Mémento de théorie de l'information

Gilles Zémor

7 décembre 2015

0 Rappels de probabilités

Espaces probabilisés. Un *espace probabilisé* (Ω, P) est un ensemble Ω muni d'une mesure de probabilité P qui est, lorsque Ω est fini, une application

$$\mathcal{P}(\Omega) \to [0,1]$$

telle que :

- $P(\Omega) = 1$
- si $A, B \subset \Omega$ avec $A \cap B = \emptyset$, alors $P(A \cup B) = P(A) + P(B)$.

Les parties de Ω sont appelés *événements*. L'événement *complémentaire* de A, noté $\overline{A}$ ou A^c, est l'événement $\Omega \setminus A$. Sa probabilité est $1 - P(A)$. Lorsque $\omega \in \Omega$, on écrit $P(\omega)$ plutôt que $P(\{\omega\})$. On dit que la probabilité P est la *probabilité uniforme* si pour tout $\omega \in \Omega$, $P(\omega) = 1/|\Omega|$.

Probabilités conditionnelles, événements indépendants. Pour deux événements quelconques A, B, on définit la *probabilité de A sachant B* :

$$P(A \,|\, B) = P(A \cap B)/P(B).$$

Si $P(A \,|\, B) = P(A)$, ou encore $P(A \cap B) = P(A)P(B)$, on dit que les événements A et B sont *indépendants*.

La fonction $P(\cdot \,|\, B)$ munit B d'une structure d'espace probabilisé. La *formule de Bayes* exprime la probabilité $P(A)$ en fonction des probabilités $P(A|B)$ et $P(A|B^c)$:

$$P(A) = P(A|B)P(B) + P(A|B^c)(1 - P(B)).$$

Exemple (Exercice) : un étudiant répond à un questionnaire à choix multiples. On propose m réponses à chaque question. On suppose que lorsque l'étudiant ne

connait pas la réponse il coche une case au hasard uniformément. Si on observe une proportion x de bonnes réponses, on souhaite évaluer la proportion y de questions auxquelles l'étudiant connait effectivement la réponse.

On appelle C l'événement «l'étudiant connait la réponse» et R l'événement «l'étudiant répond correctement». Il s'agit d'évaluer $y = P(C)$ en fonction de $x = P(R)$ et de m.

Variables aléatoires, loi. Une *variable aléatoire* est une application :

$$X : \Omega \to \mathbb{R}^n.$$

Lorsque $n = 1$ on parle de variable (aléatoire) réelle. Pour $E \subset \mathbb{R}^n$, on note $P(X \in E) = P(X^{-1}(E))$ et $P(X = x) = P(X^{-1}(x))$. Si on note $\mathcal{X}$ l'image de X, on appelle la *loi* de X la donnée des $P(x)$ lorsque x décrit $\mathcal{X}$. On dit que X suit la loi uniforme si $P(X = x) = 1/|\mathcal{X}|$ pour tout $x \in \mathcal{X}$.

Une variable est dite de *Bernoulli* si $\mathcal{X} = \{0, 1\}$. On parle également de *variable indicatrice* de l'événement $A \subset \Omega$, et on la note $\mathbf{1}_A$, la variable de Bernoulli définie par $\mathbf{1}_A(\omega) = 1$ si et seulement si $\omega \in A$.

Deux variables $X, Y : \Omega \to \mathcal{X}$ sont dites *indépendantes* si pour tous $x, y \in \mathcal{X}$,

$$P(X = x,\, Y = y) = P(X = x)P(Y = y)$$

où $\{X = x, Y = y\}$ désigne l'intersection des événements $\{X = x\}$ et $\{Y = y\}$.

De même, si $X_1, X_2, \ldots, X_n$ sont n variables $\Omega \to \mathcal{X}$, on dit qu'elles sont indépendantes (dans leur ensemble) si pour tous $x_1, x_2, \ldots, x_n$, on a :

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = P(X_1 = x_1)P(X_2 = x_2) \cdots P(X_n = x_n).$$

Attention à ne pas confondre l'indépendance deux à deux de n variables (X_i est indépendant de X_j pour tous i, j, $i \neq j$) avec l'indépendance dans leur ensemble : si X_1 et X_2 sont deux variables indépendantes, alors les trois variables $X_1, X_2, X_3 = X_1 + X_2$ sont indépendantes deux à deux mais pas dans leur ensemble.

Espérance. L'*espérance* d'une variable aléatoire X, notée $\mathbf{E}[X]$ est la quantité :

$$\begin{aligned}\mathbf{E}[X] &= \sum_{\omega \in \Omega} X(\omega)P(\omega) \\ &= \sum_{x \in \mathcal{X}} xP(X = x).\end{aligned}$$

On remarque que lorsque X est une variable de Bernoulli $\mathbf{E}[X] = P(X = 1)$.

Il vient directement de la définition de l'espérance :

Théorème 1 *Si X, Y sont deux variables aléatoires,* $\mathbf{E}[X+Y] = \mathbf{E}[X] + \mathbf{E}[Y]$. *Par ailleurs, si λ est une constante réelle on a* $\mathbf{E}[\lambda X] = \lambda \mathbf{E}[X]$.

L'additivité de l'espérance la rend plus facile à calculer que des probabilités. La méthode-type pour calculer l'espérance d'une variable X à valeurs dans $\mathbb{N}$ (variable aléatoire de comptage) est de décomposer X en somme de variables aléatoires de Bernoulli.

Exemple : calcul du nombre moyen de points fixes d'une permutation aléatoire uniforme. Soit Ω l'ensemble des permutations sur l'ensemble à n éléments $[n] = \{1, 2, \ldots, n\}$, muni de la probabilité uniforme. Soit X le nombre de points fixes de la permutation aléatoire, c'est-à-dire :

$$X(\omega) = \#\{i \in [n], \omega(i) = i\}.$$

On souhaite calculer $\mathbf{E}[X]$. On définit X_i comme la variable de Bernoulli telle que $X_i(\omega) = 1$ si et seulement si i est un point fixe de ω. On a alors :

$$X = X_1 + X_2 + \cdots + X_n.$$

Il nous suffit donc de trouver $\mathbf{E}[X_i] = P(X_i = 1)$ et d'appliquer l'additivité de l'espérance. Or il vient facilement que $P(X_i = 1) = (n-1)!/n! = 1/n$. D'où :

$$\mathbf{E}[X] = 1.$$

Variance. La *covariance* de deux variables X, Y est définie par :

$$\mathrm{cov}(X, Y) = \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])] = \mathbf{E}[XY] - \mathbf{E}[X]\,\mathbf{E}[Y].$$

La *variance* d'une variable X est définie par :

$$\mathrm{var}(X) = \mathrm{cov}(X, X) = \mathbf{E}\left[(X - \mathbf{E}[X])^2\right] = \mathbf{E}\left[X^2\right] - \mathbf{E}[X]^2.$$

L'*écart-type* de la variable X est la racine carrée de sa variance : $\sigma(X) = \sqrt{\mathrm{var}(X)}$. Le théorème suivant, appelé inégalité de Tchebitchev (Chebychov), justifie l'appelation d'écart-type en affirmant qu'il est très improbable qu'une variable X s'écarte de $\mathbf{E}[X]$ de beaucoup plus que $\sigma(X)$:

Théorème 2

$$P(|X - \mathbf{E}[X]| \geqslant \varepsilon) \leqslant \frac{\mathrm{var}(X)}{\varepsilon^2}.$$

Preuve : Il est très fréquent d'essayer de transformer le calcul d'une probabilité en le calcul d'une espérance, plus facile.

