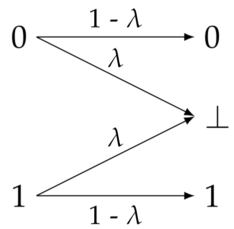
Théorie de l'information
Interrogation – solutions

14/12/2022

Exercice 1. Questions de cours : canal à effacement.

Cet exercice est un exercice de cours sur le canal à effacement de paramètre λ. Rappelons le diagramme qui le définit :



Question 1.– Donner la matrice de transition du canal à effacement.

Question 2.– Rappeler la définition de la capacité d'un canal, en nommant tous les termes engagés dans la formule.

Question 3.– Démontrer que la capacité du canal à effacement est $1-\lambda$.

Solutions de l'Exercice 1.

Cet exercice correpond à des questions déjà traitées en cours ou en TD. Si besoin de détails, s'y référer.

Solution Q1. La matrice est :

$$\begin{pmatrix} 1-\lambda & 0 \\ \lambda & \lambda \\ 0 & 1-\lambda \end{pmatrix}$$

Solution Q2. La capacité d'un canal est :

$$\max_X I(X\,;Y)$$

où X est l'entrée du canal, Y sa sortie, et $I(X\,;Y)$ l'information mutuelle entre les variables X et Y.

Solution Q3. Voici une solution détaillée (on peut aller plus vite si besoin). On utilise la formule du cours : $I(X\,;Y) = H(Y) - H(Y \mid X)$, et on calcule ces deux termes.

- Calcul de $H(Y \mid X)$. On utilise la formule générique :

$$H(Y\,|\,X) = \sum_x p(x) \underbrace{\sum_y p(y\,|\,x)\log_2 \frac{1}{p(y\,|\,x)}}_{H(Y|X=x)}$$

  Calculons d'abord $H(Y\,|\,X=0)$.

$$\begin{aligned} H(Y\,|\,X=0) &= -p_{Y|X}(0\,|\,0)\log_2 p_{Y|X}(0\,|\,0) - p_{Y|X}(\perp|\,0)\log_2 p_{Y|X}(\perp|\,0) - p_{Y|X}(1\,|\,0)\log_2 p_{Y|X}(1\,|\,0) \\ &= (1-\lambda)\log_2 \frac{1}{1-\lambda} + \lambda \log_2 \frac{1}{\lambda} + 0 = h(\lambda)\,. \end{aligned}$$

De même $H(Y \mid X=1) = h(\lambda)$. Par conséquent,

$$H(Y \mid X) = \sum_x p(x) \cdot h(\lambda) = 1 \cdot h(\lambda) = h(\lambda)\,.$$

- Calcul de $H(Y)$. Pour cela, il faut calculer $p(Y=0)$, $p(Y=\perp)$ et $p(Y=1)$:

$$\begin{aligned}
p(Y=0) &= \sum_x p(X=x)p(Y=0 \mid X=x) = \alpha(1-\lambda) \\
p(Y=\perp) &= \lambda\alpha + \lambda(1-\alpha) = \lambda \\
p(Y=1) &= (1-\lambda)(1-\alpha)
\end{aligned}$$

Puis :

$$\begin{aligned}
H(Y) &= \left(\alpha(1-\lambda)\log_2 \frac{1}{\alpha(1-\lambda)}\right) + \left(\lambda \log_2 \frac{1}{\lambda}\right) + \left((1-\alpha)(1-\lambda)\log_2 \frac{1}{(1-\alpha)(1-\lambda)}\right) \\
&= \left(\lambda \log_2 \frac{1}{\lambda}\right) + \left((\alpha + (1-\alpha))(1-\lambda)\log_2 \frac{1}{(1-\lambda)}\right) \\
&\qquad + \left(\alpha(1-\lambda)\log_2 \frac{1}{\alpha} + (1-\alpha)(1-\lambda)\log_2 \frac{1}{1-\alpha}\right) \\
&= h(\lambda) + (1-\lambda)h(\alpha)
\end{aligned}$$

Enfin, on doit maximiser $I(X\,;Y) = H(Y) - H(Y \mid X)$ en fonction de α. On a :

$$I(X\,;Y) = h(\lambda) + (1-\lambda)h(\alpha) - h(\lambda) = (1-\lambda)h(\alpha)\,.$$

On sait que la fonction $\alpha \mapsto h(\alpha)$ est maximale pour $\alpha = 1/2$, et vaut 1 dans ce cas. Donc la capacité du canal binaire à effacement est :

$$1-\lambda\,.$$

Exercice 2. Questions diverses.

