EE 376B Handout #17
Information Theory Tuesday, May 19, 2009
Prof. T. Cover

Solutions to Homework Set #5

1. **Counting**
Let $\mathcal{X} = \{1, 2, \ldots, m\}$. Show that the number of sequences $x^n \in \mathcal{X}^n$ satisfying $\frac{1}{n}\sum_{i=1}^{n} g(x_i) \geq \alpha$ is approximately equal to 2^{nH^*}, to first order in the exponent, for n sufficiently large, where

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

Solution: Counting
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 $\frac{1}{n}\sum g(x_i) \geq \alpha$. Then by Sanov's theorem, we have

$$Q^n(E) \doteq 2^{-nD(P^*||Q)}, \tag{1}$$

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

$$H^* = \max_{P: \sum_{i=1}^{m} P(i)g(i) \geq \alpha} H(P), \tag{2}$$

we have

$$Q^n(E) \doteq 2^{-n(\log m - H^*)}. \tag{3}$$

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

$$|E| \doteq 2^{-n \log m} 2^{nH^*} m^n = 2^{nH^*}. \tag{4}$$

From theorem 11.1.1 we know that the maximizing entropy distribution P^* is given by

$$P^*(x) = e^{\lambda_0 + \lambda_1 g(x)},$$

where λ_0 and λ_1 are chosen to satisfy the condition $E\left(g(X)\right) \geq \alpha$ and that P^* sums to one.

2. **Counting states**
Suppose an atom is equally likely to be in each of 6 states, $X \in \{s_1.s_2, \ldots, s_6\}$. One observes n atoms $X_1, X_2, \ldots, X_n$ independently drawn according to this uniform distribution. It is observed that the frequency of occurrence of state s_1 is twice the frequency of occurrence of state s_2.

   (a) To the first order in the exponent, what is the probability of observing this event?

   (b) Assume n large, find the conditional distribution of the state of the first atom X_1, given this observation.

Solution: Counting states

(a) Let E be the set of probability distributions such that $E\left(g(X)\right) = 0$, where

$$g(X) = \begin{cases} 1 & \text{if } X = s_1 \\ -2 & \text{if } X = s_2 \\ 0 & \text{else.} \end{cases}$$

We want to find the distribution in E which minimizes $D(P||Q)$.
Define $P^* = \text{argmin}_{P \in \tilde{E}} D(P||Q)$. Then according to equation 11.110,

$$P^*(x) = \frac{Q(x)e^{\lambda g(x)}}{\sum_{i=1}^{6} Q(i)e^{\lambda g(i)}},$$

where λ is chosen such that $E(g(X)) = 0$. Plugging in the uniform distribution for $Q(x)$ we get

$$P^*(x) = \frac{e^{\lambda g(x)}}{\sum_{i=1}^{6} e^{\lambda g(i)}}.$$

Solving for λ gives $\lambda = \frac{\ln 2}{3}$. Therefore,

$$P^*(x) = \frac{2^{g(x)/3}}{4 + 2^{1/3} + 2^{-2/3}}. \tag{5}$$

This implies that $P^* = \text{argmin}_{P \in E} D(P||Q)$. Applying Sanov's theorem then gives

$$\Pr\left\{\text{Twice as many occurrences of } s_1 \text{ than } s_2 \text{ in } X^n\right\} \doteq 2^{-nD(P^*||Q)} = \left(\frac{1}{6}\right)^n 2^{nH(P^*)}.$$

(b) Applying the conditional limit theorem we see that the conditional distribution on X_1 given the observation is simply P^*, which we found in part (a).

3. **Sanov**
Let X_i be i.i.d. $\sim N(0, \sigma^2)$.

(a) Find the exponent in the behavior of $\Pr\left\{\frac{1}{n}\sum_{i=1}^{n} X_i^2 \geq \alpha^2\right\}$. This can be done from first principles (since the normal distribution is nice) or by using Sanov's theorem.

(b) What does the data look like if $\frac{1}{n}\sum_{i=1}^{n} X_i^2 \geq \alpha^2$? That is, what is the P^* that minimizes $D(P||Q)$ subject to the appropriate constraints?

