

Partiel

Exercice 1 Entropie de Shannon¹ (60 minutes)

L'entropie de Shannon pour une variable aléatoire discrète $X = \{x_1, x_2, \dots\}$ tel que $P(X = x_i) = p_i$ est définie comme

$$H(X) = - \sum_i p_i \ln(p_i),$$

lorsque la série converge (ce que l'on ne cherchera pas à démontrer dans cet exercice). On supposera que $\forall i, p_i \neq 0$.

On considère en premier lieu les variables aléatoires finies c'est-à-dire que $i \in \{1, \dots, N\}$.

1. Soit $X \sim \text{Bernouilli}(p)$, calculer l'entropie $H(X)$.
2. Dans l'exemple précédent, pour quelle valeur de p l'entropie est elle maximale?
3. Soit X une variable aléatoire discrète de distribution uniforme sur $\{1, \dots, N\}$. Calculer l'entropie $H(X)$.
4. Démontrer que pour toute variable aléatoire discrète X telle que $H(X)$ existe, $H(X) \geq 0$ et $H(X) = 0$ si et seulement si X est une variable aléatoire constante, c'est-à-dire que $X = \alpha$ avec $P(X = \alpha) = 1$.

5. Justifier que $\forall x > 0, -x \ln(x) \leq 1 - x$ avec égalité si et seulement si $x = 1$. On rappelle que $\lim_{x \rightarrow 0^+} x \ln(x) = 0$.

6. En posant $x = Np_i$ dans l'inégalité précédente et en sommant pour i allant de 1 à N , en déduire que pour toute variables aléatoires finies X , on a $H(X) \leq H(Y)$ où Y est une variable aléatoire de distribution uniforme discrète sur $\{1, \dots, N\}$.

On considère maintenant la variable aléatoire non finie mais dénombrable $X \sim \text{Geometrique}(p)$ avec $0 < p < 1$. On admettra que $H(X)$ converge.

7. Montrer que l'entropie $H(X) = -\ln(p) - \frac{1-p}{p} \ln(1-p)$. On rappelle que $\sum_{k=0}^{+\infty} x^k = \frac{1}{1-x}$ si $|x| < 1$, et $\sum_{k=1}^{+\infty} kx^{k-1} = \frac{1}{(1-x)^2}$ si $|x| < 1$.

On considère maintenant les variables aléatoires X admettant une densité de probabilité f_X . On supposera que le support de f_X est \mathbb{R} . Dans ce cas, l'entropie de Shannon (toujours sous réserve d'existence de l'intégrale) est définie comme

$$H(X) = - \int_{-\infty}^{+\infty} f_X(x) \ln f_X(x) dx.$$

8. On pose $Y = aX + b$ avec $a > 0$ et $b \in \mathbb{R}$. Montrer que $f_Y(y) = \frac{1}{a} f_X\left(\frac{y-b}{a}\right)$.
9. En déduire l'on peut écrire $H(Y) = H(X) - \ln(a)$.

Exercice 2 Questions diverses (20 minutes)

On rappelle que $\forall z, \sum_{m=0}^{+\infty} \frac{z^m}{m!} = e^z$.

Les questions au sein de cet exercice sont toutes indépendantes.

1. En théorie de l'information (au programme du M1 E3A), l'entropie de Shannon est une fonction mathématique qui, intuitivement, correspond à la quantité d'information contenue ou fournie par une source d'information. Le calcul de l'entropie d'une source de messages donne une mesure de l'information minimale que l'on doit conserver afin de représenter ces données sans perte. On cherche ici à montrer quelques résultats concernant les valeurs minimale et maximal pour être atteinte par cette fonction.

1. La densité moyenne d'une certaine molécule dans un mètre cube d'air est égale à 100. On prend un échantillon de deux décimètres cubes d'air. On modélise le nombre de molécule par une loi de Poisson. Donner la probabilité pour que dans ce volume il y ait au moins une molécule.

2. Soit X une variable aléatoire (discrète ou à densité) telle que $X \in L^2$. Démontrer en justifiant soigneusement (c'est-à-dire sans utiliser l'homogénéité de la variance) que $Var(2X + 1) = 4Var(X)$.

