Module D4MA1C20, Compléments magistère
Devoir n^{0}1
A rendre le 9 mars 2012

1. Conditionner ne diminue pas l'information mutuelle

Soient X, Y et Z des variables aléatoires.

1– Montrer que $I(X;Y|Z) \leqslant I(X;Y,Z)$ et que $I(X;Y) \leqslant I(X;Y,Z)$.

2– Démontrer l'identité

$$I(X;Y|Z) - I(X;Y) = I(Z;Y|X) - I(Z;Y).$$

3– Donner un exemple de triplet de variables aléatoires tel que $I(X;Y|Z) < I(X;Y)$.

4– Donner un exemple de triplet de variables aléatoires tel que $I(X;Y|Z) > I(X;Y)$.

Solution :

1– Dans le cours, on trouve la formule $I(X;Y,Z) = I(X;Z) + I(X;Y|Z)$ (règle d'addition pour l'information mutuelle), qui entraîne que $I(X;Y,Z) \geqslant I(X;Y|Z)$ et que $I(X;Y,Z) \geqslant I(X;Z)$. En échangeant Y et Z, cela donne $I(X;Y,Z) \geqslant I(X;Y)$.

On peut aussi utiliser la version conditionnelle de l'expression de l'information mutuelle en fonction de l'entropie conditionnelle, $I(X;Y|Z) = H(X|Z) - H(X|Y,Z)$. Comme $H(X|Z) \leqslant H(X)$, il vient

$$I(X;Y|Z) \leqslant H(X) - H(X|Y,Z) = I(X;Y,Z).$$

De même,

$$I(X;Y) = H(X) - H(X|Y) \leqslant H(X) - H(X|Y,Z) = I(X;Y,Z).$$

2– On exprime les informations mutuelles en fonctions d'entropies conditionnelles, $I(X;Y|Z) = H(Y|Z) - H(Y|X,Z)$ et $I(X;Y) = H(Y) - H(Y|X)$, et on remplace les entropies conditionnelles par des entropies jointes (formule d'addition). On obtient

$$\begin{aligned} I(X;Y|Z) - I(X;Y) &= H(Y|Z) - H(Y|X,Z) - H(Y) + H(Y|X) \\ &= H(Y,Z) - H(Z) - H(X,Y,Z) + H(X,Z) - H(Y) + H(X,Y) - H(X), \end{aligned}$$

qui est symétrique en X, Y et Z. Donc intervertir X et Z donne le même résultat.

On peut aussi partir de la règle d'addition pour l'information mutuelle, $I(X;Y|Z) = I(Y;X|Z) = I(Y;X,Z) - I(Y;Z)$. Il vient

$$I(X;Y|Z) - I(X;Y) = I(Y;X,Z) - I(Y;Z) - I(Y;X),$$

qui est symétrique en X et Z.

3– Si $Z = Y$, $I(Z;Y) = H(Y)$ et $I(Y;Z|X) = H(Y|X)$, d'où $I(X;Y|Z) - I(X;Y) = -I(Y;X) \leqslant 0$, et qui est strictement négatif en général, et en particulier si $X = Y$ n'est pas presque sûrement constante.

4– Si Z est indépendante de Y, $I(Z;Y)=0$, d'où $I(X;Y|Z)-I(X;Y)=I(Y;Z|X)\geqslant 0$, avec de bonnes chances d'être strictement positif. C'est notamment le cas pour les variables Y = réussite au code, Z = réussite à la conduite et $X=YZ$ = réussite au permis de l'exercice 4 de la feuille 1. En effet, sachant que $X=1$, les variables conditionnées $Y|X=1$ et $Z|X=1$ ne sont pas indépendantes, puisqu'elle ne peuvent pas s'annuler simultanément, donc $I((Y|X=1);(Z|X=1))>0$, donc $I(Y;Z|X)>0$.

2. Traduire le Shadok

Les Shadoks sont des sortes d'oiseaux sans cervelle, mais qui utilisent néanmoins un langage rudimentaire. On étudie les codages uniquement décodables de cette langue.

