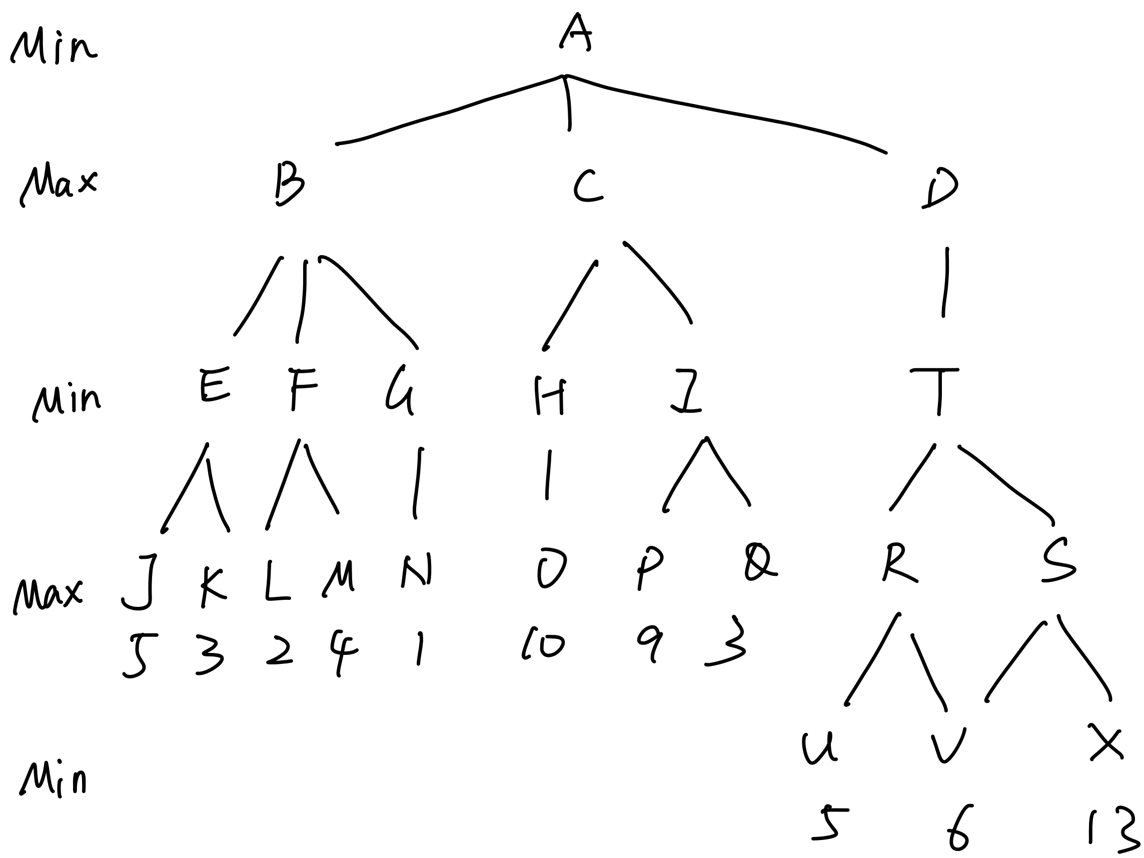
23-S1-Q2

first - minimizer

Q(i) left -> righ. α-β.

