## Information Theory (0432 L 654) – WS 2016-2017

# Problem Set 1 with Solutions

1. Prove that if $P_{X,Y,Z} = P_Z P_{X|Z} P_{Y|X}$ then

$$I(X;Y) \geq I(X;Y|Z)$$

**Solution:** Using the chain rule we have:

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

where (1) follows from the fact that $Z \to X \to Y$ is a Markov chain, therefore $Z$ and $Y$ are conditionally statistically independent given $X$. Since $I(Z;Y) \geq 0$, we have the result.

2. Provide a direct proof (without exploiting the convexity of divergence) that the entropy function $\mathcal{H}(\mathbf{p})$ (seen as function of the $|\mathcal{X}|$-dimensional probability vector $\mathbf{p}$) is a concave function of $\mathbf{p}$. *Hint: consider the matrix of second-order derivatives.*

**Solution:** Let's write the entropy as a function of the probability vector $\mathbf{p} = (p_1, \ldots, p_M)$ with $M = |\mathcal{X}|$. For simplicity, we consider logarithm base-$e$. We have

$$\mathcal{H}(\mathbf{p}) = \sum_i p_i \log \frac{1}{p_i}$$

The gradient of the entropy is the vector

$$\nabla \mathcal{H}(\mathbf{p}) = \left( \frac{\partial \mathcal{H}(\mathbf{p})}{\partial p_1}, \ldots, \frac{\partial \mathcal{H}(\mathbf{p})}{\partial p_M} \right)$$

with $i$-th element

$$\frac{\partial \mathcal{H}(\mathbf{p})}{\partial p_i} = \log \frac{1}{p_i} - 1$$

The matrix of second derivatives $\nabla \times \nabla \mathcal{H}(\mathbf{p})$ with elements $\frac{\partial^2 \mathcal{H}(\mathbf{p})}{\partial p_i \partial p_j}$ is diagonal with $i$-th diagonal elements

$$\frac{\partial^2 \mathcal{H}(\mathbf{p})}{\partial p_i^2} = -\frac{1}{p_i}$$

Hence, $\nabla \times \nabla \mathcal{H}(\mathbf{p})$ is negative definite for any probability vector $\mathbf{p}$. Since the domain of probability vectors is convex, then $\mathcal{H}(\mathbf{p})$ is a strictly concave function.

3. Let $\mathbf{p}, \mathbf{q}$ be two probability vectors defined on the same alphabet $\mathcal{X}$. Prove that $D(\mathbf{p}\|\mathbf{q})$ is a convex function of the pair $(\mathbf{p}, \mathbf{q})$. *Hint: first, you can prove the "log-sum inequality":*

$$\sum_i a_i \log \frac{a_i}{b_i} \geq \left( \sum_i a_i \right) \log \frac{\sum_i a_i}{\sum_i b_i}$$

*for non-negative numbers $\{a_i\}$ and $\{b_i\}$. Then, the results follows almost immediately as an application.*

**Solution:** We have

$$\begin{aligned}\sum_i \frac{b_i}{\sum_j b_j}\left(\frac{a_i}{b_i}\log\frac{a_i}{b_i}\right) &\geq \left(\sum_i \frac{b_i}{\sum_j b_j}\frac{a_i}{b_i}\right)\log\left(\sum_i \frac{b_i}{\sum_j b_j}\frac{a_i}{b_i}\right)\\ &= \frac{\sum_i a_i}{\sum_j b_j}\log\left(\frac{\sum_i a_i}{\sum_j b_j}\right)\end{aligned}$$

where (1) follows by applying Jensen's inequality to the convex function $f(t) = t\log t$.

