Tutorial 2

1 Exercises

1. *A fair coin is flipped until the first head occurs. Let X denote the number of flips required. Find the entropy $H(X)$ in bits. The following expressions may be useful.*

$$\sum_{n=0}^{\infty} r^n = \frac{1}{1-r}, \quad \sum_{n=0}^{\infty} nr^n = \frac{r}{(1-r)^2} \quad \textit{for } r < 1.$$

Solution: The number X of tosses til the first head has the geometric distribution with the parameter $p = 1/2$, where $P(X = n) = (1-p)^{n-1}p$, $n \in \{1, 2, \cdots\}$. Hence the entropy of X is

$$\begin{aligned}
H(X) &= -\sum_{n=1}^{\infty}(1-p)^{n-1}p\log((1-p)^{n-1}p) \\
&= -\left[\sum_{n=1}^{\infty}(1-p)^{n-1}p\log p + \sum_{n=1}^{\infty}(1-p)^{n-1}p\log(1-p)^{n-1}\right] \\
&= -\left[\sum_{m=0}^{\infty}(1-p)^{m}p\log p + \sum_{m=0}^{\infty}m(1-p)^{m}p\log(1-p)\right] \\
&= \frac{-p\log 0}{1-(1-p)} - \frac{p(1-p)\log(1-p)}{p^2} \\
&= \frac{-p\log p - (1-p)log(1-p)}{p} \\
&= \frac{H(p)}{p} \quad \text{bits.}
\end{aligned}$$

If $p = 1/2$, then $H(X) = 2$ bits. □

2. Let X be a random variable taking on a finite number of values. What is the general inequality relationship of $H(X)$ and $H(Y)$ if

(a) $Y = 2^X$?

(b) $Y = \cos X$?

Solution: Let $y = g(x)$. Then

$$p(y) = \sum_{x:y=g(x)} p(x).$$

Consider any set of $x's$ that map onto a single y. For this set

$$p(y) = \sum_{x:y=g(x)} p(x)\log p(x) \leq p(y) = \sum_{x:y=g(x)} p(x)\log p(y) = p(y)\log p(u),$$

since log is a monotone increasing function and $p(x) \leq \sum_{x:y=g(x)} p(x) = p(y)$. Extending this argument to the entire range of X and Y, we obtain

$$\begin{aligned} H(X) &= -\sum_x p(x) \log p(x) \\ &= -\sum_y \sum_{x:y=g(x)} p(x) \log p(x) \\ &\geq -\sum_y p(y) \log p(y) \\ &= H(Y), \end{aligned}$$

with equality iff g is one-to-one with probability one.

(a) $Y = 2^X$ is one-to-one and hence the entropy doesn't change, i.e., $H(X) = H(Y)$.

(b) $Y = \cos X$ is not necessarily one-to-one. Hence all that we can say is that $H(X) \geq H(Y)$, with equality if cosine is one-to-one on the range of X. □

3. Let $X_1 \to X_2 \to \cdots \to X_n$ form a Markov chain in this order; i.e. let

$$p(x_1, x_2, \cdots, x_n) = p(x_1)p(x_2|x_1) \cdots p(x_n|x_{n-1}).$$

Reduce $I(X_1; X_2, \cdots, X_n)$ to its simplest form.

Solution: By the chain rule for mutual information,

$$I(X_1; X_2, \cdots, X_n) = I(X_1; X_2) + I(X_1; X_3|X_2) + \cdots + I(X_1; X_n|X_2, \cdots, X_{n-1})$$

Note that the mutual information between two independent random variables is zero. By the Markov property, the past and the future are conditionally independent given the present and hence all term except the first are zero. Therefore,

$$I(X_1; X_2, \cdots, X_n) = I(X_1; X_2)$$

□

4. Let X, Y and Z be joint random variables. Prove the following inequalities ad find conditions for equality.

(a) $H(X, Y|Z) \geq H(X|Z)$.

(b) $I(X, Y; Z) \geq I(X; Z)$.

(c) $H(X, Y, Z) - H(X, Y) \leq H(X, Z) - H(X)$.

(b) $I(X; Z|Y) \geq I(Z; Y|X) - I(Z; Y) + I(X; Z)$.

Solution:

(a) Using the chain rule for conditional entropy,

$$H(X, Y|Z) = H(X|Z) + H(Y|X, Z) \geq H(X|Z)$$

with the equality iff $H(Y|X, Z) = 0$, that is when Y is a function of X and Z.

