10-704: Information Processing and Learning **Spring 2012**

Homework 3: Solution

Lecturer: Aarti Singh

Note: *The TA graciously thank Rafael Stern for providing all of these solutions*

3.1 Problem 1

3.1.1 (a)

If $p \leq 0.5$, $\hat{X} = Y$ and, if $p > 0.5$, $\hat{X} = (1 - Y)$. The probability of getting a given bit of X right is $\max\{p, 1-p\}$. Hence, since the bits of X are independent, the probability of getting at least one bit wrong, $P(E)$, is such that $P(E) = 1 - \max\{p, 1-p\}^n$.

3.1.2 (b)

Fano's Inequality gives:

$$H(Y|X) \leq H(E) + P(E=1)\log(2^n)$$

$$P(E=1) \geq \frac{nH(p) - H(E)}{n} \geq H(p) - \frac{H(0.5)}{n}$$

This gives us a satisfactory solution to the homework problem but we can go on to derive a much tighter bound:

Observe that $\hat{X}$ is the Bayes Estimator for X and, thus, comparing (a) and (b), Fano's Inequality is very loose in this application. Going back to the proof of Theorem 2.3 in Lecture Notes 2, observe that, at some point, we have the following bound:

$$H(X|Y) \leq H(E) + H(X|E=1,Y)P(E=1)$$

$$P(E=1) \geq \frac{H(X|Y) - H(E)}{H(X|E=1,Y)}$$

Using $H(X|E=1,Y) \leq \log(|\chi| - 1)$ is a crude approximation which completes Fano's Inequality. We can expect this approximation to be reasonably good if $X|Y$ is close to a uniform distribution. Since, for $p \neq .5$, this is not the case, the approximation turns out to be bad.

Instead, recall that:

$$H(X|Y) \geq H(X|Y,E) = P(E=1)H(X|Y,E=1) + P(E=0)H(X|Y,E=0)$$

From $H(X|Y,E=0)=0$, conclude that $H(X|Y,E=1) \leq H(X|Y)$. Plugging in to the previous bound:

$$P(E=1) \geq \frac{H(X|Y) - H(E)}{H(X|Y)} = 1 - \frac{H(E)}{nH(p)}$$

Which, at least, shows that for large sample sizes, the problem is hopeless.

3.2 Problem 2

3.2.1 (a)

1 is encoded as 1 and 3 is encoded as 11. Hence, both $(1,1)$ and (3) are encoded as 11 and the code is not uniquely decodable.

3.2.2 (b)

Assume the code is not a prefix code. Hence, there exist two codewords c^1 and c^2 respectively of sizes m and n, such that:

$$c^2 = c^1_1 c^1_2 \ldots c^1_{2m-1} c^1_{2m} c^2_{2m+1} c^2_{2m+2} \ldots c^2_{2n-1} c^2_{2n}$$

Since $c^1_{2m-1} \neq c^1_{2m}$, conclude that $c^2_{2m-1} \neq c^2_{2m}$. This is a contradiction since, by construction, for any $i < n$, $c^2_{2i-1} = c^2_{2i}$. Hence, the code is a prefix code.

3.2.3 (c)

The length of the codeword for n is $2\lceil \log((n \vee 1)+1) \rceil + 2$.

3.2.4 (d)

Take two arbitrary numbers $m < n$ and define their codewords as $c(m)$ and $c(n)$. Assume $l(m) \neq l(n)$. In order for $c(m)$ to be a prefix of $c(n)$, the codification of $l(m)$ must be a prefix of the codification for $l(n)$. By the same reasoning as in item (b) this condition cannot hold. Hence, if $l(m) \neq l(n)$, $c(m)$ is not a prefix of $c(n)$. Assume $l(m) = l(n)$. Hence, $l(c(m)) = l(c(n))$ and $c(m)$ cannot be a prefix of $c(n)$. Hence, c is a prefix code.

3.2.5 (e)

$2\lceil \log(\lceil \log((n \vee 1)+1) \rceil + 1) \rceil + 2 + \lceil \log((n \vee 1)+1) \rceil$

3.3 Problem

3.3.1 (a)

$a = 0$, $b = 10$, $c = 110$, $d = 111$.

