Théorie de l'information
Notes de cours

Elena Berardini*

Dernière mise à jour le 12/03/2024

Contents

*Ces notes contiennent les notions et les résultats vus en cours. Elles sont inspirées par d'autres notes de cours (citées en bibliographie). Elles sont mises à jour plus ou moins régulièrement. Elles sont très brouillon et très peu contrôlées par l'autrice, donc méfiez-vous et venez en cours !

Programme du cours jour par jour

1. Introduction à la théorie de l'information, représentation de l'information, rappels de probabilité discrète.

2. Entropie et divergence de Krullback.

3. Codage de source et compression, codes uniquement decodables, inégalités de Kraft (sans preuve) et de McMillan (avec preuve).

4. Longueur moyenne et entropie, premier théorème de Shannon. Codage de Huffman.

5. Canaux discrets et sans mémoire, capacité, exemples de canaux

6. Capacité des canaux binaire symétrique et à effacement. Rendement. Second théorème de Shannon (enoncé). Rappels sur les corps finis. Codes linéaires: premières définitions.

7. DSI.

8. Codes linéaires : distance de Hamming, correction d'erreurs et d'effacements, borne de Gilbert–Varshamov et borne de Singleton (sans preuve).

9. Preuve de la borne de Singleton. Second théorème de Shannon pour le canal binaire à effacements. Restitution des DSI.

10. Preuve de la borne de Gilbert–Varshamov. Second théorème de Shannon pour le canal binaire symétrique.

11. Distance de codes aléatoires et borne de Gilbert–Varshamov asymptotique. Dual d'un code. Matrices de parité, décodage par syndrome. Construction des nouveaux codes.

12. Codes de Reed–Solomon, codes de Goppa. Cryptographie à base de codes : le cryptosystème de McEliece.

1 Introduction

La théorie de l'information traite la transmission **efficace, intègre** et **sûre** de l'information. Fondée par Shannon en 1948 [5], la théorie de l'information construit et étudie de modèles mathématiques, basés essentiellement sur de probabilités, permettant de donner une bonne intuition du comportement des canaux de communication. Elle est désormais incontournable dans la conception de tout système de communication.

Ses domaines principaux sont

- le codage de l'information
  1. la compression des données (**efficacité**);
  2. la correction d'erreurs(**intégrité**);
- la cryptographie (**sécurité**).

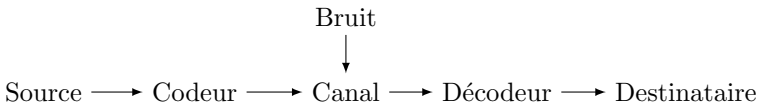

Figure 1: Représentation d'un envoie de message.
Codeur = opérations effectuées sur l'information émise par la source, avant transmission.
Bruit = le canal est généralement perturbé, ce qui peut engendrer une erreur (perte d'intégrité) ou intrusion (perte de sécurité).
Décodeur = restitue de façon acceptable l'information fournie par la source.

Le but du codeur, est double : représenter l'information émise par la source de façon concise (compression), puis protéger l'information par le bruit du canal (correction d'erreurs). Nous allons traiter ces deux tâches séparément, et parleront de codeur de source et codeur de canal (Figure 2).

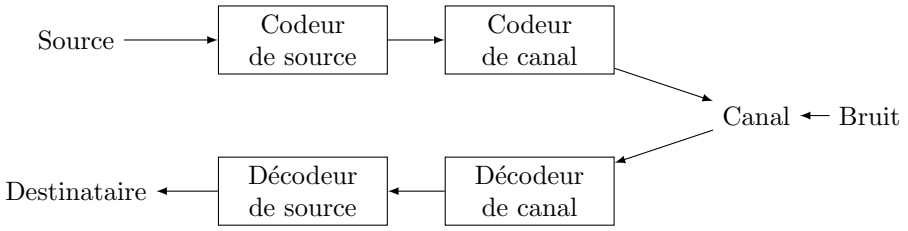

Figure 2: Codeur de source et codeur de canal.

1.1 Représentation de l'information

Définition 1.1. Un **alphabet** $\mathcal{A}$ est un ensemble fini de caractères (ou lettres, ou symboles...).

Exemple 1.2.
1. L'alphabet français, $\mathcal{A} = \{a, b, c, d...\}$, avec les majuscules, les espaces, la ponctuation, etc.
2. L'alphabet binaire $\{0, 1\}$.
3. L'alphabet octet $\{0, 1\}^8$.

Définition 1.3. Un **mot** est une suite finie de lettre de l'alphabet. On note $\mathcal{A}^n$ l'ensemble de mots de longueur n et

$$\mathcal{A}^* = \bigcup_{n \geq 0} \mathcal{A}^n.$$

Un sous ensemble $\mathcal{M}$ de $\mathcal{A}^*$ de mots (ou messages) est appelé **langue** (formelle ou naturelle).

Définition 1.4. Soit un $\mathcal{M}$ une langue dans un alphabet. Un **codage** de $\mathcal{M}$ est une application

$$c : \mathcal{M} \to \mathcal{A}^*.$$

Attention, $\mathcal{M}$ ne doit pas forcement être dans l'alphabet $\mathcal{A}$. Si c est injectif, le codage est dit **sans pertes**. On appelle $\mathcal{C}(M) \subseteq \mathcal{A}^*$ un **code**, ses éléments de **mots du code**.

Un **décodage** est une application inverse

$$D : \mathcal{C}(\mathcal{M}) \to \mathcal{M}.$$

Exemple 1.5. Le code ASCII (American Standard Code for Information Interchange) est un codage qui envoie de lettres, nombres et symboles à des octets $\{0, 1\}^8$.

Dans une langue, il y a naturellement des mots qui sont plus probables que d'autres: fixons $n = 8$, alors $m = jai\_faim$ est plus probable que $m = jdbkjshs$. On peut donc équiper $\mathcal{M} \subseteq \mathcal{A}^n$, les mots de la langue $\mathcal{M}$ de longueur n, avec une fonction de probabilité:

$$P : \mathcal{M} \to [0, 1],$$

avec $\sum_{m \in \mathcal{M}} P(m) = 1$. Le pair $(\mathcal{M}, P)$ est un **espace de probabilité**, que l'on étudiera dans la prochaine section.

Remarque 1.6. Si $\mathcal{M} = \mathcal{A}^n$ on peut avoir de mots qui ont probabilité nulle. Par exemple, pour $m = jdbkjshs$, dans la langue française on a $P(m) = 0$. Dans la théorie de l'information, au contraire, on cherche souvent un codage $\mathcal{C}(\mathcal{M})$ avec probabilité uniforme.

2 Théorie de probabilité discrète

Définition 2.1. Un espace de probabilité discrète (Ω, P) est la donné d'un ensemble fini (ou dénombrable) Ω muni d'une mesure de probabilité

$$P : \Omega \to [0, 1],$$

qui vérifie $\sum_{\omega\in\Omega} P(\omega) = 1$. La probabilité P est dite *uniforme* si $P(\omega) = 1/|\Omega|$ pour tout $\omega \in \Omega$.

Définition 2.2. Un sous-ensemble A de Ω est appelé événement. Le singleton $\{\omega\}$ est appelé événement élémentaire. On étend la mesure de probabilité sur Ω aux parties de Ω, $2^\Omega = \{A \subseteq \Omega\}$, en posant $P(\{\omega\}) = P(\omega)$. La probabilité de A est donc $P(A) = \sum_{\omega\in A} P(\omega)$. Le complémentaire de A, $\Omega \setminus A$ est noté A^c. Sa probabilité est $P(A^c) = 1 - P(A)$.

Exemple 2.3. Si la mesure de probabilité sur Ω est uniforme alors $P(A) = |A|/|\Omega|$.

Exercice 2.4. Si $A, B \subset \Omega$ et $A \cap B = \emptyset$, alors $P(A \cup B) = P(A) + P(B)$.

Remarque 2.5. Pour rappel, en théorie de la probabilité général (Ω non nécessairement fini) on définit une mesure μ sur Ω et la probabilité de $A \subset \Omega$ est $P(A) = \int_A d\mu$.

2.1 Probabilités conditionnelles

Définition 2.6. Soient A, B deux événements. On définit la probabilité de A sachant B par

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

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

Lemme 2.7. *La fonction $P(\,\cdot\,|B) : \Omega \to [0, 1]$ est une fonction de probabilité sur Ω. De plus, $(B, P(\,\cdot\,|B))$ est un espace de probabilité.*

Exercice 2.8. Prouver le Lemme 2.7.

Théorème 2.9 (Formule de Bayes)**.** *On a*

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

Démonstration. On a $A = (A \cap B) \cup (A \cap B^c)$ avec $(A \cap B) \cap (A \cap B^c) = \emptyset$. Puisque $P(A|B)P(B) = P(A \cap B)$ et $P(A|B^c)(1 - P(B)) = P(A \cap B^c)$, le résultat suit. □

Exercice 2.10. Un étudiant répond à un questionnaire à choix multiples. On propose m réponses à chaque question. On suppose que lorsque l'étudiant ne connaît 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 connaît effectivement la réponse. On appelle C l'événement "l'étudiant connaît 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.

Solution. On utilise la formule de Bayes.

$$\begin{aligned} P(R) &= P(R|C)P(C) + P(R|C^c)(1 - P(C)) \\ &= 1 \cdot y + \frac{1}{m}(1 - y). \end{aligned}$$

.

