Class no 14

Huffman code – Proof of optimality:

Claim of optimality:

Let $L_{Huffman}$ be the length of a code constructed for $X \sim (p_1, \dots, p_k)$ wlog $p_1 \geq p_2 \geq \dots \geq p_k$ according to the Huffman algorithm. Then $\bar{L}_{Huffman} = \min_{C \in \text{Prefix free codes for } X} \bar{L}_C$.

Proof:

We will denote the avg length of some code C for some random variable taking M values with probabilities $q_1, \dots, q_M$ as $L_{C_1}(q_1, \dots, q_M)$.

The optimal lengths for the same distribution is written as $L^*(q_1, \dots, q_M)$ → [note that this number does not depend on which optimal code we are choosing]

We will show that

$$L^*(p_1, \dots, p_k) = L^*(p_1, \dots, p_{k-1}+p_k) + p_{k-1} + p_k$$

(p_{k-1}, p_k: 2 least probability symbols are combined)

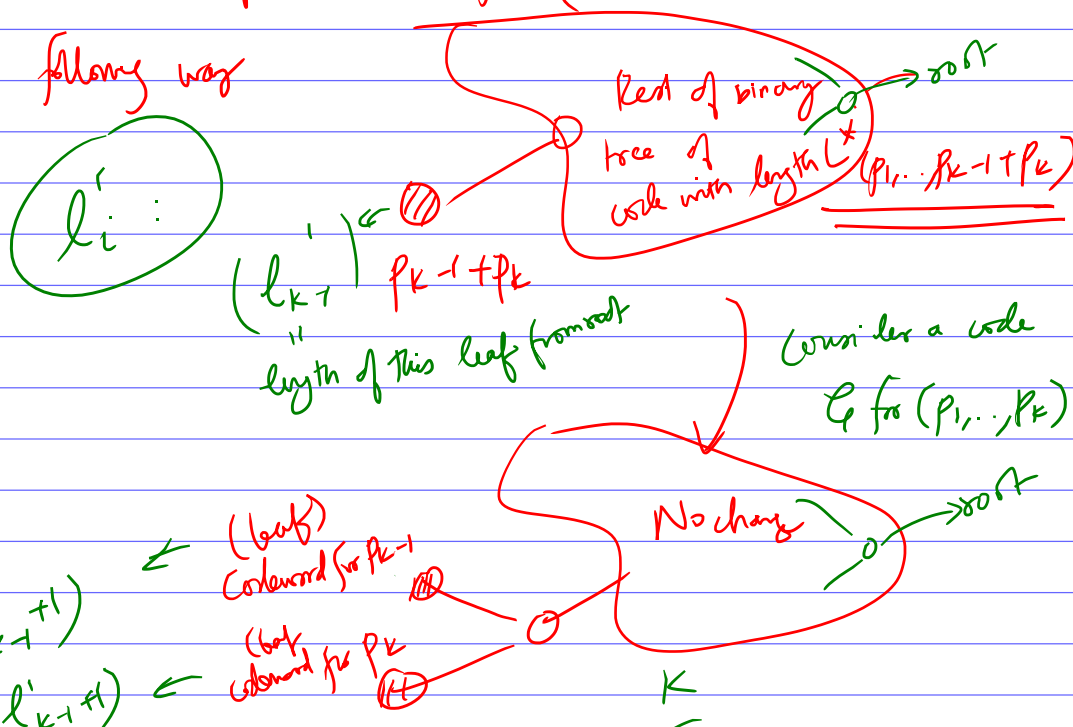
→ (A)

& then this will imply optimality of Huffman.

(To be proved later)

How?

