ÉLÉMENTS DE THÉORIE DE L'INFORMATION POUR LES COMMUNICATIONS.

La théorie de l'information est une discipline qui s'appuie non seulement sur les (télé-) communications, mais aussi sur l'informatique, la statistique, la physique statistique, la théorie de l'inférence.

Il est indéniable que le développement des communications a nourri la théorie de l'information et inversement. Une contribution fondamentale et extrêmement importante est l'article de C.E. Shannon, « A mathematical theory of communications », paru dans le journal de ATT Bell, en 1948. (Cet article a été ensuite réédité sous la forme d'un livre, curieusement cette fois-ci avec un co-auteur).

La presque totalité de ce cours est contenue dans l'article original de Shannon. On présentera d'abord l'entropie (dite de Shannon) comme mesure d'information.

À l'aide des notions d'information, ou d'entropie, on introduira ensuite trois théorèmes de Shannon.

- Le premier théorème indique qu'une source, d'entropie *H*, peut être codée de façon déchiffrable, avec des mots dont la longueur moyenne est bornée par l'entropie de la source. L'entropie *H* représente alors une limite fondamentale sur la longueur moyenne minimale des mots utilisés pour coder la source.

- Le second théorème de Shannon, ou théorème du codage de canal, ou théorème du canal bruité, est tout aussi fondamental, et peut être plus. Ce théorème indique que si l'entropie de la source est inférieure ou égale à la capacité du canal, alors il est possible de

trouver un code tel que la probabilité d'erreur lors de la transmission soit aussi faible que l'on veut.

- Le troisième théorème de Shannon, appelé théorème de la capacité d'information, permet de relier la capacité d'information à la bande du canal, la puissance transmise et la puissance du bruit additif. Il fournit en outre la capacité limite (lorsque la bande tend vers l'infini), et la limite de Shannon.

Avant de poursuivre, un point essentiel, emprunté au Professeur G. Demoment (poly Théorie de l'information, licence EEA, Orsay).

L'expéditeur et le destinataire d'un télégramme ont des attitudes différentes de celles de l'employé de la Poste : pour les deux premiers, le message a une signification. Pour l'employé de la Poste, cette signification est indifférente ; il compte les caractères qu'il transmet en vue de faire payer un service : la transmission d'une « quantité d'information » proportionnelle à la longueur du texte.

La théorie de l'information est un outil, développé pour et par l'ingénieur chargé de concevoir un système de transmission, et les notions probabilistes n'ont de sens clair que de son point de vue.

1. Incertitude, information et entropie

Le problème de référence est le suivant : on considère un alphabet E de N éléments, et on s'intéresse à un message de longueur K composé à l'aide des éléments de E.

L'ensemble des mots de longueur K, que l'on note E_k, est de cardinal N^k. On suppose que les symboles émis par la source sont indépendants ; la source est alors dite « source discrète sans mémoire ».

Chaque mot véhicule une certaine « quantité d'information », que l'on peut raisonnablement relier à la longueur du mot. On peut chercher à quantifier l'information gagnée à la réception d'un mot de E_k (où l'information transmise par l'employé de la Poste). On note I_k cette information. Celle-ci devrait vérifier deux propriétés intuitives :

(a) $I(E_{k+l}) = I(E_k) + I(E_l)$

(b) $I(E_k) \leq I(E_{k+l})$

Précisons la signification de l'exigence (a). On peut décomposer l'ensemble E_{k+l}, qui comprend N^{k+l} éléments, en N^k sous-ensembles de N^l éléments. Dans ces conditions, la sélection d'un élément de E_{k+l} peut être vue comme la sélection d'un des sous-ensembles, qui nécessite une information $I(E_k)$, suivie par la sélection d'un élément du sous-ensemble, avec l'information $I(E_l)$. L'information apportée par l'élément est alors $I(E_{k+l})=I(E_k)+I(E_l)$.

La seconde propriété indique que plus l'ensemble est grand, plus la sélection d'un élément nécessite d'information.

Ces deux propriétés suffisent à déterminer la forme prise par la mesure d'information :

$$I(E_k) = \alpha K.\log(N),$$

où N est le nombre de lettres dans l'alphabet, K la longueur du message, et α une constante arbitraire qui fixe la base du logarithme.

En communications, on s'intéresse souvent à des séquences binaires, et on impose la normalisation $I(E_2) = 1$, ce qui conduit à :

$$I(E_k) = K\log_2(N).$$

Cette dernière relation est la formule de HARTLEY, proposée dès 1928.

Le raisonnement précédent s'appuie implicitement sur une hypothèse d'équiprobabilité des mots de même longueur. Ceci ne sera évidemment pas souvent le cas, en fait dès que les lettres de l'alphabet ne seront plus équiprobables.

Considérons un ensemble de N mots $M_1, ..., M_N$, de probabilité d'occurrence $P_1, ..., P_N$, avec $\sum_{i=1}^{N} P_i = 1$.

On peut de la même façon que précédemment chercher à définir l'information portée par la connaissance d'un mot.

Notons $I(M_i)$ l'information relative à un mot de probabilité P_i. Deux exigences suffisent encore à déterminer la forme de cette information.

(a) $I(M_{kl}) = I(M_k) + I(M_l)$

(b) $I(M_i)$ est une fonction décroissante de la probabilité P_i.

La première propriété illustre le fait que l'information apportée par un mot M_{kl}, composé de deux mots M_k et M_l, supposés indépendants, est égale à la somme des informations apportées par ses parties. La seconde exigence indique que l'information apportée par un mot, ou l'information nécessaire pour le caractériser est d'autant plus importante que sa probabilité est faible.

