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 \(l_{\text{SF}}(A)\) be the Shannon–Fano codeword length for symbol \(A\), and let \(l_{\text{SF}}(B), \ldots\) 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 \(l_{\text{SF}}(A) = 2\), \(l_{\text{SF}}(B) = 2\), \(l_{\text{SF}}(C) = 2\), \(l_{\text{SF}}(D) = 3\), and \(l_{\text{SF}}(E) = 3\). 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 \(l_{\text{H}}(A) = 1\), \(l_{\text{H}}(B) = 3\), \(l_{\text{H}}(C) = 3\), \(l_{\text{H}}(D) = 3\), and \(l_{\text{H}}(E) = 3\), 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.

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

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

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