LECTURE 13

INFORMATION THEORY

Information Theory was created by C. E. Shannon in the late 40's. The management of Bell Telephone Labs wanted him to call it "Communication Theory" as that is a far more accurate name, but for obvious publicity reasons "Information Theory" has a much greater impact - this Shannon chose and so it is known to this day. The title suggests that the theory deals with information - and therefore it must be important since we are entering more and more deeply into the information age. Hence I shall go through a few main results, not with rigorous proofs of complete generality, but rather intuitive proofs of special cases, so that you will understand what information theory is and what it can and cannot do for you.

First, what is "information"? Shannon identified information with surprise. He chose the negative of the log of the probability of an event as the amount of information you get when the event of probability p happens. For example, if I tell you it is smoggy in Los Angles then p is near 1 and that is not much information, but if I tell you that it is raining in Monterey in June then that is surprising and represents more information. Because log 1 = 0 the certain event contains no information.

In more detail, Shannon believed that the measure of the amount of information should be a continuous function of the probability p of the event, and for independent events it should be additive - that what you learn from each independent event when added together should be the amount you learn from the combined event. As an example, the outcome of the roll of a die and the toss of a coin are generally regarded as independent events. In mathematical symbols, if I(p) is the amount of information you have for event of probability p, then for event x of probability p_1 and for the independent event y of probability p_2, you will get for the event of both x and y

$$I(p_1 p_2) = I(p_1) + I(p_2) \quad \text{(x and y independent events)}$$

This is the Cauchy functional equation, true for all p_1 and p_2.

To solve this functional equation suppose

$$p_1 = p_2 = p$$

then this gives

$$I(p^2) = 2I(p)$$

If $p_1 = p^2$ and $p_2 = p$, then

$$I(p^3) = 3I(p)$$

etc. Extending this process you can show, via the standard method used for exponents, that for all rational numbers m/n

$$I(p^{m/n}) = (m/n)I(p)$$

From the assumed continuity, of the information measure it follows that the log is the only continuous solution to the Cauchy functional equation.

In information theory it is customary to take the base of the log system as 2, so that a binary choice is exactly 1 bit of information. Hence information is measured by the formula

$$I(p) = -\log_2 p = \log_2(1/p).$$

Let us pause and examine what has happened so far. First, we have not defined "information", we merely gave a formula for measuring the amount. Second, the measure depends on surprise, and while it does match, to a reasonable degree, the situation with machines, say the telephone system, radio, television, computers, and such, it simply does not represent the normal human attitude towards information. Third, it is a relative measure, it depends on the state of your knowledge. If you are looking at a stream of "random numbers" from a random source then you think that each number comes as a surprise, but if you know the formula for computing the "random numbers" then the next number contains no surprise at all, hence contains no information! Thus, while the definition Shannon made for information is appropriate in many respects for machines, it does not seem to fit the human use of the word. This is the reason it should have been called "Communication Theory", and not "Information Theory". It is too late to undo the definition (which produced so much of its initial popularity, and still makes people think that it handles "information") so we have to live with it, but you should clearly realize how much it distorts the common view of information and deals with something else, which Shannon took to be surprise.

This is a point that needs to be examined whenever any definition is offered. How far does the proposed definition, for example Shannon's definition of information, agree with the original concepts you had, and how far does it differ? Almost no definition is exactly congruent with your earlier intuitive concept, but in the long run it is the definition that determines the meaning of the concept - hence the formalization of something via sharp definitions always produces some distortion.

Given an alphabet of q symbols with probabilities p_i then the average amount of information, (the expected value), in the system is

$$H(P) = \mathrm{SUM}[i=1,q;\ p_i I(p_i)] = \mathrm{SUM}[i=1,q:\ p_i \log 1/p_i)]$$

This is called the entropy of the system with the probability distribution $\{p_i\}$. The name "entropy" is used because the same mathematical form arises in thermodynamics and in statistical mechanics, and hence the word "entropy" gives an aura of importance that is not justified in the long run. The same mathematical form does not imply the same interpretation of the symbols!

The entropy of a probability distribution plays a central role in coding theory. One of the important results is Gibbs' inequality for two different probability distributions, p_i and q_i. We have to prove

$$SUM[p_i \log\{q_i/p_i\}] \leq 0$$

The proof rests on the obvious picture, Figure 13-1, that

$$\log x \leq x - 1, \qquad (0 \leq x < \infty),$$

and equality occurs only at $x = 1$. Apply the inequality to each term in the sum on the left hand side

$$SUM[p_i\{q_i/p_i - 1\}] = SUM[q_i] - SUM[p_i] = 1 - 1 = 0$$

If there are q symbols in the signaling system then picking the $q_i = 1/q$ we get from Gibbs' inequality, by transposing the q terms,

$$H(P) \leq \log q$$

This says that in a probability distribution if all the q symbols are of equal probability, $1/q$, then the maximum entropy is exactly ln q, otherwise the inequality holds.

