Partie 2, Chapitre 2 : Le Codage de Source

MODULE : TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION

W. BAHRI[1] A. GUEDDANA[1]

[1]Département Informatique
Université de Carthage, École Nationale d'Ingénieurs de Carthage - Enicarthage

Année Universitaire 2019-2020

Plan

Introduction

Dans ce chapitre, nous donnons un bref aperçu sur la théorie de l'information dans le but de l'appliquer au codage de source (compression).

Pour cela, on commence par mesurer la quantité d'information qu'apporte l'observation d'une source aléatoire (Entropie de la source). Après cela, nous établissons le théorème de Shannon pour le codage de source qui donne la borne inférieure du taux de compression. Ensuite, on établit une méthode basée sur la méthode des intervalles pour construire des codes instantannés.

Enfin, on détaille les étapes nécessaires pour la réalisation de l'algorithme de Huffman utilisé pour le codage source optimal.

Quantité d'information (1/2)

Il s'agit de caractériser la quantité d'information moyenne apportée par l'observation d'une source aléatoire. On commence par considérer le cas d'une source discrète X.

La source X est modélisée par une variable aléatoire (v.a.) discrète ayant comme alphabet $A = \{x_1, x_2, \cdots, x_M\}$. On note $p_i = P(X = x_i)$.

La quantité d'information apportée par la réalisation de $X = x_i$ est donnée par

$$Q(X = x_i) = \log_2(\frac{1}{p_i}). \quad (1)$$

Quantité d'information (2/2)

On vérifie que la quantité d'information apportée par la réalisation de l'évènement $X = x_i$ est inversement proportionnelle à sa probabilité de réalisation p_i.

En outre, si on considère deux sources X et Y indépendantes alors la quantité d'information apportée par la réalisation de $X = x_i$ et $Y = y_j$ est la somme des quantités d'information apportées par la réalisation de $X = x_i$ et $Y = y_j$:

$$\begin{aligned} Q\left(X = x_i, Y = y_j\right) &= \log_2\left(\frac{1}{P\left(X = x_i, Y = y_j\right)}\right) \\ &= \log_2\left(\frac{1}{P\left(X = x_i\right) P\left(Y = y_j\right)}\right) \quad (2) \\ &= Q\left(X = x_i\right) + Q\left(Y = y_j\right). \quad (3) \end{aligned}$$

Entropie d'une Source (1/2)

L'entropie d'une source est la quantité d'information moyenne apportée par l'observation de la source :

$$H(X) = \sum_{i=1}^{M} p_i \log_2 \left(\frac{1}{p_i} \right). \tag{4}$$

L'entropie d'une source est grande lorsque son observation apporte beaucoup d'information, on peut donc dire que l'entropie donne l'incertitude sur la source. Cette dernière définition peut être étendue au cas de sources à valeurs réelles :

$$H(X) = \int_x p_X(x) \log_2 \left(\frac{1}{p_X(x)} \right) dx. \tag{5}$$

Exemple : Entropie d'une source discrète de cardinal $M = 2$

$$H(X) = p \log_2 \left(\frac{1}{p} \right) + (1-p) \log_2 \left(\frac{1}{1-p} \right), \tag{6}$$

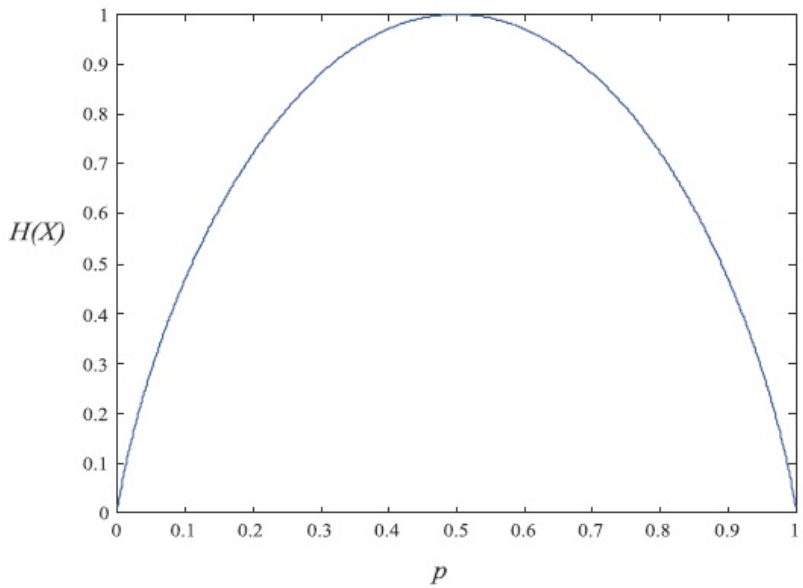
où $p = P(X = x_1)$ et $1 - p = P(X = x_2)$

