École Internationale des Sciences du Traitement de l'Information

Département "Systèmes Informatiques Formels et Intelligents"

Introduction à la Théorie de l'Information

Notes de cours - Version préliminaire

Anya Désilles

Année scolaire 2007/2008

Table des matières

Introduction

La théorie de l'information trouve ses origines dans les débuts des communications électriques. Le premier télégraphe élaboré par S. Morse entre 1832 et 1838 a permis de communiquer n'importe quel texte à l'aide de signaux électriques. Dans le code de Morse chaque lettre de l'alphabet est représentée par une séquence de

- **points** (courant électrique de courte durée) ;
- **traits** (courant de longue durée) ;
- **espaces** (absence de courant) ;

On peut remarquer que dans la version définitive du code la lettre "e", la plus fréquente dans l'anglais, est représentée par la séquence la plus courte : un seul point. À cette époque, S. Morse n'avait pas fait d'analyse théorique pour arriver à cette conclusion, mais plutôt des observations empiriques. Son but en effet était de concevoir le code tel que la saisie d'un texte par un opérateur soit, en moyenne, la plus rapide possible. Aujourd'hui, la théorie moderne de communication a montré qu'il est possible de gagner à peine 15 pour cent de temps de saisie par rapport au code de Morse.

Quelques années plus tard, une première ligne télégraphique est installée entre Washington et Baltimore. En enterrant les câbles, S. Morse rencontre une nouvelle difficulté : le milieu dans lequel ces derniers se trouvent influe sur la qualité de transmission. En particulier, il remarque que si l'opérateur saisit trop vite son code, le signal reçu est indéchiffrable. Les points, les traits, les espaces sont confondus dans un seul signal d'intensité moyenne. Ainsi apparaît le problème de perturbations dues aux conditions physiques de transmission de l'information et, comme condition de bonne réussite, la nécessité de limiter le débit d'émission de symboles.

Bien plus tard, dans les années 1940, l'ingénieur et mathématicien C. Shannon et le mathématicien Warren Weaver, ont formalisé le processus de communication et donné les outils mathématiques permettant de répondre aux questions posées empiriquement par les premiers télégraphes. Ils proposent en particulier, une représentation schématique de processus de communication, connue sous le nom de paradigme de Shannon.

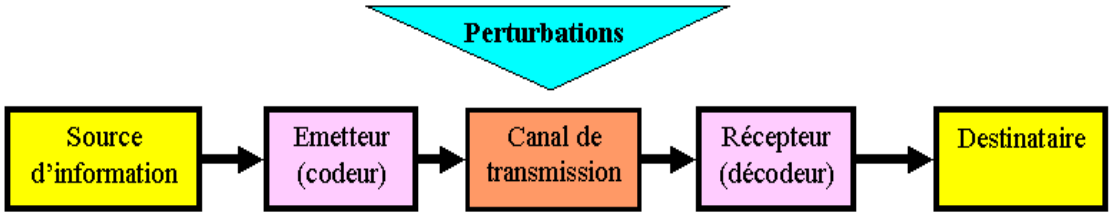

FIG. 1 – Paradigme de Shannon

Ce modèle est une **approximation linéaire de processus de communication** qui met l'accent sur les aspects purement techniques de transmission d'un message. On peut résumer ce modèle de façon suivante :

1. la source d'information choisit un message M parmi un certain nombre de messages possibles ;
2. l'émetteur transforme le message en signal S compatible physiquement avec le mode de transmission choisi. On dit qu'il **encode** le message ;
3. le signal S est alors soumis à l'entrée d'un canal de transmission ;
4. lors de la transmission des perturbations peuvent intervenir et transformer le signal envoyé ; on parle alors **de bruit de canal** B ;
5. à la sortie du canal, le signal $\tilde{S}$ éventuellement entaché d'erreurs dues au bruit, est soumis au décodeur qui le transforme en message $\tilde{M}$ lisible par le destinataire.

Voici un exemple pour illustrer ce schéma. Imaginez qu'un ami vous envoie une carte postale du lieu de ses vacances. La carte voyage sans incident jusqu'à la ville où vous habitez. Le jour où votre facteur doit enfin vous l'apporter, il la fait tomber par mégarde. Malheur ! Il pleut. L'enveloppe a pris l'eau. C'est ainsi que vous retrouvez les quelques lignes écrites par votre ami à moitié illisibles. Allez vous pouvoir reconstituer le contenu complet du message ?

Dans cet exemple, **la source d'information** est votre ami, et plus précisément, son cerveau. C'est encore lui qui joue ici le rôle de **l'émetteur**, en transformant ses idées en mots et ensuite en écrivant les mots à l'aide de lettres de l'alphabet. **Le canal de transmission** est ici représenté par les services postaux. Les dégâts causés par l'eau à la carte postale représentent **le bruit de canal**. Les questions que l'on pourrait poser dans cette situation sont les suivantes :

1. Y a-t-il un moyen de préparer le message de façon à éviter les désagréments causés par les erreurs de transmission et garantir à l'arrivée la lisibilité du message ? Par

exemple, pourrait-on écrire avec un encre indélébile ? **C'est le problème de codage de source**.

2. Y a-t-il un moyen de transporter le signal dans le canal de manière à limiter les perturbations ? Par exemple, protéger le courrier dans des enveloppes imperméables ? **C'est le problème de codage de canal.**
3. Comment évaluer l'incertitude que le récepteur a sur le contenu réel du message reçu ? **C'est le problème de mesure de l'information.**

Ce modèle, certes très simplifié, de la communication a servi de base dans le développement de la théorie de l'information à partir des années 1940. Il s'inscrit dans une description plus générale, donnée par Warren Weaver des trois niveaux de problèmes de communication :

Niveau 1 : Technique Avec quelle précision peut on transmettre les symboles de la communication ?

Niveau 2 : Sémantique Dans quelle mesure les symboles véhiculent la signification ?

Niveau 3 : Efficacité Dans quelle mesure la signification reçue influence le comportement et l'action du destinataire ?

La théorie de l'information s'intéresse uniquement aux problèmes du premier niveau de communication. En particulier, le sens des messages traités n'a aucune importance. L'information quantifiable associée à un message donné n'est pas dans son sens sémantique mais dans sa rareté. Un message, même très significatif, connu à l'avance par le récepteur ne lui apporte aucune information au sens technique du terme. Par contre, la réception d'un message très improbable mais démuni de tout sens est dans ce cadre considérée comme événement porteur d'une grande quantité d'information.

Les chapitres qui suivent ont pour objectif d'introduire les concepts principaux de la théorie de l'information et les théorèmes fondamentaux. Ces derniers établissant les limites théoriques en matière de codage de l'information et de transmission. Enfin, nous consacrerons la deuxième moitié du cours à l'étude de quelques algorithmes de base de codage de source.

Chapitre 1

Modélisation mathématique de l'information

1 Une source

1.1 Représentation mathématique d'une source

Soit une source (un émetteur) produisant des symboles d'un alphabet fini $\Omega = \{\omega_1, \ldots, \omega_m\}$. Nous supposons que chaque symbole $\omega_i,\ i = 1, 2, \ldots, m$ est émis aléatoirement avec une probabilité connue p_i de telle sorte que $\sum_{i=1}^{m} p_i = 1$.

On considère alors l'expérience aléatoire consistant à observer un symbole émis. À cette expérience nous associons une variable aléatoire, notée X à valeurs dans Ω. C'est une variable aléatoire discrète, dont la loi est définie par l'ensemble des probabilités d'émission des symboles de l'alphabet :

$$P[X = \omega_i] = p(\omega_i) = p_i,\ i = 1, \ldots, m$$

Exemple 1.1. Soit X la variable associée à une source binaire, produisant des symboles 0 et 1 avec les probabilités respectives p et $q = 1 - p$. Nous avons ici l'alphabet constitué de deux symboles $\Omega = \{0, 1\}$ et $P[X = 0] = p, \quad P[X = 1] = q$.

Exemple 1.2. Soit Y la variable associée à une source disposant d'un vocabulaire de 20 mots, $\Omega = \{$le, la, pomme, fruit, est, un, arbre, pommier, jardin, ombre, fait, soleil, vent, pluie, grandit, quand, ne, pas, mais, oiseau$\}$ et que tous les mots sont équiprobables.

Ainsi, n'importe quel mot $M_i \in \Omega$ du vocabulaire est choisi avec la probabilité $\frac{1}{20}$ et $P[X = M_i] = \frac{1}{20},\ i = 1, \ldots, 20.$

Dans la suite nous allons considérer des événements associées à la variable aléatoire X comme des sous-ensembles de Ω et la mesure de probabilité définie pour $A \subset \Omega$ par

$$P[A] = P[X \in A].$$

Si on considère l'émission de plusieurs symboles successifs, à des instants de temps précis $\{t_1, t_2, \ldots, t_n\}$ on peut y associer la suite de variables aléatoires $\{X_1, X_2, \ldots, X_n\}$, chacune correspondant à l'émission d'un symbole $s(t) \in \Omega,\ t \in \{t_1, t_2, \ldots, t_n\}$. Nous supposons pour l'instant que tous les symboles successifs sont émis de façon indépendante les uns des autres et avec les mêmes probabilités d'émissions : $p_i = P[s(t) = \omega_i]$. Cela signifie que les variables aléatoires $X_k,\ k = 1, \ldots, n$ sont **deux à deux indépendantes et identiquement distribuées**. On parle alors d'une source stationnaire et sans mémoire. Nous allons préciser la signification formelle de ces termes plus loin.

1.2 Information propre

Pour introduire l'ensemble des axiomes qui définissent l'information propre d'un événement $A \subset \Omega$, nous allons nous servir du premier exemple de source binaire ci-dessus (voir 1.1). Soit donc une source binaire d'alphabet $\Omega = \{0, 1\}$ et la variable aléatoire X associée avec les probabilités données :

$$P(0) = P[X = 0] = p \text{ et } P(1) = P[X = 1] = q = 1 - p$$

Dans ce qui suit, nous allons construire de façon axiomatique, une fonction

$$h : \mathcal{P}(\Omega) \to \mathbb{R}$$

associant à un événement $A \subset \Omega$ la mesure de la quantité d'information contenue dans A. Nous l'appellerons **"information propre d'un événement** A**"**.

Dans la discussion du chapitre 1 nous avons déjà indiqué de façon intuitive que la quantité d'information apportée par un événement devrait être d'autant plus grande que l'événement est rare, improbable. Nous pouvons donc postuler que que la quantité d'information d'un événement doit être de la forme :

$$h(A) = f\left(\frac{1}{P(A)}\right)$$

où f est une fonction croissante.

En particulier, si, dans notre exemple de source binaire $A = \{\omega\}$, avec $\omega \in \Omega$ on doit avoir

$$h(\omega) = f\left(\frac{1}{P(\omega)}\right),\ \omega \in \{0, 1\}$$

Si un événement A est certain, c'est-à-dire si $P[A] = 1$, il n'apporte aucune information. La fonction f que l'on recherche doit donc vérifier la condition limite suivante :

$$\lim_{x \to 1} f(x) = 0$$

Par ailleurs, il serait logique de s'attendre à ce que l'information apportée par deux événements indépendants, A et B, soit égale à la somme des informations apportées par chacun de ces événements :

$$h(A \cap B) = h(A) + h(B)$$

et donc

$$h(A \cap B) = f\left(\frac{1}{P(A \cap B)}\right) = f\left(\frac{1}{P(A)} \cdot \frac{1}{P(B)}\right)$$

Ainsi, la fonction f doit vérifier la propriété suivante :

$$f(xy) = f(x) + f(y)$$

L'unique fonction (à une constante multiplicative près) qui vérifie toutes ces propriétés est la fonction logarithme. Ainsi, nous définissons l'information propre d'un événement de façon suivante.

Définition 1.1 (Information propre)**.** Soient $\Omega = \{\omega_1, \ldots, \omega_m\}$ un alphabet discret et X la variable aléatoire associée. Pour tout événement $A \subset \Omega$ la quantité d'information propre de A est définie par

$$h(A) = -\log(P(A)).$$

Remarque 1.1. Dans la suite nous allons considérer le logarithme de base 2. L'unité de mesure de l'information est alors le *Shannon*. la changement de base e logarithme provoque la modification de la constante multiplicative.

Voici quelques propriétés élémentaires de la fonction $h(A)$ qui sont les conséquences immédiates des propriétés de la fonction logarithme.

Proposition 1.1 (Propriétés de l'information propre)**.** *1. La quantité d'information est toujours une grandeur positive :* $\forall A \subset \Omega$, $h(A) \geq 0$ *et* $h(A) = 0 \Leftrightarrow P(A) = 1$. *Elle est nulle si et seulement si l'événement A est certain.*

2. $\lim_{P(A) \to 0} h(A) = +\infty$.

1.3 Information conditionnelle

Nous avons déjà évoqué dans la section précédente la question d'évaluation de la quantité d'information apportée par la réalisation conjointe de deux événements A et B. Nous avons postulé que dans le cas de deux événements A et B indépendants on a

$$h(A \cap B) = h(A) + h(B).$$

Or, dans le cas où A et B ne sont pas indépendants, on peut utiliser la notion de probabilité conditionnelle définie par (voir le cours "Probabilités" de Marietta Manolessou) :

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

d'où les relations :

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

Alors, la quantité d'information propre de $A \cap B$ est :

$$h(A \cap B) = -\log(P(A)P(B|A)) = -\log(P(A)) - \log(P(B|A)) = h(A) - \log(P(B|A)).$$

Définition 1.2 (Information conditionnelle)**.** On appelle information conditionnelle de B sachant A la quantité

$$h(B|A) = -\log(P(B|A)).$$

Remarque 1.2. On peut donc écrire $h(A \cap B)$ de la façon suivante :

$$h(A \cap B) = h(A) + h(B|A).$$

On peut interpréter la quantité d'information conditionnelle comme suit : $h(B|A)$ représente l'information apportée par l'observation de l'événement B *qui n'a pas déjà été* apportée par l'observation de A.

