CS3236: 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 $\mathcal{X} = \{1,\dots,N_{\rm X}\}$ and $\mathcal{Y} = \{1,\dots,N_{\rm Y}\}$ for some alphabet sizes $N_{\rm X}$ and $N_{\rm Y}$):

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

The size of the matrix is $N_{\rm Y} \times N_{\rm X}$; 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:

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

Part I (Week 1 of 2) – 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.

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

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

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

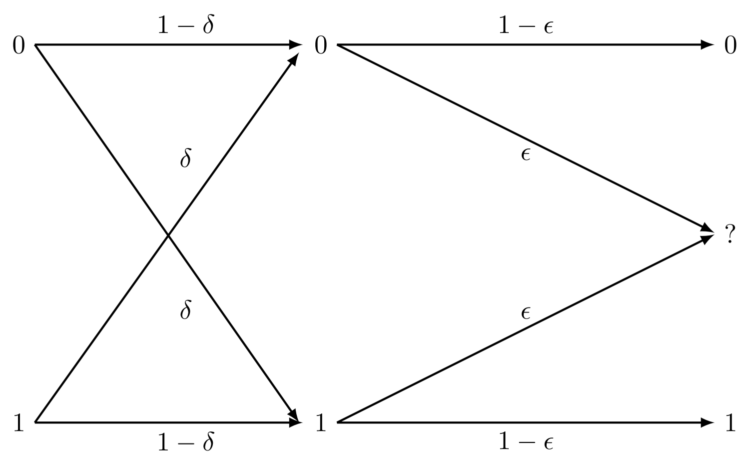
$$Y = (X + Z) \bmod m,$$

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

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

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:

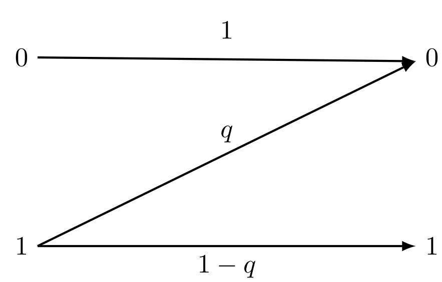


   (a) Draw the transition probabilities diagram for the new composed channel.

   (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 $H(Y)$.)

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.)

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.

   (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.

Hints for Part I

1. For (a) use the BSC analysis as a template. For (b) the idea is to get 1 bit "for free" by choosing between the first two vs. last two inputs (these two cases can be distinguished perfectly by the decoder), and then the remaining R bits using a BSC code.

2. Again use the BSC analysis as a template.

3. Instead of computing $H(Y|X)$ directly, try computing it as $H(Y|X) = H(Y,E|X) = H(E|X) + H(Y|E,X)$ where $E = \mathbf{1}\{Y = ?\}$. Similarly for $H(Y)$.

4. Encode $x = 0$ as 01 and $x = 1$ as 10. How can we then decode?

5. In (a) use the matrix multiplication idea noted at the start of the tutorial. In (b) use the hint to show that letting $p_i = 1/3$ for $i = 1,2,3$ is optimal.

Part II (Week 2 of 2) – 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.

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.

(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.

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

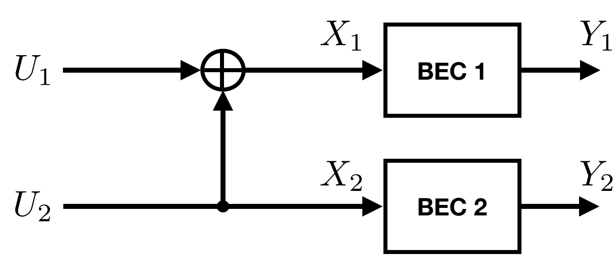
Calculate the optimal input distribution for achieving the capacity of the channel $P_{Y|X}$. You do not need to calculate the capacity itself (though an expression for it may arise in some form).

(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.)

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.)

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 $m \in \{1, \dots, M\}$, 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 $R(p_{\rm b}) = \frac{C}{1-H_2(p_{\rm b})}$. Notice that, as one would expect, this rate approaches the capacity C as the bit error probability $p_{\rm b}$ approaches zero.

    In this question, we will show that for any probability of bit error $p_{\rm b}$, rates greater than $R(p_{\rm b}) = \frac{C}{1-H_2(p_{\rm b})}$ 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: $I(\mathbf{s}; \hat{\mathbf{s}}) \le I(\mathbf{x}; \mathbf{y})$: Furthermore, by the definition of channel capacity, $I(\mathbf{x}; \mathbf{y}) \le nC$, so $I(\mathbf{s}; \hat{\mathbf{s}}) \le nC$. Assume that a system achieves a rate R and a bit error probability p_b; then the mutual information $I(\mathbf{s}; \hat{\mathbf{s}}) \ge nR(1 - H_2(p_b))$ (see below). Combining this with $I(\mathbf{s}; \hat{\mathbf{s}}) \le nC$ means that we must have $R \le \frac{C}{1-H_2(p_b)}$, 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 $I(\mathbf{x}; \mathbf{y}) \le nC$ in the lecture, but why does the inequality $I(\mathbf{s}; \hat{\mathbf{s}}) \ge nR(1 - H_2(p_b))$ hold?

    (Hint: There are quite a few steps involved. Non-standard ones include $\frac{1}{k}\sum_{i=1}^k H_2(p_i) \le H_2\big(\frac{1}{k}\sum_{i=1}^k p_i\big)$ (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.)

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 $P_{\rm e}$ satisfying

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

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 $\hat{m} \in \{1, \dots, M\}$ 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.)

(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 $n \to \infty$.)

Hints for Part II

6. Relate the capacity to $\log_2 |\mathcal{X}|$.

7. In (a) use the formula for capacity and the data processing inequality. In (b) use the identity $H(U, V) = H(U) + H(V) - I(U; V)$ and some further manipulations to show that $I(X; Y_1, Y_2) = 2\, I(X; Y_1) - I(Y_1; Y_2)$. Note that $I(Y_1; Y_2) \geq 0$.

8. Use the hint given and some direct calculations.

9. (i) For the first term, show that $I(U_1; Y_1, Y_2) = 1 - H(U_1|Y_1, Y_2)$, and compute the entropy by considering three cases: Erasure in BEC 1, Erasure in BEC 2, and erasure in neither. This should lead you to $H(U_1|Y_1, Y_2) = 1 - (1-\epsilon)^2$. (ii) For the second term, show that $I(U_2; Y_1, Y_2|U_1) = 1 - H(U_2|Y_1, Y_2, U_1)$, and consider three cases: Erasure in both BECs, no erasure in BEC 1, and no erasure in BEC 2. This should lead you to $H(U_2|Y_1, Y_2, U_1) = \epsilon^2$.

10. Hints given in the question.

11. Hints given in the question.