Data Compression

Homework 1

Student ID:

108368017

Student:

Zi-Yang Lin

Advisor:

Chiu-Ching Tuan

National Taipei University of Technology

2019

Problem 1

Let X be a random variable with an alphabet $H = \{1,2,3,4,5\}$. Please determine $H(X)$ for the following three cases of probability mass function $p(i) = prob[X = i]$. (15%)

(a) $p(1) = p(2) = \frac{1}{2}$

Ans:

$$\begin{aligned} H(X) &= -(\frac{1}{2})\log_2(\frac{1}{2}) + \frac{1}{2}\log_2(\frac{1}{2})) \\ &= -(-\frac{1}{2} - \frac{1}{2}) \\ &= 1\ bits/symbol \end{aligned}$$

(b) $p(i) = \frac{1}{4}, for\ i = 1,2,3,\ and\ p(4) = p(5) = \frac{1}{8}$

Ans:

$$\begin{aligned} H(X) &= -(3 \times \frac{1}{4}\log_2(\frac{1}{4}) + 2 \times \frac{1}{4}\log_2(\frac{1}{4})) \\ &= -(-1.5 - 0.75) \\ &= 2.25\ bits/symbol \end{aligned}$$

(c) $P(i) = 2^{-i}, for\ i = 1,2,3,4, and\ p(5) = \frac{1}{16}$

Ans:

$$\begin{aligned} H(X) &= -(\sum_{i=1}^{4} 2^{-i}\log_2 2^{-i} + \frac{1}{16}\log_2\frac{1}{16}) \\ &= -(0.5 \times (-1) + 0.25 \times (-2) + 0.125 \times (-3) + 0.0625 \times (-4) + 0.0625 \times (-4)) \\ &= 1.875\ bits/symbol \end{aligned}$$

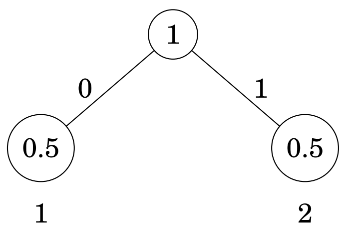
Problem 2

Design a Huffman code C for the source in Problem 1.

(a) Specify your codewords for individual pmf model in Problem 1.

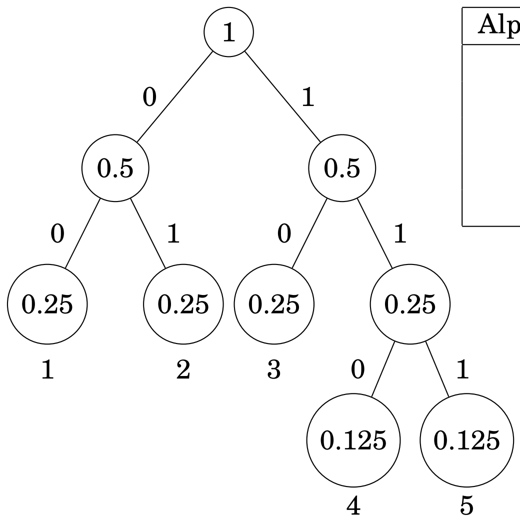
Ans:

(a)



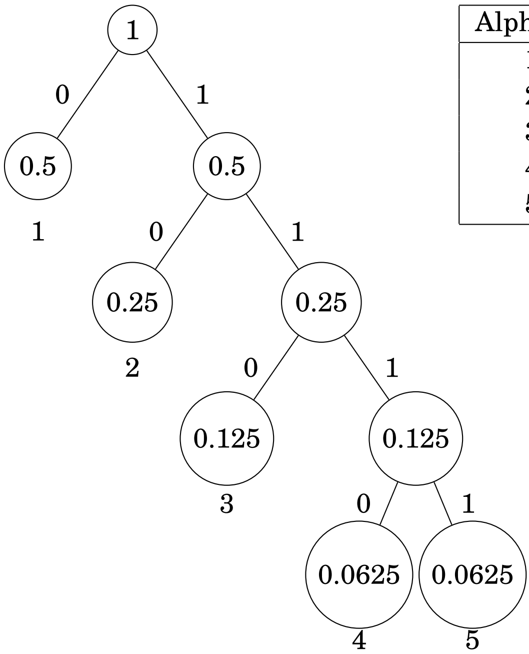
Alphabet	P	Codeword
1	0.5	0
2	0.5	1

(b)



