3/1/2026 ①

NOFDS & ML session 3 Mansoura AI 46

① Feature Scaling

0 → 1

"measures of spread"

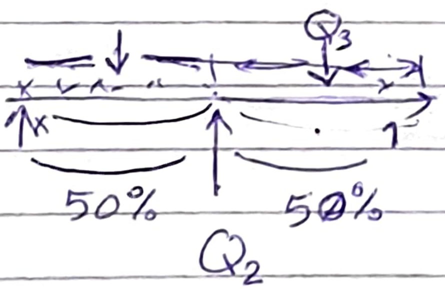
	x_1	x_2
	1	100
	2	120
	3	150

→ Range: $x_{max} - x_{min}$

→ Standard deviation $\sigma_x = \sqrt{\sigma_x^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}$

→ Interquartile range $IQR = Q_3 - Q_1$

Quartiles:



Q_0 : minimum value
Q_1 : 25% : 75%
Q_2 : median
Q_3 : 75% : 25%
Q_4 : maximum value

$Q_3 - Q_1 = \boxed{IQR}$ → Robust

122.5

100 100 | 110 | 120 125 | 125 | 130 140

Q_1 Q_1 Q_2 Q_3 Q_3

feature scaling

$$x_i \longrightarrow x_i'$$

$\Longrightarrow$ new scale (normalized)

$\sim 0 \to 1$

$\sim -1 \to 1$

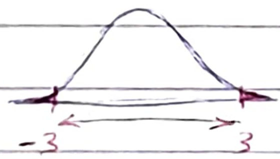
1) Range: min-max normalization

$$0 \le x_i' = \frac{x_i - x_{min}}{x_{max} - x_{min}} \le 1$$

1 2, 3, 3, 4, 5, ..., 5, 50

2) mean-normalization (standardization)

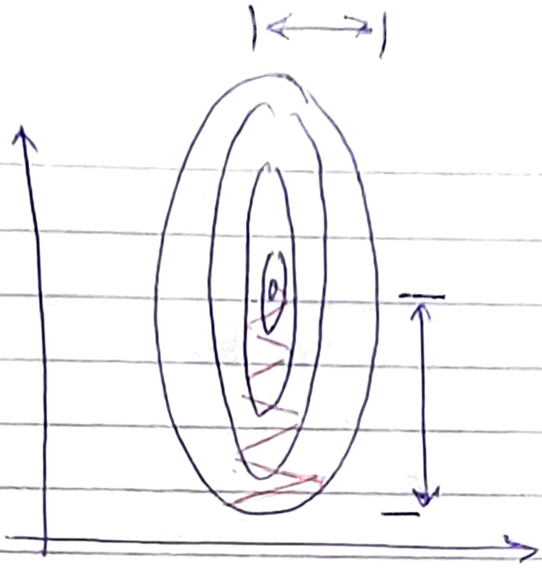
$$x_i' = \frac{x_i - \bar{x}}{\sigma_x}$$



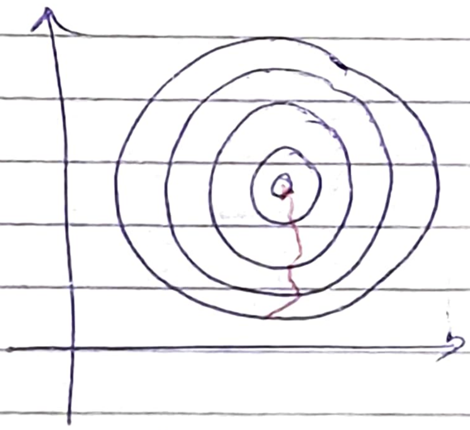
zero-mean, standard dev. = 1

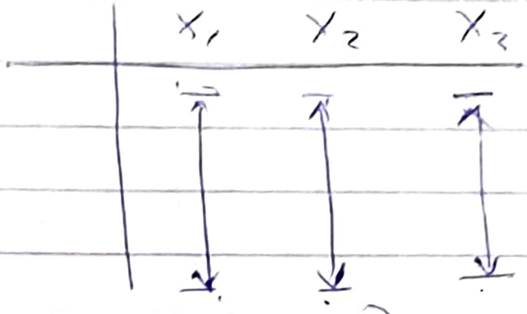
3) Robust feature scaling (Robust scaling)