Ces deux exigences, auxquelles on ajoute une exigence de continuité, conduisent à :

$$I(M_i) = -\alpha \log(P_i),$$

où la constante α est à nouveau arbitraire. Comme précédemment, on utilisera souvent le logarithme de base 2.

Remarques sur les unités :

Lorsqu'on choisit le logarithme népérien, l'unité d'information est le nat (*natural unit*). Lorsqu'on choisit le logarithme en base 10, l'unité est le dit (*decimal unit*), ou Hartley. Lorsqu'on choisit le logarithme en base 2, l'unité est le bit (*binary unit*), ou Shannon, **qu'il ne faut pas confondre** avec le bit de l'informatique (*binary digit*), et qui est simplement un chiffre binaire.

L'information établie ci-dessus n'entre pas en contradiction avec la formule de Hartley. En effet, dans l'ensemble E_k, où l'on considère tous les mots équiprobables avec la probabilité N^{-k}, on obtient

$$I(M_i) = k \log(N).$$

On s'intéresse également à l'informatique apportée en moyenne par ces mots $\{M_1, ..., M_N\}$. Cette information caractérisera par exemple l'information moyenne émise par une source. Cette information ne dépend que de la loi P. On la note $H(P)$, ou $H(P1, ..., PN)$, ou encore $H(x)$, où x représente une variable prenant ces valeurs dans $\{M_1,..., M_N\}$, avec la loi P.

$$\mathrm{H(X)} = \mathrm{E_p}\left[\mathrm{I(X)}\right] = -\sum_{i=1}^{N} P_i \log P_i$$

Cette relation a été établie par Shannon (1948), et indépendamment par Wiener (1948, dans son livre *Cybernetics*).

Le parallèle avec l'entropie de la thermodynamique et de la physique statistique n'est pas fortuit. Shannon a reconnu avoir été guidé par les résultats de Boltzmann, et l'anecdote conte que c'est von Neumann qui a conseillé à Shannon de baptiser « entropie » sa mesure d'information, en raison de la similitude des expressions, et parce que « personne ne sait vraiment ce qu'est l'entropie ».

Les relations entre l'entropie de la physique et l'entropie de la théorie de l'information ont été discutées par exemple par Brillouin (*Science and Information Theory*, Academic Press, 1956).

L'entropie possède plusieurs propriétés naturelles que nous donnons ci-après :

(i) l'entropie est maximale lorsque les événements sont équiprobables, $P_i = \frac{1}{N}$ et vaut alors $\log(N)$; c'est dans cette configuration que le système est le moins bien défini et que les messages sont les plus « surprenants » et apportent le plus d'information.

(ii) l'information est minimale lorsque l'un des événements est certain : le système est parfaitement connu, et aucun apport d'information n'est possible.

(iii) pour N événements équiprobables, l'entropie croît avec N.

(iv) l'entropie est une fonction positive, continue et symétrique en ses arguments.

(v) l'information acquise en deux étapes s'ajoute :

$$H(P_1,...,P_N) = H(Q_1,Q_2) + Q_1 H\left(\frac{P_1}{Q_1},...,\frac{P_M}{Q_1}\right) + Q_2 H\left(\frac{P_{M+1}}{Q_2},...,\frac{P_N}{Q_2}\right)$$

où $Q_1 = \sum_{i=1}^{N} P_i$, $Q_2 = \sum_{i=M+1}^{N} P_i$ sont les probabilités de deux groupes d'événements distincts 1, ..., M et M+1, ..., N. L'information moyenne de l'événement total est ainsi égale à la somme des informations moyennes apportées par les deux événements de la partition, pondérée par leurs probabilités d'apparition, et de l'information de partage.

Preuve de (i) :

On établit d'abord une propriété très importante dans ce cours, et qui est appelée dans beaucoup d'ouvrages le *lemme fondamental*.

On utilise la propriété $\ln x \le x - 1$, avec égalité pour $x = 1$.

On considère deux distributions de probabilité $P = \{P_1,...,P_N\}$, et $Q = \{Q_1,...,Q_N\}$. On considère ensuite $D\left(P\middle\|Q\right) = -\sum_{k=1}^{N} P_i \log_2 \frac{Q_i}{P_i} = -\frac{1}{\ln(2)} \sum_{k=1}^{N} P_i \ln \frac{Q_i}{P_i}$

En utilisant ln(x) ≤x-1, on a alors

$$D\left(P\middle\|Q\right) \ge -\frac{1}{\ln(2)} \sum_{i=1}^{N} P_i \left(\frac{Q_i}{P_i} - 1\right)$$

$$D\left(P\middle\|Q\right) \ge -\frac{1}{\ln(2)} \sum Q_i - P_i = 0,$$

soit $D(P\|Q) = \sum_{k=1}^{N} P_i \log_2 \frac{P_i}{Q_i} \geq 0,$

une inégalité fondamentale.

En prenant enfin $Q_i = \frac{1}{N} \forall i$, il reste

$$\sum P_i \log_2 P_i \geq -\log_2(N),$$

et enfin $$H(P) = -\sum P_i \log_2 P_i \leq \log_2(N)$$

Autre démonstration :

On cherche à maximiser l'entropie

$$H(P) = -\sum P_i \log_2 P_i ,$$

sur l'ensemble des distributions $\{P_1, \ldots, P_N\}$ telles que $\sum P_i = 1.$

Ceci est un problème de maximisation sous contrainte, pour lequel on fait appel au Lagrangien :

$$L(P, \lambda) = -\sum P_i \log P_i + \lambda\left(\sum P_i - 1\right)$$

$$\frac{\partial L(P,\lambda)}{\partial P_i} = 0 \Rightarrow -\log P_i - 1 + \lambda = 0,$$

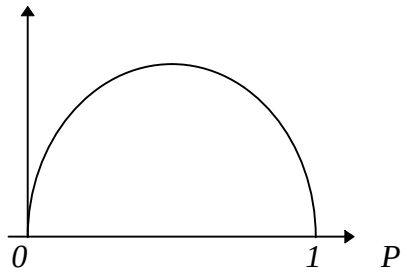
soit $P_i = e^{-\lambda+1} \forall_i$, donc tous les P_i sont égaux et équiprobables, et $P_i = \frac{1}{N}$.

