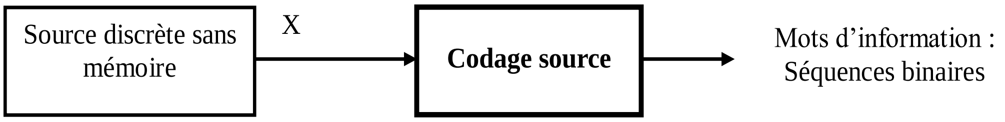
Chapitre

2

Codage des sources discrètes sans mémoire

I.1 Introduction

Cette partie s'intéresse au codage de la sortie d'une source discrète sans mémoire en une séquence binaire. En effet, ce codage devra permettre de retrouver la séquence de lettre de sortie à partir du code binaire. Ce codage est une procédure qui associe à chaque symbole de la source une séquence binaire appelé mot d'information.



$X=\{X_1, X_2, \ldots\ldots\ldots\ldots\ldots ,X_m\}$

$P(X_i)=\{P(X_1), P(X_2),\ldots\ldots\ldots\ldots\ldots\ldots,P(X_m)\}$

I.2 Différents type de codage des sources discrètes sans mémoires

I.2.1 Définition :

Code régulier : Un code est dit régulier si deux symboles distincts de la source sont codés à l'aide de deux mots d'information différents.

Longueur moyenne d'un code : Soit X une source discrète sans mémoire avec une entropie H(X), un alphabet $X=\{X_1, X_2,\ldots\ldots\ldots,X_m\}$ et ses caractéristiques d'émission régies par une loi de probabilité P : $P(Xi)= \{P(X_1), P(X_2), \ldots\ldots, P(X_m)\}$. La longueur moyenne des mots d'information qui représente le nombre moyen des bits par symbole de la source d'information utilisé dans le processus de codage est définie par l'équation suivante : $\overline{L} = \sum_{i=1}^{m} L_i P(L_i)$

Avec L_i représente la longueur du mot d'information associé à X_i.

Efficacité d'un code : L'efficacité d'un code d'une source X d'entropie H(X) est définie par : $E = \frac{H(X)}{\overline{L}}$

I.2.2 Code de longueur fixe :

Un code de longueur fixe est un code dont tous les mots d'information ont la même longueur. Si m est la taille de l'alphabet d'une source discrète sans mémoire X, il existe un code régulier de X de longueur $\overline{L}$ telle que : $\mathrm{Log}_2 \mathrm{m} \leq \overline{\mathrm{L}} < 1 + Log_2 m$

Exemple : code de longueur fixe :

Soit une source discrète sans mémoire dont l'alphabet de sortie est $X=\{X_1, X_2,\ldots\ldots\ldots,X_{10}\}$muni de la loi de probabilité uniforme.

1- Déterminer la longueur de ce code fixe.
2- Proposer un code pour cette source.
3- Calculer l'efficacité de ce code.

Réponse :

1- Longueur du code fixe :

On a m=10

$$\Rightarrow \mathrm{Log}_2 10 \leq \overline{\mathrm{L}} < 1 + Log_2 10 \Rightarrow 3.32 \leq \overline{\mathrm{L}} < 4.32 \Rightarrow \overline{L} = 4$$

2- Codage :

Symboles	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
Mots d'information	0000	0001	0010	0011	0100	0101	0110	0111	1000	1001

3- Efficacité :
Calcul de H(X)
On a $P(X_i)=1/10$

$$H(X)=-\sum_{i=1}^{10} P(X_i)Log_2(P(X_i))$$

$$=-10*\frac{1}{10}\text{Log}_2\frac{1}{10}$$

$$=3.32\,Bits/symbole$$

$$\Rightarrow E=\frac{H(X)}{\bar{L}}=\frac{3.32}{4}=0.83$$

$$=83\%$$

Ce code n'est pas optimal puisque six mots binaires sont non utilisés.

