Théorie de l'Information
Notes de Cours
2006-2007

SIC-SICOM

Maria-João Rendas

October 11, 2006

Chapter 1

Mesures d'Information et leurs propriétés

Nous pésentons dans ce Chapitre les mesures fondamentales de la Théorie de l'information (entropie, information mutuelle, entropie relative,...). Le Chapitre présente également des propriétés de ces mesures, ainsi qu'un certain nombre de relations qui seront utilisées par la suite.

Considérez une *variable aléatoire* X qui prend des valeurs dans un ensemble dénombrable $\mathcal{X}$ (X est une variable aléatoire *discrète*), et soit $p_X(x)$ sa *loi de probabilité*:

$$X \sim p_X(x). \tag{1.1}$$

Cette équation doit être lue comme : "la variable aléatoire X suit la loi $p_X(x)$," impliquant

$$\Pr\{X = x\} = p_X(x) \ \ \forall x \in \mathcal{X}.$$

La loi de probabilité $p_X(x)$ vérifie les conditions suivantes:

- $p_X(x) \geq 0, \forall x \in \mathcal{X}$,
- $\sum_{x \in \mathcal{X}} p_X(x) = 1$.

Souvent, par souci de simplicité, nous utiliserons la notation simplifiée $p(x)$ pour représenter la loi de probabilité, la variable aléatoire étant déduite de l'argument de la fonction: $p_X(x) \equiv p(x)$.

Nous allons maintenant introduire une des définitions fondamentales de la Théorie de l'Information: l'*entropie*. L'entropie d'une variable aléatoire est une mesure quantitative de l'incertitude (ou, alternativement, de la quantité d'information) associée aux valeurs prises par la variable aléatoire. Elle a été introduite par Shannon dans les années 50.

Définition 1 Entropie
*Soit $X \sim p(x)$ une variable aléatoire, $X \in \mathcal{X}$, avec $\mathcal{X}$ un ensemble dénombrable. Alors, l'***entropie** *de X, notée $H(X)$ est, par définition*

$$H(X) = \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)}. \tag{1.2}$$

△

Remarque 1 Dans la définition de $H(X)$, nous utilisons la convention $0 \log 0 = 0$.

Reprennons la définition d'une variable aléatoire X définie dans l'espace de probabilité $(\Omega, \mathcal{B}, P)$ [1] et avec des valeurs dans un espace mesurable $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$, comme une application $X : \Omega \to \mathcal{X}$ telle que

$$\text{Si } F \in \mathcal{B}_{\mathcal{X}} \Rightarrow X^{-1}(F) = \{\omega \ : \ X(\omega) \in F\} \in \mathcal{B}.$$

Quand $\mathcal{X}$ est un ensemble fini, nous pouvons associer à X une partition Q_X de Ω:

$$\Omega = \bigcup_{x \in \mathcal{X}} Q_x, \ \ x_1 \neq x_2 \Rightarrow Q_{x_1} \cap Q_{x_2} = \emptyset, \qquad \text{où } \ Q_x = \{\omega \ : \ X(\omega) = x\}$$

et $p_X(x) = P(Q_x)$. Nous utiliserons aussi la notation

$$Q_x = X^{-1}(x) = \{\omega \ : \ X(\omega) = x\}.$$

La Figure 1.1 illustre la définition de cette partition pour un exemple où $\mathcal{X} = \{x, y, z\}$, et donc

$$\Omega = Q_x \cup Q_y \cup Q_z.$$

L'entropie de la variable aléatoire est donc uniquement fonction de la partition $Q_X = \{Q_x\}_{x \in \mathcal{X}}$, et peut être écrite en termes de la mesure de probabilité originale, P:

$$H(X) = \sum_{x \in \mathcal{X}} P(Q_x) \log \frac{1}{P(Q_x)}.$$

On remarquera finalement que l'entropie de la variable aléatoire X dépend uniquement de l'ensemble des valeurs de $p_X(x)$, et pas des valeurs $x \in \mathcal{X}$ prises par la variable elle même (le "code" associé à chaque élément de la partition Q_X). Si nous considérons une nouvelle variable $Y = f(x)$, avec $f(\cdot)$ une fonction inversible, alors $H(Y) = H(X)$, car la partition de Ω déterminée par la variable aléatoire Y sera la

[1] Nous rappellons les entités qui composent un espace de probabilité: Ω est l'espace des *évènements*; $\mathcal{B}$ est une collection de sous-ensembles de Ω telle que : *(i)* $\Omega \in \mathcal{B}$, *(ii)*si $A \in \mathcal{B} \Rightarrow A^c \in \mathcal{B}$ (le complément de A est aussi dans $\mathcal{B}$), et *(iii)*si $A, B \in \mathcal{B} \Rightarrow A \bigcup B \in \mathcal{B}$; P est la *mesure de probabilité*, $P \ : \ \mathcal{B} \to [0, 1]$, telle que : *(i)* $P(\Omega) = 1$, *(ii)* $A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B)$.

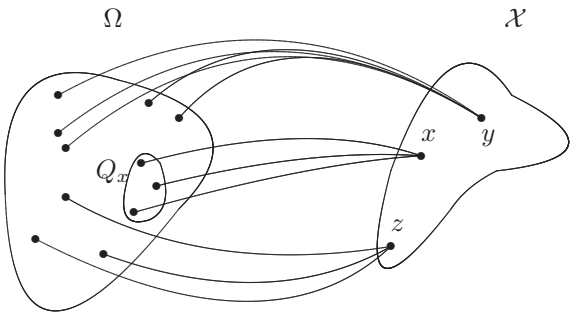

Figure 1.1: Partition de Ω associée à une variable aléatoire dicrète (prenant uniquement 3 valeurs) finie définie sur Ω.

même que celle qui est déterminée par X. Pour cette raison, nous représenterons souvent l'entropie comme une fonction de la loi de probabilité p_X : $H(X) \equiv H(p_X)$ ou encore de la partition Q_X : $H(X) \equiv H(Q_X)$.

Dans la Définition 1, nous utilisons la fonction logarihtme. Le choix de la base du logarithme (qui doit cependant être consistant!) ne modifie pas les propriétés de l'entropie (voir Propriété 2), et il détermine les unités utilisées pour la mesurer. Nous utiliserons la notation $H_a(X)$ quand nous souhaitons expliciter la base a utilisée dans le calcul de l'entropie. Des choix usuels pour a sont :

- $\mathbf{2}$: l'entropie est dans ce cas mesurée en *bits* (la justification pour cette désignation deviendra claire quand nous discuterons les relations entre entropie et codage)
- $\mathbf{e}$: l'entropie est dans ce cas mesurée en *nats* (*natural units*). Ce choix simplifie certains calculs, par exemple dans des problèmes d'optimisation où il faut dériver.
- $\mathbf{10}$: l'entropie est dans ce cas mesurée en *digits*.

La Définition 1, eq.(1.2), peut être écrite de la façon suivante, en reconnaissant la d éfinition de l'opérateur d'espérance (ou moyenne) statistique $E[\cdot]$[2] :

$$H(X) = E_X\left[\log \frac{1}{p_X(x)}\right], \tag{1.3}$$

c'est à dire, comme la moyenne de la variable aléatoire $Z = \log \frac{1}{p_X(x)}$ construite à partir de la variable aléatoire X :

$$Z(\omega) = -\log \Pr\left(\nu \in X^{-1}\left(X(\omega)\right)\right), \forall \omega \in \Omega.$$

[2]Le sous-index dans les notations $E_X[\cdot] \equiv E_{p_X}[\cdot]$ indique par rapport à quelle variable (loi) la moyenne est calculée.

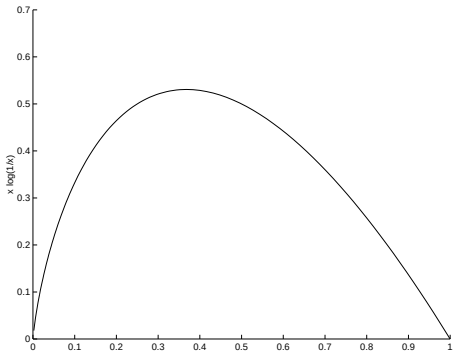

Figure 1.2: graphe de $x \log(1/x)$ sur l'intervalle unitaire.

