ELEC 6131 – Error Detecting and Correcting Codes

Appendix A:
An Intuitive View of Information Theory

Appendix A: Information Theory

- **Information Theory :** The science that provides us with bounds on the performance of transmission strategies.
- This field was established in 1948 with the ground breaking paper **A Mathematical Theory of Communication** by **Claude E. Shannon.**
- Other works by Shannon such as **Communication Theory of Secrecy Systems is a paper published** in **1949** and **Coding theorems for a discrete source with a fidelity criterion in 1959** paved the road to the modern day digital information age.
- **The role of Shannon in Modern Day Digital World.**
  - Many experts believe that many of the things that we take for granted such as Smartphone, High-speed Internet and HDTV would not have happened the way they happened and when they happened if it was not for Shannon.
- **Watch the Video about Shannon and his contributions:**
  - **https://www.youtube.com/watch?v=z2Whj_nL-x8**

Appendix A: Information Theory

Results developed by Shannon and later refined and/or generalized by many researchers in the area of information theory provide limits on not only the transmission rates feasible over different physical channels, but also the limit on how much a source such voice, audio, video can be compressed given the level of distortion that one can tolerate. So, in a course like this where we tackle both audio/video compression and data transmission a basic understanding of information theory is a must. Of course, given the amount of things that we have to cover, we cannot spend much on this topic. Therefore, I try first to present an intuitive definition of some basic concepts in information theory such as the **entropy** and **mutual information** and then give some results of information theory useful in this course without proof. I will try to clarify these results with reference to examples related to our course.

Appendix A:

An intuitive approach to Information Theory

In a formal information theory course, one first defines some of the entities used in information theory such as **mutual information** and **entropy** using abstract mathematical formulas involving probabilistic description of the source and channel and then their physical meaning is demonstrated. In this note, I start by stating these entities as symbols and then discuss their intuitive meaning, derive certain relationships between them based on "common sense" and finally express them in terms of probabilistic parameters. The latter being mainly necessary for computational purposes.

In most terms entropy is defined first and then the mutual information. I find mutual information more intuitively explainable, particularly in communications. So, I start with the concept of mutual information. The **mutual informat**ion quantifies the amount of information an event (a process) provides about another event (process). Take a process X with events $x \in X$. For example X can represent the weather in a specific season and $x \in X$ can be "cold", "hot", "very cold", "cool", etc.

Appendix A:
An intuitive approach to Information Theory

Another process Y can model the trend clothing purchase by people. So $y \in Y$ can be “coat”, “jacket”, “pants”, “shorts”, etc. The joint information between an outcome $x \in X$ and $y \in Y$ is denoted by I(x; y) and is the amount of information the knowledge of y gives about x. For example, knowing that the people buy more coats than any other type of apparel points to the possibility that the weather is cold and vice versa, i.e., the weather is going to be cold , the vendors will stock coats instead of other clothing items. The amount of information x gives about y (or y gives about x), i.e., I(x; y) should logically depend on how much y depends on x. The extreme case is when x and y are independent. In such a case it is natural to expect that $I(x;y) = 0$.

Just a brief mention of probability here: when two events are independent their joint probability mass (or density) function can be written as $p(x,y) = p(x)p(y)$. But when there are not independent $p(x,y) = p(x)p(y|x)$ or $p(y)p(x|y)$. So, in a sense mutual information is a quantification of how much p(x, y) is different from p(x)p(y).

Appendix A:
Mutual Information and Entropy

What usually is considered as the mutual information is the average of I(x; y) over all possible values of $x \in X$ and $y \in Y$:

$$I(X;\ Y) = E_{x,y}\{I(x; y)\}$$

And is called the [average] mutual information and is a measure of the average amount of information that observing the process X provides about Y and vice versa.

Now let's consider I(X; X), i.e. the amount of information the observation of X gives about X!?

I(X; X) is all you need to know (or like to know or can know) about X. Having seen X, there is no uncertainty about X. So, I(X; X) is the **uncertainty** that we have about X prior to observing it or the amount of information contained in X. It is give a special symbol H(X) nd is called the entropy of X.

In order to be useful, we expect that the mutual information between two processes be non-negative

$$I(X; Y) \geq 0$$

Appendix A: Mutual Information and Entropy

This is to say that knowing something at worst can be useless.

Conditional Mutual Information: $I(X;Y|Z)$ is the mutual information between X and given Z, i.e., the average amount of information X provides about Y given that we have already observed Z.

Conditional Entropy: $H(X|Y)$ is the uncertainty about X given that we have observed Y.