Entropie d'une Source (2/2)

La figure suivante montre l'évolution de $H(X)$ en fonction de p. On constate que l'entropie est nulle pour $p = 1$ et $p = 0$ c'est à dire lorsque la source est déterministe. L'entropie est maximale pour $p = 1/2$ c'est à dire lorsque la source est uniforme. En général, on montre que :

$$0 \leq H(X) \leq \log_2(M). \tag{7}$$

L'entropie est bien sûr nulle lorsque la source est déterministe et elle est maximale valant $\log_2(M)$ lorsque la source est uniforme.



Théorème de Shannon (1/3)

Le codage de source ou encore compression a pour but de réduire le nombre de bits utilisés pour représenter une source binaire. Cette technique de codage porte aussi le nom de codage entropique car elle utilise des statistiques de la source ou plus précisément la probabilité d'occurrence de ces différents symboles. Nous nous intéressons dans cette section à établir une borne inférieure du taux de compression.

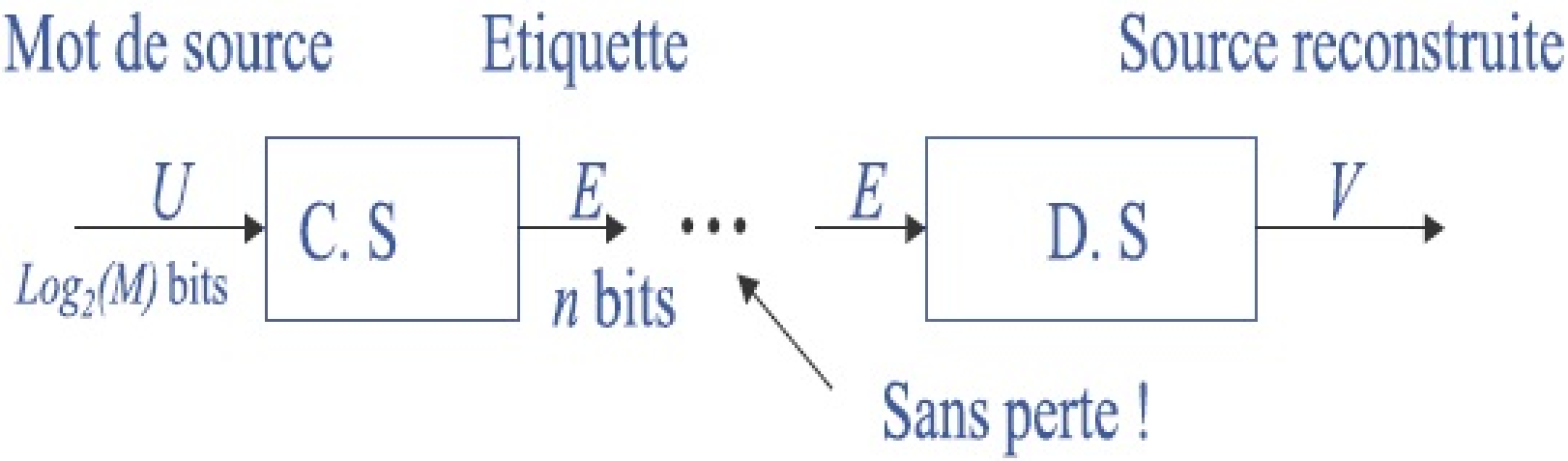
On considère une source U de cardinal M, habituellement chaque symbole U_i de cette source doit être représenté sur $log_2(M)$ bits. Le codeur source consiste à associer à chaque symbole de la source U_i une étiquette E_i formée de n bits. Le **taux de codage de source** est défini comme étant le rapport du nombre de bits des étiquettes E_i par celui des mots de source U_i :

$$R_s = \frac{n}{log_2(M)} = \frac{\sum_{i=1}^{M} p_i l_i}{\log_2(M)} \tag{8}$$

où $p_i = p(E = E_i) = p(U = U_i)$ et l_i est la longueur en nombre de bits de E_i.
Bien évidemment, plus R_s est faible, plus la compression est forte.