L'entropie possède les propriétés suivantes :

Propriété 1 *Non-negativité.*

$$H(X) \geq 0.$$

Ceci est facilement déduit du fait que tous les termes dans la somme (1.2) sont non-négatifs (voir Figure 1.2).

Propriété 2 *Changement de base.*

$$H_b(X) = \log_b(a) H_a(X).$$

Cette relation découle immédiatement de la formule pour le changement de base du logarithme:

$$x = a^{\log_a(x)} \Leftrightarrow \log_b(x) = \log_a(x) \log_b(a).$$

Exemple 1 *Entropie d'une variable aléatoire binaire.*
Considérons une variable aléatoire binaire $X \in \{a, b\}$, avec la loi suivante :

$$p_X(a) = q, \qquad p_X(b) = 1 - q,$$

où $q \in [0, 1]$. Son entropie est, par définition

$$H(X) = -q \log q - (1 - q) \log(1 - q). \tag{1.4}$$

Comme nous l'avons affirmé, l'entropie ne dépend pas des valeurs a et b pris par X, mais uniquement de la valeur de q. Pour indiquer cela, nous utiliserons suivent la notation $H(q)$ pour indiquer l'entropie d'une variable aléatoire binaire qui prend un des deux valeurs possibles avec probabilité q. Il est évident que $H(q) = H(1 - q)$, et que donc $H(q)$ est une fonction symmérique autour de $q = 1/2$. La Figure 1.3

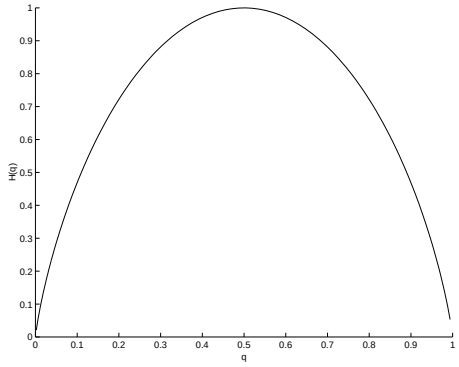

Figure 1.3: Entropie d'une variable aléatoire binaire.

représente les valeurs de l'entropie pour q dans l'intervalle unitaire. Nottons que si $q = 1/2 = 1 - q$, la valeur de l'entropie est maximale et égale à $\log_2 2 = 1$, et que pour $q = 1$ et $q = 0$ (et donc $1 - q = 1$) nous obtenons la valeur minimale zéro: $0 \leq H(q) \leq 1$. Nous verrons plus tard que ce comportement (valeur maximale quand les éléments de $\mathcal{X}$ sont équiprobables, et entropie nulle quand un des évènements est certain) est vérifié pour toutes les variables discrètes dans un alphabet fini (même de dimension supérieure à 2).

△

Exemple 2 *Entropie d'une variable dans un ensemble fini.*
Dans cet exemple nous considérons une variable aléatoire $X \in \mathcal{X}$, $|\mathcal{X}| = 4$,[3] avec loi de probabilité

$$p_X = \begin{pmatrix} 1/2 & 1/4 & 1/8 & 1/8 \end{pmatrix}.$$

Nous pouvons utiliser la Définition 1 pour calculer directement l'entropie $H(X)$:

$$H(X) = \frac{1}{2}\log 2 + \frac{1}{4}\log 4 + 2\frac{1}{8}\log 8 = 1 + \frac{3}{4} = 1.75 < \log_2 4 = 2.$$

Les valeurs p_X sont dans ce cas tous de la forme 2^{-n_i}, avec n_i un nombre naturel, et ce fait nous permet une interprétation de cette variable aléatoire comme composition d'évènements élémentaires (statistiquement indépendants) sur une variable aléatoire binaire $U \in \{0, 1\}$ avec $p_U(0) = p_U(1) = 1/2$. Nous nottons maintenant par $\mathcal{X} = \{a, b, c, d\}$ les différents valeurs pris par X.

1. Générer $u_1 \sim p_U$. Si $u_1 = 0$ prendre $X = a$ **stop**
2. Générer $u_2 \sim p_U$. Si $u_2 = 0$ prendre $X = b$ **stop**
3. Générer $u_3 \sim p_U$. Si $u_3 = 0$ prendre $X = c$ sinon $X = d$ **stop**

[3] La notation $|A|$ représente le cardinal de l'ensemble A.

Le nombre *moyen* de tirages de la variable aléatoire U nécessaires pour obtenir une valeur de X est:

$$\frac{1}{2}(1) + \frac{1}{4}(2) + \frac{1}{4}(3) = 1.75,$$

ce qui est exactement la valeur de l'entropie de X. Nous reviendrons plus tard sur cette interprétation de l'entropie d'une variable aléatoire X comme le nombre moyen de tirages sur une variable aléatoire binaire (uniforme) nécessaires pour simuler les valeurs de X.

Considérons maintenant une autre intreprétation de l'entropie, plus proche de la problématique du codage de source. Considérons le codage (avec un code de longueur variable) des valeurs de X obtenu à partir de la s équence des valeurs u_i qui leur correspondent dans l'algorithme de simulation vu précédemment:

$x \in \mathcal{X}$	code $u_1u_2u_3$
a	0
b	10
c	110
d	111

Nottons que le code obtenu est un code de *préfixe* (aucun mot de code n'est préfixe d'un autre mot de code). Ceci implique que le code est *immédiatement décodable*. Par exemple le décodage de la s équence binaire 0100100110 peut être fait au fur et à mesure que les bits sont examinés:

$$0(\rightarrow a)10(\rightarrow b)0(\rightarrow a)10(\rightarrow b)0(\rightarrow a)0(\rightarrow a)110(\rightarrow c).$$

Nous pouvons également constater (Essayez!) que n'importe quelle séquence de chifres binaires peut être d écodée comme une séquence de symboles dans $\mathcal{X}$ (le code est *complet* : il n'y a pas de mots de code non utilisées).

Soit $l(x), x \in \mathcal{X}$, la longueur du mot de code pour l'évènement x, par exemple $l(b) = 2$. Nous pouvons constater facilement que la longueur moyenne des mot de code est encore égale à 1.75, c'est à dire, à $H(X)$, inférieure au nombre de (2) bits nécessaires pour coder les éléments de $\mathcal{X}$ avec un code de longueur constante. La relation entre entropie et longueur moyenne des mots de code est un des résultats fondamentaux de la Théorie de l'Information, et sera présenté rigoureusement plus tard.

Exemple 3 *Entropie d'une variable uniforme.*
Prennons une variable aléatoire $X \in \mathcal{X}$, $|\mathcal{X}| = m$, avec distribution uniforme: $p(x) = 1/m, \forall x \in \mathcal{X}$, et calculons son entropie:

$$H(X) = \sum_{x \in \mathcal{X}} \frac{1}{m} \log m = \log m.$$

Nous pouvons maintenant constater que la valeur $H(1/2) = 1$ obtenue dans l'exemple 1 est un cas particulier de celui-ci. La propriété suivante montre que $\log m$ est en fait

une borne supérieure.

