Lecture 8: Channel Capacity, Continuous Random Variables

Lecturer: Tsachy Weissman Scribe: Augustine Chemparathy, Adithya Ganesh, Philip Hwang

1 Channel Capacity

Given a channel with inputs X and outputs Y with transition probability $P(Y|X)$:

Define: Channel capacity C is the maximal rate of reliable communication (over memoryless channel characterized by $P(Y|X)$).

Further, recall the following definition:

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

Theorem.

$$C = C^{(I)}.$$

Proof: We will see this proof in the coming lectures.

- This theorem is important because C is challenging to optimize over, whereas $C^{(I)}$ is a tractable optimization problem.

1.1 Examples

Example I. Channel capacity of a Binary Symmetric Channel (BSC).

Define alphabets $\mathcal{X} = \mathcal{Y} = \{0, 1\}$. A BSC is defined by the PMF:

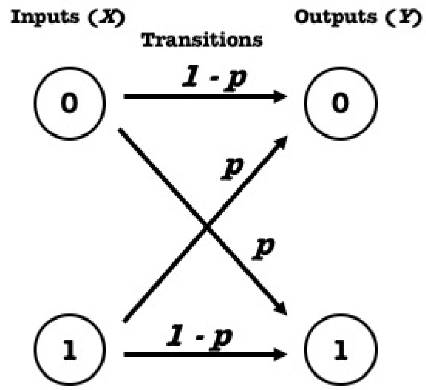
$$P_{Y|X}(y|x) = \begin{cases} p & y \neq x \\ 1-p & y = x. \end{cases}$$

This is equivalent to a channel matrix

$$\begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}$$

The rows of the matrix correspond to input symbols 0 and 1, while the columns correspond to output symbols 0 and 1.

And the graph representation



This can also be expressed in the form of additive noise.

$$Y = X \oplus_2 Z, \text{ where } Z \sim \text{Ber}(p) \text{ and } Z \text{ is independent of } X.$$

To determine the channel capacity of a BSC, by the theorem we must maximize the mutual information.

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(Y) - H(X \oplus_2 Z|X) \end{aligned}$$

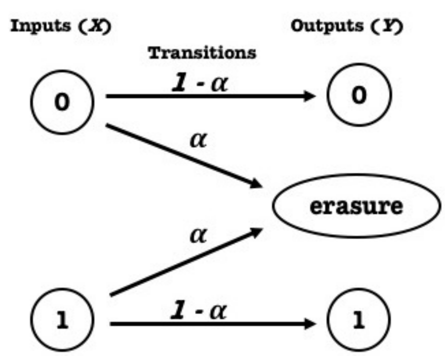
Once we condition on X, the uncertainty in $X \oplus_2 Z$ is same as the uncertainty in Z. Formally, we can simplify the second term (can also be shown by separately considering cases $X = 0$ and $X = 1$):

$$\begin{aligned} I(X;Y) &= H(Y) - H(Z) \\ &= H(Y) - h_2(p) \leq 1 - h_2(p). \end{aligned}$$

where $h_2(p)$ is the binary entropy function. Taking $X \sim \text{Ber}(\frac{1}{2})$ achieves equality: $I(X;Y) = 1 - h_2(p)$ (note: by symmetry $X \sim \text{Ber}(\frac{1}{2})$ implies $Y \sim \text{Ber}(\frac{1}{2})$). Thus, $C = 1 - h_2(p)$.

Example II. Channel capacity of a Binary Erasure Channel (BEC).

Define alphabets $\mathcal{X} = \{0,1\}$ and $\mathcal{Y} = \{0,1,e\}$ where e stands for erasure. Any input symbol X_i has a probability $1-\alpha$ of being retained in the output sequence and a probability α of being erased. Schematically, we have:



Examining the mutual information, we have that

$$\begin{aligned}
I(X;Y) &= H(X) - H(X|Y) \\
&= H(X) - [H(X|Y=e)P(Y=e) + H(X|Y=0)P(Y=0) + H(X|Y=1)P(Y=1)] \\
&= H(X) - [H(X)\cdot\alpha + 0\cdot P(Y=0) + 0\cdot P(Y=1)] \\
&= (1-\alpha)H(X)
\end{aligned}$$

Because the entropy of a binary variable can be no larger than 1:

$$(1-\alpha)H(X) \leq 1-\alpha$$

Equality is achieved when $H(X) = 1$, that is $X \sim \mathrm{Ber}(\frac{1}{2})$. Thus, the capacity of the BEC is $C = 1-\alpha$.

