6.440 Essential Coding Theory Feb 11, 2008

Lecture 2

Lecturer: Madhu Sudan *Scribe: Daniel E. Lucani*

1 Reminder

1. Problem set 1 is due Wednesday at 11pm.
2. Swastik (TA) will hold office hours on Wednesday 5-7pm.
3. Sign up early for scribe.

2 Lecture overview

This lecture will discuss some of the important points of Shannon's 1948 paper ("A Mathematical Theory of Communication"). The lecture covers:

1. An overview of Shannon's problem (model considerations and contributions).
2. The Noisy Channel Coding Theorem for the case of the Binary Symmetric Channel (BSC).
3. An outline of the converse of the Coding Theorem.

3 Overview of Shannon's Problem

Let us first look at some of entities in Shannon's model:

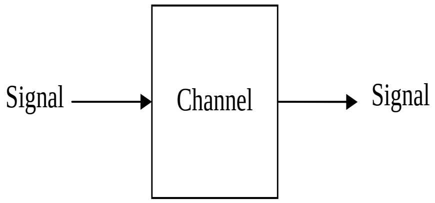
1. Source of Information: constitutes the entity that generates the information to be stored or transmitted, e.g. a camera, a satellite.
2. Channel of communication: for the example of satellite communication, the channel is the space.
3. Receiver: is the entity that should reconstruct the information generated by the source.

3.1 Contributions: Mathematical Model

Shannon gave us a mathematical model of both the source and the channel. Let us look at the details of the model.

1. Source: It is modelled as a stochastic process with a probabilistic distribution. Shannon associated a parameter called Entropy to this distribution. The entropy measures the uncertainty of a random variable produced by the distribution and is associated also to a Rate at which this uncertainty is generated.

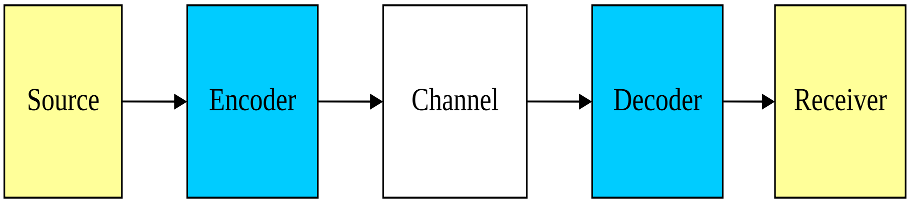
2. Channel: It is modelled as an Input - Output process through a Marginal Distribution of the output given the input of the channel. This gives a way to study a channel of communication. Also, there is an associated Capacity with every channel.



3.2 Contributions: Architecture

Shannon provided an architecture of an information transmission system, introducing the concepts of encoder and decoder which he showed to give more reliable communications.

The implementation of an encoder and decoder has been a rich source of algorithmic problems to computer science till today.



4 Meta-theorem

$\forall$ channels in "this class" $\exists C$ (Capacity), $\forall$ Source in "this class" its Rate (R) such that we can transmit information iff $R < C$.

Note:a more detailed version will be provided after some examples.

4.1 Case 1: Noiseless Channel

In this case, the channel is represented by an identity function, i.e. the output is the same as the input. So there is not much interest in studying the channel.

However, we can consider compressing the information generated by the source, and it can be done.

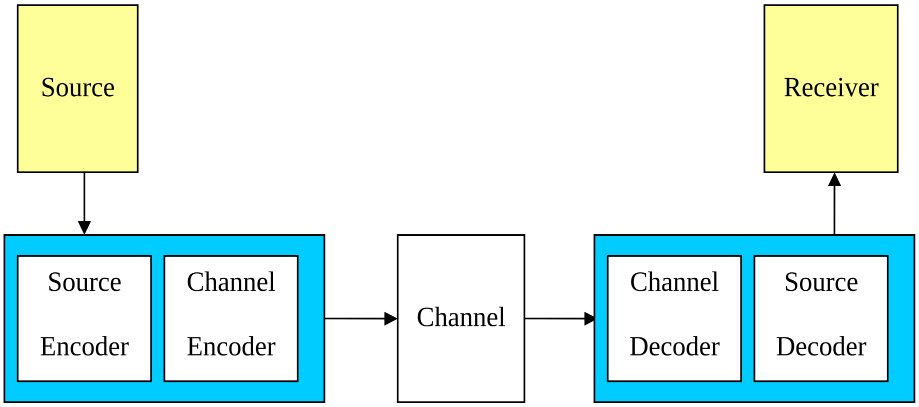
Shannon: Can compress source to its "Rate" but no better, and this compression depends only on the source.

4.2 Case 2: Source generated k-bit strings with bits uniformly distributed

With uniformly distributed bits there is not much compression to be done to the information from the source.