Théorème de Shannon (2/3)

Lorsqu'on s'intéresse à l'étude et optimisation du codage de source, on suppose que le canal est non bruité, c'est à dire que la transmission des données ou leur stockage se fait sans aucune erreur. Ainsi, le décodeur source a pour entrée la source aléatoire E et on note V sa sortie.



Théorème de Shannon (3/3)

A présent, nous allons établir la borne inférieure du taux de compression dans le cas où le codeur source **n'introduit pas de perte**. Du moment que le codeur et le décodeur de source n'introduisent pas de perte $U = V$, la **limite de Shannon** est donnée par :

$$R_s \geq \frac{H(U)}{log_2(M)}. \tag{9}$$

Ainsi l'entropie de la source apparait comme la borne inférieure du taux de compression. Plus la source présente des symboles très probables plus son entropie est faible et plus on pourra la comprimer. Le cas limite est celui d'une source déterministe dont l'entropie est nulle, dans ce cas on peut faire tendre le taux de compression vers zéro. L'autre cas limite est celui d'une source uniforme dont l'entropie est égale à $\log_2(M)$ donc le taux de compression est égal à 1 c'est à dire qu'on ne peut pas la comprimer.

Caractéristiques des Codes (1/2)

Le codeur source consiste à associer à chaque symbole de la source U_i une étiquette E_i **de sorte que le taux de compression soit minimal**. Étant donné que les étiquettes peuvent avoir des longueurs variables, on peut avoir une ambigüité lors du décodage si la concaténation de certaines étiquettes peut être interprétée de différentes façons. Par exemple, si on utilise les quatre étiquettes suivantes
$E = \{E_1 = 0,\ E_2 = 10,\ E_3 = 100,\ E_4 = 101\}$, on voit que la concaténation de E_2 et E_1 donne E_3 donc le décodeur source est incapable de faire le décodage. Un codeur ou encore un code non ambigu est aussi dit **uniquement déchiffrable** (u.d.).

Définition : Un code est u.d si toute concaténation d'étiquettes ne peut être interprétée que d'une seule façon :

$$\forall p,\ \forall k,\ \forall i_l,\ \forall\, j_m,\quad E_{i_1} \cdots E_{i_k} = E_{j_1} \cdots E_{j_p} \ \Rightarrow p = k \text{ et } E_{i_n} = E_{j_n}\ \forall n = 1, \cdots, p. \quad (10)$$

C'est cette classe de codes qu'on doit utiliser. Utiliser un code u.d. peut entrainer un retard lors du décodage du moment qu'on doit attendre la réception de plusieurs étiquettes avant de décider. Ceci augmente aussi la complexité du décodeur. Pour cela, on opte pour une sous classe des codes u.d. à savoir les codes instantanés.

Caractéristiques des Codes (2/2)

Propriété : Un code est instantané s'il vérifie la condition du préfixe : aucune étiquette ne doit être le début d'une autre. Un code instantané n'entraine pas de retards lors du décodage du moment qu'aucune étiquette n'est le début d'une autre.

Exemple :

- $E = \{E_1 = 0,\ E_2 = 10,\ E_3 = 100,\ E_4 = 101\}$ est ambigu.
- $E = \{E_1 = 10,\ E_2 = 00,\ E_3 = 11,\ E_4 = 110\}$ est u.d.
- $E = \{E_1 = 0,\ E_2 = 10,\ E_3 = 110,\ E_4 = 111\}$ est instantané.