(b) Using the chain rule of mutual information,

$$I(X,Y;Z) = I(X;Z) + I(Y;Z|X) \geq I(X;Z),$$

with the equality iff $I(Y;Z|X) = 0$, that is, when Y and Z are conditionally independent given X.

(c) Using first the chain rule for entropy and then the definition of conditional mutual information,

$$\begin{aligned} H(X,Y,Z) - H(X,Y) =& H(Z|X,Y) \\ =& H(Z|X) - I(Y;Z|X) \\ \leq& H(Z|X) \\ =& H(X,Z) - H(X), \end{aligned}$$

with the equality iff $I(Y;Z|X) = 0$, that is, when Y and Z are conditionally independent given X.

(d) Using the chain rule of mutual information,

$$I(X;Z|Y) + I(Z;Y) = I(X,Y;Z) = I(Z;Y|X) + I(X;Z),$$

and therefore

$$I(X;Z|Y) = I(Z;Y|X) + I(X;Z) - I(Z;Y).$$

We see that this inequality is actually an equality in all cases.

□

5. Let $P(X = i) = p_i$, $i = 1, 2, \cdots, m$ and let $p_1 \geq p_2 \geq p_3 \geq \cdots p_m$. The minimal probability of error predictor of X is $\hat{X} = 1$, with resulting probability of error $P_e = 1 - p_1$. Maximize $H(X)$ subject to the constraint $1 - p_1 = P_e$ to find a bound on P_e in terms of H.

Solution: The minimal probability of error predictor when there is no information is $\hat{X} = 1$, the most probable value. The probability of error in this case is $P_e = 1 - p_1$. Hence if we fix P_e, we fix p_1. We maximize the entropy of X for a given P_e to obtain an upper bound on the entropy for a given P_e. The entropy is as follows.

$$\begin{aligned} H(X) =& -\sum_{i=1}^{m} p_i \log p_i \\ =& -p_1 \log p_1 - \sum_{i=2}^{m} p_i \log p_i \\ =& -p_1 \log p_1 - \sum_{i=2}^{m} P_e \frac{p_i}{P_e} \log \frac{p_i}{P_e} P_e \\ =& -p_1 \log p_1 - \sum_{i=2}^{m} P_e \frac{p_i}{P_e} \log \frac{p_i}{P_e} - \sum_{i=2}^{m} P_e \frac{p_i}{P_e} \log P_e \\ =& -p_1 \log p_1 - P_e \sum_{i=2}^{m} \frac{p_i}{P_e} \log \frac{p_i}{P_e} - \sum_{i=2}^{m} p_i \log P_e \end{aligned}$$

Since $\sum_{i=2}^m p_i = 1 - p_1 = P_e$, $\sum_{i=2}^m \frac{p_i}{P_e} = 1$. And $\frac{p_2}{P_e}, \frac{p_3}{P_e}, \cdots, \frac{p_m}{P_e}$ is actually a distribution. Then we can rewrite $H(X)$ as

$$H(X) = - p_1 \log p_1 - P_e \log P_e + P_e H(\frac{p_2}{P_e}, \frac{p_3}{P_e}, \cdots, \frac{p_m}{P_e})$$
$$= H(P_e) + P_e H(\frac{p_2}{P_e}, \frac{p_3}{P_e}, \cdots, \frac{p_m}{P_e})$$
$$\leq H(P_e) + P_e \log(m-1)$$

since the maximum of $H(\frac{p_2}{P_e}, \frac{p_3}{P_e}, \cdots, \frac{p_m}{P_e})$ is attained by an uniform distribution. Hence any X that can be predicted with a probability of error P_e must satisfy

$$H(X) \leq H(P_e) + P_e \log(m-1)$$

This is the unconditional form of Fano's inequality. We can weaken this inequality to obtain an explicit lower bound for P_e,

$$P_e \geq \frac{H(X) - 1}{\log(m-1)}.$$

□

Note: **Fano's inequality:** Suppose we wish to estimate a random variable X with a distribution $p(x)$. We observe a random variable Y which is related to X by the conditional distribution $p(x|y)$. From Y, we calculate a function $g(Y) = \hat{X}$, which is an estimate of X. We wish to bound the probability that $\hat{X} = X$. We observe that $X - Y - \hat{X}$ forms a Markov chain. Define the probability of error $P_e = P\{\hat{X} \neq X\}$. Then $H(P_e) + P_e \log(|\mathcal{X}| - 1) \geq H(X|Y)$. This inequality can be weaken to $1 + P_e \log|\mathcal{X}| \geq H(X|Y)$.

