EL-GY 6063: Information Theory
Lecture 10

May 1, 2019

1 Channel Transition Probability

For a given channel, we often write the transition probability $P_{Y|X}$ in terms of a matrix called the transition probability matrix.

Definition 1. *For the channel characterized by $(\mathcal{X}, \mathcal{Y}, P_{Y|X})$, the transition probability matrix is defined as the matrix $P = [P_{x,y}]_{x,y \in \mathcal{X} \times \mathcal{Y}}$, where $P_{x,y} = P_{Y|X}(y|x)$.*

Note that for the transition probability matrix, the sum of each row is equal to one: $\sum_{y \in \mathcal{Y}} P_{x,y} = \sum_{y \in \mathcal{Y}} P_{Y|X}(y|x) = 1$.

Example 1. *For the binary symmetric channel with parameter p, we have*

$$P = \begin{bmatrix} 1-p & p \\ p & 1-p \end{bmatrix}.$$

For the binary erasure channel with parameter ϵ, we have:

$$P = \begin{bmatrix} 1-\epsilon & \epsilon & 0 \\ 0 & \epsilon & 1-\epsilon \end{bmatrix}.$$

Exercise. For the channel probability matrix $P = [P_{x,y}]_{(x,y) \in \mathcal{X} \times \mathcal{Y}}$, show that $\sum_{x,y \in \mathcal{X} \times \mathcal{Y}} = |\mathcal{X}|$.

2 Symmetric Channels

Symmetric channels are a category of channels for which the channel capacity formula can be easily calculated.

Definition 2. *The channel is called symmetric if every row of the transition probability matrix is a permutation of the other rows and each column is a permutation of the other columns.*

Example: A symmetric channel:

$$P = \begin{bmatrix} \frac{1}{3} & \frac{1}{2} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{3} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{6} & \frac{1}{3} \end{bmatrix}$$

Definition 3. *The channel is called weakly symmetric if if every row is a permutation of the other rows.*

A weakly symmetric channel:

$$P = \begin{bmatrix} \frac{1}{3} & \frac{1}{2} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{3} & \frac{1}{2}. \end{bmatrix}$$

Note that every symmetric channel is weakly symmetric, but the converse is not true.

Theorem 1. *For a weakly symmetric channel, the capacity is given by:*

$$C = \log|\mathcal{Y}| - H(Y|X = x), \forall x \in \mathcal{X}.$$

The capacity is achieved by taking P_X to be uniform over $\mathcal{X}$.

proof.

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) - \sum_{x\in\mathcal{X}} P_X(x)H(Y|X = x).$$

Note that $H(Y|X = x)$ is the same for all x since the channel is weakly symmetric. Take $x' \in \mathcal{X}$ arbitrarily. Then,

$$\begin{aligned} I(X;Y) = H(Y) - H(Y|X) = H(Y) - \sum_{x\in\mathcal{X}} P_X(x)H(Y|X = x) \\ = H(Y) - H(Y|X = x')\sum_{x\in\mathcal{X}} P_X(x) = H(Y) - H(Y|X = x'), \forall x' \in \mathcal{X}. \end{aligned}$$

Now let us consider the capacity formula:

$$C = max_{P_X} I(X;Y) = max_{P_X} H(Y) - H(Y|X = x') = (max_{P_X} H(Y)) - H(Y|X = x').$$

We know that $H(Y) \leq \log|\mathcal{Y}|$ with equality if and only if Y is uniformly distributed. We need to show that there exists a distribution P_X for which P_Y is uniformly distributed. Note that if we take X to be symmetric, we get $P_Y(y) = \sum_{x\in\mathcal{X}} P_X(x)P_{Y|X}(y|x) = \frac{1}{|\mathcal{X}|}\sum_{x\in\mathcal{X}} P_{Y|X}(y|x)$. Since the channel is weakly symmetric, we get that $\mathcal{X} P_{Y|X}(y|x)$ is the same for all y, so $P_Y(y)$ is the same for all y and $H(Y) = \log|\mathcal{Y}|$.

3 Properties of Channel Capacity

Property 1. $C \geq 0$ *and equality iff* $P_{Y|X} = P_Y$ *(i.e. Y is independent of X for every P_X.).*

Proof. Note that $C = maxI(X;Y)$. We know that $I(X;Y) \geq 0$ with equality iff X and Y are independent. So, C=0 iff X and Y are independent for every P_X.

Property 2. $C \leq \min(\log|X|, \log|Y|)$. *With equality iff either X is a function of Y or vice versa.*

Proof. We know that $I(X,Y) \leq min(H(X), H(Y)) \leq min(\log|X|, \log|Y|)$. This implies the above property.

