EE276: Homework #2 Solutions

1. **Data Processing Inequality.**
The random variables X, Y and Z form a Markov triplet $(X - Y - Z)$ if $p(z|y) = p(z|y,x)$, and as a corollary $p(x|y) = p(x|y,z)$. 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) $I(X;Z|Y) = 0$

   The following definition may be useful:

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

2. **Two looks.**
Let X, Y_1, and Y_2 be binary random variables. Assume that $I(X;Y_1) = 0$ and $I(X;Y_2) = 0$.

(a) Does it follow that $I(X; Y_1, Y_2) = 0$? Prove or provide a counterexample.

(b) Does it follow that $I(Y_1; Y_2) = 0$? Prove or provide a counterexample.

Solution: Two looks

(a) The answer is "no". Although at first the conjecture seems reasonable enough– after all, if Y_1 gives you no information about X, and if Y_2 gives you no information about X, then why should the two of them together give any information? But remember, it is NOT the case that $I(X; Y_1, Y_2) = I(X; Y_1) + I(X; Y_2)$. The chain rule for information says instead that $I(X; Y_1, Y_2) = I(X; Y_1) + I(X; Y_2|Y_1)$. The chain rule gives us reason to be skeptical about the conjecture.

This problem is reminiscent of the well-known fact in probability that pair-wise independence of three random variables is not sufficient to guarantee that all three are mutually independent. $I(X; Y_1) = 0$ is equivalent to saying that X and Y_1 are independent. Similarly for X and Y_2. But just because X is pairwise independent with each of Y_1 and Y_2, it does not follow that X is independent of the vector (Y_1, Y_2).

Here is a simple counterexample. Let Y_1 and Y_2 be independent fair coin flips. And let $X = Y_1$ XOR Y_2. X is pairwise independent of both Y_1 and Y_2, but obviously not independent of the vector (Y_1, Y_2), since X is uniquely determined once you know (Y_1, Y_2).

(b) Again the answer is "no". Y_1 and Y_2 can be arbitrarily dependent with each other and both still be independent of X. For example, let $Y_1 = Y_2$ be two observations of the same fair coin flip, and X an independent fair coin flip. Then $I(X; Y_1) = I(X; Y_2) = 0$ because X is independent of both Y_1 and Y_2. However, $I(Y_1; Y_2) = H(Y_1) - H(Y_1|Y_2) = H(Y_1) = 1$.

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 providing 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 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 (as we assumed with pmf p). 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 $X \sim q$ is

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

5. **The AEP and source coding.** A discrete memoryless source emits a sequence of statistically independent binary digits with probabilities $p(1) = 0.005$ and $p(0) = 0.995$. The digits are taken 100 at a time and a binary codeword is provided for every sequence of 100 digits containing three or fewer ones.

   (a) Assuming that all codewords are the same length, find the minimum length required to provide codewords for all sequences with three or fewer ones.

   (b) Calculate the probability of observing a source sequence for which no codeword has been assigned.

   (c) Use Chebyshev's inequality to bound the probability of observing a source sequence for which no codeword has been assigned. Compare this bound with the actual probability computed in part (b).

   (d) If the codewords for sequences with four or more ones were taken as simply the sequences themselves, give a bound on the expected compression rate of the code. Compare this with the entropy rate of the source.

Solution: *The AEP and source coding.*

(a) The number of 100-bit binary sequences with three or fewer ones is

$$\binom{100}{0} + \binom{100}{1} + \binom{100}{2} + \binom{100}{3} = 1 + 100 + 4950 + 161700 = 166751\,.$$

The required codeword length is $\lceil \log_2 166751 \rceil = 18$. (Note that $H(0.005) = 0.0454$, so 18 is quite a bit larger than the 4.5 bits of entropy.)

(b) The probability that a 100-bit sequence has three or fewer ones is

$$\sum_{i=0}^{3} \binom{100}{i} (0.005)^i (0.995)^{100-i} = 0.60577 + 0.30441 + 0.7572 + 0.01243 = 0.99833$$

Thus the probability that the sequence that is generated cannot be encoded is $1 - 0.99833 = 0.00167$.

(c) In the case of a random variable S_n that is the sum of n i.i.d. random variables $X_1, X_2, \ldots, X_n$, Chebyshev's inequality states that

$$\Pr(|S_n - n\mu| \geq \epsilon) \;\leq\; \frac{n\sigma^2}{\epsilon^2}\,,$$

where μ and σ^2 are the mean and variance of X_i. (Therefore $n\mu$ and $n\sigma^2$ are the mean and variance of S_n.) In this problem, $n = 100$, $\mu = 0.005$, and $\sigma^2 = (0.005)(0.995)$. Note that $S_{100} \geq 4$ if and only if $|S_{100} - 100(0.005)| \geq 3.5$, so we should choose $\epsilon = 3.5$. Then

$$\Pr(S_{100} \geq 4) \leq \frac{100(0.005)(0.995)}{(3.5)^2} \approx 0.04061\,.$$

This bound is much larger than the actual probability 0.00167.

(d) Let the random variable L be defined as the length of the resulting codeword. Then the compression rate is

$$\frac{1}{n}E(L) = \frac{1}{100}\,(18 \times 0.99833 + 100 \times 0.00167) = 0.181369. \tag{1}$$

Meanwhile, if Y is the random string of length $n = 100$ at the source, then the entropy rate is given by

$$\frac{1}{n}H(Y) = H(p) = 0.0454 \tag{2}$$

where $H(p)$ is the binary entropy.

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 n sufficiently large.

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

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 $X_1, X_2, \dots$ 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 $\delta > 0$,

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

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,

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

This along with AEP implies that $P(|p(X_1, X_2, \dots, X_n))^{\frac{1}{n}} - 2^{-H(X)}| \leq \epsilon) \to 1$ for all $\epsilon > 0$ and hence $(p(X_1, X_2, \dots, X_n))^{\frac{1}{n}}$ 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.

Alternate proof (using Strong LLN):
$X_1, X_2, \dots,$ i.i.d. $\sim p(x)$. Hence $\log(X_i)$ are also i.i.d. and

$$\begin{aligned}
\lim (p(X_1, X_2, \dots, X_n))^{\frac{1}{n}} &= \lim 2^{\log(p(X_1, X_2, \dots, X_n))^{\frac{1}{n}}} \\
&= 2^{\lim \frac{1}{n}\sum \log p(X_i)} \\
&= 2^{E(\log(p(X)))} \\
&= 2^{-H(X)}
\end{aligned}$$

where the second equality uses the continuity of the function 2^x and the third equality uses the strong law of large numbers. Thus, $(p(X_1, X_2, \dots, X_n))^{\frac{1}{n}}$ converges to $2^{-H(X)}$ alomost surely, and hence 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. AEP immediately 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 $t > H(X)$.