6.440: Essential Coding Theory Spring 2013

Lecture 2 — February 11, 2013

Prof. Madhu Sudan *Scribe: Mina Karzand*

1 Advertisement

The course 6.441 ”Information Theory” is offered this spring by Prof. Yury Polyanskiy.

2 Plan

Today we will talk about Shannon’s 1948 paper ”A mathematical theory of communications” and

- Coding theorems (source coding and channel coding)
- Converse theorems

3 Shannon’s paper

Shannon in his 1948 paper had a new perspective on errors that might happen in communications. He also tried to come up with an idea of how *reliable communications* should look like and it should be mathematically modeled. To do so, he studied two extreme cases of communication channels:

- *Noiseless Channel*: In which the goal of the communication is to transmit the message with the smallest use of resources (bits) as possible. The basic idea behind the source coding is *compression.*
- *Noisy Channel*: In which the during the transmission of the message, the output of the channel is a noisy version of the input. The goal of Channel coding is to introduce some redundant representation of the message such that reliable communication is possible even in the presence of noise.

He introduced two intermediate objects: *the encoder* at the transmission side of the communications and *the decoder* at the receiver side of the communications help achieve the above goals.

At the time that Shannon wrote his paper, the computational power to perform the encoding and specially decoding in a tractable way did not exist. Instead, he claimed that these are mathematically well-defined objects.

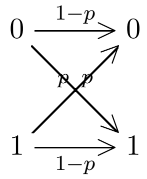
4 Noisy Channel Theorem

$$\boxed{\text{Source}} \xrightarrow{m} \boxed{\text{Encoder}} \xrightarrow{X=E(m)} \boxed{\text{Noisy Channel}} \xrightarrow{y} \boxed{\text{Decoder}} \xrightarrow{\hat{m}=D(y)} \boxed{\text{Receiver}}$$

The channel is modeled as an object whose output is a noisy version of the input. The probabilistic nature of the relationship between the output of the channel, y and input of the channel, x is modeled as $P(y|X)$.

Discrete memoryless channels are defined to be the channels this relationship between the input and output for each letter is independent of the time and the past history of the communication.

Binary Symmetric Channels, BSC(p) are the discrete memoryless channels with binary input and output alphabet in which the input is flipped with probability p at the output.



Shannon claimed that for any discrete memoryless channel (or more generally, ergodic channels), there exist limiting rat of communication, *capacity*, which he found exactly.

Capacity for BSC(p) is $C_{\text{BSC}(p)} = 1 - H(p)$ where

$$H(p) \triangleq -p \log p - (1-p) \log (1-p)$$

5 Informal Statement of the noisy channel coding theorem:

For any communication channel, there is a sharp threshold called capacity C.

If you transmit information at a rate $R < C$, then $\Pr\{\hat{m} \neq m\} \leq \exp(-n)$.

If you transmit information at a rate $R > C$, then $\Pr\{\hat{m} = m\} \leq \exp(-n)$.

As described, if the rate of the communications is lower than the capacity, the larger the block length, the probability of error at the receiver is smaller. The cost of achieving this smaller probability of error with longer blocks is that the encoding and especially the decoding would be heavier computationally and the delay to extract the message in each block is also longer.

If the rate of the communications is above the capacity, there is nothing that can be done at the encoder and the decoder to have reliable communications and small probability of error.

6 Informal Statement of Noiseless Channel Coding Theorem:

Having a source of information which generates a sequence of bits as its output and each bit is 0 with probability $1-p$ and 1 with probability p, we want to represent this sequence with as few bits as possible in a noiseless channel.

Suppose that we run this source n times. We expect to see about pn 1's in the generated sequence and $(1-p)n$ 0's. One way to represent the message (the generated sequence) is to have a two phase communication. In the first phase the number of 1's in the sequence is transmitted. In the second phase of the communications, the string with the given number o 1's is identified.

1. Send $k =$ number of 1's in the msg.
2. Use $\lceil \log_2 \binom{n}{k} \rceil$ more bits to identify which of the strings of length n with k 1's is the message.

