30/12/2025 ①

Numerical Optimization for ML & DS.

Mansoura, AI46, session 2

1) Review of Partial differentiation.

⇒ Gradient $\underline{\nabla}$ → del / nabla

2) Contour plot ←

⇒ Objective function (Review)

3) Gradient of a cost function

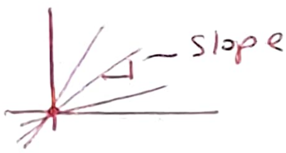
(Linear regression / Logistic Reg.)

↳ 1 - parameter

↳ 2 - parameter

⋮

↳ n - parameter



4) Problems with "Vanilla Gradient Descent" algorithm.

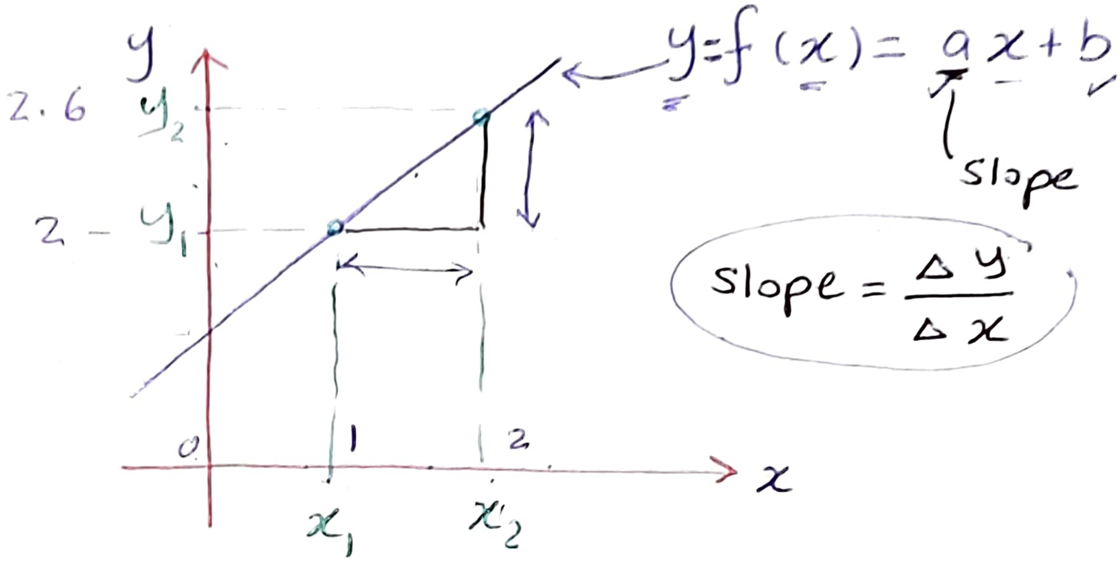
5) feature scaling

↓

6) Variants of GD algorithm.

~> derivative of a function

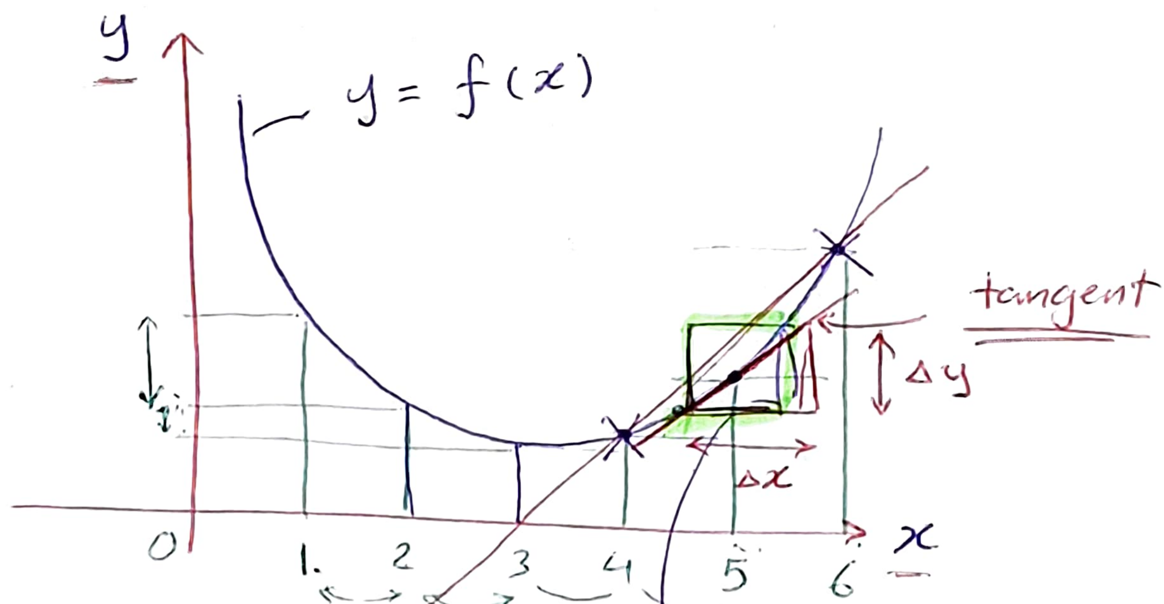
"defferintiation"



