CIE 6020 Assignment 3

WU, Chenhao 117010285

March 2, 2019

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

Answer:

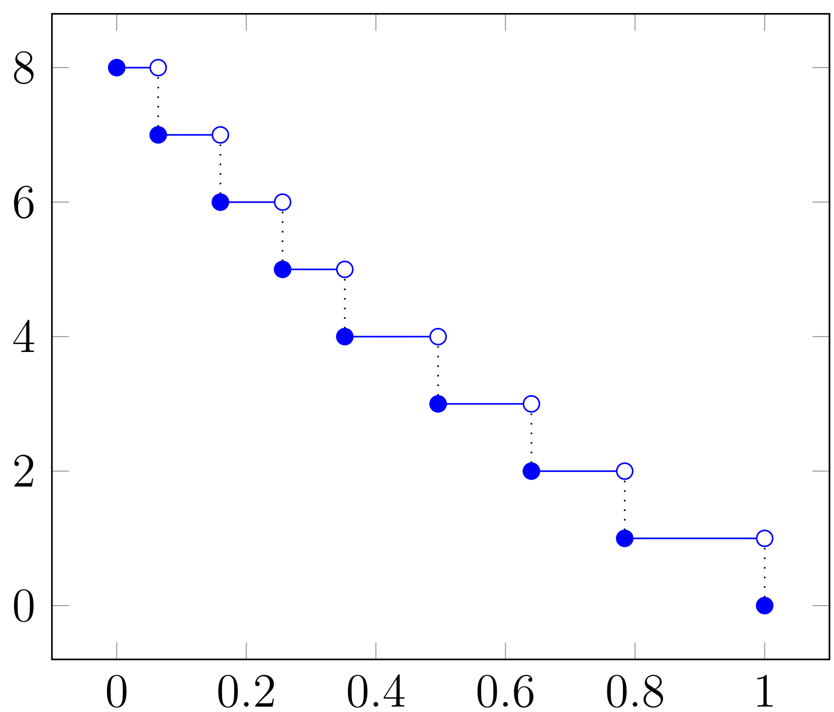
We can calculate the probability of each sequence with length equal to 3.

$$p(aaa) = 0.064$$
$$p(aab) = p(aba) = p(baa) = 0.096$$
$$p(abb) = p(bab) = p(bba) = 0.144$$
$$p(bbb) = 0.216$$

then we can obtain the smallest cardinality of 3-length block code with $\epsilon \in [0, 1]$

$$M^*(3, \epsilon) = \begin{cases} 8 & 0 \leq \epsilon < 0.064 \\ 7 & 0.064 \leq \epsilon < 0.16 \\ 6 & 0.16 \leq \epsilon < 0.256 \\ 5 & 0.256 \leq \epsilon < 0.352 \\ 4 & 0.352 \leq \epsilon < 0.496 \\ 3 & 0.496 \leq \epsilon < 0.64 \\ 2 & 0.64 \leq \epsilon < 0.784 \\ 1 & 0.784 \leq \epsilon < 1 \\ 0 & \epsilon = 1 \end{cases} \tag{1}$$

Curve of M can be drawn as



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

Answer:

From the lower bound for prefix code we can infer that

$$
\begin{aligned}
L_p^*(x) &\geq H(p) = -\sum_{i=1}^{m} p_i \log p_i \\
L_{p'}^*(x) &\geq H(p') = -\sum_{i=1}^{m-2} p_i \log p_i - (p_{m-1} + p_m) \log(p_{m-1} + p_m)
\end{aligned}
$$

in which

$$
p_{m-1} \log(p_{m-1} + p_m) + p_m \log(p_{m-1} + p_m) \geq p_{m-1} \log p_{m-1} + p_m \log p_m
$$

thus, the optimal prefix code length of p is longer than p'