An optimal code for $(p_1, \ldots, p_k)$ can be obtained from an optimal code for $(p_1, \ldots, p_{k-1}+p_k)$ in the following way



$$\text{Avg length of } \mathcal{C} = \bar{L}_{\mathcal{C}} = \sum_{i=1}^{k} p_i l_i$$

$$= \sum_{i=1}^{k-2} p_i l'_i + p_{k-1}(l'_{k-1}+1) + p_k (l'_{k-1}+1)$$

$$= \left(\bar{L}^*_{(p_1, \ldots, p_{k-1}+p_k)} + (p_{k-1}+p_k) \right)$$

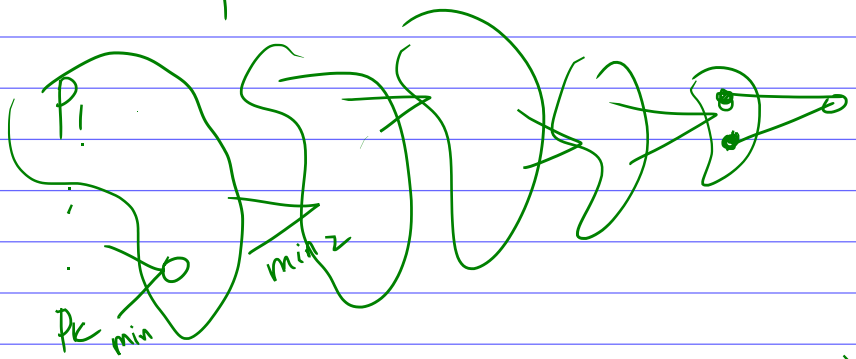
Thus is optimal by (A) for $(p_1, \ldots, p_k)$

$$(a) \quad = L^*(p_1, \ldots, p_k)$$

To get a optimal code for $(p_1, \ldots, p_{k-1}+p_k)$

we obtain an optimal code for $(k-1)$ - prob distribution obtained by combining 2 least prob symbols of $(p_1, \ldots, p_{k-1}+p_k)$

& keep repeating this process until we have only 2 - probabilities in the distribution. (When there are only 2 symbols, the lengths of the codewords will be exactly 1. ⇒ (avg length = 1))



Huffman algorithm constructs such a binary tree & assigns codewords such that at each stage the

→ avg length = (smallest 2 probabilities) + avg length of code at next stage
(sum of)
at this stage

→ So optimal length is assured by above formula since in the last stage (with 2 prob), Huffman algorithm exactly gets an optimal code

This shows that (Prop A) ⇒ $\bar{L}_{Huffman} = L^*_{(p_1 \cdots p_k)}$

Now we show Prop (A): for $K = 2$ (then the statement of prop A is trivial as $L^*_{K-1} = 0$)

$K > 2$ Let $\mathcal{C}^*_K$ be some optimal code for $(p_1, \cdots, p_k)$

& $\mathcal{C}^*_{K-1}$ be optimal code for $(p_1, \cdots, p_{k-1} + p_k)$

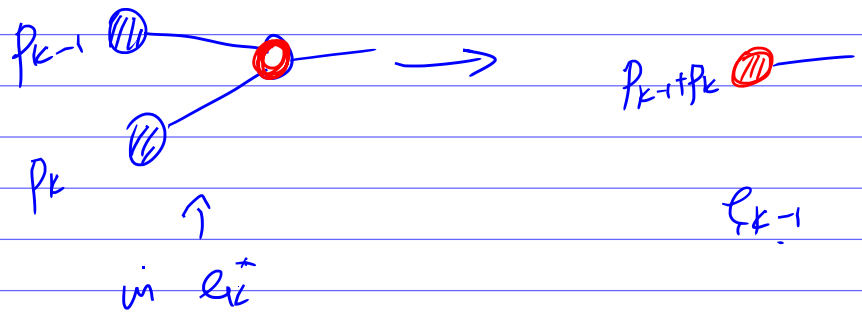
$$\bar{L}_{\mathcal{C}^*_K} = L^*_{(p_1 \cdots p_k)}$$

$$\bar{L}_{\mathcal{C}^*_{K-1}} = L^*_{(p_1, \cdots, p_{k-1} + p_k)}$$

Step 1: Now from C_k^* we will obtain a new code for $(p_1, \ldots, p_{k-1}+p_k)$ with length $\bar{L}_{C_{k-1}} = \bar{L}^*_{(p_1 \cdots p_k)} - p_{k-1} - p_k$

