ECE 154C Homework Assignment #1

(Due: Monday, April 22, 2019)

Reading assignment: Jack Wolf lecture notes pp. 1-39 and lecture slides 1-83 from spring 2014

Recommended: Solve HW #1 and HW #2 from spring 2017 for practice (appended at the end)

Programming assignment: You can use any programming language you prefer (MATLAB, Python, C/C++, or Julia, for example). Write down your code as clearly as possible and add suitable comments. Submit your code and the output plots to AFAZELIC@UCSD.EDU or turn in the hard copies during the lecture on Monday morning.

Consider a source with probability vector $pmf = \{\alpha, 1-\alpha\}$:

a) Write a program for a function $binaryHuffman(pmf, n)$ that takes a probability vector pmf and a source block-length n, constructs the binary Huffman code and then outputs the average number of code symbols per source symbols. Please note that for a source block-length of n, the corresponding pmf consists of 2^n elements.
   - Generate a plot where Y axis denotes the average coded symbols/source symbols and X axis denotes n for $n = 1,2,4,8,16$ in log scale. Assume $\alpha = 0.1$.

b) Write a program for a function $binaryShannonFano(pmf, n)$ that takes a probability vector pmf and a source block-length n, constructs the binary Shannon Fano code and then outputs the average number of code symbols per source symbols.
   - Generate a plot where Y axis denotes the average coded symbols/source symbols and X axis denotes n for $n = 1,2,4,8,16$ in log scale. Assume $\alpha = 0.1$.

c) Write a program for a function $maryHuffman(pmf, n, m)$ that takes a probability vector pmf, the coded alphabets size m, and a source block-length n as input, constructs an m-ary Huffman code and then outputs the average number of code symbols per source symbols. The program should also add phantom symbols if necessary.
   - Generate a plot where Y axis denotes the average coded symbols/source symbols and X axis denotes n for $n = 1,2,4,8,16$ in log scale. Assume $\alpha = 0.1, m = 3$.

d) Write a program for a function $binaryTunstall(pmf, 2^L)$ that takes a probability vector pmf and the number of source phrases 2^L as input, constructs a variable length to fixed length Tunstall code with 2^L binary code words and then outputs the average number of code symbols per source symbols. Note that no phantom symbols are needed for this case.
   - Generate a plot where Y axis denotes the average coded symbols/source symbols and X axis denotes L for $L = 1,2,4,8,16$ in log scale. Assume $\alpha = 0.1$.

UCSD ECE154C
Prof. Young-Han Kim

Handout #1
Thursday, April 6, 2017

Homework Set #1
(Due: Thursday, April 13, 2017)

Your answer should be as clear and readable as possible. You should not rely on a calculator or computer in solving these problems.

1. For each of the following codes, determine if it is non-singular, uniquely decodable, or instantaneous.

   (a) {01, 10}

   (b) {0, 01, 10}

   (c) {0, 10, 11}

   (d) {110, 11, 100, 00, 01}

2. Consider a source which produces an i.i.d. sequence of symbols from the alphabet $\{A, B, C\}$ with probabilities $\{0.5, 0.25, 0.25\}$ respectively. For $n = 1$, 2, and 3, find binary Huffman codes for taking n source symbols at a time. In each case compute the average number of binary code symbols per source symbol and compare the results.

3. Repeat Problem 2 with probabilities $\{0.4, 0.35, 0.25\}$.

4. Repeat Problems 2 and 3 with Shannon–Fano codes.

UCSD ECE154C Handout #2
Prof. Young-Han Kim Thursday, April 13, 2017

Solutions to Exercise Set #1

1. For each of the following codes, determine if it is non-singular, uniquely decodable, or instantaneous.

   (a) {01, 10}
   (b) {0, 01, 10}
   (c) {0, 10, 11}
   (d) {110, 11, 100, 00, 01}

   Solution:

   (a) It is non-singular, uniquely decodable, and instantaneous.
   (b) It is non-singular. It is not uniquely decodable since 010 can be parsed as $01, 0$ or $0, 10$. It is not instantaneous since it is not uniquely decodable.
   (c) It is non-singular, uniquely decodable, and instantaneous.
   (d) It is non-singular. It is not uniquely decodable since 110100 can be parsed as $110, 100$ or $11, 01, 00$. It is not instantaneous since it is not uniquely decodable.

