe.

Proposition : Soit f une application. $I(X, f(Y)) \leq I(X,Y)$ ou encore $H(X|f(Y)) \geq H(X|Y)$.

interprétation : si au lieu d'observer Y, on observe seulement $f(Y)$, on a alors plus d'indétermination sur X.

Démonstration : * si f est injective il y a égalité (les valeurs de $f(Y)$ sont toutes distinctes).

* si $f(y_k) = f(y_\ell) \neq f(y_m) \quad \forall m \neq k, \ell$ et les $f(y_m)$, $m \neq k,\ell$ tous distincts :

valeurs de $f(Y)$: $\{ f(y_m), m \neq k, \ell \text{ et } f(y_k) = f(y_\ell) \}$.

$$\mathbb{P}(f(Y) = f(y_m)) = \mathbb{P}(Y = y_m) = q_m \text{ pour } m \neq k, \ell,$$
$$\mathbb{P}(f(Y) = f(y_k)) = \mathbb{P}(Y = y_k) + \mathbb{P}(Y = y_\ell) = q_k + q_\ell .$$

$$H(X|f(Y)) = \sum_j \underbrace{q_j^f}_{\text{associé à } f(Y)} \underbrace{H(\mathcal{P}_j^f)} = \sum_{m \neq k,\ell} q_m \underbrace{H(\mathcal{P}_m)}_{H(X|Y=y_m)} + (q_k + q_\ell) \underbrace{H(\mathcal{P}_{k,\ell})}_{H(X|Y=y_k \text{ ou } Y=y_\ell)} .$$

Or $\mathbb{P}(X = x_i \mid Y = y_k \text{ ou } Y = y_\ell) = \dfrac{r_{ik} + r_{i\ell}}{q_k + q_\ell} = \dfrac{q_k p_{i|k} + q_\ell p_{i|\ell}}{q_k + q_\ell}$

donc $H(\mathcal{P}_{k,\ell}) = -\sum_i \varphi\left(\dfrac{q_k p_{i|k} + q_\ell p_{i|\ell}}{q_k + q_\ell}\right) \geq -\sum_i \dfrac{q_k \varphi(p_{i|k}) + q_\ell \varphi(p_{i|\ell})}{q_k + q_\ell}$

$$\geq \frac{q_k H(\mathcal{P}_k) + q_\ell H(\mathcal{P}_\ell)}{q_k + q_\ell}$$

ou encore $q_k H(\mathcal{P}_k) + q_\ell H(\mathcal{P}_\ell) \leq (q_k + q_\ell) H(\mathcal{P}_{k,\ell})$.

Ainsi $H(X|f(Y)) \geq \sum_m q_m H(\mathcal{P}_m) = H(X|Y)$.

* le même raisonnement est applicable si plus de valeurs de f coïncident... □

III Gain d'information

Soit $E_N = E_{N_1}^{(1)} \cup E_{N_2}^{(2)} \cup \dots \cup E_{N_n}^{(n)}$, $\operatorname{card} E_{N_i}^{(i)} = N_i = p_i N$.

Un élément choisi au hasard dans E_N peut être caractérisé de deux manières différentes :
1) soit par son numéro dans E_N, noté X.
2) soit par la donnée de E_{N_k} le contenant (k noté Y), puis de son numéro dans $E_{N_k}^{(k)}$, noté Z.

Alors $H(X) = H(Y) + H(Z|Y) = H(Y,Z)$ avec $H(X) = \log_2 N$, $H(Y) = \sum_{i=1}^{n} p_i \log_2 \frac{1}{p_i}$, $H(Z|Y) = \sum_{i=1}^{n} p_i \underbrace{\log_2 N_i}_{H(Z|Y=i)}$

Soit maintenant $E'_{N'} \subset E_N$, $E'^{(i)}_{N'_i} = E_N \cap E^{(i)}_{N_i}$ $\quad q_i = \dfrac{N'_i}{N'}$, $N'_i = \operatorname{card} E'^{(i)}_{N'_i}$.

