CS3236: Solutions to Tutorial 4
(Channel Coding)

Note 1. Throughout this tutorial, as usual $H_2(q) = q \log_2 \frac{1}{q} + (1-q) \log_2 \frac{1}{1-q}$ (binary entropy function).

Note 2. Some of the questions below represent the channel in matrix form as follows (assuming $\mathcal{X} = \{1, \ldots, N_{\mathrm{X}}\}$ and $\mathcal{Y} = \{1, \ldots, N_{\mathrm{Y}}\}$ for some alphabet sizes N_{X} and N_{Y}):

$$\begin{bmatrix} P_{Y|X}(1|1) & P_{Y|X}(1|2) & \ldots & P_{Y|X}(1|N_{\mathrm{X}}) \\ P_{Y|X}(2|1) & P_{Y|X}(2|2) & \ldots & P_{Y|X}(2|N_{\mathrm{X}}) \\ \vdots & \vdots & \ddots & \vdots \\ P_{Y|X}(N_{\mathrm{Y}}|1) & P_{Y|X}(N_{\mathrm{Y}}|2) & \ldots & P_{Y|X}(N_{\mathrm{Y}}|N_{\mathrm{X}}) \end{bmatrix}$$

The size of the matrix is $N_{\mathrm{Y}} \times N_{\mathrm{X}}$; rows correspond to output symbols, and columns correspond to input symbols. If you find this format confusing, you may want to draw the corresponding channel diagrams (and double-check that the sum of edges connected to each input is one).

Note 3. One convenient feature of the channel matrix form is that we can calculate the output distribution P_Y by multiplying the channel matrix (matrix $\times$ vector) by the input distribution vector:

$$\begin{bmatrix} P_Y(1) \\ P_Y(2) \\ \vdots \\ P_Y(N_{\mathrm{Y}}) \end{bmatrix} = \begin{bmatrix} P_{Y|X}(1|1) & P_{Y|X}(1|2) & \ldots & P_{Y|X}(1|N_{\mathrm{X}}) \\ P_{Y|X}(2|1) & P_{Y|X}(2|2) & \ldots & P_{Y|X}(2|N_{\mathrm{X}}) \\ \vdots & \vdots & \ddots & \vdots \\ P_{Y|X}(N_{\mathrm{Y}}|1) & P_{Y|X}(N_{\mathrm{Y}}|2) & \ldots & P_{Y|X}(N_{\mathrm{Y}}|N_{\mathrm{X}}) \end{bmatrix} \begin{bmatrix} P_X(1) \\ P_X(2) \\ \vdots \\ P_X(N_{\mathrm{X}}) \end{bmatrix}.$$

Part I – Finding the Channel Capacity

1. **[Four-Input Channel Capacity]**

   A channel $P_{Y|X}$ with input alphabet $\mathcal{X} = \{1, 2, 3, 4\}$ and output alphabet $\mathcal{Y} = \{1, 2, 3, 4\}$ has conditional probability matrix:

   $$Q = \begin{bmatrix} 1-\delta & \delta & 0 & 0 \\ \delta & 1-\delta & 0 & 0 \\ 0 & 0 & 1-\delta & \delta \\ 0 & 0 & \delta & 1-\delta \end{bmatrix}$$

   where $\forall j \in \mathcal{Y}, \forall i \in \mathcal{X} \ : \ Q(j,i) = \Pr(Y = j | X = i)$.

   (a) Calculate the capacity of the channel $P_{Y|X}$.

   (b) **(Harder)** Suppose that we have a binary codebook $\mathcal{C}$ of rate R that achieves error probability ϵ when used on a binary symmetric channel (BSC) with transition probability δ. Describe how to transmit over the above channel at rate $1 + R$ with error probability ϵ.

Solution. *(a) Let the input distribution to the channel be given by $P_X(1) = p_1$, $P_X(2) = p_2$, $P_X(3) = p_3$ and $P_X(4) = p_4$. Consider,*

$$\begin{aligned} H(Y|X) &= \sum_{x\in\mathcal{X}} P_X(x) H(Y|X=x) \\ &= \sum_{x\in\mathcal{X}} P_X(x) H_2(\delta) \\ &= H_2(\delta)\,. \end{aligned}$$

The capacity of the channel $P_{Y|X}$ is

$$\begin{aligned} C &= \max_{P_X} I(X;Y) \\ &= \max_{P_X} H(Y) - H(Y|X) \\ &= 2 - H_2(\delta), \end{aligned}$$

where the last step follows since $H(Y) \le 2$, and equality can be achieved by letting X be uniform (leading to uniform Y). Hence, $C = 2 - H_2(\delta)$.

(b) The number of codewords in $\mathcal{C}$ is $M = 2^{nR}$. To transmit one of $M' = 2^{n(1+R)}$ messages, identify each one with a length-n bit sequence $\mathbf{u} = (u_1, \ldots, u_n)$ and an index $m \in \{1, \ldots, M\}$. The number of combinations of $\mathbf{u}$ and m is $2^n \times 2^{nR} = 2^{n(1+R)}$, as required.

To transmit the $(\mathbf{u}, m)$ pair, do the following: In channel use i, if $u_i = 0$, then transmit $x_i = 1 + \widetilde{x}_i$, where $\widetilde{x}_i \in \{0, 1\}$ is the bit in the BSC codeword indexed by m; whereas if $u_i = 1$, then transmit $x_i = 3 + \widetilde{x}_i$.

The decoder can determine each u_i with certainty (it is 0 if $y_i \in \{1, 2\}$, and 1 if $y_i \in \{3, 4\}$). By mapping 1 and 3 to zero, and 2 and 4 to one, one recovers the binary sequence $(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ that can be decoded using the BSC decoder (with probability at most ϵ).

2. **[Capacity Calculation for Modulo Sum Channels]**

   For two positive integers k and m, let $(k \bmod m)$ be the *remainder* when k is divided by m.