3. *Huffman coding.* Consider the random variable

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

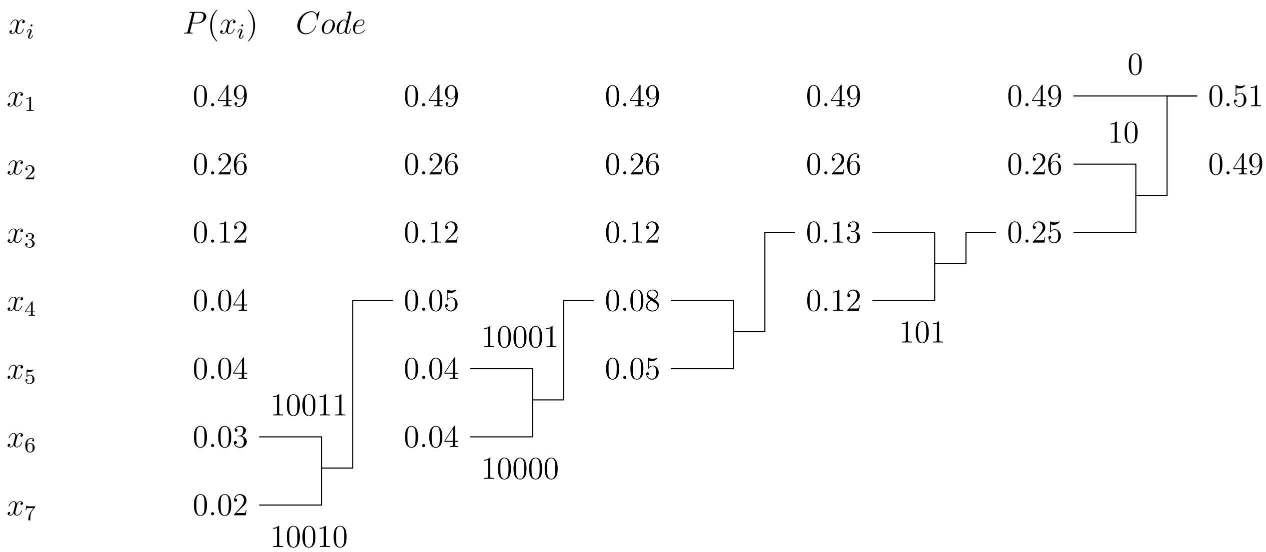
(a) Find a binary Huffman code for X.

(b) Find the expected code length for the above encoding.

Answer:

(a) From the random distribution we can obtain a Huffman Code tree as

in which the Huffman coding can be listed as



Symbol	**Codeword Length**	**Code**
x_1	1	0
x_2	2	10
x_3	3	101
x_4	5	10001
x_5	5	10000
x_6	5	10011
x_7	5	10010

(b) The expect code length can be obtained by

$$\begin{aligned} L_f(p) &= \sum_{a \in \mathcal{A}} p(a) l(a) \\ &= 2.02 (bits) \end{aligned}$$

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

Answer:

Suppose that the alphabet $\mathcal{X} = \{a_1, a_2, a_3, ..., a_{|\mathcal{X}|}\}$ and $N(a|x^n)$ be the number of times that a appears in sequence x^n.

Let P_n be the collection of all possible types of sequences of length n.

$$P_n = \{(P(a_1), P(a_2), ..., P(a_{|\mathcal{X}|}))\}$$

where $N(a_0) + N(a_1) + N(a_2) + ... + N(a_{|\mathcal{X})|}) = n$, thus $|P_n| = \binom{n+2}{|\mathcal{X}|-1}$

5. Let p be any probability distribution over a finite set $\mathcal{X}$ and c be a real number in $(0, 1)$. Prove that for any subset A of $\mathcal{X}^n$ with $p^n(A) \geq c$ and sufficiently large n,

$$|A \cap T^n_{[X]\delta}| \geq 2^{n(H(p)-\delta')}$$

where $\delta' \to 0$ as $\delta \to 0$.

Proof:

The probability that an n-length sequence $\mathbf{x}$ in subset A can be generated as

$$\begin{aligned}
p^n(\mathbf{X}^n \in A \cap T^n_{[X]\delta}) &= \sum_{x \in A \cap T^n_{[X]\delta}} p^n(x) \\
&\leq \sum 2^{-nH(p)-nD(p||A)} \\
&= \sum 2^{-nH(p)-n\sum_{x^n \in \mathcal{X}^n} p^n(x^n) \log \frac{p^n(x^n)}{p^n(A)}} \\
&\leq \sum 2^{-nH(p)+nc} \\
&= |A \cap T^n_{[X]\delta}| 2^{-nH(p)+nc}
\end{aligned}$$

Also, from De Morgan's law and **AEP** II, we can obtain that

$$
\begin{aligned}
1 &\geq p^n(\mathbf{X}^n \in A \cap T^n_{[X]\delta}) \\
&= 1 - p^n(\mathbf{X}^n \notin A) - p^n(\mathbf{X}^n \notin T^n_{[X]\delta}) \\
&\geq 1 - (1-c) - \delta \\
&= c - \delta
\end{aligned}
$$

Combine two inequalities we can derivative that

$$
\begin{aligned}
|A \cap T^n_{[X]\delta}| &\geq (c-\delta)2^{nH(p)-nc} \\
&= 2^{nH(p)-nc+\log(c-\delta)} \\
&= 2^{n(H(p)-c(1-\frac{\log(c-\delta)}{nc}))}
\end{aligned}
$$

Let $\delta' = c(1 - \frac{log(c-\delta)}{nc})$, from weak typicality we can infer that when n is sufficiently large c converges to 0 and thus, $\delta' \to 0$ when $\delta \to 0$.