1.4 Information mutuelle

Définition 1.3 (Information mutuelle)**.** On appelle information mutuelle de deux événements B et A la quantité

$$i(A;B) = h(A) - h(A|B) = \log \frac{P(A|B)}{P(A)} = \log \frac{P(A \cap B)}{P(A)P(B)}.$$

1.5 Information *vs* incertitude

La définition de l'information associée à un événement A implique que cette quantité est d'autant plus grande que l'événement est rare (c'est-à-dire que sa probabilité est proche de zéro). Ainsi, on peut assimiler l'information apportée par l'observation de A et l'incertitude que nous avons sur sa réalisation, ou encore la difficulté de prévoir A.

Considérons par exemple l'expérience de lancé d'un dé équilibré à 6 faces. Soit X la variable aléatoire associée :

$$\forall \omega \in \{1,2,3,4,5,6\}, \quad P[X=\omega] = \frac{1}{6}.$$

Imaginons que le résultat soit $X = 2$. Soit A l'événement associé. On a $P(A) = \frac{1}{6}$.

Si on essaye de deviner ce résultat, on a une chance sur 6 de trouver la réponse correcte. L'information associée est

$$h(A) = -\log\left(\frac{1}{6}\right) = \log(6) \simeq 2.585$$

Imaginons que l'on affirme que le résultat du lancé est un nombre pair. Soit B l'événement associé. On a $P(B) = \frac{1}{2}$. Alors, la quantité d'information que nous apporte ce renseignement est

$$h(B) = -\log\left(\frac{1}{2}\right) = \log(2) = 1$$

Cette information, augmente notre chance de réussite. En effet, nous avons maintenant la probabilité $P(A|B) = \frac{1}{3}$ de deviner le résultat. Et nous retrouvons la décomposition de l'information mutuelle associée à $A \cap B$ en somme de l'information apportée par l'observation de B et l'information conditionnelle $h(A|B)$:

$$h(A \cap B) = h(B) + h(A|B) = -\log\left(\frac{1}{2}\right) - \log\left(\frac{1}{3}\right) = \log(2) + \log(3) = \log(6).$$

1.6 Entropie d'une source

On considère maintenant une source stationnaire et sans mémoire, produisant à des instants de temps précis $\{t_1, t_2, \ldots, t_k\}$ des symboles $s(t) \in \Omega,\ t \in \{t_1, t_2, \ldots, t_k\}$ d'un alphabet Ω donné.

L'entropie d'une telle source représente la quantité moyenne d'information propre associée à l'observation de chacun des symboles possibles.

Supposons que l'alphabet de la source $\Omega = \{s_i,\ i = 1, \ldots, n\}$ comporte n symboles et que la variable aléatoire associée à l'observation d'un symbole émis, X, ait la distribution de probabilité donnée $\{p_i = P[X = s_i],\ i = 1, \ldots, n\}$. L'entropie de la source est alors la fonction des n probabilités $H(p_1, p_2, \ldots, p_n)$.

Tout comme pour l'information propre, l'entropie peut être définie par l'ensemble d'axiomes suivant :

1. Si tous les symboles sont équiprobables l'entropie $H(1/n, \ldots, 1/n) = f(n)$ est une fonction de n, la taille de l'alphabet. Nous souhaitons qu'elle soit la mesure de l'incertitude associée à la source, ou encore de la difficulté à prédire le symbole émis. Il est alors naturel de supposer que l'entropie augmente avec la taille de l'alphabet. Par exemple, il est plus difficile de deviner le résultat de lancé d'un dé à 6 faces que celui d'une pièce de monnaie.

   Ainsi, la fonction $f(n) = H(1/n, \ldots, 1/n)$ ci-dessus doit être **croissante.**

2. Considérons une source qui émet des couples de symboles puisés chacun dans un alphabet indépendant de tailles respectives l et m. La taille de l'alphabet produit est alors $l \cdot m$. Or, l'observation de chaque couple apporte la somme d'informations liées à chaque symbole.

   Ainsi, la fonction f doit vérifier la **propriété d'additivité** $f(l \cdot m) = f(l) + f(m)$.

3. La fonction $H(p_1, \ldots, p_n)$ **est continue** selon chacune de ses variables.

4. **Propriété de groupes.** Si l'on divise l'alphabet donné Ω en deux parties disjointes, $\Omega_1 = \{s_1, \ldots, s_r\}$ et $\Omega_2 = \{s_{r+1}, \ldots, s_n\}$ chacune peut être considérée comme alphabet. Notons $Q_1 = P(\Omega_1) = \sum_{k=1}^{r} p_k$ et $Q_2 = P(\Omega_2) = \sum_{k=r+1}^{n} p_k$. On a $Q_1 + Q_2 = 1$. On peut alors définir sur Ω_1 et sur Ω_2 les distributions des probabilités respectives $\left\{\frac{p_1}{Q_1}, \ldots, \frac{p_r}{Q_1}\right\}$ et $\left\{\frac{p_{r+1}}{Q_2}, \ldots, \frac{p_n}{Q_2}\right\}$. Alors l'entropie associée à l'alphabet Ω se décompose de la façon suivante :

   $$H(p_1, \ldots, p_r, p_{r+1}, \ldots, p_n) = H(Q_1, Q_2) + Q_1 H\left(\frac{p_1}{Q_1}, \ldots, \frac{p_r}{Q_1}\right) + Q_2\left(\frac{p_{r+1}}{Q_2}, \ldots, \frac{p_n}{Q_2}\right)$$

5. **Symétrie.** Il est naturel de supposer que la quantité moyenne d'information associée à un alphabet Ω ne dépend pas de l'ordre de numérotation des symboles. Ainsi, la permutation de deux arguments de la fonction d'entropie ne la modifie pas :

   $$H(p_1, \ldots, p_i, \ldots, p_j, \ldots, p_n) = H(p_1, \ldots, p_j, \ldots, p_i, \ldots, p_n)$$

La fonction vérifiant ces quatre axiomes est une unique à une constate multiplicative près. Après la publication en 1948 par C. Shannon de son article "The Mathematical Theory of Communication", de nombreux mathématiciens ont travaillé l'élaboration de la définition axiomatique de l'entropie. On peut citer Khinchin (1957), Fadeev(1956),

Kullback 1958, Rényi 1962. En sélectionnant des ensembles d'axiomes différents, tous ont abouti à la même fonction. D'où la définition qui suit.

Définition 1.4 (Entropie d'une source)**.** Soient $\Omega = \{\omega_1, \ldots, \omega_m\}$ l'alphabet fini d'une source et X la variable aléatoire associée t.q. $P[\omega_i] = p_i, \quad i = 1, \ldots, m$. On appelle **entropie** ou encore **quantité moyenne d'information** de la source la quantité

$$H(X) = H(p_1, p_2, \ldots, p_n) = E[h(\omega)] = -\sum_{i=1}^{m} p_i \log(p_i)$$

L'unité de mesure de cette quantité est le "bit par symbole".

L'entropie représente la quantité d'information que l'on obtient en moyenne, en observant les symboles en sortie de la source pendant suffisamment longtemps ou encore, la valeur de l'information moyenne obtenue en observant simultanément les symboles en sortie d'un grand nombre de sources identiques.

Elle mesure également le nombre de bits moyens nécessaires pour coder chaque symbole de la source. C'est une limite théorique. Nous verrons plus tard, comment en pratique approcher cette limite. Les différents algorithmes de codage donnent plus ou moins satisfaction.

Exemple 1.3 (Exemple important : entropie d'une source binaire)**.** Soit une source émettant des symboles 0 avec la probabilité p et 1 avec la probabilité $q = 1 - p$. Alors l'entropie de cette source est

$$H_2(p) = -p \log(p) - (1 - p) \log(1 - p).$$

Lorsque, par exemple, les deux symboles sont équiprobables, on a $p = 1/2$ et alors

$$H_2(1/2) = -1/2 \log(1/2) - 1/2 \log(1/2) = 1.$$

Nous pouvons interpréter ce résultat comme suit : *lorsque les symboles d'une source binaire sont équiprobables, il faut un bit par symbole en moyenne.*

Il s'agit de la valeur maximale possible. Nous allons dans la suite utiliser cette fonction pour introduire ou illustrer dans ce cas simple les propriétés fondamentales de l'entropie.

Proposition 1.2 (Propriétés de l'entropie d'une source binaire)**.**
Soit $H_2(p),\ p \in [0,1]$ *l'entropie d'une source binaire définie dans l'exemple ci-dessous. En tant que fonction de* p *elle a les propriétés suivantes :*

1. $H_2(p)$ *est une fonction continue sur* $]0,1[$ *telle que*

$$\lim_{p\to 0+} H_2(p) = 0 = \lim_{p\to 1-} H_2(p)$$

2. $H_2(p)$ *est positive sur* $H_2(p),\ p \in [0,1]$
3. $H_2(p)$ *est symétrique par rapport à* $p_0 = 0.5$ *et atteint son maximum en* p_0 *t.q.* $H_2(0.5) = 1$.
4. $H_2(p)$ *est strictement concave :*

$$\forall (p_1,p_2) \in [0,1],\ p_1 \neq p_2, \quad \forall \lambda \in]0,1[$$

$$H_2(\lambda p_1 + (1-\lambda)p_2) > \lambda H_2(p_1) + (1-\lambda)H_2(p_2)$$

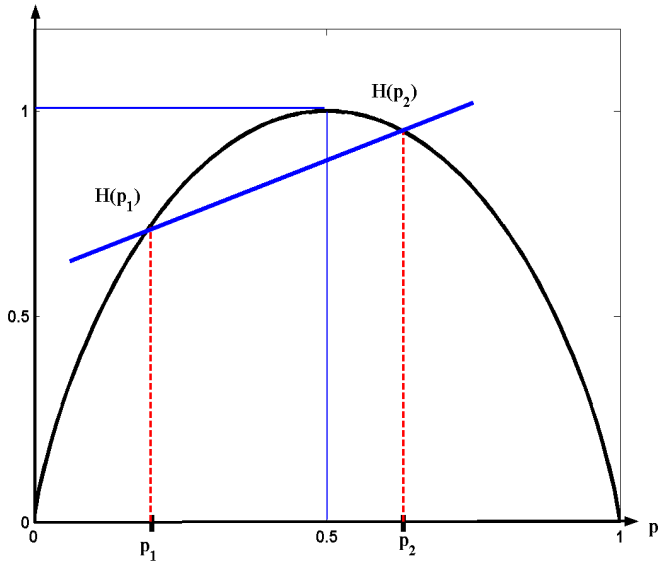

FIG. 1.1 – Fonction d'entropie d'une source binaire

Toutes ces propriétés sont faciles à vérifier avec les méthodes simples de l'analyse des fonctions à une variable réelle. Elles sont résumées par la représentation graphique de $H_2(p)$ donnée sur la figure 1.1.

1.6.1 Interprétations de la fonction d'entropie

Considérons l'exemple d'une source S dont l'alphabet est composé de 5 symboles $\Omega = \{a, b, c, d, e\}$ et la distribution de probabilité est la suivante :

X	a	b	c	d	e
P(X)	0.3	0.2	0.2	0.15	0.15

L'entropie de cette source, calculée selon la définition est

$$H(X) = -\sum_{i=1}^{5} p_i \log(p_i) = -(0.3\log(0.3) - 2 \cdot 0.2\log(0.2) - 2 \cdot 0.15\log(0.15) \simeq 2.27.$$

Entropie comme mesure d'information Nous avons déjà mentionné que l'entropie représente la quantité moyenne d'information apportée par l'observation des symboles de la source. On peut également l'interpréter comme la difficulté moyenne de prédire chaque symbole à la sortie.

Entropie comme mesure de nombre de bits pour le codage Imaginons que l'on essaye de deviner le symbole observé à la sortie de la source. Pour ce faire on peut poser des questions à un automate qui ne répond que par **"oui"** ou par **"non".** Voici un schéma représentant le déroulement du jeu :

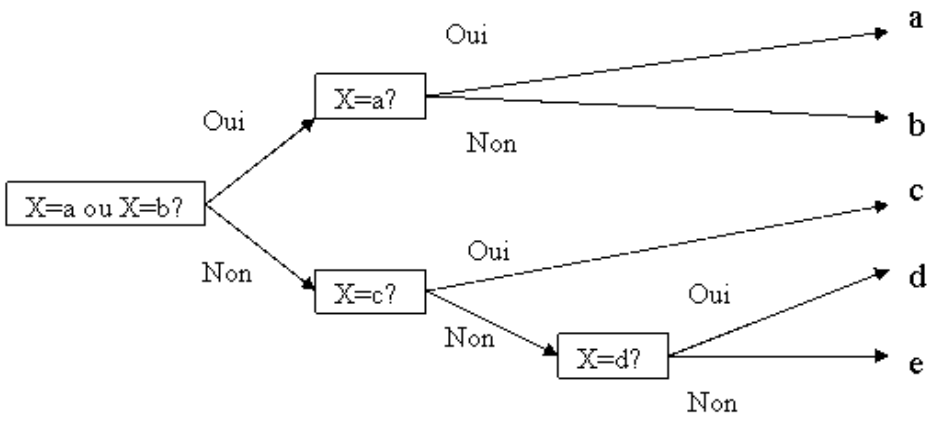

FIG. 1.2 – Questions

Il est facile à voir que le nombre de questions nécessaires pour deviner le symbole est une variable aléatoire Nb définie sur Ω de la façon suivante :

x	a	b	c	d	e
P(x)	0.3	0.2	0.2	0.15	0.15
Nb(x)	2	2	2	3	3

Alors le nombre moyen de questions nécessaires pour deviner le symbole est

$$\bar{Nb} = 2(0.3 + 0.2 + 0.2) + 3(0.15 + 0.15) = 2.3.$$

Si on utilise un bit pour coder la réponse à chaque question, cela nous permet de coder les symboles de cette source selon le schéma ci-dessus :

a	Oui-Oui	11
b	Oui-Non	10
c	Non-Oui	01
d	Non-Non-Oui	001
e	Non-Non-Non	000

Nous étudierons plus loin l'un des théorèmes fondamentaux de la théorie de l'information établissant que l'entropie d'une source est le plus petit nombre moyen de bits par symbole nécessaire pour coder les messages produits par cette source. Dans cet exemple, nous avons bien $\bar{Nb} = 2.3 > 2.27 = H(X)$.

