Mémento de théorie de l'information

Gilles Zémor

24 mars 2011

1 Grandeurs informationnelles

Soit X une variable aléatoire prenant ses valeurs dans l'ensemble fini $\mathcal{X}$ à m éléments. Soit p la loi de X, c'est-à-dire la donnée des $P(X = x)$. On écrira indifféremment $p(x)$ pour désigner $P(X = x)$, ou $p = (p_1 \dots p_m)$ si l'on convient que $\mathcal{X} = \{x_1, \dots x_m\}$ et que $p_i = P(X = x_i)$.

L'*entropie* de X ne dépend que de sa loi p et est définie par

$$\sum_{x \in \mathcal{X}} P(X = x) \log_2 \frac{1}{P(X = x)} = \sum_{i=1}^{m} p_i \log_2 \frac{1}{p_i}.$$

On la note indifféremment $H(X)$ ou $H(p)$.

Si X et Y sont deux variables aléatoires, alors le couple (X, Y) est aussi une variable aléatoire et on définie l'entropie jointe $H(X, Y)$ tout simplement comme l'entropie du couple (X, Y), soit

$$H(X, Y) = \sum_{x,y} P(X = x, Y = y) \log_2 \frac{1}{P(X = x, Y = y)}.$$

On définit la «distance» de Kullback entre deux lois p et q par

$$\mathrm{D}(p \,||\, q) = \sum_{x} p(x) \log_2 \frac{p(x)}{q(x)}.$$

Lemme 1 *On a* $\mathrm{D}(p \,||\, q) \geqslant 0$ *et* $\mathrm{D}(p \,||\, q) = 0$ *si et seulement si* $p = q$.

Preuve : Pour tout réel $z \geqslant 0$ on a $\ln z \leqslant z - 1$. On en déduit

$$\ln \frac{q(x)}{p(x)} \leqslant \frac{q(x)}{p(x)} - 1$$

$$p(x)\ln\frac{q(x)}{p(x)} \leqslant q(x)-p(x)$$

$$p(x)\ln\frac{p(x)}{q(x)} \geqslant p(x)-q(x)$$

et en sommant sur x,

$$\sum_{x\in\mathcal{X}} p(x)\ln\frac{p(x)}{q(x)} \geqslant 1-1$$

$$\mathrm{D}(p\,||\,q) \geqslant 0.$$

∎

Corollaire 2 *L'entropie d'une variable aléatoire X prenant $m=|\mathcal{X}|$ valeurs est maximale lorsqu'elle est de loi uniforme et l'on a alors $H(X)=\log_2 m$.*

Preuve : L'entropie de la loi uniforme vaut

$$\sum_{i=1}^{m}\frac{1}{m}\log_2 m = \log_2 m.$$

Pour tout autre loi p on a

$$\begin{aligned}
\log_2 m - H(p) &= \log_2 m - \sum_{i=1}^{m} p_i \log_2\frac{1}{p_i} \\
&= \sum_{i=1}^{m} p_i\log_2 m + \sum_{i=1}^{m} p_i\log_2 p_i \\
&= \sum_{i=1}^{m} p_i\log_2 p_i m \\
&= \mathrm{D}(p\,||\,1/m) \geqslant 0.
\end{aligned}$$

∎

Voici un autre exemple de maximisation de l'entropie dans le cas d'une variable prenant une infinité de valeurs.

Proposition 3 *Parmi les variables à valeurs dans $\mathbb{N}$ d'espérance finie donnée μ, l'entropie maximale est atteinte pour les variables de loi géométrique d'espérance μ.*

Preuve : La loi géométrique Γ de paramètre γ est définie par $P(X=i)=(1-\gamma)\gamma^i$ et son espérance vaut

$$\mu = (1-\gamma)\sum_{i=1}^{\infty} i\gamma^i = \frac{\gamma}{(1-\gamma)}.$$

Son entropie vaut :

$$\begin{aligned} H(\Gamma) &= -\sum_{i\in\mathbb{N}}(1-\gamma)\gamma^i \log_2(1-\gamma)\gamma^i \\ &= \log_2 \frac{1}{1-\gamma}\sum_{i\in\mathbb{N}}(1-\gamma)\gamma^i + \log_2\frac{1}{\gamma}\sum_{i\in\mathbb{N}} i(1-\gamma)\gamma^i \\ &= \log_2 \frac{1}{1-\gamma} + \mu \log_2 \frac{1}{\gamma} \end{aligned}$$

