Homework Set #1
Entropy rate, directed information

1. **Monotonicity of entropy per element.**

For a stationary stochastic process $X_1, X_2, \ldots, X_n$, show that

(a)
$$\frac{H(X_1, X_2, \ldots, X_n)}{n} \leq \frac{H(X_1, X_2, \ldots, X_{n-1})}{n-1}.$$

(b)
$$\frac{H(X_1, X_2, \ldots, X_n)}{n} \geq H(X_n|X_{n-1}, \ldots, X_1).$$

Solution: Monotonicity of entropy per element.

(a) By the chain rule for entropy,

$$\begin{aligned}\frac{H(X_1, X_2, \ldots, X_n)}{n} &= \frac{\sum_{i=1}^{n} H(X_i|X^{i-1})}{n} \\ &= \frac{H(X_n|X^{n-1}) + \sum_{i=1}^{n-1} H(X_i|X^{i-1})}{n} \\ &= \frac{H(X_n|X^{n-1}) + H(X^{n-1})}{n}. \qquad (1)\end{aligned}$$

From stationarity it follows that for all $1 \leq i \leq n$,

$$H(X_n|X^{n-1}) \leq H(X_i|X^{i-1}),$$

which further implies, by averaging both sides, that,

$$\begin{aligned}H(X_n|X^{n-1}) &\leq \frac{\sum_{i=1}^{n-1} H(X_i|X^{i-1})}{n-1} \\ &= \frac{H(X^{n-1})}{n-1}. \qquad (2)\end{aligned}$$

Combining (1) and (2) yields,

$$\begin{aligned}\frac{H(X_1,X_2,\ldots,X_n)}{n} &\le \frac{1}{n}\left[\frac{H(X_1,X_2,\ldots,X_{n-1})}{n-1}+H(X_1,X_2,\ldots,X_{n-1})\right]\\ &= \frac{H(X_1,X_2,\ldots,X_{n-1})}{n-1}.\end{aligned}$$

(b) By stationarity we have for all $1 \le i \le n$,

$$H(X_n|X^{n-1}) \le H(X_i|X^{i-1}),$$

which implies that,

$$\begin{aligned}H(X_n|X^{n-1}) &= \frac{\sum_{i=1}^{n} H(X_n|X^{n-1})}{n}\\ &\le \frac{\sum_{i=1}^{n} H(X_i|X^{i-1})}{n}\\ &= \frac{H(X_1,X_2,\ldots,X_n)}{n}.\end{aligned}$$

2. **Pairwise independence.**

Let $X_1, X_2, \ldots, X_{n-1}$ be i.i.d. random variables taking values in $\{0,1\}$, with $\Pr\{X_i = 1\} = \frac{1}{2}$. Let $X_n = 1$ if $\sum_{i=1}^{n-1} X_i$ is odd and $X_n = 0$ otherwise. Let $n \ge 3$.

(a) Show that X_i and X_j are independent, for $i \ne j, i, j \in \{1, 2, \ldots, n\}$.

(b) Find $H(X_i, X_j)$, for $i \ne j$.

(c) Find $H(X_1, X_2, \ldots, X_n)$. Is this equal to $nH(X_1)$?

Solution: Pairwise Independence.

$X_1, X_2, \ldots, X_{n-1}$ are i.i.d. Bernoulli(1/2) random variables. We will first prove that for any $k \le n-1$, the probability that $\sum_{i=1}^{k} X_i$ is odd is 1/2. We will prove this by induction. Clearly this is true for $k = 1$. Assume that it is true for $k-1$. Let $S_k = \sum_{i=1}^{k} X_i$. Then

$$\begin{aligned}P(S_k \text{ odd}) &= P(S_{k-1} \text{ odd})P(X_k = 0) && (3)\\ &+ P(S_{k-1} \text{ even})P(X_k = 1)\\ &= \frac{1}{2}\frac{1}{2}+\frac{1}{2}\frac{1}{2} && (4)\\ &= \frac{1}{2}. && (5)\end{aligned}$$