I(X; Y) is the information Y gives about X. So, it is natural to expect it to be the difference between the uncertainty that we have about X before and after observing Y, i.e.,

$$I(X;Y) = H(X) - H(X|Y)$$

and since $I(X;Y) = I(Y;X)$, we can write

$$I(X;Y) = H(Y) - H(Y|X).$$

Appendix A: Channel Capacity

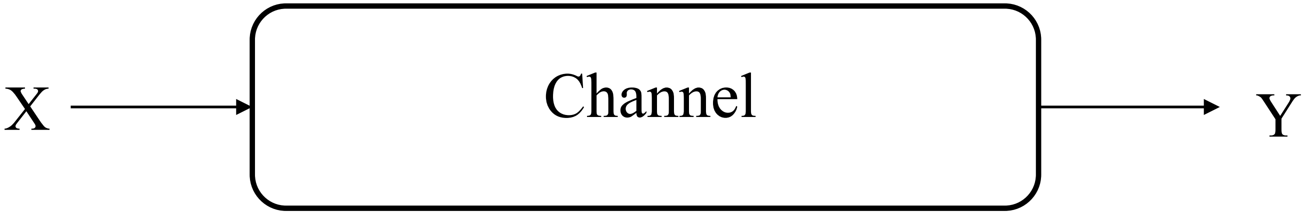
Entropy cannot increase with conditioning.

$$I(X;Y) = H(X) - H(X|Y) \geq 0.$$

So,

$$H(X|Y) \leq H(X).$$

Now, let's consider a channel with input X and output Y,



$H(X)$ is the uncertainty that we have about the input *a priori* and $H(X|Y)$ is the uncertainty about the input after observing the received signal Y. So, the quantity $H(X) - H(X|Y)$ is the amount of information carried through the channel bout the input. That is, $R = H(X) - H(X|Y) = I(X;Y)$ is the rate of information through the channel for a given input X.

Appendix A: Channel Capacity

So, if for a given channel, we find the maximum of R=I(X;Y), we have found the capacity of the channel, i.e., the highest rate at which communications is possible over the channel:

$$C = max_{p(x)}\, I(X;Y).$$

Lecture 6:
Formal Definitions

Department of Electrical & Computer Engineering

Entropy for a discrete source:

Assume that a source takes values $x \in X$ with probabilities $\{p(x)\}$. Then the entropy is defined as

$$H(X) = \sum_x p(x) log\left(\frac{1}{p(x)}\right) = -\sum_x p(x) \log(p(x)).$$

The unit of H(X) depends on the base of the logarithm. For base two logarithm, H(X) is in bits. H(X) is maximized when the probabilities of all events are the same, i.e., H(X)≤ log_2^m where m is the number of outcomes of X.

The **conditional entropy** is,

$$H(X|Y) = \sum_y p(y) H(X|Y = y)$$

$$= -\sum_y p(y) \sum_x p(x|y) log\big(p(x|y)\big)$$

$$= -\sum_x \sum_y p(x, y) log\big(p(x|y)\big)$$

Lecture 6:
Formal Definitions

Similarly,

$$H(Y|X) = -\sum_x \sum_y p(x,y) log\big(p(y|x)\big).$$

Mutual Information is:

$$I(X;Y) = H(X) - H(X|Y)$$

$$= -\sum_{x} p(x) log\big(p(x)\big) + \sum_{x}\sum_{y} p(x,y) \log\big(p(x|y)\big)$$

$$= \sum_{x}\sum_{y} p(x,y) log\left(\frac{p(x|y)}{p(x)}\right)$$

$$= \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}.$$

Lecture 6:
Entropy and Data Compression

The entropy is the least number of bits required to describe the outcome of a process. For a source it means the minimum number of bits required to encode the output of the source.

To see this let's consider a binary source X that generates two outputs, say "zero" and "one". Let the probability of getting a one is p. The probability of having zero is, obviously, 1-p. The entropy of this source is:

$$H(X) = -plog_2(\mathrm{p}) - (1-p)log_2(1-p).$$

Since, the above quantity depends on the probability and not much on the name of the variable, it is usually denoted as H(p).

Now assume that we observe n outputs of the source. If n is large enough, we get a sequence with roughly np ones and $n(1-p)$ zeros. Probability of such a sequence is $p^{np}(1-p)^{n(1-p)}$. Since most of the sequences will be of this type (they are called typical sequences), if we can encode only these sequences, we get a vanishingly small probability of observing a sequence that we cannot encode. Let the number of typical sequences be N_T.

Lecture 6: Entropy and Data Compression

It is clear that,

$$N_T p^{np}(1-p)^{n(1-p)} < 1.$$

or,

$$N_T < p^{-np}(1-p)^{-n(1-p)}.$$

So, we need,

$$k = logN_T < n[-plog(p) - (1-p)\log(1-p)] = -nH(p).$$

So, the compression rate is bounded as,

$$\frac{k}{n} < H(p) .$$

The Huffman coding technique we discussed in previous lectures is a way to attain compression close to entropy.

Exercise 6.1: Consider a source with letters A, B, …, G with probabilities {3/8, 3/16, 3/16, 1/8, 1/16, 1/32, 1/32}. Find the entropy. Compare with the mean length of the Huffman code for this source.

