

DÉPARTEMENT " INFORMATIQUE "

THÉORIE DE L'INFORMATION

Série d'exercices N°2 Corrigé.

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 $\{0, 1\}$.

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

Solution de l'exercice 1

La réponse est "d (1011)h(000)i(010)e(1010)f(100)".

Théorème 0.1 (Inégalité de Kraft). *Un code instantané de longueurs de mots données l_1, \dots, 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 : $\{4, 4, 4, 3, 2, 4, 3, 4, 3, 4\}$.

Solution de l'exercice 2

Il suffit d'utiliser le théorème. Ici la taille de l'alphabet du canal est $d = 2$ (code binaire !) :

$$6 \left(\frac{1}{2}\right)^4 + 3 \left(\frac{1}{2}\right)^3 + \left(\frac{1}{2}\right)^2 = \frac{6 + 6 + 4}{16} = 1$$

Donc l'inégalité de Kraft est vérifiée, le code existe.

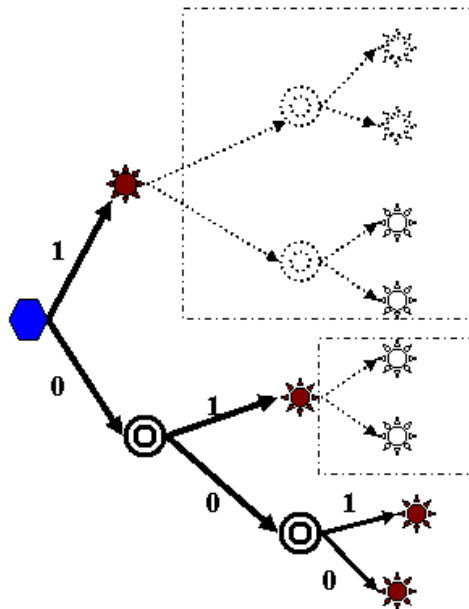


FIGURE 1 – Exemple

PARTIE II. CODES SANS PRÉFIXE ET ARBRES BINAIRES.

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

Soit $\{m_1, \dots, m_n\}$ 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 de 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 $\Omega_X = \{x_1, \dots, x_n\}$ et de distribution de probabilités $P_X = \{p_1, \dots, p_n\}$. Soit $1 \leq r < n$. On divise l'alphabet en deux sous-ensembles $A = \{x_1, \dots, x_r\}$ et $B = \{x_{r+1}, \dots, x_n\}$ 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) :

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

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

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 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 .

Solution de l'exercice 4

1. Calculer l'entropie de X .

$$H(X) = -0.1 \log 0.1 - 0.2 \log 0.2 - 0.3 \log 0.3 - 0.15 \log 0.15 - 0.25 \log 0.25 \simeq 2.23$$

2. Quelle est l'information moyenne obtenue par la réponse à une question binaire ? $H_2(p) = H(p, 1-p)$ représente la quantité moyenne d'information apportée par l'observation d'une source binaire. Ici la source est la réponse à une question et p est la probabilité de la réponse *Oui*.
3. Quelle est l'information moyenne maximale ? Et pour quelle valeur de p est-elle atteinte ? D'après les propriétés de l'entropie (voir Chapitre 1 du poly) elle atteint sa valeur maximale uniquement dans le cas d'une distribution uniforme. Dans le cas binaire, quand $p = 1 - p = 0.5$. On a alors

$$\max_{p \in [0,1]} H_2(p) = H_2(0.5) = \log 2$$

4. Quel est alors le meilleur choix de première question à poser ? Pour répondre à la question, on peut remarquer que la réponse à une question binaire partage l'ensemble de valeurs possibles de X en deux parties disjointes : celle du *Oui* et celle du *Non*. On peut alors appliquer la propriété du groupe :

$$H(X) = H_2(p) + pH_{Oui} + (1-p)H_{Non}$$

On interprète cela de façon suivante. Si $H(X)$ est l'incertitude moyenne que l'on a sur le résultat de tirage X au début du jeu, la quantité $H_2(p)$ représente l'information moyenne apportée par la réponse à la première question et la quantité $pH_{Oui} + (1-p)H_{Non}$ est donc l'incertitude moyenne qui restera sur la valeur de X une fois la réponse obtenue. On a alors intérêt de choisir la première question de manière à ce qu'elle apporte le maximum d'information. Nous cherchons donc une question qui partage l'ensemble de valeurs de X en deux parties équiprobables.

Dans le cas étudié ici c'est possible, en posant la question : $X \in \{2, 3\}$?

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. Construire un arbre représentant la stratégie obtenue. (voir la figure 2).

6. Calculer le nombre moyen de questions. Comparer à l'entropie de X .

A l'aide de graphe on trouve la distribution de probabilité pour le nombre de questions, N :

X	1	2	3	4	5
P_X	0.1	0.2	0.3	0.15	0.25
N	3	2	2	3	2

On peut alors calculer la moyenne de N :

$$E[N] = 3 \cdot 0.1 + 2 \cdot 0.2 + 2 \cdot 0.3 + 3 \cdot 0.15 + 2 \cdot 0.25 = 2.25$$

On remarque alors que la valeur obtenue est proche de l'entropie $H \simeq 2.23$. La méthode utilisée ici est la même que celle de codage de Shannon, que nous allons étudier plus tard.

