EEE 551 Information Theory (Spring 2022)

Homework 3 Solutions

---

1. Problem 7.5 from Cover-Thomas: *Using two channels at once.* Consider two discrete memoryless channels $(\mathcal{X}_1, p(y_1|x_1), \mathcal{Y}_1)$ and $(\mathcal{X}_2, p(y_2|x_2), \mathcal{Y}_2)$ with capacities C_1 and C_2, respectively. A new channel $(\mathcal{X}_1 \times \mathcal{X}_2, p(y_1|x_1) \times p(y_2|x_2), \mathcal{Y}_1 \times \mathcal{Y}_2)$ is formed in which $x_1 \in \mathcal{X}_1$ and $x_2 \in \mathcal{X}_2$ are sent simultaneously, resulting in y_1, y_2. Find the capacity of this channel.

   Solution: 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 $I(X_1, X_2; Y_1, Y_2)$. Since the joint distribution

   $$p(x_1, x_2, y_1, y_2) = p(x_1, x_2)p(y_1|x_1)p(y_2|x_2)$$

   we have the Markov chain $Y_1 \to X_1 \to X_2 \to Y_2$. Therefore

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

   where (a) follows from Markovity, and equality occurs in (b) if Y_1 and Y_2 are independent. Hence a sufficient condition for equality is that X_1 and X_2 are independent. We may bound the capacity of the product channel by

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

   Equality occurs if $p(x_1, x_2) = p^*(x_1)p^*(x_2)$, where $p^*(x_1)$ and $p^*(x_2)$ are distributions achieving C_1 and C_2 respectively. Therefore $C = C_1 + C_2$.

2. *Channel capacity.* Calculate the capacity of the following channels with the given probability transition matrices. Recall that each row of the matrix corresponds to an input symbol, and each column corresponds to an output symbol.

   (a) $\mathcal{X} = \mathcal{Y} = \{0, 1, 2\}$

   $$p(y|x) = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix}$$

(b) $\mathcal{X} = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$, $\mathcal{Y} = \{0, 1\}$

$$p(y|x) = \begin{bmatrix} 0.1 & 0.9 \\ 0.2 & 0.8 \\ 0.3 & 0.7 \\ 0.4 & 0.6 \\ 0.5 & 0.5 \\ 0.6 & 0.4 \\ 0.7 & 0.3 \\ 0.8 & 0.2 \\ 0.9 & 0.1 \end{bmatrix}$$

(c) $\mathcal{X} = \mathcal{Y} = \{0, 1, 2\}$

$$p(y|x) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}$$

Solution:

(a) This channel is weakly symmetric, so its capacity is

$$\begin{aligned} C &= \log |\mathcal{Y}| - H(\text{row of transition matrix}) \\ &= \log 3 - H(1/2, 1/2, 0) \\ &= \log 3 - 1. \end{aligned}$$

(b) For this channel, the output is binary, but there are many options for the input. Intuitively, we should select only the input letters that convey the most information. An input letters conveys more information if the corresponding row of the transition matrix that is more biased. Thus, it appears that we should use input letters 1 and 9, corresponding to rows [0.1 0.9], and [0.9 0.1]. In particular, if we choose

$$p(x) = \begin{cases} 1/2 & x = 1 \text{ or } x = 9, \\ 0 & \text{otherwise.} \end{cases}$$

This input distribution induces a uniform distribution on the output, so

$$I(X; Y) = H(Y) - H(Y|X) = 1 - H(0.1).$$

Thus, $C \geq 1 - H(0.1)$.
We also need an upper bound on the capacity. We have

$$\begin{aligned} I(X; Y) &= H(Y) - H(Y|X) \\ &\leq 1 - H(Y|X) \\ &= 1 - \sum_x p(x) H(Y|X = x) \\ &\leq 1 - \min_x H(Y|X = x) \\ &= 1 - H(0.1). \end{aligned}$$

Thus, in fact $C = 1 - H(0.1)$.

