EE276: Homework #8 Solutions

Due on Friday March 13, 11:59pm - Gradescope entry code: E6VP4X

1. **Shannon lower bound.**
   Let X be a continuous random variable with mean zero and variance σ^2. $R(D)$ is the corresponding rate-distortion function for mean-squared distortion.

   (a) Show the lower bound:

   $$h(X) - \frac{1}{2}\log(2\pi e D) \leq R(D).$$

   (b) Using the joint distribution shown in Figure 1, show the upper bound on $R(D)$:

   $$R(D) \leq \frac{1}{2}\log\frac{\sigma^2}{D}$$

   Are Gaussian random variables harder or easier to describe in bits - in the sense of achieving small mean squared error distortion - than other random variables with the same variance?

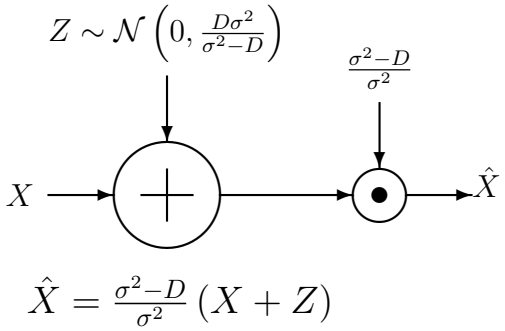

Figure 1: Joint distribution for upper bound on rate distortion function. The circle with the dot represents multiplication.

Solution:

(a) To prove the lower bound, we use the same techniques as used for the Guassian rate distortion function. Let $(X, \hat{X})$ be random variables such that $\mathbb{E}(X - \hat{X})^2 \leq$

D. Then

$$\begin{aligned} I(X;\hat{X}) &= h(X) - h(X|\hat{X}) \\ &= h(X) - h(X-\hat{X}|\hat{X}) \\ &\geq h(X) - h(X-\hat{X}) \\ &\geq h(X) - h(\mathcal{N}(0, \mathbb{E}(X-\hat{X})^2)) \\ &= h(X) - \frac{1}{2}\log(2\pi e)\mathbb{E}(X-\hat{X})^2 \\ &\geq h(X) - \frac{1}{2}\log(2\pi e)D. \end{aligned}$$

(b) Note that you could have used any constant instead of $(\sigma^2 - D)/\sigma^2$, since the mutual information in question will be the same for all choices. Here, we'll use the one in the problem.

To prove the upper bound, we consider the joint distribution as shown in Figure 1, and calculate the distortion and the mutual information between X and $\hat{X}$. Since

$$\hat{X} = \frac{\sigma^2 - D}{\sigma^2}(X + Z),$$

we have

$$\begin{aligned} \mathbb{E}(X-\hat{X})^2 &= \mathbb{E}\left(\frac{D}{\sigma^2}X - \frac{\sigma^2 - D}{\sigma^2}Z\right)^2 \\ &= \left(\frac{D}{\sigma^2}\right)^2 \mathbb{E}X^2 + \left(\frac{\sigma^2 - D}{\sigma^2}\right)^2 \mathbb{E}Z^2 \\ &= \left(\frac{D}{\sigma^2}\right)^2 \sigma^2 + \left(\frac{\sigma^2 - D}{\sigma^2}\right)^2 \frac{D\sigma^2}{\sigma^2 - D} \\ &= D, \end{aligned}$$

since X and Z are independent and zero mean. Also the mutual information is

$$\begin{aligned} I(X;\hat{X}) &= h(\hat{X}) - h(\hat{X}|X) \\ &= h(\hat{X}) - h(\frac{\sigma^2 - D}{\sigma^2}Z). \end{aligned}$$

Now

$$\begin{aligned} \mathbb{E}\hat{X}^2 &= \left(\frac{\sigma^2 - D}{\sigma^2}\right)^2 \mathbb{E}(X+Z)^2 \\ &= \left(\frac{\sigma^2 - D}{\sigma^2}\right)^2 (\mathbb{E}X^2 + \mathbb{E}Z^2) \\ &= \left(\frac{\sigma^2 - D}{\sigma^2}\right)^2 \left(\sigma^2 + \frac{D\sigma^2}{\sigma^2 - D}\right) \\ &= \sigma^2 - D. \end{aligned}$$