$$x_i' = \frac{x_i - \text{median}}{IQR} = \frac{x_i - Q_2}{Q_3 - Q_1}$$



$\Downarrow$





$\tilde{\sigma}_{x_1}^2 \uparrow$ $\tilde{\sigma}_{x_2}^2 \downarrow$

max. f. variance

$$\boxed{\kappa \approx \frac{\tilde{\sigma}_{x_1}^2}{\tilde{\sigma}_{x_2}^2}}$$

min. feature variance

Kappa
↓
κ : ~~cond~~
→ condition number of optimization

computations of Convergence ~ $O(\kappa(\quad))$

Gradient descent Algorithms

→ momentum

→ Nesterov
→ $O(\sqrt{\kappa}(\quad))$

Adaptive GD → session 4

Review and further discussion

$\boxed{\text{Gradient}}$ $(\mathcal{F}) = \nabla J = \begin{bmatrix} \partial J/\partial\theta_1 \\ \partial J/\partial\theta_2 \\ \vdots \\ \partial J/\partial\theta_n \end{bmatrix}$

scalar-valued function

vector field

vector (after subsituting for values of θ_i)

$\boxed{\text{Jacobian}}$ of a vector valued function

$$\vec{f} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_m \end{bmatrix}$$

scalar valued functions

$$J(\vec{f}) = \begin{bmatrix} \partial f_1/\partial\theta_1 & \partial f_1/\partial\theta_2 & \cdots & \partial f_1/\partial\theta_n \\ \vdots & & & \\ \partial f_m/\partial\theta_1 & \partial f_m/\partial\theta_2 & \cdots & \partial f_m/\partial\theta_n \end{bmatrix}$$

$$J(\vec{f}) = \begin{bmatrix} \nabla^T(f_1) \\ \nabla^T(f_2) \\ \vdots \\ \nabla^T(f_m) \end{bmatrix}$$

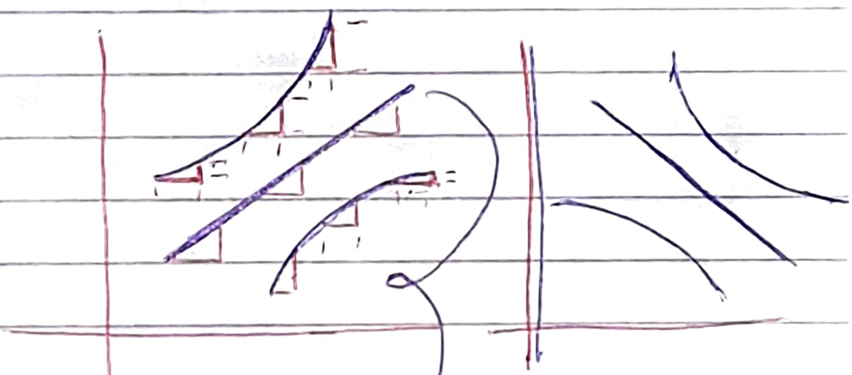
Hessian (J)

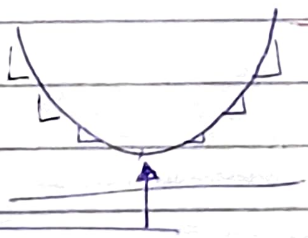
$\underline{J}$ ↳ scalar-valued function

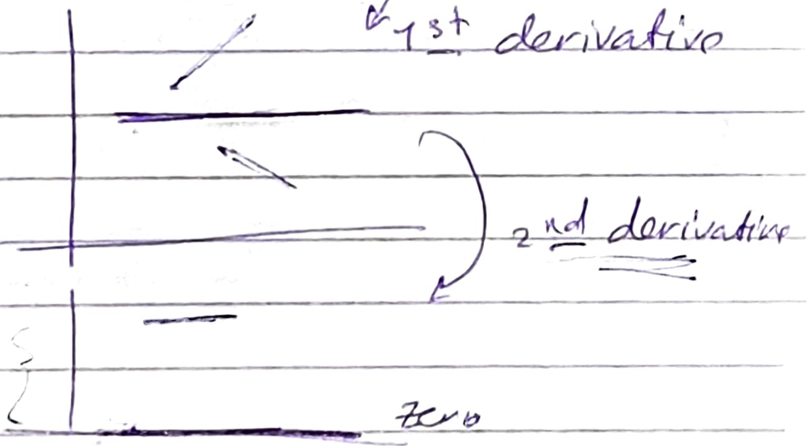
$$H(J) = \underline{\nabla^2(J)} \equiv \cancel{\nabla \nabla (J)}$$

$$H(J) = \begin{bmatrix} \frac{\partial^2 J}{\partial \theta_1^2} & \frac{\partial^2 J}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 J}{\partial \theta_1 \theta_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 J}{\partial \theta_n \partial \theta_1} & \frac{\partial^2 J}{\partial \theta_n \partial \theta_2} & \cdots & \frac{\partial^2 J}{\partial \theta_n^2} \end{bmatrix}$$

→ Convex
→ Concave
flat?
inflection?