6. How much information does the length of a sequence give about the content of a sequence? Suppose we consider a Bernoulli (1/2) process $\{X_i\}$.

   Stop the process when the first 1 appears. Let N designate this stopping time. Thus X^N is an element of the set of all finite length binary sequences $\{0,1\}^* = \{0, 1, 00, 01, 10, \cdots\}$.

   (a) Find $I(N; X^N)$.

   (b) Find $H(X^N|N)$.

   (c) Find $H(X^N)$.

   Let's now consider a different stopping time. For this part, again assume $X_i \sim$Bernoulli (1/2) but stop at time $N = 6$, with probability 1/3 and stop at time $N = 12$ with probability 2/3. Let this stopping time be independent of sequence $X_1 X_2 \cdots X_{12}$.

   (d) Find $I(N; X^N)$.

   (e) Find $I(N; X^N)$.

   (f) Find $I(N; X^N)$.

   Solution:

(a)

$$\begin{aligned} I(N; X^N) =& H(N) - H(N|X^N) \\ =& H(N) - 0 \end{aligned}$$

Note that the entropy of a geometric random variable X with parameter p is $H(X) = \frac{H(p)}{p}$. Since N is under a geometric distribution with parameter $1/2$, we have

$$I(N; X^N) = H(N) = \frac{H(1/2)}{1/2} = 2.$$

(b) Since given N we know that $X_i = 0$ for all $i < N$ and $X_N = 1$,

$$H(X^N|N) = 0.$$

(c)

$$\begin{aligned} H(X^N) =& I(X^N; N) + H(X^N|N) \\ =& I(X^N; N) + 0 \\ =& 2 \end{aligned}$$

(d)

$$\begin{aligned} I(N; X^N) =& H(N) - H(N|X^N) \\ =& H(N) - 0 \\ =& H_B(1/3) \\ =& \frac{1}{3}\log 3 + \frac{2}{3}\log\frac{3}{2} \\ =& 0.92 \end{aligned}$$

(e)

$$\begin{aligned} H(X^N|N) =& \frac{1}{3}H(X^6|N=6) + \frac{2}{3}H(X^{12}|N=12) \\ =& \frac{1}{3}H(X^6) + \frac{2}{3}H(X^{12}) \\ =& \frac{1}{3}6 + \frac{2}{3}12 \\ =& 10. \end{aligned}$$

(f)

$$\begin{aligned} H(X^N) =& I(X^N; N) + H(X^N|N) \\ =& I(X^N; N) + 10 \\ =& H_B(1/3) + 10 \\ =& 0.92 + 10 \\ =& 10.92 \end{aligned}$$

□

2 The Properties of Typical Sequences

Definition 1 *For $X \sim p(x)$ and $\epsilon \in (0,1)$, define the typical set as*

$$\mathcal{T}_\epsilon^{(n)}(X) = \{x^n : |\pi(x|x^n) - p(x)| \leq \epsilon p(x) for all x \in \mathcal{X}\}.$$

Lemma 1 **Typical Average Lemma:** *Let $x^n \in \mathcal{T}_\epsilon^{(n)}$. Then for any nonnegative function $g(x)$ on $\mathcal{X}$,*

$$(1-\epsilon)E(g(x)) \leq \frac{1}{n}\sum_{i=1}^{n} g(x_i) \leq (1+\epsilon)E(g(x)).$$

Typical sequences satisfy the following properties:

(a) Let $p(x^n) = \prod_{i=1}^{n} p_X(x_i)$. Then for each $x^n \in \mathcal{T}_\epsilon^{(n)}(X)$,

$$2^{-n(H(X)+\delta(\epsilon))} \leq p(x^n) \leq 2^{-n(H(X)-\delta(\epsilon))},$$

where $\delta(\epsilon) = \epsilon H(X)$.

(b) The cardinality of the typical set is upper bounded as

$$\left|\mathcal{T}_\epsilon^{(n)}\right| \leq 2^{n(H(X)+\delta(\epsilon))}.$$

(c) If $X_1, X_2, \cdots$ are i.i.d. with $X_i \sim p_X(x_i)$, then by the LLN,

$$\lim_{n\to\infty} P\{X^n \in \mathcal{T}_\epsilon^{(n)}\} = 1.$$

(d) The cardinality of the typical set is lower bounded as

$$\left|\mathcal{T}_\epsilon^{(n)}\right| \geq (1-\epsilon)2^{n(H(X)-\delta(\epsilon))}$$

for n sufficiently large.

