Solutions: Homework Set # 2

Problem 1

We know that

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

Now, since $H(X|Y,Z) = 0$,

$$H(X,Y) \leq H(Y,Z),$$

and thus,

$$H(X) + H(Y) - I(X;Y) \leq H(Z) + H(Y|Z) \leq H(Z) + H(Y),$$

and so

$$I(X;Y) \geq H(X) - H(Z).$$

Problem 2

(a) In this part, we will see that $I(X_1;X_2;X_3)$ can be positive or negative. This says that we can have both $I(X_1;X_2) \geq I(X_1;X_2|X_3)$ and $I(X_1;X_2) < I(X_1;X_2|X_3)$ depending on X_1, X_2, X_3. so, while conditioning reduces entropy, it can increase or decrease mutual information.

- Let X_1 and X_2 be independent. Let $X_3 = X_1 + X_2$, Then

$$\begin{aligned}&I(X_1;X_2) = 0\\ &I(X_1;X_2|X_3) = I(X_1;X_2|X_1+X_2) = H(X_1|X_1+X_2) - H(X_1|X_1+X_2,X_2)\\ &= H(X_1|X_1+X_2) - 0 > 0.\end{aligned}$$

  So in this example $I(X_1;X_2;X_3) < 0$. Note that having X_1 and X_2 independent leads to $H(X_1|X_1+X_2)$ being strictly positive.

- Let $X_1 \to X_3 \to X_2$ form a Markov chain. Then

$$\begin{aligned}&I(X_1,X_2) \geq 0\\ &I(X_1;X_2|X_3) = H(X_1|X_3) - H(X_1|X_2,X_3) = 0.\end{aligned}$$

  So in this example $I(X_1;X_2;X_3) \geq 0$

(b) We first note that $I(X_1;X_2;X_3)$ is symmetric in X_1, X_2, and X_3:

$$\begin{aligned} I(X_1;X_2;X_3) &= I(X_1;X_2) - I(X_1;X_2|X_3) \\ &= H(X_1) - H(X_1|X_2) - H(X_1|X_3) + H(X_1|X_2,X_3) \\ &= I(X_1;X_3) - I(X_1;X_3|X_2) \end{aligned}$$

and similarly,

$$\begin{aligned} I(X_1;X_2;X_3) &= I(X_1;X_2) - I(X_1;X_2|X_3) \\ &= H(X_2) - H(X_2|X_1) - H(X_2|X_3) + H(X_2|X_1,X_3) \\ &= I(X_2;X_3) - I(X_2;X_3|X_1) \end{aligned}$$

Positivity of mutual information then concludes that $I(X_1;X_2;X_3) = I(X_1;X_2) - I(X_1;X_2|X_3) \leq I(X_1;X_2)$. Similarly $I(X_1;X_2;X_3) \leq I(X_1;X_3)$ and $I(X_1;X_2;X_3) \leq I(X_2;X_3)$.

(c) Positivity of mutual information concludes that $I(X_1;X_2;X_3) = I(X_1;X_2) - I(X_1;X_2|X_3) \geq -I(X_1;X_2|X_3)$. Similarly $I(X_1;X_2;X_3) \geq -I(X_1;X_3|X_2)$ and $I(X_1;X_2;X_3) \geq -I(X_2;X_3|X_1)$.

Problem 3

Note that in general, whenever you have a random variable you also have any deterministic function of that, i.e.,

$$I(A;B) = I(A;B,f(B)). \tag{1}$$

Also by replacing a random variable by its deterministic function, the mutual information does not exceed,

$$I(A;B) = I(A;B,f(B)) = I(A;f(B)) + I(A;B|f(B)) \geq I(A;f(B)), \tag{2}$$

where the inequality holds since $I(A;B|f(B)) \geq 0$.

$$\begin{aligned} I(X;\check{X} &\overset{(a)}{\leq} I(X;TZ) \\ &\overset{(b)}{\leq} I(X;SZ) \\ &\overset{(c)}{=} I(X;SZ\hat{X}) \\ &\overset{(d)}{=} I(X;\hat{X}) + I(X;SZ) \end{aligned}$$

where

- (a) follows from (2) since $\check{X} = f_4(T,Z)$,
- (b) is again by (2) and the fact that $T = f_3(Z)$,
- (c) is due (1) since $\hat{X} = f_2(S)$,
- and (d) is by chain rule for mutual information.

