Homework Set #2
Data Compression, Huffman code and AEP

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 $X_1, X_2, \dots,$ 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

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

3. **Huffman via MATLAB or Python**

(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?

(c) Generate a sequence (using MATLAB/Python or any other software) of 10,000 symbols of X with i.i.d. probability P_X. Assume the alphabet of X is $\mathcal{X} = (0, 1, ..., 6)$.

(d) What is the percentage of each symbol $(0, 1, ..., 6)$ in the sequence that was generated. Explain this result using the law of large numbers.

(e) Represent each symbol in $\mathcal{X}$ using a simple binary representation. Namely, $X = 0$ represent as '000', $X = 1$ represent as '001',$X = 2$ represent as '010',..., $X = 6$ represent as '110'.

(f) What is the length of the simple representation. What percentage of '0' and '1' do you have in this representation?

(g) Now, compress the 10,000 symbols of X, into bits using Huffman code.

(h) What is the length of the compressed file. What percentage of '0' and '1' do you have in this representation?

(i) Explain the results.

Note: You may use the MATLAB functions randsample and strfind.
In Python respectively, you may use numpy.random.randint and String find() (notice that String find() is different than strfind).

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 $(0.5, 0.25, 0.125, 0.125)$.

8. **An AEP-like limit.** Let $X_1, X_2, \ldots$ be i.i.d. drawn according to probability mass function $p(x)$. Find

$$\lim_{n \to \infty} [p(X_1, X_2, \ldots, X_n)]^{\frac{1}{n}}.$$

9. **AEP**. Let $X_1, X_2, \ldots$ be independent identically distributed random variables drawn according to the probability mass function $p(x), x \in \{1, 2, \ldots, m\}$. Thus $p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^n p(x_i)$. We know that $-\frac{1}{n} \log p(X_1, X_2, \ldots, X_n) \to H(X)$ in probability. Let $q(x_1, x_2, \ldots, x_n) = \prod_{i=1}^n q(x_i)$, where q is another probability mass function on $\{1, 2, \ldots, m\}$.

(a) Evaluate $\lim -\frac{1}{n} \log q(X_1, X_2, \ldots, X_n)$, where $X_1, X_2, \ldots$ are i.i.d. $\sim p(x)$.

(b) Now evaluate the limit of the log likelihood ratio $\frac{1}{n} \log \frac{q(X_1,\ldots,X_n)}{p(X_1,\ldots,X_n)}$ when $X_1, X_2, \ldots$ are i.i.d. $\sim p(x)$. Thus the odds favouring q are exponentially small when p is true.

10. **Empirical distribution of a sequence** Before starting the question, below are two facts that you may consider to use:

- Stirling approximation: $n! \approx \sqrt{2\pi n}(\frac{n}{e})^n$.
- Consider a sequence of length n that consist of two different numbers. The first number appears n_1 times and the second number appears n_2 times such that $n_1 + n_2 = n$. The number of different combinations of such sequences is given by $\binom{n}{n_1 \; n_2} = \frac{n!}{n1!n2!}$.

A fair dice with 6 faces was thrown n times, where n is a very large number.

(a) Find how many different sequences there exists with an empirical pmf $(p_1, p_2, ..., p_6)$, where p_i is the portion of the sequence that is equal to $i \in \{1, 2, ..., 6\}$.
In this section you can assume that $n! \approx (\frac{n}{e})^n$ since only the power of $\frac{n}{e}$ will matter.

(b) Now, we were told that the portion of odd numbers in the sequence is 2/3 (i.e., $p_1 + p_3 + p_5 = 2/3$). For n very large, what is the most likely empirical pmf of the sequence. **Hint:** Define:

$$X = \begin{cases} 1 & p_1 \\ 3 & p_3 \\ 5 & p_5 \end{cases}, \; Y = \begin{cases} 2 & p_2 \\ 4 & p_4 \\ 6 & p_6 \end{cases}, \; Z = \begin{cases} X & \frac{2}{3} \\ Y & \frac{1}{3} \end{cases}.$$

Think why maximizing $H(Z)$ means maximizing $H(X)$, $H(Y)$.

(c) What is the cardinality of the weak typical set with respect to the pmfs that you found/given in the previous subquestions, i.e., (a) and (b)?

