Solutions des Exercices du cours de Théorie de l'Information et Codage
cours 3 du 1er mars 2011.

1. Montrer le lemme énoncé sans démonstration en cours: $c(u_1^n) = O(n/\log n)$.

   - En reprenant les notations du cours, soit $c' = c/J^3$ et $n' = n/J^3$. On a vu en cours que $n' > c' \log_J c'$. Pour n suffisament grand tel que $\sqrt{n'} \leq 2\frac{n'}{\log_J n'}$, on a
     - soit $c' < \sqrt{n'}$ et donc $c' < 2n'/\log_J n'$.
     - soit $c' \geq \sqrt{n'}$ et alors $c' < \frac{n'}{\log_J c'} \leq \frac{n'}{\log_J \sqrt{n'}} = \frac{2n'}{\log_J n'}$.

2. Montrer que pour toute v.a. X à valeurs entières de moyenne $\mathbb{E}[X]$, on a:

$$H(X) \leq (\mathbb{E}[X]+1)\log(\mathbb{E}[X]+1) - \mathbb{E}[X]\log\mathbb{E}[X],$$

   avec égalité quand X suit une loi géométrique: $\mathbb{P}(X = k) = q^k(1-q)$ pour $k \geq 0$.

   - pour Y de loi géométrique $\mathbb{P}(Y = k) = q^k(1-q)$ pour $k \geq 0$, on a $\mathbb{E}[Y] = \frac{q}{1-q}$ et :

$$\begin{aligned} H(Y) &= -\sum_k q^k(1-q)\log q^k(1-q) \\ &= -\log(1-q) - \mathbb{E}[Y]\log q, \end{aligned}$$

     on a donc bien égalité dans ce cas.
   - Soit X une v.a. discrète de loi p_k et Y de loi géométrique de même moyenne q_k. On a alors

$$\begin{aligned} H(X) &= -\sum p_k \log p_k \\ &= -\sum p_k \log\frac{p_k}{q_k} - \sum p_k \log q_k \\ &= D(p||q) - \sum p_k \log q^k(1-q) \\ &\leq -\log(1-q) - \sum k p_k \log q = H(Y), \end{aligned}$$

     où l'inégalité vient de $D(p||q) \geq 0$.

3. Soit $\{U_i\}_{i=1}^\infty$ une suite de v.a. i.i.d. à valeurs dans un ensemble fini $\mathcal{U}$ et d'entropie $H(U)$. On note $C(n)$ la v.a. égale au nombre maximal de mots distincts en lequel $U^{(n)}$ peut être découpé. C'est à dire avec les notations du cours: $C(n) = c(U_1^n)$. Nous allons montrer qu'avec probabilité 1, on a:

$$\limsup_{n\to\infty} \frac{C(n)\log C(n)}{n} \leq H(U).$$

   Ce résultat permet de démontrer l'optimalité de l'algorithme de Lempel-Ziv dans le cas particulier d'une source sans mémoire.

a) Pour un découpage de u_1^n en c mots distincts, on note c_ℓ le nombre de mots de longueur ℓ. Montrer que

$$\log \mathbb{P}(u_1, u_2, \dots, u_n) \leq -\sum_\ell c_\ell \log c_\ell.$$

b) On définit la v.a. Z par $\mathbb{P}(Z = \ell) = \frac{c_\ell}{c}$. En utilisant les deux exercices précédents, montrer que

$$\lim_{n\to\infty} \frac{c}{n} H(Z) = 0$$

c) Conclure.

a) On note le découpage: $u_1 u_2 \dots u_n = w_1 w_2 \dots w_c$. On a alors:

$$\begin{aligned}
\log \mathbb{P}(u_1, u_2, \dots, u_n) &= \sum_{i=1}^{c} \log \mathbb{P}(w_i) \\
&= \sum_\ell \sum_{i,|w_i|=\ell} \log \mathbb{P}(w_i) \\
&\leq \sum c_\ell \log \left(\sum_{i,|w_i|=\ell} \frac{\mathbb{P}(w_i)}{c_\ell} \right) \\
&\leq -\sum_\ell c_\ell \log c_\ell,
\end{aligned}$$

où la première inégalité provient de la concavité du log et la seconde du fait que le w_i étants distincs $\sum_i \mathbb{P}(w_i) \leq 1$.

b) Z est une v.a. de moyenne $\mathbb{E}[Z] = \frac{\sum_\ell \ell c_\ell}{c} = \frac{n}{c}$. On a donc:

$$\begin{aligned}
H(Z) &\leq \left(\frac{n}{c} + 1\right) \log \left(\frac{n}{c} + 1\right) - \frac{n}{c} \log \frac{n}{c} \\
&= \log \left(\frac{n}{c} + 1\right) + \frac{n}{c} \log \left(\frac{c}{n} + 1\right).
\end{aligned}$$

Donc

$$\frac{c}{n} H(Z) \leq \frac{c}{n} \log \left(\frac{n}{c} + 1\right) + \log \left(\frac{c}{n} + 1\right),$$

le résultat découle de $c = O(n/\log n)$.

c) Pour tout $u_1, \dots, u_n$, et tout découpage en c mots distincst, on a

$$\begin{aligned}
\log \mathbb{P}(u_1, \dots, u_n) &\leq -\sum_\ell c_\ell \log c_\ell \\
&= -c \log c - c \sum_\ell \frac{c_\ell}{c} \log \frac{c_\ell}{c} \\
&= -c \log c - c H(Z).
\end{aligned}$$

On a donc avec probabilité 1:

$$-\frac{1}{n}\log \mathbb{P}(U_1,\ldots,U_n) \quad \geq \quad \frac{C(n)}{n}\log C(n) - \frac{C(n)}{n}H(Z),$$

et on obtient le résultat désiré en prenant la limite $n \to \infty$.