Théorie de l'information et du codage

TD n^{o}1 – Entropie

Exercice 1 – Interprétation heuristique de l'entropie

Soit $P = (P_1, \cdots, P_N)$ une mesure de probabilité sur un espace discret $\mathcal{X} = \{x_1, \ldots, x_N\}$. Une manière de construire une variable aléatoire X de loi P consiste à se placer sur l'espace $\Omega = [0, 1)$ (muni de sa tribu borélienne et de la mesure de Lebesgue) et poser :

$$X : \omega \in \Omega \mapsto x_{n(\omega)} \in \mathcal{X},$$

où $n(\omega)$ est l'entier $1 \leq n \leq N$ tel que $\sum_{i=1}^{n-1} P_i \leq \omega < \sum_{i=1}^{n} P_i$. L'antécédent $\omega \in \Omega$ représente une information secrète, inconnue. Seule son image $X(\omega) \in \mathcal{X}$ est observable et fournit des informations sur ω. Par exemple, l'observation de l'évènement $\{X(w) = x_n\}$ révèle que ω est compris entre $\sum_{i=1}^{n-1} P_i$ et $\sum_{i=1}^{n} P_i$. On cherche à définir la "quantité d'informations" $I(p) \geq 0$ apportée par l'observation d'un évènement de probabilité p. La fonction $I : (0, 1] \to [0, \infty)$ doit vérifier

$$\forall (p, q) \in [0, 1]^2, I(pq) = I(p) + I(q).$$

Justifier heuristiquement cet axiome puis en déduire la forme générale de I.

Exercice 2 – Une quasi-distance

Soit E l'ensemble des variables aléatoires (sur un espace probabilisé donné) à valeurs dans un ensemble fini donné $\mathcal{X} = \{x_1, \ldots, x_N\}$. Pour tout $(X, Y) \in E \times E$, on pose :

$$d(X, Y) = H(X|Y) + H(Y|X).$$

1. L'application d ainsi définie est-elle une distance sur E ?
2. À quelle condition à-t'on $d(X, Y) = 0$?

Exercice 3 – Simulation parfaite à l'aide d'une pièce de monnaie

À l'aide d'une simple pièce de monnaie, on cherche à simuler de manière parfaite une variable aléatoire X de loi $P = (P_1, \cdots, P_N)$ donnée sur un ensemble $N-$aire $\mathcal{X} = \{x_1, \ldots, x_N\}$, de la manière la plus économe possible en termes du nombre moyen de lancers $\mathbb{E}[T]$.

1. Montrer que toute stratégie peut être représentée par un arbre binaire (éventuellement infini) dont les feuilles sont étiquetées par des symboles de $\mathcal{X}$. Quelle condition doit être vérifiée pour que le nombre de lancements soit presque-sûrement fini et que la loi simulée soit bien P ?
2. Étant donnée une stratégie, exprimer le nombre moyen de lancer comme l'entropie binaire d'une variable aléatoire que l'on explicitera, et en déduire que nécessairement

$$\mathbb{E}[T] \geq H_2(X).$$

3. Donner une condition nécessaire et suffisante sur P pour que cette borne puisse être atteinte.
4. Dans le cas général, proposer une stratégie qui garantisse au moins :

$$H_2(X) \leq E[T] \leq H_2(X) + 3.$$

5. En déduire une interprétation asymptotique de l'entropie pour la simulation de séquences iid.

Exercice 4 – Algorithmes de tri

On considère un algorithme de tri par comparaisons que l'on suppose capable de trier toute séquence de N éléments (deux à deux distincts) en effectuant au plus K comparaisons. Le but de l'exercice est d'établir, en utilisant l'entropie, que

$$K = \Omega(N \log N) \text{ lorsque } N \to \infty.$$

Puisque seul l'ordre des éléments importe, on peut considérer l'entrée comme une simple permutation inconnue (i.e. aléatoire) Σ de $\{1, \ldots, N\}$ dont il s'agit de calculer l'inverse Σ^{-1}. Le résultat de la $k^{\text{ème}}$ comparaison $(1 \leq k \leq K)$ est la variable binaire $T_k = \mathbb{1}_{\{\Sigma(I_k) \leq \Sigma(J_k)\}}$, où le choix des éléments $1 \leq I_k, J_k \leq N$ peut dépendre de $T_1, \ldots, T_{k-1}$. Justifier l'égalité $H(\Sigma) = H(T_1, \ldots, T_K)$ et conclure.

Exercice 5 – Compenser le biais d'une pièce truquée

Soient $(X_1, \ldots, X_n)$, n variables aléatoires de Bernoulli indépendantes de même paramètre $p \in (0, 1)$. Sans connaître p, on voudrait utiliser les $(X_1, \ldots, X_n)$ pour générer une séquence $(Y_1, \ldots, Y_K) = f(X_1, \ldots, X_n)$ de variables aléatoires de Bernoulli indépendantes de paramètre $\frac{1}{2}$. La longueur $K \geq 0$ de la séquence produite peut dépendre de l'entrée $(X_1, \ldots, X_n)$, et la stratégie $f \colon \{0, 1\}^n \to \{0, 1\}^*$ est d'autant meilleure que $\mathbb{E}[K]$ est élevé.

1. Trouver une stratégie non triviale (i.e. $\mathbb{E}[K] > 0$) pour $n = 2$.
2. Donner une bonne stratégie pour $n = 4$.
3. Montrer que toute stratégie vérifie $\mathbb{E}[K] \leq nH(p)$, avec $H(p) = -p \log_2(p) - (1 - p) \log_2(1 - p)$.

Exercice 6 – Débit entropique d'une source stationnaire

Soit $X = (X_n)_{n \geq 1}$ une suite de variables aléatoires à valeurs dans $\mathcal{X}$. On suppose que X est stationnaire au sens où pour tout entier $k \geq 1$ fixé, la loi du vecteur $(X_{n+1}, \ldots, X_{n+k})$ ne dépend pas de $n \geq 0$.

1. Montrer que la limite
$$H_\infty(X) = \lim_{n \to \infty} \frac{H(X_1, \ldots, X_n)}{n}$$
existe. On l'appelle le débit entropique de la source X.
2. Quelle signification a cette grandeur pour la compression de la suite X ?

Exercice 7 – Trahison

Vous êtes monarque dans un immense royaume et c'est le jour de la dîme : chacun de vos N sujets vous remet, dans une enveloppe à son nom, la précieuse pièce d'or qui vous revient de droit. Votre espion vous apprend qu'un traitre vous a remis une pièce creuse, plus légère que les autres. Malheureusement l'espion meurt empoisonné avant de pouvoir vous révéler l'identité du coupable. Vous disposez d'une balance comparative à deux plateaux.

1. Proposer une stratégie efficace pour déterminer le coupable.
2. En utilisant l'entropie, démontrer que cette stratégie est asymptotiquement optimale.