TD 1

Exercice 1 (Un problème de météo)**.** Dans la vallée de la mort :
- il pleut en moyenne 1 jour sur 100.
- la météo prédit 3 jours de pluie sur 100.
- chaque fois qu'il pleut, la météo l'a prévu.
- Monsieur Sûr-de-lui prévoit qu'il ne pleut jamais.

Est-il justifié de payer cher des investissements météo, alors que Monsieur Sûr-de-lui, qui ne coûte rien, se trompe moins souvent que la météo ?

Exercice 2. Soit un vecteur $(X_1, X_2, \ldots, X_N)$ de N variables aléatoires, son entropie est par définition :

$$H(X_1 X_2 \ldots X_N) = -\sum p(x_1, \ldots, x_N) \log p(x_1, \ldots, x_N)$$

1. (Cas de l'indépendance) Montrer que si $X_1, \ldots, X_n$ sont indépendantes :

$$H(X_1 X_2 \ldots X_N) = \sum_{i=1}^{N} H(X_i)$$

2. (Cas général) Montrer que (« règle de chaînage pour l'entropie »)

$$H(X_1 X_2 \ldots X_N) = H(X_N | X_1, \ldots X_{N-1}) + H(X_{N-1} | X_1 \ldots X_{N-2}) + \cdots + H(X_2 | X_1) + H(X_1)$$

Exercice 3. Soit le vecteur aléatoire (X, Y, Z), tel que

$$\begin{aligned} p_{XYZ}(0,0,0) &= \frac{1}{4} \\ p_{XYZ}(0,1,0) &= \frac{1}{4} \\ p_{XYZ}(1,0,0) &= \frac{1}{4} \\ p_{XYZ}(1,0,1) &= \frac{1}{4} \end{aligned}$$

On rappelle la formule

$$H(Y|X) = \sum_x p(x) H(Y|X = x)$$

Donner les valeurs de $H(X)$, $H(Y|X)$, $H(Z|X,Y)$. Trouver $H(X,Y,Z)$ de deux manières : par la règle de chaînage, et directement.

On rappelle que $h(1/4) \approx 0.811$. Calculer $H(Y)$, Vérifier que $H(Y|X) \leq H(Y)$. Quelle information apporte X sur Y, Y sur X ?

Exercice 4. Montrer que $H(Y|X) = 0$ si et seulement si Y est une fonction de X, c'est à dire que pour tout x, tel que $p(x) > 0$ il existe un y tel que $p(y|x) = 1$. On utilise ici la notation (ou plutôt l'abus de notation...) $p(y|x) = p(Y = y|X = x)$.

Exercice 5. Soit X une variable aléatoire et $g(x)$ une fonction. En utilisant le règle de chaînage de deux manières différentes, montrer que

$$H(g(X)) \leq H(X).$$

Exercice 6. Soit X une variable aléatoire réelle discrète d'entropie $H(X)$. Donner le lien général entre $H(Y)$ et $H(X)$

– quand $Y = 2^X$
– quand $Y = \cos X$.

Exercice 7. On suppose que l'on a une balance à deux plateaux (qui permet donc juste de comparer les poids qui sont mis sur les deux plateaux) et de n pièces. Quand une pièce est authentique, elle a un certain poids t (qui est inconnu). Quand elle est frauduleuse elle est d'un poids différent de t (elle peut aussi bien être plus légère que plus lourde). On sait que parmi ces n pièces, une pièce au plus est frauduleuse.

1. Trouver une borne inférieure sur le nombre de pesées à effectuer de manière à être sur de détecter la pièce frauduleuse et de pouvoir dire si elle est plus légère ou plus lourde. Indication : utiliser l'exercice 5...
2. On suppose que l'on dispose également d'un tas infini de pièces authentiques. Donner la stratégie optimale pour détecter une pièce frauduleuse et dire si elle est plus légère ou plus lourde pour un tas de 13 pièces. Généraliser au cas $n = \frac{3^k-1}{2}$.

