CMPSCI 691GG Applied Information Theory **Fall 2006**

Problem Set 3: Solutions

1. **[Cover and Thomas 7.1]**

   (a) Define the following notation,

$$\begin{aligned} C &= I_{p^*(x)}(X;Y) \\ &= \max_{p(x)} I_{p(x)}(X;Y) \\ \tilde{C} &= I_{\tilde{p}^*(x)}(X;\tilde{Y}) \\ &= \max_{p(x)} I_{p(x)}(X;\tilde{Y}) \end{aligned}$$

   We would like to show that $\tilde{C} = I_{\tilde{p}^*(x)}(X;\tilde{Y}) \leq I_{p^*(x)}(X;Y) = C$.

   Notice that X, Y, and $\tilde{Y}$ form a Markov chain such that $X \to Y \to \tilde{Y}$. Using the data-processing inequality (Theorem 2.8.1), we know that,

$$\begin{aligned} I_{\tilde{p}^*(x)}(X;\tilde{Y}) &\leq I_{\tilde{p}^*(x)}(X;Y) && (3.1) \\ &\leq I_{p^*(x)}(X;Y) && (3.2) \end{aligned}$$

   (b) We would like to determine under what conditions the following equality holds. Given our result in Equation 3.2, it is sufficient to show,

$$I_{\tilde{p}^*(x)}(X;\tilde{Y}) \geq I_{p^*(x)}(X;Y)$$

   We know that the following equality is true for Markov chains (see proof of Theorem 2.8.1),

$$I_{p^*(x)}(X;\tilde{Y}) = I_{p^*(x)}(X;Y) - I_{p^*(x)}(X;Y|\tilde{Y})$$

   However, $\tilde{p}^*(x)$ and $p^*(x)$ may not be the same distribution, so

$$\begin{aligned} I_{\tilde{p}^*(x)}(X;\tilde{Y}) &\geq I_{p^*(x)}(X;\tilde{Y}) && (3.3) \\ &= I_{p^*(x)}(X;Y) - I_{p^*(x)}(X;Y|\tilde{Y}) && (3.4) \end{aligned}$$

   We can show our objective inequality if $I_{p^*(x)}(X;Y|\tilde{Y}) = 0$. This occurs if $\tilde{Y} = g(Y)$ is an injective function.

2. **[Cover and Thomas 7.2]**

   Consider the behavior of this channel as depicted in Figure 3.1.

   When $|a| \neq 1$, this is a Noisy Channel with Nonoverlapping Outputs. We would like to compute the capacity of the channel in this situation,

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

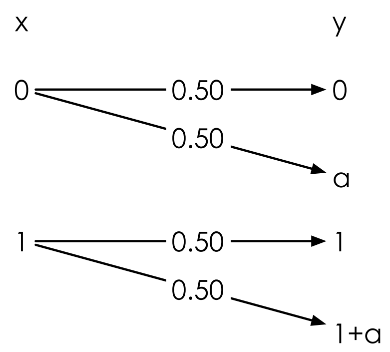

Figure 3.1: Noisy channel model for question 7.2.

Because X can be determined by Y, $H(X|Y) = 0$. Therefore,

$$\begin{aligned} C &= \max_{p(x)} H(X) \\ &= 1 \text{ bit} \end{aligned}$$

When $|a| = 1$, this is a Binary Erasure Channel. We will compute the capacity for $a = 1$. We begin by defining the conditional entropy,

$$\begin{aligned} H(X|Y) &= \sum_{y \in Y} P(Y = y) H(X|Y = y) \\ &= \frac{1}{4} H(X|Y = 0) + \frac{1}{2} H(X|Y = 1) + \frac{1}{4} H(X|Y = 2) \\ &= \frac{1}{2} H(X) \end{aligned}$$

We can now compute the capacity,

$$\begin{aligned} C &= \max_{p(x)} I(X;Y) \\ &= \max_{p(x)} H(X) - H(X|Y) \\ &= \max_{p(x)} H(X) - \frac{1}{2} H(X) \\ &= \max_{p(x)} \frac{1}{2} H(X) \\ &= \frac{1}{2} \max_{p(x)} H(X) \\ &= \frac{1}{2} \text{ bit} \end{aligned}$$

