ECE 7680
Lecture 1 – Introduction to Information Theory

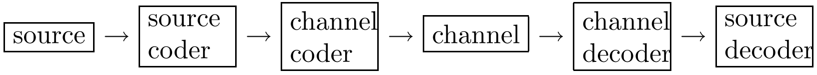
Objective: To learn what information theory is all about

Class stuff

- Syllabus
- Class time
- Course number if you have had this before

The digital communications model

In the transfer of digital information, the following framework is often used:



- The **source** is the source of (digital) data.
- The **source encoder** serves the purpose of removing as much redundancy as possible from the data. This is the **data compression** portion.
- The **channel coder** puts a modest amount of redundancy back in order to do **error detection or correction.**
- The **channel** is what the data passes through, possibly becoming corrupted along the way. There are a variety of channels of interest, including:
  - The magnetic recording channel
  - The telephone channel
  - Other bandlimited channels
  - The multi-user channel
  - Deep-space channels
  - Fading and/or jamming and/or interference channels
  - The genetic representation channel
  - Any place where there is the possibility of corruption in the data
- The **channel decoder** performs error correction or detection
- The **source decoder** undoes what is necessary to get the data back.

There are also other possible blocks that could be inserted into this model:

- A block to enforce channel-contraints. Some channels (e.g., the magnetic recording channel) have constraints on how long a run of zeros or ones can be. The constraints are enforced in what is often known as a *line coder*.
- A block to perform encryption/decryption.
- A block to perform lossy compression.

The first of these areas fall well within the scope of information theory, but unfortunately outside the scope of this class. I hope to get to the last one during the quarter.

In light of the model presented here, several questions arise of engineering interest:

- How can we measure the amount of information?
- How much can we compress?
- How do we compress?
- How do we avoid errors from affecting the performance?
- How fast can we send through the channel?
- What if data rate exceeds the capacity of the channel?

These are largely *theoretical* questions, and the answers are largely theoretical: it may take years of research to turn the answers (often expressed as existence theorems) into practical implementations.

History: Information theory was first published in 1948 by **Claude Shannon**. He suggested some fundamentla limits on the representation and transmission of information. Since that time, the results have been extended to cover a variety of problem areas and people have worked (hard!) to find ways of achieving the bounds that the theory specifies is possible. In a sense, then, information theory has provided the theoretical motivation for many of the outstanding advances in digital communications and digital storage. For example, how much information can be sent over the phone system?

Besides the (almost) practical applications of the theory, there is great beauty and elegance in the theorems, the study of which has intrinsic merit in a university education.

The fundamental concept

One of the key (and initially counter-intuitive) concepts in information theory is that information is conveyed by *randomness.* This is information as defined in some mathematical sense, which is not identical to that which humans use. For example, it is possible to measure the amount of information in a page of typewritten text. Due to the structure of the English language, the amount of information conveyed by each letter in a word is *substantially* less than the 7-bit ASCII representation used. (It is somewhere over 2 bits/letter usually). There would be more information conveyed (in the mathematical sense) if the letters were completely random, instead of structured into words.

On the other hand, it is not too difficult to make the connection between randomness and information. Consider the tossing of a coin: if you know the outcome of the coin toss before it is tossed, then learning the outcome does not give you any more information. If you have a biased coin that is heads 90% of the time, then you gain very little information when you learn it is heads. On the other hand, you gain a fair amount of information when it comes up tails; the information is thus related somewhat to the degree of "surprise" at finding out the answer. Q: what weighting of the coin gives the maximum amount of information *on the average*?

Another very important concept that we will say more about later is that of **typical sequences**. In a sequence of bits of lenght n, there are some sequences which are (in a sense to be made precise later) typical. For example, for a sequence of coin-tossing outcomes for a fair coin, such as HHTHHTHTT, we would expect the number of heads and tails to be approximately equal (since the coin is fair). For an unbiased coin, we would expect the proportion of heads to go with the bias. Sequences that do not follow this trend, such as HHHHHHHHH, are thus *atypical*. A good part of information theory is capturing this concept of typicallity as precisely as possible and using it to concluding how many bits are needed to represent sequences of data. The basic idea is to try to use bits to represent only the typical sequences, since the others don't come up very often. (Of course, when they do come up, you don't want to just throw them away.) This concept of typical sequences is what the **asymptotic equipartition property** is all about, which is the topic of Chapter 3.

Suppose we have a discrete random variable X, and x is some particular outcome what occurs with probability $p(x)$. Then we assign to that event x the information that it conveys the uncertainty measure

$$\text{uncertainty} = -\log p(x).$$

The base of the logarithms determines the units of information. If $\log_2$ is used, then the units are in *bits*. If $\log_e$ (natural log) is used, then the units are in *nats*. While nats are not as familiar to engineers, it sometimes makes the computations slightly easier. Q: how do you convert from bits to nats?

Example 1 If a random variable has two outcomes, say $\mathbf{0}$ and $\mathbf{1}$, which each occur with probability $p(\mathbf{0}) = p(\mathbf{1}) = 0.5$, then each outcome conveys 1 bit of information. (Think about it.)

On the other hand, suppose $p(\mathbf{0}) = 1$ and $p(\mathbf{1}) = 0$: that is, the outcome $\mathbf{1}$ never happens. Then the information conveyed when $\mathbf{0}$ happens is 0: we get no information from it, because we knew all along that it would happen. However, the information that we get one $\mathbf{1}$ happens is ∞, since we are totally surprised by the occurrence of something that is impossible to happen. □

What is more commonly useful is the *average* uncertainty provided by a random variable X taking values in a space $\mathcal{X}$.

Definition 1 The **entropy** $H(x)$ os a discrete random variable X is

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x).$$

□

The entropy of an r.v. is a measure of the uncertainty of the random variable. It is a measure of the amount of information required on the average to describe the random variable.

Example 2 Take a fair coin:

$$H(X) = -(0.5 \log 0.5 + 0.5 \log 0.5) = 1 \text{ bit}.$$

For a biased coin ($p(\mathbf{0}) = 0.9$),

$$H(X) = -(0.9 \log 0.9 + 0.1 \log 0.1) = 0.469 \text{ bits}.$$

□

Example 3 What about a r.v. with three outcomes? □

Notation: We shall use the operator E to denote **expectation**. If $X \sim p(x)$ (read as: X is distributed according to $p(x)$), then for some function of the random variable $g(X)$,

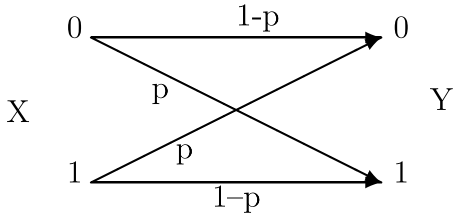
$$Eg(X) = \sum_{x \in \mathcal{X}} g(x)p(x).$$

(also known as the law of the unconcious statistician.) Recall EX, EX^2, etc. Then for $g(x) = \log 1/p(x)$,

$$H(X) = Eg(X) = E \log 1/p(x)$$

Some simple discrete channel models

The binary symmetric channel:



The parameter p is the *probability of error*. For many channels we can explicilty compute p. For example, for BPSK

$$p = Q\left(\sqrt{\frac{2E_b}{N_0}}\right)$$

The channel is characterized by its conditional probabilities:

$$P(Y = 0|X = 0) = 1 - p \qquad P(Y = 1|X = 0) = p$$

$$P(Y = 0|X = 1) = p \qquad P(Y = 1|X = 1) = 1 - p$$