Soit maintenant une loi quelconque p sur $\mathbb{N}$ d'espérance μ, c'est-à-dire telle que $\sum_{i\in\mathbb{N}} ip_i = \mu$, on a :

$$\begin{aligned} H(\Gamma) - H(p) &= \log_2 \frac{1}{1-\gamma} + \mu\log_2\frac{1}{\gamma} + \sum_{i\in\mathbb{N}} p_i \log_2 p_i \\ &= \log_2 \frac{1}{1-\gamma}\sum_{i\in\mathbb{N}} p_i + \log_2\frac{1}{\gamma}\sum_{i\in\mathbb{N}} ip_i + \sum_{i\in\mathbb{N}} p_i\log_2 p_i \\ &= \sum_{i\in\mathbb{N}} p_i\left(\log_2\frac{1}{1-\gamma} + \log_2\frac{1}{\gamma^i} + \log_2 p_i\right) \\ &= \sum_{i\in\mathbb{N}} p_i \log_2 \frac{p_i}{(1-\gamma)\gamma^i} = \mathrm{D}(p\,||\,\Gamma) \geqslant 0. \end{aligned}$$

∎

Étant données deux variables X et Y, on définit l'*entropie conditionnelle*

$$\begin{aligned} H(X|Y) &= \sum_{x,y} P(X=x, Y=y)\log_2\frac{1}{P(X=x|Y=y)} \\ &= \sum_{y} P(Y=y)\sum_{x} P(X=x|Y=y)\log_2\frac{1}{P(X=x|Y=y)}. \end{aligned}$$

Proposition 4 *On a :*

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

Preuve : On a :

$$\begin{aligned} H(Y) &= \sum_{y} P(Y=y)\log_2\frac{1}{P(Y=y)} \\ &= \sum_{x,y} P(X=x, Y=y)\log_2\frac{1}{P(Y=y)} \end{aligned}$$

donc

$$\begin{aligned} H(Y) + H(X|Y) &= \sum_{x,y} P(X=x, Y=y) \log_2 \frac{1}{P(Y=y)P(X=x|Y=y)} \\ &= \sum_{x,y} P(X=x, Y=y) \log_2 \frac{1}{P(X=x, Y=y)} \\ &= H(X,Y). \end{aligned}$$

∎

Enfin on définit l'*information mutuelle* entre deux variables X et Y par

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

D'après la proposition 4 on a

$$\begin{aligned} I(X,Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X). \end{aligned}$$

Proposition 5 *L'information mutuelle est positive. On a toujours* $I(X,Y) \geqslant 0$.

Preuve : En écrivant

$$\begin{aligned} H(X) &= \sum_{x,y} P(X=x, Y=y) \log_2 \frac{1}{P(X=x)} \\ H(Y) &= \sum_{x,y} P(X=x, Y=y) \log_2 \frac{1}{P(Y=y)} \end{aligned}$$

on obtient

$$\begin{aligned} I(X,Y) &= H(X) + H(Y) - H(X,Y) \\ &= \sum_{x,y} P(X=x, Y=y) \log_2 \frac{P(X=x, Y=y)}{P(X=x)P(Y=y)} \\ &= \mathrm{D}(p(X,Y) \,||\, p(X)p(Y)) \geqslant 0. \end{aligned}$$

∎

À retenir

- $H(X) = H(p) = \sum_i p_i \log_2 \frac{1}{p_i}$.

- $\mathrm{D}(p \,||\, q) = \sum_i p_i \log_2 \frac{p_i}{q_i} \geqslant 0$.
- L'entropie d'une loi prenant m valeurs est maximisée pour la loi uniforme et son maximum vaut $\log_2 m$.
- $$\begin{aligned} H(X|Y) &= \sum_{x,y} P(X=x, Y=y) \log_2 \frac{1}{P(X=x|Y=y)} \\ &= \sum_y P(Y=y) \sum_x P(X=x|Y=y) \log_2 \frac{1}{P(X=x|Y=y)}. \end{aligned}$$
- $H(X,Y) = H(Y) + H(X|Y)$.
- $H(X|Y) \leqslant H(X)$.
- $$\begin{aligned} I(X,Y) &= H(X) + H(Y) - H(X,Y) \\ &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X). \end{aligned}$$
- $I(X,Y) \geqslant 0$.
- Pour démontrer une inégalité informationnelle, penser à la transformer en une expression du type $\mathrm{D}(p \,||\, q) \geqslant 0$.

