1st Semester 2010/11

Solutions to Homework Set #1
Sanov's Theorem, Rate distortion

1. **Sanov's theorem:**
Prove the simple version of Sanov's theorem for the binary random variables, i.e., let $X_1, X_2, \ldots, X_n$ be a sequence of binary random variables, drawn i.i.d. according to the distribution:

$$\Pr(X = 1) = q, \quad \Pr(X = 0) = 1 - q. \tag{1}$$

Let the proportion of 1's in the sequence $X_1, X_2, \ldots, X_n$ be $p_{\mathbf{X}}$, i.e.,

$$p_{X^n} = \frac{1}{n} \sum_{i=1}^{n} X_i. \tag{2}$$

By the law of large numbers, we would expect $p_{\mathbf{X}}$ to be close to q for large n. Sanov's theorem deals with the probability that p_{X^n} is far away from q. In particular, for concreteness, if we take $p > q > \frac{1}{2}$, Sanov's theorem states that

$$-\frac{1}{n} \log \Pr\{(X_1, X_2, \ldots, X_n) : p_{X^n} \geq p\} \to p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q} = D((p, 1-p)||(q, 1-q) \tag{3}$$

Justify the following steps:

- $$\Pr\{(X_1, X_2, \ldots, X_n) : p_{\mathbf{X}} \geq p\} \leq \sum_{i=\lfloor np \rfloor}^{n} \binom{n}{i} q^i (1-q)^{n-i} \tag{4}$$

- Argue that the term corresponding to $i = \lfloor np \rfloor$ is the largest term in the sum on the right hand side of the last equation.
- Show that this term is approximately 2^{-nD}.
- Prove an upper bound on the probability in Sanov's theorem using the above steps. Use similar arguments to prove a lower bound and complete the proof of Sanov's theorem.

Solution
Sanov's theorem

- Since $n\overline{X}_n$ has a binomial distribution, we have

$$\Pr(n\overline{X}_n = i) = \binom{n}{i} q^i (1-q)^{n-i} \tag{5}$$

and therefore

$$\Pr\{(X_1, X_2, \ldots, X_n) : p_{\mathbf{X}} \geq p\} \leq \sum_{i=\lfloor np \rfloor}^{n} \binom{n}{i} q^i (1-q)^{n-i} \tag{6}$$

-

$$\frac{\Pr(n\overline{X}_n = i+1)}{\Pr(n\overline{X}_n = i)} = \frac{\binom{n}{i+1} q^{i+1}(1-q)^{n-i-1}}{\binom{n}{i} q^i (1-q)^{n-i}} = \frac{n-i}{i+1}\frac{q}{1-q} \tag{7}$$

This ratio is less than 1 if $\frac{n-i}{i+1} < \frac{1-q}{q}$,i.e., if $i > nq - (1-q)$. Thus the maximum of the terms occurs when $i = \lfloor np \rfloor$.

- From Example 11.1.3,

$$\binom{n}{\lfloor np \rfloor} \doteq 2^{nH(p)} \tag{8}$$

and hence the largest term in the sum is

$$\binom{n}{\lfloor np \rfloor} q^{\lfloor np \rfloor}(1-q)^{n-\lfloor np \rfloor} = 2^{n(-p\log p-(1-p)\log(1-p))+np\log q+n(1-p)\log(1-q)} = 2^{-nD(p||q)} \tag{9}$$

- From the above results, it follows that

$$\begin{aligned} \Pr\{(X_1, X_2, \ldots, X_n) : p_{\mathbf{X}} \geq p\} &\leq \sum_{i=\lfloor np \rfloor}^{n} \binom{n}{i} q^i (1-q)^{n-i} && (10)\\ &\leq (n - \lfloor np \rfloor)\binom{n}{\lfloor np \rfloor} q^i (1-q)^{n-i} && (11)\\ &\leq (n(1-p)+1)2^{-nD(p||q)} && (12) \end{aligned}$$

where the second inequality follows from the fact that the sum is less than the largest term times the number of terms. Taking the

logarithm and dividing by n and taking the limit as $n \to \infty$, we obtain

$$\lim_{n\to\infty} \frac{1}{n} \log \Pr\{(X_1, X_2, \ldots, X_n) : p_{\mathbf{X}} \geq p\} \leq -D(p||q) \tag{13}$$

