Information Theory
- John R Pierce

Abstract

This document contains a brief review of the book "Information theory".

The World and theories

The first chapter is a leisurely introduction in to the types of mathematical theories. It basically conveys the message that communication theory as Shannon has given it to us deals in a very broad and abstract way with certain important problems of communication and information, but it cannot be applied to all the problems which we can phrase using the words *communication* and *information* in their popular senses. Communication theory deals with certain aspects of communication which can be associated and organized in a useful and fruitful way.

The Origins of Information theory

The chapter begins by giving two reasons for studying the history of any scientific discipline.

1. Almost all the important contributions always arise from studying and refining simple man-made devices. Man learns from devices rather than the vagaries of nature. One can compress the existing knowledge by making a ton of assumptions, build a simple device and over a period of time many others can contribute to the development of powerful devices thus furthering knowledge relating to a specific field. In that sense, communication theory can be best understood from the point of view of a teletype machine.

2. One can understand with what difficulty understanding is won.

Indeed these two reasons should motivate anyone to understand the history of the subject they are learning.

The author, at the outset, requests the reader not to confuse the term *entropy* in Information theory with *entropy* in thermodynamics or communication theory. The only commonality is the mathematical representation. The usage of the term *entropy* is completely different. In thermodynamics, entropy is an indication of a reversible process, i.e. transfer of thermal energy to mechanical energy and back. In statistical mechanics, entropy is used as a measure of disorder, if entropy increases, there is more disorder. In Information theory, the term entropy is used to denote the amount of uncertainty in the message that is to be transmitted. The larger the uncertainty, the larger is the entropy. The larger the entropy, the larger is the information conveyed by a message from a source. Communication theory had its origins in the study of electrical communications, not in statistical mechanics.

While sending any message as sinusoid, there is phase delay, amplitude attenuation and hence the message received is not going to be an exact replica of the original message. However there is something that is invariant, the frequency of the sinusoid. This was the brilliant insight of Fourier. Nyquist using these observations came up with a law that gave a relationship between speed of transmission and the number of different symbols available. Nyquist showed that the number of distinct, different current values which can be sent over a circuit per second is twice the total range of frequencies used. After these developments, developments in the communication theory took a backstage until World War II. During war, it became important to detect the course of airplanes from the noisy radar data. This problem was solved by Weiner and Kolmogoroff.

There are various aspects to think about in sending messages through a channel. The signal rate of the sender, the channel capacity, the noise in the channel etc. At the receiving end, one is trying to decipher the content of the message. There are two types of situations one can come across. It is possible that an unknown message is mixed with unknown noise and is sent to the receiver. How do you tease out a message from an ensemble of messages mixed with noise? This was a problem addressed by Weiner and Kolmogoroff. Shannon's work relates to a different problem. Given a message, how do you encode it efficiently so that it can be sent via a noisy medium?. To reiterate, Weiner and Kolmogoroff solved the

problem of extracting signal from a combination of unknown signal and unknown noise. Shannon worked on the problem of effective encoding of a message so that when the encoded message combined with noise gets to the receiver, the receiver will be able to efficiently decode the message.

A Mathematical Model

Theorems in Information theory assume an ergodic source. This chapter gives an elaborate explanation of stochastic processes that are ergodic. English text is used as an example and an attempt is made to build a mathematical model that replicates English text. Firstly, it has been observed that the english text can be thought of approximately ergodic. This means that if you compare the probability of occurrence of alphabet E in all the texts vs. probability of its occurrence in a message written by a single individual, they tend to be approximately equal. Most of the literature in econometrics assumes that the process is ergodic, i.e. ensemble average is equal to the time average. By the way ensemble average need not be time average always. Lets take insurance companies for example. They get the ensemble average right and then go about setting the premium based on it. Why is ergodicity important ? Well you can pick a specific sequence and do all the statistics on the sequence and be assured that the statistics hold good for the ensemble. Ergodicity is used in math theorems and proofs. However in real world, we only see an approximate ergodicity. Hence we must be careful about applying the rigorous math theories. Having said that we can think of stochastic processes that produce English text and build a communication model. That's exactly what Shanon achieved.

Encoding and Binary Digits

The chief aim of information theory is to study how ergodic sequences of characters and signals can be most effectively coded for transmission, commonly by electronic means. To send a signal, one can use a pulse and a space. People find it easy to send it via binary representation 0 and 1. In this digitized era, most of the communication is via 0's and 1's. So, the key question is to figure out the best way of encoding the messages from a message source, a way which calls for the least number of binary digits. This chapter gives a brief idea in to this problem by taking up an example of encoding English text. There are various ways to encode the text.

