ENSEIRB **Janvier 2009**
Informatique 1ère année

Théorie de l'information - Corrigé

Exercice 1

On considère une source binaire S, émettant les symboles 0 et 1 avec probabilités respectives p et $1-p$ ($0 < p < 0.5$). On note S^k la source d'ordre k, émettant des k-uplets successifs de symboles de S ; ainsi, S^2 emet les symboles 00, 01, 10 et 11 avec probabilités respectives p^2, $p(1-p)$, $p(1-p)$ et $(1-p)^2$.

1. Donner, en fonction de p, les entropies de S, S^2 et S^3. **Question à la limite de la question de cours : la définition de l'entropie donne $H(S) = -p\log(p) - (1-p)\log(1-p)$, et, par indépendance, on peut affirmer sans calcul que l'on a $H(S^2) = 2H(S)$ et $H(S^3) = 3H(S)$; calculer les probabilités des différents symboles de S^2 et S^3, et de là les entropies, n'aboutit qu'à perdre du temps (et prendre des risques d'erreurs de calcul)**
2. Justifier l'existence d'un code uniquement déchiffrable pour S^2, d'assortiment de longueurs $\{1, 2, 3, 3\}$. Si un tel code est optimal, quels symboles peuvent être codés avec longueur 1 ? avec longueur 3 ? **On s'attend ici à l'utilisation de la fonction de Kraft : l'assortiment de longueurs proposé à pour fonction de Kraft $1/2+1/4+1/8+1/8 = 1$, donc il existe un tel code uniquement déchiffrable. Par ailleurs, on sait que dans un code optimal, les symboles doivent être codés avec des longueurs qui sont dans l'ordre inverse de leurs probabilités : le symbole codé avec longueur 1 doit donc être le plus probable (donc 11, de probabilité $(1-p)^2$, puisque $p < 0.5$), et les deux symboles codés avec longueur 3 doivent être les deux moins probables (donc 00, et soit 01, soit 10).**
3. En déduire la longueur moyenne d'un tel code, et déterminer pour quelles valeurs de p, un tel code est plus efficace qu'un code de longueur uniforme égale à 2. **La question précédente permet de calculer la longueur moyenne : $\overline{\ell} = (1-p)^2 + 2p(1-p) + 3p(1-p) + 3p^2$, qui se développe en $1 + 3p - p^2$. La longueur moyenne d'un code de longueur constante 2, est 2 (pas besoin de faire le calcul !) ; le plus efficace de deux codes est celui qui a la plus faible longueur moyenne, donc on cherche pour quelles valeurs de p on a $1 + 3p - p^2 < 2$. Le trinôme $p^2 - 3p + 1$ est positif en dehors de l'intervalle de ses racines $\frac{3\pm\sqrt{5}}{2}$, dont seule la plus petite est entre 0 et 1 ; par conséquent, le code $(1, 2, 3, 3)$ est plus efficace que le code $(3, 3, 3, 3)$ lorsque $p < \frac{3-\sqrt{5}}{2} \simeq 0.38$.**
4. Pour $p = 1/3$, déterminer un codage binaire optimal pour chacune des sources S, S^2 et S^3, et calculer son efficacité. **Puisque l'on connait p, le plus simple est ici de calculer dans chaque cas un code de Huffman. Pour S, une solution**

est $C(0) = 0$, $C(1) = 1$ **; pour** S^2, $H(00) = 000$, $H(01) = 001$, $H(10) = 01$, $H(11) = 1$ **; pour** S^3, $H(000) = 0000$, $H(001) = 0001$, $H(010) = 0010$, $H(011) = 100$, $H(100) = 0011$, $H(101) = 101$, $H(110) = 01$, $H(111) = 11$**. Les efficacités sont respectivement** 0.918, 0.972, **et** 0.979**.**

5. Justifier, sans calcul, qu'un code binaire optimal pour S^4 a une efficacité au moins égale à celle d'un code binaire optimal pour S^2. **Question volontairement un peu plus avancée ! Parmi les codes préfixes possibles pour S^4, il y a tous ceux que l'on peut obtenir en doublant les codes préfixes pour S^2 (en considérant un bloc de S^4 comme deux blocs consécutifs de S^2). La longueur moyenne sera double, tout comme l'entropie de S^4 par rapport à S^2, donc l'efficacité sera la même. Par conséquent, pour tout code préfixe de S^2, on sait construire un code préfixe pour S^4, de même efficacité : donc, un code optimal pour S^4 aura au moins la même efficacité qu'un code optimal pour S^2.**

Exercice 2

On considère une source S qui émet les symboles a, b, c, d, e, f avec probabilités respectives 0.4, 0.1, 0.06, 0.1, 0.3 et 0.04.