Similarly, using the fact the sum of the terms is larger than the largest term, we obtain

$$\begin{aligned} \Pr\{(X_1, X_2, \ldots, X_n) : p_{\mathbf{X}} \geq p\} &\geq \sum_{i=\lceil np \rceil}^{n} \binom{n}{i} q^i (1-q)^{n-i} && (14) \\ &\geq \binom{n}{\lceil np \rceil} q^i (1-q)^{n-i} && (15) \\ &\geq 2^{-nD(p||q)} && (16) \end{aligned}$$

and

$$\lim_{n\to\infty} \frac{1}{n} \log \Pr\{(X_1, X_2, \ldots, X_n) : p_{\mathbf{X}} \geq p\} \geq -D(p||q) \tag{17}$$

Combining these two results, we obtain the special case of Sanov's theorem

$$\lim_{n\to\infty} \frac{1}{n} \log \Pr\{(X_1, X_2, \ldots, X_n) : p_{\mathbf{X}} \geq p\} = -D(p||q) \tag{18}$$

2. **Strong Typicality**
Let X^n be drawn i.i.d.$\sim P(x)$. Prove that for each $x^n \in T_\delta(X)$,

$$2^{-n(H(X)+\delta')} \leq P^n(x^n) \leq 2^{-n(H(X)-\delta')}$$

for some $\delta' = \delta'(\delta)$ that vanishes as $\delta \to 0$.

Solution: Strong Typicality
From our familiar trick, we have

$$P^n(x^n) = 2^{-n\left(\sum_{a\in\mathcal{X}} P_{x^n}(a) \log \frac{1}{P(a)}\right)}$$

But, since $x^n \in T_\delta(X)$, we have $|P_{x^n}(a) - P(a)| < \frac{\delta}{|\mathcal{X}|}$ for all a. Therefore,

$$H(X) - \frac{\delta}{|\mathcal{X}|} \sum_a \log \frac{1}{P(a)} < \sum_{a\in\mathcal{X}} P_{x^n}(a) \log \frac{1}{P(a)} < H(X) + \frac{\delta}{|\mathcal{X}|} \sum_a \log \frac{1}{P(a)}$$

Hence,

$$2^{-n(H(X)+\delta')} \leq P^n(x^n) \leq 2^{-n(H(X)-\delta')}$$

where $\delta' = \frac{\delta}{|\mathcal{X}|} \sum_a \log \frac{1}{P(a)}$

3. **Weak Typicality vs. Strong Typicality**
In this problem, we compare the weakly typical set $A_\epsilon(P)$ and the strongly typical set $T_\delta(P)$. To recall, the definition of two sets are following.

$$A_\epsilon(P) = \left\{ x^n \in \mathcal{X}^n : \left| -\frac{1}{n} \log P^n(x^n) - H(P) \right| \leq \epsilon \right\}$$
$$T_\delta(P) = \left\{ x^n \in \mathcal{X}^n : \|P_{x^n} - P\|_\infty \leq \frac{\delta}{|\mathcal{X}|} \right\}$$

(a) Suppose P is such that $P(a) > 0$ for all $a \in \mathcal{X}$. Then, there is an inclusion relationship between the weakly typical set $A_\epsilon(P)$ and the strongly typical set $T_\delta(P)$ for an appropriate choice of ϵ. Which of the statement is true: $A_\epsilon(P) \subseteq T_\delta(P)$ or $A_\epsilon(P) \supseteq T_\delta(P)$? What is the appropriate relation between δ and ϵ?

(b) Give a description of the sequences that belongs to $A_\epsilon(P)$, vs. the sequences that belongs to $T_\delta(P)$, when the source is uniformly distributed, i.e. $P(a) = \frac{1}{|\mathcal{X}|}, \forall a \in \mathcal{X}$. (Assume $|\mathcal{X}| < \infty$.)