Theorem 1. *(Chernoff Bound) In a sequence of length n, where each element is* 0 *with probability* $1-p$ *and* 1 *with probability p, if k is the number of* 1*'s in the sequence, thus*

$$Pr\{k \notin [(p-\epsilon)n, (p+\epsilon)n]\} \leq \exp(-n)$$

The generated sequence from the specified source with k 1's in it, could be in one of these two cases:

- $k \in [(p-\epsilon)n, (p+\epsilon)n]$: According to the Chernoff bound, this happens with probability greater than $1 - \exp(-n)$. We want to know how $\binom{n}{k}$ looks like in this case. For simplicity, let's assume $k = pn$. Thus,

$$\binom{n}{pn} \approx \frac{n^n}{(pn)^{pn}[(1-p)n]^{(1-p)n}} = \left(\frac{1}{p^p(1-p)^{1-p}}\right)^n$$
$$\log \binom{n}{pn} = n\left[-p\log p - (1-p)\log(1-p)\right] = nH(p)$$

  Thus if the message falls into this category (which happens with probabilty $> 1 - \exp(-n)$), the number of bits required to transmit the message in noiseless channel is $\approx nH(p)(1 \pm \epsilon')$ where $\epsilon' \to 0$ as $\epsilon \to 0$.

- $k \notin [(p-\epsilon)n, (p+\epsilon)n]$: This event happens with probability $< \exp(-n)$ and in this case, the number of bits required to transmit the message $\approx \log \binom{n}{k} \leq n$

Thus, the expected length of the block of bits required to transmit the message generated from this source is $\leq nH(p)(1 \pm \epsilon') + n\exp(-n) \leq nH(p)(1 \pm \epsilon'')$.

The converse part of the coding theorem for noiseless channel says that you can not do better than this, meaning that you can not transmit the message reliably with high probability with smaller number of bits.

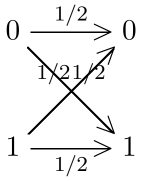
The basic idea for proving the converse is that for each $k \in [(p-\epsilon)n, (p+\epsilon)n]$, there are $\binom{n}{k}$ equiprobable strings wuth the same number of 1's. Using some pigeon-hole theorem argument, it is claimed that at least $\log \binom{n}{k} \approx nH(p)$ bits are required to distinguish these strings.

The interesting thing in this process is that in the coding procedure, it is not necessary to know the actual value of p, even though We used this value in the analysis of the optimal rate and expected block length. This implies that coding (compression) algorithm can be built oblivious to the parameters of the source or the distribution of the messages it generates. This is gonna lead to the idea of universal source coding such as Limpel-Ziv-Welch coding technique.

7 Formal Statement of Noisy channel Coding theorem

The capacity of binary symmetric channel with probability of error of p is $C_{BSC(p)} = 1 - H(p)$.

In the case that $p = 1/2$, it is intuitive that $C = 0$, meaning that no information can be transmitted reliably. Each input symbol is flipped or not with equal probability. This means that the output is independent of the input.



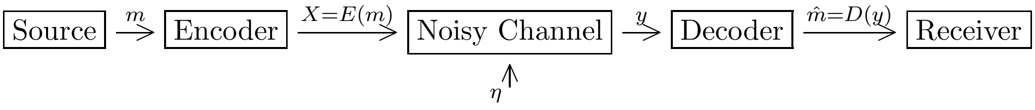
In binary symmetric channels with probability of error $p = 1/2 - \epsilon$ for very small ϵ, we can approximate $C = \Theta(\epsilon^2)$. Thus theory says that even though each input bit is flipped with a probability close to 1/2, still the capacity is nonzero and reliable communication is possible at a nonzero rate.

Shannon suggested that the coding scheme that could be used to achieve this rate is random coding. The codewords for each message should be chosen completely at random. He used probabilistic method to analyze the prove the achievability results for this coding.

7.1 Encoding of BSC(p)

Assume for any $\epsilon > 0$ and $R = 1 - H(p) + \epsilon$ and $k = Rn$. The encoder performs the function $E : \{0,1\}^k \to \{0,1\}^n$ at random. The optimal decoder performs maximum *liklihood decoding.* Given a received string ($y \in \{0,1\}^n$), the ML decoder sees how far this string is far all possible transmitted signals and chooses the closes one as the estimate of the transmitted message.

For our proof, a simpler decoder would suffice. Given the received string y, our proposed decoder generates the set $S_y = \{\tilde{m} | \delta(E(\tilde{m}), y) < (p + \epsilon/2)n\}$. If $|S_y| = 1$, then the estimated message is $\hat{m} \in S_y$. Unless it declares an error in the transmission. We can prove that the probability of error for a source with equiprobable message with completely random encoder and the specified decoder can be made arbitrarily small. It is also assumed that the decoder has complete knowledge of how the encoder function.



The proof of the theorem uses the following lemma.

