A Glance at Discrete Memoryless Channels

Miranda Christ and Rohan Jha

February 3, 2020

This work is based on Claude Shannon's *A Mathematical Theory of Communication* [1]. We also based much of the second section on Cover and Thomas' *Elements of Information Theory* [2].

1 Introduction to entropy

Consider a scenario in which Alice wishes to send a message to Bob. Available to her is a *channel*, through which she can transmit her message to Bob in the form of bits. Bob then decodes those bits to obtain Alice's original message.

This channel can be limited in terms of both accuracy and speed; the channel can accommodate a certain number of bits per second, and some subset of these bits may become corrupted. The question Shannon wishes to answer given these constraints is:

How much information can Alice send to Bob?

In Shannon's model, we think of individual messages as elements of a specified message set M. Notice that if this message set contains only two messages; for example, $M = \{hello, goodbye\}$, Alice can encode *hello* as 0 and *goodbye* as 1 and send only a single bit, despite the messages themselves appearing much longer. Notice also that if there were three possible messages in our message set, Alice would need to use more than one bit.

Shannon proposes a metric for the information contained in such a message set. As demonstrated in the above example, this metric depends not on the messages themselves but rather on the size of the message set and the distribution of the messages, as we will see.

Shannon adds another layer to the message set: the messages appear according to some known probability distribution. As an analogy, think of words in the English language; *the* appears much more often than *defenestrate*. It then makes sense to assign more frequent messages shorter encodings, since Alice will need to send them more often. Observe that this distribution also affects how laborious it is for Alice to communicate her messages. For example, consider the message space $M = \{hello, goodbye, ttyl\}$. If *hello* and *goodbye* appear with 49% probability and *ttyl* appears with 2% probability, Alice can encode *hello* as 0, *goodbye* as 1, and *ttyl* as 01. She then expects to transmit roughly 1 bit per message on average. In contrast,

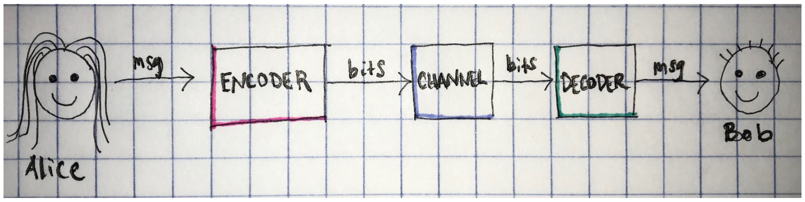

Figure 1: Alice and Bob's communication via a channel

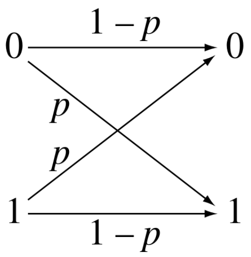

Figure 2: Binary symmetric channel

if *hello*, goodbye, and *ttyl* occur with equal probability, Alice expects to transmit well over 1 bit per message on average.

Therefore, Shannon's metric for information should depend on the distribution of the message space in addition to its size. Generally, more messages, each with significant likelihood, increases the amount of information as we have seen in the above examples. This leads us to Shannon's definition of entropy. We denote the entropy of a random variable X over a sample space of size n as $H(X)$, defined:

$$H(X) = \sum_{i=1}^{n} p_i \log p_i$$

Where p_i is the probability of the ith value. H is in fact the only such function satisfying three desirable properties:

1. H is continuous in the p_i.

2. For uniform distributions with probabilities $p_1, ..., p_n$ where all the p_i's are equal, H is monotonically increasing with n. In other words, increasing the size of the message space generally increases entropy.

3. If a random variable X is expressed in terms of a series of choices from other random variables, H should be the weighted sum of the entropies of these other variables. See Shannon's paper for the precise statement, which is a bit long. This requirement essentially says that H is consistent under equivalent expressions of X, and H is easily calculable given the entropies of the random variables in terms of which X is represented.

