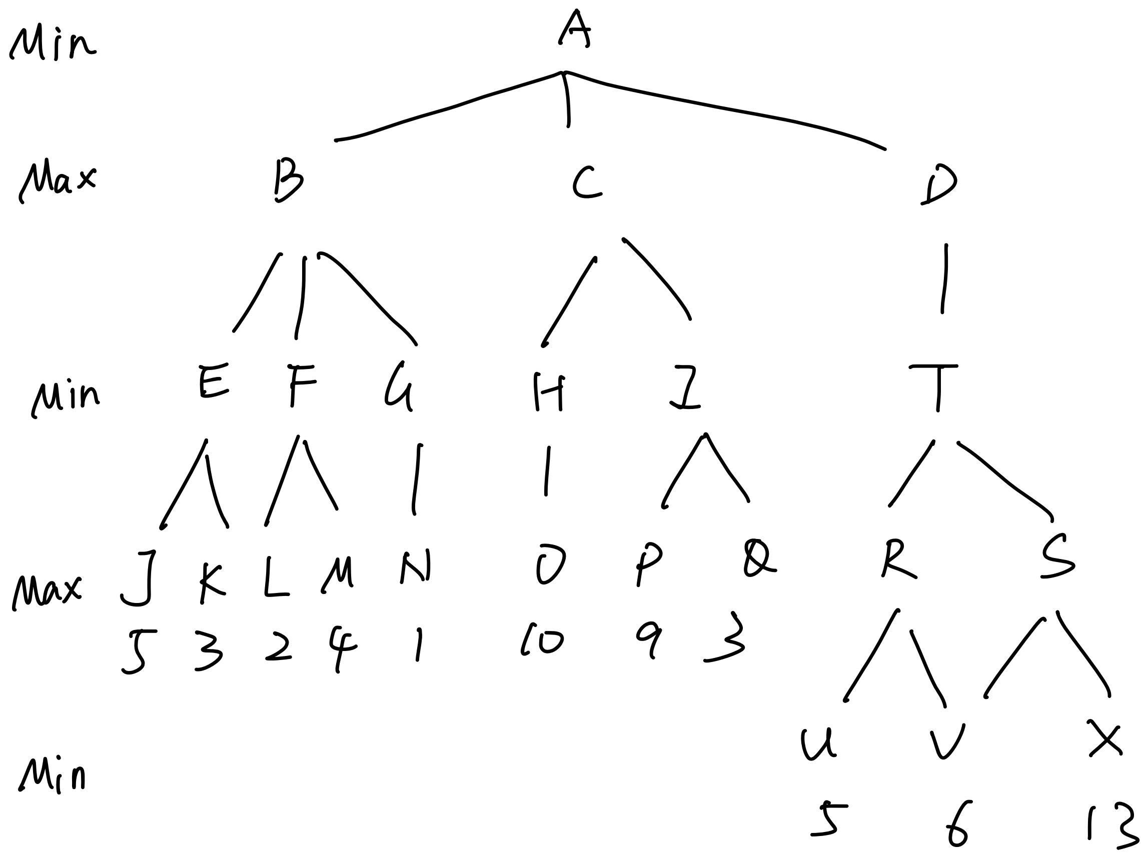
23-S1-Q2

first - minimizer

Q(i) left -> righ. α-β.