Shannon: Can encode (or add redundancy) to the message so as to recover the message reliably at the receiver, provided the channel capacity is large enough.

4.3 Combining case 1 and 2

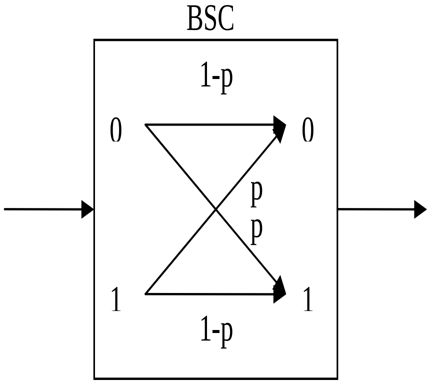


The first case corresponds to the problem of compress the information from the source to convert it into a uniformly distributed sequence, while case 2 refers to adding redundancy to a uniformly distributed sequence.

Combining both cases gives the coding theorem (Our meta-theorem).

Shannon: Proved that breaking the problems of source and channel encoding/decoding is optimal.

5 Binary Symmetric Channel (BSC)

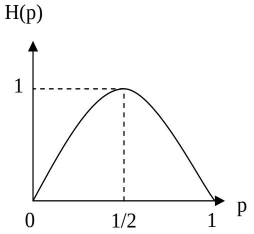


The binary symmetric channel with flipping probability p, BSC(p), is a binary channel in which the input has equal probability p to be flipped from 1 to 0, and from 0 to 1. This means that an incoming bit has probability $1 - p$ to go through the channel unchanged and probability p to flip its value.

Shannon's Theorem: The capacity of BSC(p) is $1 - H(p)$.

where $H(p) = -p \log_2 p - (1 - p) \log_2(1 - p)$.

Note that for $p = 0$, $H(0) = 0$, which means that for the case of no randomness $C = 1$, i.e. you can get 1 bit through per time unit. Note also that$H(1/2) = 1$ assures that there is no correlation between the input and the output, and it characterizes a channel with capacity $C = 0$, i.e. you cannot get information through that channel.



5.1 Noisy Coding Theorem for BSC

Motivation If you inject a sequence $X = (x_1, ..., x_n)$ through a channel we receive $X + Y$, where $Y = (y_1, ..., y_n)$.

Now, what is the distribution of Y?

Informal Intuition: a typical Y is uniformly distributed over a set of size $\begin{pmatrix} n \\ pn \end{pmatrix} \approx 2^{H(p)n}$.

Theorem $\forall p \in (0, 1/2)$ without loss of generality, $\epsilon > 0$, $\exists \delta > 0, n_0$. $\forall n \geq n_0$ $\exists k$ and functions:

$E : \{0,1\}^k \longrightarrow \{0,1\}^n$ $D : \{0,1\}^n \longrightarrow \{0,1\}^k$

such that

1.
$$P_r\left[D(E(m) + Y) \neq m\right] \leq exp(-\delta n) \tag{1}$$
for message $m \in \{0,1\}^k$ and $Y \leftarrow BSC^n(p)$.

2.
$$k \geq (1 - H(p + \epsilon))n \tag{2}$$

Note: taking $p \in (0, 1/2)$ is sufficient. For $p = 0$ it is the noiseless channel. For $p = 1/2$ there is no correlation between input and output of the channel. For the range of values of $p \in (1/2, 1)$, consider that it is sufficient to change our decision criteria at the receiver to be in $p \in (0, 1/2)$, i.e. when a 1 outputs the channel the receiver will interpret it as a 0, and vice versa.

Another question that arises is: where do the encoding (E) and decoding (D) functions come from?

Informal Interpretation: There is an encoder and decoder such that the probability of having an error when decoding is exponentially small.

5.2 Formal Statements

We first recall the Chernoff Bound.

Chernoff Bound If $y_1, ..., y_n \in [0, 1]$ are i.i.d (independent and identically distributed) with $Ey_i = p$, then

$$P_r\left[\left|\frac{\sum_{i=1}^n y_i}{n} - p\right| \geq \epsilon\right] \leq e^{-\epsilon^2/2n} \tag{3}$$

Notation: $wt(Y)$ For $Y \in \{0,1\}^n$ with $Y = y_1 ... y_n$, we define its Hamming weight $wt(Y)$ to be $\sum_i y_i$.

