Entropie de Shannon et codage

Exercice 1 Entropie de Shannon et codage

Soit $A = \{\alpha_1, ..., \alpha_D\}$ un alphabet et $M = \{x_1, ..., x_k\}$ un ensemble d'objets appelés messages. Un codage est une application de M dans A^* qui à un message x_i associe son code c_i de longueur l_i.

Un codage est dit ***préfixe*** s'il ne contient pas deux mots dont l'un est préfixe de l'autre.

▷ 1. *(Inégalité de Kraft)* Montrer que si le codage h est préfixe alors

$$\sum_{i=1}^{k} D^{-l_i} \leqslant 1$$

▷ 2. Supposons l'inégalité précédente satisfaite. Montrer que pour tout alphabet A de D lettres et tout ensemble de k messages il existe un codage h préfixe et dont les codes sont de longueurs (l_i).

▷ 3. *(Borne de Shannon sur la longueur moyenne d'un code préfixe)* Supposons maintenant que les messages proviennent d'une source aléatoire, le message x_i ayant la probabilité p_i. La longueur moyenne de codage est alors définie par :

$$L(h) = \sum_{i=1}^{k} p_i l_i$$

Soit L_{inf} la plus petite de ces longueurs pour un code préfixe. Nous allons estimer L_{inf}.
Shannon définit l'entropie d'un événement comme la quantité de surprise $-\log p_i$ qu'aurait un observateur lorsqu'il découvre la réalisation de cet événement. Plus cet événement est improbable plus l'observateur sera surpris. Si on fait la moyenne sur tous les événements possibles on obtiendra l'entropie du système :

$$H_D(p) = -\sum_{i=1}^{k} p_i \log_D p_i$$

C'est la quantité d'information relative à la distribution $p = (p_1, ..., p_k)$.
Montrer que

$$H \leqslant L_{inf} \leqslant H + 1$$

Exercice 2 L'algorithme de Shannon

On fixe $D = 2$ et on suppose $p_1 \geqslant p_2 \geqslant ... \geqslant p_k$. On pose $q_s = \sum_{i=1}^{s-1} p_i$. La méthode de Shannon consiste à écrire pour chaque i un développement dyadique de q_i :

$$q_i = \frac{a_1(i)}{2^1} + \frac{a_2(i)}{2^2} + \frac{a_3(i)}{2^3} + \cdots$$

et à prendre pour code de x_i : $a_1(i)...a_{l_i}(i)$ avec $l_i = \lceil -\log_2 p_i \rceil$.

En d'autres termes, x_i est l'écriture en base 2 de q_i, tronquée à l_i chiffres.

▷ 1. Donner le code obtenu pour

p_1	p_2	p_3	p_4	p_5	p_6
0,27	0,23	0,2	0,15	0,1	0,05

.

Calculer L et la comparer à l'entropie.

▷ 2. Montrer que ce code est préfixe.

▷ 3. Montrer que la longueur moyenne L vérifie $H_2(p) \leqslant L \leqslant H_2(p) + 1$.

▷ 4. Ce codage est-il optimal ?

Exercice 3 L'algorithme de Fano

Fixons $D = 2$ et supposons $p_1 \geqslant p_2 \geqslant ... \geqslant p_k$. On regroupe les u premiers objets où u est le plus petit entier tel que

$$p_1 + ... + p_u \geqslant \frac{1}{2}$$

Nous obtenons ainsi une partition M de l'ensemble des objets en deux sous-ensemble M_0 et M_1. Soient π_0 et π_1 les probabilités respectives de ces ensembles. Nous appliquons à M_0 l'algorithme de dichotomie pour obtenir une partition $M_0 = M_{00} + M_{01}$ où $M_{00} = \{x_1, ..., x_v\}$ et v est le plus petit entier tel que $p_1 + ...p_v \leqslant \pi_0/2$. On applique à M_1 ce même algorithme et on obtient la partition $M_1 = M_{10} + M_{11}$. On continue l'algorithme ainsi de suite jusqu'à ce qu'on ne puisse plus appliquer la dichotomie, ce qui se produit quand un ensemble M_a ne comporte plus qu'un élément et a est alors le code de cet élément.

▷ 1. Reprendre les questions 2.1 à 2.4.

Exercice 4 L'algorithme mystère

▷ 1. Donner un algorithme pour obtenir un codage préfixe optimal. Le faire tourner sur l'exemple de la question 2.1 et comparer la longueur moyenne L obtenue avec l'entropie de la distribution.