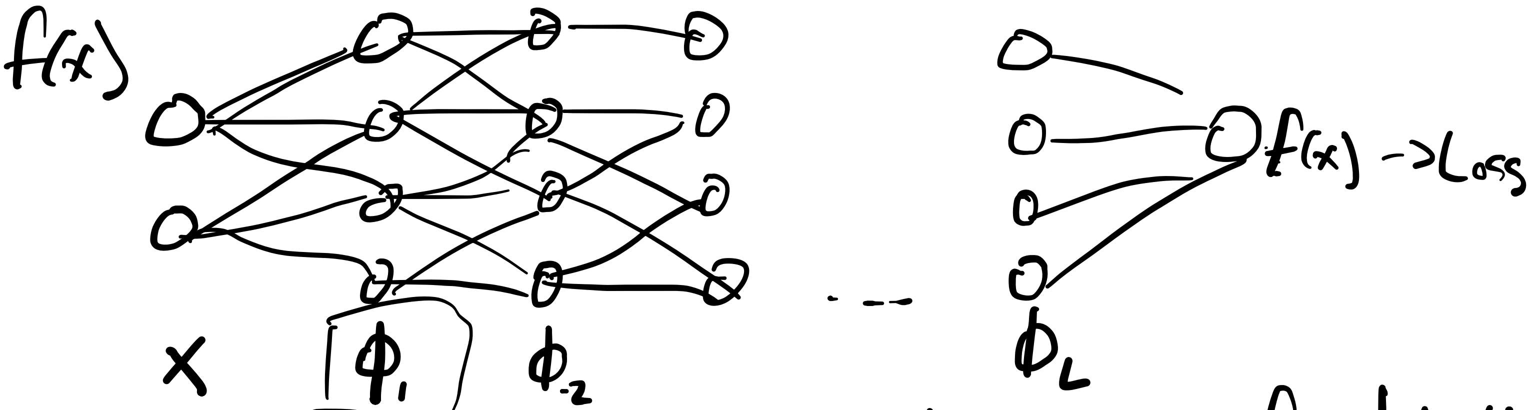



$\phi_1 = \sigma(x^T w_1 + b)$

$\phi_L = \sigma(\phi_{L-1} W_L + b)$ $\quad f = \phi_L w + b$

$\frac{d\phi_1}{dw_1}$

$\phi_2 = \sigma(\phi_1^T w_2 + b) \cdots$

$L = Loss(f)$

$$\frac{dL}{dw_1} = \underbrace{x\sigma'(xw_1+b)}_{\frac{d\phi_1}{dw_1}} \frac{d\phi_2}{d\phi_1}\cdot\frac{d\phi_3}{d\phi_2}\cdots\frac{d\phi_L}{d\phi_{L-1}}\cdot\frac{df}{d\phi_L}\cdot\frac{\partial L}{\partial f}$$

$$\frac{dL}{dw_1} = x\sigma'(xw_1+b)\left(\prod_{l=1}^{L}\frac{d\phi_l}{d\phi_{l-1}}\right)\cdot\frac{df}{d\phi_L}\cdot\frac{dL}{df}$$

$$\left|\frac{dL}{dw_1}\right| = \underbrace{|x|}\,|\sigma'(xw_1+b)|\,\underbrace{\prod_{l=1}^{L}\left|\frac{d\phi_l}{d\phi_{l-1}}\right|}\cdot\ldots$$

$$\left|\frac{d\phi_L}{d\phi_{L-1}}\right| \approx M \rightarrow \left|\frac{dL}{dw_1}\right| \approx |x||\sigma'|\underbrace{M^L}$$

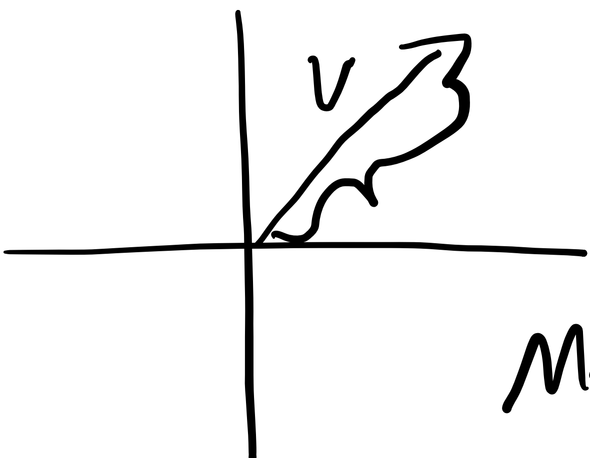
If $M > 1 \rightarrow \left|\frac{dL}{dw}\right| >> 1$