Given a uniquely decodable code we have the Kraft inequality

$$K = SUM[1/2^{l}i] \leq 1$$

Now if we now define the pseudo probabilities

$$Q_i = 2^{-l}i/K$$

where of course $SUM[Q_i] = 1$, it follows from the Gibbs' inequality,

$$SUM[i=1,q;\ p_i \log\{1/(Kp_i 2^{l}i\}] \leq 0$$

after some algebra (remember that $K \leq 1$ so we can drop the log term and perhaps strengthen the inequality further),

$$H(p) \leq \log K + SUM[p_i l_i] \leq L = \text{average code length}$$

Thus the entropy is a lower bound for any encoding, symbol to symbol, for the average code length L. This is the noiseless coding theorem of Shannon.

We now turn to the main theorem on the bounds on signaling systems that use encoding of a bit stream of independent bits and go symbol to symbol in the presence of noise, meaning that there is a probability that a bit of information is correct, $P > 1/2$, and the corresponding probability $Q = 1 - P$ that it is altered when it is transmitted. For convenience assume that the errors are independent and are the same for each bit sent, which is called "white noise".

We will encode a long stream of n bits into one encoded message, the n-th extension of a one bit code, where the n is to be determined as the theory progresses. We regard the message of n bits as a point in an n-dimensional space. Since we have an n-th extension, for simplicity we will assume that each message has the same probability of occurring, and we will assume that there are M messages, (M also to be determined later), hence the probability of each initial message is

$$1/M$$

We next examine the idea of the <u>channel capacity</u>. Without going into details the channel capacity is defined as the maximum amount of information that can be sent through the channel reliably, maximized over all possible encodings, hence there is no argument that more information can be sent reliably than the channel capacity permits. It can be proved that for the binary symmetric channel (which we are using) the capacity C, per bit sent, is given by

$$C = 1 - H(P) = 1 - H(Q)$$

where, as before, P is the probability of no error in any bit sent. For the n independent bits sent we will have the channel capacity

$$nC = n\{1 - H(P)\}$$

If we are to be near channel capacity then we must send almost that amount of information for each of the symbols a_i, $i = 1, \ldots, M$, and all of probability $1/M$, and we must have

$$I(a_i) = n\{C - e_1\}$$

when we send any one of the M equally likely messages a_i. We have, therefore

$$M = 2^{n(C - e_i)} = 2^{nC}/2^{ne_1}$$

With n bits we expect to have nQ errors. In practice we will have, for a given message of n bits sent, approximately nQ errors in the received message. For large n the relative spread, (spread = width, √variance) of the distribution of the number of errors will be increasingly narrow as n increases.

From the sender's point of view I take the message a_i to be sent and draw a sphere about it of radius

$$r = (Q + e_2)n \qquad (e_2 > 0,\ Q + e_2 < 1/2)$$

which is slightly larger by e_2 than the expected number of errors, Q, Figure 13-2. If n is large enough then there is an arbitrarily small probability of there occurring a received message point b_j that falls outside this sphere. Sketching the situation as seen by me, the sender, we have along any radii from the chosen signal, a_i, to the received message, b_j, with the probability of an error is (almost) a normal distribution, peaking up at nQ, and with any given e_2 there is an n so large that the probability of the received point, b_j, falling outside my sphere is as small as you please.

Now looking at it from your end, Figure 13-3, as the receiver, there is a sphere S(r) of the same radius r about the received point, b_j, in the space, such that if the received message, b_i, is inside my sphere then the original message a_i sent by me is inside your sphere.

How can an error arise? An error can occur according to the following table:

case	a_i in S(r)	another in S(r)	meaning
1	yes	yes	error
2	yes	no	no error
3	no	yes	error
4	no	no	error

Here we see that if there is at least one other original message point in the sphere about your received point then it is an error since you cannot decide which one it is. The sent message is correct only if the sent point is in the sphere and there is no other code point in it.