   Find the capacity of the m-input discrete memoryless channel in which

   $$Y = (X + Z) \bmod m,$$

   where the input and output alphabets are $\mathcal{X} = \mathcal{Y} = \{0, 1, \ldots, m-1\}$, and

   $$\Pr(Z = 1) = \frac{3}{4}, \quad \Pr(Z = 0) = \frac{1}{4}.$$

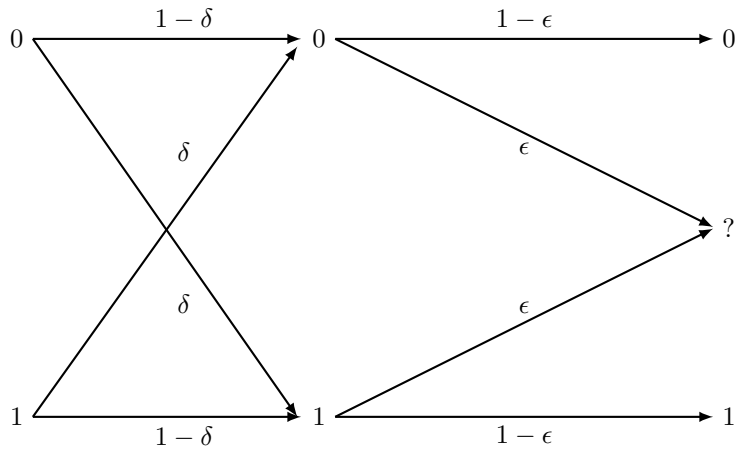
   Solution. *Consider*

   $$I(X;Y) = H(Y) - H(Y|X) = H(Y) - H_2(1/4)$$

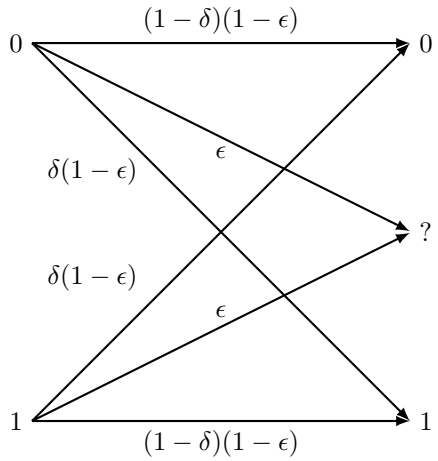
   Note that $H(Y)$ is maximized at the value $\log_2 m$, i.e., when Y is uniform on $\{0, 1, \ldots, m-1\}$. It is easy to check that this output distribution can be attained using the uniform input distribution: $P_X(x) = 1/m$ for all $x \in \{0, 1, \ldots, m-1\}$. Hence, the capacity is $\log_2 m - H_2(1/4)$.

3. **[Composition of Channels]**

   Let $\delta, \epsilon \in (0, 1)$. Consider a composition of a Binary Symmetric Channel followed by a Binary Erasure Channel with input alphabet $\mathcal{X} = \{0, 1\}$ and output alphabet $\mathcal{Y} = \{0, ?, 1\}$ and transition probabilities as shown below:



(a) Draw the transition probabilities diagram for the new composed channel.
Solution.



(b) Calculate the capacity of the new composed channel.
(Hint: Instead of computing $H(Y|X)$ directly, try letting let $E = \mathbf{1}\{Y =?\}$ and using $H(Y|X) = H(Y, E|X) = H(E|X) + H(Y|E, X)$. Similarly for $H(Y)$.)
Solution. *For the input distribution $\{p, 1-p\}$, the probabilities of Y given X are always $(1-\delta)(1-\epsilon)$, ϵ, and $\delta(1-\epsilon)$ (in some order). From this, we could compute $H(Y|X)$ directly, but it's simpler to let $E = \mathbf{1}\{Y =?\}$ and use $H(Y|X) = H(Y, E|X) = H(E|X) + H(Y|E, X)$ to get*

$$H(Y|X) = H_2(\epsilon) + (1-\epsilon)H_2(\delta),$$

which is also an instance of the decomposition property proved in Tutorial 1. We note that $H(Y|X)$ does not depend on p.
Now we find the distribution of Y:

$$\Pr(Y=0) = \Pr(X=0)\Pr(Y=0|X=0) + \Pr(X=1)\Pr(Y=0|X=1) = (1-\epsilon)(\delta + p - 2p\delta).$$
$$\Pr(Y=?) = \Pr(X=0)\Pr(Y=?|X=0) + \Pr(X=1)\Pr(Y=?|X=1) = \epsilon.$$
$$\Pr(Y=1) = \Pr(X=0)\Pr(Y=1|X=0) + \Pr(X=1)\Pr(Y=1|X=1) = (1-\epsilon)(1-\delta-p+2p\delta).$$

Again, we let $E = \mathbf{1}\{Y = ?\}$ *and use* $H(Y) = H(Y, E) = H(E) + H(Y|E)$ *to get*

$$H(Y) = H_2(\epsilon) + (1-\epsilon)H_2(\delta + p - 2p\delta).$$

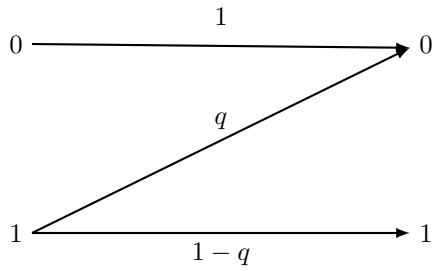
We see that $H(Y)$ *is maximized at* $p = 1/2$ *(to make the second* $H_2(\cdot)$ *have argument* $\frac{1}{2}$*) with maximum value* $H_2(\epsilon) + (1-\epsilon)$*. Thus, the capacity of the new composed channel is*

$$\begin{aligned} C &= \max_{P_X} I(X;Y) \\ &= \max_{p} H(Y) - H(Y|X) \\ &= H_2(\epsilon) + (1-\epsilon) - H_2(\epsilon) - (1-\epsilon)H_2(\delta) \\ &= (1-\epsilon)(1 - H_2(\delta)) \ . \end{aligned}$$

Note that when $\delta = 0$*, the above analysis gives an alternative proof for the BEC capacity based on* $I(X;Y) = H(Y) - H(Y|X)$ *(whereas in the lecture we used* $I(X;Y) = H(X) - H(X|Y)$*).*

4. **[Z Channel]**

Consider the Z channel: Input alphabet $\mathcal{X} = \{0, 1\}$, Output alphabet $\mathcal{Y} = \{0, 1\}$ and the transition probabilities are given by the following figure:

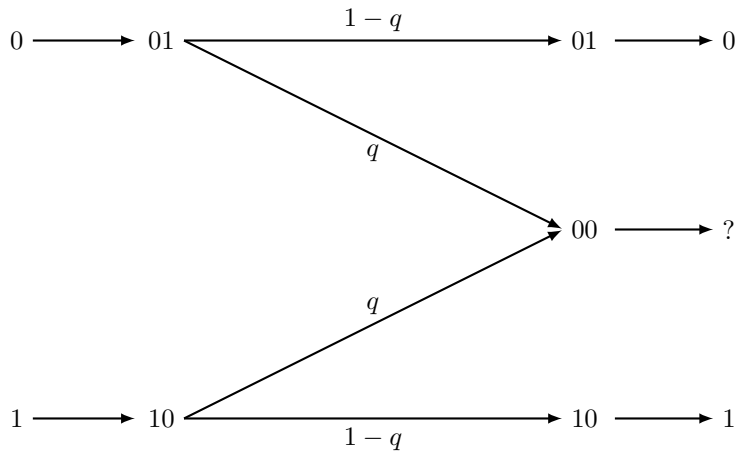


Show that two uses of a Z channel can be made to emulate one use of an erasure channel, and state the erasure probability of that erasure channel. Hence show that the capacity of the Z channel, C_Z, satisfies $C_Z \geq (1-q)/2$ bits.

(Note: If you want to take this question further, try calculating the exact capacity of the Z channel.)

Solution. *We encode* $x = 0$ *as* 01 *and* $x = 1$ *as* 10*. The 2-bit codeword is passed through two uses of a Z channel and we get* y*.*

We decode $y = 01$ *as* 0*,* $y = 00$ *as* ?*, and* $y = 10$ *as* 1*. The outcome* 11 *is impossible.*



Now, the overall probabilities are: With probability $1-q$, both bits remain the same, which equivalent to no erasure; and with probability q, one bit flip happens, which equivalent to an erasure.

Thus, the capacity of the Z channel satisfies

$$2 \times C_{\mathrm{Z}}(q) \;\geq\; C_{\mathrm{BEC}}(q) \;=\; 1-q\,,$$

$$C_{\mathrm{Z}}(q) \;\geq\; \frac{1}{2}(1-q)\,.$$

5. **[Yet Another Capacity Calculation]**

A channel $P_{Y|X}$ with input alphabet $\mathcal{X} = \{1,2,3\}$ and output alphabet $\mathcal{Y} = \{1,2,3,4\}$ has conditional probability matrix:

$$Q = \begin{bmatrix} 1/3 & 1/3 & 0 \\ 1/3 & 1/3 & 1/3 \\ 0 & 1/3 & 1/3 \\ 1/3 & 0 & 1/3 \end{bmatrix}$$

where $\forall j \in \mathcal{Y}, \forall i \in \mathcal{X} \;:\; Q(j,i) = \Pr(Y=j|X=i)$.

(a) Let the input distribution to the channel be given by $P_X(1) = 1/2, P_X(2) = 1/4, P_X(3) = 1/4$. Calculate the mutual information between random variables X and Y.

Solution. *We calculate the output distribution P_Y as follows:*

$$P_Y = Q \times P_X = \begin{bmatrix} 1/3 & 1/3 & 0 \\ 1/3 & 1/3 & 1/3 \\ 0 & 1/3 & 1/3 \\ 1/3 & 0 & 1/3 \end{bmatrix} \begin{bmatrix} 1/2 \\ 1/4 \\ 1/4 \end{bmatrix} = \begin{bmatrix} 1/4 \\ 1/3 \\ 1/6 \\ 1/4 \end{bmatrix}$$

Hence,

$$\begin{aligned} H(Y) \;&=\; \sum_{y \in \mathcal{Y}} P_Y(i) \log_2 \frac{1}{P_Y(y)} \\ &=\; \frac{1}{4}\log_2(4) + \frac{1}{3}\log_2(3) + \frac{1}{6}\log_2(6) + \frac{1}{4}\log_2(4) \\ &=\; \frac{7}{6} + \frac{1}{2}\log_2(3)\,. \end{aligned}$$

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= \frac{7}{6} + \frac{1}{2}\log_2(3) - \sum_{x\in\mathcal{X}} P_X(x) H(Y|X=x) \\ &= \frac{7}{6} + \frac{1}{2}\log_2(3) - \sum_{x\in\mathcal{X}} P_X(x)\log_2(3) \\ &= \frac{7}{6} - \frac{1}{2}\log_2(3)\,. \end{aligned}$$

(b) Calculate the capacity of the channel $P_{Y|X}$.

(Hint: (i) Let the input distribution be $P_X = (p_1, p_2, p_3)$. It is useful to express each entry of the output distribution in terms of some value of the form $1 - p_i$, rather than some sum $p_i + p_j$; (ii) When maximizing the function $a\log_2\frac{1}{a} + b\log_2\frac{1}{b} + c\log_2\frac{1}{c}$ with respect to non-negative integers (a, b, c) subject to the constraint $a + b + c = S$, the optimizing values are $a = b = c = \frac{S}{3}$. When $S = 1$, then recovers the property that the uniform distribution maximizes entropy.

Solution. *Let the input distribution to the channel be given by $P_X(1) = p_1$, $P_X(2) = p_2$ and $P_X(3) = p_3$. We can calculate distribution of output P_Y by*

$$P_Y = Q \times P_X = \begin{bmatrix} 1/3 & 1/3 & 0 \\ 1/3 & 1/3 & 1/3 \\ 0 & 1/3 & 1/3 \\ 1/3 & 0 & 1/3 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} = \begin{bmatrix} (1-p_3)/3 \\ 1/3 \\ (1-p_1)/3 \\ (1-p_2)/3 \end{bmatrix},$$

where we applied $p_1 + p_2 + p_3 = 1$ in each entry.