Armed with the log-sum inequality, take two paris of probability vectors $(\mathbf{p}_1, \mathbf{q}_1)$ and $(\mathbf{p}_2, \mathbf{q}_2)$ and $\lambda \in [0,1]$, and write:

$$\begin{aligned}D(\lambda\mathbf{p}_1 + (1-\lambda)\mathbf{p}_2 \| \lambda\mathbf{q}_1 + (1-\lambda)\mathbf{q}_2) &= \sum_i (\lambda p_{1,i} + (1-\lambda)p_{2,i})\log\left(\frac{\lambda p_{1,i} + (1-\lambda)p_{2,i}}{\lambda q_{1,i} + (1-\lambda)q_{2,i}}\right)\\ &\leq \sum_i \left(\lambda p_{1,i}\log\frac{\lambda p_{1,i}}{\lambda q_{1,i}} + (1-\lambda)p_{2,i}\log\frac{(1-\lambda)p_{2,i}}{(1-\lambda)q_{2,i}}\right)\\ &= \lambda D(\mathbf{p}_1\|\mathbf{q}_1) + (1-\lambda)D(\mathbf{p}_2\|\mathbf{q}_2)\end{aligned}$$

where in (1) we applied the log-sum inequality to each $i$-th term of the sum with respect to $i$.

4. Let $\mathbf{p}$ denote a probability vector on $\mathcal{X}$, with $x$-th element $P_X(x)$ for all $x \in \mathcal{X}$, and $\mathbf{P}$ denote a probability transition matrix on $\mathcal{X} \times \mathcal{Y}$, i.e., a matrix with $(x,y)$-th element $P_{Y|X}(y|x)$, for all pairs $(x,y) \in \mathcal{X} \times \mathcal{Y}$. Let $I(X;Y) = \mathcal{I}(\mathbf{p}, \mathbf{P})$ be the mutual information, seen as a function of $\mathbf{p}$ and $\mathbf{P}$. Prove (without neglecting any steps) that $I(\mathbf{p}, \mathbf{P})$ is concave with respect to $\mathbf{p}$ for fixed $\mathbf{P}$, and convex with respect to $\mathbf{P}$ for fixed $\mathbf{p}$.

   **Solution:** Let $p_i$ denote the $i$-th element of $\mathbf{p}$ and $P_{i,j} = P_{Y|X}(y=j|x=i)$ denote the $(i,j)$ element of $\mathbf{P}$. Let also $\mathbf{q}$ denote the probability vector corresponding to $P_Y$, i.e., $\mathbf{q} = \mathbf{pP}$, with $j$-th element

$$q_j = \sum_i p_i P_{i,j}. \tag{1}$$

   Then, we have

$$\begin{aligned}\mathcal{I}(\mathbf{p}, \mathbf{P}) &= H(Y) - H(Y|X)\\ &= \mathcal{H}(\mathbf{q}) + \sum_i p_i \sum_j P_{i,j}\log P_{i,j}.\end{aligned}$$

   We notice that $\mathcal{H}(\mathbf{q})$ is concave in $\mathbf{q}$ and that $\mathbf{q} = \mathbf{pP}$ is linear in $\mathbf{p}$ for fixed $\mathbf{P}$. Therefore, since the composition of a concave function and a linear function is concave, we have that $\mathcal{H}(\mathbf{q})$ is a concave function of $\mathbf{p}$. Furthermore, the second term in (2) for fixed $\mathbf{P}$ is a linear function of $\mathbf{p}$. Since the sum of a concave function and a linear function is concave, we conclude that $\mathcal{I}(\mathbf{p}, \mathbf{P})$ is concave in $\mathbf{p}$ for fixed $\mathbf{P}$.

   In order to show convexity in $\mathbf{P}$ for fixed $\mathbf{p}$, we write

$$\begin{aligned}\mathcal{I}(\mathbf{p}, \mathbf{P}) &= H(X) + H(Y) - H(X,Y)\\ &= \mathcal{H}(\mathbf{p}) + \mathcal{H}(\mathbf{q}) + \sum_{i,j} p_i P_{i,j}\log(p_i P_{i,j}).\end{aligned}$$

