Fundamentals of Information Theory

Homework 3

Name: Zhou Shouchen
Student ID: 2021533042

Due 23:59 (CST), Nov. 2, 2024

Problem 1

5.12 Shannon codes and Huffman codes. Consider a random variable X that takes on four values with probabilities $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{4}, \frac{1}{12}\right)$.

(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 $\left\lceil \log \frac{1}{p(x)} \right\rceil$.

Solution

(a) The Huffman tree and each variable's code are shown below.

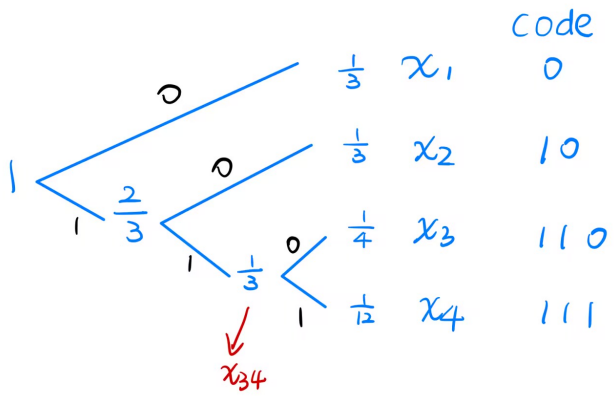

Figure 1. Huffman tree 1

(b) After merging x_3, x_4 and generate x_{34}, the probabilities of x_1, x_2, x_{34} are all $\frac{1}{3}$. Thus, we could merge any of one node in $\{x_1, x_2\}$ with x_{34} first as Figure 1, the length of the codes in that tree are $(1, 2, 3, 3)$; or we can merge x_1, x_2 first, then merge x_{12} with x_{34} as Figure 2, the length of the codes in this tree are $(2, 2, 2, 2)$. Both of these trees are valid Huffman trees, so these codewords are both optimal.

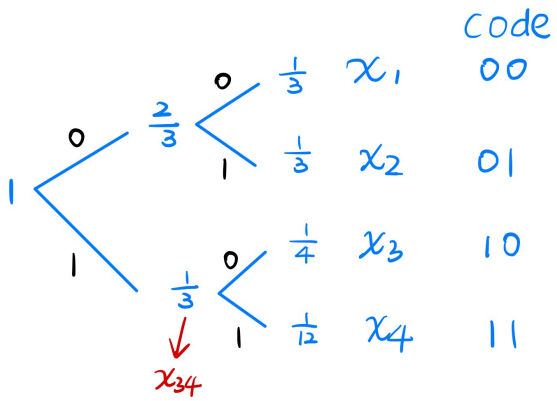

Figure 2. Huffman tree 2

And we can compute the average length of the codewords in these two trees, suppose the average length of the codewords in the first tree is $\overline{L}_1$, and the average length of the codewords in the second tree is $\overline{L}_2$,

then we have

$$\overline{L}_1 = \frac{1}{3} \times 1 + \frac{1}{3} \times 2 + \frac{1}{4} \times 3 + \frac{1}{12} \times 3 = 2 \text{ bits}$$
$$\overline{L}_2 = \frac{1}{3} \times 2 + \frac{1}{3} \times 2 + \frac{1}{4} \times 2 + \frac{1}{12} \times 2 = 2 \text{ bits}$$

And $\overline{L}_1 = \overline{L}_2$, which also shows that the codes are both optimal.

(c) For the second way to construct the Huffman tree in Figure 1, we could see that $l(x_3) = 3$ bits, but according to the Shannon's code, $\left\lceil \log \frac{1}{p(x_3)} \right\rceil = \lceil \log 4 \rceil = 2$ bits. And Huffman code is optimal, so we have for x_3, $l(x_3) > \left\lceil \log \frac{1}{p(x_3)} \right\rceil$ exceeds the Shannon code length.

Problem 2

5.20 Huffman codes with costs. Words such as "Run!", "Help!", and "Fire!" are short, not because they are used frequently, but perhaps because time is precious in the situations in which these words are required. Suppose that $X = i$ with probability $p_i, i = 1, 2, \ldots, m$. Let l_i be the number of binary symbols in the codeword associated with $X = i$, and let c_i denote the cost per letter of the codeword when $X = i$. Thus, the average cost C of the description of X is $C = \sum_{i=1}^{m} p_i c_i l_i$.

(a) Minimize C over all $l_1, l_2, \ldots, l_m$ such that $\sum 2^{-l_i} \leq 1$. Ignore any implied integer constraints on l_i. Exhibit the minimizing $l_1^*, l_2^*, \ldots, l_m^*$ and the associated minimum value C^*.

