Assignment 3

Due: 23:55, 3 March 2019

1. *Source coding.* Let p be a distribution on $\{a, b\}$ with $p(a) = 0.4$ and $p(1) = 0.6$. Draw the curve of $M^*(3, \epsilon)$ for $\epsilon \in [0, 1]$. Specify all the discontinuous points.

Solution: When $n = 2$, the probabilities of the four 2-length inputs are

$$p(0,0) = 0.09,\ p(0,1) = 0.21,\ p(1,0) = 0.21,\ p(1,1) = 0.49.$$

If a 2-length code has only one message, the error probability is at least 0.51 (when encoding only $(1, 1)$). Therefore, $M^*(2, 0.3) > 1$. As the 2-length code with messages $\{(1, 1), (1, 0)\}$ has the error probability 0.3, $M^*(2, 0.3) = 2$.

When $n = 3$, the probabilities of the four 3-length inputs are

$$p(0,0,0) = 0.027,\ p(0,0,1) = p(0,1,0) = p(1,0,0) = 0.063,$$
$$p(1,1,1) = 0.343,\ p(0,1,1) = p(1,0,1) = p(1,1,0) = 0.147.$$

If a 3-length code has only 3 messages, the error probability is all larger than 0.3. Therefore, $M^*(3, 0.3) > 3$. As the 4-length code with messages

$$\{(1,1,1), (1,1,0), (1,0,1), (0,1,1)\}$$

has the error probability $0.216 < 0.3$, $M^*(2, 0.3) = 4$.

Last, $\frac{1}{2} \log M^*(2, 0.3) = 1$ and $\frac{1}{3} \log M^*(3, 0.3) = 2/3$.

2. *Prefix codes.* Consider a probability distribution $p = (p_1, p_2, \ldots, p_m)$ with $p_1 \geq p_2 \geq \cdots \geq p_m$. Let $p' = (p_1, p_2, \ldots, p_{m-2}, p_m + p_{m-1})$. What is the difference between the optimal prefix code lengths for p and p'?

Solution: Huffman codes are optimal prefix codes. For a Huffman code C for p, the codewords for p_m and p_{m-1} are of the longest and of the same length l with difference only in the last symbol. Consider a code C' for p', where the codeword of p_i, $1 \leq i < m - 1$, is the same as the one for p_i in C, and the codeword of $p_m + p_{m-1}$ is the first $l - 1$ symbols of the codeword for p_m in C. According to the Huffman procedure, C' is also a Huffman code. The difference between the codeword lengths of C and C' is

$$p_m l + p_{m-1} l - (p_m + p_{m-1})(l-1) = p_m + p_{m-1}.$$

3. (Huffman coding) Consider the random variable

$$X = \begin{bmatrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 \\ 0.49 & 0.26 & 0.12 & 0.04 & 0.04 & 0.03 & 0.02 \end{bmatrix}$$

   (a) Find a binary Huffman code for X.
   (b) Find the expected code length for the above encoding.

Solution:

(a)

$$\begin{aligned} x_1 &\to 0 \\ x_2 &\to 10 \\ x_3 &\to 110 \\ x_4 &\to 11100 \\ x_5 &\to 11101 \\ x_6 &\to 11110 \\ x_7 &\to 11111 \end{aligned}$$

(b) The expected code length is $1 \cdot 0.49 + 2 \cdot 0.26 + 3 \cdot 0.12 + 5 \cdot (0.04 + 0.04 + 0.03 + 0.02) = 2.02$ bits.

4. Count the exact number of different types in $\mathcal{X}^n$, where $\mathcal{X}$ is a finite set.

Solution: Let $s = |\mathcal{X}|$. For $i = 1, \ldots, s$, let N_i be an non-negative integer such that $\sum_{i=1}^{s} N_i = n$. The number of types is the same as the number of sequence $(N_1, N_2, \ldots, N_s)$, where the latter can be determined as follows: Consider n balls and $s-1$ separators. We apply distinguishable permutation of these $n+s-1$ objects, i.e., all the balls (separators) are treated as the same. In a permutation, the number of balls before the first separator is N_1, the number of balls between the ith and $(i+1)$th separators is N_i, and the number of remaining balls is N_s. The number of distinguishable permutation is $\frac{(n+s-1)!}{n!(s-1)!}$.

5. Let $X^n = (X_1, \ldots, X_n)$ be an i.i.d. sequence of random variables, each of which has a distribution p over a finite set $\mathcal{X}$ and let c be a real number in $(0,1)$. Prove that for any subset A of $\mathcal{X}^n$ with $\Pr\{X^n \in A\} \geq c$ and sufficiently large n,

$$|A \cap W_\delta^n| \geq 2^{n(H(p)-\delta')},$$

where $\delta' \to 0$ as $\delta \to 0$. (Hint: the converse of the block source coding theorem.)

Solution: Let $T = T^n_{[X]\delta}$. On the one hand, we have

$$\begin{aligned} P(A \cap T) &= P(A) - P(A \cap T^c) \\ &\geq P(A) - P(T^c) \\ &\geq P(A) - \sum_{a \in \mathcal{X}: p(a) > 0} \frac{|\mathcal{X}|^2}{4n\delta^2} \\ &\geq c/2, \end{aligned} \tag{1}$$

where the second last inequality is obtained using the similar steps of proving Strong AEP 2, and the last inequality holds when n is sufficiently large.

On the other hand, by Strong AEP I,

$$P(A \cap T) \leq |A \cap T| 2^{-n(H(X) - \eta)}, \tag{2}$$

where $\eta > 0$ and $\eta \to 0$ as $\delta \to 0$. By (1) and (2), we have

$$|A \cap T| \geq \frac{c}{2} 2^{n(H(X) - \eta)} = 2^{n(H(X) - \eta + \frac{\log c/2}{n})}.$$

The proof is completed by letting $\delta' = \eta - \frac{\log c/2}{n}$.