(c) Can you explain why $T_\delta(P)$ is called **strongly** typical set and $A_\epsilon(P)$ is called **weakly** typical set?

Solution: Weak Typicality vs. Strong Typicality

(a) From Problem 2, we can see that if $x^n \in T_\delta(P)$, then

$$\left| -\frac{1}{n} \log P^n(x^n) - H(P) \right| \leq \delta' = \frac{\delta}{|\mathcal{X}|} \sum_a \log \frac{1}{P(a)}$$

Therefore, we can see that if $\epsilon > \frac{\delta}{|\mathcal{X}|} \sum_a \log \frac{1}{P(a)}$, then $A_\epsilon(P) \supseteq T_\delta(P)$. To show that the other way does not hold, see the next part.

(b) When P is a uniform distribution, we can see that $A_\epsilon(P)$ includes every possible sequences x^n. To see this, we can easily see that $P^n(x^n) = (\frac{1}{|\mathcal{X}|})^n$, and thus, $-1/n \log P^n(x^n) = \log |\mathcal{X}| = H(P)$. Thus, $x^n \in A_\epsilon(P)$ for all x^n, for all $\epsilon > 0$. However, obviously, among those sequences in $A_\epsilon(P)$, only those who have the type

$$\frac{1}{|\mathcal{X}|} - \frac{\delta}{|\mathcal{X}|} \leq P_{x^n}(a) \leq \frac{1}{|\mathcal{X}|} + \frac{\delta}{|\mathcal{X}|}$$

for each letter $a \in \mathcal{X}$, are in $T_\delta(P)$. Therefore, for sufficiently small value of δ, there always exist some sequences that are in $A_\epsilon(P)$, but not in $T_\delta(P)$.

(c) From above questions, we can see that the strongly typical set is contained in the weakly typical set for appropriate choice of δ and ϵ. Thus, we can see that the definition of strong typical set is stronger than that of the weakly typical set.

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

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

Solution: Rate distortion for uniform source with Hamming distortion. X is uniformly distributed on the set $\{1, 2, \ldots, m\}$. The distortion measure is

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

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

$$H(X|\hat{X}) \leq H(D) + D \log(m-1), \tag{19}$$

and hence

$$\begin{aligned} I(X; \hat{X}) &= H(X) - H(X|\hat{X}) && (20) \\ &\geq \log m - H(D) - D \log(m-1). && (21) \end{aligned}$$

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

$$p(\hat{x}|x) \begin{cases} = 1 - D & \text{if } \hat{x} = x \\ = D/(m-1) & \text{if } \hat{x} \neq x. \end{cases} \tag{22}$$

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

$$R(D) \begin{cases} = \log m - H(D) - D\log(m-1) & \text{if } 0 \leq D \leq 1 - \frac{1}{m} \\ 0 & \text{if } D > 1 - \frac{1}{m}. \end{cases} \tag{23}$$

5. **Erasure distortion**
Consider $X \sim \text{Bernoulli}(\frac{1}{2})$, and let the distortion measure be given by the matrix

$$d(x, \hat{x}) = \begin{bmatrix} 0 & 1 & \infty \\ \infty & 1 & 0 \end{bmatrix}. \tag{24}$$

Calculate the rate distortion function for this source. Can you suggest a simple scheme to achieve any value of the rate distortion function for this source?

Solution: Erasure distortion
The infinite distortion constrains $p(0,1) = p(1,0) = 0$. Hence by symmetry the joint distribution of $(X, \hat{X})$ is of the form shown in Figure 1.

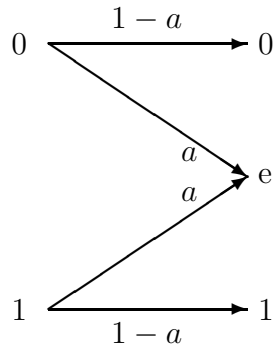

Figure 1: Joint distribution for erasure rate distortion of a binary source.