ex. slope $= \frac{y_2 - y_1}{x_2 - x_1} = \frac{2.6 - 2}{2 - 1} = \frac{0.6}{1} = 0.6$

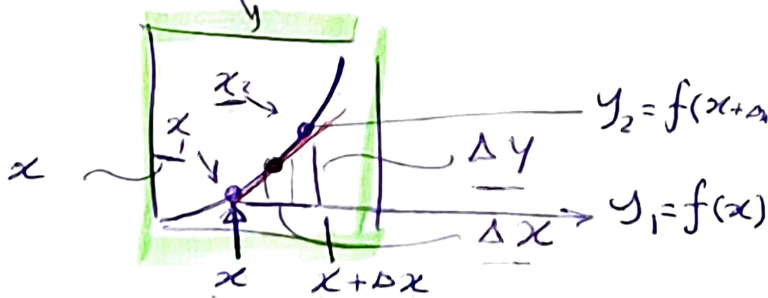
rate of change = 0.6

↳ of y value w.r.t. x value

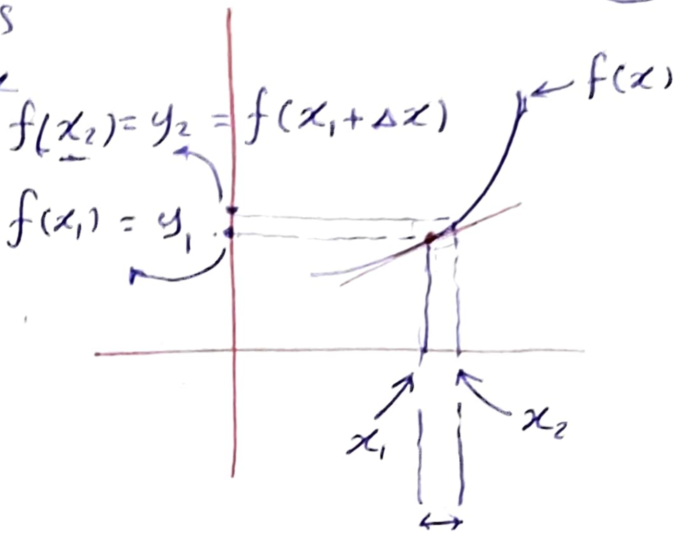


$\left(\frac{\Delta y}{\Delta x}\right)_{\Delta x \to 0}$

$\Delta x \to 0$; dx



→ for continuous functions



$$\frac{\Delta y}{\Delta x} = \frac{y_2 - y_1}{x_2 - x_1}$$

$$= \frac{f(x_2) - f(x_1)}{\Delta x}$$

$\Delta x = x_2 - x_1$

$x_2 = x_1 + \Delta x$

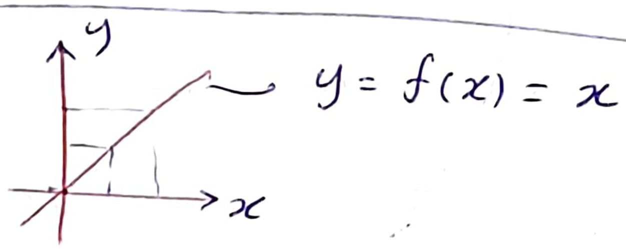
$$\text{slope}\Big|_{x=x_1} \approx \frac{f(x_1+\Delta x) - f(x_1)}{\Delta x}$$

$$\frac{dy}{dx} = \lim_{\Delta x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x}$$

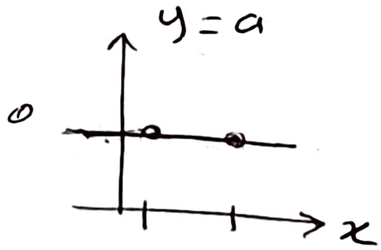
$dy \doteq f(x+\Delta x) - f(x), \Delta x \to 0$