(c) Consider any input distribution $p(x)$, and let $q = p_X(3)$. Note that $p_Y(3) = q/3$. Thus we may upper bound the mutual information by

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(p_X(1) + q/3, p_X(2) + q/3, q/3) - q\log 3 \\ &= H(q/3, 1-q/3) + (1-q/3)H\left(\frac{p_X(1)+q/3}{1-q/3}, \frac{p_X(3)+q/3}{1-q/3}\right) - q\log 3 \\ &\le H(q/3) + (1-q/3) - q\log 3. \end{aligned}$$

If we let $r = q/3$, then

$$I(X;Y) \le H(r) + 1 - r - 3r\log 3 = -r\log r - (1-r)\log(1-r) + 1 - r - 3r\log 3.$$

To find the value of r that maximizes this expression, we differentiate by r:

$$-\log r + \log(1-r) - 1 - 3\log 3 = \log\frac{1-r}{r} - 1 - 3\log 3.$$

The maximum value will be attained when this derivative is 0, so

$$\log\frac{1-r}{r} = 1 + 3\log 3 = \log 2 + \log 3^3 = \log 54.$$

Thus we need $\frac{1-r}{r} = 54$, so $r = 1/55$. Thus,

$$I(X;Y) \le H(1/55) + 1 - \frac{1}{55} - \frac{3}{55}\log 3.$$

Note that $r = 1/55$ corresponds to $q = 3/55$. This suggests that the optimal distribution for X is where $p_X(3) = 3/55$. Indeed, we can get a matching lower bound by choosing

$$p_X = \left[\frac{26}{55}, \frac{26}{55}, \frac{3}{55}\right].$$

This induces the marginal distribution for Y of

$$p_Y = \left[\frac{27}{55}, \frac{27}{55}, \frac{1}{55}\right].$$

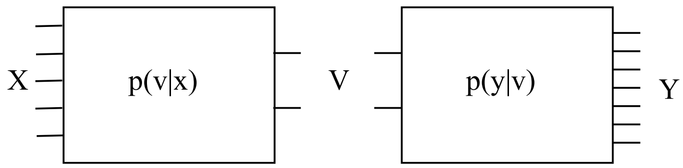
Thus

$$I(X;Y) = H(Y) - H(Y|X) = H\left(\frac{27}{55}, \frac{27}{55}, \frac{1}{55}\right) - \frac{1}{55}\log 3 = H(1/55) + 1 - \frac{1}{55} - \frac{3}{55}\log 3.$$

Therefore, in fact

$$C = H(1/55) + 1 - \frac{1}{55} - \frac{3}{55}\log 3 \approx 1.0265.$$

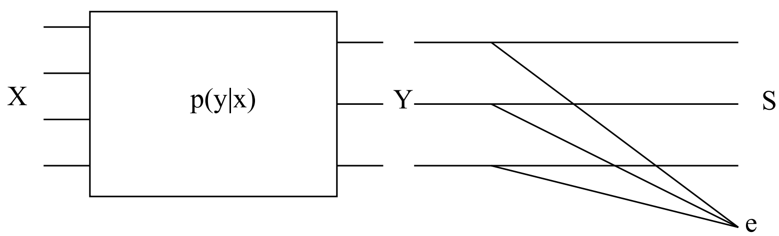
3. Problem 7.25 from Cover-Thomas: *Bottleneck channel.* Suppose a signal $X \in \mathcal{X} = \{1, 2, \ldots, m\}$ goes through an intervening transition $X \longrightarrow V \longrightarrow Y$:



where $\mathcal{X} = \{1, 2, \ldots, m\}$, $\mathcal{Y} = \{1, 2, \ldots, m\}$, and $\mathcal{V} = \{1, 2, \ldots, k\}$. Here $p(v|x)$ and $p(y|v)$ are arbitrary and the channel has transition probability $p(y|x) = \sum_v p(v|x)p(y|v)$. Show that $C \leq \log k$.

Solution: By the data processing inequality, $I(X;Y) \leq I(V;Y) = H(V) - H(V|Y) \leq H(V) \leq \log k$. Thus $C = \max_{p(x)} I(X;Y) \leq \log k$.

4. Problem 7.27 from Cover-Thomas: *Erasure channel.* Let $\{\mathcal{X}, p(y|x), \mathcal{Y}\}$ be a discrete memoryless channel with capacity C. Suppose that this channel is cascaded immediately with an erasure channel $\{\mathcal{Y}, p(s|y), \mathcal{S}\}$ that erases α of its symbols.



Specifically, $\mathcal{S} = \{y_1, y_2, \ldots, y_m, e\}$ and

$$\begin{aligned} \Pr\{S = y|X = x\} &= (1-\alpha)p(y|x), \qquad y \in \mathcal{Y}, \\ \Pr\{S = e|X = x\} &= \alpha. \end{aligned}$$

Determine the capacity of this channel.