Concluding convexity from the above expression is not immediate, since $\mathcal{H}(\mathbf{p})$ is just a constant (for fixed $\mathbf{p}$), $\mathcal{H}(\mathbf{q})$ is concave in $\mathbf{q}$ and $\mathbf{q}$ is linear in $\mathbf{P}$, therefore $\mathcal{H}(\mathbf{q})$ is concave in $\mathbf{P}$, however, $\sum_{i,j} p_i P_{i,j} \log(p_i P_{i,j})$ is convex in $\mathbf{P}$ since the function $t \log t$ is convex. Hence, we have the sum of a concave and a convex function, from which we cannot conclude anything a priori. Hence, we consider the matrix of second derivatives with respect to $\mathbf{P}$. It is not difficult to show that

$$\frac{\partial^2}{\partial P_{r,s} \partial P_{\ell,m}} \mathcal{H}(\mathbf{q}) = \left\{ \begin{array}{ll} -\frac{p_\ell^2}{\sum_i p_i P_{i,m}} & \text{for } (r,s) = (\ell, m) \\ 0 & \text{for } (r,s) \neq (\ell, m) \end{array} \right. \tag{2}$$

and that

$$\frac{\partial^2}{\partial P_{r,s} \partial P_{\ell,m}} \sum_{i,j} p_i P_{i,j} \log(p_i P_{i,j}) = \left\{ \begin{array}{ll} \frac{p_\ell}{P_{\ell,m}} & \text{for } (r,s) = (\ell, m) \\ 0 & \text{for } (r,s) \neq (\ell, m) \end{array} \right. \tag{3}$$

Hence, the matrix of second derivatives is diagonal, with diagonal elements

$$\frac{\partial^2}{\partial P_{\ell,m}^2} \mathcal{I}(\mathbf{p}, \mathbf{P}) = p_\ell^2 \left( \frac{1}{p_\ell P_{\ell.m}} - \frac{1}{\sum_i p_i P_{i,m}} \right) \geq 0. \tag{4}$$

This is clearly non-negative definite, and therefore the function $\mathcal{I}(\mathbf{p}, \mathbf{P})$ is convex fin $\mathbf{P}$ for fixed $\mathbf{p}$.

A more elegant proof based on the convexity of divergence is provided in Cover and Thomas textbook at page 33.

5. A fair coin is flipped until the first "head" occurs. Let $X$ denote the resulting number of flips. a) Find $H(X)$; b) Find an "efficient" sequence of yes-no questions of the form:

   *Is $X$ contained in the set $S$?*

   in order to determine the value of $X$. Compare $H(X)$ (expressed in bits, i.e., use log base-2) with the *expected* number of questions in your scheme, necessary to determine $X$.

   **Solution:** Let $q = 1 - p$, then the distribution of $X$ is $P_X(x) = pq^{x-1}$ for $x = 1, 2, 3 \ldots$. This is a Geometric distribution with success probability $p$. We have

$$\begin{aligned} H(X) &= -\sum_{x=1}^{\infty} q^{x-1} p \log(q^{x-1} p) \\ &= -p \log(p) \sum_{x=1}^{\infty} q^{x-1} - p \log(q) \sum_{x=1}^{\infty} (x-1) q^{x-1} \\ &= -p \log(p) \sum_{n=0}^{\infty} q^n - p \log(q) \sum_{n=0}^{\infty} n q^n \\ &= -p \log(p) \frac{1}{1-q} - p \log(q) \frac{q}{(1-q)^2} \\ &= \frac{-1}{p} \left[ p \log p + (1-p) \log(1-p) \right] \\ &= \frac{1}{p} \mathcal{H}(p) \end{aligned}$$

   where the last line follows by applying the sum formulas provided as hints in the problem. For a fair coin flip, we have $p = 1/2$ and $\mathcal{H}(1/2) = 1$ bit, therefore $H(X) = 2$ bits.

In order to answer the second question, we notice that question of the type
Q1: "Is $X = 1$"
Q2: "Is $X = 2$"
Q3: "Is $X = 3$"
etc ... arrive at determining the value of $X$ in $X$ steps, therefore the average number of questions is $\mathbb{E}[X]$. In general, $\mathbb{E}[X]$ and $H(X)$ have no particular relation, but in this case, for a fair coin flip, we have $\mathbb{E}[X] = 2 = H(X)$. Hence, this sequence of questions is as efficient as possible to determine $X$ (it is in fact a Huffman code for $X$).

