EISTI. **Année scolaire 2008-2009**

DÉPARTEMENT " INFORMATIQUE "

THÉORIE DE L'INFORMATION

Série d'exercices N^{o}2

PARTIE I. CODES SANS PRÉFIXE.

Nous allons considérer une source d'alphabet

$$\Omega = \{a, b, c, d, e, f, g, h, i, j\}$$

et un canal d'alphabet binaire $\{0, 1\}$.

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 ;

Voici les définitions importantes.

Code. Un code binaire pour l'alphabet donné Ω de taille n est un ensemble de n mots binaires $\{m_1, \ldots, m_n\}$. Un code est régulier si les mots correspondants aux différents caractères de l'alphabet sont différents. Dans la suite nous étudions uniquement les codes réguliers.

Exemple. Pour notre alphabet un code régulier peut être

s_i	a	b	c	d	e	f	g	h	i	j
m_i	0	1	00	01	11	10	100	101	110	111

Code déchiffrable Un code binaire est déchiffrable si toute séquence de bits peut être décodée de façon unique.

Exemple. Le code ci-dessus n'est pas déchiffrable. En effet, la séquence de bits "110" peut être interprétée comme "i" ou comme "bf" ou même comme "ea".

Code sans préfixe. Un code est sans préfixe ou instantané si aucun mot code n'est le préfixe d'un autre. Le code ci-dessus n'est pas un code sans préfixe. En effet, le mot code 0 est le début des mots du code 00, 01. Le code suivant est sans préfixe :

s_i	a	b	c	d	e	f	g	h	i	j
m_i	1111	1110	110	1011	1010	100	011	000	010	001

Décodage pas à pas d'un code sans préfixe. Le principe de décodage est simple. Il suffit de lire la séquence codée de gauche à droite jusqu'à ce qu'on trouve un mot du code. On est alors certain que ce mot correspond sans ambiguïté à un seul caractère de l'alphabet. On enregistre le caractère et on recommence la lecture.

Exercice 1. Appliquez la procédure de décodage pas à pas à la séquence 10110000101010100.

Théorème 0.1 (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.

Exercice 2. Vérifier s'il existe un code sans préfixe de longueurs de mots données : $\{4, 4, 4, 3, 2, 4, 3, 4, 3, 4\}$.

PARTIE II. CODES SANS PRÉFIXE ET ARBRES BINAIRES.

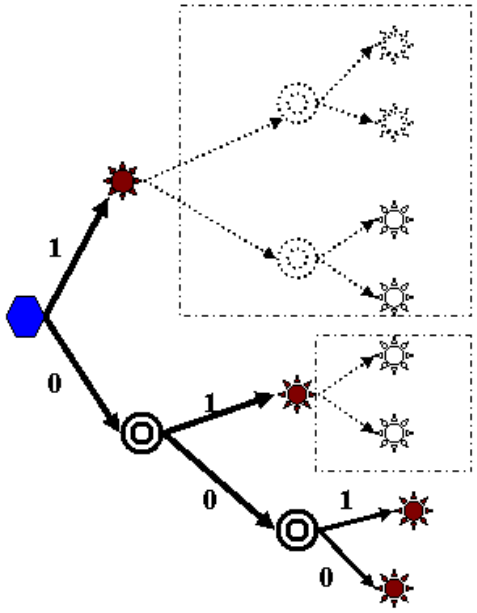

FIGURE 1 – Exemple

Vous trouverez dans le polycopié du cours les rappels nécessaires de vocabulaire associé aux arbres binaires. (voir page 44).

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 1).

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.

Exercice 3.

1. Construire l'arbre associé au code ci-dessous

s_i	a	b	c	d	e	f	g	h	i	j
m_i	1111	1110	110	1011	1010	100	011	000	010	001

2. Construire le code sans préfixe de longueurs de mots données dans l'exercice 2 à l'aide d'un arbre binaire, en procédant à l'élagage d'un arbre complet.

PARTIE III. ENTROPIE ET LE CODAGE

Exercice 4 (Vers le codage de Shannon.)**.** On rappelle ici la propriété de groupe de l'entropie, déjà vu dans le TD précédent.

