27/12/2025

Numerical Optimization for ML

Learning outcomes

- understanding Convexity
- " optimization
- " Gradient Descent algorithm

\# • Vanilla GD
  • Variants GD

- momentum-based Algorithms
  • ADAM Adaptive Momentum
  • Adam
- Newton's method
- be able to apply these algorithm.

Session 1

Machine Learning

model ~~params~~ Parameters.

$$y = \theta_0 + \theta_1 x$$

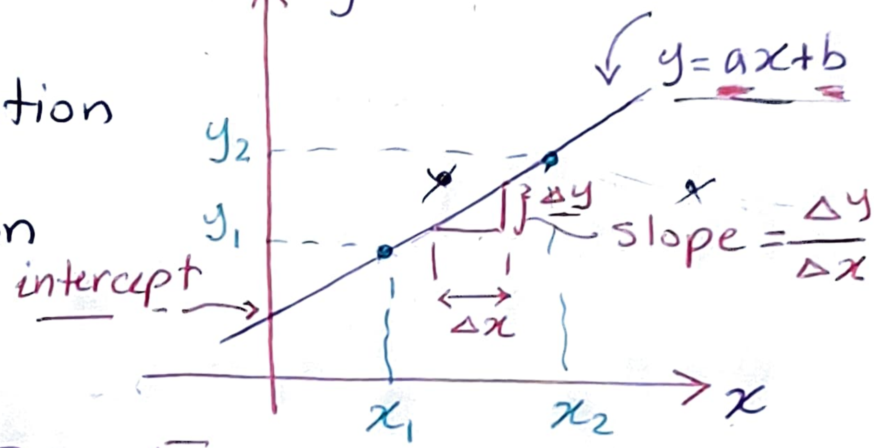
model

e.g. Linear Regression

Case #1 exact solution

→ Analytic solution

↳ Cannot be generalized



$$y_1 = a x_1 + \boxed{b}$$

$$y_2 = a x_2 + \boxed{b}$$ → 2 equations, 2 unknowns

Model Parameters

⇒ unique, exact solution

" Analytic "

data matrix

i	x	y
1	x_1	y_1
2	x_2	y_2

features

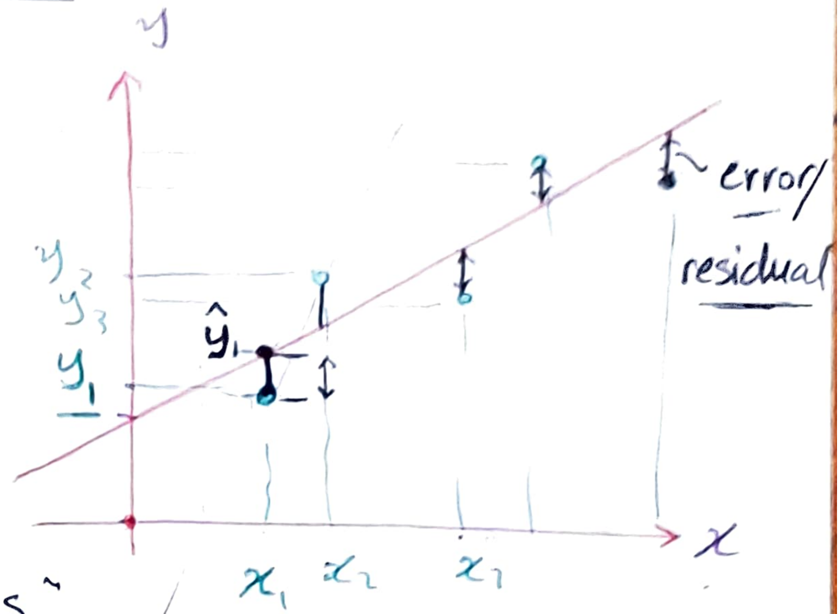
i	x_1	x_2	...	x_n	y
1					
2					
3					
⋮					
m					

approximate solution

- Linear model

$$y = \theta_0 + \theta_1 x$$

$$y_i = \theta_0 \times 1 + \theta_1 x_i$$



e.g., "least squares"

example, using pseudo inverse. $m > n$

"overdetermined system"

