École Normale Supérieure — Traitement du signal
16 avril 2015

# Feuille d'exercices n°9

## Corrigé

**Exercice 1**

Soit $\eta_1, ..., \eta_n$ une suite de zéros et de uns. Notons $k$ le nombre de uns qu'elle contient. On a :

$$P(X_1 = \eta_1, ..., X_n = \eta_n) = p^{n-k}(1-p)^k = p^n \left(\frac{1-p}{p}\right)^k$$

La suite $(\eta_1, ..., \eta_n)$ est $\epsilon$-typique si :

$$H(X) - \epsilon \leq -\frac{\log_2(P(X_1 = \eta_1, ..., X_n = \eta_n))}{n} \leq H(X) + \epsilon$$

L'entropie de $X$ vaut $H(X) = -p\log_2(p) - (1-p)\log_2(1-p)$. Puisque $\frac{\log_2(P(X_1=\eta_1,...,X_n=\eta_n))}{n} = \log_2(p) + \frac{k}{n}\log_2\left(\frac{1-p}{p}\right)$, la suite $(\eta_1, ..., \eta_n)$ est $\epsilon$-typique si et seulement si :

$$(1-p)\log_2\left(\frac{1-p}{p}\right) - \epsilon \leq \frac{k}{n}\log_2\left(\frac{1-p}{p}\right) \leq (1-p)\log_2\left(\frac{1-p}{p}\right) + \epsilon$$

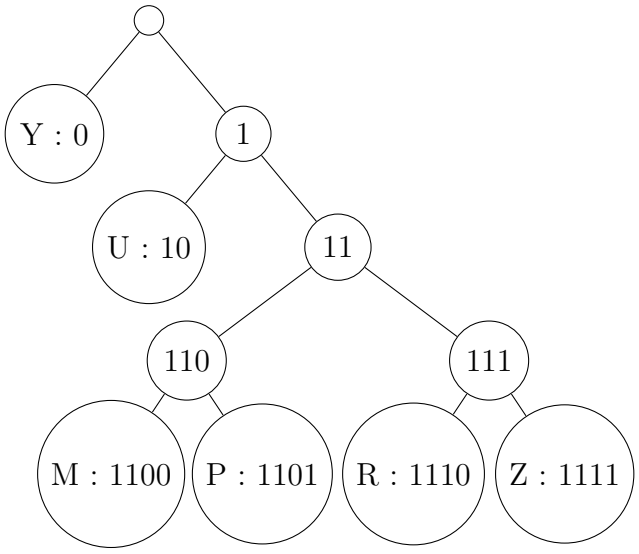
$$\Leftrightarrow \quad n\left((1-p) - \epsilon\left|\log_2\left(\frac{1-p}{p}\right)\right|^{-1}\right) \leq k \leq n\left((1-p) + \epsilon\left|\log_2\left(\frac{1-p}{p}\right)\right|^{-1}\right)$$

En effet, on a pu diviser par $\log_2\left(\frac{1-p}{p}\right)$ car, comme $p \neq 1/2$, ce nombre est non-nul.
Le nombre de uns dans une suite $\epsilon$-typique est donc de l'ordre de $n(1-p)$.

**Exercice 2**

1. $H(X) = -\left(\frac{1}{16}\log_2\left(\frac{1}{16}\right) + \frac{1}{16}\log_2\left(\frac{1}{16}\right) + \frac{1}{16}\log_2\left(\frac{1}{16}\right) + \frac{1}{4}\log_2\left(\frac{1}{4}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{1}{16}\log_2\left(\frac{1}{16}\right)\right)$ donc $H(X) = 2$.

2. a)




b) Toutes les lettres de probabilité 1/16 sont codées par 4 bits, celle de probabilité 1/4 par 2 bits et celle de probabilité 1/2 par 1 bits. La longueur moyenne est donc :

$$\frac{1}{16}.4.4 + \frac{1}{4}.2 + \frac{1}{2}.1 = 2 = H(X)$$

3. On procède par récurrence sur $n$. Pour $n = 0$, c'est vrai. Supposons qu'on l'a montré jusqu'à $n$ et montrons-le pour $n + 1$.

- Si $\epsilon_1 = 0$ : par l'hypothèse de récurrence, il existe un mot $m$ dont le code commence par $\epsilon_2...\epsilon_n$. Alors le code de $Ym$ commence par $\epsilon_1...\epsilon_n$.
- Si $\epsilon_1 = 1, \epsilon_2 = 0$ : soit $m$ un mot dont le code commence par $\epsilon_3...\epsilon_n$. Alors le code de $Um$ commence par $\epsilon_1...\epsilon_n$.
- Si $\epsilon_1 = \epsilon_2 = 1$ : il existe une lettre $l$ dont le code est $\epsilon_1\epsilon_2\epsilon_3\epsilon_4$. Soit $m$ un mot dont le code commence par $\epsilon_5...\epsilon_n$. Alors le code de $lm$ commence par $\epsilon_1...\epsilon_n$.

4. a) 0101100
b) Si on change le premier bit, on obtient 1101100. Ceci est le code du mot $PUY$.