$$P[|X - \mathbf{E}[X]| > \varepsilon] = \mathbf{E}\left[\mathbf{1}_{|X-\mathbf{E}[X]|>\varepsilon}\right] \leqslant \mathbf{E}\left[\mathbf{1}_{|X-\mathbf{E}[X]|>\varepsilon} \frac{|X - \mathbf{E}[X]|^2}{\varepsilon^2}\right]$$
$$\leqslant \mathbf{E}\left[\frac{|X - \mathbf{E}[X]|^2}{\varepsilon^2}\right] = \frac{\mathrm{var}(X)}{\varepsilon^2}.$$

∎

Lemme 3 *Si X et Y sont deux variables réelles indépendantes alors*

$$\mathbf{E}[XY] = \mathbf{E}[X]\,\mathbf{E}[Y].$$

Preuve :

$$\begin{aligned}
\mathbf{E}[X]\,\mathbf{E}[Y] &= \sum_x xP(X=x)\sum_y yP(Y=y) \\
&= \sum_{x,y} xyP(X=x)P(Y=y) \text{ et par indépendance de } X, Y, \\
&= \sum_{x,y} xyP(X=x, Y=y) \\
&= \sum_z z \sum_{x,y,xy=z} P(X=x, Y=y) \\
&= \sum_z zP(XY=z)
\end{aligned}$$

car $\{XY=z\} = \bigcup_{x,y,z=xy}\{X=x, Y=y\}$ et cette réunion est disjointe. ∎

Corollaire 4 *Si $X_1, X_2, \ldots, X_n$ sont n variables indépendantes deux à deux, alors*

$$\mathrm{var}(X_1 + X_2 + \cdots + X_n) = \mathrm{var}(X_1) + \mathrm{var}(X_2) + \cdots + \mathrm{var}(X_n).$$

Preuve :

$$\begin{aligned}
\mathrm{var}(X_1 + \cdots + X_n) &= \mathbf{E}\left[(X_1 + X_2 + \cdots + X_n)^2\right] - \mathbf{E}[X_1 + X_2 + \cdots X_n]^2 \\
&= \mathbf{E}\left[(X_1 + \cdots + X_n)^2\right] - (\mathbf{E}[X_1] + \cdots + \mathbf{E}[X_n])^2 \\
&= \mathbf{E}\left[X_1^2 + \cdots + X_n^2 + 2\sum_{i<j} X_iX_j\right] \\
&\quad - (\mathbf{E}[X_1]^2 + \cdots + \mathbf{E}\left[X_n^2\right] + 2\sum_{i<j} \mathbf{E}[X_i]\,\mathbf{E}[X_j]) \\
&= \sum_{i=1}^n (\mathbf{E}\left[X_i^2\right] - \mathbf{E}[X_i]^2) + 2\sum_{i<j} (\mathbf{E}[X_iX_j] - \mathbf{E}[X_i]\,\mathbf{E}[X_j]) \\
&= \sum_{i=1}^n (\mathbf{E}\left[X_i^2\right] - \mathbf{E}[X_i]^2) \\
&= \sum_{i=1}^n \mathrm{var}(X_i). \qquad \blacksquare
\end{aligned}$$

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» (ou divergence) 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 5 *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 6 *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 7 *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 8 *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 8 on a

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

Proposition 9 *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

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

∎

Entropie et coefficients binomiaux

Dans le cas d'une loi de Bernoulli $p = (\lambda, 1-\lambda)$, on notera souvent $h(\lambda) = h_2(\lambda)$ l'entropie $H(p)$ de la loi p. La fonction d'une variable réelle

$$\begin{array}{rccl} h: & [0,1] & \to & [0,1] \\ & x & \mapsto & h(x) = x \log_2 \frac{1}{x} + (1-x) \log_2 \frac{1}{1-x} \end{array}$$

est appelée *entropie binaire*. Cette fonction est très utile pour évaluer le comportement asymptotique des coefficients binomiaux :

Lemme 10 *Pour tout* $\lambda \leqslant 1/2$ *on a la majoration*

$$\sum_{i \leqslant \lambda n} \binom{n}{i} \leqslant 2^{nh(\lambda)}. \tag{1}$$

Pour $\lambda \geqslant 1/2$ *on a :*

$$\sum_{i \geqslant \lambda n} \binom{n}{i} \leqslant 2^{nh(\lambda)}. \tag{2}$$

Preuve : Notons que (1) se déduit de (2) par les égalités $\binom{n}{i} = \binom{n}{n-i}$. Pour démontrer (2) écrivons, pour tout $r \geqslant 0$,

$$(1+2^r)^n = \sum_{i=0}^{n} 2^{ir}\binom{n}{i} \geqslant \sum_{i \geqslant \lambda n} 2^{ir}\binom{n}{i} \geqslant 2^{\lambda n r} \sum_{i \geqslant \lambda n}\binom{n}{i},$$

ce qui donne

$$\sum_{i \geqslant \lambda n}\binom{n}{i} \leqslant 2^{-\lambda n r}(1+2^r)^n.$$

Posons maintenant $r = \log_2 \frac{\lambda}{1-\lambda} \geqslant 0$ pour tout $\lambda \geqslant 1/2$. On obtient

$$\sum_{i \geqslant \lambda n}\binom{n}{i} \leqslant \left(\frac{1-\lambda}{\lambda}\right)^{\lambda n}\left(1+\frac{\lambda}{1-\lambda}\right)^{n} = \left(\frac{1}{\lambda}\right)^{\lambda n}\left(\frac{1}{1-\lambda}\right)^{n-\lambda n} = 2^{n h_2(\lambda)}.$$

∎

Le lemme 10 nous donne en particulier la majoration $\binom{n}{\lambda n} \leqslant 2^{n h_2(\lambda)}$. On peut obtenir une majoration plus fine grâce à la formule de Stirling, dont une des formes s'énonce :

$$\sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{1}{12n+1}} \leqslant n! \leqslant \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{1}{12n}}.$$

Appliquée au coefficient binomial

$$\binom{n}{\lambda n} = \frac{n!}{(\lambda n)!(n-\lambda n)!},$$

on obtient :

$$\binom{n}{\lambda n} \leqslant \frac{1}{\sqrt{2\pi\lambda(1-\lambda)n}} 2^{n h(\lambda)}. \tag{3}$$

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 $\log_2 m$.

- $$H(X|Y) = \sum_{x,y} P(X=x, Y=y) \log_2 \frac{1}{P(X=x|Y=y)}$$
$$= \sum_y P(Y=y) \sum_x P(X=x|Y=y) \log_2 \frac{1}{P(X=x|Y=y)}.$$
- $H(X,Y) = H(Y) + H(X|Y)$.
- $H(X|Y) \leqslant H(X)$.
- $$\begin{aligned} I(X,Y) &= H(X) + H(Y) - H(X,Y) \\ &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X). \end{aligned}$$
- $I(X,Y) \geqslant 0$.
- Pour démontrer une inégalité informationnelle, penser à la transformer en une expression du type $\mathrm{D}(p \,||\, q) \geqslant 0$.

