2.2. Shannon Source Coding Theorem

The previous section derived an expectation that calculates the "rate" of a symbol-code: $$E[\ell(s)] = \sum_{s \in \Sigma} \ell(s) \cdot p(X = s)$$ This raises a natural question of what is the best possible rate that we could hope to achieve for a given distribution?

In this section, we derive a lower bound on this rate called the "Shannon Source Coding Theorem". We will focus on the case for symbol-codes and on the case the sources are i.i.d. Neither is a real restriction in practice, but the proofs are far more technical. To calculate this lower bound, we're going to start with a bit of circular logic. Let's define a function $H(X)$ that does the following:

$$H(X) = -\sum_{s \in \Sigma} \log_2 p(X = s) \cdot p(X = s)$$

Gibb's inequality says that functionals like the one above (called entropic functionals) are always upper bounded:

$$H(X) \leq -\sum_{s \in \Sigma} \log_2 q(X = s) \cdot p(X = s)$$

when $q(X = s) \neq p(X = s)$, where q is an arbitrary probability distribution over symbols. Since it holds for an arbitrary distribution, it certainly holds for one particular distribution. Let's define a probability distribution over symbols as follows:

$$q(X = s) = \frac{a^{-\ell(s)}}{C} \quad C = \sum_{s \in \Sigma} a^{-\ell(s)} \quad a = |\Sigma|$$

It follows that:

$$H(X) \leq -\sum_{s \in \Sigma} \log_2 \frac{a^{-\ell(s)}}{C} \cdot p(X = s)$$

Doing some algebra, we can see that:

$$H(X) \leq -\sum_{s \in \Sigma} (\log_2 (a^{-\ell(s)}) - \log_2 C) \cdot p(X = s)$$

Simplifying further, we notice this is exactly the expression for "rate" that we derived above:

$$H(X) \leq E[\ell(s)] \log_2 a + \log_2 C$$

Notice that $C = \sum_{s \in \Sigma} a^{-\ell(s)}$. Since codes have to be at least of length 1 $C \leq a \cdot a^{-1} \leq 1$ implying that $\log_2 C$ is less than 0. This means that:

$$\frac{H(X)}{\log_2 |\Sigma|} \leq E[\ell(s)]$$

2.2.1. Uniquely Decodeable/Prefix-Free Codes

It turns out that the above analysis is subtly vaccuous but it is a good starting point to get a lower bound. You can convince yourself that $\frac{H(x)}{\log_2 |\Sigma|}$ will always be less than 1, which doesn't make a lot of sense.

Essentially, this comes down to an overly optimistic analysis of how we can allocate bits to symbols. In fact, certain types of symbol codes may not be "uniquely decodeable" when concatenated. Let's look at the example from the previous section.

Color	key
Red	0
Green	11
Blue	10
Black	1

Suppose, we observed the sequence of bits [1,0]. Does that decode to “Black, Red” or does that decode to just “Blue”? A class of uniquely decodeable codes are called “prefix” codes. A prefix code is a type of code system distinguished by its possession of the “prefix property”, which requires that there is no whole code word in the system that is a prefix (initial segment) of any other code word in the system.

For the example above, the prefix-free code would give us:

Color	key
Red	0
Green	110
Blue	111
Black	10

Even though the lengths are longer, working out the math we can still see that it will be better than a fixed-length code:

$$0.8 * 1 + 0.15 * 2 + 0.03 * 3 + 0.02 * 3 = 1.25 \leq 2$$

For a prefix-free (binary) code, we have an additional property. Kraft’s inequality suggests the following inequality holds:

$$\sum_{s \in \Sigma} 2^{-\ell(s)} \leq 1$$

We’re not going to prove the above statement but it can be easily shown by drawing out a binary tree. This statement is key because it allows us to get a more realistic analysis of a lower bound. This allows us to take the above analysis and ammend it as follows:

$$q(X = s) = \frac{a^{-\ell(s)}}{C} \quad C = \sum_{s \in \Sigma} a^{-\ell(s)} \quad a = 2$$

If we follow the same steps, we see that this results in:

$$H(X) \leq E[\ell(s)] \blacksquare$$

$H(X)$ is exactly the lower bound on the expected rate of a prefix-free code!

2.2.2. Intuition for Engineers

Stated in another way, we can squint out the log terms and see that the lower bound is $\mathcal{O}(H(x))$ (even if we don’t use prefix-free codes). What does this mean? If we have two distributions over the same set of symbols the one with a lower $H(x)$ value has a better (lower) “rate”. This means its more compressible. Let’s see some examples.

Let’s take the skewed example in the previous section. Red with probability (0.8), Green with probability (0.02), Blue with probability (0.03), and Black with probability (0.15). What is $H(x)$?

$$-(0.8 * log(0.8) + 0.15 * log(0.15) + 0.02 * log(0.02) + 0.03 * log(0.03)) = 0.932$$

Let's compare this to the case when they are uniformly distributed. What is $H(x)$?

$$-(0.25 * log(0.25) + 0.25 * log(0.25) + 0.25 * *log(0.25) + 0.25 * log(0.25)) = 2$$

$H(x) = \log |\sigma|$ when the distribution is uniform (which is exactly the expected rate for the fixed-length code). This means that the inequality above is tight for prefix free codes.

By Sanjay Krishnan