Shannon's noiseless coding theorem

We are working with messages written in an alphabet of symbols $x_1, \ldots, x_n$ which occur with probabilities $p_1, \ldots, p_n$. We have defined the *entropy* E of this set of probabilities to be

$$E = -\sum_{i=1}^{n} p_i \log_2 p_i.$$

Theorem For any uniquely decipherable encoding of $x_1, \ldots, x_n$ as binary code words (e.g. strings of 0s and 1s) the average length of a word must be greater than E.

Our proof of this theorem will involve two lemmas.

Lemma 1 (Gibbs' inequality). Suppose $p_1, \ldots, p_n$ is a *probability distribution* (i.e. each $p_i \geq 0$ and $\sum_i p_i = 1$). Then for any other probability distribution $q_1, \ldots, q_n$ with the same number of elements,

$$-\sum_{i=1}^{n} p_i \log_2 p_i \leq -\sum_{i=1}^{n} p_i \log_2 q_i.$$

Proof: Since $\log_2 p_i = \frac{\ln p_i}{\ln 2}$ and $\ln 2 > 0$ it is enough to prove the inequality with $\log_2$ replaced by ln wherever it occurs.We use the following property of the natural logarithm:

$$\ln x \leq x - 1 \text{ for all } x > 0, \text{ and } \ln x = x - 1 \text{ only when } x = 1.$$

In order to avoid zero denominators in the following calculation, we set $I = \{i | p_i > 0\}$, the set of indices for which p_i is non-zero. Then we write

$$-\sum_{i \in I} p_i \ln \frac{q_i}{p_i} \geq -\sum_{i \in I} p_i (\frac{q_i}{p_i} - 1) = -\sum_{i \in I} q_i + \sum_{i \in I} p_i = -\sum_{i \in I} q_i + 1 \geq 0.$$

Since $\ln \frac{q_i}{p_i} = \ln q_i - \ln p_i$, this last inequality becomes

$$-\sum_{i \in I} p_i \ln q_i \geq -\sum_{i \in I} p_i \ln p_i.$$

Now $-\sum_{i \in I} p_i \ln p_i = -\sum_{i=1}^{n} p_i \ln p_i$ since the new terms all have $p_i = 0$; and $-\sum_{i \in I} p_i \ln q_i \leq -\sum_{i=1}^{n} p_i \ln q_i$ since new terms are ≤ 0. I.e.

$$-\sum_{i=1}^{n} p_i \ln q_i \geq -\sum_{i \in I} p_i \ln q_i \geq -\sum_{i \in I} p_i \ln p_i = -\sum_{i \in I} p_i \ln p_i$$

yielding Gibbs' inequality.