- If you take 26 letters of English alphabet and a space, you need 5 bits per symbol $2^5 > 27$. You will need 27.5 binary digits for a word(assuming an average of 5.5 alphabets per word)
- If you expand the gamut of characters to 50, you will need 6 bits per symbol $2^6 > 50$. You will need $6 * 5.5 = 33$ binary digits for a word.
- If you take 16384 frequently occurring words , then you will need 14 bits ($2^{14} > 16384$).
- If you do block encoding containing three alphabets then you will need 17 bits per 3 symbols, i.e less than 6 per symbol. You will need 31.2 binary digits for a word.
- If you spell out , each word has approximately 5.5 characters, you will need $5.5 * 5 = 27.5$ bits per word.

All this leads to the obvious question, *Is there an efficient way to encode ?*

Entropy

One needs to forget the definition of entropy that is given in physics to better understand its usage in information theory. The entropy of communication theory is measured in bits and we refer to message containing so many bits /symbol or bits/ message etc. The definition of entropy for an ergodic source is given by

$$H = -\sum_i p_i \log p_i$$

where i refers to the different symbols in the message and p_i refers to the probability of occurrence of the message. To send the outcome of fair coin,you will need one bit. To send the outcome of a biased coin, you will need less than one bit. The definition says that if a source has to choose between 2^N symbols that are equally possible, then its entropy is N bits. How do we use this measure ?

What about using it on English words. Zif's law is an empirical law that relates the probability of occurrence of a word to its rank. In order to create a suitable probability distribution, Shannon took the first 8727 words and computed the entropy to be 9.14 bits per word.

Shannon showed that the number of binary digits required to transmit a message is just the entropy in bits per symbol times the number of symbols. If you send alphabet by alphabet, the entropy will be higher as you are not exploiting the structure. However if you chunk up the message in to long blocks, then the dependency of each block with respect to previous blocks goes down. Hence one can write the entropy of an ergodic English source as

$$H_N = -\sum_{i,j} p(B_i) p_{B_i}(S_j) \log p_{B_i}(S_j)$$

where $P(B_i)$ is the marginal probability of the block B_i and $p_{B_i}(S_j)$ is the conditional probability that you S_j will be preceded by B_i, and N is the length of the block. As $N \to \infty$, the entropy tends towards the entropy of the source. To code the blocks, one can use Huffman code so that entropy of an ergodic source measured in bits is closer to Shannons' entropy.

At the end of this chapter, a reader gets a fair idea of entropy; the entropy of a signal source in bits per symbol or second gives the average number of binary digits per symbol or second.With connection to a message source, we think of entropy as a measure of choice with the source to send from a set. With connection to message recipient, the entropy is measure of uncertainty with which the message is received. The remarkable thing about thinking in such a mode is that we can characterize the capacity of the channel based on the largest possible entropy of a message that can be transmitted.

Language and Meaning

A brief digression in to discussing language from the point of information theory. I think nothing is lost by skipping this chapter from the book.

Efficient Encoding

Why is there a need for efficient encoding ? If the source produces a message that has random symbols with each symbol occurring with a specific probability, then one can measure the entropy of the information source in bits/symbol or bits/second;This gives us a benchmark for the number of binary digits/symbol or binary digits/second that we might use for transmitting a message.

> One way to achieve a number close to entropy is to divide the message in to successive blocks of characters, to each of which a probability of occurrence can be attached, and by encoding these blocks into these binary digits by means of the Huffman code; the number of digits used per character approaches the entropy as the block of characters are made longer and longer

So, why not follow this approach for every communication problem ? Because you can better this mathematical approach to the problem by studying the source carefully. Every one knows the solution to a cubic equation but seldom does one tends to use it in practice. Some approximation to the class of cubics is used. In the same way even though the approach of blocking and Huffman coding is always there as a mathematical framework, in practice there is a symmetry in the message that can be exploited for better transmission rates. One can always send English text in a scanned form, the downside is that the transmission is not taking in to consideration the obvious structure in English language. So, if we proceed mechanically to encode pairs, triads, etc , we might encode many sequences which aren't English words. Hence it seems logical to go with a larger unit of English text, the word. Can this be taken to the next level ? Why not use grammatical rules to create even more efficient encoding? Now this is where we hit some kind of limit. The trouble is we ourselves do not know the rules of grammar to encode the text. However it is important to at least get an idea of what *could* be accomplished. Shannon carried out this experiment and found that somewhere between 0.6 and 1.3 binary digits per character is required.

