# EE276: Homework #2 Solutions

Due on Friday Jan 23, 6pm - Gradescope entry code: E6VP4X

1. **Data Processing Inequality.**
   The random variables $X$, $Y$ and $Z$, belonging to alphabets $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$ respectively, form a Markov triplet $(X - Y - Z)$ if $p(z|y) = p(z|y, x)$, or, equivalently, if $p(x|y) = p(x|y, z)$. If $X$, $Y$, $Z$ form a Markov triplet $(X - Y - Z)$, show that:

   (a) $H(X|Y) = H(X|Y, Z)$ and $H(Z|Y) = H(Z|X, Y)$

   (b) $H(X|Y) \leq H(X|Z)$

   (c) $I(X; Y) \geq I(X; Z)$ and $I(Y; Z) \geq I(X; Z)$

   (d) $I(X; Z) \leq \log |\mathcal{Y}|$

   (e) $I(X; Z|Y) = 0$

   where the *conditional mutual information* of random variables $X$ and $Y$ given $Z$ is defined by

$$
\begin{aligned}
I(X; Y|Z) &= H(X|Z) - H(X|Y, Z) \\
&= \sum_{x,y,z} p(x, y, z) \log \frac{p(x, y|z)}{p(x|z)p(y|z)}
\end{aligned}
$$

**Solution: Data Processing Inequality.**

(a)

$$
\begin{aligned}
H(X|Y) &= \sum_{x,y} -p(x, y) \log(p(x|y)) \\
&= \sum_{x,y,z} -p(x, y, z) \log(p(x|y)) \\
&= \sum_{x,y,z} -p(x, y, z) \log(p(x|y, z)) \\
&= H(X|Y, Z)
\end{aligned}
$$

where the third equality uses the fact that $X$ and $Z$ are conditionally independent given $Y$. A similar argument can be used to show $H(Z|Y) = H(Z|X, Y)$.

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

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

(d) $I(X; Z) \leq I(X; Y) \leq H(Y) \leq \log |\mathcal{Y}|$

(e) We showed that $H(X|Y) = H(X|Z, Y)$, therefore, $I(X; Z|Y) = H(X|Y) - H(X|Z, Y) = 0$.

2. **Conditional mutual information vs. unconditional mutual information.** Give examples of joint random variables $X$, $Y$ and $Z$ such that

   (a) $I(X;Y \mid Z) < I(X;Y)$,

   (b) $I(X;Y \mid Z) > I(X;Y)$.

   **Solution:** *Conditional mutual information vs. unconditional mutual information.*

   (a) The last corollary to Theorem 2.8.1 in the text states that if $X \to Y \to Z$, that is, if $p(x,y \mid z) = p(x \mid z)p(y \mid z)$, then $I(X;Y) \geq I(X;Y \mid Z)$. Equality holds if and only if $I(X;Z) = 0$ or $X$ and $Z$ are independent.

   A simple example of random variables satisfying the inequality conditions above is, $X$ is a fair binary random variable and $Y = X$ and $Z = Y$. In this case,

$$I(X;Y) = H(X) - H(X \mid Y) = H(X) = 1$$

   and,

$$I(X;Y \mid Z) = H(X \mid Z) - H(X \mid Y, Z) = 0.$$

   So that $I(X;Y) > I(X;Y \mid Z)$.

   (b) This example is also given in the text. Let $X, Y$ be independent fair binary random variables and let $Z = X + Y$. In this case we have that,

$$I(X;Y) = 0$$

   and,

$$I(X;Y \mid Z) = H(X \mid Z) = 1/2.$$

   So $I(X;Y) < I(X;Y \mid Z)$. Note that in this case $X, Y, Z$ are not markov.

3. **Prefix and Uniquely Decodable codes**
   Consider the following code:

| $u$ | Codeword |
|---|---|
| a | 1 0 |
| b | 0 0 |
| c | 1 1 |
| d | 1 1 0 |

   (a) Is this a Prefix code?

   (b) Argue that this code is uniquely decodable, by describing an algorithm for the decoding.

   **Solution: Prefix and Uniquely Decodable**

(a) No. The codeword of $c$ is a prefix of the codeword of $d$.

(b) We decode the encoded symbols from left to right. At any stage,

- If the next two bits are 10, output $a$ and move to the third bit.
- If the next two bits are 00, output $b$ and move to the third bit.
- If the next two bits are 11, look at the third bit:
  - If it is 1, output $c$ and move to the third bit
  - If it is 0, count the number of 0's after the 11:
    * If even (say $2m$ zeros), decode to $cb \dots b$ with $m$ $b$'s and move to the bit after the 0's.
    * If odd (say $2m+1$ zeros), decode to $db \dots b$ with $m$ $b$'s and move to the bit after the 0's.