Théorème de Mac Millan et Théorème de Kraft

Théorème

Le théorème de Mac Millan

Ce théorème donne une condition nécessaire que doivent vérifier les longueurs l_i des étiquettes E_i pour que le codeur soit u.d.

Un code est u.d. alors $\sum_{i=1}^{M} 2^{-l_i} \leq 1$ où M est le cardinal de la source.

Théorème

Le théorème de Kraft

Si $\sum_{i=1}^{M} 2^{-l_i} \leq 1$ alors on peut construire un code instantané dont les étiquettes ont pour longueur $\{l_i\}_i$.

Construction des Codes Instantanés (1/2)

On peut utiliser la méthode des intervalles pour construire des codes instantanés. On note E_i^j le j-ème bit de l'étiquette E_i :

$$E_i = E_i^1 E_i^2 \cdots E_i^{l_i}. \tag{11}$$

On associe à E_i le réel suivant

$$\overline{E}_i = E_i^1 2^{-1} + E_i^2 2^{-2} + \cdots E_i^{l_i} 2^{-l_i}. \tag{12}$$

Les réels associés aux étiquettes qui commencent par E_i appartiennent à l'intervalle débutant par $\overline{E}_i$ et se terminant à

$$\overline{E}_i + \sum_{j=l_i+1}^{+\infty} 2^{-j} = \overline{E}_i + 2^{-l_i}. \tag{13}$$

Ainsi, un codeur est instantané ssi les intervalles suivants ne se recouvrent pas :

$$I_i = [\overline{E}_i \ \overline{E}_i + 2^{-l_i}[\ \forall i = 1, \cdots, M. \tag{14}$$

Construction des Codes Instantanés (2/2)

La construction d'un codeur instantané n'est possible que si les longueur l_i vérifient le théorème de Kraft. Ensuite, il suffit de suivre les étapes suivantes :
1) On place d'abord $\overline{E}_1 = 0$.
2) i=1.
3) On construit le i-ème intervalle : I_i.
4) On déduit $\overline{E}_{i+1} = \overline{E}_i + 2^{-l_i}$.
5) $i = i+1$ puis revenir à 3) tant que $i < M$.

Bien évidemment, on déduit facilement les étiquettes E_i à partir des $\overline{E}_i$ puisqu'on connait les l_i.

Pour les longueurs suivantes $l = \{1,2,3,3\}$, elles vérifient bien le théorème de Kraft : $\sum_{i=1}^{M} 2^{-l_i} = 1$. Suite à l'utilisation de la méthode des intervalles, les étiquettes obtenues sont $E = \{E_1 = 0, E_2 = 10, E_3 = 110, E_4 = 111\}$. Pour passer d'une étiquette à la suivante, il suffit de rajouter 1 en dernière position puis de compléter éventuellement par des zéros afin d'avoir la bonne longueur.

Exercice : Construire un code instantané ayant pour longueur $l = \{1,3,5,5,5,6,7,7,7\}$.

Conditions Nécessaires sur les Longueurs Optimales

On cherche à construire un codeur source optimal c'est à dire dont les étiquettes possèdent des longueurs minimisant le taux de compression :

$$R_s = \frac{\sum_{i=1}^{M} p_i l_i}{\log_2(M)}. \tag{15}$$

où $p_i = p(E = E_i) = p(U = U_i)$ et l_i est la longueur en nombre de bits de E_i.

Bien évidemment, il faut que le codeur soit instantané pour faciliter le décodage. Ainsi, les l_i doivent aussi vérifier le théorème de Kraft :

$$\sum_{i=1}^{M} 2^{-l_i} \leq 1. \tag{16}$$

Puisqu'on s'intéresse à la construction de codeurs entropiques sans pertes, la borne inférieure du taux de compression est donnée par

$$R_s = \frac{\sum_{i=1}^{M} p_i l_i}{\log_2(M)} \geq \frac{H(U)}{\log_2(M)} = \frac{\sum_{i=1}^{M} p_i \log_2\left(\frac{1}{p_i}\right)}{\log_2(M)}. \tag{17}$$

Pour minimiser le taux de compression, il suffit d'associer aux symboles de source les plus probables les plus courtes étiquettes.