Hence, we have

$$\begin{aligned} I(X;\hat{X}) &= h(\hat{X}) - h(\frac{\sigma^2 - D}{\sigma^2}Z) \\ &= h(\hat{X}) - h(Z) - \log\frac{\sigma^2 - D}{\sigma^2} \\ &\leq h(\mathcal{N}(0, \sigma^2 - D)) - \frac{1}{2}\log(2\pi e)\frac{D\sigma^2}{\sigma^2 - D} - \log\frac{\sigma^2 - D}{\sigma^2} \\ &= \frac{1}{2}\log(2\pi e)(\sigma^2 - D) - \frac{1}{2}\log(2\pi e)\frac{D\sigma^2}{\sigma^2 - D} - \frac{1}{2}\log\left(\frac{\sigma^2 - D}{\sigma^2}\right)^2 \\ &= \frac{1}{2}\log\frac{\sigma^2}{D}, \end{aligned}$$

which combined with the definition of the rate distortion function gives us the required upper bound.

For a Gaussian random variable, $h(X) = \frac{1}{2}\log(2\pi e)\sigma^2$ and the lower bound is equal to the upper bound. For any other random variable, the lower bound is strictly less than the upper bound and hence non-Gaussian random variables cannot require more bits to describe to the same accuracy than the corresponding Gaussian random variables. This is not surprising, since the Gaussian random variable has the maximum entropy and we would expect that it would be the most difficult to describe.

2. **Rate distortion for uniform source with Hamming distortion.**
Consider a source X uniformly distributed on the set $\{1, 2, ..., m\}$. Find the rate distortion function for this source with Hamming distortion, i.e.,

$$d(x, \hat{x}) = \begin{cases} 0, & x = \hat{x} \\ 1, & x \neq \hat{x} \end{cases}$$

via the following steps:

   (a) Argue that $R(D) = 0$ when $D \geq 1 - \frac{1}{m}$.

   (b) Show that for $D \leq 1 - \frac{1}{m}$, $I(X;\hat{X}) \geq \log_2 m - h_2(D) - D\log_2(m-1)$ for any joint distribution $(X, \hat{X})$ satisfying the distortion constraint D.
   Hint: Fano's inequality.

   (c) Find distribution $p(\hat{x}|x)$ that achieves the above lower bound when $0 \leq D \leq 1 - \frac{1}{m}$.
   Hint: Consider the form below.

$$p(\hat{x}|x) = \begin{cases} a, & x = \hat{x} \\ b, & x \neq \hat{x} \end{cases}$$

   (d) Use the above parts to write down the rate-distortion function $R(D)$ for $D \geq 0$.

Solution:

X is uniformly distributed on the set $\{1, 2, ..., m\}$. The distortion measure is

$$d(x, \hat{x}) = \begin{cases} 0, & x = \hat{x} \\ 1, & x \neq \hat{x} \end{cases}$$

For (a), it's enough to see that setting $\hat{X} = 1$ independently of X achieves distortion $1 - 1/m$.

For (b), consider any joint distribution that satisfies the distortion constraint D. Since $D \geq Pr(X \neq \hat{X})$, we have by Fano's inequality

$$H(X|\hat{X}) \leq h_2(P_e) + P_e \log(m-1) \leq h_2(D) + D \log(m-1) \qquad \text{for } 0 \leq P_e \leq 1 - \frac{1}{m}$$

and hence

$$\begin{aligned} I(X; \hat{X}) &= H(X) - H(X|\hat{X}) \\ &\geq \log m - h_2(D) - D \log(m-1) \end{aligned}$$

For (c), we can achieve this lower bound by choosing $p(\hat{x})$ to be the uniform distribution, and the conditional distribution of $p(\hat{x}|x)$ to be

$$p(\hat{x}|x) = \begin{cases} 1 - D, & x = \hat{x} \\ \frac{D}{m-1}, & x \neq \hat{x} \end{cases}$$

It is easy to verify that this gives the right distribution on X and satisfies the bound with equality for $D < 1 - 1/m$.