Lecture 6:
Capacity of Gaussian Channel

We discussed only about discrete sources and channels. Here, without going into details, we present some results about one of the important continuous channels.

An Additive White Gaussian Noise Channel (AWGN) is a channel where the input X is corrupted with a noise Z that consists of independent, identically distributed samples of a Gaussian process. AWGN models any noise source that combines the effects of a large number of events. Thermal noise in electronic circuits that is the result of movement of a huge number of particles is an example.

Capacity of an AWGN channel with signal power P and noise power N and bandwidth W is given as,

$$C = Wlog\left(1+\frac{P}{N}\right) \text{ bps.}$$

Most often, the capacity is normalized with frequency and is presented in bits/Hz.

Lecture 6:
Capacity of Gaussian Channel

According to Shannon's channel coding theorem, we can transmit error free as long as our transmission rate is less than C. He showed that theoretically, we can get as close as we wish to this limit. Only recently, practice has proven him right. On the other hand, he proved that we can not exceed this rate and expect low error rate.

Now let's see how we can relate this result to what we achieve using a given transmission strategy (coding and modulation techniques we use)

First note the P is the power, i.e., energy per second. So, if we transmit at the rate R=C, our energy per bit is

$$E_b = \frac{P}{C} = \frac{P}{R}.$$

Also, the noise density will be,

$$N_0 = \frac{N}{W}.$$

So, we can write the capacity as,

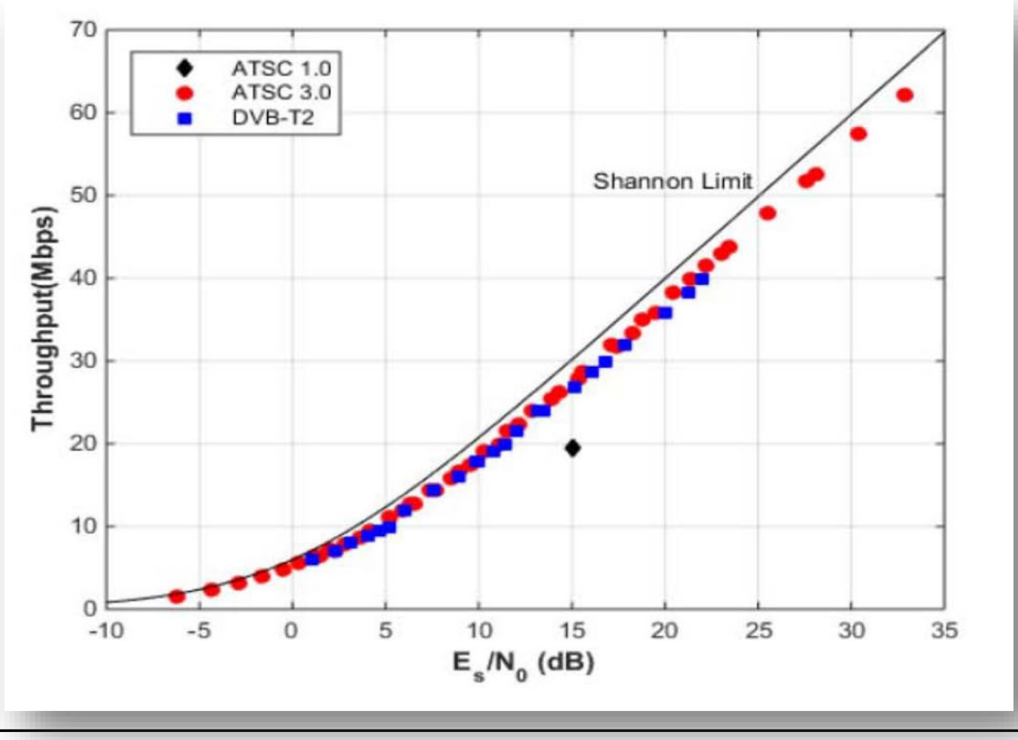
$$R = Wlog\left(1 + \frac{R}{W}\frac{E_b}{N_0}\right).$$

Lecture 6: Capacity of Gaussian Channel

The ratio $\eta = \frac{R}{W}$ is the bandwidth efficiency in bits/second/Hz.

We can write the above equation as $\eta = \log\left(1 + \eta\,\frac{E_b}{N_0}\right)$ or,

$$\frac{E_b}{N_0} = \frac{2^{\eta}-1}{\eta}.$$



Lecture 6: Capacity of Gaussian Channel

Exercise 6.2: Assume that the bandwidth available to you is 2 MHz. For a carrier to noise ratio of $\frac{P}{N} = 15\ dB$ determine the maximum bit rate possible? Compare with what you get with M-PSK modulation with roll-off factor 0.1. Consider transmission with a bit error rate of 10^{-5} as error-free.