WK2 Sec 4: Gradient Descent to learn w, b

$$\hat{y} = \sigma(w^T x + b), \quad \sigma(z) = \frac{1}{1+e^{-z}}$$

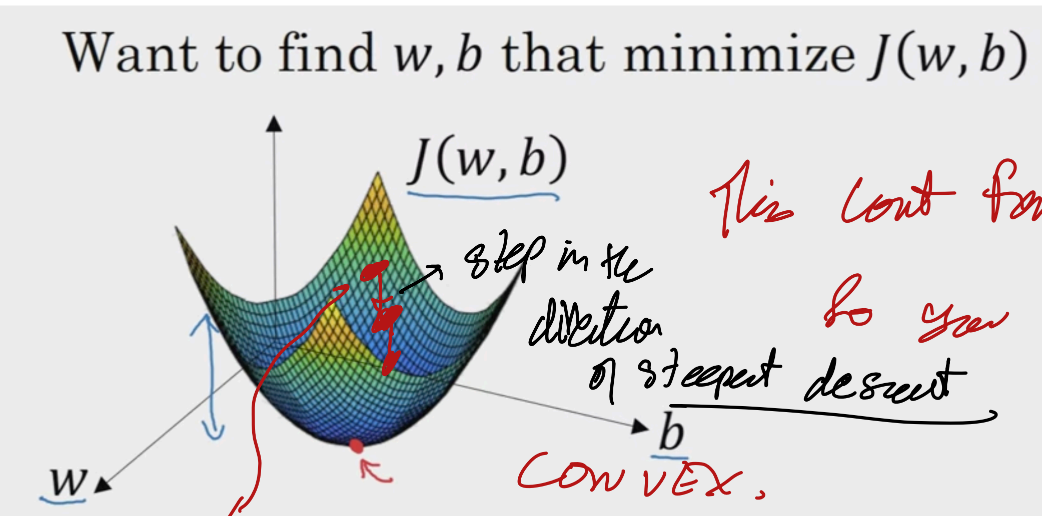
$$J(w,b) = \frac{1}{m}\sum_{i=1}^{m} \mathcal{L}(\hat{y}^{(i)}, y^{(i)})$$

$$= \frac{-1}{m}\sum_{i=1}^{m} y^{(i)} \log \hat{y}^{(i)} + (1-y^{(i)})\log(1-\hat{y}^{(i)})$$

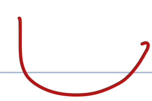
Want to minimize w, b to minimize $J(w,b)$.

Imagine for now $w \in \mathbb{R}$, $b \in \mathbb{R}$,

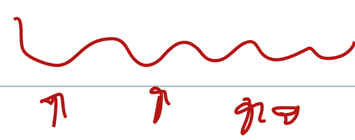
$J(w,b)$



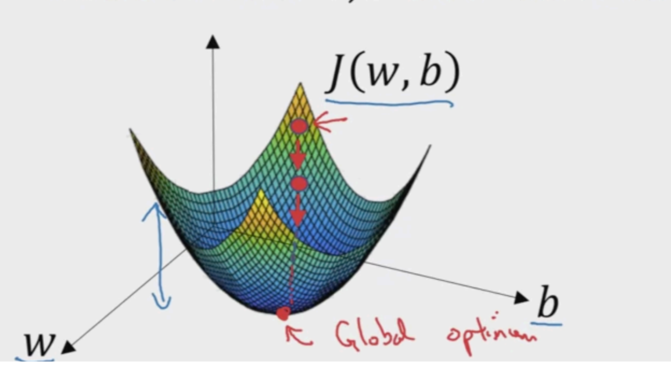
This cost fn is a convex fn so you have one min



initialize w, b as random (usually 0)