$\boxed{\text{Algebraic solution}}$ can be used to find approximate solution,

$$\vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = \left[\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \right] \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}$$

$$\vec{y} = X_{m\times 2}\, \vec{\theta}$$

$$X^T \vec{y} = X^T_{2\times m} X_{m\times 2}\, \vec{\theta}$$

$$(X^T X)^{-1} X^T \vec{y} = \cancel{(X^T X)^{-1} (X^T X)}\, \vec{\theta}$$

$$\underbrace{(X^T X)^{-1}(X^T X)}_{I}$$

$$\vec{\theta} = (X^T X^{-1}) X^T \vec{y}$$

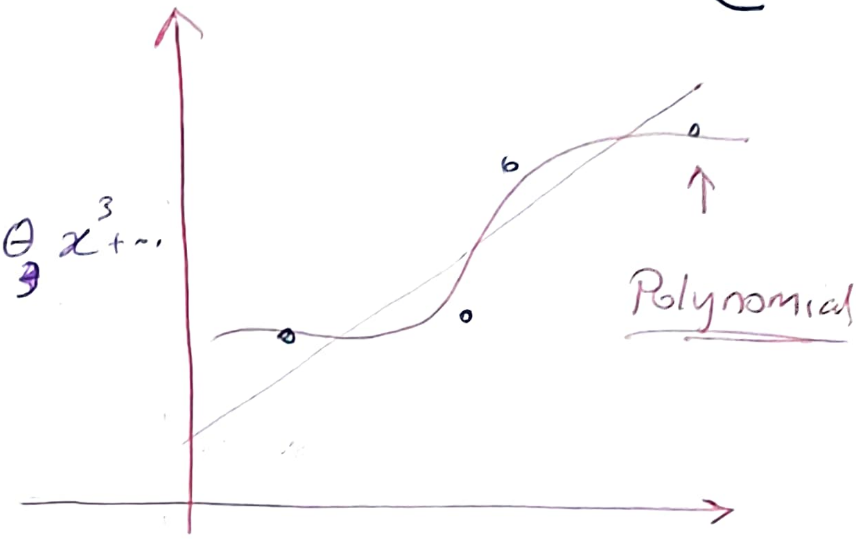
$$\overbrace{\hphantom{xxxxxxxx}}^{n}$$

$$\begin{array}{c|cc} i & & \\ \hline 1 & x_1 & y_1 \\ 2 & x_2 & y_2 \\ \vdots & \vdots & \vdots \\ m & x_m & y_m \\ & X_{m\times 2} & \vec{y} \end{array}$$

polynomial model

$$y : \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \dots$$

$\vec{y} =$ θ_0 0



Least norm method;

"Algebraic solution"

$n > m$

underdetermined system.