Question 1.– Le code $\{10, 111, 11111, 0, 110\}$ est-il uniquement décodable ? Justifier.

Question 2.– Quelle est la longueur moyenne du code de Shannon–Fano sur une source X dont la distribution de probablité est :

$$p_X = \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{16}, \frac{1}{16}, \frac{1}{32}, \frac{1}{32}, \frac{1}{32}, \frac{1}{32}\right) \ ?$$

Question 3.– Soit X une variable aléatoire à valeurs dans $\mathbb{R}$.

1. Comparer $H(X^3+1)$ et $H(X)$.
2. **[plus difficile]** Démontrer que $H(X \mid X^2) \leq 1$ et donner un cas d'égalité.

Solutions de l'Exercice 2.

Solution Q1. Vérifions si le code satisfait l'inégalité de Kraft : on a

$$2^{-2} + 2^{-3} + 2^{-5} + 2^{-1} + 2^{-3} = \frac{33}{32} > 1$$

Il ne la satisfait pas, donc le code n'est pas uniquement décodable.

Solution Q2. Le code de Shannon–Fano construit sur la source X de distribution $(p_i)_i$ sera constitué de mots de longueurs $n_i = \lceil -\log_2(p_i) \rceil$, c'est-à-dire ici respectivement :

$$(1, 2, 4, 4, 5, 5, 5, 5)$$

La longueur moyenne de ce code est donc :

$$\sum_i n_i p_i = 1 \times \frac{1}{2} + 2 \times \frac{1}{4} + 2 \times 4 \times \frac{1}{16} + 4 \times 5 \times \frac{1}{32} = \frac{17}{8}\,.$$

Solution Q3.

1. Soit $g : x \mapsto x^3 + 1$. D'après le principe de non-création d'information, on a
$$H(X) \geq H(g(X)) = H(X^3 + 1)\,.$$
2. Pour montrer que $H(X \mid X^2) \leq 1$, on peut se poser la question suivante : quelle (ou combien d') information manque-t-il pour déterminer un réel x à partir de son carré x^2 ? On peut se rendre compte que la donnée du signe de x est l'unique information manquante : avec $y = x^2$, on peut obtenir $\pm x = \sqrt{y}$ (ces deux valeurs étant égales pour $x = 0$).

   Maintenant, soyons formels et introduisons la fonction signe $s(x) = \frac{x}{|x|} \in \{-1, 1\}$ avec comme convention $s(0) = 0$. La fonction signe prend au plus deux valeurs, donc $H(s(X)) \leq \log_2(2) = 1$. Il reste à montrer que :
$$H(X \mid X^2) \leq H(s(X))$$
   Pour cela, notons que la discussion précédente affirme que
$$f : x \mapsto (x^2, s(x))$$
   est une application bijective. Ainsi, par le principe de non-création d'information on a :
$$H(X) = H(f^{-1}(f(X))) \leq H(f(X)) = H(X^2, s(X)) \leq H(X^2) + H(s(X))$$
   ce qui donne, grâce à la règle de chaînage,
$$H(X \mid X^2) = H(X, X^2) - H(X^2) = H(X) - H(X^2) \leq H(s(X)) \leq 1\,.$$
   Un cas d'égalité est donné, par exemple, par une variable X uniforme sur $\{-1, 1\}$. Dans ce cas, X^2 est déterministe égale à 1, donc $H(X \mid X^2) = H(X) - H(X^2) = 1 - 0 = 1$.

Exercice 3. Code de Huffman d'une source équilibrée.

Question 1.– On considère une source suivant une loi dont la distribution est :
$$(0.30, 0.27, 0.23, 0.20)\,.$$
Donner le code de Huffman correspondant. On précisera les étapes de la construction du code.

Considérons maintenant un cas plus général où X est une source sur un alphabet de taille $m = 2^k$, et dont la distribution satisfait $p_1 \geq \cdots \geq p_m$ et
$$p_{m-1} + p_m \geq p_1\,.$$

Question 2.– Décrire l'état de la distribution « temporaire » et des arbres « temporaires » obtenus après $m/2$ étapes de l'algorithme de Huffman.

