Théorie de l'information

Exercice 1 Application de la distance de Kullback-Leibler

Soit $\mathcal{X}$ un ensemble fini, $\mathcal{P}(\mathcal{X})$ l'ensemble des mesures de probabilité sur $\mathcal{X}$.

Applications en théorie de l'information

1. Exprimer l'information mutuelle $I(X, Y)$ de deux variables aléatoires discrètes X et Y comme une divergence de Kullback-Leibler.

Les codages étudiés jusqu'ici requièrent la connaissance exacte de la loi μ de la source, ce qui est irréalisable en pratique : on ne dispose que d'une certaine estimation à priori, notée ν.

2. Calculer la perte de performance induite dans le cas où ν est dyadique, puis en donner un encadrement dans le cas où ν est quelconque.

Second principe thermodynamique

Soient $(X_n)_{n\geq 0}$ et $(Y_n)_{n\geq 0}$ deux chaînes de Markov de même matrice de transition sur $\mathcal{X}$.

3. Montrer que $D(\mathcal{L}(X_n)||\mathcal{L}(Y_n))$ décroît avec n. Qu'en déduire lorsque la chaîne admet la loi uniforme comme mesure stationnaire ?

Exercice 2 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

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

Indication : on pourra d'abord montrer que l'application signe vérifie $sgn(x) \leq 2^x - 1$ pour x entier.

5. Dans le cas général, montrer que

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

Exercice 3 Test d'hypothèse

Soient P et Q deux distributions de probabilité sur un espace fini $\mathcal{U}$. On souhaite distinguer P et Q à l'aide d'un échantillon de taille k, c'est-à-dire à partir du résultat de k tirages indépendants dans $\mathcal{U}$. Le test est défini par un ensemble $A \subseteq \mathcal{U}^k$: si l'échantillon $(U_1, \ldots, U_k) \in A$, alors le test retourne P sinon, il retourne Q.

On désire que la probabilité d'erreur du test soit au plus ϵ si P est la vraie distribution ($P(A) \geq 1 - \epsilon$) et minimiser la probabilité d'erreur lorsque Q est la vraie distribution. On pose

$$\beta(k, \epsilon) = \min\{Q(A) \mid A \subseteq \mathcal{U} \text{ et } P(A) \geq \epsilon\}.$$

Pour ce faire, on définit l'entropie relative de deux distributions p et q, ou distance de Kullback-Leibler, par

$$D(p||q) = \mathbf{E}_p\left[\log \frac{P(U)}{Q(U)}\right] = \sum_{u\in\mathcal{U}} p(u)\log\frac{p(u)}{q(u)},$$

avec les conventions $0\log\frac{0}{0} = 0$, $0\log\frac{0}{q} = 0$ et $p\log\frac{p}{0} = +\infty$.

On suppose désormais que pour tout $u \in \mathcal{U}$, $p(u)q(u) > 0$, et on définit pour tout $\delta > 0$ et tout $n \in \mathbb{N}$ l'ensemble $A_\delta^{(n)}(p||q)$ par l'ensemble des suites $(u_1, \dots u_n) \in \mathcal{U}^n$ telles que

$$2^{-n(D(p||q)+\delta)} \leq \frac{q(u_1,\dots,u_n)}{p(u_1,\dots,u_n)} \leq 2^{-n(D(p||q)-\delta)}.$$

1. Montrer que pour tout $(u_1,\dots,u_n) \in A_\delta^{(n)}(p||q)$,

$$D(p||q) - \delta \leq \frac{1}{n}\log\frac{p(u_1,\dots,u_n)}{q(u_1,\dots,u_n)} \leq D(p||q) + \delta.$$

2. Montrer que pour tout $\epsilon > 0$ et tout n suffisamment grand, $P(A_\delta^{(n)}(p||q)) \geq 1-\epsilon$.

3. En déduire pour que $\epsilon > 0$ et tout n suffisamment grand,

$$(1-\epsilon)2^{-n(D(p||q)+\delta)} \leq Q(A_\delta^{(n)}(p||q)) \leq 2^{-n(D(p||q)-\delta)}.$$

4. Conclure en montrant que pour tout $\epsilon \in (0,1)$,

$$\lim_{k\to\infty}\frac{1}{k}\log\beta(k,\epsilon) = -\sum_{u\in\mathcal{U}} p(u)\log\frac{p(u)}{q(u)}.$$

`Exercice 4  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

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

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.