Problem Set 4 Solution

Solutions courtesy of Joy A. Thomas, with editing by Frank R. Kschischang.

7.1 *Preprocessing 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 statistician 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 the 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.

(Recall that in the proof of the data processing inequality, assuming $X \to Y \to Z$ forms a Markov chain, we have $I(X;Y,Z) = I(X;Z) + I(X;Y \mid Z) = I(X;Y) + I(X;Z \mid Y)$. The Markov chain property implies that $I(X;Z \mid Y) = 0$, so we get $I(X;Z) + I(X;Y \mid Z) = I(X;Y)$. To have $I(X;Z) = I(X;Y)$ we need $I(X;Y \mid Z) = 0$, i.e., $X \to Z \to Y$ must form a Markov chain.)

7.3 *Channels with memory have higher capacity.* We have

$$\begin{aligned}
I(X_1,\ldots,X_n;Y_1,\ldots,Y_n) &= H(Y_1,\ldots,Y_n) - H(Y_1,\ldots,Y_n \mid X_1,\ldots,X_n) \\
&= H(Y_1,\ldots,Y_n) - H(X_1+Z_1,\ldots,X_n+Z_n \mid X_1,\ldots,X_n) \\
&= H(Y_1,\ldots,Y_n) - H(Z_1,\ldots,Z_n \mid X_1,\ldots,X_n) \\
&= H(Y_1,\ldots,Y_n) - H(Z_1,\ldots,Z_n) \\
&\leq n - H(Z_1,\ldots,Z_n),
\end{aligned}$$

where equality is achieved by choosing $p(x_1,\ldots,x_n)$ so that $Y_1,\ldots,Y_n$ are independent and uniformly distributed. Such a distribution over $(Y_1,\ldots,Y_n)$ is achieved by choosing $X_1,\ldots,X_n$ to be independent and uniformly distributed. The capacity of the channel is then $C = n - H(Z_1,\ldots,Z_n)$. Since $H(Z_1,\ldots,Z_n) \leq H(Z_1) + \cdots + H(Z_n) = nH(p,1-p)$ we see that $C \geq n(1-H(p,1-p))$. Thus 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 information from the past samples of the noise to combat the present noise.

7.5 *Using two channels at once.* To find the capacity of the product channel, we must find the distribution $p(x_1, x_2)$ on the input alphabet $\mathcal{X}_1 \times \mathcal{X}_2$ that maximizes $I(X_1, X_2; Y_1, Y_2)$, where the channel law is given as $p(y_1, y_2 \mid x_1, x_2) = p(y_1 \mid x_1)p(y_2 \mid x_2)$. Note that the joint probability distribution factors as

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

which implies that $Y_1 \to X_1 \to X_2 \to Y_2$ forms a Markov chain. It follows also that $Y_2 \to X_2 \to X_1 \to Y_1$ forms a Markov chain. Now

$$
\begin{aligned}
I(X_1, X_2; Y_1, Y_2) &= H(Y_1, Y_2) - H(Y_1, Y_2 \mid X_1, X_2) \\
&= H(Y_1, Y_2) - H(Y_2 \mid X_1, X_2) - H(Y_1 \mid X_1, X_2, Y_2) \\
&= H(Y_1, Y_2) - H(Y_2 \mid X_2) - H(Y_1 \mid X_1) \\
&\le H(Y_1) + H(Y_2) - H(Y_1 \mid X_1) - H(Y_2 \mid X_2) \\
&= I(X_1; Y_1) + I(X_2; Y_2) \\
&\le C_1 + C_2
\end{aligned}
$$

where equality in the first inequality occurs if and only if Y_1 and Y_2 are independent (achieved when X_1 and X_2 are independent) and equality is achieved in the second inequality by choosing the marginal distributions $p(x_1)$ and $p(x_2)$ to be the capacity-achieving input distributions for their respective channels.

7.8 *Z-channel.* Let $p = \Pr(X = 1)$, and denote $I(X; Y)$ as $I(p)$. Then

$$
\begin{aligned}
I(p) &= H(Y) - H(Y \mid X) \\
&= \mathcal{H}(p/2) - H(Y \mid X = 0)(1 - p) - H(Y \mid X = 1)p \\
&= \mathcal{H}(p/2) - p
\end{aligned}
$$

Since $I(p) = 0$ when $p = 0$ and $p = 1$, the maximum mutual information is obtained for some value of $p \in (0, 1)$. Using calculus we find that

$$
\frac{\mathrm{d}}{\mathrm{d}p} I(p) = \frac{1}{2} \log_2 \left(\frac{2 - p}{p} \right) - 1
$$

which is equal to zero when $p = 2/5$. Thus the capacity of Z-channel is $C = \mathcal{H}(1/5) - 2/5 \approx 0.321928$ bit/channel use. It is reasonable that $p < 1/2$ because $X = 1$ is the noisy input to the channel.

7.28 *Choice of channels.* We can communicate information to the destination via the choice of channel (in addition to the information we can transmit over either channel). Furthermore, since the output alphabets are distinct, the information embedded in the choice of channel is received perfectly.