2. Consider a source which produces an i.i.d. sequence of symbols from the alphabet $\{A, B, C\}$ with probabilities $\{0.5, 0.25, 0.25\}$ respectively. For $n = 1$, 2, and 3, find binary Huffman codes for taking n source symbols at a time. In each case compute the average number of binary code symbols per source symbol and compare the results.

   Solution:

   The entropy of the source with probabilities $\{0.5, 0.25, 0.25\}$ is given as

   $$H(S) = \sum_{i=1}^{3} p_i \log \frac{1}{p_i} = 1.5 \text{ bits}.$$

   Continued on scanned pages.

3. Repeat Problem 2 with probabilities $\{0.4, 0.35, 0.25\}$.

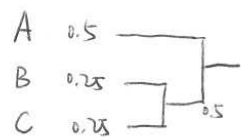
Solution:

The entropy of the source with probabilities $\{0.4, 0.35, 0.25\}$ is given as

$$H(S) = \sum_{i=1}^{3} p_i \log \frac{1}{p_i} = 1.559 \text{ bits.}$$

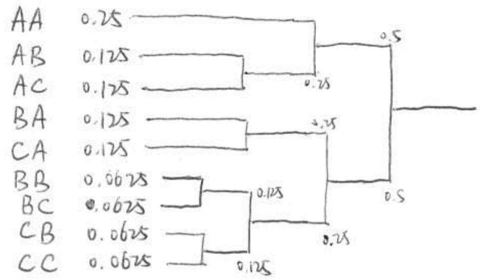
Continued on scanned pages.

