Class no 11 : Fixed to variable length source coding :-

$X \sim p_X \quad X \in \{A, B, C, D\}$

The set $\{0, 1, 10, 11\}$ = Code
each is a codeword

Suppose the code is $A \to 0,\ B \to 1,\ C \to 10,\ D \to 11$

Then the source sequences 'BA', 'C' are not uniquely decodable (ie) they result in the same codeword seq 10, this is not acceptable as this creates a nonzero prob of error which is not ok.

One Solution is to use A PREFIX-FREE CODE

Defn of P-F Code :-

A code $\mathcal{C}$ [Set of all codewords] is called a P-F code if no codeword in $\mathcal{C}$ is a prefix of another codeword in $\mathcal{C}$

[(ie) For any $\underline{c}, \underline{c}' \in \mathcal{C}$ $(\underline{c} \neq \underline{c}')$, $\underline{c}' \neq \underline{c} \underbrace{\sim\sim}_{\text{some binary string}}$]

Eg: In the code $\{0, 1, 10, 11\}$, the codeword 1 is a prefix of codeword 10. (or) 11.

Terminology :-

'codeword' is the string resulting from some particular input / source sequence

Code : = Set of all codewords

Neither Code or codewords are source sequences. They refer the mapped source sequences

So this is not a prefix-free code.

Ex: for P-F code:

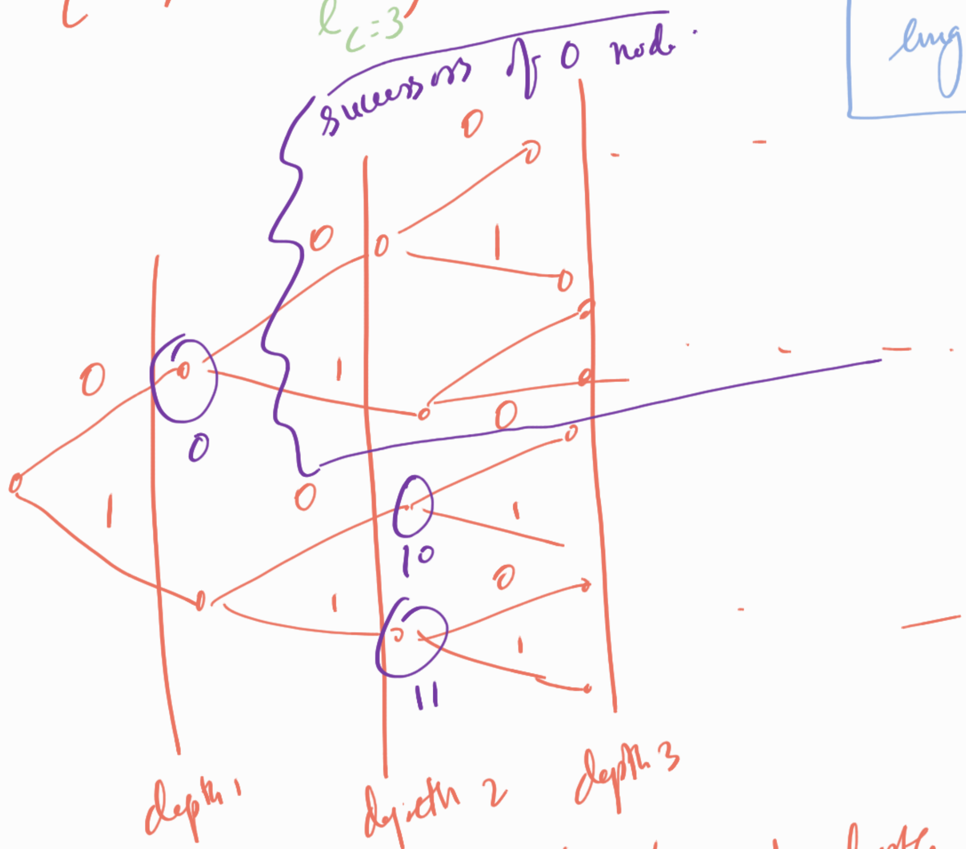
		A	B	C	D
ℓ_1	(1)	10,	11,	01,	00

$\{10, 11, 01, 00\}$ $l_C = 2$

ℓ_2 (2) $\{10, 110, 1110, 11110\}$ $l_C = 4$

ℓ_3 (3) $\{0, 10, 111, 110\}$ $l_C = 3$

> Aside: A 'variable' length code does not mean lengths of codewords are all different.
> Whereas in a Fixed-length code, all codewords MUST be of the same length



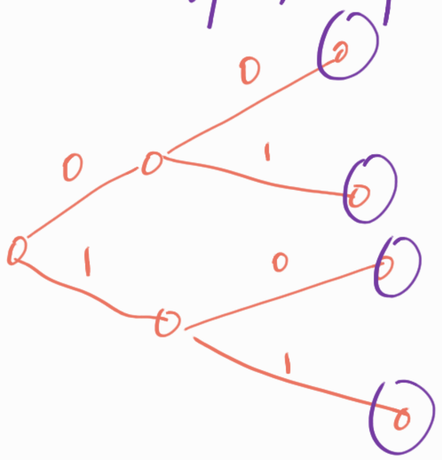
No of nodes of the tree at depth i : 2^i.