Remark: There are many other ways to prove this inequality, e.g., by expanding the mutual information as the difference of entropy functions. However, any correct solution needs to incorporate the properties of entropy or mutual information acting on functions, such as (1) and (2).

Note that although $H(f(B)|B) = 0$, but $H(B|f(B))$ can be positive (e.g., a constant function), and it is zero if and only if the function $f(\cdot)$ is injective.

Problem 4

(a) Let $P = (p_1, \cdots, p_i, \cdots, p_j, \cdots, p_n)$, $P_1 = \left(p_1, \cdots, \frac{p_i+p_j}{2} \cdots, \frac{p_i+p_j}{2}, \cdots p_n\right)$ and denote their entropies by $H(P)$ and $H(P_1)$ respectively.

$$\begin{aligned} H(P_1) - H(P) &= -2\frac{p_i+p_j}{2}\log\frac{p_i+p_j}{2} + (p_i \log p_i + p_j \log p_j) && (3) \\ &= -2\left(\frac{p_i+p_j}{2}\log\frac{p_i+p_j}{2} - \left(\frac{1}{2}p_i \log p_i + \frac{1}{2}p_j \log p_j\right)\right) && (4) \\ &\overset{(a)}{\geq} 0 && (5) \end{aligned}$$

where (a) follows by Jensen's inequality since $x \log x$ is a convex function function.

(b) Let $Q = PA = (q_1, \cdots, q_n)$. So each $q_i = \sum_j p_j a_{j,i}$.

$$\begin{aligned} H(P) &= -\sum_{i=1}^{n} p_i \log p_i && (6) \\ &\overset{(a)}{=} -\sum_{i=1}^{n}\sum_{j=1}^{n} a_{i,j} p_i \log p_i && (7) \\ &\overset{(b)}{=} -\sum_{j=1}^{n}\sum_{i=1}^{n} a_{i,j} p_i \log p_i && (8) \\ &\overset{(c)}{\leq} -\sum_{j=1}^{n}\sum_{i=1}^{n} a_{i,j} p_i \log \left(\sum_{i=1}^{n} a_{i,j} p_i\right) && (9) \\ &= -\sum_{j=1}^{n} q_j \log q_j && (10) \\ &= H(PA). && (11) \end{aligned}$$

- (a) follows from $\sum_{j=1}^{n} a_{i,j} = 1, \ \forall i = 1, \cdots, n$.
- (b) is obtained by changing the order of $\sum_i$ and $\sum_j$.
- (c) follows by Jensen's inequality and convexity of $x \log x$: $\sum_{i=1}^{n} a_{i,j} p_i \log p_i \geq q_j \log q_j$ (Note that $\sum_{i=1}^{n} a_{i,j} = 1, \forall j = 1, \cdots, n$).

Problem 5

(a) Since $T(x)$ is a statistic it is a function of X so it is completely determined by knowing X. This means that formally we can write

$$\Pr(T(x)|X, \theta) = \Pr(T(x)|X).$$

So we have the following Markov chain: $\theta \leftrightarrow X \leftrightarrow T(X)$. By the data processing inequality we can write

$$I(\theta; T(X)) \leq I(\theta; X),$$

for all distribution on the random variable θ.

(b) Let us define $\mathbf{X} \triangleq (X_1, \dots, X_n)$. To show that $T(\mathbf{X}) = (m, M)$ is a sufficient statistics for our problem, we will use the *Fisher-Neyman* factorization theorem. We will state the theorem for both discrete and continuous cases but only prove it for the discrete case as the proof for the continuous case needs some technicality.