Alphabet	P	Codeword
1	0.25	00
2	0.25	01
3	0.25	10
4	0.125	110
5	0.125	111

(c)



Alphabet	P	Codeword
1	0.5	0
2	0.25	10
3	0.125	110
4	0.0625	1110
5	0.0625	1111

(b) Compute the expected codeword length and compare with the entropy for your codes in (a).

Ans:

(a)

$$\begin{aligned} expected\ codeword\ length &= 0.5 \times 1 + 0.5 \times 1 \\ &= 1\ bits/symbol\ \textbf{(Equal Entropy)} \end{aligned}$$

(b)

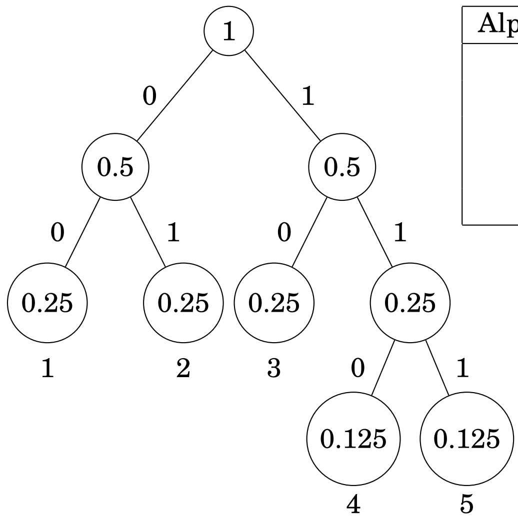
$$\begin{aligned} expected\ codeword\ length &= 0.25 \times 2 + 0.25 \times 2 + 0.25 \times 2 + 0.125 \times 3 + 0.125 \times 3 \\ &= 2.25\ bits/symbol\ \textbf{(Equal Entropy)} \end{aligned}$$

(c)

$$\begin{aligned} expected\ codeword\ length &= 0.5 \times 1 + 0.25 \times 2 + 0.125 \times 3 + 0.0626 \times 4 + 0.0625 \times 4 \\ &= 4.125\ bits/symbol\ \textbf{(NOT Equal Entropy)} \end{aligned}$$

(c) Design a code with minimum codeword length variance for the pmf model in Problem 1.(b)

Ans:



Alphabet	P	Codeword
1	0.25	00
2	0.25	01
3	0.25	10
4	0.125	110
5	0.125	111

Problem 3

Empirical distribution. In the case a probability model is not known, it can be estimated from empirical data. Lets say the alphabet is $H = \{1,2,3,\dots,m\}$. Given a set of observations of length N , the empirical distribution is given by $p = total\ number\ of\ symbol\ 1/N,\ for\ i = 1,2,3,...,m$. Please determine the empirical distribution for **santaclaus.txt**, which is an ASCII file with only lower-cased English letters (i.e., $a \sim z$), space and CR (carriage return), totally 28 symbols. The file can be found on the class web site. Compute the entropy.

Ans:

Problem 4

Write a program that designs a Huffman code for the given distribution in Problem 3.

Ans:

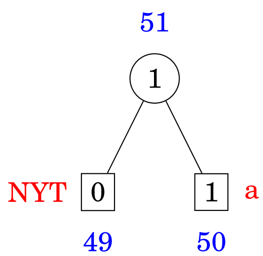
Problem 5

Let X be a random variable with an alphabet H, i.e., the 26 lower-case letters. Use adaptive Huffman tree to find the binary code for the sequence
a a b b a.
You are asked to use the following 5 bits fixed-length binary code as the initial codewords for the 26 letters. That is
a: 00000
b: 00001
⋮
z: 11001
Note: Show the Huffman tree during your coding process.