Les Longueurs Optimales (1/2)

D'après (17), si p_i est l'inverse d'une puissance de 2, on pourra atteindre la limite de Shannon en prenant $l_i = \log_2\left(\frac{1}{p_i}\right)$. On déduit ensuite les étiquettes grâce à la méthode des intervalles.

Exemple 1 : On considère une source de cardinal 4 ayant pour distribution de probabilité $p = \{p_1 = 1/2,\ p_2 = 1/4,\ p_3 = p_4 = 1/8\}$. On en déduit les longueurs optimales $l = \{1,\ 2,\ 3,\ 3\}$. Puis grâce à la méthode des intervalles les étiquettes $E = \{0,\ 10,\ 110,\ 111\}$. Dans ce cas, on vérifie qu'on a bien atteint la limite de Shannon : $R_s = H/2 = 7/8$.

Lorsque p_i n'est pas l'inverse d'une puissance de 2, on peut songer à choisir l_i comme étant l'entier immédiatement supérieur à $\log_2\left(\frac{1}{p_i}\right)$: $l_i = \left\lceil \log_2\left(\frac{1}{p_i}\right)\right\rceil$. Le jeux de longueurs qu'on trouve peut ne pas être l'optimal.

Les Longueurs Optimales (2/2)

Exemple 2 : On considère une source de cardinal 5 ayant pour distribution de probabilité $p = \{p_1 = 1/4,\ p_2 = 1/4,\ p_3 = 0.2,\ p_4 = 0.15,\ p_5 = 0.15\}$. Les longueurs déduites de la relation $l_i = \left\lceil \log_2\left(\frac{1}{p_i}\right)\right\rceil$ sont égales à $l = \{2,\ 2,\ 3,\ 3,3\}$. Le taux de compression vaut alors $R_s = 2.5/\log_2(5)$. Or, la limite de Shannon vaut $H/\log_2(M) = 2.2855/\log_2(5)$. On constate que le jeux de longueurs n'est pas optimal.

En effet, le meilleur choix est $l = \{2,\ 2,\ 2,\ 3,3\}$ dont le taux de compression est le plus proche de la borne de Shannon : $R_s = 2.3/\log_2(5)$.

Proposition 4 : A l'optimum, le taux de compression est borné par

$$\frac{H(U)}{\log_2(M)} \leq R_s < \frac{H(U)}{\log_2(M)} + \frac{1}{\log_2(M)} \tag{18}$$

Algorithme de Huffman (1/2)

L'algorithme de Huffman a été inventé en 1952, il permet la construction d'un codeur source optimal et instantané. Il est basé sur la réduction de Huffman qui permet de passer d'un problème P d'ordre M dont les probabilités sont classées par ordre décroissant : $p_1 \geq p_2 \geq \cdots \geq p_M$ au problème P' d'ordre $M-1$ suivant

$$p'_1 = p_1,\ p'_2 = p_2, \cdots,\ p'_{M-2} = p_{M-2},\ p'_{M-1} = p_{M-1} + p_M. \tag{19}$$

Proposition 5 : Les étiquettes E' sont optimales pour le problème $P' \Leftrightarrow$ Les étiquettes E suivantes :

$$E_i = E'_i,\ \forall i = 1 \cdots M-2,\ E_{M-1} = [E'_{M-1}0] \text{ et } E_M = [E'_{M-1}1]$$

sont optimales pour le problème P.

Algorithme de Huffman (2/2)

L'algorithme de Huffman comporte les étapes suivantes :

1) Classez les probabilités par ordre décroissant : $p_1 \geq p_2 \geq \cdots \geq p_M$

2) Faire plusieurs réductions de Huffman jusqu'au problème d'ordre 2. Après chaque réduction, il faut **classer de nouveau** les probabilités par ordre décroissant.