$$\vec{y} = X\vec{\theta}$$

→ infinite number of solutions

$$\boxed{\theta = X^T(XX^T)^{-1}y}$$

$X_{m \times n}$

data matrix

m rows: # of data examples

n: columns: # of features

$X_{m \times n}$

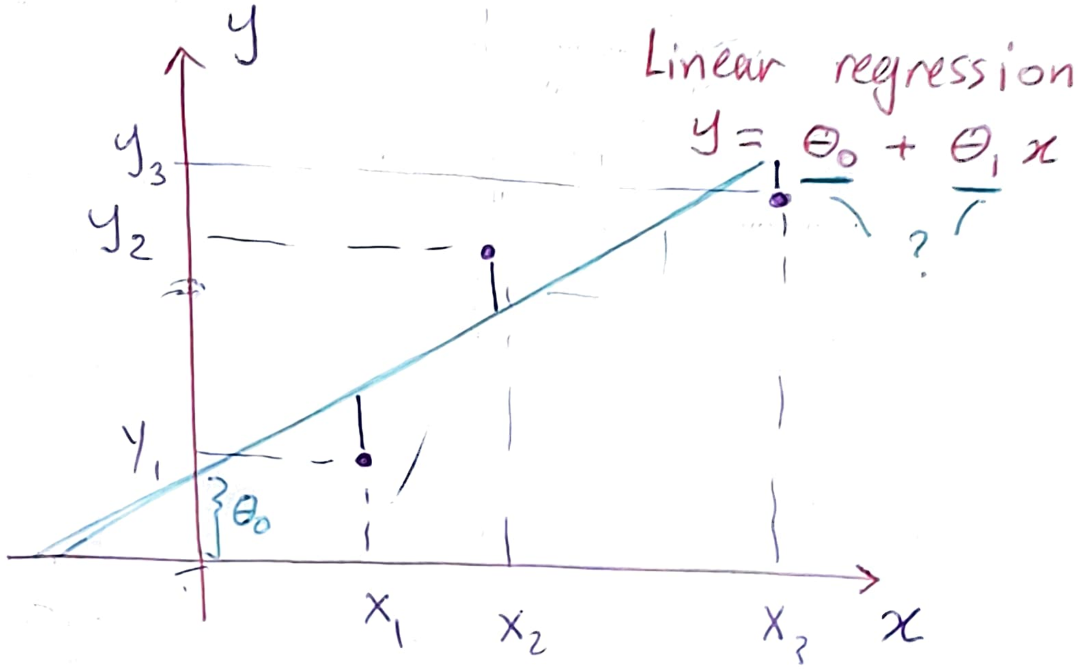
e.g. 2

$$Y = \theta_0 \times 1 + \theta_1 \times x_1 + \dots$$

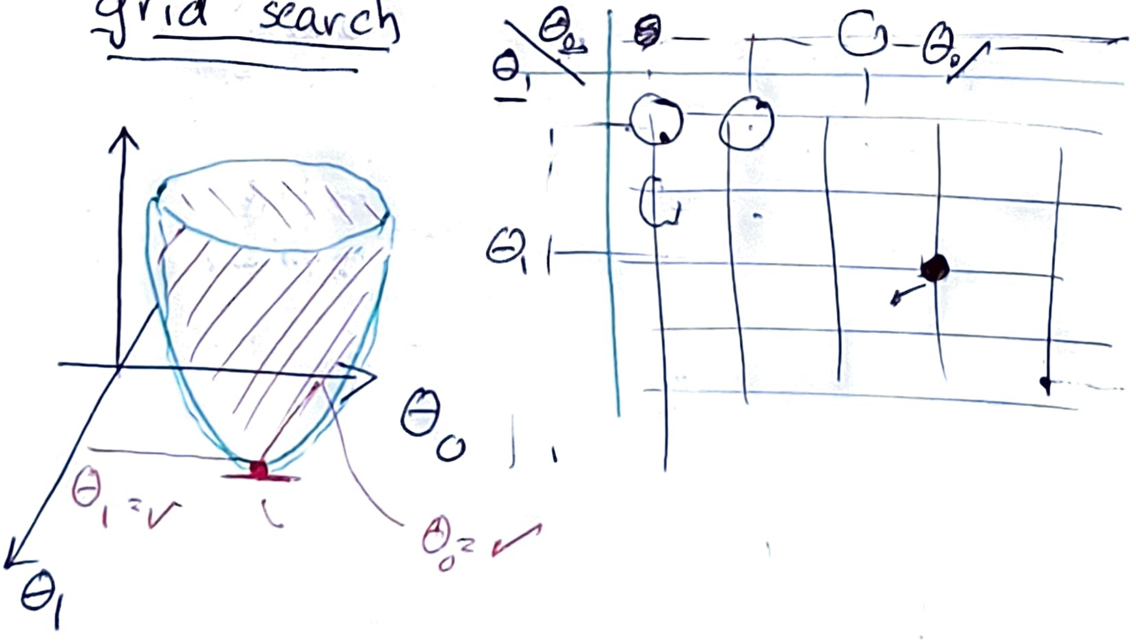
$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1n} \\ 1 & x_{21} & \dots & x_{2n} \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{bmatrix}$$