$dx \doteq \Delta x \quad , \Delta x \to 0$

$y = x$



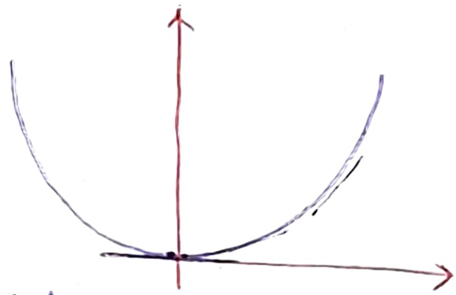
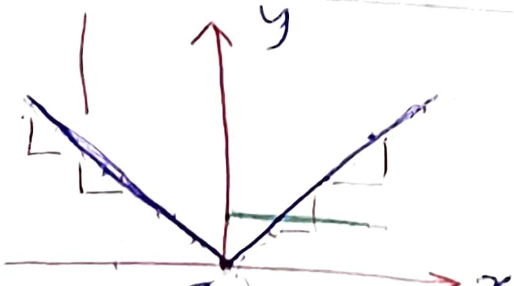
$$\frac{dy}{dx} = 1$$



$y = \underline{a}\, x$

$$\frac{dy}{dx} = \underline{a}$$

$$\frac{dy}{dx} = \underline{0}$$

$y = f(x)$	$\frac{df(x)}{dx} \equiv \frac{dy}{dx}$
$y \cancel{x} = \text{const.}$	$dy/dx = 0$
$y = ax$	$dy/dx = a$
$y = ax + b$	$dy/dx = \frac{d(ax)}{dx} + \frac{d(b)}{dx}$ $dy/dx = a$
$y = ax^2$ $y = b.x^K$	$dy/dx = a(2x)$  $dy/dx = b.K\,x^{K-1}$
$y = f(g(x))$ $y = (ax+b)^2$	$\frac{dy}{dx} = \frac{df}{dx} \cdot \frac{dg(x)}{dx}$ $\frac{dy}{dx} = 2(ax+b) \cdot \frac{d(ax+b)}{dx}$
$y = \lvert x \rvert$	 Piecewise continuous

derivative is not defined ← ← discontinuity

Partial Derivatives

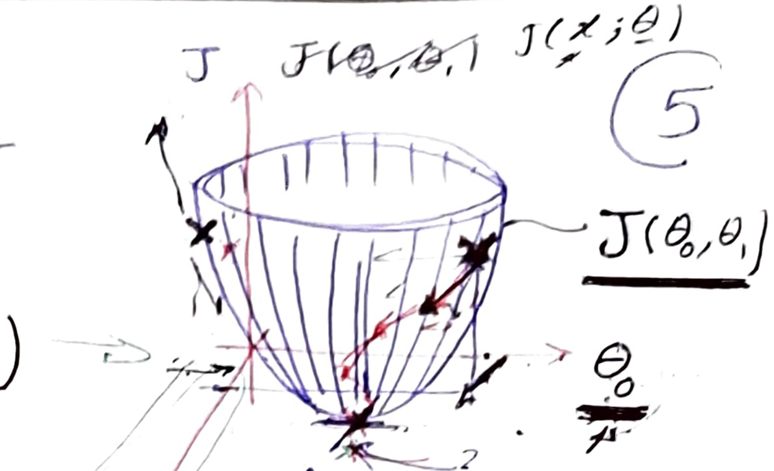
~~Gradient~~ ∇J

Partial derivatives

$$\frac{\partial J}{\partial \theta_0}$$

$$\frac{\partial J}{\partial \theta_1}$$

$J(\theta_0, \theta_1)$



$\hat{y} = \theta_0 + \theta_1 x$

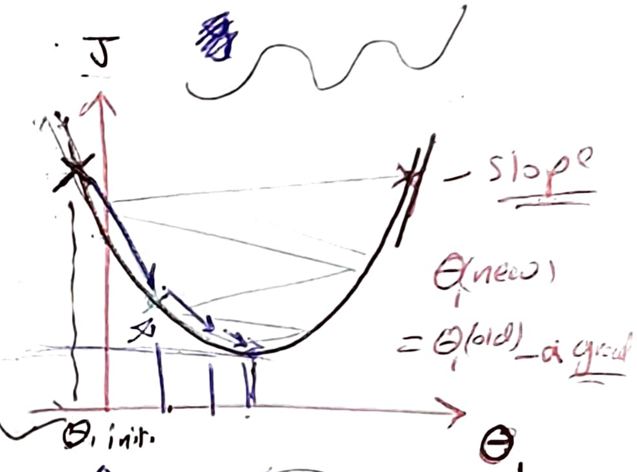
intercept

- → initialize θ_1 (initial)
- → find gradient at θ_1 of $J(\theta_1)$
- → update ~~$\theta_1^{(new)} = \theta_1^{(old)} - \alpha\, \text{gradient}$~~