The computation for $a = -1$ is similar

3. **[Cover and Thomas 7.3]**

$$I(\vec{X}; \vec{Y}) = H(\vec{X}) - H(\vec{X}|\vec{Y})$$

However, because this is a binary symmetric channel, the uncertainty about X_i and Z_i is equivalent given Y_i. We can replace $\vec{X}$ with $\vec{Z}$,

$$I(\vec{X};\vec{Y}) = H(\vec{X}) - H(\vec{Z}|\vec{Y}) \tag{3.5}$$

Now we will derive a bound for $H(\vec{Z}|\vec{Y})$ using properties of conditional entropy (Theorems 2.6.5 and 2.6.6).

$$\begin{aligned} H(\vec{Z}|\vec{Y}) &\leq H(\vec{Z}) \\ &\leq \sum_{i=1}^{n} H(Z_i) \\ &= nH(p) \end{aligned}$$

Replacing $H(\vec{Z}|\vec{Y})$ with $nH(p)$ in Equation 3.5 will reduce the right hand side. This gives us the following inequality.

$$I(\vec{X};\vec{Y}) \geq H(\vec{X}) - nH(p) \tag{3.6}$$

Define the following notation,

$$H_{\tilde{p}^*(x)}(\vec{X}) - nH(p) = \max_{p(x)} H_{p(x)}(\vec{X}) - nH(p)$$

This defines the maximum value of the right hand side of Equation 3.6. Assuming that $H(\vec{X}) = \sum H(X_i)$, the maximizing distribution, $\tilde{p}^*(x)$, is uniform. This means that

$$\begin{aligned} H_{\tilde{p}^*(x)}(\vec{X}) - nH(p) &= n - nH(p) \\ &= n(1 - H(p)) \\ &= nC \end{aligned}$$

We are interested in the capacity of the channel with memory.

$$\max_{p(\vec{x})} I_{p(\vec{x})}(\vec{X};\vec{Y}) \geq nC$$

4. **[Cover and Thomas 7.8]**

We define our set of distributions as,

$$\begin{aligned} p(x) &= \begin{bmatrix} 1-\lambda \\ \lambda \end{bmatrix} \\ p(y|x) &= \begin{bmatrix} 1 & 0 \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} \\ p(y) &= \begin{bmatrix} 1-\lambda \\ \lambda \end{bmatrix}^T \times \begin{bmatrix} 1 & 0 \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} \\ &= \begin{bmatrix} 1-\frac{\lambda}{2} & \frac{\lambda}{2} \end{bmatrix} \end{aligned}$$

First, we compute the entropy, $H(Y)$,

$$\begin{aligned} H(Y) &= -1 \times \begin{bmatrix} 1-\frac{\lambda}{2} & \frac{\lambda}{2} \end{bmatrix} \times \log \begin{bmatrix} 1-\frac{\lambda}{2} \\ \frac{\lambda}{2} \end{bmatrix} \\ &= -1 \times \left(\left(1 - \frac{\lambda}{2}\right) \log\left(1 - \frac{\lambda}{2}\right) + \left(\frac{\lambda}{2}\right) \log\left(\frac{\lambda}{2}\right) \right) \end{aligned}$$

Next, we compute the conditional entropy, $H(Y|X)$,

$$
\begin{array}{rcccccccc}
H(Y|X) & = & p(X=0) & H(Y|X=0) & + & p(X=1) & H(Y|X=1) \\
 & = & (1-\lambda) & H(\begin{bmatrix} 1 & 0 \end{bmatrix}) & + & \lambda & H(\begin{bmatrix} \frac{1}{2} & \frac{1}{2} \end{bmatrix}) \\
 & = & (1-\lambda) & 0 & + & \lambda & 1 \\
 & = & & & \lambda & &
\end{array}
$$