En conclusion $y = (x - 1/m)/(1 - 1/m)$. ■

2.2 Variables aléatoires

Idée : une variable aléatoire est une fonction définie sur Ω. Sa valeur est déterminée par la sortie de l'expérience, donc on peut associer de probabilités à ses valeurs.

Définition 2.11. Soit (Ω, P) un espace de probabilités. Une variable aléatoire est une fonction

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

Si $n = 1$ on parle de variable aléatoire *réelle*.

On note $P(X = x) = P(X^{-1}(x)) = P(\{\omega \in \Omega \mid X(\omega) = x\})$. Pour $S \subseteq \mathbb{R}^n$, on pose $P(X \in S) = P(X^{-1}(S)) = P(\{\omega \in \Omega \mid X(\omega) \in S\})$.

Soit $\mathcal{X} \subseteq \mathbb{R}^n$ l'image de X. On appelle loi de X la donnée de $P(x)$ pour $x \in \mathcal{X}$. On dit que X suit la loi uniforme si $P(X = x) = 1/|\mathcal{X}|$ pour tout $x \in \mathcal{X}$.

Une variable aléatoire est dite de Bernoulli si $\mathcal{X} = \{0, 1\}$.

Définition 2.12. Deux variables $X, Y : \Omega \to \mathcal{X}$ sont *indépendantes* si pour tout $x, y \in \mathcal{X}$

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

où $\{X = x, Y = y\}$ est l'intersection des événements $\{X = x\}$ et $\{Y = y\}$. Cette définition se généralise à plusieurs variables $X_1, \ldots, X_n$.

Définition 2.13. Soit (Ω, P) un espace de probabilités. L'*espérance* d'une variable aléatoire X est

$$\mathbf{E}(X) = \sum_{\omega \in \Omega} X(\omega)P(\omega) = \sum_{x \in \mathcal{X}} xP(X = x).$$

Il s'agit d'une "moyenne" ou somme pondérée par la loi de probabilité.

Exemple 2.14. Soit X une variable de Bernoulli. Alors $\mathbf{E}[X] = P(X = 1)$ (il suffit d'appliquer la définition).

Exercice 2.15. Pour un événement A, on définit sa fonction indicatrice $\mathbf{1}_A$ par

$$\mathbf{1}_A = \begin{cases} 1 \text{ si } A \text{ a lieu} \\ 0 \text{ si } A^c \text{ a lieu.} \end{cases}$$

Calculer $\mathbf{E}[\mathbf{1}_A]$.

Solution. $\mathbf{E}[\mathbf{1}_A] = 1 \cdot P(\mathbf{1}_A = 1) + 0 \cdot P(\mathbf{1}_A = 0) = P(\mathbf{1}_A = 1) = P(A)$. ■

Remarque 2.16. Dans le cadre non discret, l'espérance de la variable X est $\int_\Omega X(\omega) d\mu(\omega)$.

Théorème 2.17. *Soient X, Y deux variables aléatoires. Alors*

- $\boldsymbol{E}[X + Y] = \boldsymbol{E}[X] + \boldsymbol{E}[Y]$,
- $\boldsymbol{E}[\lambda X] = \lambda \boldsymbol{E}[X]$ *pour tout* $\lambda \in \mathbb{R}$.

Définition 2.18. Soient X, Y deux variables aléatoires.
La *covariance* de X, Y est

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

La *variance* de X est définie par

$$\operatorname{var}(X) = \mathbf{E}[(X - \mathbf{E}[X])^2].$$

L'*écart-type* est

$$\sigma(X) = \sqrt{\operatorname{var}(X)}.$$

Lemme 2.19. *On a* $\operatorname{var}(X) = \operatorname{cov}(X, X) = \boldsymbol{E}[X^2] - \boldsymbol{E}[X]^2$.

Démonstration. Vue en cours. □

Théorème 2.20 (Inégalité de Tchebitchev)**.** *Soit X une variable aléatoire, $\epsilon \in \mathbb{R}$. Alors,*

$$P(|X - \boldsymbol{E}(X)| \geq \epsilon) \leq \frac{\operatorname{var}(X)}{\epsilon^2}.$$

Démonstration.

$$\begin{aligned} P(|X - \mathbf{E}(X)| \geq \epsilon) &= \mathbf{E}\left[\mathbf{1}_{|X - \mathbf{E}(X)| \geq \epsilon}\right] \\ &\leq \mathbf{E}\left[\mathbf{1}_{|X - \mathbf{E}(X)| \geq \epsilon} \frac{|X - \mathbf{E}(X)|^2}{\epsilon^2}\right] \\ &\leq \mathbf{E}\left[\frac{|X - \mathbf{E}(X)|^2}{\epsilon^2}\right] = \frac{\operatorname{var}(X)}{\epsilon^2} \end{aligned}$$

□

L'inégalité de Tchebitchev montre qu'il est improbable qu'une variable aléatoire s'écarte de son espérance de beaucoup plus que $\sigma(X)$, d'où le nom *écart-type* pour ce dernier.

Lemme 2.21. *Soient $X, Y : \Omega \to \mathcal{X}$ deux variables indépendants. Alors*

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

Démonstration.

$$\begin{aligned}
\mathbf{E}[X]\mathbf{E}[Y] &= \sum_{x\in\mathcal{X}} xP(X=x) \sum_{x\in\mathcal{X}} yP(Y=y) \\
&= \sum_{x,y\in\mathcal{X}} xyP(X=x)P(Y=y) \\
&= \sum_{x,y\in\mathcal{X}} xyP(X=x,Y=y) \\
&= \sum_{z} z \sum_{xy=z} P(X=x,Y=y) \\
&= \sum_{z} zP(XY=z) = \mathbf{E}[XY],
\end{aligned}$$

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

Corollaire 2.22. *Soient $X_1, \ldots, X_n$ des variables deux à deux indépendants. Alors*

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

Démonstration. Exercice. Il s'agit d'utiliser le lemme précédent et l'additivité de l'espérance. □

3 Entropie

3.1 Rappels sur le logarithme

Dans la suite, on utilisera les propriétés suivantes du logarithme :

- Pour tout réel $z \geq 0$ on a $\log_2 z \leq z - 1$.
- $\log_2(a \cdot b) = \log_2(a) + \log_2(b)$,
- $\log_2 \frac{a}{b} = \log_2 a - \log_2 b$.

3.2 Entropie et divergence de Kullback

Dans ce qui suit, X dénote une variable aléatoire de loi p. Soit $\mathcal{X} = \{x_1, \ldots, x_m\}$ son image, et $p_i = P(X = x_i)$, alors $p = (p_1, \ldots, p_m)$.

Définition 3.1. L'*entropie* de X est définie par

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

L'entropie ne dépend que de p, la loi de X, et est donc notée aussi $H(p)$. On peut la voir aussi comme l'espérance de la variable $F = -\log_2(P(X = x))$.

Remarque 3.2. Si $p_i = 0$, on définit $p_i \log_2 \frac{1}{p_i} = 0$. Ceci est cohérent avec la limite $\lim_{t\to 0} t \log_2 \frac{1}{t} = 0$.

L'entropie, introduite par Shannon, mesure la quantité d'information d'un espace de probabilité. On appelle un bit d'information l'unité d'entropie.

Lemme 3.3. *Soit X avec la loi uniforme, soit $m = |\mathcal{X}|$. Alors, $H(X) = \log_2 m$*

Démonstration. Soit $|\mathcal{X}| = m$, alors la loi uniforme est $p_i = 1/m$ pour chaque i. Par définition d'entropie nous obtenons alors

$$H(X) = \sum_{i=1}^{m} p_i \log_2 \frac{1}{p_i} = \sum_{i=1}^{m} \frac{1}{m} \log_2 m = \log_2 m.$$

□

Définition 3.4. Pour X, Y deux variables aléatoires, on définit l'entropie du couple (X, Y) par

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

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

Définition 3.5. On définit la *divergence* de Kullback entre deux lois p et q par

$$D(p \parallel q) = \sum_{x\in\mathcal{X}} p(x) \log_2 \frac{p(x)}{q(x)}$$

Remarque 3.6. Souvent on parle de *distance* de Kullback. Cependant, la définition précédente ne définit pas une distance. Pourquoi ? Une distance d doit respecter par définition les propriétés suivantes :

1. $d(x, y) \geq 0$ avec égalité si et seulement si $x = y$;
2. $d(x, y) = d(y, x)$;
3. $d(x, y) \geq d(x, z) + d(z, y)$.

Tout d'abord, $D(p \parallel q) = \sum_x p(x) \log_2 \frac{p(x)}{q(x)}$ et $D(q \parallel p) = \sum_x q(x) \log_2 \frac{q(x)}{p(x)}$ sont différentes. De plus, la divergence de Kullback ne satisfait pas non plus l'inégalité triangulaire. Cependant, elle satisfait la première propriété d'une distance, comme montré dans le lemme qui suit.

Lemme 3.7. *On a $D(p \parallel q) \geq 0$, avec égalité si, et seulement si, $p = q$.*

Démonstration. Nous avons la suite d'inégalité suivante :

$$\begin{aligned}
\log_2 \frac{q(x)}{p(x)} &\leq \frac{q(x)}{p(x)} - 1 \\
p(x) \log_2 \frac{q(x)}{p(x)} &\leq q(x) - p(x) \\
p(x) \log_2 \frac{p(x)}{q(x)} &\geq p(x) - q(x).
\end{aligned}$$

En sommant su x, on obtient

$$\sum_{x \in \mathcal{X}} p(x) \log_2 \frac{p(x)}{q(x)} \geq \sum_{x \in \mathcal{X}} p(x) - \sum_{x \in \mathcal{X}} q(x) = 1 - 1,$$

d'où $D(p \parallel q) \geq 0$.