Solution: Sanov

(a) • **Using Sanov's theorem:**
We will make use of the continuous version of Sanov's theorem. The first step is to find the P^* which minimizes $D(P||Q)$ subject to the constraint $E(X^2) \geq \alpha^2$, where Q is the $N(0, \sigma^2)$ distribution. Making use of equation 11.110 we get

$$\begin{aligned} P^*(x) &= \frac{Q(x)e^{\lambda x^2}}{\int Q(y)e^{\lambda y^2}dy} \\ &= Ce^{\frac{-x^2}{2\sigma^2}}e^{\lambda x^2} \\ &= Ce^{-\frac{1}{2}(\frac{1}{\sigma^2}-2\lambda)x^2}. \end{aligned}$$

So we see that P^* is the $N\left(0, \frac{\sigma^2}{1-2\lambda\sigma^2}\right)$ distribution. Since we have the constraints that $E(X^2) \geq \alpha^2$, we know that λ is chosen such that P^* is the $N(0, \alpha^2)$ distribution.
Therefore,

$$\begin{aligned} D(P^*||Q) &= \int P^*(x)\log\left(\frac{\sqrt{2\pi\sigma^2}}{\sqrt{2\pi\alpha^2}}\frac{e^{\frac{-x^2}{2\alpha^2}}}{e^{\frac{-x^2}{2\sigma^2}}}\right)dx \\ &= \int P^*(x)\log\left(\frac{\sigma}{\alpha}e^{\frac{-x^2}{2}(\frac{1}{\alpha^2}-\frac{1}{\sigma^2})}\right)dx \\ &= \log\left(\frac{\sigma}{\alpha}\right)\int P^*(x)dx + \int P^*(x)\log(e)\frac{-x^2}{2}\left(\frac{1}{\alpha^2}-\frac{1}{\sigma^2}\right)dx \\ &= \log\left(\frac{\sigma}{\alpha}\right) + \log(e)\left(\frac{1}{2\sigma^2}-\frac{1}{2\alpha^2}\right)\int x^2P^*(x)dx \\ &= \log\left(\frac{\sigma}{\alpha}\right) + \log(e)\left(\frac{\alpha^2}{2\sigma^2}-\frac{1}{2}\right). \end{aligned}$$

Applying Sanov's theorem we get

$$\begin{aligned}
\Pr\left\{\frac{1}{n}\sum_{i=1}^{n} X_i^2 \geq \alpha^2\right\} &\doteq 2^{-nD(P^*||Q)} \\
&= 2^{-n\log\left(\frac{\sigma}{\alpha}\right)-n\log(e)\left(\frac{\alpha^2}{2\sigma^2}-\frac{1}{2}\right)} \\
&= \left(\frac{\sigma}{\alpha}\right)^{-n} e^{-n\left(\frac{\alpha^2}{2\sigma^2}-\frac{1}{2}\right)}.
\end{aligned}$$

- **From first principles:**
Begin by observing that $\frac{1}{\sigma}\sum_{i=1}^{n} X_i^2$ has the Chi-square distribution with n degrees of freedom. We know that the cumulative distribution function of a Chi-squared random variable with m degrees of freedom is given by

$$1 - \frac{\Gamma(m/2, x/2)}{\Gamma(m/2)}$$

where $\Gamma(\cdot)$ is the gamma function and $\Gamma(\cdot,\cdot)$ is the upper incomplete gamma function,

$$\Gamma(m) = \int_0^\infty t^{m-1}e^{-t}dt$$

and

$$\Gamma(m,x) = \int_x^\infty t^{m-1}e^{-t}dt.$$

For m a positive integer,

$$\Gamma(m) = (m-1)!$$

and

$$\Gamma(m,x) = (m-1)!e^{-x}\sum_{i=0}^{m-1}\frac{x^i}{i!}.$$

Therefore

$$\Pr\left\{\frac{1}{\sigma^2}\sum_{i=1}^{m} X_i^2 \geq x\right\} = \frac{\Gamma(m/2, x/2)}{\Gamma(m/2)}.$$

With this information at our disposal, let us examine the probability in question,

$$\begin{aligned}
\Pr\left\{\frac{1}{n}\sum_{i=1}^{n} X_i^2 \geq \alpha^2\right\} &= \Pr\left\{\frac{1}{\sigma^2}\sum_{i=1}^{n} X_i^2 \geq \frac{n\alpha^2}{\sigma^2}\right\} \\
&= \frac{\Gamma\left(\frac{n}{2}, \frac{n\alpha^2}{2\sigma^2}\right)}{\Gamma\left(\frac{n}{2}\right)}.
\end{aligned}$$

(b) From the conditional limit theorem, we have that conditional distribution approaches P^*. From part (a), P^* is the $N(0,\alpha^2)$ distribution.