1.6.2 Propriétés générales de la fonction d'entropie

Proposition 1.3 (Positivité)**.** *Soit S une source sans mémoire et stationnaire d'alphabet $\Omega = \{\omega_i\}_{i=1}^n$. Soit X la variable aléatoire associée de distribution de probabilité P donnée $P[X = \omega_i] = p_i, \quad i = 1, \ldots, n$. Notons $H(p_1, \ldots, p_n)$ sa fonction d'entropie. Alors*

$$H(p_1, p_2, \ldots, p_n) \geq 0.$$

En plus, l'égalité a lieu uniquement si l'une des probabilités p_i est égale à 1 et les autres sont nulles.

Preuve de la proposition 1.3

Sachant que $\forall i = 1, \ldots, n, \quad 0 \leq p_i \leq 1$ on a $-p_i \log(p_i) \geq 0$. Si si l'un des symboles de l'alphabet est certain ($p_i = 1$) alors tous les autres sont forcément impossibles ($\forall k \neq i, \quad p_k = 0$).
On a alors : $p_i \log(p_i) = 0$ et $\forall k \neq i,\ p_k \log(p_k) = 0$ (par continuité) et donc $H = 0$. **C.Q.F.D**

Lemme 1.1 (Inégalité de Gibbs)**.** *Soient $P = (p_i)_{i=1}^n$ et $Q = (q_i)_{i=1}^n$ deux distributions de probabilités sur le même univers fini $\Omega = \{\omega_i\}_{i=1}^n$. Alors*

$$\sum_{i=1}^{n} p_i \log\left(\frac{p_i}{q_i}\right) \leq 0 \tag{1.1}$$

*et l'égalité a lieu **si et seulement si** $\forall i = 1, \ldots, n, p_i = q_i$.*

Théorème 1.1 (Maximum de la fonction d'entropie)**.**

$$H(p_1, p_2, \ldots, p_n) \leq log(n) \tag{1.2}$$

et l'égalité a lieu si et seulement si $\forall i = 1, \ldots, n, \quad p_i = \frac{1}{n}$.

Preuve du théorème 1.1

Soient $P = (p_i)_{i=1}^n$ une distribution de probabilité sur l'univers fini

$$\Omega = \{\omega_i\}_{i=1}^n$$

associé à une source et $Q = \left(\frac{1}{n}\right)_{i=1}^n$ la distribution uniforme sur le même univers. Appliquons l'inégalité de Gibbs à P et Q :

$$\sum_{i=1}^n p_i \log\left(\frac{p_i}{q_i}\right) \leq 0 \;\Leftrightarrow\; \sum_{i=1}^n p_i \log\left(np_i\right) \leq 0$$

d'où

$$0 \geq \sum_{i=1}^n p_i(\log(n)+\log(p_i)) = \log(n)\sum_{i=1}^n p_i + \sum_{i=1}^n p_i \log(p_i) = \log(n) - H(p_1, \ldots, p_n)$$

D'après le lemme précédent, l'égalité a lieu si et seulement si $\forall i = 1, \ldots, n, p_i = q_i = \frac{1}{n}$.
C.Q.F.D

Remarque 1.3. Nous pouvons interpréter ce résultat en disant que l'incertitude sur le symbole observé à la sortie d'une source est maximale lorsque tous les symboles sont équiprobables. Nous retrouvons ici le cas particulier de l'entropie d'une source binaire $H_2(p)$ étudiée dans la proposition 1.2.

Proposition 1.4 (Concavité). *Soient $P = (p_i)_{i=1}^n$ et $Q = (q_i)_{i=1}^n$ deux distributions de probabilités sur le même univers fini $\Omega = \{\omega_i\}_{i=1}^n$ associé à une source S. Alors*
Alors, $\forall \lambda \in [0,1]$

$$H(\lambda P + (1-\lambda)Q) \geq \lambda H(P) + (1-\lambda)H(Q)$$

ou encore

$$H(\lambda p_1 + (1-\lambda)q_1, \ldots, \lambda p_n + (1-\lambda)q_n \geq \lambda H(p_1, \ldots, p_n) + (1-\lambda)H(q_1, \ldots, q_n).$$

De plus, l'égalité a lieu si et seulement si

$$\lambda \in \{0,1\} \text{ ou } P = Q$$

Remarque 1.4. Nous pouvons interpréter cette propriété comme suit. L'entropie d'une moyenne de deux sources est supérieure à la moyenne de leurs entropies.

Exemple 1.4. Considérons trois sources binaires S_1, S_2 et S_3 sur l'alphabet $\Omega = \{a, b\}$. Supposons que les distributions de probabilités de S_1 et S_2 sont les suivantes

$$P_1(a) = p_1 = \frac{1}{3},\ P_1(b) = (1-p_1) = \frac{2}{3}$$

$$P_2(a) = p_2 = \frac{1}{2},\ P_2(b) = (1-p_2) = \frac{1}{2}$$

Supposons que la distribution de S_3 soit la moyenne des deux premières

$$p_3 = \frac{p_1 + p_2}{2} = \frac{5}{12}$$

Calculons les entropies des trois sources

$$H(S_1) = H_2(p_1) = -p_1 \log(p_1) - (1-p_1)\log(1-p_1) = -\frac{2}{3} + \log(3)$$

$$H(S_2) = H_2(p_2) = -p_2 \log(p_2) - (1-p_2)\log(1-p_2) = 1$$

$$H(S_3) = H_2(p_3) = -p_3 \log(p_3) - (1-p_3)\log(1-p_3) \simeq 0.979868$$

On constate que l'entropie de la troisième source est supérieure à la moyenne des entropies des deux premières :

$$H(S_3) = 0.979868 > \frac{H(S_1) + H(S_2)}{2} = \frac{1 - \frac{2}{3} + \log(3)}{2} \simeq 0.9591479$$

1.7 Entropie conjointe - Entropie conditionnelle

1.8 Quelques notions utiles de probabilités

Soit X une variable aléatoire discrète à support fini définie sur un espace probabilisé fini. Une telle variable est définie par la donnée de son support (ensemble de valeurs possibles) $X \in \{x_i\}_{i=1}^n$ et par sa distribution de probabilité,

$$P : \{x_i\}_{i=1}^n \to [0,1], \quad x_i \mapsto P(x_i)$$

telle que $\forall i = 1, \ldots, n 0 \leq P(x_i) \leq 1$ et $\sum_{i=1}^{n} P(x_i) = 1$.

À un couple de variables aléatoires X à valeurs dans $\{x_i\}_{i=1}^n$ et Y à valeurs dans $\{y_j\}_{j=1}^m$ on associe **la distribution conjointe de probabilités**

$$P_{XY} : \{(x_i, y_j),\ (i,j) \in |[1,n]| \times |[1,m]|\} \to [0,1],\ \ (x_i, y_j) \mapsto P(x_i, y_j) = P[X = x_i \text{ et } Y = y_j]$$

On a les propriétés suivantes :

$$\begin{array}{l} \forall (i,j),\ 0 \leq P_{XY}(x_i, y_j) \leq 1 \\ \displaystyle\sum_{i=1}^{n} \sum_{j=1}^{m} P(x_i, y_j) = 1 \end{array}$$

Étant donnée la distribution conjointe d'un couple de variables aléatoires (X, Y), $P(x_i, y_j)$, on définit les **distributions marginales** de X et de Y respectivement par les relations :

$$\begin{cases} P_X : \{x_i\}_{i=1}^n \to [0,1], \quad x_i \mapsto P_X(x_i) = \displaystyle\sum_{j=1}^{m} P(x_i, y_j) \\ \\ P_Y : \{x_j\}_{j=1}^m \to [0,1], \quad y_j \mapsto P_Y(y_j) = \displaystyle\sum_{i=1}^{n} P(x_i, y_j) \end{cases}$$

Soient X et Y deux variables aléatoires discrètes. On associe à l'événement $Y = y_j$ une distribution conditionnelle $P[X|y_j]$ de X sachant y_j définie par

$$P : x_i \mapsto P(x_i|y_j) = P[X = x_i | Y = y_j].$$

1.9 Définitions et propriétés essentielles

Définition 1.5 (Entropie conjointe)**.** Soient $X = \{x_i\}_{i=1}^n$ et $Y = \{y_j\}_{j=1}^m$ deux variables aléatoires discrètes définies sur un même univers.
Soit $P(i,j) = P[X = x_i \text{ et } Y = y_j]$ leur distribution conjointe.
Alors **l'entropie conjointe de X et Y** est définie par

$$H(X,Y) = -\sum_{i=1}^{n}\sum_{j=1}^{n} P(i,j)\log(P(i,j)). \tag{1.3}$$

Cette définition peut être généralisée à un nombre quelconque r de variables aléatoires $X_1, \ldots, X_2$ à travers leur distribution de probabilité conjointe : $P(i_1, i_2, \ldots, i_r) = P[X_{i_1} = x_{i_1} \text{et } \ldots X_{i_r} = x_{i_r}]$:

$$H(X_1, X_2, \ldots, X_r) = -\sum_{i_1=1}^{n_1} \ldots \sum_{i_r=1}^{n_r} P(i_1, i_2, \ldots, i_r)\log(P(i_1, i_2, \ldots, i_r)).$$

Cette notion, que nous introduisons ici dans un contexte volontairement plus général de variables aléatoires, peut être interprétée de différentes façons selon l'application. Voici quelques exemples.

Exemple 1.5 (Étude des associations de symboles d'une source)**.** Soit une source S d'alphabet $\Omega = \{a, e, o, c, p, r, t\}$. Supposons que tous les symboles soient équiprobables. L'entropie de cette source est alors

$$H(S) = \log(7) \simeq 2.80 \text{ bits par symbole.}$$

Pour coder chaque symbole d'un message on ne peut pas utiliser en moyenne moins de 2.80 bits par symbole.

On peut se poser la question de distribution de différentes associations possibles de symboles entre eux. Par exemple, on peut étudier les syllabes formées d'une consonne et d'une voyelle. Soit X la variable aléatoire associée aux voyelles $\{a, e, o\}$ et Y la v.a. associée à l'émission de consonnes $\{c, p, r, t\}$. On s'intéresse aux probabilités d'apparition simultanée d'une de ces voyelles et d'une de ces consonnes dans une même syllabe. Supposons que ces probabilités sont données par le tableau suivant :

$Y \backslash X$	a	e	o	$P_Y(y)$
c	0.05	0.05	0.1	0.2
p	0.1	0.15	0.05	0.3
r	0.05	0.1	0.1	0.25
t	0.15	0.05	0.05	0.25
$P_X(x)$	0.35	0.35	0.3	

L'entropie conjointe de X et Y est alors

$$H(X,Y) = -(6 \cdot 0.05 \log(0.05) + 4 \cdot 0.1 \log(0.1) + 2 \cdot 0.15 \log(0.15)) \simeq 3.44$$

On peut interpréter cette quantité comme entropie d'une source $Z = (X,Y)$ dont l'alphabet est formé de couples de symboles $(consonne, voyelle)$. Nous remarquons alors que pour coder chaque couple il suffirait en moyenne 3.44 bits par couple au lieu de $2 \times 2.80 = 5.6$ bits, si l'on codait chaque symbole du couple séparément.

Exemple 1.6 (Etude d'un couple émetteur-récepteur)**.** Dans la suite de ce cours nous allons étudier en détail les problèmes de transmission de messages émis par une source d'information via un canal de transmission. Nous allons donc associer à la source, d'alphabet $\Omega_X = \{x_1, \ldots, x_n\}$ une variable aléatoire X dont les valeurs sont les symboles émis, et au récepteur, d'alphabet $\Omega_Y = \{y_1, \ldots, y_m\}$, une variable aléatoire Y dont les valeurs sont les symboles reçus. Alors la distribution conjointe de X et Y décrit le canal de transmission et son entropie est une caractéristique importante du canal.

Définition 1.6 (Entropie conditionnelle moyenne)**.** Soient $X = \{x_i\}_{i=1}^n$ et $Y = \{y_j\}_{j=1}^m$ deux variables aléatoires discrètes définies sur un même univers.
Soit $P(i,j) = P[X = x_i \text{et } Y = y_j]$ leur distribution conjointe.
Posons $P(i|j) = P[X = x_i | Y = y_j] = \dfrac{P(i,j)}{P[Y = y_j]}$.
Alors l'**entropie conditionnelle moyenne** de X sachant Y est définie par

$$H(X|Y) = -\sum_{i=1}^{n} \sum_{j=1}^{m} P(i,j) \log(P(i|j)). \tag{1.4}$$

Remarque 1.5. Il est utile d'expliciter le terme "entropie conditionnelle moyenne". Supposons que $Y = y_j$ et considérons la distribution conditionnelle $P[X|y_j]$ de X sachant $Y = y_j$ définie par

$$P[X = x_i | y_j] = P[X = x_i | Y = y_j],\ i = 1, \ldots, n$$

On lui associe l'entropie conditionnelle de X sachant y_j comme l'information conditionnelle moyenne de X sachant y_j :

$$H(X|y_j) = -\sum_{i=1}^{n} P(x_i|y_j) \log(P(x_i|y_j)) = -\sum_{i=1}^{n} \frac{P(x_i,y_j)}{P(y_j)} \log(P(x_i|y_j))$$

Alors l'entropie moyenne de X sachant Y est définie par

$$\begin{aligned} H(X|Y) &= \sum_{j=1}^{m} P(y_j) H(X|y_j) \\ &= -\sum_{j=1}^{m} P(y_j) \sum_{i=1}^{n} \frac{P(x_i, y_j)}{P(y_j)} \log(P(x_i|y_j) \\ &= -\sum_{i=1}^{n} \sum_{j=1}^{m} P(i,j) \log(P(i|j)) \end{aligned}$$

On appelle souvent $H(X)$ *l'entropie a priori* de X et $H(X|Y)$ *l'entropie (moyenne) a posteriori* de X.