gradient exploding

If $M < 1 \rightarrow \left|\frac{dL}{dw_1}\right| \approx 0$

Vanishing gradients

## Gradient Clipping




$$||V||_2 = \sqrt{\sum_{i=1}^{d} v_i^2}$$

Max length : $\varepsilon$

$$V \rightarrow \frac{V}{||V||_2} \qquad \varepsilon \frac{V}{||V||_2}$$

length 1      length $\varepsilon$

Clipping by norm

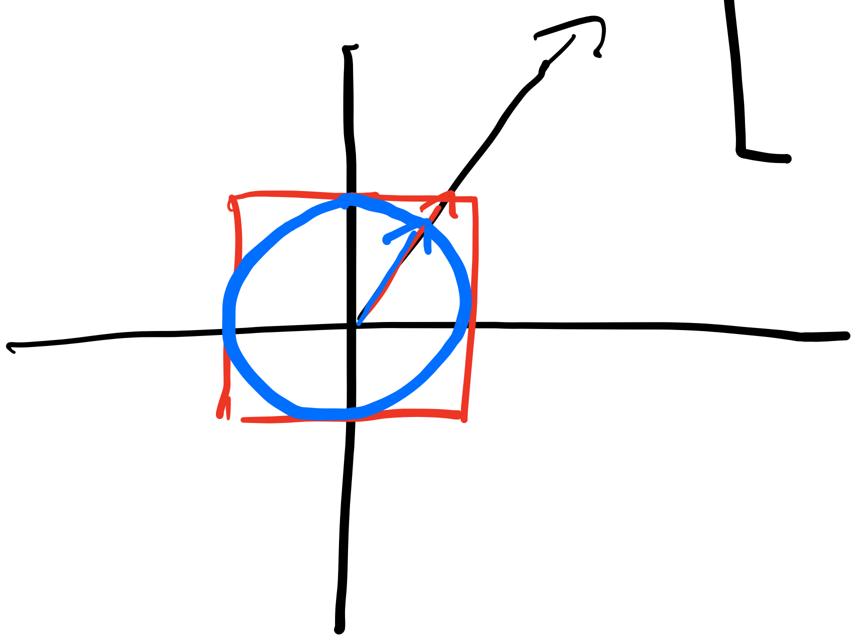
$Clip_{norm}\left(\frac{dL}{dw}\right)$      $dL$

$$\rightarrow \text{if } \left\|\frac{dL}{dw}\right\| > \varepsilon \rightarrow \frac{\frac{dL}{dw}}{\left\|\frac{dL}{dw}\right\|} \cdot \varepsilon$$

$$\text{Otherwise} \rightarrow \frac{dL}{dw}$$

## Clipping by Value

$$\text{Clip}_{value}\left(\frac{dL}{dw}\right) = \begin{bmatrix} \max\left(\min\left(\frac{dL}{dw_1}, \varepsilon\right), -\varepsilon\right) \\ \vdots \\ \vdots \end{bmatrix}$$



## Input normalization

$x \sim \text{Data}$

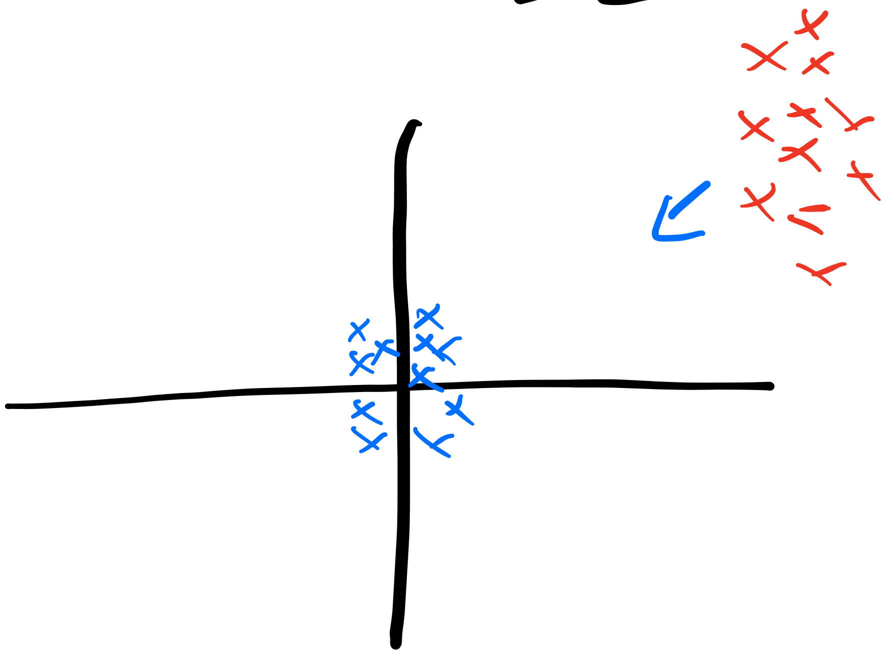
Issues

$\mathbb{E}[x] \gg 0$

$\text{Var}[x] \gg 1$

normalize(x)