$n=1$



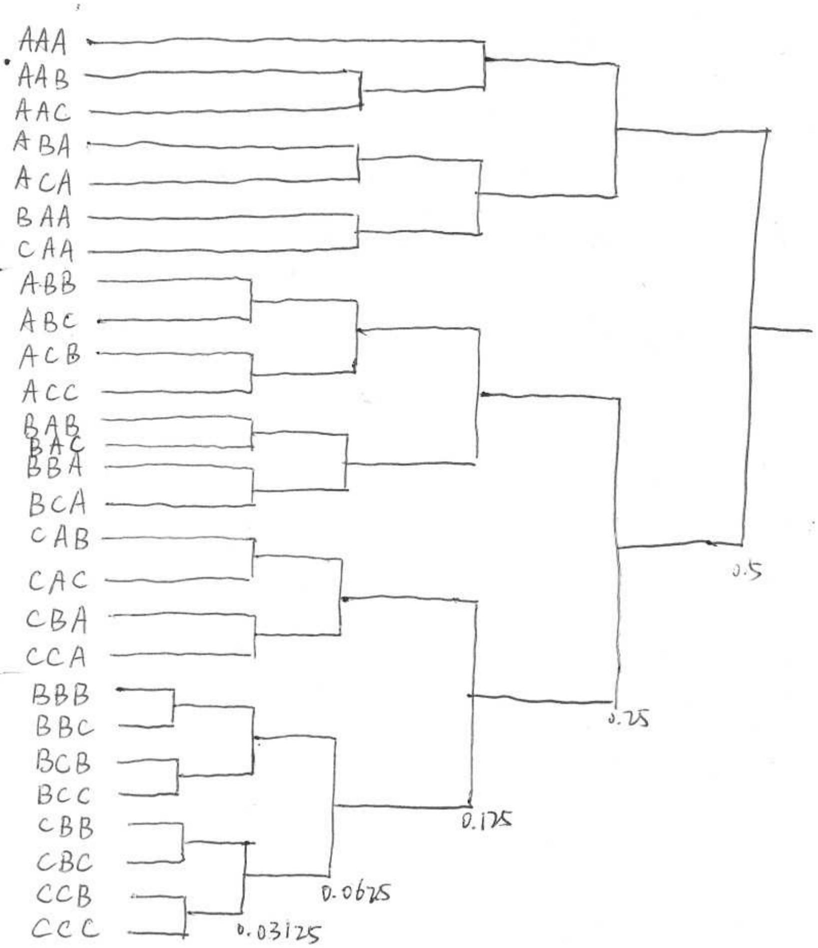
$\bar{L}_1 = 1.5$

$n=2$



$\bar{L}_2 = 3$ $\frac{\bar{L}_2}{2} = 1.5$

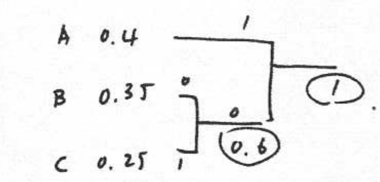
$n=3$



$\bar{L}_3 = 4.5$

$\frac{\bar{L}_3}{3} = 1.5$

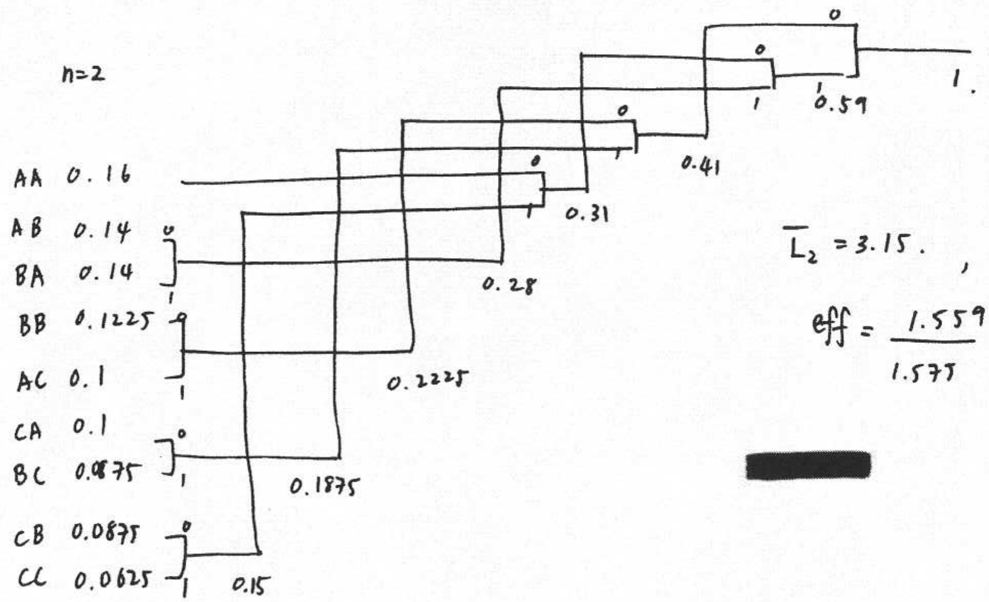
1. $n=1$



A = 1
B = 00
C = 01

$\bar{L}_1 = 1.6$, $\text{eff} = \frac{1.559}{1.6} = 97.4\%$

$n=2$

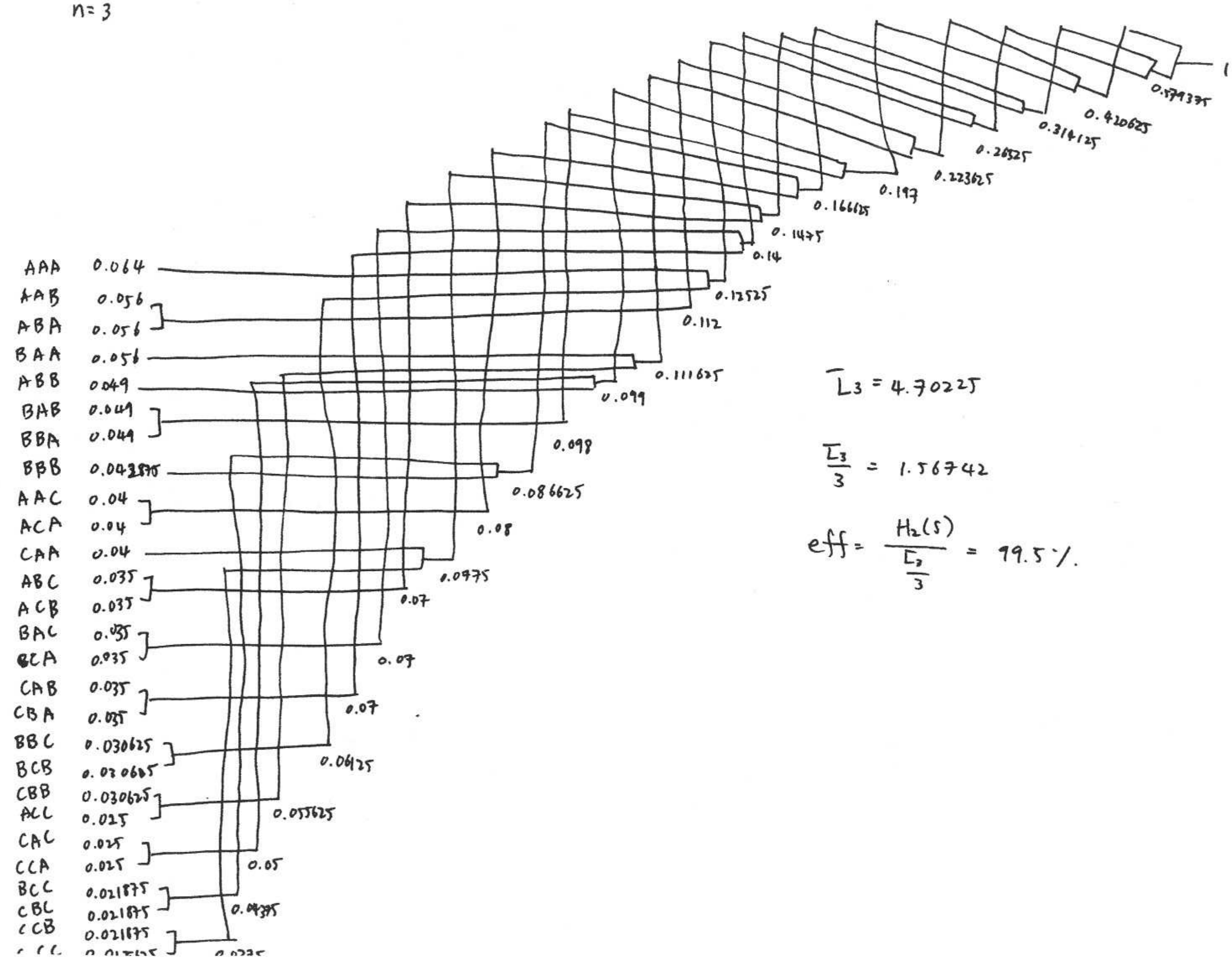


$\bar{L}_2 = 3.15$, $\frac{\bar{L}_2}{2} = 1.575$

$\text{eff} = \frac{1.559}{1.575} = 98.98\%$

[redacted]

$n=3$



$$\bar{L}_3 = 4.70225$$

$$\frac{\bar{L}_3}{3} = 1.56742$$

$$\text{eff} = \frac{H_2(S)}{\frac{\bar{L}_3}{3}} = 99.5\,\%$$

Homework Set #2

1. Construct a source on alphabet $\{A, B, C, D, E\}$ such that the average codeword length of the Shannon–Fano code is strictly larger than that of the Huffman code.

2. Consider a source on alphabet $\{A, B, C, D\}$ with probabilities $p_A \geq p_B \geq p_C \geq p_D$. Find the sufficient and necessary condition that the binary Huffman code is $\{00, 01, 10, 11\}$.

3. In the lecture notes, we defined a *uniquely decodable* code as one such that every distinct concatenation of n codewords is distinct for every n; in other words, if
$$x_1 \cdots x_n \neq y_1 \cdots y_n,$$
then
$$C(x_1) \cdots C(x_n) \neq C(y_1) \cdots C(y_n)$$
for every n. Show that the following more general statement holds for a uniquely decodable code. If
$$x_1 \cdots x_m \neq y_1 \cdots y_n,$$
then
$$C(x_1) \cdots C(x_m) \neq C(y_1) \cdots C(y_n)$$
for every m, n.

4. Consider an i.i.d. source with alphabet $\{A, B, C, D\}$ and probabilities $\{0.1, 0.2, 0.4, 0.3\}$.

   (a) Compute the entropy of the source base 2, 3, and 4.

   (b) Find a binary Huffman code for encoding 2 source symbols at a time. Compute the average number of binary code symbols per source symbol and compare it to the entropy (with the appropriate base).