Si $p = q$, alors $\log_2 \frac{p(x)}{q(x)} = 0$, donc $D(p \parallel q) = 0$. Le fait que $D(p \parallel q) = 0$ implique $p = q$ a été laissé comme exercice en cours. □

Corollaire 3.8. *L'entropie maximale est atteinte avec la loi uniforme.*

Démonstration. Nous avons vu que si p est la loi uniforme, alors $H(p) = \log_2 m$. Pour tout autre loi q on a

$$\begin{aligned}
\log_2 m - H(q) &= \log_2 m - \sum_{i=1}^{m} q_i \log_2 \frac{1}{q_i} \\
&= \sum_{i=1}^{m} q_i \log_2 m + \sum_{i=1}^{m} q_i \log_2 q_i \\
&= \sum_{i=1}^{m} q_i \log_2 \frac{q_i}{1/m} \\
&= D(p \parallel 1/m) \geq 0,
\end{aligned}$$

où la dernière inégalité repose sur le Lemme 3.7. □

3.3 Entropie conditionnelle

Définition 3.9. Soient X, Y deux variables aléatoires. On définit l'*entropie conditionnelle* comme

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

Proposition 3.10. *On a*

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

Démonstration. On a

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

Il s'en suit

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

□

Définition 3.11. On définit l'*information mutuelle* entre X et Y par

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

D'après la dernière proposition, on peut réécrire l'information mutuelle comme

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

Proposition 3.12. *On a $I(X,Y) \geq 0$.*

Démonstration. On écrit

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

et on obtient

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

où la dernière inégalité provient du Lemme 3.7 □

3.4 Loi de Bernoulli et entropie

Rappels : Une variable aléatoire est dite de Bernoulli si $\mathcal{X} = \{0, 1\}$. Dans ce cas, la loi de X est donné par $p = (\lambda, 1 - \lambda)$.
Notation: on notera l'entropie $H(p)$ de la loi p de Bernoulli par $h(p)$.

Définition 3.13. L'*entropie binaire* est la fonction

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

L'entropie binaire est utile pour évaluer le comportement asymptotique des coefficients binomiaux, comme suit.

Lemme 3.14. *Pour tout $\lambda \leq \frac{1}{2}$ on a*

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

Pour tout $\lambda \geq \frac{1}{2}$ on a

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

Démonstration. On commence par démontrer (2). Pour tout $r \geq 0$ on écrit

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

On en déduit

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

On pose maintenant $r = \log_2 \frac{\lambda}{\lambda - 1} \geq 0$ pour tout $\lambda \geq 1/2$ et on obtient

$$\begin{aligned} \sum_{i \geq \lambda n} \binom{n}{i} &\leq \left(\frac{1-\lambda}{\lambda}\right)^{\lambda n} \left(1 + \frac{\lambda}{1-\lambda}\right)^n \\ &= \left(\frac{1}{\lambda}\right)^{\lambda n} \left(\frac{1}{1-\lambda}\right)^{-\lambda n} \left(\frac{1}{1-\lambda}\right)^n \\ &= \left(\frac{1}{\lambda}\right)^{\lambda n} \left(\frac{1}{1-\lambda}\right)^{n - \lambda n} = 2^{nh(\lambda)}. \end{aligned}$$

On peut déduire (1) de (2) en utilisant $\binom{n}{i} = \binom{n}{n-i}$. □

4 Codage de source

Dans cette section on commence l'étude du codage de source (voir Figure 2). **Rappels :** Pour un alphabet $\mathcal{A}$ nous avons défini $\mathcal{A}^n$ comme l'ensemble de mots de longueur n et

$$\mathcal{A}^* = \bigcup_{n \geq 0} \mathcal{A}^n.$$

Par la suite, on va considérer $\mathcal{A} = \{0, 1\}$ l'alphabet binaire, mais la théorie se généralise à tout alphabet.

Si $X_1, \ldots, X_n$ sont n variables aléatoires à valeurs dans $\mathcal{X}$, on dit qu'elles sont *indépendantes* si

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

Définition 4.1. Un *code compressif* (ou simplement un *code*) est un ensemble fini de mots $\mathcal{C} \subset \{0, 1\}^*$. Pour un mot $c \in \mathcal{C}$ on note $\ell(c)$ sa longueur qui est le nombre de symboles binaires qui le constituent.

Définition 4.2. Soit X une variable aléatoire d'image $\mathcal{X}$. Un *codage* de X (ou de $\mathcal{X}$) est une application

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

Exemple 4.3. Soit X à valeurs dans $\mathcal{X} = \{1, 2, 3, 4\}$ et de loi $p = (1/2, 1/4, 1/8, 1/8\}$. L'application c définie par

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

est un codage de X. La longueur des mots est $1, 2, 3$ et 3, respectivement.

Soit $X^n = X_1 \ldots X_n$ une suite de variables indépendantes de même loi que X. La suite X^n se traduit en une suite de symboles de $\mathcal{X}$, *i.e.,* $x_1 \ldots x_n \in \mathcal{X}^n$. Par concaténation, et en utilisant un codage c de $\mathcal{X}$, à partir de la suite de symboles de $\mathcal{X}$ on obtient une suite de mots de $\mathcal{C}$, *i.e.,* $c_1 \ldots c_n \in \mathcal{C}^n$. Cette dernière est une suite de symboles binaires. En somme, on obtient l'application

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

définie par $c^*(x_1 \ldots x_n) = c(x_1) \ldots c(x_n)$.

Définition 4.4. Un codage c (équiv. le code $\mathcal{C}$) est dit *sans perte* si c est injectif. Un codage c (équiv. le code $\mathcal{C}$) est dit *uniquement décodable* (aussi *uniquement déchiffrable*) si son extension c^* est sans perte, équiv. si toute suite binaire de $\mathcal{C}^*$ se décompose de manière unique en une concaténation de mots de $\mathcal{C}$.

Exemple 4.5. Soit $X = \{A, B, C\}$ et soit c le codage

$$c(A) = 0$$
$$c(B) = 1$$
$$c(C) = 01.$$

Alors c est sans perte, mais pas uniquement décodable car $c^*(AB) = 01 = c^*(C)$.

Exemple 4.6. Le code $\mathcal{C}_1 = \{0, 01\}$ est uniquement décodable. Le code $\mathcal{C}_2 = \{0, 11, 010\}$ est uniquement décodable. Le code $\mathcal{C}_3 = \{0, 01, 001\}$ n'est pas uniquement décodable car on peut écrire $0001 \in \mathcal{C}^*$ comme $0 \cdot 001$ et $0 \cdot 0 \cdot 01$.

4.1 Codes préfixes

Une classe de codes uniquement décodable sont les codes préfixes.

Définition 4.7. Un code est dit *préfixe* si aucun de ses mots n'est le préfixe d'un autre mot de code.

Lemme 4.8. *Un code préfixe est uniquement décodable.*

Il existe de codes uniquement décodables qui ne sont pas préfixes, comme $\{0, 01\}$. On observe aussi que le seul code préfixe contenant 0 et 1 est le code $\{0, 1\}$.

Remarque 4.9. La notion de code préfixe est asymétrique. On peut de façon équivalent définir un code suffixe comme un code dont aucun de ses mots n'est le suffixe d'un autre mot du code. De nouveau, on peut montrer qu'un code suffixe est uniquement décodable.

Exemple 4.10 (Différence entre code préfixe et suffixe)**.** Soit $\mathcal{X} = \{A, B, C\}$, C_1 un code suffixe et C_2 un code préfixe, définis comme suit

$$C_1(A) = 0, C_1(B) = 11, C_1(C) = 01,$$
$$C_2(A) = 0, C_2(B) = 11, C_2(C) = 10.$$

Si $C_1^\star(x_1 \dots x_n) = 00111 \dots$, on doit lire jusqu'à la fin pour savoir si le message commence par ACB ou AAB. Par contre, si $C_2^\star(x_1, \dots, x_n) = 00111 \dots$ on sait dès le début de la réception que le message commence par AAB, c'est-à.dire, on peut décoder de gauche à droite sans attendre la fin de la transmission.

Lemme 4.11. *Soit $c : \mathcal{X} \to \mathcal{A}^n \subset \mathcal{A}^*$ un codage de longueur fixée. Alors, c est préfixe si et seulement s'il est suffixe, si et seulement s'il est sans perte.*

Démonstration. Exercice. □

Proposition 4.12 (Inégalité de Kraft)**.** *Soit $\mathcal{X} = \{1, \dots, m\}$ et soit $\ell_1, \dots, \ell_m$ une suite d'entier positifs. Il existe un code préfixe $\mathcal{C}$ et un encodage $c : X \to \mathcal{C}$ tel que $\ell(c(i)) = \ell_i$ si, et seulement, si,*

$$\sum_{i=1}^{m} 2^{-\ell_i} \leq 1. \tag{3}$$

Démonstration. □

La même proposition reste vraie si on remplace l'hypothèse d'être préfixe par celle plus faible d'être uniquement décodable, comme suit.

Proposition 4.13 (Inégalité de McMillan)**.** *Soit $\mathcal{X} = \{1, \dots, m\}$ et soit $\ell_1, \dots, \ell_m$ une suite d'entier positifs. Il existe un code uniquement déchiffrable $\mathcal{C}$ et un encodage $c : X \to \mathcal{C}$ tel que $\ell(c(i)) = \ell_i$ si, et seulement, si,*

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

Démonstration. Si une suite (ℓ_i) satisfait Equation (3), alors on sait d'après la Proposition 4.12 qu'il existe un code préfixe, donc uniquement décodable, associé à la suite.