2 Codage de source. Compression

Un *code* (compressif) est un ensemble fini de mots $C \subset \{0,1\}^*$. La *longueur* $\ell(c)$ d'un mot $c \in C$ est le nombre de symboles binaires qui constituent le mot c. Un *codage* d'une variable aléatoire X prenant ses valeurs dans $\mathcal{X}$ est une application

$$\mathbf{c} \ : \mathcal{X} \to C.$$

Une suite $X_1 \dots X_n$ de copies indépendantes de même loi que X se traduit par une suite de symboles de $\mathcal{X}$. Elle se traduit, par concaténation, en une suite de mots de C, qui elle-même se traduit en une suite de symboles binaires. En d'autres termes l'application $\mathcal{X} \to C$ donne naissance, par concaténation, à l'application

$$\mathbf{c}^* \ : \mathcal{X}^* \to C^*.$$

Le code C est dit *uniquement déchiffrable* si toute suite binaire de C^* se décompose de manière unique en une concaténation de mots de C. Par exemple le code $C = \{0, 01\}$ est uniquement déchiffrable, mais $C = \{0, 01, 001\}$ ne l'est pas, car $0001 = 0 \cdot 001 = 0 \cdot 0 \cdot 01$.

À une variable aléatoire X et son codage par $\mathbf{c}$, on associe sa *longueur moyenne* $\overline{\ell}(\mathbf{c})$ égale au nombre moyen de chiffres binaires par symbole de $\mathcal{X}$ codé

$$\overline{\ell}(\mathbf{c}) = \sum_{x \in \mathcal{X}} P(X = x)\ell(\mathbf{c}(x)).$$

Exemple. Soit $\mathcal{X} = \{1, 2, 3, 4\}$ et soit X à valeurs dans X de loi $p_1 = 1/2, p_2 = 1/4, p_3 = 1/4, p_4 = 1/8$ où $p_i = P(X = i)$. On considère le codage défini par

$$\begin{aligned} \mathbf{c}(1) &= 0 \\ \mathbf{c}(2) &= 10 \\ \mathbf{c}(3) &= 110 \\ \mathbf{c}(4) &= 111. \end{aligned}$$

On a :

$$\overline{\ell} = \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{4} \cdot 2 + \frac{1}{8} \cdot 3 = \frac{15}{8}.$$

On constate que $\ell(\mathbf{c}) = H(X)$.

2.1 Codes préfixes et codes uniquement déchiffrables

Parmi les codes uniquement déchiffrables on distingue les codes préfixes. Un code est dit *préfixe* si aucun de ses mots n'est le préfixe d'un autre mot de code. Il est clair qu'un code préfixe est uniquement déchiffrable. Il existe, par contre des codes uniquement déchiffrables qui ne sont pas préfixes, par exemple $\{0, 01\}$.

Un code préfixe peut être décrit par un *arbre binaire* dont les feuilles sont associées aux mots du code, comme illustré sur la figure 1.

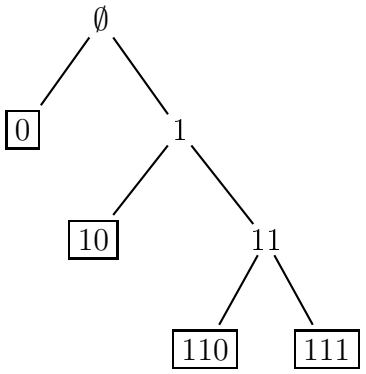

FIG. 1 – L'arbre associé au code préfixe $\{0, 10, 110, 111\}$

Proposition 6 (Inégalité de Kraft) *Soit $\mathcal{X} = \{1, 2, \ldots, m\}$ et soit $\ell_1 \ldots \ell_m$ une suite d'entiers positifs. Il existe un code préfixe C et un encodage $\mathbf{c} : \mathcal{X} \to C$ tels que $\ell(\mathbf{c}(i)) = \ell_i$, si et seulement si*

$$\sum_{i=1}^{m} 2^{-\ell_i} \leqslant 1.$$

Preuve : Soit C un code préfixe à m mots de longueurs $\ell_1, \ldots, \ell_m$. Supposons d'abord que tous les ℓ_i sont égaux à une même longueur ℓ. Comme les mots de code sont associés aux feuilles d'un arbre binaire, le code C a au maximum 2^ℓ mots et l'inégalité de Kraft $m2^{-\ell} \leqslant 1$ est clairement satisfaite. Supposons maintenant que la longueur des mots n'est plus constante. Soit $\ell = \ell_{\max}$ la longueur maximale d'un mot. Considérons un mot de longueur i associé à une feuille de l'arbre de profondeur i. Remplaçons cette feuille par un arbre binaire de profondeur $\ell - i$ à $2^{\ell-i}$ feuilles. On constate que $2^{\ell-i}2^{-\ell} = 2^{-i}$ de telle sorte que la valeur de la somme