$$\theta_1^{(new)} = \theta_1^{(old)} - \alpha\ \text{gradient}\,(J(\theta_{old}))$$

learning rate

- → repeat until convergence



$\hat{y} = \theta_1 x$

slope ~~intercept~~?

learning

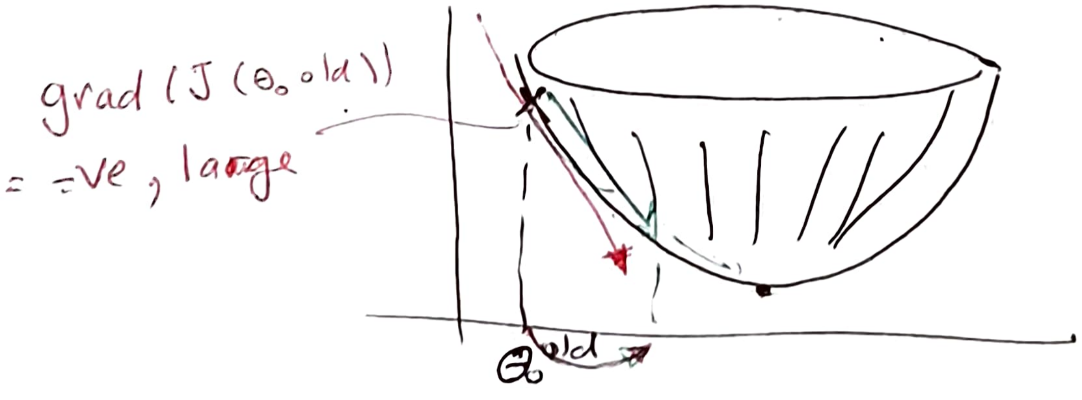
linear regression

" ~~slope~~ intercept = 0 "

$\Theta_{new} = \Theta_{old}$ + positive value

→ ve.

$$\Theta_{new} = \theta_0^{old} - \alpha\ \text{gradient}\,(J(\theta_0^{old}))$$

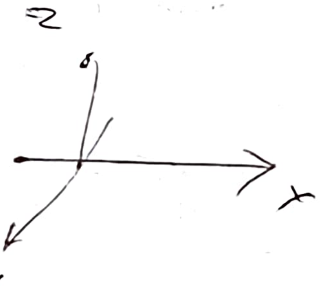


∇ : Del, Nabla operator

$$\nabla = \begin{bmatrix} \frac{\partial}{\partial \theta_0} \\ \frac{\partial}{\partial \theta_1} \\ \vdots \\ \frac{\partial}{\partial \theta_n} \end{bmatrix}$$

---

e.g., 3D problem



$$f(x, y, z) = 2x + 3y^2 - z^3$$

$$\frac{\partial f}{\partial x}, \quad \frac{\partial f}{\partial y}, \quad \frac{\partial f}{\partial z}$$

gradient "Vector" → ∇f = vector

f: Scalar function

$$\boxed{\vec{\nabla} = \begin{bmatrix} \partial/\partial x \\ \partial/\partial y \\ \partial/\partial z \end{bmatrix}} \Rightarrow \nabla f = \begin{bmatrix} \partial f/\partial x \\ \partial f/\partial y \\ \partial f/\partial z \end{bmatrix}$$

gradient (f)
grad (f)

$$\underbrace{\vec{\nabla} f}_{\text{gradient}} = \begin{bmatrix} \partial/\partial x\,(f) \\ \partial/\partial y\,(f) \\ \partial/\partial z\,(f) \end{bmatrix} \qquad f = 2x + 3y^2 - z^3$$

$$= \begin{bmatrix} \frac{\partial}{\partial x}\,(2x + 3y^2 - z^3) \\ \frac{\partial}{\partial y}\,(\qquad) \\ \partial/\partial z\,(\qquad) \end{bmatrix}$$

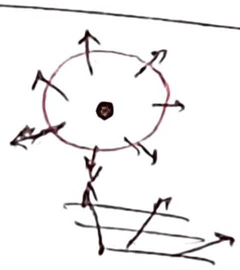
$$\underbrace{\nabla(f(x,y,z))}_{\text{gradient}} = \begin{bmatrix} 2 \\ 6y \\ -3z^3 \end{bmatrix}$$

---

note

$$\underbrace{\vec{\nabla} \cdot \vec{v}}_{} = \text{scalar}$$

divergence of $\vec{v}$



cross product

$$\underbrace{\vec{\nabla} \times \vec{u}}_{} = \text{vector}$$

curl of $\vec{u}$

magnetic monopoles ?

