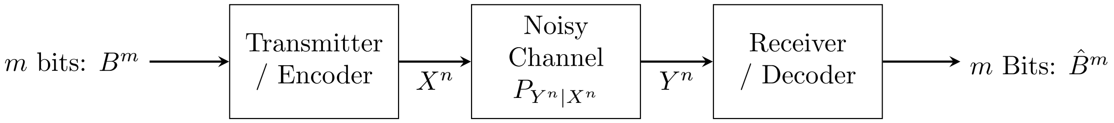
EE276 Information Theory Lecture 7 - 1/30/24

Lecture 7: Communication and Channel Capacity

Lecturer: Tsachy Weissman

1 Communication and Channel Capacity

We want to communicate bits from some source to some target. We use the following model to represent how to transmit this data across a channel that can have noise that corrupts the data. In order to ensure good transmission quality, we will try to encode and decode the data in such a way that reduces the probability of error in the received signal.



We have $B^m = (B_1, B_2, \ldots, B_m)$, and $\hat{B}^m = (\hat{B}_1, \hat{B}_2, \ldots, \hat{B}_m)$. Notice that m is not necessarily equal to n.

Here, let $B_1, B_2, ..., B_m$ be i.i.d. bits $\sim Ber(\frac{1}{2})$. The conditional distribution of the signal which the noisy channel emits given the transmitted signal, $P_{Y^n|X^n}(y^n|x^n)$, is given. In order to transmit B^m across the channel, we first encode these bits into a new vector, $X_1, X_2, ..., X_n$. This is the information that is sent across the noisy channel. Then, the receiver will receive the transformed vector, $Y_1, Y_2, ..., Y_n$. The decoder's job is now to to transform the received bits into a vector $\hat{B}^m$ that closely resembles B^m.

To work with this problem, we first need to make some definitions.

Definition 1. *Scheme:*

$$\textit{Scheme} \triangleq (m, n, \textit{encoder}, \textit{decoder}) \tag{1}$$

A scheme is first characterized by the number of bits we are trying to send (m) and the number of channel uses (n). Once we have picked these, we pick an encoder, that changes the (m) bits into an encoded (n)-tuple. Then, we pick the decoder that converts encoded (n)-tuple into, hopefully, the same (m) bits that were sent.

Definition 2. *Rate:*

$$\textit{rate} \triangleq \frac{m}{n} \frac{\textit{bits}}{\textit{channel use}} \tag{2}$$

The higher the rate, the better we are at communicating our information.

Definition 3. *Probability of Error:*

$$P_e^{(n)} \triangleq P(\hat{B}^m \neq B^m) \tag{3}$$

Definition 4. *Achievable Rate: R is an achievable rate if there exists a sequence of schemes $\{Scheme_n\}_{n\geq 1}$ with rate equal to or greater than R that transmits with vanishing probability of error, such that*

$$P_{e,Scheme_n}^{(n)} \xrightarrow{n\to\infty} 0 \tag{4}$$

Note that m has to be scaling and growing with n as it goes to infinity.

With the notion of achievable rate, we can talk about channel capacity.

Definition 5. *Channel Capacity: this is the maximal rate of reliable communication:*

$$C \triangleq \sup\{R | R \textit{ is achievable}\} \tag{5}$$

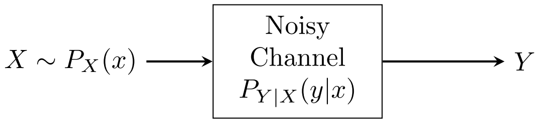
All these definitions are valid for any kind of channel. However, to get started, we choose to work with simpler channels. This brings us to the notion of a memoryless channel, which is a surprisingly common assumption to make in communication.

Definition 6. *Memoryless Channel: The conditional distribution of the output given the input is the product of the same conditional distribution of an output symbol given the input symbol, for every bit sent or received. In other words,*

$$P_{Y^n|X^n}(y^n|x^n) \triangleq \prod_{i=1}^{n} P_{Y|X}(y_i|x_i) \tag{6}$$

A memoryless channel corresponds to n independent uses of a channel that transmits a single symbol.

Now let's consider the "single-letter" channel.



Given the input distribution on X and the conditional distribution of Y given X, we can compute the joint distribution and quantities like the mutual information between the input and the output. In fact, since we can modify P_X, we can find a way to maximize the mutual information between the input and the output. Intuitively, maximizing the mutual information $I(X;Y)$ minimizes the error in the channel because it increases the amount by which the input informs the output. Consider then the following quantity, obtained by taking a maximum over the probability distributions of X:

$$C^{(I)} \triangleq \max_{P_X} I(X;Y) \tag{7}$$

Theorem 7. *The maximum mutual information is the channel capacity*

$$C = C^{(I)} \tag{8}$$

This is profound because it relates how much we can physically transmit over a channel reliably to the mutual information between input and output. This makes the complicated problem of finding channel capacity a clean optimization problem which involves finding the input distribution that maximizes the mutual information between the input and the output. We will see the 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:

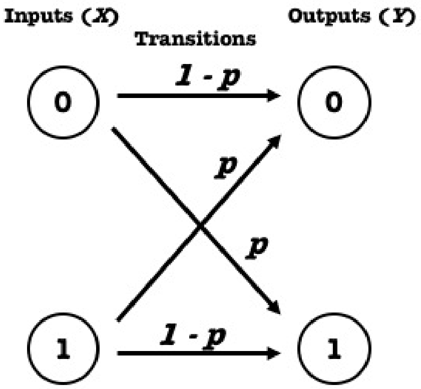
$$P_{Y|X}(y|x) = \begin{cases} p & y \neq x \\ 1-p & y = x. \end{cases}$$

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.

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

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

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

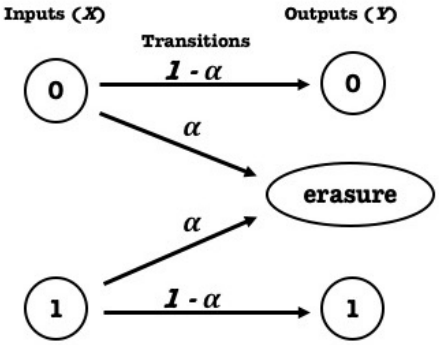
Once we condition on X, the uncertainty in $X \oplus_2 Z$ is same as the uncertainty in Z. Formally, we can simplify the second term (can also be shown by separately considering cases $X = 0$ and $X = 1$):

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

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

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

Equality is achieved when $H(X) = 1$, that is $X \sim \text{Ber}(\frac{1}{2})$. 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.