CSE 533: Error-Correcting Codes **(Autumn 2006)**

Lecture 3: Shannon's Theorem

October 9, 2006

Lecturer: Venkatesan Guruswami *Scribe: Widad Machmouchi*

1 Communication Model

The communication model we are using consists of a source that generates digital information. This information is sent to a destination through a channel. The communication can happen in the spatial domain (i.e., we need to send information over a physical distance on a channel) or in the time domain (i.e., we want to retrieve data that we stored at an earlier point of time).

The channel can be associated with noise. So we have two cases :

- **Noiseless case:** The channel in this case transmits symbols without causing any errors. One would need to exploit the redundancy in the source to economize the length of the transmission. This is done through data compression, also called source coding. The information is decompressed at destination.

- **Noisy case:** The channel in this case introduces noise that causes errors in the received symbols at the destination. To reduce the errors incurred due to noise, one should add systematic redundancy to the information to be sent. This is done through channel coding.

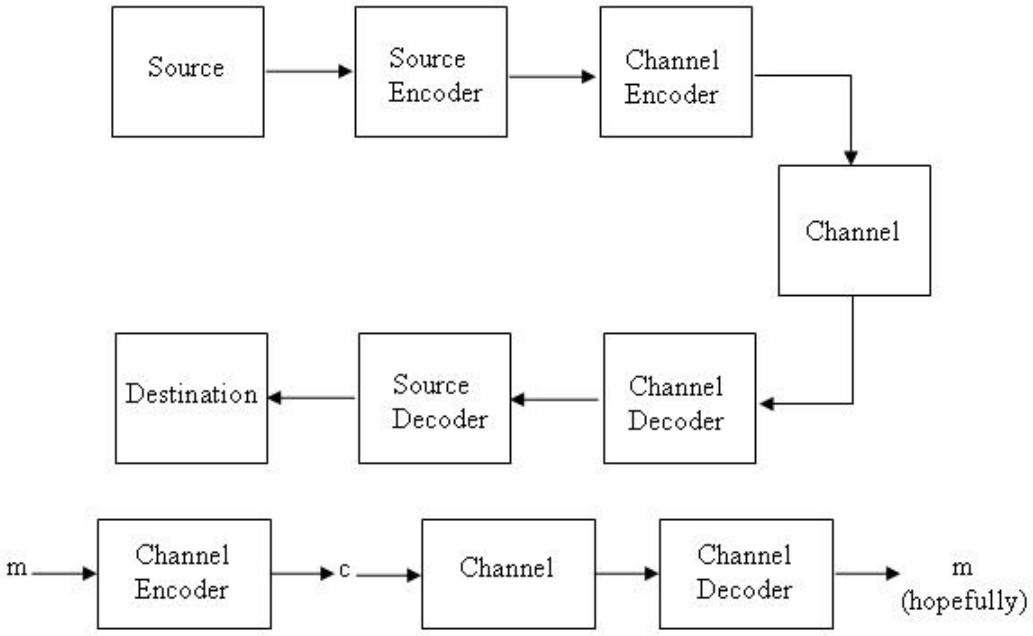
It is known that reliable communication is possible in the above model if the entropy of the source, i.e., the amount of non-redundant information it generates per unit of time, is less than the "capacity" of the channel, i.e., the maximum number of information bits that can be communicated reliable per channel use.

We will focus on the channel coding problem in the presence of noise, assuming that the information is generated at the source is "source-coded" or compressed into a string of non-redundant symbols over a finite alphabet prior to communication on the channel. We can communicate such a two-stage scheme as in the following figure, and treat source and channel coding in isolation, due to what is known as the *Source-Channel Coding* theorem.

The following diagram shows the modules of the communication model:

Source-Channel Coding Theorem: For a source with entropy no greater than the capacity of the channel, dividing the transmission process to source coding followed by channel coding can achieve a probability of error tending to zero for a large block length.

