Aix-Marseille Université
Théorie de l'information

Master FSI	**Exercices pour le DM 1**	2021–2022

Entropie et codes de compression

1. **a.** Déterminer si le codage suivant est *i*) sans pertes, *ii*) instantané, et *iii*) uniquement décodable (avec justification) :

x	A	B	C	D
$C(x)$	000	001	01	1

b. Calculer l'entropie de $\mathcal{X} = \{\mathtt{A}, \mathtt{B}, \mathtt{C}, \mathtt{D}\}$ avec fonction de probabilité

$$p(\mathtt{A}) = p(\mathtt{B}) = 0{,}125,\ p(\mathtt{C}) = 0{,}250,\ \text{et } p(\mathtt{D}) = 0{,}500,$$

et déterminer si C est optimal.

2. Soit $\mathcal{X} = \{\mathtt{A}, \mathtt{B}, \mathtt{C}, \mathtt{D}, \mathtt{E}, \mathtt{F}, \mathtt{G}, \mathtt{H}\}$ l'espace de probabilité avec les probabilités suivantes :

x	A	B	C	D	E	F	G	H
$p(x)$	0,25	0,25	0,20	0,14	0,10	0,04	0,01	0,01

a. Calculer l'entropie $H(\mathcal{X})$.

b. Construire l'arbre de codage associé à un codage de Huffman C pour X, et donner les codes $C(x)$ pour chaque élément x.

c. Trouver l'espérance mathématique de la fonction longueur de $C(x)$, et vérifier le théorème *Codage source* de Shannon.

3. Soit $\mathcal{A} = \{\mathtt{A}, \mathtt{B}, \mathtt{C}, \mathtt{D}, \mathtt{E}\}$ un alphabet avec probabilités

$$p(\mathtt{A}) = 0{,}40,\ p(\mathtt{B}) = 0{,}1,\ p(\mathtt{C}) = p(\mathtt{D}) = 0{,}20,\ \text{et } p(\mathtt{E}) = 0{,}1,$$

et $m = \mathtt{AAACDCBDCEAADADACEDAEADCABEDADDCECAAAAAD}$.

a. Trouver un codage instantané $C : \mathcal{A} \longrightarrow \{0,1\}^*$, avec $E(\ell_C) \leq H(\mathcal{A}) + 1$.

b. Comparer la longueur de m multiplié par $\log_2(5)$ et la longueur de $C(m)$.

4. Soit $\mathcal{A} = \{\mathtt{A}, \mathtt{B}, \mathtt{C}\}$, et supposons que les digrammes suivent les probabilités :

	$p(xy)$	A (y)	B (y)	C (y)
	A	0	4/15	1/15
x	B	8/27	8/27	0
	C	1/27	4/135	1/135

a. Calculer les probabilités $p(x)$ pour chaque lettre x dans $\mathcal{A}$.

b. Déterminer $H(\mathcal{A}^2)$, $H(\mathcal{A})$ et vérifier $H(\mathcal{A}^2) \leq 2H(\mathcal{A})$.

c. Trouver des codages $C_1 : \mathcal{A} \longrightarrow \{0,1\}^*$, et $C_2 : \mathcal{A}^2 \longrightarrow \{0,1\}^*$, qui sont uniquement décodable, avec l'espérance des longueurs bornée par $H(\mathcal{A}) + 1$ et $H(\mathcal{A}^2) + 1$.

d. Trouver les longueurs des codes $C_1(m)$ et $C_2(m)$ où

$$m = \texttt{ABBABABABABABABABBBABBBBBABABABABABABBBACACABBABBBBABBABACBBBABA}.$$

5. On considère un texte source formé à partir de 5 symboles distincts $(\texttt{a}, \texttt{b}, \texttt{c}, \texttt{d}, \texttt{r})$ avec les fréquences d'apparition suivantes :

$$p(\texttt{a}) = 0{,}43,\ p(\texttt{b}) = 0{,}20,\ p(\texttt{c}) = 0{,}10,\ p(\texttt{d}) = 0{,}09,\ p(\texttt{r}) = 0{,}18.$$

a. Générer un arbre de Huffman binaire et proposer le codage correspondant en affectant les 0 aux probabilités les plus grandes.

b. Coder le texte suivant et calculer le gain de compression par rapport à un code binaire de longueur fixe et minimale : `abracadabrabracadabra`.

Rendement des codes correcteurs

6. Déterminer le rendement du codage $C : \{\texttt{A}, \texttt{B}, \texttt{C}, \texttt{D}\} \to \{0,1\}^2$, donné par

$$C(\texttt{A}) = 00,\ C(\texttt{B}) = 01,\ C(\texttt{C}) = 10,\ C(\texttt{D}) = 11.$$

7. Soit C le codage $\mathcal{X} = \{\texttt{A}, \texttt{B}, \texttt{C}, \texttt{D}\} \to \{0,1\}^*$, donné par

$$C(\texttt{A}) = 0, C(\texttt{B}) = 10, C(\texttt{C}) = 110, C(\texttt{D}) = 111.$$

a. Remplir le tableau suivant, où $r_n = \dfrac{\log_2(|C(\mathcal{X}^*) \cap \mathcal{A}^n|)}{\log_2 |\mathcal{A}^n|}$.

n	$\lvert C(\mathcal{X}^*) \cap \mathcal{A}^n \rvert$	$\lvert \mathcal{A}^n \rvert$	r_n
1			
2			
3			
4			
5			
6			

b. Quel est le rendement de C ?

8. Trouver le rendement pour les codages suivants : $x_1x_2x_3x_4 \mapsto x_1x_2x_3x_4y_1y_2y_3y_4$ où les bits y_i sont définés par

a.
$$y_1 = x_1,\quad y_2 = x_2,\quad y_3 = x_3,\quad y_4 = x_4$$

b.
$$y_1 = x_1,\quad y_2 = x_1 + x_2,\quad y_3 = x_3,\quad y_4 = x_3 + x_4$$

c.
$$y_1 = x_1 + x_2 + x_3,\quad y_2 = x_1 + x_2 + x_4,\quad y_3 = x_1 + x_3 + x_4,\quad y_4 = x_2 + x_3 + x_4$$

Combien d'erreurs peuvent être détectés et corrigeés par chaque codage ?