CS 229r Information Theory in Computer Science Feb 14, 2019

Lecture 6

Instructor: Madhu Sudan *Scribe: Xingchi Yan*

1 Overview

1.1 Outline for today

Channel Coding (Or Error Correction)

- Definitions: Rate, Capacity
- Coding Theorem for Binary Symmetric Channel (BSC)
- Coding Theroem for general channels
- Converse

2 Binary Symmetric Channel

Today we will move to a new topic which is channeling coding. Channeling coding is for correcting errors. And this is the second part of Shannon's 48 paper. The first part was about compressing information.

The simplest example is the Binary Symmetric Channel. The BSC(p) has some parameter where $0 \leq p < \frac{1}{2}$. What this channel does in communication is that it sends messages in bit X, what it receives is X with probability $1-p$ and $1-X$ with probability p. And this happens independently on every use of the channel.

In order to use such a channel, Shannon's idea was to encode the information before you send it and decode it afterwards. Suppose you have a message m that you want to send, you encode it and get some sequence X, maybe n such symbol X, get Y and then you want to decode.

And what we want is

$$\Pr[m \neq \hat{m}] \to 0, as\ n \to \infty$$

Intuitively, if you want to send longer and longer message, the error probability should be increasing. But here there is one thing that we want to use length of message to improve the reliability. And now we would like to understand how many uses of the channel do we have to do? What is the largest R?

The encoding sends k bits to n bits, our message would be some k bit string and encoding turns it to n bit string. The decoding takes n bits and brings it back down to k bits. $E_n : \{0,1\}^k \to \{0,1\}^n$, $D_n : \{0,1\}^n \to \{0,1\}^k$ where $k \leq n$.

Definition 1. *The rate would be* $Rate = \frac{k}{n}$

What we would really want to understand is the capacity of the channel. In this case the capacity of the channel would be

Definition 2.

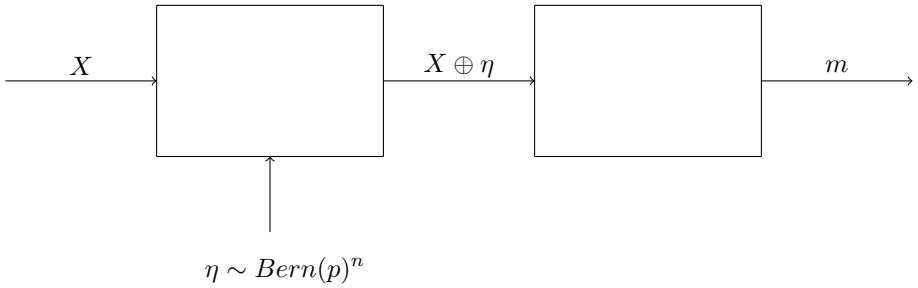
$$Capacity(BSC(p)) = \sup_{R}\{\lim_{\varepsilon\to 0}\lim_{n\to\infty}\{Rate\ R\ of\ E_n, D_n\}\}$$

such that

$$\Pr_{m,y|x,X=E(m)}[D(y^n) \neq m] \leq \varepsilon$$

So what is the best rate you can get? Allow to use n as large as you get but have to make sure that error goes to zero when n goes to infinity. This is the general quantity we want to understand for any channel.
Remark The capacity of binary symmetric channel is $Capacity(BSC(p)) = 1 - h(p)$ where $h(p) = plog\frac{1}{p} + (1-p)log\frac{1}{1-p}$ which is the entropy of Bernoulli p random variable.

This is a little striking theorem. Why we get the $1 - h(p)$? Roughly the idea is the following, X came in the channel with a sequence of Bernoulli random variables η distributed according to $Bern(p)$, after the decoding, suppose you are able to reproduce the original message m, then you can easily use this to determine also η. This channel is sending for free after the decoding n Bernoulli random bits that should require $nh(p)$ uses of the channel. What is remaining to use is $1 - h(p)$ and that's what we're going to use to convey X.



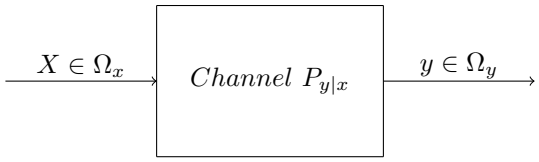
If you try to draw the capacity of the channel $1 - h(p)$ and $h(p)$. When $p = 0$ you get capacity 1 which means that for every use of the channel you can send 1 bit through which makes sense. When $p = \frac{1}{2}$ you get no capacity which also makes sense when you send 0 or 1 the receiver receives unbiased bit so there is no correlation between sending and receiving.

