

Sophomore College

Mathematics of the Information Age

Crossing the Channel

Shannon's greatest contribution was to apply his measure of information to the question of communication in the presence of noise. In fact, in Shannon's theory a *communication channel* can be abstracted to a system of inputs and outputs, where the outputs depend *probabilistically* on the inputs. Noise is present in any real system and its effects are felt in what happens, probabilistically, to the inputs when they become outputs.

In the early days (the early 1940's) it was naturally assumed that increasing the transmission rate (now thought of in terms of bits per second, though they didn't use bits then) across a given communications channel increased the probability of errors in transmission. Shannon proved that this was *not* true *provided* that the transmission rate was below what he termed the *channel capacity*.

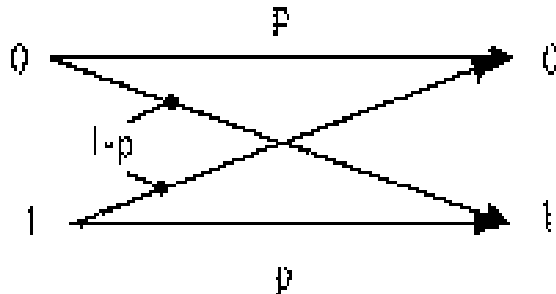
One way of stating Shannon's theorem is:

Shannon's Channel Coding Theorem A channel with capacity C is capable, with suitable coding, of transmitting at any rate less than C bits per symbol with vanishingly small probability of error. For rates greater than C the probability of error cannot be made arbitrarily small.

There's a lot to define here. Patience.

Channel capacity and Entropy: The symmetric binary channel

We send 0's and 1's, and only 0's and 1's. In a perfect world there is never an error in transmission. In an imperfect world, errors might occur in the process of sending our message; a 1 becomes a zero or vice versa. Let's say that, with probability p a transmitted 1 is received as a 1. Then, of course, with probability $1 - p$ a transmitted 1 is received as a 0. Symmetrically, let's say that the same thing applies to transmitting a 0; with probability p a transmitted 0 is received as a 0, and with probability $1 - p$ a transmitted 0 is received as a 1. This describes the channel completely. It's called a *binary symmetric channel*, and it's represented schematically by the following drawing.



We've essentially worked with this before. It's winning or losing at roulette – but this time with the two statements:

- The symbol we received is correct
- The symbol we received is in error

The definition of the binary symmetric channel says that the first message occurs with probability p and the second occurs with probability $1-p$. The entropy of this source is the weighted average of the information associated with each of the two messages – a measure of our average uncertainty in getting one message or the other, or our average uncertainty about the outcome of a transmission of symbols over a binary symmetric channel. The formula for the entropy is, as before,

$$H = p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p}.$$

Remember what this function looks like as a function of p . $H(p)$ is ≤ 1 and equal to 1 when $p = 1/2$. When $p = 1/2$ either outcome is equally likely and the channel is completely unreliable. When $p = 0$ or $p = 1$ the channel is perfect. Otherwise, the entropy is strictly less than 1, and there is uncertainty – possible errors in transmission.

It's as if, in a perfect world, the 0 and 1 were each worth a full bit before we sent them, but the noise in the channel, measured by the probability of error in transmission, takes away H bits, leaving us with only $1-H$ bits worth of information in transmitting each 0 or 1. That's a limit on the reliability of the channel to transmit information. If we were in a defining mood we might define the capacity for a binary symmetric channel to be

$$C = 1 - \left(p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p} \right) = 1 - H$$

In fact, we'll define the channel capacity in general as a maximum of 'mutual information' between a set of messages to be selected at one end and a set of messages that are received at the other. The 'noise' will be described by conditional probabilities relating the selected and received messages. With this general definition, *the* classic example of the calculation of channel capacity is for the binary symmetric channel, and the formula does come out just as above, $C = 1 - H$.