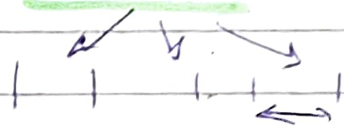
$$\boxed{\underline{K} = \frac{\lambda_{max}(H)}{\lambda_{min}(H)}}$$

↳ ~ 1

$O(\underline{K}(\quad)) -$

GD variants

Review Vanilla GD / Batch GD



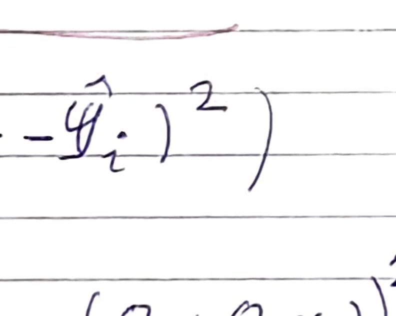
repeat until convergence

$$\begin{cases} \theta^{new} = \theta^{old} \\ \vec{\theta}^{(K+1)} = \vec{\theta}^{(K)} - \alpha \nabla J(\vec{\theta}) \end{cases}$$

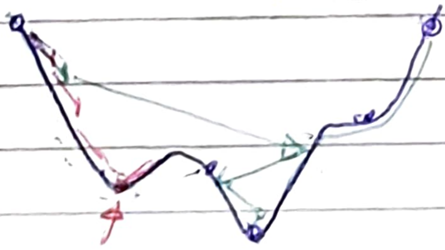
for MSE : $\nabla\left(\frac{1}{2m}\sum_{i=1}^{m}(y_i-\hat{y}_i)^2\right)$

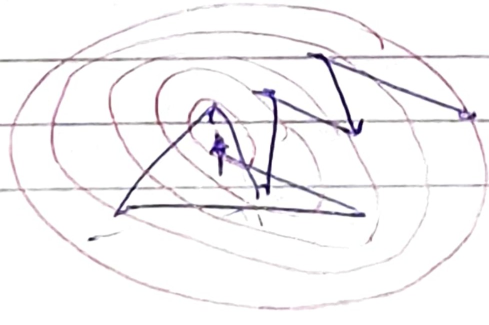
e.g., in Linear Regression $\nabla\left(\frac{1}{2m}\sum_{i=1}^{m}\left(y_i-(\theta_0+\theta_1 x_i)\right)^2\right)$

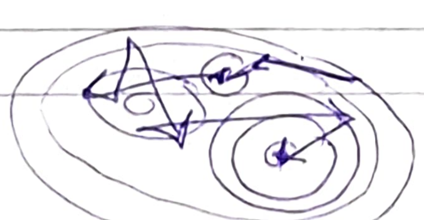
→ Stochastic GD "SGD"



→ $\boxed{\text{mini-batch GD}}$

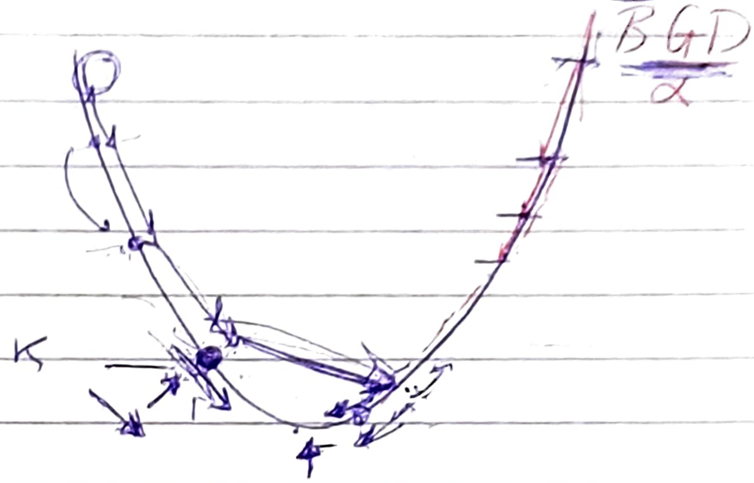






- Momentum

→ "heavy-ball"



$$~~\Theta^{(K+1)}~~ = \Theta^{(new)} = \Theta^{(K)} - \alpha \nabla J(\Theta^{(K)})$$

$$\Theta^{(K+1)} = \Theta^{(K)} + \boxed{\mu \underset{\text{current}}{\text{Velocity}} - \underset{\text{current}}{\text{gradient}}}$$

$$\boxed{V^{(K+1)} = \underset{\text{OR } \beta}{\mu} V^{(K)} - \underset{\text{OR } \eta}{\alpha} \nabla J(\Theta^{(K)})}$$