1– La langue des Shadoks ne comporte que 4 mots, *ga*, *bu*, *zo* et *meu*. Leurs voisins Gibis l'ont traduite en utilisant seulement deux lettres de leur alphabet, ♡ et ♢. Donner un exemple de codage qui satisfasse à ces spécifications, avec une fonction longueur la plus petite possible.

2– Les Shadoks sont très bêtes. Quand on leur pose une question, la plupart du temps, il ne répondent rien du tout. Le silence constitue donc le 5ème mot de leur langue. Les Gibis ont dû coder ce mot supplémentaire en faisant appel à la troisième lettre de leur alphabet, ♠. Peut on le faire en n'utilisant toujours que 2 lettres par mot au plus ?

3– Quand ils sont mécontents, les Shadoks émettent une sorte de piaillement impossible à transcrire de façon phonétique. Pendant longtemps, les Gibis ont cru qu'il s'agissait d'un 6ème mot, et l'ont codé ♠♠♠ (ce qui est totalement imprononçable en Gibi). Les Gibis auraient ils pu être plus économes et coder chacun des 6 mots par une suite d'au plus 2 lettres parmi les trois disponibles, ♡, ♢ et ♠ ?

4– Les Gibis ont observé les fréquences suivantes dans la langue Shadok.

ga	*bu*	*zo*	*meu*	silence	piaillement
1/27	1/27	2/27	3/27	14/27	6/27

On sait seulement que le piaillement se traduit par ♠♠♠ en Gibi, et que les traductions des autres mots ont au plus deux lettres chacune. Donner un minorant optimal de la longueur moyenne de la traduction Gibi.

5– Existe t'il un codage n'utilisant que les deux lettres ♡ et ♢ et de longueur moyenne inférieure à celle de la traduction Gibi ?

6– Comparer la borne donnée par le Théorème du codage de source aux bornes obtenues dans les questions précédentes.

Solution :

1– Le codage $ga\mapsto ♡♡$, $bu\mapsto ♡♢$, $zo\mapsto ♢♡$, $meu\mapsto ♢♢$ est uniquement décodable. En effet, tous les mots ont 2 lettres. Une suite de lettres de longueur paire se découpe de façon unique en mots qui correspondent aux 4 mots Shadoks. La consigne de rendre la longueur la plus petite possible est un peu vague. Voilà une remarque pertinente mais insuffisante. Le Théorème de codage de source, appliqué à la distribution de probabilité uniforme sur les 4 mots, affirme que la longueur moyenne est au moins égale à l'entropie de la source, qui vaut 2. Donc si on code un mot Shadok avec une seule lettre, il faut en coder un autre avec au moins 3 lettres.

2– Oui. Le Théorème de Kraft prédit qu'il existe un codage sans préfixe dont les longueurs sont 1, 1, 2, 2, 2. En effet, $3^{-1}+3^{-1}+3^{-2}+3^{-2}+3^{-2}=1$. Par exemple, le codage $ga\mapsto ♡$, $bu\mapsto ♢$, $zo\mapsto ♠♡$, $meu\mapsto ♠♢$, silence $\mapsto ♠♠$ est sans préfixe, et donc uniquement décodable.

3– Oui. Tout codage non singulier dont les mots-codes ont tous le même nombre de lettres (ici, 2) est sans préfixe donc uniquement décodable. Or il y a 9 suites de 2 lettres disponibles. Le codage suivant convient : $ga \mapsto \heartsuit\heartsuit$, $bu \mapsto \heartsuit\diamondsuit$, $zo \mapsto \diamondsuit\heartsuit$, $meu \mapsto \diamondsuit\diamondsuit$, silence $\mapsto \spadesuit\spadesuit$, piaillement $\mapsto \heartsuit\spadesuit$.

4– Montrons que parmi les 6 traductions, il y a au plus un mot d'une lettre. Soit x le nombre de mots d'une lettre parmi les 6. Alors le nombre de mots de deux lettres est $5-x$. Le codage, dans un alphabet à 3 éléments, étant uniquement décodable, l'inégalité de Kraft énonce que $3^{-1}x + (5-x)3^{-2} + 3^{-3} \leqslant 1$, i.e. $6x \leqslant 11$. Cela entraîne que $x \leqslant 1$.