I.2.3 Code de longueur variable :

Le code de longueur variable est un code utilisé afin d'augmenter l'efficacité à 100%, c'est-à-dire la longueur moyenne du code soit la plus faible possible.

I.2.3.1 Principe

Pour introduire les notions essentielles nous allons utiliser un exemple de trois codage possible pour une source discrète sans mémoire qui comporte 4 symbole X= {A, B, C, D}.

Symboles	Probabilité	Code I	Code II	Code III
A	1/2	1	0	0
B	1/4	00	01	10
C	1/8	01	011	110
D	1/8	10	111	111

Supposons que nous cherchons à transmettre le message BDC

Interprétation du code I :
Message envoyé : 001001
On suppose qu'il n'ya pas de bruit dans la transmission
Donc le message reçu avant le décodeur est : 001001.

Interprétation du message reçu : 00 1 00 1
B A B A

On remarque qu'il y a une ambigüité. En effet ce codage n'est pas décodable de manière unique. Ceci est du au fait que le « 1 », code attribuer à « A » est le début du code attribué à « D ». Pour éviter cette ambigüité, Il ne faut jamais qu'un code soit le début d'un autre code.
Définition :

- Un code est dit uniquement décodable si toute séquence émies est interprétable de manière unique.
- Un code vérifie la condition de préfixe si aucun mot d'information n'est le début d'un autre mot d'information.

Interprétation du code II :
Message envoyé : 01111011
On suppose qu'il n'ya pas de bruit dans la transmission
Donc le message reçu avant le décodeur est : 01111011.

0 111 1
A D 10
Interprétation du message reçu : 101
1011
Ne sont pas des mots d'information

Dans cette interprétation du message reçu on revient en arrière pour trouver le bon message. On dit que le code n'est pas décodable de manière instantanée. De plus ce code n'est pas préfixe donc non déchiffrable de façon unique.

Interprétation du code III :
Message envoyé : 10111110
On suppose qu'il n'ya pas de bruit dans la transmission
Donc le message reçu avant le décodeur est : 10111110.
Interprétation du message reçu : 10 111 110
B D C

Ce code est préfixe avec les deux propriétés souhaitées décodable de manière unique et de manière instantanée.

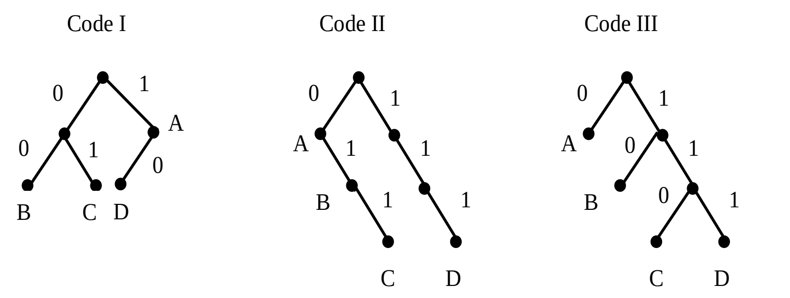
I.2.3.2 Arbre de décision d'un code

Pour tout code préfixe, il existe un arbre (arbre de décision) dont tous les mots d'information sont des feuilles.

Construction de l'arbre de décision :

- Un déplacement à gauche correspond à un « 0 »
- Un déplacement à droite correspond à un « 1 »
- Chaque déplacement crée un nœud et une branche
- Un nœud qui présente aucun descendant est une feuille

Représentation des arbres de décision :



A n'est pas une feuille $\Rightarrow$ Code non déchiffrable

A et B ne sont pas des feuilles $\Rightarrow$ Code non déchiffrable

Tous les symboles sont des feuilles $\Rightarrow$ Code déchiffrable

I.2.3.3 Inégalité de Kraft

Soit X une source discrète sans mémoire avec un alphabet $X=\{X_1, X_2,\ldots\ldots\ldots,X_m\}$. On suppose que la longueur d'un mot d'information binaire associé à X_i soit L_i. Une condition nécessaire pour l'existence d'un code binaire instantané, appelée l'inégalité de kraft est définit par :

$$K = \sum_{i=1}^{m} 2^{-L_i} \leq 1$$

Exemple :

