$$J(w) = \underbrace{R(w)}_{\text{avg loss}} + \underbrace{\lambda \|w\|^2}_{\text{Regularizer}}$$

$$\nabla J(w) = \nabla R(w) + 2\lambda w$$

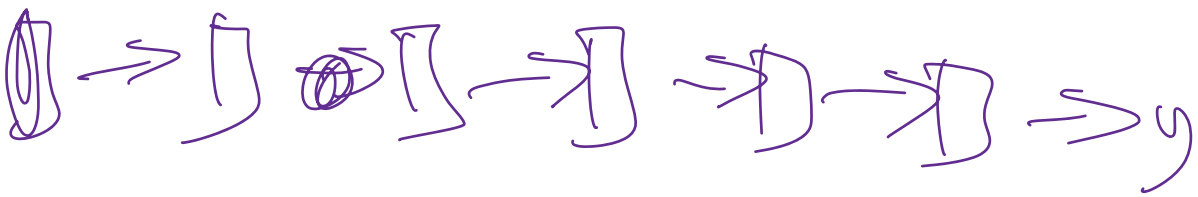
Gradient Step:

$$w^{new} = w^{old} - \alpha(\nabla J(w^{old}))$$
$$= w^{old} - \alpha \nabla R(w) + 2\alpha\lambda w^{old}$$

Say you gradient step only on regularizer

$$w^{new} = w^{old} - 2\alpha\lambda w^{old}$$
$$= w^{old}\underbrace{(1 - 2\alpha\lambda)}_{\text{"shrinks" } w^{old}}$$

Many layers



$W^{(2)}$ . . . . .

$$\frac{\partial R}{\partial W^{(2)}} = \frac{\partial W^{(3)}}{\partial f^{(3)}} \frac{\partial f^{(3)}}{\partial W^{(2)}}$$

All small numbers, makes final gradient small

Layer Normalization

~~Algo~~

Normally:

$$h^1 = f(W^1 x)$$

$$h^2 = f(W^2 h^1)$$

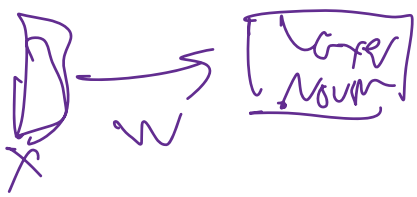
---

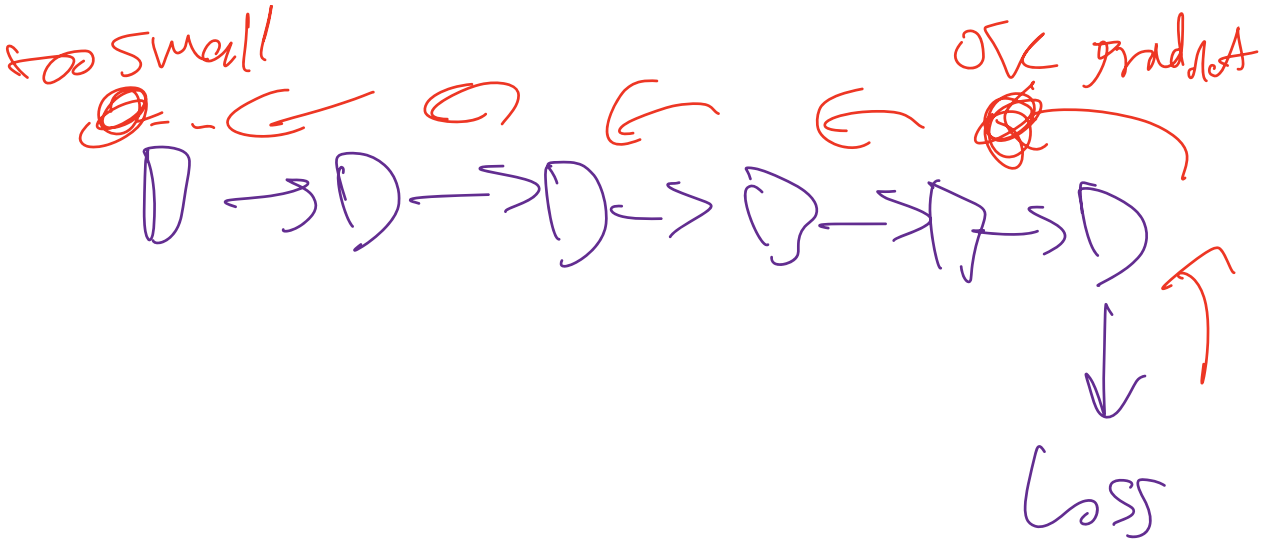
with layer norm

$$h^1 = \frac{f(W^1 x)}{\| f(W^1 x) \|}$$

so values are merely scal

etc.





→ forward pass

→ backward pass: Chain rule