Soit $\mathcal{C}$ un code uniquement décodable. Notons $\mathcal{C}^k$ l'ensemble de suites binaires obtenues par concaténation de exactement k mots de C. Si $c = c_1 c_2 \dots c_k$ est la concaténation de k mot de $\mathcal{C}$, alors $\ell(c) = \sum \ell(c_i)$. Puisque le code est uniquement décodable, on sait que chaque mot de $\mathcal{C}^k$ est associé à exactement une suite de k mots de $\mathcal{C}$ dont il est la concaténation (car le codage c dont $\mathcal{C}$ est l'image est injectif), donc

$$\begin{aligned}\left(\sum_{c\in\mathcal{C}} 2^{-\ell(c)}\right)^k &= \sum_{c_1,\dots,c_k\in C} 2^{-(\ell(c_1)+\dots+\ell(c_k))} \\ &= \sum_{c\in C^k} 2^{-\ell(c)}.\end{aligned}$$

Soit ℓ_{max} la longueur maximale d'un mot de $\mathcal{C}$. Alors la longueur maximale d'un mot de $\mathcal{C}^k$ est $k\ell_{max}$. Soit A_i le nombre de mots de $\mathcal{C}^k$ de longueur i. On obtient

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

Comme $A_i \leq 2^i$ on en déduit

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

et donc

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

Puisque la dernière inégalité est vraie pour tout k entier, il s'en suit que

$$\sum_{c\in\mathcal{C}} 2^{-\ell(c)} \leq 1.$$

□

4.2 Longueur moyenne et premier théorème de Shannon

Définition 4.14. Soit X une variable aléatoire, c un codage de X. La *longueur moyenne* de c, $\bar{\ell}(c)$, est définie par

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

Exemple 4.15. La longueur moyenne du codage de l'Exemple 4.3 est

$$\bar{\ell}(c) = \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{8} \cdot 3 + \frac{1}{8} \cdot 3 = \frac{7}{4}.$$

On remarque que dans l'exemple précédent on a $\bar{\ell}(c) = H(X)$.

Théorème 4.16 (Premier théorème de Shannon)**.** *Soit X une variable aléatoire d'image finie $\mathcal{X}$.*

1. *Soit $c : \mathcal{X} \to \mathcal{C}$ un codage de X par un code uniquement déchiffrable. Alors la longueur moyenne de ce codage vérifie*

$$\bar{\ell}(c) \geq H(x).$$

2. *Il existe un codage $c : \mathcal{X} \to \mathcal{C}$ dont la longueur moyenne vérifie*

$$\bar{\ell}(c) < H(x) + 1.$$

Démonstration. Posons $\mathcal{X} = \{x_1, \ldots, x_m\}$ et $\mathcal{C} = \{c_1, \ldots, c_m\}$ de telle sorte que $c(x_i) = c_i$. On commence par prouver (1). Considérons

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

D'après la Proposition 4.13, on sait que $Q \leq 1$. 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 $D(p \parallel q) \geq 0$ s'écrit

$$\sum_i p_i \log_2 \frac{p_i}{q_i} \geq 0,$$

soit

$$\begin{aligned} -H(X) - \sum_i p_i \log_2 q_i &\geq 0 \\ \sum_i p_i \ell(c_i) + \sum_i p_i \log_2 Q &\geq H(X) \\ \bar{\ell}(c) &\geq H(X) - \log_2 Q, \end{aligned}$$

ce qui conclut la preuve de (1) puisque $Q \leq 1$.

Soit $p_i = P(X = x_i)$ une loi sur X. Pour $i = 1, \ldots, m$, posons

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

On a

$$\log_2 \frac{1}{p_i} \leq \ell_i \Rightarrow -\ell_i \leq \log_2 p_i \Rightarrow 2^{-\ell_i} \leq p_i \Rightarrow \sum_i 2^{-\ell_i} \leq 1.$$

Donc, on sait qu'il existe un codage de X par un code $\mathcal{C}$ de distribution des longueurs ℓ_i pour $i = 1, \ldots, m$. Par ailleurs, on a

$$\ell_i < \log_2 \frac{1}{p_i} + 1,$$

et donc

$$\sum_i p_i \ell_i < H(X) + \sum_i p_i = H(X) + 1,$$

ce qui conclut la preuve de (2). □

5 Codage de Huffman

D'après le premier théorème de Shannon on sait que chaque codage de X uniquement decodable a une longueur moyenne d'au moins $H(X)$, et qu'il existe au moins un codage dont la longueur moyenne est strictement inférieure à $H(X)+1$. De plus, d'après l'inégalité de Kraft et de McMillan, on sait que pour chaque code uniquement decodable il existe un code préfixe de même distribution de longueurs. On peut donc chercher le *code optimal*, c'est-à-dire qui minimise la longueur moyenne, parmi les codes préfixes. Le codage de Huffman [3], inventé par David Albert Huffman lors de sa thèse de doctorat au MIT, est un exemple d'un code optimal : sa longueur moyenne $\overline{\ell}$ satisfait $H(X) \leq \overline{\ell} < H(X) + 1$.

5.1 Arbre de codage

Un arbre est un graphe acyclique et connexe, avec un sommet distingué appelé racine. Les sommets terminales qui n'ont qu'un seul arrêt, sont appelés *racines*. Un arbre est appelé binaire si tout sommet a au plus deux fils, un fils gauche et un fils droit.

Soit X une variable aléatoire à valeurs dans $\mathcal{X}$. Un *arbre de codage* Γ est un arbre fini tel que chaque feuille est associé à un certain $x \in \mathcal{X}$, et chaque arête d'un sommet est associée à un symbole différent de $\mathcal{A}$. Si $\mathcal{A} = \{0, 1\}$ comme dans notre contexte, on obtient alors un arbre binaire. Le mot du code $c(x)$ correspond à la suite des éléments de $\{0, 1\}$ qui apparaissent sur le chemin qui porte de la racine de l'arbre à la feuille associée à x. La longueur d'un mot du code $c(x)$ est la profondeur de x, c'est-à-dire le nombre d'arêtes qui lient la racine à x.

Si chaque $x \in \mathcal{X}$ apparaît une et une seule fois dans une feuille de Γ, on dit que Γ est un *arbre de codage de* X. En particulier, on remarque qu'un code associé à un arbre de codage binaire est préfixe.

5.2 L'algorithme de Huffman

Soit X une variable aléatoire à valeurs dans $\mathcal{X} = \{x_1, \ldots, x_m\}$ et avec loi de probabilité $p = (p_1, \ldots, p_m)$. Sans perte de généralité, on peut supposer $p_1 \geq p_2 \geq \cdots \geq p_m$. L'algorithme procède par récurrence, en construisant l'arbre binaire à partir des feuilles. L'arbre est obtenu comme suit :

- on associe deux feuilles aux valeurs de probabilité plus baisses, x_m et x_{m-1}, issues d'un sommet père commun, soit x'_{m-1}.
- on calcule l'arbre de Huffman associé à la variable aléatoire X' à valeurs dans $\mathcal{X}' = \{x_1, x_2, \ldots, x_{m-2}, x'_{m-1}\}$ et de loi $p' = (p_1, p_2, \ldots, p_{m-2}, p_{m-1} + p_m)$;
- on rajoute deux fils issus de x'_{m-1} qui sont étiqueté par x_{m-1} et x_m.

Exemple 5.1. Soit X une variable aléatoire d'image $\mathcal{X} = \{x_1, x_2, x_3, x_4, x_5, x_6\}$ et de loi $p = (0.4, 0.04, 0.14, 0.18, 0.18, 0.06)$. On peut permuter les x_i pour obtenir la loi $p = (0.4, 0.18, 0.18, 0.14, 0.06, 0.04)$ qui respecte $p_1 \geq \cdots \geq p_6$. On applique l'algorithme de Huffman. La première étape consiste à joindre les sommets x_6 et x_5 et à créer un sommet intermédiaire x_5^1 de probabilité $p_5^1 = p_6 + p_5 = 0.1$. Maintenant on considère la variable aléatoire $X^1 = \{x_1, x_2, x_3, x_4, x_5^1\}$ de loi $p = (0.4, 0.18, 0.18, 0.14, 0.1)$. On itère le processus avec x_4 et x_5^1 et on crée ainsi un sommet x_4^2 de probabilité $p_4^2 = p_4 + p_5^1 = 0.24$. Maintenant on considère la variable aléatoire $X^2 = \{x_1, x_2, x_3, x_4^2\}$ de loi $p = (0.4, 0.18, 0.18, 0.24)$. On itère le processus avec x_2 et x_3 qu'on joint au sommet x_2^3 de probabilité 0.36. On va après joindre les sommets x_4^2, de probabilité 0.24 et x_2^3 de probabilité 0.36, en le sommet x_2^4 de probabilité 0.6. Pour finir, on joint le sommet x_1 et le sommet x_2^4 au dernier nouveau sommet, noté $\emptyset$, de probabilité 1.