On en déduit $H_{\max}(P) = \log(N)$.

Exemple : loi de BERNOULLI.

On considère l'ensemble $\{a_1, a_2\}$, muni des probabilités *p* et *(1-p)*.

$$H(X) = -\sum_{i=1}^{2} P_i \log P_i = -p \log p - (1-p)\log(1-p)$$



- *p=0 ;* 1-*p*=1 → événement certain, $H(X) = 0,$
- *p*=1, événement certain, $H(X) = 0,$
- $p = \frac{1}{2}$, équiprobabilité, *H(X)* maximale.

2. Codage de source

Un problème très important en communications est la représentation efficace des données générées par une source. Ce processus de représentation est le codage de source. Pour être efficace, l'encodeur doit s'appuyer sur les caractéristiques probabilistes de la source. Les mots-codes les plus courts seront par exemple affectés aux messages les plus fréquents, i.e. de plus forte probabilité.

Ceci est le cas du code Morse. En code Morse, les lettres et les chiffres sont codés en succession de marque « . » et d'espaces « _ ». La lettre la plus fréquente, le E, est ainsi codé « . » alors que la lettre la plus rare, Q (en anglais), est codée « _ _ . _ _ ».

On s'intéresse à encoder des messages M_K en une séquence b_K, de longueur l_K. La probabilité du message M_K est désignée par P_K.

La longueur moyenne d'un mot-code vaut alors

$$\overline{L} = \sum_{K=1}^{N} P_K l_K .$$

Cette longueur moyenne représente alors le nombre moyen de bits (binary digits) par symbole source, utilisé par le processus de codage (si l'alphabet est binaire).

Les mots codes $\{C_1,...,C_N\}$ ont même probabilité que les messages $\{M_1,...,M_N\}$. On désigne par $\{Q_1,...,Q_L\}$ les probabilités associées à chacune des lettres de l'alphabet $\{a_1,...,a_L\}$. L'entropie de la source, qui est aussi l'entropie du code, vaut alors

$$H(S) = H(C) = -\sum_{i=1}^{L} P_i \log P_i .$$

L'entropie de l'alphabet vaut quant-à-elle

$$H(A) = -\sum_{i=1}^{L} Q_i \log Q_i$$

L'information moyenne par mot code est aussi donnée par le produit entre le nombre moyen de lettres dans le mot et l'information moyenne des lettres :

$$H(C) = \overline{L}.H(A)$$

Or, $H(A)$ est borné par $\log(L)$ (toutes les lettres équiprobables). On en déduit donc

$$H(C) \le \overline{L}.\log(L),$$

soit

$$\boxed{\overline{L} \ge \frac{H(C)}{\log(L)}}.$$

L'entropie de la source impose donc une limite fondamentale à la longueur moyenne d'un mot code utilisé pour représenter les symboles émis par cette source.
Ce résultat, le *théorème du codage de source*, indique qu'il est possible de représenter les symboles émis à l'aide de mots, dont la longueur moyenne la plus faible est bornée par l'entropie de la source. Tout code dont la longueur moyenne serait plus faible ne pourrait représenter, sans erreur de décodage, les différents symboles associés à cette source.

En général, l'alphabet de la source et l'alphabet du canal sont différents. Un des premiers buts du codage est de passer de l'un à l'autre. Un autre point est que le code « naturel » associé à une source est souvent très redondant. L'objectif du codage de source est alors d'éliminer cette redondance, pour aboutir à un code dont la longueur moyenne est la plus proche possible du rapport entre l'entropie de la source et l'entropie maximale de l'alphabet.

Exemples : codes irréductibles

Les codes irréductibles sont des codes déchiffrables (ou à décodage unique), qui peuvent être décodés sans utiliser les mots suivants ou précédents. Ce sont des codes instantanés. Ils sont construits en utilisant la condition du préfixe.
Condition du préfixe : il n'existe aucun mot-code qui soit le commencement d'un autre mot-code.
Exemple :

M_i	P_i	*I*	*II*	*III*
M_0	$1/2$	0	0	0
M_1	$1/4$	1	10	01
M_2	$1/8$	00	110	011
M_3	$1/8$	11	111	0111

Le code n°*II* est ici un code à préfixe.

Construction d'un code irréductible (binaire)

- On divise $M = \{M_1,..., M_K, M_{K+1},..., M_N\}$ en deux sous ensembles $M_1 = \{M_1,..., M_K\}$ et $M_2 = \{M_{K+1},..., M_N\}$ et on attribue comme première lettre « 0 » à tous les messages de M_1, et comme première lettre « 1 » à tous les messages de M_2.
- On réitère cette opération en divisant M_1 en deux sous ensembles M_{11} et M_{12}, puis M_2 en M_{21} et M_{22}.

- On continue jusqu'à ce qu'il ne reste plus qu'un élément dans chaque sous ensemble.

Le code obtenu ainsi est irréductible. L'arbitraire de la construction du code est contenu dans le choix des divisions en sous ensembles.