2 Paris et distance de Kullback

On considère une course à m chevaux numérotés $1, 2, \ldots, m$. Le cheval i a une probabilité p_i de gagner la course. Une mise de 1€ sur le cheval i rapporte au parieur une somme de g_i euros. Supposons que le parieur dispose d'une somme de S_0 euros et qu'il décide de la répartir suivant une loi $\pi = (\pi_1, \pi_2, \ldots, \pi_m)$, c'est-à-dire qu'il mise $S_0\pi_k$ sur le cheval k. Soit X le numéro du cheval gagnant : le parieur dispose donc, après la course, de la somme

$$S_1 = S_0 \pi_X g_X.$$

Quelle est la meilleure stratégie π pour parier ? Si on calcule l'espérance de S_1 on obtient :

$$\mathbf{E}\,[S_1] = S_0 \sum_{k=1}^{m} p_k \pi_k g_k.$$

On peut être tenté de maximiser cette quantité, mais ceci veut dire que le parieur mise tout ($\pi_k = 1$) sur le cheval k qui maximise $p_k g_k$. Si le cheval k ne gagne pas, le parieur a tout perdu. Une autre stratégie apparait si on suppose que l'on puisse jouer pendant n courses successives, indépendantes et identiquement distribuées. Appelons $X_1, X_2, \ldots, X_n$ les numéros des chevaux gagnants des courses

$1, 2, \ldots, n$. Au bout des n courses le parieur dispose de la somme

$$S_n = S_0 \prod_{i=1}^{n} D(X_i)$$

où l'on a noté $D(X) = \pi_X g_X$. On peut écrire :

$$\frac{1}{n} \log S_n = \frac{1}{n} \log S_0 + \frac{1}{n} \sum_{i=1}^{n} \log D(X_i)$$

et la loi des grands nombres nous dit que $\frac{1}{n} \log S_n$ converge vers la quantité :

$$W(p, \pi) = \mathbf{E} \left[\log D(X) \right] = \sum_{k=1}^{m} p_k \log \pi_k g_k.$$

En d'autres termes, après n courses le parieur dispose d'une somme

$$S_n \approx 2^{nW(p,\pi)}$$

et l'on voit que le problème du parieur est donc de choisir la loi π afin de

$$\boxed{\text{Maximiser la quantité } W(p, \pi)}$$

La solution optimale est donnée par :

Théorème 11 *La loi $\pi = p$ maximise la valeur de $W(p, \pi)$ et l'on a*

$$W^\star(p) = W(p, p) = \sum_{k=1}^{m} p_k \log g_k - H(p).$$

Preuve :

$$\begin{aligned} W(p, \pi) &= \sum_{k=1}^{m} p_k \log \pi_k g_k \\ &= \sum_{k=1}^{m} p_k \log g_k p_k \frac{\pi_k}{p_k} \\ &= \sum_{k=1}^{m} p_k \log g_k - H(p) - \mathrm{D}(p \,||\, \pi). \end{aligned}$$

Or $\mathrm{D}(p \,||\, \pi) \geqslant 0$ et $\mathrm{D}(p \,||\, \pi) = 0$ si et seulement si $\pi = p$, d'où le résultat. ∎

En pratique on ne connait pas forcément la loi p et on ne sait pas choisir $\pi = p$ exactement. Considérons le cas particulier où les gains (g_k) sont distribués équitablement, c'est-à-dire que l'on a :

$$\sum_{k=1}^{m} \frac{1}{g_k} = 1.$$

Posons $r = (r_k)$ et $r_k = 1/g_k$, on peut écrire :

$$\begin{aligned} W(p,\pi) &= \sum_{k=1}^{m} p_k \log \pi_k g_k \\ &= \sum_{k=1}^{m} p_k \log \frac{p_k}{r_k}\frac{\pi_k}{p_k} \\ &= \mathrm{D}(p \,||\, r) - \mathrm{D}(p \,||\, \pi). \end{aligned}$$

Ceci s'interprète ainsi : le parieur réussit à obtenir un taux de croissance $W(p,\pi)$ positif lorsqu'il son estimation π de la loi p est meilleure que l'estimation r de p faite par l'organisateur de la course.

3 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

$$\overline{\ell}(\mathbf{c}) = \mathbf{E}\left[\ell(\mathbf{c}(X))\right] = \sum_{x \in \mathcal{X}} P(X = x)\ell(\mathbf{c}(x)).$$

Exemple. Soit $\mathcal{X} = \{1, 2, 3, 4\}$ et soit X à valeurs dans $\mathcal{X}$ de loi $p_1 = 1/2, p_2 = 1/4, p_3 = 1/8, p_4 = 1/8$ 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 :

$$\overline{\ell} = \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{8} \cdot 2 + \frac{1}{8} \cdot 3 = \frac{14}{8}.$$

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

3.1 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 $\{0, 01\}$, ou encore $\{0, 11, 010\}$.

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.

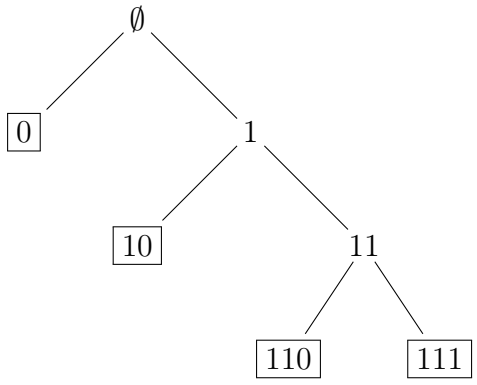

FIGURE 1 – L'arbre associé au code préfixe $\{0, 10, 110, 111\}$

Proposition 12 (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 13 (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. ∎

3.2 Longueur moyenne et entropie

Proposition 14 *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 5) 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 15 *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. ∎

3.3 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 16 *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

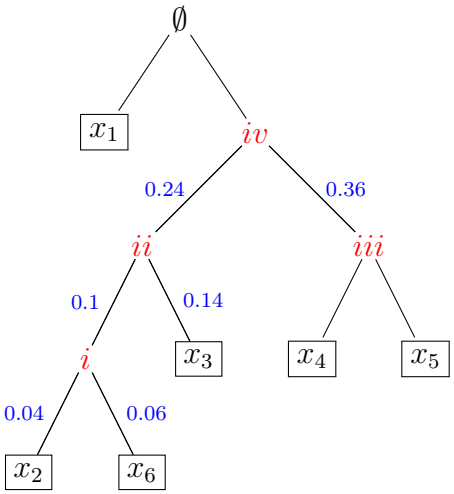

FIGURE 2 – Arbre obtenu par application de l'algorithme de Huffman

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 16. 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 16, 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.

4 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 les deux hypothèses :

– *caractère sans mémoire.*

$$P(Y_n = y_n | X^n = x^n, Y^{n-1} = y^{n-1}) = P(Y_n = y_n | X_n = x_n).$$

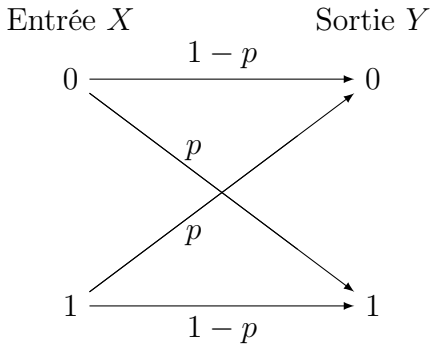

FIGURE 3 – Le canal binaire symétrique

– Le canal est utilisé *sans rétroaction* (feedback). Le n-ième symbole d'entrée X_n ne dépend pas des sorties précédentes :

$$P(X_n = x_n | X^{n-1} = x^{n-1}, Y^{n-1} = y^{n-1}) = P(X_n = x_n | X^{n-1} = x^{n-1}).$$

Ces deux hypothèses impliquent la décomposition

$$P(Y^n = y^n | X^n = x^n) = \prod_{i=1}^{n} P(Y_i = y_i | X_i = x_i). \tag{4}$$

L'équation (4) se démontre par récurrence : on peut écrire

$$\begin{aligned} &P(Y^n = y^n | X^n = x^n) \\ &\quad = P(Y_n = y_n | X^n = x^n, Y^{n-1} = y^{n-1}) P(Y^{n-1} = y^{n-1} | X^n = x^n) \\ &\quad = P(Y_n = y_n | X_n = x_n) P(Y^{n-1} = y^{n-1} | X^n = x^n) \end{aligned} \tag{5}$$

d'après le caractère sans mémoire. Par ailleurs

$$\begin{aligned} P(Y^{n-1} = y^{n-1} | X^n = x^n) &= \frac{P(Y^{n-1} = y^{n-1}, X^n = x^n)}{P(X^n = x^n)} \\ &= \frac{P(X_n = x_n | Y^{n-1} = y^{n-1}, X^{n-1} = x^{n-1}) P(Y^{n-1} = y^{n-1}, X^{n-1} = x^{n-1})}{P(X^n = x^n)} \\ &= \frac{P(X_n = x_n | X^{n-1} = x^{n-1}) P(Y^{n-1} = y^{n-1}, X^{n-1} = x^{n-1})}{P(X^n = x^n)} \end{aligned}$$

d'après l'hypothèse sans rétroaction. On obtient donc :

$$\begin{aligned} P(Y^{n-1} = y^{n-1} | X^n = x^n) &= \frac{P(Y^{n-1} = y^{n-1}, X^{n-1} = x^{n-1})}{P(X^{n-1} = x^{n-1})} \\ &= P(Y^{n-1} = y^{n-1} | X^{n-1} = x^{n-1}). \end{aligned}$$

