Information and communication

What is called 'information theory' today began as a “Mathematical theory of communication presented in two papers by C.E.Shannon in 1948 in two long papers in the Bell System Technical Journal. Because the ideas have been taken well beyond the original context, not always with the original author's approval, it is worth beginning with communication. This is both because of the intrinsic interest of Shannons work, the wide use of ICT today, and also so as to appreciate the appropriateness, or othewise, of later uses.

The fundamental idea is simple to state and hard to disagree with. If I tell you something you already know, then no information has been conveyed. We need a measure of the ignorance or uncertainty before a message is received, to assess how much we have gained by receiving the message. One must ask – ignorance of what? In the Shannon model, this is taken care of by postulating an ensemble – a collection of possible messages, each associated with a probability. A very elementary example is the anxious father who sees the nurse bringing a bundle towards him. We have two options, boy or girl, with equal proabaiity. Surely, this should be the unit of information. Since it can also be represented as a zero or a one with equal probability, it is called the 'binary digit' or 'bit' for short. Bits require no introduction to anyone who pays a mobile data bill.

More generally, if there are W alternatives, each with probability $p=1/W$, then the uncertainty associated with this situtation is defined to be $H=\log(W) =-\log p$. All logarithms are to the base 2. Why not use W itself as a measure? The reason is that when we have many independent situations, W gets multiplied. Six fathers viewing six nurses carrying six bundles corresponds to $W=2^6 =64$. However, if we take the log to the base 2, we have H=6 bits, which seems reasonable – H is additive while W is multiplicative.

What about the case when the alternatives are not equally probable? Shannons proposal, and that of Boltzmann well before him, goes as follows. The “entropy' associated with a discrete probability distribution p_i , where i goes from 1 to s (the number of 'states') is given by $H=-\sum_i p_i \log p_i$ It's not unreasonable, it is the average of $-\log p$, using the same probability distribution as weight. This expression has a formal derivation: One can take N independent trials of which $N_i=N p_i$ result in the outcome i . W is then the number of ways of dividing N into s groups, of size N_i . Then one takes the logarithm, this will be N times the formula given earlier for H.

This expression has some desirable properties. First let us put all the probabilities equal to $1/W$. We then get back the previous expression $\log W$. Further, any event with zero probability contributes zero to this expression And if one of the probabilities is 1, this also contributes zero, and so do all the other alternatives since their p's are zero. So a situation with a certain outcome carries zero entropy and zero information.

Another consistency check is that if we have two distributions, relating to two independent situations, then the entropies add. In symbols, we have two probabiity distributions, $p_i, 1\le i\le s$ and $q_j, 1\le j\le t$ which have entropies H_P and H_Q associated with them. When we consider both the distributions jointly, the probability of the event (ij) is $p_i q_j$ We can calculate the entropy of this joint probability distribution, and it comes out to be $H_{PQ}=H_P+H_Q$. All these properties are reasonable, and in fact if any one of them failed, our measure of uncertainty wojuld be unreasonable.

The Russian probabililist Khinchin has a book, reproduced by Dover, called Mathematical foundations of information theory, which tries to actually derive this expression from desirable properties, rather than any cobinatorics.

We now consider a situation when the events P and Q are not independent. In that case, we have always, $H_{PQ} \leq H_P + H_Q$ (It requires proof, deferred). This is a very interesting test of correlation. It tells you that the occurence of P gives you information about Q, So our ingnorance of Q is reduced by our knowledge of P. It does not care whether the two are correlated, or anticorrelated, or even something more complicated. In that sense, $H_P + H_Q - H_{PQ}$ goes beyond the usual measures of correlation that one learns about in statistics. In the extreme case that we can uniquely predict Q once we know that P has occurred, we have $H_{PQ} = H_P$ Notice the lack of symmetry – it is not necessarily true that P is uniquely predictable from Q. so in general, $H_Q \leq H_P$ And if it also happens to be true that each is uniquely predictable from the other, then the equality sign holds, $H_P = H_Q$.