Hence for all $k \leq n-1$, the probability that S_k is odd is equal to the probability that it is even. Hence,

$$P(X_n = 1) = P(X_n = 0) = \frac{1}{2}. \tag{6}$$

(a) It is clear that when i and j are both less than n, X_i and X_j are independent. The only possible problem is when $j = n$. Taking $i = 1$ without loss of generality,

$$\begin{aligned}
P(X_1 = 1, X_n = 1) &= P(X_1 = 1, \sum_{i=2}^{n-1} X_i \text{ even}) & (7)\\
&= P(X_1 = 1) P(\sum_{i=2}^{n-1} X_i \text{ even}) & (8)\\
&= \frac{1}{2}\frac{1}{2} & (9)\\
&= P(X_1 = 1) P(X_n = 1) & (10)
\end{aligned}$$

and similarly for other possible values of the pair X_1, X_n. Hence X_1 and X_n are independent.

(b) Since X_i and X_j are independent and uniformly distributed on $\{0, 1\}$,

$$H(X_i, X_j) = H(X_i) + H(X_j) = 1 + 1 = 2 \text{ bits.} \tag{11}$$

(c) By the chain rule and the independence of $X_1, X_2, \ldots, X_{n_1}$, we have

$$\begin{aligned}
H(X_1, X_2, \ldots, X_n) &= H(X^{n-1}) + H(X_n|X^{n-1}) & (12)\\
&= \sum_{i=1}^{n-1} H(X_i) + 0 & (13)\\
&= n - 1, & (14)
\end{aligned}$$

since X_n is a function of the previous X_i's. The total entropy is not n, which is what would be obtained if the X_i's were all independent. This example illustrates that pairwise independence does not imply complete independence.

3. **Cesáro mean.**
Prove that if $\lim a_n = a$ and $b_n = \sum_{i=1}^{n} a_i$, then $\lim b_n = a$.

Solution: Cesáro mean.

Let $\varepsilon > 0$. Since $a_n \to a$, there exists a number $N(\varepsilon)$ such that $|a_n - a| \leq \varepsilon$ for all $n \geq N(\varepsilon)$. Furthermore,

$$\begin{aligned} |b_n - a| &= |\frac{1}{n}\sum_{i=1}^{n}(a_i - a)| \\ &\leq \frac{1}{n}\sum_{i=1}^{n}|(a_i - a)| \\ &\leq \frac{1}{n}\sum_{i=1}^{N(\varepsilon)}|(a_i - a)| + \frac{n - N(\varepsilon)}{n}\varepsilon \\ &\leq \frac{1}{n}\sum_{i=1}^{N(\varepsilon)}|(a_i - a)| + \varepsilon \end{aligned}$$

for all $n \geq N(\varepsilon)$. Since the first term goes to 0 as $n \to \infty$, we can make $|b_n - a| \leq 2\varepsilon$ by taking n large enough. Hence, $b_n \to a$ as $n \to \infty$.

4. **Stationary processes.**
Let $\ldots, X_{-1}, X_0, X_1, \ldots$ be a stationary (not necessarily Markov) stochastic process. Which of the following statements are true? State true or false. Then either prove or provide a counterexample. Warning: At least one answer is false.

(a) $H(X_n|X_0) = H(X_{-n}|X_0)$.

(b) $H(X_n|X_0) \geq H(X_{n-1}|X_0)$.

(c) $H(X_n|X_1^{n-1}, X_{n+1})$ is nonincreasing in n.

Solution: Stationary processes.

(a) $H(X_n|X_0) = H(X_{-n}|X_0)$.
This statement is true, since

$$\begin{aligned} H(X_n|X_0) &= H(X_n, X_0) - H(X_0) && (15) \\ H(X_{-n}|X_0) &= H(X_{-n}, X_0) - H(X_0) && (16) \end{aligned}$$

and $H(X_n, X_0) = H(X_{-n}, X_0)$ by stationarity.

(b) $H(X_n|X_0) \geq H(X_{n-1}|X_0)$.
Solution: This statement is not true in general, though it is true for first order Markov chains. A simple counterexample is a periodic process with period n. Let $X_0, X_1, X_2, \ldots, X_{n-1}$ be i.i.d. uniformly distributed binary random variables and let $X_k = X_{k-n}$ for $k \geq n$. In this case, $H(X_n|X_0) = 0$ and $H(X_{n-1}|X_0) = 1$, contradicting the statement $H(X_n|X_0) \geq H(X_{n-1}|X_0)$.