Considérons une source binaire sans mémoire avec un alphabet $X=\{X_1, X_2, X_3, X_4\}$. Pour cette source on propose les codes suivants comme indiqué dans le tableau suivant :

Symboles	Code C_1	Code C_2	Code C_3	Code C4
X_1	00	0	0	0
X_2	01	10	11	100
X_3	10	11	100	110
X_4	11	110	110	111

1- Montrer est ce que ces codes vérifient l'inégalité de Kraft.
2- Montrer est ce que ces codes sont décodable de manière unique.

Réponse :

1- Calcul de l'inégalité de Kraft :

Code C_1 : $L_1=L_2=L_3=L_4=2$.

$$\Rightarrow K = -\sum_{i=1}^{4} 2^{-L_i} = 4 * 2^{-2} = 1$$

$\Rightarrow$ Le code C_1 vérifie l'inégalité de Kraft.

Code C_2 : $L_1=1$, $L_2=L_3=2$, $L_4=3$.

$$\Rightarrow K = -\sum_{i=1}^{4} 2^{-L_i} = 2^{-1} + 2 * 2^{-2} + 2^{-3} = \frac{9}{8}$$

$\Rightarrow$ Le code C_2 ne vérifie pas l'inégalité de Kraft.

Code C_3 : $L_1=1$, $L_2=2$, $L_3=L_4=3$.

$$\Rightarrow K = -\sum_{i=1}^{4} 2^{-L_i} = 2^{-1} + 2^{-2} + 2 * 2^{-3} = 1$$

$\Rightarrow$ Le code C_3 vérifie l'inégalité de Kraft.

Code C_4 : $L_1=1$, $L_2=L_3=L_4=3$.

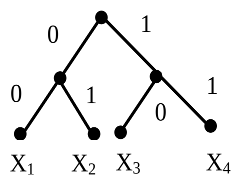
$$\Rightarrow K = -\sum_{i=1}^{4} 2^{-L_i} = 2^{-1} + 3 * 2^{-3} = \frac{7}{8}$$

$\Rightarrow$ Le code C_4 vérifie l'inégalité de Kraft.

2- Le Code C_2 ne vérifie pas l'inégalité de Kraft donc c'est un code qui n'est pas décodable de manière unique.

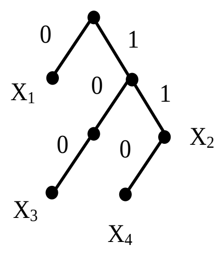
Pour les codes C_1, C_3 et C_4, l'analyse de leurs natures est assurée par l'interprétation de leurs arbres de décision :

Code C1



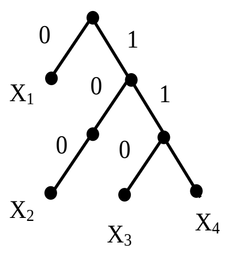
Tous les symboles sont des feuilles
$\Rightarrow$ Code déchiffrable

Code C3



X_2 n'est pas une feuille
$\Rightarrow$ Code non déchiffrable

Code C4



Tous les symboles sont des feuilles
$\Rightarrow$ Code déchiffrable

I.2.3.4 Théorème de Shannon

Soit une source binaire discrète sans mémoire $X=\{X_1, X_2, \ldots\ldots\ldots, X_m\}$ munie d'une loi de probabilité P : $P(Xi)= \{P(X_1), P(X_2), \ldots\ldots, P(X_m)\}$.

