Théorie de l'Information et Codage: Fiche d'exercices 1

à rendre pour le cours du 26 mars 2013.

Instructions: merci à chacun de rendre une copie manuscrite. Si vous avez réfléchi à plusieurs sur un problème, mettez les noms de vos collaborateurs.

Problème 1 (4 points)

On considère une source ayant un alphabet de $M = 4$ symboles ayant des probabilités $p_1 \geq p_2 \geq p_3 \geq p_4 > 0$.

1. Montrer que si $p_1 = p_3 + p_4$, il existe un code de Huffman dont tous les mots-code ont même longueur et un autre avec un mot-code de longueur 1, un de longueur 2 et 2 mots-code de longueur 3.
2. Quelle est la plus grande valeur de p_1, notée $p_{\max}$, telle que $p_1 = p_3 + p_4$?
3. Quelle est la plus petite valeur de p_1, notée $p_{\min}$, telle que $p_1 = p_3 + p_4$?
4. Montrer que si $p_1 > p_{\max}$, alors tout code de Huffman a un mot-code de longueur 1.
5. Montrer que si $p_1 > p_{\max}$, alors tout code optimal instantané a un mot-code de longueur 1.
6. Montrer que si $p_1 < p_{\min}$, alors tous les mots-code ont longueur 2 pour tout code de Huffman.
7. On suppose $M > 4$. Trouver la plus petite valeur $p'_{\max}$ telle que $p_1 > p'_{\max}$ garantisse qu'un code de Huffman aura un mot-code de longueur 1.

Problème 2 (3 points)

On considère une source avec M symboles équiprobables.

1. Soit $k = \lceil \log M \rceil$. Montrer que pour un code de Huffman, les seules longueurs possibles des mots-code sont k et $k - 1$.
2. Quelle est la longueur moyenne $\overline{L}$ des mots-code en bits par symbole de la source?
3. Soit $y = M/2^k$. Trouver le maximum de $\overline{L} - \log M$ comme fonction de $1/2 \leq y \leq 1$. Qu'en dire par rapport à la borne $\overline{L} \leq H[X] + 1$?

Problème 3 (3 points)

Une source est dite stationnaire si $P(U_1 = u_1, \ldots, U_L = u_L) = P(U_{j+1} = u_1, \ldots, U_{j+L} = u_L)$ pour tout L et j. Dans ce cadre, il y a deux notions naturelles d'entropies:

$$H_L(U) = \frac{1}{L} H(U_1, U_2, \ldots, U_L) \text{ et } H(U_L|U_1, \ldots, U_{L-1}).$$

Montrer que pour une source stationnaire avec $H(U_1) < \infty$, on a:

1. $H(U_L|U_1, \ldots, U_{L-1})$ est décroissant en L.
2. $H_L(U) \geq H(U_L|U_1, \ldots, U_{L-1})$.
3. $H_L(U)$ est décroissant en L.
4. $\lim_{L\to\infty} H_L(U) = \lim_{L\to\infty} H(U_L|U_1, \ldots, U_{L-1})$.

Problème 4 (4 points)

Dans un casino, un jeu consiste à parier sur un tirage aléatoire d'une variable X à valeur dans $\{1, \ldots, K\}$. La distribution de X est $p(x)$. Si $X = k$, le casino multiplie la somme misée sur k par $1/p(k)$ et toutes les autres mises sont perdues. Une stratégie q d'un joueur est de mettre de coté (c'est à dire ne pas miser) une fraction $q(0)$ de son capital et pour le reste de miser une fraction $q(k)$ de son capital sur la valeur k. Donc une stratégie q est telle que $q(k) \geq 0$ pour tout $k \geq 0$ et $\sum_{k=0}^{K} q(k) = 1$. On suppose que la distribution $p(x)$ est connue du joueur.

1. On considère une stratégie q avec $q(0) > 0$. Montrer qu'il existe une stratégie $\hat{q}$ avec $\hat{q}(0) = 0$ qui est aussi performante que q, dans le sens où le joueur aura la même somme d'argent quelle que soit la valeur prise par X pour les deux stratégies.
   On suppose pour la suite que $q(0) = 0$. On définit:

   $$R_n = \frac{1}{n} \log \frac{C_n}{C_0},$$

   le taux de retour du joueur où C_0 est le capital initial et C_n est le capital après n tours dans le jeu.
2. En utilisant la loi des grands nombres, calculer $r = \lim_{n\to\infty} R_n$ en fonction des distributions p et q.
3. On suppose maintenant que avant chaque tour i, le joueur a une information donnée par Y_i qui est corrélée à X. La distribution de (X_i, Y_i) est $p(x, y)$. La stratégie du joueur au tour i, étant donné l'information $Y_i = y$ est de parier une fraction notée $q(k|y)$ de son capital sur la valeur k. Recalculer $r = \lim_{n\to\infty} R_n$.
4. Trouver la stratégie $q(x|y)$ qui maximise r.

