Problem sheet 1, Information Theory, HT 2022

Designed for the first tutorial class

Question 1 We are given a deck of n cards in order $1, 2, \cdots, n$. Then a randomly chosen card is removed and placed at a random position in the deck. What is the entropy of the resulting deck of card?

Answer 1 There are evenly n cards be picked up, and n places to be placed evenly. So there are $1/n^2$ different actions with even probability $1/n^2$ and some of them result in the same outcome as following:

(a) The original order can be resulted by any card be picked up and placed at its original place, so the probability of the original order is $1/n$;

(b) a swap of two adjacent card can be resulted by two different operations. There are $n-1$ of these results, each with probability $2/n^2$.

(c) a card is moved at least 2 positions away: there are $n \times (n-3) + 2$ possible results, each with probability $1/n^2$.

So there are $1 + (n-1) + (n^2 - 3n + 2)$ different results with probabilities as above, whose entropy is

$$\begin{aligned} H(\text{“}deck\text{”}) &= \frac{1}{n}\log(n) + (n-1)\frac{2}{n^2}\log(n^2/2) + (n^2-3n+2)\frac{1}{n^2}\log(n^2) \\ &= \frac{1}{n^2}\left[n\log(n) + 2(n-1)(2\log(n)-1) + (n^2-3n+2)2\log(n)\right] \\ &= \frac{2n-1}{n}\log(n) + \frac{2n-2}{n^2}. \end{aligned}$$

Question 2 (Polling inequalities) Let $a \geq 0, b \geq 0$ are given with $a+b > 0$. Show that

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

and that the first inequality becomes an equality iff $ab = 0$, the second inequality becomes an equality iff $a = b$.

Answer 2 Denote $p = \frac{a}{a+b}$. Divided by $a + b$ and then add $\log(a + b)$ on all three terms, the equalities are equivalent to

$$0 \leq -p\log(p) - (1-p)\log(1-p)) \leq -\log(\frac{1}{2}),$$

which is obvious according to the first basic property of entropy.

Question 3 Let X, Y, Z be discrete random variables. Prove or provide a counterexample to the following statements:

(a) $H(X) = H(42X)$;

(b) $H(X|Y) \geq H(X|Y,Z)$;

(c) $H(X,Y) = H(X) + H(Y)$.

Answer 3 The first one is true : $f(x) = -42x$ is a bijective.

The second is true: $H(X|Y) - H(X|Y,Z) = I(X,Z|Y)$, and the interpretation by information/surprise works.

The third is wrong: By the chain rule, $H(X,Y) = H(Y|X) + H(X)$, and $H(Y|X) = H(Y)$ if and only if X, Y are independent. An easy counter example is when $Y = X$ and $H(X) > 0$, we have $H(X,Y) = H(X,X) = H(X) < H(Y) + H(X)$.

Question 4 Does there exist a discrete random variable X with a distribution such that $H(X) = +\infty$? If so, describe it as explicitly as possible.

Answer 4 Obviously, $H(X) < +\infty$ for any case with finite image space. So we assume the image space is the natural nubmers. Here is an counter example: $\mathbb{P}(X = n) = \frac{c}{n\log^2(n)}$ with $c = \frac{1}{\sum_n n\log^2 n} > 0$. then $H(X) = \sum_n \frac{c}{n\log^2(n)}[\log(n\log^2(n)) - \log(c)] = \sum_n \left[\frac{2c}{n\log(n)} + \frac{2c(\log(\log(n))}{n\log^2(n)}\right] - \log(c) = +\infty$ since $\log(\log(n)) \to +\infty$ and $\sum_n \frac{1}{n\log(n)} = +\infty$.

Question 5 Let $\mathcal{X}$ be a finite set, f a real-valued function $f : \mathcal{X} \mapsto \mathbb{R}$ and fix $\alpha \in \mathbb{R}$. We want to maximise the entropy $H(X)$ of a random variable X taking values in $\mathcal{X}$ subject to the constraint

$$\mathbb{E}[f(X)] \leq \alpha. \tag{1}$$

Denote by U a uniformly distributed random variable over $\mathcal{X}$. Prove the following optimal solutions for the maximisation.

(a) If $\alpha \in [\mathbb{E}[f(U)],\ \max_{x\in\mathcal{X}} f(x)]$, then the entropy is maximised subject to (1) by the uniformly distributed random variable U.

(b) If f is non-constant and $\alpha \in [\min_{x\in\mathcal{X}} f(x),\ \mathbb{E}[f(U)]]$, then the entropy is maximised subject to (1) by the random variable Z given by

$$P(Z = x) = \frac{e^{\lambda f(x)}}{\sum_{y\in\mathcal{X}} e^{\lambda f(y)}} \qquad \text{for } x \in \mathcal{X}.$$

where $\lambda < 0$ is chosen such that $\mathbb{E}[f(Z)] = \alpha$.

(c) (Optional) Prove that under the assumptions of (b), the choice for λ is unique and we have $\lambda < 0$.