Propriété 3 *Borne supérieure de l'entropie (alphabet fini).*
Si $X \in \mathcal{X}$, où $|\mathcal{X}| = m$, alors

$$H(X) \leq \log m.$$

△

Cette inégalité peut être obtenue de plusieurs façons. En particulier, elle découle de certaines inégalités fondamentales de la Théorie de l'Information, comme nous le verrons plus tard (Propriété 15, page 23). Nous pouvons l'obtenir directement comme la solution d'un problème d'*optimisation sous contraintes*:

$$\max_{p_X} \quad H(X), \qquad \text{s.c.} \sum_{x\in\mathcal{X}} p_X(x) = 1.$$

Pour résoudre ce problème nous utilisons la méthode des multiplicateurs de Lagrange, et formons la fonctionnelle

$$L = H(X) + \lambda(\sum_{x\in\mathcal{X}} p_X(x) - 1)$$

Si nous égalons à zéro la dérivée par rapport à chaque $p_X(x)$ (nous considérons ici que $\log \equiv \log_e$)

$$\frac{\partial L}{\partial p_X(x)} = -\log p_X(x) - 1 - \lambda = 0,$$

ce qui nous permet de conclure que les valeurs optimaux (qui maximisent $H(X)$ sous la constrainte de somme unitaire) de $p_X(x)$ sont indépendants de x. Comme leur somme doit être égale à 1, nous obtenons $p_X(x) = 1/m$, comme nous voulions démontrer.

La Figure 1.4 illustre la variation de l'entropie dans le simplex probabiliste de dimension $m = 3$.[4] Comme nous pouvons constater, la valeur maximale de $H(X)$ est obtenue pour la distribution uniforme, au centre du simplex: $p_1 = p_2 = p_3 = 1/3$. Nous démontrerons plus tard (Propriété 19, page 25) que H est une *fonction concave* dans le simplex, ce qui est apparent de la Figure 1.4 (considérez les valeurs de la fonction le long d'un segment qui joint deux points sur la même ligne de niveau de $H(X)$).

Définition 2 *Entropie conjointe.*
Soient $X \in \mathcal{X}$ et $Y \in \mathcal{Y}$ deux variables aléatoires avec distribution conjointe $p_{XY}(x, y)$, $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Leur entropie conjointe est

$$H(X, Y) = \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p(x, y) \log \frac{1}{p(x, y)}. \tag{1.5}$$

[4] Les triples (p_1, p_2, p_3), tels que $p_1 + p_2 + p_3 = 1$ et $p_i \in [0, 1], i = 1, 2, 3$.

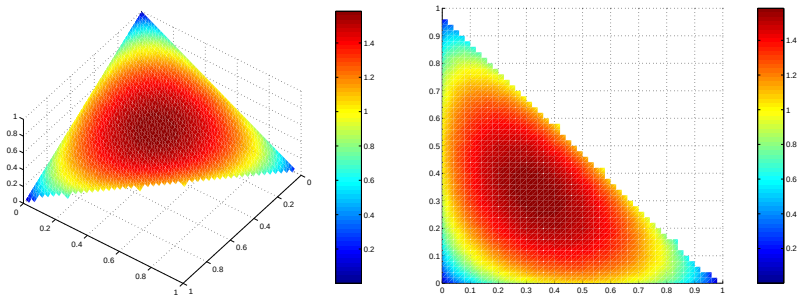

Figure 1.4: Entropie des distributions ternaires (vue tridimensionnelle, à gauche, et représentation sur le simplex, à droite).

Cette définition est une application de la définition originale à l'ensemble de toutes les paires possibles des valeurs de X et Y, le produit $\mathcal{X} \times \mathcal{Y}$.
Nottons que l'entropie conjointe est symmétrique, i.e., $H(X,Y) = H(Y,X)$.

Propriété 4 *Borne inférieure de l'entropie conjointe.*

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

△

Cette inégalité découle directement du fait que la probabilité conjointe de deux évènements est toujours inf érieure ou ègale à la probabilité de chaque évènement : $p(x,y) \leq p(x)$, $\forall x \in \mathcal{X}, y \in \mathcal{Y}$:

$$\begin{aligned} H(X,Y) &= -\sum_{x,y} p(x,y) \log p(x,y) = -\sum_{x,y} p(y|x)p(x) \log p(y|x)p(x) \\ &\geq -\sum_{x,y} p(y|x)p(x) \log p(x) = -\sum_{x} \left(\sum_{y} p(y|x) \right) \log p(x) \\ &= -\sum_{x} p(x) \log p(x) = H(X) \end{aligned}$$

Exemple 4 *Entropie conjointe entre entrée et sortie d'un canal binaire.*
Considérons la transmission d'une source binaire dans un canal avec du bruit, voir Figure 1.5. La source S suit la loi de probabilité suivante:

$$p_s(0) = q = 1/4, \;\; p_s(1) = 1 - q = 3/4.$$

La sortie du canal O peut être en erreur avec probabilité $\varepsilon = 1/8$, et donc

$$\begin{aligned} p_O(0) &= q(1-\varepsilon) + (1-q)\varepsilon = 0.3125 \\ p_O(1) &= (1-q)(1-\varepsilon) + q\varepsilon = 0.6875 = 1 - 0.3125. \end{aligned}$$

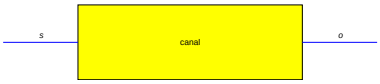

Figure 1.5: Transmission d'une source par un canal de communication.

Nous avons donc les valeurs suivants pour les entropies de la source $H(S)$ et de la sortie du canal $H(O)$:

$$H(S) = H(q) = 0.8113 \text{ bits}, \quad H(O) = H(0.6875) = 0.8960 \text{ bits}.$$

La loi de probabilité conjointe de l'entrée (S) et de la sortie (O) est

$$p_{S,0} = [q(1-\varepsilon) \;\; (1-q)\varepsilon \;\; q\varepsilon \;\; (1-q)(1-\varepsilon)], \tag{1.6}$$

où nous avons ordonnée les quatre évènements possibles de la façon suivante $\{(s = 0, o = 0), (s = 1, o = 0), (s = 0, o = 1), (s = 1, o = 1)\}$. L'entropie de cette loi est

$$H(S,O) = H([0.2188 \;\; 0.0938 \;\; 0.0313 \;\; 0.6563]) = 1.3548 \text{ bits}.$$

La Figure 1.6 illustre la variation de l'entropie conjointe pour les valeurs de q et ε dans l'intervalle unitaire. Nottez que pour $\varepsilon = 0$ (ou $\varepsilon = 1$), la courbe coïncide avec

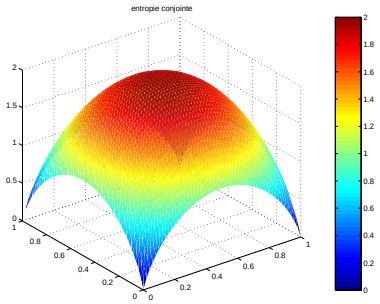

Figure 1.6: Entropie conjointe de l'entrée et de la sortie.

l'entropie de la source pour la valeur de q correspondante. Cette Figure montre donc que l'entropie conjointe est toujours supérieure à l'entropie de la source S, comme l'affirme la Propriété 4.

Définition 3 *Entropie conditionnelle.*
Soient $X \in \mathcal{X}$ et $Y \in \mathcal{Y}$ deux variables aléatoires avec distribution conjointe $p_{XY}(x, y)$, $(x, y) \in \mathcal{X} \times \mathcal{Y}$. L'entropie conditionnelle de Y sachant X est

$$H(Y|X) = E_X [H(Y|x)], \tag{1.7}$$

où

$$H(Y|x) = \sum_{y \in \mathcal{Y}} p_{Y|x}(y|x) \log \frac{1}{p_{Y|x}(y|x)}. \tag{1.8}$$

Cette dernière équation (1.8) est l'entropie de la distribution conditionnelle de Y sachant que $X = x$, pour une valeur particulière de X. Elle dépend donc de la valeur particulière qui a été fixée. L'entropie conditionnelle est obtenue (eq. (1.7)) en considérant la valeur moyenne par rapport à la valeur de X.

Propriété 5 *Non-négativité de l'entropie conditionnelle.*

$$H(Y|X) \geq 0.$$

△

Cette inégalité découle directement du fait que l'entropie de chaque distribution conditionnelle (chaque terme de (1.7)) est non-négative. Nous remarquons que $H(X|Y) = 0$ si et seulement si $H(Y|x) = 0, \forall x \in \mathcal{X}$. Mais l'entropie est nulle uniquement quand toute la probabilité est concentrée dans un seul évènement, et donc, $H(Y|X) = 0 \Leftrightarrow Y = f(X)$.