Then the part of the previous scheme we'll be considering is:



Shannon put forth a stochastic model of the channel. For us, it suffices to talk about discrete memoryless channels. Such channels have an input alphabet $\mathcal{X}$, an output alphabet $\mathcal{Y}$ and a probability transition matrix describing the output distribution for every input. Each symbol is sent over the channel independently of the previous symbols sent. Thus the channel is prescribed by a $|\mathcal{X}| \times |\mathcal{Y}|$ stochastic matrix where each row sums to 1.
Probability transition matrix:

$$|\mathcal{X}| \left\{ \overbrace{\begin{pmatrix} \quad p(y/x) \quad \end{pmatrix}}^{|\mathcal{Y}|} \right.$$

2 Examples of Channels

Channels are often described by input-output diagrams.

2.1 Binary Symmetric Channel (BSC)

The BSC takes as input one bit (0/1) and flips it with probability $p, 0 \leq p \leq \frac{1}{2}$. p is called the crossover probability; we write BSC_p to denote the binary symmetric channel with crossover probability p.

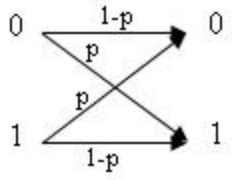


Figure 1: Diagram for BSC_p

2.2 Binary Erasure Channel (BEC)

The BEC takes as input one bit (0/1)and erases it to ? with probability ε , $0 \leq \varepsilon \leq 1$, . ε is called the erasure probability; we write BEC_ε to denote the binary erasure channel with erasure probability ε.

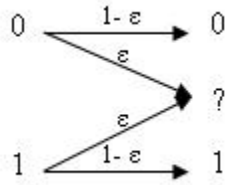


Figure 2: Diagram for BEC_ε

2.3 Continuous Output Channel

The continuous output channel takes as input a symbol from a finite alphabet and maps it, according to a specific noise distribution, to a real number. One example is the Binary Input Additive White Gaussian Noise (BIAWGN) channel, where the noise has a normal distribution and acts additively.

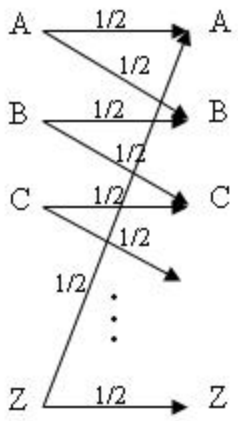
$$\begin{array}{ccccc} \Sigma = \{-1, 1\} & \rightarrow & \boxed{\text{Channel}} & \rightarrow & \mathbb{R} \\ x & & \longrightarrow & & y = x + z;\ z \in N(0, \sigma^2) \end{array}$$

Hence the conditional probability density function of the channel output y on input x is given by :

$$\Pr(y|x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-x)^2}{2\sigma^2}\right)$$

2.4 Noisy Typewriter Channel

The noisy typewriter channel is given by the following diagram:



Zero-error communication is possible if we just send A, C, E,..., Y or B, D, F,..., Z. So we will end up with only 13 possibilities for the sent symbols. Thus the capacity of the channel is at least $\log_2 13$ bits. In fact, one can prove that this rate is the maximum possible and the capacity of the channel is exactly $\log_2 13$ bits.

Zero-error communication at a positive rate is not possible with BSC_p, since for every pair of strings x and y, there is a positive probability that x gets distorted into y. This probability of miscommunication can be reduced arbitrarily by high enough order repetition code.
Consider mapping 0 to m zeroes and 1 to m ones. At the destination, decoding is done by majority, i.e. if the number of received 0's is greater than $\frac{m}{2}$, we decide on a 0, else we decide on a 1. Hence, the probability of error is given by: $\Pr(error) = \sum_{i=\frac{m}{2}}^{m} \binom{m}{i} p^i (1-p)^{m-i}$. This probability tends to 0 as m tends to ∞ but this causes the rate to tend to zero!!
Can we achieve any desired probability of error while maintaining a positive rate ? The answer is yes. In fact, the largest possible rate was precisely characterized and described in Shannon's work.

3 Shannon Capacity Theorem

We will start by defining the binary entropy function.

Definition 3.1. *For $0 \leq x \leq 1$, the entropy binary function, denoted $H(x)$ is given by*

$$H(x) = x \log_2 \frac{1}{x} + (1-x) \log_2 \frac{1}{1-x}.$$

Note that we have $2^{-H(x)n} = x^{xn}(1-x)^{(1-x)n}$.