(ii) not be examined. nodes



(b) NN

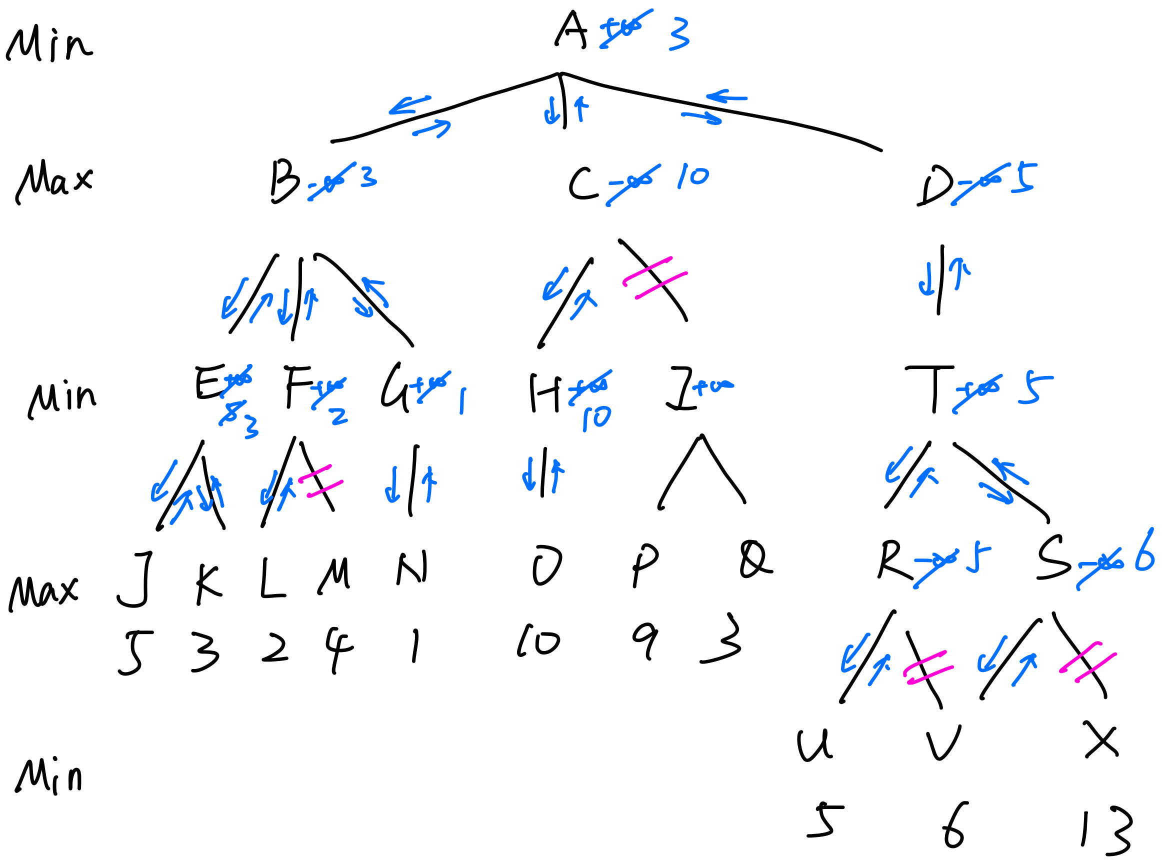
(i) Oj

(ii)

(c)(i) gradient vanishing

(ii) methods

Solution(a)



(ii) not examinate nodes

M I P Q X

(b)(i) O_j

unit$_j$	Net input net$_j$	output O_j
1		1
2	$0.8 \times 1 = 0.8$	0.8
3	$0.4 \times 1 = 0.4$	0.4
4	$0.5 \times 0.8 + 0.3 \times 0.4 + 1.0 \times 1 = 1.52$	1.52
5	$0.5 \times 0.8 + 0.2 \times 0.4 + 0.1 \times 1 = 0.58$	0.58
6	$0.3 \times 1.52 + 0.2 \times 0.58 = 0.572$	0.572
7	$0.2 \times 1.52 + 0.5 \times 0.58 = 0.594$	0.594

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

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

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

(iii) $\delta_4 = \sigma'(net_4)\left(\delta_6 \times 0.3 + \delta_7 \times 0.2\right)$

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

$= 1 \times (0.8 - 0.572)$

$= 0.228$

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

$= 1 \times (0.2 - 0.594)$

$= -0.394$

$\delta_4 = 1 \times (0.228 \times 0.3 + (-0.394) \times 0.2)$

$= -0.0104$

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

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

$= 0.1 \times (-0.0104) \times 1$

$= -0.00104$

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

② Products of many $\sigma'(x)$ terms drive δ toward zero, so earlior layers learn extremly 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 weighe initialisation

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.