Supposons que l'on sache d'un élément pris au hasard dans E_N qu'il est dans $E'_{N'}$.
Combien d'information cela fournit-il sur Y ?
La distribution a priori de Y était $\mathcal{P} = (p_1 \cdots p_n)$. Après l'information disant que l'élément est dans E', Y a la distribution a posteriori $\mathcal{Q} = (q_1 \cdots q_n)$.
On note alors $I(\mathcal{Q} \| \mathcal{P})$ le gain d'information pour Y sachant que l'élément est dans Y.

On a $\underbrace{\log_2 \frac{N}{N'}}_{\substack{\text{info sur } X \\ \text{sachant } x \in E'}} = \underbrace{I(Q \| \mathcal{P})}_{\substack{\text{info sur } Y \\ \text{sachant } x \in E'}} + \underbrace{\sum_{i=1}^{m} q_i \log_2 \frac{N_i}{N'_i}}_{\substack{\text{info que donne } x \in E' \\ \text{sur } Z \text{ connaissant } Y}}$.

Or $N_i = p_i N$, $N'_i = q_i N'$ donc $\frac{N N'_i}{N' N_i} = \frac{q_i}{p_i}$ d'où :

Définition : gain d'information du remplacement de $\mathcal{P}$ par Q : $\boxed{I(Q \| \mathcal{P}) = \sum_{i=1}^{m} q_i \log_2 \frac{q_i}{p_i}.}$
(Information de Kullback de Q par rapport à $\mathcal{P}$)

Exemples : 1) $\mathcal{P} = (p_1, \dots, p_m)$ distribution quelconque, $\mathcal{E}_n = (\frac{1}{n}, \dots, \frac{1}{n})$ distribution uniforme

$$I(\mathcal{P} \| \mathcal{E}_n) = \sum_{i=1}^{m} p_i \log_2(n p_i) = \log_2 n + \sum_{i=1}^{m} p_i \log_2 p_i = H(\mathcal{E}_n) - H(\mathcal{P}) .$$

$I(\mathcal{P} \| \mathcal{E}_n)$ est le gain d'information obtenu en remplaçant la distribution uniforme par une quelconque $\to$ diminution de l'indétermination.
En général on n'a pas $I(Q \| \mathcal{P}) = H(Q) - H(\mathcal{P})$.

2) $I(X,Y) = I(\mathcal{R} \| \mathcal{P} \otimes Q)$ avec $\mathcal{R} = (r_{ij})_{\substack{1 \le i \le m \\ 1 \le j \le n}}$, $\mathcal{P} \otimes Q = (p_i q_j)_{\substack{1 \le i \le m \\ 1 \le j \le n}}$

Proposition : $I(X,Y) = \sum_{j=1}^{n} q_j \, I(\mathcal{P}_j \| \mathcal{P})$ où $\mathcal{P}_j = (p_{1|j}, \dots, p_{m|j})$

Démonstration : $I(X,Y) = \sum_{i,j} r_{ij} \log_2 \frac{r_{ij}}{p_i q_j} = \sum_{j=1}^{n} q_j \underbrace{\sum_{i=1}^{m} p_{i|j} \log_2 \frac{p_{i|j}}{p_i}}_{I(\mathcal{P}_j \| \mathcal{P})}$. □

Proposition : $I(Q \| \mathcal{P}) \geq 0$

Démonstration : $I(Q \| \mathcal{P}) = \sum_{i=1}^{m} p_i \, \varphi\!\left(\frac{q_i}{p_i}\right) \geq \varphi\!\left(\sum_{i=1}^{m} p_i \frac{q_i}{p_i}\right) = \varphi(1) = 0$ (convexité). □

Proposition : $I(Q_1 \otimes Q_2 \| \mathcal{P}_1 \otimes \mathcal{P}_2) = I(Q_1 \| \mathcal{P}_1) + I(Q_2 \| \mathcal{P}_2)$