Exemple 5 *Entropie conditionnelle de l' entrée d'un canal binaire sachant sa sortie.*
Nous considérons encore la transmission d'une source binaire dans un canal avec des erreurs de l'exemple 4, Figure 1.5. Nous avons calculé la loi conjointe de l'entrée et la sortie du canal, voir eq. (1.6). Par application de la loi de Bayes, nous obtenons la loi de l'entrée sachant les valeurs de la sortie:

$$p_{S|O}(s|O = o) = \frac{p(S = s, O = o)}{p_O(o)} = \frac{p(O = o|S = s)p_S(S = s)}{p_O(o)},$$

où le comportement du canal est décrit par les loi conditionnelles suivantes:

$$p_{O|S=0}(o|S = 0) = \begin{cases} 1 - \varepsilon, & o = 0 \\ \varepsilon, & o = 1 \end{cases}$$
$$p_{O|S=1}(o|S = 1) = \begin{cases} \varepsilon, & o = 0 \\ 1 - \varepsilon, & o = 1 \end{cases}$$

Nous obtenons les entropies suivantes :

$$H(S|O = 0) = 0.8813 \text{ et } H(S|O = 1) = 0.5746 \text{ bits.}$$

L'entropie conditionnelle est donc

$$H(S|O) = p_O(0)H(S|O = 0) + p_O(1)H(S|O = 1) = 0.4588,$$

qui est inférieure à l'entropie de la source, $H(S) = 0.8113$. La Figure 1.7 illustre la variation de l'entropie conditionnelle $H(S|O)$ avec ε sur l'intervalle unitaire.

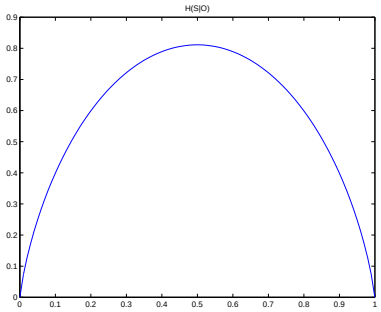

Figure 1.7: Entropie conditionnelle $H(S|0)$ en fonction de la probabilité d'erreur.

Remarquez que pour $\varepsilon = 0$ ou $\varepsilon = 1$, c'est à dire, quand la sortie est une fonction déterministe de l'entrée, cette entropie conditionnelle devient nulle. Sa valeur maximale, égale à l'entropie de la source (vérifiez cette affirmation et interprétez ce fait), est obtenue pour $\varepsilon = 1/2$.

Remarque 2
Comme cas particulier, nous pouvons conclure que

$$H(X|X) = 0.$$

Propriété 6 *Règle de chaîne pour l'entropie conjointe (deux variables).*

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y). \tag{1.9}$$

△

La démonstration de cette relation est immédiate à partir de la définition de l'entropie conjointe, en utilisant $p(x,y) = p(y|x)p(x)$:

$$\begin{aligned} H(X,Y) &= \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} p(y|x)p(x)\frac{1}{p(y|x)p(x)} \\ &= \sum_{x\in\mathcal{X}} p(x)\log\frac{1}{p(x)}\sum_{y\in\mathcal{Y}} p(y|x) + \sum_{x\in\mathcal{X}} p(x)\sum_{y\in\mathcal{Y}} p(y|x)\log\frac{1}{p(y|x)} \\ &= H(X) + H(Y|X). \end{aligned}$$

L'équation alternative en termes de $H(X|Y)$ est obtenue en factorisant la loi conjointe comme $p(x,y) = p(x|y)p(y)$.

Cette dernière propriété nous permet de décomposer en deux pas l'entropie conjointe de deux variable al éatoires : on observe d'abord la valeur de X (ou Y), avec une entropie $H(X)$ (respectivement $H(Y)$). L'observation de Y (X) a alors une incertitude qui est quantifiée par l'entropie conditionnelle $H(Y|X)$ ($H(X|Y)$).

Remarque 3
À partir de cette décomposition nous pouvons obtenir la Propriété 4, qui découle du fait que l'entropie conidtionnelle $H(Y|X)$ est non-négative (Propriété 5).

Exemple 6 L'application de cette formule nous permet de calculer plus facilement l'entropie conditionnelle de l'exemple précédant:

$$H(S|O) = H(S,O) - H(O) = 1.3548 - 0.8960 = 0.4588.$$

La Propriété 6 peut être étendue à un ensemble dénombrable de variables aléatoires :

Propriété 7 *Règle de chaîne pour l'entropie conjointe (n variables).*

$$H(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X^{i-1}). \tag{1.10}$$

où la notation X^i représente l'ensemble $\{X_1, \ldots, X_i\}$, et, par convention, $X^0 = \emptyset$. △

Cette relation découle de la factorisation de la densité conjointe de la façon suivante:

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i x^{i-1}).$$

Avec les définitions d'entropie et d'entropie conditionnelle, nous pouvons maitenant introduire la définition d'*information mututelle*, qui joue un rôle déterminant dans la notion de capacité de canal, comme nous le verrons plus tard.

Définition 4 *Information mutuelle.*
Soient $X \in \mathcal{X}$ et $Y \in \mathcal{Y}$ deux variables conjointement distribuées. L'information mutuelle entre X et Y est, par définition,

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(X) - H(X|Y) \end{aligned} \tag{1.11}$$

Il est facile de démontrer que l'information mutuelle peut encore être écrite comme

Propriété 8

$$I(X;Y) = H(X) + H(Y) - H(X;Y). \tag{1.12}$$

△

Remarque 4 De la Définition 4 et la remarque 2

$$I(X;X) = H(X), \tag{1.13}$$

et donc nous pouvons interpréter l'entropie comme l'information d'une variable sur soit même.

Remarque 5 *Diagrammes de Venn, algèbre d'information*
Les relations (1.9), (1.11) et (1.12) peuvent être résumées par le diagramme de la Figure 1.8, où l'entropie conjointe $H(X,Y)$, qui n'est pas représentée, correspond à l'union des ensembles représentés.

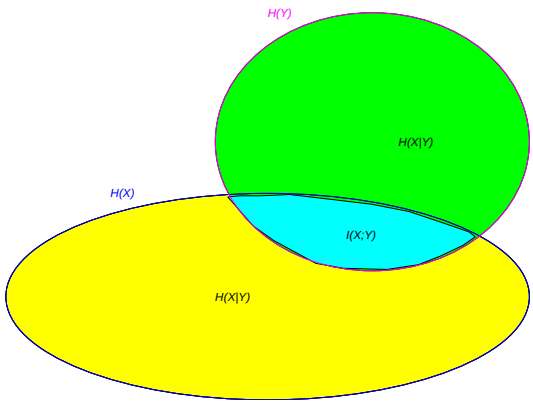

Figure 1.8: Algèbre de l'information. Relations entre entropie conjointe, conditionnelle et information mutuelle.

Exemple 7 Si nous calculons l'information mutuelle pour l'exemple du canal binaire bruité, nous obtenons:

$$I(S,O) = H(S) - H(S|O) = 0.1432.$$

La Figure 1.9 représente l'information mutuelle entre l'entrée et la sortie en fonction de la probabilit é d'erreur du canal. Nous pouvons constater que l'information mutuelle est maximale pour $\varepsilon = 0$ ou $\varepsilon = 1$ et nulle quand la probabilité d'erreur est 1.2.

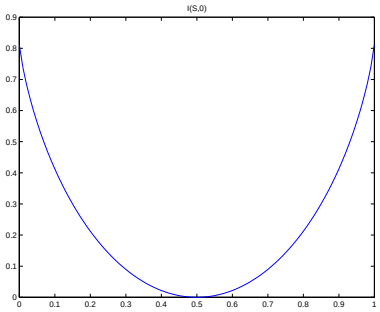

Figure 1.9: Information mutuelle $I(S;0)$ en fonction de la probabilité d'erreur du canal.

