A Property of the Entropy

For any two probability distributions, $p_1, \ldots, p_q$ and $p'_1, \ldots, p'_q$:

$$\sum_{i=1}^{q} p_i \log_r \left(\frac{1}{p_i}\right) \;\le\; \sum_{i=1}^{q} p_i \log_r \left(\frac{1}{p'_i}\right)$$

Proof:

First, note that for all $x > 0$, $\ln x \le x-1$ (see Jones & Jones, p. 40). So $\log_r x \le (x-1)/\ln r$.

We can now show that the LHS-RHS above is:

$$\sum_{i=1}^{q} p_i \left[\log_r \left(\frac{1}{p_i}\right) - \log_r \left(\frac{1}{p'_i}\right)\right] \;=\; \sum_{i=1}^{q} p_i \log_r \left(\frac{p'_i}{p_i}\right)$$

$$\le\; \frac{1}{\ln r} \sum_{i=1}^{q} p_i \left(\frac{p'_i}{p_i} - 1\right) \;=\; \frac{1}{\ln r}\left(\sum_{i=1}^{q} p'_i - \sum_{i=1}^{q} p_i\right) \;=\; 0$$

Proving We Can't Compress to Less Than the Entropy

We can use this result to prove that any uniquely decodable r-ary code for $\mathcal{S}$ must have average length at least $H_r(\mathcal{S})$:

Proof:

Let the codeword lengths be $l_1, \ldots, l_q$, and define $K = \sum_{i=1}^{q} r^{-l_i}$ and $p'_i = r^{-l_i}/K$.

The p'_i can be seen as probabilities, so

$$H_r(\mathcal{S}) \;=\; \sum_{i=1}^{q} p_i \log_r \left(\frac{1}{p_i}\right) \;\le\; \sum_{i=1}^{q} p_i \log_r \left(\frac{1}{p'_i}\right)$$

$$=\; \sum_{i=1}^{q} p_i \log_r (r^{l_i} K) \;=\; \sum_{i=1}^{q} p_i (l_i + \log_r K)$$

Since the code is uniquely decodable, $K \le 1$ and hence $\log_r K \le 0$. We conclude that the the average code length, $\sum p_i l_i$, is at least as great as the entropy, $H_r(\mathcal{S})$.

Shannon-Fano Codes

Lengths of optimal codes are hard to figure out, but it's easy to find the lengths of the *almost* optimal Shannon-Fano codes.

We make an r-ary code for symbols with probabilities $p_1, \ldots, p_q$ using codewords of lengths

$$l_i \;=\; \lceil \log_r (1/p_i) \rceil$$

Here, $\lceil x \rceil$ is the smallest integer greater than or equal to x.

The McMillan inequality says such a code exists, since

$$\sum_{i=1}^{q} \frac{1}{r^{l_i}} \;\le\; \sum_{i=1}^{q} \frac{1}{r^{\log_r (1/p_i)}} \;=\; \sum_{i=1}^{q} p_i \;=\; 1$$

Example with $r = 2$:

p_i:	0.4	0.3	0.2	0.1
$\log_2(1/p_i)$:	1.32	1.74	2.32	3.32
$l_i = \lceil \log_2(1/p_i) \rceil$:	2	2	3	4
Codeword:	00	01	100	1100

Average Lengths of Shannon-Fano Codes

The average length of a Shannon-Fano code for source $\mathcal{S}$ with symbols probabilities $p_1, \ldots, p_q$ is

$$\begin{aligned}
\sum_{i=1}^{q} p_i l_i \;&=\; \sum_{i=1}^{q} p_i \lceil \log_r (1/p_i) \rceil \\
&\le\; \sum_{i=1}^{q} p_i \,(1 + \log_r (1/p_i)) \\
&=\; \sum_{i=1}^{q} p_i \;+\; \sum_{i=1}^{q} p_i \log_r (1/p_i)) \\
&=\; 1 + H_r(\mathcal{S})
\end{aligned}$$

Proof of Shannon's Noiseless Coding Theorem

Consider coding the n-th extension of a source $\mathcal{S}$, whose symbols have probabilities $p_1, \ldots, p_q$, using an r-ary Shannon-Fano code.

The Shannon-Fano code for blocks of n symbols will have average length, L_n, no greater than $1 + H_r(\mathcal{S}^n) = 1 + nH_r(\mathcal{S})$.

The average length per original source source symbol will therefore be no greater than

$$\frac{L_n}{n} = \frac{1 + nH_r(\mathcal{S})}{n} = H_r(\mathcal{S}) + \frac{1}{n}$$

By choosing n to be large enough, we can make this as close to the entropy, $H_r(\mathcal{S})$, as we wish.

An End and a Beginning

This is a mathematically satisfying result. From a practical point of view, though, we still have two problems:

- How can we compress data to nearly the entropy **in practice**?

  The number of possible blocks of size n is q^n — huge when n is large. And n may need to be large to get close to the entropy.

  One solution: A technique known as *arithmetic coding*.

- Where do the probabilities $p_1, \ldots, p_q$ come from? And are they really constant?

  This is the problem of *source modeling*.