WITTEN'S DERIVATION OF THE SHANNON SOURCE CODING THEOREM

Aidan Rocke
aidanrocke@gmail.com

May 7, 2021

ABSTRACT

The Shannon source coding theorem shows that in the asymptotic limit it is impossible to compress i.i.d. data such that the average number of bits per symbol is less than the Shannon entropy of the source. This elegant derivation, due to Edward Witten, is entirely combinatorial in nature.

1 A special case

1.1 The case of two letters

Let's suppose we receive a message encoded with two letters:

$$abbaabbb... \tag{1}$$

and let's suppose that the zeros occur with probability p and the ones occur with probability $1-p$. How many bits of information can a person expect to receive from a long message i.e. large N?

For large N, the message will consist of approximately $p \cdot N$ zeros and $(1-p) \cdot N$ ones so we have:

$$\frac{N!}{(p \cdot N)! \cdot ((1-p) \cdot N)!} \sim \frac{N^N}{(p \cdot N)^{p \cdot N}((1-p) \cdot N)^{(1-p) \cdot N}} = 2^{N \cdot S} \tag{2}$$

where S is the Shannon entropy of each symbol:

$$S = -p \cdot \log p - (1-p) \cdot \log(1-p) \tag{3}$$

It follows that in the asymptotic limit, the total number of messages of length N given our knowledge of the relative probability of symbols one and zero is approximately:

$$2^{N \cdot S} \tag{4}$$

so the number of bits of information one might gain from observing such a message is roughly NS.

1.2 Maximum entropy

It is worth adding that the Shannon entropy $S(p) = -p \cdot \log p - (1-p) \cdot \log(1-p)$ is maximised when $\frac{\partial S}{\partial p} = 0$. So we find that in the case of an alphabet with two letters:

$$S \leq \log 2 \tag{5}$$

so the maximum entropy distribution is the one where each letter appears with equal probability so $p = (1-p)$.

2 The General case

2.1 The case of k letters

Let's suppose more generally that our message is taken from an alphabet of k letters $\{a_i\}^k$ with probability distribution $P = \{p_i\}_{i=1}^k$. For large N, the symbol a_i will appear approximately $N \cdot p_i$ times and the number of such messages is asymptotically:

$$\frac{N!}{\prod_{i=1}^k (p_i \cdot N)!} \sim \frac{N^N}{\prod_{i=1}^k (p_i \cdot N)^{p_i \cdot N}} = 2^{N \cdot S} \tag{6}$$

where the entropy per letter is:

$$S = -\sum_{i=1}^k p_i \cdot \log p_i \tag{7}$$

so the number of bits of information that we can gain from a message with N symbols is generally $N \cdot S$.

2.2 Maximum entropy

As demonstrated in the previous section, the maximum entropy distribution in the case of an alphabet with k letters is the one where each letter appears with equal probability:

$$S \leq \log k \tag{8}$$

References

[1] Edward Witten. A Mini-Introduction To Information Theory. 2018.

[2] Edwin Jaynes. Information Theory and Statistical Mechanics I. The Physical Review. 1957.

[3] Edwin Jaynes. Information Theory and Statistical Mechanics II. The Physical Review. 1957.