Code Length and Source Coding Theorem

September 11, 2013

• Source Coding

The conversion of the output of a Discrete Memoryless Source (DMS) into a sequence of binary symbols is called source coding.

The aim of source coding is to minimize the average bit rate required for representation of the source by reducing redundancy of the information source.

• Average Code Length

For a DMS with a set of alphabet \( \{x_1, x_2, ..., x_m\} \) with corresponding probability of occurrence \( \{p_1, p_2, ..., p_m\} \) and code length \( \{l_1, l_2, ..., l_m\} \),

The average code length \( L \) per source symbol is thus

\[
L = \sum_{i=1}^{m} p_i l_i
\]

• Source Coding Theorem

For a DMS with finite entropy \( H \), the average code length \( L \) per source symbol has a lower bound

\[
L \geq H
\]

i.e.

\[
L_{\text{min}} = H
\]

• Code Efficiency \( \eta \) and Code Redundancy \( \gamma \)

\[
\eta = \frac{L_{\text{min}}}{L} = \frac{H}{L}
\]

\[
\gamma = 1 - \eta = 1 - \frac{H}{L}
\]
• Entropy Bound

For $m$ symbol $x_i$ with occurrence probability $p_i$

The entropy is thus

$$ H = \sum p_i I_i = \sum p_i \log_2 \frac{1}{p_i} $$

This entropy $H$ actually has an upper bound and lower bound

$$ 0 \leq H \leq \log_2 m $$

Proof of the left hand side inequality $0 \leq H$

$$ p_i \in [0, 1] \Rightarrow \frac{1}{p_i} \geq 1 \Rightarrow \log_2 \frac{1}{p_i} \geq 0 \Rightarrow p_i \log_2 \frac{1}{p_i} \geq 0 \Rightarrow \sum p_i \log_2 \frac{1}{p_i} \geq 0 \Rightarrow H \geq 0 $$

Proof of the right hand side inequality $H \leq \log_2 m$

Consider an inequality (can be proved simply by differentiation)

$$ \ln x \leq x - 1 $$

Consider two group of probability $\{p_i\}, \{q_i\}$ on $\{x_i\}$

By axiom of probability distribution, sum of probability should be 1

$$ \sum q_i = \sum p_i = 1 $$

$$ \sum p_i \log_\frac{q_i}{p_i} = \sum p_i \ln_\frac{q_i}{p_i} = \frac{1}{\ln 2} \sum p_i \ln_\frac{q_i}{p_i} $$

Using the inequality

$$ \sum p_i \log_\frac{q_i}{p_i} = \frac{1}{\ln 2} \sum p_i \ln_\frac{q_i}{p_i} \leq \frac{1}{\ln 2} \sum p_i \left( \frac{q_i}{p_i} - 1 \right) 
= \frac{1}{\ln 2} \sum q_i - p_i 
= \frac{1}{\ln 2} \left( \sum q_i - \sum p_i \right) $$

By axiom of probability distribution

$$ = 0 $$

Thus

$$ \sum p_i \log_\frac{q_i}{p_i} = \frac{1}{\ln 2} \sum p_i \ln_\frac{q_i}{p_i} 
\leq \frac{1}{\ln 2} \sum p_i \left( \frac{q_i}{p_i} - 1 \right) 
= 0 $$

2
Therefore

$$\sum p_i \log \frac{q_i}{p_i} \leq 0$$

Let

$$q_i = \frac{1}{m} \quad \text{(Equal probability distribution for m-message)}$$

$$\sum p_i \log \frac{q_i}{p_i} \leq 0$$

$$\iff \sum p_i \log \frac{1}{p_i m} \leq 0$$

$$\iff \sum p_i \log_2 \frac{1}{p_i} - \sum p_i \log_2 m \leq 0$$

$$\iff \sum p_i \log_2 \frac{1}{p_i H} - \log_2 m \sum p_i \leq 0$$

$$H \leq \log_2 m$$

Thus

$$0 \leq H \leq \log_2 m$$

• **Source Coding Theorem**

The entropy $H$ is the optimal lower bound of the average code length

$$L \geq H$$

i.e. When the coding is optimal (using shortest amount of code to represent information)

$$L_{\text{min}} = H$$

Consider the inequality

$$\sum p_i \log \frac{q_i}{p_i} \leq 0$$

And let $q_i = \frac{1}{2^l_i}$, notice that

$$\sum q_i = \sum_{i=1}^{m} \frac{1}{2^{l_i}} = \frac{1}{\sum_{i=1}^{m} \frac{1}{2^{l_i}}} = 1$$

And thus

$$\sum p_i \log \frac{q_i}{p_i} \leq 0$$

$$\iff \sum p_i \log \left( \frac{1}{p_i \sum_{i=1}^{m} \frac{1}{2^{l_i}}} \right) \leq 0$$
\[ \iff \sum p_i \left[ \log_2 \left( \frac{1}{p_i} \right) + \log_2 \frac{1}{2^{-l_i}} - \log_2 \sum_{i=1}^{m} \frac{1}{2^{l_i}} \right] \leq 0 \]

\[ \iff \sum p_i \log_2 \left( \frac{1}{p_i} \right) - \sum_{i} \log_2 \frac{m}{2^{l_i}} \leq 0 \]

\[ \iff H - L - \log_2 \sum_{i=1}^{m} \frac{1}{2^{l_i}} \leq 0 \]

Using Kraft Inequality

\[ \sum_{i=1}^{m} \frac{1}{2^{l_i}} \leq 1 \]

Thus

\[ \log_2 \sum_{i=1}^{m} \frac{1}{2^{l_i}} \leq \log_2 1 = 0 \]

Therefore

\[ H - L \leq \log_2 \sum_{i=1}^{m} \frac{1}{2^{l_i}} \]

\[ H - L \leq 0 \]

Thus

\[ L \geq H \]

Equality holds when

\[ L_{\text{min}} = H \]

\[ -\text{END}- \]