- Pour toute source d'entropie H(X) codée au moyen d'un code déchiffrable de longueur moyenne $\overline{L}$ On a : $\overline{L} \geq H(X)$
- Pour toute source d'entropie H(X), il existe code préfixe dont la longueur moyenne $\overline{L}$ est telle que : $H(X) \leq \overline{L} < H(X)+1$

L'inégalité est obtenue lorsque : $P(X_i) = 2^{-L_i} \forall i$

$$H(X) = -\sum_{i=1}^{m} P(X_i) Log_2(P(X_i))$$

$$\Rightarrow H(X) = -\sum_{i=1}^{m} P(X_i) Log_2 2^{-L_i}$$

On a : $\Rightarrow H(X) = \sum_{i=1}^{m} P(X_i) * L_I Log_2 2$

$$H(X) = \sum_{i=1}^{m} L_i P(X_i)$$

$$H(X) = \overline{L}$$

Efficacité :

$$H(X) \leq \overline{L} < H(X)+1 \quad \text{avec E} = \frac{\text{H(X)}}{\overline{\text{L}}} \Rightarrow \overline{L} = \frac{H(X)}{E}$$

$$\Rightarrow H(X) \leq \frac{H(X)}{E} < H(X)+1$$

$$\frac{1}{H(X)+1} < \frac{E}{H(X)} \leq \frac{1}{H(X)}$$

$$\Rightarrow \frac{H(X)}{H(X)+1} < E \leq 1$$

$$\Rightarrow 1 - \frac{1}{H(X)+1} < E \leq 1$$

Pour toute source discrète sans mémoire, il existe un code déchiffrable dont l'efficacité est arbitrairement proche de 1.

I.3 Codage entropique : Codage de Huffman

La conception d'un code à longueur variable de telle façon que sa longueur moyenne s'approche de l'entropie de la source est un codage entropique. Le codage de Huffman est un codage entropique qui permet d'élaborer un code déchiffrable et optimal.

I.3.1 Arbre de codage

L'arbre de Huffman permet de développer un code pour chaque symbole de l'alphabet de la source en fonction de sa fréquence.
La procédure du codage de Huffman suit les étapes suivantes :

- Ranger les symboles de la source avec un ordre croissant des probabilités.
- Combiner successivement deux à deux les symboles de plus faible probabilité P_i, P_j.
- Former un nœud dont la probabilité représente la somme P_i+P_j.
- Arriver à la racine de l'arbre dont la probabilité est égale à 1.

- Etiqueter les branches par des zéros à gauche et des 1 à droite.
- Elaborer tous les mots d'information de chaque symbole de la source par une lecture des étiquettes de l'arbre. La lecture des séquences binaires s'effectuent à partir de la racine de l'arbre jusqu'aux symboles source.

I.3.2 Exemple de codage de Huffman

Une source binaire sans mémoire X possède cinq symboles X_1, X_2, X_3, X_4 et X_5 avec les probabilités suivantes : $P(X_1)= 0.1$, $P(X_2)=0.1$, $P(X_3)= 0.25$, $P(X_4)=0.35$ et $P(X_5)=0.2$.

1- Calculer l'entropie de cette source.
2- Proposer un codage régulier de longueur fixe de cette source et calculer son efficacité E_1.
3- Elaborer un code de Huffman de cette source et calculer son efficacité E_2.

Réponse :
1- Calcul de l'entropie de la source :

$$H(X) = -\sum_{i=1}^{6} P(X_i) Log_2(P(X_i))$$

$$= -P(X_1)Log_2P(X_1) - P(X_2)Log_2P(X_2) - P(X_3)Log_2P(X_3) - P(X_4)Log_2P(X_4)$$
$$- P(X_5)Log_2P(X_5)$$
$$= -2*0.1 \ \mathrm{Log}_2 0.1 - 0.25 \mathrm{Log}_2 0.25 - 0.35 \mathrm{Log}_2 0.35 - 0.2 \mathrm{Log}_2 0.2$$
$$= 2.158 \ Bits/symbole$$

