COMPSCI 650: Applied Information Theory
Exam II: Apr. 28, 2016

Instructor: Arya Mazumdar

ADD THIS PAGE TO YOUR SOLUTION. TAKE HOME EXAM. OPEN BOOK, OPEN NOTES. CALCULATORS OK. DISCUSSIONS WITH ANY LIVING BEING NOT ALLOWED. PLEASE BE RIGOROUS AND PRECISE AND SHOW YOUR WORK. THIS EXAM CONSISTS OF FIVE PROBLEMS THAT CARRY A TOTAL OF 100 POINTS, AS MARKED (PERFECT SCORE = 100). RETURN EXAM BY: 11 AM MONDAY MAY 2 AT INSTRUCTOR MAILBOX LOCATED AT CS MAIN OFFICE. USE EXTRA PAGES IF NEEDED. GOOD LUCK!

NAME (LAST, FIRST):
Student Id Number:

	Points Available	Points Achieved
Problem 1 :	30	
Problem 2 :	25	
Problem 3 :	15	
Problem 4 :	15	
Problem 5 :	15	
Totals :	100	

Problem 1: ($6 \times 5 = 30$ points Clustering) This might be a hard problem to some. But you should not be deterred by that. This problem is here to enhance your ability to understand a research scenario.

Consider a clustering problem where k clusters are present. At some stage of the clustering algorithm we need to assign an element v to one of the k clusters. Each cluster at this stage has m elements (that is, a total of $k \times m$ elements are already clustered and equally distributed among the k clusters).

Suppose u is an element that is already clustered (one of the km elements). One is allowed to make a query of the following form: "Are u and v in the same cluster?" The 'yes/no' answer to such query is incorrect with probability p. We can make many such queries before assigning v to one of the k clusters.

In this problem we try to answer the following question: What is the necessary number of queries one should make so that with probability at least 0.9, v is assigned to the correct cluster?

Given the answers to the queries, this is simply a hypothesis testing problem with k possible hypotheses.

Consider the case $k = 2$. Hence there are only two clusters, cluster 1 and cluster 2. Here Hypothesis 1 is "v belongs to cluster 1." Here Hypothesis 2 is "v belongs to cluster 2."

1) Let a total t queries have been asked involving v. The answers to these queries are random variables $X_1, X_2, \ldots, X_t$. The probability distribution of $X_1, X_2, \ldots, X_t$ can be two different possibilities: P^t if Hypothesis 1 is true, and Q^t if Hypothesis 2 is true.
   Find out $D(P^t \| Q^t)$. (Hint: Think about the two cases, when v in cluster 1 and when v in cluster 2).
2) Write down Le Cam's identity for probability of error for binary hypothesis testing for this case.
3) Use Pinsker's inequality to get a lower bound on probability of error in terms of $D(P^t \| Q^t)$.
4) Use the value of $D(P^t \| Q^t)$ found in the first part, to derive a lower bound on probability of error in terms of p and t.
5) If p is 0.2 find t that guarantees 90% accuracy of assignment. This many queries must be made for clustering each elements.
6) Can you guess then how many queries one should make for 90% accuracy of assignment is there are k clusters?

Problem 2: (5+10+10 points Expander codes and erasures) Consider the parity check matrix H, of dimension $n - k \times n$, of a binary linear error-correcting code. Let the number of nonzero elements in each row of H is d and the number of nonzero elements on each column of H is c.

Further consider the factor graph (the bipartite graph) for this parity check matrix. Let the set of variable nodes is V_1 and the set of check nodes is V_2. Let this be an expander graph with the following property: any set $S \subset V_1$ such that $|S| \leq \alpha n$, must have total number of neighbors at least $\frac{3c}{4}|S|$.

1) What is the rate of the code k/n in terms of c and d?
2) Suppose we want to use this code over a binary erasure channel (i.e., correct erasures). Write down an iterative erasure correction algorithm (Hint: Recall what has been done in class).
3) Show how to correct any αn erasures with your iterative algorithm.

Problem 3: (5+10 points Compressed Sensing) Consider the following sampling matrix Φ:

$$\begin{bmatrix} 1 & 0 & 0.5 \\ -2 & 1 & 1 \end{bmatrix}$$

We observe samples of an 1-sparse vector x:

$$y = \Phi x = \begin{bmatrix} 0 \\ 5 \end{bmatrix}.$$

Find out x.

Next, suppose we observe:

$$z = \Phi x = \begin{bmatrix} 10.1 \\ -20.1 \end{bmatrix}.$$

For a general vector x. Find out x such that x is approximately sparse (use intuition).

Problem 4: (3+2+5+5 points Data Compression) Suppose, $\mathcal{X} = \{A, B, C, D\}$. A source produces i.i.d. $\sim X$ symbols from this source, with $\Pr(X = A) = p_A, \Pr(X = B) = p_B, \Pr(X = C) = p_C, \Pr(X = D) = p_D$. You are given a file $\mathcal{F} = AACDDBBBBBCAABCDAABAADCB$ generated by this source.

1) What is your best guess for p_A, p_B, p_C, p_D? Reason.
2) What is the entropy (in bits) of the probability distribution you guessed?
3) What is the Huffman code for the probability distribution that you guessed? What is the average number of bits per symbol?
4) Encode the file $\mathcal{F}$ with the Huffman code you have designed. What is the length of the encoded binary file? What is the average number of bits that have been used for a symbol in this file?

Problem 5: (7+8 points Channel Capacity)

1) Consider the following binary-input binary-output channel called the Z-channel. The input-output transition probabilities are given by:

$$p(y|x) = \begin{cases} 1, & x = y = 0 \\ p, & x = 1, y = 0 \\ 1 - p, & x = 1, y = 1 \\ 0, & \text{otherwise.} \end{cases}$$

   Calculate the channel capacity of this Z-channel (why is it called Z-channel by the way?).
2) Consider a binary symmetric channel with transition probability p. The output of this channel is fed to the input of a binary erasure channel with erasure probability p'. What is the capacity of this composite channel?