Costs and Optimality of Codes

Definition: i) For a code $C = \{c_1, c_2, \ldots, c_m\}$ and a probability distribution $P = \{p_1, p_2, \ldots, p_m\}$, $p_i \geq 0$ for $1 \leq i \leq m$, $\sum_{i=1}^{m} p_i = 1$, we define the costs of C under P by

$$\mathcal{L}(C, P) = \sum_{i=1}^{m} p_i |c_i|.$$

ii) For a probability distribution $P = \{p_1, p_2, \ldots, p_m\}$, $p_i \geq 0$ for $1 \leq i \leq m$, $\sum_{i=1}^{m} p_i = 1$, and an alphabet X, we set

$$\mathcal{L}_X(P) = \inf \mathcal{L}(C, P),$$

where the infimum is taken over all codes with m elements over X. A code C' over X is called optimal for P, if $\mathcal{L}(C', P) = \mathcal{L}_X(P)$ holds.

Optimal Codes I

Theorem: For each distribution P whose probabilities are all positive and each alphabet X, there is a (prefix) code over X which is optimal for P.

Theorem: For each distribution $P = \{p_0, p_1, \ldots, p_m\}$ and every alphabet X with $\#(X) = n$, we have

$$\sum_{i=1}^{m} p_i \log_n \left(\frac{1}{p_i}\right) \leq \mathcal{L}_X(P) \leq 1 + \sum_{i=1}^{m} p_i \log_n \left(\frac{1}{p_i}\right),$$

where the equality $\mathcal{L}_X(P) = \sum_{i=1}^{m} p_i \log_n \left(\frac{1}{p_i}\right)$ holds if and only if $\log_n(p_i)$ are integers for $1 \leq i \leq m$.

Optimal Codes II

Theorem: Let $C = \{c_1, c_2, \ldots, c_m\} \subseteq \{0,1\}^+$ be an optimal prefix code for a distribution $P = \{p_1, p_2, \ldots p_m\}$. Further, let

$$p_j = q_0 + q_1$$

and $p_1 \geq p_2 \geq \ldots \geq p_{j-1} \geq p_j \geq p_{j+1} \geq \ldots \geq p_m \geq q_0 \geq q_1$. Then

$$C' = \{c_1, c_2, \ldots, c_{j-1}, c_{j+1}, \ldots, c_m, c_j0, c_j1\}$$

is an optimal prefix code for the distribution

$$P' = \{p_1, p_2, \ldots, p_{j-1}, p_{j+1}, \ldots, p_m, q_0, q_1\}.$$