EEE 551 Information Theory (Spring 2022)

Homework 2 Solutions

---

1. *The Binary Typical Set.* Let $X \sim \text{Bern}(p)$ (i.e. it is binary with $p_X(1) = p,\ p_X(0) = 1-p$). Given a sequence $x^n \in \{0,1\}^n$, let n_1 be the number of 1s that appear in the sequence (e.g. if $x^n = (1,1,0,1)$, then $n_1 = 3$).

   (a) Write $p(x^n)$ in terms of only p, n, and n_1.

   (b) Find a condition (again in terms of only p, n, and n_1) for the sequence x^n being in the typical set $A_\epsilon^{(n)}$.

   (c) If $\epsilon = 0$, is the typical set empty?

   (d) Suppose $p = 1/3$, $n = 20$, and $\epsilon = 0.05$. Calculate $\Pr\{A_\epsilon^{(n)}\}$. *Hint:* If $X^n \overset{\text{iid}}{\sim} \text{Bern}(p)$, then n_1 is a binomial random variable.

   Solution:

   (a) $p(x^n) = \prod_{i=1}^n p(x_i) = p^{n_1}(1-p)^{n-n_1}$.

   (b) We use the fact that $x^n \in A_\epsilon^{(n)}$ iff

$$\left| -\frac{1}{n}\log p(x^n) - H(X) \right| \le \epsilon.$$

   We may rewrite the quantity in the left-hand side are

$$\begin{aligned}
\left| -\frac{1}{n}\log p(x^n) - H(X) \right| &= \left| -\frac{1}{n}\log\left(p^{n_1}(1-p)^{n-n_1}\right) - H(X) \right| \\
&= \left| -\frac{n_1}{n}\log p - \frac{n-n_1}{n}\log(1-p) + p\log p + (1-p)\log(1-p) \right| \\
&= \left| \frac{n_1}{n}\log\frac{1-p}{p} - \log(1-p) + p\log p + (1-p)\log(1-p) \right| \\
&= \left| \frac{n_1}{n}\log\frac{1-p}{p} + p\log p - p\log(1-p) \right| \\
&= \left| \left(\frac{n_1}{n} - p\right)\log\frac{1-p}{p} \right|.
\end{aligned}$$

   Thus $x^n \in A_\epsilon^{(n)}$ iff

$$\left| \left(\frac{n_1}{n} - p\right)\log\frac{1-p}{p} \right| \le \epsilon.$$

   Multiplying by $n/|\log\frac{1-p}{p}|$, this condition can be written

$$|n_1 - np| \le \frac{n\epsilon}{\left|\log\frac{1-p}{p}\right|}.$$

   Hence $x^n \in A_\epsilon^{(n)}$ iff

$$np - \frac{n\epsilon}{\left|\log\frac{1-p}{p}\right|} \le n_1 \le np + \frac{n\epsilon}{\left|\log\frac{1-p}{p}\right|}.$$

(c) From the condition found in part (a), if $\epsilon = 0$ then for x^n to be in $A_\epsilon^{(n)}$ we need $n_1 = np$. Since n_1 must be an integer, this occurs only if np is an integer. Thus, the typical set is empty iff np is not an integer.

(d) For these parameters, from part (b), $x^n \in A_\epsilon^{(n)}$ iff

$$5.67 \leq n_1 \leq 7.67.$$

Since n_1 can only be an integer, $x^n \in A_\epsilon^{(n)}$ iff $n_1 = 6$ or $n_1 = 7$. Since n_1 is a binomial random variable,

$$\Pr(n_1 = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Thus

$$\Pr\{A_\epsilon^{(n)}\} = \Pr\{n_1 = 6\} + \Pr\{n_1 = 7\} = \binom{20}{6}(1/3)^6(2/3)^{14} + \binom{20}{7}(1/3)^7(2/3)^{13} = 0.3643.$$

2. Problem 3.9 from Cover-Thomas: *AEP*. Let $X_1, X_2, \ldots$ be independent, identically distributed random variables drawn according to the probability mass function $p(x), x \in \{1, 2, \ldots, m\}$. Thus $p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^n p(x_i)$. We know that $-\frac{1}{n} \log p(X_1, X_2, \ldots, X_n) \to H(X)$ in probability. Let $q(x_1, x_2, \ldots, x_n) = \prod_{i=1}^n q(x_i)$, where q is another probability mass function on $\{1, 2, \ldots, m\}$.

(a) Evaluate $\lim -\frac{1}{n} \log q(X_1, X_2, \ldots, X_n)$, where $X_1, X_2, \ldots$ are i.i.d. $\sim p(x)$.