Démonstration : $I(Q_1 \otimes Q_2 \| \mathcal{P}_1 \otimes \mathcal{P}_2) = \sum_{i,j} q_{1i} q_{2j} \log_2 \frac{q_{1i} q_{2j}}{p_{1i} p_{2j}} = \sum_j q_{2j} \sum_i q_{1i} \log_2 \frac{q_{1i}}{p_{1i}} + \underbrace{\sum_i q_{1i}}_{1} \sum_j q_{2j} \log_2 \frac{q_{2j}}{p_{2j}}$
$= I(Q_1 \| \mathcal{P}_1) + I(Q_2 \| \mathcal{P}_2)$. □ 1

Expression symétrisée (due à Jeffreys) : $J(\mathcal{P}, Q) = I(\mathcal{P} \| Q) + I(Q \| \mathcal{P}) = \boxed{\sum_{i=1}^{m} (p_i - q_i) \log_2 \frac{p_i}{q_i}}$.

Extension au cas des v.a. continues

Soit X une v.a.c. On pose $X_n = \frac{[nX]}{n}$, X_n v.a.d. à valeurs dans $\frac{1}{n}\mathbb{Z}$. Soit $p_i^{(n)} = \mathbb{P}(X_n = \frac{i}{n}) = \int_{\frac{i}{n}}^{\frac{i+1}{n}} f_X(x)\,dx$

h Poter f
$H(X_n) = \log_2 n + \sum_{i \in \mathbb{Z}} \int_{\frac{i}{n}}^{\frac{i+1}{n}} f(x) \log \frac{1}{f(x)} dx$

$H(X_n) = \sum_{i \in \mathbb{Z}} p_i^{(n)} \log_2 \frac{1}{p_i^{(n)}} = \sum_{i \in \mathbb{Z}} p_i^{(n)} \log_2 \frac{1}{n p_i^{(n)}} + \log_2 n = \log_2 n + \int_{-\infty}^{+\infty} f_n(x) \log_2 \frac{1}{f_n(x)} dx$ avec $f_n(x) = n p_i^{(n)}$ sur $[\frac{i}{n}, \frac{i+1}{n}[$

On a $f_n(x) = \sum_{i \in \mathbb{Z}} n \int_{\frac{i}{n}}^{\frac{i+1}{n}} f_X(y)\,dy \; \mathbb{1}_{[\frac{i}{n}, \frac{i+1}{n}[}(x) \to f_X(x)$ et $H(X_n) - \log_2 n \to \int_{-\infty}^{+\infty} f_X(x) \log_2 \frac{1}{f_X(x)} dx$

D'où la définition $H(X) = \int_{-\infty}^{+\infty} f_X(x) \log_2 \frac{1}{f_X(x)} dx = \mathbb{E}\left[\log_2 \frac{1}{f_X(X)}\right]$

puis $H(X,Y) = \iint f_{(X,Y)}(x,y) \log_2 \frac{1}{f_{(X,Y)}(x,y)} dx\,dy$, $H(X|Y) = \iint f_{(X,Y)}(x,y) \log_2 \frac{1}{f_{(X|Y=y)}(x,y)} dx\,dy = \int H(X|Y=y) f_Y(y)\,dy$

$I(X;Y) = \iint f_{(X,Y)}(x,y) \log_2 \frac{f_X(x) f_Y(y)}{f_{(X,Y)}(x,y)} dx\,dy$.

III bis Information de Kullback (complément à Gain d'information)

Soit X, Y deux v.a. On définit une entropie relative (information de Kullback) de Y par rapport à X

Définition : $I(Y\|X) = \begin{cases} \sum_i q_i \log_2 \frac{q_i}{p_i} & \text{si } \mathbb{P}_X = \{p_i, i \in \mathbb{N}\},\ \mathbb{P}_Y = \{q_i, i\in\mathbb{N}\} \text{ et } p_i = 0 \Rightarrow q_i = 0 \\ \int g(x) \log_2 \frac{g(x)}{f(x)} dx & \text{si } f = f_X \text{ et } g = f_Y \text{ sont telles que } f(x) = 0 \Rightarrow g(x) = 0 \end{cases}$

avec la convention $\log_2 \frac{a}{b} = 0$ si $a = 0$ ou $b = 0$.