We can use $H(Y)$ and $H(Y|X)$ to compute the capacity of the channel,

$$
\begin{aligned}
C &= \max_{p(x)} I(X;Y) \\
&= \max_{p(x)} H(Y) - H(Y|X) \\
&= \max_{p(x)} -1 \times \left(\left(1 - \frac{\lambda}{2}\right) \log\left(1 - \frac{\lambda}{2}\right) + \left(\frac{\lambda}{2}\right) \log\left(\frac{\lambda}{2}\right) \right) - \lambda
\end{aligned}
$$

Notice that we are maximizing a function of λ,

$$
f(\lambda) = -1 \times \left(\left(1 - \frac{\lambda}{2}\right) \log\left(1 - \frac{\lambda}{2}\right) + \left(\frac{\lambda}{2}\right) \log\left(\frac{\lambda}{2}\right) \right) - \lambda
$$

To find the maximum of this function, we differentiate with respect to *lambda*,

$$
\begin{aligned}
\frac{df}{d\lambda} &= -\left(1 - \frac{\lambda}{2}\right) \times \frac{1}{1 - \frac{\lambda}{2}} \times \left(-\frac{1}{2}\right) \\
&\quad - \left(-\frac{1}{2}\right) \times \log\left(1 - \frac{\lambda}{2}\right) \\
&\quad - \frac{\lambda}{2} \times \frac{1}{\frac{\lambda}{2}} \times \frac{1}{2} \\
&\quad - \frac{1}{2} \times \log\left(\frac{\lambda}{2}\right) \\
&\quad - 1 \\
&= \frac{1}{2} \log\left(1 - \frac{\lambda}{2}\right) - \frac{1}{2} \log\left(\frac{\lambda}{2}\right) - 1
\end{aligned}
$$

Setting this to zero, we can derive the maximum,

$$
\begin{aligned}
\frac{1}{2} \log\left(1 - \frac{\lambda}{2}\right) - \frac{1}{2} \log\left(\frac{\lambda}{2}\right) - 1 = 0 \\
\lambda = \frac{2}{5} \\
f(\lambda) = \log 5 - 2 \text{ bits} \\
\approx 0.3219 \text{ bits}
\end{aligned}
$$

5. **[Cover and Thomas 7.13]**

(a) Given the following distributions,

$$\begin{aligned}
p(x) &= \begin{bmatrix} 1-\lambda \\ \lambda \end{bmatrix} \\
p(y|x) &= \begin{bmatrix} 1-\alpha-\epsilon & \epsilon & \alpha \\ \epsilon & 1-\alpha-\epsilon & \alpha \end{bmatrix} \\
p(y) &= \begin{bmatrix} 1-\lambda \\ \lambda \end{bmatrix}^T \begin{bmatrix} 1-\alpha-\epsilon & \epsilon & \alpha \\ \epsilon & 1-\alpha-\epsilon & \alpha \end{bmatrix} \\
&= \begin{bmatrix} 1-\alpha-\epsilon+\alpha\lambda+2\epsilon\lambda-\lambda \\ \epsilon-2\epsilon\lambda+\lambda-\alpha\lambda \\ \alpha \end{bmatrix}^T
\end{aligned}$$

We would like to compute the capacity of this channel,

$$\begin{aligned}
C &= \max_\lambda I(X;Y) \\
&= \max_\lambda H(Y) - H(Y|X)
\end{aligned}$$

However, we can show that $H(Y|X)$ does not depend on λ,

$$\begin{array}{rclccccc}
H(Y|X) &=& p(X=0) & H(Y|X=0) & + & p(X=1) & H(Y|X=1) \\
&=& (1-\lambda) & H(\begin{bmatrix}1-\alpha-\epsilon & \epsilon & \alpha\end{bmatrix}) & + & \lambda & H(\begin{bmatrix}\epsilon & 1-\alpha-\epsilon & \alpha\end{bmatrix}) \\
&=& & & H(\begin{bmatrix}1-\alpha-\epsilon & \epsilon & \alpha\end{bmatrix}) & &
\end{array}$$

This means we only need to find the λ maximizing $H(Y)$. We could differentiate using our calculation of $p(y)$. Instead, we use the method from Section 7.1.5, defining E be the event that $\{Y = e\}$. We can use this to derive $H(Y)$.