(b) Now evaluate the limit of the log likelihood ratio $\frac{1}{n} \log \frac{q(X_1,\ldots,X_n)}{p(X_1,\ldots,X_n)}$ when $X_1, X_2, \ldots$ are i.i.d. $\sim p(x)$. Thus, the odds favoring q are exponentially small when p is true.

Solution:

(a) Since $X_1, X_2, \ldots, X_n$ are i.i.d., so are $q(X_1), q(X_2), \ldots, q(X_n)$, and hence we can apply the law of large numbers to obtain

$$\begin{aligned}
-\frac{1}{n} \log q(X_1, X_2, \ldots, X_n) &= -\frac{1}{n} \sum_{i=1}^n \log q(X_i) \\
&\to -E[\log q(X)] \\
&= -\sum_{x \in \mathcal{X}} p(x) \log q(x) \\
&= \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \cdot \frac{1}{p(x)} \\
&= \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} - \sum_{x \in \mathcal{X}} p(x) \log p(x) \\
&= D(p\|q) + H(p).
\end{aligned}$$

(b) Again applying the law of large numbers

$$\begin{aligned}
\frac{1}{n} \log \frac{q(X_1, X_2, \ldots, X_n)}{p(X_1, X_2, \ldots, X_n)} &= \frac{1}{n} \sum_{i=1}^n \log \frac{q(X_i)}{p(X_i)} \\
&\to E\left[\log \frac{q(X)}{p(X)}\right] \\
&= \sum_{x \in \mathcal{X}} p(x) \log \frac{q(x)}{p(x)} \\
&= -\sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \\
&= -D(p\|q).
\end{aligned}$$

3. *Optimal Fixed-to-Fixed Source Code.* Consider the following distribution:

x	a	b	c
$p(x)$	0.5	0.3	0.2

This problem requires you to find the fixed-to-fixed source code that minimizes the probability of error for a given number of compressed bits. This optimal code will not necessarily be the code based on the typical set. A fixed-to-fixed code compresses a sequences of length n to a fixed number ℓ bits. That is, there is an encoding function

$$f : \mathcal{X}^n \to \{0,1\}^\ell$$

and a decoding function

$$g : \{0,1\}^\ell \to \mathcal{X}^n.$$

For each of the following values of n and ℓ, find the code that minimizes the probability of error, and the associated probability of error.

(a) $n = 1, \ell = 1$

(b) $n = 2, \ell = 1$

(c) $n = 2, \ell = 2$

(d) $n = 2, \ell = 3$

Solution: To minimize the probability of error, we always want the code to be such that the most likely sequences are decoded correctly. Given an output length ℓ, we can ensure that 2^ℓ sequences are decoded correctly, so we should always choose the 2^ℓ most likely sequences.

(a) Since $\ell = 1$, we can correctly decode $2^\ell = 2$ sequences. Thus we should correctly decode the two most likely letters, a, b. The following code achieves this:

$$f(\mathsf{a}) = 0,\ f(\mathsf{b}) = 1,\ f(\mathsf{c}) = 0,$$
$$g(0) = \mathsf{a},\ g(1) = \mathsf{b}.$$

The probability of error is $P_e = p(\mathsf{c}) = 0.2$.

(b) For $n = 2$, the joint PMF of x_1, x_2 is given by (sorted from most likely to least likely):

$$p(\mathsf{aa}) = 0.25,\ p(\mathsf{ab}) = 0.15,\ p(\mathsf{ba}) = 0.15,\ p(\mathsf{ac}) = 0.1,\ p(\mathsf{ca}) = 0.1,$$
$$p(\mathsf{bb}) = 0.09,\ p(\mathsf{bc}) = 0.06\ p(\mathsf{cb}) = 0.06\ p(\mathsf{cc}) = 0.04$$

For $\ell = 1$, we should correctly decode the 2 most likely sequences aa, ab. The following code achieves this:

$$f(\mathsf{aa}) = 0,\ f(\mathsf{ab}) = 1,\ f(x_1x_2) = 0 \text{ for any other } (x_1, x_2),$$
$$g(0) = \mathsf{aa},\ g(1) = \mathsf{ab}.$$

The probability of error is $P_e = 0.6$.

(c) Since $\ell = 2$, we should correctly decode the $2^2 = 4$ most likely sequences $\mathsf{aa}, \mathsf{ab}, \mathsf{ba}, \mathsf{ac}$. The following code achieves this:

$$f(\mathsf{aa}) = 00,\ f(\mathsf{ab}) = 01,\ f(\mathsf{ba}) = 10,\ f(\mathsf{ac}) = 11,\ f(x_1x_2) = 0 \text{ for any other } (x_1, x_2),$$
$$g(00) = \mathsf{aa},\ g(01) = \mathsf{ab},\ g(10) = \mathsf{ba},\ g(11) = \mathsf{ac}.$$

The probability of error is $P_e = 0.35$.