Lemma 2. $\forall \epsilon > 0$, *assuming uniform distribution over the messages, completely random encoder and η being the error sequence in the BSC(p):*

$$\Pr_{E,m,\eta}\{D(E(m) + \eta) \neq \hat{m}\} \leq \exp(-n)$$

In the decoding process, there are three types of events that could result in an error in the decoding process.

1. $|S_y| = 0$
2. $|S_y| \geq 2$
3. $m \notin S_y$

The proof of the theorem consists of three parts:

(L1) We use Chernoff bound and the fact that the expected number of 1's in the error sequence (η) is pn to claim

$$Pr\{m \notin S_y\} \leq \exp(-n)$$

(L2a) For given received sequence y and fixed $\hat{m} \neq m$,

$$\begin{aligned}\Pr_{E(\hat{(}m))} \{\hat{(}m) \in S_y\} &\leq \frac{\sum_{i=0}^{(p+\epsilon/2)n} \binom{n}{i}}{2^n} \\ &\approx \frac{2^{nH(p)(1+\epsilon')}}{2^n} \\ &\approx 2^{-n[1-H(p)+\epsilon'']}\end{aligned}$$

(L2) Union bound is used to prove

$$\begin{aligned}Pr\{\exists \hat{m} \neq m, \hat{m} \in S_y\} &\leq 2^k 2^{-n[1-H(p)+\epsilon']} \\ &= 2^{n[1-H(p)-\epsilon]} 2^{-n[1-H(p)+\epsilon'']} \\ &\rightarrow 2^{-\epsilon''' n}\end{aligned}$$

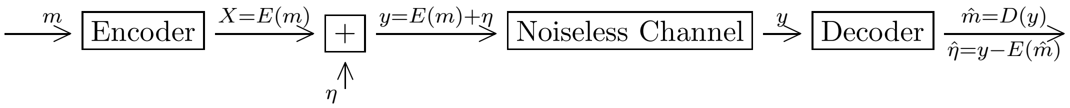
Theorem 3. *$\forall p \in (0, 1/2), \forall \epsilon > 0, \exists \delta > 0, n_0 < \infty$ s.t. $\forall n > n_0$ There exist $E : \{0,1\}^k \rightarrow \{0,1\}^n$ where $k \geq n(1 - H(p) - \epsilon)$ and $D : \{0,1\}^n \rightarrow \{0,1\}^k$ such that*

$$\Pr_{m,\eta \in BSC(p)} \{D(E(m) + \eta) \neq m\} \leq 2^{-\delta n}$$

Note that the ball of radius pn around any point has size $\approx 2^{nH(p)}$.

Theorem 4 (Converse)**.** *$\forall p, \forall \epsilon > 0, \exists \delta > 0, n_0 < \infty$ s.t. $\forall n \geq n_0$ and $\forall E : \{0,1\}^k \rightarrow \{0,1\}^n, D : \{0,1\}^n \rightarrow \{0,1\}^k$ where $k = \lfloor (1 - H(p) + \epsilon)n \rfloor$*

$$\Pr_{m,\eta} \{D(E(m) + \eta) = m\} \leq 2^{-\delta n}$$



If we can recover $\hat{m} = m$, then $\hat{\eta} = \eta$ too. Using the source coding theorem, we know that to transmit η in the noiseless channel we need $H(p)$ bits/symbol. Thus, if we transmit information at a rate $R > 1 - H(p)$, then we are expecting the noiseless channel to transmit information at rate above 1 bit/symbol which is impossible.

The combinatorics way of looking at the converse theorem: We model the input-output relationship by using a bipartite graph. There are 2^k messages which is the number of nodes at the right side of the graph. The nodes at the left side of the graph correspond to the received sequences. Having transmitted a message if a typical error happens, a subset of the output sequences could be received. A typical error has roughly pn number of 1's in it. Thus, assuming only thinking about the typical errors, the degree of the nodes on the left side of the graph is $D = 2^{nH(p)}$, because there are roughly $\binom{n}{np} \approx 2^{nH(p)}$ sequences which correspond to the typical errors. The decoding procedure hopes to be able to recover the message using the noisy received signal. Thus, if $D2^k \gg 2^n$, the decoding procedure fails. Decoding correctly is possible if the degree of the nodes at the right side of the graph is usually 1. Average degree of the nodes at the right side of the graph is $D2^{k-n}$.

$$\Pr\{\text{a random edge lands on vertex of degree} < 2^{k-n}\} < \tau$$