EE/Stats 376A - Information Theory
Sample Problems for Midterm

Here are some problems you can work on to practice for the midterm. The total number of questions here is NOT indicative of the length of the actual midterm.

1. X, Y, Z are three discrete-valued random variables taking values in alphabets $\mathcal{X}$, $\mathcal{Y}$,$\mathcal{Z}$ respectively. Are these statements necessarily True?

(a) $H(X, Y) \geq H(X)$.

(b) $H(f(X)|g(X)) = 0$ for any functions f, g defined on $\mathcal{X}$.

(c) $I(X; Y, Z) = I(X; Y) + I(X; Y|Z)$.

(d) $H(X|Y = y) \leq H(X) \qquad \forall y \in \mathcal{Y}$.

(e) $I(X; Z|Y = y) \geq 0 \qquad \forall y \in \mathcal{Y}$.

(f) If $p(y|x)$ is a DMC with input in $\mathcal{X}$ and output in $\mathcal{Y}$ and capacity C_1, and $q(z|y)$ is a DMC with input in $\mathcal{Y}$ and output in $\mathcal{Z}$ and capacity C_2, then the DMC:

$$r(z|x) := \sum_{y \in \mathcal{Y}} q(z|y)p(y|x)$$

with input in $\mathcal{X}$ and output in $\mathcal{Z}$ *can* have capacity greater than $\max(C_1, C_2)$.

2. Let $X_1, X_2, \ldots$ be an i.i.d. sequence of binary random variables, each equally likely to be 0 or 1.

(a) Define:

$$Y_n := |\{j : 1 \leq j \leq n, X_j = 1\}|, \qquad n = 1, 2, \ldots$$

Compute the entropy rate of the process $\{Y_n\}$. (Notation: $|\S|$ is the number of elements in the set $\S$.)

(b) Define:

$$Z_n = \begin{cases} 0 & \text{if } n = 1 \\ |X_n - X_{n-1}| & \text{if } n > 1 \end{cases}$$

Compute the entropy rate of the process $\{Z_n\}$. (Notation: $|x|$ is the absolute value of a number x.)

3. Suppose X is a random variable which takes on m values $\{1, 2, \ldots, m\}$ with probabilities $p_1, \ldots, p_m$. The *Shannon code* is a prefix code that assigns binary codewords of length $l_i = \lceil \log_2 1/p_i \rceil$ to letter i. You have analyzed this code in one of your homeworks, but we need not worry about how the code is constructed here.

(a) What does "prefix code" mean?

(b) Argue why a prefix code with these codeword lengths always exists.

(c) Argue that the Huffman code cannot beat the Shannon code by more than 1 bit in expected codeword length.

(d) Suppose now the Shannon code is used to compress an i.i.d. binary source $\{Y_n\}$ (with probability of $Y_n = 1$ being p) by coding over blocks of length k. Can the expected codeword length per source symbol using the Shannon code approach the entropy rate of the source as the block length $k \to \infty$? Explain.

(e) Suppose the Shannon code in part (d) is designed incorrectly thinking that the probability of $Y_n = 1$ is q (although the true probability is p). Quantify the loss in performance compared to the entropy rate of the source, as the block length $k \to \infty$.

4. Kedar, Mikel and Naroa have been instructed to record the outcomes of a coin toss experiment. Consider the coin toss experiment $X_1, X_2, X_3, ...$ where X_i are i.i.d. $Bern(p)$ (probability of a H (head) is p), $p = 15/16$.

(a) Kedar decides to use Huffman coding to represent the outcome of each coin toss separately. What is the resulting scheme? What compression rate does it achieve?

(b) Mikel suggests he can do a better job by applying Huffman coding on blocks of r tosses. Will his scheme approach the optimum expected number of bits per description of source symbol (coin toss outcome) with increasing r? How does the space required to represent the codebook increase as we increase r?

(c) Naroa suggests that, as the occurrence of T is so rare, we should just record the number of tosses it takes for each T to occur.
To be precise, if Y_k represents the number of trials until the k^{th} T occurred (inclusive), then Naroa records the sequence:

$$Z_k = Y_k - Y_{k-1}, \ k \geq 1, \tag{1}$$

where $Y_0 = 0$

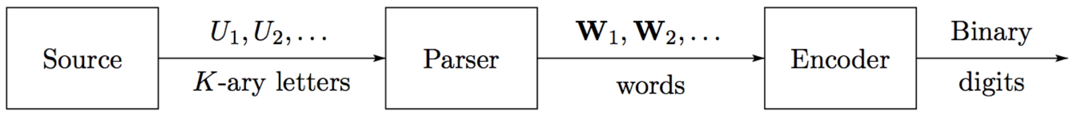
i. What is the distribution of Z_k, i.e., $P(Z_k = j), j \in 1, 2, 3, ...$?

ii. Compute the entropy and expectation of Z_k.

iii. How does the ratio between the entropy and the expectation of Z_k compare to the entropy of X_k? Give an operational interpretation.