Solution: The capacity of this channel is $C = \max_{p(x)} I(X;S)$. Define a random variable E, where $E = 1$ if $S = e$ and $E = 0$ otherwise. Since E is a function of S, $H(E|S) = 0$. Moreover, E is independent of X, and $\Pr(E = 1) = \alpha$. Thus

$$\begin{aligned} I(X;S) &= H(X) - H(X|S) \\ &= H(X|E) + I(X;E) - H(X|S,E) + I(X;E|S) \\ &\overset{(a)}{=} H(X|E) - H(X|S,E) \\ &= I(X;S|E) \\ &= \alpha I(X;S|E=1) + (1-\alpha)I(X;S|E=0) \end{aligned}$$

where in (a) $I(X;E) = 0$ because X and E are independent, and $I(X;E|S) = 0$ because E is a function of S. When $E = 1$, $S = e$, so $I(X;S|E=1) = 0$. When $E = 0$, $S = Y$, so $I(X;S|E=0) = I(X;Y)$. Thus $I(X;S) = (1-\alpha)I(X;Y)$ and

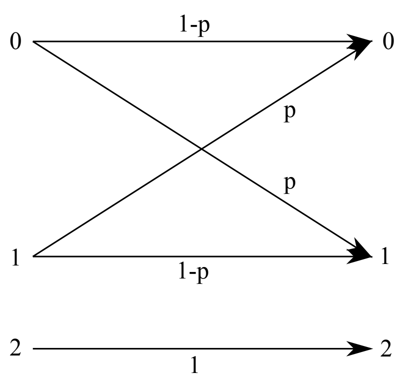
$$C = \max_{p(x)} I(X;S) = (1-\alpha) \max_{p(x)} I(X;Y).$$

That is, the capacity of the erasure channel is $(1-\alpha)$ times the capacity of the original channel.

5. Problem 7.28 from Cover-Thomas, parts (a) and (c): *Choice of channels.* Find the capacity C of the union of two channels $(\mathcal{X}_1, p_1(y_1|x_1), \mathcal{Y}_1)$ and $(\mathcal{X}_2, p_2(y_2|x_2), \mathcal{Y}_2)$ [with capacities C_1 and C_2 respectively], where at each time, one can send a symbol over channel 1 or channel 2 but not both. Assume that the output alphabets are distinct and do not intersect.

   (a) Show that $2^C = 2^{C_1} + 2^{C_2}$.

   (c) Use the above result to calculate the capacity of the following channel.



Solution:

(a) Define random variable θ where $\theta = 1$ if $X \in \mathcal{X}_1$ and $\theta = 2$ if $X \in \mathcal{X}_2$. Also let $\alpha = \Pr(\theta = 1)$. Note that θ is a function of X. Moreover, θ is also a function of Y, since whether Y is in $\mathcal{Y}_1$ or $\mathcal{Y}_2$ determines θ. We have

$$\begin{aligned}
I(X;Y) &= H(Y) - H(Y|X) \\
&= H(Y,\theta) - H(\theta|Y) - H(Y|X,\theta) - I(Y;\theta|X) \\
&\overset{(a)}{=} H(Y,\theta) - H(Y|X,\theta) \\
&= H(\theta) + H(Y|\theta) - H(Y|X,\theta) \\
&= H(\theta) + I(X;Y|\theta) \\
&= H(\alpha) + \alpha I(X;Y|\theta=1) + (1-\alpha) I(X;Y|\theta=2)
\end{aligned}$$

where in (a) we have used that θ is a function of both X and Y. Note that conditioned on $\theta = i$ for $i \in \{1, 2\}$, X and Y are contained in $\mathcal{X}_i$ and $\mathcal{Y}_i$ respectively, and they are related by the channel $p_i(y_i|x_i)$. Thus

$$I(X;Y|\theta = i) \leq \max_{p(x_i)} I(X_i;Y_i) = C_i$$

where in the above mutual information X_i and Y_i have distribution $p(x_i)p_i(y_i|x_i)$. Equality occurs in the above if the distribution of X given $\theta = i$ is $p^*(x_i)$, the optimal input distribution to channel i. Therefore

$$C = \max_{p(x)} I(X;Y) = \max_{\alpha \in [0,1]} H(\alpha) + \alpha C_1 + (1-\alpha) C_2.$$

Calculus can be used to find that the optimal α in the above optimization is $\frac{2^{C_1}}{2^{C_1}+2^{C_2}}$ and therefore $2^C = 2^{C_1} + 2^{C_2}$.