Theorem 3.2 (Shannon Capacity Theorem for the BSC). *For every $p, 0 \leq p < \frac{1}{2}$, and $0 < \varepsilon < 1/2 - p$, there exists $\delta > 0$ such that for all large n, there exist an encoding function $E : \{0,1\}^k \to \{0,1\}^n$ and a decoding function $D : \{0,1\}^n \to \{0,1\}^k$ for $k = (1 - H(p+\varepsilon))n$ such that $\forall m \in \{0,1\}^k$,*

$$\Pr_{\text{noise of the } BSC_p} D(E(m) + noise \neq m) \leq 2^{-\delta n}.$$

The occurence of the entropy function $H(p)$ in the statement of the capacity theorem arises since we will see that $2^{H(p)n}$ is an accurate asymptotic estimate of the volume of a Hamming ball of radius pn.

Lemma 3.3. *For* $0 \leq p \leq \frac{1}{2}$, $\mathrm{Vol}_2(B(0,pn)) = \sum_{i=0}^{pn} \binom{n}{i} \leq 2^{H(p)n}$.

Proof.

$$\begin{aligned}
1 &= (p + (1-p))^n \\
&\geq \sum_{i=0}^{pn} \binom{n}{i} p^i (1-p)^{n-i} \\
&= \sum_{i=0}^{pn} \binom{n}{i} (1-p)^n \left(\frac{p}{1-p}\right)^i \\
&\geq \sum_{i=0}^{pn} \binom{n}{i} (1-p)^n \left(\frac{p}{1-p}\right)^{pn} \\
&= \sum_{i=0}^{pn} \binom{n}{i} p^{pn} (1-p)^{(1-p)n} \\
&= \sum_{i=0}^{pn} \binom{n}{i} 2^{-H(p)n} .
\end{aligned}$$

□

We will first prove the converse of Shannon theorem to give an intuition why $1 - H(p)$ is the best rate one can hope for. For this purpose we will use the lower bound $\mathrm{Vol}_2(B(0,pn)) \geq \binom{n}{pn} \geq 2^{H(p)n - o(n)}$. This fact follows from applying Stirling approximation of $n!$ given by:

$$n! = \sqrt{2\pi n} \left(\frac{n}{e}\right)^n \left(1 + \Theta\left(\frac{1}{n}\right)\right) .$$

We will also need the Chernoff bound.

Chernoff bound If $X_1, X_2, \ldots X_n$ are i.i.d 0/1 random variables with $\Pr[X_i = 1] = p$, then $\forall\ 0 < \varepsilon < 1$

$$\Pr[\sum_{i=1}^{n} X_i \geq (p+\varepsilon)n] \leq 2^{\frac{-\varepsilon^2 n}{3}}$$

$$\Pr[\sum_{i=1}^{n} X_i \leq (p-\varepsilon)n] \leq 2^{\frac{-\varepsilon^2 n}{3}}$$

Proof sketch of the converse of Shannon theorem
Let $E : \{0,1\}^k \to \{0,1\}^n$ be an encoding function, and let D be the decoding function D :

$\{0,1\}^n \to \{0,1\}^k$. For a message $m \in \{0,1\}^k$, let S_m be the inverse image of m under D, i.e., $S_m = \{y \mid D(y) = m\}$. When m is transmitted, by the Chernoff bound, with high probability $y = E(m) +$ noise lies in a shell $\mathcal{S}$ of radii $((p-\varepsilon)n, (p+\varepsilon)n)$ around $E(m)$. To achieve small decoding error probability, most of the strings in $\mathcal{S}$ must be decoded to m, i.e., belong to S_m. We know that $|\mathcal{S}| \geq \binom{n}{pn} \geq 2^{H(p)n-o(n)}$. It follows that $|S_m| \geq 2^{H(p)n-o(n)}$ for every m. This implies that $2^n \geq \sum_{m\in\{0,1\}^k} |S_m| \geq 2^{k+H(p)n-o(n)}$, which yields $k \leq (1-H(p)+o(1)n$.

We will continue the proof of Shannon theorem next lecture.