1. Quelle est l'entropie de S ? Quelle est l'efficacité maximale d'un code binaire de longueur constante pour S ? **Le calcul de l'entropie (binaire !) de S, donne** 2.14**. Le code ayant 6 symboles, le code de longueur constante minimale est de longueur** 3**, d'où une efficacité de** $2.14/3 = 0.714$**.**
2. Alice prétend avoir construit pour S un code binaire préfixe de longueur moyenne 2.1. Est-ce possible ? Et si le code est ternaire, est-ce possible ? **La longueur moyenne proposée par Alice est inférieure à l'entropie : son code ne peut pas être uniquement déchiffrable, donc il ne peut pas être préfixe (ou alors sa longueur moyenne n'est pas** 2.1**). Avec un codage ternaire, l'entropie ternaire n'est que de** 1.35**, donc rien ne s'oppose à ce qu'un tel code existe.**
3. Bob propose le code suivant pour S : $C(a) = 0$, $C(b) = 100$, $C(c) = 1100$, $C(d) = 101$, $C(e) = 111$, $C(f) = 1101$. Ce code est-il uniquement déchiffrable ? Décoder 1101011000101111. Ce code est-il plus efficace qu'un code binaire de longueur constante le plus efficace possible ? **Le code est préfixe, donc uniquement déchiffrable, et le décodage donne FACADE. La longueur moyenne est de** 2.3**, ce qui est inférieur à** 3 **(longueur moyenne du code de longueur constante** 3**), donc il est effectivement plus efficace qu'un tel code.**
4. On souhaite coder S (de manière uniquement déchiffrable) avec des mots ayant tous une longueur d'au plus 3.
   (a) un mot d'un tel code peut-il avoir longueur 1 ? **On vérifie aisément qu'un assortiment de longueurs, contenant une longueur 1 aucune plus grande**

que 3, aura au minimum une fonction de Kraft de $1/2 + 5/8 > 1$; par conséquent un tel code ne peut pas être uniquement déchiffrable.

(b) en déduire l'assortiment des longueurs que doit avoir un tel code pour être le plus efficace possible. Donner un tel code ; est-il plus efficace que celui de Bob ? **Puisque l'on ne peut pas avoir de longueur 1, il faut faire avec des longueurs 2 et 3. L'assortiment $(2, 2, 3, 3, 3, 3)$ est celui qui utilise le plus de 2 sans dépasser la limite de 1 pour la fonction de Kraft, c'est donc lui qui donne la meilleure efficacité. Le plus efficace est alors atteint en codant avec longueur 2 les deux symboles les plus probables, par exemple $C(a) = 00$, $C(e) = 01$, $C(b) = 100$, $C(c) = 101$, $C(d) = 110$, $C(f) = 111$. La longueur moyenne est de 2.3, donc l'efficacité est la même que celle du code de Bob.**

5. Construire un code de Huffmann pour S, et donner son efficacité. **Une possibilité est $C(a) = 0$, $C(b) = 1100$, $C(c) = 1110$, $C(d) = 1101$, $C(e) = 10$, $C(f) = 1111$, pour une longueur moyenne de 2.2 et une efficacité de 0.972.**

Exercice 3

On rappelle qu'un mot v est un *suffixe* d'un mot w, s'il existe un mot u tel que l'on ait $w = u.v$ (le produit étant un produit de concaténation). On note $\tilde{u}$ le *miroir* d'un mot u, c'est-à-dire le mot u "lu à l'envers" : si $u = u_1 u_2 \dots u_\ell$ (où les u_i sont des lettres de l'alphabet), $\tilde{u} = u_\ell u_{\ell-1} \dots u_1$.

Un code est dit *suffixe* si, parmi les mots du code, il n'en existe pas deux codant des lettres distinctes et dont l'un soit un suffixe de l'autre.

1. Rappeler la définition d'un code préfixe. **Question de cours pure : un code est préfixe si aucun mot de code n'est un préfixe d'un autre mot de code.**
2. Pour tout code C, on note $\tilde{C}$ le code formé des miroirs des mots de code de C. Montrer que $\tilde{C}$ est un code suffixe si et seulement si C est un code préfixe. **On commence par montrer qu'un mot v est suffixe de u, si et seulement si $\tilde{v}$ est préfixe de $\tilde{u}$. En effet, si $u = w.v$, on a bien $\tilde{u} = \tilde{v}.\tilde{w}$, et réciproquement, si $\tilde{u} = \tilde{v}.w'$, on a $u = \tilde{w'}.v$. À partir de là, il existe deux mots de code suffixes l'un de l'autre dans $\tilde{C}$, si et seulement si il en existe deux (leurs miroirs, en l'occurrence) qui sont préfixes l'un de l'autre dans C. Autrement dit, $\tilde{C}$ est non suffixe si et seulement si C est non préfixe, ce qui termine la preuve.**
3. Une source S admet-elle toujours un code suffixe qui soit optimal (parmi tous les codes uniquement déchiffrables) ? **Étant donné un code préfixe C, le code suffixe $\tilde{C}$ code les lettres avec les mêmes longueurs, donc a la même longueur moyenne et la même efficacité que C. On ne peut toutefois pas affirmer de but en blanc que le miroir d'un code préfixe optimal (dont on sait qu'il en existe) est un code suffixe optimal ; il faut argumenter que, si C est un**

code préfixe optimal, $\tilde{C}$ est un code suffixe, de même efficacité, et qu'il est donc optimal (car aucun code uniquement déchiffrable ne peut avoir une longueur moyenne plus petite que celle de C). Donc il est bien vrai qu'une source admet toujours un code suffixe optimal.

4. Donner une raison pour laquelle les codes suffixes sont moins utilisés que les codes préfixes (on pourra penser au code ternaire défini par $C(a) = 0$, $C(b) = 01$, $C(c) = 11$ et au décodage des mots 01^{1000} et 01^{1001}). **La raison pratique pour favoriser les codes préfixes par rapport aux codes suffixes est tout simplement une question de délai : avec un code préfixe, on peut toujours identifier les k premières lettres du mot codé dès que l'on a lu leur codage ; avec un code suffixe, cela n'est pas vrai. Avec le code suffixe proposé, le mot 01^k se décode en $bc^{(k-1)/2}$ si k est impair, et en $ac^{k/2}$ si k est pair : il est impossible d'identifier la première lettre du mot codé (a ou b) avant d'avoir lu l'intégralité du mot.**