(c) Since the first channel is a BSC with crossover probability p, $C_1 = 1 - H(p)$. The second channel has only one input/output letter, so $C_2 = 0$. Thus $C = \log(2^{C_1} + 2^{C_2}) = \log(2^{1-H(p)} + 1)$.

6. *Polar codes for the BSC.* Consider using a polar code for the binary symmetric channel (rather than the binary erasure channel, which we covered in class).

(a) Consider a BSC(p). For the basic (i.e., single generation) polar transform, calculate the channel capacities of W^- and W^+. That is, calculate

$$\begin{aligned} I(W^-) &= I(U_1; Y_1, Y_2) \\ I(W^+) &= I(U_2; Y_1, Y_2, U_1). \end{aligned}$$

Confirm that $I(W^-) + I(W^+) = 2I(W)$, and $I(W^-) < I(W) < I(W^+)$, where $I(W)$ is the capacity of the original channel.

(b) The following table lists the features of three polar encoders. Given is the generation number (a polar code of genertaion t has blocklength $n = 2^t$), and which U bits are frozen to 0 and which are used as message bits.

Generation	Frozen Bits	Message Bits
2	U_1, U_2, U_3	U_4
3	U_1, U_2, U_3, U_4, U_6	U_5, U_7, U_8
3	U_1, U_2, U_3, U_4, U_5	U_6, U_7, U_8

For each of these polar codes:

i. List all the codewords associated with the code.

ii. In the following, a "single bit flip" means an error pattern in which exactly 1 bit is changed. For a blocklength n code, there are n different single bit flip error patterns. For each code, determine which of the following is true (and explain why):
(A) The code cannot correct any single bit flip
(B) The code can correct some (but not all) single bit flips
(C) The code can correct any single bit flip

Solution:

(a) To simplify the calculations, we let $Z_i \sim \text{Bern}(p)$, and then $Y_i = X_i \oplus Z_i$. Since in the basic polar transformation $X_1 = U_1 \oplus U_2$ and $X_2 = U_2$, we have

$$\begin{aligned} I(W^-) &= I(U_1; Y_1, Y_2) \\ &= I(U_1; X_1 \oplus Z_1,\ X_2 \oplus Z_2) \\ &= I(U_1; U_1 \oplus U_2 \oplus Z_1,\ U_2 \oplus Z_2) \\ &= H(U_1 \oplus U_2 \oplus Z_1,\ U_2 \oplus Z_2) - H(U_1 \oplus U_2 \oplus Z_1,\ U_2 \oplus Z_2 | U_1) \\ &= H(U_2 \oplus Z_2) + H(U_1 \oplus U_2 \oplus Z_1 | U_2 \oplus Z_2) - H(U_2 \oplus Z_1,\ U_2 \oplus Z_2 | U_1) \\ &= H(U_2 \oplus Z_2) + H(U_1 \oplus U_2 \oplus Z_1 | U_2 \oplus Z_2) - H(U_2 \oplus Z_1,\ U_2 \oplus Z_2) \\ &= H(U_2 \oplus Z_2) + H(U_1 \oplus U_2 \oplus Z_1 | U_2 \oplus Z_2) - H(U_2 \oplus Z_2) - H(U_2 \oplus Z_1 | U_2 \oplus Z_2) \\ &= H(U_1 \oplus U_2 \oplus Z_1 | U_2 \oplus Z_2) - H(U_2 \oplus Z_1 | U_2 \oplus Z_2) \\ &= H(U_1 \oplus Z_1 \oplus Z_2 | U_2 \oplus Z_2) - H(Z_1 \oplus Z_2 | U_2 \oplus Z_2) \end{aligned}$$

$$
\begin{aligned}
&= H(U_1 \oplus Z_1 \oplus Z_2) - H(Z_1 \oplus Z_2) \\
&= 1 - H(2p(1-p)).
\end{aligned}
$$