$$x \to \frac{x - E[x]}{\sqrt{Var[x] + \varepsilon}} \qquad \varepsilon \ll 1$$



$E[x] = ?$ $\qquad Var[x] = ?$

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

$$s^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2$$

(unbiased est. of Var)

$$s^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2$$

(biased est. of Var)

$$\text{Normalize}(x) = \frac{x - \bar{x}}{\sqrt{s^2 + \varepsilon}}$$

# Batch Normalization

$$BN(x) = \frac{x - \mathbb{E}[x]}{\sqrt{Var[x] + \varepsilon}}$$

$$\phi_l = \sigma(\phi_{l-1} W + b) \quad \text{w/o BatchNorm}$$

$$\phi_l = \sigma(BN(\phi_{l-1}) W + b) \quad \text{w/ BatchNorm}$$

or

$$\phi_L = \sigma(BN(\phi_{L-1} W + b))$$

Est:

$$BN(x) = \frac{x - \bar{x}}{\sqrt{s^2 + \varepsilon}}$$

## Gradient descent

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

## Stochastic gradient descent

Minibatch of B obs.

$$\bar{x} = \frac{1}{B}\sum_{i=1}^{B} x_i \,, \quad s^2 = \frac{1}{B-1}\sum_{i=1}^{B}(x_i - \bar{x})^2$$

$B > 1$ $$BN(x) = \frac{x - \bar{x}}{\sqrt{s^2 + \varepsilon}}$$

Training

## Test time

while training

$$\underbrace{\bar{\mu}^{(k+1)}}_{\substack{SS \\ E[x]}} \longleftarrow \beta\bar{\mu}^{(k)} + (1-\beta)\,\bar{x}^{(k)}$$

$$\bar{\sigma}^{2(k+1)} \leftarrow \beta \bar{\sigma}^{2(k)} + (1-\beta) s^{2(k)}$$

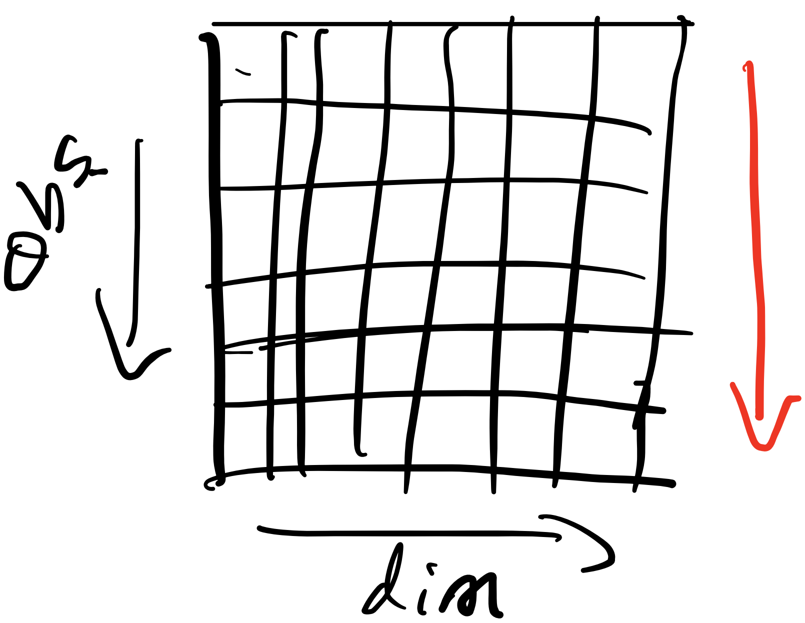
$$\overset{ss}{\mathrm{Var}}[x]$$

$$BN_{test}(\underline{x}) = \frac{x - \bar{\mu}}{\sqrt{\bar{\sigma}^2 + \epsilon}}$$

## Layer normalization

$$LN(\underline{x}) = \frac{\underline{x} - \frac{1}{d}\sum_{i=1}^{d} x_i}{\sqrt{\frac{1}{d-1}\sum_{i=1}^{d}(x_i - \bar{x}) + \epsilon}} = \frac{\underline{x} - \bar{x}}{\sqrt{s^2 + \epsilon}}$$

$$\bar{x} = \frac{1}{d}\sum_{i=1}^{d} x_i$$




Batch Norm
Normalize columns

→ LayerNorm Normalize rows