4 Converse Proof

Theorem 2 (Shannon's Channel Coding Theorem)**.** *For the channel coding problem characterized by* $(\mathcal{X}, \mathcal{Y}, P_{Y|X})$, *the channel capacity is given as:*

$$C = max_{P_X} I(X;Y).$$

Lemma 1 (Single-letter Fano's Inequality)**.** *Let X and Y be two discrete random variables with joint distribution $P_{X,Y}$. Assume that Y is observed and the reconstruction $\hat{X} = f(Y)$ is produced, where $f : \mathcal{Y} \to \mathcal{X}$. Define $P_e = P(\hat{X} \neq X)$. Then,*

$$H(X|Y) \leq H(P_e, 1 - P_e) + P_e \log(|\mathcal{X}| - 1).$$

proof. Define E as the binary random variable which is equal to 1 if $X = \hat{X}$ and is equal to 0 otherwise. Then, E is clearly a Bernoulli random variable with parameter P_e. We have:

First, note that $H(E, X|Y) = H(X|Y) + H(E|X, Y)$. Note that $E = \nVdash(X = f(Y))$ is a function of X and Y, so $H(E|X, Y) = 0$. So, $H(E, X|Y) = H(X|Y)$. As a result, we have:

$$H(X|Y) = H(E, X|Y) = H(E|Y) + H(X|E, Y) \leq H(E) + H(X|E, Y),$$

where in the last inequality, we have used the fact that conditioning reduces entropy. Now, we have $H(X|E, Y) = P(E = 1)H(X|E = 1, Y) + P(E = 0)H(X|E = 0, Y)$. Note that if $E = 0$, then $X = \hat{X} = f(Y)$. So, $H(X|E = 0, Y) = 0$. Hence,

$$H(X|Y) = H(E, X|Y) = H(E|Y) + H(X|E, Y) \leq H(E) + P(E = 1)H(X|E = 1, Y),$$

Note that $H(X|E = 1, Y) \leq \log(|\mathcal{X}| - 1)$ since given that $E = 1$, X takes values from the alphabet $\mathcal{X} - \{f(Y)\}$ which has $|\mathcal{X}|$ elements. This completes the proof.

Lemma 2 (Multi-letter Fano's Inequality)**.** *Let X^n and Y^n be two correlated sequences of i.i.d random vectors with joint distribution $P_{X,Y}$. Assume that Y^n is observed and the reconstruction $\hat{X}^n = f(Y^n)$ is produced, where $f : \mathcal{Y}^n \to \mathcal{X}^n$. Define $P_e = P(\hat{X}^n \neq X^n)$. Then,*

$$\frac{1}{n}H(X^n|Y^n) \leq \frac{1}{n}H(P_e, 1 - P_e) + P_e \log|\mathcal{X}|.$$

The proof is similar to that of Lemma 1.

We are set to prove the converse of the channel coding theorem.

Proof. Fix $R > 0$. Assume that the rate R is achievable for the channel $(\mathcal{X}, \mathcal{Y}, P_{Y|X})$. This implies that for any $\epsilon > 0$, there exists $k, n \in \mathbb{N}$ such that $R \leq \frac{k}{n}$ and a coding strategy (e_n, f_n) such that $P_e \leq \epsilon$. From the multi-letter Fano's inequality, we have:

$$\frac{1}{k}H(Z^k|Y^n) \leq \frac{1}{k}H(P_e, 1 - P_e) + P_e \log|\mathcal{Z}| \triangleq \delta_\epsilon.$$

Obeserve that $\delta_\epsilon \to 0$ as $P_e \to 0$. Also, $H(Z^k) = k \geq nR$. Hence,

$$nR \leq H(Z^k) = H(Z^k|Y^n) + I(Z^k; Y^n) \leq k\delta_\epsilon + I(Z^k; Y^n). \tag{1}$$

Note that $X^n = e_n(Z^k)$. So, $I(Z^k; Y^n) = I(X^n, Z^k; Y^n)$. On the other hand from the channel model, we have the Markov chain $Z^k \leftrightarrow X^n \leftrightarrow Z^n$. So, $I(X^n, Z^k; Y^n) = I(X^n; Y^n)$. Combining these two facts, we get $I(Z^k; Y^n) = I(X^n; Y^n)$. So,

$$I(Z^k; Y^n) = I(X^n; Y^n) = H(Y^n) - H(Y^n|X^n) = H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i),$$

where in the last equality we have used the stationary and memoryless property of the channel. Also, note that from previous lectures, $H(Y^n) \leq \sum_{i=1}^{n} H(Y_i)$. So,

$$I(Z^k; Y^n) \leq \sum_{i=1}^{n}(H(Y_i) - H(Y_i|X_i)) = \sum_{i=1}^{n} I(Y_i; X_i).$$