We may also write

$$
\begin{aligned}
I(W^+) &= I(U_2; Y_1, Y_2, U_1) \\
&= I(U_2; Y_1, Y_2 | U_1) \\
&= I(U_2; U_1 \oplus U_2 \oplus Z_1,\ U_2 \oplus Z_2 | U_1) \\
&= H(U_1 \oplus U_2 \oplus Z_1,\ U_2 \oplus Z_2 | U_1) - H(U_1 \oplus U_2 \oplus Z_1,\ U_2 \oplus Z_2 | U_1, U_2) \\
&= H(U_2 \oplus Z_1,\ U_2 \oplus Z_2 | U_1) - H(Z_1, Z_2 | U_1, U_2) \\
&= H(U_2 \oplus Z_1,\ U_2 \oplus Z_2) - H(Z_1, Z_2) \\
&= H(U_2 \oplus Z_1) + H(U_2 \oplus Z_2 | U_2 \oplus Z_1) - 2H(p) \\
&= 1 + H(Z_1 \oplus Z_2 | U_2 \oplus Z_1) - 2H(p) \\
&= 1 + H(Z_1 \oplus Z_2) - 2H(p) \\
&= 1 + H(2p(1-p)) - 2H(p).
\end{aligned}
$$

Combining these two calculations, we have

$$
I(W^-) + I(W^+) = \big(1 - H(2p(1-p))\big) + \big(1 + H(2p(1-p)) - 2H(p)\big) = 2 - 2H(p) = 2I(W).
$$

In addition,

$$
I(W) - I(W^-) = \big(1 - H(p)\big) - \big(1 - H(2p(1-p))\big) = H(2p(1-p)) - H(p).
$$

If $p \le 1/2$, then $1/2 \ge 2p(1-p) \ge p$, so $H(2p(1-p)) > H(p)$, so $I(W) > I(W^-)$. Similarly,

$$
I(W^+) - I(W) = H(2p(1-p)) - H(p) > 0
$$

so $I(W^+) > I(W)$.

(b) For the first code, with $t = 2$ and message bit U_4, the codewords are:

0000
1111

Thus, this code is simply a length-4 repetition code. If any single bit flip occurs, it can be corrected by considering the majority of the bits. Thus (C) holds.

For the second code, with $t = 3$ and message bits U_5, U_7, U_8, the codewords are:

00000000
10001000
10101010
00100010
11111111
01110111

01010101

11011101

This code cannot correct any single bit flip. For example, if 00000000 is sent but 10000000 is received, then the decoder could not tell if 00000000 or 10001000 was transmitted. However, the code can correct some single bit flips. For example, if 01000000 is received, then the unique closest codeword is 00000000, so this bit flip can be corrected. Thus (B) holds.

For the third code, with $t = 3$ and message bits U_6, U_7, U_8, the codewords are:

00000000

11001100

10101010

01100110

11111111

00110011

01010101

10011001

This code can correct any single bit flip. To see why, we may write the codeword as a function of the message bits U_6, U_7, U_8 as follows:

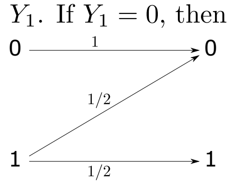
$$U_6 \oplus U_7 \oplus U_8,\ U_6 \oplus U_8,\ U_7 \oplus U_8,\ U_8,\ U_6 \oplus U_7 \oplus U_8,\ U_6 \oplus U_8,\ U_7 \oplus U_8,\ U_8.$$

In particular, the first 4 bits and the second 4 bits are the same. Any single bit flip will cause a discrepancy between one of these 4 bits. The decoder can simply ignore any bit with a discrepancy and trust the remaining 3 bits to decode. For example, if there is a discrepancy in $U_6 \oplus U_7 \oplus U_8$, then the remaining 3 bits:

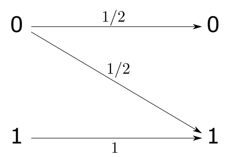
$$U_6 \oplus U_8,\ U_7 \oplus U_8,\ U_8$$

can be used to recover U_6, U_7, U_8. This is true for any three of the 4 repeated bits, so we can always recover from any single bit flip. Thus (C) holds.

7. *Channel with memory and feedback.* Consider the following binary-input binary-output channel *with memory*; meaning that the output at a given time depends on not only the input at that time but also previous inputs. At time 1, the channel acts like a BSC(1/2); that is, the output Y_1 has a uniform distribution on $\{0, 1\}$ independently of the input. At time 2, the behavior of the channel depends on Y_1. If $Y_1 = 0$, then the channel at time 2 behaves like a Z-channel, as shown below:



Recall that the optimal input distribution for a Z-channel sends 1 with probability 2/5. If $Y_1 = 1$, then the channel at time 2 behaves like an inverted Z-channel, as shown below:



For every two subsequent transmissions, the channel replicates these two steps independently (i.e. the odd transmissions are BSC(1/2), the even transmissions are Z-channel or inverted Z-channel depending on the previous output).