$$\sum_{i=1}^{m} 2^{-\ell_i}$$

est inchangée si l'on remplace l'arbre associé à C par un arbre de profondeur constante ℓ en développant chaque sommet jusqu'à la profondeur ℓ. On est ainsi ramené au cas précédent. La réciproque se démontre par un raisonnement analogue. ∎

La proposition précédente reste vraie si l'on remplace l'hypothèse «préfixe» par l'hypothèse plus faible «uniquement déchiffrable».

Théorème 7 (McMillan) *Soit $\mathcal{X} = \{1, 2, \ldots, m\}$ et soit $\ell_1 \ldots \ell_m$ une suite d'entiers positifs. Il existe un code uniquement déchiffrable C et un encodage $\mathbf{c} : \mathcal{X} \to C$ tels que $\ell(\mathbf{c}(i)) = \ell_i$, si et seulement si*

$$\sum_{i=1}^{m} 2^{-\ell_i} \leqslant 1.$$

Preuve : Si une distribution de longueurs (ℓ_i) satisfait l'inégalité de Kraft nous savons déjà qu'il existe un code préfixe, donc uniquement déchiffrable, de distribution (ℓ_i). Soit maintenant un code C uniquement déchiffrable. Notons C^k l'ensemble des suites binaires obtenues par concaténation d'exactement k mots de C. Si $c = c_1 c_2 \ldots c_k$ est la concaténation des k mots $c_1, \ldots c_k$ de C on a $\ell(c) = \ell(c_1) + \cdots + \ell(c_k)$ et donc, par unique déchiffrabilité,

$$\begin{aligned}\left(\sum_{c \in C} 2^{-\ell(c)}\right)^k &= \sum_{c_1 \ldots c_k \in C} 2^{-\ell(c_1)\cdots-\ell(c_k)} \\ &= \sum_{c \in C^k} 2^{-\ell(c)}\end{aligned}$$

puisque chaque mot de C^k est associé à exactement une suite de k mots de C dont il est la concaténation. La longueur maximale $\ell(c)$ d'un mot de C^k est $k\ell_{\max}$ où $\ell_{\max}$ est la longueur maximale d'un mot de C. On a donc

$$\left(\sum_{c\in C} 2^{-\ell(c)}\right)^k = \sum_{i=1}^{k\ell_{\max}} 2^{-i} A_i$$

où A_i est le nombre de mots de C^k de longueur i. Comme $A_i \leqslant 2^i$ on en déduit

$$\left(\sum_{c\in C} 2^{-\ell(c)}\right)^k \leqslant k\ell_{\max}$$

ce qui implique

$$\sum_{c\in C} 2^{-\ell(c)} \leqslant \ell_{\max}^{1/k} k^{1/k}.$$

Comme cette dernière inégalité doit être vraie pour tout k on en déduit

$$\sum_{c\in C} 2^{-\ell(c)} \leqslant 1$$

ce qui démontre le théorème. ∎

2.2 Longueur moyenne et entropie

Proposition 8 *Soit X une variable aléatoire prenant ses valeurs dans un ensemble fini $\mathcal{X}$. Soit $\mathbf{c}$ un codage de X par un code uniquement déchiffrable C. La longueur moyenne $\overline{\ell}(\mathbf{c})$ de ce codage vérifie*

$$\overline{\ell}(\mathbf{c}) \geqslant H(X).$$

Preuve : Posons $\mathcal{X} = \{x_1, \ldots, x_m\}$ et $C = \{c_1, \ldots, c_m\}$ de telle sorte que $\mathbf{c}(x_i) = c_i$. Considérons

$$Q = \sum_{i=1}^{m} 2^{-\ell(c_i)}.$$

On a, d'après le théorème de McMillan, $Q \leqslant 1$. Posons

$$q_i = \frac{2^{-\ell(c_i)}}{Q}$$

de telle sorte que $(q_1 \ldots q_m)$ est une distribution de probabilités (i.e. $\sum_i q_i = 1$). Soit $p_i = P(X = x_i)$. L'inégalité $\mathrm{D}(p \,||\, q) \geqslant 0$ (Lemme 1) s'écrit