4. **Large deviations**
Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables drawn according to the geometric distribution

$$\Pr\{X = k\} = p^{k-1}(i-p), k = 1, 2, \ldots.$$

Find good estimates (to the first order in the exponent) of

(a) $\Pr\left\{\frac{1}{n}\sum_{i=1}^{n} X_i \geq \alpha\right\}$
(b) $\Pr\left\{X_1 = k \,\middle|\, \frac{1}{n}\sum_{i=1}^{n} X_i \geq \alpha\right\}$
(c) Evaluate (a) and (b) for $p = \frac{1}{2}$, $\alpha = 4$.

Solution: Large deviations

(a) Let E be the set of distributions such that $E(X) \geq \alpha$. Define $P^* = \text{argmin}_{P \in E} D(P||Q)$. Then according to equation 11.110,

$$P^*(x) = \frac{Q(x)e^{\lambda x}}{\sum_{i=1}^{\infty} Q(i)e^{\lambda i}},$$

where λ is chosen such that $E(X) \geq \alpha$ and Q is the geometric distribution with parameter p. Plugging in the geometric distribution we get

$$P^*(x) = \frac{p^{x-1}e^{\lambda x}}{\sum_{i=1}^{\infty} p^{i-1}e^{\lambda i}}.$$

Therefore

$$\begin{aligned} E(X) &= \sum_{x=1}^{\infty} x P^*(x) \\ &= \frac{e^{\lambda}}{(1-pe^{\lambda})^2} \frac{1-pe^{\lambda}}{e^{\lambda}} \\ &= \frac{1}{1-pe^{\lambda}}. \end{aligned}$$

Setting $E(X)$ equal to α and solving for λ we get $\lambda = \ln\left(\frac{\alpha-1}{p\alpha}\right)$. Therefore

$$\begin{aligned}
P^*(x) &= \frac{p^{x-1}\left(\frac{\alpha-1}{p\alpha}\right)^x}{\sum_{i=1}^{\infty} p^{i-1}\left(\frac{\alpha-1}{p\alpha}\right)^i} \\
&= p^{x-1}\left(\frac{\alpha-1}{p\alpha}\right)^x \frac{1-p\left(\frac{\alpha-1}{p\alpha}\right)}{\left(\frac{\alpha-1}{p\alpha}\right)} \\
&= p^{x-1}\left(\frac{\alpha-1}{p\alpha}\right)^{x-1}\left(1-p\left(\frac{\alpha-1}{p\alpha}\right)\right) \\
&= p^{x-1}\left(\frac{\alpha-1}{p\alpha}\right)^{x-1}\frac{1}{\alpha}\left(1-p\left(\frac{\alpha-1}{p\alpha}\right)\right) \\
&= \left(\frac{\alpha-1}{\alpha}\right)^{x-1}\frac{1}{\alpha} \\
&= \left(1-\frac{1}{\alpha}\right)^{x-1}\frac{1}{\alpha}.
\end{aligned}$$

Hence P^* is a geometric distribution with parameter $1 - 1/\alpha$. Observe that P^* is independent of the original geometric distribution parameter p. Therefore,

$$D(P^*||Q) = (\alpha-1)\log\left(\frac{\alpha-1}{\alpha p}\right) - \log(\alpha) - \log(1-p).$$

Using Sanov's theorem we get

$$\begin{aligned}
\Pr\left\{\frac{1}{n}\sum_{i=1}^{n} X_i \geq \alpha\right\} &\doteq 2^{-nD(P^*||Q)} \\
&\doteq \alpha^n(1-p)^n\left(\frac{\alpha p}{\alpha-1}\right)^{n(\alpha-1)}
\end{aligned}$$

(b) Applying the conditional limit theorem we see that the conditional distribution is simply the P^* distribution from part (a). Therefore

$$\Pr\left\{X_1 = k \,\middle|\, \frac{1}{n}\sum_{i=1}^{n} X_i \geq \alpha\right\} \doteq \frac{1}{\alpha}\left(1-\frac{1}{\alpha}\right)^{k-1}.$$