Propriété 9 *Borne supérieure de l'information mutuelle.*

$$I(X;Y) \leq \min\left(H(X), H(Y)\right). \tag{1.14}$$

△

Cette inégalité découle de la définition d'information mutuelle, Définition 4, et la non-négativité de l'entropie conditionnelle, Propriété 5. L'intréprétation de la Figure 1.8 confirme cette relation: l'information mutuelle, étant l'intersection des ensembles qui représentent $H(X)$ et $H(Y)$, ne peut pas être supérieure à chacun des ensembles considéré d'une façon séparée.

Propriété 10 *Symmétrie de l'information mutuelle.*

$$I(X;Y) = I(Y;X). \tag{1.15}$$

△

L'équation (1.11), ainsi que la Propriété 8, eq. (1.12), montrent que les rôles de X et Y peuvent être interchangés dans la définition de l'information mutuelle. Cependant, nous le d émontrons directement, ce qui nous permetra d'obtenir un résultat qui nous sera utile par la suite.
La démonstration est faite en écrivant $p(y)$ comme la marginale de la distribution conjointe

$$p(y) = \sum_{x \in \mathcal{X}} p(x,y),$$

dans la définition d'information mututelle:

$$I(X;Y) \quad = \quad \sum_{y} p(y) \log \frac{1}{p(y)} - \sum_{x,y} p(y|x)p(x) \log \frac{1}{p(y|x)}$$

$$
\begin{aligned}
&= \sum_{xy} p(y,x) \log \frac{1}{p(y)} - \sum_{x,y} p(y,x) \log \frac{p(x)}{p(y,x)} \\
&= \sum_{x,y} p(y,x) \log \frac{p(y,x)}{p(x)p(y)} \\
&= E_{X,Y} \left[\log \frac{p(y,x)}{p(x)p(y)} \right] \qquad (1.16)
\end{aligned}
$$

Cette dernière expression montre bien que $I(X;Y)$ est une fonction symmétrique des deux variables al éatoires.

Nous allons maintenant introduire une autre mesure fondamentale de la Théorie de l'Information: l'entropie relative (aussi appellée divergence – ou même "distance" – de Kullback-Leibler, car elle a été introduite par S. Kullback pour des applications de la théorie de l'information à des problèmes de statistique).

Définition 5 *Entropie relative*.
Soient $p(x)$ et $s(x)$ deux lois de probabilité sur le même alphabet dénombrable $\mathcal{X}$. L'entropie relative de p par rapport à s est, par définition :

$$D(p||s) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{s(x)} = E_p \left[\log \frac{p(x)}{s(x)} \right]. \qquad (1.17)$$

Remarque 6
La définition précédante n'a de sens que si $p(x) = 0$ pour tous les x pour lesquels $s(x) = 0$:

$$s(x) = 0 \Rightarrow p(x) = 0,$$

c'est à dire, si la mesure p est *absolument continue* par rapport à s, que nous nottons $s >> p$. Quand ce n'est pas le cas, nous définissons $D(p||s) = \infty$. La loi s est désignée par *mesure de réf érence*.

Propriété 11 *Entropie relative est non-négative*.
Soient $p(x)$ et $s(x)$ deux lois de probabilité sur le même alphabet dénombrable $\mathcal{X}$, et $D(p||s)$ l'entropie relative de p par rapport à s. Alors

$$D(p||s) \geq 0. \qquad (1.18)$$

$\triangle$

Cette inégalité fondamentale de la Théorie de l'Information découle directement de l'inégalité de Jensen.

Avant de présenter l'inégalité de Jensen, nous rappelons la notion de fonctions convexes (concaves).

Définition 6 *Fonction convexe.*
Soit $f(\cdot)$ une fonction avec des valeurs en $\Re$:

$$\begin{array}{rccc} f & : & \Re^n & \rightarrow & \Re \\ & & x & \rightarrow & f(x) \end{array}$$

On dit que $f(\cdot)$ est une fonction convexe si et seulement si $\forall x_1, x_2 \in \Re^n$, $\lambda \in [0,1]$,

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2). \tag{1.19}$$

Si l'inégalité est satisfaite avec $<$ à la place de $\leq$, nous dirons que f est *strictement convexe*.

Ceci veut dire que le segment de droite qui joint les points $(x_1, f(x_1))$ et $(x_2, f(x_2))$ en $\Re^{n+1}$ est au- dessus de la surface de la fonction, voir Figure 1.10. Des exemples de fonctions convexes sont

- x (qui n'est pas strictement convexe!)
- $|x|$ (idem)
- e^x
- $\log \frac{1}{x}$, $x > 0$

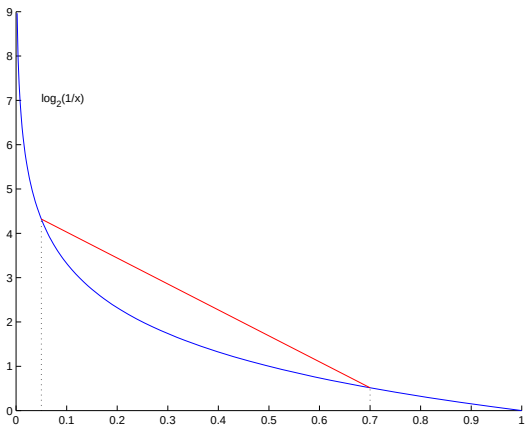

Figure 1.10: Illustration de la définition de fonction convexe.

Si la deuxième dérivée de $f(\cdot)$ existe, alors une condition sufisante pour que f soit convexe est que son Hessian soit une matrice définie non-négative (ne pas confondre, dans l'équation suivante, l'Hessian H_f – matrice des dérivées partielles de f – avec l'entropie !!)

$$H_f(x) \geq 0, \ \ \forall x \in \Re^n \Rightarrow f \text{ est convexe}, \qquad [H]_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}, , i, j = 1, \ldots, n.$$

Définition 7 *Fonction concave.*
Nous dirons qu'une fonction $f(\cdot)$ est concave si $-f(\cdot)$ est convexe.

Propriété 12 *Inégalité de Jensen.*
Soit f une fonction convexe, $f : \Re^n \rightarrow \Re$, et X un vecteur aléatoire en $\mathcal{X} \subset \Re^n$, avec loi de probabilité p_X. Alors

$$E_{p_X}\left[f(x)\right] \geq f\left(E_{p_X}\left[x\right]\right). \tag{1.20}$$

△

La démonstration de cette inégalité découle directement, pour des fonctions doublement diférentiables, de la définition de fonction convexe. Nous la présentons dans le cas scalaire. Considérons l'expansion de $f(x)$, pour $x \in \mathcal{X}$ arbitraire, autour d'un point fixé x_0:

$$f(x) = f(x_0) + (x - x_0)f'(x_0) + \frac{1}{2}(x - x_0)^2 f''(x^\star),$$

où nous avons utilisé le théorème du point intermédiaire ($x^\star \in [x, x_0]$). On remarque que le dernier terme du membre droit de cette équation est toujours non-négatif, vu que f est convexe, et donc sa deuxième dérivée est toujours non-négative:

$$(x - x_0)^2 f''(x^\star) \geq 0.$$

Nous pouvons alors écrire

$$f(x) \geq f(x_0) + (x - x_0)f'(x_0).$$

Si nous appliquons maintenant l'opérateur valeur moyenne nous obtenons

$$E_{p_X}\left[f(x)\right] \geq f(x_0) + E_{p_X}\left[(x - x_0)\right] f'(x_0) = f(x_0) + (E_{p_X}\left[x\right] - x_0)f'(x_0),$$

car x_0 est une constante fixée. Si nous prennons $x_0 = E[x]$, le deuxième terme devient nul, et nous obtenons directement l'inégalité de Jensen:

$$E_{p_X}\left[f(x)\right] \geq f(E_{p_X}[x]).$$

Propriété 13
Si f est *strictement convexe*, et l'inégalité de Jensen se vérifie avec égalité, c'est à dire, si

$$E_{p_X}\left[f(x)\right] = f(E_{p_X}[x]),$$

alors X est une *constante*, c'est à dire, sa loi de probabilité est concentrée dans un seul point $x^\star \in \mathcal{X}$:

$$p_X(x) = \delta(x - x^\star)$$

$\triangle$

Remarque 7
L'inégalité de Jensen nous permet de prouver facilement que la variance est toujours non-négative (x^2 est une fonction convexe!):

$$var(X) = E[x^2] - E[x]^2 \geq 0.$$

Nous revenons maintenant à la démonstration de la non-négativité de l'entropie relative, Propriété 11. De la définition de l'entropie relative

$$-D(p||s) = \sum_{x\in\mathcal{X}} p(x) \log \frac{s(x)}{p(x)} = E_p\left[\log \frac{s(x)}{p(x)}\right]$$