Consider

$$\begin{aligned} H(Y) &= \sum_{y\in\mathcal{Y}} P_Y(y)\log_2\frac{1}{P_Y(y)} \\ &= \frac{1}{3}\log_2(3) + \frac{1-p_1}{3}\log_2\frac{3}{1-p_1} + \frac{1-p_2}{3}\log_2\frac{3}{1-p_2} + \frac{1-p_3}{3}\log_2\frac{3}{1-p_3}. \end{aligned}$$

Observe that the last three terms take the form $a\log_2\frac{1}{a} + b\log_2\frac{1}{b} + c\log_2\frac{1}{c}$, and their values $a+b+c$ must sum to $\frac{2}{3}$, since $p_1 + p_2 + p_3 = 1$. Therefore, by the second part of the hint, the maximum is achieved when each $\frac{1-p_i}{3} = \frac{2}{9}$, which simplifies to $p_i = \frac{1}{3}$, and the maximum value is

$$\max_{p_1,p_2,p_3} H(Y) = \log_2 3 + \frac{2}{3}\log_2\frac{3}{2}.$$

Thus, the capacity of the channel $P_{Y|X}$ is,

$$\begin{aligned} C &= \max_{P_X} I(X;Y) \\ &= \max_{P_X} H(Y) - H(Y|X) \\ &= \max_{P_X} H(Y) - \sum_{x\in\mathcal{X}} P_X(x) H(Y|X=x) \\ &= \max_{P_X} H(Y) - \sum_{x\in\mathcal{X}} P_X(x)\log_2(3) \\ &= \max_{P_X} H(Y) - \log_2(3) \\ &= \frac{2}{3}\log_2(3/2)\,. \end{aligned}$$

Part II – Properties and Proofs

6. **[Possible Capacity Value]**

   State whether the following statement is TRUE or FALSE: There exists a discrete memoryless channel (DMC) with a binary (i.e., $|\mathcal{X}| = 2$ symbols) input alphabet and a ternary (i.e., $|\mathcal{Y}| = 3$ symbols) output alphabet such that its capacity is equal to $C = 1.5$ bits/channel use.

   Solution. *FALSE. This is because* $I(X;Y) \leq \min\{\log_2|\mathcal{X}|, \log_2|\mathcal{Y}|\}$ *so*

$$C \leq \min\{\log_2|\mathcal{X}|, \log_2|\mathcal{Y}|\} = \min\{1, \log_2 3\} = 1.$$

7. **[Futile Capacity Improvements]**

   Professor Xavier told you that he has found some ways to increase the capacity of a channel.

   (a) He says he has invented an algorithm $\mathcal{G}$ that changes the channel output by forming $\tilde{Y} = \mathcal{G}(Y)$ to obtain new channel $\hat{\mathcal{Q}} = \mathcal{Q} \circ \mathcal{G}$. He claims that this will strictly improve the capacity C of channel $\mathcal{Q}$ to capacity $\tilde{C}$ of channel $\hat{\mathcal{Q}}$ i.e $\tilde{C} > C$.

   Show that he is wrong.

   Solution. *Since* $\tilde{Y} = \mathcal{G}(Y)$, $X \longrightarrow Y \longrightarrow \tilde{Y}$ *forms a Markov chain. We can apply the data-processing inequality, to get* $I(X;Y) \geq I(X;\tilde{Y})$.

   Let P_X^* *be the distribution that maximizes* $I(X;\tilde{Y})$*, and let* $I(X;\tilde{Y})_{P_X^*}$ *denote mutual information when the* X *marginal has distribution* P_X^**. Then,*

$$\begin{aligned} C &= \max_{P_X} I(X;Y) \\ &\geq I(X;Y)_{P_X^*} \\ &\geq I(X;\tilde{Y})_{P_X^*} \\ &= \max_{P_X} I(X;\tilde{Y}) \\ &= \tilde{C}\ . \end{aligned}$$

   Thus, his first claim is wrong.

   (b) He also says that with the help of his friend, Professor Charles, that takes two independent observations at the output of the channel, he can strictly improve the capacity.

   Let Y_1 and Y_2 be two independent observations of same channel $P_{Y|X}$. This means that $\Pr(Y_1 = y_1, Y_2 = y_2|X = x) = \Pr(Y_1 = y_1|X = x)\Pr(Y_2 = y_2|X = x)$ where $\Pr(Y_1 = y_1|X = x)$ and $\Pr(Y_2 = y_2|X = x)$ both follow the conditional law $P_{Y|X}$. First, show that $I(X;Y_1,Y_2) = 2\,I(X;Y_1) - I(Y_1;Y_2)$. Let the capacity of the single observation channel $X \longrightarrow Y_1$ be C, and the capacity of the double observation channel $X \longrightarrow (Y_1,Y_2)$ be C'. Use the formula for channel capacity to show that $C' \leq 2\,C$.

   Show that his claim here is also wrong.

   Solution. *We have*

$$\begin{aligned} I(X;Y_1,Y_2) &= H(Y_1,Y_2) - H(Y_1,Y_2|X) \\ &\overset{(a)}{=} H(Y_1) + H(Y_2) - I(Y_1;Y_2) - H(Y_1|X) - H(Y_2|X) \\ &\overset{(b)}{=} I(X;Y_1) + I(X;Y_2) - I(Y_1;Y_2) \\ &\overset{(c)}{=} 2\,I(X;Y_1) - I(Y_1;Y_2)\ . \end{aligned}$$

where (a) uses $H(U,V) = H(U) + H(V) - I(U;V)$ and the fact that Y_1 and Y_2 are conditionally independent given X, (b) is just the definition of mutual information, and and (c) holds since Y_1 and Y_2 are conditionally identically distributed given X.

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

$$\begin{aligned} C' &= \max_{P_X} I(X; Y_1, Y_2) \\ &= \max_{P_X} 2\, I(X; Y_1) - I(Y_1; Y_2) \\ &\leq \max_{P_X} 2\, I(X; Y_1) \\ &= 2\, C. \end{aligned}$$

Thus, two independent observations cannot be more than twice as good as one observation, and once again, his claim is wrong.

8. **[Non-Uniform Capacity-Achieving Input Distribution]**

A channel $P_{Y|X}$ with input alphabet $\mathcal{X} = \{1, 2, 3\}$ and output alphabet $\mathcal{Y} = \{1, 2, 3\}$ has conditional probability matrix:

$$Q = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2/3 & 1/3 \\ 0 & 1/3 & 2/3 \end{bmatrix}$$

where $\forall j \in \mathcal{Y}, \forall i \in \mathcal{X} \ : \ Q(j,i) = \Pr(Y = j | X = i)$.