Proof:

(a) Let's define a nonnegative function $g(x)$ as $g(x) = -\log p(x)$ due to $p(x) \in [0,1]$. According to typical average lemma, we have

$$(1-\epsilon)E(-\log p(x)) \leq \frac{1}{n}\sum_{i=1}^{n} -\log p(x_i) \leq (1+\epsilon)E(-\log p(x))$$

Note that $E(-\log p(x)) = -\sum_{x\in\mathcal{X}} p(x)\log p(x) = H(X)$, and $\sum_{i=1}^{n}\log p(x_i) = \log\prod_{i=1}^{n} p(x_i) = \log p(x^n)$ because $p(x^n) = \prod_{i=1}^{n} p_X(x_i)$. We have

$$\begin{aligned}(1-\epsilon)H(X) \leq & -\frac{1}{n}\log p(x^n) \leq (1+\epsilon)H(X)\\ -n(H(X)+\epsilon H(X)) \leq & \log p(x^n) \leq -n(H(X)-\epsilon H(X))\\ 2^{-n(H(X)+\delta(\epsilon))} \leq & \ p(x^n) \leq 2^{-n(H(X)-\delta(\epsilon))}\end{aligned}$$

where $\delta(\epsilon) = \epsilon H(X)$.

(b) Note that $\mathcal{T}_\epsilon^{(n)} \subset \mathcal{X}^n$.

$$\begin{aligned}1 =& \sum_{x^n \in \mathcal{X}^n} p(x^n)\\ \geq& \sum_{x^n \in \mathcal{T}_\epsilon^{(n)}} p(x^n)\\ \geq& \sum_{x^n \in \mathcal{T}_\epsilon^{(n)}} 2^{-n(H(X)+\delta(\epsilon))}\\ =& \left|\mathcal{T}_\epsilon^{(n)}\right| 2^{-n(H(X)+\delta(\epsilon))}\end{aligned}$$

Thus, $\left|\mathcal{T}_\epsilon^{(n)}\right| \leq 2^{n(H(X)+\delta(\epsilon))}$.

(c) According to the definition of typical set, to show that $\lim_{n\to\infty} P\{X^n \in \mathcal{T}_\epsilon^{(n)}\} = 1$ is equivalent to show that $\lim_{n\to\infty} P\left\{|\pi(x|x^n) - p(x)| \geq \epsilon\right\} = 0$ for all $x \in \mathcal{X}$.

For any $x \in \mathcal{X}$, let us define Y_i as $Y_i = \mathcal{I}_x(X_i)$, where $\mathcal{I}_x$ is the indicator function of x. Then we have

$$\pi(x|x^n) = \frac{Y_1 + Y_2 + \cdots + Y_n}{n}$$

and

$$\mu_Y = E[Y] = p(x)\mathcal{I}_x(X = x) + (1 - p(x))\mathcal{I}_x(X \neq x) = p(x).$$

According to LLN,

$$\lim_{n\to\infty} P\left\{|\pi(x|x^n) - p(x)| \geq \epsilon\right\} = \lim_{n\to\infty} P\left\{\left|\frac{Y_1 + Y_2 + \cdots + Y_n}{n} - \mu_Y\right| \geq \epsilon\right\} = 0$$

(d) According to property 3, there exists some $\epsilon > 0$ such that $P\{X^n \in \mathcal{T}_\epsilon^{(n)}\} = 1 - \epsilon$ when n is sufficiently large. Note that for each $x^n \in \mathcal{T}_\epsilon^{(n)}$, $p(x^n) \leq 2^{-n(H(X)-\delta(\epsilon))}$. We have

$$\begin{aligned}1 - \epsilon =& P\{X^n \in \mathcal{T}_\epsilon^{(n)}\}\\ =& \sum_{x^n \in \mathcal{X}^n} p(x^n)\\ \leq& \sum_{x^n \in \mathcal{X}^n} 2^{-n(H(X)-\delta(\epsilon))}\\ =& \left|\mathcal{T}_\epsilon^{(n)}\right| 2^{-n(H(X)-\delta(\epsilon))}\end{aligned}$$

Thus, $\left|\mathcal{T}_\epsilon^{(n)}\right| \geq (1 - \epsilon)2^{n(H(X)-\delta(\epsilon))}$.