Week 6

1 Single random variable source coding

Let $\underline{c}(x)$ be the codeword assigned to $x \in \mathcal{X}$.

$l(x)$ be the length of codeword assigned to x.

Here we are coding a single random variable and all codewords are binary strings. (Fixed-variable length source coding)

$$\mathscr{C} = \{\underline{c}(x) : x \in \mathcal{X}\}$$

$$L_{\mathscr{C}} = \sum_{x \in \mathcal{X}} p(x) l(x)$$

Our goal is to design a code which has minimum $L_{\mathscr{C}}$, L^*

$$L^* = \min_{\mathscr{C}} L_{\mathscr{C}}$$

1.1 Kraft inequality

Let $\mathscr{C}$ be any prefix-free (binary) code. Then,

$$\sum_{x \in \mathcal{X}} 2^{-l(x)} \leq 1$$

Proof: We know that any prefix-free code can be represented using a binary tree which has 'leaves' as the codeword. In the binary tree corresponding to the code, the depth of the tree would be the length of the largest codeword in the code. ($l_{\max} = max_{x \in \mathcal{X}} l(x)$)

Suppose there is a codeword $\underline{c}$ of length l represented by a node at depth l($l \leq l_{\max}$).

Then $\underline{c}$ has $2^{l_{\max} - l}$ successors at level $l_{\max}$. Also, none of these are codewords as the code is prefix-free. Now,

$$\sum_{\underline{x} \in \mathcal{X}} 2^{l_{\max} - l(\underline{x})} \leq 2^{l_{\max}} \tag{1}$$

This is true if distinct codewords $\underline{c1}$, $\underline{c2}$ don't have common successors at $l_{\max}$ level. Any successor of $\underline{c1}$ at $l_{\max}$ has $\underline{c1}$ as a prefix. Similarly in the other case as well.

Suppose $l(\underline{c1}) \leq l(\underline{c2})$, so there can be a common successor $\underline{v}$ at $l_{\max}$.

Then, $\underline{v}$ has first $l(\underline{c1})$ places as $\underline{c1}$ and first $l(\underline{c2})$ places as $\underline{c2}$.

$\Rightarrow$ $\underline{c1}$ should be a prefix of $\underline{c2}$ which isn't true as $\mathscr{C}$ is a prefix-free code. Hence, no pair of codewords in $\mathscr{C}$ have any common successors.

Hence (1) is true. And dividing both sides by $2^{l_{\max}}$ gives us the Kraft inequality.

1.2 Lemma

$$L^* \geq H(X)$$

Any prefix-free code for X has average length of at least $H(X)$.

Proof:

$$L - H(X) = \sum_{x \in \mathcal{X}} p(x) l(x) - \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)}$$

We know,

$$D(p||q) = \sum_{x \in supp(P_X)} p(x) \log \frac{p(x)}{q(x)}$$

Let

$$q(x) := \frac{2^{-l(x)}}{\sum_{x \in \mathcal{X}} 2^{-l(x)}}$$

$$\begin{aligned}
D(p_X||q_X) &= \sum_{x \in supp(P_X)} p(x) \log \frac{p(x)}{\frac{2^{-l(x)}}{\sum_{x' \in \mathcal{X}} 2^{-l(x')}}} \\
&= -\sum p(x) \log \frac{1}{p(x)} + \sum p(x) \log \frac{1}{\frac{2^{-l(x)}}{\sum_{x' \in \mathcal{X}} 2^{-l(x')}}} \\
&= -H(X) + \sum_{x \in supp(p_x)} p(x) \log 2^{l(x)} + \sum_{x \in supp(p_x)} p(x) \log \sum_{x'} 2^{-l(x')} \\
&= -H(X) + L - \epsilon \qquad \text{(from Kraft's)} \\
&\leq -H(X) + L
\end{aligned}$$

$$\Rightarrow L_{\mathscr{C}} - H(X) \geq 0$$
$$L^* \geq H(X)$$

Note: Equality happens only iff $p_x = q_x$ and ϵ is zero.