Calculate the optimal input distribution for achieving the capacity of the channel $P_{Y|X}$. You do not need to calculate the capacity itself (though an expression for it may arise in some form).

(Hint: You may assume that the capacity-achieving P_X satisfies $P_X(2) = P_X(3)$ due to the symmetry. Hence, let $P_X = (1 - 2p, p, p)$ for some p, then find $I(X;Y)$ and maximize it by differentiating with respect to p and setting to zero.)

Solution. *Let the input distribution to the channel be given by $P_X(1) = 1 - 2p$, $P_X(2) = p$ and $P_X(3) = p$, in accordance with the hint. We can calculate distribution of output P_Y by*

$$P_Y = Q \times P_X = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2/3 & 1/3 \\ 0 & 1/3 & 2/3 \end{bmatrix} \begin{bmatrix} 1-2p \\ p \\ p \end{bmatrix} = \begin{bmatrix} 1-2p \\ p \\ p \end{bmatrix}$$

Consider,

$$\begin{aligned} H(Y) &= \sum_{y \in \mathcal{Y}} P_Y(y) \log_2 \frac{1}{P_Y(y)} \\ &= 2p \log_2 \frac{1}{p} + (1-2p) \log_2 \frac{1}{1-2p}. \end{aligned}$$

$$\begin{aligned} H(Y|X) &= \sum_{x \in \mathcal{X}} P_X(x) H_2(Y|X = x) \\ &= pH_2\left(\frac{1}{3}\right) + (1-2p) H_2(0) + pH_2\left(\frac{1}{3}\right) \\ &= 2pH_2\left(\frac{1}{3}\right). \end{aligned}$$

The capacity of the channel $P_{Y|X}$ is

$$\begin{aligned} C &= \max_{P_X} I(X;Y) \\ &= \max_{P_X} H(Y) - H(Y|X) \\ &= \max_{p} \left[2p \log_2 \frac{1}{p} + (1-2p) \log_2 \frac{1}{1-2p} - 2pH_2\left(\frac{1}{3}\right) \right] . \end{aligned}$$

Denote the right-hand side by $I(p)$. To find the value of p at which $I(p)$ is maximized, we have to differentiate C with respect to p and equate it to 0. We have:

$$\begin{aligned} \frac{dI}{dp} &= 2\log_2 \frac{1}{p} + 2p \frac{d}{dp}\left(\log_2 \frac{1}{p}\right) + (-2)\log_2 \frac{1}{1-2p} + (1-2p)\frac{d}{dp}\left(\log_2 \frac{1}{1-2p}\right) - 2H_2\left(\frac{1}{3}\right) \\ &= 2\log_2 \frac{1}{p} - \frac{2}{\ln 2} + (-2)\log_2 \frac{1}{1-2p} + \frac{2}{\ln 2} - 2H_2\left(\frac{1}{3}\right) \\ &= 2\log_2 \frac{1-2p}{p} - 2H_2\left(\frac{1}{3}\right) . \end{aligned}$$

Setting to zero gives $\log_2 \frac{1-2p}{p} = H_2\left(\frac{1}{3}\right)$, or equivalently $\frac{1-2p}{p} = 2^{H_2(1/3)}$. Multiplying p on both sides and re-arranging gives the optimal value

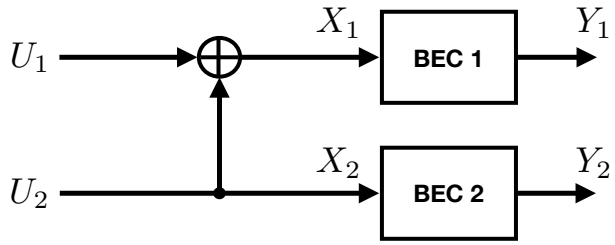
$$p = \frac{1}{2 + 2^{H_2(1/3)}},$$

and then the capacity achieving input distribution simply has probabilities $(1-2p, p, p)$.

9. **(Fairly Advanced) [A Step Towards Polar Codes]**

Consider the setup shown in the following illustration, where:

- The random variables U_1, U_2, X_1, X_2 take values on $\{0, 1\}$, whereas Y_1 and Y_2 take values on $\{0, e, 1\}$ with e representing an "erasure";
- U_1 and U_2 are independent, and equal 0 or 1 with probability $\frac{1}{2}$ each;
- We have $X_2 = U_2$, and $X_1 = U_1 \oplus U_2$, with $\oplus$ denoting modulo-2 addition;
- "BEC 1" and "BEC 2" are binary erasure channels, each having transition law $\mathbb{P}[Y_i = X_i] = 1 - \epsilon$ and $\mathbb{P}[Y_i = e] = \epsilon$ (for some $\epsilon \in (0, 1)$) with independence between the two channels.



We can express the joint mutual information $I(U_1, U_2; Y_1, Y_2)$ using the chain rule as

$$I(U_1, U_2; Y_1, Y_2) = I(U_1; Y_1, Y_2) + I(U_2; Y_1, Y_2 | U_1),$$

By carefully using the assumptions in the above four dot points, find exact expressions for both $I(U_1; Y_1, Y_2)$ and $I(U_2; Y_1, Y_2|U_1)$, writing your answer in terms of the erasure probability ϵ.

(Note: The answer shows that one of the mutual information terms is strictly above $1-\epsilon$ (the BEC capacity), and the other is strictly below $1-\epsilon$. This can be interpreted as forming one "stronger" channel and one "weaker" channel. By recursively applying this idea, we get something called a polar code (invented in the late 2000's). Roughly, the mappings from various U's to Y's create "channels", and compared to the original BEC's, some of those channels' capacity has increased and others have decreased. Remarkably, in the asymptot ic limit, a fraction $1-\epsilon$ of the channels approach a perfect channel (output = input), and a fraction ϵ of them approach a useless channel (output is independent of input). See `https://www.youtube.com/watch?v=VhyoZSB9g0w` for an excellent summary.)

