Carnegie Mellon University
15-859: Information Theory and Applications in TCS — Spring 2013

PROBLEM SET 2
Due by Thursday, February 28

INSTRUCTIONS

- You are allowed to collaborate with up to two other students taking the class in solving problem sets. But here are some rules concerning such collaboration:
  1. You should think about each problem by yourself for at least 30 minutes before commencing any collaboration.
  2. Collaboration is defined as discussion of the lecture material and solution approaches to the problems. Please note that *you are not allowed to share any written material and you must write up solutions on your own without any "collaboration notes" as an aid.*
  3. You must clearly acknowledge your collaborator(s) in the write-up of your solutions.
  4. Of course, if you prefer, you can also work alone (see the last bullet item for some "credit" for doing so).
- Solutions typeset in LaTeX are strongly preferred.
- You should *not* search for solutions on the web. More generally, you are urged to try and solve the problems without consulting *any* reference material other than the course notes and what we cover in class. If for some reason you feel the need to consult some source, *please acknowledge the source* and try to articulate the difficulty you couldn't overcome before consulting the source and how it helped you overcome that difficulty. Alternatively, before turning to any such material, we encourage you to ask us for hints or clarifications.
- Please start work on the problem set early. The problem set has **four** problems and is worth a total of 100 points. As a rather rough guess/estimate, scoring around 80% of the points, or 70% of the points if you work by yourself, might suffice for an A on this problem set.

---

1. *Fun with Fano* (20 points)

   (a) Consider the following two pairs of correlated random variables:

   i. X is uniform on $\{0,1\}^n$, Y equals the first $n/2$ bits of X.

   ii. With probability α, $\alpha \in (0,1)$, X is uniform on $\{0,1\}^n$ and $Y = X$; and with probability $1-\alpha$, X is uniform on $\{0,1\}^n$ and Y is the all 0s string.

   Suppose we observe Y and estimate $\hat{X} = g(Y)$. What is the minimum possible value of $\Pr[\hat{X} \neq X]$ in the above two examples? What lower bound does Fano's inequality give in the two examples?

   (b) Suppose X and Y are two correlated random variables taking values in $\{0,1\}^n$. For $i = 0,1,\ldots,n$, define $\theta_i = \Pr[\Delta(X,Y) = i]$. Prove that

   $$H(X|Y) \le \sum_{i=0}^{n} \theta_i \log_2\left(\binom{n}{i}\frac{1}{\theta_i}\right) .$$

   Hint: Define the random variable $E = \Delta(X,Y)$, and mimic steps from the proof of Fano's inequality.

2. *Maximum Likelihood Decoding* (25 points)

   (a) Consider a *symmetric* **binary input** channel W with finite output alphabet $\mathcal{Y}$ and given by transition probabilities $p(y|0)$ and $p(y|1)$ for $y \in \mathcal{Y}$. Prove that the capacity of the channel W equals
   $$\frac{1}{2}\sum_{b\in\{0,1\}}\sum_{y\in\mathcal{Y}} p(y|b) \lg \frac{2p(y|b)}{p(y|0)+p(y|1)} \, .$$

   (b) Consider the channel W above. Suppose a bit $c \in \{0,1\}$ was transmitted and we receive $y \in \mathcal{Y}$. We decode y into the bit 0 if $p(y|0) > p(y|1)$, and 1 otherwise. Prove that the probability we make a decoding error (i.e., our estimate differs from c) is at most
   $$Z(W) = \sum_{y\in\mathcal{Y}} \sqrt{p(y|0)\cdot p(y|1)} \, .$$

   (c) Suppose a code $C \subseteq \{0,1\}^n$ is used for transmitting a sequence of n bits on the discrete memoryless channel W. Consider the following maximum likelihood decoding rule at the receiver: if $\mathbf{y} \in \mathcal{Y}^n$ is received, output a codeword $\mathbf{c} \in C$ for which $p(\mathbf{y}|\mathbf{c}) = \prod_{i=1}^n p(y_i|c_i)$ is maximum, ties broken arbitrarily.

   Prove that if a codeword $\mathbf{c_0} \in C$ was transmitted on the channel, and $\mathbf{y_0} \in \mathcal{Y}^n$ was received, the probability that the above decoding rule outputs a codeword different from $\mathbf{c_0}$ is at most
   $$\sum_{j=1}^n d_j \cdot Z(W)^j$$
   where d_j is the number of codewords in C at Hamming distance j from $\mathbf{c_0}$.