Let Z be a random variable with alphabet $\{1,2\}$, where $Z=1$ indicates that channel 1 is selected for the current transmission, $Z=2$ indicates channel 2 is selected, and $\Pr\{\mathrm{Z}=1\}=\alpha$. Let X be a random variable (representing the channel input) with alphabet $\{\mathcal{X}_1 \cup \mathcal{X}_2\}$, and similarly let Y be a random variable (representing the channel output) with alphabet $\{\mathcal{Y}_1 \cup \mathcal{Y}_2\}$.

(a) We communicate to the destination via Z and X. We have that

$$\begin{aligned}
I(X,Z;Y) &= H(Z,X) - H(X,Z|Y) \\
&= H(Z) + H(X|Z) - H(X|Y) \\
&= \mathcal{H}(\alpha) + \alpha H(X_1) + (1-\alpha)H(X_2) - H(X|Y) \\
&= \mathcal{H}(\alpha) + \alpha H(X_1) + (1-\alpha)H(X_2) - \alpha H(X_1|Y_1) - (1-\alpha)H(X_2|Y_2) \\
&= \mathcal{H}(\alpha) + \alpha I(X_1;Y_1) + (1-\alpha)I(X_2;Y_2).
\end{aligned}$$

Now, we obtain the capacity by maximizing over the choice of input distribution, which involves maximizing over $p(x_1)$, $p(x_2)$ and α. Note that $I(X_1;Y_1)$ depends only on $p(x_1)$ and similarly for $I(X_2;Y_2)$ and $p(x_2)$. Thus,

$$C = \max_{\alpha} \left(\mathcal{H}(\alpha) + \alpha C_1 + (1-\alpha)C_2\right).$$

Now, it is an exercise in calculus to show that this quantity is maximized when

$$\alpha = \frac{2^{C_1}}{2^{C_1}+2^{C_2}},$$

and that

$$2^C = 2^{C_1} + 2^{C_2}$$

for the optimal choice of α; the details are given below.

(b) We have that the effective alphabet size for noise-free communication is the sum of the effective alphabets of the two constituent channels. In particular, note that the condition that the output symbols be disjoint is analogous to the condition in Problem 2.10 that the random variables have disjoint ranges.

(c) The capacity of the BSC is $1-\mathcal{H}(p)$, and that of the single-input single-output channel is zero, thus

$$C = \log_2\left(1 + 2^{1-\mathcal{H}(p)}\right).$$

Now for the details. Let $f(\alpha) = \mathcal{H}(\alpha) + \alpha C_1 + (1-\alpha)C_2$. Recall that $\frac{\mathrm{d}}{\mathrm{d}p}\mathcal{H}(p) = \log_2\left(\frac{1-p}{p}\right)$. Thus

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} f(\alpha) = \log_2\left(\frac{1-\alpha}{\alpha}\right) + C_1 - C_2.$$

which passes through zero when $\log_2((1-\alpha)/\alpha) = C_2 - C_1$, i.e., when $1/\alpha = 1 + 2^{C_2-C_1}$, or

$$\alpha = \frac{1}{1+2^{C_2-C_1}} = \frac{2^{C_1}}{2^{C_1}+2^{C_2}}.$$

We have, for this optimizing value of α, that

$$\begin{aligned}
2^C &= 2^{\mathcal{H}(\alpha)} 2^{\alpha C_1 + (1-\alpha) C_2} \\
&= \frac{2^{\alpha C_1} 2^{(1-\alpha) C_2}}{\alpha^\alpha (1-\alpha)^{1-\alpha}} \\
&= \left(\frac{2^{C_1}}{\alpha}\right)^\alpha \left(\frac{2^{C_2}}{1-\alpha}\right)^{1-\alpha} \\
&= \left(2^{C_1} + 2^{C_2}\right)^\alpha \left(2^{C_1} + 2^{C_2}\right)^{1-\alpha} \\
&= 2^{C_1} + 2^{C_2}.
\end{aligned}$$

7.9 *Suboptimal codes.* From the proof of the channel coding theorem, it follows that using a random code with codewords generated according to probability $p(x)$, we can send information at a rate $I(X;Y)$ corresponding to that $p(x)$ with an arbitrarily low probability of error. For the Z-channel described in Problem 7.8, we can calculate $I(X;Y)$ for a uniform distribution on the input. The distribution on Y is $(3/4, 1/4)$, and therefore

$$I(X;Y) = H(Y) - H(Y \mid X) = \mathcal{H}\left(\frac{1}{4}\right) - \frac{1}{2} \approx 0.311278 \text{ bit/channel use},$$

which is smaller than capacity of 0.321928 bit/channel use.

7.14 *Channels with dependence between the letters.*

(a) First,

$$\begin{aligned}
I(X_1, X_2; Y_1, Y_2) &= H(X_1, X_2) - H(X_1, X_2 | Y_1, Y_2) \\
&= H(X_1, X_2),
\end{aligned}$$

where the last equality follows from the fact that (Y_1, Y_2) uniquely identifies (X_1, X_2).

(b)

$$\begin{aligned}
C &= \max_{p(x_1, x_2)} H(X_1, X_2) \\
&= \log_2(4) = 2,
\end{aligned}$$

when the four input symbols are used equiprobably.