Remark 1 *The weak typical set is the typical set we learned in the AEP lecture.*

11. **drawing a codebook** Let X_i be a r.v. i.i.d distributed according to $P(x)$. We draw codebook of 2^{nR} codewords of X^n independently using $P(x)$ and i.i.d.. We would like to answer the question: what is the probability that the first codeword would be identical to another codeword in the codebook as n goes to infinity.

- Let x^n be a sequence in the typical set $A_\epsilon^n(X)$. What is the asymptotic probability (you may provide an upper and lower bound) as $n \to \infty$ that we draw a sequence X^n i.i.d distributed according to $P(x)$ and we get x^n.
- Using your answer from the previous sub-question find an $\overline{\alpha}$ such that if $R < \overline{\alpha}$ the probability that the first codewaord in the codebook appears twice or more in the codebook goes to zero as $n \to \infty$.
- Find an $\underline{\alpha}$ such that if $R > \underline{\alpha}$ the probability that the first codewaord in the codebook appears twice or more in the codebook goes to 1 as $n \to \infty$. (Hint: you may use Bernoulli's inequality $(1+x)^r \le e^{rx}$ for all real numbers $r \ge 0, x \ge -1$)

12. **Saving the princess** A princess was abducted and was put in one of K rooms. Each room is labeled by a number 1,2,...,K. Each room is of a size s_i where $i = 1, 2, ..., k$. The probability of the princess to be in room i, p_i, is proportional to the size of the room, namely $p_i = \alpha s_i$ where α is a constant.

(a) Find α

(b) In order to save the princess you need to find in which room she is. You may ask the demon an yes/no question. Like is she in room number 1 or is she in room 2 or 5 or is she in a room of odd number, and so on. You will save the princess if only if the expected number of questions is the minimum possible. What would be the questions you should ask the demon to save the princess?

13. **Lossless source coding with side information.**
Consider the lossless source coding with side information that is available at the encoder and decoder, where the source X and the side information Y are i.i.d. $\sim P_{X,Y}(x,y)$.

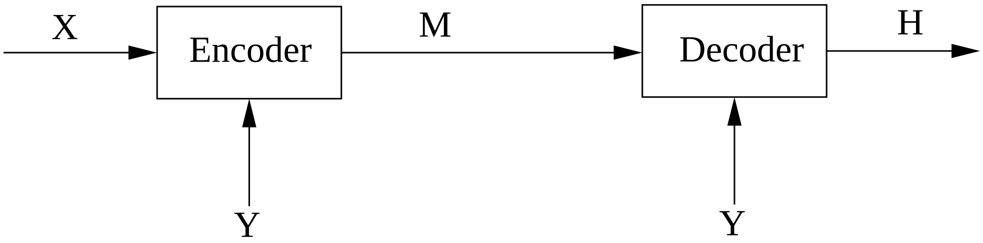

Figure 1: Lossless source coding with side information at the encoder and decoder.

Show that a code with rate $R < H(X|Y)$ can not be achievable, and interpret the result.

Hint: Let $T \triangleq f(X^n, Y^n)$. Consider

$$\begin{aligned} nR &\geq H(T) \\ &\geq H(T|Y^n), \end{aligned} \tag{4}$$

and use similar steps, including Fano's inequality, as we used in the class to prove the converse where side information was not available.

14. **Challenge: Optimal code for an infinite alphabet**
This is a bonus question! Solving this question would add **10 points** to the final grade of the course.

Let X be a r.v. with a discrete and infinite alphabet, in particular assume that $\mathcal{X} = 1, 2, 3, ...$ The pmf is given by an infinite vector $[p_1, p_2, p_3, ...]$ where $p_i \geq p_j$ if $i > j$. Find an optimal prefix code for X.

15. **Conditional Information Divergence**

(a) Let X, Z be random variables jointly distributed according to $P_{X,Z}$. We define the conditional informational divergence as follows:

$$D\big(P_{X|Z}\big|\big|Q_{X|Z}\big|P_Z\big) = \sum_{(x,z)\in\mathcal{X}\times\mathcal{Z}} P_{X,Z}(x,z)\log\left(\frac{P_{X|Z}(x|z)}{Q_{X|Z}(x|z)}\right).$$

With respect to this definition, prove for each relation if it is **true** or **false**:

For any pair of random variables A, B that are jointly distributed according to $P_{A,B}$, i.

$$D\big(P_{A,B}\big|\big|Q_{A,B}\big) = D\big(P_A\big|\big|Q_A\big) + D\big(P_{B|A}\big|\big|Q_{B|A}\big|P_A\big).$$

ii.