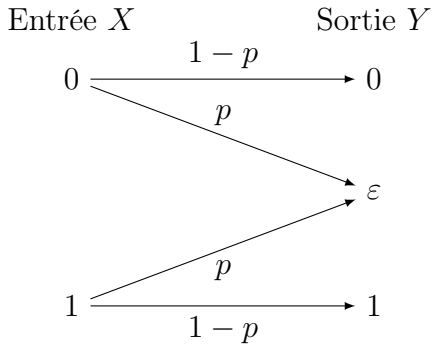

FIGURE 4 – Le canal binaire à effacements

On déduit donc de (5)

$$P(Y^n = y^n | X^n = x^n) = P(Y_n = y_n | X_n = x_n) P(Y^{n-1} = y^{n-1} | X^{n-1} = x^{n-1})$$

ce qui permet d'en déduire (4) par récurrence.

On fait en général l'hypothèse supplémentaire 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. On les appellent *probabilités de transition* et elles caractérisent le canal qu'il est commode de représenter alors par un diagramme : les figures 3 et 4 illustrent deux exemples classiques et utiles, le *canal binaire symétrique* et le *canal 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 récepteur 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 $I(X^n, Y^n) = H(Y^n) - H(Y^n|X^n)$. La décomposition (4) implique

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

Donc

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

On a donc

$$C^{(n)} \leqslant \max_{i=1..n} I(X_i, Y_i)$$

soit $C^{(n)} \leqslant C$ ou l'on a posé :

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

car Y_i ne dépend que de X_i. La quantité C mérite donc d'être étudiée et 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_m par :

$$\lambda_m = P(g(Y^n) \neq m \,|\, X^n = f(m)).$$

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

$$\lambda^n = \max_{m \in \mathcal{M}} \lambda_m$$

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

$$P_e^n = \frac{1}{M} \sum_{m=1}^{M} \lambda_m.$$

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 17 (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* $\limsup R(C_n) \leqslant C$.

5 Codes linéaires

5.1 Encodage, matrice génératrice

Nous nous intéressons maintenant au cas où l'alphabet $\mathcal{X}$ est de cardinal $q = p^m$ pour p un nombre premier, ce qui veut dire que $\mathcal{X}$ peut être muni d'une structure de code fini $\mathbb{F}_q$. Le code $C \subset \mathbb{F}_q^n$ est dit *linéaire* si c'est un sous-espace vectoriel de $\mathbb{F}_q^n$. On s'intéressera tout particulièrement au cas $q = 2$, auquel cas on parle de code linéaire *binaire* : dans ce cas dire qu'un code $C \subset \mathbb{F}_2^n$ est linéaire veut tout simplement dire qu'il est stable par addition dans $\mathbb{F}_2^n$.

L'ensemble $\mathcal{M}$ des messages est en général identifié à $\{0,1\}^k$, ce qui nous donne $|\mathcal{M}| = M = 2^k$. On appelle *matrice génératrice* du code C une matrice $\mathbf{G}$ dont les

lignes constituent une base de l'espace vectoriel C. Si C est un espace vectoriel de dimension k, une matrice génératrice $\mathbf{G}$ est donc une matrice $k \times n$. Il arrive que l'on continue à appeler «matrice génératrice» de C une matrice $k' \times n$ avec $k' > k$, dont une sous-matrice $k \times n$ est une matrice génératrice au sens strict.

Si $\mathbf{G}$ est une matrice génératrice $k \times n$ du code C, une fonction d'encodage est donnée par :

$$\begin{array}{rcl} f : \mathcal{M} & \to & \mathbb{F}_2^n \\ \mathbf{x} & \mapsto & \mathbf{xG}. \end{array}$$

Une matrice génératrice G est dite sous forme *systématique* si elle s'écrit :

$$\mathbf{G} = [\mathbf{I}_k \mid \mathbf{A}].$$

Dans ce cas la fonction d'encodage est de la forme :

$$(x_1, \ldots, x_k) \mapsto (x_1, \ldots, x_k, x_{k+1}, \ldots, x_n).$$

Les k premiers symboles d'un mot de code sont alors appelés bits ou symboles d'*information* et les $n - k$ symboles supplémentaires *symboles de parité*.

Exemples.

1. *Code de parité.* Ce code est constitué des mots de poids pair dans $\mathbb{F}_2^n$. La dimension du code est $k = \dim C = n - 1$.
2. *Code à répétition.* Ce code a pour matrice génératrice $1 \times n$ la matrice $\mathbf{G} = [1, 1, \cdots, 1]$. On a $k = \dim C = 1$, le code contient deux mots.
3. *Code de parité double.* Considérons l'encodage systématique suivant :

$$\begin{array}{rcl} f : \mathbb{F}_2^4 & \to & \mathbb{F}_2^9 \\ (x_1, \ldots, x_4) & \mapsto & (x_1, \ldots, x_9) \end{array}$$

où les symboles de parité $x_5 \ldots x_9$ sont définis de telle sorte que toutes les lignes et toutes les colonnes du tableau ci-dessous soient de poids pairs.

x_1	x_2	x_5
x_3	x_4	x_6
x_7	x_8	x_9

Ce code est de dimension 4 par construction, et une erreur dans une position arbitraire est corrigible car la position erronnée (i, j) se décèle en repérant

que la ligne i et la colonne j sont de poids impairs. On constate également qu'une configuration arbitraire de trois effacements est toujours corrigible. Une matrice génératrice du code est donnée par :

$$\mathbf{G} = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}.$$

5.2 Distance de Hamming, correction d'erreurs et d'effacements

Définition 18 *La* distance de Hamming *entre deux vecteurs de $\mathbb{F}_q^n$ est le nombre de coordonnées où les deux vecteurs diffèrent.*

Le poids de Hamming *d'un vecteur de $\mathbb{F}_q^n$ est le nombre de ses coordonnées non nulles.*

La distance minimale *d'un code C est la plus petite distance non nulle entre deux mots du code C. Quand le code C est linéaire c'est aussi le plus petit poids $|\mathbf{c}|$ d'un mot non nul $\mathbf{c}$ de C.*

La distance de Hamming $d(\cdot, \cdot)$ est une distance au sens mathématique. Elle vérifie les propriétés $d(\mathbf{x}, \mathbf{x}) = 0$, $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$, et $d(\mathbf{x}, \mathbf{y}) \leqslant d(\mathbf{x}, \mathbf{z}) + d(\mathbf{z}, \mathbf{y})$ pour tous vecteurs $\mathbf{x}, \mathbf{y}, \mathbf{z}$.

Les *paramètres* $[n, k, d]$ d'un code C sont sa longueur i.e. la dimension de l'espace vectoriel ambiant $\mathbb{F}_q^n$, la dimension de C (en tant que $\mathbb{F}_q$-espace vectoriel), et sa distance minimale. Le code C contient q^k mots de code.