(c) For $p = \frac{1}{2}$ and $\alpha = 4$ we get

$$\begin{aligned}
\Pr\left\{\frac{1}{n}\sum_{i=1}^{n} X_i \geq 4\right\} &\doteq 2^n\left(\frac{2}{3}\right)^{3n} = \left(\frac{16}{27}\right)^n \\
\Pr\left\{X_1 = k \,\middle|\, \frac{1}{n}\sum_{i=1}^{n} X_i \geq 4\right\} &\doteq \frac{1}{4}\left(1-\frac{1}{4}\right)^{k-1}.
\end{aligned}$$

5. **A relation between $D(P \parallel Q)$ and Chi-square**
   Show that the χ^2 statistic

$$\chi^2 = \Sigma_x \frac{(P(x) - Q(x))^2}{Q(x)}$$

is (twice) the first term in the Taylor series expansion of $D(P \parallel Q)$ about Q. Thus $D(P \parallel Q) = \frac{1}{2}\chi^2 + \ldots$

Hint: Write $\frac{P}{Q} = 1 + \frac{P-Q}{Q}$ and expand the log.

Solution: A relation between $D(P \parallel Q)$ and Chi-square
There are many ways to expand $D(P||Q)$ in a Taylor series, but when we are expanding about $P = Q$, we must get a series in $P - Q$, whose coefficients depend on Q only. It is easy to get misled into forming another series expansion, so we will provide two alternative proofs of this result.

- Expanding the log.
  Writing $\frac{P}{Q} = 1 + \frac{P-Q}{Q} = 1 + \frac{\Delta}{Q}$, and $P = Q + \Delta$, we get

$$D(P||Q) = \int P \ln \frac{P}{Q} \tag{6}$$

$$= \int (Q + \Delta) \ln \left(1 + \frac{\Delta}{Q}\right) \tag{7}$$

$$= \int (Q + \Delta) \left(\frac{\Delta}{Q} - \frac{\Delta^2}{2Q^2} + \ldots\right) \tag{8}$$

$$= \int \Delta + \frac{\Delta^2}{Q} - \frac{\Delta^2}{2Q} + \ldots. \tag{9}$$

  The integral of the first term $\int \Delta = \int P - \int Q = 0$, and hence the first non-zero term in the expansion is

$$\frac{\Delta^2}{2Q} = \frac{\chi^2}{2}, \tag{10}$$

  which shows that locally around Q, $D(P||Q)$ behaves quadratically like χ^2.

- By differentiation.
  So construct the Taylor series expansion for a single variable function f as

$$f(x) = f(c) + f'(c)(x - c) + f''(c)\frac{(x - c)^2}{2} + \ldots \tag{11}$$

  We can do the same expansion for $D(P||Q)$ as a multivariate function of P around the point Q. We get, get

$$D(P||Q)_{P=Q} = 0, \tag{12}$$

$$\frac{\partial}{\partial P_i} D(P||Q)|_{P=Q} = (\ln \frac{P_i}{Q_i} + 1)|_{P=Q} = 1, \tag{13}$$

and

$$\frac{\partial^2}{\partial P_i \partial P_j} D(P||Q)|_{P=Q} = \left(\frac{1}{P_i}\right)|_{P=Q} = \frac{1}{Q_i}\delta(i=j). \tag{14}$$

Hence the Taylor series is

$$\begin{aligned} D(P||Q) &= 0 + \sum_i 1(P_i - Q_i) + \sum_i \frac{1}{Q_i}\frac{(P_i-Q_i)^2}{2} + \dots & (15)\\ &= \frac{1}{2}\chi^2 + \dots . & (16)\end{aligned}$$

and we get $\frac{\chi^2}{2}$ as the first non-zero term in the expansion.

6. *Stein's lemma.* Consider the two hypothesis test

$$H_1 : f = f_1 \quad \text{vs.} \quad H_2 : f = f_2$$

Find $D(f_1 \parallel f_2)$ if

(a) $f_i(x) = N(0, \sigma_i^2), i = 1, 2$

(b) $f_i(x) = \lambda_i e^{-\lambda_i x}, x \geq 0, i = 1, 2$

(c) $f_1(x)$ is the uniform density over the interval [0,1] and $f_2(x)$ is the uniform density over $[a, a+1]$. Assume $0 < a < 1$.

(d) f_1 corresponds to a fair coin and f_2 corresponds to a two-headed coin.

Solution.
Stein's lemma.