Solution. *(i) For the first term, we write*

$$I(U_1; Y_1, Y_2) = H(U_1) - H(U_1|Y_1, Y_2) = 1 - H(U_1|Y_1, Y_2)$$

and observe the following:

- *If an erasure occurs in BEC 1, then the value of U_1 has no impact on either of the outputs. This implies that $H(U_1|Y_1 = e, Y_2 = y_2) = H(U_1) = 1$.*
- *Suppose an erasure occurs in BEC 2. Then given $Y_1 = y_1$ and $Y_2 = e$, U_1 is always equally likely to be 0 or 1, because whatever pair (u_1, u_2) produced the output y_1 would have also been produced by $(u_1', u_2') = (1 - u_1, 1 - u_2)$ (and both U_1 and U_2 are uniform). Hence, we again have $H(U_1|Y_1 = y_1, Y_2 = e) = 1$.*
- *Suppose that neither BEC 1 nor BEC 2 have an erasure. Then given $(Y_1, Y_2) = (y_1, y_2)$, we trivially have $X_1 = y_1$ and $X_2 = y_2$, from which we can deterministically produce $U_2 = X_2$ and $U_1 = X_1 \oplus X_2$. Hence, the outputs determine U_1, so we have $H(U_1|Y_1 = y_1, Y_2 = y_2) = 0$.*

Combining these cases, and noting that the third case has probability $(1-\epsilon)^2$, we obtain

$$H(U_1|Y_1, Y_2) = \sum_{y_1, y_2} P_{Y_1 Y_2}(y_1, y_2) H(U_1|Y_1 = y_1, Y_2 = y_2) = 1 - (1-\epsilon)^2$$

and hence $I(U_1; Y_1, Y_2) = (1-\epsilon)^2$.

(ii) For the second term, we use the independence of U_1 and U_2 to write

$$I(U_2; Y_1, Y_2|U_1) = H(U_2|U_1) - H(U_2|Y_1, Y_2, U_1) = 1 - H(U_2|Y_1, Y_2, U_1)$$

and observe the following:

- *Suppose an erasure occurs in both BEC 1 and BEC 2. Then the value of U_2 has no impact on the output, so its conditional distribution given (y_1, y_2, u_1) is uniform, and the corresponding conditional entropy is 1.*
- *Suppose that no erasure occurs in BEC 2. Then we must have $X_2 = Y_2$, from which we get $U_2 = X_2$, so that Y_2 determines U_2.*
- *Suppose that no erasure occurs in BEC 1. Then we must have $X_1 = Y_1$, from which we get $U_2 = X_1 \oplus U_1$, so that the pair (U_1, Y_1) determines U_2.*

Since the first case occurs with probability ε^2, we deduce that $H(U_2|Y_1, Y_2, U_1) = \epsilon^2$, which implies that $I(U_2; Y_1, Y_2|U_1) = 1 - \epsilon^2$.

Observe that the two mutual information terms sum to $1 - \epsilon^2 + (1-\epsilon)^2 = 2(1-\epsilon)$, the same value as $I(X_1, X_2; Y_1, Y_2)$. The fact that $I(U_1, U_2; Y_1, Y_2) = I(X_1, X_2; Y_1, Y_2)$ could have also been used to do only one of the above two calculations and then easily infer the other term via the chain rule.

10. **(Advanced) [Converse Bound for Bit Error Probability]**

This is Exercise 10.1, page 168 in MacKay's textbook.

In a variant of the noisy channel coding theorem described in this book, instead of generic messages $m \in \{1, \ldots, M\}$, we consider the message to be a sequence of $k = nR$ bits. The notion of error probability shown in the lecture corresponds to *block error probability*, meaning we get an error if *any* of the k bits comes out wrong. A less stringent notion is the *bit error probability* p_b, which is the proportion of bits flipped on average.

It can be shown (see MacKay's book) that if a probability of bit error p_b is acceptable, then rates up to $R(p_b)$ are achievable, where $R(p_\mathrm{b}) = \frac{C}{1-H_2(p_\mathrm{b})}$. Notice that, as one would expect, this rate approaches the capacity C as the bit error probability p_b approaches zero.

In this question, we will show that for any probability of bit error p_b, rates greater than $R(p_\mathrm{b}) = \frac{C}{1-H_2(p_\mathrm{b})}$ are not achievable.

Argument: Let $\mathbf{s} \in \{0,1\}^k$ be the string of bits, and $\hat{\mathbf{s}}$ its estimate. The source, encoder, noisy channel and decoder define a Markov chain: $\mathbf{s} \to \mathbf{x} \to \mathbf{y} \to \hat{\mathbf{s}}$.

The data processing inequality must apply to this chain: $I(\mathbf{s};\hat{\mathbf{s}}) \leq I(\mathbf{x};\mathbf{y})$: Furthermore, by the definition of channel capacity, $I(\mathbf{x};\mathbf{y}) \leq nC$, so $I(\mathbf{s};\hat{\mathbf{s}}) \leq nC$. Assume that a system achieves a rate R and a bit error probability p_b; then the mutual information $I(\mathbf{s};\hat{\mathbf{s}}) \geq nR(1 - H_2(p_b))$ (see below). Combining this with $I(\mathbf{s};\hat{\mathbf{s}}) \leq nC$ means that we must have $R \leq \frac{C}{1-H_2(p_b)}$, or in other words, it is impossible to have $R > \frac{C}{1-H_2(p_b)}$.

Fill in the details in the preceding argument. We already established $I(\mathbf{x};\mathbf{y}) \leq nC$ in the lecture, but why does the inequality $I(\mathbf{s};\hat{\mathbf{s}}) \geq nR(1 - H_2(p_b))$ hold?

(Hint: There are quite a few steps involved. Non-standard ones include $\frac{1}{k}\sum_{i=1}^{k} H_2(p_i) \leq H_2\big(\frac{1}{k}\sum_{i=1}^{k} p_i\big)$ (proved in an earlier tutorial, and can also be seen via Jensen's inequality) and $H(s_i|\hat{s}_i) = H_2(p_i)$ (should be easy to see). More standard ones include chain rule and conditioning reducing entropy.)

