EE 376A
Information Theory
Prof. T. Cover

Handout #28
Tuesday, March 3, 2009

Solutions to Homework Set #7

1. **Postprocessing the output.**
   One is given a communication channel with transition probabilities $p(y \mid x)$ and channel capacity $C = \max_{p(x)} I(X;Y)$. A helpful statistician postprocesses the output by forming $\tilde{Y} = g(Y)$, yielding a channel $p(\tilde{y}|x)$. He claims that this will strictly improve the capacity.

   (a) Show that he is wrong.

   (b) Under what conditions does he not strictly decrease the capacity?

Solution: Postprocessing the output.

(a) The statistician calculates $\tilde{Y} = g(Y)$. Since $X \to Y \to \tilde{Y}$ forms a Markov chain, we can apply the data processing inequality. Hence for every distribution on x,

$$I(X;Y) \geq I(X;\tilde{Y}).$$

Let $\tilde{p}(x)$ be the distribution on x that maximizes $I(X;\tilde{Y})$. Then

$$C = \max_{p(x)} I(X;Y) \geq I(X;Y)_{p(x)=\tilde{p}(x)} \geq I(X;\tilde{Y})_{p(x)=\tilde{p}(x)} = \max_{p(x)} I(X;\tilde{Y}) = \tilde{C}.$$

Thus, the helpful suggestion is wrong and processing the output does not increase capacity.

(b) We have equality (no decrease in capacity) in the above sequence of inequalities only if we have equality in data processing inequality, i.e., for the distribution that maximizes $I(X;\tilde{Y})$, we have $X \to \tilde{Y} \to Y$ forming a Markov chain. Thus, $\tilde{Y}$ should be a sufficient statistic.

2. **Noisy typewriter.**
   Consider a 26-key typewriter.

(a) If pushing a key results in printing the associated letter, what is the capacity C in bits?

(b) Now suppose that pushing a key results in printing that letter or the next (with equal probability). Thus $A \to A$ or B, and $Z \to Z$ or A. What is the capacity?

(c) What is the highest rate code with block length one that you can find that achieves *zero* probability of error for the channel in part (b) .

Solution: Noisy typewriter.

(a) If the typewriter prints out whatever key is struck, then the output Y is the same as the input X and

$$C = \max I(X;Y) = \max H(X) = \log 26, \tag{1}$$

attained by a uniform distribution over the letters.

(b) In this case, the output is either equal to the input (with probability $\frac{1}{2}$) or equal to the next letter (with probability $\frac{1}{2}$). Hence $H(Y|X) = \log 2$ independent of the distribution of X, and hence

$$C = \max I(X;Y) = \max H(Y) - \log 2 = \log 26 - \log 2 = \log 13, \tag{2}$$

which is attained for a uniform distribution over the output, which in turn is attained by a uniform distribution on the input.

(c) A simple zero error block length one code is the one that uses every alternate letter, say A,C,E,...,W,Y. In this case, none of the codewords will be confused, since A will produce either A or B, C will produce C or D, etc. The rate of this code,

$$R = \frac{\log(\#\text{ codewords})}{\text{Block length}} = \frac{\log 13}{1} = \log 13. \tag{3}$$

In this case, we can achieve capacity with a simple code with zero error. Note that the uniform distribution over the output is attained also by this input distribution.

3. **The Z Channel.**
The Z-channel has binary input and output alphabets and transition probabilities $p(y|x)$ given by the following matrix:

$$Q = \begin{bmatrix} 1 & 0 \\ 1/2 & 1/2 \end{bmatrix} \qquad x, y \in \{0, 1\}$$

Find the capacity of the Z-channel and the maximizing input probability distribution.

Solution: The Z channel.
First we express $I(X;Y)$, the mutual information between the input and output of the Z-channel, as a function of $\alpha = \Pr(X=1)$:

$$\begin{aligned} H(Y|X) &= \Pr(X=0)\cdot 0 + \Pr(X=1)\cdot 1 = \alpha \\ H(Y) &= \mathbf{H}(\Pr(Y=1)) = \mathbf{H}(\alpha/2) \\ I(X;Y) &= H(Y) - H(Y|X) = \mathbf{H}(\alpha/2) - \alpha \end{aligned}$$

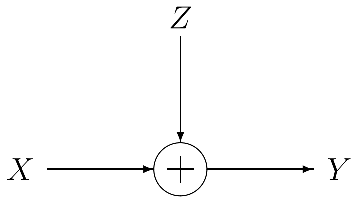
Since $I(X;Y)$ is strictly concave on α (why?) and $I(X;Y) = 0$ when $\alpha = 0$ and $\alpha = 1$, the maximum mutual information is obtained for some value of α such that $0 < \alpha < 1$.