Some examples with their decoding:

- 11011. It is not possible to split this string as $11-0-11$ because there is no codeword "0" . Therefore the only way is: $110-11$.
- 1110. It is not possible to split this string as $1-11-0$ or $1-110$ because there is no codeword "0" or "1" . Therefore the only way is: $11-10$.
- 110010. It is not possible to split this string as $110-0-10$ because there is no codeword "0" . Therefore the only way is: $11-00-10$.

For a more elaborate discussion on this topic read Problem 5.27[1]. In this problem,the *Sardinas-Patterson* test of unique decodability is explained.

4. **Relative entropy and the cost of miscoding.** Let the random variable $X$ be defined on $\{1,2,3,4,5,6\}$ according to pmf $p$. Let $p$ and another pmf $q$ be

| Symbol | $p(x)$ | $q(x)$ | $C_1(x)$ | $C_2(x)$ |
|---|---|---|---|---|
| 1 | 1/2 | 1/2 | 0 | 0 |
| 2 | 1/8 | 1/4 | 100 | 10 |
| 3 | 1/8 | 1/16 | 101 | 1100 |
| 4 | 1/8 | 1/16 | 110 | 1101 |
| 5 | 1/16 | 1/16 | 1110 | 1110 |
| 6 | 1/16 | 1/16 | 1111 | 1111 |

(a) Calculate $H(X)$, $D(p||q)$ and $D(q||p)$.

(b) The last two columns above represent codes for the random variable. Verify that codes $C_1$ and $C_2$ are optimal under the respective distributions $p$ and $q$.

(c) Now assume that we use $C_2$ to code $X$. What is the average length of the codewords? By how much does it exceed the entropy $H(X)$, i.e., what is the redundancy of the code?

(d) What is the redundancy if we use code $C_1$ for a random variable $Y$ with pmf $q$?

**Solution:**

[1]from: T.M. Cover and J.A. Thomas, "Elements of Information Theory", Second Edition,2006.

(a) For $X \sim p$

$$
\begin{aligned}
H(X) &= \frac{1}{2}\log 2 + \frac{1}{8}\log 8 + \frac{1}{8}\log 8 + \frac{1}{8}\log 8 + \frac{1}{16}\log 16 + \frac{1}{16}\log 16 \\
&= \frac{1}{2} + \frac{3}{8} + \frac{3}{8} + \frac{3}{8} + \frac{4}{16} + \frac{4}{16} \\
&= 2.125.
\end{aligned}
$$

For $X \sim q$

$$
\begin{aligned}
H(X) &= \frac{1}{2}\log 2 + \frac{1}{4}\log 4 + \frac{1}{16}\log 16 + \frac{1}{16}\log 16 + \frac{1}{16}\log 16 + \frac{1}{16}\log 16 \\
&= \frac{1}{2} + \frac{2}{4} + \frac{4}{16} + \frac{4}{16} + \frac{4}{16} + \frac{4}{16} \\
&= 2.
\end{aligned}
$$

Lets calculate $D(p||q)$,

$$
\begin{aligned}
D(p||q) &= \frac{1}{2}\log 1 + \frac{1}{8}\log\frac{1}{2} + \frac{1}{8}\log 2 + \frac{1}{8}\log 2 + \frac{1}{16}\log 1 + \frac{1}{16}\log 1 \\
&= \frac{1}{8}\log\frac{1}{2} + \frac{1}{8}\log 2 + \frac{1}{8}\log 2 \\
&= 1/8.
\end{aligned}
$$

Similarly

$$
\begin{aligned}
D(q||p) &= \frac{1}{2}\log 2 + \frac{1}{4}\log 2 + \frac{1}{16}\log\frac{1}{2} + \frac{1}{16}\log\frac{1}{2} + \frac{1}{16}\log 1 + \frac{1}{16}\log 1 \\
&= \frac{1}{4}\log 2 + \frac{1}{16}\log\frac{1}{2} + \frac{1}{16}\log\frac{1}{2} \\
&= \frac{1}{4} - \frac{1}{16} - \frac{1}{16} \\
&= \frac{1}{8}.
\end{aligned}
$$

(b) For $X \sim p$, the expected length of $C_1$ is

$$
\begin{aligned}
E[\ell(X)] &= \frac{1}{2} + \frac{3}{8} + \frac{3}{8} + \frac{3}{8} + \frac{4}{16} + \frac{4}{16} \\
&= 2.125 \\
&= H(X)
\end{aligned}
$$

and for $X \sim q$ , the expected length of $C_2$ is

$$
\begin{aligned}
E[\ell(X)] &= \frac{1}{2} + \frac{2}{4} + \frac{4}{16} + \frac{4}{16} + \frac{4}{16} + \frac{4}{16} \\
&= 2 \\
&= H(X)
\end{aligned}
$$

and thus both $C_1$ and $C_2$ are optimal codes.