(d) Since $\ell = 3$, we should correctly decode the $2^3 = 8$ most likely sequences, which is all sequences except cc. The following code achieves this:

$$\begin{aligned} f(\mathsf{aa}) = 000,\ f(\mathsf{ab}) = 001,\ f(\mathsf{ba}) = 010,\ f(\mathsf{ac}) = 011, f(\mathsf{ca}) = 100, \\ f(\mathsf{bb}) = 101,\ f(\mathsf{bc}) = 110,\ f(\mathsf{cb}) = 111,\ f(\mathsf{cc}) = 000, \\ g(000) = \mathsf{aa},\ g(001) = \mathsf{ab},\ g(010) = \mathsf{ba},\ g(011) = \mathsf{ac}, \\ g(100) = \mathsf{ca},\ g(101) = \mathsf{bb},\ g(110) = \mathsf{bc},\ g(111) = \mathsf{cb}. \end{aligned}$$

The probability of error is $P_e = 0.04$.

4. *Coding for Two Different Distributions.* Consider two distributions $p(x)$ and $q(x)$ on the same alphabet $\mathcal{X}$. We want to design a source code that works with *either* distribution. Let $R > \max\{H_p(X), H_q(X)\}$, where H_p and H_q refer to the entropies under $p(x)$ and $q(x)$ respectively. Find a fixed-to-fixed source code with rate R such that the probability of error is less than ϵ if the underlying distribution is either $p(x)$ or $q(x)$.

Solution: Let $A_\epsilon^{(n)}[p]$ be the typical set with respect to distribution p, and $A_\epsilon^{(n)}[q]$ the same for q. We know that

$$|A_\epsilon^{(n)}[p]| \le 2^{n(H_p(X)+\epsilon)}, \qquad |A_\epsilon^{(n)}[q]| \le 2^{n(H_q(X)+\epsilon)}.$$

Let $B = A_\epsilon^{(n)}[p] \cup A_\epsilon^{(n)}[q]$; that is, B is the set of sequences that are typical with respect to either $p(x)$ or $q(x)$. Thus

$$|B| \le |A_\epsilon^{(n)}[p]| + |A_\epsilon^{(n)}[q]| \le 2^{n(H_p(X)+\epsilon)} + 2^{n(H_q(X)+\epsilon)} = 2^{nR}\left[2^{n(-R+H_p(X)+\epsilon)} + 2^{n(-R+H_q(X)+\epsilon)}\right].$$

If we assume that $R > \max\{H_p(X), H_q(X)\}$, then as long as ϵ is small enough, $-R + H_p(X) + \epsilon < 0$ and $-R + H_q(X) + \epsilon < 0$. Thus, for n sufficiently large,

$$2^{n(-R+H_p(X)+\epsilon)} + 2^{n(-R+H_q(X)+\epsilon)} \le 1$$

which means that $|B| \le 2^{nR}$. We now form a fixed-to-fixed source code as follows. For each $x^n \in B$, let $\phi(x^n)$ be a unique index in $\{1, \ldots, 2^{nR}\}$. Then define the encoder and decoder as

$$\begin{aligned} f(x^n) &= \begin{cases} \phi(x^n), & x^n \in B \\ 1, & \text{otherwise} \end{cases} \\ g(m) &= \phi^{-1}(m). \end{aligned}$$

This code will not make any errors if $X^n \in B$. Since B contains the typical sets for both $p(x)$ and $q(x)$, the error will be no more than ϵ if the underlying distribution is either $p(x)$ or $q(x)$.

5. Problem 5.9 from Cover-Thomas: *Optimal code lengths that require one bit above entropy.* The source coding theorem shows that the optimal code for a random variable X has an expected length less than $H(X) + 1$. Give an example of a random variable for which the expected length of the optimal code is close to $H(X) + 1$ [i.e., for any $\epsilon > 0$, construct a distribution for which the optimal code has $L(C) > H(X) + 1 - \epsilon$]. *Hint:* Try a Bernoulli distribution.