---

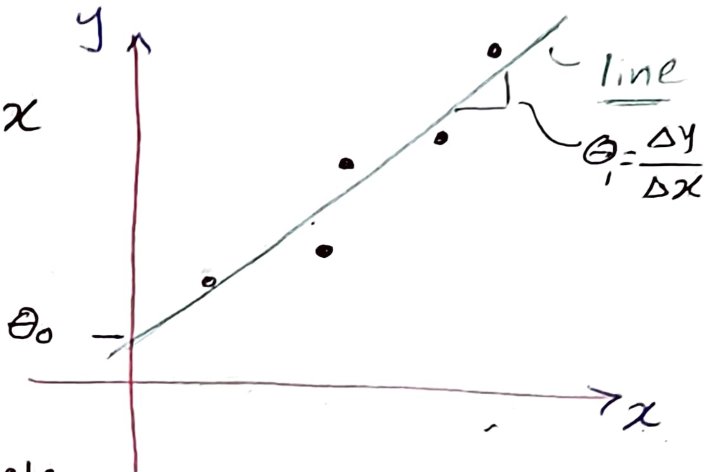
are they used in ML ?

→ Linear Regression (two parameters)

$f(x;\theta_0,\theta_1) \equiv h_\theta(x) = \hat{y} = \theta_0 + \theta_1 x$

↑ hypothesis — model

$\hat{y}$: predicted y



→ loss ; $\ell(\cdot)$; for a ~~certain~~ single example (1-data points)

→ cost ; $\underbrace{\sum_{i=1}^{m} \ell(\cdot)}_{\text{sum of losses}} = J(\theta)$?

objective : minimize $J(\theta)$ over θ_0, θ_1 ~?

→ let ; MSE

m-data points

m-examples

$$J(\theta_0,\theta_1) = \frac{1}{2m} \underbrace{\sum_{i=1}^{m} (\hat{y}_i - y_i)^2}_{RSS}$$

?

$$\equiv \frac{1}{2m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$$

$$\frac{1}{m}\sqrt{\sum_{i=1}^{m} (y_i - \hat{y}_i)^2} = (\|\vec{y} - \vec{\hat{y}}\|)^2/m$$

ex

$\hat{y} = \theta_0 + \theta_1 x$

$\vec{\hat{y}} = \theta_0 + \theta_1 \vec{x}$

$$\vec{\Theta} \underbrace{:=}_{\text{update}} \vec{\Theta} - \underbrace{\alpha}_{\text{learning rate}} \vec{\nabla}(J(\vec{\Theta}))$$

"or η"

$$\vec{\Theta} = \begin{bmatrix} \Theta_0 \\ \Theta_1 \\ \vdots \\ \Theta_n \end{bmatrix}$$

$$\vec{\Theta}^{(new)} = \vec{\Theta}^{(old)} - \alpha \vec{\nabla}(J(\Theta^{(old)}))$$

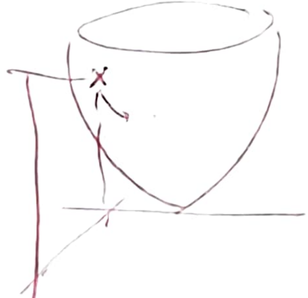
repeat until convergence

$$\vec{\Theta}(\underbrace{K+1}_{\text{time index / step index}}) = \vec{\Theta}(K) - \alpha \vec{\nabla}(J(\Theta(K)))$$

$$\Theta_0^{new} := \Theta_0^{old} - \alpha \frac{\partial}{\partial \Theta_0} J(\Theta^{old})$$

and

$$\Theta_1^{new} := \Theta_1^{old} - \alpha \frac{\partial}{\partial \Theta_1} J(\Theta^{old})$$



↳ all parameters $(\Theta_0, \Theta_1, \dots, \Theta_n)$ are updated simultaneously!