The Channel: Mutual Information

Here's the set-up. Suppose we have messages s_1, s_2, \dots, s_N to select send at one end of a communication channel, and messages r_1, r_2, \dots, r_N that are received at the other end. At the one end, the s_j 's are selected according to some probability distribution, say the probability that s_j is selected is $P(s_j)$. Then the relationship between the sending and receiving ends, between selecting and receiving, is described by the *conditional* probabilities

$$P(s_j|r_k) = \text{Probability that } s_j \text{ was sent given that } r_k \text{ was received}$$

As a 'conditional probability' this is read: 'the probability of s_j given r_k '. Think of it as answering the question: Given the output (r_k), how probable is it that this output resulted from a given input (s_j)? Answer: $P(s_j|r_k)$. One might say that the channel *is* the assignment of conditional probabilities to the selected and received messages.

Aside on conditional probabilities A formula for the conditional probability that event A occurs given that B has occurred is

$$P(A|B) = P(A \text{ and } B)/P(B).$$

One way to think of this is as follows. Imagine doing an experiment N times (a large number) where both A and B can be observed outcomes. Then the probability that A occurs given that B has occurred is approximately the number of times that A occurs in the runs of the experiment when B has occurred, *i.e.*,

$$P(A|B) \approx \frac{\text{Number of occurrences of } A \text{ and } B}{\text{Number of occurrences of } B}.$$

Now write this as

$$\begin{aligned} P(A|B) &\approx \frac{\text{Number of occurrences of } A \text{ and } B}{\text{Number of occurrences of } B} \\ &= \frac{\frac{\text{Number of occurrences of } A \text{ and } B}{N}}{\frac{\text{Number of occurrences of } B}{N}} \approx \frac{P(A \text{ and } B)}{P(B)}. \end{aligned}$$

If A and B are independent events then $P(A \text{ and } B) = P(A)P(B)$ and so

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} = P(A),$$

which makes sense – if A and B are independent, and you're A , who cares whether B happened.

You might find it helpful to think of this as an $N \times N$ matrix. That is, write

$$P_{jk} = P(s_j|r_k).$$

and

$$\mathbf{P} = \begin{pmatrix} P_{11} & P_{12} & \dots & P_{1N} \\ P_{21} & P_{22} & \dots & P_{2N} \\ \dots & \dots & \dots & \dots \\ P_{N1} & P_{N2} & \dots & P_{NN} \end{pmatrix}$$

The k 'th column,

$$\begin{pmatrix} P(s_1|r_k) \\ P(s_2|r_k) \\ P(s_3|r_k) \\ \vdots \\ P(s_N|r_k) \end{pmatrix} = \begin{pmatrix} P_{1k} \\ P_{2k} \\ P_{3k} \\ \vdots \\ P_{Nk} \end{pmatrix}$$

is all about what might have led to the received message r_k : It could have come from s_1 with probability P_{k1} ; it could have come from s_2 with probability P_{k2} ; it could have come from s_3 with probability P_{k3} , and so on. Message r_k had to have come from somewhere, so observe that

$$\sum_{j=1}^N P_{kj} = 1.$$

If s_j and r_k are independent, then

$$P(s_j|r_k) = P(s_j)$$

and the channel matrix looks like

$$\mathbf{P} = \begin{pmatrix} P(s_1) & P(s_1) & \dots & P(s_1) \\ P(s_2) & P(s_2) & \dots & P(s_2) \\ \dots & \dots & \dots & \dots \\ P(s_N) & P(s_N) & \dots & P(s_N) \end{pmatrix}$$

What's the channel matrix for a perfect channel, where there is no noise, no uncertainty? That must be that r_j comes from s_j and no other message, so that

$$P_{jk} = \begin{cases} 1, & j = k \\ 0, & j \neq k. \end{cases}$$

Thus \mathbf{P} is the $N \times N$ identity matrix.