Proposition 1.5 (Additivité)**.** *L'entropie conjointe de deux variables aléatoires X et Y est égale à la somme de l'entropie de l'une d'elles et à l'entropie conditionnelle moyenne de l'autre.*

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

Corollaire 1.1.

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

Proposition 1.6 (Règles de chaînage)**.** *Soient $(X_i)_{i=1}^r$ r variables aléatoires discrètes. Alors*

1. *La loi de probabilités conjointes de $(X_i)_{i=1}^r$ peut être développée comme suit :*

$$\begin{aligned} P(X_1, \ldots, X_r) &= P(X_1) \cdot P(X_2|X_1) \cdot P(X_3|X_2, X_1) \ldots P(X_r|X_{r-1}, \ldots, X_1) \\ &= \prod_{k=1}^{r} P(X_k|X_{k-1}, \ldots, X_1) \end{aligned} \tag{1.5}$$

2. *L'entropie conjointe de $(X_i)_{i=1}^r$ vérifie*

$$\begin{aligned} H(X_1, \ldots, X_r) &= H(X_1) + H(X_2|X_1) + \ldots H(X_r|X_{r-1}, \ldots, X_1) \\ &= \sum_{k=1}^{r} H(X_k|X_{k-1}, \ldots, X_1) \end{aligned} \tag{1.6}$$

Proposition 1.7 (Propriétés)**.**

1.

$$H(X,Y) \geq 0, \quad H(X|Y) \geq 0$$

L'entropie conjointe $H(X,Y)$ est nulle ssi une seule des combinaisons (x_i, y_j) est possible. L'entropie conditionnelle moyenne $H(X,Y)$ est nulle ssi X est une fonction de Y.

2.

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

3.

$$H(X,Y) \leq H(X) + H(Y) \leq 2H(X+Y)$$

1.10 Information mutuelle moyenne

Définition 1.7 (Information mutuelle moyenne)**.** Soient $X = \{x_i\}_{i=1}^n$ et $Y = \{y_j\}_{j=1}^m$ deux variables aléatoires discrètes définies sur un même univers. Soit $P(i,j) = P[X = x_i \text{et } Y = y_j]$ leur distribution conjointe. L'information mutuelle moyenne de X et Y est définie par

$$I(X;Y) = \sum_{i=1}^{n} \sum_{j=1}^{m} P(i,j) \log \left(\frac{P(i,j)}{P(x_i)P(y_j)} \right). \tag{1.7}$$

Remarque 1.6. À partir des définitions données ci-dessus on déduit facilement ces relations importantes :

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y) \tag{1.8}$$

Dans le cas où les variables X et Y représentent respectivement le message émis et le message reçu par le destinataire, cette relation signifie que l'information mutuelle moyenne est égale à :

- l'information émise $H(X)$, diminuée de l'incertitude sur le symbole x émis quand le symbole y reçu est connu, $H(X|Y)$;
- et de façon symétrique, l'information reçue, diminuée de l'incertitude sur le symbole reçu y quand le symbole émis x est connu, $H(Y|X)$.

2 Un canal

2.1 Description mathématique d'une communication

Un canal de transmission peut être vu comme un système qui reçoit en entrée des symboles émis par une source d'alphabet $\Omega_X = \{x_i\}_{i=1}^n$ et qui donne en sortie des symboles de l'alphabet du récepteur $\Omega_Y = \{x_i\}_{i=1}^m$ (éventuellement différent de Ω_X). Lors de la transmission des erreurs peuvent se produire de façon aléatoire. Ainsi, le lien entre un symbole émis et un symbole reçu est incertain. On peut alors parler de **canal avec bruit.**

Dans ce qui suit, nous ne nous intéressons pas à la nature de ce bruit ni à ses causes possibles. *La théorie de l'information s'intéresse aux conséquences de l'incertitude introduite par ce bruit sur l'exactitude de l'information reçue à la sortie d'un canal de transmission.*

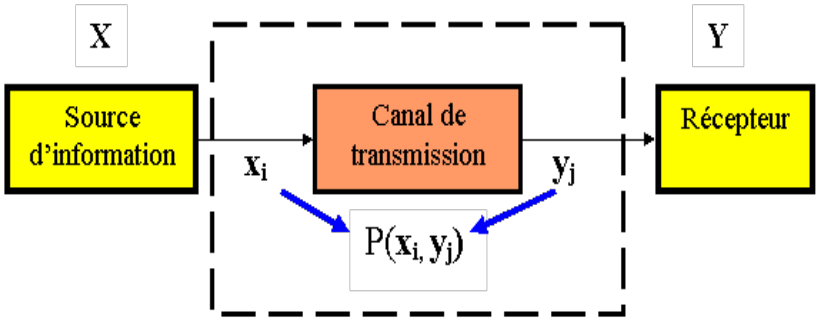

FIG. 1.3 – Canal de transmission

Quelles sont les mesures de quantité d'information qui peuvent être associées à l'ensemble "source-canal-récepteur" ? Commençons par un exemple simple.

Soient une source binaire X d'alphabet $\Omega = \{0, 1\}$ et de distribution de probabilité uniforme : $P(1) = P(0) = 0.5$. Soit un récepteur Y de même alphabet. Imaginons, qu'à chaque symbole émis, le canal de transmission a la probabilité de $\frac{1}{4}$ de faire une erreur. On peut représenter alors le fonctionnement de ce canal par le schéma suivant :

Soient X et Y les variables aléatoires associées respectivement à la source et au récepteur. La probabilité de l'erreur donnée permet d'établir les probabilités conditionnelles suivantes :

$$P(Y = 0|X = 0) = \frac{3}{4} \quad P(Y = 1|X = 0) = \frac{1}{4}$$

$$P(Y = 0|X = 1) = \frac{1}{4} \quad P(Y = 1|X = 1) = \frac{3}{4}$$

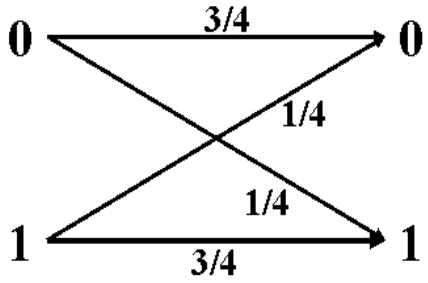

FIG. 1.4 – Canal de transmission

En connaissant également la distribution de probabilité de la source X on peut en déduire la distribution conjointe. En effet,

$$P_{XY}(0,0) = P(X=0)P(Y=0|X0) = \frac{3}{4}\cdot\frac{1}{2} = \frac{3}{8} \quad P_{XY}(0,1) = P(X=0)P(Y=1|X=0) = \frac{1}{8}$$

$$P_{XY}(1,0) = P(X=1)P(Y=0|X=1) = \frac{1}{8} \qquad P_{XY}(1,1) = P(X=1)P(Y=1|X=1) = \frac{3}{8}$$

Enfin, on peut calculer les distributions marginales de X et Y :

$$P_X(0) = P_X(1) = \frac{1}{2}, \quad P_Y(0) = P_Y(1) = \frac{3}{8} + \frac{1}{8} = \frac{1}{2}$$

À partir de ces différentes distributions des probabilités, on peut maintenant calculer les entropies $H(X),\ H(Y),\ H(X,Y),\ H(X|Y)$ et enfin l'information mutuelle $I(X;Y)$:

$$H(X) = H(Y) = 1$$

L'entropie de chacune des variables X et Y décrit la difficulté moyenne de prédire le symbole émis (ou reçu).

$$H(X,Y) = -\left(2\frac{1}{8}\log\left(\frac{1}{8}\right) + 2\frac{3}{8}\log\left(\frac{3}{8}\right)\right) \simeq 1.8$$

L'entropie mutuelle décrit l'information moyenne apportée par l'observation d'un couple de symboles x, le symbole émis et y, le symbole reçu.

$$H(Y|X) = -\left(2\frac{1}{8}\log\left(\frac{1}{4}\right) + 2\frac{3}{8}\log\left(\frac{3}{4}\right)\right) \simeq 0.8$$

L'entropie conditionnelle exprime la difficulté moyenne de prévoir le symbole reçu lorsqu'on connaît celui qui a été émis.

$$I(X;Y) = H(X) - H(Y|X) = 1 - 0.8 = 0.2$$

2.2 Canal discret stationnaire, sans mémoire

Supposons que l'alphabet de la source, en entrée est composé de n symboles $\Omega_X = \{x_i,\ i = 1, \ldots, n\}$ et que l'alphabet du récepteur, en sortie, est composé de m symboles $\Omega_Y = \{y_j,\ j = 1, \ldots, m\}$.

Dans la suite de ce cours, nous allons considérer seulement un cas particulier de canaux de communication, **les canaux discrets sans mémoire et stationnaires**. Un tel canal est décrit par la donnée de la matrice de probabilités conditionnelles, appelée **matrice de transition**

$$p_{i,j} = P[y_j|\, x_i],\ i = 1, \ldots, n,\ j = 1, \ldots, m$$

Cette matrice décrit les propriétés du bruit dans le canal.

- Le terme **"sans mémoire"** signifie que la réception à tout instant d'un symbole Y ne dépend que du symbole émis X. En particulier, chaque symbole reçu est indépendant des symboles reçus précédemment.
- Le terme **"stationnaire"** signifie que les caractéristiques probabilistes du bruit sont indépendantes du temps. Ainsi, à tout instant de transmission, cette matrice est la même.

En pratique, pour décrire les caractéristiques du bruit d'un canal on dispose souvent de la matrice de transition $P_{Y|X} : n \times m$, c'est à dire de la distribution conditionnelle de la sortie sachant l'entrée. Dans ces cas là on connaît également la distribution de probabilité de la source $P_X = (P(x_i))_{i=1}^n$. En connaissant ces deux distributions on peut déduire toutes les autres distributions associées au couple (X, Y) de l'émetteur et récepteur :

- Distribution de probabilité conjointe :

$$\forall (i, j),\quad P(x_i, y_j) = P(x_i)P(y_j|x_i)$$

- Distribution marginale de Y :

$$\forall j,\quad P(y_j) = \sum_{i=1}^{n} P(y_j|x_i)P(x_i)$$

- Distribution conditionnelle $P[X|Y]$:

$$\forall (i, j),\quad P(x_i|y_j) = P(x_i, y_j)/P(y_j)$$

Lorsque toutes ces distributions sont connues, on peut associer à un système de communication "source - canal - récepteur" différentes entropies :

$H(X)$**. L'entropie de la source** Elle représente l'information moyenne par symbole de la source ou encore la difficulté moyenne de prédire le symbole émis.

$H(Y)$**. L'entropie du récepteur** Elle représente l'information moyenne par symbole reçu ou encore la difficulté moyenne de prédire le symbole reçu.

$H(X, Y)$**. L'entropie conjointe "source-récepteur"** Elle représente l'incertitude moyenne du système de communication dans son ensemble ou encore la quantité de l'information moyenne par paire "symbole émis - symbole reçu".

$H(X|Y)$**. L'entropie conditionnelle de la source, sachant le symbole reçu** Elle représente l'incertitude moyenne sur le symbole émis lorsqu'on connaît le symbole reçu.

$H(Y|X)$**. L'entropie conditionnelle du récepteur, sachant le symbole émis** Elle représente l'incertitude moyenne sur le symbole reçu lorsqu'on connaît le symbole émis.

$I(X;Y)$**. L'information mutuelle moyenne** Elle représente la quantité moyenne d'information par symbole transmise à travers le canal.

2.2.1 Cas particulier : canal sans bruit

L'absence de bruit dans un canal signifie que la transmission est exacte, sans erreurs. Le canal peut alors être vu comme une mise en correspondance bi-univoque (bijective) des deux alphabets. Cela se traduit par une matrice de transition identité :

$$P[Y|X] = I_n = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & \vdots \\ \vdots & \cdots & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix}$$

La matrice de probabilités conjointes est alors diagonale

$$P[X,Y] = I_n = \begin{pmatrix} p(x_1,y_1) & 0 & \cdots & 0 \\ 0 & p(x_2,y_2) & \cdots & \vdots \\ \vdots & \cdots & \ddots & 0 \\ 0 & \cdots & 0 & p(x_n,y_n) \end{pmatrix}$$

Il est facile de vérifier que les entropies sont :

$$H(X,Y) = H(X) = H(Y) = -\sum_{i=1}^{n} p(x_i,y_i) \log p(x_i,y_i)$$

et que

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

Enfin,

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

On peut facilement interpréter ces relations. En effet, dans un canal sans bruit à chaque symbole émis correspond un et une seul symbole reçu. Alors lorsque l'on connaît le symbole émis, on connaît **sans aucune incertitude** le symbole reçu et *vice versa*. Ainsi, les entropies conditionnelles moyennes sont nulles car elle mesurent précisément l'incertitude moyenne sur le symbole émis (resp. reçu) sachant le symbole reçu (resp. émis). C'est aussi pour cette raison que les incertitudes moyennes par symbole à la source, $H(X)$, et à la réception, $H(Y)$, sont les mêmes.

