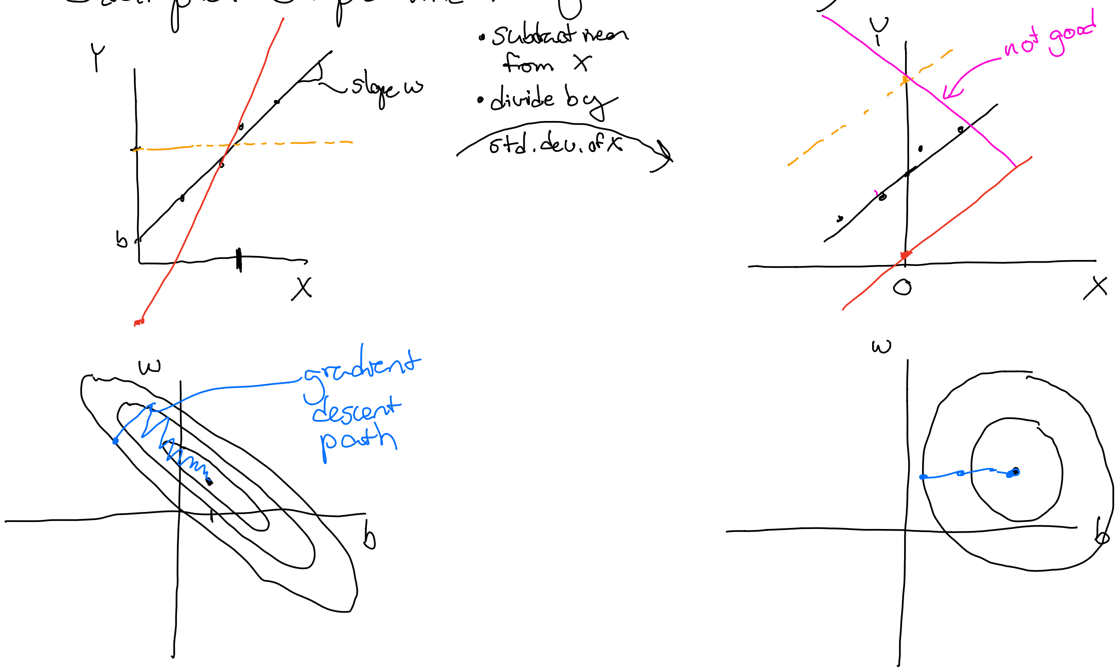
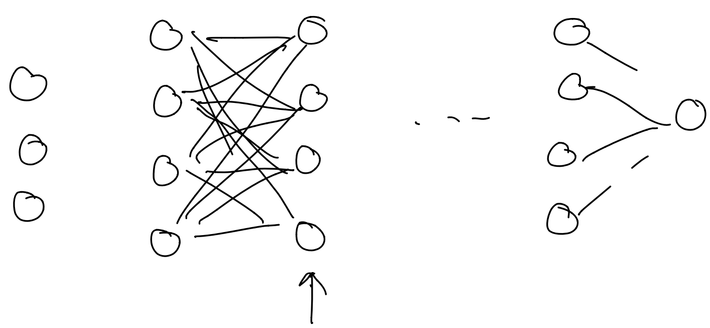
Reminder of normalizing network inputs:

Example: Simple linear regression: one X, numeric Y



Larger network:



for this layer:

$$z^{[l]} = b^{[l]} + (W^{[l]})^T \boxed{a^{[l-1]}}$$

$$a^{[l]} = g^{[l]}(z^{[l]})$$

Idea of batch normalization:

- we standardize $a^{[l-1]}$, this fixes geometry of loss for $b^{[l]}$ and $W^{[l]}$.

Normaling inputs:

$$\mu^{[0]} = \frac{1}{m} \sum_{i=1}^{m} a^{(i)[0]}$$ ← this is $x^{(i)}$

$$\sigma^{2[0]} = \frac{1}{m-1} \sum_{i=1}^{m} \left(a^{(i)[0]} - \mu^{[0]}\right)^2$$

$$\tilde{x}^{(i)} = \frac{x^{(i)} - \mu^{[0]}}{\sqrt{\sigma^{2[0]}}}$$ } average of $\tilde{x}^{(i)}$'s is 0; std. dev. of $\tilde{x}^{(i)}$'s is 1.