Question 3.– Démontrer que $p_m + p_{m-1} + p_{m-2} + p_{m-3} \geq p_1 + p_2$. En déduire l'état de la distribution et des arbres « temporaires » après $3m/4$ étapes de l'algorithme de Huffman.

Question 4.– Conclure sur la forme de l'arbre binaire retourné à la fin de l'exécution de l'algorithme de Huffman.

Question 5.– En déduire que $H(X) > k - 1$ où $k = \log_2(m)$.

Solutions de l'Exercice 3.

Solution Q1. On obtient la séquence d'arbres suivante :

symbole	a	b	c	d
proba.	0.3	0.27	0.23	0.2
arbre	a	b	c	d

symbole	a	b	c ou d
proba.	0.3	0.27	0.43
arbre	a	b	c d

symbole	a ou b	c ou d
proba.	0.57	0.43
arbre	a b	c d

symbole	a ou b ou c ou d
proba.	1.00
arbre	a b c d

Solution Q2. La première étape de l'algorithme de Huffman sélectionne les symboles de probabilité p_{m-1} et p_m, et les rassemble dans un arbre.

À la deuxième étape, comme $p_{m-1} + p_m \geq p_1$, les symboles dont la probabilité est la plus petite sont p_{m-3} et p_{m-2}. Par ailleurs, comme $p_{m-3} \geq p_{m-1}$ et $p_{m-2} \geq p_m$, on a

$$p_{m-3} + p_{m-2} \geq p_{m-1} + p_m \geq p_1 .$$

En suivant le procédé, au bout de $m/2$ étapes, l'algorithme aura sélectionné tous les symboles par paires de probabilités (p_{2i-1}, p_{2i}) pour $1 \leq i \leq m/2$, et aura créé les arbres suivants :

symbole	x_1 ou x_2	x_3 ou x_4	$\cdots$	x_{m-1} ou x_m
proba.	$p_1 + p_2$	$p_3 + p_4$	$\cdots$	$p_{m-1} + p_m$
arbre	p_1 p_2	p_3 p_4	$\cdots$	p_{m-1} p_m

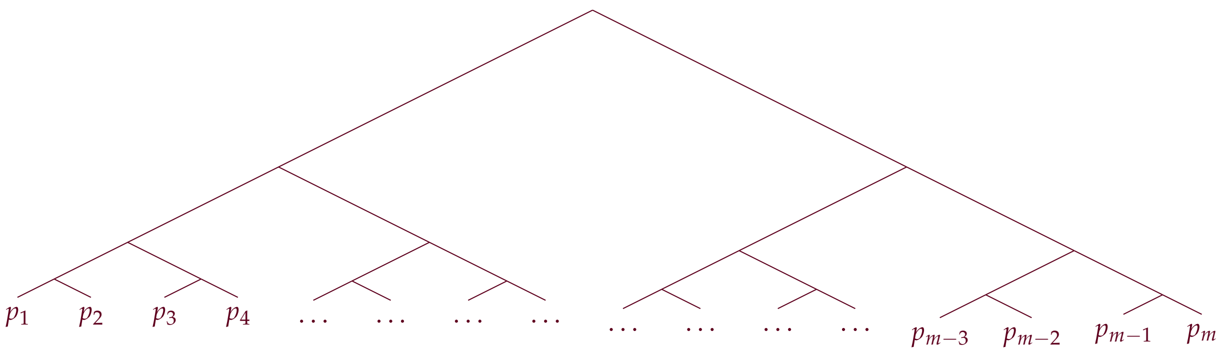
Solution Q3. Posons $p'_i = p_{2i-1} + p_{2i}$ pour tout $1 \leq i \leq m/2$. On observe que les probabilités satisfont maintenant $p'_1 \geq p'_2 \geq \cdots \geq p'_{m/2}$. Par ailleurs, grâce à un calcul entamé dans la question précédente, on a

$$p'_{m-1} + p'_{m/2} = p_{m-3} + p_{m-2} + p_{m-1} + p_m \geq p_1 + p_1 \geq p_1 + p_2 .$$