(b) How would you use the Huffman code procedure to minimize C over all uniquely decodable codes? Let C_{Huffman} denote this minimum.

(c) Can you show that

$$C^* \leq C_{\text{Huffman}} \leq C^* + \sum_{i=1}^{m} p_i c_i$$

Solution

(a) Ignore the integer constrains for the length l_i, we can construct the optimization problem as

$$\begin{aligned} \min_{l_1 \ldots, l_m} \quad & C = \sum_{i=1}^{m} p_i c_i l_i \\ \text{s.t.} \quad & \sum_{i=1}^{m} 2^{-l_i} \leq 1 \end{aligned}$$

The Lagrangian of the optimization problem is

$$L(l_1, \ldots, l_m, \lambda) = \sum_{i=1}^{m} p_i c_i l_i + \lambda \left(\sum_{i=1}^{m} 2^{-l_i} - 1 \right)$$

According to the KKT condition, we have the first order derivative of L with respect to l_i is 0, which is

$$\frac{\partial L}{\partial l_i} = p_i c_i - \lambda 2^{-l_i} \log_e 2 = 0, \quad \forall i = 1, \ldots, m$$

i.e.

$$(\lambda \log_e 2) \cdot 2^{-l_i^*} = p_i c_i$$

Thus, to have an optimal code, we set the inequality to be an equality, then we sum all the equations above, we have

$$\begin{aligned} & (\lambda \log_e 2) \sum_{i=1}^{m} 2^{-l_i^*} = \sum_{i=1}^{m} p_i c_i \\ \Rightarrow & \lambda \log_e 2 = \sum_{i=1}^{m} p_i c_i \\ \Rightarrow & 2^{-l_i^*} = \frac{p_i c_i}{\sum_{i=1}^{m} p_i c_i} \\ \Rightarrow & l_i^* = -\log \left(\frac{p_i c_i}{\sum_{i=1}^{m} p_i c_i} \right) \end{aligned}$$

So the minimum value is

$$\begin{aligned} C^* &= \sum_{i=1}^{m} p_i c_i \left(-\log \left(\frac{p_i c_i}{\sum_{j=1}^{m} p_j c_j} \right) \right) \\ &= -\sum_{i=1}^{m} (p_i c_i) \log (p_i c_i) + \left(\sum_{i=1}^{m} p_i c_i \right) \log \left(\sum_{i=1}^{m} p_i c_i \right) \end{aligned}$$

(b) Just let $q_i = \dfrac{p_i c_i}{\sum_{i=1}^{m} p_i c_i}$ to be the new probability distribution, and apply the Huffman code procedure to encode this new distribution q.

(c) It is obvios that $C^* \leq C_{\text{Huffman}}$, since the C^* is the minimum value of the cost function C over all uniquely decodable codes without integer constrains.

Since we have known that C_{Huffman} is the optimal for the distribution q if we add the integer constrain to the optimization problem in (a), with the Shannon's code, we have $l_i = \left\lceil \log \left(\dfrac{1}{q_i} \right) \right\rceil$, then we have

$$\begin{aligned} C_{\text{Shannon}} &= \sum_{i=1}^{m} p_i c_i \left\lceil \log \left(\frac{1}{q_i} \right) \right\rceil \\ &< \sum_{i=1}^{m} p_i c_i \left(\log \left(\frac{1}{q_i} \right) + 1 \right) \\ &= \sum_{i=1}^{m} p_i c_i \left(\log \left(\frac{\sum_{j=1}^{m} p_j c_j}{p_i c_i} \right) + 1 \right) \\ &= C^* + \sum_{i=1}^{m} p_i c_i \end{aligned}$$

Thus, we have

$$C_{\text{Huffman}} \leq C_{\text{Shannon}} \leq C^* + \sum_{i=1}^{m} p_i c_i$$

i.e. we have proved that

$$C^* \leq C_{\text{Huffman}} \leq C^* + \sum_{i=1}^{m} p_i c_i$$

Problem 3

5.30 Relative entropy is cost of miscoding. Let the random variable X have five possible outcomes $\{1, 2, 3, 4, 5\}$. Consider two distributions $p(x)$ and $q(x)$ on this random variable.

Symbol	$p(x)$	$q(x)$	$C_1(x)$	$C_2(x)$
1	$\frac{1}{2}$	$\frac{1}{2}$	0	0
2	$\frac{1}{4}$	$\frac{1}{8}$	10	100
3	$\frac{1}{8}$	$\frac{1}{8}$	110	101
4	$\frac{1}{16}$	$\frac{1}{8}$	1110	110
5	$\frac{1}{16}$	$\frac{1}{8}$	1111	111

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

(b) The last two columns represent codes for the random variable. Verify that the average length of C_1 under p is equal to the entropy $H(p)$. Thus, C_1 is optimal for p. Verify that C_2 is optimal for q.