   (c) Repeat part (b) for ternary codes. Here the code alphabet is $\{0, 1, 2\}$.

   (d) Repeat part (b) for quaternary codes. Here the code alphabet is $\{0, 1, 2, 3\}$.

UCSD ECE154C | Handout #7
Prof. Young-Han Kim | Thursday, April 27, 2017

Solutions to Homework Set #2

1. Construct a source on alphabet $\{A, B, C, D, E\}$ such that the average codeword length of the Shannon–Fano code is strictly larger than that of the Huffman code.

   Solution: Let $l_{\mathrm{SF}}(A)$ be the Shannon–Fano codeword length for symbol A, and let $l_{\mathrm{SF}}(B), \ldots$ be defined similarly. Consider the following probability assignment

   $$(p_A, p_B, p_C, p_D, p_E) = (0.4, 0.15, 0.15, 0.15, 0.15).$$

   It is easy to show that $l_{\mathrm{SF}}(A) = 2$, $l_{\mathrm{SF}}(B) = 2$, $l_{\mathrm{SF}}(C) = 2$, $l_{\mathrm{SF}}(D) = 3$, and $l_{\mathrm{SF}}(E) = 3$. Thus, the average codeword length of the Shannon–Fano code is

   $$(0.4 + 0.15 + 0.15) \times 2 + 0.15 \times 3 \times 2 = 2.3.$$

   For the Huffman code, we have $l_{\mathrm{H}}(A) = 1$, $l_{\mathrm{H}}(B) = 3$, $l_{\mathrm{H}}(C) = 3$, $l_{\mathrm{H}}(D) = 3$, and $l_{\mathrm{H}}(E) = 3$, and the corresponding average codeword length is

   $$0.4 \times 1 + 0.15 \times 3 \times 4 = 2.2.$$

2. Consider a source on alphabet $\{A, B, C, D\}$ with probabilities $p_A \geq p_B \geq p_C \geq p_D$. Find the sufficient and necessary condition that the binary Huffman code is $\{00, 01, 10, 11\}$.

   Solution: The sufficient and necessary condition that the binary Huffman codeword lengths are all 2 is $p_C + p_D > p_A$.

   Let's consider necessary part first. Suppose $p_C + p_D < p_A$ and this implies that in the second iteration of the Huffman code construction, it combines the two least likely symbols p_B and $p_C + p_D$ into one symbol. Thus, the codeword length of A must be 1. If $p_C + p_D = p_A$, in the second iteration of the Huffman algorithm, it is still possible to combine symbols p_B and $p_C + p_D$ into one symbol and the codeword length of A is 1.

   For the sufficient part, if $p_C + p_D > p_A$, following similar arguments above, the codeword length of A is 2.

   We can consider the problem from another angle. Let the codeword length of A be l_A, and let l_B, l_C, and l_D be defined similarly. The only two possible sets of codeword lengths are $(l_A, l_B, l_C, l_D) = (1, 2, 3, 3)$ and $(2, 2, 2, 2)$. Therefore, the equivalent question is when we have $l_A = 2$. We can easily establish the following sufficient and necessary conditions (check!): (a) If $p_A > \frac{2}{5}$, $l_A = 1$. (b) If $p_A < \frac{1}{3}$, $l_A = 2$.

3. In the lecture notes, we defined a *uniquely decodable* code as one such that every distinct concatenation of n codewords is distinct for every n; in other words, if

$$x_1 \cdots x_n \neq y_1 \cdots y_n,$$

then

$$C(x_1) \cdots C(x_n) \neq C(y_1) \cdots C(y_n)$$

for every n. Show that the following more general statement holds for a uniquely decodable code. If

$$x_1 \cdots x_m \neq y_1 \cdots y_n,$$

then

$$C(x_1) \cdots C(x_m) \neq C(y_1) \cdots C(y_n)$$

for every m, n.

Solution: We provide a proof by contradiction. Suppose there exist $x^m := x_1 \cdots x_m$ and $y^n := y_1 \cdots y_n$ such that $x^m \neq y^n$ yet $C(x^m) := C(x_1) \cdots C(x_m) = C(y_1) \cdots C(y_n) =: C(y^n)$. By the unique decodability, it must be that neither x^m nor y^n is a prefix of the other (why?).