Codage de Huffman :

Le codage de Huffman est un code à préfixe, qui utilise les statistiques de la source, l'idée de base étant d'obtenir des ensembles de même rang $\{M_1\ M_2\}$, $\{M_{11}\ M_{12}\}$, de probabilités les plus proches possibles.

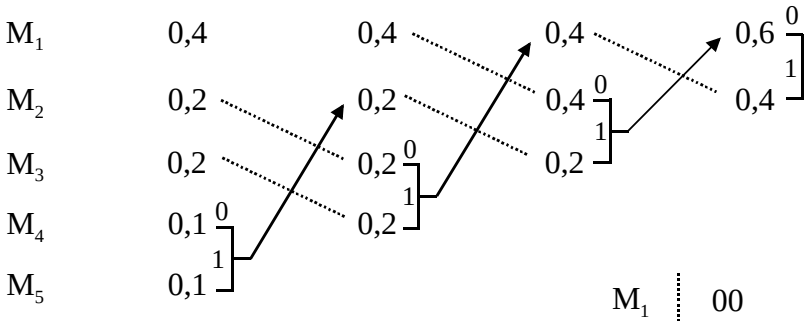
L'algorithme est le suivant :

- Les symboles sont classés par ordre de probabilité décroissante. On assigne alors aux deux symboles de probabilité la plus faible les lettres 0 et 1.
- Ces deux symboles sont combinés en un nouveau symbole, fictif, dont la probabilité est la somme des probabilités des symboles élémentaires. La liste des symboles est alors mise à jour. Cette liste est toujours classée, et comporte un élément de moins.
- La procédure est réitérée, jusqu'à ce que la liste finale ne contienne plus que deux symboles, auxquels on assigne les lettres « 0 » et « 1 ».

Le code pour chaque symbole initial est alors établi en repartant « à l'envers » et en notant la suite de 0 et de 1.

Exemple :

Symbole Probabilité



soit :

M_1	00
M_2	10
M_3	11
M_4	010
M_5	011

Il est clair que ce codage n'est pas unique. En effet, il y a un arbitraire dans l'affectation des 0 et des 1 à chaque étape, il y a un arbitraire dans le classement des symboles lorsque plusieurs possèdent la même probabilité, et un arbitraire dans le placement du nouveau

symbole lorsque d'autres ont la même probabilité. Deux stratégies peuvent alors être employées : placer le nouveau symbole aussi haut ou aussi bas que possible.

Entropie de la source :

$$H(S) = -\sum_{i=1}^{5} P_i \log_2 P_i = 2.12193 \text{ bits}$$

Longueur moyenne du code :

$$\overline{L} = \sum_{i=1}^{5} P_i l_i = 2.2$$

On retrouve ici la relation $\overline{L} \geq H(S)$ (théorème du codage de source), qui indique que le code de Huffman permet ici de s'approcher à 0,078 bits de la longueur moyenne limite pour cette source.

3. Entropies et canaux de communication

Nous avons discuté précédemment de la caractérisation des sources et du problème de codage de source. Nous en arrivons maintenant au second aspect de la communication, à savoir la transmission du message sur le canal de communication, avec la caractérisation du canal, et la possibilité de transmettre sans erreur un message sur un canal imparfait. Ceci est l'objectif du codage de source.

3.1. Canal discret sans mémoire

Un canal discret sans mémoire est caractérisé par un ensemble discret de messages source - un alphabet d'entrée, et par un ensemble discret de messages de sortie :

$$X = \{x_1, \ldots, x_I\}$$
$$Y = \{y_1, \ldots, y_J\}$$

On dispose en outre d'un ensemble de probabilités de transition,

$$p(y_i | x_i) = P(Y = y_i | X = x_i) \qquad \forall i, j$$

La transmission des messages de X le long du canal peut alors être décrit par la distribution de probabilité conjointe

$$p(x_i, y_i) = p(y_i | x_i) p(x_i)$$

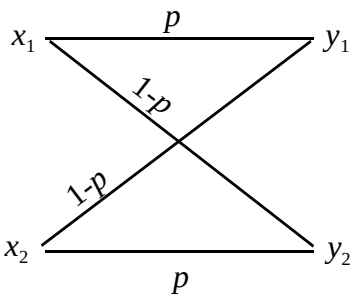
La distribution de probabilité de la sortie peut être obtenue en marginalisant la distribution conjointe :

$$p(y_j) = \sum_{i=1}^{I} p(x_i, y_j) = \sum_{i=1}^{I} p(y_j|x_i)p(x_i)$$

Une représentation commode d'un canal discret sans mémoire est l'utilisation d'une matrice de transition :

$$\begin{bmatrix} p(y_1) \\ - \\ - \\ - \\ p(y_J) \end{bmatrix} = \begin{bmatrix} p(y_1|x_1) & - & - & - & - & - & p(y_1|x_I) \\ p(y_2|x_1) & & & & & & \\ & - & & & & & \\ & - & & & & & \\ p(y_J|x_1) & - & - & - & - & - & p(y_J|x_I) \end{bmatrix} \begin{bmatrix} p(x_1) \\ - \\ - \\ - \\ p(x_I) \end{bmatrix}$$

Exemple : canal binaire symétrique (CBS)



$$\begin{bmatrix} p(y_1) \\ p(y_2) \end{bmatrix} = \begin{bmatrix} p & 1-p \\ 1-p & p \end{bmatrix} \begin{bmatrix} p(x_1) \\ p(x_2) \end{bmatrix}$$

3.2. Entropies pour un canal discret.

On définit aisément, pour un canal discret, trois entropies :

