31.3.2020

$$\delta^{(l-1)} = \mathrm{diag}\left(\varphi'(z^{(l-1)})\right) \cdot W^{(l)} \cdot \delta^{(l)}$$

$$\nabla_{W^{(l)}} \ell(\cdot) = \delta^{(l)} \cdot v^{(l-1)}$$

$v^{(l-1)}$ ← activations from layer $l-1$

$$v^{(l)} = \varphi\left(W^{(l)} v^{(l-1)}\right)$$

Failure mode:

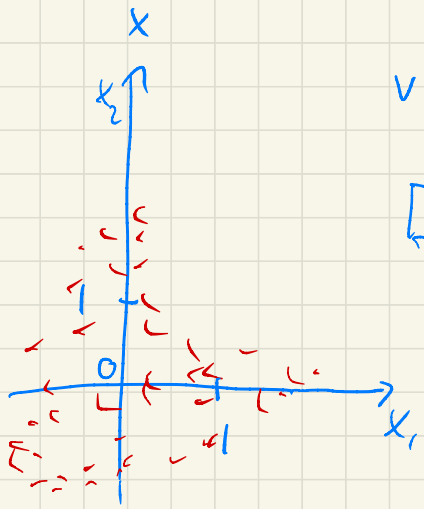
(1) vanishing gradients $\quad \nabla_{W^{(l)}} \ell \to 0$

(2) exploding gradients $\quad \nabla_{W^{(l)}} \ell \to \infty$

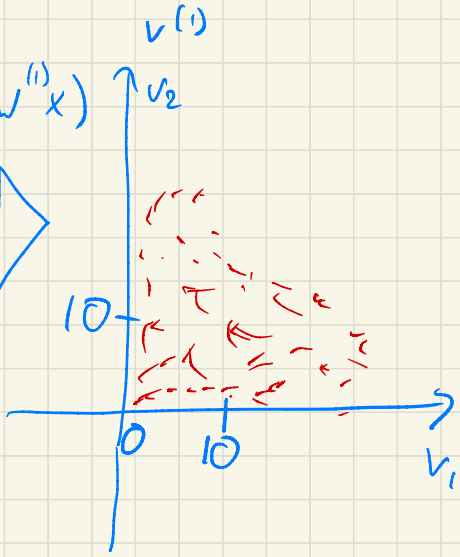
One possible reason for (1)/(2) is $v^{(l)} \to 0 / \infty$

He-initialization tries to ensure that $\mathrm{Var}(v^{(l)}) = \text{const}$ at initialization

Batch-normalization aims to ensure standardized activations throughout training

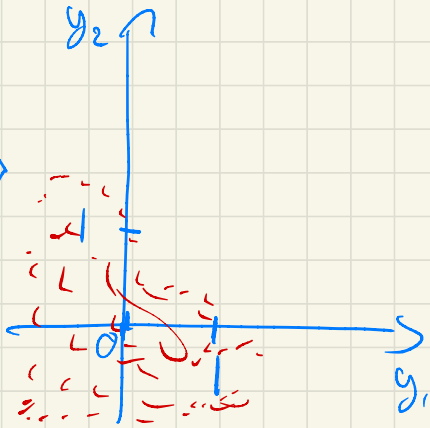


$v = \varphi(v^{(1)} x)$



$y^{(1)} = BN(v^{(1)})$

BN



Effect of batch normalization