Université Paris 8 Année 2022–2023

M1 Mathématiques et applications, parcours ACC

Théorie de l'information – Solutions feuille de TD 1

21/09/2022

Retrouvez le sujet du TD et d'autres exercices à l'adresse :

`www.math.univ-paris13.fr/~lavauzelle/teaching/2022-23/theorie-information.html`

($\star$) exercice fondamental ($\star\star$) pour s'entraîner ($\star\star\star$) pour aller plus loin 💻 sur machine

Exercice 1. ($\star$) Applications numériques.

Question 1.– Un tirage de la loterie consiste à choisir uniformément et sans remise 5 boules parmi 49, puis, de manière indépendante, à tirer un numéro complémentaire entre 1 et 10. Calculer l'entropie du tirage (on donnera une valeur approchée).

Question 2.– Comparer les entropies des variables aléatoires représentant :

(i) la valeur de 3 dés à 4 faces,
(ii) la valeur de 2 dés à 6 faces,
(iii) la valeur d'1 dé à 12 faces.

On supposera que les dés sont équilibrés.

Question 3.– On considère la distribution conjointe suivante :

$p_{X,Y}(x,y)$	x_1	x_2
y_1	$\frac{1}{4}$	0
y_2	$\frac{1}{8}$	$\frac{5}{8}$

Comparer les valeurs de $H(X)$, $H(Y)$, $H(X \mid Y)$ et $H(Y \mid X)$.

Solutions de l'Exercice 1.

Solution Q1. Comme les deux tirages sont indépendants, on a

$$H(\texttt{loterie}) = H(\texttt{boules}) + H(\texttt{numéro})\,.$$

Pour le tirages des 5 boules, on tire un sous-ensemble de 5 éléments parmi 49. Comme le tirage est uniforme, l'entropie est

$$H(\texttt{boules}) = \log_2 \binom{49}{5} = \log_2 \left(\frac{49 \cdot 48 \cdot 47 \cdot 46 \cdot 45}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} \right) = \log_2(1906884) \text{ bits.}$$

Le tirage du numéro complémentaire a une entropie de $H(\texttt{numéro}) = \log_2(10)$ bits.

Au total, on obtient donc $H(\texttt{loterie}) = \log_2(1906884) + \log_2(10) \simeq 24{,}18$ bits.

Solution Q2. On obtient les entropies suivantes.

(i) 3 dés à 4 faces : $H = 3\log_2 4 = 6$ bits ;
(ii) 2 dés à 6 faces : $H = 2\log_2 6 = 2 + 2\log_2 3 \simeq 5{,}2$ bits ;
(iii) 1 dé à 12 faces : $H = \log_2 12 = 2 + \log_2 3 \simeq 3{,}58$ bits ;

Le triplet de dés à 4 faces est donc la plus grande source d'aléa.

Solution Q3. On note $h(t) = -t\log t - (1-t)\log(1-t)$. On a respectivement :

— $H(X) = -p_{x_1}\log_2(p_{x_1}) - p_{x_2}\log_2(p_{x_2}) = h(\frac{3}{8}) = 3 - \frac{3}{8}\log 3 - \frac{5}{8}\log 5 \simeq 0{,}954$ bit.

— $H(Y) = h(\frac{1}{4}) = 2 - \frac{3}{4}\log 3 \simeq 0{,}811$ bit.

— $H(X \mid Y = y_1) = 0$ (lorsque $Y = y_1$, la valeur de X est déterminée) et $H(X \mid Y = y_2) = h(\frac{1}{6}) = 1 + \log 3 - \frac{5}{6}\log 5 \simeq 0{,}650$ bit. Donc, $H(X|Y) = \frac{1}{4} \times 0 + \frac{3}{4}h(\frac{1}{6}) = \frac{3}{4} + \frac{3}{4}\log 3 - \frac{5}{8}\log 5 = 0{,}486$ bit.

— $H(Y|X = x_1) = h(\frac{1}{3}) = 2 - \log 3$ et $H(Y|X = x_2) = 0$. Donc $H(Y|X) = \frac{3}{8}h(1/3) \simeq 0{,}344$ bit.

Exercice 2. ($\star$) Rappels sur les distributions.