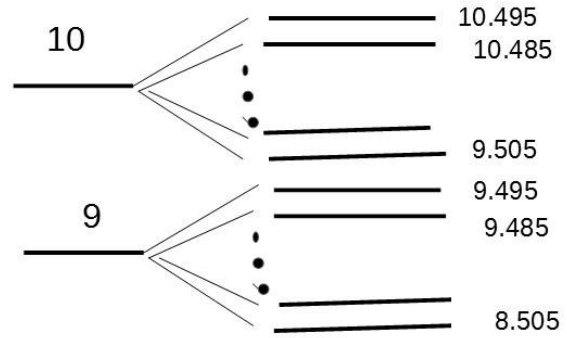
So far, we just have “preparation” and “propeties” but no “uses”. The first use we look at is compression., illustrated by the following example. Let us say we have a message of N characters, each of which comes from a two letter alphabet, A and B. If they occurred with equal probability, W would be 2^N , and we would have $H = N$ bits – the information per character would be 1 bit. Now supose the we change the nature of the source, so that A occurs with probability ¾ and B with ¼ . According to the Shannon measure, the information per character is now $-0.25\log(0.25) - 0.75\log(.75) \simeq 0.81$. If this is to be taken seriously, a message of 100 characters carries only 81 bits of information. The Shannon coding theorem states, informally, that it is indeed possible to transmit such a message using 81 bits, using a suitable code (conditions apply) . Let us put aside the question of constructing the code, for the moment. The fact that we have 100 two-valued characters conveying 81 bits of information means that there is redundancy in the original message, and this is exploited to compress it. This is not such a strange idea – we use it all the time. If a mobile number was 9999999999, we would simply say ten nines, and not bother to list out all the digits. But the real achievement is to quantify and prove that one can exploit the redundancy.

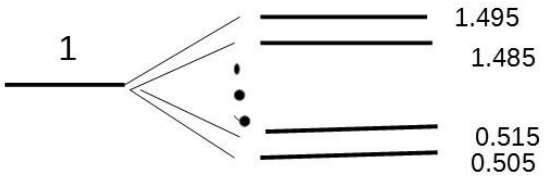
In his papers, Shannon took a more difficult example, a text in the English language. The logarithm of 26 to the base 2 is around 4.7, but of course we have to include puncutation, spaces, case etc as additional characters, so we are talking of more than this 4.7 bits per character. However, the letters neither occur with equal probability, nor are uncorrelated. It is possible to use known texts to estimate the actual entropy which turns out to be about one fourth of this number of bits. This is of course what is exploited by people who use SMS English. A nice illustration is provided by texts in which the first and last letters of words are preserved, and the remainder are permuted, they are still quite surprisingly readable! Pictures are even more compressible- large areas of blank sky can be conveyed quite concisely, and when we talk of a movie, successive frames are highly correlated. All this reduces the entropy below the naive value that we would get from the size of the alphabet and the length of the message, *N log s* . (For 24 frames per second, each of one megapixel with say four bits of colour and intensity information, the naive rate is more than 100 Mbps, so 3.6 Gb for a one hour movie. Real movie files are much smaller). The point of the Shannon coding theorem is that it defines the limits to this process of compression.

All this technology was introduced well after 1948. It is worth keeping in mind that what we call the “Shannon limit” for compression of a message or picture depends very much on the underlying probability distribution that we assume for the source. The proof of the coding theorem is 'existential', - it says that given an ϵ , however small, we can find a coding scheme that for large enough N the

error rate is less than ϵ It bypasses specific algorithms for compression The creation and improvement of compression / coding algorithms especially for speech and pictures continue to employ engineers / computer scientists to the present day. It is a sobering thought that all our profound conversations and the even more profound creations of film directors are well approximated as random processes for the purposes of compression.