Exercice 8 (Le problème du mot de passe)**.** Un individu (probablement mal intentionné) cherche à accéder à un service protégé par un mot de passe qu'il ne connaît pas. Soit $\mathcal{M} = \{0,1\}^m$ l'ensemble des mots de passe possibles. Nous supposons que le système d'authentification est parfait et que la seule possibilité d'action pour l'attaquant consiste à essayer les mots de passe un par un.

On suppose ensuite que le mot de passe est choisi dans $\mathcal{M}$ selon une loi d'entropie $h \leq m$. Nous notons p_i les probabilités des lettres de $\mathcal{M}$ dans l'ordre décroissant (le mot le plus probable a pour probabilité p_1, le suivant pour probabilité p_2, ...).

1. Montrer que la meilleure stratégie consiste à tester les mots dans l'ordre des probabilités décroissantes. Exprimez le nombre moyen d'essais, $\mathcal{N}(p)$, en fonction des p_i.
2. Soient deux lois de probabilité $p = (p_i)_{i\geq 1}$ et $q = (q_i)_{i\geq 1}$ telles que les suites p_i et q_i soient décroissantes avec $q_i > 0$ pour tout $i \geq 1$ (en revanche p_i peut être nul à partir d'un certain rang). On rappelle que la distance de Kullback de p par rapport à q est donnée par

$$D(p \parallel q) = \sum_{i\geq 1} p_i \log \frac{p_i}{q_i}. \tag{1}$$

   Nous posons $q_i = (1-\alpha)\alpha^{i-1}$ pour un certain réel $0 < \alpha < 1$. On suppose que les entropies de $H(p)$ et $H(q)$ de p et q sont bien définies (c'est à dire que les sommes $-\sum_{i\geq 1} p_i \log p_i$ et $-\sum_{i\geq 1} q_i \log q_i$ sont bien définies). Montrer que si $H(p) = H(q)$, alors $\sum_{i\geq 1} ip_i \geq \sum_{i\geq 1} iq_i$. *Indication : on pourra tirer profit de la positivité de la distance de Kullback $D(p \parallel q) \geq 0$.*
3. Calculer l'entropie $H(q)$ de la loi q en fonction de α. Nous noterons H_α cette quantité. On rappelle les identités $\sum_{i\geq 1} \alpha^{i-1} = 1/(1-\alpha)$ et $\sum_{i\geq 1} i\alpha^{i-1} = 1/(1-\alpha)^2$
4. En déduire que pour tout réel $0 < \alpha < 1$ nous avons $1 < (1-\alpha)2^{H_\alpha} < e$, où e est la base du logarithme népérien.
5. Déduire du résultat précédent que que $\mathcal{N}(p) > c_1 2^h$ (on s'efforcera de donner une valeur à c_1). Interprétez le résultat.

Exercice 9. Montrer que

$$\binom{n}{t} \leq 2^{nh(t/n)}$$

Exercice 10. [Lemme de Fano- lien entre la probabilité d'erreur d'un estimateur et l'entropie conditionnelle] Soit X et Y deux variables aléatoires (avec X prenant ses valeurs dans un alphabet de taille a). On estime X par une certaine fonction $\hat{X}$ de Y. On note P_e la probabilité de l'estimateur, c'est à dire $P_e = p(\hat{X} \neq X)$. Montrer que

$$h(P_e) + P_e \log_2(a-1) \geq H(X|Y).$$

Indication : introduire une variable aléatoire E définie par

$$E = \begin{cases} 1 & \text{si } \hat{X} \neq X \\ 0 & \text{si } \hat{X} = X \end{cases}$$

puis écrire $H(E, X|Y)$ de deux manières différentes.

Exercice 11. Soient X et Y deux variables aléatoires à valeurs dans un groupe $(G, +)$. Soit la variable aléatoire $Z = X + Y$.

1. Montrer que $H(Z|X) = H(Y|X)$.
2. Montrer que si X et Y sont indépendantes $H(Y) \leq H(Z)$ et $H(X) \leq H(Z)$ (utiliser la positivité de l'information mutuelle).
3. Donner un exemple de deux variables aléatoires X et Y telles que $H(X) > H(Z)$ et $H(Y) > H(Z)$.