Proposition : $\forall X, Y, \quad I(Y\|X) \geqslant 0.$

Démonstration : $\varphi(x) = x \log_2 x$ est convexe donc $I(Y\|X) = \sum_i p_i \varphi\left(\frac{q_i}{p_i}\right) \geqslant \varphi\left(\sum p_i \frac{q_i}{p_i}\right) = \varphi(1) = 0.$ $\square$

Autre dém : $I(Y\|X) = \sum_i \left(\frac{q_i}{p_i}\log_2\frac{q_i}{p_i} + 1 - \frac{q_i}{p_i}\right) p_i = \sum_i \psi\left(\frac{q_i}{p_i}\right) p_i$ avec $\psi(x) = x\log_2 x + (1-x)\log_2 e$, ψ convexe $\geqslant 0$.

Théorème :

1) Soit E l'ensemble des v.a.d. (resp. v.a.c.) à valeurs dans $\{1, \dots, n\}$ (resp. $[a, b]$). Alors $H : E \to \mathbb{R}$ est maximale pour la loi uniforme $\mathcal{U}\{1, \dots, n\}$ (resp. $\mathcal{U}[a,b]$).

2) Soit E_m l'ensemble des v.a.d. (resp. v.a.c.) à valeurs dans $\mathbb{N}^*$ (resp. $\mathbb{R}^+$) d'espérance fixée $\mathbb{E}(X) = m$. Alors $H : E_m \to \mathbb{R}$ est maximale pour la loi géométrique $\mathcal{G}\left(\frac{1}{m}\right)$ (resp. exp. $\mathcal{E}\left(\frac{1}{m}\right)$).

3) Soit $E_{m,\sigma}$ l'ensemble des v.a.c. telles que $E(X) = m$, $var(X) = \sigma^2 > 0$. Alors $H : E_{m,\sigma} \to \mathbb{R}$ est maximale pour la loi normale $\mathcal{N}(m, \sigma^2)$.

Démonstration : Calculons d'abord les entropies correspondantes :

$X : \mathcal{U}\{1, \dots, n\} \qquad H(X) = \log_2 n \qquad\qquad X : \mathcal{U}[a,b] \qquad H(X) = \log_2(b-a)$

$X : \mathcal{G}(p) \qquad H(X) = \frac{1}{p}\left[p \log_2 \frac{1}{p} + (1-p)\log_2 \frac{1}{1-p}\right] \qquad X : \mathcal{E}(\lambda) \qquad H(X) = \log_2 \frac{e}{\lambda}$

$X : \mathcal{N}(m, \sigma^2) \qquad H(X) = \log_2(\sqrt{2\pi e}\,\sigma)$

Soit $Y \in E$.

1) $I(Y\|X) = \begin{Bmatrix} \sum_{i=1}^{n} q_i \log_2 n \\ \int_a^b g(x) \log_2(b-a)\, dx \end{Bmatrix} - H(Y) = H(X) - H(Y) \geqslant 0 \Rightarrow \forall_E Y : H(Y) \leqslant H(X)$

2) $I(Y\|X) = \begin{Bmatrix} -\sum_{i=1}^{\infty} q_i \log_2(p(1-p)^{i-1}) \\ -\int_0^\infty g(x)\log_2(\lambda e^{-\lambda x})\,dx \end{Bmatrix} - H(Y) = \begin{Bmatrix} -\log_2 p \times \sum_{i=1}^{\infty} q_i - \log_2(1-p) \overbrace{\sum_{i=1}^{\infty}(i-1)q_i}^{\frac{1}{p}-1} \\ -\log_2 \lambda \int_0^\infty g(x)\,dx + \lambda \log_2 e \underbrace{\int_0^\infty x g(x)\,dx}_{1/\lambda} \end{Bmatrix} - H(Y)$

$\begin{pmatrix} p = \frac{1}{m} \\ \text{ou } \lambda = \frac{1}{m} \end{pmatrix}$

