Codes pour des sources discrètes

Exercice 1 **Bouteille empoisonnée**

Vous avez 8 bouteilles de vin dont une exactement est empoisonnée avec les probabilités suivantes : $p_1 = p_2 = 1/45, p_3 = p_4 = 6/45, p_5 = 7/45, p_6 = p_7 = p_8 = 8/45$. Vous voulez déterminer quelle bouteille est empoisonnée en les testant sur des rats. Un rat qui boit du poison meurt instantanément et un rat qui boit du vin n'est plus fiable pour une nouvelle dégustation ! Vous devez donc utiliser chaque rat au plus une fois.

Déterminer quelle bouteille est empoisonnée en minimisant le nombre moyen de rats utilisés.

Exercice 2 **Simulation parfaite à l'aide d'une pièce de monnaie**

Soit $\mathcal{X}$ un ensemble fini et P une loi de probabilité sur $\mathcal{X}$. Le but de cet exercice est de simuler une variable aléatoire X selon la probabilité P exactement à l'aide d'une pièce de monnaie (non biaisée). On veut aussi minimiser le nombre moyen de lancers, c'est-à-dire $\mathbf{E}[T]$, où T est le nombre de lancers effectués pour une simulation.

1. Montrer que toute stratégie peut être représentée par un arbre binaire (éventuellement infini) dont les feuilles sont étiquetées par des symboles de $\mathcal{X}$. Quelle condition doit être vérifiée pour que le nombre de lancements soit presque sûrement fini et que la loi simulée soit bien P ?

2. Étant donnée une stratégie, exprimer le nombre moyen de lancers comme l'entropie d'une variable aléatoire que l'on explicitera, et en déduire que nécessairement $\mathbf{E}[T] \geq H(X)$.

3. Donner une condition nécessaire et suffisante sur P pour que cette borne puisse être atteinte.

4. Dans le cas général, proposer une stratégie qui garantisse au moins $H(X) \leq \mathbf{E}[T] \leq H(X) + 2$.

5. En déduire une interprétation asymptotique de l'entropie pour la simulation de suites i.i.d.

Exercice 3 **Codage de Shannon-Fano-Ellias**

On considère une source générant des lettres de l'alphabet $\mathcal{X} = \{1, 2, \ldots, m\}$ selon une loi de probabilité p. On note $F(x) = \sum_{y \leq x} p(y)$ et $\overline{F}(x) = \sum_{y<x} p(y) + \frac{1}{2}p(x)$. On considère le code qui à x associe le mot code correspondant aux $\ell(x)$ premiers bits de $\overline{F}(x)$.

Comment choisir $\ell(x)$ pour avoir un code uniquement décodable et ayant une longueur moyenne la plus petite possible ? Donner une borne supérieure sur la longueur moyenne du code ?

Exercice 4 **Codage de Shannon et optimalité**

On considère une source générant des lettres de l'alphabet $\mathcal{X} = \{x_1, \ldots, x_k\}$ selon une loi de probabilité p. On suppose les lettres de X ordonnées de telle manière que $p(x_i) \geq p(x_j)$ si $i < j$. Pour tout $i \in \{1, \ldots, k\}$, on définit $F(x_i) = \sum_{j<i} p(x_j)$. On souhaite utiliser le code ϕ suivant : le mot $\phi(x_i)$ codant x_i est donné par les $\lceil -\log p(x_i) \rceil$ premiers bits du développement de $F(x_i)$.

1. Montrer que le code de Shannon est instantané et que sa longueur moyenne L vérifie

$$H(X) \leq L \leq H(X) + 1.$$

2. Le code de Shannon est-il optimal ?

3. Soit $\ell(x) = \lceil -\log p(x_i) \rceil$ les longueurs des mots code associés au code de Shannon et soit $\ell'(x)$ les longueurs associées à un code uniquement décodable. Montrer que

$$\mathbf{P}\left(\ell(X) \geq \ell'(X) + c\right) \leq 2^{-c+1}.$$

4. Montrer que si pour tout $x \in \mathcal{X}$, $p(x)$ est un rationnel dyadique alors

$$\mathbf{P}\left(\ell(X) < \ell'(X)\right) \geq \mathbf{P}\left(\ell(X) > \ell'(X)\right)$$

5. Dans le cas général, montrer que

$$\mathbf{E}\left(sgn(\ell(X) - \ell'(X) - 1)\right) \leq 0.$$

`Exercice 5 Optimalité de l'algortihme de Lempel-Ziv pour une source sans mémoire`

Soit $(U_i)_{i \in \mathbb{N}-\{0\}}$ uns suite de variables aléatoires i.i.d. à valeurs dans un ensemble fini $\mathcal{U}$ et d'entropie $H(U)$. On note $C(n)$ la v.a. égale au nombre maximal de mots distincts en lequel $U^{(n)} = U_1 \cdots U_n$ peut être découpé, c'est-à-dire, avec les notations du cours, $C(n) = c(U(n))$. Le but de cet exercice est de montrer qu'avec probabilité 1, on a

$$\limsup_{n \to \infty} \frac{C(n) \log C(n)}{n} \leq H(U).$$

Ce résultat permet de montrer l'optimalité de l'algorithme de Lempel-Ziv dans le cas particulier d'une source sans mémoire.

1. Pour un découpage de $u^n = u_1 \cdots u_n \in \mathcal{U}^n$ et c mots distincts, on note c_ℓ le nombre de mots de longueur ℓ. Montrer que

$$\log P(u_1, u_2, \ldots, u_n) \leq -\sum_{\ell} c_\ell \log c_\ell.$$

2. Montrer que pour toute v.a. X à valeurs entières et de moyenne $\mathbf{E}[X]$ on a

$$H(X) \leq (\mathbf{E}[X] + 1) \log(\mathbf{E}[X] + 1) - \mathbf{E}[X] \log \mathbf{E}[X],$$

avec égalité si et seulement si X suit une loi géométrique : $\mathbf{P}(X = k) = q^k(1 - q)$.

On admet (comme en cours) que $c(u^n) = O(n/\log n)$. On définit la variable aléatoire Z par $\mathbf{P}(Z = \ell) = \frac{c_\ell}{c}$.

3. Montrer que $\lim_{n \to \infty} \frac{c}{n} H(Z) = 0$.

4. Conclure.