Note that $I(Y_i; X_i) \leq max_{P_{X_i} I(X_i;Y_i)=C}$. So, $I(Z^k; Y^n) \leq nC$. As a result, from (1), we have:

$$nR \leq k\delta_\epsilon + nC \Rightarrow R \leq \frac{k}{n}\delta_\epsilon + C \Rightarrow R \leq \delta_\epsilon + C.$$

Note that $\delta_\epsilon \to 0$ as $\epsilon \to 0$. This implies $R \leq C$.

5 The Blahut-Arimoto Algorithm

So far, we have computed the capacity formula for the BSC, BEC and weakly symmetric channels. However, for more general channel transition probability matrices, the optimization is not as straightforward. For such cases, there are numerical algorithms which compute the capacity given a specific transition probability. The Blahut-Arimoto (BA) algorithm is an iterative algorithm which is used to compute the capacity formula.

In finding channel capacity, the goal is to maximize the function $I(P_X) = I(X;Y)$ over P_X. Note that:

$$I(P_X) = I(X;Y) = \sum_{x,y} P_{X,Y}(x,y) \log P_{Y|X(y|x)} P_Y(y) = \sum_{x,y} P_X(x) P_{Y|X}(y|x) \log \frac{P_{Y|X(y|x)}}{\sum_{x' \in \mathcal{X}} P_X(x') P_{Y|X}(y|x')},$$

where $P_{Y|X}$ is a fixed function.

The optimization over P_X is often computationally challenging. In the BA algorithm, an alternative way of expressing the objective function is used where the mutual information is written as $I(P_X, Q_{X|Y}) = I(X;Y)$. More precisely,

$$I(P_X, Q_{X|Y}) = \sum_{x,y} P_X(x) P_{Y|X}(x) \log \frac{Q_{X|Y}(x|y)}{P_X(x)}.$$

The algorithm optimizes the objective function on P_X and $Q_{X|Y}$ iteratively. This is explained in more detail next.
The BA algorithm operates in steps. In the first step, it fixes an initial $P_X^{(1)}$, and considers the optimization:

$$\max_{Q_{X|Y}} I(P_X^{(1)}, Q_{X|Y}) = \sum_{x,y} P_X^{(1)}(x) P_{Y|X}(x) \log \frac{Q_{X|Y}(x|y)}{P_X^{(1)}(x)}.$$

Let $Q_{X|Y}^{(2)}$ be an optimizing distribution in the above equation. In the second step, the BA algorithm fixes $Q_{X|Y}^{(2)}$ and optimizes

$$\max_{P_X} I(P_X, Q_{X|Y}^{(2)}) = \sum_{x,y} P_X(x) P_{Y|X}(x) \log \frac{Q_{X|Y}^{(2)}(x|y)}{P_X(x)}.$$

Let $P_X^{(2)}$ be an optimizing distribution in the above equation. In the third step, the algorithm fixes $P_X^{(2)}$ and optimizes over $Q_{X|Y}$. The process is iteratively repeated until the outputs of the optimization converge (e.g. until $|P_X^{(n)} - P_X^{(n-1)}| \leq \epsilon$ and $|Q_{X|Y}^{(n)} - Q_{X|Y}^{(n-1)}| \leq \epsilon$ for some $\epsilon > 0$).

The next three lemmas both prove the correctness of the BA algorithm and provide a simple method to perform the optimization:

Lemma 3. *Given a fixed input distribution $P_X^{(i)}$ and conditional distribution $P_{Y|X}$, for any conditional distribution $Q_{X|Y}$,*

$$I(P_X^{(i)}, Q_{X|Y}) \leq I(P_X^{(i)}) = \sum_{x,y} P_X^{(i)}(x) P_{Y|X}(y|x) \log \frac{P_{Y|X}(y|x)}{P_Y^{(i)}(y)},$$

where $P_Y^{(i)}(y) = \sum_x P_X^{(i)}(x) P_{Y|X}(y|x)$. Furthermore, equality holds if and only if:

$$Q_{X|Y} = P_{X|Y}^{(i)} = \frac{P_X^{(i)} P_{Y|X}}{P_Y^{(i)}}$$

proof.

$$I(P_X^{(i)}) - I(P_X^{(i)}, Q_{X|Y}) = \sum_{x,y} P_X^{(i)}(x) P_{Y|X}(y|x) \log \frac{P_{X|Y}^{(i)}(x|y)}{Q_{X|Y}(x|y)} = D(P_{X|Y}^{(i)} || Q_{X|Y}) \geq 0,$$

and equality iff $P^{(i)}_{X|Y} = Q_{X|Y}$ from the Gibbs Lemma.