$$D\big(P_{A,B}\big|\big|P_A P_B\big) = D\big(P_{B|A}\big|\big|P_B\big|P_A\big).$$

iii.

$$I(A;B) = D\big(P_{B|A}\big|\big|P_B\big|P_A\big).$$

iv.

$$D\big(P_{A|B}\big|\big|Q_{A|B}\big|P_B\big) = \sum_{b\in\mathcal{B}} P_B(b) D\big(P_{A|B=b}\big|\big|Q_{A|B=b}\big).$$

(b) Consider the setting in Fig. 2.

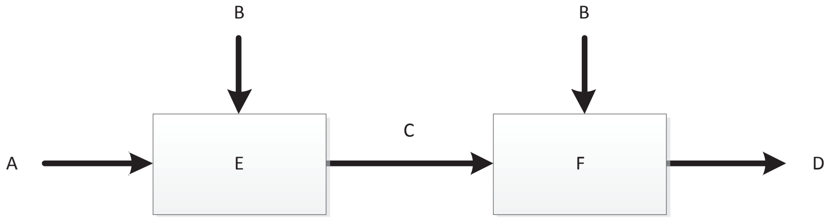

Figure 2: Source coding with side information.

We would like to compress the source sequence X^n losslessly using a prefix code with side information Z^n which is available to the encoder and the decoder. The sources (X^n, Z^n) are distributed i.i.d. according to $P_{X,Z}$ and that all the distribution and conditional distributions are dyadic (i.e., P_X is dyadic if $P_X(x) = 2^{-i}$, for some i, for all $x \in \mathcal{X}$). We denote the average number of bits per symbol needed to compress the source X^n as L.

i. What is the minimal L?

ii. Although the distribution of (X^n, Z^n) is $P_{X,Z}$, the distribution that is used design the optimal prefix code is $Q_{X|Z}P_Z$. What is the actual L (average bits per symbol) of this code?

iii. Now, the distribution that is used to design the prefix code is $Q_{X,Z}$. What is the actual L now?

16. **True or False of a constrained inequality**:

Given are three discrete random variables X, Y, Z that satisfy $H(Y|X,Z) = 0$.

(a) Copy the next relation and write **true** or **false** (If true, prove the statement, and if not provide a counterexample).

$$I(X;Y) \geq H(Y) - H(Z)$$

(b) What are the conditions for which the equality $I(X;Y) = H(Y) - H(Z)$ holds.

(c) Assume that the conditions for $I(X;Y) = H(Y) - H(Z)$ are satisfied. There exists a function such that $Z = g(Y)$.

(d) Assume that the conditions for $I(X;Y) = H(Y) - H(Z)$ are satisfied. It is *always* true that $Z = g(Y)$.

17. **True or False**: Copy each relation and write **true** or **false**.

(a) Let $X - Y - Z - W$ be a Markov chain, then the following holds:

$$I(X;W) \leq I(Y;Z).$$

(b) For two probability distributions, p_{XY} and q_{XY}, that are defined on $\mathcal{X} \times \mathcal{Y}$, the following holds:

$$D(p_{XY}||q_{XY}) \geq D(p_X||q_X).$$

(c) If X and Y are dependent and also Y and Z are dependent, then X and Z are dependent.

18. **Huffman Code** : Let X^n be a an i.i.d. source that is distributed according to p_X:

x	0	1	2	3
$p_X(x)$	0.5	0.25	0.125	0.125

(a) Find $H(X)$.

(b) Build a binary Huffman code for the source X.

(c) What is the expected length of the resulting compressed sequence.

(d) What is the expected number of zeros in the resulting compressed sequence.

(e) Let $\tilde{X}^n$ be an another source distributed i.i.d. according to $p_{\tilde{X}}$.

$\tilde{x}$	0	1	2	3
$p_{\tilde{X}}(\tilde{x})$	0.3	0.4	0.1	0.2

What is the expected length of compressing the source $\tilde{X}$ using the code constructed in (b).

(f) Answer (d) for the code constructed in (b) and the source $\tilde{X}^n$.

(g) Is the relative portion of zeros (the quantity in (d) divided by the quantity in (c)) after compressing the source X^n and the source $\tilde{X}^n$ different? For both sources, explain why there is or there is not a difference.