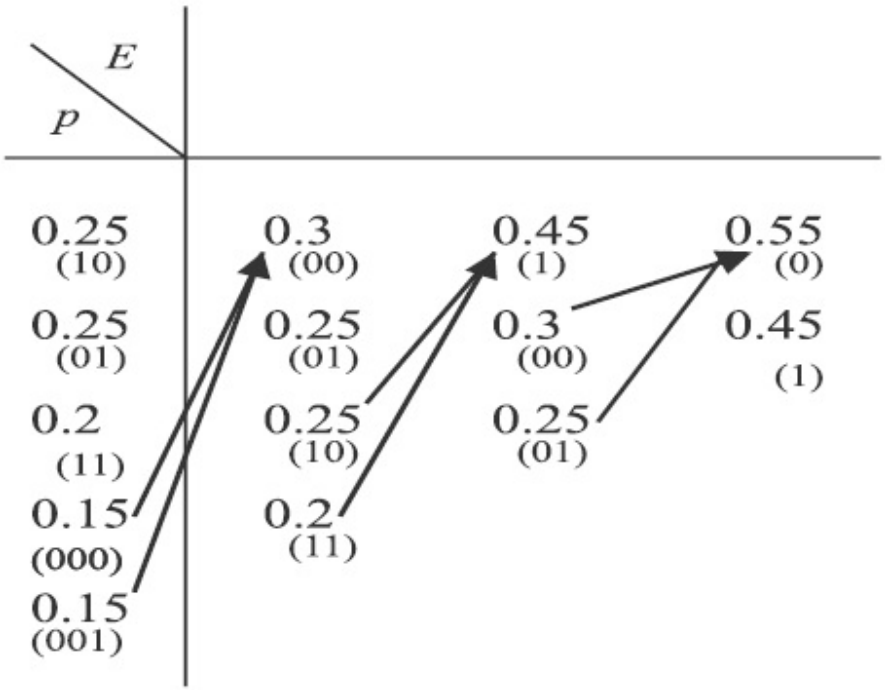
3) Quand on arrive au problème d'ordre 2, on attribue par exemple les étiquettes 0 et 1 respectivement au symbole le plus probable et au symbole le moins probable.

Ensuite, on déduit les étiquettes du problème d'ordre 3 grâce à la proposition 2.5 et ainsi de suite jusqu'à aboutir au problème initial :

- si E'_{M-1} est l'étiquette du symbole de probabilité $p'_{M-1} = p_{M-1} + p_M$ alors $E_{M-1} = [E'_{M-1}\ 0]$ et $E_M = [E'_{M-1}\ 1]$ sont respectivement les étiquettes des symboles de probabilité p_{M-1} et p_M

Exemple d'Application

On considère une source de cardinal 5 ayant pour distribution de probabilité $p = \{p_1 = 1/4,\ p_2 = 1/4,\ p_3 = 0.2,\ p_4 = 0.15,\ p_5 = 0.15\}$. En utilisant l'algorithme de Huffman, on obtient les étiquettes $E = \{10, 01, 11, 000, 001\}$. On vérifie aisément que ce codeur est instantané et qu'il est optimal : $R_s = 2.3/\log_2(5)$.



Question : vérifier le théorème de shannon pour le codage de source.