$$\sum_i p_i \log_2 \frac{p_i}{q_i} \geqslant 0$$

soit

$$\begin{aligned} -H(X) - \sum_i p_i \log_2 q_i &\geqslant 0 \\ \sum_i p_i \ell(c_i) + \sum_i p_i \log_2 Q &\geqslant H(X) \\ \overline{\ell}(\mathbf{c}) &\geqslant H(X) - \log_2 Q \end{aligned}$$

ce qui prouve la proposition puisque $Q \leqslant 1$. ∎

Proposition 9 *Soit X une variable aléatoire prenant ses valeurs dans un ensemble fini $\mathcal{X}$. Il existe un codage $\mathbf{c}$ de X dont la longueur moyenne $\overline{\ell}(\mathbf{c})$ vérifie*

$$\overline{\ell}(\mathbf{c}) \leqslant H(X) + 1.$$

Preuve : Soit $\mathcal{X} = \{x_1, \ldots, x_m\}$ et soit $p_i = P(X = x_i)$. Posons

$$\ell_i = \left\lceil \log_2 \frac{1}{p_i} \right\rceil.$$

On a :

$$\begin{aligned} \log_2 \frac{1}{p_i} &\leqslant \ell_i \\ -\ell_i &\leqslant \log_2 p_i \\ 2^{-\ell_i} &\leqslant p_i \\ \sum_i 2^{-\ell_i} &\leqslant 1. \end{aligned}$$

Il existe donc un codage de X par un code C de distribution des longueurs (ℓ_i). Par ailleurs on a :

$$\begin{aligned} \ell_i &< \log_2 \frac{1}{p_i} + 1 \\ \sum_i \ell_i p_i &< H(X) + \sum_i p_i = H(X) + 1 \end{aligned}$$

ce qu'il fallait démontrer. ∎

2.3 Codage de Huffman

Le théorème de McMillan et l'inégalité de Kraft nous disent que pour tout code uniquement déchiffrable il existe un code préfixe de même distribution des

longueurs. On peut donc chercher le code optimal (qui minimise la longueur moyenne) parmi les codes préfixes. L'*algorithme de Huffman* permet de trouver un code préfixe optimal.

Algorithme. Soit X une variable aléatoire prenant ses valeurs dans

$$\mathcal{X} = \{x_1, x_2, \ldots, x_m\}$$

de distribution de probabilités $(p_1 \ldots p_m)$. Quitte à réordonner les x_i, on peut supposer que les p_i sont en ordre décroissant, de telle sorte que les p_i les plus faibles sont p_{m-1} et p_m. L'algorithme procède par récurrence en construisant l'arbre binaire à partir de ses feuilles. À x_{m-1} et x_m sont associées deux feuilles issues d'un père commun que l'on peut appeler x'_{m-1}. L'arbre de Huffman est obtenu en

- calculant l'arbre de Huffman associé à la variable aléatoire X' prenant ses valeurs dans l'ensemble $\mathcal{X}' = \{x_1, x_2, \ldots, x_{m-2}, x'_{m-1}\}$ et de distribution de probabilité $(p_1, p_2, \ldots, p_{m-2}, p'_{m-1} = p_{m-1} + p_m)$,
- et en rajoutant deux fils issus de x'_{m-1} qui seront associés aux valeurs x_{m-1} et x_m.

Exemple. Soit X une variable prenant ses valeurs dans $\mathcal{X} = \{x_1, x_2, \ldots, x_6\}$ et de loi $(p_1 = 0.4, p_2 = 0.04, p_3 = 0.14, p_4 = 0.18, p_5 = 0.18, p_6 = 0.06)$. L'arbre obtenu par l'algorithme de Huffman est représenté sur la figure 2. La première étape consiste à joindre les sommets terminaux (feuilles) x_2 et x_6 associés aux probabilités p_2 et p_6 les plus faibles et à créer ainsi un sommet intermédiaire i de l'arbre associé à la probabilité $p_i = p_2 + p_6 = 0.1$. Puis on recommence la procédure sur l'ensemble $\mathcal{X}' = \{x_1, x_3, x_4, x_5, i\}$ pour la loi $(p_1 = 0.4, p_3 = 0.14, p_4 = 0.18, p_5 = 0.18, p_i = 0.1)$. Les probabilités les plus faibles sont p_3 et p_i, on joint donc x_3 et i en un sommet père ii de probabilité $p_{ii} = 0.24$. La procédure se termine par l'arbre de la figure.