On peut calculer la longueur moyenne d'un codage associé à cette arbre:

$$\overline{\ell} = 0.04 \cdot 4 + 0.06 \cdot 4 + 0.14 \cdot 3 + 0.18 \cdot 3 + 0.18 \cdot 3 + 0.4 \cdot 1 = 2.3.$$

Aussi, on calcule aisément qu'on a $H(X) \simeq 2.24...$. Donc on vérifier bien que

$$H(x) \leq \overline{\ell} < H(X) + 1.$$

Lemme 5.2. *Soit X une variable aléatoire à valeurs dans $\mathcal{X} = \{x_1, \ldots, x_m\}$ et avec loi de probabilité $p = (p_1, \ldots, p_m)$ avec $p_1 \geq p_2 \geq \cdots \geq p_m$. Parmi les codages optimaux de X il existe un codage préfixe c dont l'arbre encode x_{m-1} et x_m par des feuilles de profondeur maximale et ayant un même père.*

Démonstration. Soit C un codage de X. Si x_m n'est pas associé à un sommet de profondeur maximale, alors il suffit de l'échanger avec un sommet x_i de

profondeur maximale. Ainsi, la longueur moyenne du codage associé ne peut que diminuer. Donc on peut toujours se ramener au cas où x_{m-1} et x_m sont des feuilles de profondeur maximale. Le sommet x_m ne peut pas être le seul fils du sommet père x'_{m-1}, sinon on enlève x'_{m-1} de l'arbre et on associe x_m au nœud de x'_{m-1}. Il y a donc un autre sommet, une feuille, soit x_j, pour $j \neq m$, qui a x'_{m-1} comme père. En particulier, puisque $j \neq m$ on $p_j \geq p_{m-1}$, donc en échangeant x_j et x_{m-1} la longueur moyenne du codage ne peut que diminuer. En conclusion, on peut toujours se ramener au cas où x_m et x_{m-1} ont le même sommet père. □

Proposition 5.3. *L'algorithme de Huffman donne un codage optimal de X.*

Démonstration. Soit $\mathcal{C}_m^*$ un code préfixe optimal associé à la loi p de X tel que l'arbre de codage associé encode x_{m-1} et x_m par des feuilles de profondeur maximale et ayant un même père. Soit $\mathcal{C}_{m-1}^*$ un code préfixe optimal associé à la loi $p' = (p_1, p_2, \ldots, p_{m-2}, p_{m-1} + p_m)$. Soit C_{m-1} le code préfixe (ou arbre) associé à loi p' obtenu à partir de $\mathcal{C}_m^*$ par réduction de Huffman, c'est-à-dire on supprime les sommets x_m et x_{m-1} et on rajoute le sommet père x'_{m-1} de probabilité $p_{m-1} + p_m$. Enfin, soit C_m le code préfixe (ou arbre) obtenu à partir de C_{m-1}^* en rajoutant au sommet de probabilité $p_{m-1} + p_m$ deux feuilles, nommées x_{m-1} et x_m.

La longueur moyenne de C_m^* est

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

Puisque $\ell_{m-1} = \ell_m$, d'après le Lemme 5.2 on a

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

c'est-à-dire

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

De la même façon on a

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

et

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

et on obtient

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

En additionnant les deux égalités on obtient

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

et donc

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

Par optimalité de C^*_m et C^*_{m-1} le terme de gauche est positif et le terme de droite est négatif, donc les deux sont forcement nuls. On en déduit que les codes C_m et C_{m-1} sont optimaux pour les lois p et p'. On en déduit par récurrence sur m que les réductions de Huffman successives mènent un code préfixe optimal. □

6 Canaux discrets sans mémoire et leurs capacité

6.1 Canaux discrets sans mémoire

Pour définir un canal de transmission, il faut décrire l'ensemble des entrées et des sorties possibles du canal, ainsi que le bruit qui perturbera la transmission. Dans le cas d'un canal discret, l'exemple plus simple, l'ensemble des entrées et celui des sorties sont des ensembles finis $\mathcal{X}$ et $\mathcal{Y}$, vu comme les images de deux variables aléatoires X et Y, et le bruit est modélisé par la donnée d'une loi de probabilité conditionnelle de Y sachant X.

Définition 6.1. Soient X et Y deux variables aléatoires à valeurs dans $\mathcal{X}$ et $\mathcal{Y}$, respectivement. Un *canal (de communication)* est la donnée de $(\mathcal{X}, \mathcal{Y}, P)$ où $\mathcal{X}$ et $\mathcal{Y}$ sont appelés alphabet d'entrée et de sortie, respectivement, et $P : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}, (y, x) \mapsto P(Y = y|X = x)$ est une fonction de probabilité conditionnelle.

Remarque 6.2. A chaque canal on peut associer une collection d'espaces de probabilités, comme suit. Soit $x \in \mathcal{X}$ un élément de l'alphabet d'entrée. On peut alors définir $P_x : 2^{\mathcal{Y}} \to \mathbb{R}$ par $P_x(\mathcal{Y}') = \sum_{y \in \mathcal{Y}'} P(Y = y|X = x)$, pour $\mathcal{Y}' \in 2^{\mathcal{Y}}$. Alors $(\mathcal{Y}, P_x)$ est un espace probabilisé.

Définition 6.3. Soient $\mathcal{X}, \mathcal{Y} = \{0, 1\}$ et soit $0 \leq p \leq 1$ un réel. Le *canal binaire symétrique* $(\mathcal{X}, \mathcal{Y}, P)$ avec probabilité p est défini par

$$P(y|x) = \begin{cases} 1 - p \text{ si } x = y \\ p \text{ si } x \neq y. \end{cases}$$

Définition 6.4. Soient $\mathcal{X} = \{0, 1\}$, soit $\mathcal{Y} = \mathcal{X} \cup ?$, et soit $0 \leq p \leq 1$ un réel. Le *canal binaire à effacement* $(\mathcal{X}, \mathcal{Y}, P)$ avec probabilité d'effacement p est défini par

$$P(y|x) = \begin{cases} 1 - p \text{ si } x = y \\ p \text{ si } x = ? \\ 0 \text{ sinon.} \end{cases}$$

Le symbole ? est appelé symbole d'effacement.

Définition 6.5. Soit $(\mathcal{X}, \mathcal{Y}, p)$ un canal, et n un entier positif. Soient $X^n = (X_1, \ldots, X_n)$ et $Y^n = (Y_1, \ldots, Y_n)$ deux n-uples de variables aléatoires à valeurs dans $\mathcal{X}$ et $\mathcal{Y}$, respectivement. On définit le *canal discret sans mémoire* par $(\mathcal{X}^n, \mathcal{Y}^n, P)$ où P respecte les deux hypothèses qui suivent.

- Le canal est *sans mémoire*, c'est à dire

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

- Le canal est utilisé *sans rétroaction*, c'est-à-dire le n-ième symbole X_n ne dépend pas de sorties précédentes :

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

"Sans mémoire" indique le fait que ce qui se passe dans une utilisation du canal n'influence pas ce qui se passe dans une autre utilisation. En effet, les deux hypothèses impliquent en particulier la décomposition du lemme suivant.

Lemme 6.6. *Soit P la fonction de probabilité du canal discret sans mémoire. Alors*

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

Démonstration.

$$\begin{aligned} P(Y^n = y^n | X^n = x^n) &= P(Y_n = y_n | X^n = x^n, Y^{n-1} = y^{n-1}) P(Y^{n-1} = y^{n-1} | X^n = x^n) \\ &= P(Y_n = y_n | X_n = x_n) P(Y^{n-1} = y^{n-1} | X^n = x^n), \end{aligned} \tag{4}$$

en utilisant que le canal est sans mémoire. De plus

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

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

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

En utilisant cette dernière égalité dans (4) on déduit

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

On conclut enfin par récurrence. □

6.2 Capacité d'un canal

Quel est le maximum d'information qu'on peut faire passer sur le canal ?

L'émetteur choisi la loi du n-uple X^n. S'il veut que le destinataire puisse reconstituer X^n à partir du n-uplet reçu Y^n il faut que l'entropie conditionnelle $H(Y^n|X^n)$ soit négligeable, voir nulle. L'émetteur veut donc optimiser l'information mutuelle

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

L'objectif est d'obtenir $I(X^n, Y^n) = H(X^n)$, ce qui correspond à zero perte d'information. On pose

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

On peut écrire $I(X^n, Y^n) = H(Y^n) - H(X^n|Y^n)$. D'après le Lemme 6.6 on peut écrire

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

et donc

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

On obtient donc

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

soit

$$C^{(n)} \leq C := \max_{p(X)} I(X, Y),$$

puisque Y_i ne dépend que de X_i. Ici $p(X)$ désigne toutes les lois d'émissions possibles. On appelle C la *capacité du canal*. Elle majore la quantité d'information fiable par symbole qu'il est possible de transmettre sur un canal.

6.3 Exemples de calcul de capacité

6.3.1 Le canal binaire symétrique

On écrit $I(X, Y) = H(Y) - H(Y|X)$. 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 $h(p)$ désigne l'entropie d'une loi de Bernoulli $(p, 1-p)$. On a donc

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

Lorsque la loi sur X est uniforme, la loi sur Y l'est aussi. En effet, pour $x \in \{0,1\} = \mathcal{Y}$, on a

$$\begin{aligned} P(Y=x) &= P(X=0, Y=x) + P(X=1, Y=x) \\ &= P(X=0)P(Y=x|X=0) + P(X=1)P(Y=x|X=1) \\ &= \frac{1}{2}p + \frac{1}{2}(1-p) = \frac{1}{2}. \end{aligned}$$

En particulier, la loi uniforme maximise $I(X,Y)$. On a donc

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

6.3.2 Le canal binaire à effacements.

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

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

On a

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

Pour $x = 0, 1$ on a

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

Il s'en suite que $H(X|Y=?) = H(X)$ et donc

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

La loi de X qui maximise cette quantité est la loi uniforme pour laquelle $H(X) = 1$. En conclusion

$$C = 1 - p.$$

6.4 Rendement et théorème de Shannon pour les canaux

Soit $\mathcal{M}$ un ensemble de messages. Nous avons modélise un système de communication comme suit:

- un message $m \in \mathcal{M}$ est transformé en un n-uple de $\mathcal{X}^n$ par une fonction d'*encodage* $c : \mathcal{M} \to \mathcal{X}^n$. L'image de la fonction c, $C \subseteq \mathcal{X}^n$, est un *code* ;
- l'n-uple est envoyé sur un canal discret sans mémoire et devient un élément de $\mathcal{Y}^n$;
- ensuite avec une fonction de *décodage* $d : \mathcal{Y}^n \to \mathcal{M}$ on transforme le n-uple reçu en un message de $\mathcal{M}$.