Step 2: also, from C^*_{k-1}, we will obtain a new code C_k for $(p_1, \ldots, p_k)$. with length $\bar{L}_{C_k} = \bar{L}^*_{(p_1 \cdots p_{k-1}+p_k)} + p_{k-1} + p_k$

Doing Step 1: We assume that the code C_k^* is picked so that Lemma 3 is satisfied



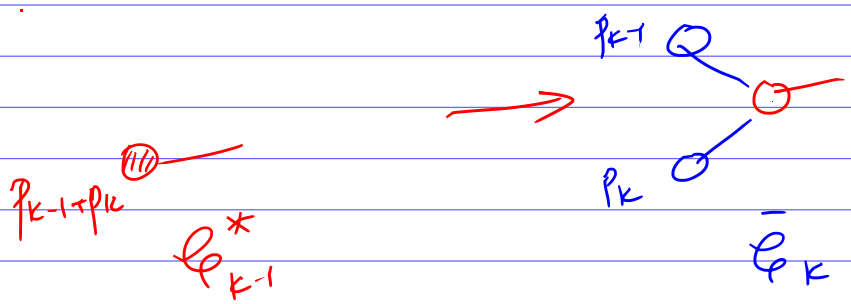
From the above construction, we see that

$$\bar{L}_{C_{k-1}} = \bar{L}_{C_k^*} - p_{k-1} - p_k$$

$$\Rightarrow \bar{L}_{C_k^*} - \bar{L}_{C_{k-1}} = p_{k-1} + p_k$$

$$\Rightarrow \bar{L}^*_{(p_1 \cdots p_k)} - \bar{L}_{C_{k-1}} = p_{k-1} + p_k \rightarrow (1)$$

Doing Step 2:



$$\bar{L}_{C_k} = \bar{L}_{C^*_{k-1}} + p_{k-1} + p_k$$

$$\bar{L}_{C_k} = \bar{L}^*_{(p_1, \cdots p_{k-1}+p_k)} + p_{k-1} + p_k$$

$$\bar{L}_{e_K} - \bar{L}^{*}_{(p_1, \ldots, p_{K-1}+p_K)} = p_{K-1} + p_K \rightarrow (2)$$

By (1) & (2) as RHS is same, LHS must be same

$$\Rightarrow \bar{L}_{e_K} - \bar{L}^{*}_{(p_1 \ldots p_{K-1}+p_K)} = \bar{L}^{*}_{p_1 \ldots p_K} - \bar{L}_{e_{K-1}}$$

$$\Rightarrow \bar{L}_{e_K} - \bar{L}^{*}_{(p_1, \ldots, p_K)} = \bar{L}^{*}_{(p_1, \ldots, p_{K-1}+p_K)} - \bar{L}_{e_{K-1}}$$

$\bar{L}_{e_K} - \bar{L}^{*}_{(p_1, \ldots, p_K)} \geq 0$

$\bar{L}^{*}_{(p_1, \ldots, p_{K-1}+p_K)} - \bar{L}_{e_{K-1}} \leq 0$

$\Rightarrow$ both sides $= 0$

$\Rightarrow \bar{L}_{e_K} = \bar{L}^{*}_{p_1 \ldots p_K} \rightarrow (3)$

& $\bar{L}^{*}_{p_1, \ldots, p_{K-1}+p_K} = \bar{L}_{e_{K-1}} \rightarrow (4)$

$\Rightarrow$ Proof A is true (Take (3) & substitute in (2))

This completes the proof of optimality of Huffman coding algorithm.