Note that if we knew exactly which positions were going to be erased, we could communicate at this rate by sending the input bits at exactly those positions (since the expected fraction of erasures is $1-\alpha$). The fact that $C = 1-\alpha$ indicates that we can achieve this rate even when we do not know which positions are going to be erased.

2 Information of Continuous Random Variables

The channel capacity theorem also holds for continuous valued channels, which are very important in a number of practical scenarios, e.g., in wireless communication. But before studying such channels, we need to extend notions like entropy and mutual information for continuous random variables.

Definition: The relative entropy between two probability density functions f and g is given by

$$D(f||g) = \int f(x)\log\frac{f(x)}{g(x)}\,dx.$$

Exercise: Show that $D(f||g) \geq 0$ with equality if and only if $f = g$.

Proof. Observe that that

$$\begin{aligned}
D(f||g) &= \int f(x)\log\frac{f(x)}{g(x)}dx \\
&= -\int f(x)\log\frac{g(x)}{f(x)}dx \\
&= -\mathbb{E}\left[\log\frac{g(x)}{f(x)}\right] \\
&\geq -\log\mathbb{E}\left[\frac{g(x)}{f(x)}\right] \\
&= -\log\int f(x)\frac{g(x)}{f(x)}\,dx \\
&= -\log 1 \\
&= 0.
\end{aligned}$$

Equality occurs in the manner of Jensen's when $f = g$.

Definition: The mutual information between X and Y that have a joint probability density function $f_{X,Y}$ is

$$I(X;Y) = D(f_{X,Y}||f_X f_Y).$$

Definition: The differential entropy of a continuous random variable X with probability density function f_X is

$$h(X) = -\int f_X(x) \log f_X(x)\, dx = \mathbb{E}\left[-\log f_X(X)\right]$$

If X, Y have joint density $f_{X,Y}$, the conditional differential entropy is

$$h(X|Y) = -\int f_{X,Y}(x,y) \log f_{X|Y}(x|y)\, dx\, dy = \mathbb{E}[-\log f_{X|Y}(X|Y)],$$

and the joint differential entropy is

$$h(X,Y) = -\int f_{X,Y}(x,y) \log f_{X,Y}(x,y)\, dx\, dy = \mathbb{E}[-\log f_{X,Y}(X,Y)].$$

2.1 Exercises

Exercise 1 . Show that

$$h(X|Y) \leq h(X)$$

with equality iff X and Y are independent.

Proof. This follows from exercise 2 below combined with the fact that $I(X;Y) \geq 0$ which holds since the relative entropy is non-negative (see exercise above). The equality holds exactly when $f_{X,Y}(x,y) = f_X(x)f_Y(y)$, i.e. when X and Y are independent.

Exercise 2. Show that

$$\begin{aligned} I(X;Y) &= h(X) - h(X|Y) \\ &= h(Y) - h(Y|X) \\ &= h(X) + h(Y) - h(X,Y). \end{aligned}$$

Proof.

$$\begin{aligned} I(X;Y) &= \int f_{X,Y}(x,y) \log \frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)} dxdy \\ &= \int f_{X,Y}(x,y) \log \frac{f_{X,Y}(x,y)}{f_X(x)} dxdy - \int f_{X,Y}(x,y) \log f_Y(y) dxdy \\ &= \int f_X(x) \left[\int f_{Y|X}(y|x) \log f_{Y|X}(y|x) dy\right] dx - \int f_Y(y) \log f_Y(y) dy \\ &= h(Y) - h(Y|X). \end{aligned}$$

Symmetrically the same can be shown for $I(X;Y) = H(X) - H(X|Y)$. Also

$$\begin{aligned} I(X;Y) &= \int f_{X,Y}(x,y) \log \frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)} dxdy \\ &= \int f_{X,Y}(x,y) \log f_{X,Y}(x,y) dxdy - \int f_{X,Y}(x,y) \log f_X(x) dxdy - \int f_{X,Y}(x,y) \log f_Y(y) dxdy \\ &= h(X,Y) - h(X) - h(Y). \end{aligned}$$

Exercise 3. Show that

$$h(X + c) = h(X).$$

and

$$h(c \cdot X) = h(X) + \log|c|, c \neq 0.$$

Proof. We will combine them and prove that $h(aX + b) = h(X) + \log|a|$ whenever $a \neq 0$. Let $Y = aX + b$. Then the pdf for Y can be written as

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)$$

Thus,

$$\begin{aligned} h(aX + b) = h(Y) &= -\int f_Y(y) \log f_Y(y)\, dy \\ &= -\int \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right) \log\left(\frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)\right) dy \end{aligned}$$