(c) $H(X_n|X_1^{n-1}, X_{n+1})$ is non-increasing in n.
This statement is true, since by stationarity $H(X_n|X_1^{n-1}, X_{n+1}) = H(X_{n+1}|X_2^n, X_{n+2}) \geq H(X_{n+1}|X_1^n, X_{n+2})$ where the inequality follows from the fact that conditioning reduces entropy.

5. **Directed Information and causal conditioning**

Directed information is denoted as $I(X^n \to Y^n)$ and is defined as

$$I(X^n \to Y^n) \triangleq \sum_{i=1}^{n} I(X^i; Y_i | Y^{i-1}).$$

Causal conditioning is denoted as $p(y^n||x^{n-d})$ and is defined as

$$p(y^n||x^{n-d}) \triangleq \prod_{i=1}^{n} p(y_i|y^{i-1}, x^{i-d})$$

a. Prove that

i.
$$I(X^n \to Y^n) \triangleq E\left[\log \frac{p(y^n||x^n)}{p(y^n)}\right]$$

ii.
$$p(y^n, x^n) = p(y^n||x^n)p(x^n||y^{n-1}),$$

iii.
$$I(X^n \to Y^n) \geq 0$$

and equals zero if and only if $p(y^n||x^n) = p(y^n)$

iv.
$$I(X^n; Y^n) = I(X^n \to Y^n) + I(0Y^{n-1} \to X^n),$$

where the term , $\emptyset Y^{n-1}$, denotes the concatenation of 0 (Null) to the sequence Y^{N-1}, i.e. $(\emptyset, Y_1, Y_2, ...Y_{n-1})$.

b. Prove that in general

$$I(X^n \to Y^n) \leq I(X^n; Y^n)$$

and equality holds if and only if $p(x^n||y^{n-1}) = p(x^n)$.

c. Suggest (without proof) properties similar to those of mutual information that should hold for directed information.

Solution: Directed Information and causal conditioning.

(a) i.

$$\begin{aligned} I(X^n \to Y^n) &= H(Y^n) - H(Y^n||X^n) \\ &= E\left[\log \frac{1}{P(y^n)}\right] - E\left[-\log P(y^n||x^n)\right] \\ &= E\left[\log \frac{P(y^n||x^n)}{P(y^n)}\right] \end{aligned}$$

ii.

$$\begin{aligned} p(y^n, x^n) &= \prod_{i=1}^{n} p(y_i, x_i | y^{i-1}, x^{i-1}) \\ &= \prod_{i=1}^{n} p(x_i | y^{i-1}, x^{i-1}) p(y_i | y^{i-1}, x^i) \\ &= p(x^n || y^{n-1}) p(y^n || x^n) \end{aligned}$$

iii. Directed information is a sum of non-negative terms, i.e. $I(X^i; Y_i | Y^{i-1}) \geq 0$ and therefore it is nonnegative. From the definition we see that directed information equals zero if and only if $p(y^n||x^n) = p(y^n)$ for all (x^n, y^n).

iv.

$$\begin{aligned} I(X^n; Y^n) &= E\left[\log \frac{p(Y^n, X^n)}{p(Y^n)p(X^n)}\right] \\ &= E\left[\log \frac{p(Y^n||X^n)p(X^n||X^{n-1})}{p(Y^n)p(X^n)}\right] \\ &= E\left[\log \frac{p(Y^n||X^n)}{p(Y^n)}\right] + E\left[\log \frac{p(X^n||Y^{n-1})}{p(X^n)}\right] \\ &= I(X^n \to Y^n) + I(Y^{n-1} \to X^n) \end{aligned}$$

(b) The inequality $I(X^n \to Y^n) \leq I(X^n; Y^n)$ follows immediately from the exercise in (a) and equality holds if and only if $p(x^n || y^{n-1}) = p(x^n)$. This condition is equivalent to the condition that feedback is not used in the encoding scheme.