Claim If $y \in \{0,1\}^n$ with $wt(y) = i \in \{(p-\epsilon)n, ..., (p+\epsilon)n\}$, then

$$\Pr_Y[Y = y] \leq \frac{1}{\begin{pmatrix} n \\ i \end{pmatrix}} \leq 2^{-H(p-\epsilon)n} \tag{4}$$

with $Y \leftarrow BSC^n(p)$. This follows from symmetry, as all error patterns with i 1s are equally likely. Note that the size of the set $\{y : wt(y) = pn\}$ is $\begin{pmatrix} n \\ pn \end{pmatrix} \approx 2^{H(p)n}$.

5.2.1 Back to the coding theorem: Proof

Encoding function E: Let us pick $E : \{0,1\}^k \to \{0,1\}^n$ at random uniformly from all such functions. Then, $\forall m \neq m'$, $E(m)$ is a random n-bit string, independent of $E(m')$.

Decoding function D: Let us define our decoding function $D(X) =$ if $\exists! m \in \{0,1\}^k$ such that $\Delta(E(m), X) \leq (p+\epsilon/2)n$ produces as output message m. Otherwise, the output is that "Too many errors" occurred.

Note that it is not efficient algorithmically.

Now let us look at two bad events:

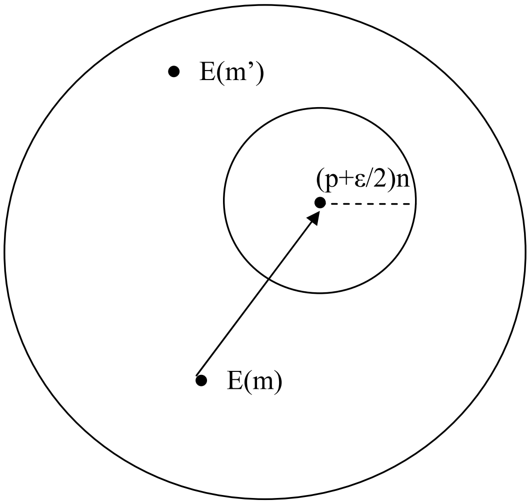
1. E1: "Too many errors" defined as "$wt(Y) \geq (p+\epsilon/2)n$".
2. E2: "Error Pattern/Encoding Bad " defined as "$\exists m' \neq m$ such that $\Delta(E(m)+Y, E(m')) \leq (p+\epsilon/2)n$"

To prove the theorem we need to prove that:
A. If neither E1 nor E2 happen, then decoding is successfully.
B. $P_r[E1]$ is small.
Proof: By Chernoff Bound, $P_r[E1] \leq e^{-\frac{\epsilon^2}{8}n}$
C. $P_r[E2]$ is small.
Proof: Initially fix $E(m), Y, m' \neq m$ and pick $E(m')$ at random. Then,

$$\begin{aligned} P_r\left[\Delta(E(m'), E(m)+Y) \leq (p+\epsilon/2)n\right] &\leq \frac{\sum_{i=0}^{(p+\epsilon/2)n} \begin{pmatrix} n \\ i \end{pmatrix}}{2^n} && (5) \\ &\leq \frac{n 2^{H(p+\epsilon/2)n}}{2^n} && (6) \\ &\doteq 2^{(H(p+\epsilon/2)-1)n} && (7) \end{aligned}$$

Now, we take the union bound over all $m_1 \neq m$.

$$\begin{aligned} P_r[E2] &\leq \sum_{m' \neq m} P_r\left[\Delta(E(m'), E(m)+Y) \leq (p+\epsilon/2)n\right] && (8) \\ &\leq 2^k 2^{(H(p+\epsilon/2)-1)n} && (9) \end{aligned}$$



Note that if $k = (1 - H(p + \epsilon))n$, then

$$P_r[E2] \quad \leq 2^{(H(p+\epsilon/2)-H(p+\epsilon))n} \approx e^{-n} \tag{10}$$

Since both B and C hold for $k = (1-H(p+\epsilon))n$, $P_r[DecodingError] = \tau \leq e^{-n}$ for parameters m,Y,E. Thus, $\exists$ E such that $P_r[DecodingError] \leq \tau \leq e^{-n}$ as function of m,Y.

Note that the statement in B is proven based on properties of Y, while C depends on properties of the encoding function (E) itself.

6 Some Comments on the Converse

The converse of the coding theorem allows us to answer the following questions:

Is the rate $k = (1 - H(p))n$ the best possible? or Can we do better?

Shannon's Answer: NO

In other words, the converse let us state that the rate $k = (1 - H(p))n$ is the best you can achieve.

The converse will be proven in detail in next lecture. The idea behind the proof is that the source will be generating k-bit strings, while the channel introduces errors. The number of typical errors is $2^{H(p)n}$, while the number of possible messages is 2^k. The output of the channel has 2^n possible values. If we want no two messages to be confused with one another, we would want that $2^k 2^{H(p)n} \leq 2^n$, i.e., $k \leq (1 - H(p))n$.