(c) Average length of the codeword when $C_2$ is assigned to $X \sim p$ is

$$\begin{aligned} E[\ell(X)] &= \frac{1}{2} + \frac{2}{8} + \frac{4}{8} + \frac{4}{8} + \frac{4}{16} + \frac{4}{16} \\ &= 2.25 \\ &= H(X) + .125 \\ &= H(X) + D(p||q)! \end{aligned}$$

(d) Similarly the average length of the codeword when $C_1$ is assigned to $Y \sim q$ is

$$\begin{aligned} E[\ell(Y)] &= \frac{1}{2} + \frac{3}{4} + \frac{3}{16} + \frac{3}{16} + \frac{4}{16} + \frac{4}{16} \\ &= 2.125 \\ &= H(Y) + .125 \\ &= H(Y) + D(q||p)! \end{aligned}$$

5. **Shannon code.** Consider the following method for generating a code for a random variable $X$ which takes on $m$ values $\{1, 2, \ldots, m\}$ with pmf $p$ having probabilities $p_1, p_2, \ldots, p_m$. Assume that the probabilities are ordered so that $p_1 \geq p_2 \geq \cdots \geq p_m$. Define

$$F_i = \sum_{k=1}^{i-1} p_k,$$

i.e. the sum of the probabilities of all symbols less than $i$. Then the codeword for $i$ is the number $F_i \in [0, 1]$ rounded off to $l_i$ bits, where $l_i = \lceil \log \frac{1}{p_i} \rceil$.

(a) Show that the code constructed by this process is prefix-free and the average length satisfies

$$H(X) \leq L < H(X) + 1.$$

(b) Construct the code for the probability distribution $(0.5, 0.25, 0.125, 0.125)$.

(c) Now, suppose the code in (a) is used on a random variable $\tilde{X}$ taking values in $\{1, 2, \ldots, m\}$ distributed with pmf $q$ having probabilities $q_1, q_2, \ldots q_m$. Show that the average length satisfies

$$H(\tilde{X}) + D(q||p) \leq L < H(\tilde{X}) + D(q||p) + 1$$

**Solution:**

*Shannon code.*

(a) Since $l_i = \lceil \log \frac{1}{p_i} \rceil$, we have

$$\log \frac{1}{p_i} \leq l_i < \log \frac{1}{p_i} + 1$$

which implies that

$$H(X) \leq L = \sum p_i l_i < H(X) + 1.$$

The difficult part is to prove that the code is a prefix code. By the choice of $l_i$, we have

$$2^{-l_i} \leq p_i < 2^{-(l_i-1)}.$$

Thus $F_j$, $j > i$ differs from $F_i$ by at least $2^{-l_i}$, and will therefore differ from $F_i$ is at least one place in the first $l_i$ bits of the binary expansion of $F_i$. Thus the codeword for $F_j$, $j > i$, which has length $l_j \geq l_i$, differs from the codeword for $F_i$ at least once in the first $l_i$ places. Thus no codeword is a prefix of any other codeword.

(b) We build the following table

| Symbol | Probability | $F_i$ in decimal | $F_i$ in binary | $l_i$ | Codeword |
|---|---|---|---|---|---|
| 1 | 0.5 | 0.0 | 0.0 | 1 | 0 |
| 2 | 0.25 | 0.5 | 0.10 | 2 | 10 |
| 3 | 0.125 | 0.75 | 0.110 | 3 | 110 |
| 4 | 0.125 | 0.875 | 0.111 | 3 | 111 |

The Shannon code in this case achieves the entropy bound (1.75 bits) and is optimal.