6. Let $X$ denote a discrete random variable taking on values in the set $\mathcal{X}$. Let $g(x)$ be a (deterministic) function defined on $\mathcal{X}$. We wish to show that the entropy of a function of a random variable is always not larger than the entropy of the random variable itself. In order to do so, justify the following equality/inequality:

$$H(X, g(X)) = H(X)$$
$$H(X, g(X)) \geq H(g(X))$$

**Solution:** For the first equality, use the chain rule of entropy $H(X, g(X)) = H(X) + H(g(X)|X)$ and notice that when conditioning on $X$, $g(X)$ is a constant, such that $H(g(X)|X) = 0$. For the second inequality, use the chain rule in the other order, and the fact that entropy is non-negative.

7. An urn contains $r$ red, $w$ white, and $b$ blue balls. Let $X$ denote the result of drawing $k \geq 2$ balls from the urn with replacement (i.e., after picking a ball, the ball is put back in the urn before the next pick), and $Y$ denote the result of drawing $k \geq 2$ balls from the urn without replacement (i.e., after picking a ball, this is eliminated from the urn). Compare the entropies $H(X)$ and $H(Y)$. Which is larger?

**Solution:** Define the experiment of drawing $k$ balls from the urn with or without replacement. The problem is somehow ambiguous, since we have to define a random variable corresponding to this random experiment. In particular, there are two natural ways of defining this random variable: ordered sequence of balls, or unordered sequence of balls. In the first case, two sequences with the same composition but different order are counted as two different outcomes. In the second case, they are counted as the same outcome. Since the problem does not specify, we are entitled to choose the easiest definition, and we shall choose the case of *ordered* sequences.

Let $n = r + w + b$. Define $X^k = (X_1, X_2, \ldots, X_k)$ as the ordered sequence of balls with replacement Since we draw with replacement, the successive draws from the urn are statistically independent. Therefore,

$$H(X^k) = \sum_{i=1}^{k} H(X_i) = kH(X_1)$$

where

$$H(X_1) = \mathcal{H}\left(\frac{r}{n}, \frac{w}{n}, \frac{b}{n}\right).$$

Next, define $Y^k = (Y_1, Y_2, \ldots, Y_k)$ denote the ordered sequence of drawings *without replacement.* Now, the successive draws from the urn are statistically dependent. Hence, we have

$$H(Y^k) = \sum_{i=1}^{k} H(Y_i|Y_{i-1}) \leq \sum_{i=1}^{k} H(Y_i).$$

Clearly, $P_{Y_1} = P_{X_1}$ since the initial state of the urn is the same in both cases. Consider step 2 of the sampling without replacement. We sample from the urn having $n - 1$. The domain of each $Y_i$ is $\{R, W, B\}$, where we assume $k \leq n$ (otherwise, we have to introduce a dummy "empty" symbol for $k > n$ when we sample from an empty urn. This case is easily handled and it is not considered here). We have

$$\begin{aligned} P_{Y_2}(R) &= P_{Y_2|Y_1}(R|R)P_{Y_1}(R) + P_{Y_2|Y_1}(R|W)P_{Y_1}(W) + P_{Y_2|Y_1}(R|B)P_{Y_1}(B) \\ &= \frac{r-1}{n-1}\frac{r}{n} + \frac{r}{n-1}\frac{w}{n} + \frac{r}{n-1}\frac{b}{n} \\ &= \frac{(r-1)r + rw + rb}{n(n-1)} \\ &= \frac{r}{n} = P_{Y_1}(R) \end{aligned}$$

Similarly, we find $P_{Y_2}(W) = P_{Y_1}(W)$ and $P_{Y_2}(B) = P_{Y_1}(B)$. By induction, we have that for any $1 \leq i \leq k$, with $k \leq n$, we have $P_{Y_i} = P_{X_i}$. Therefore, we have

$$H(Y^k) \leq \sum_{i=1}^{k} H(Y_i) = \sum_{i=1}^{k} H(X_i) = H(X^k).$$

We conclude that the sequence generated by sampling with replacement has a higher entropy than the sequence generated by sampling without replacement.

