2. 4. Source et codage de source

2.4.1. Modélisation d'une source d'information

Source discrète : système émettant régulièrement des symboles issus d'un alphabet fini.

Alphabet : ensemble fini des symboles de la source.

Source aléatoire : les symboles sont émis aléatoirement suivant les probabilités :

Source sans mémoire : source aléatoire dont les symboles émis sont statistiquement indépendants.

Source discrète : une loi de probabilité donnée associée à une variable aléatoire discrète.

Source continue : une loi de densité de probabilité associée à une variable aléatoire continue.

2.4.2. Codage de source

Le codage de source ou compression des données sert à fournir une représentation efficace des données (un taux de compression important) tout en préservant l'information essentielle qu'elles portent. Il est employé pour le stockage ou la transmission de ces données (on appelle données le résultat de la numérisation de signaux comme ceux de parole ou d'images ou plus généralement les données disponibles sur un fichier d'ordinateur) .
Le codage de source est d'autre part connecté à d'autres applications techniques telles que la classification d'images, la reconnaissance vocale,...

<u>Exemple1 :</u>

Soit une source d'information qui fournit comme information l'une des quatre lettres a1,a2,a3,a4. Supposons que le codage de source transforme cette information discrète en symboles binaires. Nous donnons deux exemples de codages différents.

Codage 1	**Codage 2**
a1→00 a2→01 a3→10 a4→11	a1→0 a2→10 a3→110 a4 → 111

Dans la première méthode, deux symboles binaires sont générés pour chaque lettre émise, alors que dans la seconde le nombre de symboles est variable.
Si les quatre lettres sont équiprobables, alors la première méthode est la meilleure : 2 symboles par lettre en moyenne au lieu de 2,25.
En revanche si l'on a : P(a1) = ½, P(a2) =1/4, P(a3) = P(a4) =1/8,

Alors la méthode 1 nécessite toujours 2 symboles binaires par lettre en moyenne alors que la méthode 2 qui n'en nécessite que 1,75. Elle est dans ce cas la plus économique.
Il est donc important pour coder correctement une source de connaitre son comportement statistique.

Exemple 2 :

On cherche à comprimer l'image 4x4 suivante où chaque pixel est caractérisé par une couleur parmi 4 : R = « rouge », J = « jaune », B = « bleu », V = « vert ».

R	R	R	R
R	B	R	B
R	B	V	B
R	J	J	J

Codage sans perte :

La méthode triviale pour transmettre cette information sous forme binaire est d'associer un mot de code sur 2 bits à chaque couleur.
Ex : R = « 00 » B = « 01 » V = « 10 » J = « 11 »
L'image peut être codée sur 32 bits en parcourant les lignes de l'image de haut en bas et de gauche à droite :

00 00 00 00, 00 01 00 01, 00 01 10 01, 00 11 11 11

Une méthode plus efficace tiendra compte des probabilités de chaque couleur en opérant l'affectation de mots de taille différente à chaque couleur R = « 1 » B = « 01 » J = « 001 » V = « 000 » ce qui conduit à coder l'image sur 28 bits.

1 1 1 1 1 0 1 1 0 1 1 0 1 0 0 0 0 1 1 0 0 1 0 0 1 0 0 1

Codage avec perte :

Pour chaque sous bloc 2x2 de l'image on garde la couleur dominante, c'est-à-dire
:

R	R
R	J

Ce qui donne après le codage « trivial »
00 00 00 11 (8 bits).
Et après le codage « entropique »
1 1 1 001 (6 bits)
Mais bien sur on ne peut pas revenir à l'image initiale avec ce codage.

2.4.3. Entropie d'une source

Définition : **l'entropie** d'une source sans mémoire, d'alphabet *A*, est l'espérance mathématique de la quantité d'information prise comme variable aléatoire.

$$H(A) = E[I(s_k)]$$
$$= \sum_{k=0}^{K-1} p_k I(s_k)$$
$$H(A) = -\sum_{k=0}^{K-1} p_k \log_2 p_k$$

Remarques :

- L'entropie est une mesure de l'information moyenne par symbole issue de la source.
- l'unité de l'entropie est le bit/symbole.