(c) Just as in (a), we have

$$\log \frac{1}{p_i} \leq l_i < \log \frac{1}{p_i} + 1$$

which implies that

$$\sum_i q_i \log \frac{1}{p_i} \leq \sum_i q_i l_i < \sum_i q_i \log \frac{1}{p_i} + 1$$

$$\sum_i q_i \log \frac{1}{q_i} + \sum_i q_i \log \frac{q_i}{p_i} \leq L < \sum_i q_i \log \frac{1}{q_i} + \sum_i q_i \log \frac{q_i}{p_i} + 1$$

$$H(\tilde{X}) + D(q||p) \leq L < H(\tilde{X}) + D(q||p) + 1$$

6. **AEP.** Let $X_i$ for $i \in \{1, \ldots, n\}$ be an i.i.d. sequence from the p.m.f. $p(x)$ with alphabet $\mathcal{X} = \{1, 2, \ldots, m\}$. Denote the expectation and entropy of $X$ by $\mu := \mathbb{E}[X]$ and $H := -\sum p(x) \log p(x)$ respectively.

For $\epsilon > 0$, recall the definition of the typical set

$$A_\epsilon^{(n)} = \left\{ x^n \in \mathcal{X}^n : \left| -\frac{1}{n} \log p(x^n) - H \right| \leq \epsilon \right\}$$

and define the following set

$$B_\epsilon^{(n)} = \left\{ x^n \in \mathcal{X}^n : \left| \frac{1}{n} \sum_{i=1}^n x_i - \mu \right| \leq \epsilon \right\}.$$

In what follows, $\epsilon > 0$ is fixed.

(a) Does $\mathbb{P}\left(X^n \in A_\epsilon^{(n)}\right) \to 1$ as $n \to \infty$?

(b) Does $\mathbb{P}\left(X^n \in A_\epsilon^{(n)} \cap B_\epsilon^{(n)}\right) \to 1$ as $n \to \infty$?

(c) Show that for all $n$,

$$|A_\epsilon^{(n)} \cap B_\epsilon^{(n)}| \leq 2^{n(H+\epsilon)}.$$

(d) Show that for all $n$ sufficiently large:

$$|A_\epsilon^{(n)} \cap B_\epsilon^{(n)}| \geq \left(\frac{1}{2}\right) 2^{n(H-\epsilon)}.$$

**Solution: AEP**

(a) Yes, by the AEP for discrete random variables the probability $X^n$ is typical goes to 1.

(b) Yes, by the Law of Large Numbers $P(X^n \in B_\epsilon^{(n)}) \to 1$. So there exists $\epsilon > 0$ and $N_1$ such that $P(X^n \in A_\epsilon^{(n)}) > 1 - \frac{\epsilon}{2}$ for all $n > N_1$, and there exists $N_2$ such that $P(X^n \in B_\epsilon^{(n)}) > 1 - \frac{\epsilon}{2}$ for all $n > N_2$. So for all $n > \max(N_1, N_2)$:

$$\begin{aligned} P(X^n \in A_\epsilon^{(n)} \cap B_\epsilon^{(n)}) &= P(X^n \in A_\epsilon^{(n)}) + P(X^n \in B_\epsilon^{(n)}) - P(X^n \in A_\epsilon^{(n)} \cup B_\epsilon^{(n)}) \\ &> 1 - \frac{\epsilon}{2} + 1 - \frac{\epsilon}{2} - 1 \\ &= 1 - \epsilon \end{aligned}$$

So for any $\epsilon > 0$ there exists $N = \max(N_1, N_2)$ such that $P(X^n \in A_\epsilon^{(n)} \cap B_\epsilon^{(n)}) > 1 - \epsilon$ for all $n > N$, therefore $P(X^n \in A_\epsilon^{(n)} \cap B_\epsilon^{(n)}) \to 1$.

(c) By the law of total probability $\sum_{x^n \in A_\epsilon^{(n)} \cap B_\epsilon^{(n)}} p(x^n) \leq 1$. Also, for $x^n \in A_\epsilon^{(n)}$, from Theorem 3.1.2 in the text, $p(x^n) \geq 2^{-n(H+\epsilon)}$. Combining these two equations gives $1 \geq \sum_{x^n \in A_\epsilon^{(n)} \cap B_\epsilon^{(n)}} p(x^n) \geq \sum_{x^n \in A_\epsilon^{(n)} \cap B_\epsilon^{(n)}} 2^{-n(H+\epsilon)} = |A_\epsilon^{(n)} \cap B_\epsilon^{(n)}| 2^{-n(H+\epsilon)}$. Multiplying through by $2^{n(H+\epsilon)}$ gives the result $|A_\epsilon^{(n)} \cap B_\epsilon^{(n)}| \leq 2^{n(H+\epsilon)}$.

