EE276: Homework #7 Solutions

Due on Friday March 6, 6pm - Gradescope entry code: E6VP4X

1. **Method of Types and Constrained Sets.**
Consider a finite alphabet $\mathcal{X}$. Given $D \geq 0$ and a weight function $\rho : \mathcal{X} \to \mathbb{R}_+$, define

$$B_n(\rho, D) := \left\{ x^n : \frac{1}{n} \sum_{i=1}^{n} \rho(x_i) \leq D \right\}.$$

(a) Show that

$$B_n(\rho, D) = \bigcup_{p \in \mathbb{P}_n : \langle p, \rho \rangle \leq D} T(p)$$

where

$$\langle p, \rho \rangle := \sum_{x \in \mathcal{X}} p(x)\rho(x).$$

(b) Show that

$$|B_n(\rho, D)| \doteq 2^{n \max_{p : \langle p, \rho \rangle \leq D} H(p)},$$

where $\doteq$ denotes equality up to first order in the exponent, i.e., $\alpha_n \doteq \beta_n$ means that $\lim_{n\to\infty} \frac{1}{n} \log \frac{\alpha_n}{\beta_n} = 0$. (Use the expression derived in (a) to obtain lower and upper bounds on $|B_n(\rho, D)|$ that match up to first order in exponent.)

(c) Specializing the result of (b), show that for $D \in [0, 1/2]$,

$$\left| \left\{ y^n \in \{0,1\}^n : \frac{1}{n} \sum_{i=1}^{n} y_i \leq D \right\} \right| \doteq 2^{n h_2(D)}$$

where h_2 is the binary entropy function.

Solution:

(a) Let $p_{x^n} \in \mathbb{P}^n$. We claim that $\langle \rho, p_{x^n} \rangle = \frac{1}{n} \sum_{i=1}^{n} \rho(x_i)$. Indeed,

$$\begin{aligned}
\langle \rho, p_{x^n} \rangle &= \sum_{x \in \mathcal{X}} \rho(x) p_{x^n}(x) \\
&= \frac{1}{n} \sum_{x \in \mathcal{X}} \rho(x) N(x|x^n) \\
&= \frac{1}{n} \sum_{x \in \mathcal{X}} \rho(x) \sum_{i=1}^{n} \mathbf{1}_{\{x_i = x\}} \\
&= \frac{1}{n} \sum_{i=1}^{n} \sum_{x \in \mathcal{X}} \rho(x) \mathbf{1}_{\{x_i = x\}} \\
&= \frac{1}{n} \sum_{i=1}^{n} \rho(x_i).
\end{aligned}$$

Hence,

$$\begin{aligned} B_n(\rho, D) &= \bigcup_{x^n : \frac{1}{n}\sum_{i=1}^n \rho(x_i) \le D} \{x^n\} \\ &= \bigcup_{p_{x^n} \in \mathbb{P}_n : \langle p, \rho \rangle \le D} T(p) \end{aligned}$$

(b) Recall that in class we showed that

$$|T(p)| \doteq 2^{nH(p)}.$$

So the claim is saying that the size of the union is dominated by the size of the largest set. To prove this, we use the relation in (a) and upper bound

$$\begin{aligned} \left| \bigcup_{p \in \mathbb{P}_n : \langle p, \rho \rangle \le D} T(p) \right| &\le \sum_{p \in \mathbb{P}_n : \langle p, \rho \rangle \le D} |T(p)| \\ &\le |\mathbb{P}^n| \max_{p \in \mathbb{P}_n : \langle p, \rho \rangle \le D} |T(p)| \\ &\doteq (n+1)^{r-1} \max_{p \in \mathbb{P}_n : \langle p, \rho \rangle \le D} 2^{nH(p)} \\ &\doteq 2^{n \max_{p \in \mathbb{P}_n : \langle p, \rho \rangle \le D} H(p)}. \end{aligned}$$

Similarly, for the lower bound we can obtain

$$\begin{aligned} \left| \bigcup_{p \in \mathbb{P}_n : \langle p, \rho \rangle \le D} T(p) \right| &\ge \max_{p \in \mathbb{P}_n : \langle p, \rho \rangle \le D} |T(p)| \\ &\doteq 2^{n \max_{p \in \mathbb{P}_n : \langle p, \rho \rangle \le D} H(p)}. \end{aligned}$$

Along with part (a), this proves the required claim.

(c) Taking $\mathcal{X} = \{0, 1\}$ in (b), and $x^n = y^n$, $\rho(x) = x$ gives the result after noticing that the max in part (b) is given by

$$\max_{q \in [0, D]} h_2(q) = h_2(D)$$

when $D \le 1/2$.