On peut donc appliquer **exactement** le même raisonnement qu'à la question précédente sur la ditribution $\boldsymbol{p}' = (p'_1, \ldots, p'_{m/2})$. Après $m/2 + m/4 = 3m/4$ itérations, l'algorithme de Huffman produit donc les arbres suivants :

symbole	x_1 ou x_2 ou x_3 ou x_4	x_5 ou x_6 ou x_7 ou x_8	$\cdots$	x_{m-3} ou x_{m-2} ou x_{m-1} ou x_m
proba.	$p_1 + p_2 + p_3 + p_4$	$p_5 + p_6 + p_7 + p_8$	$\cdots$	$p_{m-3} + p_{m-2} + p_{m-1} + p_m$
arbre	p_1 p_2 p_3 p_4	p_5 p_6 p_7 p_8	$\cdots$	p_{m-3} p_{m-2} p_{m-1} p_m

Solution Q4. On itère le raisonnement des deux questions précédentes pour $m - 1 = m/2 + m/4 + m/8 + \cdots + m/m$ étapes. On en déduit que l'arbre final est de la forme

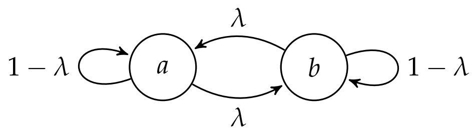


Solution Q5. [Il y avait une erreur dans l'énoncé originel : on montre bien que $H(X) > k - 1$ et non $H(X) > m - 1$] Tous les mots du code de Huffman ont longueur k. La longueur moyenne du code est donc k. D'après le cours, on sait que la longueur moyenne du code de Huffman est compris strictement plus petite que $H(X) + 1$. Par conséquent,

$$H(X) > k - 1 .$$

Exercice 4. Codage par plage d'un processus stochastique.

Étant donné un paramètre $\lambda \in [0, 1]$, on considère le graphe orienté suivant.



On fixe une variable aléatoire X_0 sur $\{a, b\}$, correspondant à la position initiale sur le graphe. À chaque pas de temps, cette position est modifiée selon les probabilités de transition indiquées sur les arêtes du graphe. On note X_n la position après n pas de temps, et $\boldsymbol{X} = (X_n)_{n\in\mathbb{N}}$ le processus associé.

Question 1.– Le processus $\boldsymbol{X}$ est-il sans mémoire ? Est-il markovien ? Est-il homogène ? Justifier rapidement chaque réponse.

Question 2.– Sous quelle(s) condition(s) le processus $\boldsymbol{X}$ est-il stationnaire ? Quel est le taux d'entropie associé ?

On considère maintenant le codage RLE (codage par plage) de la source $\boldsymbol{X}$. Ce codage forme une suite de couples (n_i, c_i) où n_i est un entier strictement positif et $c_i \in \{a, b\}$ est le caractère de la plage. On note ensuite N_i la variable aléatoire décrivant n_i, et C_i celle décrivant c_i. Enfin, on définit les processus $\boldsymbol{N} = (N_i)_{i\in\mathbb{N}}$ et $\boldsymbol{C} = (C_i)_{i\in\mathbb{N}}$.

Question 3.– Un exemple de séquence représentant les premières réalisations du processus $\boldsymbol{X}$ est le suivant :

$$aaabbabbbba$$

Donner le codage RLE de cette séquence, en précisant les valeurs prises par les variables N_i et C_i.

On revient au cas général.

Question 4.– Le processus $\boldsymbol{N}$ est-il sans mémoire ? Est-il markovien ? Est-il homogène ? Justifier rapidement chaque réponse.

Question 5.– Déterminer la loi de N_i pour tout $i \geq 0$.

On souhaite maintenant réencoder le codage RLE en binaire (c'est-à-dire, réencoder les couples (n_i, c_i) avec des bits 0 ou 1). Pour cela, on va réencoder séparément les processus $\boldsymbol{C}$ et $\boldsymbol{N}$. On appelle RLE_{bin} ce réencodage binaire.

Question 6.– Expliquer pourquoi il suffit d'un bit pour encoder tout le processus $\boldsymbol{C}$.

On considère un code uniquement décodable $\mathcal{E}$ sur les entiers naturels. On note alors $\mathcal{E}(N_i)$ l'encodage binaire de N_i et $\ell(\mathcal{E}(N_i))$ sa longueur (en nombre de bits). D'après les questions précédentes, on peut donc maintenant supposer que l'encodage binaire des k premiers couples $[(N_0, C_0), (N_1, C_1), \ldots, (N_{k-1}, C_{k-1})]$ du codage RLE est de longueur $1 + \sum_{i=0}^{k-1} \ell(\mathcal{E}(N_i))$.

