Homework Set #3
Data Compression and Huffman code

1. **Huffman coding.**
   Consider the random variable

$$X = \begin{pmatrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 \\ 0.50 & 0.26 & 0.11 & 0.04 & 0.04 & 0.03 & 0.02 \end{pmatrix}$$

   (a) Find a binary Huffman code for X.

   (b) Find the expected codelength for this encoding.

   (c) Extend the Binary Huffman method to Ternarry (Alphabet of 3) and apply it for X.

2. **Codes.**
   Let $X_1, X_2, \ldots,$ i.i.d. with

$$X = \begin{cases} 1, & \text{with probability } 1/2 \\ 2, & \text{with probability } 1/4 \\ 3, & \text{with probability } 1/4. \end{cases}$$

   Consider the code assignment

$$C(x) = \begin{cases} 0, & \text{if } x = 1 \\ 01, & \text{if } x = 2 \\ 11, & \text{if } x = 3. \end{cases}$$

   (a) Is this code nonsingular?

   (b) Uniquely decodable?

   (c) Instantaneous?

   (d) Entropy Rate is defined as

$$H(\mathcal{X}) \triangleq \lim_{n\to\infty} \frac{H(X^n)}{n}. \tag{1}$$

   What is the entropy rate of the process

$$Z_1 Z_2 Z_3 \ldots = C(X_1)C(X_2)C(X_3)\ldots?$$

3. **Compression**

   (a) Give a Huffman encoding into an alphabet of size D = 2 of the following probability mass function:

$$\left(\frac{1}{2}, \frac{1}{8}, \frac{1}{8}, \frac{1}{16}, \frac{1}{16}, \frac{1}{16}, \frac{1}{16}\right)$$

   (b) Assume you have a file of size 1,000 symbols where the symbols are distributed i.i.d. according to the pmf above. After applying the Huffman code, what would be the pmf of the compressed binary file (namely, what is the probability of '0' and '1' in the compressed file), and what would be the expected length?

4. **Entropy and source coding of a source with infinite alphabet**
   Let X be an i.i.d. random variable with an infinite alphabet, $\mathcal{X} = \{1, 2, 3, ...\}$. In addition let $P(X = i) = 2^{-i}$.

   (a) What is the entropy of the random variable?

   (b) Find an optimal variable length code, and show that it is indeed optimal.

5. **Bad wine.**
   One is given 6 bottles of wine. It is known that precisely one bottle has gone bad (tastes terrible). From inspection of the bottles it is determined that the probability p_i that the i^{th} bottle is bad is given by $(p_1, p_2, \ldots, p_6) = (\frac{7}{26}, \frac{5}{26}, \frac{4}{26}, \frac{4}{26}, \frac{3}{26}, \frac{3}{26})$. Tasting will determine the bad wine.

   Suppose you taste the wines one at a time. Choose the order of tasting to minimize the expected number of tastings required to determine the bad bottle. Remember, if the first 5 wines pass the test you don't have to taste the last.

   (a) What is the expected number of tastings required?

   (b) Which bottle should be tasted first?

   Now you get smart. For the first sample, you mix some of the wines in a fresh glass and sample the mixture. You proceed, mixing and tasting, stopping when the bad bottle has been determined.

(c) What is the minimum expected number of tastings required to determine the bad wine?

(d) What mixture should be tasted first?

6. **Relative entropy is cost of miscoding.**
Let the random variable X have five possible outcomes $\{1, 2, 3, 4, 5\}$. Consider two distributions on this random variable

Symbol	$p(x)$	$q(x)$	$C_1(x)$	$C_2(x)$
1	1/2	1/2	0	0
2	1/4	1/8	10	100
3	1/8	1/8	110	101
4	1/16	1/8	1110	110
5	1/16	1/8	1111	111

(a) Calculate $H(p)$, $H(q)$, $D(p||q)$ and $D(q||p)$.

(b) The last two columns above represent codes for the random variable. Verify that the average length of C_1 under p is equal to the entropy $H(p)$. Thus C_1 is optimal for p. Verify that C_2 is optimal for q.

(c) Now assume that we use code C_2 when the distribution is p. What is the average length of the codewords. By how much does it exceed the entropy $H(p)$?

(d) What is the loss if we use code C_1 when the distribution is q?

7. **Shannon code.** Consider the following method for generating a code for a random variable X which takes on m values $\{1, 2, \ldots, m\}$ with probabilities $p_1, p_2, \ldots, p_m$. Assume that the probabilities are ordered so that $p_1 \geq p_2 \geq \cdots \geq p_m$. Define

$$F_i = \sum_{k=1}^{i-1} p_i, \tag{2}$$

the sum of the probabilities of all symbols less than i. Then the codeword for i is the number $F_i \in [0, 1]$ rounded off to l_i bits, where $l_i = \lceil \log \frac{1}{p_i} \rceil$.

(a) Show that the code constructed by this process is prefix-free and the average length satisfies

$$H(X) \leq L < H(X) + 1. \tag{3}$$

(b) Construct the code for the probability distribution $(0.5, 0.25, 0.125, 0.125)$.