Solution. *To complete the argument, we want to show that $I(\boldsymbol{s};\hat{\boldsymbol{s}}) \;\geq\; k\,(1 - H_2(p_b))$, where $k = n\,R$ is the block size of $\boldsymbol{s}$ and $\hat{\boldsymbol{s}}$.*

Let s_i be the i-th bit of $\boldsymbol{s}$, $\hat{s}_i$ be the i-th bit of $\hat{\boldsymbol{s}}$, p_i be the probability that $s_i \neq \hat{s}_i$, and the bit error probability

$$p_b \;=\; \frac{1}{k}\sum_{i=1}^{k} \mathbb{P}[s_i \neq \hat{s}_i] \;=\; \frac{1}{k}\sum_{i=1}^{k} p_i.$$

Thus

$$\begin{aligned}
I(\boldsymbol{s}\,;\,\hat{\boldsymbol{s}}) &= H(\boldsymbol{s}) - H(\boldsymbol{s}\,|\,\hat{\boldsymbol{s}}) \\
&\overset{(a)}{=} k - H(\boldsymbol{s}\,|\,\hat{\boldsymbol{s}}) \\
&\overset{(b)}{=} k - \sum_{i=1}^{k} H(s_i\,|\,\hat{\boldsymbol{s}}, s_1, \ldots, s_{i-1}) \\
&\overset{(c)}{\geq} k - \sum_{i=1}^{k} H(s_i\,|\,\hat{s_i}) \\
&\overset{(d)}{\geq} k - \sum_{i=1}^{k} H_2(\,p_i\,) \\
&\overset{(e)}{=} k - k\cdot\frac{1}{k}\sum_{i=1}^{k} H_2(\,p_i\,)
\end{aligned}$$

$$\stackrel{(f)}{\geq} k - k\, H_2\left(\frac{1}{k}\sum_{i=1}^{k} p_i\right)$$
$$\stackrel{(g)}{=} k - k\, H_2(\, p_b\,)$$
$$= k\,(1 - H_2(p_b)),$$

where (a) is since $\boldsymbol{s}$ *is chosen uniformly, (b) is the chain rule, (c) is since conditioning reduces entropy, (d) is by the first hint, (e) is trivial, (f) is by the second hint, and (g) is by the definition of* p_b*.*

11. **(Advanced) [Alternative Proof of Channel Coding Achievability]**

In class, we saw how to do typical set decoding and proved that for all rates R smaller than capacity $C = \max_{P_X} I(X;Y)$, there exist codes of (some) length n with $M = 2^{nR}$ codewords and arbitrarily small error probability. Here, we consider an alternative proof that has the advantage of extending immediately to continuous-alphabet channels (and, although we will not show it, can provide refined asymptotics quantifying how fast the rate can converge to C as the block length n increases).

Let $\mathcal{X}$ and $\mathcal{Y}$ be the input and output alphabets of a channel. Unlike with the analysis in class, the alphabets here need not be finite. Let $P_{Y|X}$ be a channel from $\mathcal{X}$ to $\mathcal{Y}$, and let $P_{\mathbf{Y}|\mathbf{X}}$ be the joint conditional distribution when using the channel n times.

(a) Show that there exists a code with M codewords with average error probability P_{e} satisfying

$$P_{\mathrm{e}} \leq \Pr\left(\log_2 \frac{P_{\mathbf{Y}|\mathbf{X}}(\mathbf{Y}|\mathbf{X})}{P_{\mathbf{Y}}(\mathbf{Y})} \leq \log_2 M + \gamma\right) + 2^{-\gamma}.$$

for any choice of $\gamma > 0$ and any distribution P_X, where $P_{\mathbf{Y}}(\mathbf{y}) = \sum_{\mathbf{x}} P_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) P_{\mathbf{X}}(\mathbf{x})$ and $P_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} P_X(x_i)$.
(Hint: Randomly generate the codewords independently using P_X*, like in class. Instead of using typical set decoding, let the decoder output* $\hat{m} \in \{1, \ldots, M\}$ *if it is the unique one satisfying*

$$\log_2 \frac{P_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}^{(\hat{m})})}{P_{\mathbf{Y}}(\mathbf{y})} \geq \log_2 M + \gamma$$

If there is no $\hat{m}$ *satisfying the above condition, or if multiple exist, then we adopt a pessimistic view and assume that an error occurred. The analysis to arrive at the bound above is quite similar to typical set decoding, but getting the* $2^{-\gamma}$ *term requires some thought; try using the fact that* $\log_2 \frac{P_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}^{(\hat{m})})}{P_{\mathbf{Y}}(\mathbf{y})} \geq \log_2 M + \gamma$ *is equivalent to* $P_{\mathbf{Y}}(\mathbf{y}) \leq P_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}^{(\hat{m})}) \times \frac{2^{-\gamma}}{M}$*. A stronger version of this bound (for maximum error) was shown by Feinstein.)*
Solution. *As provided in the hint, we generate* M *codewords* $\mathbf{x}^{(m)}$ *with symbols drawn independently from* P_X*. To send message* m*, transmit codeword* $\mathbf{x}^{(m)}$*. Decode using the rule given above. Without loss of generality, suppose that* $m = 1$*. We make an error if and only if one or more of the following events occurs:*

$$\mathcal{E}_1 := \left\{\log_2 \frac{P_{\mathbf{Y}|\mathbf{Y}}(\mathbf{Y}|\mathbf{X}^{(1)})}{P_{\mathbf{Y}}(\mathbf{Y})} < \log_2 M + \gamma\right\}$$
$$\mathcal{E}_2 := \left\{\exists\, \tilde{m} \neq 1 : \log_2 \frac{P_{\mathbf{Y}|\mathbf{Y}}(\mathbf{Y}|\mathbf{X}^{(\tilde{m})})}{P_{\mathbf{Y}}(\mathbf{Y})} \geq \log_2 M + \gamma\right\}.$$