Question 7.– Donner un encadrement de la longueur moyenne $\overline{\ell}(\mathcal{E}(N_i))$ en fonction d'une grandeur informationnelle liée à N_i.

On **admet** dans la suite que la longueur moyenne de $\mathcal{E}$ atteint la borne inférieure donnée à la question précédente.

Question 8.– [plus difficile] Démontrer que la longueur moyenne par bit de source de $\text{RLE}_{\text{bin}}(\boldsymbol{X})$ est :

$$\lim_{k\to+\infty} \frac{1 + kH(N_0)}{k\mathbb{E}(N_0)}$$

On rappelle (vu en exercice) que la loi géométrique Γ de paramètre γ vérifie $\gamma H(\Gamma) = h_2(\gamma)\mathbb{E}(\Gamma)$.

Question 9.– [plus difficile] En déduire la longueur moyenne par bit de source de $\text{RLE}_{\text{bin}}(\boldsymbol{X})$, puis comparer au taux d'entropie de $\boldsymbol{X}$ calculé à la **Question 2**. Commenter.

Solutions de l'Exercice 4.

Solution Q1. Le processus $\boldsymbol{X}$:

- n'est pas sans mémoire puisque les X_i ne sont généralement pas indépendantes (la position X_i dépend de la position X_{i-1}),
- est markovien car la dépendance de X_i par rapport à son passé $X_{i-1}, \ldots, X_0$ peut être réduite à une dépendance par rapport à X_{i-1} seulement (« X_i ne dépend que de X_{i-1} »),
- homogène car les transitions ne dépendent par de i.

Solution Q2. Considérons la matrice de transition du processus :

$$\boldsymbol{M} = \begin{pmatrix} 1-\lambda & \lambda \\ \lambda & 1-\lambda \end{pmatrix}.$$

Le processus $\boldsymbol{X}$ est stationnaire lorsque la distribution initiale $\boldsymbol{p}$ de X_0 vérifie $\boldsymbol{M}\boldsymbol{p} = \boldsymbol{p}$. Résolvons ce système avec $\boldsymbol{p} = (x, y)$. On obtient :

$$\begin{cases} (1-\lambda)x & + & \lambda y & = & x \\ \lambda x & + & (1-\lambda)y & = & y \end{cases} \iff x = y$$

Comme la distribution est sujette à la contrainte $x + y = 1$, on en déduit que $\boldsymbol{X}$ est stationnaire si et seulement si X_0 est uniforme.

Dans ce cas, le taux d'entropie vaut $H(\boldsymbol{X}) = H(X_1 \mid X_0)$. Un calcul montre alors que

$$\begin{aligned} H(X_1 \mid X_0) &= p(X_0 = a)H(X_1 \mid X_0 = a) + p(X_0 = b)H(X_1 \mid X_0 = b) \\ &= \frac{1}{2}h_2(\lambda) + \frac{1}{2}h_2(\lambda) \\ &= h_2(\lambda)\,. \end{aligned}$$

Solution Q3. Le codage RLE de la séquence est :

$$(3, a)\ (2, b)\ (1, a)\ (4, b)\ (1, a)$$

Solution Q4. Oui, le processus $\boldsymbol{N}$ est sans mémoire, car le nombre de fois que X_i reste consécutivement sur un même état est indépendant de ses passages sur les autres états. Le processus $\boldsymbol{N}$ est donc également markovien et homogène (a fortiori).

Solution Q5. Pour $n \geq 1$, on a $N_i = n$ si et seulement si il n'y a pas eu de transition d'état pendant exactement $n - 1$ pas de temps. La loi de N_i est donc la suivante :

$$p(N_i = n) = \lambda(1-\lambda)^{n-1}$$

C'est une loi géométrique de paramètre $1 - \lambda$.

Solution Q6. On observe que processus $\boldsymbol{C}$ passe successivement de a à b puis de b à a. Il suffit donc d'avoir la première valeur ($C_0 = a$ ou b) pour connaître l'entièreté du processus. Cette première valeur est binaire, elle peut donc être codée sur 1 bit.

Solution Q7. D'après le cours, on a :

$$\overline{\ell}(\mathcal{E}(N_i)) \geq H(N_i)$$

Solution Q8. à faire

Solution Q9. à faire