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$$

$$\frac{\partial J}{\partial \theta_0} = \frac{\partial}{\partial \theta_0}\left(\frac{1}{2m} \sum_{i=1}^{m} (y_i - (\theta_0 + \theta_1 x_i))^2\right)$$

$$\cancel{\frac{\partial J}{\partial \theta_1}} = \frac{1}{\cancel{2}m} \sum_{i=1}^{m} \cancel{2}\left(y_i - (\theta_0 + \theta_1 x_i)\right) \times (-1)$$

$$\frac{\partial J}{\partial \theta_0} = \frac{1}{m} \sum_{i=1}^{m} (\theta_0 + \theta_1 x_i - y_i)$$

if $J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (\hat{y}_i - y_i)^2$

$$\frac{\partial J}{\partial \theta_0} = \frac{1}{2m} \frac{\partial}{\partial \theta_0} \sum_{i=1}^{m} (\theta_0 + \theta_1 x_i - y_i)^2$$

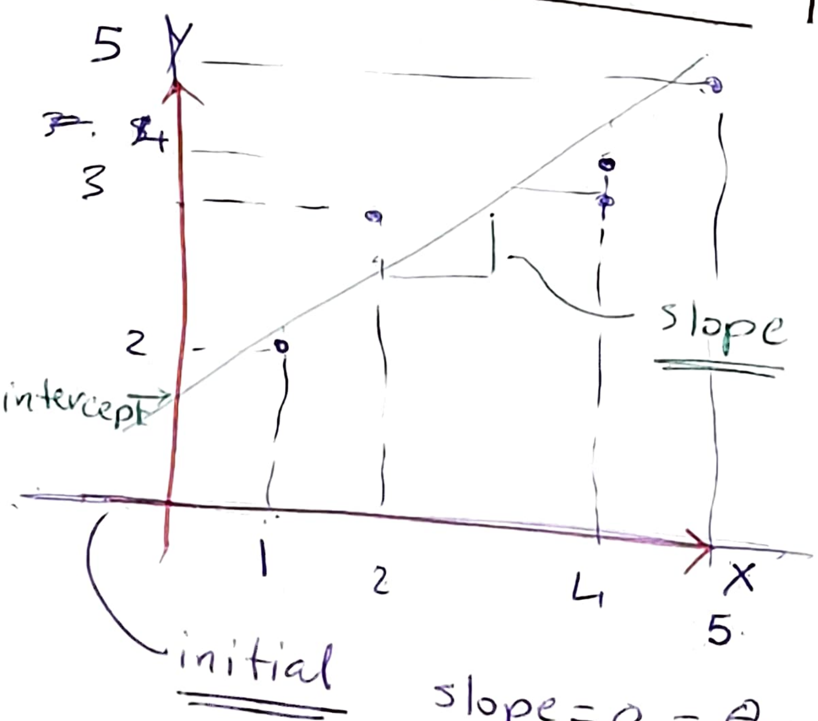
$$\frac{\partial J}{\partial \theta_0} = \frac{1}{m} (\theta_0 + \theta_1 x_i - y_i)(+1)$$

$$\frac{\partial J}{\partial \theta_1} = \frac{1}{2m}\sum_{i=1}^{m} \frac{\partial}{\partial \theta_1}\left(y_i - (\theta_0 + \theta_1 x_i)\right)^2$$

$$\frac{\partial J}{\partial \theta_1} = \frac{1}{m}\sum_{i=1}^{m} \left(y_i - (\theta_0 + \theta_1 x_i)\right)(-x_i)$$

$$\nabla J(\Theta) = \begin{bmatrix} \frac{\partial J(\theta)}{\partial \theta_0} \\ \frac{\partial J(\theta)}{\partial \theta_1} \end{bmatrix}$$

i	x	y
1	1	2
2	2	3
3	4	4
4	5	5



m = 4 : 4 - examples

slope = 0 = θ_1

intercept = 0 = θ_0

$$\vec{\hat{y}} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad \vec{y} = \begin{bmatrix} 2 \\ 3 \\ 4 \\ 5 \end{bmatrix} \Leftarrow \quad \vec{\Theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix} \quad \vec{\theta}_{(init)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\overrightarrow{error} = \begin{bmatrix} -2 \\ -3 \\ -4 \\ -5 \end{bmatrix}$$