Comme $\log(\cdot)$ est une fonction concave,

$$-D(p||s) \leq \log E_p\left[\frac{s(x)}{p(x)}\right] = \log \sum_{x\in\mathcal{X}} p(x)\frac{s(x)}{p(x)} = \log 1 = 0$$

et nous obtenons donc le résultat prétendu :

$$D(p||s) \geq 0.$$

Comme $\log(\cdot)$ est une fonction *strictement* concave, nous pouvons encore affirmer que

$$D(p||s) = 0 \Leftrightarrow \forall x \in \mathcal{X} \ : \ \frac{s(x)}{p(x)} = c,$$

où c est une constante, c'est à dire, nous devons avoir

$$s(x) = cp(x), \forall x \in \mathcal{X}$$

pour que $D(p||s) = 0$. Mais, comme les deux lois doivent avoir une somme égale à 1, la seule solution possible est $c = 1$, et donc,

$$p(x) = s(x), \forall x \in \mathcal{X}.$$

Nous venons de prouver

Propriété 14 Soient $p(x)$ et $s(x)$ deux lois de probabilité sur le même alphabet $\mathcal{X}$ et $D(p||s)$ l'entropie relative de p par rapport à s. Alors

$$D(p||s) = 0 \Leftrightarrow p = s. \tag{1.21}$$

$\triangle$

Remarque 8
Une démonstration alternative (et plus simple) de la non-négativité de l'entropie relative est obtenue à partir de l'inégalité

$$\log x \leq x - 1, \tag{1.22}$$

avec égalité si et seulement si $x = 1$ (voir Figure 1.11). Alors nous pouvons écrire

$$-D(p||s) = \sum_{x\in\mathcal{X}} p(x) \log \frac{s(x)}{p(x)} \leq \sum_{x\in\mathcal{X}} p(x) \left(\frac{s(x)}{p(x)} - 1\right) = \sum_{x\in\mathcal{X}} s(x) - \sum_{x\in\mathcal{X}} p(x) = 0,$$

avec égalité si et seulement si $p(x) = s(x), \forall x \in \mathcal{X}$. D'où nous concluons

$$D(p||s) \geq 0.$$

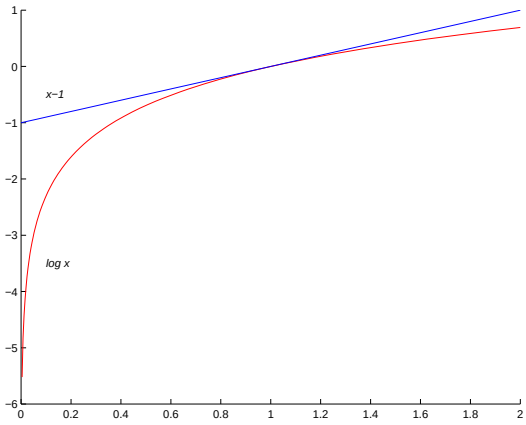

Figure 1.11: Illustration de (1.22).

Les Propriétés 11 and 14 justifient l'intréprétation de l'entropie relative comme une *distance* (ou mesure de similarité) entre des lois de probabilité. Cependant, elle n'est pas une vraie distance, au sens mathématique du terme, car elle ***n'est pas symmétrique*** :

$$D(p||s) \neq D(s||p),$$

et ***ne vérifie pas l'inégalité triangulaire*** :

$$D(p||s) \not\leq D(p||t) + D(t||s).$$

Comme elle est non-négative, et égale à zéro uniquement quand les deux lois coincident, l'entropie relative est une mesure de *distortion*, qui peut être considérée comme une généralisation de distance.

L'interprétation de D comme une distance est souvent utile pour analyser le son comportement, et nous verrons qu'elle exibe des relation avec des mesures de distance ordinaires entre des lois de probabilité. Elle joue un r ôle fondamentale en théorie de l'information, et comme nous le verrons, toutes les autres mesures que nous utiliserons (entropie, information mutuelle, et ses versions conditionnelles) peuvent être exprimées comme une divergence.

Il y a trois manières de voir l'**entropie comme un cas particulier de divergence**. La *première* consiste à permettre à la mesure s de ne pas avoir masse unitaire (s est une mesure générale, pas une mesure *de probabilité*), et prendre pour s la mesure $\bar{s}$ qu'attribue une masse unitaire à chaque point de $\mathcal{X}$ (s est donc la *fonction indicatrice* de $\mathcal{X}$) :

$$\forall x \in \mathcal{X}, \bar{s}(x) = 1, \qquad \Rightarrow\ D(p||\bar{s}) = \sum_{x \in \mathcal{X}} p(x) \log p(x) = -H(X).$$

L'entropie est donc le symmétrique de la divergence de p par rapport à cette mesure unitaire $\bar{s}$. (Remarquez que comme $\bar{s}$ n'est pas une mesure de probabilité la Propriété 11 ne s'applique plus.)

Deuxièmement, si nous prennons $s = u$, la mesure (de probabilité) uniforme en $\mathcal{X}$,

$$\forall x \in \mathcal{X},\ \ u(x) = \frac{1}{|\mathcal{X}|},$$

nous concluons facilement que

$$D(p||u) = \log |\mathcal{X}| - H(X), \tag{1.23}$$

et donc l'entropie $H(X)$ est le logarithme de la taille de l'alphabet $\mathcal{X}$ moins la divergence de p_X par rapport à la mesure uniforme u:

$$H(X) = \log |\mathcal{X}| - D(p||u),$$

ou encore

$$H(X) = -D(u||\bar{s}) - D(p||u),$$

où nous avons utilisé $\log|\mathcal{X}| = H(u) = -D(u||\bar{s})$, où $\bar{s}$ est la mesure uniforme introduite plus haut.

Finalement, nous pouvons encore établir une *troisième* relation entre entropie et entropie relative, en faisant appel à notion de mesure produit. Soient p et q deux mesures de probabilité définies dans un m ême alphabet $\mathcal{X}$. Nous allons définir, à partir de la loi p, deux mesures de probabilité dans le produit $\mathcal{X} \times \mathcal{X}$ de la forme suivante:

- p_0 est la mesure "diagonale":

$$p_0(x,y) = \begin{cases} p(x), & \text{si } x = y \\ 0, & \text{si } x \neq y \end{cases}$$

  C'est la loi conjointe de deux variables aléatoires X et Y parfaitement correlés: avec probabilité 1 les variables prennent la même valeur : toutes les réalisations sont de la forme $(x,x), x \in \mathcal{X}$.

- $p \cdot q$ est la mesure produit usuelle:

$$p \cdot q(x,y) = p(x)q(y),$$

  qui correspond au cas où les valeurs de x et y sont statistiquement indépendants.

Il est facile de démontrer que

$$H(X) = D(p_0||p \cdot p),$$

et donc l'entropie peut encore être intréprétée comme la divergence entre ces deux mesures produit extrêmes, qui correspondent aux cas de correlation parfaite (p_0) et indépendance statistique ($p \cdot p$).