Using elementary calculus, we determine that

$$\frac{d}{d\alpha} I(X;Y) = \frac{1}{2}\log_2 \frac{1-\alpha/2}{\alpha/2} - 1\,,$$

which is equal to zero for $\alpha = 2/5$. (It is reasonable that $\Pr(X=1) < 1/2$ since $X = 1$ is the noisy input to the channel.) So the capacity of the Z-channel in bits is $H(1/5) - 2/5 = 0.722 - 0.4 = 0.322$.

4. **An additive noise channel.**
Find the channel capacity of the following discrete memoryless channel:



where $\Pr\{Z=0\} = \Pr\{Z=a\} = \frac{1}{2}$. The alphabet for x is $\mathcal{X} = \{0, 1\}$. $Y = X + Z$ with real addition. Assume that Z is independent of X.

Observe that the channel capacity depends on the value of a.

Solution: An additive noise channel.

$$Y = X + Z \qquad X \in \{0,1\}, \quad Z \in \{0, a\}$$

We have to distinguish various cases depending on the values of a.

$a = 0$ In this case, $Y = X$, and $\max I(X;Y) = \max H(X) = 1$. Hence the capacity is 1 bit per transmission.

$a \neq 0, \pm 1$ In this case, Y has four possible values $0, 1, a$ and $1 + a$. Knowing Y, we know the X which was sent, and hence $H(X|Y) = 0$. Hence $\max I(X;Y) = \max H(X) = 1$, which is achieved for an uniform distribution on the input X.

$a = 1$ In this case Y has three possible output values, $0, 1$ and 2, and the channel is identical to the binary erasure channel discussed in class, with $\alpha = 1/2$. As derived in class, the capacity of this channel is $1 - \alpha = 1/2$ bit per transmission.

$a = -1$ This is similar to the case when $a = 1$ and the capacity here is also $1/2$ bit per transmission.

5. **Channels with memory have higher capacity.**
Consider a binary symmetric channel with $Y_i = X_i \oplus Z_i$, where $\oplus$ is mod 2 addition, and $X_i, Y_i \in \{0, 1\}$.

Suppose that $\{Z_i\}$ has constant marginal probabilities $P\{Z_i = 1\} = p = 1 - P\{Z_i = 0\}$, but that $Z_1, Z_2, \ldots, Z_n$ are not necessarily independent. Let $C = 1 - H(p, 1-p)$. Show that $\max\limits_{p(x_1, x_2, \ldots, x_n)} I(X_1, X_2, \ldots, X_n; Y_1, Y_2, \ldots, Y_n) \geq nC$. Comment on the implications.

Solution: Channels with memory have higher capacity.

$$Y_i = X_i \oplus Z_i,$$

where

$$Z_i = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}$$

and Z_i are not independent.

When $X_1, X_2, \ldots, X_n$ are chosen i.i.d. $\sim$ Bern($\frac{1}{2}$),

$$\begin{aligned}
& I(X_1, X_2, \ldots, X_n; Y_1, Y_2, \ldots, Y_n) \\
&\quad = H(X_1, X_2, \ldots, X_n) - H(X_1, X_2, \ldots, X_n | Y_1, Y_2, \ldots, Y_n) \\
&\quad = H(X_1, X_2, \ldots, X_n) - H(Z_1, Z_2, \ldots, Z_n | Y_1, Y_2, \ldots, Y_n) \\
&\quad \geq H(X_1, X_2, \ldots, X_n) - H(Z_1, Z_2, \ldots, Z_n) \\
&\quad \geq H(X_1, X_2, \ldots, X_n) - \sum H(Z_i) \\
&\quad = n - nH(p).
\end{aligned}$$

Hence, the capacity of the channel with memory over n uses of the channel is

$$\begin{aligned}
nC^{(n)} &= \max_{p(X_1, X_2, \ldots, X_n)} I(X_1, X_2, \ldots, X_n; Y_1, Y_2, \ldots, Y_n) \\
&\geq I(X_1, X_2, \ldots, X_n; Y_1, Y_2, \ldots, Y_n)_{p(x_1, x_2, \ldots, x_n) = \text{Bern}(\frac{1}{2})} \\
&\geq n(1 - H(p)) \\
&= nC.
\end{aligned}$$

Hence, channels with memory have higher capacity. The intuitive explanation for this result is that the correlation between the noise decreases the effective noise; one could use the information from the past samples of the noise to combat the present noise.

