EE5585 Data Compression	January 22, 2013
Lecture 10	
Instructor: Arya Mazumdar	*Scribe: Mandy Ding*

Solutions for Homework 1

Problem 1 Suppose, $\mathcal{X} = \{A, B, C, D\}$. A source produces i.i.d. X symbols from this source,with $Pr(X = A) = P_A, Pr(X = B) = P_B, Pr(X = C) = P_C, Pr(X = D) = P_D$. You are given a file $F = AACDDBBBBBCAABCDAABAADCB$ generated by this source.
(a) What is your best guess for p_A, p_B, p_C, p_D? Reason.
(b) What is the entropy(in bits) of the probability distribution you guess?
(c) What is the Huffman code for the probability distribution that you guessed? What is the average number of bits per symbol?
(d) Encode the file F with the Huffman code you have designed. What is the length of the encoded binary file?What is the average number of bits that have been used for a symbol in this file?

Solution
(a)
The length of file F is 24 bits, and it has 8 of A, 8 of B, 4 of C and 4 of D. Hence we have

$$P_A = \frac{8}{24} \quad P_B = \frac{8}{24} \quad P_C = \frac{4}{24} \quad P_C = \frac{4}{24}$$

Next we prove that this guess is valid by solving the following maximization problem.

$$\begin{aligned} \text{maximize} \quad & f = {P_A}^8 {P_B}^8 {P_C}^4 {P_D}^4 \\ \text{subject to} \quad & 0 \leqslant P_A P_B P_C P_D \leqslant 1 \\ & P_A + P_B + P_C + P_D = 1 \\ & P_D = 1 - P_A - P_B - P_C \end{aligned}$$

Thus, rewrite the cost funciton

$$\begin{aligned} f(P_A, P_B, P_C) &= {P_A}^8 (P_B)^8 (P_C)^4 (1 - P_A - P_B - P_C)^4 \\ g(P_A, P_B, P_C) &= \log f(P_A, P_B, P_C) \\ &= 8 \log P_A + 8 \log P_B + 4 \log P_C + 4 \log P_D \end{aligned}$$

Get the partial derivatives of g

$$\begin{aligned} \frac{\partial g}{\partial P_A} &= \frac{8}{P_A} - \frac{4}{1 - P_A - P_B - P_C} = 0 \\ \frac{\partial g}{\partial P_B} &= \frac{8}{P_B} - \frac{4}{1 - P_A - P_B - P_C} = 0 \\ \frac{\partial g}{\partial P_C} &= \frac{8}{P_C} - \frac{4}{1 - P_A - P_B - P_C} = 0 \end{aligned}$$

By solving the above equations, we get the solutions

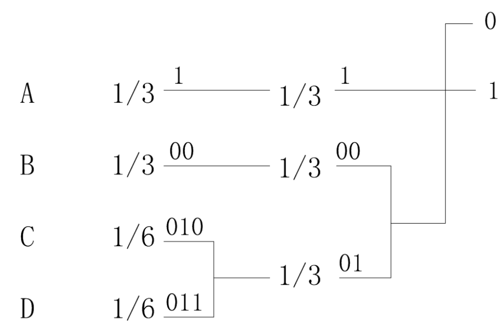
$$P_A = \frac{1}{3} \quad P_B = \frac{1}{3} \quad P_C = \frac{1}{6} \quad P_D = \frac{1}{6}$$

which are the same as we guess.

(b)

$$H(X) = -\frac{1}{3}\log\frac{1}{3} - \frac{1}{3}\log\frac{1}{3} - \frac{1}{6}\log\frac{1}{6} - \frac{1}{6}\log\frac{1}{6} = 1.92 \quad \text{bits/symbol}$$

(c)



Thus, the Huffman code set is

$$\begin{array}{ll} A & 1 \\ B & 00 \\ C & 010 \\ D & 011 \end{array}$$

The average codeword length is $\frac{1}{3} \times 1 + \frac{1}{3} \times 2 + \frac{1}{6} \times 3 + \frac{1}{6} \times 3 = 2 \quad bits/symbol$

(d)
Encode the file with the Huffman code we designed in (c), we get the binary sequence

$$110100110110000000000010110001001111001101101000$$

Length of this binary sequence is $1 \times 8 + 2 \times 8 + 3 \times 4 + 3 \times 4 = 48$ bits
Average number of bits per symbol is $\frac{48}{24} = 2$ bits/symbol

Problem 2 Consider the code $0, 01$. Is this code uniquely decodable?Why?Is it instantaneous?

Solution
The code is uniquely decodable but not instantaneous.

Problem 3 Suppose $\chi = \{0, 1\}$. The random variable(source) X takes value in χ, with $Pr(X = 0) = \frac{3}{4}$ and $Pr(X = 0) = \frac{1}{4}$. What is the probability that the source produce a sequence 0000011111?

Solution

$$P = (\frac{3}{4})^5 (\frac{1}{4})^5$$

Problem 4 Write the Lempel-Ziv parsing for the file F in Problem 1. What is the number of bits that you need to write the entire compressed file(with LZ algorithm).