For (d),

$$R(D) = \begin{cases} \log m - h_2(D) - D \log(m-1), & 0 \leq D \leq 1 - \frac{1}{m} \\ 0, & D > 1 - \frac{1}{m} \end{cases}$$

3. **Rate distortion for two independent sources.** Can one simultaneously compress two independent sources better than by compressing the sources individually? This problem addresses this question. Let $\{X_i\}$ be iid $\sim p(x)$ with distortion $d_1(x, \hat{x})$ and rate distortion function $R_X(D)$. Similarly, let $\{Y_i\}$ be iid $\sim p(y)$ with distortion $d_2(y, \hat{y})$ and rate distortion function $R_Y(D)$.

   Suppose the $\{X_i\}$ process and the $\{Y_i\}$ process are independent of each other.

   Suppose we now wish to describe the process $\{(X_i, Y_i)\}$ subject to distortions $\mathbb{E}[d_1(X, \hat{X})] \leq D_1$ and $\mathbb{E}[d_2(Y, \hat{Y})] \leq D_2$. Thus a rate $R_{X,Y}(D_1, D_2)$ is sufficient, where

$$R_{X,Y}(D_1, D_2) = \min_{p(\hat{x}, \hat{y}|x,y): \mathbb{E}[d_1(X,\hat{X})] \leq D_1, \mathbb{E}[d_2(Y,\hat{Y})] \leq D_2} I(X, Y; \hat{X}, \hat{Y})$$

(a) Show that
$$I(X,Y;\hat{X},\hat{Y}) \geq I(X;\hat{X}) + I(Y;\hat{Y})$$

(b) Hence conclude
$$R_{X,Y}(D_1,D_2) \geq R_X(D_1) + R_Y(D_2).$$

(c) Show that the above actually holds with equality.

Solution:

(a) Given that X and Y are independent, we have
$$p(x,y,\hat{x},\hat{y}) = p(x)p(y)p(\hat{x},\hat{y}|x,y)$$

Then
$$\begin{aligned} I(X,Y;\hat{X},\hat{Y}) &= H(X,Y) - H(X,Y|\hat{X},\hat{Y}) \\ &= H(X) + H(Y) - H(X|\hat{X},\hat{Y}) - H(Y|X,\hat{X},\hat{Y}) \\ &\geq H(X) + H(Y) - H(X|\hat{X}) - H(Y|\hat{Y}) \\ &= I(X;\hat{X}) + I(Y;\hat{Y}) \end{aligned}$$

where the inequality follows from the fact that conditioning reduces entropy.

(b) From part (a), we can write
$$\begin{aligned} R_{X,Y}(D_1,D_2) &= \min_{p(\hat{x},\hat{y}|x,y):Ed(X,\hat{X})\leq D_1, Ed(Y,\hat{Y})\leq D_2} I(X,Y;\hat{X},\hat{Y}) \\ &\geq \min_{p(\hat{x},\hat{y}|x,y):Ed(X,\hat{X})\leq D_1, Ed(Y,\hat{Y})\leq D_2} \Big(I(X;\hat{X}) + I(Y;\hat{Y}) \Big) \\ &= \min_{p(\hat{x}|x):Ed(X,\hat{X})\leq D_1} I(X;\hat{X}) + \min_{p(\hat{y}|y):Ed(Y,\hat{Y})\leq D_2} I(Y;\hat{Y}) \\ &= R_X(D_1) + R_Y(D_2) \end{aligned}$$

Thus we cannot simultaneously compress two independent sources better than by compressing the sources individually.