**Exercice 3**

1. On ordonne les éléments de $\mathcal{E}$ par longueur de code croissante : $\pi(1)$ est l'élément de $\mathcal{E}$ tel que $l(\phi(\pi(1)))$ est minimale, $\pi(2)$ est l'élément de $\mathcal{E}$ tel que $l(\phi(\pi(2)))$ est la plus petite possible après $l(\phi(\pi(1)))$ etc.
Pour tout $k$, $l(\phi(\pi(k))) \geq [\log_2(k)]$. En effet, sinon, $l(\phi(\pi(k))) \leq [\log_2(k)] - 1$ et, puisque $s \to l(\phi(\pi(s)))$ est croissante, on a, pour tout $s \in \{1, ..., k\}$ :

$$l(\phi(\pi(s))) \leq [\log_2(k)] - 1$$

Le nombre d'éléments de $\mathcal{B}$ dont la longueur est inférieure ou égale à $[\log_2(k)]-1$ est $2^{[\log_2(k)]}-1$. Ce nombre est inférieur ou égal à $2^{\log_2(k)} - 1 = k-1$. C'est absurde car $\phi(\pi(1)), ..., \phi(\pi(k))$ sont $k$ éléments de $\mathcal{B}$ distincts dont la longueur est inférieure ou égale à $[\log_2(k)] - 1$.

Ainsi :

$$\begin{aligned}\mathbb{E}(l(\phi(X))) &= \sum_{k=1}^{n} P(X=\pi(k)) l(\phi(\pi(k))) \\ &\geq \sum_{k=1}^{n} P(X=\pi(k))[\log_2(k)]\end{aligned}$$

2. D'après la question précédente, il suffit de montrer :

$$\begin{aligned}\sum_{k=1}^{n} P(X=\pi(k))[\log_2(k)] &\geq H(X) - 1 - \log_2(1+\ln(n)) \\ &= -\sum_{k=1}^{n} P(X=\pi(k)) \log_2(P(X=\pi(k))) - 1 - \log_2(1+\ln(n))\end{aligned}$$

On a :

$$\begin{aligned}\sum_{k=1}^{n} P(X=\pi(k))[\log_2(k)] &\geq \sum_{k=1}^{n} P(X=\pi(k)) \left(\log_2(k) - 1\right) \\ &= \sum_{k=1}^{n} P(X=\pi(k)) \log_2(k) - \sum_{k=1}^{n} P(X=\pi(k)) \\ &= \sum_{k=1}^{n} P(X=\pi(k)) \log_2(k) - 1\end{aligned}$$

Il suffit donc de démontrer :

$$\sum_{k=1}^{n} P(X=\pi(k)) \log_2(k) \geq -\sum_{k=1}^{n} P(X=\pi(k)) \log_2(P(X=\pi(k))) - \log_2(1+\ln(n))$$

soit :

$$\log_2(1+\ln(n)) \geq \sum_{k=1}^{n} P(X=\pi(k)) \log_2\left(\frac{1}{kP(X=\pi(k))}\right)$$

Or, d'après la concavité de la fonction $\log_2$ et puisque $\sum_{k=1}^{n} P(X=\pi(k)) = 1$ :

$$\begin{aligned}\sum_{k=1}^{n} P(X=\pi(k)) \log_2\left(\frac{1}{kP(X=\pi(k))}\right) &\leq \log_2\left(\sum_{k=1}^{n} P(X=\pi(k)) \frac{1}{kP(X=\pi(k))}\right) \\ &= \log_2\left(\sum_{k=1}^{n} \frac{1}{k}\right) \\ &\leq \log_2(1+\ln(n))\end{aligned}$$

3. Puisque les variables aléatoires $X_1, X_2, ...$ sont indépendantes les unes des autres et de même loi :

$$\forall k \in \mathbb{N}^*, \quad H((X_1, ..., X_k)) = kH(X_1)$$

Si on note $N = n^k$ le nombre d'éléments de $\mathcal{E}^k$, on a, d'après la question précédente :

$$\mathbb{E}(l(\phi_k(X_1, ..., X_k))) \geq H((X_1, ..., X_k)) - 1 - \log_2(1 + \ln(N))$$

donc :

$$\begin{aligned}\frac{1}{k}\mathbb{E}(l(\phi_k(X_1, ..., X_k))) &\geq \frac{1}{k}\big(H((X_1, ..., X_k)) - 1 - \log_2(1 + \ln(N))\big) \\ &= H(X_1) - \frac{1}{k} - \frac{1}{k}\log_2(1 + k\ln(n))\end{aligned}$$