Soit $X : \Omega \to \mathcal{X} = \{x_1, \dots, x_n\}$ et $Y : \Omega \to \mathcal{Y} = \{y_1, \dots, y_m\}$ deux variables aléatoires. On note $\boldsymbol{u} = (p_X(x_1), \dots, p_X(x_n))$ et $\boldsymbol{v} = (p_Y(y_1), \dots, p_Y(y_m))$ leurs distributions de probabilité, représentées sous la forme de vecteurs lignes.

Alors, la distribution de la variable conjointe (X,Y) peut être vue comme une matrice $\boldsymbol{M}$ de taille $n \times m$, où le coefficient d'indices (i,j) est

$$M_{i,j} = p_{X,Y}(x_i,y_j)\,.$$

On définit également une matrice $\boldsymbol{N}$ de taille $n \times m$, dont les entrées dont les probabilités condionnelles

$$N_{i,j} = p_{X|Y}(x_i,y_j)\,.$$

Question 1.– Parmi les assertions suivantes, lesquelles sont exactes (justifier) ?

1. Tous les coefficients de $\boldsymbol{M}$ sont dans $[0,1]$.
2. L'égalité $\boldsymbol{M} = \boldsymbol{u}^\top \cdot \boldsymbol{v}$ est toujours vérifiée.
3. La somme de tous les coefficients de $\boldsymbol{M}$ vaut 1.
4. La somme de tous les coefficients de $\boldsymbol{N}$ vaut 1.
5. La somme des coefficients de chaque colonne de $\boldsymbol{M}$ vaut 1.
6. La somme des coefficients de chaque colonne de $\boldsymbol{N}$ vaut 1.
7. Si $M_{i,j} = 0$, alors $N_{i,j} = 0$.
8. On a $M_{i,j} \leq N_{i,j}$ pour tous i,j.

Question 2.– On suppose maintenant que $m = n$. Démontrer que le vecteur $(1,1,\dots,1)^\top$ est un vecteur propre de $\boldsymbol{N}^\top$. Quelle est la valeur propre associée ?

Solutions de l'Exercice 2.

Solution Q1.

1. Vrai : ce sont des probabilités discrètes.
2. Faux : $\boldsymbol{M} = \boldsymbol{u}^\top \cdot \boldsymbol{v}$ est équivalent à dire que $p_{X,Y}(x_i,y_j) = p_X(x_i)p_Y(y_j)$. Ce n'est vrai que si X et Y sont indépendantes.
3. Vrai, car (X,Y) est une variable aléatoire, donc $\sum_{i,j} p_{X,Y}(x_i,y_j) = 1$.
4. Faux, ce sont les colonnes de $\boldsymbol{N}$ dont la somme des coefficients vaut 1 (voir le point numéro 6). On a donc $\sum_{i,j} N_{i,j} = m$.
5. Faux, conséquence du point numéro 3.
6. Vrai. En effet, pour tout y_j fixé, on a $\sum_{i=1}^n p_{X|Y}(x_i, y_j) = 1$, car c'est la somme des probabilités de la variable X conditionnée à $Y = y_j$.
7. Vrai si l'on suppose que les $p(y_j)$ sont tous non-nuls (pas vraiment précisé dans l'énoncé), car $M_{i,j} = p_{X,Y}(x_i,y_j) = p_{X|Y}(x_i,y_j)p(y_j) = N_{i,j}p(y_j)$.
8. Vrai, grâce à la formrule précédente, car $p(y_j) \leq 1$.

Solution Q2. Calculons le produit $\boldsymbol{N}^\top \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$. Le j-ème coefficient de ce vecteur vaut :

$$\sum_{i=1}^n N_{i,j} = \sum_{i=1}^n p_{X|Y}(x_i, y_j) = 1$$

On a donc $N^\top \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$. La valeur propre associée est donc 1.

Exercice 3. ($\star\star$) Entropie et tirages uniformes.

Soit X une variable aléatoire de distribution $(p_1, p_2, p_3, p_4) = (\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8})$.

Question 1.– Quelle est l'entropie de X ? Quelle serait l'entropie maximale d'une variable aléatoire à image dans un ensemble à 4 élements ?