(ii) not be examined. nodes



(b) NN

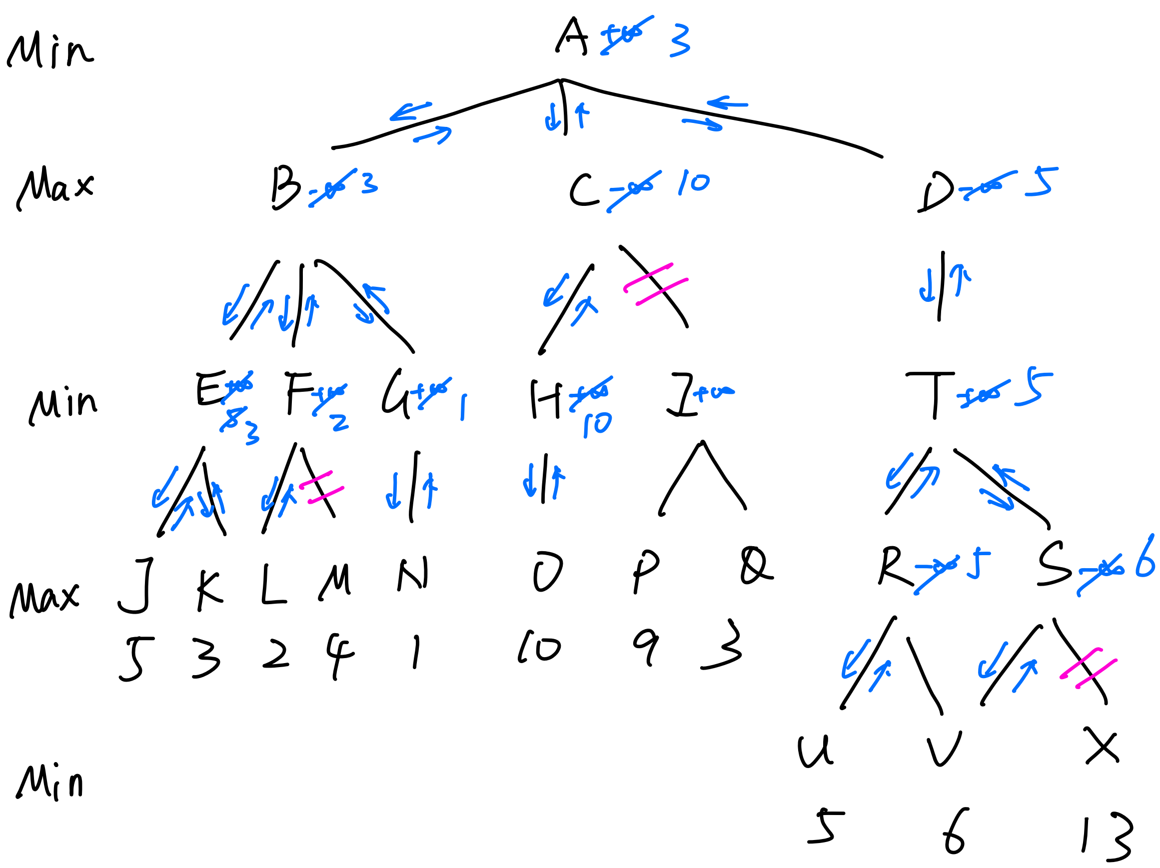
(i) O_j

(ii)

(c)(i) gradient vanishing

(ii) methods

Solution (a)



(ii) not examinate nodes

M I P Q X

(b) (i) O_j bias = [0.2 0.6 0.2 -0.9 -1.0 0.8]

unit_j	Net input net_j	output O_j
1		1
2	$0.8\times1+0.2\times1=1$	1
3	$0.4\times1+0.6\times1=1$	1
4	$1\times1+0.5\times1+0.3\times1+0.2\times1=2$	2
5	$0.5\times1+0.2\times1+0.1\times1-0.9\times1=-0.1$	0
6	$0.3\times2+0.2\times0-1\times1=-0.4$	0
7	$0.2\times2+0.5\times0+0.8\times1=1.2$	1.2

(ii) $\delta_j = \sigma'(net_j)\sum_k \delta_k w_{kj}$

$\delta_k = \sigma'(net_k)(t_k - O_k)$

$$\sigma'(x) = \begin{cases} 1, & x>0 \\ 0, & x<0 \end{cases}$$

(iii) $\delta_6 = \sigma'(net_6)(t_6 - O_6)$

$= 0\times(0.8-0)$

$= 0$

$\delta_7 = \sigma'(net_7)(t_7 - O_7)$

$= 1\times(0.2-1.2)$

$= -1$

$$\delta_4 = \sigma'(net_4)\left(\delta_6 \times W_{64} + \delta_7 \times W_{74}\right)$$

$$= 1 \times (0 \times 0.3 + (-1) \times 0.2)$$

$$= -0.2$$

(iv) ① $\Delta w_{ji} = \eta \delta_j O_i$

$$\Delta w_{41} = \eta \delta_4 O_1$$

$$= 0.1 \times (-0.2) \times 1$$

$$= -0.02$$

(c) (i) ① Gradients shrink exponentially while BP through many layer, because $\sigma'(x) \le 0.25$ for a sigmoid.

② Products of many $\sigma'(x)$ terms drive δ toward zero, so earlier layers learn exteremly slowly or stop learning altogether.

(ii) ① use ReLU / Leaky-ReLU to keep $\sigma'(x) \approx 1, x>0$

② maintain variance of activation and gradients by proper weight initialization

③ Batch / Layer Normalization rescales activations keeping them in regions with healthy derivative

④ Use Residual connection

⑤ Use Gradient-clipping or adaptive optimizers e.g. Adam, RMSProp prevent tiny update after many layers.