Si $\mathbf{c}$ est un mot du code C de distance minimale d, et si l'on a $d(\mathbf{c}, \mathbf{x}) \leqslant t$ avec $t < d/2$ pour un certain mot $\mathbf{x}$, alors pour tout mot de code $\mathbf{c}' \in C$, $\mathbf{c}' \neq \mathbf{c}$, on a $d(\mathbf{c}', \mathbf{x}) > d(\mathbf{c}, \mathbf{x})$. En effet, l'inégalité triangulaire

$$d \leqslant d(\mathbf{c}, \mathbf{c}') \leqslant d(\mathbf{c}, \mathbf{x}) + d(\mathbf{x}, \mathbf{c}')$$

implique que si $d(\mathbf{c}, \mathbf{x}) < d/2$ alors $d(\mathbf{x}, \mathbf{c}') > d/2$.

On en déduit que si un mot $\mathbf{c}$ subit t erreurs avec $t < d/2$ et est transformé en le vecteur $\mathbf{x}$, alors le mot de code le plus proche (pour la distance de Hamming) de $\mathbf{x}$ est le mot de code émis. On peut donc énoncer :

Proposition 19 *Un code de distance minimale d peut corriger n'importe quelle configuration de $\lceil d/2 \rceil - 1$ erreurs.*

Intéressons-nous maintenant au nombre d'effacements corrigibles par un code correcteur de distance minimale d. Appelons *vecteur effacement* le vecteur de $\{0,1\}^n$ dont le support (l'ensemble des indices des coordonnées non nulles) correspond à l'ensemble des coordonnées effacées. Écrivons $\mathbf{x} \subset \mathbf{y}$ pour signifier que le support du vecteur $\mathbf{x}$ est inclus dans le support du vecteur $\mathbf{y}$, auquel cas on dira que le vecteur $\mathbf{y}$ couvre le vecteur $\mathbf{x}$. Nous pouvons énoncer :

Proposition 20 *Si C est un code linéaire, une configuration de positions effacées $E \subset \{1, 2, \ldots, n\}$ est corrigible si et seulement si le vecteur effacement correspondant $\mathbf{v}_E$ ne couvre aucun mot non nul de C, c'est-à-dire si $\mathbf{c} \not\subset \mathbf{v}_E$ pour tout $\mathbf{c} \in C$, $\mathbf{c} \neq 0$.*

Preuve : Dire qu'une configuration d'effacements est non corrigible veut dire qu'il existe au moins deux mots de code distincts qui coïncident en dehors des positions effacées. Mais dans ce cas la différence de ces deux mots est un mot de code couvert par le vecteur effacement. ∎

Corollaire 21 *Un code linéaire de distance minimale d peut corriger n'importe quelle configuration de $d-1$ effacements.*

Remarque. Une configuration de t erreurs avec $t \geqslant d/2$ ou bien de t effacements avec $t \geqslant d$ peut parfaitement être corrigible. Par exemple le code de parité double de paramètres $[9, 4, 4]$ permet de corriger la configuration de cinq effacements :

⋆	⋆	
⋆		
	⋆	⋆

5.3 Preuve du théorème de Shannon pour le canal binaire à effacements

Lemme 22 *Soit M une matrice $k \times (k+x)$ sur $\mathbb{F}_q$ tirée aléatoirement et uniformément : alors*

$$P(\operatorname{rg} M < k) \leqslant \frac{1}{q^x}$$

où $\operatorname{rg} M$ désigne le rang de M sur $\mathbb{F}_q$.

Preuve : On procède par récurrence sur k. Pour $k = 1$ on a clairement que $\operatorname{rg} M < 1$ si et seulement si la matrice M est la ligne identiquement nulle, ce qui arrive avec probabilité $1/q^{1+x} < 1/q^x$. Soit M une matrice $k \times (k+x)$ et supposons la proposition démontrée pour toute matrice $(k-1) \times y$ pour tout y. Soit M_{k-1} la matrice constituée des $k-1$ premières lignes de M. On peut écrire :

$$\begin{aligned} P(\operatorname{rg} M < k) &= P(\operatorname{rg} M_{k-1} < k-1) \\ &\quad + P(\operatorname{rg} M < k \,|\, \operatorname{rg} M_{k-1} = k-1) P(\operatorname{rg} M_{k-1} = k-1) \\ &\leqslant P(\operatorname{rg} M_{k-1} < k-1) + P(\operatorname{rg} M < k \,|\, \operatorname{rg} M_{k-1} = k-1). \end{aligned}$$

Par hypothèse de récurrence on a $P(\operatorname{rg} M_{k-1} < k-1) \leqslant 1/q^{1+x}$ et par ailleurs, la probabilité que la dernière ligne appartienne à un sous-espace donné de dimension au plus $k-1$ vaut :

$$P(\operatorname{rg} M < k \,|\, \operatorname{rg} M_{k-1} = k-1) \leqslant \frac{q^{k-1}}{q^{k+x}} = \frac{1}{q^{1+x}}$$

d'où $P(\operatorname{rg} M < k) \leqslant 1/q^{1+x} + 1/q^{1+x} \leqslant 1/q^x$. ∎

Lemme 23 *Soient $X_1, \ldots X_n$ une suite de variables aléatoires de Bernoulli, indépendantes et de même paramètre α. Soit β, $\alpha < \beta < 1$. Alors,*

$$P(X_1 + \cdots + X_n \geqslant \beta n) \leqslant 2^{-n\mathrm{D}(\beta, 1-\beta \,||\, \alpha, 1-\alpha)}.$$

Preuve : Nous avons :

$$P(X_1 + \cdots + X_n \geqslant \beta n) = \sum_{i \geqslant \beta n} \binom{n}{i} \alpha^i (1-\alpha)^{n-i}.$$

Nous pouvons utiliser la même astuce que pour la démonstration du Lemme 10. Écrivons, pour tout $r \geqslant 0$,

$$(2^r \alpha + 1 - \alpha)^n = \sum_{i=0}^{n} 2^{ri} \binom{n}{i} \alpha^i (1-\alpha)^{n-i} \geqslant 2^{r\beta n} \sum_{i \geqslant \beta n} \binom{n}{i} \alpha^i (1-\alpha)^{n-i}$$

ce qui donne :

$$\sum_{i \geqslant \beta n} \binom{n}{i} \alpha^i (1-\alpha)^{n-i} \leqslant 2^{-r\beta n} (1-\alpha)^n \left(2^r \frac{\alpha}{1-\alpha} + 1 \right)^n$$

En posant

$$r = \log_2 \left(\frac{1-\alpha}{\alpha} \frac{\beta}{1-\beta} \right) \geqslant 0$$

pour $\beta \geqslant \alpha$, on obtient :

$$\sum_{i \geqslant \beta n} \binom{n}{i} \alpha^i (1-\alpha)^{n-i} \leqslant \left(\frac{\alpha}{\beta}\right)^{\beta n} \left(\frac{1-\alpha}{1-\beta}\right)^{n-\beta n}$$

ce qui est exactement la majoration annoncée. ∎

Nous pouvons maintenant énoncer une version du théorème de Shannon pour le canal à effacements.

Théorème 24 *Soit un canal binaire à effacements de probabilité de transition p, et soit $R = 1 - p - \varepsilon$ avec $\varepsilon > 0$ fixés. Soit C le code binaire engendré par une matrice $\mathbf{G}$ à n colonnes et $k = Rn$ lignes, obtenue* aléatoirement *uniformément parmi toutes les matrices $k \times n$. La probabilité P_e, sur à la fois l'action du canal et le choix de la matrice $\mathbf{G}$, que la configuration d'effacements $E \subset \{1, 2, \ldots, n\}$ soit non corrigible tend vers 0 lorsque n tend vers l'infini. De plus, cette convergence est une fonction exponentielle de la longueur n, c'est-à-dire*

$$\liminf \frac{1}{n} \log \frac{1}{P_e} \geqslant f(\varepsilon) > 0.$$

Preuve : Appelons $\mathbf{G}_E$ et $\mathbf{G}_{\overline{E}}$ les sous-matrices de $\mathbf{G}$ constituées des coordonnées effacées et des coordonnées non effacées respectivement. D'après la proposition 20, le caractère non corrigible de la configuration d'effacement E ne dépend pas du mot de code émis : E est non corrigible si et seulement s'il existe un mot de de code non nul de support inclus dans E. Autrement dit, E est non corrigible si et seulement s'il existe une combinaison linéaire non triviale des lignes de $\mathbf{G}_{\overline{E}}$ qui égale zéro, c'est-à-dire si et seulement si $\operatorname{rg} \mathbf{G}_{\overline{E}} < k$.

