Université Paris 8 Année 2020–2021

Master Mathématiques et applications

Théorie de l'information – Interrogation 1 – Solutions

sujet du 23 octobre 2020
corrigé du 5 février 2021

Exercice 1 – Entropie et code de Shannon-Fano.

Soit $m \geq 1$ et $M = 2^m$. On considère un alphabet $\mathcal{X} = \{x_1, \ldots, x_M\}$.

Question 1. Donner une majoration de l'entropie d'une variable aléatoire à valeurs dans $\mathcal{X}$. Quel type de variable atteint cette borne ?

Solution : Pour toute variable aléatoire X sur $\mathcal{X}$ on a $H(X) \leq \log_2(|\mathcal{X}|) = m$. La borne est atteinte si X suit une loi uniforme.

Soit X une source sur l'alphabet $\mathcal{X}$. On suppose que X est uniforme.

Question 2. Donner la forme de l'arbre binaire du code de Shannon-Fano associé à la source X.

Solution : Les longueurs des mots du code sont toutes égales à

$$\left\lceil -\log_2\left(\frac{1}{|\mathcal{X}|}\right)\right\rceil = m\,.$$

On a donc un arbre parfait : toutes les feuilles sont au même niveau, celui de profondeur m.

Question 3. Calculer la longueur moyenne de ce code. Est-il optimal ?

Solution : Comme tous les mots ont longueur m, le code a longueur moyenne m. Il est optimal car $H(X) = m$ est une borne inférieure sur la longueur moyenne d'un code sur X.

Exercice 2 – Information mutuelle.

Lors du tirage du loto, des boules numérotées de 1 à 49 sont tirées **successivement**, **uniformément** et **sans remise**.

On note X_1 la variable aléatoire correspondant à la valeur de la boule obtenue au premier tirage, et X_2 celle obtenue au second tirage.

Question 1. Calculer $H(X_1)$ et $H(X_2)$.

Solution : $H(X_1) = H(X_2) = \log_2(49) = 2\log_2(7)$ car ce sont des variables uniformes sur $\mathcal{X} = \{1, \ldots, 49\}$.

Question 2. Calculer l'information mutuelle $I(X_1\,; X_2)$.

Solution : On a $I(X_1; X_2) = H(X_1) + H(X_2) - H(X_1X_2)$. Par ailleurs, la variable X_1X_2 est uniforme sur l'ensemble

$$\{(x_1, x_2) \in \{1, \ldots, 49\} \mid x_1 \neq x_2\}$$

qui est de taille 49×48. On a donc :

$$I(X_1; X_2) = 2\log_2(49) - \log_2(49 \times 48) = \log_2(49) - \log_2(48) = 2\log_2(7) - \log_2(3) - 4\,.$$

Exercice 3 – Codes uniquement décodables.

Pour les codes binaires suivants, déterminer **en justifiant** s'ils sont uniquement décodables.

Question 1. Le code dont les mots sont :

$$\{\,0,\ 001,\ 10,\ 101,\ 11111\,\}.$$

Solution : Le code n'est pas uniquement décodable car ses longueurs ne vérifient pas l'inégalité de Kraft :

$$\frac{1}{2} + \frac{1}{8} + \frac{1}{4} + \frac{1}{8} + \frac{1}{32} = \frac{33}{32} > 1\,.$$

Question 2. Le code dont les mots sont :

$$\{\,01,\ 10,\ 111,\ 1101\,\}.$$

Solution : Le code est uniquement décodable car ses mots vérifient la condition du préfixe.

Question 3. Le code B défini sur l'ensemble $\mathbb{N}$ des entiers naturels, tel que le mot associé à l'entier $n = \sum_{i=0}^{k} n_i 2^i$ est :

$$B(n) = (n_0, n_1, \ldots, n_k),$$

où $k = \lfloor \log_2(n) \rfloor$ si $n \geq 1$ et $k = 0$ sinon.

Solution : Le code n'est pas uniquement décodable car le codage associé n'est pas injectif. Par exemple, 101 est le mot associé aux messages $(2, 1)$ et $(1, 0, 1)$.

