Least Square Gradient Descent

→ first lets compute gradient wrt single param θ_i & single observation (x, y) & then vectorize soln to update params simultaneously

→ one obs at a time? → iterative / online gradient learning.

for single obs, error $\Rightarrow J = (y - x\theta)^2$

$$\Rightarrow \frac{d}{d\theta_i}\left(y^2 - 2yx\theta + (x\theta)^2\right)$$

$$= 0 - 2yx_i + 2x\theta \cdot x_i$$

$$\boxed{\frac{d}{d\theta_j} = 2(x\theta - y)x_i}$$

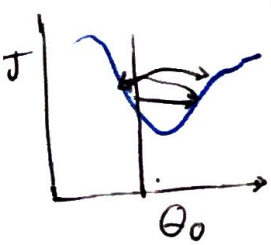
$$\frac{dJ}{d\theta} = \begin{bmatrix} 2(x\theta - y)x_0 \\ 2(x\theta - y)x_1 \\ \vdots \\ 2(x\theta - y)x_n \end{bmatrix} = X^T$$

$$\boxed{\frac{dJ}{d\theta} = 2X^T(x\theta - y)}$$

$$\boxed{\theta = \theta - 2X^T(x\theta - y)} \quad *$$

Learning Rate:

$$\theta = \frac{dJ}{d\theta}$$



if we move params according to $\frac{dJ}{d\theta}$

~~we~~ can go too slowly → too much time

go too fast → overshoot desired minima/maxima

(basically step size)

∴ we use η called learning rate that controls how much to go in direction of gradient.

(making step-size dynamic sorta)

* → standardize the data or η goes crazy!!