$= \begin{Bmatrix} \log_2 \frac{1}{p} + \frac{1-p}{p}\log_2\frac{1}{1-p} \\ \log_2 \frac{e}{\lambda} \end{Bmatrix} - H(Y) = H(X) - H(Y) \geqslant 0 \Rightarrow \forall Y \in E, H(Y) \leqslant H(X)$

3) $I(Y\|X) = -\int_{-\infty}^{+\infty} g(x) \log_2 \frac{e^{-\frac{(x-m)^2}{2\sigma^2}}}{\sqrt{2\pi}\,\sigma}\,dx - H(Y) = \frac{\log_2 e}{2\sigma^2} \underbrace{\int_{-\infty}^{+\infty}(x-m)^2 g(x)\,dx}_{\sigma^2} + \log_2\sqrt{2\pi}\,\sigma \int_{-\infty}^{+\infty} g(x) - H(Y)$

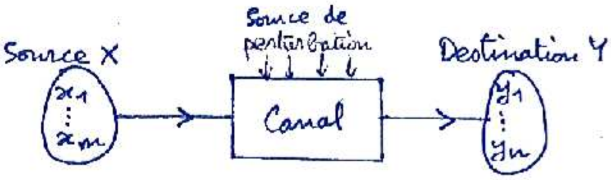
$= \log_2 \sqrt{2\pi e}\,\sigma - H(Y) = H(X) - H(Y) \geqslant 0 \Rightarrow \forall Y \in E,\ H(Y) \leqslant H(X).$ $\square$

IV Application à la transmission de l'information

Une source X émet des signaux $(x_1, \dots, x_m)$ avec probabilités $(p_1, \dots, p_m)$ à travers un canal.

A la sortie de ce canal on réceptionne les signaux Y : $(y_1, \dots, y_n)$ avec probabilités $(q_1, \dots, q_n)$.

Le canal a été source de perturbation.



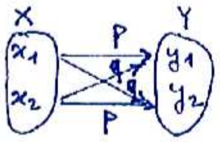
$H(X)$: entropie à l'entrée du canal
$H(Y)$: entropie à la sortie du canal
$H(X,Y)$: entropie entrée - sortie
$H(X|Y)$: Équivoque : mesure l'équivoque sur la source X connaissant la réception Y
$H(Y|X)$: erreur moyenne : mesure l'incertitude de la réception Y connaissant l'émission X.

La matrice $P_{Y|X} = \left(\mathbb{P}(Y=y_j | X=x_i)\right)_{\substack{1\le i\le m \\ 1\le j\le n}}$ s'appelle la matrice de bruit du canal.

$I(X,Y)$: valeur moyenne de l'information mutuelle = information que l'on obtient sur la source par la réception de la sortie = information transmise à travers le canal.

S'il n'y a pas de perturbation dans le canal, $H(X|Y) = H(Y|X) = 0$, $I(X,Y) = H(X) = H(Y)$ $(X=Y)$.
S'il y a un maximum de perturbation, X et Y sont indépendantes, $I(X,Y) = 0$.
Capacité du canal : $C = \max\limits_{X} I(X,Y)$. Plus C est grand, plus l'écart entre $H(Y)$ et $H(Y|X)$ est grand ce qui signifie que X a influence d'autant plus grande si Y.

Ex : Le canal binaire symétrique



Matrice de bruit $P_{Y|X} = \begin{pmatrix} p & q \\ q & p \end{pmatrix}$. $p+q=1$.

$p_i = \mathbb{P}(X=x_i)$, $q_i = \mathbb{P}(Y=y_i)$ $i=1,2$ $p_1+p_2 = q_1+q_2 = 1$.