The probability of error can be bounded as

$$\Pr(\mathcal{E}) \leq \Pr(\mathcal{E}_1) + \Pr(\mathcal{E}_2)$$

Now note that $(\mathbf{X}^{(1)}, \mathbf{Y}) \sim P_{\mathbf{X}} P_{\mathbf{Y}|\mathbf{X}}$ *and so* $\Pr(\mathcal{E}_1)$ *gives the first term in the bound we have to show. We simply have to show that* $\Pr(\mathcal{E}_2) \leq 2^{-n\gamma}$*. For this consider*

$$\begin{aligned}
\Pr(\mathcal{E}_2) &= \Pr\left(\exists \tilde{m} \neq 1 : \log_2 \frac{P_{\mathbf{Y}|\mathbf{Y}}(\mathbf{Y}|\mathbf{X}^{(\tilde{m})})}{P_{\mathbf{Y}}(\mathbf{Y})} \geq \log_2 M + \gamma\right) \\
&\overset{(a)}{\leq} \sum_{\tilde{m}=2}^{M} \Pr\left(\log_2 \frac{P_{\mathbf{Y}|\mathbf{Y}}(\mathbf{Y}|\mathbf{X}^{(\tilde{m})})}{P_{\mathbf{Y}}(\mathbf{Y})} \geq \log_2 M + \gamma\right) \\
&\overset{(b)}{=} \sum_{\tilde{m}=2}^{M} \sum_{\mathbf{x},\mathbf{y}} P_{\mathbf{X}}(\mathbf{x}) P_{\mathbf{Y}}(\mathbf{y}) \mathbf{1}\left\{\log_2 \frac{P_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})}{P_{\mathbf{Y}}(\mathbf{y})} \geq \log_2 M + \gamma\right\} \\
&\overset{(c)}{\leq} \sum_{\tilde{m}=2}^{M} \sum_{\mathbf{x},\mathbf{y}} P_{\mathbf{X}}(\mathbf{x}) P_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) \frac{2^{-\gamma}}{M} \mathbf{1}\left\{\log_2 \frac{P_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})}{P_{\mathbf{Y}}(\mathbf{y})} \geq \log_2 M + \gamma\right\} \\
&\overset{(d)}{\leq} \sum_{\tilde{m}=2}^{M} \sum_{\mathbf{x},\mathbf{y}} P_{\mathbf{X}}(\mathbf{x}) P_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) \frac{2^{-\gamma}}{M} \\
&\overset{(e)}{\leq} 2^{-\gamma}
\end{aligned}$$

where (a) is due to the union bound, (b) due to the fact that for $\tilde{m} \neq 1$*, the codeword* $\mathbf{X}^{(\tilde{m})}$ *and channel output* $\mathbf{Y}$ *are independent, (c) due to the fact that we're only summing over all* $(\mathbf{x}, \mathbf{y})$ *such that* $\log_2 \frac{P_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})}{P_{\mathbf{Y}}(\mathbf{y})} \geq \log_2 M + \gamma$ *(see also the hint), (d) simply drops the indicator, and (e) use the fact that* $\sum_{x,y} P_{\mathbf{X}}(\mathbf{x}) P_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) = 1$ *and there are* $M - 1$ *terms in the outer sum.*

(b) Based on part (a), prove the channel coding theorem for finite $\mathcal{X}, \mathcal{Y}$ and memoryless channels. *(Hint: Choose* P_X *to be a capacity-achieving input distribution* $P_X \in \arg\max_{P_X} I(X;Y)$*. Also note that taking the product of* $P_{\mathbf{X}}$ *and* $P_{\mathbf{Y}|\mathbf{X}}$ *gives* $P_{\mathbf{XY}}(\mathbf{x}, \mathbf{y}) = \prod_{i=1}^{n} P_{XY}(x_i, y_i)$*; writing this as* $P_{\mathbf{XY}}(\mathbf{x}, \mathbf{y}) = \big(\prod_{i=1}^{n} P_Y(y_i)\big)\big(\prod_{i=1}^{n} P_{X|Y}(x_i|y_i)\big)$ *and summing over all* $\mathbf{x}$ *gives* $P_{\mathbf{Y}}(\mathbf{y}) = \prod_{i=1}^{n} P_Y(y_i)$*. Set* γ *above to be* $n\gamma'$ *for some* $\gamma' > 0$*. Set* $\log_2 M = n(C - 2\gamma')$*. Apply the law of large numbers to the first term to see that there exist codes with* $2^{n(C-2\gamma')}$ *codewords and vanishing average error probability as* $n \to \infty$*.)*

Solution. *Choose* $P_{\mathbf{X}}$ *to be the n-fold product distribution corresponding to a capacity-achieving input distribution* $P_X \in \arg\max_{P_X} I(X;Y)$*. Since channel is a DMC, both* $P_{\mathbf{Y}|\mathbf{X}}$ *and* $P_{\mathbf{Y}}$ *have a product distribution (see the hint), we have*

$$\begin{aligned}
\Pr\left(\log_2 \frac{P_{\mathbf{Y}|\mathbf{X}}(\mathbf{Y}|\mathbf{X})}{P_{\mathbf{Y}}(\mathbf{Y})} \leq \log_2 M + \gamma\right) &= \Pr\left(\sum_{i=1}^{n} \log_2 \frac{P_{Y|X}(Y_i|X_i)}{P_Y(Y_i)} \leq \log_2 M + n\gamma'\right) \\
&= \Pr\left(\sum_{i=1}^{n} \log_2 \frac{P_{Y|X}(Y_i|X_i)}{P_Y(Y_i)} \leq n(C - 2\gamma') + n\gamma'\right) \\
&= \Pr\left(\frac{1}{n}\sum_{i=1}^{n} \log_2 \frac{P_{Y|X}(Y_i|X_i)}{P_Y(Y_i)} \leq C - \gamma'\right).
\end{aligned}$$

Since

$$\mathbb{E}\left[\log_2 \frac{P_{Y|X}(Y_i|X_i)}{P_Y(Y_i)}\right] = I(X;Y) = C$$

for all i*, we have that the probability above tends to zero by the weak law of large numbers. Clearly, the second term in the bound* $2^{-\gamma} = 2^{-n\gamma'}$ *also tends to zero because* $\gamma' > 0$*. So we have*

demonstrated the existence of codes for which $P_{\mathrm{e}} \to 0$ *and the code rate is* $C - 2\gamma'$ *(recall that* $R = \frac{1}{n}\log_2 M$*, and we chose* $\log_2 M = n(C - 2\gamma')$*) which is arbitrarily close to* C *since* γ' *may be arbitrarily small.*