For this joint distribution, it is easy to calculate the distortion $D = a$ and that $I(X; \hat{X}) = H(X) - H(X|\hat{X}) = 1 - a$. Hence we have $R(D) = 1 - D$ for $0 \leq D \leq 1$. For $D > 1$, $R(D) = 0$.

It is very see how we could achieve this rate distortion function. If D is rational, say k/n, then we send only the first $n - k$ of any block of n bits. We reproduce these bits exactly and reproduce the remaining bits as erasures. Hence we can send information at rate $1 - D$ and achieve a distortion D. If D is irrational, we can get arbitrarily close to D by using longer and longer block lengths.

6. **Rate distortion.**
A memoryless source U is uniformly distributed on $\{0, \ldots, r-1\}$. The following distortion function is given by

$$d(u, v) = \begin{cases} 0, & u = v, \\ 1, & u = v \pm 1 \mod r, \\ \infty, & \text{otherwise.} \end{cases}$$

Show that the rate distortion function is

$$R(D) = \begin{cases} \log r - D - h_2(D), & D \leq \frac{2}{3}, \\ \log r - \log 3, & D > \frac{2}{3}. \end{cases}$$

Solution: Rate distortion

From the symmetry of the problem, we can assume the conditional distribution of $p(v|u)$ as

$$p(v|u) = \begin{cases} 1 - p & u = v, \\ p/2 & u = v \pm 1 \mod r, \\ 0 & \text{otherwise} \end{cases}$$

Then, $E(d(U, V)) = p$. Therefore, the rate distortion function is

$$R(D) = \min_{p \leq D} I(U; V).$$

Now, we know that

$$\begin{aligned} I(U; V) &= H(V) - H(V|U) \\ &= \log r - H(1 - p, p/2, p/2), \end{aligned}$$

since U is uniform, and due to symmetry, V is also uniform. We know that $H(1-p,p,p) \leq \log 3$, and this is achieved when $p = 2/3$. Therefore,

$$R(D) = \log r - \log 3, \quad \text{if } D > 2/3.$$

Now, let's consider the case when $D \leq 2/3$. Denote

$$\begin{aligned} f(p) = H(1-p, p/2, p/2) &= -(1-p)\log(1-p) - p/2 \log p/2 \times 2 \\ &= -(1-p)\log(1-p) - p\log p + p, \end{aligned}$$

We know that $f(p)$ is a concave function. By differentiating with respect to p,

$$\frac{df(p)}{dp} = \log(1-p) + 1 - \log p - 1 + 1 = \log \frac{1-p}{p} + 1$$

and setting $f(p) = 0$, $f(p)$ becomes maximum when $p = 2/3$. Therefore, if $D \leq 2/3$, $f(p)$ is an increasing function of p. Thus,

$$R(D) = \log 3 - D - h_2(D), \quad \text{if } D \leq 2/3.$$

.

7. **Adding a column to the distortion matrix**. Let $R(D)$ be the rate distortion function for an i.i.d. process with probability mass function $p(x)$ and distortion function $d(x, \hat{x})$, $x \in \mathcal{X}$, $\hat{x} \in \hat{\mathcal{X}}$. Now suppose that we add a new reproduction symbol $\hat{x}_0$ to $\hat{\mathcal{X}}$ with associated distortion $d(x, \hat{x}_0)$, $x \in \mathcal{X}$. Can this increase $R(D)$? Explain.

   Solution: Adding a column

   Let the new rate distortion function be denoted as $\tilde{R}(D)$, and note that we can still achieve $R(D)$ by restricting the support of $p(x, \hat{x})$, i.e., by simply ignoring the new symbol. Thus, $\tilde{R}(D) \leq R(D)$.

   Finally note the duality to the problem in which we added a row to the channel transition matrix to have no smaller capacity (Problem 7.22).