8. Let $X_1$ and $X_2$ denote two discrete random variables distributed according to their own probability mass functions $P_{X_1}$ and $P_{X_2}$, defined on the disjoint sets $\mathcal{X}_1$ and $\mathcal{X}_2$, respectively. Let

$$X = \left\{ \begin{array}{ll} X_1 & \text{with probability } \alpha \\ X_2 & \text{with probability } 1 - \alpha \end{array} \right.$$

(such type of random variable is called a "disjoint mixture" of $X_1$ and $X_2$ with mixing distribution $(\alpha, 1 - \alpha)$). a) Find $H(X)$ in terms of $H(X_1)$, $H(X_2)$ and $\alpha$. b) Maximize the obtained expression with respect to $\alpha$ and show that

$$2^{H(X)} \leq 2^{H(X_1)} + 2^{H(X_2)}$$

**Solution:** The key observation is that the two alphabets $\mathcal{X}_1$ and $\mathcal{X}_2$ are disjoint. Hence, we can extend $P_{X_1}(\cdot)$ to take value zero on symbols $x \in \mathcal{X}_2$ and, similarly, $P_{X_2}(\cdot)$ to take value zero on the symbols $x \in \mathcal{X}_1$. It follows that the pmf of $X$ can be written as

$$P_X(x) = \alpha P_{X_1}(x) + (1 - \alpha)P_{X_2}(x)$$

defined on the union alphabet $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2$. Then, we can write

$$\begin{aligned} H(X) &= \sum_{x \in \mathcal{X}} P_X(x) \log \frac{1}{P_X(x)} \\ &= \sum_{x \in \mathcal{X}_1} \alpha P_{X_1}(x) \log \frac{1}{\alpha P_{X_1}(x)} + \sum_{x \in \mathcal{X}_1} (1 - \alpha)P_{X_2}(x) \log \frac{1}{(1 - \alpha)P_{X_2}(x)} \\ &= \alpha H(X_1) + (1 - \alpha)H(X_2) + \mathcal{H}_2(\alpha) \end{aligned}$$

where $\mathcal{H}_2(\alpha) = \alpha \log(1/\alpha) + (1 - \alpha) \log(1/(1 - \alpha))$ is the binary entropy function.

Differentiating and setting the derivative to zero, we find the solution

$$\alpha^* = \frac{2^{H(X_1)}}{2^{H(X_1)} + 2^{H(X_2)}}$$

Notice that this corresponds necessarily to a maximum, since it is a stationary point of $H(X)$ with respect to $\alpha = (0, 1)$, and it is easy to see (by usual concavity arguments) that $H(X)$ is concave in $\alpha \in [0, 1]$.

Replacing $\alpha^*$ into the expression of $H(X)$, after some algebra, we obtain

$$\max_{\alpha} H(X) = \log\left(2^{H(X_1)} + 2^{H(X_2)}\right).$$

Since obviously we have that, for any fixed value of $\alpha$, $H(X) \leq \max_\alpha H(X)$, by taking exponential of both sides we obtain the upper bound:

$$2^{H(X)} \leq 2^{H(X_1)} + 2^{H(X_2)}.$$

9. Let $X_1 \to X_2 \to X_3$ be a Markov chain, and let $X_1, X_2, X_3$ be discrete random variables defined on alphabets of size $n, k$ and $m$ respectively, with $k \leq \min\{n, m\}$. Show that $I(X_1; X_3) \leq \log_2 k$. What happens for $k = 1$? Can you interpret this fact?

   **Solution:** By the data processing inequality, we have

   $$I(X_1; X_3) \leq I(X_2; X_3) = H(X_2) - H(X_2|X_3) \leq H(X_2) \leq \log k$$

   Clearly, for $k = 1$ we must have $I(X_1; X_3) = 0$. Interpretation: since for $k = 1$ $X_2$ is a deterministic constant, $X_1$ and $X_3$ in this case are statistically independent.