(c) If
$$p(x,y,\hat{x},\hat{y}) = p(x)p(y)p(\hat{x}|x)p(\hat{y}|y),$$

then
$$\begin{aligned} I(X,Y;\hat{X},\hat{Y}) &= H(X,Y) - H(X,Y|\hat{X},\hat{Y}) \\ &= H(X) + H(Y) - H(X|\hat{X},\hat{Y}) - H(Y|X,\hat{X},\hat{Y}) \\ &= H(X) + H(Y) - H(X|\hat{X}) - H(Y|\hat{Y}) \\ &= I(X;\hat{X}) + I(Y;\hat{Y}) \end{aligned}$$

Let $p(x,\hat{x})$ be a distribution that achieves the rate distortion $R_X(D_1)$ at distortion D_1 and let $p(y,\hat{y})$ be a distribution that achieves the rate distortion $R_Y(D_2)$ at

distortion D_2. Then for the product distribution $p(x, y, \hat{x}, \hat{y}) = p(x, \hat{x})p(y, \hat{y})$, where the component distributions achieve rates $(D_1, R_X(D_1))$ and $(D_2, R_X(D_2))$, the mutual information corresponding to the product distribution is $R_X(D_1) + R_Y(D_2)$. Thus

$$R_{X,Y}(D_1, D_2) = \min_{p(\hat{x},\hat{y}|x,y):Ed(X,\hat{X})\leq D_1, Ed(Y,\hat{Y})\leq D_2} I(X, Y; \hat{X}, \hat{Y}) = R_X(D_1) + R_Y(D_2)$$

Thus by using the product distribution, we can achieve the sum of the rates.

Therefore the total rate at which we encode two independent sources together with distortions D_1 and D_2 is the same as if we encoded each of them separately.

4. In what follows, all random variables have finite alphabet, and all pmfs are defined on finite alphabets. First, recall the definitions from class:

   Strongly typical. A sequence $x^n \in \mathcal{X}^n$ is *strongly δ-typical* with respect to a probability mass function $P \in \mathcal{M}(\mathcal{X})$ if

   $$|P_{x^n}(a) - P(a)| \leq \delta \cdot P(a), \quad \forall a \in \mathcal{X}$$

   where $P_{x^n}(a)$ is the empirical probability of seeing a based on sequence x^n.

   In words, a sequence is strongly δ-typical with respect to P if its empirical distribution is close to the probability mass function P. (δ is some fixed number, typically small.) The *strongly δ-typical set* (ie. strongly typical set) of P, $T_\delta(P)$, is defined as the set of all sequences that are strongly δ-typical with respect to P, i.e.

   $$T_\delta(P) = \{x^n : |P_{x^n}(a) - P(a)| \leq \delta \cdot P(a), \forall a \in \mathcal{X}\}$$

   Recall: the *weakly ϵ-typical set*, which you are familiar with, of an IID source P is defined as

   $$A_\epsilon(P) := \{x^n : \left| -\frac{1}{n} \log P(x^n) - H(P) \right| \leq \epsilon\}.$$

   (Strongly) Jointly typical. In the following, we refer to the sequences $x^n = (x_1, x_2, \ldots, x_n)$, $x_i \in \mathcal{X}$ and $y^n = (y_1, y_2, \ldots, y_n)$, $y_i \in \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are finite alphabets.
   The *joint empirical distribution* of (x^n, y^n) is:

   $$P_{x^n,y^n}(x, y) = \frac{1}{n} N(x, y|x^n, y^n)$$

   where

   $$N(x, y|x^n, y^n) := \sum_{i=1}^{n} \mathbf{1}_{\{x_i=x, y_i=y\}}.$$

   (x^n, y^n) is *jointly δ-typical* with respect to $P \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})$ if

   $$|P_{x^n,y^n}(x, y) - P(x, y)| \leq \delta \cdot P(x, y), \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}.$$

   The *jointly δ-typical set* with respect to $P \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})$ is

   $$T_\delta(P) = \{(x^n, y^n) : (x^n, y^n) \text{ is jointly } \delta\text{-typical with respect to } P\}.$$

(a) Let Z be a random variable with pmf p_z and alphabet $\mathcal{Z}$. Let $T_\delta(p_z)$ be the set of strongly δ-typical sequences with respect to p_z.
Show that for any nonnegative $g : \mathcal{Z} \to \mathbb{R}^+$, and $z^n \in T_\delta(p_z)$, we have

$$\left| \frac{1}{n} \sum_{i=1}^{n} g(z_i) - \mathbb{E}[g(Z)] \right| \leq \delta \mathbb{E}[g(Z)].$$

In what follows, it may be useful to invoke part (a) in your arguments.