Pour démontrer l'optimalité de l'algorithme de Huffman nous utiliserons le lemme suivant.

Lemme 10 *Soit X une variable prenant ses valeurs dans $\mathcal{X} = \{x_1, \ldots, x_m\}$ de loi $(p_1, \ldots, p_m)$ où l'on a ordonné les x_i de telle sorte que la suite des p_i décroisse. Parmi les codages optimaux de X il existe un code préfixe C dont l'arbre encode x_{m-1} et x_m par des feuilles*

- *de profondeur maximale,*
- *ayant un même père.*

Preuve : Si x_m n'est pas associé à un sommet de profondeur maximale, alors il suffit d'échanger x_m avec le x_i associé à un sommet de profondeur maximale et la longueur moyenne du nouvel encodage de X ne peut que diminuer. Ceci démontre le premier point. Par ailleurs, le sommet x_m ne peut pas être l'unique sommet

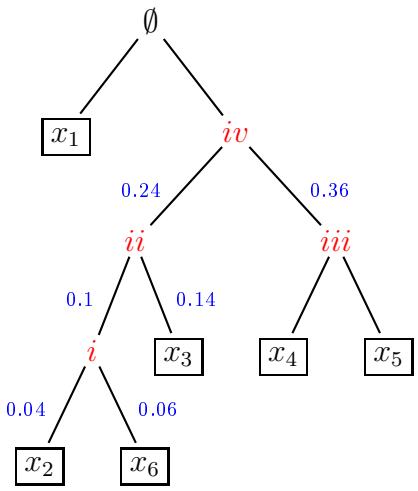

FIG. 2 – Arbre obtenu par application de l'algorithme de Huffman

fils du sommet père π de x_m : sinon on enlève de l'arbre la feuille associée à x_m et on associe x_m au sommet π. Il existe donc un second sommet fils du sommet π, associé à x_j, pour un certain $j \neq m$. Mais $j \neq m$ implique $p_j \geqslant p_{m-1}$. En échangeant x_j et x_{m-1} la longueur moyenne de l'encodage de peut que diminuer : ceci démontre le second point. ∎

Soit X une variable prenant m valeurs, soit $\mathcal{X} = \{x_1, \ldots, x_m\}$. Montrons par récurrence sur m que l'algorithme de Huffman débouche sur un codage optimal de X. Nous supposons $p_1 \geqslant \ldots \geqslant p_{m-1} \geqslant p_m$.

Appelons C_m^* un code préfixe optimal associé à la loi p et vérifiant les propriétés du Lemme 10. Appelons C_{m-1}^* un code préfixe optimal associé à la loi $p' = (p_1, p_2, \ldots, p_{m-2}, p_{m-1} + p_m)$. Appelons C_{m-1} le code préfixe (ou l'arbre) associé à la loi p' obtenu à partir de C_m^* par réduction de Huffman, c'est-à-dire qu'on le construit en prenant l'arbre C_m^*, en supprimant les sommets nommés x_{m-1} et x_m associés aux probabilités p_{m-1} et p_m, et en attribuant la probabilité $p_{m-1} + p_m$ au sommet père de x_{m-1} et x_m. Enfin, appelons C_m le code préfixe (ou l'arbre) obtenu à partir de C_{m-1}^* par l'augmentation inverse, c'est-à-dire que l'on rajoute au sommet associé à la probabilité $p_{m-1}+p_m$ deux feuilles, que l'on nomme x_{m-1} et x_m.

Calculons maintenant la longueur moyenne de C_{m-1}. L'expression de la longueur moyenne de C_m^* étant,

$$\begin{aligned}\overline{\ell}(C_m^*) &= \sum_{i=1}^{m} \ell_i p_i \\ &= \sum_{i=1}^{m-2} \ell_i p_i + \ell_m (p_{m-1} + p_m)\end{aligned}$$

puisque $\ell_{m-1} = \ell_m$ d'après le Lemme 10, on obtient

$$\overline{\ell}(C_{m-1}) = \sum_{i=1}^{m-2} \ell_i p_i + (\ell_m - 1)(p_{m-1} + p_m).$$

Autrement dit,

$$\overline{\ell}(C_{m-1}) = \overline{\ell}(C_m^*) - p_{m-1} - p_m.$$

Par un argument similaire on a :

$$\overline{\ell}(C_m) = \overline{\ell}(C_{m-1}^*) + p_{m-1} + p_m.$$

En additionnant ces deux dernières égalités on obtient :

$$\overline{\ell}(C_{m-1}) + \overline{\ell}(C_m) = \overline{\ell}(C_m^*) + \overline{\ell}(C_{m-1}^*)$$

ou encore :

$$\overline{\ell}(C_{m-1}) - \overline{\ell}(C_{m-1}^*) = \overline{\ell}(C_m^*) - \overline{\ell}(C_m).$$