- L'entropie de source $H(X) = -\sum_{i=1}^{I} p_i \log p_i$
- L'entropie de sortie $H(Y) = -\sum_{j=1}^{J} p_j \log p_j$
- L'entropie conjointe entrée-sortie $H(X,Y) = -\sum_{i} \sum_{j} p_{ij} \log p_{ij}$

où on a noté $p_i = p(x_i) = P(X = x_i)$

$p_j = p(y_j) = P(Y = y_j)$

$p_{ij} = p(x_i, y_j) = P(X = x_i, Y = y_j)$

3.3. Relations entre les entropies.

a) lorsque X et Y sont indépendantes, i.e. $P(X,Y) = P(X)P(Y)$, alors

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

En communications, l'indépendance entre X et Y signifierait que le bruit sur le canal est tellement important qu'il supprime toute liaison entre Y et X.

b) entropie conditionnelle

$$H(X,Y) = -\sum_i \sum_j P(X,Y) \log_2 P\left(X = x_i, Y = y_j\right)$$

or $$P\left(X = x_i, Y = y_j\right) = P\left(Y = y_j \middle| X = x_i\right) P\left(X = x_i\right)_j$$

alors

$$H(X,Y) = -\sum_i \sum_j P\left(Y = y_j, \middle| X = x_i\right) P\left(X = x_i\right) \left[\log_2 P\left(X = x_i\right) + \log_2 P\left(Y = y_j \middle| X = x_i\right)\right]$$

soit $$H(X,Y) = H(X) - \sum_i P\left(X = x_i\right) H\left(Y \middle| X = x_i\right)$$

On pose par définition

$$H(Y|X) \overset{def}{=} -\sum_i P\left(X = x_i\right) H\left(Y \middle| X = x_i\right)$$

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

Pour des raisons de symétrie, on a de la même façon :

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

$H(X|Y)$ représente une incertitude moyenne sur l'entrée lorsque la sortie est connue. C'est l'information qui serait encore nécessaire pour caractériser X alors que Y est connue. On l'appelle *l'équivoque*.

$H(Y|X)$ représente l'incertitude moyenne sur la sortie lorsque l'entrée X est connue. On l'appelle parfois erreur moyenne.

3.4. Information mutuelle

En sommant les deux relations précédentes, on obtient l'inégalité

$$2H(X,Y) \geq H(X) + H(Y).$$

À partir de ces deux relations liant l'entropie conjointe et les entropies conditionnelles, on obtient

$$H(Y) - H(Y|X) = H(X) - H(X|Y)$$

Chacun des deux membres de cette égalité représente une diminution de l'information apportée par Y (respectivement X) lorsque X (respectivement Y) est connu.

On appelle ces différences information relative donnée par Y sur X (par X sur Y), ou plus simplement, l'information mutuelle, $I(X,Y)$. C'est l'information transmise par le canal.

- Dans le cas sans bruit, $H(X|Y) = 0$ et $I(X,Y) = H(X)$,

- dans le cas hyper bruité, $H(X|Y) = H(X)$ et l'information mutuelle $I(X,Y) = 0$.

Afin d'optimiser la transmission dans le canal, il faudra donc chercher à maximiser l'information échangée, c'est-à-dire l'information mutuelle. Lorsque la source est fixée, maximiser l'information mutuelle revient à minimiser l'équivoque *H(X|Y)*, c'est-à-dire à minimiser l'incertitude sur *X* lorsque la sortie *Y* est connue. D'un autre coté, lorsque la source est « libre », maximiser l'information mutuelle, c'est aussi rechercher la source qui rende l'information émise en moyenne *H(X)* maximale.

On voit facilement que l'information mutuelle s'exprime également par

$$I_M(X,Y) = H(X) - H(X|Y) = H(X) - \big(H(X,Y) - H(Y)\big),$$

soit

$$I_M(X,Y) = H(X) + H(Y) - H(X,Y).$$

Cette inégalité conduit immédiatement à

$$I_M(X,Y) = \sum_i \sum_j p(x_i, y_j) \log_2 \frac{p(x_i, y_j)}{p(x_i)p(y_j)}$$

Cette dernière expression n'est autre que l'entropie relative ou le gain d'information, entre la distribution conjointe et le produit $P(X)P(Y)$.

C'est en quelque sorte une distance entre la distribution conjointe, et la distribution conjointe qu'on obtiendrait dans le cas indépendant.

3.5. Relations entre les entropies

Rappelons le lemme fondamental : nous avons montré que

$$\sum_j p_j \log \frac{p_j}{q_j} \geq 0.$$

En appliquant ici cette inégalité à l'information mutuelle, on obtient

$$I_M(X, Y \geq 0)$$

et par conséquent
- $H(X) + H(Y) \geq H(X, Y)$
- $H(X) \geq H(X|Y)$
- $H(Y) \geq H(Y|X)$

La première inégalité signifie que l'entropie conjointe est inférieure à la somme des entropies des variables considérées indépendamment, à cause de la liaison entre X et Y.

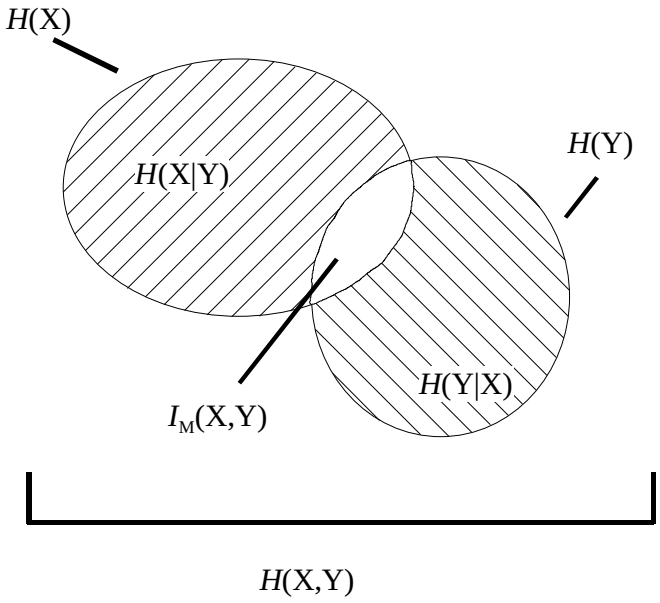
Les deux inégalités suivantes signifient que l'information portée par X (respectivement Y) est inférieure à l'information portée par X, lorsque Y est connu.