2- Code de longueur fixe :
Calcul de la longueur du code fixe :
On a m=5

$$\Rightarrow \mathrm{Log}_2 5 \le \overline{\mathrm{L}} \le 1 + Log_2 5 \Rightarrow 2.32 \le \overline{\mathrm{L}} \le 3.32 \Rightarrow \overline{L}_1 = 3$$

Codage proposé :

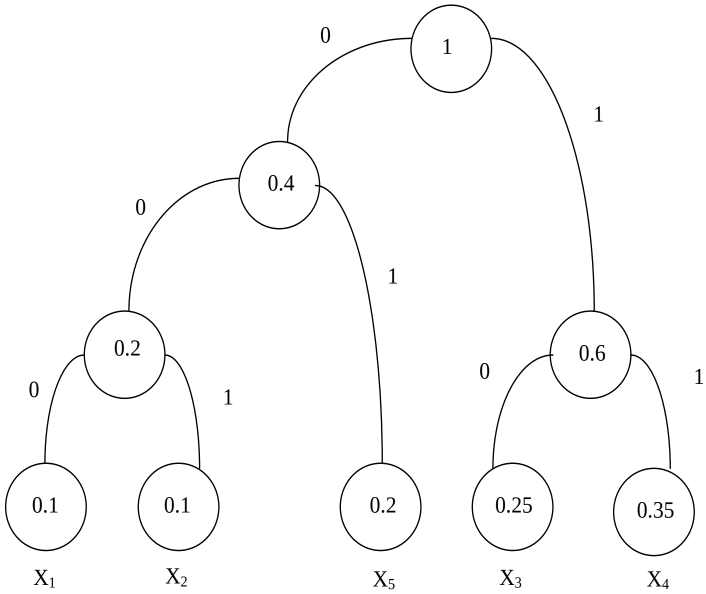
Symboles	X_1	X_2	X_3	X_4	X_5
Mots d'information	000	001	010	011	100

Calcul de l'efficacité E_1:

$$\Rightarrow E_1 = \frac{H(X)}{\overline{L_1}} = \frac{2.32}{3} = 0.72$$
$$= 72\%$$

3- Code de longueur Variable : Code de Huffman :

Symboles	Probabilités	Codes $C(X_i)$	Longueur L_i
X_1	0.1	000	3
X_2	0.1	001	3
X_3	0.25	10	2
X_4	0.35	11	2
X_5	0.2	01	2



Calcul de la longueur moyenne du code de Huffman :

$$\overline{L}_2 = \sum_{i=1}^{5} L_i P(L_i)$$

$$\Rightarrow \overline{L}_2 = 2*[P(X_3) + P(X_4) + P(X_5)] + 3*[P(X_3) + P(X_4)]$$

$$\Rightarrow \overline{L}_2 = 2.2 Bits/symbole$$

Calcul de l'efficacité E_2:

$$\Rightarrow E_2 = \frac{H(X)}{\overline{L_2}} = \frac{2.32}{2.2} = 0.98$$

$$= 98\%$$

Les résultats montrent que le code de Huffman est plus efficace que le code de longueur fixe ⇒ le code de Huffman est un code optimal.

I.1 Extension de source

On appelle extension de source d'ordre K, la source qui s'obtient en groupant les symboles de la source par paquet de K symboles consécutif.

Détermination des symboles de l'extension de la source discrète

1- Soit X une source discrète sans mémoire d'alphabet $X=\{X_1, X_2\}$. Déterminer les symboles de l'extension d'ordre 2 et d'ordre 3 de cette source.

2- Soit X une source discrète sans mémoire d'alphabet $Y=\{Y_1, Y_2\ Y_3\}$. Déterminer les symboles de l'extension d'ordre 2 de cette source.

Réponse :

1- Extension de la source $X=\{X_1, X_2\}$.