Answer 5 (a) Since the uniform distribution achieves the maximal entropy without any constrained, so we just need to verify it satisfies the constraint (1), which is obvious.

(b) Recall the Gibbs' inequality that for any pmf p and q,

$$-\sum_{x\in\mathcal{X}} p(x)\log(p(x)) \leq -\sum_{x\in\mathcal{X}} p(x)\log(q(x)).$$

So we can try to write $\mathbb{E}[f(X)]$ into the form of $-\sum_{x\in\mathcal{X}} p(x)\frac{\log(q(x))+c}{-\lambda}$ for some constant $\lambda < 0$ and c with $p(\cdot)$ being the pmf of X, for which we should write

$$\lambda f(x) = \log(q(x)) + c(\lambda) \iff e^{c(\lambda)}e^{\lambda f(x)} = q(x).$$

With the fact that q is a pmf, we have

$$q(x) = \frac{e^{\lambda f(x)}}{\sum_{x\in\mathcal{X}} e^{\lambda f(x)}} \text{ and } c(\lambda) = -\log(\sum_{x\in\mathcal{X}} e^{\lambda f(x)}).$$

So for any $\lambda < 0$, define the pmf $q(x) := \frac{e^{\lambda f(x)}}{\sum_{x\in\mathcal{X}} e^{\lambda f(x)}}$, then

$$\mathbb{E}[f(X)] = \frac{1}{\lambda}\sum_{x\in\mathcal{X}} p(x)\log(q(x)) + \frac{c(\lambda)}{\lambda}.$$

With $\lambda < 0$, we have that $\mathbb{E}[f(X)] \leq \alpha$ is equivalent to $-\sum_{x\in\mathcal{X}} p(x)\log(q(x)) \leq -\lambda\alpha + c(\lambda)$. Hence $H(X) \leq -\sum_{x\in\mathcal{X}} p(x)\log(q(x)) \leq -\lambda\alpha + c(\lambda)$, and the equality holds iff $p(x) = q(x)$, i.e., $\mathbb{P}(X = x) = q(x) = \frac{e^{\lambda f(x)}}{\sum_{x\in\mathcal{X}} e^{\lambda f(x)}}$ and $\alpha = \sum_{x\in\mathcal{X}} p(x)f(X) = \sum_{x\in\mathcal{X}} f(x)\frac{e^{\lambda f(x)}}{\sum_{y\in\mathcal{X}} e^{\lambda f(y)}}$.

To make sure the existence of $\lambda < 0$ such that $\mathbb{E}[f(X)] = \sum_{x\in\mathcal{X}} q(x)f(x) = \alpha$, we leave the proof to part (c).

Question 5b: to show that equality constraint must be satisfied.

$$
\begin{aligned}
& \mathbb{E}_p[f(x)] \leq \alpha \\
\iff & \sum_x p(x) f(x) \leq \alpha \\
\iff & -\sum_x p(x) \lambda f(x) \leq -\lambda\alpha && \text{since } \lambda < 0 \\
\iff & -\sum_x p(x) \log\Big(\exp\big(\lambda f(x)\big)\Big) \leq -\lambda\alpha \\
\iff & -\sum_x p(x) \log\Big(\frac{\exp\big(\lambda f(x)\big)}{\sum_x \exp\big(\lambda f(x)\big)}\Big) \leq -\lambda\alpha + \log\Big(\sum_x \exp\big(\lambda f(x)\big)\Big) && \text{Note } \frac{\exp\big(\lambda f(x)\big)}{\sum_x \exp\big(\lambda f(x)\big)} \text{ is a probability distribution}
\end{aligned}
$$

Thus, for the solution given by the Lagrangian multiplier (shown in the class), we have

$$
H(X) = -\sum_x p(x) \log\Big(\frac{\exp\big(\lambda f(x)\big)}{\sum_x \exp\big(\lambda f(x)\big)}\Big) \leq -\lambda\alpha + \log\Big(\sum_x \exp\big(\lambda f(x)\big)\Big)
$$

There exists a unique λ such that the equality holds and we denote it as λ_0, i.e., $\sum_x p(x) f(x) = \alpha$ (the proof is given in question 3c). For all other $\lambda' \neq \lambda_0$ and $\lambda' < 0$ (i.e., $\sum_x p'(x) f(x) < \alpha$), we have

$$
\begin{aligned}
H_{\lambda'}(X) &= -\sum_x p'(x) \log\Big(\frac{\exp\big(\lambda' f(x)\big)}{\sum_x \exp\big(\lambda' f(x)\big)}\Big) \\
&< -\sum_x p'(x) \log\Big(\frac{\exp\big(\lambda f(x)\big)}{\sum_x \exp\big(\lambda f(x)\big)}\Big) && \text{Gibb's inequality} \\
&< -\lambda\alpha + \log\Big(\sum_x \exp\big(\lambda f(x)\big)\Big) \\
&= H_\lambda(X) && \text{The entropy for all other } \lambda \text{ is strictly smaller than } H_\lambda(X)
\end{aligned}
$$

Therefore, the equality holds $\sum_x p(x) f(x) = \alpha$.