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

On a $*$ $H(Y|X) = \sum\limits_{i=1}^{2} \mathbb{P}(X=x_i)\, H(Y|X=x_i) = p_1 H(Y|X=x_1) + p_2 H(Y|X=x_2)$

or $H(Y|X=x_1) = H(Y|X=x_2) = p \log_2 \frac{1}{p} + q \log_2 \frac{1}{q}$

donc $H(Y|X) = p \log_2 \frac{1}{p} + q \log_2 \frac{1}{q}$ (indépendant de X)

$*$ $\mathbb{P}_Y = (q_1\ q_2) = (p_1\ p_2)\begin{pmatrix} p & q \\ q & p \end{pmatrix} = (p_1 p + p_2 q, p_1 q + p_2 p) = ((p-q)p_1 + q, -(p-q)p_1 + p)$

$H(Y) = [(p-q)p_1 + q] \log_2 \frac{1}{(p-q)p_1+q} + [p - (p-q)p_1] \log_2 \frac{1}{p-(p-q)p_1} = f(p_1)$ $(= 1$ si $p = q = \frac{1}{2})$

D'où $C = \max\limits_{X} H(Y) - H(Y|X) = \max\limits_{p_1 \in [0,1]} f(p_1) + p \log_2 p + q \log_2 q$

si $p \neq \frac{1}{2}$, $f(p_1) = x \log_2 \frac{1}{x} + (1-x) \log_2 \frac{1}{1-x}$, $x = (p-q)p_1 + q = q_1 \in [0,1]$
$\max\limits_{p_1 \in [0,1]} f(p_1) = \max\limits_{x \in [0,1]} \left[x \log_2 \frac{1}{x} + (1-x) \log_2 \frac{1}{1-x}\right] = 1$ (vrai aussi pour $p = \frac{1}{2}$)

$$\boxed{C = 1 + p \log_2 p + q \log_2 q.}$$

Annexe : autres fonctions d'information

1) Mesure d'ordre α de l'information pour X finie ou dénombrable :

$$H_\alpha(X) = \frac{1}{1-\alpha} \log_2 \sum_{i=1}^{n \text{ ou } +\infty} p_i^\alpha \quad , \quad \alpha \neq 1$$

On a $\lim_{\alpha \to -\infty} H_\alpha(X) = \log_2 \frac{1}{\min_{1\le i\le n} p_i}$, $\lim_{\alpha \to +\infty} H_\alpha(X) = \log_2 \frac{1}{\max_{1\le i\le n} p_i}$, $\lim_{\alpha \to 1} H_\alpha(X) = \sum_{i=1}^n p_i \log_2 \frac{1}{p_i}$.

En effet : * supposons e.g. $p_n = \max_{1\le i\le n} p_i$. $H_\alpha(X) = \frac{1}{1-\alpha}\left[\alpha \log_2 p_n + \log_2\left(1+\sum_{i=1}^{n-1}\left(\frac{p_i}{p_n}\right)^\alpha\right)\right] \xrightarrow[\alpha\to+\infty]{} -\log p_n$

$$* \quad \frac{\log_2 \sum_{i=1}^n p_i^\alpha}{1-\alpha} \underset{\alpha\to 1}{\sim} -\frac{\sum_{i=1}^n p_i^\alpha - 1}{(\alpha-1)\ln 2} \xrightarrow[\alpha\to 1]{} -\frac{d}{d\alpha}\left(\sum_{i=1}^n p_i^\alpha\right)_{\alpha=1} \times \frac{1}{\ln 2} = -\sum_{i=1}^n p_i \log_2 p_i .$$

Si $\mathcal{P}$ est incomplète i.e. $\sum p_i \le 1$ on pose $H_\alpha(\mathcal{P}) = \frac{1}{1-\alpha} \log_2 \frac{\sum p_i^\alpha}{\sum p_i}$.