8. **Simplification**. Suppose $\mathcal{X} = \{1,2,3,4\}$, $\hat{\mathcal{X}} = \{1,2,3,4\}$, $p(i) = \frac{1}{4}$, $i = 1,2,3,4$, and $X_1, X_2, \ldots$ are i.i.d. $\sim p(x)$. The distortion matrix

$d(x, \hat{x})$ is given by

	1	2	3	4
1	0	0	1	1
2	0	0	1	1
3	1	1	0	0
4	1	1	0	0

(a) Find $R(0)$, the rate necessary to describe the process with zero distortion.

(b) Find the rate distortion function $R(D)$.
(*Hint*: The distortion measure allows to simplify the problem into one you have already seen.)

(c) Suppose we have a nonuniform distribution $p(i) = p_i$, $i = 1, 2, 3, 4$. What is $R(D)$?

Solution: Simplification

(a) We can achieve 0 distortion if we output $\hat{X} = 1$ if $X = 1$ or 2, and $\hat{X} = 3$ if $X = 3$ or 4. Thus if we set $Y = 1$ if $X = 1$ or 2, and $Y = 2$ if $X = 3$ or 4, we can recover Y exactly if the rate is greater than $H(Y) = 1$ bit. It is also not hard to see that any 0 distortion code would be able to recover Y exactly, and thus $R(0) = 1$.

(b) If we define Y as in the previous part, and $\hat{Y}$ similarly from $\hat{X}$, we can see that the distortion between X and $\hat{X}$ is equal to the Hamming distortion between Y and $\hat{Y}$. Therefore if the rate is greater than the Hamming rate distortion function $R(D)$ for Y, we can recover X to distortion D. Thus $R(D) = 1 - H(D)$.

(c) If the distribution of X is not uniform, the same arguments hold and Y has a distribution $(p_1 + p_2, p_3 + p_4)$, and the rate distortion function is $R(D) = H(p_1 + p_2) - H(D)$,

9. **Rate distortion for two independent sources.** Can one simultaneously compress two independent sources better than compressing the sources individually? The following problem addresses this question. Let the pair $\{(X_i, Y_i)\}$ be iid $\sim p(x, y)$. The distortion measure for X is $d(x, \hat{x})$ and its rate distortion function is $R_X(D)$. Similarly, the

distortion measure for Y is $d(y, \hat{y})$ and its rate distortion function is $R_Y(D)$.

Suppose we now wish to describe the process $\{(X_i, Y_i)\}$ subject to distortion constraints $\lim_{n\to\infty} Ed(X^n, \hat{X}^n) \leq D_1$ and $\lim_{n\to\infty} Ed(Y^n, \hat{Y}^n) \leq D_2$. Our rate distortion theorem can be shown to naturally extend to this setting and imply that the minimum rate required to achieve these distortion is given by

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

Now, suppose the $\{X_i\}$ process and the $\{Y_i\}$ process are independent of each other.

(a) Show
$$R_{X,Y}(D_1, D_2) \geq R_X(D_1) + R_Y(D_2).$$

(b) Does equality hold?

Now answer the question.

Solution: Rate distortion for two independent sources.

(a) Given that X and Y are independent, we have

$$p(x, y, \hat{x}, \hat{y}) = p(x)p(y)p(\hat{x}, \hat{y}|x, y) \tag{25}$$

Then

$$\begin{aligned}
I(X, Y; \hat{X}, \hat{Y}) &= H(X, Y) - H(X, Y|\hat{X}, \hat{Y}) && (26)\\
&= H(X) + H(Y) - H(X|\hat{X}, \hat{Y}) - H(Y|X, \hat{X}, \hat{Y}) && (27)\\
&\geq H(X) + H(Y) - H(X|\hat{X}) - H(Y|\hat{Y}) && (28)\\
&= I(X; \hat{X}) + I(Y; \hat{Y}) && (29)
\end{aligned}$$