Soit U la variable aléatoire uniforme sur $\{0, 1\}$. On considère l'algorithme suivant :

1. Tirer $u_1 \leftarrow U$. Si $u_1 = 0$, retourner $x = 1$.
2. Sinon, tirer un nouvel élément $u_2 \leftarrow U$. Si $u_2 = 0$, retourner $x = 2$.
3. Sinon, tirer un nouvel élément $u_3 \leftarrow U$. Si $u_3 = 0$, retourner $x = 3$. Sinon, retourner $x = 4$.

On suppose que les tirages successifs de U sont indépendants.

Question 2.– Démontrer que l'algorithme retourne la valeur i avec probabilité p_i, où p_i est défini plus haut.

Question 3.– Quelle est le nombre moyen de tirages de U lors d'une exécution de l'algorithme ? Comparer avec l'entropie de X.

Solutions de l'Exercice 3.

Solution Q1. Ici, $H(X) = \frac{1}{2}\log(2) + \frac{1}{4}\log(4) + \frac{1}{8}\log(8) + \frac{1}{8}\log(8) = \frac{7}{4}$. L'entropie maximale d'une variable aléatoire prenant potentiellement 4 valeurs est $\log_2(4) = 2$.

Solution Q2. Notons A la variable aléatoire de sortie de l'algorithme. L'algorithme retourne 1 si et seulement si le premier tirage de U vaut 0, donc $p(A = 1) = p(U_1 = 0) = 1/2$. Ensuite, on a $A = 2$ si le premier tirage de U est 1 et le second tirage de U vaut 0. Comme ces tirages sont indépendants, on a

$$p(A = 2) = p(U_1 = 1 \text{ et } U_2 = 0) = p(U_1 = 1)p(U_2 = 0) = \frac{1}{4}.$$

En raisonnant de manière similaire, on obtient $p(A = 3) = p(A = 4) = 1/8$.

Solution Q3. Soit N la variable aléaroire qui compte le nombre de tirages de U. On a $p(N = 1) = p(A = 1) = 1/2$, $p(N = 2) = p(A = 2) = 1/4$ et $p(N = 3) = p(A = 3 \text{ ou } A = 4) = 1/4$. Donc,

$$\mathbb{E}(N) = \frac{1}{2} + 2 \times \frac{1}{4} + 3 \times \frac{1}{4} = \frac{7}{4} = H(X).$$

L'entropie de la variable aléatoire X correspond au nombre moyen de tirages binaires uniformes nécessaires pour simuler X.

Exercice 4. ($\star\star$) Entropie conditionnelle.

On considère une rencontre sportive en deux manches gagnantes, entre deux adversaires A et B. On suppose que, pour toutes les manches de la rencontre, l'évènement « A gagne la manche » suit une loi de Bernoulli de paramètre $0 < t < 1$. On suppose également que ces variables sont indépendantes.

On note :

— Ω l'ensemble des scores possibles de la rencontre (qu'on peut écrire sous la forme de chaînes de 2 ou 3 caractères, comme $'aa'$ pour signifier que A a gagné la rencontre en deux manches), et p sa loi de probabilité associée,

— $\mathcal{X} = \{A, B\}$ l'ensemble des gagnant·e·s possibles de la rencontre, et X la variable associée,

— $\mathcal{N} = \{2, 3\}$ l'ensemble correspondant au nombre de manches du match, et N la variable associée,

— $\mathcal{Y} = \{a^*, b^*\}$ l'ensemble représentant le·a gagnant·e de la première manche, et Y la variable associée (a^* signifie que A a gagné la première manche).

Question 1.– Donner la distribution de probabilités de p, ainsi que des lois des variables X, Y et N.

Question 2.– Calculer et comparer les entropies $H(X)$, $H(Y)$ et $H(X,Y)$.

Question 3.– Rappeler comment se comparent génériquement $H(N)$ et $H(N\,|\,Y)$. Calculer $H(N)$, $H(N\,|\,Y=a^*)$ et $H(N\,|\,Y=b^*)$. Commenter.

Solutions de l'Exercice 4.