Le poids de E étant fixé, la probabilité que $\operatorname{rg} \mathbf{G}_{\overline{E}} < k$ est bien estimée par le lemme 22. Appliquons la formule de Bayes pour écrire :

$$\begin{aligned} P[\operatorname{rg} \mathbf{G}_{\overline{E}} < k] \leqslant & P[\operatorname{rg} \mathbf{G}_{\overline{E}} < k \,|\, |E| > (p+\varepsilon/2)n] P[|E| > (p+\varepsilon/2)n] + \\ & \sum_{e \leqslant (p+\varepsilon/2)n} P[\operatorname{rg} \mathbf{G}_{\overline{E}} < k \,|\, |E| = e] P[|E| = e] \end{aligned}$$

$$P[\operatorname{rg} \mathbf{G}_{\overline{E}} < k] \leqslant P[|E| > (p+\varepsilon/2)n] + \max_{e \leqslant (p+\varepsilon/2)n} P[\operatorname{rg} \mathbf{G}_{\overline{E}} < k \,|\, |E| = e].$$

Or,

$$\max_{e \leqslant (p+\varepsilon/2)n} P[\operatorname{rg} \mathbf{G}_{\overline{E}} < k \,|\, |E| = e] = \max_{x \geqslant \varepsilon n/2} P[\operatorname{rg} \mathbf{G}_{\overline{E}} < k \,|\, |\overline{E}| \geqslant Rn + x]$$

ce qui, d'après le lemme 22, est $\leqslant 1/2^{n\varepsilon/2}$. Et d'après le lemme 23,

$$P[|E| > (p+\varepsilon/2)n] \leqslant 1/2^{n \mathrm{D}(p+\varepsilon/2 \,||\, p)}.$$

On en déduit donc :

$$P[\mathrm{rg}\,\mathbf{G}_{\overline{E}} < k] \leqslant \frac{1}{2^{n\varepsilon/2}} + \frac{1}{2^{n\mathrm{D}(p+\varepsilon/2\,||\,p)}} \leqslant 2\frac{1}{2^{n\min(\frac{\varepsilon}{2}\,,\,\mathrm{D}(p+\varepsilon/2\,||\,p))}}$$

ce qui démontre le résultat. ∎

Exposant d'erreur des codes linéaires aléatoires.

L'exposant d'erreur est la quantité $\liminf \frac{1}{n}\log P_e$ définie au théorème 24. On peut l'estimer plus précisément que dans la démonstration du théorème 24 ci-dessus, en écrivant :

$$\begin{aligned}
P[\mathrm{rg}\,\mathbf{G}_{\overline{E}} < k] &= \sum_{i=0}^{n} P[\mathrm{rg}\,\mathbf{G}_{\overline{E}} < k\,|\,|E| = i]P[|E| = i] \\
&\leqslant pnP[\mathrm{rg}\,\mathbf{G}_{\overline{E}} < k\,|\,|E| = pn] \\
&\quad + \sum_{i=pn}^{(p+\varepsilon)n} P[\mathrm{rg}\,\mathbf{G}_{\overline{E}} < k\,|\,|E| = i]P[|E| = i] \\
&\quad + (1-p-\varepsilon)nP[|E| \geqslant (p+\varepsilon)n] \\
&\leqslant n \max_{i=pn}^{(p+\varepsilon)n} P[\mathrm{rg}\,\mathbf{G}_{\overline{E}} < k\,|\,|E| = i]P[|E| = i] \\
&\leqslant n \sup_{0\leqslant\lambda\leqslant\varepsilon} \frac{1}{2^{(\varepsilon-\lambda)n}} \frac{1}{2^{\mathrm{D}(p+\lambda\,||\,p)n}}
\end{aligned}$$

ce qui permet la minoration de l'exposant d'erreur

$$\liminf \frac{1}{n}\log_2 P[\mathrm{rg}\,\mathbf{G}_{\overline{E}} < k] \geqslant \inf_{0\leqslant\lambda\leqslant\varepsilon} \left[\varepsilon - \lambda + \mathrm{D}(p+\lambda\,||\,p)\right].$$

Or, comme la dérivée en $\lambda = 0$ de la fonction $\mathrm{D}(p+\lambda\,||\,p)$ vaut 0, il vient que pour tous les $\varepsilon > 0$ suffisamment petits,

$$\inf_{0\leqslant\lambda\leqslant\varepsilon} \left[\varepsilon - \lambda + \mathrm{D}(p+\lambda\,||\,p)\right] = \mathrm{D}(p+\varepsilon\,||\,p).$$

Par ailleurs, nous pouvons écrire

$$P[\mathrm{rg}\,\mathbf{G}_{\overline{E}} < k] \geqslant P(|\overline{E}| < Rn) = \sum_{i>(p+\varepsilon)n} \binom{n}{i} p^i (1-p)^{n-i} \tag{6}$$

car si $|\overline{E}| < Rn$, alors forcément $\mathrm{rg}\,\mathbf{G}_{\overline{E}} \leqslant |\overline{E}| < Rn$, ceci d'ailleurs, quelle que soit la matrice $\mathbf{G}$ de dimension $k \times n$. La formule de Stirling montre que la somme dans (6) se comporte comme $2^{-n\mathrm{D}(p+\varepsilon\,||\,p)+o(n)}$, et nous avons donc montré que, pour $\varepsilon > 0$ suffisamment petit,

$$\liminf \frac{1}{n}\log\frac{1}{P_e} = \liminf \frac{1}{n}\log_2 P[\mathrm{rg}\,\mathbf{G}_{\overline{E}} < k] = \mathrm{D}(p+\varepsilon\,||\,p).$$

Ceci est remarquable, car l'argument menant à (6) montre que *n'importe quelle* famille de codes linéaires de rendement R proche de la capacité $1-p$ ne peut avoir un meilleur (plus élevé) exposant d'erreur. En ce sens *les codes linéaires aléatoires sont donc optimaux pour le canal à effacements.*

5.4 Preuve du théorème de Shannon pour le canal binaire symétrique

Décodage au maximum de vraisemblance. Lorsqu'on reçoit un vecteur $\mathbf{y}$, le problème du décodage consiste à rechercher le mot de code $\mathbf{c}$ le plus probable, c'est-à-dire qui maximise la probabilité conditionnelle

$$P(X = \mathbf{c} \,|\, Y = \mathbf{y}) \tag{7}$$

où les variables X et Y désignent les vecteurs aléatoires émis et reçus respectivement. Nous pouvons écrire :

$$P(X = \mathbf{c} \,|\, Y = \mathbf{y}) = \frac{P(X = \mathbf{c}, Y = \mathbf{y})}{P(Y = \mathbf{y})} = \frac{P(X = \mathbf{c})}{P(Y = \mathbf{y})} P(Y = \mathbf{y} \,|\, X = \mathbf{c}).$$

Si nous faisons l'hypothèse que le mot de code émis X suit une loi uniforme sur le code C, la quantité $P(X = \mathbf{c})/P(Y = \mathbf{y})$ est une constante pour tout $\mathbf{y}$ reçu donné, et maximiser la probabilité conditionnelle (7) équivaut à maximiser la probabilité conditionnelle $P(Y = \mathbf{y} \,|\, X = \mathbf{c})$. Or, si p est la probabilité de transition du canal binaire symétrique, nous avons

$$P(Y = \mathbf{y} \,|\, X = \mathbf{c}) = p^{|\mathbf{y}+\mathbf{c}|}(1-p)^{n-|\mathbf{y}+\mathbf{c}|}.$$

Comme nous supposons $p \leqslant 1/2$, cette quantité est maximale lorsque $|\mathbf{y} + \mathbf{c}|$ est minimum. Nous avons donc démontré :

Proposition 25 *Le mot de code* $\mathbf{c}$ *le plus vraisemblable est le mot de code le plus proche, au sens de la distance de Hamming, du mot reçu* $\mathbf{y}$*.*

Remarque. Il n'est pas impossible qu'il n'y ait pas un unique mot de code le plus proche du vecteur reçu, mais qu'il y en ait plusieurs, donc également vraisemblables.