(c) When the input symbols are equiprobable, it is easy to show that

$$p(x_1 \mid y_1) = p(x_1).$$

Therefore,

$$\begin{aligned}
I(X_1; Y_1) &= H(X_1) - H(X_1 \mid Y_1) \\
&= H(X_1) - H(X_1) \\
&= 0.
\end{aligned}$$

7.19 *Capacity of the carrier pigeon channel.*

(a) Since a pigeon arrives at the destination every 5 minutes, there are 12 pigeons per hour that arrive safely. Each pigeon carries an 8 bit message, thus the capacity of the link is 96 bits/hour.

Effectively, we get to use an errorless channel (with 256 inputs and 256 outputs, and thus capacity $C_1 = \log_2(256) = 8$ bits/channel-use) 12 times per hour

(b) We can model this problem using an erasure channel with 256 inputs and 257 outputs (256 output symbols that match the inputs, and one erasure); the input is received correctly with probability $1-\alpha$, and an erasure occurs with probability α.

Just as in the binary-input erasure channel, it is easy to show that the capacity is

$$C_2 = (1-\alpha)\log_2(256) = 8(1-\alpha) \text{ bit/channel use},$$

which is equivalent to $96(1-\alpha)$ bit/hour.

(c) We can model this problem using a 256-ary symmetric channel (just like the BSC, but with 256 inputs and 256 outputs). A pigeon arrives safely with probability $1-\alpha$, and if the pigeon is shot down, the probability that the dummy carries the intended message is $\frac{1}{256}$. Therefore, the message is received correctly with probability

$$1-\alpha+\alpha\left(\frac{1}{256}\right) = 1-\frac{255\alpha}{256}.$$

Furthermore, when an error occurs (i.e., the pigeon gets shot, and the dummy carries the wrong message), all of the incorrect messages are equiprobable.

Finally, by symmetry, it is clear that the uniform input distribution is capacity-achieving. Let X represent the channel input, and Y represent the channel output. Then, for equiprobable inputs,

$$\begin{aligned}
C_3 &= H(Y) - H(Y \mid X) \\
&= \log_2(256) - H\left(1-\frac{255\alpha}{256}, \frac{\alpha}{256}, \frac{\alpha}{256}, \ldots, \frac{\alpha}{256}\right) \\
&= 8 - \mathcal{H}\left(\frac{255\alpha}{256}\right) - \frac{255\alpha}{256}\log_2(255),
\end{aligned}$$

and the link capacity is $12C_3$ bits/hour.

7.27 *Erasure channel.* Expanding $I(X;Y,S)$ in two ways,

$$\begin{aligned}
I(X;Y,S) &= I(X;Y) + I(X;S \mid Y) \\
&= I(X;S) + I(X;Y \mid S).
\end{aligned}$$

Since $X \to Y \to S$, we have $I(X;S \mid Y) = 0$, and thus

$$I(X;S) = I(X;Y) - I(X;Y \mid S).$$

Now

$$
\begin{aligned}
I(X;Y \mid S) &= H(Y \mid S) - H(Y \mid X,S) \\
&= \alpha H(Y \mid S=?) + (1-\alpha)H(Y \mid S=Y) \\
&\quad - \alpha H(Y \mid X, S=?) - (1-\alpha)H(Y \mid X, S=Y) \\
&= \alpha H(Y) + (1-\alpha)\cdot 0 - \alpha H(Y \mid X) - (1-\alpha)\cdot 0 \\
&= \alpha(H(Y) - H(Y \mid X)) \\
&= \alpha I(X;Y).
\end{aligned}
$$

Thus $I(X;S) = (1-\alpha)I(X;Y)$ and therefore $C_{erasure} = (1-\alpha)C$, where C is the capacity of the first channel. This capacity is easily achieved with the help of a feedback channel. Whenever $S=?$, the receiver could ask for retransmission of X, and otherwise pass S to a decoder for the $X \to Y$ channel, which would simply see a sequence of received Y symbols without any erasures at all. The fraction of "wasted" (i.e., erased) channel uses is α, thus the $(1-\alpha)$ fraction of the remaining channel uses implement the $X \to Y$ channel.

7.35 *Capacity.*

(a) Due to the structure of $\hat{\mathcal{P}}$ (in particular, the off-diagonal zeros), we can interpret $\hat{\mathcal{P}}$ to be made up of two distinct channels with non-overlapping inputs and outputs, and at each channel use, the user must communicate using one of the two channels. In other words, the user selects a channel (and thus conveys info by doing so), and then transmits a symbol over the given channel (conveying more info).

Following our solution to Problem 7.28, we know that the effective capacity is

$$C = \log_2(2^{C_p} + 2^{C_a}),$$

where C_p is the capacity of $\mathcal{P}$, and C_a is the capacity of the noiseless unary channel. Clearly, $C_a = 0$, therefore

$$C_= \log_2(2^{C_p} + 1).$$

(b) Similarly, since I_k represents a noiseless k-ary channel, its capacity is $C_a = \log_2 k$, and we have

$$C = \log_2(2^{C_p} + k).$$