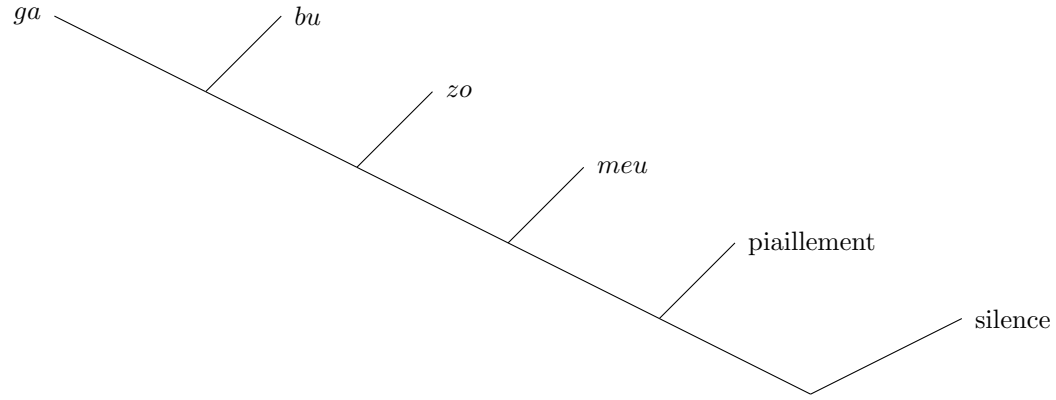
Pour minimiser la longueur moyenne, il faut affecter un mot d'une lettre au silence, qui a la probabilité la plus élevée. Cela donne l'inégalité

$$\mathbb{E}(\ell(X)) \geqslant 3.\frac{6}{27} + 2.\frac{1}{27} + 2.\frac{1}{27} + 2.\frac{2}{27} + 2.\frac{3}{27} + 1.\frac{14}{27} = \frac{46}{27} = 1.7037...$$

Cette minoration est atteinte par le codage $ga \mapsto \diamondsuit\heartsuit$, $bu \mapsto \diamondsuit\diamondsuit$, $zo \mapsto \spadesuit\heartsuit$, $meu \mapsto \spadesuit\diamondsuit$, silence $\mapsto \heartsuit$, piaillement $\mapsto \spadesuit\spadesuit\spadesuit$, qui est sans préfixe, donc uniquement décodable.

5– Non. Pour la longueur moyenne d'un codage uniquement décodable à deux lettres, le Théorème de codage de source donne la borne inférieure $\mathbb{E}(\ell(X)) \geqslant H(X) = 1.9561 > 1.7037....$

On peut aussi utiliser le fait que le minimum de la longueur moyenne d'un codage uniquement décodable à deux lettres est donné par un codage de Huffman. L'arbre de Huffman de la distribution donnée est



Arbre de Huffman

ce qui donne le codage

ga	*bu*	*zo*	*meu*	silence	piaillement
$\heartsuit\heartsuit\heartsuit\heartsuit$	$\heartsuit\heartsuit\heartsuit\diamondsuit$	$\heartsuit\heartsuit\diamondsuit$	$\heartsuit\diamondsuit$	$\diamondsuit$	$\heartsuit\diamondsuit$

et la longueur moyenne

$$\mathbb{E}(\ell(X)) = 5.\frac{1}{27} + 5.\frac{1}{27} + 4.\frac{2}{27} + 3.\frac{3}{27} + 1.\frac{14}{27} + 2.\frac{6}{27} = \frac{53}{27} = 1.963... > 1.7037...$$

6– Pour un codage à 3 lettres, le Théorème de codage de source donne

$$\mathbb{E}(\ell(X)) \geqslant \frac{H(X)}{\log_2 3} = \frac{1.9561}{1.585} = 1.2341,$$

qui est nettement inférieur aux bornes ci-dessus. La contrainte de n'utiliser qu'un mot de longueur > 2 est forte. Néanmoins, la borne supérieure $\frac{H(X)}{\log_2 3} + 1$ est satisfaite par ces codages contraints.

NB. La longueur moyenne minimale d'un codage uniquement décodable à 3 lettres pour la distribution observée vaut $\frac{4}{3}$.