2.2.2 Cas particulier : canal avec entrée et sortie indépendantes

Il s'agit d'un canal d'alphabets d'entrée et de sortie respectivement $\Omega_X = \{x_i,\ i = 1, \ldots, n\}$ et $\Omega_Y = \{y_j,\ j = 1, \ldots, m\}$ tel que, quel que soit le symbole émis, on peut recevoir n'importe quel symbole y_j avec équiprobabilité :

$$p(y_j|x_i) = p(y_j) = \frac{1}{m}$$

Ainsi ma matrice de transition est de la forme :

$$P[Y|X] = \begin{pmatrix} 1/m & 1/m & \cdots & 1/m \\ 1/m & 1/m & \cdots & 1/m \\ \vdots & \cdots & \ddots & 1/m \\ 1/m & \cdots & 1/m & 1/m \end{pmatrix}$$

Étant donnée la distribution de probabilité de la source : $P(x_i) = p_i,\ i = 1, \ldots, n \sum_{i=1}^{n} p_i = 1$ on en déduit les probabilités conjointes :

$$p(x_i, y_j) = p(x_i)p(y_j) = \frac{p_i}{m} = q_i, \quad i = 1, \ldots, n, \quad j = 1, \ldots, m$$

La matrice de probabilités conjointes a alors m colonnes identiques

$$P[X, Y] = \begin{pmatrix} q_1 & q_1 & \cdots & q_1 \\ q_2 & q_2 & \cdots & q_2 \\ \vdots & \vdots & \ddots & \vdots \\ q_n & q_n & \cdots & q_n \end{pmatrix}$$

La distribution conditionnelle de la source sachant le récepteur se calcule comme suit :

$$p(x_i|y_j) = p(x_i) = m \cdot q_i$$

Pour les entropies on a

$$\begin{aligned} H(X,Y) &= -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j) log p(x_i, y_j) = -m \sum_{i=1}^{n} q_i \log q_i \\ H(X) &= -\sum_{i=1}^{n} p_i \log p_i \\ H(Y) &= \log m \\ H(X|Y) &= H(X) = -\sum_{i=1}^{n} p_i \log p_i \\ H(Y|X) &= H(Y) = \log m \\ I(X,Y) &= H(X) - H(X|Y) = 0 \end{aligned}$$

L'interprétation de ces formules est la suivante : un tel canal, ne transporte aucune information entre la source et le récepteur ($I(X;Y) = 0$). Si un canal sans bruit peut être considéré comme le cas idéal d'un canal sans pertes alors un canal à entrée et sortie indépendantes comme ayant la perte maximale d'informations.

2.2.3 Exemple complet

Soient une source d'un alphabet de 5 symboles $\Omega_X = \{x_1, x_2, x_3, x_4, x_5\}$ et un récepteur d'alphabet ayant quatre symboles $\Omega_Y = \{y_1, y_2, y_3, y_4\}$. Supposons que la matrice de probabilités conjointes associée à un canal est connue :

$$P(X,Y) = \begin{array}{c|cccc} & y_1 & y_2 & y_3 & y_4 \\ \hline x_1 & 0.25 & 0 & 0 & 0 \\ x_2 & 0.1 & 0.3 & 0 & 0 \\ x_3 & 0 & 0.05 & 0.10 & 0 \\ x_4 & 0 & 0 & 0.05 & 0.10 \\ x_5 & 0 & 0 & 0.05 & 0 \end{array}$$

Calculons toutes les autres distributions de probabilités associées :

Distribution marginale de la source On utilise la définition

$$\forall i = 1, \ldots, 5,\ p(x_i) = \sum_{j=1}^{4} p(x_i, y_j)$$

Ainsi, en faisant les sommes des éléments de chaque **ligne** de la matrice $P(X,Y)$ on trouve :

$$\begin{array}{ll} p(x_1) = 0.25 & p(x_2) = 0.1 + 0.3 = 0.4 \\ p(x_3) = 0.05 + 0.10 = 0.15 & p(x_4) = 0.05 + 0.10 = 0.15 \\ p(x_5) = 0.05 & \end{array}$$

Distribution marginale du récepteur On utilise la définition

$$\forall j = 1, \ldots, 4,\ p(y_j) = \sum_{i=1}^{5} p(x_i, y_j)$$

Ainsi, en faisant les sommes des éléments de chaque **colonne** de la matrice $P(X,Y)$ on trouve :

$$\begin{array}{ll} p(y_1) = 0.25 + 0.10 = 0.35 & p(y_2) = 0.3 + 0.05 = 0.35 \\ p(y_3) = 0.10 + 0.05 + 0.05 = 0.2 & p(y_4) = 0.10 \end{array}$$

Distribution conditionnelle $P(X|Y)$ On utilise la définition

$$\forall i = 1, \ldots, 5,\ \forall j = 1, \ldots, 4,\ p(x_i|y_j) = \frac{p(x_i, y_j)}{p(y_j)}$$

On trouve la matrice

$$P(X|Y) = \begin{pmatrix} 5/7 & 0 & 0 & 0 \\ 2/7 & 6/7 & 0 & 0 \\ 0 & 1/7 & 1/2 & 0 \\ 0 & 0 & 1/4 & 1 \\ 0 & 0 & 1/4 & 0 \end{pmatrix}$$

Distribution conditionnelle $P(Y|X)$ On utilise la définition

$$\forall i = 1, \ldots, 5,\ \forall j = 1, \ldots, 4,\ p(y_j|x_i) = \frac{p(x_i, y_j)}{p(x_i)}$$

On trouve la matrice

$$P(Y|X) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1/4 & 3/4 & 0 & 0 \\ 0 & 1/3 & 2/3 & 0 \\ 0 & 0 & 1/3 & 2/3 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

Calculons maintenant les entropies.

$H(X)$**. L'entropie de la source**

$$\begin{aligned} H(X) &= -\sum_{i=1}^{5} p_i \log p_i = \\ &= -0.25 \log 0.25 - 0.4 \log 0.4 - 2 \cdot 0.15 \log 0.15 - 0.05 \log 0.05 \simeq 2.066 \end{aligned}$$

$H(Y)$**. L'entropie du récepteur**

$$\begin{aligned} H(Y) &= -\sum_{i=1}^{4} p_i \log p(y_i) = \\ &= -2 \cdot 0.35 \log 0.35 - \cdot 0.20 \log 0.20 \\ &- 0.10 \log 0.10 \simeq 1.856 \end{aligned} \tag{1.9}$$

$H(X, Y)$**. L'entropie conjointe "source-récepteur"**

$$\begin{aligned} H(X, Y) &= -\sum_{i=1}^{5} \sum_{j=1}^{4} p(x_i, y_j) \log p(x_i, y_j) = \\ &= -0.25 \log 0.25 - 3 \cdot 0.1 \log 0.1 \\ &- 3 \cdot 0.05 \log 0.05 - 0.3 \log 0.3 \simeq 2.665 \end{aligned}$$

$H(X|Y)$. L'entropie conditionnelle de la source, sachant le symbole reçu

$$\begin{aligned} H(X|Y) &= -\sum_{i=1}^{5}\sum_{j=1}^{4} p(x_i,y_j)\log\frac{p(x_i,y_j)}{p(y_j)} = \\ &= -0.25\log\frac{5}{7} - 0.1\log\frac{2}{7} - 0.1\log\frac{1}{2} \\ &- 0.05\left(\log\frac{1}{7}+\log\frac{1}{4}+\log\frac{1}{4}\right) - 0.3\log\frac{6}{7} \simeq 0.809 \end{aligned}$$

$H(Y|X)$. L'entropie conditionnelle du récepteur, sachant le symbole émis

$$\begin{aligned} H(Y|X) &= -\sum_{i=1}^{5}\sum_{j=1}^{4} p(x_i,y_j)\log\frac{p(x_i,y_j)}{p(x_i)} = \\ &= -0.1(\log\tfrac{1}{4} + 2\log\tfrac{2}{3}) - 2\cdot 0.05\log\tfrac{1}{3} - 0.3\log\tfrac{3}{4} = 0.6 \end{aligned}$$

2.3 Capacité d'un canal

Nous avons donné plus haut la définition de l'information mutuelle moyenne $I(X;Y)$ comme mesure de l'information transmise en moyenne par symbole à travers le canal. On peut remarquer que cette quantité dépend du canal mais aussi de la source d'information. On introduit alors une nouvelle mesure, qui ne décrit que le canal : la capacité de canal.

Définition 1.8 (Capacité d'un canal)**.** La capacité d'un canal est définie par

$$C = \max_{P(X)} I(X;Y) = \max_{P(X)}(H(X) - H(X|Y))$$

Le maximum est pris sur toutes les distributions de probabilité possibles de la source.

Exemple 1.7 (Capacité d'un canal sans bruit)**.** Considérons un canal sans bruit et une source d'alphabet $\Omega = \{x_i,\ i = 1,\ldots,n\}$. Nous avons déjà vu (voir 2.2.1) que dans ce cas $I(X;Y) = H(X)$. Alors

$$C = \max_{P(X)} I(X;Y) = \max H(X)$$

Or nous avons également vu que l'entropie d'une source est maximale lorsque tous les symboles de l'alphabet sont équiprobables. Sa valeur maximale est alors égale à $\log n$. Ainsi on en déduit que pour un canal sans bruit d'alphabet de n caractères

$$C = \log n$$

Chapitre 2

Théorèmes fondamentaux de la théorie de l'information.

1 Codage

D'une manière générale le codage peut être vu comme une transformation de symboles d'un alphabet donné $\Omega_1 = \{s_1, \ldots, s_n\}$ en suites de symboles d'un autre alphabet $\Omega_2 = \{c_1, \ldots, c_d\}$.

Dans un schéma de communication, la source et le canal de transmission n'utilisent pas forcément le même alphabet. Par exemple, la source peut être un texte anglais, utilisant les 26 lettres de l'alphabet latin, et le canal peut être tout support numérique, utilisant l'alphabet binaire. Il se pose donc le problème de **codage** comme "traduction" entre l'alphabet de la source et celui du canal. Si le récepteur utilise le même alphabet que la source il est également de **décoder** le message avant de pouvoir le lire, c'est-à-dire, appliquer la transformation inverse. Toute transformation candidate doit vérifier un certain nombre de critères que nous allons détailler plus loin. Étant donné qu'il peut exister une grande quantité de codages possibles il se posera la question de savoir lequel est meilleur que tous les autres. Cela dépendra de critère d'optimalité que l'on fixera. Nous allons dans cette perspective considérer deux types d'applications de codage :

1. **Codage de source ou encore codage sans bruit.** Nous allons supposer que la communication se fait via un canal sans bruit. Cela signifie que tout message est transmis de façon exacte. Sous cette hypothèse le meilleur code sera celui qui permettra la transmission la plus rapide possible. **Le premier théorème de Shannon** donne la solution à ce problème. *A noter* que ce cours sera dans la suite consacré essentiellement à l'étude des méthodes de codage de source.

2. **Codage de canal ou encore codage en présence de bruit.** En supposant qu'il existe des perturbations pouvant engendré des erreurs à la réception, on cherchera une méthode de codage permettant une transmission aussi rapide que possible tout en minimisant la probabilité des erreurs. **Le second théorème de Shannon** donne la solution à ce problème.

2 Premier théorème fondamental

2.1 Codage de source

Nous allons dans un premier temps considérer un modèle de communication avec un canal discret sans bruit. Soient $\Omega_S = \{s_1, \ldots, s_n\}$ l'alphabet de la source et $\Omega_C = \{c_1, \ldots, c_d\}$ l'alphabet du canal. Supposons que la distribution de probabilité de l'alphabet de la source est connue :

$$P_S : \Omega_S \to [0,1], \quad s_i \mapsto p_S(s_i)$$

On peut alors en déduire l'entropie de la source

$$H(S) = -\sum_{i=1}^{n} p_i \log p_i.$$

Nous utiliserons dans la suite la terminologie suivante :

Lettre, symbole ou caractère Tout élément d'un alphabet donné ;

Message ou mot Une séquence fini m de caractères d'un alphabet donné ;

Longueur de mot le nombre $l(m)$ de caractères d'un mot m ;

Le codage consiste à faire correspondre à chaque symbole s_i de la source une séquence $m(s_i)$ de symboles de l'alphabet du canal, appelée **mot-code**. Une telle association peut donc être représentée par un ensemble $\{m_1, m_2, \ldots, m_n\}$ de n mots codes correspondants chacun à un symbole de l'alphabet de la source : $\forall\ i = 1, \ldots, n,\ m_i = m(s_i)$. Si l'on note l_i les longueurs des mots m_i du code. Soit L la variable aléatoire dont la valeur est la longueur du mot-code associé à un symbole émis par la source S. Alors cette variable aléatoire prend les valeurs l_i avec les probabilités $p_i = p_S(s_i), \quad i = 1, \ldots, n$. On définit alors la longueur moyenne du code par

$$\bar{L} = E[L] = \sum_{i=1}^{n} p_i l_i$$

Cette quantité est importante pour l'analyse des caractéristiques d'un code donné. Elle représente le nombre moyen de caractères de l'alphabet Ω_C pour coder un symbole émis par la source S.

Supposons que le temps de transmission est le même pour tous les caractères de l'alphabet du canal. Alors, le temps moyen de transmission d'un symbole de l'alphabet de

la source est proportionnel à la longueur moyenne des mots du code, $\bar{L}$. C'est pour cette raison que nous allons dans la suite étudier en détail ce paramètre.

Un code doit vérifier un certain nombre de propriétés garantissant la possibilité de reconstituer tout message codé à la réception. Ces propriétés sont :

Régularité Un code $\{m_1, m_2, \ldots, m_n\}$ est dit régulier si tous les mots qui le composent sont distincts : $m_i \neq m_k, \quad \forall i \neq k$. Cette condition garantit au moins que tout message d'un seul caractère de l'alphabet de la source peut être décodé. Un code qui n'est pas régulier est dit singulier ou irréversible.

Déchiffrabilité Un code régulier est dit déchiffrable (ou encore à décodage unique) si pour toute suite de mots de code $m^1 m^2 \ldots m^k$ il est possible de distinguer sans ambiguïté tous les mots et donc identifier les symboles $s^j, \ j = 1, \ldots, k$ composant le message.

Voici un exemple de plusieurs codes possibles pour un même alphabet.

Exemple 2.1. Soit $\Omega_S = \{a, b, c, d\}$ de distribution de probabilité $P_S = \{0.4, 0.3, 0.2, 0.1\}$. L'entropie de cette source est $H(S) \simeq 1.85$. Supposons que le canal est binaire. Le tableau ci-dessous propose quelques codes et leurs longueurs moyennes de mots :

S	Proba	Code 1	Code 2	Code 3	Code 4	Code 5	Code 6
a	0.4	1	00	0	0	0	0
b	0.3	0	01	10	01	10	11
c	0.2	1	10	01	011	110	100
d	0.1	0	11	010	0111	1110	101
	Long. Moy.	1	2	1.7	2	2	1.9

Le code 1 n'est pas régulier. En effet, le caractère 0 correspond à deux caractères différents dans l'alphabet initial. Tous les autres codes du tableau sont régulier. Le code 2 est un code **de longueur fixe** : tous les mots du code sont de même longueur. Les codes 3-6 sont de longueur variable.