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(y_i - \hat{y}_i \right)^2$$

$$= \frac{\cancel{2}\,1}{8}$$

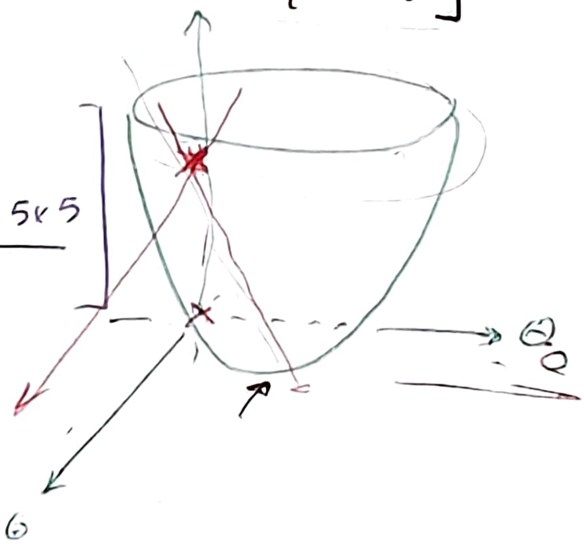
$$\nabla J(\theta) = \begin{bmatrix} \partial J / \partial \theta_0 \\ \partial J / \partial \theta_1 \end{bmatrix}$$

for old θ

$$\underset{\text{evaluated at old } \theta}{\nabla J(\theta)} = \begin{bmatrix} \frac{1}{m} \sum_{i=1}^{m} \left(\cancel{\theta_0}^{\,0} + \cancel{\theta_1}^{\,0} x_i - y_i \right) \\ \frac{1}{m} \sum_{i=1}^{m} \left(\cancel{\theta_0}^{\,0} + \cancel{\theta_1}^{\,0} x_i - y_i \right) x_i \end{bmatrix}$$

$$= \begin{bmatrix} -14/4 \\ -\frac{1 \times 2 + 2 \times 3 + 4 \times 4 + 5 \times 5}{4} \end{bmatrix}$$

$$= \begin{bmatrix} -14/4 \\ -49/4 \end{bmatrix}$$



$$\frac{\partial J}{\partial \theta_0} = -3.5 \qquad \frac{\partial J}{\partial \theta_1} = -12.25$$

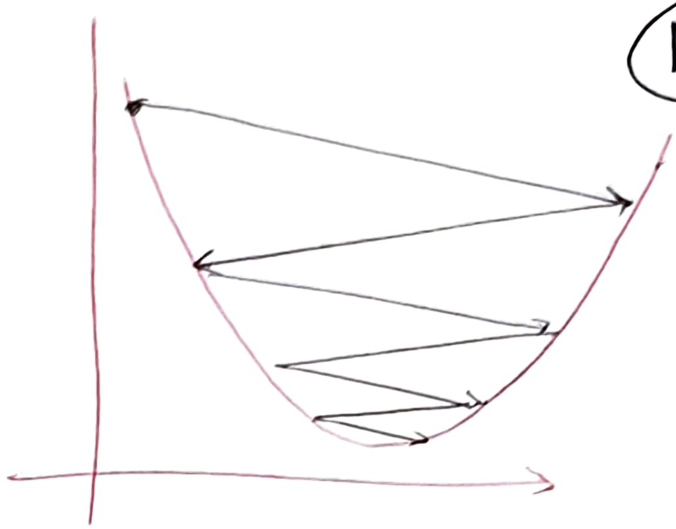
$\Rightarrow$ update $\vec{\theta}$; update simultaneously θ_0, θ_1

