CSC 310, Spring 2002 — Solutions to Assignment #1

Question 1: 30 marks, 2 for each answer.

A) $\{xx,\ xz,\ y,\ zz,\ xyz\}$

1) Yes, the code satisfies the inequality, since

$$\frac{1}{3^2}+\frac{1}{3^2}+\frac{1}{3^1}+\frac{1}{3^2}+\frac{1}{3^3} = \frac{3}{27}+\frac{3}{27}+\frac{9}{27}+\frac{3}{27}+\frac{1}{27} = \frac{19}{27} \le 1$$

2) Yes, the code is instantaneously decodable.

3) Yes, the code is uniquely decodable, since it is instantaneously decodable.

B) {000, 10, 00, 11}

1) Yes, the code satisfies the inequality, since

$$\frac{1}{2^3}+\frac{1}{2^2}+\frac{1}{2^2}+\frac{1}{2^2} = \frac{1}{8}+\frac{2}{8}+\frac{2}{8}+\frac{2}{8} = \frac{7}{8} \le 1$$

2) No, the code is not instantaneously decodable, since 00 is a prefix of 000.

3) No, the code is not uniquely decodable. For example, the sequence 000000 could be decoded as 00, 00, 00 or as 000, 000.

C) {100, 101, 0, 11}

1) Yes, the code satisfies the inequality, since

$$\frac{1}{2^3}+\frac{1}{2^3}+\frac{1}{2^1}+\frac{1}{2^2} = \frac{1}{8}+\frac{1}{8}+\frac{4}{8}+\frac{2}{8} = \frac{8}{8} \le 1$$

2) Yes, the code is instantaneously decodable.

3) Yes, the code is uniquely decodable, since it is instantaneously decodable.

D) {01, 100, 011, 00, 111, 1010, 1011, 1101}

1) No, the code does not satisfy the inequality, since

$$\frac{1}{2^2}+\frac{1}{2^3}+\frac{1}{2^3}+\frac{1}{2^2}+\frac{1}{2^3}+\frac{1}{2^4}+\frac{1}{2^4}+\frac{1}{2^4} = \frac{4}{16}+\frac{2}{16}+\frac{2}{16}+\frac{4}{16}+\frac{2}{16}+\frac{1}{16}+\frac{1}{16}+\frac{1}{16} = \frac{17}{16} > 1$$

2) No, the code is not instantaneously decodable, since 01 is a prefix of 011.

3) No, the code is not uniquely decodable. For example, the sequence 01100 could be decoded as 01, 100 or as 011, 00.

E) {01, 111, 011, 00, 010, 110}

1) Yes, the code satisfies the inequality, since

$$\frac{1}{2^2}+\frac{1}{2^3}+\frac{1}{2^3}+\frac{1}{2^2}+\frac{1}{2^3}+\frac{1}{2^3} = \frac{2}{8}+\frac{1}{8}+\frac{1}{8}+\frac{2}{8}+\frac{1}{8}+\frac{1}{8} = \frac{8}{8} \le 1$$

2) No, the code is not instantaneously decodable, since 01 is a prefix of 011, and also of 010.

3) Yes, the code is uniquely decodable. Applying the procedure of Section 1.2 of Jones & Jones (see also the slides for lecture 2a), we see that

$$\begin{aligned} \mathcal{C} = \mathcal{C}_0 &= \{01,\ 111,\ 011,\ 00,\ 010,\ 110\} \\ \mathcal{C}_1 &= \{1,0\} \\ \mathcal{C}_2 &= \{11,10,1,0\} \\ \mathcal{C}_3 &= \{11,10,1,0\} \end{aligned}$$

and $\mathcal{C}_n = \{11,10,1,0\}$ for all larger n. From this, we can find that $\mathcal{C}_\infty = \{11,10,1,0\}$, which is disjoint from $\mathcal{C}$. The Sardinas-Patterson Theorem therefore guarantees that the code is uniquely decodable.

One can also show that the code is uniquely decodable by noting that the code obtained by reversing all the codewords is instantaneously decodable. One can therefore easily decode a message by scanning it from the end to the beginning.

Question 2: 30 marks.

This can be proved by induction on the value of k.

First, for $k = 1$, it is obvious that any Huffman code with only $2^k = 2$ symbols will have codewords that are all of length $k = 1$, since the two symbols will be given codewords of "0" and "1".

Next, we assume that codewords in Huffman codes for alphabets of size 2^{k-1} with probabilities that satisfy $p_i/p_j < 2$ are all of length $k-1$, and we try to then show that the codewords are all of length k in a Huffman code for an alphabet of size 2^k with probabilities satisfying $p_i/p_j < 2$.