Changing variable to $x = (y - b)/a$ (and reversing limits if $a < 0$),

$$\begin{aligned} h(aX + b) &= -\int f_X(x) \log\left(\frac{1}{|a|} f_X(x)\right) dx \\ &= -\int f_X(x) \log\left(\frac{1}{|a|}\right) dx - \int f_X(x) \log\left(f_X(x)\right) dx \\ &= h(X) + \log|a| \end{aligned}$$

□

2.2 Examples

Example I: **Differential entropy of a uniform random variable $U \sim \mathbf{U}(a, b)$.**

- The pdf of an uniform random variable is

$$f_X(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

The differential entropy is simply:

$$h(X) = \mathbb{E}[-\log f_X(X)] = \log(b - a)$$

- Notice that the differential entropy can be negative or positive depending on whether $b - a$ is less than or greater than 1. In practice, because of this property, differential entropy is usually used as means to determine mutual information and does not have much operational significance by itself.

Example II: **Differential entropy of a Gaussian random variable $X \sim \mathcal{N}(0, \sigma^2)$.**

- The pdf of a Gaussian random variable is $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}x^2}$.

The differential entropy is:

$$h(X) = \mathbb{E}[-\log f(X)]$$

For simplicity, convert the base to e:

$$\begin{aligned} h(X) &= \frac{1}{\ln 2}\mathbb{E}[-\ln f(X)] \\ &= \frac{1}{\ln 2}\mathbb{E}\left[\frac{1}{2}\ln 2\pi\sigma^2 + \frac{1}{2\sigma^2}X^2\right] \\ &= \frac{1}{\ln 2}\left[\frac{1}{2}\ln 2\pi\sigma^2 + \mathbb{E}\left[\frac{1}{2\sigma^2}X^2\right]\right] \\ &= \frac{1}{\ln 2}\left[\frac{1}{2}\ln 2\pi\sigma^2 + \frac{1}{2\sigma^2}\sigma^2\right] \\ &= \frac{1}{\ln 2}\left[\frac{1}{2}\ln 2\pi e\sigma^2\right] = \frac{1}{2}\log 2\pi e\sigma^2 \end{aligned}$$

- Per Exercise 3, differential entropy is invariant to constant shifts. Therefore this expression represents the differential entropy of all Gaussian random variables regardless of mean.
- *Claim:* The Gaussian distribution has maximal differential entropy, i.e., if $X \sim f_X$ with second moment $E[X^2] \leq \sigma^2$ and $G \sim \mathcal{N}(0, \sigma^2)$ then $h(X) \leq h(G)$. Equality holds if and only if $X \sim \mathcal{N}(0, \sigma^2)$.

  Proof:

$$\begin{aligned} 0 \leq D(f_X \| G) &= \mathbb{E}\left[\log \frac{f_X(X)}{f_G(X)}\right] \\ &= -h(X) + \mathbb{E}\left[\log \frac{1}{f_G(X)}\right] \\ D(f_X \| G) &= -h(X) + \mathbb{E}\left[\log \frac{1}{\sqrt{2\pi\sigma^2}} + \frac{\frac{X^2}{2\sigma^2}}{\ln 2}\right] \end{aligned}$$

  Because the second moment of X is upper bounded by the second moment of G:

$$\begin{aligned} 0 \leq D(f_X \| G) &\leq -h(X) + \mathbb{E}\left[\log \frac{1}{\sqrt{2\pi\sigma^2}} + \frac{\frac{G^2}{2\sigma^2}}{\ln 2}\right] \\ &\leq -h(X) + \mathbb{E}\left[\log \frac{1}{f_G(G)}\right] = -h(X) + h(G) \end{aligned}$$

  Rearranging:

$$h(X) \leq h(G)$$

  Equality holds when $D(f_X \| G) = 0$, i.e., when $X \sim \mathcal{N}(0, \sigma^2)$. □

Example III: Channel capacity of an Additive White Gaussian Noise channel (AWGN) that is restricted by power P

The AWGN channel with parameter σ^2 has real input and output related as $Y_i = X_i + W_i$, where W_i's are iid $\sim \mathcal{N}(0, \sigma^2)$ (and W_i's are independent of X_i's).

- Power constraint: $\frac{1}{n}\sum_{i=1}^{n} X_i^2 \leq P$.
- The Channel Coding Theorem in this setting states that:

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

  Where $C(P)$ represents the 'capacity'; the maximal rate of reliable communication when constrained to power P. This maximization problem will be addressed in the next lecture.