2 Discrete channel with noise

Again, we can think of a channel as a system for communicating information. In the real world, this might be a radio wave or a telephone signal, where the source can become distorted. In this section we consider the case of a discrete (the inputs and output are discrete) channel with noise (there's some probabilistic mapping from inputs to outputs). We'll introduce a concrete example. Consider a channel in which the information source is trying to transmit a sequence of zeros and ones and some of the bits are flipped. By this, we mean that some zeros are received as ones and some ones are received as zeros. This can be visualized with Figure 2, which is an example of a binary symmetric channel (BSC) where bits are corrupted with probability p.

Shannon is interested in the question of how much information can be transmitted error-free over a noisy channel. He introduces the definition of capacity and shows that it provides an upper bound on the rate of transmission over the channel. I here is the mutual information, which is loosely a measure of the extent of correlation between two random variables. This corresponds with our intuition that a channel in which the output is not very correlated with the input should have a low capacity.

Definition (Capacity)**.** Let X be an information source and Y the recipient. Then

$$C = max(I(X;Y)) = H(Y) - H(Y|X), \tag{1}$$

where the maximum is taken over probability distributions for X.

Theorem (Capacity of BSC)**.** *Assume a BSC as in Figure 2. Its capacity is* $1 - H(p)$*, where* $H(p) = -(p \log p + (1-p)\log(1-p))$*.*

Proof.

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(Y) - P(X=0)H(Y|X=0) - P(X=1)H(Y|X=1) \\ &= H(Y) - H(p) \\ &\leq 1 - H(p) \text{ with equality when } P(X=0) = P(X=1) = \frac{1}{2} \end{aligned}$$

□

We compute the capacities of specific BSC's and verify that the definition corresponds with our intuition. If $p = 0$ (meaning there isn't noise), the capacity is 1. If $p = 0.5$ (meaning Y is independent of X), the capacity is 0. And if $p = 0.1$, the capacity is somewhere in the middle (0.531). Noisier channels have lower capacity, as we would expect. The central claim, now, of the channel coding theorem is that if the entropy of the source (the rate) is below this capacity, then error-free communication is possible in the limit (as we encode larger and larger blocks of messages). More precisely:

Theorem (Channel Coding)**.** *Let a discrete channel have the capacity* C *and a discrete source the entropy per second* H*. If* $H \leq C$ *there exists a coding system such that the output of the source can be transmitted over the channel with an arbitrarily small frequency of errors.*

We'll give a rough sketch of the proof, with the emphasis on intuition. In the proof, we consider encoding and sending sequences of messages rather than individual messages. To motivate the use of sequence, we consider the following example. If we have 3 possible messages m_1, m_2, m_3 each occurring with equal probability, we require an average of $\frac{4}{3}$ bits to encode them individually, since without loss of generality we must use 1 bit for m_1, 1 bit for m_2, and 1 bit for m_3. However, if we instead encode them as pairs (e.g., $(m_1, m_1) \rightarrow 0, (m_1, m_2) \rightarrow 1, ..., (m_3, m_3) \rightarrow 111$), we require on average $\frac{2 \cdot 1 + 4 \cdot 2 + 3 \cdot 3}{9} = \frac{19}{9}$ bits per pair and roughly 2.05 bits per symbol. By block coding sequences of messages in this fashion, we decrease the expected encoded length.

Another consequence of working with sequences of messages rather than individual messages is that many of the possible sequences become increasingly unlikely as the sequence length increases, by the law of large numbers. We'll start by discussing the remaining set of reasonably likely sequences, called the jointly typical set. If we assume a sequence of i.i.d. random variables, we can consider its 'empirical entropy' to be $-\frac{1}{n} \log p(x^n)$. We note the following:

$$\begin{aligned} -\frac{1}{n} \log p(x^n) &= -\frac{1}{n} \sum \log p(x_i) \\ &\rightarrow -\mathbb{E}(\log(p(x)) \text{ as n gets large by the LLN} \\ &= H(X) \end{aligned}$$