Le code 3 est régulier mais pas déchiffrable. En effet, la déchiffrabilité signifie que toute séquence de mots du code correspond à au plus un message. Dans le cas du code 3, la séquence 010 correspond à la fois à trois messages différents : "d", "ca" et "ab". Elle ne peut donc pas être décodée correctement. Nous allons maintenant analyser le problème de déchiffrabilité en détail.

2.1.1 Le problème de décodage unique

Il existe plusieurs façons de garantie qu'un code soit déchiffrable :

Codes de longueur fixe Un code régulier de longueur fixe peut toujours être décodé sans ambiguïté car il suffit pour cela de découper la séquence en mots de longueur connue. Cette solution présente néanmoins un désavantage. On peut l'observer dans le tableau ci-dessus. Le code 1, de longueur fixe a la longueur moyenne de code égale à 2 tandis qu'il existe dans la même table des codes avec une longueur moyenne inférieure.

Utilisation d'un séparateur Il est possible de consacrer un symbole de l'alphabet du canal comme séparateur de mots du code. Par exemple, pour un canal binaire, on peut coder le i-ème symbole de la source s_i à l'aide de i caractères "1" et utiliser "0" comme séparateur. Dans le cas de l'exemple 2.1 cela donnerait le code suivant

S	Proba	Code 1
a	0.4	10
b	0.3	110
c	0.2	1110
d	0.1	11110
	Long. Moy.	3

et la séquence "abc" donnerait "101101110"

On constate que la longueur moyenne qui tient compte du séparateur est plus élevée que tous les autres codes.

Codes sans préfixe On dit qu'un mot W est **un préfixe** d'un autre mot V s'il existe un mot U tel que $V = WU$. Autrement dit, le mot V commence par le mot W. Dans ce cas le mot U s'appelle *suffixe*.

On dit qu'un code donné est **sans préfixe** ou **instantané** si aucun mot du code n'est un préfixe d'un autre.

Dans notre exemple 2.1 le code 6 est sans préfixe. Un tel code est toujours déchiffrable. En effet, pour décoder une séquence quelconque de mots du code il suffit lire la séquence caractère par caractère de gauche à droite. Dès qu'un mot du code m est formé on sait qu'il n'est pas un début d'un autre mot. On peut donc séparer le mot et recommencer la lecture. Cela donne une procédure de décodage "pas à pas". Prenons une séquence $W = 011010011101$ du code 6. Le décodage de la séquence se passe de la façon suivante :

Pas 1 Le premier mot du code formé en lisant de gauche à droite est $m^1 =$ "0". Donc le premier symbole est $s^1 = a$. On sépare le mot m^1 de la séquence. On obtient la nouvelle séquence $W_1 = 11010011101$.

Pas 2 $m^2 =$ "11" $\Rightarrow$ $s^2 = b$ et $W_2 = 010011101$.

Pas 3 $m^3 =$ "0" $\Rightarrow$ $s^3 = a$ et $W_3 = 10011101$.

Pas 4 $m^4 =$ "100" $\Rightarrow$ $s^4 = c$ et $W_2 = 11101$.

Pas 5 $m^5 =$ "11" $\Rightarrow$ $s^5 = b$ et $W_2 = 101$.

Pas 6 $m^6 =$ "101" $\Rightarrow$ $s^6 = d$ et $W_6 = \emptyset$.

On obtient en symboles de l'alphabet de la source : *abacbd*.

On remarque que dans la famille de codes à longueur variable et sans séparateur les codes sans préfixe ou instantanés représentent un intérêt. Il est évident que tout code instantané est déchiffrable.

La réciproque n'est pas vraie. Soit en effet une source binaire d'alphabet $\Omega_S = \{a, b\}$ et un canal binaire d'alphabet $\Omega_C = \{0, 1\}$. Le code suivant $m_1 = m(a) = 0, \quad m_2 = m(b) = 01$ n'est pas instantané car m_1 est un préfixe de m_2. Il est tout de même déchiffrable. Pour décoder une séquence de mots d ece code il suffit de repérer d'abord les positions de "1". Chaque "1" est obligatoirement précédé de "0" et donne le caractère "b". les autres "0" correspondent à "a". C'est comme cela qu'on trouve que la séquence "000101000100" correspond à "aabbaabaa".

On voit alors la signification du terme "instantané" et l'intérêt principal de ces codes. Un code sans préfixe peut être décodé pas à pas, par séparation successive des mots. Cela peut être fait au fur et à mesure de la réception du message. Tandis qu'un code déchiffrable mais non instantané nécessite un traitement plus long et spécifique au code pour être déchiffré.

Nous allons nous intéresser maintenant au problème d'existence de codes instantanés. Ce problème se formule de la façon suivante

> Soient l'alphabet de la source $\Omega_S = \{s_1, \ldots, s_n\}$ de taille n et et l'alphabet du canal $\Omega_C = \{c_1, \ldots, c_d\}$ de taille d. Étant donnés n nombres entiers positifs $(l_1, l_2, \ldots, l_n) \in \mathbb{Z}_+^*$ existe-t-il un code régulier instantané de n mots $\{m_1, \ldots, m_n\}$ tel que chaque nombre l_i soit la longueur du mot de code m_i ?

Le théorème suivant donne la condition nécessaire et suffisante d'existence de tels codes.

Théorème 2.1 (Inégalité de Kraft)**.** *Un code instantané de longueurs de mots données $l_1, \ldots, l_n$ existe si et seulement si*

$$\sum_{i=1}^{n} d^{-l_i} \leq 1$$

où d est la taille de l'alphabet du canal.

Historiquement, l'inégalité de Kraft a d'abord été démontrée par McMillan, comme condition nécessaire et suffisante d'existence de codes déchiffrables de longueurs de mots

données. Voici le théorème qui est une extension du théorème précédent sur la classe entière de codes déchiffrables :

Théorème 2.2 (Condition de McMillan)**.** *Un code déchiffrable de longueurs de mots données $l_1, \ldots, l_n$ existe si et seulement si*

$$\sum_{i=1}^{n} d^{-l_i} \leq 1$$

où d est la taille de l'alphabet du canal.

Remarque 2.1. Ces deux théorèmes montrent qu'il existe un code déchiffrable de longueurs de mots données $l_1, \ldots, l_n$ si et seulement si il existe un code instantané de mêmes longueurs de mots.

Nous avons maintenant pourvoir poser le problème de codage de source ou encore de codage sans bruit :

Soit une source S d'alphabet $\Omega_S = \{s_1, \ldots, s_n\}$ de taille n et de distribution de probabilités $P_S = \{p_1, \ldots, p_n\}$. Soit un canal d'alphabet $\Omega_C = \{c_1, \ldots, c_d\}$ de taille d, sans bruit, stationnaire et sans mémoire.
Existe-t-il un code qui minimise la longueur moyenne de mots $\bar{L}$?

2.2 Le premier théorème de Shannon

Nous allons approcher la solution du problème de codage de source en trois étapes. D'abord, nous allons énoncer la borne inférieure pour la longueur moyenne de mots de code. Ensuite, nous allons proposer une borne supérieure. Et enfin, énoncer le premier théorème fondamental de la théorie de l'information qui montre qu'il est possible d'approcher la borne inférieure avec autant de précision que l'on souhaite.

2.2.1 Borne inférieure de longueur moyenne de code

Théorème 2.3 (Borne inférieure)**.** *Soit une source S d'alphabet $\Omega_S = \{s_1, \ldots, s_n\}$ de taille n et de distribution de probabilités $P_S = \{p_1, \ldots, p_n\}$. Soit un canal d'alphabet $\Omega_C = \{c_1, \ldots, c_d\}$ de taille d, sans bruit, stationnaire et sans mémoire. Soit un code déchiffrable $\{m_1, \ldots, m_n\}$ de longueurs de mots $\{l_1, \ldots, l_n\}$.*
Alors la longueur moyenne de mots de code vérifie :

$$\bar{L} = \sum_{i=1}^{n} p(s_i) l_i \geq \frac{H(S)}{\log(d)}$$

L'égalité n'est possible que si $\forall i = 1, \ldots, n, \quad p_i = d^{-l_i}$.

Preuve du théorème 2.3

Puisque le code que nous avons est déchiffrable il vérifie l'inégalité de Kraft :

$$0 < Q = \sum_{i=1}^{n} d^{-l_i} \leq 1$$

On peut alors définir les nombres

$$q_i = \frac{d^{-l_i}}{Q}, \quad i = 1, \ldots, n$$

On a alors : $\sum_{i=1}^{n} q_i = 1$ et donc on peut appliquer l'inégalité de Gibbs (voir lemme 1.1)

$$\sum_{i=1}^{n} p_i \log\left(\frac{p_i}{q_i}\right) \leq 0$$

et l'égalité a lieu **si et seulement si** $\forall i = 1, \ldots, n, p_i = q_i$.

On en déduit alors :

$$\begin{aligned} -\sum_{i=1}^{n} p_i \log p_i &\leq -\sum_{i=1}^{n} p_i \log q_i && (2.1)\\ &= -\sum_{i=1}^{n} p_i \log \frac{d^{-l_i}}{Q} = && (2.2)\\ &= -\sum_{i=1}^{n} p_i \log d^{-l_i} + \sum_{i=1}^{n} p_i \log Q && (2.3)\\ &= \sum_{i=1}^{n} p_i l_i \log d + \log Q \sum_{i=1}^{n} p_i && (2.4) \end{aligned}$$

En remarquant que $\sum_{i=1}^{n} p_i l_i = \bar{L}$ et que $\sum_{i=1}^{n} p_i = 1$ on a

$$H(S) = -\sum_{i=1}^{n} p_i \log p_i \leq \bar{L} + \log Q$$

Enfin, d'après l'inégalité de Kraft que nous avons mentionné au début, $Q \leq 1$. Donc $\log Q \leq 0$ d'où l'inégalité que nous devions démontrer.
C.Q.F.D

Remarque 2.2. On peut noter le fait que la quantité $\frac{H(S)}{\log(d)}$ représente l'entropie de la source calculée par rapport à la base d :

$$\frac{H(S)}{\log(d)} = -\sum_{i=1}^{n} p_i \frac{\log(p_i)}{\log(d)} = -\sum_{i=1}^{n} p_i \log_d(p_i) = H_d(S)$$

Un code dont la longueur moyenne de mots atteint la borne inférieure s'appelle **absolument optimal**. Un exemple de code absolument optimal est donné par la tableau suivant

S	Proba	Code
a	0.5	0
b	0.25	10
c	0.125	110
d	0.125	111

et on a $H(S) = \bar{L} = \frac{7}{4}$.

2.2.2 Borne supérieure de longueur moyenne de code

Cependant, un code absolument optimal n'est pas toujours réalisable. En effet, pour atteindre la borne inférieure, les longueurs de mots doivent vérifier $p_i = d^{-l_i}$ et donc $l_i = \dfrac{logp_i}{logd}$. Or ces nombres ne sont pas forcément entiers. Dans ces cas la meilleure solution consiste à choisir les longueurs de mots de telle sorte que

$$\forall\ i = 1, \ldots, n, \quad -\frac{\log p_i}{\log(d)} \leq l_i < -\frac{\log p_i}{\log d} + 1$$

Théorème 2.4 (Borne supérieure)**.** *Soit une source S d'alphabet $\Omega_S = \{s_1, \ldots, s_n\}$ de taille n et de distribution de probabilités $P_S = \{p_1, \ldots, p_n\}$. Soit un canal d'alphabet $\Omega_C = \{c_1, \ldots, c_d\}$ de taille d, sans bruit, stationnaire et sans mémoire.*
Alors il existe un code déchiffrable dont la longueur moyenne de mots de code vérifie :

$$\frac{H(S)}{\log(d)} \leq \bar{L} = \sum_{i=1}^{n} p(s_i) l_i < \frac{H(S)}{\log(d)} + 1$$

Preuve du théorème 2.4

Choisissons les longueurs de mots comme précisé ci-dessus :

$$\forall\ i = 1, \ldots, n, \quad -\frac{\log p_i}{\log d} \leq l_i < -\frac{\log p_i}{\log d} + 1$$

Tout d'abord, montrons que l'inégalité de Kraft est vérifiée. En effet, de l'inégalité ci-dessus on déduit que

$$\forall\ i = 1, \ldots, n, \quad \log p_i \geq -l_i \log d \quad \Rightarrow \quad p_i \geq d^{-l_i}$$

Alors pour la somme on a

$$\sum_{i=1}^{n} d^{-l_i} \leq \sum_{i=1}^{n} p_i = 1$$

Alors il existe un code instantané de longueurs de mots l_i que nous avons choisies. Il reste à étudier sa longueur moyenne. On a :

$$\bar{L} = \sum_{i=1}^{n} p_i l_i < -\sum_{i=1}^{n} p_i \frac{\log p_i}{\log d} + 1 = -\frac{1}{\log d} \sum_{i=1}^{n} p_i \log p_i + 1 = \frac{H(S)}{\log(d)} + 1$$

et

$$\bar{L} = \sum_{i=1}^{n} p_i l_i \geq -\sum_{i=1}^{n} p_i \frac{\log p_i}{\log d} = \frac{H(S)}{\log(d)}$$

C.Q.F.D

2.2.3 Extension de source et le premier théorème de Shannon

Nous avons déjà vu que, pour un alphabet et une distribution de probabilité donnés, il est possible de trouver un code déchiffrable est proche de la borne inférieure $\frac{H(S)}{\log(d)}$ à un bit de base d près. La dernière question qu'il reste à éclaircir est de savoir s'il est possible d'approcher cette borne avec une précision arbitraire.

Pour répondre à cette question nous allons élargir notre point de vue sur le problème d'encodage de messages d'une source. Nous allons introduire l'idée de codage par blocks. Commençons par un exemple.