Problème 5 (6 points)

Soit X une variable aléatoire à valeur dans $\mathcal{N} = \{1, 2, \dots\}$ et de distribution $p(x)$. Un encodage binaire de X est une injection $\phi : \mathcal{N} \to \{0,1\}^*$, de $\mathcal{N}$ dans l'ensemble des mots binaires finis (y compris le mot vide). La longueur moyenne de l'encodage ϕ est: $\ell(\phi) = \sum_{u\in\mathbb{N}} |\phi(x)|p(x)$, où $|\phi(x)|$ est la longueur du mot $\phi(x)$. Dans certain cas, s'il existe un symbole encodant la fin d'un message, il n'est pas nécessaire de considérer des codes ayant la propriété du préfixe. On définit donc:

$$\mathcal{L}_{1-1}(X) = \min\{\ell(\phi) : \phi \text{ est un encodage de } X\}.$$

1. Montrer que $\mathcal{L}_{1-1}(X) \leq H(X)$.

2. Montrer que pour toute variable aléatoire U à valeur dans $\{0, 1, \dots\}$, on a: $H(U) \leq \log(E[U] + 1) + \log e$.

3. En déduire que: $H(X) \leq \mathcal{L}_{1-1}(X) + \log(\mathcal{L}_{1-1}(X) + 1) + \log e$ puis une borne inférieure pour $\mathcal{L}_{1-1}(X)$. (On pourra utiliser l'axiome (A8) des notes de cours (Section 1.7) dans la définition axiomatique de l'entropie pour une partition bien choisie de la distribution $p(x)$).

On considère maintenant le problème suivant: (X, Y) est un couple de variables aléatoires à valeur dans l'espace dénombrable $\mathcal{X} \times \mathcal{Y}$ de distribution $p(x, y)$. Alice connaît X, Bob connaît Y et veut connaître X. On suppose qu'Alice peut communiquer vers Bob sans erreur; que Bob doit pouvoir déterminer la fin d'un message d'Alice; qu'Alice et Bob se sont mis d'accord sur un protocole déterministe qui peut dépendre de p. Le but est de trouver le nombre moyen de bits qu'Alice doit envoyer à Bob.

On définit le support $S = \{(x, y),\ p(x, y) > 0\}$ et $x \neq x'$ sont ambigus si il existe y tel que $(x, y), (x', y) \in S$. Un protocole pour des entrées restreintes est une fonction $\phi : \mathcal{X} \to \{0,1\}^*$ telle que pour x et x' ambigus, $\phi(x)$ n'est ni égal à, ni un préfixe de $\phi(x')$. On définit le nombre de bits moyens pour ϕ comme précédement: $\ell(\phi) = \sum_x |\phi(x)|p(x)$. On définit alors

$$\overline{L} = \min\{\ell(\phi) : \phi \text{ est un protocole pour entrées restreintes } (X, Y)\}.$$

4. Montrer que

$$H(X|Y) \leq \overline{L} \leq H(X) + 1,$$

que ces bornes sont les meilleures possibles et qu'elles peuvent être arbitrairement éloignées l'une de l'autre.

Clairement, $\overline{L}$ ne dépend de (X, Y) que par S et la distribution $p(x)$ (les valeurs de $p(y|x)$ n'interviennent pas dans les définitions). On définit donc le graphe G dont l'ensemble des sommets est $\mathcal{X}$ et deux sommets distincts x et x' sont connectés si ils sont ambigus. Le graphe probabiliste (G, X) est défini par le graphe G et la distribution de probabilité sur ses sommets $p(x)$. $\overline{L}$ ne dépend que de (G, X).

Si X est une variable aléatoire à valeur dans $\mathcal{X}$ et c est une fonction définie sur $\mathcal{X}$ alors $c(X)$ est une variable aléatoire d'entropie:

$$H[c(X)] = -\sum_{\gamma\in c(\mathcal{X})} p[c^{-1}(\gamma)] \log p[c^{-1}(\gamma)],$$

où c^{-1} est l'inverse de c et la probabilité d'un ensemble est la somme des probabilités de ses éléments. On définit l'entropie chromatique d'un graphe probabiliste (G, X) par:

$$H(G, X) = \min\{H[c(X)],\ c \text{ est un coloriage de } G\}.$$

5. Donner l'entropie chromatique pour: le graphe vide, le graphe complet, le pentagone avec la distribution uniforme sur ses sommets, le cycle avec $p_0 = 0.3, p_1 = p_2 = p_3 = 0.2$ et $p_4 = 0.1$.

6. Un protocole pour des entrées non-restreintes est une fonction $\phi : \mathcal{X} \to \{0,1\}^*$ telle que pour $x \neq x'$, $\phi(x)$ n'est pas un préfixe propre de $\phi(x')$ et si x et x' sont ambigus, $\phi(x) \neq \phi(x')$. Soit $\overline{\mathcal{L}} = \min\{\ell(\phi) :\ \phi \text{ est un protocole pour entrées non-restreintes } (X, Y)\}$. Montrer que

$$H(G, X) \leq \overline{\mathcal{L}} \leq H(G, X) + 1.$$

7. En utilisant la première partie de l'exercice, montrer que:

$$H(G, X) - \log[H(G, X) + 1] - \log e \leq \overline{L} \leq H(G, X) + 1.$$