Relations entre les entropies : résumé

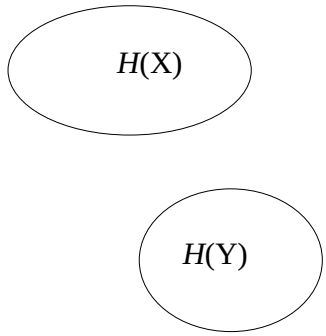
$$2H(X, Y) \geq H(X) + H(Y) \geq H(X, Y) \geq H(Y) \geq H(Y|X)$$

$$2H(X, Y) \geq H(X) + H(Y) \geq H(X, Y) \geq H(X) \geq H(X|Y)$$

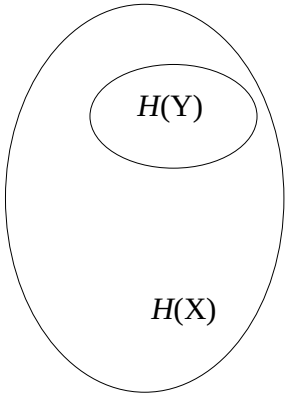
Représentation graphique :



Les différentes entropies représentent les surfaces indiquées



Cas indépendant



Liaison Y=*f*(X)

3.6. Capacité d'un canal

Nous avons vu que l'optimisation de la transmission dans le canal passait par la maximisation de l'information échangée ; ceci débouche sur la notion de capacité du canal. La capacité d'un canal est définie comme la valeur maximale de l'information mutuelle entrée-sortie.

$$C \overset{def}{=} \max I(X,Y) = \max H(X) - H(X|Y) = \max H(Y) - H(Y|X)$$

L'information mutuelle dépend à la fois des probabilités de transition et de la distribution de la source.

Lorsque le canal, c'est-à-dire les probabilités de transition sont fixées, on ne peut plus qu'agir sur la distribution de probabilité de la source pour maximiser l'information mutuelle. C'est aussi l'objet du codage de source.

Exemple : le canal binaire symétrique

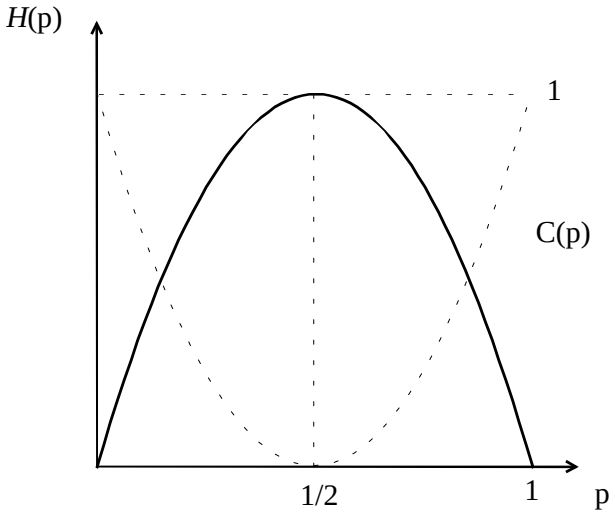
$$H(Y|X) = -p\log p - (1-p)\log(1-p)$$
$$I(X,Y) = H(Y) - H(Y|X)$$

$H(Y)$ est max si $p(y_1) = p(y_2) = 1/2$

En raison de la symétrie du canal, il faut que $p(x_1) = p(x_2) = 1/2$, et $H(Y) = 1$.

La capacité du canal vaut alors

$$C = 1 + p\log p + (1-p)\log(1-p).$$



3.7. Le théorème du canal bruyant.

On définit le taux moyen d'information par

$$R = H(X)/T_S \text{ en bits/s (débit de la source)}$$

(on suppose que la source émet un symbole toutes les T_S secondes).

On suppose que le canal peut être utilisé toutes les T_C secondes. On définit alors la capacité par unité de temps par C/T_C (bits/s).

Le théorème est en deux parties :

(i) si $\dfrac{H(X)}{T_S} \le \dfrac{C}{T_C}$,

alors il existe un code tel que la sortie de la source puisse être transmise sur le canal et reconstruite avec une probabilité d'erreur arbitrairement faible. Le paramètre C/T_C est appelé le *taux critique*. Le débit vaut au maximum le taux critique.

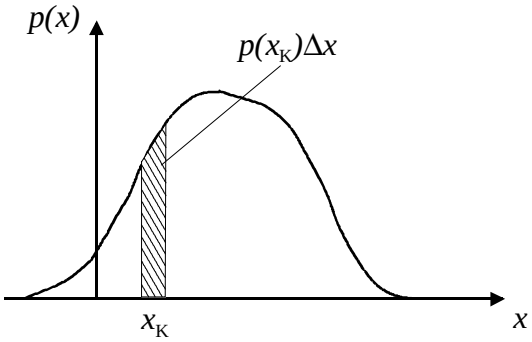
(ii) au contraire, si $\dfrac{H(X)}{T_S} > \dfrac{C}{T_C}$,

il n'est pas possible, en moyenne, de transmettre d'information sans erreur.