Now consider two *distinct* sequences $x^m y^n$ and $y^n x^m$ of length $m+n$. The corresponding codewords are the same $C(x^m)C(y^n) = C(y^n)C(x^m)$, which violates the assumption that the code is uniquely decodable. Thus, we have the desired contradiction, and $C(x^m)$ and $C(y^n)$ must be distinct.

4. Consider an i.i.d. source with alphabet $\{A, B, C, D\}$ and probabilities $\{0.1, 0.2, 0.4, 0.3\}$.

   (a) Compute the entropy of the source base 2, 3, and 4.

   (b) Find a binary Huffman code for encoding 2 source symbols at a time. Compute the average number of binary code symbols per source symbol and compare it to the entropy (with the appropriate base).

   (c) Repeat part (b) for ternary codes. Here the code alphabet is $\{0, 1, 2\}$.

   (d) Repeat part (b) for quaternary codes. Here the code alphabet is $\{0, 1, 2, 3\}$.

   Solution:

   (a) We can calculate the entropy of the source with probabilities $\{0.1, 0.2, 0.4, 0.3\}$

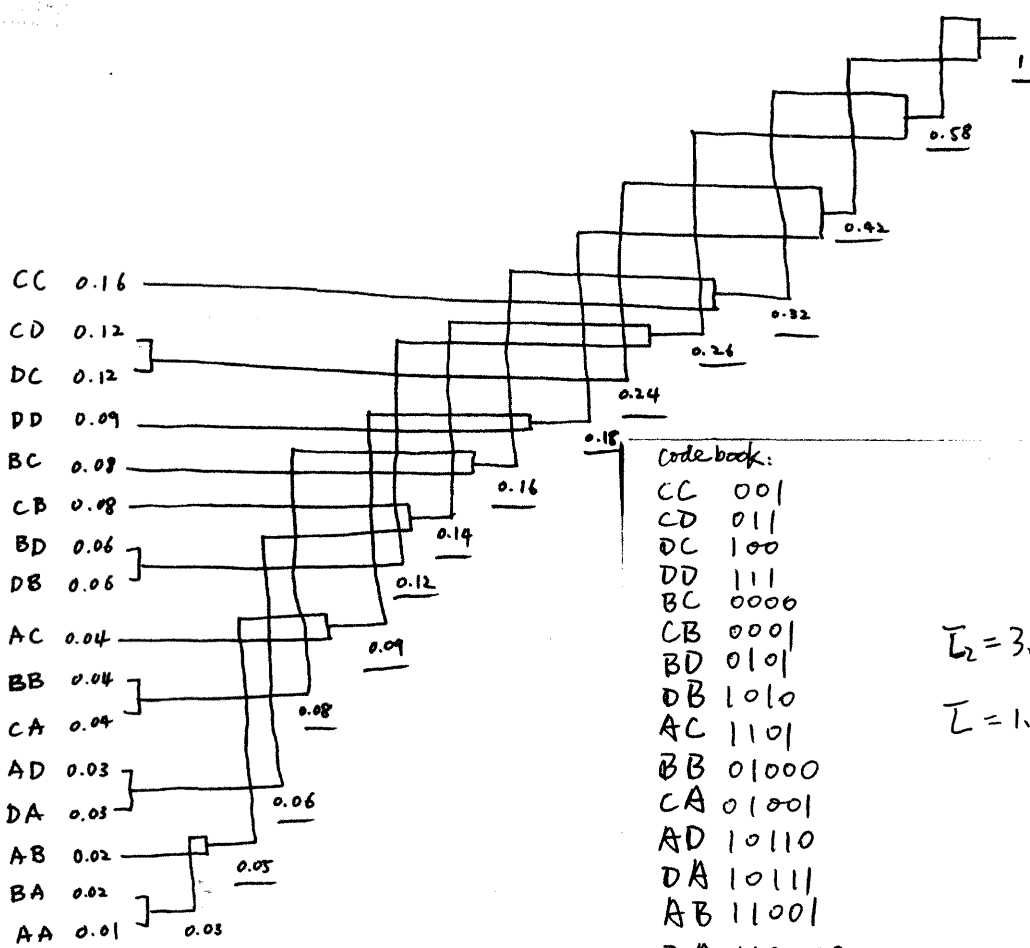
with appropriate bases.