On définit la probabilité conditionnelle

$$\lambda_m = P(d(Y^n) \neq m | X^n = c(m)).$$

On définit deux probabilités différentes d'une erreur de décodage.

Définition 6.7. La *probabilité maximale* d'une erreur de décodage est

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

La *probabilité moyenne* d'une erreur de décodage est

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

Nous avons $P_e^n \leq \lambda^n$.

Définition 6.8. Le *rendement* d'un code $C \subseteq \mathcal{X}^n$ binaire ($\mathcal{X} = \{0, 1\}$) est

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

Théorème 6.9 (Deuxième théorème de Shannon)**.** *Pour tout canal discret sans mémoire de capacité C et pour tout $R < C$, il existe une suite (C_n) de codes $C_n \subset \mathcal{X}^n$ de rendement supérieur ou égale à R et pour laquelle $\lambda^n \to 0$ quand $n \to \infty$. Réciproquement, si pour une suite (C_n) de codes $\lambda^n \to 0$ quand $n \to \infty$, alors $\limsup_{n\to\infty} R(C_n) \leq C$.*

Ce théorème montre qu'il existe des codes permettant de réaliser un code dont le rendement est aussi proche qu'on veut de la capacité du canal. Il nous indique aussi, par la réciproque, qu'il est inutile de chercher des codes de rendement supérieur à la capacité du canal.

7 Codes linéaires

Notations: on notera

- $\mathbf{0}$ le vecteur $(0, \ldots, 0) \in \mathbb{F}_q^n$;
- $\boldsymbol{I}_n$ la matrice identité de taille $n \times n$.

Définition 7.1. Un *code linéaire* $C \subseteq \mathbb{F}_q^n$ est un sous-espace vectoriel de $\mathbb{F}_q^n$.

Quand $q = 2$ on parle de code binaire.

Définition 7.2. La *matrice génératrice* G d'un code linéaire $C \subseteq \mathbb{F}_q^n$ de dimension k est une matrice de taille $k \times n$ à coefficient dans $\mathbb{F}_q$ dont les lignes constituent une base de l'espace vectoriel C.

Si G est une matrice génératrice du code C, une fonction d'encodage de l'ensemble des messages $\mathcal{M}$ est donnée par

$$\begin{aligned} c : \mathcal{M} &\to \mathbb{F}_q^n \\ x &\mapsto xG. \end{aligned}$$

Définition 7.3. Une matrice génératrice est dite sous forme *systématique* si elle s'écrit comme

$$G = [\boldsymbol{I}_k | A]$$

avec A une matrice $k \times (n-k)$.

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 premiers k symboles sont appelé *bit d'information* et les autres $n-k$ symboles *bits de parité*.

7.1 Distance de Hamming et capacité de correction et détection

Le support d'un mot $x \in \mathbb{F}_q^n$, noté $\sigma(x)$, est l'ensemble d'indices sur lesquels il a de cordonnées non nulles, c'est-à-dire

$$\sigma(x) = \{i \in \{1, \ldots, n\} \,|\, x_i \neq 0\}.$$

Pour deux mots x et y on note $x \subset y$ si $\sigma(x) \subset \sigma(y)$. Dans ce cas on dit que "y couvre x".

Définition 7.4. On définit la distance de Hamming entre $x, y \in C \subseteq \mathbb{F}_q^n$ par

$$d_H(x, y) = \#\{i \in \{1, \ldots, n\} \,|\, x_i \neq y_i\}.$$

Le poids de Hamming de $x \in C$ est le nombre de ses cordonnées non nulles, c'est-à-dire

$$\mathrm{wt}(x) = |\sigma(x)|.$$

Si le code C est linéaire, alors en particulier $\mathrm{wt}(x) = d(x, 0)$.

Définition 7.5. La *distance minimale* $d(C)$ d'un code C est la plus petite distance non nulle entre deux mots du code C. Quand le code $C \subseteq \mathbb{F}_q^n$ est linéaire, alors on a

$$d(C) = \min\{\omega(x) \,|\, x \in C, x \neq \mathbf{0}\}.$$

Par convention, si $|C| = 1$ alors $d(C) = n + 1$, où n est la longueur du code.

Remarque 7.6.
- La distance de Hamming $d(\cdot, \cdot)$ est une distance. La vérification est laissée comme exercice.
- Si $D \subseteq C \subseteq \mathbb{F}_q^n$ alors $d(D) \geq d(C)$.

On dit que C est un (n, M, d) code si C a longueur n, distance minimale d et $|C| = M$. Si C est linéaire, on parle de $[n, k, d]_q$–code pour indiquer que C est de longueur n, que sa *dimension* en tant qu'espace vectoriel est k, et que sa distance minimale est d. Un $[n, k, d]_q$- code linéaire C contient q^k mots. En effet, soit $\{c_1, \ldots, c_k\}$ une base de C. Alors tous les éléments du code sont de combinaisons linéaires des c_i, et celles-ci sont q^k.

Exemple 7.7.
1. **Codes triviaux.** $\mathbb{F}_q^n$ et tous les ensembles $C \subseteq \mathbb{F}_q^n$ de cardinalité 1 sont de codes triviaux de longueur n. Ils ont dimension 1, et distance minimale 1 et $n + 1$, respectivement. Les seules codes linéaires triviaux sont $\mathbb{F}_q^n$ et $\{0\}$.

2. **Code à répétition.**

$$C = \{(\alpha, \ldots, \alpha) \,|\, \alpha \in \mathbb{F}_q\} \subseteq \mathbb{F}_q^n.$$

   Il a dimension 1 (quelle est une base ? Et sa matrice génératrice ?), et distance minimale n.

3. **Code de parité.** Pour $n \geq 2$ l'ensemble

$$C = \left\{ (x_1, \ldots, x_n) \in \mathbb{F}_q^n \,|\, x_n = -\sum_{i=1}^{n-1} x_i \right\} \subseteq \mathbb{F}_q^n,$$

   définit un code de longueur n et dimension $n - 1$. Quelle est sa distance minimale ?

Nous allons maintenant explorer la relation entre la distance minimale et la capacité de détecter et corriger des erreurs.

Proposition 7.8. *Un code de distance minimale d peut détecter jusqu'à $d - 1$ erreurs.*

Démonstration. Soit $c \in C$ un mot du code envoyé sur un canal. On suppose qu'il subisse $t < d$ erreurs et soit transformé en $x \in \mathbb{F}_q^n$. Alors $d(c, x) = t < d$. Puisque la distance minimale est d on détecte que $x \notin C$. □

Proposition 7.9. *Un code de distance minimale d peut corriger jusqu'à $\lfloor \frac{d-1}{2} \rfloor$ erreurs.*

Démonstration. Soit C un code de distance minimale d. Supposons que le mot du code $c \in C$ soit émit et subit t erreurs, avec $t < d/2$. Soit $x \in \mathbb{F}_q^n$ le mot reçu. On a alors $d(c, x) = t$, et pour tout autre mot de code $c' \in C$, $c' \neq c$, on a $d(c', x) > d(c, x)$. En effet, l'inégalité triangulaire donne

$$d \leq d(c, c') \leq d(c, x) + d(x, c'),$$

et donc comme $d(c, x) < d/2$ alors $d(x, c') > d/2$. En conclusion, le mot du code le plus proche à x, qui est unique si $t < d/2$, est le mot c qui a été émis. Donc on corrige x à c. □

On s'intéresse maintenant à la capacité de correction d'effacements d'un code.

Définition 7.10. On appelle *vecteur d'effacement* un vecteur v_E de $\{0, 1\}^n$ dont le support correspond à l'ensemble E des coordonnées effacés dans un mot de code émis.

Proposition 7.11. *Un code linéaire C peut corriger les effacements dans un ensemble E si et seulement si le vecteur d'effacement v_E ne couvre aucun mot non nul de C, c'est-à-dire si pour tout $c \in C, c \neq \mathbf{0}$ on a $c \not\subset v_E$.*

Démonstration. Soit E une configuration d'effacement qui n'est pas corrigible par le code C. Alors, il existe au moins deux mots de codes $c, c' \in C$, $c \neq c'$ qui coïncident en dehors des positions effacées. Alors, le vecteur $c - c'$, qui est un mot du code C par linéarité, est couvert par v_E, *i.e.*, $c - c' \subset v_E$. □

Corollaire 7.12. *Un code linéaire de distance minimale d peut corriger jusqu'à $d - 1$ effacements.*

Il s'en suit qu'on aimerait avoir de codes avec une dimension et une distance minimale aussi large que possible. Nous allons étudier de bornes pour les paramètres de codes linéaires.