$(~~\mu - 1~~)$
$(1-\mu)$
$(1-\beta)$

$K=0 \qquad v^{(0)} = 0$

$K=1 \qquad v^{(1)} = \mu \times 0 - \alpha \text{ gradient}(\Theta^{(0)})$

$K=2 \qquad v^{(2)} = \mu(v^{(1)}) - \alpha \text{ gradient}(\Theta^{(1)})$

$\qquad\qquad = \boxed{\mu(-\alpha \text{ grad}(\Theta^{(0)})) - \alpha \text{ grad}(\Theta^{(1)})}$

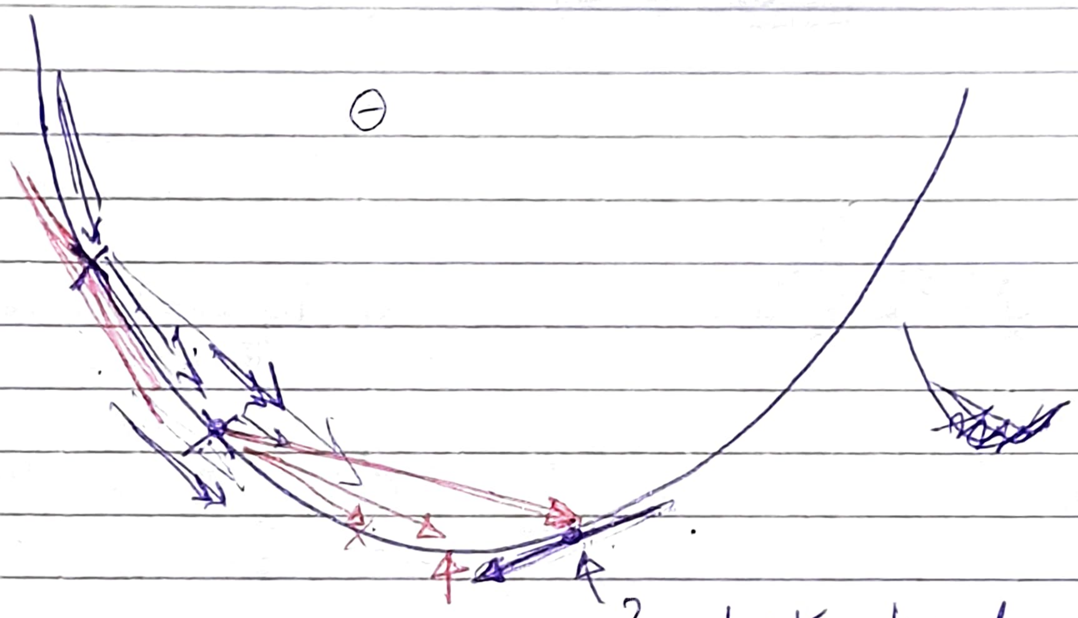
$K=3 \qquad v^{(3)} = \mu(v^{(2)}) - \alpha \text{ grad}(\Theta^{(2)})$

$\qquad\qquad = \boxed{\mu(\mu(-\alpha \text{ grad}(\Theta^{0})) - \alpha \text{ grad}(\Theta^{(1)}))}$
$\qquad\qquad - \alpha \text{ grad } \Theta^{(2)}$

↑ EWA
EWMA

Nesterov method

$\Rightarrow$ look ahead !



gradient at $\Theta|_{\text{lookahead}}$

$v^{(k)}$

β optimally ~ 0.9

$0.8 < \beta < 0.999$

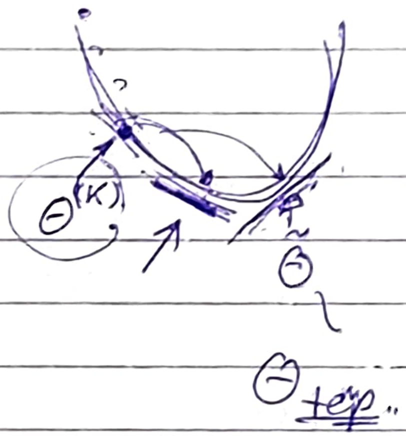
Nesterov

- look-ahead parameter
- temp. value of Θ ; $\tilde{\Theta}^{(K)}$

$$\boxed{\tilde{\Theta}^{(K)} = \Theta^{(K)} + \mu v^{(K)}}$$

extrapolated Value of θ

$\cancel{\nabla J(}$



$$v^{(K+1)} = v^{(K)} - \alpha \nabla J(\tilde{\Theta}^{(K)})$$

$$\Theta^{(K+1)} = \Theta^{(K)} + v^{(K+1)}$$