As a corollary, we have that $\max_{P_X, Q_{X|Y}} I(X;Y) = \max_{P_X} I(X;Y) = C$. So the BA algorithm does compute the capacity provided that the result does converge to $\max_{P_X, Q_{X|Y}} I(X;Y)$.

Lemma 4. *Given a fixed conditional distribution $Q^{(i)}_{X|Y}$ and conditional distribution $P_{Y|X}$, for any distribution P_X,*

$$I(P_X, Q^{(i)}_{X|Y}) \leq \log \sum_x 2^{\sum_y P_{Y|X}(y|x) \log Q^{(i)}_{X|Y}(x|y)},$$

with equality if and only if:

$$P_X(x) = \frac{2^{\sum_y P_{Y|X}(y|x) \log Q^{(i)}_{X|Y}(x|y)}}{\sum_{x' \in \mathcal{X}} 2^{\sum_y P_{Y|X}(y|x') \log Q^{(i)}_{X|Y}(x'|y)}}$$

The proof follows by using the Lagrange multipliers method to find the optimizing P_X distribution.

From the above lemmas, we can simplify the BA algorithm as follows:

Initialization: Choose $P^{(0)}_X$ arbitrarily.
Step n: Set

$$Q^{(n)}_{X|Y}(y|x) = \frac{P^{(n-1)}_X(x) P_{Y|X}(y|x)}{P^{(n-1)}_Y(y)}, \qquad P^{(n)}_X(x) = \frac{2^{\sum_y P_{Y|X}(y|x) \log Q^{(n)}_{X|Y}(x|y)}}{\sum_{x' \in \mathcal{X}} 2^{\sum_y P_{Y|X}(y|x') \log Q^{(n)}_{X|Y}(x'|y)}}.$$

Termination: terminate if $|I(P^{(n)}_X, Q^{(n)}_{X|Y}) - I(P^{(n-1)}_X, Q^{(n)}_{X|Y})| < \epsilon$ for a suitably chosen $\epsilon > 0$.

6 Gaussian Channels

Definition 4. *Given the input and output alphabets $\mathcal{X} = \mathbb{R}$ and $\mathcal{Y} = \mathbb{R}$, a stationary and memoryless continuous channel (without feedback) is characterized by the transition probability distribution function $f_{Y|X}$.*

A Gaussian channel is specific continuous channel where the channel noise is additive and Gaussian.

Definition 5. *Let N be a Gaussian random variable with variance σ^2 and zero mean. The continuous channel characterized by $Y = X + N$ where X and N are independent is called the Gaussian channel.*

Often when transmitting data over a Gaussian channel we are given a power constraint. The power constraint states that $\frac{1}{n}\sum_{i=1}^n X_i^2 \leq P$, where $P > 0$. Similar to the discrete case, we could derive the following expression for the capacity of the Gaussian channel using weak typicality.

Theorem 3. *For the Gaussian Channel coding problem with noise variance σ^2 and power constraint P, the channel capacity is given as:*

$$C = max_{\mathbb{E}(X^2) \leq P} I(X;Y).$$

Note that $I(X;Y) = h_d(Y) - h_d(Y|X) = h_d(Y) - h_d(X+N|X) = h_d(Y) - h_d(N|X) = h_d(Y) - h_d(N)$. As a result, $C = max_{\mathbb{E}(X^2) \leq P} I(X;Y) = (max_{\mathbb{E}(X^2) \leq P} h_d(Y)) - h_d(N)$.

As a reminder, the entropy of a continuous random variable is upper bounded as follows $h_d(Y) \leq \frac{1}{2}\log 2\pi e \sigma_Y^2$. Assume that X is zero mean (otherwise it can be shown that f_X does not maximize the capacity formula), Note that $\sigma_Y^2 = \mathbb{E}((X+N)^2) = \mathbb{E}(X^2) + \mathbb{E}(N^2) = P + \sigma^2$. Taking X to be Gaussian with variance P, we get that $h_d(Y) = \frac{1}{2}\log 2\pi e \sigma_Y^2$ and $\mathbb{E}(X^2) \leq P$. So, this is the optimizing distribution for the capacity formula:

$$C = \frac{1}{2}\log 2\pi e(P + \sigma^2) - \frac{1}{2}\log 2\pi e \sigma^2 = \frac{1}{2}\log{(1 + \frac{P}{\sigma^2})}.$$