numerical solutions

"optimization" → minimization
→ maximization



grid search



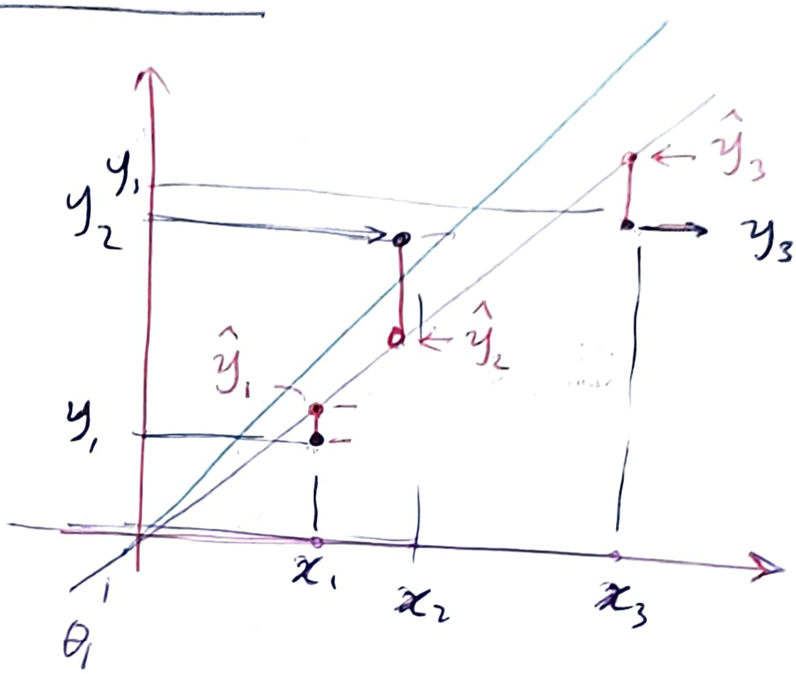
Linear regression problem

1- parameter

intercept = 0

$\theta_1 = ?$

$\hat{y} = \theta_1 x$



Error ; residuals $\quad y_i - \hat{y}_i = y_i - \theta_1 x_i$

Sum. of Absolute values of Residuals

| Sum. of squared values of Residuals. |

sum of squared error

Sum of Squared Residuals $\equiv$ ~~M~~ SSE

RSS $\quad$ Residual squares sum.

$$\sum_{i=1}^{m} (y_i - \hat{y}_i)^2 = \underbrace{\sum_{i=1}^{m} (\underbrace{y_i - \theta_1 x_i}_{\text{loss}})^2}_{\text{cost function}}$$

optimization problem

$$\underset{\theta_1}{\text{minmize}} \, (RSS) = \underbrace{\min_{\theta_1} \sum_{i=1}^{m} (y_i - \theta_1 x_i)^2}_{\text{objective fun}}$$

- For linear regression model

$$\hat{y}_i = \Theta_0 + \Theta_1 x_i$$

find optimum values of parameters Θ

the minimize prediction error $|(\vec{y}_i - \hat{\vec{y}}_i)|$

→ steps involve :

~~assume~~ select certain model (linear regression)

→ select certain error function (SSR, MAE, MSE, SAT)

1) Parameter initialization (assume $\Theta = 0$ or any other initial value.)

2) Predict output $\hat{y}$ using model

3) Evaluate prediction error $|y_i - \hat{y}_i|$

4) Find direction & magnitude of changing of model parameters (Θ)
-ve, or +ve

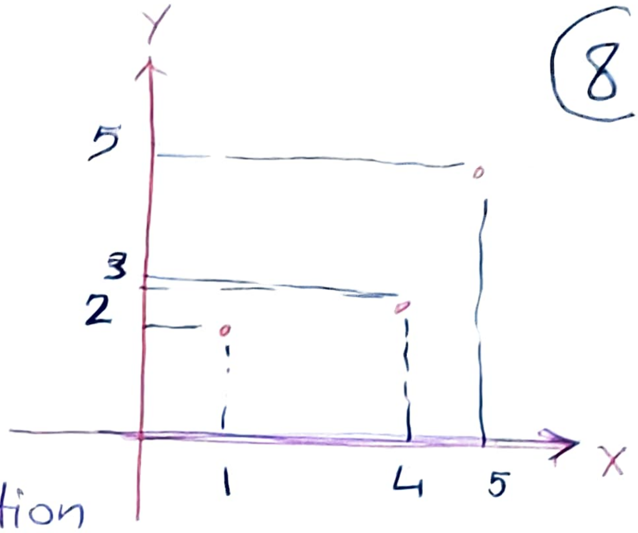
5) update model parameters

6) go to step 2 (repeat untill convergence)

