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

à rendre pour le cours du 13 mars 2012.

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)

1. Une source a un alphabet de 4 lettres a_1, a_2, a_3, a_4 avec des probabilités $p(a_1) > p(a_2) = p(a_3) = p(a_4)$. Trouver le plus petit q tel que $p(a_1) > q$ implique que $n_1 = 1$ où n_1 est la longueur du mot code correspondant à a_1 dans le code de Huffman binaire.
2. Montrer que si $p(a_1) = q$, il est possible de trouver un code de Huffman avec $n_1 > 1$.
3. On suppose maintenant que la source a un alphabet de K lettres et $p(a_1) > p(a_2) \geq \cdots \geq p(a_K)$. Est-ce que $p(a_1) > q$ implique encore $n_1 = 1$?
4. Si maintenant $p(a_1) \geq p(a_2) \geq \cdots \geq p(a_K)$, trouver le plus grand q' tel que $p(a_1) < q'$ implique $n_1 > 1$.

Problème 2 (2 points)

Une variable aléatoire prend ses valeurs dans un alphabet de K lettres et chaque lettre a la même probabilité. Ces lettres sont encodées dans des mots binaires de façon à minimiser la longueur moyenne des mots code. On définit l'entier j et $1 \leq x < 2$ par $K = x2^j$.

1. Montrer que tous les mots code ont pour longueur j ou $j + 1$.
2. Quelle est la longueur moyenne d'un mot code?

Problème 3 (2 points)

On considère un système de cryptographie: P est le message à transmettre, C est le message crypté. Il est obtenu à partir de P et d'une clé aléatoire K, donc il existe une fonction déterministe telle que $C = f(P, K)$. Le récepteur a accès à la clé K et au message crypté C et doit pouvoir retrouver le message original. Il existe donc une fonction déterministe telle que $g(C, K) = P$. La question est alors: quelle doit être la longueur minimale de la clé aléatoire pour rendre le message crypté indéchiffrable par un tiers? On pourra montrer que si P et C sont indépendants alors $H(K) \geq H(P)$.

Problème 4 (6 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.

Paradoxe de St Petersbourg: on considère le jeu suivant: pour un prix d'entrée de c euros, un joueur recoit 2^k euros avec probabilité 2^{-k}. Certains disent qu'ils ont 'intérêt' à jouer quelque soit la valeur de c. Pourquoi? Bernoulli dans Specimen theoriae novae de mensura sortis (1738) propose de résoudre le paradoxe en intoduisant une fonction d'utilité logarithmique pour l'argent: une somme s a une utilité $\ln s$. L'utilité moyenne devient alors $E[\ln X] = \ln 4$ et donc Bernoulli accepte de jouer uniquement si $c < 4$. Le choix de la fonction d'utilité logarithmique étant arbitraire, nous allons étudier une autre solution de ce paradoxe.

5. On suppose maintenant que le joueur peut acheter une part du jeu. Par exemple, s'il investit $c/2$ euros dans le jeu, il recoit $X/2$ avec $P(X = 2^k) = 2^{-k}$. On suppose les $X_1, X_2, \ldots$ i.i.d. et que le joueur est obligé de réinvestir la totalité de sa fortune à chaque étape. Soit $F_n(c)$ sa fortune au temps n. On suppose que $F_0(c) = 1 \leq c$. Montrer qu'il existe c^* tel que pour $c < c^*$ sa fortune tend vers l'infini et si $c > c^*$ elle tend vers 0. Calculer la valeur du prix 'équitable' c^* et pour une distribution différente de X.

6. Etudier le cas où le joueur peut choisir de ne miser qu'une fraction de sa fortune. Pour quelle valeur de c, le joueur va-t-il choisir de parier toute sa fortune?

7. Question bonus: Si on a $P(X = 2^{2^k-1}) = 2^{-k}$, faut-il investir la totalité de sa fortune pour toute valeur de c?

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:

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