Exemple 2.2. Soit une source X d'alphabet $\omega_X = \{a, b\}$ et de distribution de probabilité $P_X = \{3/4, 1/4\}$. L'entropie de cette source est

$$H(X) = -\frac{3}{4}\log\left(\frac{3}{4}\right) - \frac{1}{4}\log\left(\frac{1}{4}\right) = -\frac{3}{4}\log(3) + \frac{3}{4}\log(4) + \frac{1}{4}\log(4) = -\frac{3}{4}\log(3) + 2 \simeq 0.811$$

En codant cette source caractère par caractère, on peut envisager le code suivant : $m(a) = 0,\ m(b) = 1$ de longueur moyenne de mots $\bar{L}_1 = 1$ bit par caractère. Un message T de longueur initiale $l(T)$ sera alors codé par une suite binaire de $l(T)$ bits.

Et si on voulait coder les messages de cette source en associant un code au couples de caractères ? Tout d'abord, formalisons cette idée. Avec un alphabet de taille 2 il est possible de former $2^2 = 4$ couples de caractères différents. On peut considérer cela comme un nouvel alphabet, d'une nouvelle source, Y. La variable aléatoire Y associée à cette nouvelle source correspond à l'observation de deux caractères émis indépendamment par la source initiale X. Ainsi, $Y = (X_1, X_2)$ où X_1 et X_2 sont deux variables aléatoires indépendantes ayant la même distribution que X. Compte tenu de l'indépendance de X_1 et X_2 on peut construire la distribution de probabilité de Y :

Y	aa	ab	ba	bb
P_Y	$p(a)^2 = 9/16$	$p(a)p(b) = 3/16$	$p(a)p(b) = 3/16$	$p(b)^2 = 1/16$

Enfin, l'entropie de Y peut aussi être déduite de l'indépendance de X_1 et X_2 :

$$H(Y) = H(X_1, X_2) = H(X_1) + H(X_2) = 2H(X)$$

Maintenant, si l'on considère Y comme une source à part entière, on peut appliquer le théorème 2.4. Il est possible de trouver un code déchiffrable dont la longueur moyenne de mots de code vérifie (ici $d = 2$ et donc $\log(d) = 1$) :

$$H(Y) \leq \bar{L}_2 =< H(Y) + 1 \quad \Leftrightarrow \quad 2H(X) \leq \bar{L}_2 =< 2H(X) + 1$$

On a alors la relation :

$$H(X) \leq \frac{\bar{L}_2}{2} =< H(X) + \frac{1}{2}$$

Il reste à remarquer que $\bar{L}_2$ est mesurée en "bits par symbole" de l'alphabet de Y. Or, tout symbole de l'alphabet de Y est un couple de caractères de l'alphabet de X. Ainsi, la quantité $\frac{\bar{L}_2}{2}$ représente la longueur moyenne de mot de code par symbole de X. Nous avons alors, un code dont la longueur moyenne de mot de code se trouve dans un intervalle plus petit (de longueur de 1/2 au lieu de 1) autour de la borne inférieure.

Voici l'exemple d'un tel code

$Y = (X_1, X_2)$	Proba	Code
aa	9/16	0
ab	3/16	10
ba	3/16	110
bb	1/16	111

et on a $H(Y) = 2H(X) \simeq 1.622$ et $\bar{L}_2 = \frac{27}{17}$ bits par valeur de Y.

Enfin, il est très important de remarquer que la longueur moyenne $\bar{L}_2$ représente le nombre moyen de bits pour coder un "symbole" de l'alphabet de la source Y. Donc il s'agit de $\frac{27}{17}$ bits par symbole de Y. or, chaque symbole de l'alphabet de Y est un couple de symboles de l'alphabet de X. Ainsi, en unités de mesure de la source X on a $\bar{L}_2 = \frac{27}{2 \cdot 17} \simeq 0.8$ bits par symbole de X.

Cela signifie qu'un message T de longueur $l(T)$ sera codé avec le code 2, en moyenne par $0.8l(T)$ bits binaires au lieu de $l(T)$ bits avec le premier code.

Ainsi, en introduisant le codage par blocks nous avons trouvé la possibilité de s'approcher de la borne inférieure. En généralisant l'idée de l'exemple précédent on peut formuler et démontrer le premier théorème fondamental de la théorie de l'information.

Définition 2.1 (Extension de source)**.** Soit une source X d'alphabet $\Omega_X = \{x_1, \ldots, n\}$. On appelle extension d'ordre s de la source X la source $Y = (X_1, \ldots, X_s)$ où $X_i,\ i = 1, \ldots, s$ sont les variables aléatoires indépendantes et identiquement distribuées selon la distribution de X.

Théorème 2.5 (Premier théorème de Shannon)**.** *Soit une source X d'alphabet $\Omega_X = \{x_1, \ldots, x_n\}$ de taille n et de distribution de probabilités $P_X = \{p_1, \ldots, p_n\}$. Soit un canal d'alphabet $\Omega_C = \{c_1, \ldots, c_d\}$ de taille d, sans bruit, stationnaire et sans mémoire.*
Alors il existe un procédé de codage déchiffrable dont la longueur moyenne de mots de code est aussi voisine que l'on souhaite de la borne inférieure $\dfrac{H(S)}{\log(d)}$.

Preuve du théorème 2.5

Soit la suite $(Y_s)_{s=1}^{\infty}$ d'extensions d'ordre s de la source Y. On alors $\forall\ s \geq 1,\ \ H(Y_s) = sH(X)$ et il existe un code déchiffrable dont la longueur moyenne $\bar{L}_s$ de mots de code vérifie :

$$\frac{H(Y)}{\log(d)} \leq \bar{L}_s =< \frac{H(Y)}{\log(d)} + 1 \quad \Leftrightarrow \quad s\frac{H(X)}{\log(d)} \leq \bar{L}_s =< s\frac{H(X)}{\log(d)} + 1$$

On a alors la relation :

$$\frac{H(X)}{\log(d)} \leq \frac{\bar{L}_s}{s} =< \frac{H(X)}{\log(d)} + \frac{1}{s}$$

Il reste à remarquer que $\bar{L}_s$ est mesurée en "bits par symbole" de l'alphabet de Y_s. Or, tout symbole de l'alphabet de Y_s est un s-uplet de caractères de l'alphabet de X. Ainsi, la quantité $\dfrac{\bar{L}_s}{s}$ représente la longueur moyenne de mot de code par symbole de X.
Alors en passant à la limite quand s tend vers ∞ on a

$$\lim_{s\to\infty} \frac{\bar{L}_s}{s} = \frac{H(X)}{\log(d)}$$

C.Q.F.D

3 Second théorème de Shannon

3.1 Codage de canal

Dans cette partie nous allons poser et étudier le problème de codage dans le contexte de transmission via un canal avec bruit. La différence par rapport au codage sans bruit est dans le fait que lors de la transmission des messages des erreurs peuvent se produire

de façon aléatoire. Ces erreurs sont caractérisées par la donnée de la matrice de transition du canal.

Notre problème sera alors d'évaluer la probabilité d'erreur de transmission. Nous allons montrer que cette probabilité dépend non seulement des caractéristiques probabilistes du canal et de la source mais aussi du choix du codage et de décodage.

Alors le problème de choix de meilleur code en présence de bruit sera reformulé pour prendre en compte non seulement la vitesse de transmission moyenne mais aussi la probabilité d'erreur.

3.1.1 Règle de décodage d'un canal avec bruit.

En absence de bruit de transmission la fonction de décodage était naturellement définie comme la transformation inverse de la fonction de codage. C'était par définition l'unique façon de retrouver exactement le message émis.

En présence de bruit il existe une incertitude sur le message émis lorsqu'on observe le message reçu. Autrement dit, un message reçu peut correspondre à plusieurs messages en entrée avec une distribution de probabilité qui peut être déduite à partir de la matrice de transition. Pour traiter ce problème et quantifier la probabilité d'erreur de transmission nous allons introduire la notion de règle de décodage de canal.

Définition 2.2 (Information propre)**.** Soit un canal avec un alphabet d'entrée $\Omega_X = \{x_1, \ldots, x_n\}$ et un alphabet de sortie $\Omega_Y = \{y_1, \ldots, y_d\}$. La règle de décodage de canal est une fonction déterministe

$$g : \{y_1, \ldots, y_D\} \to \{x_1, \ldots, x_M\}$$

qui à chaque symbole reçu y_j associe un symbole de l'alphabet d'entrée $x_j^* = g(y_j)$.

On peut interpréter cela sous forme de règle de décision : si le symbole y_j est reçu il est décodé comme x_j^*. Soient les variables aléatoires X, Y et $Z = g(Y)$ représentant respectivement les symboles émis, reçu et décodé. Considérons un exemple. Soit l'alphabet d'entrée de taille 3 et celui de sortie de taille 3. Supposons que l'alphabet d'entrée a la distribution de probabilité suivante

$$p(x_1) = 1/2,\ p(x_2) = p(x_3) = 1/4$$

Les probabilités de transition et la règle de décodage sont illustrées par le schéma sur la figure 2.1.

Considérons la règle de décision g définie par

$$g(y_1) = x_1,\ g(y_1) = x_2,\quad g(y_3) = x_2$$

Le seul cas où une erreur a lieu est l'émission de symbole x_2. Ainsi la probabilité d'erreur est égale à la probabilité d'émission de x_2 donc à 1/4.

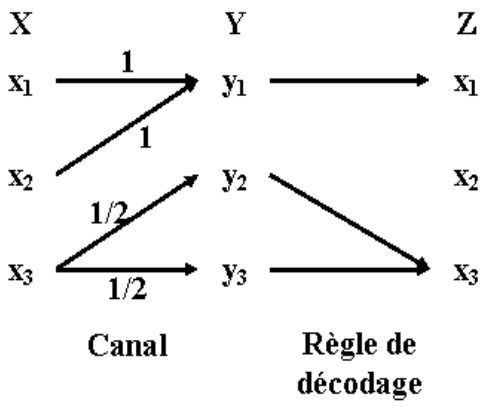

FIG. 2.1 – Un exemple de règle de décision

Dans le cas général soit E l'événement correspondant à un erreur lors de transmission d'un seul symbole. Comment peut on calculer $P[E]$? Soient les variables aléatoires X, Y et $Z = g(Y)$ représentant respectivement les symboles émis, reçu et décodé.

Tout d'abord, on peut décomposer cette probabilité selon la formule de Bays :

$$P(E) = \sum_{i=1}^{n} p(x_i)p(E|x_i)$$

Sachant que le symbole x_i est transmis, l'erreur se produit lorsque la fonction foncion de décodage renvoi un caractère différent de x_i. Ainsi dans ce cas l'événement E équivaut à $g(Y) \neq x_i$. On a donc $p(E|x_i) = P(g(Y) \neq x_i|x_i)$. Cette dernière probabilité se décompose en somme selon les différentes valeurs de Y :

$$p(E|x_i) = P(g(Y) \neq x_i|x_i) = \sum_{j=1}^{d} p(Y = y_j \text{ et } g(y_j) \neq x_i|x_i)$$

Etant donné que la règle de décodage est déterministe on a

$$p(Y = y_j \text{ et } g(y_j) \neq x_i|x_i) = \begin{cases} p(y_j|x_j), & \text{si } g(y_j) \neq x_i \\ 0 & \text{si } g(y_j) = x_i \end{cases} = p(y_j|x_j)(1 - \delta_{g(y_j),\, x_i})$$

Ici $\delta_{ik} = \begin{cases} 1, & \text{si } i \neq k \\ 0 & \text{si } i = k \end{cases}$ est le symbole de Kronecker.

Nous obtenons ainsi la probabilité d'erreur conditionnelle pour la transmission d'un seul symbole :

$$p(E|x_i) = \sum_{j=1}^{d} p(y_j|x_j)(1 - \delta_{g(y_j),\, x_i}).$$

3.1.2 Notion de code de canal.

On peut généraliser les raisonnements de la section précédente au cas où il s'agit de transmettre des messages par blocks de longueur donné l.

Dans la suite un canal est modélisé par le triplet $X, Y, P(Y|X)$ où X est la variable aléatoire correspondante à l'émission d'un symbole de l'alphabet d'entrée $\Omega_X = \{x_1, \ldots, x_n\}$ et Y est la variable correspondante à l'observation d'un symbole reçu dans l'alphabet de sortie $\Omega_Y = \{y_1, \ldots, y_d\}$. $P(Y|X)$ est la matrice de transition du canal.

On peut associer à La transmission d'un mot de longueur l une variable aléatoire $X^{(l)}$ à valeurs dans Ω_X^l. Cette variable peut être interprétée comme observation simultanée de l variables aléatoires indépendantes et toutes distribuées comme X :

$$X^{(l)} = (X_1, \ldots, X_l)$$

De même, on peut associer à la réception d'un mot de l caractères la variable aléatoire $Y^{(l)}$ à valeurs dans Ω_Y^l. Cette variable peut être interprétée comme observation simultanée de l variables aléatoires indépendantes et toutes distribuées comme Y :

$$X^{(l)} = (Y_1, \ldots, Y_l)$$

On peut considérer alors la transmission d'un message de longueur l comme un nouveau canal $X^{(l)}, Y^{(l)}, P(X^{(l)}|X^{(l)})$, appelé **$l$ème extension de canal** initial. la matrice de transition de ce nouveau canal peut être déduite de

$$P((y_1, \ldots, y_l)|(x_1, \ldots, x_l)) = \prod_{k=1}^{l} p(y_k|x_k)$$

Définition 2.3 (Code de canal)**.** Soit un canal $X, Y, P(Y|X)$. Un code (n, l) pour ce canal est un couple (W, g) où

1. $W = \{w_1, \ldots, w_n\}$ est un ensemble de mots de longueur l dans l'alphabet d'entrée Ω_X, appelés mots du code.
2. g est une règle de décodage

   $$g : (\Omega_Y)^l \to W$$

   qui associe à toute séquence reçue de longueur l dans l'alphabet de sortie Ω_Y un des mots du code.