10. The World Series is a seven-game series that terminates as soon as one team has won four games. Let $X$ be the random variable that represents the outcome of a World Series sequence between teams $A$ and $B$. Possible outcomes for $X$ are, for example, $(AAAA)$, $(ABAAA)$, $(BABABAB)$, etc ... Let $Y$ be the number of games played, which ranges from 4 to 7. Assuming that teams $A$ and $B$ are equally likely to win games, and games are statistically independent events, calculate (in bits) $H(X), H(Y), H(Y|X)$ and $H(X|Y)$.

    **Solution:** The key of the problem is to calculate the joint probability distribution $P_{X,Y}(x, y)$, from which all the other probability distributions can be obtained and, as a consequence, we can calculate the entropies $H(X), H(Y), H(Y|X)$ and $H(X|Y)$. We can use simple counting arguments. We have $2^7$ possible sequences of results. For a sequence $x = (x_1, x_2, \ldots, x_7)$ of A's and B's we define $y(x)$ as the smallest index $i$ such that the subsequence $(x_1, \ldots, x_i)$ contains four A's or four B's. It is clear that $y(\bar{x}) = y(x)$, where $\bar{x}$ denote the sequence obtained by complementing all elements of $x$ (i.e., turning A into B and B into A).

    Hence, we shall focus on the case where we get four A's. A sequence $x$ such that $y(x) = k$ for $k = 4, 5, 6, 7$ and such that $(x_1, \ldots, x_k)$ contains four A's must have $x_k = A$ and the remaining three A's arranged in some way into the remaining $k - 1$ positions. Hence, we have

    $$\binom{k-1}{3}$$

    such arrangements with $y(x) = k$ when the team A wins, and the same number in the case team $B$ wins. Hence, the total number of such arrangements with $y(x) = k$ is $2\binom{k-1}{3}$. For each such arrangements, the remaining (virtual) results from $k+1$ to 7 can be arranged in any order, therefore the number of sequences $x$ with $y(x)=k$ are

$$2\binom{k-1}{3}2^{7-k}$$

Since we have a total of $2^7$ equiprobable sequences of results, we have

$$\mathbb{P}(y(X)=y)=\frac{2\binom{y-1}{3}2^{7-y}}{2^7}=\binom{y-1}{3}2^{-y+1}$$

This results in the following probabilities for the random variable $Y=y(X)$:

$$\begin{aligned}\mathbb{P}(Y=4) &= \binom{3}{3}2^{-3}=\frac{1}{8}\\ \mathbb{P}(Y=5) &= \binom{4}{3}2^{-4}=\frac{1}{4}\\ \mathbb{P}(Y=6) &= \binom{5}{3}2^{-5}=\frac{5}{16}\\ \mathbb{P}(Y=7) &= \binom{6}{3}2^{-6}=\frac{5}{16}\end{aligned}$$

Note that $\mathbb{P}(Y=y)=0$ for $0\le y\le 3$ since you can not possibly have 4 wins if you play less than 4 games.

Note also that $P_{X|Y}(x|y)$ where $y\in\{4,5,6,7\}$ can be calculated by observing that for a given length $y$, the possible sequences $x$ are equiprobable. Hence,

$$P_{X|Y}(x|y)=\frac{1}{2\binom{y-1}{3}}\,1\{y(x)=y\}.$$

Finally, the joint distribution is given by

$$P_{X,Y}(x,y)=2^{-y}\,1\{y(x)=y\}.$$

This means that for any valid sequence $x$ with length $y(x)=y$ (i.e., such that it has the $y$-th symbol equal to $A$ and 3 previous A's and less than 4 B's, or the $y$-th symbol equal to $B$ and 2 previous B's and less than 4 A's), the probability is simply $2^{-y}$. Otherwise, the probability is equal to zero, since such combination of $x$ and $y$ is not possible.