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

$$\begin{aligned}
R_{X,Y}(D_1, D_2) &= \min_{p(\hat{x},\hat{y}|x,y):Ed(X,\hat{X})\leq D_1, Ed(Y,\hat{Y})\leq D_2} I(X, Y; \hat{X}, \hat{Y}) && (30)\\
&\geq \min_{p(\hat{x},\hat{y}|x,y):Ed(X,\hat{X})\leq D_1, Ed(Y,\hat{Y})\leq D_2} \left(I(X; \hat{X}) + I(Y; \hat{Y})\right) && (31)\\
&= \min_{p(\hat{x}|x):Ed(X,\hat{X})\leq D_1} I(X; \hat{X}) + \min_{p(\hat{y}|y):Ed(Y,\hat{Y})\leq D_2} I(Y; \hat{Y}) && (32)\\
&= R_X(D_1) + R_Y(D_2) && (33)
\end{aligned}$$

(b) If

$$p(x, y, \hat{x}, \hat{y}) = p(x)p(y)p(\hat{x}|x)p(\hat{y}|y), \tag{34}$$

then

$$\begin{aligned} I(X, Y; \hat{X}, \hat{Y}) &= H(X, Y) - H(X, Y|\hat{X}, \hat{Y}) && (35)\\ &= H(X) + H(Y) - H(X|\hat{X}, \hat{Y}) - H(Y|X, \hat{X}, \hat{Y}) && (36)\\ &= H(X) + H(Y) - H(X|\hat{X}) - H(Y|\hat{Y}) && (37)\\ &= I(X; \hat{X}) + I(Y; \hat{Y}) && (38) \end{aligned}$$

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

$$R_{X,Y}(D_1, D_2) = \min_{p(\hat{x},\hat{y}|x,y): Ed(X,\hat{X}) \le D_1, Ed(Y,\hat{Y}) \le D_2} I(X, Y; \hat{X}, \hat{Y}) = R_X(D_1) + R_Y(D_2) \tag{39}$$

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.

10. **One bit quantization of a single Gaussian random variable.** Let $X \sim Norm(0, \sigma^2)$ and let the distortion measure be squared error. Here we do not allow block descriptions. Show that the optimum reproduction points for 1 bit quantization are $\pm\sqrt{\frac{2}{\pi}}\sigma$, and that the expected distortion for 1 bit quantization is $\frac{\pi-2}{\pi}\sigma^2$.

    Compare this with the distortion rate bound $D = \sigma^2 2^{-2R}$ for $R = 1$.

    Solution: One bit quantization of a single Gaussian random variable

    With one bit quantization, the obvious reconstruction regions are the positive and negative real axes. The reconstruction point is the centroid

of each region. For example, for the positive real line, the centroid a is given as

$$\begin{aligned} a &= \int_0^\infty x \frac{2}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}\, dx \\ &= \int_0^\infty \sigma \sqrt{\frac{2}{\pi}} e^{-y}\, dy \\ &= \sigma\sqrt{\frac{2}{\pi}}, \end{aligned}$$

using the substitution $y = x^2/2\sigma^2$. The expected distortion for one bit quantization is

$$\begin{aligned} D &= \int_{-\infty}^0 \left(x + \sigma\sqrt{\frac{2}{\pi}} \right)^2 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}\, dx \\ &\quad + \int_0^\infty \left(x - \sigma\sqrt{\frac{2}{\pi}} \right)^2 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}\, dx \\ &= 2\int_{-\infty}^\infty \left(x^2 + \sigma^2 \frac{2}{\pi} \right) \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}\, dx \\ &\quad - 2\int_0^\infty \left(-2x\sigma\sqrt{\frac{2}{\pi}} \right) \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}\, dx \\ &= \sigma^2 + \frac{2}{\pi}\sigma^2 - 4\frac{1}{\sqrt{2\pi}}\sigma^2\sqrt{\frac{2}{\pi}} \\ &= \sigma^2 \frac{\pi - 2}{\pi} \\ &\approx .3634\,\sigma^2, \end{aligned}$$

which is much larger than the distortion rate bound $D(1) = \sigma^2/4$.