(c) Now assume that we use code C_2 when the distribution is p. What is the average length of the codewords. By how much does it exceed the entropy p?

(d) What is the loss if we use code C_1 when the distribution is q?

Solution

(a)

$$H(p) = H\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{16}\right) = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{4}\log_2\frac{1}{4} + \frac{1}{8}\log_2\frac{1}{8} + \frac{1}{16}\log_2\frac{1}{16} + \frac{1}{16}\log_2\frac{1}{16}\right) = \frac{15}{8} \text{ bits}$$

$$H(q) = H\left(\frac{1}{2}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}\right) = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{8}\log_2\frac{1}{8} + \frac{1}{8}\log_2\frac{1}{8} + \frac{1}{8}\log_2\frac{1}{8} + \frac{1}{8}\log_2\frac{1}{8}\right) = 2 \text{ bits}$$

$$D(p\|q) = \sum_x p(x)\log_2\frac{p(x)}{q(x)} = \frac{1}{8} \text{ bits}$$

$$D(q\|p) = \sum_x q(x)\log_2\frac{q(x)}{p(x)} = \frac{1}{8} \text{ bits}$$

(b) Let the average length of C_1 under p be $\overline{L}_1$, the average length of C_2 under q be $\overline{L}_2$, and let $l_i(x) = |C_i(x)|$, then we have

$$\overline{L}_1 = \sum_x p(x)l_1(x) = \frac{15}{8} \text{ bits} = H(p)$$

$$\overline{L}_2 = \sum_x q(x)l_2(x) = 2 \text{ bits} = H(q)$$

Which means that C_1 is optimal for p, and C_2 is optimal for q.

(c) Suppose the average length of the codewords is $\overline{L}'$ when use code C_2 when the distribution is p. Then we have

$$\overline{L}_1' = \sum_x p(x)l_2(x) = 2 \text{ bits}$$

Since we could discover that $l_1(x) = \log\left(\frac{1}{p(x)}\right), l_2(x) = \log\left(\frac{1}{1(x)}\right)$ The exceed parts of the entropy p is

$$\begin{aligned}
\Delta_1 &= \overline{L}_1' - H(p) \\
&= \sum_x p(x) l_2(x) - \sum_x p(x) l_1(x) \\
&= \sum_x p(x) \log\left(\frac{1}{q(x)}\right) - \sum_x p(x) \log\left(\frac{1}{p(x)}\right) \\
&= \sum_x p(x) \log\left(\frac{p(x)}{q(x)}\right) \\
&= D\left(p \| q\right) \\
&= \frac{1}{8} \text{ bits}
\end{aligned}$$

(d) Similarly with (c), we could get that the loss is that

$$\Delta_2 = \overline{L}_2' - H(q) = D\left(q \| p\right) = \frac{1}{8} \text{ bits}$$

Problem 4

3.1 Markov's inequality and Chebyshev's inequality

(a) (Markov's inequality) For any nonnegative random variable X and any $t > 0$, show that

$$\Pr\{X \geq t\} \leq \frac{EX}{t}$$

Exhibit a random variable that achieves this inequality with equality.

(b) (Chebyshev's inequality) Let Y be a random variable with mean μ and variance σ^2. By letting $X = (Y - \mu)^2$, show that for any $\epsilon > 0$,

$$\Pr\{|Y - \mu| > \epsilon\} \leq \frac{\sigma^2}{\epsilon^2}.$$

(c) (Weak law of large numbers) Let $Z_1, Z_2, \ldots, Z_n$ be a sequence of i.i.d. random variables with mean μ and variance σ^2. Let $\overline{Z}_n = \frac{1}{n}\sum_{i=1}^{n} Z_i$ be the sample mean. Show that

$$\Pr\left\{\left|\overline{Z}_n - \mu\right| > \epsilon\right\} \leq \frac{\sigma^2}{n\epsilon^2}.$$

Thus, $\Pr\left\{\left|\overline{Z}_n - \mu\right| > \epsilon\right\} \to 0$ as $n \to \infty$. This is known as the weak law of large numbers.