Solution
The file F is parsed as following

$$F = \underset{1}{A}|\underset{2}{AC}|\underset{3}{D}|\underset{4}{DB}|\underset{5}{B}|\underset{6}{BB}|\underset{7}{BC}|\underset{8}{AA}|\underset{9}{BCD}|\underset{10}{AAB}|\underset{11}{AAD}|\underset{12}{C}|\underset{13}{B}$$

Therefore, the corresponding codewords are

$$(0,\mathrm{A})(1,\mathrm{C})(0,\mathrm{D})(3,\mathrm{B})(0,\mathrm{B})(5,\mathrm{B})(5,\mathrm{C})(1,\mathrm{A})(7,\mathrm{D})(8,\mathrm{B})(8,\mathrm{D})(0,\mathrm{C})$$

Generate initial dictionary

A	00
B	01
C	10
D	11

By encoding, we get the binary sequence

$$\underset{000000}{(0,A)} \quad \underset{000110}{(1,C)} \quad \underset{000011}{(0,D)} \quad \underset{001101}{(3,B)} \quad \ldots$$

As shown above, each phrase can be decoded into 6 bits, and we have 13 phrases, thus the length of the entire file is $6 \times 13 = 78 \quad bits/symbol$

Rate Distortion Theory

In the previous class, we have proved $R(D) = 1 - h(D)$. If there is a n-bits sequence allowing nD bits distortion, what is the optimal rate of compression?

Information Theory

Entropy

$$H(X) = -\sum_x p(x) \log p(x)$$

where $p(x) = Pr(X = x)$

Conditional Entropy

$$\begin{aligned} H(X|Y) &= \sum_y H(X|Y = y) \\ &= -\sum_x \sum_y p(y)p(x|y) \log p(x|y) \\ &= -\sum_x \sum_y p(x,y) \log p(x|y) \end{aligned}$$

which is the conditional entropy of X given Y

Joint Entropy

$$H(X,Y) = -\sum_x \sum_y p(x,y) \log p(x,y)$$

Claim 1 $H(X,Y) = H(Y) + H(X|Y)$

$$\begin{aligned} H(Y) + H(X|Y) &= -\sum_y p(y)\log p(y) - \sum_x \sum_y p(x,y)\log p(x|y) \\ &= -\sum_y p(y)\log p(y) - \sum_x \sum_y p(x,y)\log p(x|y) \\ &= -\sum_x \sum_y p(x,y)\log p(y) - \sum_x \sum_y p(x,y)\log p(x|y) \\ &= -\sum_{x,y} p(x,y)\log p(x,y) \\ &= H(X,Y) \end{aligned}$$

Claim 2 $H(Y) \geqslant H(Y|X)$ with equality when X and Y are independent.

$$\begin{aligned} H(Y) - H(Y|X) &= -\sum_y p(y)\log p(y) + \sum_x \sum_y p(x,y)\log p(y|x) \\ &= -\sum_x \sum_y p(x,y)\log p(y) + \sum_x \sum_y p(x,y)\log p(y|x) \\ &= \sum_x \sum_y p(x,y)\log \frac{p(x,y)}{p(x)p(y)} \\ &= D(p(x,y)\|p(x)p(y)) \end{aligned}$$

By using the property of divergence, $D(X\|Y) \geqslant 0$ where X,Y are random variable, we get

$$H(Y) \geqslant H(Y|X)$$

<u>Mutual Information</u>

$$I(X,Y) = H(Y) - H(Y|X) = H(Y) - H(X|Y)$$

In the definition of mutual information, $H(Y)$ means the uncertainty in Y and $H(Y|X)$ means the uncertainty in Y if X is known. This equality implies that the amount of information of X gives about Y is the same as the that Y gives about X.

Property 1 $I(X,Y) = I(Y,X)$

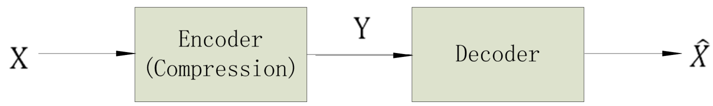
$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$

$$H(X) - H(X|Y) = H(Y) - H(Y|X)$$

$$I(X,Y) = I(Y,X)$$

<u>Rate Distortion</u>

Suppose we have $X_1, X_2, X_3 \dots X_n$ which are $i.i.d$ random variables, and then compressed them into $Y_1, Y_2, Y_3 \dots Y_k$. At the receiver, $Y_1, Y_2, Y_3 \dots Y_k$ are decoded into $\hat{X}_1, \hat{X}_2, \hat{X}_3 \dots \hat{X}_n$



Let R(D) represent the optimal compression rate given D is the normalized distortion, we have $\frac{1}{n}Ed(X^n, \hat{X^n}) \leqslant D$ and define

$$R(D) = \min_{p(\hat{x}|x):\sum p(\hat{x},x)d(x,\hat{x})\leq D} I(X;\hat{X})$$