7.2 Bornes sur les paramètres

La borne de Gilbert–Varshamov montre qu'il existe toujours un code de cardinalité et distance minimale "assez grandes".

Définition 7.13. On définit la balle de Hamming de rayon $0 \leq r \leq n$ centré en $x \in \mathbb{F}_q^n$ par

$$B(x, r) = \{c \in \mathbb{F}_q^n \mid d(c, x) \leq r\} \subseteq \mathbb{F}_q^n.$$

La cardinalité de $B(x, r)$ ne dépend pas du choix de x. En effet, on a

$$|B(x, r)| = \sum_{i=0}^{r} \binom{n}{i} (q-1)^i.$$

Proposition 7.14 (Borne de Gilbert-Varshamov)**.** *Il existe un code $C \in \mathbb{F}_q^n$ de distance minimale $d(C) \geq d$ pour $0 < d < n+1$ entier, et*

$$|C| \geq \frac{q^n}{\sum_{i=0}^{d-1} \binom{n}{i}(q-1)^i} = \frac{q^n}{|B(x, d-1)|}.$$

Démonstration. On commence par remarquer que si $d = n+1$ alors $|B(x,n)| = q^n$ et donc l'enoncé du théorème devient $|C| \geq 1$, pour lequel on peut prendre $C = \{0\}$. Donc pour la suite on supposera $d \leq n$.

Soit C le code de cardinalité maximale parmi les codes de distance minimale $\geq d$. Pour tout $x \in \mathbb{F}_q^n$ il existe un $c \in C$ tel que $d(x,c) \leq d-1$, car sinon le code $C \cup \{x\}$ aurait cardinalité plus grande que C et distance minimale $\geq d$. Donc $\mathbb{F}_q^n$ est contenu dans l'union de balle de Hamming de rayon $d-1$ centré en les mots du code de C, *i.e.,*

$$\mathbb{F}_q^n \subseteq \bigcup_{c \in C} B(c, d-1).$$

En conclusion on a

$$q^n = |\mathbb{F}_q^n| \leq |\bigcup_{c \in C} B(c, d-1)| \leq \sum_{c \in C} |B(c, d-1)| = |C| \sum_{i=0}^{d-1} \binom{n}{i}(q-1)^i,$$

d'où $|C| \geq \frac{q^n}{|B(x,d-1)|}$. □

Proposition 7.15 (Borne de Singleton)**.** *Pour un (n, M, d) code $C \subseteq \mathbb{F}_q^n$ on a $M \leq q^{n+1-d}$. En particulier, si C est linéaire, alors*

$$d + k \leq n + 1.$$

Démonstration. Paul fera la preuve en cours le 12/03. □

Définition 7.16. Un code dont les paramètres atteignent la borne de Singleton est appelé MDS (Maximum Distance Separable).

Proposition 7.17 (Borne de Hamming)**.** *Soit $C \subseteq \mathbb{F}_q^n$ un code de cardinalité ≥ 2 et distance minimale $d(C) \geq d$. Soit $t = \lfloor (d-1)/2 \rfloor$. Alors*

$$|C| \leq \frac{q^n}{\sum_{i=0}^{t} \binom{n}{i}(q-1)^i}.$$

Démonstration. Les balles de Hamming de rayon t centrées en les mots du code C sont deux à deux disjointes. En effet, si $x \in B(c,t) \cap B(d,t)$, alors on aurait

$$d(c,x) + d(x,c') \leq 2t < d \leq d(C).$$

Du coup, on a

$$\sum_{c \in C} |B(c,t)| = |\bigcup B(c,t)| \leq |\mathbb{F}_q^n| = q^n.$$

Puisque $\sum_{c \in C} |B(c,t)| = |C| \sum_{i=0}^{t} \binom{n}{i}(q-1)^i$, on conclut. □

Définition 7.18. Un code dont les paramètres atteignent la borne de Hamming est dit *parfait.*

7.3 Théorème de Shannon pour le canal binaire à effacements

Lemme 7.19. *Soit M une matrice de taille $k \times (k+x)$ à coefficients dans $\mathbb{F}_q$, tirée aléatoirement et uniformément. Alors*

$$P(\text{rg } M < k) \leq \frac{1}{q^x},$$

où rg M *désigne le rang de M.*

Démonstration. Par induction sur k.

Soit $k = 1$. Alors $\text{rg}M < 1 = k$ si et seulement si la matrice M est le vecteur nul de longueur $1 + x$. Cela arrive avec probabilité $1/q^{1+x} < 1/q^x$.

Soit M une matrice de taille $k \times (k + x)$ et supposons que l'enoncé est vrai pour les matrices de taille $(k-1)\times y$ pour tout y. Soit M_{k-1} la matrice constituée des $k - 1$ premières lignées de M. Alors, on peut écrire

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

Par hypothèse de récurrence on a

$$P(\text{rg } M_{k-1} < k-1) \leq 1/q^{1+x}.$$

La probabilité que la dernière ligne de M appartienne à un sous-espace donné de dimension au plus $k - 1$ vaut

$$P(\text{rg } M < k | \text{rg } M_{k-1} = k-1) \leq \frac{q^{k-1}}{q^{k+x}} = \frac{1}{q^{1+x}}.$$

En conclusion, on a obtenu

$$P(\text{rg } M < k) \leq \frac{1}{q^{1+x}} + \frac{1}{q^{1+x}} \leq \frac{1}{q^x}.$$

□

Lemme 7.20. *Soient $X_1, \ldots, X_n$ une suite de variables aléatoires de Bernoulli, indépendantes et de même paramètre α. Soit $\alpha < \beta < 1$. Alors*

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

Démonstration. Très similaire à l'exercice 3 du DSI !

On a

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

Pour tout $r \geq 0$, on peut écrire

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

ce qui donne

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

En posant

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

pour $\beta \geq \alpha$ on obtient

$$\sum_{i\geq\beta n}\binom{n}{i}\alpha^i(1-\alpha)^{n-i} \leq \left(\frac{\alpha}{\beta}\right)^{\beta n}\left(\frac{1-\alpha}{1-\beta}\right)^{n-\beta n}.$$

□

Rappel : la capacité du canal à effacement de probabilité de transition p est $1-p$. Dans ce canal une erreur correspond à une effacement, donc la probabilité maximale d'une erreur de décodage λ_n est la probabilité maximale qu'une configuration d'effacements $E \subseteq \{1, 2\ldots, n\}$ soit non corrigible.

Théorème 7.21 (Deuxième théorème de Shannon pour le canal à effacements)**.** *On considère un canal binaire à effacements de probabilité de transition p. Soit $R = 1 - p - \epsilon$ pour $\epsilon > 0$ fixé. Soit $C \subseteq \mathbb{F}_2^n$ le code binaire engendré par une matrice $\boldsymbol{G}$ de taille $k \times n$ (lignes - colonnes) avec $k = Rn$, 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 $\boldsymbol{G}$, que la configuration d'effacements $E \subseteq \{1, 2\ldots, n\}$ soit non corrigible tend vers 0 lorsque $n \to \infty$.*

Démonstration. Soient $\boldsymbol{G}_E$ et $\boldsymbol{G}_{\overline{E}}$ les sous-matrices de G constituées des coordonnées effacées et non effacées, respectivement. D'après la Proposition 7.11, la configuration E est non corrigible si et seulement sil existe un mot de code non nul de support inclut dans E. Donc, E est non corrigible si et seulement s'il existe une combinaison linéaire non triviale des lignes de $\boldsymbol{G}_{\overline{E}}$ qui égale 0, c'est-à-dire si et seulement si $\operatorname{rg} G_{\overline{E}} \leq k$. Nous allons donc estimer la probabilité d'avoir $\operatorname{rg} G_{\overline{E}} \leq k$.

$$\begin{aligned}
P(\operatorname{rg} G_{\overline{E}} \leq k) =& P(\operatorname{rg} G_{\overline{E}} < k || E| > (p+\epsilon/2)n)P(|E| > (p+\epsilon/2)n) + \\
&+ P(\operatorname{rg} G_{\overline{E}} < k || E| \leq (p+\epsilon/2)n)P(|E| \leq (p+\epsilon/2)n) \\
\leq& P(\operatorname{rg} G_{\overline{E}} < k || E| > (p+\epsilon/2)n)P(|E| > (p+\epsilon/2)n) + \\
&+ \sum_{e\leq(p+\epsilon/2)n} P(\operatorname{rg} G_{\overline{E}} < k || E| = e)P(|E| = e),
\end{aligned}$$

et donc

$$P(\text{rg } G_{\overline{E}} \leq k) \leq P(|E| > (p+\epsilon/2)n) + \max_{e \leq (p+\epsilon/2)n} P(\text{rg } G_{\overline{E}} < k || E| = e).$$

On peut écrire

$$\max_{e \leq (p+\epsilon/2)n} P(\text{rg } G_{\overline{E}} < k || E| = e) = \max_{x \geq \epsilon n/2} P(\text{rg } G_{\overline{E}} < k || \overline{E}| \geq Rn + x).$$

En effet, si $|E| = e$, alors $|\overline{E}| = n - e$. Donc $|E| \leq (p + \epsilon/2)n$ implique $|\overline{E}| \geq n - (p+\epsilon/2)n$. En utilisant que $p = 1 - R - \epsilon$ on obtient $|\overline{E}| \geq Rn + \epsilon n/2$. Or, $G_{\overline{E}}$ est la matrice des colonnes de G qui ne sont pas effacées par E, donc est de taille $k \times y$ avec $y \geq Rn + x = k + x$. Donc, d'après le Lemme 7.19 on sait que

$$\max_{x \geq \epsilon n/2} P(\text{rg } G_{\overline{E}} < k || \overline{E}| \geq Rn + x) \leq \frac{1}{q^x} \leq \frac{1}{2^{n\epsilon/2}},$$

et d'après le Lemme 7.20 on a

$$P(|E| > (p + \epsilon/2)n) \leq \frac{1}{2^{nD(p+\epsilon/2 \| p)}}.$$

En conclusion, on en déduit

$$P(\text{rg } G_{\overline{E}} \leq k) \leq \frac{1}{2^{n\epsilon/2}} + \frac{1}{2^{nD(p+\epsilon/2 \| p)}} \leq 2 \cdot \frac{1}{2^{n \cdot \min\{n\epsilon/2, nD(p+\epsilon/2 \| p)\}}},$$

□

qui pour $n \to \infty$ va a zéro.