Extension d'ordre 2 :

m=2, K=2 ⇒ $m^K = 2^2 = 4$ Symboles : $\{X_1X_1, X_1X_2, X_2X_1, X_2X_2\}$

Extension d'ordre 3 :

m=2, K=3 ⇒ $m^K = 2^3 = 8$ Symboles : $\{X_1X_1X_1, X_1X_1X_2, X_1X_2X_1, X_1X_2X_2, X_2X_1X_1, X_2X_1X_2, X_2X_2X_1, X_2X_2X_2\}$

2- Extension de la source $Y=\{Y_1, Y_2\ Y_3\}$.

Extension d'ordre 2 :

m=3, K=2 ⇒ $m^K = 3^2 = 9$ Symboles : $\{X_1X_1, X_1X_2, X_1X_3, X_2X_1, X_2X_2, X_2X_3, X_3X_1, X_3X_2, X_3X_3\}$

Etude de l'efficacité de l'extension d'une source discrète

Soit X une source discrète sans mémoire d'alphabet $X=\{X_1, X_2, X_3\}$ munie d'une loi de probabilité P : $P(Xi)= \{0.1, 0.2, 0.7\}$.

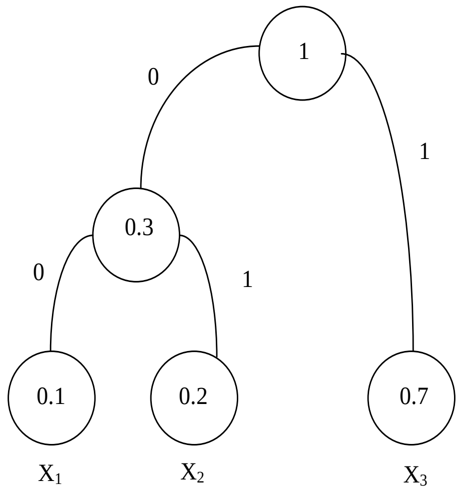
1- Construire le code d Huffman de la source X et calculer son efficacité.

2- Construire le code d Huffman de l'extension d'ordre 2 de la source X et calculer son efficacité.

Réponse :

1- Code de Huffman de la source X:

Symboles	Probabilités	Codes $C(X_i)$	Longueur L_i
X_1	0.1	00	2
X_2	0.2	01	2
X_3	0.7	1	1



Calcul de L'entropie:

$$H(X) = -\sum_{i=1}^{3} P(X_i) Log_2(P(X_i))$$

$$= -P(X_1)Log_2P(X_1) - P(X_2)Log_2P(X_2) - P(X_3)Log_2P(X_3)$$

$$= -0.1 \text{ Log}_2 0.1 - 0.2 \text{ Log}_2 0.2 - 0.7 \text{ Log}_2 0.7$$

$$= 1.157 \ Bits/symbole$$

Calcul de la longueur moyenne:

$$\bar{L} = \sum_{i=1}^{3} L_i P(L_i)$$

$$\Rightarrow \bar{\text{L}} = 2*[\text{P}(\text{X}_1) + \text{P}(\text{X}_2)] + 1*\text{P}(\text{X}_3)$$

$$\Rightarrow \bar{\text{L}} = 2*[0.1 + 0.2] + 1*0.7$$

$$\Rightarrow \bar{\text{L}} = 1.3 Bits/symbole$$

Calcul de l'efficacité E:

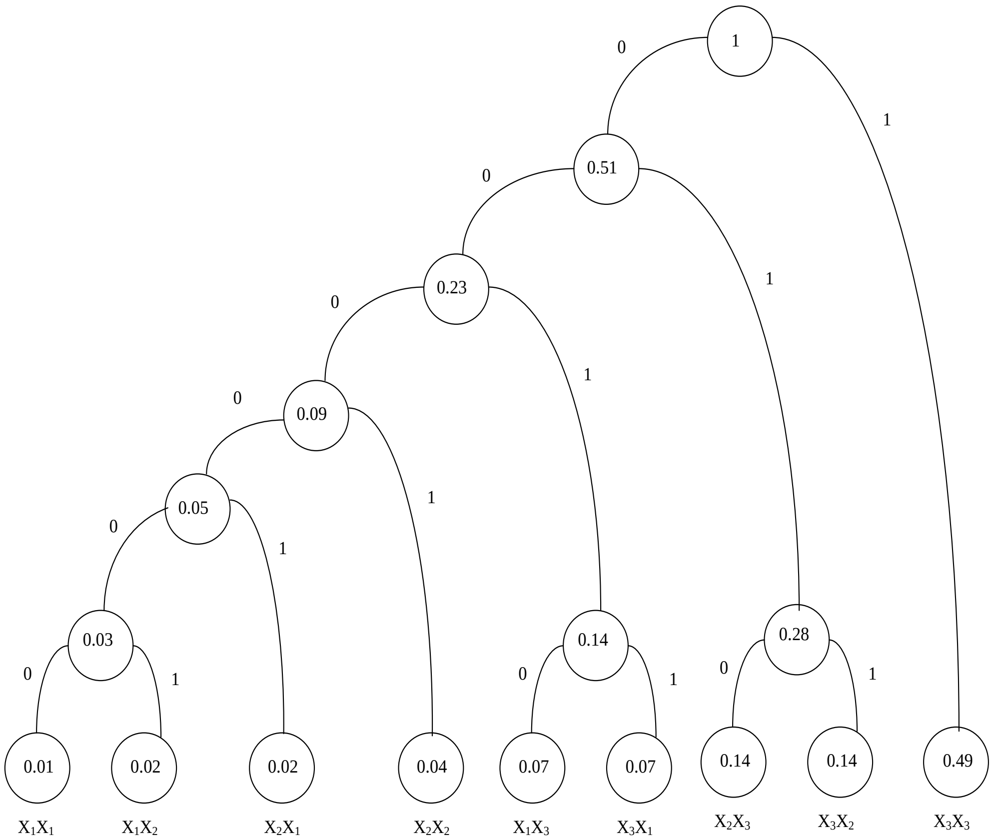
$$\Rightarrow E = \frac{H(X)}{\bar{L}} = \frac{1.157}{1.3} = 0.89$$

$$= 89\%$$

2- Code de Huffman de l'extension d'ordre 2 de la source X:

m=3, K=2 $\Rightarrow m^K = 3^2 = 9$ Symboles : $\{X_1X_1, X_1X_2, X_1X_3, X_2X_1, X_2X_2, X_2X_3, X_3X_1, X_3X_2, X_3X_3\}$

Symboles	Probabilités	Codes $C(X_i)$	Longueur L_i
X_1X_1	0.01	000000	6
X_1X_2	0.02	000001	6
X_1X_3	0.07	0010	4
X_2X_1	0.02	00001	5
X_2X_2	0.04	0001	4
X_2X_3	0.14	010	3
X_3X_1	0.07	0011	4
X_3X_2	0.14	011	3
X_3X_3	0.49	1	1



Calcul de L'entropie:

$$H\left(X^2\right) = 2 * H(X)$$

$$= 2.314 \ Bits / symboles$$

Calcul de la longueur moyenne:

$$\overline{L}_2 = 6*[0.01+0.02]+5*0.02+4*[0.07+0.04+0.07]+3*[0.14+0.14]+1*0.49$$

$$\Rightarrow \overline{\mathrm{L}}_2 = 2.33\, Bits / symboles$$

Calcul de l'efficacité E_2:

$$\Rightarrow E_2 = \frac{H(X^2)}{\overline{L_2}} = \frac{1.314}{2.33} = 0.993$$

$$= 99.3\%$$

Les résultats obtenus montrent que le codage selon l'extension de la source permet d'améliorer l'efficacité.