Soit X une source d'alphabet $\Omega_X = \{x_1, \ldots, x_n\}$ et de distribution de probabilités $P_X = \{p_1, \ldots, p_n\}$. Soit $1 \leq r < n$. On divise l'alphabet en deux sois-ensembles $A = \{x_1, \ldots, x_r\}$ et $B = \{x_{r+1}, \ldots, x_n\}$ de telle sorte que $\Omega_X = A \cup B$ et $A \cap B = \emptyset$. Soit $p = P(A)$ et $q = 1 - p = P(B)$. Alors on a la relation suivante (propriété du groupe) :

$$H(X) = H(p_1, \ldots, p_n) = H(p, 1-p) + pH\left(\frac{p_1}{p}, \ldots, \frac{p_r}{p}\right) + (1-p)H\left(\frac{p_{r+1}}{1-p}, \ldots, \frac{p_n}{1-p}\right)$$

On peut interpréter cette propriété de la façon suivante. Supposons que nous avons une partition de l'ensemble de valeurs possibles de X en deux parties complémentaires, A et B. Alors l'incertitude moyenne (l'entropie) que nous avons sur X est composée de :

1. l'incertitude que nous avons sur le choix de l'une des deux parties A et B ; c'est $H(p, 1-p)$;
2. la moyenne des incertitudes associées à chacune des parties séparément ; c'est $pH\left(\frac{p_1}{p}, \ldots, \frac{p_r}{p}\right) + (1-p)H\left(\frac{p_{r+1}}{1-p}, \ldots, \frac{p_n}{1-p}\right)$.

Soit X une source d'alphabet $\Omega = \{1, 2, 3, 5, 4\}$ de distribution de probabilité $P = \{0.1, 0.2, 0.3, 0.15, 0.25\}$. Supposons que l'on doit deviner le symbole émis par la source et que l'on a droit de poser des questions binaires (réponses possibles "oui" et "non"). On cherche à construire la stratégie qui, en moyenne, permet de trouver la réponse en un nombre minimal de questions.

Remarquons qu'une question binaire induit sur l'ensemble Ω une partition en deux sous-ensembles A et B correspondants aux réponses "oui" et "non". Par exemple, si l'on demande "est que le nombre est impair ?" la partition sera $A = \{1, 3, 5\}$ et $B = \{2, 4\}$. Soit $p = P(A)$ la probabilité de la réponse "oui" à une question donnée.

1. Calculer l'entropie de X.
2. Quelle est l'information moyenne obtenue par la réponse à une question binaire ?
3. Quelle est l'information moyenne maximale ? Et pour quelle valeur de p est est atteinte ?
4. Quel est alors le meilleur choix de première question à poser ? **Indication :** cherchez à diminuer autant que possible votre incertitude, en posant une question.
5. Appliquer le même raisonnement récursivement pour choisir la meilleure deuxième question selon la réponse à la première. Continuer jusqu'à arriver à identifier chaque symbole.
6. Construire un arbre représentant la stratégie obtenue. Si on associe le code 1 à la réponse "oui" et le code 0 à "non", cet arbre définit un code. Est il sans préfixe ?
7. Calculer le nombre moyen de questions (la longueur moyenne des mots de code ?). Comparer à l'entropie de X.

Exercice 5 (Vers le codage de Huffman.)**.** Reprenons la même source que dans l'exercice **??** X d'alphabet $\Omega = \{1, 2, 3, 5, 4\}$ de distribution de probabilité $P = \{0.1, 0.2, 0.3, 0.15, 0.25\}$.

Supposons que l'on doit deviner le symbole émis par la source et que l'on a droit de poser des questions binaires (réponses possibles "oui" et "non"). On cherche à construire la stratégie qui, en moyenne, permet de trouver la réponse en un nombre minimal de questions.

Cette fois, nous recherchons une stratégie dans laquelle à chaque question on "élimine" un symbole. Autrement dit, chaque question est de la forme "Est ce que $X = x_i$? Soit $p_i = P(x_i)$.

1. Quelle est l'information moyenne obtenue par la réponse à une question de type $X = x_i$?
2. Appliquer la propriété de groupe de l'entropie dans ce cas particulier ?
3. Par quel symbole a-t-on intérêt de commencer : le plus probable ou le moins probable ?
4. Construire un arbre représentant la stratégie obtenue.
5. Calculer le nombre moyen de questions. Comparer à l'entropie de X.