Theorem 1 (Fisher-Neyman Factorization Theorem)**.** *Suppose that $\mathbf{X} = (X_1, \dots, X_n)$ has a joint density or frequency function $f(\mathbf{x}; \theta)$, $\theta \in \Theta$. A statistic $T = T(\mathbf{X})$ is sufficient for θ if and only if*

$$f(\mathbf{x}; \theta) = g(T(\mathbf{x}), \theta) h(\mathbf{x}).$$

Proof. First, suppose that T is a sufficient statistics. We can write

$$f(\mathbf{x}; \theta) = \Pr[\mathbf{X} = \mathbf{x}] = \sum_t \Pr[\mathbf{X} = \mathbf{x}, T = t].$$

But T is a function of $\mathbf{x}$ so $\Pr[\mathbf{X} = \mathbf{x}, T = t]$ is non-zero only if $t = T(\mathbf{x})$. Then we can write

$$f(\mathbf{x}; \theta) = \Pr[\mathbf{X} = \mathbf{x}, T = T(\mathbf{x})] = \Pr[T = T(\mathbf{x})] \Pr[\mathbf{X} = \mathbf{x} | T = T(\mathbf{x})].$$

Since T is sufficient, $\Pr[\mathbf{X} = \mathbf{x} | T = T(\mathbf{x})]$ is independent of θ and so we have $f(\mathbf{x}; \theta) = g(T(\mathbf{x}); \theta) h(\mathbf{x})$.

Now suppose that $f(\mathbf{x}; \theta) = g(T(\mathbf{x}); \theta) h(\mathbf{x})$. Let us assume $T(\mathbf{x}) = t$ and write

$$\Pr[\mathbf{X} = \mathbf{x} | T = t] = \frac{\Pr[\mathbf{X} = \mathbf{x}, T = t]}{P[T = t]}.$$

Because $T(\mathbf{x})$ is a function of $\mathbf{x}$ we can write $\Pr[\mathbf{X} = \mathbf{x}, T = t] = \Pr[\mathbf{X} = \mathbf{x}] \cdot 1_{\{T(\mathbf{x}) = t\}}$. This let us write

$$\begin{aligned} \Pr[\mathbf{X} = \mathbf{x} | T = t] &= \frac{\Pr[\mathbf{X} = \mathbf{x}] \cdot 1_{\{T(\mathbf{x}) = t\}}}{P[T = t]} \\ &= \frac{g(T(\mathbf{x}); \theta) h(\mathbf{x}) \cdot 1_{\{T(\mathbf{x}) = t\}}}{\sum_{\mathbf{y}: T(\mathbf{y}) = t} g(T(\mathbf{y}); \theta) h(\mathbf{y})} \\ &= \frac{h(\mathbf{x}) \cdot 1_{\{T(\mathbf{x}) = t\}}}{\sum_{\mathbf{y}: T(\mathbf{y}) = t} h(\mathbf{y})}, \end{aligned}$$

which does not depend on θ and we are done. □

Back to our problem, for each random variable X_i we can write the density function as follows

$$f_{X_i}(x_i; \theta) = u(\theta + 1 - x_i) \cdot u(x_i - \theta),$$

where $u(x)$ is the step function

$$u(x) = \begin{cases} 1 & x \geq 0, \\ 0 & x < 0. \end{cases}$$

Because observations are independent from each other ($X_1, \dots, X_n$, are independent from each other) for the joint density function $f_{\mathbf{X}}(\mathbf{x};\theta)$ we have

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x};\theta) &= \prod_{i=1}^{n} f_{X_i}(x_i;\theta) \\ &= \prod_{i=1}^{n} u(\theta+1-x_i)\cdot u(x_i-\theta) \\ &= \underbrace{u(\theta+1-\max(\mathbf{x}))\cdot u(\min(\mathbf{x})-\theta)}_{g(T(\mathbf{x});\theta)}\cdot \underbrace{1}_{h(\mathbf{x})}, \end{aligned}$$

so by the Fisher-Neyman factorization theorem $T(\mathbf{x}) = (m, M)$ is a sufficient statistics for θ.