Appelons $\mathbf{e} \in \{0,1\}^n$ le vecteur erreur, de telle sorte que si le mot de code $\mathbf{c}$ est émis, le mot reçu est $\mathbf{y} = \mathbf{c} + \mathbf{e}$. Conformément à l'analyse ci-dessus, nous dirons qu'il n'y a pas d'erreur de décodage s'il n'existe pas de mot de code $\mathbf{c}' \neq \mathbf{c}$ tel que

$$d(\mathbf{c}', \mathbf{y}) \leqslant d(\mathbf{c}, \mathbf{y}) \tag{8}$$

Comme $d(\mathbf{c}, \mathbf{y}) = d(\mathbf{c}, \mathbf{c} + \mathbf{e}) = d(\mathbf{0}, \mathbf{e})$ et $d(\mathbf{c}', \mathbf{c} + \mathbf{e}) = d(\mathbf{c} + \mathbf{c}', \mathbf{e})$, la condition (8) équivaut à la non-existence d'un mot de code $\mathbf{x} \neq \mathbf{0}$ tel que

$$d(\mathbf{x}, \mathbf{e}) \leqslant d(\mathbf{0}, \mathbf{e}). \tag{9}$$

Soulignons que cette condition ne dépend pas du choix particulier du mot de code émis, mais ne dépend que du vecteur erreur $\mathbf{e}$, et de la nature du code C tout entier.

Notons $B(\mathbf{z}, t)$ la boule de Hamming centrée en $\mathbf{z}$ et de rayon t. Notons $V(t)$ le *volume* d'une telle boule, de telle sorte que nous avons

$$V(t) = 1 + \binom{n}{1} + \binom{n}{2} + \cdots + \binom{n}{t}.$$

Notons que la condition (9) se reformule en :

$$C \cap B(\mathbf{e}, |\mathbf{e}|) = \{\mathbf{0}\}. \tag{10}$$

Nous adoptons la même stratégie que pour le canal à effacements, c'est-à-dire que nous allons étudier la probabilité de l'événement (10) lorsque le code C est choisi aléatoirement. Pour cela nous modifions légèrement notre manière d'engendrer une matrice aléatoire $\mathbf{G}$, génératrice du code C. Nous prenons $\mathbf{G}$ sous la forme

$$\mathbf{G} = [\mathbf{I}_k \mid \mathbf{A}] \tag{11}$$

où $\mathbf{A}$ est choisie uniformément dans l'espace des matrices $\{0,1\}^{k\times(n-k)}$, ce qui forme une matrice $\mathbf{G}$ de dimension $k \times n$.

Lemme 26 *Soit $\mathbf{x} \in \{0,1\}^n$, $\mathbf{x} \neq \mathbf{0}$. La probabilité que $\mathbf{x}$ soit un mot de code est au plus $P(\mathbf{x} \in C) \leqslant 1/2^{n-k}$.*

Preuve : Les k premières coordonnées $(x_1, \ldots, x_k)$ du vecteur $\mathbf{x}$ déterminent entièrement la seule combinaison linéaire des lignes de $\mathbf{G}$ susceptible d'engendrer $\mathbf{x}$: autrement dit, nous avons $\mathbf{x} \in C$ si et seulement si $(x_1, \ldots x_k)\mathbf{G} = \mathbf{x}$, ou encore, si et seulement si :

$$(x_1, \ldots x_k)\mathbf{A} = (x_{k+1}, x_{k+2}, \ldots, x_n).$$

Nous pouvons distinguer deux cas :
- $(x_1, \ldots x_k) = \mathbf{0}$ et $(x_{k+1}, x_{k+2}, \ldots, x_n) \neq \mathbf{0}$, auquel cas $P(\mathbf{x} \in C) = 0$.
- $(x_1, \ldots x_k) \neq \mathbf{0}$, auquel cas $\mathbf{x} \in C$ si et seulement si le vecteur aléatoire $(x_1, \ldots x_k)\mathbf{A}$ égale le vecteur fixé $(x_{k+1}, x_{k+2}, \ldots, x_n)$. Comme $(x_1, \ldots x_k)\mathbf{A}$ est distribué uniformément dans $\{0,1\}^{n-k}$, nous avons

$$P(\mathbf{x} \in C) = \frac{1}{2^{n-k}}.$$

∎

Nous pouvons en déduire l'énoncé suivant :

Proposition 27 *Pour tout $t \leqslant n/2$,*

$$P(C \cap B(\mathbf{e}, |\mathbf{e}|) \neq \{\mathbf{0}\} \mid |e| = t) \leqslant 2^{nh(\frac{t}{n})-(n-k)}.$$

Preuve : L'événement $C \cap B(\mathbf{e}, |\mathbf{e}|) \neq \{\mathbf{0}\}$ est la réunion des événements $\{\mathbf{x} \in C\}$ pour tous les $\mathbf{x} \neq 0$ de la boule $B(\mathbf{e}, |\mathbf{e}|)$. D'après le lemme 26 on a donc

$$P(C \cap B(\mathbf{e}, |\mathbf{e}|) \neq \{\mathbf{0}\} \mid |e| = t) \leqslant V(t) \frac{1}{2^{n-k}}.$$

On obtient le résultat en obtenant la majoration de $V(t)$ donnée par le lemme 10. ∎

Nous pouvons maintenant énoncer une version du théorème de Shannon pour le canal binaire symétrique

Théorème 28 *Soit un canal binaire symétrique de probabilité de transition $p < 1/2$, et soit $R < 1 - h(p)$ fixé. Soit C le code de rendement R engendré par une matrice $\mathbf{G}$ aléatoire de type* (11)*. La probabilité P_e, sur à la fois l'action du canal et le choix de la matrice $\mathbf{G}$, que le décodage au maximum de vraisemblance échoue tend vers* 0 *lorsque n tend vers l'infini. De plus, cette convergence est une fonction exponentielle de la longueur n, c'est-à-dire*

$$\liminf \frac{1}{n} \log \frac{1}{P_e} \geqslant f(R, p) > 0.$$

Preuve : Souvenons-nous que la fonction entropie binaire h est strictement croissante sur l'intervalle $[0, 1/2]$. Nous pouvons donc définir θ comme l'unique réel, $0 < \theta < 1/2$, tel que $R = 1 - h(\theta)$, et, puisque $R < 1 - h(p)$, nous avons $p < \theta$. Écrivons la formule de Bayes, en dénotant l'événement $\{C \cap B(\mathbf{e}, |\mathbf{e}|) \neq \{\mathbf{0}\}\}$ par ED (erreur de décodage),

$$\begin{aligned} P[ED] &= \sum_{t=0}^{n} P[ED \mid |\mathbf{e}| = w] P[|\mathbf{e}| = w] \\ &\leqslant \max_{t \leqslant n\left(p + \frac{\theta - p}{2}\right)} P[ED \mid |\mathbf{e}| = t] + P\left[|\mathbf{e}| > n\left(p + \frac{\theta - p}{2}\right)\right] \end{aligned}$$

