1st Semester 2010/11

Homework Set #1
Sanov's Theorem, Typicality and 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
$$-\frac{1}{n}\log\Pr\left\{(X_1, X_2, \ldots, X_n) : p_{X^n} \geq p\right\} \to p\log\frac{p}{q}+(1-p)\log\frac{1-p}{1-q} = D((p,1-p)||(q,1-q)) \tag{3}$$
Justify the following steps:

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

2. **Strong Typicality**
   Let X^n be drawn i.i.d.$\sim P(x)$. Prove that for each $x^n \in T_\delta(X)$,
$$2^{-n(H(X)+\delta')} \le P^n(x^n) \le 2^{-n(H(X)-\delta')}$$
   for some $\delta' = \delta'(\delta)$ that vanishes as $\delta \to 0$.

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

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

5. **Erasure distortion**
   Consider $X \sim \text{Bernoulli}(\frac{1}{2})$, and let the distortion measure be given by the matrix
$$d(x, \hat{x}) = \begin{bmatrix} 0 & 1 & \infty \\ \infty & 1 & 0 \end{bmatrix}. \quad (5)$$

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?

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

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

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.

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 $R(D)$?

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.

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 $R = 1$.

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 $R_{U|S}(D)$?