$$\begin{aligned}
H(Y) &= H(Y, E) \\
&= H(E) + H(Y|E) \\
&= H(E) + (1-\alpha)H(Y|E=0)
\end{aligned}$$

where the last line follows from the fact that $H(Y|E=1) = 0$. Because $H(E)$ is not a function of λ, we can leave it here. Therefore, we want to maximize $H(Y|E=0)$. So we need to compute $p(y|E=0)$,

$$\begin{aligned}
p(Y=0|E=0) &= \frac{P(E=0|Y=0)P(Y=0)}{P(E=0)} \\
&= \frac{1+\alpha-\epsilon+\alpha\lambda+2\epsilon\lambda-\lambda}{1-\alpha} \\
p(Y=1|E=0) &= \frac{P(E=0|Y=1)P(Y=1)}{P(E=0)} \\
&= \frac{\epsilon-2\epsilon\lambda+\lambda-\alpha\lambda}{1-\alpha} \\
&= 1 - P(Y=0|E=0)
\end{aligned}$$

Again, we could differentiate $H(Y|E=0)$ with respect to λ but that's hairy. Instead, we'll recall that $H(Y|E=0) \le 1$ with equality when $p(Y=0|E=0) = p(Y=1|E=0)$.

$$\begin{aligned}
p(Y=0|E=0) &= p(Y=1|E=0) \\
\frac{1-\alpha-\epsilon+\alpha\lambda^*+2\epsilon\lambda^*-\lambda^*}{1-\alpha} &= \frac{\epsilon-2\epsilon\lambda^*+\lambda^*-\alpha\lambda^*}{1-\alpha} \\
\lambda^* &= \frac{1}{2}
\end{aligned}$$

The channel capacity is,

$$\begin{array}{rcccccc} C & = & \max_\lambda I(X;Y) & & & & \\ & = & H(E) & + & (1-\alpha)H_{\lambda^*}(Y|E=0) & - & H(Y|X) \\ & = & H\left(\begin{bmatrix}\alpha & 1-\alpha\end{bmatrix}\right) & + & (1-\alpha) & - & H(\begin{bmatrix}1-\alpha-\epsilon & \epsilon & \alpha\end{bmatrix}) \end{array}$$

(b) In the situation where $\alpha = 0$,

$$\begin{array}{rcccccc} C & = & H\left(\begin{bmatrix}0 & 1\end{bmatrix}\right) & + & 1 & - & H(\begin{bmatrix}1-\epsilon & \epsilon & 0\end{bmatrix}) \\ & = & & & 1 & - & H(\begin{bmatrix}1-\epsilon & \epsilon\end{bmatrix}) \end{array}$$

(c) In the situation where $\epsilon = 0$,

$$\begin{array}{rcccccc} C & = & H\left(\begin{bmatrix}\alpha & 1-\alpha\end{bmatrix}\right) & + & (1-\alpha) & - & H(\begin{bmatrix}1-\alpha & 0 & \alpha\end{bmatrix}) \\ & = & & & 1-\alpha & & \end{array}$$

6. **[Cover and Thomas 7.15]**

Given the following distributions,

$$\begin{aligned} p(x) &= \begin{bmatrix}\frac{1}{2} \\ \frac{1}{2}\end{bmatrix} \\ p(y) &= \begin{bmatrix}\frac{1}{2} & \frac{1}{2}\end{bmatrix} \\ p(x,y) &= \begin{bmatrix}0.45 & 0.05 \\ 0.05 & 0.45\end{bmatrix} \\ p(y|x) &= \begin{bmatrix}0.90 & 0.10 \\ 0.10 & 0.90\end{bmatrix} \end{aligned}$$

(a)

$$\begin{aligned} H(X) &= 1 \text{ bit} \\ H(Y) &= 1 \text{ bit} \\ H(X,Y) &\approx 1.469 \text{ bits} \\ I(X;Y) &= H(X) + H(Y) - H(X,Y) \\ &\approx 0.531 \ \text{ bits} \end{aligned}$$

(b) For X^n,

$$\begin{aligned} \left| -\frac{1}{n} \log p(x^n) - H(X) \right| &< \epsilon \\ \left| -\frac{1}{n} \log \left(\frac{1}{2}\right)^2 - H(X) \right| &< \epsilon \\ 0 &< \epsilon \end{aligned}$$

Therefore, all X^n are typical. The proof for Y^n is similar.