L'entropie relative a plusieurs propriétés attractives. En particulier, si nous considérons le sous-ensemble du simplex probabiliste des lois qui sont à une distance inférieure ou égale à α d'une certaine distribution s:

$$\mathcal{A} = \{p : D(p||s) \leq \alpha\}$$

il peut être démontré que $\mathcal{A}$ est un *ensemble convexe* (voir la Figure 1.12, où la loi s est indiquée par une étoile jaune), ce qui est une propriété importante dans des problèmes d'optimisation.

De l'équation (1.23) et de la non-négativité de l'entropie relative découle directement une borne sup érieure pour l'entropie de variables aléatoires dans un alphabet fini:

Propriété 15 *Borne supérieure de H.*
Soit X une variable aléatoire dans un alphabet fini $\mathcal{X}$, $|\mathcal{X}| = m$. Alors

$$H(X) \leq \log m. \tag{1.24}$$

△

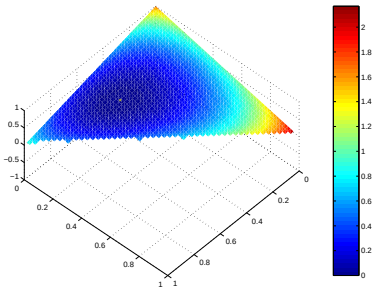

Figure 1.12: Entropie relative par rapport à la mesure indiquée par une étoile (jaune).

Nous avions déjà identifé cette borne dans les exemples 1 et 2.

L'information mutuelle, introduite dans la Définition 4, et comme le montre l'équation (1.16), peut aussi être écrite comme une entropie relative:

Propriété 16 *Information mutuelle et entropie relative.*

$$I(X;Y) = D(p(x,y)||p(x)p(y)). \tag{1.25}$$

△

L'information mutuelle est donc une *généralisation de la covariance* comme mesure quantitative de d épendance statistique (nous rappelons que pour des variables Gaussiennes, la corrélation est effectivement une mesure d'indépendance statistique: des variables Gaussiennes sont statistiquement indépendantes si et seulement si elles son non-correlées). De cette relation découle immédiatement

Propriété 17 *Non-négativité de l'information mutuelle.*

$$I(X;Y) \geq 0 \tag{1.26}$$
$$I(X;Y) = 0 \Leftrightarrow p(x,y) = p(x)p(y) \Leftrightarrow X, Y \text{ stat. ind.} \tag{1.27}$$

△

De la même manière, pour l'entropie conditionnelle, nous pouvons affirmer que

Propriété 18 *Borne supérieure de l'entropie conditionnelle.*

$$H(X|Y) \leq H(X), \tag{1.28}$$

avec égalité si et seulement si les deux variables sont statistiquement indépendantes. Ceci est une cons équence immédiate de la définition d'information mutuelle, eq. (1.11), et de la Propriété 11. △

La propriété 18 peut être interpretée comme " les observations sont toujours utiles" : la connaissance d'une autre variable Y ne peut que faire diminuer l'incertitude existante sur la variable X.

Remarque 9
Cette dernière intréprétation suppose, bien-sûr, une utilisation *optimale* de la nouvelle information !

Remarque 10
Notez que la borne présentée dans la Propriété 18 concerne l'entropie conditionnelle $H(X|Y)$. Elle peut être violée pour l'entropie de la loi conditionnelle $p_{X|y}(x|y)$: pour certaines valeurs particulières de y nous pouvons avoir $H(X|Y = y) > H(X)$! En moyenne, cependant, l'entropie doit décroitre.

Propriété 19 *Concavité de l'entropie.*
L'entropie $H(p)$ est une fonction concave de p. Soient p_1 et p_2 deux lois de probabilité définies dans le même alphabet $\mathcal{X}$, et $\lambda \in [0, 1]$. Alors

$$H(\lambda p_1 + (1 - \lambda)p_2) \geq \lambda H(p_1) + (1 - \lambda)H(p_2). \tag{1.29}$$

△

Cette Propriété affirme donc que l'*entropie du mélange* de deux distributions *est supérieure à la moyenne* pondérée (avec les mêmes coefficients) *de l'entropie des deux lois*. Elle peut, évidement, être étendue au mélange d'un ensemble dénombrable de lois de probabilité. La démonstration de cette inégalité peut être faite directement en calculant les deuxièmes dérivées. Nous présentons une démonstration alternative, qui fait appel à la régle de chaîne de l'entropie, établie dans la Propriété 7, page 14.
Désignons par $X \in \mathcal{X}$ et $Y \in \mathcal{X}$ des variables aléatoires qui suivent les lois p_1 et p_2, respectivement. Soit Θ une variable aléatoire binaire $\Theta \in \{1, 2\}$, avec

$$p_\Theta(1) = Pr\{\Theta = 1\} = \lambda, \qquad p_\Theta(2) = Pr\{\Theta = 2\} = 1 - \lambda.$$

Nous construisons une nouvelle variable aléatoire Z, selon la régle suivante. On génère d'abord $\theta \sim p_\Theta$. Si $\theta = 1$, nous procédons à un tirage selon p_1 : $z \sim p_1(X)$; sinon nous utilisons la loi p_2 : $z \sim p_2(Y)$. La loi de la nouvelle variable est

$$p_Z(x) = \lambda p_1(x) + (1 - \lambda)p_2(x), \forall x \in \mathcal{X}.$$

c'est à dire, p_Z est la loi de mélange de p_1 et p_2 avec coefficents $(\lambda, 1-\lambda)$. L'entropie conditionnelle de Z sachant Θ est, par application de la Définition 3, page 12,

$$H(Z|\Theta) = \lambda H(Z|\Theta = 1) + (1-\lambda)H(Z|\Theta = 2).$$

Mais

$$H(Z|\Theta = 1) = H(p_1), \qquad \text{et} \qquad H(Z|\Theta = 2) = H(p_2).$$

De la borne supérieure pour l'entropie conditionnelle, nous obtenons finalement

$$H(Z|\Theta) \leq H(Z) \Leftrightarrow \lambda H(p_1) + (1-\lambda)H(p_2) \leq H(p_Z) = H(\lambda p_1 + (1-\lambda)p_2).$$

De la non-négativité de l'information mutuelle et de l'équation (1.12), que nous répétons ici :

$$0 \leq I(X;Y) = H(X) + H(Y) - H(X;Y),$$

nous pouvons établir

Propriété 20

$$\max\left(H(X), H(Y)\right) \leq H(X,Y) \leq H(X) + H(Y), \tag{1.30}$$

où la dernière inégalité est stricte si et seulement si X et Y ne sont pas statistiquement indépendantes.

△

Propriété 21 *Règle de la chaîne pour l'information mutuelle.*

$$I(X^n;Y) = \sum_{i=1}^{n} X(X_i;Y|X^{i-1}).$$

△

Cette propriété et la suivante sont facilement obtenues en utilisant récursivement la loi de Bayes pour exprimer les lois conjointes comme le produit d'une loi marginale et des lois conditionnelles :

Propriété 22 *"Règle de la chaîne" pour l'entropie relative.*

$$D(p(x,y)||s(x,y)) = D(p(x)||s(x)) + E_X\left[D(p(y|X=x)||s(y|X=x))\right].$$

△

Finalement, nous établissons une relation qui nous sera utile plus tard. Pour cela nous introduisons d'abord la notion de *raffinement* d'une partition.

Définition 8 *Raffinement d'une partition.*
Soient $\mathcal{Q}$ et $\mathcal{R}$ deux partitions d'un ensemble Ω. Nous disons que $\mathcal{R}$ est un raffinement de $\mathcal{Q}$ si tous les éléments de $\mathcal{Q}$ peuvent être écrits comme une union des éléments de $\mathcal{R}$. Nous nottons cette relation par $\mathcal{Q} < \mathcal{R}$.