(b) Let (X, Y) be random variables with joint pmf $p_{x,y}$. Let $T_\delta(p_{x,y})$ be the set of jointly δ-typical sequences with respect to $p_{x,y}$. Show that

$$\frac{1}{n} \sum_{i=1}^{n} d(x_i, y_i) \leq (1 + \delta)\mathbb{E}[d(X, Y)]$$

for any distortion function $d(x, y)$ and $(x^n, y^n) \in T_\delta(p_{x,y})$.

(c) Let $A_\epsilon(p)$ denote the (weakly) typical set with respect to p, and $T_\delta(p)$ be the set of δ-typical sequences with respect to p. Show that

$$T_\delta(p) \subseteq A_\epsilon(p)$$

for $\epsilon = \delta H(p)$.

(d) Let Q and P be pmfs over $\mathcal{X}$, and $T_\delta(P)$ be the set of δ-typical sequences with respect to P. Assuming $D(P\|Q)$ is finite, show that

$$Q(T_\delta(P)) \doteq 2^{-n(D(P\|Q)-\alpha(\delta))},$$

where $\alpha(\delta) \geq 0$ and $\alpha(\delta) \to 0$ as $\delta \to 0$.

Solution:

(a) From the definition of strong typicality, we know

$$|p_{z^n}(z) - p_z(z)| \leq \delta p_z(z) \qquad \forall z \in \mathcal{Z}.$$

With nonnegative function g,

$$\sum_{z \in \mathcal{Z}} -g(z)\delta p_z(z) \leq \sum_{z \in \mathcal{Z}} g(z) \left(p_{z^n}(z) - p_z(z)\right) \leq \sum_{z \in \mathcal{Z}} g(z)\delta p_z(z).$$

Plugging in $\mathbb{E}[g(Z)] = \sum_{z \in \mathcal{Z}} g(z)p_z(z)$, we get

$$-\delta \mathbb{E}[g(Z)] \leq \sum_{z \in \mathcal{Z}} g(z)p_{z^n}(z) - \mathbb{E}[g(Z)] \leq \delta \mathbb{E}[g(Z)].$$

Since

$$\frac{1}{n} \sum_{i=1}^{n} g(z_i) = \sum_{z \in \mathcal{Z}} g(z)p_{z^n}(z),$$

we conclude

$$-\delta \mathbb{E}[g(Z)] \leq \frac{1}{n} \sum_{i=1}^{n} g(z_i) - \mathbb{E}[g(Z)] \leq \delta \mathbb{E}[g(Z)].$$

(b) Let $Z = (X, Y)$ and apply part (a) with $g(z) = d(x, y)$.

(c) Let $X \sim p$. Applying part (a) to $Z = X$ and $g(z) = -\log p(z)$ gives

$$(1-\delta)H(p) \le -\frac{1}{n}\sum_i \log p(x_i) \le (1+\delta)H(p)$$

for any $x^n \in T_\delta(p)$. Rearranging and raising to the exponent then gives

$$2^{-n(1+\delta)H(p)} \le p(x^n) \le 2^{-n(1-\delta)H(p)}$$

which implies $x^n \in A_\epsilon(p)$ with $\epsilon = \delta H(p)$.

(d) For the upper bound, we have via a union bound

$$\begin{aligned} Q(T_\delta(p)) &= Q\left(\bigcup_{\hat{p}\in\mathbb{P}_n:|\hat{p}-p|\le\delta p} \{x^n : p_{x^n} = \hat{p}\}\right) \\ &\le \sum_{\hat{p}\in\mathbb{P}_n:|\hat{p}-p|\le\delta p} Q(T(\hat{p})) \\ &\dot{\le} \sum_{\hat{p}\in\mathbb{P}_n:|\hat{p}-p|\le\delta p} 2^{-nD(\hat{p}\|q)} \end{aligned}$$

Since there are at most $|\mathbb{P}_n|$ terms in the sum, we further bound this by

$$\sum_{\hat{p}\in\mathbb{P}_n:|\hat{p}-p|\le\delta p} 2^{-nD(\hat{p}\|q)} \le |\mathbb{P}_n| \max_{\hat{p}\in\mathbb{P}_n:|\hat{p}-p|\le\delta p} 2^{-nD(\hat{p}\|q)}.$$