Mais d'après l'optimalité de C_m^* et C_{m-1}^* le terme gauche doit être positif et le terme de droite doit être négatif. Les deux termes de l'égalité ne peuvent donc être que nuls, et on en déduit que les deux codes préfixes C_m et C_{m-1} sont optimaux pour les lois p et p' respectivement. On en déduit par récurrence sur m que les réductions de Huffman successives mènent à un code préfixe optimal.

3 Canaux discrets sans mémoire, capacité

Un *canal discret sans mémoire* est un modèle simple d'un canal de transmission qui prend en entrée des n-uples aléatoires $X^n = (X_1, \ldots, X_n)$ de variables X_i prenant leurs valeurs dans l'alphabet fini (discret) $\mathcal{X}$ et qui sort des n-uples aléatoires $Y^n = (Y_1, \ldots Y_n)$ où chaque Y_i prend ses valeurs dans l'alphabet $\mathcal{Y}$.

On fait l'hypothèse que les probabilités $P(Y_i = y | X_i = x)$ sont invariantes dans le temps, c'est-à-dire ne dépendent pas de l'indice i, et que les lois conditionnelles $Y_i|X_i$ sont indépendantes, c'est-à-dire que l'on a

$$P(Y^n = y^n | X^n = x^n) = \prod_{i=1}^{n} P(Y_i = y_i | X_i = x_i),$$

c'est le caractère *sans mémoire.*

Un canal discret sans mémoire est donc caractérisé par ses *probabilités de transition.* Il est commode de le représenter par un diagramme : les figures 3 et 4 illustrent deux exemples classiques et utiles, le *canal binaire symétrique* et le *canal à effacements.*

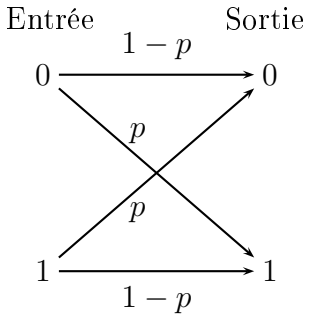

FIG. 3 – Le canal binaire symétrique

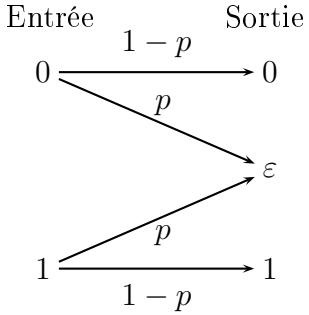

FIG. 4 – Le canal binaire à effacements

Quel est le maximum d'information que l'on peut faire passer sur le canal ? L'émetteur a le choix de la loi du n-uple X^n. S'il souhaite que le recepteur puisse reconstituer X^n à partir du n-uple reçu Y^n, il faut que l'entropie conditionnelle $H(X^n|Y^n)$ soit nulle ou négligeable. Dans ce cas la quantité que l'on souhaite optimiser, soit $H(X^n)$, sous la condition $H(X^n|Y^n) \approx 0$, est égale à l'information mutuelle

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

Posons :

$$C^{(n)} = \frac{1}{n} \max_{p(X^n)} I(X^n, Y^n).$$

Écrivons $I(X^n, Y^n) = H(Y^n) - H(Y^n|X^n)$. Par l'indépendance des lois conditionnelles $(Y_i|X_i)$ on a

$$H(Y^n|X^n) = \sum_{i=1}^{n} H(Y_i|X_i).$$

Donc

$$\begin{aligned} I(X^n, Y^n) &= \sum_{i-1}^{n} H(Y_i|Y_1, \ldots, Y_{i-1}) - \sum_{i=1}^{n} H(Y_i|X_i) \\ &\leqslant \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) \\ &\leqslant n \cdot C^{(1)}. \end{aligned}$$

On a donc

$$C^{(n)} \leqslant C^{(1)}.$$

La quantité $C^{(1)}$ égale à

$$C = \max_{p(X)} I(X, Y)$$

mérite donc d'être étudiée est on l'appelle *capacité* du canal. La discussion précédente prouve qu'il s'agit d'un majorant de la quantité d'information fiable par symbole qu'il est possible de faire transiter par le canal. Le théorème de Shannon énoncé plus loin affirme qu'il est effectivement possible de véhiculer de manière fiable une quantité d'information par symbole arbitrairement proche de la capacité, ceci pour tout canal discret sans mémoire.