We do this by visualizing how the Huffman code will be created. First, the two symbols with the smallest probabilities, say p_a and p_b, will be merged to create a combined symbol with probability $p_a + p_b$. This combined probability will be greater than the probability of any single symbol, say p_c, because $p_c/p_a < 2$ and $p_c/p_b < 2$ imply that $p_a > p_c/2$ and $p_b > p_c/2$, which in turn imply that $p_a + p_b > p_c/2 + p_c/2 = p_c$. The next step in the Huffman procedure will therefore merge another two of the original symbols (not the combined symbol formed in the previous step). The probability for this combined symbol will also be greater than the probability of any single symbol. Continuing, the Huffman procedure will merge the original symbols in pairs until no original symbols are left. Since 2^k is an even number, there will not be a symbol left over after the last such pair is merged.

At this point, we have 2^{k-1} merged symbols. Consider two such merged symbols, with merged probabilities $p_a + p_b$ and $p_c + p_d$, where p_a, p_b, p_c, and p_d are probabilities of symbols in the original alphabet. The merged symbol probabilities will satisfy $(p_a + p_b) / (p_c + p_d) < 2$, since $p_a/p_c < 2$ and $p_b/p_d < 2$ imply that $p_a < 2p_c$ and $p_b < 2p_d$, which in turn imply that $p_a + p_b < 2p_c + 2p_d = 2(p_c + p_d)$. The probabilities for these 2^{k-1} merged symbols therefore satisfy the requirements for applying our inductive hypothesis, which allows us to conclude that the Huffman procedure will assign each of these merged symbols a codeword of length $k - 1$. The codewords for each of the original symbols that were merged to form them will be one bit longer, for a total of k bits.

We have therefore proved that the statement is true for alphabets of size 2, and it is true for alphabets of size 2^k if it is true for alphabets of size 2^{k-1}. It is therefore true for alphabets of size 2^k, for any positive integer k.

Question 3: 40 marks total.

Part A: 15 marks.

Here is a table of the possible blocks of size two, their probabilities, and the codewords assigned to them by the Huffman procedure. (Other Huffman codes are also possible, but they will have the same average length as this one.)

Block	Probability	Codeword
`WW`	$0.99 \times 0.99 = 0.9801$	`0`
`WB`	$0.99 \times 0.01 = 0.0099$	`11`
`BW`	$0.01 \times 0.99 = 0.0099$	`100`
`BB`	$0.01 \times 0.01 = 0.0001$	`101`

The average codeword length for this code is

$$0.9801 \times 1 \;+\; 0.0099 \times 2 \;+\; 0.0099 \times 3 \;+\; 0.0001 \times 3 \;=\; 1.0299$$

Part B: 15 marks.

Here is a table of the possible blocks of size three, their probabilities, and the codewords assigned to them by the Huffman procedure. (Other Huffman codes are also possible, but they will have the same average length.)

Block	Probability	Codeword
`WWW`	$0.99 \times 0.99 \times 0.99 = 0.970299$	`0`
`WWB`	$0.99 \times 0.99 \times 0.01 = 0.009801$	`100`
`WBW`	$0.99 \times 0.01 \times 0.99 = 0.009801$	`101`
`BWW`	$0.01 \times 0.99 \times 0.99 = 0.009801$	`110`
`WBB`	$0.99 \times 0.01 \times 0.01 = 0.000099$	`11100`
`BWB`	$0.01 \times 0.99 \times 0.01 = 0.000099$	`11101`
`BBW`	$0.01 \times 0.01 \times 0.99 = 0.000099$	`11110`
`BBB`	$0.01 \times 0.01 \times 0.01 = 0.000001$	`11111`

The average codeword length for this code is

$$0.970299 \times 1 \;+\; 0.009801 \times (3+3+3) \;+\; 0.000099 \times (5+5+5) \;+\; 0.000001 \times 5 \;=\; 1.06$$

Part C: 10 marks.

(3 marks) The binary entropy is $0.99 \log_2(1/0.99) + 0.01 \log_2(1/0.01) \;=\; 0.08079$ bits.

(3 marks) The average bits per symbol using a Huffman code for blocks of length two is $1.0299/2 \;=\; 0.51495$, and for blocks of length three, it is $1.06/3 \;=\; 0.35333$. Both are quite a ways above the entropy.

(4 marks) A Huffman code for blocks of length n must have average length at least 1 (since codewords are at least one bit long), so the average number of bits per symbol using blocks of size n will be at least $1/n$. Ten percent above the entropy is $1.1 \times 0.080179 = 0.08819$. We must therefore use a block size for which $1/n \leq 0.08819$, which implies $n \geq 1/0.08819 = 11.339$. Since n must be an integer, we will need to use blocks of at least $n = 12$ symbols.