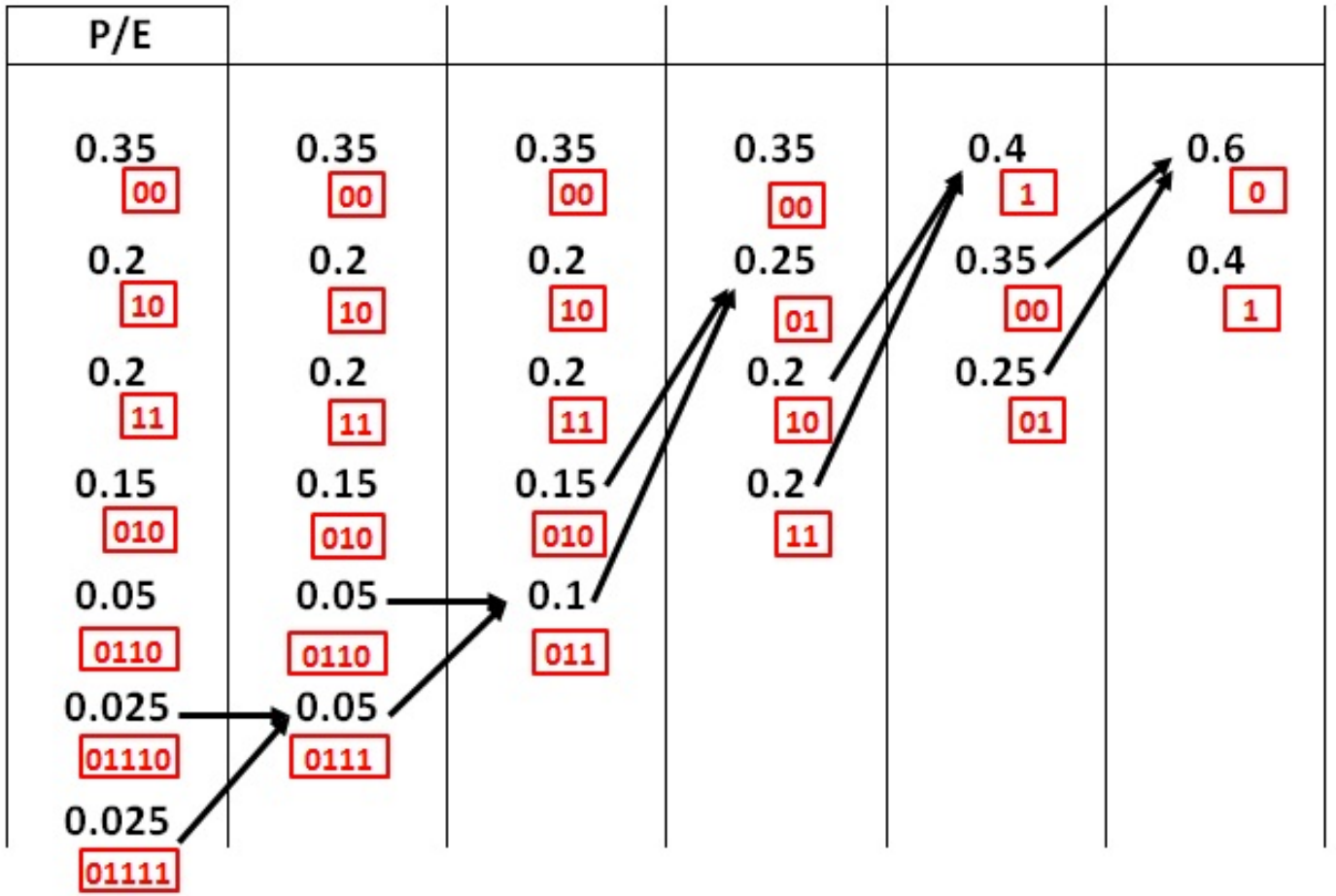
Exercice

On considère une source de cardinal 7 ayant la distribution de probabilité suivante :