3. Le joueur A possède deux dés à 6 faces et un dé à 12 faces, et le joueur B possède un dé à 24 faces (les 4 dés sont équilibrés). Le joueur qui fait le plus grand score remporte la mise (match nul si égalité). Calculer la probabilité pour que le joueur A gagne. On utilisera la fonction $\mathbb{I}_{\{p>q\}} = \begin{cases} 1 & \text{si } p > q \\ 0 & \text{sinon} \end{cases}$ pour calculer le nombre d'éléments de l'espace :

$$\Theta = \{(i, j, k, m) \text{ tel que } 1 \leq i \leq 6; 1 \leq j \leq 6; 1 \leq k \leq 12; 1 \leq m \leq 24; \text{ et } i + j + k > m\}$$

4. Soit X une V.A. de distribution de Poisson de paramètre θ , montrer que $\frac{1}{(X+1)(X+2)(X+3)} \in L^1$ et calculer $\mathbb{E}\left(\frac{1}{(X+1)(X+2)(X+3)}\right)$.

Exercice 3 Distribution de Pareto² (40 minutes)

Soit $\alpha > 0$ et $\beta > 0$. Soit X une variable aléatoire admettant pour densité de probabilité la fonction $f_X(x) = \frac{\beta^\alpha \alpha}{x^{\alpha+1}}$ si $x \geq \beta$ et $f_X(x) = 0$ si $x < \beta$.

- Démontrer que $f_X(x)$ respecte les propriétés d'une densité de probabilité.
- Calculer la fonction de répartition de X (ainsi que son domaine de définition)
- Pour quelles valeurs de α a-t-on $X \in L^1$? Pour quelles valeurs de α a-t-on $X \in L^2$?
- Calculer la moyenne de X dans le cas où α est tel que $X \in L^1$.
- Montrer que la variance de X dans le cas où α est tel que $X \in L^2$ s'écrit $Var(X) = \frac{\beta^2 \alpha}{(\alpha-1)^2(\alpha-2)}$.
- Pour une variable aléatoire réelle admettant une densité de probabilité, on appelle médiane³ le nombre, noté m , tel que $P(X \leq m) = P(X \geq m) = \frac{1}{2}$. Calculer la médiane de la variable aléatoire X .
- On pose $Y = X^2 - 1$. Montrer en justifiant soigneusement que la densité de probabilité $f_Y(y)$ de la variable aléatoire Y s'écrit $f_Y(y) = \frac{\alpha \beta^\alpha}{2(y+1)^{\frac{\alpha+2}{2}}}$ sur son domaine de définition que l'on précisera.

Exercice 4 Perte de paquets de données (30 minutes)

La perte de paquets de données se produit lorsqu'un ou plusieurs paquets de données transitant par un réseau informatique n'arrivent pas à destination. La perte de paquets est causée soit par des erreurs de transmission de données, généralement sur des réseaux sans fil soumis au bruit, soit par une congestion du réseau. Dans les médias en streaming et les applications de jeux en ligne, la perte de paquets peut affecter la qualité d'expérience d'un utilisateur et connaître les probabilités que de tels événements se produisent est donc un enjeu crucial. Bien qu'il ne s'agisse pas d'un événement rare, on peut supposer que le nombre N de paquets de données envoyés à travers un réseau chaque jour par une entreprise suit une distribution de Poisson avec un paramètre θ . Les paquets de données sont transmis de manière indépendante les uns des autres. La probabilité qu'un paquet de données soit perdu lors de sa transmission est égale à p . Nous

2. Cette distribution est un outil fondamental en gestion de la qualité. Elle est aussi utilisée en réassurance. La théorie des files d'attente s'est intéressée à cette distribution, lorsque des recherches des années 90 ont montré que cette loi régissait aussi nombre de grandeurs observées dans le trafic Internet (et plus généralement sur tous les réseaux de données à grande vitesse). Ce phénomène a de sévères répercussions sur les performances des systèmes (routeurs en particulier).