We have, therefore, the mathematical equation for a probability p_E of an error, if the message sent is a_i,

$$P_E = P\{a_i \text{ not in } S(r)\} + P\{a_i \text{ is in } S(r)\}$$

$$\times P\{\text{at least one more } a_j \text{ is in } S(r)\}$$

We can drop the first factor in the second term by setting it equal to 1, thus making an inequality

$$P_E \leq P\{a_i \text{ not in } S(r)\} + P\{\text{at least one more } a_j \text{ is in } S(r)\}$$

But using the obvious fact that

$$P\{E_1 \text{ and/or } E_2\} = P\{E_1\} + P\{E_2\} - P\{E_1E_2\}$$

hence

$$P\{E_1 \text{ and/or } E_2\} \leq P\{E_1\} + P\{E_2\}$$

applied repeatedly to the last term on the right

$$P_E \leq P\{a_i \text{ not in } S(r)\} + \text{SUM}[\text{all } a_j \text{ except } a_i ; P\{a_j \text{ in } S(r)\}]$$

By making n large enough the first term can be made as small as we please, say less than some number d. We have, therefore,

$$P_E \leq d + \mathrm{SUM}[P(a_j \text{ in } S(r)\}]$$

We now examine how we can make the code book for the encoding of the M messages, each of n bits. Not knowing how to encode, error correcting codes not having been invented as yet, Shannon chose a random encoding. Toss a penny for each bit of the n bits of a message in the code book, and repeat for all M messages. There are nM tosses hence

$$2^{nM}$$

possible code books, all books being of the same probability $1/2^{nM}$. Of course the random process of making the code book means that there is a chance that there will be duplicates, and that there may be code points that are close to each other and hence will be a source of probable errors. What we have to prove is that this does not occur with a probability above any positive small level of error you care to pick - provided n is made large enough.

The decisive step is that Shannon <u>averaged over all possible code books</u> to find the average error! We will use the symbol Av[.] to mean average over the set of all possible random code books. Averaging over the constant d of course gives the constant, and we have, since for the average each term is the same as any other term in the sum,

$$\mathrm{Av}[P_E\} \leq d + (M-1)\mathrm{Av}[P(a_j \text{ in } S(r)\}]$$

which can be increased, (M-1 goes to M),

$$\mathrm{Av}[P_E] \leq d + M\ \mathrm{SUM}[\text{for all } a_j \text{ not } a_i;\ P(a_i \text{ in } S(r)\}]$$

For any particular message, when we average over all code books, the encoding runs through all possible values, hence the average probability that a point is in the sphere is the ratio of the volume of the sphere to the total volume of the space. The volume of the sphere is

$$1 + C(n,1) + C(n,2) + \ldots + C(n,ns)$$

where $s = Q + e_2 < 1/2$, and ns is supposed to be an integer.

The largest term in this sum is the last (on the right). We first estimate the size of it, via Stirling's formula for the factorials. We then look at the rate of fall off to the next term before it, note that this rate increases as we go to the left, and hence we can: (1) <u>dominate</u> the sum by a geometric progression with this initial rate, then (2) extend the geometric progression from ns terms to an infinite number, (3) sum the geometric progression (all standard algebra of no great importance) and we finally get (4) the bound (for n large enough)

$$1 + C(n,1) + C(n,2) + \dots + C(n,ns) \le 2^{nH(s)} \qquad (s < 1/2)$$

Note how the entropy H(s) has appeared in a binomial identity.

We have now to assemble the parts, note that the Taylor series expansion of $H(s) = H(Q + e_2)$ gives a bound when you take only the first derivative term and neglect all others, to get the final expression

$$Av[P_E] \le d + 2^{-n(e_1 - e_3)}$$

where

$$e_3 = e_2 \ln\{(1 - Q)/Q\} \qquad (Q < 1/2)$$

All we have to do now is pick an e_2 so that $e_3 < e_1$ and the last term will get as small as you please with sufficiently large n. Hence the average error of P_E can be made as small as you please while still being as close to channel capacity C as you please.

If the average over all codes has a suitably small error, then at least one code must be suitable - hence there exists at least one suitable encoding system. This is Shannon's important result, the "noisy coding theorem", though let it be noted that he proved it in much greater generality than the simple binary symmetric channel I used. The mathematics is more difficult in the general case, but the ideas are not so much different, hence the very particular case used suffices to show you the true nature of the theorem.

Let us critique the result. Again and again we said, "For sufficiently large n." How large is this n? Very, very large indeed if you want to be both close to channel capacity and reasonably sure you are right! So large, in fact, that you would probably have to wait a very long time to accumulate a message of that many bits before encoding it, let alone the size of the random code books (which being random cannot be represented in a significantly shorter form than the complete listing of all Mn bits, both n and M being very large).