3. *Lossy compression* (30 points)

   In *lossy source coding*, the goal is to compress a sequence $\underline{X} := (X_1,\ldots,X_n)$ emitted from an iid source $X \in \mathcal{X}$ into Rn bits so that it is possible to *approximate* the original sequence from the compressed information. In this model, an encoder is a deterministic function
   $$f\colon \mathcal{X}^n \to \{0,1\}^{Rn},$$
   and the decoder is set to be
   $$g\colon \{0,1\}^{Rn} \to \mathcal{X}^n.$$
   The quality of the approximation is determined by a non-negative *distortion* function
   $$\Delta\colon \mathcal{X}\times\mathcal{X} \to \mathbb{R},$$
   which can be extended to sequences $\underline{x} := (x_1,\ldots,x_n)$ and $\underline{y} := (y_1,\ldots,y_n)$ as follows
   $$\Delta(\underline{x},\underline{y}) := \frac{1}{n}\sum_{i=1}^n \Delta(x_i,y_i).$$
   We assume that Δ is always upper bounded by some real number $d_{\max}$. The approximation is acceptable for a distortion parameter $D \geq 0$ if the expected distortion is bounded as follows
   $$\mathbb{E}[\Delta(\underline{X}, g(f(\underline{X})))] \leq D.$$
   The goal of this exercise is to prove an upper bound on the achievable rates R using a random coding argument similar to the achievability part of the channel coding theorem.

(a) We define jointly typical sequences similarly to what we saw in the lectures except for an additional constraint. Namely, for a distortion parameter D and jointly distributed random variables $(X, Y) \leftarrow p(x, y)$, we define

$$\begin{aligned}
A^n_{\epsilon,\Delta} :=\{(\underline{x}, \underline{y})\colon & \\
& \big| - \frac{\log p(\underline{x})}{n} - H(X)\big| < \epsilon, \\
& \big| - \frac{\log p(\underline{y})}{n} - H(Y)\big| < \epsilon, \\
& \big| - \frac{\log p(\underline{x}, \underline{y})}{n} - H(X, Y)\big| < \epsilon, \\
& \big|\Delta(\underline{x}, \underline{y}) - \mathbb{E}[\Delta(X, Y)]\big| < \epsilon\}.
\end{aligned}$$

Suppose $(X_1, Y_1), \dots (X_n, Y_n)$ are drawn iid from the joint distribution of (X, Y). Show that $\Pr[(\underline{X}, \underline{Y}) \in A^n_{\epsilon,\Delta}] \to 1$ as $n \to \infty$.

(b) Let $(\underline{x}, \underline{y}) \in A^n_{\epsilon,\Delta}$. Using the definition of $A^n_{\epsilon,\Delta}$, show that

$$p(\underline{y}) \geq p(\underline{y}|\underline{x}) 2^{-n(I(X;Y)+3\epsilon)}.$$

(c) Let $\underline{X}$ be drawn iid according to $p(x)$ and suppose we wish to encode the sequence at rate R and distortion bounded by D with respect to a distortion measure Δ. Fix a joint distribution $p(x, y)$, to be determined later, over $\mathcal{X} \times \mathcal{X}$. We assume the distribution has the property that, for $(X, Y) \sim p(x, y)$,

$$\mathbb{E}[\Delta(X, Y)] \leq D.$$

We generate a random codebook consisting of 2^{nR} codewords, indexed from 1 to 2^{nR}. Each codeword is of length n and generated iid according to the marginal distribution $p(y)$. Now we define the encoder $f(\underline{x})$ as follows.

> Encode $\underline{x}$ to the index of a codeword $\underline{y}$ such that $(\underline{x}, \underline{y}) \in A^n_{\epsilon,\Delta}$. If there is no such codeword, encode to 1 and if there is more than one suitable choice, encode to the least possible index.

The decoder $g(i)$, given an index i, simply outputs the ith codeword. Show that, for this coding scheme,

$$\mathbb{E}[\Delta(\underline{X}, g(f(\underline{X}))] \leq D + \epsilon + d_{\max} p_0,$$

where $p_0 := \Pr[(\underline{X}, \underline{Y}) \notin A^n_{\epsilon,\Delta}]$.