3 General Channel

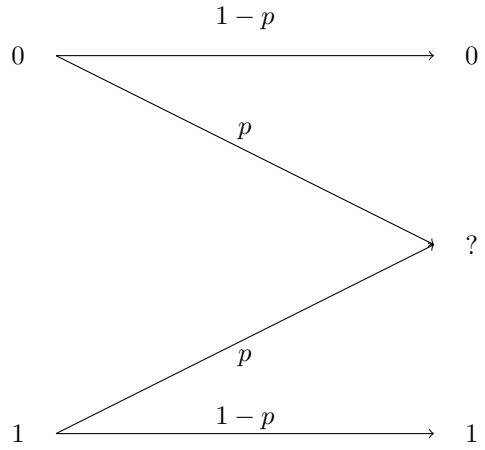
It turns out when you are talking about general channel, you get a even nicer connection. We would like to find the rate and capacity for the general channel. First we would like to specify the encoder for general channel. For today general channel means memoryless and here we're just taking some arbitrary channel which acts independently on every single bit being transmitted.

3.1 Introduction

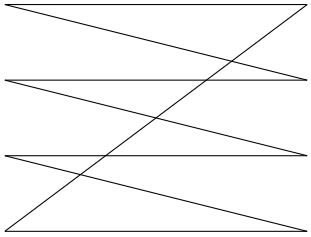
The input X is some element of some universe Ω_x. The output some element of some universe Ω_y. The universe means not to be related. We want think about stochastic channels which make a lot sense that channel are given by a bunch of conditional distributions.



Example 3. *One very simple example of this may be called erasure channel. The following figure shows a Binary Erasure Channel(BEC). It produces a output of* 0 *or* 1 *or* ?.



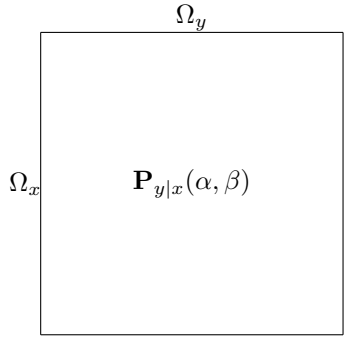
Example 4. *People in information theory may also think about something called noisy typewriter. The channel input is either unchanged with probability* $\frac{1}{2}$ *or transformed into the next letter with probability* $\frac{1}{2}$ *in the output.*



Exercise 5. *Take a binary erasure channel with parameter* p *and a binary erasure channel with parameter* q *and try to find an reasonable relationship between* p *and* q.

Remark The matrix $\mathbf{P}_{y|x}(\alpha, \beta)$ specifies the channel.

$$\mathbf{P}_{y|x}(p|\alpha) = \Pr[Channel\ outputs\ \beta \in \Omega_y, information\ input\ \alpha \in \Omega_x]$$



For the binary symmetry channel case, we have

$1-p$	p
p	$1-p$

Exercise 6. *What's the rate of this binary symmetry channel?*

3.2 Coding Theorem

An encoding function would be a map from $\{0,1\}^n$ to Ω_x^n, the decoding function would be a map from Ω_y^n to $\{0,1\}^k$. And rate is still defined as $Rate = \frac{k}{n}$. And we still have the probability of error and we still can talk about capacity of channel. For any general channel, the capacity is

Definition 7.

$$Capacity(Channel\ P_{y|x}) = \sup_R\{\lim_{\varepsilon\to 0}\lim_{n\to\infty}\{Rate\ R\ of\ E_n, D_n\}\}$$

Here we will get the remarkable theorem once again. It turns out the capacity of the channel are given by this joint distribution.

Theorem 8.

$$Capacity(P_{y|x}) = \sup_{P_x} I(x;y)$$

Remark Once X distribution is specified, I get a joint distribution on (X,Y). The information that Y conditioned on X is the capacity of this channel. This completely characterize every memoryless channel of communication. Given Y, figuring out X. Then mutual information is the right characterization.

Let's prove half of the statement first, half of that statement next, and then other half. Here is what we're going to do with the encode, pick n large enough, let $k \geq (1-h(p)-\varepsilon)n$, let $E_n : \{0,1\}^k \to \{0,1\}^n$ be completely random. The decoding function is, not changed, given some sequence Y, we're going to look at the m which maximizes the probability y^n conditioned on x^n, that is,

$$D_n(\beta^n) = \operatorname*{argmax}_m\{P_{y^n|x^n}(\beta^n, E(m)\}$$

where $\beta_n \in \Omega_y^n$. It could be used to talk any channel, for instance, the Markov channels. The following theorem is nor hard to prove. We have

Theorem 9.

$$\Pr_{E_n,m,y|E_n(m)}[D_n(y) \neq m] \leq \varepsilon$$

$$\Rightarrow \exists E_n \Pr_{m,y|E_n(m)}[D_n(y) \neq m] \leq \varepsilon$$

Remark

$$\Pr_{E_n,m,BCS_p}[D_n(BSC_p(E_n(m))) \neq m] \leq \varepsilon$$

$$\Rightarrow \exists E_n \Pr_{m,BSC_p}[D(BSC_p(E(m))) \neq m] \leq \varepsilon$$

Proof: There are two types of errors

- Error of type 1(E1): Too many error happen. X^n and Y^n differ in more than $(P+\varepsilon)n$ coordinates. It's straightforward to show that $\Pr[E_1] \leq exp(-n)$. (If you don not see Chernoff bounds like this, do send me a note and I will put some further readings.)