Error correcting codes escape this waiting for a very long message and then encoding it via a very large encoding book, along with the corresponding large decoding book, because they avoid code books and adopt regular (computable) methods. In the simple theory they tend to lose the ability to come very near to the channel capacity and still keep an arbitrarily low error rate but when a large number of errors are corrected by the code they can do well. Put into other words, if you provide a capacity for some level of error correction then for efficiency you must use this ability most of the time or else you are wasting capacity, and this implies a high number of errors corrected in each message sent.

But the theorem is not useless! It does show, in so far as it is relevant, that efficient encoding schemes must have very elaborate encodings of very long strings of bits of information. We see this accomplished in the satellites that passed the outer planets; they corrected more and more errors per block as they got farther and farther from both the Earth and the Sun (which for some satellites supplied the solar power of about 5 watts at most, others used atomic power sources of about the same power). They had to use high error correcting codes to be effective, given the low power of the source, their small dish size, the limited size of the receiving dishes on Earth as seen from their position in space, and the enormous distances that the signal had to travel.

We return to the n-dimensional space which we used in the proof. In the discussion of n-dimensional space we showed that almost all the volume of a sphere lay near the outer surface - thus for the very slightly (relatively) enlarged sphere about the received signal it is almost certain that the original sent signal lies in it. Thus the error correction of an arbitrarily large number of errors, nQ, with arbitrarily close to no errors after decoding is not surprising. What is more surprising is that the M spheres can be packed with almost no overlap - again an overlap as small as you please. Insight as to why this is possible comes from a closer examination of the channel capacity than we have gone into, but you saw for the Hamming error correcting codes that the spheres had <u>no</u> overlap. The many almost orthogonal directions in n-dimensional space indicates why we can pack the M sphere into the space with little overlap. By allowing a slight, arbitrarily small amount, of overlap which can lead to only a very few errors in your decoding you can get this dense packing. Hamming guaranteed a certain level; Shannon only a probably small error but as close to the channel capacity as you wish, which Hamming codes do not do.

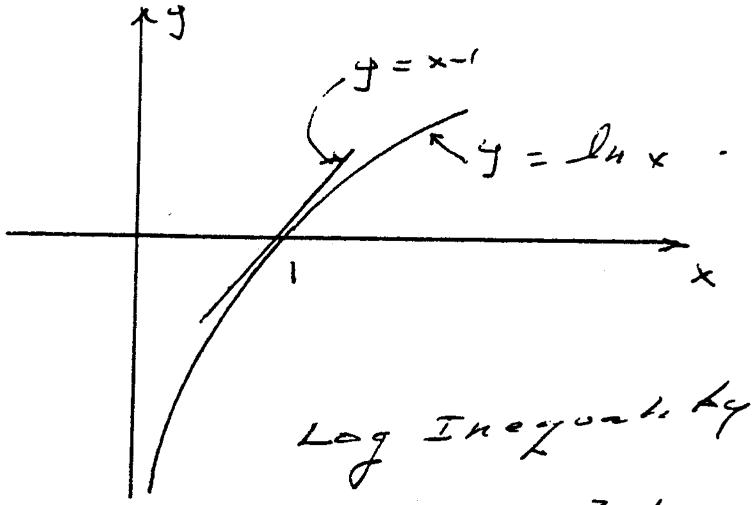
Information theory does not tell you much about how to design, but it does point the way towards efficient designs. It is a valuable tool for engineering communication system between machine-like things, but as noted before it is not really relevant to human communication of information. The extent to which biological inheritance, for example, is machine-like, and hence you can apply information theory to the genes, and to what extent it is not and hence the application is irrelevant, is simply not known at present. So we have to try, and the success will show the machine-like character, while the failure will point towards the fact that other aspects of information are important.

We now abstract what we have learned. We have seen that all initial definitions, to a larger or smaller extent, should get the essence of our prior beliefs, but they always have some degree of distortion and hence non-applicability to things as we thought they were. It is traditional to accept, in the long run, that the definition we use actually defines the thing defined; but of course it only tells us how to handle things, and in no way actually tells us any meaning. The postulational approach, so strongly favored in mathematical circles, leaves much to be desired in practice.