Exemple 8 Considérez l'ensemble suivant $\mathcal{X} = \{1, 2, 3, 4, \ldots, 100\}$, et les partitions

$$\begin{aligned}
\mathcal{P}_1 &= \{\{1, 3, 5, \ldots, 99\}, \{2, 4, 6, \ldots, 100\}\} \\
\mathcal{P}_2 &= \{\{1, 2, 3, \ldots, 50\}, \{51, 52, \ldots, 100\}\} \\
\mathcal{P}_3 &= \{\{1, 2, 3, \ldots, 50\}, \{51, 52, \ldots, 75\}, \{76, 77, \ldots, 100\}\}
\end{aligned}$$

Nous pouvons constater que $\mathcal{P}_3$ est un raffinement de $\mathcal{P}_2$ mais pas de $\mathcal{P}_1$:

$$\mathcal{P}_2 < \mathcal{P}_3, \qquad \mathcal{P}_1 \not< \mathcal{P}_3.$$

Propriété 23 Soient P et M deux mesures définies dans le même espace mesurable $(\Omega, \mathcal{B})$, et soient $\mathcal{Q}$ et $\mathcal{R}$ deux partitions finies, avec $\mathcal{R}$ un raffinement de $\mathcal{Q}$: $\mathcal{Q} < \mathcal{R}$. Désignons par $P_\mathcal{Q}$ la loi de probabilité des éléments de la partition $\mathcal{Q}$ induite par la mesure P (voir discussion de la relation entre lois de probabilité et partitions, page 5). Alors

$$D(P_\mathcal{Q}||M_\mathcal{Q}) \leq D(P_\mathcal{R}||M_\mathcal{R}),$$

et

$$H(P_\mathcal{Q}) \leq H(P_\mathcal{R}).$$

△

C'est à dire, cette Propriété nous dit qu'un raffinement de la partition conduit à une augmentation de l'entropie et de la distance entre lois de probabilité. Ceci veut dire, par exemple, que si nous diminuons le nombre de bits qui codent (en regroupant, par exemple, les niveaux deux à deux) chaque pixel d'une image, sont entropie doit d écroitre.

Nous allons maintenant démontrer l'inégalité sur les entropies relatives. La démonstration de l'inégalité sur les entropies peut être faite selon la même approche.

Si le raffinement de la partition conduit à des lois qui ne sont pas absolument continues (cela veut dire que certains évènements de mesure nulle selon $M_\mathcal{R}$ ont une probabilité positive selon $P_\mathcal{R}$), alors $D(P_\mathcal{R}||M_\mathcal{R}) = \infty$ et la propriété est triviellement satisfaite.

Si $D(P_{\mathcal{Q}}||M_{\mathcal{Q}}) = \infty$, cela veut dire qu'il existe au moins un élément $\mathcal{Q}_i \in \mathcal{Q}$ tel que $M(\mathcal{Q}_i) = 0$ mais $P(\mathcal{Q}_i) \neq 0$. Alors, il existe un $\mathcal{R}_j \subset Q_i$ tel que $M(\mathcal{R}_j) = 0$ et $P(\mathcal{R}_j) > 0$, c'est à dire P n'est pas absolument continue par rapport à M et donc $D(P_{\mathcal{R}}||M_{\mathcal{R}}) = \infty$,et donc l'inégalité est satisfaite avec égalité ($\infty = \infty$).

Il nous reste le cas $D(P_{\mathcal{R}}||M_{\mathcal{R}}) \neq \infty$ et $D(P_{\mathcal{Q}}||M_{\mathcal{Q}}) \neq \infty$. Considérons la différence entre les deux entropies relatives :

$$D(P_{\mathcal{R}}||M_{\mathcal{R}}) - D(P_{\mathcal{Q}}||M_{\mathcal{Q}}) = \sum_j P(\mathcal{R}_j) \log \frac{P(\mathcal{R}_j)}{M(\mathcal{R}_j)} - \sum_i P(\mathcal{Q}_i) \log \frac{P(\mathcal{Q}_i)}{M(\mathcal{Q}_i)} \ .$$

Comme $\mathcal{Q} < \mathcal{R}$, nous pouvons regrouper la somme sur les éléments de $\mathcal{R}$ en considérant tous ceux qui appartiennent à un même élément de $\mathcal{Q}$:

$$D(P_{\mathcal{R}}||M_{\mathcal{R}}) - D(P_{\mathcal{Q}}||M_{\mathcal{Q}}) = \sum_i \left[\sum_{j:\mathcal{R}_j \subset \mathcal{Q}_i} P(\mathcal{R}_j) \log \frac{P(\mathcal{R}_j)}{M(\mathcal{R}_j)} - P(\mathcal{Q}_i) \log \frac{P(\mathcal{Q}_i)}{M(\mathcal{Q}_i)} . \right]$$

Nous pouvons maintenant démontrer que chaque terme entre parenthèses est non-négatif.

Fixons une valeur de i. Comme $\mathcal{Q}_i = \bigcup \mathcal{R}_j$, si $P(\mathcal{Q}_i) = 0 \Rightarrow P(\mathcal{R}_j) = 0, \forall j$: $\mathcal{R}_j \subset \mathcal{Q}_i$, et donc le terme correspondant est nul. Si $P(\mathcal{Q}_i) \neq 0$, nous pouvons re-écrire le terme correspondant comme

$$P(\mathcal{Q}_i) \left[\sum_{j:\mathcal{R}_j \subset \mathcal{Q}_i} \frac{P(\mathcal{R}_j)}{P(\mathcal{Q}_i)} \log \frac{P(\mathcal{R}_j)/P(\mathcal{Q}_i)}{M(\mathcal{R}_j)/M(\mathcal{Q}_i)} \right]$$

où nous avons utilisé le fait que $D(P_{\mathcal{Q}}||M_{\mathcal{Q}}) \neq \infty$ et donc $P(\mathcal{Q}_i) \neq 0 \Rightarrow M(\mathcal{Q}_i) \neq 0$. Pour les valeurs de j dans chaque terme, $\mathcal{R}_j \subset \mathcal{Q}_i$, et donc

$$\mathcal{R}_j = \mathcal{R}_j \cap \mathcal{Q}_i \Rightarrow \frac{P(\mathcal{R}_j)}{P(\mathcal{Q}_i)} = \frac{P(\mathcal{R}_j \bigcap \mathcal{Q}_i)}{P(\mathcal{Q}_i)} = P(\mathcal{R}_j|\mathcal{Q}_i).$$

Pour les autres valeurs de j, pour lesquelles $\mathcal{R}_j \not\subset Q_i$, alors $\mathcal{R}_j \cap \mathcal{Q}_i = \emptyset$, et

$$P(\mathcal{R}_j|\mathcal{Q}_i) = 0 \ .$$

Des expressions équivalentes peuvent être établies pour la mesure M. Une exprèssion équivalente de chaque terme entre parenthèses dans l'expression de la différence des entropies relatives est donc

$$P(\mathcal{Q}_i) \left[\sum_j P(\mathcal{R}_j|\mathcal{Q}_i) \log \frac{P(\mathcal{R}_j|\mathcal{Q}_i)}{M(\mathcal{R}_j|\mathcal{Q}_i)} \right] = P(\mathcal{Q}_i) D\left(P(\mathcal{R}|\mathcal{Q}_i)||M(\mathcal{R}|\mathcal{Q}_i)\right) \geq 0 \ ,$$

où nous avons utilisé la non-négativité de la mesure P et de l'entropie relative. De ce résultat découle donc

$$D(P_{\mathcal{R}}||M_{\mathcal{R}}) - D(P_{\mathcal{Q}}||M_{\mathcal{Q}}) = \sum_i P(\mathcal{Q}_i) D\left(P(\mathcal{R}|\mathcal{Q}_i)||M(\mathcal{R}|\mathcal{Q}_i)\right) \geq 0 \ ,$$

et nous obtenons l'expression souhaitée :

$$D(P_{\mathcal{R}}||M_{\mathcal{R}}) \geq D(P_{\mathcal{Q}}||M_{\mathcal{Q}}) \; .$$