- Error of type 2(E2): Fix message m you just transmitted, fix the encoding of this message $E_n(m)$, that is an random variable but it's fixed and fix $y = BSC_p(E_n|m)$, let $E_n^1 : \{0,1\}^k - \{m\} \to \{0,1\}^n$ is not fixed and be random. E_2 is the event that there exists m' such that encoding of $E_n(m')$ and y disagree in less than or equal to $(p+\varepsilon)n$ coordinates.

  We want neither of these events happen then the most likely message is the one we transmitted.

$$\Pr[D_n(BSC(E_n(m))) \neq m] \leq \Pr[E_1 \cup E_2] \leq \Pr[E_1] + \Pr[E_2]$$

  $\Pr[E_1]$ is exponentially small thus we are left with probability E_2, the bound relys on the probability of E_2 which is simple to calculate.

  Lemma 10.

$$\forall m', \Pr[m' \in Ball\ of\ radius\ (P+\varepsilon)n\ around\ y] = \frac{Volume(Ball)}{2^n} \approx \frac{2^{(h(p)+\varepsilon)n}}{2^n}$$

  Remark

$$Ball_r(y) = \{Z \in \{0,1\}^n | Z\ and\ y\ differ\ in \leq r\ coordinates\}$$

  The volume of this ball or the size of this set is

$$|Ball_r(y)| = \sum_{i=0}^{r} \binom{n}{i} \approx 2^{h(\frac{r}{n})n}$$

  That's just one single message, if you want any valid message

$$Pr[\exists m', st.\ E2] \leq \frac{2^k 2^{(h(p)+\varepsilon)n}}{2^n}$$

  This is where we see the quantity that we want, $\frac{k}{n} + h(p) + \varepsilon < 1$ and why need one minus entropy.

Exercise 11. *Show* $\Pr[E_1] \leq exp(-n)$

That proved part of what we want to prove today. It says about the capacity of the binary symmetric channel is

$$Capacity\ of\ BSC \geq \lim_{\varepsilon \to 0}\{1 - h(p) - \varepsilon\}$$

Now let's see if we can show the capacity of general channel is at least $\lim_{\varepsilon\to 0} \sup_{P_x}\{I(X;Y)\} - \varepsilon$.

We are going to pick n to be large enough and k large enough. We're going to fix some distribution P_x. Now here we are going to choose the encoding of $E_n(m)_i \sim P_x$, iid over all (m,i), $m \in \{0,1\}^k$ and $i \in [n]$. The decoding function is still the same maximum likelihood. Now there is a question what is the error of type 1 and type 2 look like? There's no notion like error anymore.

So What we're going to do instead is to start talking about typical sequences. Let's recall asymptotic equipartition principle (AEP):

Lemma 12. *If* $Z_1, ..., Z_n$, $Z_i \sim P_z$ *iid, then*

$$\exists S \subseteq \Omega_z^n, \frac{1}{|S|^{1+\varepsilon}} \leq \Pr[Z_1...Z_n = r_1...r_n] \leq \frac{1}{|S|^{1-\varepsilon}}$$

$$\Pr[(Z_1...Z_n) \notin S] \leq \varepsilon$$

Here S="typical set for P_z*"*

We are going to do the usual three step analysis. The first step is to pick the encoding of the message $E(m)$, the second step is to pick what you received conditioned on what you send $y|E(m)$ and the third step is to look at $E(m')$ where $m' \neq m$. There are three kinds of typicality errors. If $x = E(m)$ is not be typical for P_x, we say we have error of type 1. y may not be typical for P_y. The most important thing is we also want (X, Y) to be typical. But (X, Y) may not be typical for P_{xy}. Probability of any of these errors happens is negligible by AEP.

But we have to get the mutual information out of these. Let's change the decoding algorithm.

Remark The decoding is some function of β and if β is not typical for P_y, then declare an error. If there exists a unique m such that $E(m)$ is typical for P_x and $(E(m), y)$ are typical for P_{xy}, then output m.

We are going to look at $E(m')$ and want to analyze the probability that $E(m')$ is typical and $E(m', y)$ is also typical. Let's fix the m' and ask the question. The fact $E(m')$ is typical is going to happen in a very high probability. We are more interested in the probability that $(E(m'), y)$ is jointly typical. This is the crucial question. This is actually distributed according to $P_x \times P_y$.

Here is the lemma that I will state which immediately imply.

Lemma 13. *Let* $Z_n \sim P^n$,

$$\Pr[Z^n \textit{ is typical for some distribution } Q] \leq 2^{-D(Q||P)n}$$

This is another fundamental reason to understand the divergence between these two distributions. You can apply to very simple things. In the current case we are looking into $D(P_{xy}||P_x \times P_y) = I(X, Y)$. When you combine these two facts, apply here, it turns out that any particular message is going to be jointly typical, the probability is approximately two to the minus mutual information. It tells the rate is at least the mutual information for this distribution. Next lecture we will try to prove the upper bounds.