The second, and possible more far reaching use of the entropy concept is in setting the limits to transmission of messages in the presence of noise. Before presenting the general picture, we take a simple, though artificial example. Imagine that the information source uses an alphabet of 10 letters, occuring with equal probability, which are encoded as voltages of electrical signals, 1 V for 1, ….10 for 10. These are sent down a wire (the general name is channel) which adds noise to the signal Let the noise have a uniform probability distribution between -0.495 V and +0.495, in steps of 0.01 V. (Since we havent talked about continous distributions, let us imagine for convenience that all voltages are multiples of 0.005 V) In this case, as the diagram below shows, we are able to reconstruct what was sent, even in the presence of noise, with full accuracy. If the received signal is V volts, we simply find the integer closest to V, which is the transmitted voltage. The received signal can vary from 0.505 to 10.495 volts, in steps of 0.01 V so it can take 1000 values – the entropy per character is $\log(1000) \simeq 9.97$ However, the entropy of the input is only $\log(10) \simeq 3.32$ What has happened is that the noise – which has 100 values in the range of -0.495 to +0.495 V, has added entropy of $\log(100) \simeq 6.65$ In this example, it is clear that this is the marginal or critical situation.





If our noise were less, we could either use lower signal voltages or send a bigger alphabet. If the noise were greater, we would have a situation where the same received signal would correspond to more than one possible transmitted signal, and our goal of beng error free is not fultfiled.

So in this example we have have the criterion for maximum noise free transmission, that the entropy of noise should not be greater than entropy of received signal minus the entropy of the transmitted signal. In symbols, $H_T = H_R - H_N$ is the maximum rate of error free transmission, the so called channel capacity. This capacity would be log(10 000)-log(10) =log(10000/10)=log(received signal / noise)., per character. This form of the result – log of a ratio- tells us that it is independent of the subdivision into units of 0.01 V which we made for convenience.

As an intuitive principle this is not surprising. It is telling you that you can transmit more information by either increasing the signal or decreasing the noise. But the achievement is to turn it into a theorem, much more general that the baby example we have given. To appreciate this, just imagine that the distribution of noise was gaussian.. Then the received signals would overlap for any two transmitted signals – we could never be sure at the single character level about what was transmitted. The secret of Shannons success is not to stick to single characters but think of whether the different possible messages as a whole, suitably encoded, can be distinguished with a probability approaching within ϵ of unity, for long enough messages. The not at all obvious answer is yes. The strategy of seeing that this is possible, and proving it, is deferred for the moment.

We do have to consider another factor, viz how many characters can we send per second on a wireless channel. Intuitively, this would depend on how rapidly the signal could change, which in turn depends on the bandwidth, denoted by *B* and measured in Hertz or cycles per second. The precise result is that a signal occupying a bandwidth *B* with a time duration of T has 2BT independent degrees of freedom – independent real numbers characterising it. In fact, although time is continuous, the signal can be reconstructed fro samples separated by 1/2B. This famous result – the 'sampling theorem' predates Shannon – it is called the Nyquist formula and -no surprise – Nyquist was a staff membe at Bell Telephone Laboratories in the 1920's. K.S.Krishnan had some interesting things to say about in in Narure in the 1940's

The famous formula for channel capacity reads $C = B \log(1 + S/N)$ bits per second. Hence the importance of the two things that mobile operators fight each other for – bandwidth and signal power (i.e more towers). It is clear that the dependence on the signal strength is only logarithmic, while the dependence on the bandwidth is linear. The steady progress form 2G to 3G to 4G is measured in terms of bandwidth – of course superb electronics and software plays its role as well. The fact that this leaves less and less bandwidth for radio astronomy is an unfortunate corollary. Whatever you enjoy on the mobile and internet owes everything to the coding and capacity formulae, both of which work with entropy.

Note that Shannon chose a model for the communicaton process which was amenable to reasoning based on probabiity theory. At the root lies the ensemble of possible messages, of which we receive one. This has to be given a priori, and so sidesteps two tricky questions – what the transmitter means and what the receiver understands. In Shannon's memorable words – messages may sometimes have meaning. Most people of an older generation like me would feel that the quantity of meaning is not proportional to the amount of kilo/mega/giga bits received, but actually gets buried in these megabits. The signal has become noise. This is related to notions of simplicity and complexity of ones view of the world. Interestingly, one of the later offshoots of information theory has something to say even on this question of meaning!