(d) Since from (b) $P\{X^n \in A_\epsilon^{(n)} \cap B_\epsilon^{(n)}\} \to 1$, there exists $N$ such that $P\{X^n \in A_\epsilon^{(n)} \cap B_\epsilon^{(n)}\} \geq \frac{1}{2}$ for all $n > N$. From Theorem 3.1.2 in the text, for $x^n \in A_\epsilon^{(n)}$, $p(x^n) \leq 2^{-n(H-\epsilon)}$. So combining these two gives $\frac{1}{2} \leq \sum_{x^n \in A_\epsilon^{(n)} \cap B_\epsilon^{(n)}} p(x^n) \leq \sum_{x^n \in A_\epsilon^{(n)} \cap B_\epsilon^{(n)}} 2^{-n(H-\epsilon)} = |A_\epsilon^{(n)} \cap B_\epsilon^{(n)}| 2^{-n(H-\epsilon)}$. Multiplying through by $2^{n(H-\epsilon)}$ gives the result $|A_\epsilon^{(n)} \cap B_\epsilon^{(n)}| \geq (\frac{1}{2}) 2^{n(H-\epsilon)}$ for $n$ sufficiently large.

7. **An AEP-like limit and the AEP (Bonus)**

(a) Let $X_1, X_2, \ldots$ be i.i.d. drawn according to probability mass function $p(x)$. Find the limit in probability as $n \to \infty$ of

$$p(X_1, X_2, \ldots, X_n)^{\frac{1}{n}}.$$

(b) Let $X_1, X_2, \ldots$ be an i.i.d. sequence of discrete random variables with entropy $H(X)$. Let

$$C_n(t) = \{x^n \in \mathcal{X}^n : p(x^n) \geq 2^{-nt}\}$$

denote the subset of $n$-length sequences with probabilities $\geq 2^{-nt}$.

i. Show that $|C_n(t)| \leq 2^{nt}$.

ii. What is $\lim_{n\to\infty} P(X^n \in C_n(t))$ when $t < H(X)$? And when $t > H(X)$?

**Solution: An AEP-like limit and the AEP.**

(a) By the AEP, we know that for every $\delta > 0$,

$$\lim_{n\to\infty} P\left(-H(X) - \delta \leq \frac{1}{n}\log p(X_1, X_2, \ldots, X_n) \leq -H(X) + \delta\right) = 1$$

Now, fix $\epsilon > 0$ (sufficiently small) and choose $\delta = \min\{\log(1 + 2^{H(X)}\epsilon), -\log(1 - 2^{H(X)}\epsilon)\}$. Then, $2^{-H(X)}(2^{\delta} - 1) \leq \epsilon$ and $2^{-H(X)}(2^{-\delta} - 1) \geq -\epsilon$. Thus,

$$\begin{aligned}
& -H(X) - \delta \leq \frac{1}{n}\log p(X_1, X_2, \ldots, X_n) \leq -H(X) + \delta \\
\implies & 2^{-H(X)}2^{-\delta} \leq (p(X_1, X_2, \ldots, X_n))^{\frac{1}{n}} \leq 2^{-H(X)}2^{\delta} \\
\implies & 2^{-H(X)}(2^{-\delta} - 1) \leq (p(X_1, X_2, \ldots, X_n))^{\frac{1}{n}} - 2^{-H(X)} \leq 2^{-H(X)}(2^{\delta} - 1) \\
\implies & -\epsilon \leq (p(X_1, X_2, \ldots, X_n))^{\frac{1}{n}} - 2^{-H(X)} \leq \epsilon
\end{aligned}$$

This along with AEP implies that $P(|p(X_1, X_2, \ldots, X_n))^{\frac{1}{n}} - 2^{-H(X)}| \leq \epsilon) \to 1$ for all $\epsilon > 0$ and hence $(p(X_1, X_2, \ldots, X_n))^{\frac{1}{n}}$ converges to $2^{-H(X)}$ in probability. This proof can be shortened by directly invoking the continuous mapping theorem, which says that if $Z_n$ converges to $Z$ in probability and $f$ is a continuous function, then $f(Z_n)$ converges to $f(Z)$ in probability.

(b) i.

$$\begin{aligned}
1 &\geq \sum_{x^n \in C_n(t)} p(x^n) \\
&\geq \sum_{x^n \in C_n(t)} 2^{-nt} \\
&= |C_n(t)|2^{-nt}
\end{aligned}$$

Thus, $|C_n(t)| \leq 2^{nt}$.

ii. Given the size of $C_n(t)$ from part (i), AEP directly implies that $\lim_{n\to\infty} P(X^n \in C_n(t)) = 0$ for $t < H(X)$ and $\lim_{n\to\infty} P(X^n \in C_n(t)) = 1$ for $t > H(X)$.