Solution Q1. Écrivons d'abord la distribution de p :

ω	aa	aba	abb	baa	bab	bb
$p(\omega)$	t^2	$t^2(1-t)$	$t(1-t)^2$	$t^2(1-t)$	$t(1-t)^2$	$(1-t)^2$

On peut alors calculer les tables des distributions des variables X, Y et N :

x	A	B
$p_X(x)$	$t^2(3-2t)$	$(1+2t)(1-t)^2$

y	a^*	b^*
$p_Y(y)$	t	$1-t$

n	2	3
$p_N(n)$	$t^2+(1-t)^2$	$2t(1-t)$

Solution Q2. Pour $H(X)$ et $H(Y)$, le calcul est immédiat :

$$H(Y) = -t\log_2(t) - (1-t)\log_2(1-t) = h(t)$$

et

$$H(X) = h(t^2(3-2t))$$

où la fonction $h(\cdot)$ est la fonction d'entropie binaire.

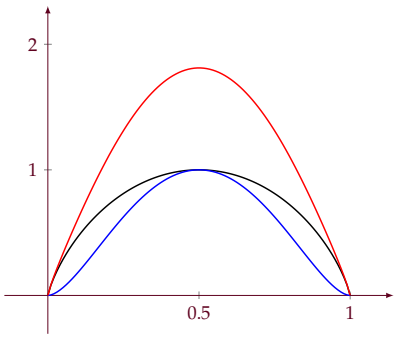
Pour la variable conjointe X,Y, présentons d'abord sa table de probabilités :

$p_{X,Y}(x,y)$	$x=A$	$x=B$
$y=a^*$	$t^2(2-t)$	$t(1-t)^2$
$y=b^*$	$(1-t)t^2$	$(1-t)^2(1+t)$

Puis on obtient

$$\begin{aligned}
H(X,Y) &= -2t^2(2-t)\log_2(t) - t^2(2-t)\log_2(2-t) - t(1-t)^2\log_2(t) - 2t(1-t)^2\log_2(1-t)\\
&\quad - 2(1-t)t^2\log_2(t) - (1-t)t^2\log_2(1-t) - 2(1-t)^2(1+t)\log_2(1-t) - (1-t)^2(1+t)\log_2(1+t)\\
&= t(1+4t-3t^2)\log_2\left(\frac{1}{t}\right) + (1-t)(1+4(1-t)-3(1-t)^2)\log_2\left(\frac{1}{1-t}\right)\\
&\quad + t^2(2-t)\log_2\left(\frac{1}{2-t}\right) + (1-t)^2(1+t)\log_2\left(\frac{1}{1+t}\right).
\end{aligned}$$

En fonction de t, les entropies calculées ont la forme suivante ($H(Y)$ en noir, $H(X)$ en bleu, $H(X,Y)$ en rouge) :



Solution Q3. On a toujours $H(N\,|\,Y) \leq H(N)$. Ici,

$$H(N) = h(t^2+(1-t)^2)\,.$$

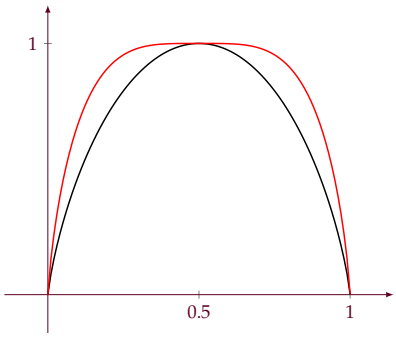
Par ailleurs, on a $p(N=2\,|\,Y=a^*) = \frac{p(N=2 \text{ et } Y=a^*)}{p(Y=a^*)} = \frac{t^2}{t} = t$ donc $p(N=3\,|\,Y=a^*) = (1-t)$ et il en résulte que :

$$H(N\,|\,Y=a^*) = h(t)\,.$$

Similairement, comme $p(N = 2 \,|\, Y = b^*) = 1 - t$, on obtient

$$H(N \,|\, Y = b^*) = h(1-t) = h(t)\,.$$

On en conclut que $H(N \,|\, Y) = h(t)$. On peut également vérifier graphiquement que $H(N \,|\, Y) \leq H(N)$ (où $H(N)$ est en rouge, $H(N \,|\, Y)$ est en noir) :



Exercice 5. ($\star\star\star$) Entropie et probabilité maximale.

Soit X une variable aléatoire de distribution $(p_1, \ldots, p_n)$. On suppose que tous les p_i sont non-nuls et on note $p^* := \max\{p_i\}_{1 \leq i \leq n}$.