2. **Counting.**
Let $\mathcal{X} = \{1, 2, \ldots, m\}$. Show that the number of sequences $x^n \in \mathcal{X}^n$ satisfying

$$\frac{1}{n} \sum_{i=1}^n g(x_i) \ge \alpha$$

is approximately equal to 2^{nH^*}, to first order in the exponent, for n sufficiently large, where

$$H^* = \max_{P:\sum_{i=1}^m P(i)g(i)\geq\alpha} H(P).$$

Solution:

We wish to count the number of sequences satisfying a certain property. Instead of directly counting the sequences, we will calculate the probability of the set under a uniform distribution. Since the uniform distribution puts a probability of $\frac{1}{m^n}$ on every sequence of length n, we can count the sequences by multiplying the probability of the set by m^n.

The probability of the set can be calculated easily from Sanov's theorem. Let Q be the uniform distribution, and let E be the set of sequences of length n satisfying

$$\frac{1}{n}\sum_{i=1}^{n} g(x_i) \geq \alpha.$$

Then by Sanov's theorem, we have

$$Q^n(E) \doteq 2^{-nD(P^*\|Q)},$$

where P^* is the type in E that is closest to Q. Since Q is the uniform distribution, $D(P\|Q) = \log m - H(P)$, and therefore P^* is the type in E that has maximum entropy. Therefore, if we let

$$H^* = \max_{P:\sum_{i=1}^m P(i)g(i)\geq\alpha} H(P),$$

we have

$$Q^n(E) \doteq 2^{-n(\log m - H^*)}.$$

Multiplying this by m^n to find the number of sequences in this set, we obtain

$$|E| \doteq 2^{-n\log m}\, 2^{nH^*}\, m^n = 2^{nH^*}.$$

3. **Convexity of rate distortion function.**
Assume $(X, Y) \sim p(x, y) = p(x)p(y|x)$. In this problem, you will show that for fixed $p(x)$, $I(X;Y)$ is a convex function of $p(y|x)$.

   (a) The log sum inequality states that for n positive numbers $a_1, a_2, \cdots, a_n$, and $b_1, b_2, \cdots, b_n$, we have

   $$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i\right) \log \left(\frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i}\right)$$

   with equality if and only if $\frac{a_i}{b_i}$ =constant. Using this inequality (you don't have to prove this inequality), show that $D(p||q)$ is convex in (p, q), i.e.,

   $$\lambda D(p_1||q_1) + (1-\lambda)D(p_2||q_2) \geq D(\lambda p_1 + (1-\lambda)p_2||\lambda q_1 + (1-\lambda)q_2)$$

(b) Let $p_1(y|x)$ and $p_2(y|x)$ be two different conditional distributions. For $i \in \{1, 2\}$, let $p_i(x, y) = p_i(y|x)p(x)$, i.e., their corresponding joint distributions. For $0 \leq \lambda \leq 1$, let $p_\lambda(y|x) \stackrel{\Delta}{=} \lambda p_1(y|x) + (1-\lambda)p_2(y|x)$. Show that

$$p_\lambda(y) = \lambda p_1(y) + (1-\lambda)p_2(y)$$

(c) The mutual information between random variables X and Y can be alternatively written as

$$I(X;Y) = D(p(x,y)||p(x)p(y))$$

Using this in addition to the results of the previous parts show that for fixed $p(x)$, $I(X;Y)$ is convex in $p(y|x)$.

(d) Using the previous part, show that the rate distortion function $R^{(I)}(D)$ is convex in the distortion parameter D.

Solution:

(a) By definition,

$$\begin{aligned}
&D(\lambda p_1 + (1-\lambda)p_2||\lambda q_1 + (1-\lambda)q_2) \\
&\qquad = \sum_{x\in\mathcal{X}} (\lambda p_1(x) + (1-\lambda)p_2(x)) \log \frac{\lambda p_1(x) + (1-\lambda)p_2(x)}{\lambda q_1(x) + (1-\lambda)q_2(x)} \\
&\qquad \stackrel{(a)}{\leq} \sum_{x\in\mathcal{X}} \lambda p_1(x) \log \frac{\lambda p_1(x)}{\lambda q_1(x)} + (1-\lambda)p_2(x) \log \frac{(1-\lambda)p_2(x)}{(1-\lambda)q_2(x)} \\
&\qquad = \sum_{x\in\mathcal{X}} \lambda p_1(x) \log \frac{p_1(x)}{q_1(x)} + (1-\lambda)p_2(x) \log \frac{p_2(x)}{q_2(x)} \\
&\qquad = \lambda D(p_1||q_1) + (1-\lambda)D(p_2||q_2)
\end{aligned}$$

where (a) is because of the log-sum inequality.