Puisque $\frac{1}{k} + \frac{1}{k}\log_2(1 + k\ln(n)) \to 0$ quand $k \to +\infty$, l'égalité suivante est vraie pour tout $k$ assez grand :

$$\frac{1}{k}\mathbb{E}(l(\phi_k(X_1, ..., X_k))) \geq H(X_1) - \epsilon$$

**Exercice 4**

1. Soit $r$ le réel codant le mot de $n$ lettres.
On décode par récurrence : si $n = 0$, on renvoie le mot vide.
Supposons qu'on a déjà décodé les $m$ premières lettres, avec $m < n$. Alors on peut calculer $y_k^{\inf}$ et $y_k^{\sup}$ pour tout $k \leq m$. La $m+1$-ème lettre est l'unique $x_{m+1}$ tel que $r \in [y_m^{\inf} + a_{x_{m+1}-1}(y_m^{\sup} - y_m^{\inf}); y_m^{\inf} + a_{x_{m+1}}(y_m^{\sup} - y_m^{\inf})]$.
2. Pour toute suite de lettres $x_1...x_n$, on note $[y_n^{\inf}(x_1...x_n), y_n^{\sup}(x_1...x_n)[$ l'intervalle correspondant.
On vérifie par récurrence que sa largeur est exactement $p_{x_1}...p_{x_n}$. Cet intervalle contient donc un réel de la forme $M2^{-m}$ où $M$ est un entier et $m = \lceil -\log_2(p_{x_1}...p_{x_n})\rceil$.
Puisqu'un réel compris entre 0 et 1 de la forme $M2^{-m}$ se code en $m$ bits, la longueur du code de $x_1...x_n$ est au plus :

$$\lceil -\log_2(p_{x_1}...p_{x_n})\rceil \leq 1 - \sum_{s\leq n}\log_2(p_{x_s})$$

On obtient :

$$\begin{aligned}l_n &\leq \sum_{x_1,...,x_n} p(x_1)...p(x_n)\left(1 - \sum_{s\leq n}\log_2(p_{x_s})\right) \\ &= 1 - \sum_{s\leq n}\sum_{x_s} p_{x_s}\log_2(p_{x_s}) \\ &= 1 + nH(X)\end{aligned}$$

donc $l_n/n \leq H(X) + 1/n$.
3. a) $H(X) = -(1-\epsilon)\log_2(1-\epsilon) - \epsilon\log_2\epsilon$
b) 0 est codé par 0 et 1 est codé par 1.
c) Un mot de $n$ lettres est toujours codé par un code de longueur $n$. On a donc $l_n/n = 1$. Pourtant, si $\epsilon$ est proche de 0 ou de 1, $H(X) \ll 1$ (car la fonction $x \to x\log_2(x)$ tend vers 0 en $x = 0$ et en $x = 1$).

**Exercice 5**

1.

$$\begin{aligned} H(X) &= -\sum_{k\geq 1} 2^{-k} \log_2\left(2^{-k}\right) \\ &= \sum_{k\geq 1} k2^{-k} \\ &= \sum_{k\geq 1}\sum_{s\geq k} 2^{-s} \\ &= \sum_{k\geq 1} 2^{-k+1} \\ &= \sum_{k\geq 0} 2^{-k} = 2 \end{aligned}$$

2. On code le nombre $k$ par le mot 1...10, qui contient $(k-1)$ zéros (c'est-à-dire que 1 est codé par 0, 2 par 10, 3 par 110 etc.).
Le nombre moyen de bits pour un symbole est (puisque le symbole $k$ est codé par $k$ bits) :

$$\sum_k P(X=k).k = \sum_k k2^{-k} = H(X)$$

Donc le code est optimal.

3. Dans ce cas, les symboles $1, ..., N$ voient leur code entier envoyé. En revanche, les codes des symboles $N+1, N+2, ...$ sont tronqués : pour un tel symbole, on n'envoie que $x = 1...1$ ($N$ fois le bit 1). On suppose que lorsque le $N$-uplet reçu est $x$, on le décode par $N+1$.
Dans ce cas, le nombre de bits moyen est :

$$\begin{aligned} \sum_{k\geq 1} P(X=k)\min(k,N) &= \sum_{k\geq 1} \min(k,N)2^{-k} \\ &= \sum_{k=1}^{N}\sum_{s\geq k} 2^{-s} \\ &= \sum_{k=1}^{N} 2^{-(k-1)} \\ &= \sum_{k=0}^{N-1} 2^{-k} \\ &= 2.(1-2^{-N}) \end{aligned}$$

La distance entre le symbole $k$ à transmettre et le symbole décodé est 0 si $k \leq N+1$ et

$k-(N+1)$ si $k > N+1$. L'erreur quadratique moyenne est donc :

$$\begin{aligned}
\sum_{k\geq N+1} P(X=k)(k-(N+1))^2 &= \sum_{k\geq N+1} 2^{-k}(k-(N+1))^2 \\
&= 2^{-(N+1)} \sum_{k\geq 0} k^2 2^{-k} \\
&= 2^{-(N+1)} \sum_{k\geq 0} (1+3+5+...+(2k-1))2^{-k} \\
&= 2^{-(N+1)} \sum_{k\geq 1} (2k-1) \left( \sum_{s\geq k} 2^{-s} \right) \\
&= 2^{-(N+1)} \sum_{k\geq 1} (2k-1) 2^{-(k-1)} \\
&= 2^{-(N+1)} \left( 2. \sum_{k\geq 0} k 2^{-k} + \sum_{k\geq 0} 2^{-k} \right) \\
&= 2^{-(N+1)} (2.2+2) = 3.2^{-N}
\end{aligned}$$