(d) Define

$$A(\underline{x}, \underline{y}) := \begin{cases} 1 & \text{if } (\underline{x}, \underline{y}) \in A^n_{\epsilon,\Delta}, \\ 0 & \text{otherwise.} \end{cases}$$

Show that

$$p_0 = \sum_{\underline{x}} p(\underline{x}) \Big[1 - \sum_{\underline{y}} p(\underline{y}) A(\underline{x}, \underline{y})\Big]^{2^{nR}}.$$

(e) Use part (b) to show that

$$p_0 \leq \sum_{\underline{x}} p(\underline{x}) \Big(1 - 2^{-n(I(X;Y)+3\epsilon)} \sum_{\underline{y}} p(\underline{y}|\underline{x}) \cdot A(\underline{x}, \underline{y})\Big)^{2^{nR}}.$$

(f) Deduce from the previous part that $p_0 \leq \epsilon$, for large enough n, provided that $R > I(X;Y) + 3\epsilon$. (Hint: you may find the inequality $(1 - xy)^n \leq 1 - x + e^{-yn}$ useful).

(g) Conclude that the encoder/decoder pair can achieve rates arbitrarily close to

$$R^* := \min_{p(y|x):\ \sum_{x,y} p(x)p(y|x)\Delta(x,y) \leq D} I(X;Y)$$

and expected distortion arbitrarily close to D.

(h) Now consider the Hamming distortion measure over the binary alphabet $\mathcal{X} = \{0, 1\}$

$$\Delta(x,y) := \begin{cases} 0 & \text{if } x = y, \\ 1 & \text{if } x \neq y. \end{cases}$$

Suppose $X \sim \text{Bernoulli}(p)$. Show that in this case $R^* \leq h(p) - h(D)$, concluding that X can be compressed at rates arbitrarily close to $h(p) - h(D)$ within an expected Hamming distortion of D.

Remark: In fact $R^* = h(p) - h(D)$ in this case, but you don't have to prove the "converse" coding theorem showing optimality of compression rate of $h(p) - h(D)$.

4. *Achieving capacity using linear codes* (25 points)

(a) Give a brief but convincing argument (full proof not required) why the proof of Shannon's theorem that we did in class holds for the binary symmetric channel (showing the capacity of BSC_p to be at least $1 - h(p)$ for $p \in [0, 1/2]$) even if the encoding function E is restricted to be linear. (Recall that a linear encoding function E is given $E(x) = Gx$ for all $x \in \{0,1\}^k$ for some $G \in \{0,1\}^{n \times k}$ where the operations are modulo 2.)

(b) Recall the binary erasure channel (BEC) with erasure probaility α: the input to this channel is a bit $x \in \{0, 1\}$, and the output is x with probability $1 - \alpha$, and an erasure '?' with probability α. For a linear code $C = \{Gx \mid x \in \{0,1\}^k\}$ generated by an $n \times k$ matrix G over $\{0, 1\}$, let $D : \{0, 1, ?\}^n \to C \cup \{\mathsf{fail}\}$ be the following decoder:

$$D(y) = \begin{cases} c & \text{if } y \text{ agrees with exactly one } c \in C \text{ on the unerased entries} \\ \mathsf{fail} & \text{otherwise} \end{cases}$$

i. For positive integers $k \leq m$, show that less than a fraction 2^{k-m} of the matrices M in $\{0,1\}^{m \times k}$ have rank less than k (over the field $\{0, 1\}$ of integers mod 2).

ii. For a set $J \subseteq \{1, 2, \ldots, n\}$, let $P_{\text{err}}(G|J)$ be the probability (over the channel noise and choice of a random message) that D outputs fail conditioned on the erasures being indexed by J. Prove that the average value of $P_{\text{err}}(G|J)$ taken over all $G \in \{0,1\}^{n \times k}$ is less than $q^{k-n+|J|}$.

iii. Let $P_{\text{err}}(G)$ be the decoding error probability of the decoder D for communication using the code generated by G on BEC with erasure probability α. Show that when $k = Rn$ for $R < 1 - \alpha$, the average value of $P_{\text{err}}(G)$ over all $n \times k$ matrices G is exponentially small in n.

iv. Using the above parts, conclude that one can reliably communicate on the BEC with erasure probability α at any rate less than $1 - \alpha$ using a *linear* code.