4. Cas continu

On s'intéresse maintenant à des variables à valeurs réelles, et non plus à valeurs discrètes. Il s'agit souvent des signaux présents en pratique à l'entrée et à la sortie du canal. De plus, le bruit additif est souvent continu et fait perdre le caractère éventuellement discret des signaux auxquels nous nous intéressons.

Il est tentant de redéfinir les entropies conjointes, conditionnelles pour des variables réelles en remplaçant les sommes discrètes par des intégrales. Une telle manipulation est incorrecte. En effet, en découpant de plus en plus finement un intervalle, l'entropie discrète diverge :



Si on pose $x_K = K\Delta x$, avec $K = -\infty,\ldots+\infty$

et $\Delta x \to 0$,

$$H(X) = \lim_{\Delta x \to 0} - \sum p(x_K)\Delta x \quad \log p(x_K)\Delta x$$

où $p(x)$ est la densité de probabilité de X

$$= \lim_{\Delta x \to O} - \sum p(x_K)\Delta x \mid \log p(x_K) \quad - \sum p(x_K) \quad - \sum p(x_K)\Delta x \quad \log \Delta x$$

$$= -\int_{-\infty}^{+\infty} p(x) \log p(x) dx \quad - \lim_{\Delta x \to 0} \quad \log(\Delta x) \int_{-\infty}^{+\infty} p(x) dx$$

$$H(X) = h_C(X) - \lim_{\Delta x \to 0} \log(\Delta x).$$

Il est par contre une quantité qui converge sans problème, à savoir l'information mutuelle. On retrouve donc la notion d'information mutuelle et de capacité. L'inégalité fondamentale s'étend également sans (trop) de difficulté (on peut être amené à introduire des éléments de théorie de la mesure).

$$I(X,Y) = \int_{-\infty}^{+\infty} p(x,y) \quad \log \frac{p(x,y)}{p(x)p(y)} \quad dxdy$$

On a ici défini la densité conjointe par rapport à la mesure de Lebesgue.

Exemple : distribution uniforme

On considère une variable aléatoire répartie de façon uniforme sur l'intervalle $[0,a]$

$$p(X)=\frac{1}{a} \qquad \text{si } x\in[0,a]$$

$$p(X)=0 \qquad \text{sinon.}$$

L'entropie de la loi uniforme vaut alors

$$\begin{aligned} H(X) &= \int_{-\infty}^{+\infty}\frac{1}{a}\prod_{a}^{0}\log\frac{1}{a}dx \\ &= \int_{0}^{a}\frac{1}{a}\log\frac{1}{a}dx=\log a. \end{aligned}$$

Pour $a\in[0,1]$, l'entropie est négative. On rappelle que le passage à la limite du cas discret au cas continu fait apparaître quelques problèmes...

exercice : montrer que si une variable aléatoire est contrainte à appartenir à $[-a,a]$, alors la loi à maximum d'entropie est la loi uniforme sur $[-a,a]$.

solution :

On écrit le Lagrangien,

$$L(p,\lambda)=-\int_{-a}^{+a}p(x)\log p(x)dx+\lambda\left(\int_{-a}^{+a}p(x)dx-1\right),$$

en dérivant sous l'intégrale (osé), on trouve $p(x)=\exp(1-\lambda)$.

La condition de normalisation fournit ensuite $\lambda=1+\log(2a)$. $p(x)=\frac{1}{2a}$ sur l'intervalle $[-a,a]$.

4.1. Le cas gaussien.

La distribution gaussienne est très importante, à la fois parce qu'elle permet de calculer, et ensuite parce qu'elle justifie du théorème centrale limite. Elle apparaît en outre comme la distribution à maximum d'entropie sous une contrainte de variance. Elle est très souvent une modélisation commode pour les bruits rencontrés en pratique.

La distribution gaussienne est définie par

$$p(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x-m)^2}{2\sigma^2}\right\},$$

où *m* est la moyenne et σ^2 la variance.

L'entropie attachée à une distribution gaussienne est donnée par

$$-\log_2(e)\int p(x)\log p(x)dx = -\int p(x)\left[-\frac{(x-m)^2}{2\sigma^2}-\log\sqrt{2\pi}\sigma\right]dx$$

$$=\left[\frac{\sigma^2}{2\sigma^2}-\log\sqrt{2\pi}\sigma\right]\log_2 e$$

$$=\left[\frac{1}{2}\log e-\log\sqrt{2\pi}\sigma\right]\log_2 e$$

$$H(X)=\log_2\left(\sqrt{2\pi\ e}\sigma\right)$$

4.2. Maximisation de l'entropie sous contrainte de variance et moyenne.

Le lemme fondamental permet d'écrire,

$$\int_{-\infty}^{+\infty} p_x(x)\log\frac{p(x)}{q(x)}\geq 0$$

dans le cas continu également.

On en déduit que

$$\int_{-\infty}^{+\infty} p(x)\log p(x)dx \geq \int_{-\infty}^{+\infty} p(x)\log q(x)dx$$

soit $$H(P)\leq\int_{-\infty}^{+\infty} p(x)\log q(x)dx$$

On prend $q\approx \mathrm{N}\left(m,\sigma^2\right)$.