11. **Side information.**
A memoryless source generates i.i.d. pairs of random variables (U_i, S_i), $i = 1, 2, \ldots$ on finite alphabets, according to

$$p(u^n, s^n) = \prod_{i=1}^n p(u_i, s_i).$$

We are interested in describing U, when S, called *side information*, is known both at the encoder and at the decoder. Prove that the rate distortion function is given by

$$R_{U|S}(D) = \min_{p(v|u,s):E[d(U,V)]\leq D} I(U;V|S).$$

Compare $R_{U|S}(D)$ with the ordinary rate distortion function $R_U(D)$ without any side information. What can you say about the influence of the correlation between U and S on $R_{U|S}(D)$?

Solution: Side information

Since both encoder and decoder have the common side information S, we can be helped from the correlation of our source U and the side information S. The idea for the achievability is, we use the different rate distortion code for the source symbols that have different corresponding side information. Therefore, for the symbols that has side information $S = s$ will have the rate distortion code at rate

$$R_{U|S=s}(D) = \min_{p(v|u,S=s):E(d(U,V))\leq D} I(U;V|S=s)$$

and expected distortion $\leq D$. Now, the total rate will be

$$R = \sum_{s=1}^{|\mathcal{S}|} P_{S^n}(s) R_{U|S=s}(D),$$

where $P_{S^n}(s)$ is the empirical distribution of S^n. As n becomes large enough, we know that $P_{S^n}(s)$ will be very close to the true distribution of S, $p(s)$, or with high probability, S^n will be in the strong typical set $T_P(\delta)$, where

$$T_P(\delta) = \left\{ s^n : \max_{s\in\mathcal{S}} |P_{s^n}(s) - p(s)| < \frac{\delta}{|\mathcal{S}|} \right\}.$$

Therefore, if n sufficiently large, and δ sufficiently small, we will achieve the rate

$$R = \sum_{s\in\mathcal{S}} p(s) I(U;V|S=s) = I(U;V|S)$$

It is clear that the overall rate distortion code has the expected distortion $\leq D$.

For the converse, we have

$$\begin{aligned}
nR \geq & H(V^n|S^n) \\
= & I(U^n; V^n|S^n) \\
= & H(U^n|S^n) - H(U^n|V^n, S^n) \\
= & \sum_{i=1}^{n} \big(H(U_i|S_i) - H(U_i|U^{i-1}, V^n, S^n)\big) \\
\geq & \sum_{i=1}^{n} \left(H(U_i|S_i) - H(U_i|V_i)\right) \\
= & \sum_{i=1}^{n} I(U_i; V_i|S_i) \\
\geq & \sum_{i=1}^{n} R_{U|S}(Ed(U_i, V_i)) \\
\geq & R_{U|S}(Ed(U^n, V^n)) \\
\geq & R_{U|S}(D)
\end{aligned}$$

where each step has the identical reason as the original converse. Now, to compare $R_{U|S}(D)$ and $R_U(D)$, consider following:

$$\begin{aligned}
R_{U|S}(D) = & \min_{p(v|u,s):E(d(U,V))\leq D} I(U; V|S) \\
\leq & \min_{p(v|u):E(d(U,V))\leq D} I(U; V|S) \qquad (40) \\
\leq & \min_{p(v|u):E(d(U,V))\leq D} I(U; V) \qquad (41) \\
= & R_U(D).
\end{aligned}$$

Here, the inequality in (40) is from the fact that the minimization is held on the smaller set. The inequality in (41) is from the fact that $S \to U \to V$ forms a Markov chain with the joint distribution that we get from the minimization in (40). Also, we have,

$$\begin{aligned}
I(U, S; V) = & I(U; V) + I(S; V|U) \\
= & I(S; V) + I(U; V|S).
\end{aligned}$$

Since $I(S;V|U) = 0$ and $I(S;V) \geq 0$, we have the inequality in (41). Therefore, we can only gain from the the side information. Intuitively, this also makes sense, because we can always just ignore the side information and get the original rate distortion code, but the rate may be worse. The two rates will be equal if and only if U and S are independent, which achieves the equality in both (40) and (41).