Solution: Let $X \sim \text{Bern}(p)$, where $0 < p < 1$. Since both possible outcomes for X occur with positive probability, the only reasonable code (and the optimal code) is the one that assigns a length-1 bit string to each outcome. That is, the simple code where

$$C(0) = 0, \qquad C(1) = 1.$$

Thus, we have $L(C) = 1$. On the other hand, the entropy is

$$H(X) = -p \log p - (1-p) \log(1-p).$$

For p either very close to 0 or very close to 1, this entropy can be made arbitrarily close to 0. Thus, the gap between $L(C)$ and $H(X)$ can be made arbitrarily close to 1.

6. Problem 5.12 from Cover-Thomas: *Shannon codes and Huffman codes.* Consider a random variable X that takes on four values with probabilities $(\frac{1}{3}, \frac{1}{3}, \frac{1}{4}, \frac{1}{12})$.

   (a) Construct a Huffman code for this random variable.

   (b) Show that there exist two different sets of optimal lengths for the codewords; namely, show that codeword length assignments $(1, 2, 3, 3)$ and $(2, 2, 2, 2)$ are both optimal.

   (c) Conclude that there are optimal codes with codeword lengths for some symbols that exceed the Shannon code length $\lceil \log \frac{1}{p(x)} \rceil$.

   Solution:

   (a) The first step in the Huffman algorithm is to merge symbols 3 and 4, forming a combined symbol with probability $\frac{1}{4} + \frac{1}{12} = \frac{1}{3}$. At this point, the three remaining symbols each have probability $1/3$, so there are multiple choices: we may next combine symbols 1 and 2, which yields the following code:

   $$C(1) = 00, \quad C(2) = 01, \quad C(3) = 10, \quad C(4) = 11.$$

   Or, we may next combine 2 with the $(3, 4)$ symbol, which yields the following code:

   $$C(1) = 0, \quad C(2) = 10, \quad C(3) = 110, \quad C(4) = 111.$$

   Either of these codes is an optimal Huffman code.

   (b) As seen above, both length assignments $(1, 2, 3, 3)$ or $(2, 2, 2, 2)$ may result from the Huffman algorithm, so they are both optimal. Another way to see it is to explicitly calculate the expected length, and indeed the expected length is 2 for each code.

   (c) We may compare the two Huffman code lengths with the Shannon code length as follows:

Probability	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{12}$
Huffman #1 length	1	2	3	3
Huffman #2 length	2	2	2	2
Shannon length	2	2	2	4

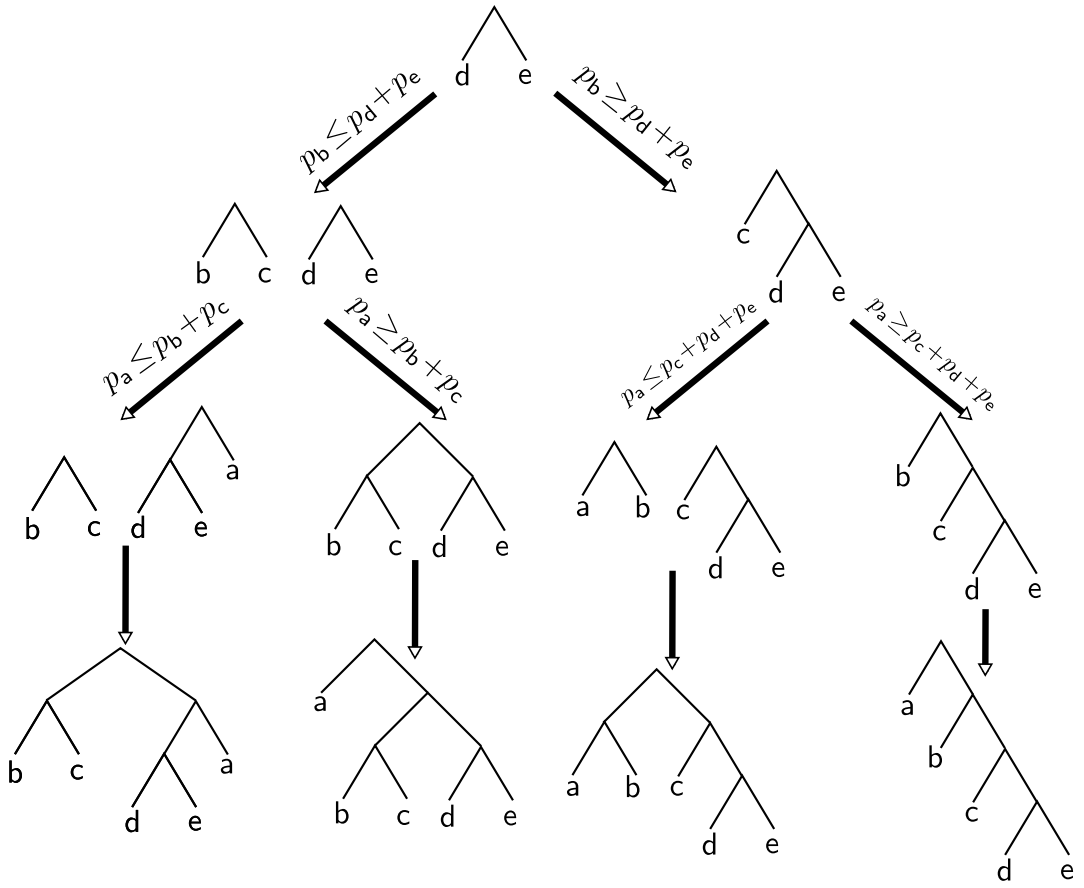
   Indeed, for Huffman #1, the length for the symbol with probability $1/4$ is larger than that for the Shannon code.