Solution

(a)

$$\begin{aligned}
\Pr\{X \geq t\} &= \int_t^{+\infty} p(x)dx \\
&\leq \int_t^{+\infty} \frac{x}{t}p(x)dx \qquad (x \geq t, \text{ since x integral from } t) \\
&= \frac{1}{t}\int_t^{+\infty} xp(x)dx \\
&\leq \frac{1}{t}\int_0^{+\infty} xp(x)dx \qquad (x, p(x) \geq 0) \\
&= \frac{1}{t}\mathbb{E}[X] \qquad (\text{definition of expectation, and } X \geq 0)
\end{aligned}$$

For a fixed t, to make the inequality to be an equality, we set X to be discrete random variable. Let $X = \begin{cases} t & \text{w.p. } \theta \\ 0 & \text{w.p. } 1-\theta \end{cases}$, where $\theta \in [0, 1]$.

Then, $\mathbb{E}[X] = \theta t$ and $\Pr\{X \geq t\} = \Pr\{X = t\} = \theta$.

In this case, we have $\Pr\{X \geq t\} = \frac{\mathbb{E}[X]}{t}$.

(b) From the Markov inequality, we have

$$\Pr\left\{X \geq \epsilon^2\right\} \leq \frac{\mathbb{E}[X]}{\epsilon^2}$$

Let $X = (Y - \mu)^2$, then

$$\Pr\left\{(Y-\mu)^2 > \epsilon^2\right\} \leq \Pr\left\{(Y-\mu)^2 \geq \epsilon^2\right\} \leq \frac{\mathbb{E}[(Y-\mu)^2]}{\epsilon^2}$$

From the definition of variance, we have

$$\sigma^2 = \mathbb{E}[(Y - \mathbb{E}(Y))^2] = \mathbb{E}[(Y-\mu)^2]$$

Combine all these together, we have

$$\begin{aligned}\Pr\{|Y-\mu|>\epsilon\} &= \Pr\left\{(Y-\mu)^2>\epsilon^2\right\} \\ &\leq \frac{\mathbb{E}[(Y-\mu)^2]}{\epsilon^2} \\ &= \frac{\sigma^2}{\epsilon^2}\end{aligned}$$

(c) Since $Z_1,\ldots,Z_n$ are i.i.d. random variables, and $\overline{Z}_n = \frac{1}{n}\sum_{i=1}^{n} Z_i$, from the linearity of expectation and variance(among independent variables), we have

$$\begin{aligned}\mathbb{E}\left(\overline{Z}_n\right) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(Z_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu \\ \mathrm{Var}\left(\overline{Z}_n\right) &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left(Z_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n}\end{aligned}$$

So from Chebyshev's inequality, we have

$$\Pr\left\{|\overline{Z}_n-\mu|>\epsilon\right\} \leq \frac{\frac{\sigma^2}{n}}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$$

Problem 5

3.9 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 $\overset{i.i.d.}{\sim} p(x)$.

(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 $\overset{i.i.d.}{\sim} p(x)$. Thus, the odds favoring q are exponentially small when p is true.

Solution

(a)

$$\begin{aligned}
\lim_{n\to\infty} -\frac{1}{n} \log q(X_1, X_2, \ldots, X_n) &= \lim_{n\to+\infty} -\frac{1}{n} \log \prod_{i=1}^{n} q(X_i) \\
&= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \log \frac{1}{q(X_i)} \\
&= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} Y_i \qquad (\text{Let } Y = \log \frac{1}{q(X)}) \\
&= \mathbb{E}(Y) \quad \text{w.p. } 1 \qquad (\text{Strong Law of Large Numbers}) \\
&= \mathbb{E}_{X\sim p(x)} \left[\log \frac{1}{q(X)} \right] \\
&= \sum_x p(x) \log \frac{1}{q(x)}
\end{aligned}$$

Which is the cross entropy between p and q.

(b)

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{n} \log \frac{q(X_1, \ldots, X_n)}{p(X_1, \ldots, X_n)} &= \lim_{n\to\infty} \frac{1}{n} \log \frac{\prod_{i=1}^{n} q(X_i)}{\prod_{i=1}^{n} p(X_i)} \\
&= \lim_{n\to\infty} -\frac{1}{n} \sum_{i=1}^{n} \log \frac{p(X_i)}{q(X_i)} \\
&= -\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} Y_i \qquad (\text{Let } Y = \log \frac{p(X)}{q(X)}) \\
&= -\mathbb{E}(Y) \quad \text{w.p. } 1 \qquad (\text{Strong Law of Large Numbers}) \\
&= -\mathbb{E}_{X\sim p(x)} \left[\log \frac{p(X)}{q(X)} \right] \\
&= -\sum_x p(x) \log \frac{p(x)}{q(x)} \\
&= -D(p||q)
\end{aligned}$$

i.e. when $n \to \infty$,

$$q(X_1, \ldots, X_n) = p(X_1, \ldots, X_n) e^{-nD(p\|q)}$$

Which shows that the odds favoring q are exponentially small when p is true.