Suppose we are marking off codewords of the PF code on the nodes of the binary tree.

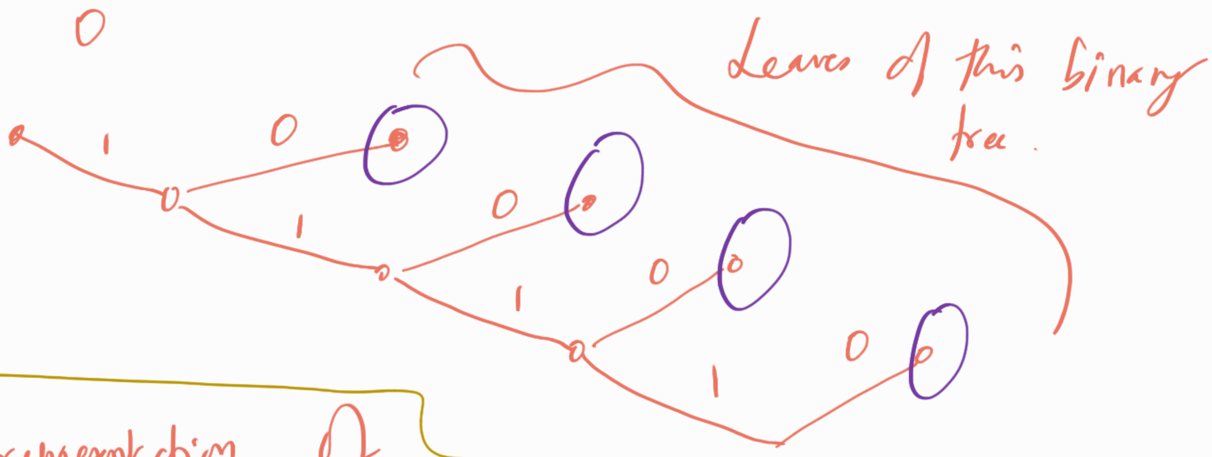
→ Then no successor nodes of any codeword node will correspond to a codeword node for any PF code [necessary & sufficient]

Tree representation of the P-F codes

C_1 :



C_2 :



This tree representation of collection of nodes which are not successors of each other gives us a P-F code

$\rightarrow$ Use this to understand why PF code will give zero prob of errors.

Hint: Each seq of source symbol has a unique codeword sequence

Terminology :- 'Leaves' of the binary tree

$\downarrow$

Nodes in the tree with no children

In a binary tree representing a PF code, the leaves will be the codewords.

Suppose $X \in \mathcal{X}$, $X \sim p_X$, & $|\text{supp}(p_X)| = s$ (X can take s possible values)

source

Let $l_i : i = 1, \ldots, s$ be the length of the binary codewords associated to the i^{th} symbol in $\text{supp}(p_X)$.

Expected length of the code $\bar{L} = \sum p_i l_i$

Suppose $X \in \{A, B, C, D\}$ $\quad p_X(A) = \frac{1}{4} = p_X(B)$

$p_X(C) = \frac{3}{8}, \; p_X(D) = \frac{1}{8}$

(a) $\bar{L}(\mathcal{C}_1) = \sum_{i=1}^{4} p_i l_i = \frac{1}{4}(2) + \frac{1}{4}(2) + \frac{3}{8}(2) + \frac{1}{8}(2)$

Exp length of code $\mathcal{C}_1$

$= 2$

(b) $\bar{L}(\mathcal{C}_2) = \sum p_i l_i = \frac{1}{4}(2) + \frac{1}{4}(3) + \frac{3}{8}(4) + \frac{1}{8}(5)$

$= 27/8 \approx 3.4$

(c) $\bar{L}(\mathcal{C}_3) = ?$

The goal of P-F variable length source coding is to minimize $\bar{L}$ over all possible choice of P-F codes.

Let the minimum possible $\bar{L}$ among all PF codes be denoted $\bar{L}^*$ ($*$ → optimality)

We will show that

$$H(X)+1 \geq \bar{L}^* \geq H(X)$$

No matter what code we use, avg length is at least $H(X)$.

We will show a particular coding scheme with length $H(X)+1$