(c)

$$\begin{aligned} H(X,Y,Z) &= H(X,Y) + H(Z|X,Y) \\ &= H(X,Y) \end{aligned} \qquad \begin{aligned} H(X,Y,Z) &= H(X,Z) + H(Y|X,Z) \\ &= H(X,Z) \end{aligned}$$

$$
\begin{aligned}
H(X,Z) &= H(X) + H(Z|X) \\
&= H(X) + H(Z) && \text{since } Z \text{ is independent of } X \\
&= H(X,Y)
\end{aligned} \tag{3.7}
$$

We also know that z^n is typical. Therefore,

$$
\begin{aligned}
\epsilon &> \left| -\frac{1}{n} \log p(z^n) - H(Z) \right| && z^n \text{ is typical} \\
&= \left| -\frac{1}{n} \log p(z^n) - H(Z) + \left(-\frac{1}{n} \log p(x^n) - H(X) \right) \right| && \text{shown in part b} \\
&= \left| -\frac{1}{n} \left(\log p(z^n) + \log p(x^n) \right) - \left(H(Z) + H(X) \right) \right| \\
&= \left| -\frac{1}{n} \left(\log p(z^n) p(x^n) \right) - H(X,Y) \right| && \text{Equation 3.7} \\
&= \left| -\frac{1}{n} \left(\log p(x^n, y^n) \right) - H(X,Y) \right| && \text{Equation 7.161 in the text}
\end{aligned}
$$

Therefore, (x^n, y^n) is jointly typical.

(d) By inspecting $p(x,y)$, above, we know that,

$$p(x) = \begin{bmatrix} 0.90 & 0.10 \end{bmatrix}$$

By the definition of typicality, we know that if z^n is in the set, then

$$
\begin{aligned}
H(Z) - \epsilon < & \quad -\tfrac{1}{n} \log p(z^n) & < H(Z) + \epsilon \\
H(\begin{bmatrix} 0.90 & 0.10 \end{bmatrix}) - 0.20 < & \quad -\tfrac{1}{n} \log p(z^n) & < H(\begin{bmatrix} 0.90 & 0.10 \end{bmatrix}) + 0.20 \\
0.269 < & \quad -\tfrac{1}{n} \log p(z^n) & < 0.669
\end{aligned}
$$

This corresponds to $k = 1, 2, 3, 4$. Therefore $|A_{0.10}^{25}(Z)| = 15275$.

(e)

$$
\begin{aligned}
Pr((x^n(i), Y^n) \in A_\epsilon^n(X,Y)) &= Pr(Y^n - x^n(i) \in A_\epsilon^n(Z)) \\
&= Pr(x^n(i) + Z^n - x^n(i) \in A_\epsilon^n(Z)) \\
&= Pr(Z^n \in A_\epsilon^n(Z)) \\
&= \sum_{z^n \in A_\epsilon^n(Z)} p(z^n) \\
&= \sum_{z^n \in A_\epsilon^n(Z)} p^k (1-p)^{n-k} \\
&= \sum_{k=1}^{4} \binom{n}{k} p^k (1-p)^{n-k} \\
&\approx 0.8302
\end{aligned}
$$

(f)

$$\begin{aligned} Pr((X^n, y^n) \in A_\epsilon^n(X,Y) &= Pr(y^n - X^n \in A_\epsilon^n(Z)) \\ &= \sum_{x^n} Pr(y^n - x^n \in A_\epsilon^n(Z)) \\ &= \sum_{z^n \in A_\epsilon^n(Z)} p(x^n) \\ &= \sum_{z^n \in A_\epsilon^n(Z)} \frac{1}{2^n} \\ &= \frac{|A_\epsilon^n(Z)|}{2^n} \end{aligned}$$

(g)

(h)