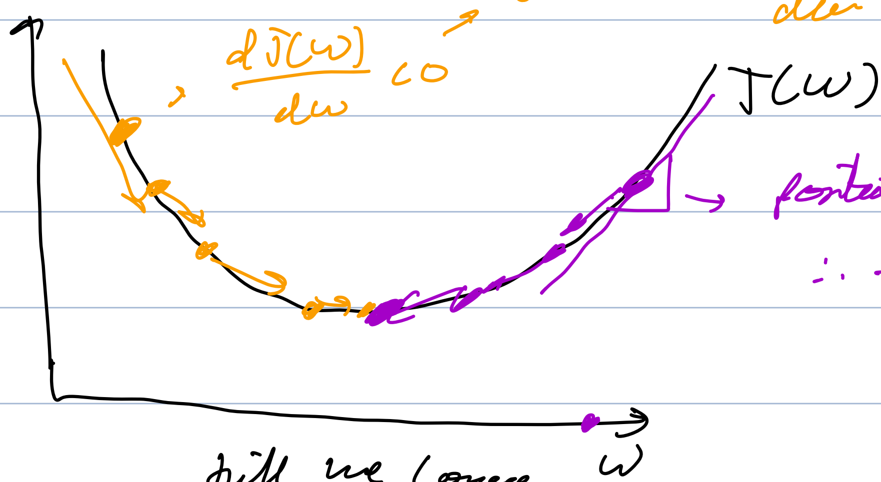


Fn is convex, so no matter where you start, you get the optimum



Gradient Descent. Example only w is parameter

so $-\alpha \frac{dJ(w)}{dw}$ → step to the right



→ partial derivative

$\therefore -\alpha \frac{dJ(w)}{dw}$, take step to left.

← w

till we converge

Repeat { update w

$$w := w - \alpha \frac{dJ(w)}{dw}$$

}

update or change to w.

learning rate (how big our step is

"dw" in code to represent change

$$w := w - \alpha dw$$

Or if $J(w, b)$ ⇒

$$w := w - \alpha \frac{\partial J(w,b)}{\partial w}$$

dw

partial

how much for slope in w dir.

$$b := b - \alpha \frac{\partial J(w,b)}{\partial b}$$

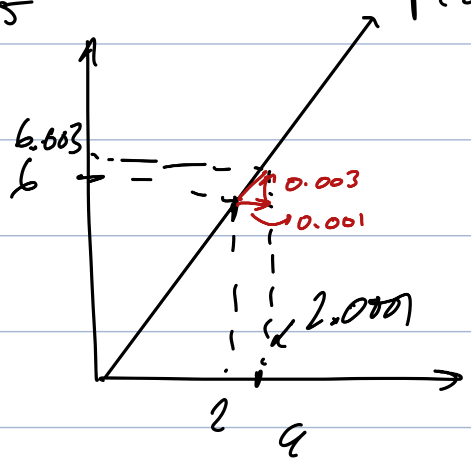
↳ db.

what update you perform in the loop.

Lec 5 and 6 : Derivatives

Lec 5

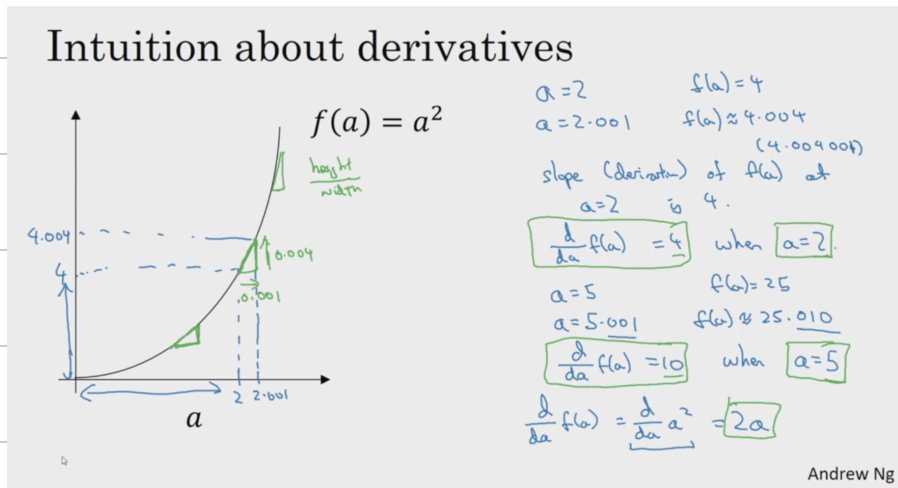
$f(a) = 3a$



$a = 2 \quad f(a) = 6$

$a = 2.001 \quad f(a) = 6.003$

$\frac{0.003}{0.001} \frac{\text{height}}{\text{width}}$ = Slope (derivative) of $f(a)$ at $a = 2$ is 3



Lec 6.

$f(a) = a^2$

$f'(a) = 2a.$

$\frac{d}{da} f(a) = \frac{d}{da} a^2 = 2a.$

↳ if you plug in value, get slope at that point

$\frac{d}{da} \log(a) = \frac{1}{a}.$