$$H_2(S) = \sum_{i=1}^{4} p_i \log_2 \frac{1}{p_i} = 1.846.$$

$$H_3(S) = \sum_{i=1}^{4} p_i \log_3 \frac{1}{p_i} = 1.165.$$

$$H_4(S) = \sum_{i=1}^{4} p_i \log_4 \frac{1}{p_i} = 0.923.$$

(b)



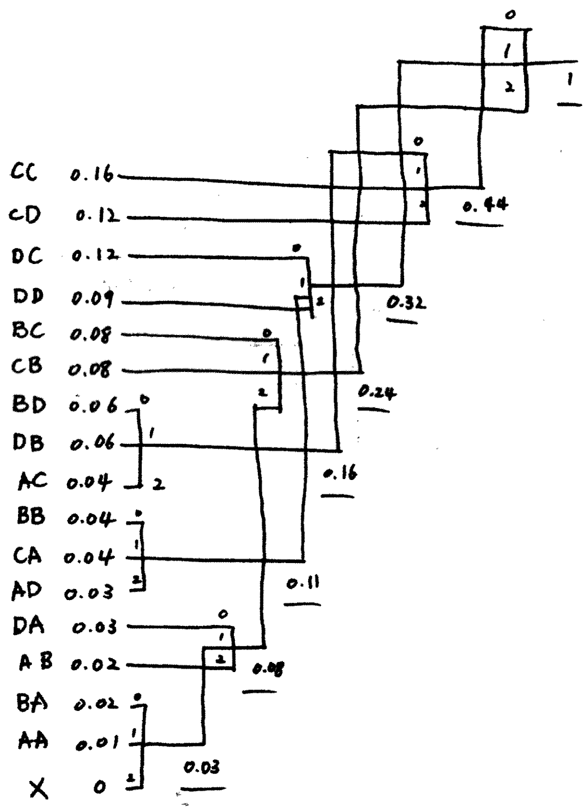
Code book:

CC 001
CD 011
DC 100
DD 111
BC 0000
CB 0001
BD 0101
DB 1010
AC 1101
BB 01000
CA 01001
AD 10110
DA 10111
AB 11001
BA 110000
AA 110001.

$L_2 = 3.73.$

$L = 1.865.$

(c)



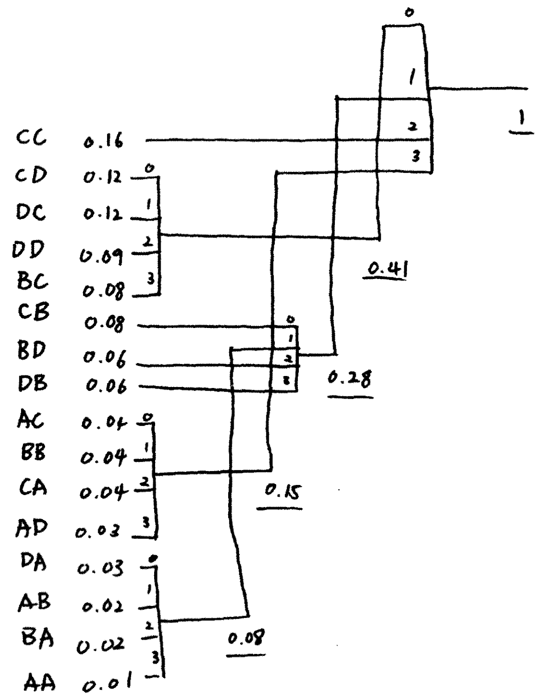
codebook:

CC	01
CD	02
DC	10
DD	12
BC	20
CB	21
BD	000
DB	001
AC	002
BB	110
CA	111
AD	112
DA	221
AB	222
BA	2200
AA	2202.

$\bar{L}_2 = 2.38$

$\bar{L} = 1.19.$

(d)



Codebook:

CC	2
CD	00
DC	01
DD	20
BC	10
CB	11
BD	12
DB	13
AC	30
BB	31
CA	32
AD	33
DA	030
AB	031
BA	032
AA	033.

$\bar{L}_2 = 1.92.$

$\bar{L} = 0.96.$