Définition 2.4 (Probabilité d'erreur conditionnelle)**.** Soient un canal $X, Y, P(Y|X)$ et un code (n, l) pour ce canal (W, g). Pour chaque mot du code w_i on définit la probabilité d'erreur conditionnelle

$$\lambda_i^{(l)} = P[E|w_i] = \sum_{(y_1,\ldots,y_l)\in(\Omega_Y)^l} P((y_1,\ldots,y_l)|w_i)(1 - \delta_{g(y_1,\ldots,y_l),w_i}).$$

Définition 2.5 (Probabilité d'erreur moyenne)**.** Soient un canal $X, Y, P(Y|X)$ et un code (n, l) pour ce canal (W, g). On définit la probabilité d'erreur moyenne (algébrique) du code par :

$$\lambda^{(l)} = \frac{1}{n}\sum_{i=1}^{n} \lambda_i^{(l)}.$$

Définition 2.6 (Probabilité d'erreur maximale)**.** Soient un canal $X, Y, P(Y|X)$ et un code (n, l) pour ce canal (W, g). On définit la probabilité d'erreur maximale du code par :

$$\lambda^{(l)} = \max_{i=1,\ldots,n} \lambda_i^{(l)}.$$

Définition 2.7 (Débit de communication d'un code)**.** Soient un canal $X, Y, P(Y|X)$ et un code (n, l) pour ce canal (W, g). On définit le débit de communication du code par :

$$R = \frac{log(n)}{l}.$$

l'unité de mesure est Shannon par symbole transmis.

3.2 Second théorème de Shannon

Nous avons maintenant pourvoir poser le problème de codage en présence de bruit :

Étant donné un canal $X, Y, P(Y|X)$ de capacité C est il possible de transmettre des messages avec un débit aussi proche que possible de C et avec une probabilité d'erreur aussi petite que possible ?

La réponse à ce problème est donnée par le théorème suivant :

Théorème 2.6 (Second théorème de Shannon)**.** *Soit un canal $X, Y, P(Y|X)$ de capacité $C > 0$. Pour tout $R < C$ il est possible de trouver un code de canal avec un débit R et une probabilité d'erreur aussi petite que possible. Plus précisément, il existe une suite de codes $(M(l), l)$ tels que $M(l) = 2^{lR}$ telle que*

$$\lim_{l \to \infty} \lambda^{(l)} = 0$$

Chapitre 3

Construction de codes optimaux

Le premier théorème de Shannon, vu dans le chapitre précédent établit l'*existence* de codes dont la longueur moyenne de mots est aussi proche de la borne inférieure que l'on veut. Il ne propose cependant pas de méthode pour construire un tel code.

Rappelons qu'un code est dit absolument optimal pour une source donnée (représentée par son alphabet et sa distribution de probabilité) si sa longueur moyenne de mots atteint la borne inférieure. Nous avons mentionné dans le chapitre précédent que tels codes n'existent pas toujours pour une source donnée.

Définition 3.1 (Code optimal)**.** Soit une source X d'alphabet $\Omega_X = \{x_1, \ldots, n\}$ et de distribution de probabilité P_X donnés. On dit qu'un code est **optimal** dans une classe de codes pour cette source si sa longueur de mots de code est minimale dans la classe considérée.

Dans ce chapitre nous nous intéressons au problème de construction effective de **codes optimaux**. Pour faciliter la compréhension, nous allons nous restreindre à l'étude de **codes binaires sans préfixe** (instantanés). La généralisation des idées que nous allons exposer ici au cas d'alphabets de codes q-aires (de taille q) n'est pas difficile. De plus, les codes binaires occupent une grande place parmi les codes actuellement utilisés tout simplement à cause des principes de fonctionnement des ordinateurs.

1 Codes binaires instantanés et arbres

1.1 Quelques rappels sur les arbres

Nous rappelons ici quelques notions utiles dans la suite sur les arbres. Etant donné que nous avons choisi de présenter dans ce chapitre seulement les codes binaires, nous n'aurons besoin que d'*arbres binaires*. On peut consulter pour plus de détails le cours sur les graphes de Marietta Manolessou dans la matière "Algorithmique II".

Un arbre binaire est un cas particulier de graphe qui peut être facilement défini de façon recursive : *un arbre binaire est soit vide soit composé d'un nœud particulier, appelé racine, d'un sous-arbre gauche et d'un sous-arbre droit qui sont eux mêmes des arbres binaires disjoints.* Nous allons tirer de cette définition quelques propriétés essentielles et ne description plus détaillée de la structure d'un arbre binaire.

1. Comme tout graphe, un arbre binaire est un couple (N, R) où N est un ensemble de nœuds et $R \subset N \times N$ est un ensemble d'arcs reliant certains sommets. Un arbre binaire est graphe orienté dans lequel les arcs sont considérés comme des relations "père-fils".
2. La racine de l'arbre est l'unique nœud qui n'a pas de père.
3. Dans un arbre binaire, chaque nœud a au plus deux fils et chaque nœud sauf la racine a exactement un père.
4. Les nœuds qui n'ont pas de fils s'appellent feuilles de l'arbre.
5. les nœuds qui ne sont ni feuilles ni racine s'appellent internes.
6. Un chemin entre deux nœuds est une suite d'arcs consécutifs (deux arcs (i, j) et (k, l) sont consécutifs si l'extrémité du premier est l'origine de l'autre : $j = k$). La longueur d'un chemin est le nombre de branches qui le constituent.
7. Tout nœud est relié à la racine par un unique chemin. On dit qu'un nœud est de niveau n si le chemin qui le relie à la racine est de longueur n.
8. On appelle hauteur ou profondeur d'un arbre la longueur du plus long chemin partant de la racine.
9. Un arbre binaire complet de profondeur n est un arbre dont tous les nœuds sauf les feuilles ont exactement 2 fils. Autrement dit, toutes les feuilles appartiennent au même niveau et tous les nœuds ont exactement 2 ou 0 fils. Un tel arbre a exactement 2^n feuilles et pour tout $i \leq n$ il y a exactement 2 nœuds de niveau i.
10. On appelle arbre binaire incomplet un arbre binaire obtenu à partir d'un arbre complet en supprimant tous les successeurs d'un certain nombre de nœuds ainsi que toutes les branches qui touchent ces nœuds. Les nœuds dont a supprimé les successeurs deviennent les feuilles. Attention ! Il existe des arbres qui ne sont ni complets, ni incomplet. Un arbre incomplet est un arbre dont tous les nœuds ont 2 ou 0 fils. A la différence d'un arbre complet, il n'est pas exigé que toutes les feuilles soient de même niveau. **Attention !** Il existe des arbres qui ne sont ni complets ni incomplets.

11. Dans un arbre binaire complet ou incomplet le nombre de feuilles F et le nombre de branches B vérifient la relation

$$B = 2(F - 1)$$

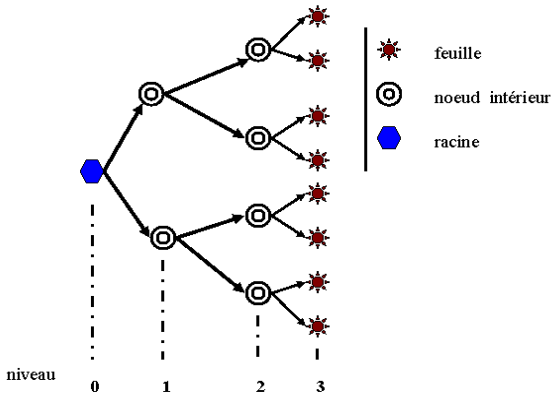

FIG. 3.1 – Un arbre binaire complet de profondeur 3.

Sur la figure 3.1 on peut voir un arbre binaire complet de profondeur 3 et plus loin sur la figure 3.2un arbre incomplet obtenu à partir du premier en supprimant les descendants de deux noeuds.

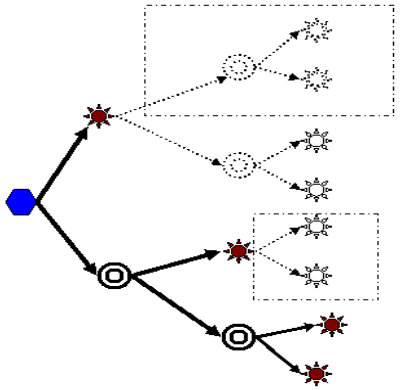

FIG. 3.2 – Un arbre binaire incomplet

1.2 Représentation de codes instantanés par les arbres

Soit $\{m_1, \ldots, m_n\}$ un code binaire sans préfixes de longueur maximale l. Il est évident que chaque mot de ce code peut être représenté par un chemin partant de la racine d'un

arbre binaire complet de profondeur l. il suffit pour cela d'étiqueter les arcs du l'arbre avec 0 et 1. Supposons qu'à un mot m_i de longueur $l_i \leq l$ on vient d'associer un chemin de l_i arcs en partant de la racine. Le chemin s'arrête alors à un noeud de niveau l_i. Comme aucun autre mot du code ne peut avoir celui ci comme préfixe, on peut supprimer tous les descendants du noeud final du chemin. Ce dernier devient alors une feuille. Ainsi on peut associer à tout code sans préfixes un arbre binaire incomplet. L'arbre vu précédemment, est ainsi associé au code $\{1, 01, 001, 000\}$ (voir la figure **??**).

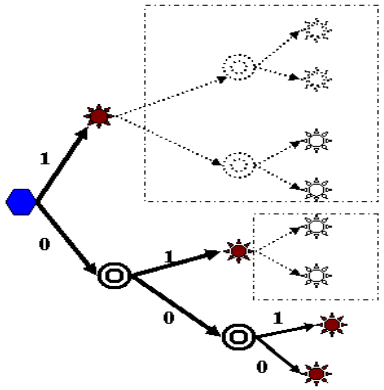

FIG. 3.3 – Représentation d'un code par un arbre binaire incomplet

Ainsi dans un arbre correspondant à un code binaire, les feuilles correspondent aux mots du code. Si ce dernier est associé à l'alphabet d'une source, il est également possible d'associer aux feuilles les probabilités des symboles correspondants.

2 Méthode de Huffman de construction de codes optimaux

Nous allons maintenant présenter la méthode de Huffman de construction de codes instantanés optimaux pour une source donnée. Nous commençons par condition nécessaire d'optimalité que nous présentons ici sans démonstration mais dans le but de faciliter la compréhension de la procédure de Huffman.

Lemme 3.1. *(Condition nécessaire d'optimalité) Soit une source X représentée d'alphabet $\Omega_X = \{x_1, \ldots, x_n\}$ et de distribution de probabilités $P_X = \{p_1, \ldots, p_n\}$. Soit $C = \{m_1, \ldots, m_n\}$ un code instantané associé à la source de longueurs de mots $\{l_1, \ldots, l_n\}$. Supposons que les symboles de l'alphabet sont numérotés dans l'ordre décroissant de leurs probabilités :*

$$p_1 \geq p_2 \geq \cdots \geq p_n.$$

Si plusieurs symboles ont la même probabilité, ils sont numérotés dans l'ordre croissant de longueurs de mots de code correspondants.

Si C est optimal pour la source X dans la classe de codes instantanés alors il vérifie les propriétés suivantes :

1. *Les symboles les plus probables ont les longueurs de mots de code plus petites :*

$$p_i \geq p_j \quad \Rightarrow \quad l_i \leq l_j$$

2. *Les deux derniers symboles ont les longueurs de mots de code égales :*

$$l_n = l_{n-1}$$

3. *Parmi les symboles dont la longueur de mots de code est l_n il y a au moins deux dont les mots de code ont tous les chiffres identiques sauf le dernier.*

La méthode de Huffman (1952) procède de façon recursive, en construisant l'arbre binaire du code à partir des feuilles du niveau le plus élevé. D'après le lemme ci-dessus ces feuilles doivent correspondre aux symboles les moins probables x_n et x_{n-1}. Le principe général de la construction est de regrouper les deux symboles les moins probables en un seul, $w_{n,n-1}$ en additionnant leurs probabilités et de considérer une nouvelle source de $n-1$ symboles. On parle alors de réduction de source. On applique ainsi le principe récursivement jusqu'à ce qu'il ne reste que deux symboles. On leur attribue alors le code $\{0,1\}$.

Le parcours inverse permet de construire le code de manière également recursive. Soit C_{n-1} le code associé à la source de $n-1$ symboles dont le dernier est $w_{n,n-1}$ a été obtenu en regroupant x_n et x_{n-1}. Alors le code C_n associé à la source de n symboles est obtenu de façon suivante. Les mots du code associés aux symboles x_n et x_{n-1} sont construits en ajoutant au mot code du symbole $w_{n,n-1}$ respectivement 0 et 1.

Voici un exemple.

Etape 1		Etape 2		Etape 3		Etape 4		Etape 5	
x_1	0.3	x_1	0.3	x_1	0.3	$x_{3,4,5,6}$	0.45	$x_{3,4,5,6}$	0.45
x_2	0.25	x_2	0.25	x_2	0.25	x_1	0.3	$x_{1,2}$	0.55
x_3	0.2	x_3	0.2	$x_{4,5,6}$	0.25	x_2	0.25		
x_4	0.1	$x_{5,6}$	0.15	x_3	0.2				
x_5	0.1	x_4	0.1						
x_6	0.05								

L'arbre représentant le code résultant est représenté sur la figure suivante :

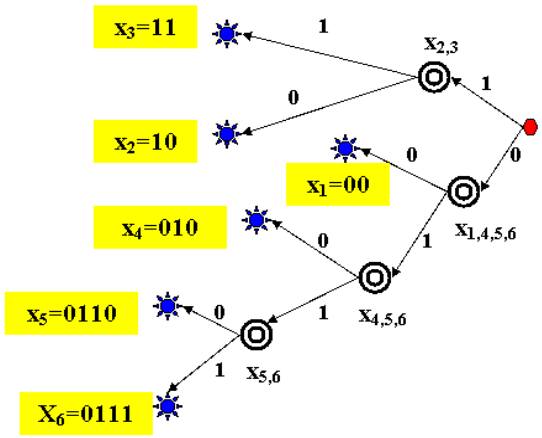

FIG. 3.4 – Arbre de code de Huffman

On y retrouve les codes associés à tous les symboles de l'alphabet initial.

Index