3.3.2 (b)

Not unique. For example, another Huffman code would be $a = 00$, $b = 01$, $c = 10$, $d = 11$. The expected length of the code is 2.

3.3.3 (c)

The length of a particular codeword using Huffman encoding might be larger than that for a Shannon code. For example, in the encoding in $3a$, $l(c) = 3$. On the other hand, the length for c using a Shannon code would be $-\log(.25) = 2$.

3.4 Problem 4

3.4.1 (a)

Let there be a symbol, ξ, which occurs with probability $p > \frac{2}{5}$. Assume all codewords have length greater or equal to 2. Hence, there must be at least 4 symbols. Look at the Huffman coding scheme when only 4 groups are left. Label these groups as a, b, c, d in such a way that, in the next steps, a merges with b and, after that, c merges with d. This labeling is possible since all codewords have length greater or equal to 2. By Huffman Coding, conclude that: $P(a \cup b) \geq \max\{P(c), P(d)\}$ and that $\min\{P(c), P(d)\} \geq \max\{P(a), P(b)\} \geq \frac{P(a \cup b)}{2}$.

Assume ξ is in a or in b. Hence, $P(c) > 0.4$ and $P(d) > 0.4$. Obtain that $P(a) + P(c) + P(d) > 3 * 0.4 > 1$, a contradiction. Assume ξ is in c or d. By Huffman coding, conclude that $P(a \cup b) > P(\{\xi\}) = 0.4$. Using the relation in the previous paragraph, $\min\{P(c), P(d)\} > 0.2$. Hence $P(a) + P(b) > 0.4$, $\max\{P(c), P(d)\} > 0.4$ and $\min\{P(c), P(d)\} > 0.2$. Conclude that, $P(a) + P(b) + P(c) + P(d) > 1$, a contradiction. Hence, there must be a symbol with codeword of length equal to 1.

3.4.2 (b)

In the last stage of Huffman's coding scheme, there are three groups left. If there is a symbol with codeword 1, there is a symbol ξ such that these groups are $a = \{\xi\}$, b and c and b is merged with c. Hence, by Huffman's coding, $P(b) < P(a) < \frac{1}{3}$ and $P(c) < P(a) < \frac{1}{3}$. Thus, $P(a) + P(b) + P(c) < 1$, a contradiction. Conclude that, if all symbols have probability lesser than $\frac{1}{3}$, no codeword has length 1.

3.5 Problem 5

3.5.1 (a)

Using arithmetic coding, the message must be codified by a number in $[\frac{1}{3}+\frac{1}{9}, \frac{1}{3}+\frac{1}{9}+\frac{1}{81})$. In ternary notation, the middle point of this interval is $0.1100111\ldots$. The Shannon Information content of this sequence is $\log_3(81) = 4$. Hence, using Shannon-Fano-Elias rounding, the codeword would be 11001.

3.5.2 (b)

Using arithmetic coding, the first block (bb) must be codified by a number in $[\frac{1}{3}+\frac{1}{9}, \frac{1}{3}+\frac{2}{9}]$. In ternary notation, the middle point of this interval is $0.111\ldots$. The Shannon Information content of this block is $\log_3(9) = 2$. Hence, using Shannon-Fano-Elias rounding, the codeword would be 111.

Similarly, the second block (aa) must be codified by a number in $[0, \frac{1}{9}]$. In ternary notation, the middle point of this interval is $0.00111\ldots$. The Shannon Information content of this block is $\log_3(9) = 2$. Hence, using Shannon-Fano-Elias rounding, the codeword would be 001.

Hence, the encoding would be 111001.

3.5.3 (c)

Consider the messages c and d. The first one must be encoded in the interval $[\frac{2}{3}, \frac{2}{3}+\frac{1}{4}]$ and the second in $[\frac{2}{3}+\frac{1}{4}, 1]$. Hence, using the lower end of the interval, the codeword for c would be 2 and the codeword for d would start with 21. Hence, using the lower end of the intervals does not generate a prefix code.

3.5.4 (d)

Since 111001 starts with a 1, we know that the first letter must be b. Similarly, 11 indicates that there is another b. Hence the first block is bb. Next, a 0 indicates that the first symbol of the next block must be a. Similarly, 01 indicates that the next symbol must also be a. Thus, the second block is aa and the entire message is $bbaa$.