ce qui nous donne, d'après la proposition 27 d'une part et le lemme 23 d'autre part,

$$P[ED] \leqslant \frac{1}{2^{n\left[h(\theta) - h\left(p + \frac{\theta - p}{2}\right)\right]}} + \frac{1}{2^{n\mathrm{D}\left(p + \frac{\theta - p}{2} \,||\, p\right)}}$$

ce qui démontre le résultat. ∎

5.5 Distance minimale typique des codes linéaires, borne de Gilbert-Varshamov

Reprenons le même modèle de code aléatoire C qu'à la section précédente, c'est-à-dire engendré par une matrice $\mathbf{G}$ de la forme (11). La probabilité que la distance

minimale de C soit strictement supérieure à t est égale à la probabilité qu'aucun mot non nul de la boule centrée en $\mathbf{0}$ et de rayon t soit dans C, soit

$$\begin{aligned} P(C \cap B(\mathbf{0},t) = \{\mathbf{0}\}) &\geqslant 1 - \sum_{\mathbf{x}\in B(\mathbf{0},t), \mathbf{x}\neq\mathbf{0}} P(x \in C) \\ &\geqslant 1 - V(t)\frac{1}{2^{n-k}} \\ &\geqslant 1 - 2^{n\left(h(\frac{t}{n})-1+R\right)} \end{aligned}$$

où $V(t)$ désigne le volume de la boule et en ayant appliqué la majoration du lemme 10. On en déduit que pour tout R, $0 \leqslant R \leqslant 1$, et pour tout $\varepsilon > 0$, si la famille de codes C est de rendement R,

$$\lim_{n\to\infty} P(d(C) \geqslant n(h^{-1}(1-R) - \varepsilon)) = 1.$$

On en déduit donc :

Théorème 29 (*Borne de Gilbert-Varshamov.) Pour tout rationnel* $0 \leqslant R \leqslant 1$, *il existe une famille de codes linéaires* C *de rendement* R *tels que la quantité* $\delta = \liminf_{n\to\infty} \frac{d(C)}{n}$ *vérifie*

$$R \geqslant 1 - h(\delta).$$

On peut montrer également que pour la famille des codes linéaires aléatoires de rendement R définis par (11), avec probabilité 1 on a $\liminf \frac{d(C)}{n} \geqslant h^{-1}(1-R)$ [1] et même $\liminf \frac{d(C)}{n} = h^{-1}(1-R)$.

5.6 Code dual, matrice de parité, syndrome

On appelle *produit scalaire* de deux vecteurs $\mathbf{x} = (x_1, \ldots, x_n)$ et $\mathbf{y} = (y_1, \ldots, y_n)$ de $\mathbb{F}_q^n$ la quantité

$$\mathbf{x}.\mathbf{y} = x_1y_1 + x_2y_2 + \cdots + x_ny_n \in \mathbb{F}_q.$$

Lorsque $\mathbf{x}.\mathbf{y} = 0$, on dit que $\mathbf{x}$ et $\mathbf{y}$ sont *orthogonaux*. Le code C étant un code linéaire de longueur n et de dimension k, on appelle *code orthogonal* ou *code dual* de C le code constitué des vecteurs orthogonaux à C

$$C^\perp = \{x \in \mathbb{F}_q^n, \forall \mathbf{c} \in C, \mathbf{x}.\mathbf{c} = 0\}.$$

Proposition 30 *Si* $\mathbf{G} = [\mathbf{I}_k \mid \mathbf{A}]$ *est une matrice génératrice du code* C *de longueur* n *et de dimension* k, *alors le code* $C^\perp$ *admet* $\mathbf{H} = [-{}^t\mathbf{A} \mid \mathbf{I}_{n-k}]$ *comme matrice génératrice.*

1. Appliquer le lemme de Borel-Cantelli.

Preuve : Il est facile de vérifier que le produit scalaire de la i-ième ligne de $\mathbf{G}$ et de la j-ième ligne de $\mathbf{H}$ vaut $-\mathbf{G}_{ij} + \mathbf{G}_{ij} = 0$. Donc le code engendré par $\mathbf{H}$ est bien inclus dans $C^\perp$. Par ailleurs, si $\mathbf{x} \in C^\perp$, on peut former

$$\mathbf{y} = \mathbf{x} - \sum_{j=n-k+1}^{n} x_j \mathbf{H}_j$$

où $\mathbf{H}_j$ désigne la j-ième ligne de $\mathbf{H}$, de telle sorte que
– $\mathbf{y} \in C^\perp$,
– pour tout j, $n-k+1 \leqslant j \leqslant n$, $y_j = 0$.
De l'orthogonalité de $\mathbf{y}$ avec toutes les lignes de $\mathbf{G}$, on déduit alors que $(y_1, \ldots, y_k)$ est orthogonal aux vecteurs de la base canonique de $\mathbb{F}_q^k$, soit

$$\mathbf{e}_1 = (1, 0, \ldots, 0), \mathbf{e}_2, \ldots, \mathbf{e}_k,$$

et donc que $y_1 = y_2 = \cdots = 0$, i.e. $\mathbf{y} = \mathbf{0}$. Le vecteur $\mathbf{x}$ est donc engendré par la matrice $\mathbf{H}$. ∎

Corollaire 31 *Pour tout code linéaire C de longueur n on a* $\dim C + \dim C^\perp = n$.

Une matrice $\mathbf{H}$ du code dual $C^\perp$ d'un code C est appelée *matrice de parité* ou *matrice de contrôle* de C. On définit l'application *syndrome* associée à la matrice $\mathbf{H}$ de dimension $r \times n$:

$$\begin{array}{rcl} \sigma : \mathbb{F}_q^n & \to & \mathbb{F}_q^r \\ \mathbf{x} & \mapsto & \sigma(\mathbf{x}) = \mathbf{H}\,{}^t\mathbf{x} \end{array}$$

L'application syndrome est une application linéaire est elle caractérise le code C par :

Proposition 32 *Le code C est l'ensemble des vecteurs de $\mathbb{F}_q^n$ de syndrome nul.*

Il est commode de retenir l'expression du syndrome sous la forme :

$$\sigma(\mathbf{x}) = \sum_{i=1}^{n} x_i \mathbf{h}_i$$

où $\mathbf{x} = (x_1 \ldots x_n)$ et $\mathbf{h}_1 \ldots \mathbf{h}_n$ sont les colonnes de la matrice $\mathbf{H}$. En se souvenant que la distance minimale d de C est le plus petit nombre de coordonnées non nulles d'un mot $\mathbf{x}$ de C, on obtient la caractérisation suivante de d :

Proposition 33 *La distance minimale de C est le plus nombre de colonnes de $\mathbf{H}$ sommant à zéro.*

Décodage par syndrome. Pour trouver le plus proche mot de code de $\mathbf{x}$, on calcule $\mathbf{s} = \sigma(\mathbf{x})$, puis on cherche le plus petit ensemble $I \subset \{1, 2, \dots n\}$ tel qu'il existe des $\lambda_i \in \mathbb{F}_q$ non nuls, $i \in I$, et

$$\sum_{i \in I} \lambda_i \mathbf{h}_i = \mathbf{s}.$$

Sur le corps $\mathbb{F}_2$, cette expression se réduit à

$$\sum_{i \in I} \mathbf{h}_i = \mathbf{s}.$$

Comme $\sigma(\mathbf{e}_i) = \mathbf{h}_i$, le mot de code le plus proche de $\mathbf{x}$ est

$$\mathbf{c} = \mathbf{x} - \sum_{i \in I} \lambda_i \mathbf{e}_i.$$