Batch normalization for later layers:

$$\mu^{[l-1]} = \frac{1}{m} \sum_{i=1}^{m} a^{(i)[l-1]}$$

$$\sigma^{2[l-1]} = \frac{1}{m-1} \sum_{i=1}^{m} \left(a^{(i)[l-1]} - \mu^{[l-1]}\right)^2$$

$$a_{std}^{(i)[l-1]} = \frac{a^{(i)[l-1]} - \mu^{[0]}}{\sqrt{\sigma^{2[l-1]} + \varepsilon}}$$ } Average of $a_{std}^{(i)[l-1]}$'s is 0; std dev. of $a_{std}^{(i)[l-1]}$ is ≈ 1

← to prevent division by 0

$$\tilde{a}^{(i)[l-1]} = \beta^{[l-1]} + \gamma^{[l-1]} * a_{std}^{(i)[l-1]}$$ } average of $\tilde{a}_j^{(i)[l-1]}$ is $\beta_j^{[l-1]}$; std. dev. of $\tilde{a}_j^{(i)[l-1]}$ is $\approx \sqrt{\gamma^{[l-1]}}$

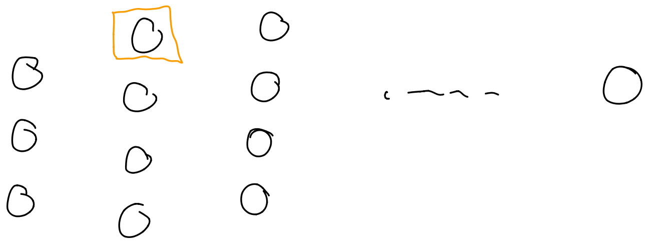
shape $(n_{l-1}, 1)$

these calculations occur within a Batch normalization layer

the layer parameters are β and γ.

Keras function is layers.BatchNormalization()

A second motivation for Batch Normalization:
helps reduce dependence among parameters in different layers of network



Suppose we have $m=4$ obs.

$a_1^{(1)[1]} = 2$, $a_1^{(2)[1]} = 0$, $a_1^{(3)[1]} = 2$, $a_1^{(4)[1]} = 4$

BN: $\mu^{[1]}$ is a $(4,1)$ shape array

$$\mu_1^{[1]} = \frac{2+0+2+4}{4} = 2$$

$$\sigma_1^{2[1]} = \frac{(2-2)^2 + (0-2)^2 + (2-2)^2 + (4-2)^2}{4} = \frac{8}{4} = 2$$

(should be 4-1)

$a_{1,std}^{(1)[1]} = \frac{\cancel{2-2}}{\sqrt{2}} \to 0$ $(+\varepsilon)$ $\quad a_{1,std}^{(2)[1]} = \frac{0-2}{\sqrt{2}} \sim -\sqrt{2}$ $\quad a_{1,std}^{(3)[1]} = \frac{\cancel{2-2}}{\sqrt{2}} \to 0$ $\quad a_{1,std}^{(4)[1]} = \frac{4-2}{\sqrt{2}} \sim \sqrt{2}$

$\tilde{a}_1^{(1)[1]} = \beta \quad \tilde{a}_1^{(2)[1]} = \beta - \sqrt{2}\gamma \quad \tilde{a}_1^{(3)[1]} = \beta \quad \tilde{a}_1^{(4)[1]} = \beta + \sqrt{2}\gamma$

What happens if $W^{[1]}$ changes and

$a_1^{(1)[1]} = 200$, $a_1^{(2)[1]} = 0$, $a_1^{(3)[1]} = 200$, $a_1^{(4)[1]} = 400$.

↳ after B.N.,

$\tilde{a}_1^{(1)[1]} = \beta$, $\tilde{a}_1^{(2)[1]} = \beta - \sqrt{2}\gamma$, $\tilde{a}_1^{(3)[1]} = \beta$, $\tilde{a}_1^{(4)[1]} = \beta + \sqrt{2}\gamma$

we don't need a big adjestment to param's for layer 2 to accont for a big change in activation outputs from layer 1.