We will now take up an example where a definition still bothers us, namely IQ. It is as circular as you could wish. A test is made up that is supposed to measure "intelligence", it is revised to make it as consistent internally as we can, and then it is declared, when calibrated by a simple method, to measure "intelligence" which is now normally distributed (via the calibration curve). All definitions should be inspected, not only when first proposed, but much later when you see how they are going to enter into the conclusions drawn. To what extent were the definitions framed as they were to get the desired result? How often were the definitions framed under one condition and are now being applied under quite different conditions? All too often these are true! And it will probably be more and more true as we go farther and farther into the softer sciences, which is inevitable during your life time.

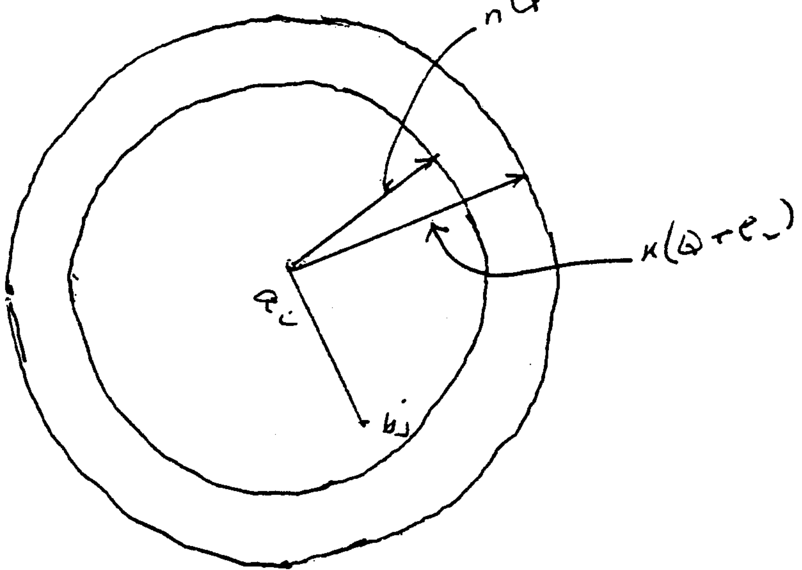
Thus one purpose of this presentation of information theory, besides its usefulness, is to sensitize you to this danger, or if you prefer, how to use it to get what you want! It has long been recognized that the initial definitions determine what you find, much more than most people care to believe. The initial definitions need your careful attention in any new situation, and they are worth reviewing in fields in which you have long worked so you can understand the extent that the results are a tautology and not real results at all.

There is the famous story by Eddington about some people who went fishing in the sea with a net. Upon examining the size of the fish they had caught they decided that there was a minimum size to the fish in the sea! Their conclusion arose from the tool used and not from reality.



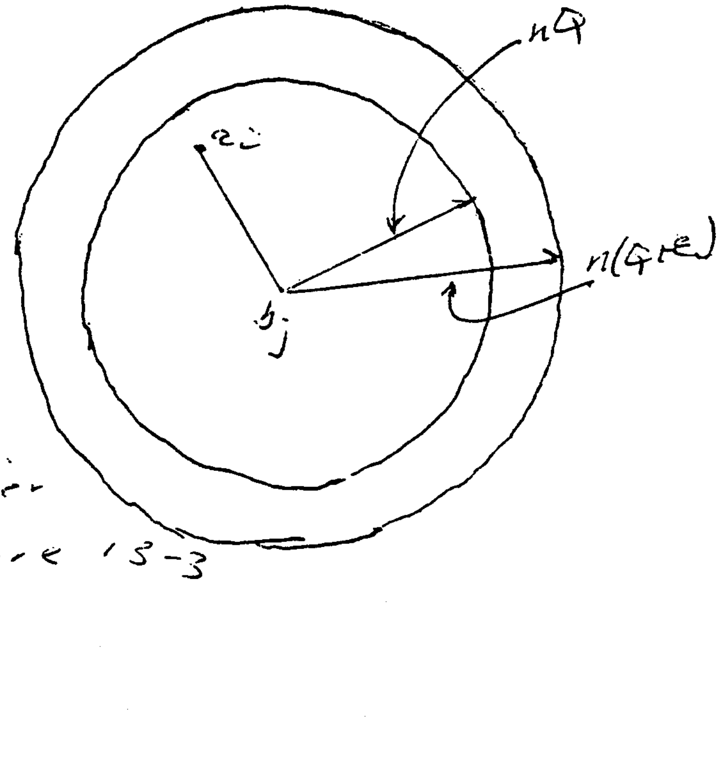
Log Inequality

Figure 13-1



Sender

Figure 13-2



Receiver

Figure 13-3