Remarque 7.22. La convergence du théorème est une fonction exponentielle de la longueur n :

$$\liminf_{n \to \infty} \frac{1}{n} \log \frac{1}{P_e} \geq f(\epsilon) > 0.$$

On peut mieux estimer cet exposant d'erreur (voire [6]).

7.4 Second théorème de Shannon pour le canal binaire symétrique

Rappel : la capacité du canal binaire symétrique de probabilité de transition p est $1 - h(p)$, où $h(p)$ désigne l'entropie d'une loi de Bernoulli $(p, 1-p)$. Par la suite, on suppose $p \leq \frac{1}{2}$.

Le but de section est de prouver le second théorème de Shannon pour le canal binaire symétrique. Avant, on doit décrire l'erreur maximal de décodage dans ce canal.

Décodage au maximum de vraisemblance Dans le canal binaire symétrique une erreur correspond à un bit-flip: $0 \to 1$, $1 \to 0$. Soient X et Y les variables aléatoires à valeurs dans $\mathcal{X}^n$ et $\mathcal{Y}^n$, respectivement, qui désignent les messages émis et reçus. Lorsqu'on reçoit un message $y \in \mathcal{Y}^n$, le problème du décodage consiste à rechercher le mot du code $c \in \mathcal{C} \subseteq \mathcal{X}^n$ le plus probable, *i.e.,* qui maximise

$$P(X = c|Y = y).$$

Proposition 7.23. *Soit y le message reçu. On suppose que le mot du code émis X suit une loi uniforme sur le code $\mathcal{C}$. Alors, le mot du code $c \in \mathcal{C}$ le plus vraisemblable est le mot du code le plus proche du mot reçu y, au sens de la distance de Hamming.*

Démonstration. On a

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

Sous les hypothèses de la proposition $\frac{P(X=c)}{P(Y=y)}$ est constante. Par conséquence, maximiser $P(X = c|Y = y)$ revient à maximiser $P(Y = y|X = c)$. Nous avons

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

Puisque $p \leq 1/2$, cette quantité est maximale lorsque $d(y, c)$ est minimale. □

Remarque 7.24. Il est possible qu'il y a ait plusieurs mots du codes à distance minimale de y, et donc plusieurs mots du code également vraisemblable.

Théorème 7.25. *On considère 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 aléatoire G de taille $k \times n$ avec $k = Rn$, sous forme systématiques, i.e., de la forme $[I_k|A]$, où A est choisie de manière uniforme dans l'espace des matrices binaires de taille $k \times (n-k)$. La probabilité P_e sûr à la fois l'action du canal et le choix de G, que le décodage au maximum de vraisemblance échoue, tend vers 0 lorsque $n \to \infty$.*

Par la suite, on notera $e \in \{0, 1\}^n$ un vecteur d'erreur, et $y = c + e$ le mot reçu quand le mot du code c est envoyé. D'après la Proposition 7.23, il n'y a pas d'erreur de décodage si

$$d(y, c') \leq d(y, c) \, \forall c' \in \mathcal{C}. \tag{5}$$

Puisque $d(y,c) = d(c+e,c) = d(e,0)$ et $d(y,c') = d(c+e,c') = d(e,c+c')$, la condition (5) équivaut à dire qu'il n'existe pas un mot du code non nul x tel que

$$d(x,e) \leq d(0,e) = \mathrm{wt}(e). \tag{6}$$

Définition 7.26. On définit la balle de Hamming de rayon $0 \leq r \leq n$ centré en $x \in \mathbb{F}_q^n$ par

$$B(x,r) = \{c \in \mathbb{F}_q^n \,|\, d(c,x) \leq r\} \subseteq \mathbb{F}_q^n.$$

La cardinalité de $B(x,r)$ ne dépend pas du choix de x. En effet, on a

$$|B(x,r)| = \sum_{i=0}^{r} \binom{n}{i} (q-1)^i.$$

Remarque 7.27. En particulier, pour $q = 2$ on obtient

$$|B(x,r)| = \sum_{i=0}^{r} \binom{n}{i}.$$

De plus, en utilisant le Lemme 3.14 on obtient l'estimation suivante

$$|B(x,r)| = \sum_{i=0}^{r} \binom{n}{i} \leq 2^{nh(t/n)}.$$

La condition (6) peut donc être récrite comme

$$\mathcal{C} \cap B(e, \mathrm{wt}(e)) = \{0\}.$$

On s'intéresse donc à la probabilité de cet événement lorsque $\mathcal{C}$ est un code aléatoire. Nous considérerons de matrices génératrices G de la forme $[I_k|A]$.

Lemme 7.28. *Soit $x \in \{0,1\}^n$ un vecteur non-nul. Alors $P(x \in \mathcal{C}) \leq \frac{1}{2^{n-k}}$.*

Démonstration. Soient $(x_1, \ldots, x_k)$ les premières k coordonnés de x. Alors, $x \in \mathcal{C}$ si et seulement si $(x_1, \ldots, x_k)G = x$, si et seulement si

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

Nous avons alors deux cas. Soit $x_1, \ldots, x_k) = \mathbf{0}$ et $(x_{k+1}, x_{k+1}, \ldots, x_n) \neq \mathbf{0}$, et donc $P(x \in \mathcal{C}) = 0$. Soit $x_1, \ldots, x_k) \neq \mathbf{0}$, donc $x \in \mathcal{C}$ si et seulement si le vecteur aléatoire $(x_1, \ldots, x_k)A$ est égale au vecteur fixé $(x_{k+1}, x_{k+1}, \ldots, x_n)$, ce qui arrive avec probabilité $\frac{1}{2^{n-k}}$ puisque A est choisie de façon uniforme dans $M_{k \times n-k}(\mathbb{F}_2)$. □

Proposition 7.29. *Pour tout $t \leq n/2$ on a*

$$P(C \cap B(e, \mathrm{wt}(e)) \neq \{0\} | \mathrm{wt}(e) = t) \leq 2^{nh(t/n)-(n-k)}.$$

Démonstration. L'événement $C \cap B(e, \mathrm{wt}(e)) \neq \{0\}$ est la réunion des événements $x \in \mathcal{C}$ pour $x \in B(e, \mathrm{wt}(e))$. On a donc

$$P(C \cap B(e, \mathrm{wt}(e)) \neq \{0\} | \mathrm{wt}(e) = t) \leq |B(e,t)| \frac{1}{2^{n-k}} \leq 2^{nh(t/n)-(n-k)}.$$

□

Démonstration (Second théorème de Shannon). La fonction d'entropie binaire h est strictement croissante sur l'intervalle $[0, 1/2]$. On peut définit alors θ comme l'unique réel $0 < \theta < 1/2$ tel que $R = 1 - h(\theta)$, et puisque $R < h(p)$, nous avons $p < \theta$. Alors

$$\begin{aligned} P(C \cap B(e, \mathrm{wt}(e)) \neq \{0\}) &= \sum_{t=0}^{n} P(C \cap B(e, \mathrm{wt}(e)) \neq \{0\} | \mathrm{wt}(e) = t) P(\mathrm{wt}(e) = t) \\ &\leq \max_{t \leq n(p + \frac{\theta - p}{2})} P(C \cap B(e, \mathrm{wt}(e)) \neq \{0\} | \mathrm{wt}(e) = t) + \\ &\quad + P\left(\mathrm{wt}(e) > n\left(p + \frac{\theta - p}{2}\right)\right). \end{aligned}$$

En utilisant la Proposition 7.29 et le Lemme 7.20, on obtient

$$P(C \cap B(e, \mathrm{wt}(e)) \neq \{0\}) \leq \frac{1}{2^{n\left[h(\theta) - h(p + \frac{\theta - p}{2})\right]}} + \frac{1}{2^{nD(p + \frac{\theta - p}{2} \| p)}},$$

qui quand n tend vers l'infini donne le résultat attendu. □

Bibliographie

[1] Venkatesan Guruswami, Atri Rudra, et Madhu Sudan. *Essential coding theory.* Draft available at http://www. cse. buffalo. edu/atri/courses/coding-theory/book 2.1 (2012).

[2] Venkatesan Guruswami et Mahdi Cheraghchi. Information Theory and its applications in theory of computation, (2013).

[3] David A. Huffman. *A method for the construction of minimum-redundancy codes.* Proceedings of the IRE 40.9 (1952): 1098-1101.

[4] Nicolas Sendrier, Introduction à la théorie de l'information (2007).

[5] Claude Elwood Shannon. *A mathematical theory of communication.* The Bell System Technical Journal 27.3 (1948): 379-423.

[6] Gilles Zèmor, Mémento de théorie de l'information (2015).