2) Mesure d'ordre α de l'information pour X absolument continue :

$$H_\alpha(X) = \frac{1}{1-\alpha} \log_2 \int_{-\infty}^{+\infty} f(x)^\alpha dx \quad \text{pour } \alpha \neq 1$$

$$H_1(X) = \int_{-\infty}^{+\infty} f(x) \log_2 \frac{1}{f(x)} dx = \mathbb{E}\left[\log_2 \frac{1}{f(X)}\right] \quad \text{(information de Shannon)}$$

(obtenue à partir de la v.a. discrète $X_N = \frac{[NX]}{N}$ et $\lim_{N\to+\infty}\left[H_\alpha(X_N) - \log_2 N\right] = H_\alpha(X)$.)

e.g. $H_1(X_N) - \log_2 N = -\sum_{i=-\infty}^{\infty} p_{i,N} \log_2 (N p_{i,N}) = -\int_{-\infty}^{\infty} f_N(x) \log_2 f_N(x)\, dx \xrightarrow[N\to\infty]{} \int_{-\infty}^{\infty} f(x) \log_2 \frac{1}{f(x)} dx$ avec $f_N(x) = N p_{i,N}$ sur $\left[\frac{i}{N}, \frac{i+1}{N}\right[$, $f_N(x) \xrightarrow[N\to\infty]{} f(x)$, $p_{i,N} = P(X_N = \frac{i}{N})$

3) Si Q est absolument continue par rapport à $\mathcal{P}$:

$$I_\alpha(Q \| \mathcal{P}) = \frac{1}{\alpha-1} \log_2 \int_\Omega \left(\frac{dQ}{d\mathcal{P}}\right)^\alpha d\mathcal{P} \quad \text{pour } \alpha \neq 1$$

$$I_1(Q \| \mathcal{P}) = \int_\Omega \left(\frac{dQ}{d\mathcal{P}} \log_2 \frac{dQ}{d\mathcal{P}}\right) d\mathcal{P}$$

ex : * $\mathcal{P} = (p_1 \cdots p_n)$, $Q = (q_1 \cdots q_n)$ $\quad$ Q absolument continue par rapport à $\mathcal{P} \Leftrightarrow \forall i, q_i \ge 0$
telle que $\forall i, p_i > 0$

et alors $\frac{dQ}{d\mathcal{P}} = \left(\frac{q_1}{p_1}, \dots, \frac{q_n}{p_n}\right)$, donc $I_1(Q \| \mathcal{P}) = \sum q_i \log_2 \frac{q_i}{p_i}$

$$I_\alpha(Q \| \mathcal{P}) = \frac{1}{1-\alpha} \log_2 \sum \frac{q_i^\alpha}{p_i^{\alpha-1}} \quad \text{pour } \alpha \neq 1$$

* $\frac{d\mathcal{P}}{dx}(x) = p(x)$, $\frac{dQ}{dx}(x) = q(x)$ telle que $\forall x, p(x) > 0$ $\quad$ Q absolument continue par rapport à $\mathcal{P}$ $\Leftrightarrow$ pour presque tout x, $q(x) \ge 0$.

$\frac{dQ}{d\mathcal{P}}(x) = \frac{q(x)}{p(x)}$ et donc $I_1(Q \| \mathcal{P}) = \int_{-\infty}^{+\infty} q(x) \log_2 \frac{p(x)}{q(x)} dx$

$$I_\alpha(Q \| \mathcal{P}) = \frac{1}{1-\alpha} \log_2 \int_{-\infty}^{+\infty} \frac{q(x)^\alpha}{p(x)^{\alpha-1}} dx .$$

4) Complément : Soit X de loi $X(P) = \sum_{i=1}^n p_i \delta_{a_i}$, $p_i = P\{X = a_i\} > 0$, soit $L_i(x) = \prod_{j\neq i} \frac{x-a_j}{a_i-a_j}$ (lagrange)

Posons $\mathcal{P} = \prod_{i=1}^n p_i^{L_i(X)}$. $\mathcal{P}$ est une v.a.r. telle que $P(\mathcal{P} = p_i) = P(X = a_i) = p_i$

$\mathbb{E}(\mathcal{P}) = \sum_{i=1}^n p_i^2 = H_2(X)$ et $H_1(X) = \mathbb{E}\left(\log_2 \frac{1}{\mathcal{P}}\right)$.

Bibliographie

. Rényi : Calcul des probabilités 1966 (appendice)

Spâtaru : Fondements de la théorie de la transmission de l'info.

Karlin & Taylor : a first course in stoch. processes