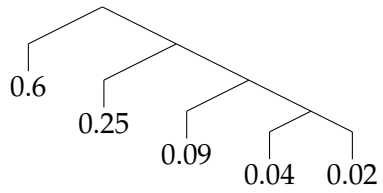
Exercice 4 – Codes de Huffman.

Question 1. On considère une source suivant une loi dont la distribution est :

$$(0.6, 0.25, 0.09, 0.04, 0.02)\,.$$

Donner le codage de Huffman correspondant.

Solution : On obtient l'arbre suivant :



Soit maintenant X une source de distribution $p_1 \geq \cdots \geq p_m$. On suppose que

$$p_i \leq \frac{1}{2}\, p_{i-1}$$

pour tout $i \in \{2, \ldots, m\}$.

Question 2. Soit $i \geq 2$. Comparer les valeurs de p_{i-1}, p_i et $\sum_{j=i+1}^{m} p_j$.

Solution : Pour $j \geq i + 1$, on a les inégalités suivantes :

$$p_j \leq \frac{1}{2} p_{j-1} \leq \frac{1}{2^2} p_{j-2} \leq \cdots \leq \frac{1}{2^{j-i}} p_i\,.$$

Par conséquent,

$$\sum_{j=i+1}^{m} p_j \leq \sum_{j=i+1}^{m} \frac{1}{2^{j-i}} p_i = p_i \sum_{j=1}^{m-i} \frac{1}{2^j} < p_i\,.$$

Donc on obtient $p_{i-1} > p_i > \sum_{j=i+1}^{m} p_j$.

Question 3. À quelle étape de l'algorithme d'Huffman la probabilité p_i sera-t'elle sélectionnée pour la construction de l'arbre ? Justifier.

Solution : À chaque appel récursif, l'algorithme de Huffman sélectionne les probabilités les plus faibles de la distribution passée en paramètre. Démontrons par récurrence sur $k = m - i$, que pour $k \geq 1$, la probabilité p_i est sélectionnée à la k-ème étape.

— $[k = m - i = 1]$ À la première étape, les probabilités p_{m-1} et p_m sont sélectionnées, et remplacées par la probabilité $p_m + p_{m-1}$ dans l'appel récursif suivant.

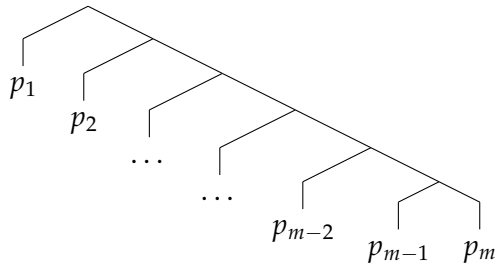
— $[k \to k+1]$ Soit $k \leq m - i - 1$ et supposons le résultat vrai pour tout $k' \leq k$. Alors, à l'étape $k+1$, on passe en paramètre de l'appel récursif la distribution

$$\left(p_1, \ldots, p_{k-1}, \sum_{j=k}^{m} p_j \right)$$

puisqu'aucun des $\{p_j\}_{j \leq k-1}$ n'a été sélectionné lors des k premières étapes (par hypothèse de récurrence). D'après la Question 2., on sait également que $p_1 > p_2 > \cdots > p_{k-1} > \sum_{j=k}^{m} p_j$. Ainsi, p_{k-1} et $p'_k = \sum_{j=k}^{m} p_j$ sont les probabilités les plus faibles de la distribution, donc sont sélectionnées lors de cet appel récursif.

Question 4. En déduire la forme de l'arbre binaire du code de Huffman associé à la source X. Quelle est la longueur maximale d'un mot de code ? La longueur minimale ?

Solution : D'après la Question 3., pour $k \leq m - 1$, la feuille correspondant à la probabilité p_k devra être positionnée à la hauteur k, tandis que celle correspondant à p_m est à hauteur $m - 1$. On a donc un arbre de la forme suivante :



La longueur maximale de ses mots est donc $m - 1$, et leur longueur minimale est 1.