6. **Channel capacity.**
Consider the channel $Y = X + Z$ (mod 13), where

$$Z = \begin{pmatrix} 1, & 2, & 3, & 4 \\ \frac{1}{4}, & \frac{1}{4}, & \frac{1}{4}, & \frac{1}{4} \end{pmatrix}$$

and $X \in \{0, 1, \ldots, 12\}$.

(a) Find the capacity.

(b) What is the maximizing $p^*(x)$?

Solution: Channel capacity.

$$Y = X + Z \pmod{13}$$

where

$$Z = \begin{cases} 1 & \text{with probability } 1/4 \\ 2 & \text{with probability } 1/4 \\ 3 & \text{with probability } 1/4 \\ 4 & \text{with probability } 1/4 \end{cases} .$$

Then,

$$H(Y|X) = H(Z|X) = H(Z) = \log 4,$$

which is independent of the distribution of X. Hence the capacity of the channel is

$$\begin{aligned} C &= \max_{p(x)} I(X;Y) \\ &= \max_{p(x)} H(Y) - H(Y|X) \\ &= \max_{p(x)} H(Y) - \log 4 \\ &= \log 13 - \log 4, \end{aligned}$$

which is attained when Y has a uniform distribution, which occurs (by symmetry) when X has a uniform distribution.

(a) The capacity of the channel is $\log \frac{13}{4}$ bits/transmission.

(b) The capacity is achieved by a uniform distribution on the inputs, that is,

$$p(X = i) = \frac{1}{13} \qquad \text{for } i = 0, 1, \ldots, 12.$$

7. **Using two channels at once.**
Consider two discrete memoryless channels $(\mathcal{X}_1, p(y_1 \mid x_1), \mathcal{Y}_1)$ and $(\mathcal{X}_2, p(y_2 \mid x_2), \mathcal{Y}_2)$ with capacities C_1 and C_2 respectively. A new channel $(\mathcal{X}_1 \times \mathcal{X}_2, p(y_1 \mid x_1) \times p(y_2 \mid x_2), \mathcal{Y}_1 \times \mathcal{Y}_2)$ is formed in which $x_1 \in \mathcal{X}_1$ and $x_2 \in \mathcal{X}_2$, are *simultaneously* sent, resulting in y_1, y_2. Find the mutual information maximizing $p^*(x_1, x_2)$ in terms of the individual maximizing distributions $p^*(x_1)$ and $p^*(x_2)$. Find the capacity of this channel.

Solution: Using two channels at once.
To find the capacity of the product channel $(\mathcal{X}_1 \times \mathcal{X}_2, p(y_1, y_2|x_1, x_2), \mathcal{Y}_1 \times \mathcal{Y}_2)$, we have to find the distribution $p(x_1, x_2)$ on the input alphabet $\mathcal{X}_1 \times \mathcal{X}_2$ that maximizes $I(X_1, X_2; Y_1, Y_2)$. Since the transition probabilities are given as $p(y_1, y_2|x_1, x_2) = p(y_1|x_1)p(y_2|x_2)$,

$$\begin{aligned} p(x_1, x_2, y_1, y_2) &= p(x_1, x_2)p(y_1, y_2|x_1, x_2) \\ &= p(x_1, x_2)p(y_1|x_1)p(y_2|x_2), \end{aligned}$$

Therefore, $Y_1 \to X_1 \to X_2 \to Y_2$ forms a Markov chain and

$$\begin{align*}
I(X_1, X_2; Y_1, Y_2) &= H(Y_1, Y_2) - H(Y_1, Y_2|X_1, X_2) \\
&= H(Y_1, Y_2) - H(Y_1|X_1, X_2) - H(Y_2|X_1, X_2) \tag{4} \\
&= H(Y_1, Y_2) - H(Y_1|X_1) - H(Y_2|X_2) \tag{5} \\
&\leq H(Y_1) + H(Y_2) - H(Y_1|X_1) - H(Y_2|X_2) \tag{6} \\
&= I(X_1; Y_1) + I(X_2; Y_2),
\end{align*}$$

where Eqs. (4) and (5) follow from Markovity, and Eq. (6) is met with equality if X_1 and X_2 are independent and hence Y_1 and Y_2 are independent. Therefore

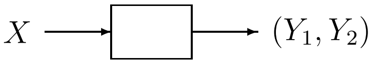
$$\begin{align*}
C &= \max_{p(x_1,x_2)} I(X_1, X_2; Y_1, Y_2) \\
&\leq \max_{p(x_1,x_2)} I(X_1; Y_1) + \max_{p(x_1,x_2)} I(X_2; Y_2) \\
&= \max_{p(x_1)} I(X_1; Y_1) + \max_{p(x_2)} I(X_2; Y_2) \\
&= C_1 + C_2.
\end{align*}$$

with equality iff $p(x_1, x_2) = p^*(x_1)p^*(x_2)$ and $p^*(x_1)$ and $p^*(x_2)$ are the distributions that maximize C_1 and C_2 respectively.