3. La médiane est un indicateur de tendance centrale. Par comparaison avec la moyenne, elle est insensible aux valeurs extrêmes mais son calcul est un petit peu plus complexe. Prenons l'exemple des salaires. En France, en 2019, le salaire mensuel net moyen (en équivalent temps plein) était de 2 424 euros dans le secteur privé. Or, cette série contient de nombreuses valeurs extrêmes, c'est-à-dire éloignées de la moyenne. Par exemple, 1 % des employés du secteur privé ont un salaire supérieur à 9 103 euros par mois. Ces valeurs extrêmes gonflent ici, artificiellement, la moyenne et la rendent, d'une manière générale, peu représentative de la distribution effective des données. Dans ce cas, calculer la médiane est utile : le salaire mensuel net médian (en équivalent temps plein) était de seulement 1 940 euros. Autrement dit, 50 % des salariés du secteur privé touchaient plus de 1 940 euros par mois et 50 % gagnaient moins. La médiane a, de plus, l'avantage de toujours exister (contrairement à la moyenne). Les enseignants ne devraient donc pas donner la médiane des notes obtenues à un examen pour une promotion plutôt que la classique « moyenne de classe » ?

nous intéressons aux paquets de données envoyés un jour donné : N est la variable aléatoire dénombrant le nombre de paquets de données envoyés ; X est la variable aléatoire dénombrant le nombre de paquets de données perdus.

1. Si $k > n$, donner la probabilité $P(X = k | N = n)$.
2. On suppose maintenant dans toute la suite que $n \geq k$. En justifiant soigneusement, expliquer que la distribution de la variable aléatoire $X = k | N = n$ est une loi Binomiale. Donner ensuite les paramètres de cette loi.
3. Montrer, en justifiant soigneusement, que $P(X = k) = \sum_{n=k}^{+\infty} P(X = k | N = n) P(N = n)$.
4. Montrer ensuite que l'on peut écrire $P(X = k) = \alpha \sum_{n=k}^{+\infty} \frac{(\theta(1-p))^n}{(n-k)!}$ où α est une fonction de θ , k et p que l'on détaillera. Est-ce que cette série converge ?
5. Sachant que $\sum_{m=0}^{+\infty} \frac{z^m}{m!} = e^z$, à l'aide d'un changement de variable bien choisi, montrer que le nombre de paquets de données perdus suit, lui aussi, une loi de Poisson dont on précisera le paramètre.
6. A l'aide du même raisonnement donner directement la loi de la variable aléatoire Y dénombrant le nombre de paquets de données arrivé à destination.

Exercice 5 Problème (30 minutes)

Dans cet exercice, toutes les formules doivent être démontrées avant calcul numérique en utilisant les lettres bien choisies liées aux événements aléatoires de l'énoncé. Pour les applications numérique, les résultats seront arrondis au millième.

Une agence de voyage propose des itinéraires touristiques pour lesquels chaque client effectue obligatoirement un aller et un retour en utilisant soit un bateau, soit un train touristique. Le choix du mode de transport peut changer entre l'aller et le retour. À l'aller, le bateau est choisi dans 65% des cas. **Sachant que** le bateau est choisi à l'aller, il l'est également pour le retour fois 9 sur 10. **Sachant que** le train a été choisi à l'aller, le bateau est préféré pour le retour dans 70% des cas. Tous les clients sont supposés indépendants.

On choisit au hasard un client de l'agence.

1. Calculer la probabilité que le client fasse l'aller-retour en bateau.
2. Montrer que la probabilité que le client utilise les deux moyens de transport est de 0,31.
3. **Sachant que** le bateau est choisi au retour, calculer la probabilité que le bateau ait été choisi à l'aller.

On choisit au hasard 20 clients de cette agence. On note X la variable aléatoire qui compte le nombre de clients qui utilisent les deux moyens de transport.

4. Donner la distribution de X ainsi que ses paramètres.
5. Déterminer la probabilité qu'exactement 12 clients utilisent les deux moyens de transport différents.
6. Déterminer la probabilité qu'il y ait au moins 2 clients qui utilisent les deux moyens de transport différents.

Le coût d'un trajet aller **ou** d'un retour est de 1560€ en bateau, le coût d'un trajet aller **ou** d'un retour est de 1200€ en train. On note Y la variable aléatoire qui associe, à un client pris au hasard, le coût en euro de son trajet aller-retour.

7. Calculer la distribution de probabilité de Y .
8. Calculer l'espérance mathématique de Y .