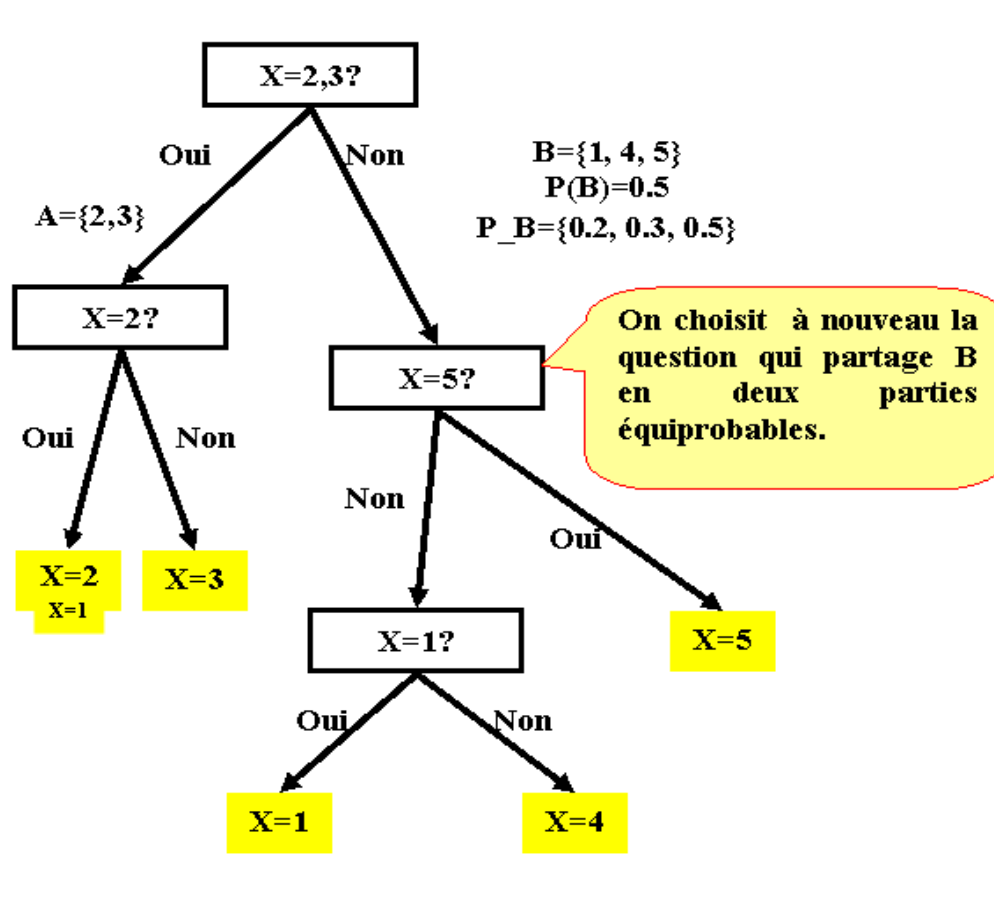


FIGURE 2 – $X_0 = 3$.

Exercice 5 (Vers le codage de Huffman.). Reprenons la même source que dans l'exercice 4 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 .

Solution de l'exercice 5

1. Quelle est l'information moyenne obtenue par la réponse à une question de type $X = x_i$?
Une telle question divise l'alphabet de la source en deux sous-ensembles complémentaires :

$$A = \{x_i\}, \quad B = \{x_k, k \neq i\}$$

A cette partition on associe la distribution de probabilité

$$P(A) = P(\text{"oui"}) = p_i, \quad P(B) = P(\text{"non"}) = 1 - p_i$$

Alors l'information moyenne apportée par la réponse à une telle question est

$$H_2(p_i) = -p_i \log_2(p_i) - (1 - p_i) \log_2(1 - p_i)$$

2. Appliquer la propriété de groupe de l'entropie dans ce cas particulier ?

$$H(X) = H_2(p_i) + p_i H_{Oui} + (1 - p_i) H_{Non} = H_2(p_i) + (1 - p_i) H_{Non}$$

Car l'ensemble A est un singleton.

3. Par quel symbole a-t-on intérêt de commencer : le plus probable ou le moins probable ?
On a intérêt à commencer par le caractère le plus probable.
4. Construire un arbre représentant la stratégie obtenue.

Le principe de construction est simple. A chaque nouvelle question on isole le caractère le plus probable de l'ensemble des caractères restants. Voici le code binaire de l'arbre (1="Oui", 0="non")

x_i	1	2	3	4	5
code	0000	001	1	0001	01

5. Calculer le nombre moyen de questions. Comparer à l'entropie de X .
On utilise le tableau pour calculer :

$$E[N] = 4 * 0.1 + 3 * 0.2 + 1 * 0.3 + 4 * 0.15 + 2 * 0.25 = 2.4$$