7. *Huffman Codes on Five Letters.* Consider a random variable with alphabet $\mathcal{X} = \{\mathsf{a}, \mathsf{b}, \mathsf{c}, \mathsf{d}, \mathsf{e}\}$, where

   $$p_{\mathsf{a}} \geq p_{\mathsf{b}} \geq p_{\mathsf{c}} \geq p_{\mathsf{d}} \geq p_{\mathsf{e}}.$$

   Let $(l_{\mathsf{a}}, l_{\mathsf{b}}, l_{\mathsf{c}}, l_{\mathsf{d}}, l_{\mathsf{e}})$ be the codeword lengths for the Huffman code for this distribution. Determine the Huffman codeword lengths as a function of the distribution; i.e. give all possible Huffman word lengths that could occur, and for each one, say what conditions on the distribution need to hold in order for this to be the Huffman code.

   Solution: The following diagram shows the possible steps in building the Huffman tree for 5 symbols, and the conditions on the probabilities under which different trees occur. Note that in all cases we start out by merging symbols d and e, because they are the least likely. The next step depends on whether $p_{\mathsf{b}} \leq p_{\mathsf{d}} + p_{\mathsf{e}}$: if it is, then b and c are now the least likely symbols, so they are merged; otherwise, c and the combined symbol (d, e) are the least likely symbols. The full process is shown in the diagram.



As shown above, there are four possible trees that can occur. The codeword lengths can be determined from the trees. Note that even though there are four trees, there are only three different sets of codeword lengths (the first and third trees both give $(2, 2, 2, 3, 3)$). Therefore the codeword lengths are determined by the probabilities as follows:

$$(l_a, l_b, l_c, l_d, l_e) = \begin{cases} (2,2,2,3,3) & \text{if } p_b \le p_d + p_e \text{ and } p_a \le p_b + p_c, \text{ or } p_b \ge p_d + p_e \text{ and } p_a \le p_c + p_d + p_e \\ (1,3,3,3,3) & \text{if } p_b \le p_d + p_e \text{ and } p_a \ge p_b + p_c \\ (1,2,3,4,4) & \text{if } p_b \ge p_d + p_e \text{ and } p_a \ge p_c + p_d + p_e \end{cases}$$

8. *Coding for the Wrong Distribution.* Let $p(x)$ and $q(x)$ be two different distributions on the same alphabet $\mathcal{X}$. Suppose you construct a code for distribution $q(x)$, but the true distribution is $p(x)$. Consider the fixed-to-variable code with Shannon code length for $q(x)$, i.e., the length is $\ell(x) = \lceil \log \frac{1}{q(x)} \rceil$. Prove that the expected length (under distribution $p(x)$) satisfies

$$H(X) + D(p\|q) \le L(C) < H(X) + D(p\|q) + 1.$$

Solution: The length $\ell(x) = \lceil \log \frac{1}{q(x)} \rceil$ satisfies

$$\log \frac{1}{q(x)} \le \ell(x) < \log \frac{1}{q(x)} + 1.$$

Thus

$$\begin{aligned} L(C) &= \mathbb{E}[\ell(X)] \\ &= \sum_{x \in \mathcal{X}} p(x)\ell(x) \\ &\geq \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{q(x)} \\ &= \sum_{x \in \mathcal{X}} p(x) \log \left(\frac{1}{p(x)} \cdot \frac{p(x)}{q(x)} \right) \\ &= \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)} + \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \\ &= H(X) + D(p\|q). \end{aligned}$$

By an identical calculation,

$$L(C) < \sum_{x \in \mathcal{X}} p(x) \left(\log \frac{1}{q(x)} + 1 \right) = H(X) + D(p\|q) + 1.$$