(c) Denote $g(\lambda) := \sum_{x\in\mathcal{X}} f(x)\frac{e^{\lambda f(x)}}{\sum_{y\in\mathcal{X}} e^{\lambda f(y)}}$. Then g is a differentiable function with

$$\begin{aligned} g'(\lambda) &= \sum_{x\in\mathcal{X}} f(x)^2 \frac{e^{\lambda f(x)}}{\sum_{y\in\mathcal{X}} e^{\lambda f(y)}} - \sum_{x\in\mathcal{X}} f(x) \frac{e^{\lambda f(x)}}{(\sum_{y\in\mathcal{X}} e^{\lambda f(y)})^2} \sum_{y\in\mathcal{X}} f(y) e^{\lambda f(y)} \\ &= \mathbb{E}[f(X)^2] - (\mathbb{E}[f(X)])^2. \end{aligned}$$

Since f is not a constant, so $g'(\lambda) > 0$, which means g is a strictly increasing and continuous function. Furthermore, $g(0) = \mathbb{E}[f(U)]$, $g(-\infty) = \min_{x\in\mathcal{X}} f(x)$. So $g(\lambda) = \alpha \in (\min f(x), \mathbb{E}[f(U)])$ admits a unique solution $\lambda < 0$.

Question 6 *(A revision on strong law of large numbers (SLLN) in probability theory, please take this question as a reference) Let X be a real-valued random variable.*

(a) *Assume additionally that X is non-negative. Show that for every $x > 0$, we have*

$$\mathbb{P}(X \geq x) \leq \frac{\mathbb{E}[X]}{x}.$$

(b) *Let X be a random variable of mean μ and variance σ^2. Show that*

$$\mathbb{P}(|X - \mu| > \varepsilon) \leq \frac{\sigma^2}{\epsilon^2}.$$

(c) *Let $(X_n)_{n\geq 1}$ be a sequence of i.i.d random variables with mean μ and variance σ^2. Show that $\frac{1}{m}\sum_{n=1}^m X_n$ converges to μ in probability, i.,e., for every $\varepsilon > 0$,*

$$\lim_{m\to+\infty} \mathbb{P}\left(\left| \frac{1}{m}\sum_{n=1}^m X_n - \mu \right| > \epsilon \right) = 0.$$

`This is a weak version of SLLN. It can be strengthen by Borel-Cantelli lemma to the often-used version:` $\mathbb{P}(\lim_{m\to+\infty} \frac{1}{m}\sum_{n=1}^m X_n = \mu) = 1$.

Answer 6 (a) $\mathbb{E}[X] = \mathbb{E}[X\mathbf{1}_{X\geq x}] + \mathbb{E}[X\mathbf{1}_{X<x}] \geq \mathbb{E}[X\mathbf{1}_{X\geq x}] \geq \mathbb{E}[x\mathbf{1}_{X\geq x}] = x\mathbb{P}(X \geq x)$, so we have the inequality.

(b) Similar to part (a), for any random variable Y and constant $\varepsilon > 0$, $\mathbb{P}(|Y| > \varepsilon) \leq \frac{\mathbb{E}[Y^2]}{\varepsilon^2}$. Apply $Y = X - \mu$ in this inequality, we get the one in the question.

(c) For any integer m, denote $Y_m = \frac{1}{m}\sum_{n=1}^m X_n - \mu$, then $\mathbb{E}[Y_m] = 0, \mathrm{Var}(Y_m) = \frac{\sigma^2}{m}$. Hence $\mathbb{P}(|Y_m| > \varepsilon) \leq \frac{\sigma^2}{m\varepsilon} \overset{m\to+\infty}{\longrightarrow} 0$.

Question 7 *(Optional)* Partition the interval $[0, 1]$ into n disjoint sub-intervals of length $p_1, \cdots, p_n$. Let $X_1, X_2, \cdots$ be i.i.d. random variables, uniformly distributed on $[0, 1]$, and $Z_m(i)$ be the number of the $X_1, \cdots, X_m$ that lie in the i^{th} interval of the partition. Show that the random variables

$$R_m = \Pi_{i=1}^n p_i^{Z_m(i)} \text{ satisfy } \frac{1}{m}\log(R_m) \overset{m\to+\infty}{\longrightarrow} \sum_{i=1}^n p_i \log(p_i) \text{ with probability 1.}$$

Answer 7 Denote I_i as the i^{th} subinterval.

It is easy to see that $\frac{1}{m}\log(R_m) = \frac{1}{m}\sum_{i=1}^n Z_m(i)\log(p_i) = \sum_{i=1}^m \frac{\sum_{j=1}^n \mathbf{1}_{X_j\in I_i}}{m}\log(p_i)$. Since $\mathbb{P}(\lim_{n\to+\infty} \frac{\sum_{j=1}^n \mathbf{1}_{X_j\in I_i}}{m} = p_i) = 1$, the conclusion follows.