(a) Find the capacity of this channel. *Hint:* Considering two transmissions at a time, this is a standard discrete memoryless channel.

(b) Find the capacity of this channel with feedback. In particular, before the second transmission the encoder knows the output of the first transmission.

Solution:

(a) Following the hint, we can consider the DMC made up of two transmissions; this DMC has a two-bit input (X_1, X_2) and a two-bit output (Y_1, Y_2). The conditional distribution $p(y_1, y_2|x_1, x_2)$ describing this two-bit channel is shown below, where as usual each row is an element of $\mathcal{X}$ and each column is an element of $\mathcal{Y}$:

$$\begin{array}{cc} & \begin{array}{cccc} 00 & 01 & 10 & 11 \end{array} \\ \begin{array}{c} 00 \\ 01 \\ 10 \\ 11 \end{array} & \left[\begin{array}{cccc} 1/2 & 0 & 1/4 & 1/4 \\ 1/4 & 1/4 & 0 & 1/2 \\ 1/2 & 0 & 1/4 & 1/4 \\ 1/4 & 1/4 & 0 & 1/2 \end{array}\right] \end{array}.$$

Note that $p(y_1, y_2|x_1, x_2)$ does not depend on X_1. Thus we can ignore it, and in fact this is a channel just from X_2 to (X_1, X_2). The conditional distribution $p(y_1, y_2|x_2)$ for this slightly simpler channel is given by:

$$\begin{array}{cc} & \begin{array}{cccc} 00 & 01 & 10 & 11 \end{array} \\ \begin{array}{c} 0 \\ 1 \end{array} & \left[\begin{array}{cccc} 1/2 & 0 & 1/4 & 1/4 \\ 1/4 & 1/4 & 0 & 1/2 \end{array}\right] \end{array}.$$

Since this new channel represents two uses of the original channel, its capacity is twice the capacity C of the original channel. Thus we have

$$\begin{aligned} 2C &= \max_{p(x_2)} I(X_2; Y_1, Y_2) \\ &= \max_{p(x_2)} H(Y_1, Y_2) - H(Y_1, Y_2|X_2) \\ &= \max_{p(x_2)} H(Y_1, Y_2) - H(1/2, 1/4, 1/4, 0) \\ &= \max_{p(x_2)} H(Y_1, Y_2) - \frac{3}{2} \end{aligned}$$

where we have used the fact that both rows of the probability transition matrix have distribution $(1/2, 1/4, 1/4, 0)$. It is easy to see that the distribution on X_2 maximizing $H(Y_1, Y_2)$ is the uniform one, in which case Y_1, Y_2 has distribution $(3/8, 1/8, 1/8, 3/8)$ and entropy $H(1/4) + 1$. Thus

$$2C = H(1/4) + 1 - \frac{3}{2} = H(1/4) - \frac{1}{2}$$

and

$$C = \frac{H(1/4)}{2} - \frac{1}{4} \approx 0.156.$$

(b) With feedback, the encoder knows the value of Y_1 before it sends X_2, which means it knows whether the channel from X_2 to Y_2 will be a Z-channel or an inverted Z-channel. Thus the encoder can shape the distribution of X_2 depending on which channel will occur. In particular, if $Y_1 = 0$, the encoder chooses $X_2 \sim \text{Bern}(2/5)$, and if $Y_1 = 1$, the encoder chooses $X_2 \sim \text{Bern}(3/5)$. In either case, the encoder's second transmission can achieve the capacity of the Z-channel or the inverted Z-channel (which has the same capacity). Therefore, with two channel uses of the original channel, the encoder can send as many bits as in the capacity of the Z-channel. Thus, recalling that the capacity of the Z-channel is $H(1/5) - 2/5$, we have

$$2C_{\text{FB}} = H(1/5) - \frac{2}{5}.$$

and therefore

$$C_{\text{FB}} = \frac{H(1/5)}{2} - \frac{1}{5} \approx 0.161.$$

Note that for this channel, feedback does increase capacity.