Ans:

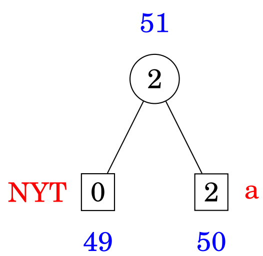
1. Initial step:
   $Total\ nodes = 2m - 1 = 26 \times 2 - 1 = 51$

2. **a** encoded:



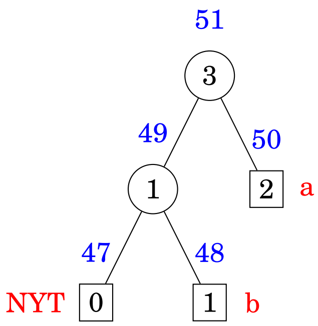
00000
a

3. **a a** encoded:



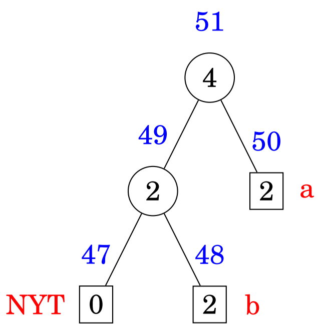
00000 1
a **a**

4. **a a b** encoded:



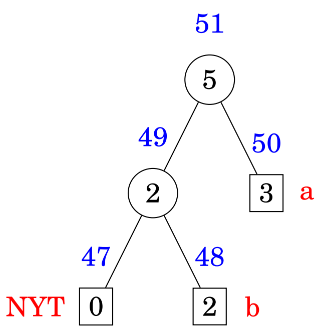
00000	1	0	00001
a	**a**	**NYT**	**b**

5. **a a b b** encoded:



00000	1	0	00001	01
a	**a**	**NYT**	**b**	**b**

6. **a a b b a** encoded:



00000	1	0	00001	01	1
a	**a**	**NYT**	**b**	**b**	**a**

Problem 6

(a) Find the Golomb code of n=21 when m=4.

Ans:

$$2^{\lceil \log_2^m \rceil} - m = 2^2 - 4 = 0$$
$$encoding\ 21 = 21 \div 4 = 5 \ldots 1 = 111110\ 01$$

(b) Find the Golomb code of n=14 when m=4.

Ans:

$$2^{\lceil \log_2^m \rceil} - m = 2^2 - 4 = 0$$
$$encoding\ 14 = 14 \div 4 = 3 \ldots 2 = 1110\ 10$$

(c) Find the Golomb code of n=21 when m=5.

Ans:

$$2^{\lceil \log_2^m \rceil} - m = 2^3 - 5 = 3$$
$$encoding\ 21 = 21 \div 5 = 2 \ldots 1 = 110\ 01$$

(d) Find the Golomb code of n=14 when m=5.

Ans:

$$2^{\lceil \log_2^m \rceil} - m = 2^3 - 5 = 3$$
$$encoding\ 14 = 14 \div 5 = 2 \ldots 4 = 110\ 111$$

(e) A two-integer sequence is encoded by Golomb code with m=4 to get the bitstream 11101111000. Whats the decoded two-integer sequence?

Ans:

$$2^{\lceil \log_2^m \rceil} - m = 2^2 - 4 = 0$$

$$\underset{3}{\underline{1110}} \quad \underset{3}{\underline{11}} \quad \underset{2}{\underline{110}} \quad \underset{0}{\underline{00}}$$

15 8

sequence: 15, 8

(f) A two-integer sequence is encoded by Golomb code with m=5 to get the bitstream 11101111000 (the same bitstream as that in (e)). Whats the decoded two-integer sequence?
Hint: The unary code for a positive integer q is simply q 1s followed by a 0.

Ans:

$$2^{\lceil \log_2^m \rceil} - m = 2^3 - 5 = 3$$

$$\begin{array}{cccc} \underline{1110} & \underline{111} & \underline{10} & \underline{00} \\ 3 & 7\text{-}3 = 4 & 2 & 0 \end{array}$$

$$\begin{array}{cc} 19 & 10 \end{array}$$

sequence: 19, 10