Exemples de calculs de capacité

Le canal binaire à effacements.

Considérons le canal binaire à effacements de la figure 4. Écrivons :

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

On a :

$$\begin{aligned} H(X|Y) &= \sum_{y} P(Y = y) H(X|Y = y) \\ &= P(Y = \varepsilon) H(X|Y = \varepsilon). \end{aligned}$$

Or

$$\begin{aligned} P(Y = \varepsilon) &= P(X = 0, Y = \varepsilon) + P(X = 1, Y = \varepsilon) \\ &= P(X = 0) P(Y = \varepsilon|X = 0) + P(X = 1) P(Y = \varepsilon|X = 1) \\ &= P(X = 0)p + P(X = 1)p = (P(X = 0) + P(X = 1))p \\ &= p. \end{aligned}$$

Par ailleurs, pour $x = 0, 1$ on a :

$$\begin{aligned} P(X = x|Y = \varepsilon) &= \frac{P(X = x, Y = \varepsilon)}{P(Y = \varepsilon)} \\ &= \frac{P(X = x)P(Y = \varepsilon|X = x)}{P(Y = \varepsilon)} \\ &= P(X = x). \end{aligned}$$

Ainsi $H(X|Y = \varepsilon) = H(X)$ et l'on a :

$$I(X, Y) = H(X) - pH(X) = H(X)(1 - p).$$

La loi de X qui maximise cette quantité est celle qui maximise $H(X)$, soit la loi uniforme pour laquelle on a $H(X) = 1$, et donc :

$$C = 1 - p.$$

Le canal binaire symétrique.

Considérons le canal binaire à effacements de la figure 3. Écrivons cette fois :

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

On a :

$$H(Y|X) = P(X = 0)H(Y|X = 0) + P(X = 1)H(Y|X = 1)$$

et l'on constate que

$$H(Y|X = 0) = H(Y|X = 1) = h(p)$$

où $h(p)$, fonction du paramètre p, désigne l'entropie d'une loi de Bernoulli $(p, 1 - p)$. On a donc :

$$I(X, Y) = H(Y) - h(p).$$

Par ailleurs il est ainsi de constater que lorsque la loi de X est uniforme, alors la loi de Y est uniforme aussi et maximise $I(X, Y)$. On a donc :

$$C = 1 - h(p).$$

Le théorème de Shannon

Considérons un ensemble de messages $\mathcal{M} = \{0, 1, \ldots, M\}$. Un système de communication est modélisé ainsi : une fonction $f : \mathcal{M} \to \mathcal{X}^n$ transforme le message en un n-uple de $\mathcal{X}^n$ qui est envoyé sur un canal discret sans mémoire. La fonction

f est la fonction *d'encodage*. Une fonction $g : \mathcal{Y}^n \to \mathcal{M}$ transforme le n-uple reçu en un message de $\mathcal{M}$, c'est la fonction de *décodage*. Un *code* $C \subset \mathcal{X}^n$ est un sous-ensemble de $\mathcal{X}^n$ qui peut être défini comme l'image d'une fonction d'encodage f.

On définit la probabilité conditionnelle λ_i par :

$$\lambda_i = P(g(Y^n) = i \,|\, X^n = f(i)).$$

On définit de deux manières la probabilité d'une erreur de décodage, la probabilité *maximale* d'une erreur de décodage vaut :

$$\lambda^n = \max_{i \in \mathcal{M}} \lambda_i$$

et la probabilité *moyenne* d'une erreur de décodage est :

$$P_e^n = \frac{1}{M} \sum_{i=1}^{M} \lambda_i.$$

On a clairement $P_e^n \leqslant \lambda^n$.

Le *rendement* d'un code C de $\mathcal{X}^n$ de cardinal M est :

$$R = R(C) = \frac{1}{n} \log_2 M.$$

Théorème 11 (Shannon) *Pour tout canal discret sans mémoire de capacité C, et pour tout $R < C$, il existe une suite (C_n) de codes où $C_n \subset \mathcal{X}^n$ est de rendement $\geqslant R$ et pour laquelle $\lambda^n \to 0$ quand $n \to \infty$. Réciproquement, si $\lambda^n \to 0$ pour une suite (C_n) de codes, alors* $\limsup R(C_n) \leqslant C$.