Informally, the typical set is the set of sequences with empirical entropy close to the true entropy, and by the LLN, the probability a random sequence is in the typical set approaches 1. The *jointly* typical set A_ϵ^n, is defined (informally) for a sequence of length n of i.i.d pairs (x^n, y^n) (but X and Y need not be independent) as the set of sequences for which the empirical entropy of x^n is close to $H(X)$, that of y^n is close to $H(Y)$, and that of (x^n, y^n) is close to $H(X,Y)$. The probability, again, that a randomly chosen pair is in the jointly typical set approaches 1. Further, we state without proof the following from [2]:

$$P((X,Y) \in A_\epsilon^n) \approx 2^{-nI(X;Y)} \text{ where X and Y are independent} \tag{2}$$

$$|A_\epsilon^n| \leq 2^{n(H(X,Y)+\epsilon)} \tag{3}$$

(2) means that *independent* pairs of X and Y are very unlikely to be jointly typical (while *observed* pairs of X and Y are almost certain to be jointly typical). To make sense of this, let's return to the BSC. We can consider repeated uses of our BSC to be pairs (x^n, y^n), where x^n is the input and y^n is the output. For less noisy channels, then, X^n and Y^n will be more correlated, mutual information will be higher (as we discussed earlier), and there's a lower probability that independently chosen pairs will be jointly typical. This makes sense; if we consider a channel that is not very noisy, then certain pairs (x^n, y^n) will be relatively more likely, the typical set will be smaller (as stated by (3)), and a given y^n will be jointly typical with fewer x^n.

Above we've established that a relatively small subset of input-output pairs x^n and y^n occur together with reasonable probability (with larger subsets for noisier channels). In other words, given an output from the channel y^n, there are few inputs that likely induced that output. And the number of likely inputs depends on the mutual information; if the source has lower mutual information, there are more likely input sequences. We can then in theory construct a coding scheme where we decode an output by choosing the input with which it is jointly typical, requiring longer transmission sequences for noisier channels.

We'll make the above intuition concrete and assume that we construct a system in which we're able to send a message from a set of size 2^{nR} with n uses of the channel. R is the *rate* of the channel, and it describes how the number of possible messages scales with respect to the number of uses of the channel. Note that in the noiseless setting, we'll be able to communicate at rate 1.

Now, we'll briefly sketch the proof, now, of the following theorem, which is closely related to Shannon's version of the theorem above.

Theorem. *If $R < C$ then error-free communication is possible, as n (the number of uses of the channel) becomes large.*

First, a discussion of how it's related to Shannon's paper. Assume we can achieve error-free communication for 2^{nR} distinct messages. Now if we have an information source with entropy $H \leq R < C$, there will only be $2^{nH} < 2^{nR}$ typical sequences, so error-free communication will be possible for these typical sequences and in general because the probability of an observed sequence being typical converges to 1.

Now, we'll describe a communication scheme and provide the outline of the argument that it's asymptotically error-free if $R < C$ (as the number of uses becomes large).

We consider the following (naive) scheme. A codebook is constructed that maps each of the 2^{nR} messages to some random binary code. After an output is received, we decode it by choosing the input with which it is jointly typical. Possible sources of error then include:

1. A given input induces an output with which it is not jointly typical.

2. There are multiple inputs that are jointly typical with a given output.

Because the probability that x^n and y^n are jointly typical converges to 1, the probability that (1) happens goes to 0.

We can ensure that (2) happens rarely if we choose long enough sequences to describe few enough messages. In other words, we want the sets of typical y^n and $\hat{y}^n$ corresponding to any two inputs x^n and $\hat{x}^n$ to be disjoint. The size of the set of typical x^n converges to $2^{nH(X)}$, and the size of the set of typical y^n converges to $2^{nH(Y)}$. The size of the set of jointly typical x^n with a given y^n is $\frac{|A_\epsilon^n|}{\#\text{ typical } y^n}$, which converges to $2^{n(H(X,Y)-H(Y))} = 2^{nH(X|Y)}$ by definition of the jointly typical set. Therefore, the number of sequences we can use as codewords before any of their typical outputs intersect is upper bounded by the number of typical x^n divided by the number of jointly typical x^n with a given y^n:

$$\frac{2^{nH(X)}}{2^n H(X|Y)} = 2^{n(H(X)-H(X|Y))} = 2^{nI(X;Y)} = 2^{nC}$$

This argument can be extended to show that as long as $R < C$, we can communicate our messages with arbitrarily low probability of error.

References

[1] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

[2] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006.