What's the channel matrix for the binary symmetric channel? Here $\{s_1, s_2\} = \{0, 1\}$ and $\{r_1, r_2\}$ likewise is $\{0, 1\}$, and

$$\mathbf{P} = \begin{pmatrix} p & 1-p \\ 1-p & p \end{pmatrix}$$

Now, by definition (of the channel), the probability that a message s_j was sent given that r_k was received is $P(s_j|r_k)$. On the other hand, the message s_j had an *a priori* probability of being selected; call this $P(s_j)$. On receiving r_k there is thus a 'change in status' of the probability associated with s_j , from an *a priori* $P(s_j)$ to an *a posteriori* (after reception of r_k) $P(s_j|r_k)$. By the same token, according to how we think of information (quantitatively) we can say that there has been a 'change of status' of the information associated with s_j on receiving r_k , from an *a priori* $\log(1/P(s_j))$ to an *a posteriori* $\log(1/P(s_j|r_k))$. The former, $\log(1/P(s_j))$, is what we've always called the information associated with s_j , the latter, $\log(1/P(s_j|r_k))$, is something we haven't used or singles out, but it's along the same lines. What's really important here is the *change* in the information associated with s_j , and so we set

$$I(s_j, r_k) = \log \frac{1}{P(s_j)} - \log \frac{1}{P(s_j|r_k)} = \log \frac{P(s_j|r_k)}{P(s_j)}.$$

and call this the *mutual information* of s_j given r_k . *It is a measure of the (amount of) information the channel transmits about s_j if r_k is received.*

If s_j and r_k are independent then $P(s_j|r_k) = P(s_j)$. Intuitively, there has been no change of status in the probability associated with s_j . To say it differently there has been ‘nothing more learned’ about s_j as a result of receiving r_k , or there has been no change in the information associated with s_j as a result of receiving r_k . In terms of mutual information, this is exactly the statement that

$$I(s_j, r_k) = 0,$$

the *mutual information* is zero. The mutual information is only positive when $P(s_j|r_k) > P(s_j)$, *i.e.*, when we are ‘more certain’ of s_j by virtue of having received r_k .

Just as entropy of a source proved to be a more widely useful concept than the information of a particular message, here too it is useful to define a *system mutual information* by finding an average of the mutual informations of the individual messages. This depends on all of the source messages, S , and all of the received messages R . It can be expressed in terms of the entropy $H(S)$ of the source S , the entropy, $H(R)$ of the received messages R , both of which use our usual definition of entropy, and a ‘conditional entropy’ $H(S|R)$ which is something new. For completeness, the joint entropy is defined by

$$H(S|R) = \sum_{j,k} P(s_j \text{ and } r_k) \log \frac{1}{P(s_j|r_k)}$$

This isn’t so hard to motivate, but enough is enough – it shouldn’t be hard to believe there’s *something* like a conditional entropy – and let’s just use it to define the system mutual information, which is what we really want. This is

$$I(S, R) = H(S) - H(S|R).$$

This does look very much like an averaged form of the mutual information, above. And if the mutual information $I(s_j, r_k)$ is supposed to measure the amount of information the channel transmits about s_j given that r_k was received, then $I(S, R)$ should be a measure of the amount of information (on average) of all the source messages S that the channel transmits given all the received messages R . It’s a measure of the change of status of the entropy of the source due to the conditional probabilities between the source messages and received messages.

So now, finally, *what is the channel capacity?* It’s the most information a channel can ever transmit. This Shannon defines to be

$$C = \max I(S, R)$$

where the maximum is taken over the possible probabilities for the source messages S . (That is, the maximum is taken over all possible ways (in terms of probabilities) of selecting the source messages.) It is in terms of this that Shannon proved his famous channel coding theorem. We state it again:

Shannon’s Channel Coding Theorem A channel with capacity C is capable, with suitable coding, of transmitting at any rate less than C bits per symbol with vanishingly small probability of error. For rates greater than C the probability of error cannot be made arbitrarily small.

Whew!

I must report that we *do not* have the tools available for a proof. In fact, we cannot even show that $C = 1 - H$ for the binary symmetric channel. Sorry.