(b) We have

$$\begin{aligned}
p_\lambda(y) &= \sum_{x\in\mathcal{X}} p_\lambda(y|x)p(x) \\
&= \sum_{x\in\mathcal{X}} (\lambda p_1(y|x) + (1-\lambda)p_2(y|x))p(x) \\
&= \lambda \sum_{x\in\mathcal{X}} p_1(y|x)p(x) + (1-\lambda) \sum_{x\in\mathcal{X}} p_2(y|x)p(x) \\
&= \lambda p_1(y) + (1-\lambda)p_2(y).
\end{aligned}$$

(c) Let $p_1(y|x)$ and $p_2(y|x)$ be two different conditional distributions, and let $I_1(X;Y)$ and $I_2(X;Y)$ denote that respective mutual information between X and Y when

$p(x)$ is fixed. Note that

$$\begin{aligned}
&\lambda I_1(X;Y) + (1-\lambda) I_2(X;Y) \\
&\quad = \lambda D(p_1(x,y)||p(x)p_1(y)) + (1-\lambda) D(p_2(x,y)||p(x)p_2(y)) \\
&\quad \geq D(\lambda p_1(x,y) + (1-\lambda)p_2(x,y)||\lambda p(x)p_1(y) + (1-\lambda)p(x)p_2(y)) \\
&\quad = D(p_\lambda(x,y)||p(x)p_\lambda(y)) \\
&\quad = I_\lambda(X;Y).
\end{aligned}$$

where $I_\lambda(X;Y)$ corresponds to the mutual information between X and Y when the conditional distribution of Y given X is $p_\lambda(y|x)$.

(d) Consider distortions D_1 and D_2. We need to show that

$$R^{(I)}(\lambda D_1 + (1-\lambda)D_2) \leq \lambda R^{(I)}(D_1) + (1-\lambda) R^{(I)}(D_2)$$

for any $\lambda \in [0,1]$. To show this, consider the joint distributions achieving the rate-distortion optimum at D_1 and D_2, $p_1(x,\hat{x}) = p(x)p_1(\hat{x}|x)$ and $p_2(x,\hat{x}) = p(x)p_2(\hat{x}|x)$. Also consider the distribution $p_\lambda = \lambda p_1 + (1-\lambda)p_2$. Since distortion is a linear function of the joint probability distribution, the distortion for p_λ is at most $\lambda D_1 + (1-\lambda)D_2$. By definition of $R^{(I)}(D)$,

$$\begin{aligned}
R^{(I)}(\lambda D_1 + (1-\lambda)D_2) &\leq I_\lambda(X;\hat{X}) \\
&\leq \lambda I_1(X;\hat{X}) + (1-\lambda) I_2(X;\hat{X}) \\
&= \lambda R^{(I)}(D_1) + (1-\lambda) R^{(I)}(D_2)
\end{aligned}$$

where I_λ, I_1 and I_2 denote the mutual informations when the distribution is p_λ, p_1 and p_2, respectively. The second inequality uses the convexity of mutual information proved in part (c).

4. **Properties of** $R(D)$. Consider a discrete source $X \in \mathcal{X} = \{1, 2, \ldots, m\}$ with distribution $p_1, p_2, \ldots, p_m$ and a distortion measure $d(i,j)$. Let $R(D)$ be the rate distortion function for this source and distortion measure. Let $d'(i,j) = d(i,j) - w_i$ be a new distortion measure and let $R'(D)$ be the corresponding rate distortion function. Show that $R'(D) = R(D + \bar{w})$, where $\bar{w} = \sum p_i w_i$, and use this to show that there is no essential loss of generality in assuming that $\min_{\hat{x}} d(i,\hat{x}) = 0$, i.e., for each $x \in \mathcal{X}$, there is one symbol $\hat{x}$ which reproduces the source with zero distortion.

Solution:

By definition,

$$R'(D') = \min_{p(\hat{x}|x):\sum_{x,\hat{x}} p(\hat{x}|x)p(x)d'(x,\hat{x}) \leq D'} I(X;\hat{X}).$$

For any conditional distribution $p(\hat{x}|x)$, we have

$$\begin{aligned} D' &= \sum_{x,\hat{x}} p(x)p(\hat{x}|x)d'(x,\hat{x}) \\ &= \sum_{x,\hat{x}} p(x)p(\hat{x}|x)\big(d(x,\hat{x}) - w_x\big) \\ &= \sum_{x,\hat{x}} p(x)p(\hat{x}|x)d(x,\hat{x}) - \sum_{x} p(x)w_x \sum_{\hat{x}} p(\hat{x}|x) \\ &= D - \sum_{x} p(x)w_x \\ &= D - \bar{w}, \end{aligned}$$

or $D = D' + \bar{w}$.

Hence,

$$\begin{aligned} R'(D') &= \min_{p(\hat{x}|x):\sum_{x,\hat{x}} p(\hat{x}|x)p(x)d'(x,\hat{x}) \le D'} I(X;\hat{X}) \\ &= \min_{p(\hat{x}|x):\sum_{x,\hat{x}} p(\hat{x}|x)p(x)d(x,\hat{x}) \le D'+\bar{w}} I(X;\hat{X}) \\ &= R(D' + \bar{w}). \end{aligned}$$

For any distortion matrix, we can set

$$w_i = \min_{\hat{x}} d(i,\hat{x}),$$

hence ensuring that

$$\min_{\hat{x}} d'(x,\hat{x}) = 0$$

for every x. This produces only a shift in the rate–distortion function and does not change the essential theory. Hence, there is no essential loss of generality in assuming that for each $x \in \mathcal{X}$, there exists a symbol $\hat{x}$ that reproduces it with zero distortion.