It follows that:

$$H(X)=\sum_{k=4}^{7}k\binom{k-1}{3}2^{-(k-1)}=\frac{4}{8}+\frac{5}{4}+\frac{30}{16}+\frac{35}{16}=\frac{93}{16}.$$

$$H(Y)=\frac{1}{8}\log 8+\frac{1}{4}\log 4+\frac{5}{16}\log\frac{16}{5}+\frac{5}{16}\log\frac{16}{5}=\frac{54}{16}-\frac{10}{16}\log 5.$$

$$
\begin{aligned}
H(X|Y) &= \sum_{k=4}^{7} \binom{k-1}{3} 2^{-(k-1)} \sum_{x:y(x)=k} \frac{1}{2\binom{k-1}{3}} \log\left(2\binom{k-1}{3}\right) \\
&= \sum_{k=4}^{7} 2^{-k} \sum_{x:y(x)=k} \log\left(2\binom{k-1}{3}\right) \\
&= \sum_{k=4}^{7} \binom{k-1}{3} 2^{-(k-1)} \log\left(2\binom{k-1}{3}\right) \\
&= \frac{39}{16} + \frac{10}{16}\log 5.
\end{aligned}
$$

Obviously, since $Y$ is a function of $X$, we have $H(Y|X) = 0$. Notice that, by using the chain rule in two ways, we have

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

Hence, as a quick check, we have:

$$H(X) = H(Y) + H(X|Y) = \frac{54}{16} - \frac{10}{16}\log 5 + \frac{39}{16} + \frac{10}{16}\log 5 = \frac{93}{16}.$$

11. Let $P_X(x)$ denote a probability mass function. Prove that, for all $d \geq 0$, the following inequality holds:

$$\mathbb{P}(P_X(X) \leq d) \log \frac{1}{d} \leq H(X)$$

**Solution:** First, notice that for $d \geq 1$ the inequality is trivially true. Then, we consider only the non-trivial case of $0 < d < 1$. In this case, we can re-write the above inequality as

$$\mathbb{P}\left(\log \frac{1}{P_X(X)} \geq \log \frac{1}{d}\right) \leq \frac{H(X)}{\log \frac{1}{d}}$$

Then, we notice that this is just an application of Markov inequality: for any non-negative function $g(X)$, we have

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

12. Let $X, Y, Z$ be jointly distributed discrete random variables. Prove the following inequalities:

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

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

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

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

**Solution:** We have:
Inequality a) Using the chain rule we can write

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

with equality when $Y$ is a deterministic function of $X, Z$.
Inequality b) We can write

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

with equality when $Y$ and $Z$ are conditionally independent given $X$, i.e., $Y \to X \to Z$ is a Markov chain.
Inequality c) Using the chain rule we can write

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

which is obviously true, and equality is achieved when $Z$ is independent of $Y$ given $X$.
Inequality d) Rearranging terms and using the chain rule we obtain

$$\begin{aligned} I(X;Z|Y) + I(Z;Y) &\geq I(Z;Y|X) + I(X;Z) \\ I(X,Y;Z) &\leq I(X,Y;Z) \end{aligned}$$

where equality holds always.

13. How much information does the length of a sequence give about its content? Suppose $X_i$, for $i = 1, 2, 3, \ldots$ are i.i.d. Bernoulli-1/2 random variables. Consider the process $\{X_i\}$ that stops when the first 1 appears. Let $N$ denote the length of the resulting sequence (stopping time), such that the realization of the stopped process is denotes by $X^N$.

    a) Find $I(N; X^N)$.
    b) Find $H(X^N|N)$.
    c) Find $H(X^N)$.

    **Solution:** Let $X_i$ be independent Bernoulli-$p$ random variables (the problem deals with $p = 1/2$ but everything can be done in general here), and let $N$ the stopping time of the first occurrence of a "1". We have
    a)
    $$I(N; X^N) = H(N) - H(N|X^N) = H(N) = \mathcal{H}(p)/p$$
    where this follows form the fact that $N$ is a function of $X^N$ (it is its length) and that $N$ is geometrically distributed with probability of success $p$.
    b)
    $$H(X^N|N) = 0$$
    since $X^N$ takes on values in the set $\{1, 01, 001, 0001, \ldots\}$ and therefore it is a function of $N$.
    c) Obviously, we have $H(X^N) = I(N; X^N) = \mathcal{H}(p)/p$.