On suppose que *p* a mêmes moyennes et variance que *q*.

$$H(P)\leq\int_{-\infty}^{+\infty} p(x).\left[-\frac{(x-m)^2}{2\sigma^2}+\log\sqrt{2\pi}\sigma\right]dx*\log_2(e)$$

$$H(P)\leq\left[\frac{-\sigma^2}{2\sigma^2}+\log\sqrt{2\pi}\sigma\right]\log_2 e$$

soit $$H(P)\leq\log_2\left(\sqrt{2\pi\ e}\sigma\right).$$

On en déduit que parmi toutes les lois de mêmes moyenne et variance, la loi normale est celle qui possède l'entropie maximale.

Exercice : montrer que la loi normale est la loi à entropie maximale sous contrainte de moyenne et variance, en procédant par la technique des Lagrangiens.

4.3. Information mutuelle

L'information mutuelle dans le cas continu possède les mêmes propriétés et interprétations que dans le cas discret.

$$I(X,Y) = I(Y,X)$$
$$I(X,Y) \geq 0$$
$$I(X,Y) = H(X) - H(X|Y)$$
$$I(X,Y) = H(Y) - H(Y|X)$$
$$I(X,Y) = H(X) + H(Y) - H(X,Y)$$

4.4. Le théorème de la capacité d'information

Le théorème de la capacité d'information permet d'utiliser la notion d'information mutuelle dans le cas d'un canal à bande limitée, d'une source de puissance limitée, et d'un bruit additif gaussien.

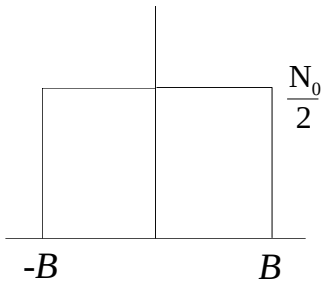
On considère un processus aléatoire $X(t)$ à bande limitée B. On échantillonne ce processus sur une durée T à la fréquence de Shannon $2B$.

On obtient alors $K = 2B\mathrm{T}$ échantillons $x_1, \ldots x_K$. Ces échantillons sont transmis sur un canal, également à bande limitée B, perturbée par un bruit additif gaussien.

$$y(n) = x(n) + \xi(n).$$

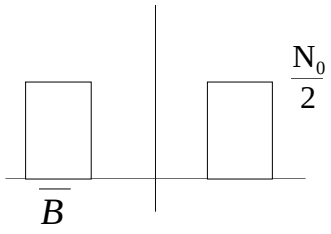
On prend un bruit blanc dans la bande, de densité spectrale $N_0/2$.

$$\sigma^2 = N_0/2 \times 2B = N_0 B$$



La source est de puissance limitée :

$$E\left[X^2(n)\right] = P_X$$



La capacité d'information est le maximum de l'information mutuelle, sous contrainte de puissance pour la source :

$$C = \max\left\{I(X,Y): E\left[X^2\right] = P_X\right\}.$$

L'information mutuelle peut être exprimée comme

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

Dans la mesure où X et ξ sont indépendants, on a

$p(y|x) = p(\xi)$, et

$H(Y|X) = H(\xi)$.

Maximiser l'information mutuelle revient donc à maximiser $H(Y)$;

la puissance en sortie est fixée ; en effet,

$$E\left[Y^2\right] = E\left[(X+\xi)^2\right] = E\left[X^2\right] + E\left[\xi^2\right].$$

La distribution qui maximise l'entropie de la sortie est donc une distribution gaussienne, d'entropie

$$H(Y) = \frac{1}{2}\log_2\left(2\pi\ e\left[-P_X + \sigma^2\right]\right)$$

La sortie étant gaussienne, l'entrée l'est forcément, (loi stable : gauss + gauss = gauss), et son entropie vaut

$$H(X) = \frac{1}{2}\log_2\left(2\pi\ eP_X\right)$$

Au total, on obtient donc l'expression de la capacité d'information :

$$C = H(Y) - H(\xi) = \frac{1}{2}\log_2\left(2\pi\ e\frac{P_X + \sigma^2}{\sigma^2}\right),$$

soit encore $$C = \frac{1}{2}\log_2\left(2\pi\ e\left(1 + \frac{P}{\sigma^2}\right)\right)$$

La capacité par unité de temps (en bits/sec) vaut ici

$$C = B\log_2\left(1+\frac{P}{N_0 B}\right)\text{bits/s}$$

Le théorème de la capacité d'information est donc simplement :
la capacité d'information, pour un canal continu, à bande limitée B, perturbé par un bruit blanc additif gaussien de densité spectrale $N_0/2$, et limité en bande à B, est

$$\boxed{C = B\log_2\left(1+\frac{P}{N_0 B}\right)}\text{bits/s,}$$

où P est la puissance transmise moyenne.
Le théorème indique que l'on peut transmettre sans erreur sur un canal de ce type, pourvu que le débit vérifie $R \le C \quad \left[R = \frac{H(X)}{T_S}\right]$.

4.5. Conséquences du théorème sur la capacité d'information

On s'intéresse au système idéal, où

$R_b = R = C$ (taux critique en bits/s).

On exprime la puissance comme

$P = E_b \cdot C$ E_b : homogène à une énergie/bit.

On a alors

$$\boxed{\frac{C}{B} = \log_2\left(1+\frac{E_b}{N_0}\cdot\frac{C}{B}\right)}.$$

De façon équivalente,

$$\frac{E_b}{N_0} = \frac{2^{C/B}-1}{C/B}$$

- $\frac{E_b}{N_0}$ est un *rapport signal-à-bruit* (par bit)
- $\frac{C}{B}$ est *l'efficacité spectrale*

Pour une bande infinie, $\left(\frac{E_b}{N_0}\right)_\infty = \lim_{B\to\infty}\frac{2^{C/B}-1}{C/B}$

$$= \log(2) = 0{,}693 = -1{,}6013 \text{ dB}.$$