8. *Jointly typical sequences.* Let X and Y be jointly binary random variables. Let $X \sim \text{Bern}(1/2)$ be the input to a binary symmetric channel with crossover probability p and Y the output. That is, the joint distribution $p(x, y)$ is given by

$X \setminus Y$	0	1
0	$\frac{1}{2}(1-p)$	$\frac{1}{2}p$
1	$\frac{1}{2}p$	$\frac{1}{2}(1-p)$

Note that the marginal distribution of Y is also Bern(1/2). Recall that the jointly typical set $A_\epsilon^{(n)}$ is defined as the set of pairs (x^n, y^n) that satisfy the three conditions

$$\left| -\frac{1}{n} \log p(x^n) - H(X) \right| \le \epsilon,$$
$$\left| -\frac{1}{n} \log p(y^n) - H(Y) \right| \le \epsilon,$$
$$\left| -\frac{1}{n} \log p(x^n, y^n) - H(X, Y) \right| \le \epsilon.$$

For a given pair of sequences (x^n, y^n), let k be the number of places in which the sequence x^n differs from y^n (i.e. the number of times a bit flip occurred). Prove that there exist numbers $k_{\min}$ and $k_{\max}$ where $(x^n, y^n) \in A_\epsilon^{(n)}$ if and only if

$$k_{\min} \le k \le k_{\max}.$$

What are these numbers $k_{\min}$ and $k_{\max}$?

Solution: Consider the first condition, concerning X. Since X is distributed according to Bern(1/2), $H(X) = 1$ and for each sequence $p(x^n) = 2^{-n}$. Thus

$$-\frac{1}{n}\log p(x^n) = -\frac{1}{n}\log 2^{-n} = -\frac{1}{n}(-n) = 1.$$

Hence, for all x^n, $-\frac{1}{n}\log p(x^n) = H(X)$, meaning the first condition actually holds for all x^n (i.e. all x^n sequences are typical). Similarly, the Y condition holds for all y^n.

Now consider the joint condition. Note that $H(X,Y) = H(X) + H(Y|X) = 1 + H(p)$. Moreover, since $p(y_i|x_i) = p$ if $y_i \neq x_i$ and $p(y_i|x_i) = 1-p$ if $y_i = x_i$, we have

$$\begin{aligned}
p(x^n, y^n) &= \prod_{i=1}^{n} p(x_i, y_i) \\
&= \prod_{i=1}^{n} p(x_i)p(y_i|x_i) \\
&= \prod_{i=1}^{n} \frac{1}{2}p(y_i|x_i) \\
&= 2^{-n}\left(\prod_{i:x_i\neq y_i} p\right)\left(\prod_{i:x_i=y_i}(1-p)\right) \\
&= 2^{-n}p^k(1-p)^{n-k}
\end{aligned}$$

where we have used the fact that k is the number of times where $x_i \neq y_i$, meaning that $n-k$ is the number of times where $x_i = y_i$. Thus for the third condition to hold we need

$$\begin{aligned}
\epsilon &\geq \left|-\frac{1}{n}\log p(x^n, y^n) - H(X,Y)\right| \\
&= \left|-\frac{1}{n}\log(2^{-n}p^k(1-p)^{n-k}) - 1 - H(p)\right| \\
&= \left|-\frac{1}{n}[-n + k\log p + (n-k)\log(1-p)] - 1 - H(p)\right| \\
&= \left|1 + \frac{k}{n}\log\frac{1-p}{p} - \log(1-p) - 1 - H(p)\right| \\
&= \left|\frac{k}{n}\log\frac{1-p}{p} - \log(1-p) + p\log p + (1-p)\log(1-p)\right| \\
&= \left|\frac{k}{n}\log\frac{1-p}{p} + p\log p - p\log(1-p)\right| \\
&= \left|\left(\frac{k}{n} - p\right)\log\frac{1-p}{p}\right|.
\end{aligned}$$

Therefore this condition holds if and only if

$$\left|\frac{k}{n} - p\right| \leq \frac{\epsilon}{\left|\log\frac{1-p}{p}\right|}$$

or equivalently $k_{\min} \leq k \leq k_{\max}$ where

$$\begin{aligned} k_{\min} &= np - \frac{n\epsilon}{\left|\log \frac{1-p}{p}\right|}, \\ k_{\max} &= np + \frac{n\epsilon}{\left|\log \frac{1-p}{p}\right|}. \end{aligned}$$