$$p = \{0.35,\ 0.2,\ 0.2,\ 0.15,\ 0.05,\ 0.025,\ 0.025\}.$$

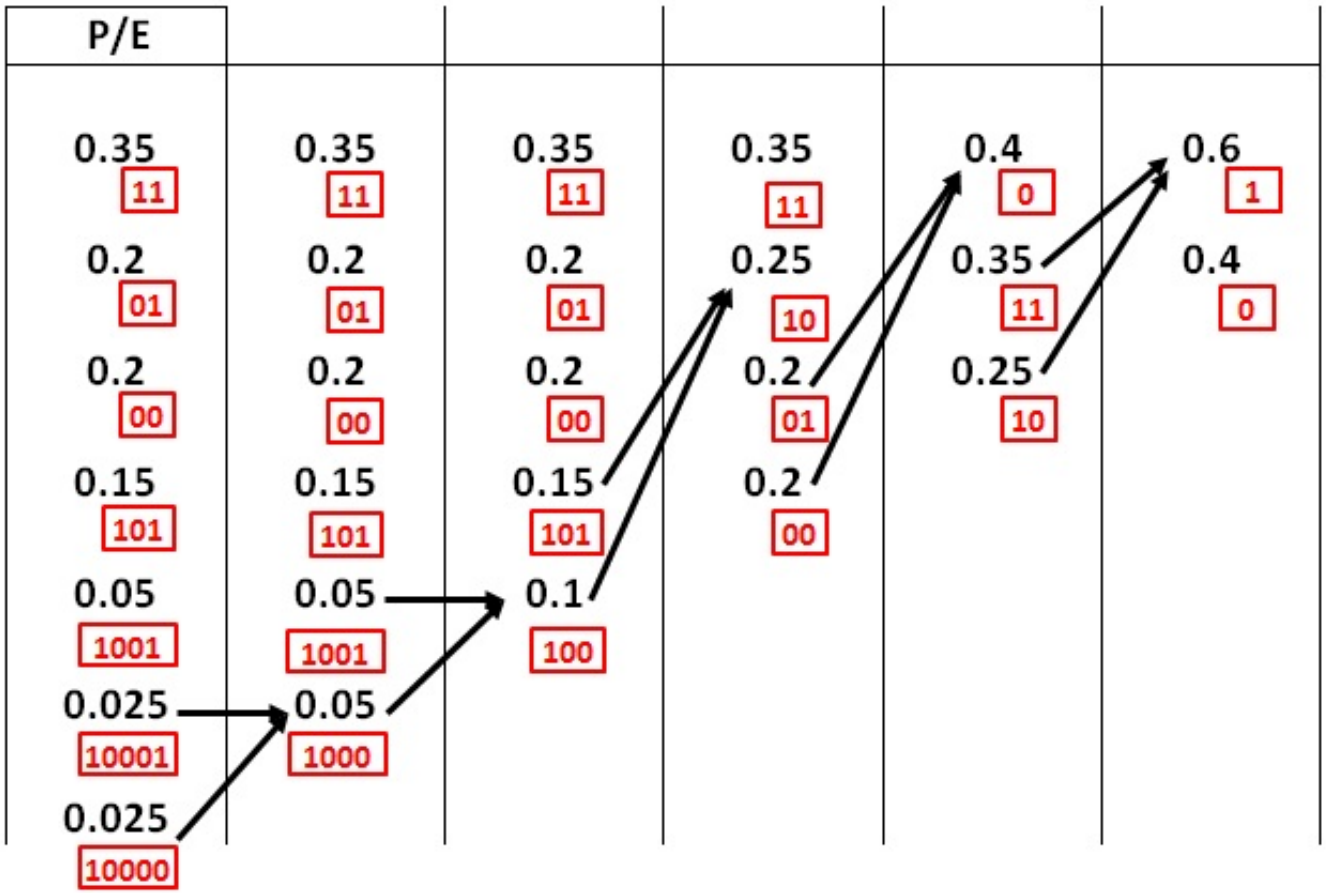
1- En utilisant l'algorithme de Huffman, déterminer les étiquettes E
2- Est ce que ce codeur est optimal ?
3- Vérifier si le théorème de shannon pour le codage de source est vérifié ou non.

Solution 1



$$E = \{00, 10, 11, 010, 0110, 01110, 01111\}$$

Solution 2



$$E = \{11, 01, 00, 101, 1001, 10001, 10000\}$$

Solution

Selon les étiquettes obtenues, on a :

$$E = \{E_i\}_{1 \leq i \leq 7} = \{11, 01, 00, 101, 1001, 10001, 10000\}$$

$$l = \{l_i\}_{1 \leq i \leq 7} = \{2, 2, 2, 3, 4, 5, 5\}$$

$$p = \{p_i\}_{1 \leq i \leq 7} = \{0.35, 0.2, 0.2, 0.15, 0.05, 0.025, 0.025\}$$

Après calcul, on a :

$$\frac{H(U)}{\log_2(M)} = 0.538$$

$$R_s = \frac{\sum_{i=1}^{M} p_i l_i}{\log_2(M)} = 0.857$$

$$\left(\frac{H(U)}{\log_2(M)} + \frac{1}{\log_2(M)}\right) = 0.895$$

Le taux de compression obtenu $R_s = 0.857$ est optimal parceque :

$$\frac{H(U)}{\log_2(M)} \leq R_s < \left(\frac{H(U)}{\log_2(M)} + \frac{1}{\log_2(M)}\right)$$