Using $|\mathbb{P}_n| \le (n+1)^{|\mathcal{X}|-1} \le 2^{n\alpha}$ for some small α, we get

$$|\mathbb{P}_n| \max_{\hat{p}\in\mathbb{P}_n:|\hat{p}-p|\le\delta p} 2^{-nD(\hat{p}\|q)} \dot{\le} 2^{-n\min_{\hat{p}\in\mathbb{P}_n:|\hat{p}-p|\le\delta p} D(\hat{p}\|q)}.$$

For the lower bound, we bound the probability of the union by the probability of the most probable set:

$$\begin{aligned} Q(T_\delta(p)) &= Q\left(\bigcup_{\hat{p}\in\mathbb{P}_n:|\hat{p}-p|\le\delta p} \{x^n : p_{x^n} = \hat{p}\}\right) \\ &\ge \max_{\hat{p}\in\mathbb{P}_n:|\hat{p}-p|\le\delta p} Q(T(\hat{p})) \\ &\dot{\ge} \max_{\hat{p}\in\mathbb{P}_n:|\hat{p}-p|\le\delta p} 2^{-nD(\hat{p}\|q)} \\ &= 2^{-n\min_{\hat{p}\in\mathbb{P}_n:|\hat{p}-p|\le\delta p} D(\hat{p}\|q)}. \end{aligned}$$

This shows that

$$Q(T_\delta(p)) \doteq 2^{-n\min_{\hat{p}\in\mathbb{P}_n:|\hat{p}-p|\le\delta p} D(\hat{p}\|q)}.$$

Now let p^* be the minimizer in the above expression. Since the type of p^* is in the typical set, we have by (a) that

$$\begin{aligned}|D(p^*\|q) - D(p\|q)| &= |\mathbb{E}_{p^*}[\log(p^*/q)] - \mathbb{E}_p[\log(p/q)]| \\ &\leq |\mathbb{E}_{p^*}[\log(p^*/q)] - \mathbb{E}_{p^*}[\log(p/q)]| + |\mathbb{E}_{p^*}[\log(p/q)] - \mathbb{E}_p[\log(p/q)]|.\end{aligned}$$

The first term can be bounded by

$$|\mathbb{E}_{p^*}[\log(p^*/q)] - \mathbb{E}_{p^*}[\log(p/q)]| = |\mathbb{E}_{p^*}[\log(p^*/p)]| \leq \max\{|\log(1+\delta)|, |\log(1-\delta)|\}.$$

This converges to zero as $\delta \to 0$. The second term can be bounded by

$$\begin{aligned}|\mathbb{E}_{p^*}[\log(p/q)] - \mathbb{E}_p[\log(p/q)] &\leq |\mathbb{E}_{p^*}[\log(p)] - \mathbb{E}_p[\log(p)] + |\mathbb{E}_{p^*}[\log(q)] - \mathbb{E}_p[\log(q)] \\ &\leq \delta(\mathbb{E}_p[-\log(p)] + \mathbb{E}_p[-\log(q)]),\end{aligned}$$

where in the last step we apply (a) to $-\log(p)$ and $-\log(q)$. Since the alphabet is finite, $\mathbb{E}_p[-\log(p)]$ is finite; since $D(p\|q)$ is finite, $\mathbb{E}_p[-\log(q)])$ is finite. Thus, thi term also converges to 0 as $\delta \to 0$. Combing these two, we have shown that

$$|D(p^*\|q) - D(p\|q)| \to 0$$

as $\delta \to 0$.

Meanwhile, by definition,

$$D(p^*\|q) \leq D(p\|q),$$

hence we must have

$$D(p^*\|q) = D(p\|q) - \alpha(\delta),$$

where $\alpha(\delta) \geq 0$ and $\alpha(\delta) \to 0$ as $\delta \to 0$.

Note: Students can argue for the above expression without showing all the math by just saying that $D(p^*\|q) \leq D(p\|q)$ and as $\delta \to 0$, p^* gets close to p.