(a) $f_1 = \mathcal{N}(0, \sigma_1^2)$, $f_2 = \mathcal{N}(0, \sigma_2^2)$,

$$\begin{aligned} D(f_1||f_2) &= \int_{-\infty}^{\infty} f_1(x)\left[\frac{1}{2}\ln\frac{\sigma_2^2}{\sigma_1^2} - \left(\frac{x^2}{2\sigma_1^2} - \frac{x^2}{2\sigma_2^2}\right)\right] dx & (17)\\ &= \frac{1}{2}\left[\ln\frac{\sigma_2^2}{\sigma_1^2} + \frac{\sigma_1^2}{\sigma_2^2} - 1\right]. & (18)\end{aligned}$$

(b) $f_1 = \lambda_1 e^{-\lambda_1 x}$, $f_2 = \lambda_2 e^{-\lambda_2 x}$,

$$\begin{aligned} D(f_1||f_2) &= \int_0^{\infty} f_1(x)\left[\ln\frac{\lambda_1}{\lambda_2} - \lambda_1 x + \lambda_2 x\right] dx & (19)\\ &= \ln\frac{\lambda_1}{\lambda_2} + \frac{\lambda_2}{\lambda_1} - 1. & (20)\end{aligned}$$

(c) $f_1 = U[0,1],\ f_2 = U[a, a+1]$,

$$\begin{aligned} D(f_1||f_2) &= \int_0^1 f_1 \ln \frac{f_1}{f_2} && (21)\\ &= \int_0^a f_1 \ln \infty + \int_a^1 f_1 \ln 1 && (22)\\ &= \infty. && (23)\end{aligned}$$

In this case, the Kullback Leibler distance of ∞ implies that in a hypothesis test, with high probability the sequence will reveal the correct hypothesis with certainty.

(d) $f_1 = \text{Bern}\left(\frac{1}{2}\right)$ and $f_2 = \text{Bern}(1)$,

$$D(f_1||f_2) = \frac{1}{2}\ln\frac{\frac{1}{2}}{1} + \frac{1}{2}\ln\frac{\frac{1}{2}}{0} = \infty. \tag{24}$$

The implication is the same as in part (c).

7. *Error exponent for universal codes.* A universal source code of rate R achieves a probability of error $P_e^{(n)} \doteq e^{-nD(P^*\|Q)}$, where Q is the true distribution and P^* achieves $\min D(P \parallel Q)$ over all P such that $H(P) \geq R$.

(a) Find P^* in terms of Q and R.

(b) Now let X be binary. Find the region of source probabilities $Q(x),\ x \in \{0,1\}$, for which rate R is sufficient for the universal source code to achieve $P_e^{(n)} \to 0$.

Solution.
Error exponent for universal codes.

(a) We have to minimize $D(p||q)$ subject to the constraint that $H(p) \geq R$. Rewriting this problem using Lagrange multipliers, we get

$$J(p) = \sum p \log \frac{p}{q} + l \sum p \log p + \nu \sum p. \tag{25}$$

Differentiating with respect to $p(x)$ and setting the derivative to 0, we obtain

$$\log \frac{p}{q} + 1 + l \log p + l + \nu = 0, \tag{26}$$

which implies that

$$p^*(x) = \frac{q^\mu(x)}{\sum_a q^\mu(a)}. \tag{27}$$

where $\mu = \frac{1}{1+l}$ is chosen to satisfy the constraint $H(p^*) = R$. We have to first check that the constraint is active, i.e., that we really need equality in the constraint. For this we set $l = 0$ or $\mu = 1$, and we get $p^* = q$. Hence if q is such that $H(q) \geq R$, then the maximizing p^* is q. On the other hand, if $H(q) < R$, then $l \neq 0$, and the constraint must be satisfied with equality.

Geometrically it is clear that there will be two solutions for l of the form (27) which have $H(p^*) = R$, corresponding to the minimum and maximum distance to q on the manifold $H(p) = R$. It is easy to see that for $0 \leq \mu \leq 1$, $p^*_\mu(x)$ lies on the geodesic from q to the uniform distribution. Hence, the minimum will lie in this region of μ. The maximum will correspond to negative μ, which lies on the other side of the uniform distribution as in the figure.

(b) For a universal code with rate R, any source can be transmitted by the code if $H(p) < R$. In the binary case, this corresponds to $p \in [0, h^{-1}(R))$ or $p \in (1 - h^{-1}(R), 1]$, where h is the binary entropy function.