Corrigé TD 1 : Un codage optimal

1. (i) 0 110 111 10 0 110 = $C(1)\ C(3)\ C(4)\ C(2)\ C(1)\ C(3) \mapsto 134213$
   (ii) La formule de l'entropie $H(X)$ est donnée par $H(X) = -\sum_{x\in\mathcal{A}} p(x)\log_2(p(x))$.

$$H(X) = \frac{1}{2}\log_2 2 + \frac{1}{4}\log_2 4 + \frac{1}{8}\log_2 8 + \frac{1}{8}\log_2 8 = \frac{1}{2} + \frac{1}{2} + \frac{3}{4} = \frac{7}{4}$$

   Ensuite, la longueur moyenne $L(C)$ est :

$$L(C) = \frac{1}{2}\times 1 + \frac{1}{4}\times 2 + \frac{1}{8}\times 3 + \frac{1}{8}\times 3 = \frac{7}{4}.$$

   Donc ce codage est optimal !
2.

Code 1
- Non-ambigu : Oui.
- Instantané : Non, car $C(2) = 010$ a le préfixe $C(1) = 0$.
- Uniquement décodable : Non, par exemple $C(1)C(4) = C(2)$.

Code 2
- Non-ambigu : Oui.
- Instantané : Non, car $C(3)$ est préfixe de $C(4)$.
- Uniquement décodable : Oui. Si w a 0 en préfixe, cela ne peut être que par le codage de 2. Considérons un préfixe w de la forme 1 n fois puis 0. Si $n = 2k$, w est peut uniquement être codé par $k-1$ fois 3 puis 4. Si $n = 2k+1$, par k fois 3 puis 1.

Code 3
- Non-ambigu : Oui.
- Instantané : Oui.
- Uniquement décodable : Oui, il suffit de regarder le nombre de 1 consécutifs.

3. On a :

$$L(C) = p(a_i)\ell(a_i) + p(a_j)\ell(a_j) + \sum_{t\neq i,j} p(a_t)\ell(a_t)$$
$$L(C') = p(a_i)\ell(a_j) + p(a_j)\ell(a_i) + \sum_{t\neq i,j} p(a_t)\ell(a_t)$$

   La différence est :

$$L(C') - L(C) = (p(a_i) - p(a_j))(\ell(a_j) - \ell(a_i)) < 0$$

   Puisque $p(a_i) < p(a_j)$ et $\ell(a_i) < \ell(a_j)$, on a donc $L(C') < L(C)$ et cet échange réduit $L(C)$.
4. Supposons que $C(a_k)$ est le seul mot de longueur $\ell(a_k)$. Quitte à échanger 0 et 1, on a $C(a_k) = w\cdot 0$ avec $|w| = \ell(a_k) - 1$. On définit alors le codage C' dans lequel tous les caractères de $\mathcal{A}$ sont codés de façon identique sauf pour $C'(a_k) = w$. Montrons que c'est un codage instantané, ce qui mène a une contradiction car sa longueur moyenne

est strictement inférieure à $L(C)$. Cela signifie qu'il existe un autre mot dans le codage, disons $w \cdot 0$, avec $|w \cdot 0| \leq \ell(a_k) - 1$.
D'un part, aucun mots $C(a_i)$ avec $i \leq k-1$ ne peut être préfixe de w car sinon il serait préfixe $w \cdot 0$. Ce qui contredirait l'instantanéité de C. D'autre part, si w était préfixe d'un mot $C(a_i)$, nécessairement, comme ce mot est de longueur inférieure ou égale à $\ell(a_k) - 1$ par hypothèse, alors $C(a_i) = w$. Donc on aurait $C(a_i)$ préfixe de $C(a_k)$ ce qui n'est pas possible.
Dans tous les cas, on arrive à une contradiction. Donc il existe un autre caractère a_i dont le codage a la même longueur que $C(a_k)$. Par la question précédente, si $p(a_i) < p(a_j)$ on a $\ell(a_j) \geq \ell(a_i)$ donc $\ell(a_j) = \ell(a_k)$.

Par ailleurs, si w et w' les préfixes de longueur $\ell(a_k)-1$ de $C(a_k)$ et $C(a_i)$ étaient différents pour tous i tel que $\ell(a_i) = \ell(a_k)$, le code C' défini par $C'(a_k) = w$ et $C'(a_{k-1}) = w'$ serait instantané par les même arguments. Ce qui prouve la dernière partie de l'énoncé.

5. Pour montrer que C' est optimal, il suffit d'observer que si $C'(a_i)$ est préfixe de $C'(a_{k-1})$ pour un $1 \leq i \leq k-2$, il est aussi préfixe de $C(a_k)$. Par ailleurs, $C'(a_{k-1})$ a longueur maximale donc si il était préfixe d'un autre codage, les deux codages seraient égaux, ce qui n'est pas possible.

   La relation sur les taux de compression
   $$L(C) = L(C') + p(a'_{k-1})$$
   découle du fait que pour les caractères a_{k-1} et a_k, dont la probabilité conjointe est $p(a'_{k-1})$, le codage $L(C)$ a une lettre de plus que pour a'_{k-1}. Cela induit que si C' est optimal C l'est aussi.

6. $$P(X=1) = \frac{1}{3}, \quad P(X=2) = \frac{4}{15}, \quad P(X=3) = \frac{1}{5}, \quad P(X=4) = \frac{1}{5}$$
   En utilisant l'algorithme de Huffman, on obtient le code suivant :
   $$C(1) = 0, \quad C(2) = 10, \quad C(3) = 110, \quad C(4) = 111$$
   Calculons $H(X)$ et $L(C)$:
   $$H(X) = -\left(\frac{1}{3}\log_2\frac{1}{3} + \frac{4}{15}\log_2\frac{4}{15} + 2 \times \frac{1}{5}\log_2\frac{1}{5}\right)$$
   $$H(X) \approx 1.846$$
   $$L(C) = \frac{1}{3} \times 1 + \frac{4}{15} \times 2 + 2 \times \frac{1}{5} \times 3 = 1.867$$

7. Pour n variables i.i.d., on a :
   $$H(X_1, \ldots, X_n) = nH(X)$$
   et $L(C_n) < nH(X) + 1$. Cela montre qu'en groupant les caractères, il est possible de s'approcher arbitrairement de $H(X)$ en termes de taux de compression.