TD 1 : Un codage optimal

Un code C pour une variable aléatoire X sur l'alphabet $\mathcal{A}$ est une fonction de $\mathcal{A}$ vers $\mathcal{B}^*$ (l'ensemble des mots finis sur $\mathcal{B}$).

— $C(x)$ est le mot code correspondant à x, et $\ell(x)$ est sa longueur.

— La longueur moyenne $L(C)$ d'un code pour la variable aléatoire X de distribution de probabilité $p(x)$ est :

$$L(C) = \sum_{x \in \mathcal{A}} p(x)\ell(x)$$

Dans la suite, on supposera pour simplifier sans grande perte de généralité : $\mathcal{B} = \{0, 1\}$.

1. Sur un alphabet binaire, si il est injectif :

$$P(X=1) = \frac{1}{2} \qquad C(1) = 0 \qquad\qquad P(X=3) = \frac{1}{8} \qquad C(3) = 110$$
$$P(X=2) = \frac{1}{4} \qquad C(2) = 10 \qquad\qquad P(X=4) = \frac{1}{8} \qquad C(4) = 111$$

   (i) décoder 0110111100110,
   (ii) calculer $H(X)$ et $L(C)$.

Un code C est dit non-ambigu si :

$$x \neq y \Rightarrow C(x) \neq C(y)$$

et instantané si aucun mot du code n'est le préfixe d'un autre mot. Le code C s'étend à $\mathcal{A}^*$ en définissant pour toute suite $x_1, \ldots, x_n$ le code concaténé

$$C(x_1 \ldots x_n) = C(x_1) \ldots C(x_n).$$

On dit que C est uniquement décodable si son extension est non-ambigu.

2. Dans les codes suivants, lesquels sont non-ambigus, lesquels sont instantanés et lesquels sont uniquement décodables.

X	Code 1	Code 2	Code 3
1	0	10	0
2	010	00	10
3	01	11	110
4	10	110	111

Un théorème de MacMillan affirme que les codes instantanés sont aussi performant pour leur taux de compression que les codes uniquement décodables. On cherche donc un code instantané optimal.

Procédons par analyse. Supposons que l'on a un code C tel que $L(C)$ est minimal sur tous les codes non-ambigu instantanés. Posons $\mathcal{A} = \{a_1, \dots, a_k\}$ avec $p(a_1) \geq \cdots \geq p(a_k)$.

3. Montrer que si il existe $1 \leq i, j \leq k$ tel que $p(a_i) < p(a_j)$ et $\ell(a_i) < \ell(a_j)$ alors en définissant un code C' dans lequel on inverse le codage de a_i et a_j on a $L(C') < L(C)$.

On suppose à présent $\ell(a_k) = \max_i \ell(a_i)$.

4. Montrer que $C(a_k)$ ne peut pas être le seul mot de longueur $\ell(a_k)$ et que (quitte à réordonner des caractères de même probabilité) c'est $C(a_{k-1})$ cet autre mot. De plus, montrer que les préfixes de $C(a_k)$ et $C(a_{k-1})$ de longueur $\ell(a_k) - 1$ sont identiques

On a donc réduit le problème de construction d'un code optimal à celui de construire $C(a_1), \dots, C(a_{k-2})$ et trouver les $\ell(a_k) - 1$ premiers chiffres de $C(a_k)$. On définit maintenant l'ensemble réduit $\mathcal{A}' = \{a'_1, \dots, a'_{k-1}\}$ avec la variable aléatoire $\mathcal{A}'$ associée

$$p(a'_j) = \begin{cases} p(a_j) & \text{si } j \leq k-2 \\ p(a'_{k-1}) = p(a_k) + p(a_{k-1}) & \text{si } j = k-1 \end{cases}$$

À un code instantané pour $\mathcal{A}$ pour lesquels $C(a_k)$ et $C(a_{k-1})$ ne diffèrent que par le dernier chiffre, on peut associer un code C' défini par $C'(a_i) = C(a_i)$ pour $0 \leq i \leq k-2$ et $C'(a_{k-1})$ est donné par le préfixe commun à $C(a_k)$ et $C(a_{k-1})$ de longueur $\ell(a_k) - 1$.

5. Montrer C' est aussi instantané et que si il est optimal C l'est aussi et $L(C) = L(C') + p(a'_{k-1})$.
6. En déduire de manière itérative que le codage de la question 1 est optimal.
7. Construire un codage optimal pour la loi suivante

$$P(X=1) = \frac{1}{3} \qquad P(X=3) = \frac{1}{5}$$
$$P(X=2) = \frac{4}{15} \qquad P(X=4) = \frac{1}{5}$$

   puis calculer $H(X)$ et $L(C)$.
8. Proposer un algorithme en pseudocode pour donner un codage optimal.
9. Montrer que pour un codage optimal $L(C) \leq H(X) + 1$.
10. Tirons n caractères $X_1, \dots, X_n$ i.i.d. avec la loi de X. Montrer que $H(X_1, \dots, X_n) = nH(X)$ en déduire que quitte à grouper les caractères, il existe un codage au taux de compression arbitrairement proche de $H(X)$ par caractère.