Question 1.– Montrer que $H(X) \geq -\log p^*$.

Question 2.– Vérifier que cela ne contredit pas la borne $H(X) \leq \log n$ vue en cours.

Question 3.– Démontrer que, pour tous $a, b \geq 0$ avec $a + b > 0$, on a :

$$-(a+b)\log(a+b) \leq -a\log a - b\log b \leq -(a+b)\log\frac{a+b}{2}\,.$$

Démontrer également que la première égalité se produit si et seulement si $a = 0$ ou $b = 0$, et que la seconde se produit si et seulement si $a = b$.

Question 4.– Grâce à la question précédente, montrer que $H(X) \geq h_2(p^*)$, où $h_2(t) := -t\log_2 t - (1-t)\log_2(1-t)$ est la fonction d'entropie binaire.

Question 5.– En déduire que $H(X) \geq 2(1 - p^*)$.

Solutions de l'Exercice 5.

Solution Q1. Comme $-\log p_i \geq -\log p^*$ pour tout i, on obtient :

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i \geq -\sum_{i=1}^{n} p_i \log p^* = -\log p^*\,.$$

Solution Q2. Il suffit d'observer que, comme $p^* = \max_{1 \leq i \leq n}\{p_i\}$, on a bien $p^* \geq 1/n$. En effet, si ce n'était pas le cas, alors la somme $\sum_{i=1}^{n} p_i \leq np^*$ serait strictement plus petite que 1, ce qui est impossible (elle vaut 1 par définition d'une loi de probabilité).

Solution Q3. Soit $t = \frac{a}{a+b}$. On sait que $0 \leq h_2(t) \leq 1 = \log 2$, où $h_2(t) = -t\log t - (1-t)\log(1-t)$ est la fonction d'entropie binaire. On obtient donc :

$$0 \leq -\frac{a}{a+b}\log\left(\frac{a}{a+b}\right) - \frac{b}{a+b}\log\left(\frac{b}{a+b}\right) \leq \log 2$$

ce qui est équivalent à

$$-\log(a+b) \leq -\frac{a}{a+b}\log a - \frac{b}{a+b}\log b \leq -\log(a+b) + \log 2\,.$$

Le résultat voulu en résulte en multipliant par $a + b$.

Concernant les cas d'égalité :

— la première égalité se produit lorsque $h_2(t) = 0$, c'est-à-dire pour $t \in \{0,1\}$ ce qui correspond à $a = 0$ ou $b = 0$;
— la seconde égalité se produit lorsque $h_2(t) = 0$, autrement dit pour $t = 1/2$, ce qui correspond à $a = b$.

Solution Q4. Quitte à réordonner les indices, supposons que $p_1 = p^*$. D'après la question précédente, on a

$$-\sum_{i=2}^{n} p_i \log p_i \geq -\left(\sum_{i=2}^{n} p_i\right) \log\left(\sum_{i=2}^{n} p_i\right).$$

On obtient donc :

$$\begin{aligned} H(X) &= -p_1 \log p_1 - \sum_{i=2}^{n} p_i \log p_i \\ &\geq -p_1 \log p_1 - \left(\sum_{i=2}^{n} p_i\right) \log\left(\sum_{i=2}^{n} p_i\right) \\ &= -p^* \log p^* - (1-p^*) \log(1-p^*). \end{aligned}$$

Solution Q5. Pour cette question, on distingue deux cas.

— Si $p^* \geq \frac{1}{2}$, comme la fonction $t \mapsto h_2(t)$ est strictement concave sur $[\frac{1}{2}, 1]$, son graphe se situe au-dessus de la droite l'intersectant en $t = \frac{1}{2}$ et $t = 1$. Cette droite a pour équation $y = 2 - 2x$ (car $h_2(\frac{1}{2}) = 1$ et $h_2(1) = 0$). On obtient donc, grâce à la question précédente :

$$H(X) \geq h_2(p^*) \geq 2 - 2p^*.$$

— Si $p^* \leq \frac{1}{2}$, on commence par établir que $-\log_2 t \geq 2(1-t)$ pour tout $t \in [0, \frac{1}{2}]$ (la vérification est élémentaire). Puis, grâce à la question 1 on obtient

$$H(X) \geq -\log_2(p^*) \geq 2 - 2p^*.$$