$$\theta_0^{(new)} = \theta_0^{(old)} - \alpha(-3.5) = 0.35$$

$$\theta_1^{(new)} = \theta_1^{(old)} - \alpha(-12.25) = 1.225$$

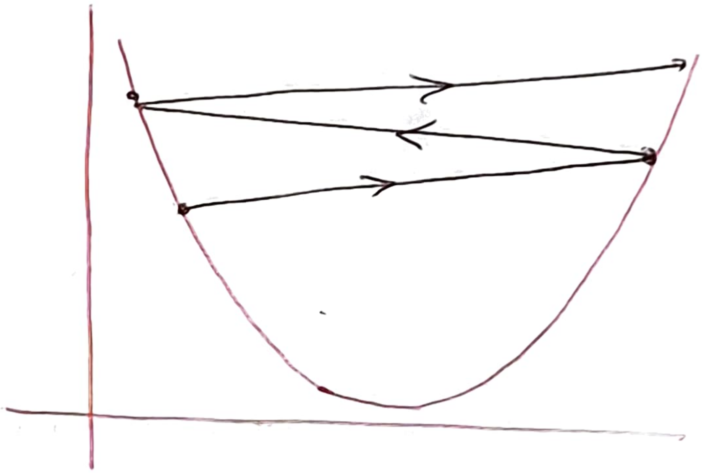
$\boxed{\text{let } \alpha = 0.1}$

learning rate α ↑

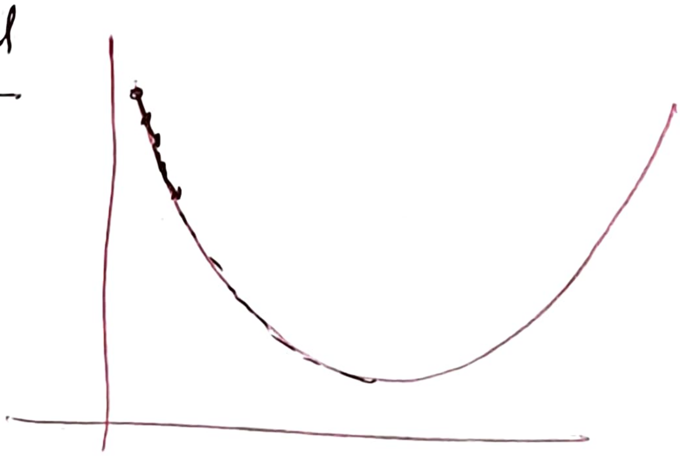


learing rate too big

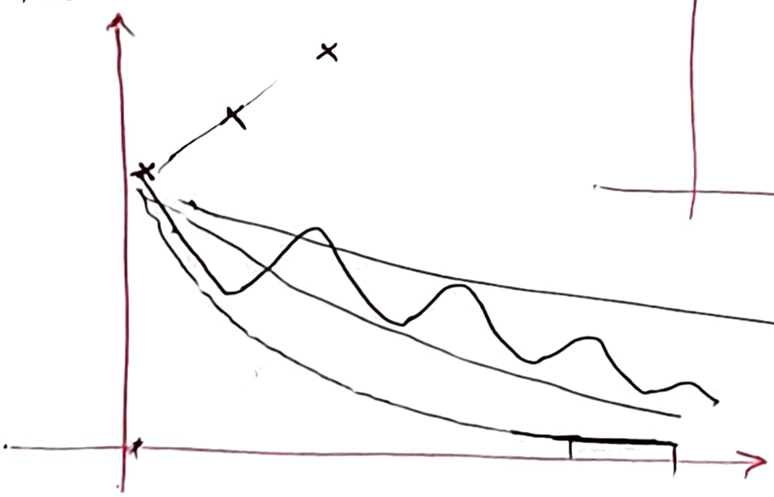
(diverge)



learning rate too small



error



dataset $m \times n$ matrix

m rows : # of samples

n columns : # of features

	x_0	x_1 (area)	x_2	x_3 ...	x_n	y (Price)
$i=1$	1	✓ 100	✓ 1	✓	...	
$i=2$	1	90	2			
$\vdots$	$\vdots$	150	3			
		120	2			
$i=m$	1					

linear regression :

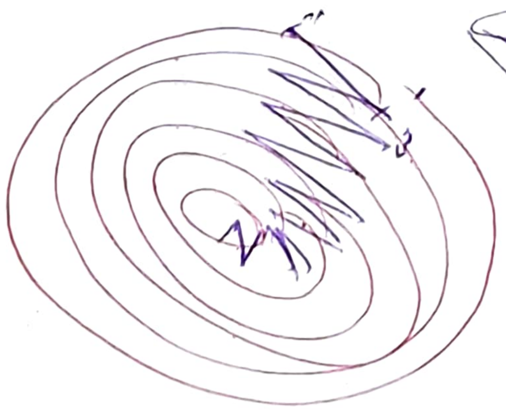
$$\hat{y} = h_{\Theta}(x_1, \dots, x_n) = \underbrace{\overset{\text{intercept}}{\Theta_0} + \Theta_1 x_1 + \Theta_2 x_2 + \dots + \Theta_n x_n}_{n+1 \text{ parameters}}$$

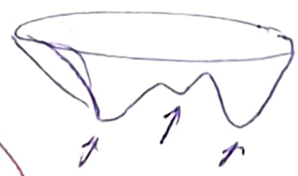
$$\hat{y}_i = \begin{bmatrix} \Theta_0 \\ \Theta_1 \\ \Theta_2 \\ \vdots \\ \Theta_n \end{bmatrix} \cdot \begin{bmatrix} 1 \\ x_{i_1} \\ x_{i_2} \\ \vdots \\ x_{i_n} \end{bmatrix}$$

add one extra feature

---

feature scaling





contour plot