Given today's transmission rates, sending text is no longer an issue. The bigger problems are sending high quality messages for voice transmission and TV/video transmission. How do we go about computing the entropy of an audio source ? video source ? Does it make sense to talk about such a thing at all? In these kind of problems, we are moving in to a continuous world and hence the first problem we face is discretization. In the case of audio transmission and video transmission, the systems should take in to account the limitations on the source and destination of the signal, i.e. human voice, ear, eye capabilities to send just enough information so that there is no perceivable difference. What's the point in sending an accurate voice message when human ear can only make sense of certain frequencies ? Think of image compression for example. You can find a good basis for the pixel data, compress the data, send it to the destination, apply inverse transform to deliver the image to the receiver. If done properly, the human eye cannot detect the compressed image.
As we see, there are three important principles in encoding signals efficiently

1. Don't encode the signal one sample or one character a time; encode a considerable stretch of a signal at a time
2. Take in to account the limitations on the source of the signal
3. Take in to account any inabilities of the eye or eat to detect error in a reconstruction of the signal

While the quest for the efficient encoding is an ongoing one, there is something that needs to be kept in mind - The noise in the channel. Even the best transmissions can be corrupted by noise. Hence the balancing force of redundancy in the message should always be kept in mind. Efficient encoding and redundancy in the message should be balanced intelligently. Does that mean that the useful relationship between entropy and the number of binary digits equal for transmission has to tempered down, while applying it in practice ?

Noisy Channel

The more efficient the message encoding, the lesser is the redundancy in the transmission. The lesser the redundancy in the message, the more the noise can play havoc with it. A naive approach is to send the symbols in the message twice or thrice or four times. If you send the each symbol twice, you will be able to only detect the error. If you send the symbol thrice, you will be able to detect and correct the message(Think about why?). The downside to mere repetition is that the length of the message is increased and subsequently transmission rate are reduced. We need a measure to quantify the entropy when there is noise in the channel. If we denote x as the character sent by the source and y as the character received, there are four quantities of interest :

- $H(x)$: The uncertainty as to x, that is, as to which character will be transmitted.
- $H(y)$: The uncertainty as to y, that is, as to which character will be received.
- $H_x(y)$: The uncertainty of x being transmitted and y being received.
- $H_y(x)$: The uncertainty of receiving y when x is transmitted.

The uncertainty as to which symbol was transmitted when a given symbol is received,$H_y(x)$ seems a natural measure of the information lost in transmission. This is termed as *equivocation*. The rate of transmission R can be represented as $H(x) - H_y(x)$ bits per second. This is where the rubber meets the road. You have to choose the message source in such a way that R is maximized.Turn the statement around and we can define the maximum possible rate of transmission for the channel as the *channel capacity*. Shannon's fundamental theorem of noisy channels says that if the entropy of the source H is less than the channel capacity C, then the message can be transmitted over the channel with an arbitrary small frequency of errors. If the entropy of the channel capacity is greater than that of the source, efficient encoding can be done to reach the lower bound of equivocation $C - H + \epsilon$. This theorem lead to a massive research on error correcting codes(Hamming, Golay, Convolution coding).

The key shift in perspective via Shannon's second theorem is this: Block encoding is being done to introduce redundancy to non redundant messages so that the messages can be transmitted over noisy channels. This is different from block transmission introduced to reduce the entropy of the source. For example, using blocks of english alphabets reduces the entropy as compared to sending alphabets. However to send it via a noisy channel, one needs to add redundancy in the form of block encoding.

> The whole problem of efficient and error-free communication turns out to be that of removing from messages the somewhat inefficient redundancy which they have and then adding redundancy of the right sort in order to allow correction of errors made in transmission.

Many Dimensions

This chapter uses the geometry of an n dimensional sphere to derive the following relation:

$$C = W \log(1 + P/N)$$

where C is the channel capacity, W is the signal bandwidth, P, N are the powers of signal and noise. This relationship says that increasing signal power is the not the only way to achieve maximum transmission rate. Many combinations of W, P, N can produce the same transmission rate. A highlight of this chapter is the use of multi-dimensional space to attack a real life problem in communication theory. No one knows when abstract math finds life in applied math!.

The last four chapters relate the principles of Information theory to Physics, Psychology, Cybernetics and Art. The book concludes by delving in to some of the areas where Information theory has been used since Shannon's work. The author qualifies *Kelly criterion* as the only mathematical established interpretation, other than those concerned with the rate of generation of probable messages and their efficient encoding for transmission, that anyone has discovered.