7. **[Cover and Thomas 7.20]**

(a)

$$\begin{aligned} I(X;Y_1,Y_2) &= H(Y_1,Y_2) - H(Y_1,Y_2|X) \\ &= H(Y_1) + H(Y_2|Y_1) - H(Y_1|X) \\ &\quad - H(Y_2|Y_1,X) \\ &= H(Y_1) + H(Y_2) - I(Y_1,Y_2) - H(Y_1|X) \\ &\quad - H(Y_2|Y_1,X) \\ &= H(Y_1) - H(Y_1|X) + H(Y_2) - H(Y_2|X) \\ &\quad + H(Y_2|X) - H(Y_2|Y_1,X) - I(Y_1,Y_2) \\ &= I(Y_1;X) + I(Y_2;X) + I(Y_2,Y_1|X) - I(Y_1,Y_2) \\ &= I(Y_1;X) + I(Y_2;X) - I(Y_1,Y_2) && Y_1 \text{ and } Y_2 \text{ conditionally independent given } X \\ &= 2I(Y_1;X) - I(Y_1,Y_2) && Y_1 \text{ and } Y_2 \text{ identically distributed given } X \end{aligned}$$

(b)

$$\begin{aligned} C_{X\to(Y_1,Y_2)} &= \max_{p(x)} I(X;Y_1,Y_2) \\ &= \max_{p(x)} (2I(X;Y_1) - I(Y_1,Y_2)) \\ &\le \max_{p(x)} 2I(X;Y_1) && \text{since } I(Y_1;Y_2) \ge 0 \\ &= 2\max_{p(x)} I(X;Y_1) \\ &= 2C_{X\to Y_1} \end{aligned}$$

8. **[Cover and Thomas 7.30]**

(a)

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

We cleverly select $\mathcal{Z}$ so that $H(X|Y) = 0$. This occurs when $\mathcal{Z}$ results in a channel with nonoverlapping outputs. One such set of values is $\mathcal{Z} = \{4, 8, 12\}$.

We pick the uniform distribution over $\mathcal{X}$ to maximize $H(X)$. The entropy for this distribution is $\log |\mathcal{X}| = 2$ bits. This is also our maximum channel capacity.

(b)

$$
\begin{aligned}
H(X,Y,Z) &= H(X,Y,Z) \\
H(X,Y|Z) + H(Z) &= H(X,Z|Y) + H(Y) \\
H(X|Z) + H(Y|X,Z) + H(Z) &= H(X|Y) + H(Z|X,Y) + H(Y) \\
H(X|Z) + H(Z) &= H(X|Y) + H(Y) \\
H(X) + H(Z) &= H(X|Y) + H(Y) \\
H(X) - H(X|Y) &= H(Y) - H(Z) \\
I(X;Y) &= H(Y) - H(Z) \\
I(X;Y) &= H(Y) - \log 3
\end{aligned}
$$

Therefore $\min I(X;Y) = \min H(Y)$. The minimum entropy for Y occurs when $|\mathcal{Y}|$ is small. This occurs when $\mathcal{Z}$ is a set of 3 consecutive integers. In this case, $\mathcal{Y}$ is a set of six consecutive integers. In the case of $\mathcal{Z} = \{0, 1, 2\}$, we have

$$
\begin{aligned}
P(Y=0) &= \frac{1}{3}\lambda_0 \\
P(Y=1) &= \frac{1}{3}(\lambda_0 + \lambda_1) \\
P(Y=2) &= \frac{1}{3}(\lambda_0 + \lambda_1 + \lambda_2) \\
P(Y=3) &= \frac{1}{3}(1 - \lambda_0) \\
P(Y=4) &= \frac{1}{3}(1 - (\lambda_0 + \lambda_1)) \\
P(Y=5) &= \frac{1}{3}(1 - (\lambda_0 + \lambda_1 + \lambda_2))
\end{aligned}
$$

where $p(i) = \lambda_i$ and $\lambda_3 = 1 - \sum_{i=0}^{2} \lambda_i$. We need to find the values of λ_i which maximize $H(Y)$. This occurs when Y is uniformly distributed or,

$$
\begin{bmatrix} \lambda_0 \\ \lambda_1 \\ \lambda_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \\ 0 \\ 0 \end{bmatrix}
$$

In this case, $H(Y) = \log 6$ and $C = \log 6 - \log 3 = 1$.