Propriétés :
Pour une source discrète sans mémoire, l'entropie est bornée :

$$0 \leq H(A) \leq \log_2 K$$

1. $H(A) = 0$ ssi $p_k = 1$ pour un k donné, les autres probabilités étant nulles.
=>*aucune incertitude*
2. $H(A) = \log_2 K$ ssi $p_k = 1/K$ pour tout k .
=> *incertitude maximale*

2.5. Canaux et codage de canaux

2.5.1. Modélisation du canal de transmission

Pour modéliser un canal de transmission, il est nécessaire de spécifier l'ensemble des entrées et l'ensemble des sorties possibles. Le cas le plus simple est celui du canal discret sans mémoire.

- L'entrée une lettre $\in$ A= $\{a_1,...,a_n\}$
- La sortie une lettre $\in$ B= $\{b_1,...,b_m\}$

Ces lettres sont émises en séquence. Le canal est sans mémoire si chaque lettre de la séquence reçue ne dépend statistiquement que de la lettre émise de même position.

2.5.2. Canaux discrets sans mémoire

Définition : un *canal discret sans mémoire* est un *modèle statistique* comportant une entrée *X* (v.a.) et une sortie *Y* (v.a.) qui est une version bruitée de *X*.

A chaque unité de temps, la source émet un symbole issu de l'alphabet $X= \{a_1,a_2,\ldots,a_k\}$

La sortie du canal discret sans mémoire est un symbole issu de l'alphabet $Y= \{ b_1,b_2,\ldots,b_j \}$.

Le canal est aussi caractérisé par des *probabilités de transition* :

$P_{Y/X}$, i.e ; une matrice stochastique qui est la matrice du canal

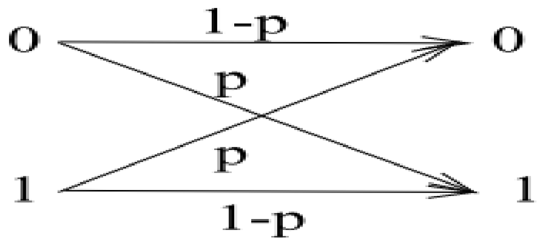
$$\Pi= \begin{vmatrix} P(b_1/ a_1) & P(b_2/ a_1) & \ldots\ldots & P(b_j/ a_1) \\ \ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ \ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ P(b_1/ a_k) & \ldots\ldots & \ldots\ldots & P(b_j/ a_k) \end{vmatrix}$$

Le canal est sans mémoire si pour tout ($x_1,\ldots,x_n$) transmis et ($y_1,\ldots, y_n$) reçu, on a :

$$P(y_1,\ldots, y_n/ x_1,\ldots,x_n) = P(y_1/x_1)\ldots P(y_n/x_n)$$

Exemple : Le canal binaire symétrique

Le plus connu est le canal binaire symétrique défini par $X = Y=\{0,1\}$ et dont les probabilités de transitions sont données par la figure suivante.



P est appelé probabilité de transition ou probabilité d'erreur du canal.

2.5.3. Capacité d'un canal

Définition : $H(A) - H(A \mid B)$ est appelée *information mutuelle du canal* :

$$I(A,B) = H(A) - H(A \mid B)$$

Propriétés :

- $I(A,B) = I(B,A)$ Symétrie

- $I(A,B) \geq 0$ on ne peut perdre d'information

- $I(A,B) = H(A) + H(B) - H(A,B)$, avec

- $H(A,B) = -\sum_{j=0}^{J-1}\sum_{k=0}^{K-1} p(x_j, y_k)\log_2 p(x_j, y_k)$

Remarques :

- Le calcul de l'information mutuelle $I(A, B)$ nécessite la connaissance de la distribution *a priori* de X ;

- En conséquence, l'information mutuelle d'un canal dépend non seulement du canal mais également de la manière dont il est utilisé ;

- Or la distribution *a priori* est (évidemment) indépendante du canal ;
- On peut donc chercher à maximiser l'information mutuelle par rapport à $\{p(x_j)\}$

Définition : la *capacité* d'un canal discret sans mémoire est le maximum de l'information mutuelle *I*(*A*, *B*) moyenne obtenu pour l'ensemble des symboles émis, la maximisation étant opérée sur toutes les distributions *a priori* possibles $\{p(x_j)\}$ sur *A*..

$$C = \max_{\{p(x_j)\}} I(A,B)$$

C se mesure en bits par utilisation du canal (*bits per channel use*).

Remarque : le calcul de *C* implique la maximisation sur *J* variables (les probabilités d'entrée) sous deux contraintes :

$$p(x_j) \geq 0 \forall j$$

$$\sum_{j=0}^{J-1} p(x_j) = 1$$