5. Tunstall Codes

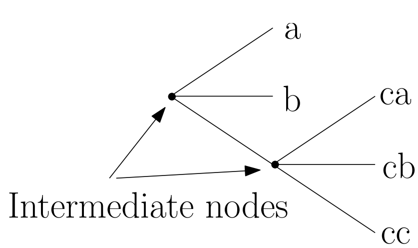
Consider an approach for compression of a memoryless K-ary source $U_1, U_2, \ldots$ distributed according to the discrete random variable U, with $|\mathcal{U}| = K$, illustrated in the following figure:



Instead of encoding the source symbols directly, we first parse them into words with the help of a dictionary of size M. A dictionary is a prefix-free collection of words $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_M$, where each $\boldsymbol{w}_m$ is a sequence of source letters, with the property that any stream of source symbols can be constructed as a concatenation of words from the dictionary. Each dictionary entry is then represented by a binary codeword of the same constant length b. These binary representations of the dictionary words that make up the source stream comprise the encoder output.

Given a dictionary with words $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_M$, the parser picks the longest prefix of the source sequence $U_1, U_2, \ldots$ that is in the dictionary (say, $U_1, \ldots, U_L$), and then continues in the same fashion with $U_{L+1}, U_{L+2}, \ldots$, etc.

Consider, for example, $\mathcal{U} = \{a, b, c\}$, and $\boldsymbol{w}_1 =$"a", $\boldsymbol{w}_2 =$"b", $\boldsymbol{w}_3 =$"ca", $\boldsymbol{w}_4 =$"cb", $\boldsymbol{w}_5 =$"cc". The tree associated with this dictionary is given by



(a) Given a dictionary represented by binary codewords of fixed length b, what should the words of the dictionary satisfy such that the overall compression scheme is optimal (i.e., it achieves the entropy)? Specifically, what should M satisfy, and what should be the probabilities of the words that comprise the dictionary?

Consider the following Tunstall code, for $\mathcal{U} = \{a, b, c\}$, and $P_U(a) = 0.6$, $P_U(b) = 0.3$, and $P_U(c) = 0.1$:

Word	Codeword
b	000
c	001
ab	010
ac	011
aaa	100
aab	101
aac	110

(b) Compute the entropy of the source $H(U)$.

(c) Compute the entropy of the dictionary words, denoted by $H(W)$.

(d) Compute the expected length of the dictionary words, denoted by $E[L_W]$.

(e) What is the overall compression rate, i.e., the average number of bits of description per source symbol under this compression scheme? Compare this with the compression achieved if the dictionary words were given by “a”, “’b” and “c”, and the associated binary codewords were of length 2. Which scheme is better?

(f) Given the dictionary in the above table, can you come up with a better assignment of codewords (not necessarily of the same length) for the same dictionary? The code should be a prefix code. Please construct the best such code. What is the overall compression achieved, i.e., the expected codeword length per source symbol in this case? How does it compare to that of the Tunstall code above? How does it compare to the entropy of the source? Explain.

6. The set $A_\epsilon^{(n)}$ of jointly typical n-length sequences with respect to the distribution $p(x,y)$ is defined by:

$$A_\epsilon^{(n)} := \{(x^n, y^n) : \quad \begin{aligned} &|-\tfrac{1}{n}\log \Pi_i p(x_i) - H(X)| < \epsilon, \\ &|-\tfrac{1}{n}\log \Pi_i p(y_i) - H(Y)| < \epsilon, \\ &|-\tfrac{1}{n}\log \Pi_i p(x_i, y_i) - H(X,Y)| < \epsilon\} \end{aligned}$$

(a) Let (X^n, Y^n) be a pair of random sequences length n drawn according to

$$\Pi_{i=1}^n p(x_i, y_i).$$

Argue that $\Pr((X^n, Y^n) \in A_\epsilon^{(n)}) \to 1$ as $n \to \infty$.

(b) Argue that

$$H(X,Y) - \epsilon \leq \lim_{n\to\infty} \frac{1}{n} \log |A_\epsilon^{(n)}| \leq H(X,Y) + \epsilon.$$

(You can assume the limit exists.)

7. Modulo Channel

(a) Consider the DMC defined as follows: Output $Y = X \oplus_2 Z$ where X, taking values in $\{0, 1\}$, is the channel input, $\oplus_2$ is the modulo-2 summation operation, and Z is binary channel noise uniformly distributed over $\{0, 1\}$ and independent of X. What is the capacity of this channel?

(b) Consider the channel of the previous part, but suppose that instead of modulo-2 addition $Y = X \oplus_2 Z$, we perform modulo-3 addition $Y = X \oplus_3 Z$. Now what is the capacity?

(c) Now suppose the noise Z is no longer independent of the input X, but is instead described by the following conditional distribution:

$$p(Z = z|X = 0) = \begin{cases} 1/4 \text{ if } z = 0 \\ 3/4 \text{ if } z = 1, \end{cases}$$

and

$$p(Z = z|X = 1) = 1/2 \quad \text{both for } z = 0 \text{ and } z = 1.$$

What is the capacity?