8. **A channel with two independent looks at Y.**
Let Y_1 and Y_2 be conditionally independent and conditionally identically distributed given X. Thus $p(y_1, y_2|x) = p(y_1|x)p(y_2|x)$.

   (a) Show $I(X; Y_1, Y_2) = 2I(X; Y_1) - I(Y_1; Y_2)$.

   (b) Conclude that the capacity of the channel



   is less than twice the capacity of the channel

   (c) How about 3 independent looks? Compare $I(X; Y_1, Y_2, Y_3)$ to $3I(X; Y_1)$.

Solution: A channel with two independent looks at Y.

(a) First note that Y_1 and Y_2 are identically distributed since $p(y_1|x) = p(y_2|x)$ and hence $p(y_1) = \sum_x p(y_1|x)p(x) = \sum_x p(y_2|x)p(x) = p(y_2)$. Therefore,

$$\begin{aligned} I(X;Y_1,Y_2) &= H(Y_1,Y_2) - H(Y_1,Y_2|X) \\ &= H(Y_1) + H(Y_2) - I(Y_1;Y_2) - H(Y_1,Y_2|X) \\ &= H(Y_1) + H(Y_2) - I(Y_1;Y_2) - H(Y_1|X) - H(Y_2|X) \qquad (7) \\ &= 2H(Y_1) - 2H(Y_1|X) - I(Y_1;Y_2) \qquad (8) \\ &= 2I(X;Y_1) - I(Y_1;Y_2), \end{aligned}$$

where Eq. (7) follows from the fact that Y_1 and Y_2 are conditionally independent given X and Eq. (8) follows from the fact that Y_1 and Y_2 are identically distributed and conditionally identically distributed given X.

(b) The capacity of the single look channel $X \to Y_1$ is

$$C_1 = \max_{p(x)} I(X;Y_1).$$

The capacity of the channel $X \to (Y_1, Y_2)$ is

$$\begin{aligned} C_2 &= \max_{p(x)} I(X;Y_1,Y_2) \\ &= \max_{p(x)} 2I(X;Y_1) - I(Y_1;Y_2) \\ &\le \max_{p(x)} 2I(X;Y_1) \\ &= 2C_1. \end{aligned}$$

Hence, two independent looks cannot be more than twice as good as one look.

(c) Observe for Y^n conditionally independent given X that

$$\begin{aligned} I(X;Y_k|Y^{k-1}) &= H(Y_k|Y^{k-1}) - H(Y_k|Y^{k-1},X) \\ &= H(Y_k|Y^{k-1}) - H(Y_k|X) \qquad (9) \\ &\le H(Y_k) - H(Y_k|X) \qquad (10) \\ &= I(X;Y_k) \end{aligned}$$

where Eq. (9) follows from the fact that Y_k and Y^{k-1} are conditionally independent given X and Eq. (10) follows from the fact that conditioning reduces entropy.

Using the above relationship,

$$\begin{aligned} I(X;Y_1,Y_2,Y_3) &= I(X;Y_1) + I(X;Y_2|Y_1) + I(X;Y_3|Y_1,Y_2) \\ &\le I(X;Y_1) + I(X;Y_2) + I(X;Y_3) \qquad (11) \\ &= 3I(X;Y_1) \qquad (12) \end{aligned}$$

where Eq. (11) follows from the relationship shown above and Eq. (12) follows from the fact that Y_1, Y_2 and Y_3 are identically distributed.

Thus it can be shown that the capacity C_3 of three looks is less than three times the capacity C_1 of one look, $C_3 < 3C_1$.

9. **Can signal alternatives lower capacity?**
Show that adding a row to a channel transition matrix does not decrease capacity.

Solution: Can signal alternatives lower capacity?
Adding a row to the channel transition matrix is equivalent to adding a symbol to the input alphabet $\mathcal{X}$. Suppose there were m symbols and we add an $(m+1)$st. We can always choose not to use this extra symbol.

Alternatively, let C_m and C_{m+1} denote the capacity of the original channel and the new channel, respectively. Since using the original m symbols is same as using $m+1$ symbols with the newly added symbol assigned zero probablity,

$$\begin{aligned} C_{m+1} &= \max_{p(x_1,\ldots,x_{m+1})} I(X;Y) \\ &\geq \max_{p(x_1,\ldots,x_m,0)} I(X;Y) \\ &= C_m. \end{aligned}$$