- linear regression

→ $\hat{y} = \theta_1 x$

→ using SSR



1) parameter initialization

let $\theta_1 = \text{zero}$ (or any initial value(s))

$\hat{y} = \text{zero} \times x = 0$

2) Predict output

$\theta_1 = 0$

i	x	y	$\hat{y} = \theta_1 x$	error
1	1	2	0	2
2	4	3	0	3
3	5	5	0	5

3) error $= |y_i - \hat{y}_i|$

$= |\hat{y}_i - y_i|$

"SSR"

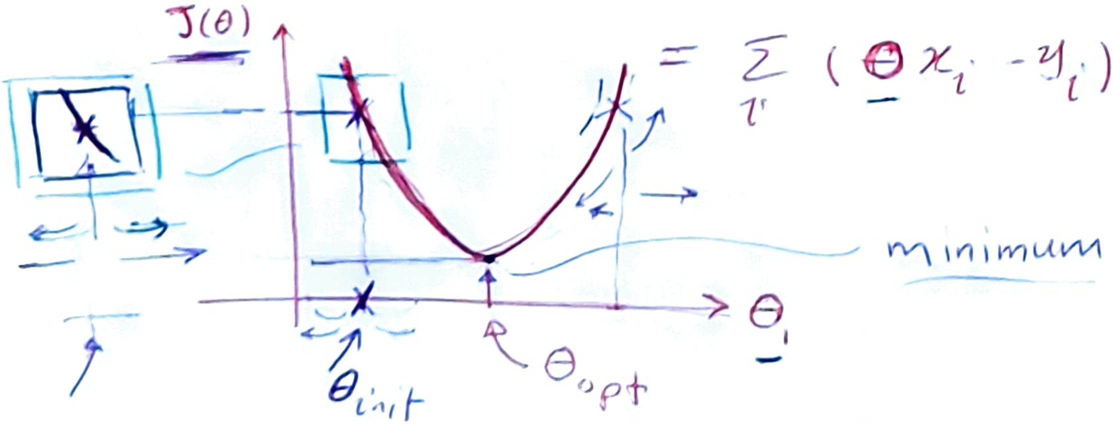
4) $J(\theta) = \sum_{i=1}^{m} (|\hat{y}_i - y_i|)^2 = 2^2 + 3^2 + 5^2$

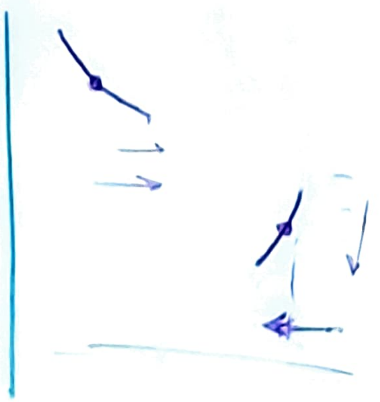
5) update θ; $\theta_{new} = \theta_{old} - \text{const.} \cdot \text{gradient}$

learning rate

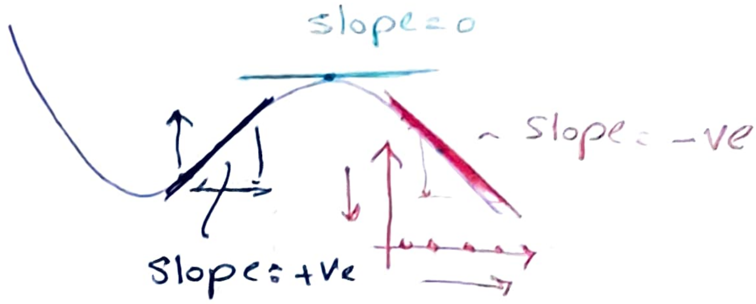
SSR : $J(\theta) = \sum_i (\hat{y}_i - y_i)^2$

$= \sum_i (\theta x_i - y_i)^2$





slope ≡ derivative of curve ≡ grad.



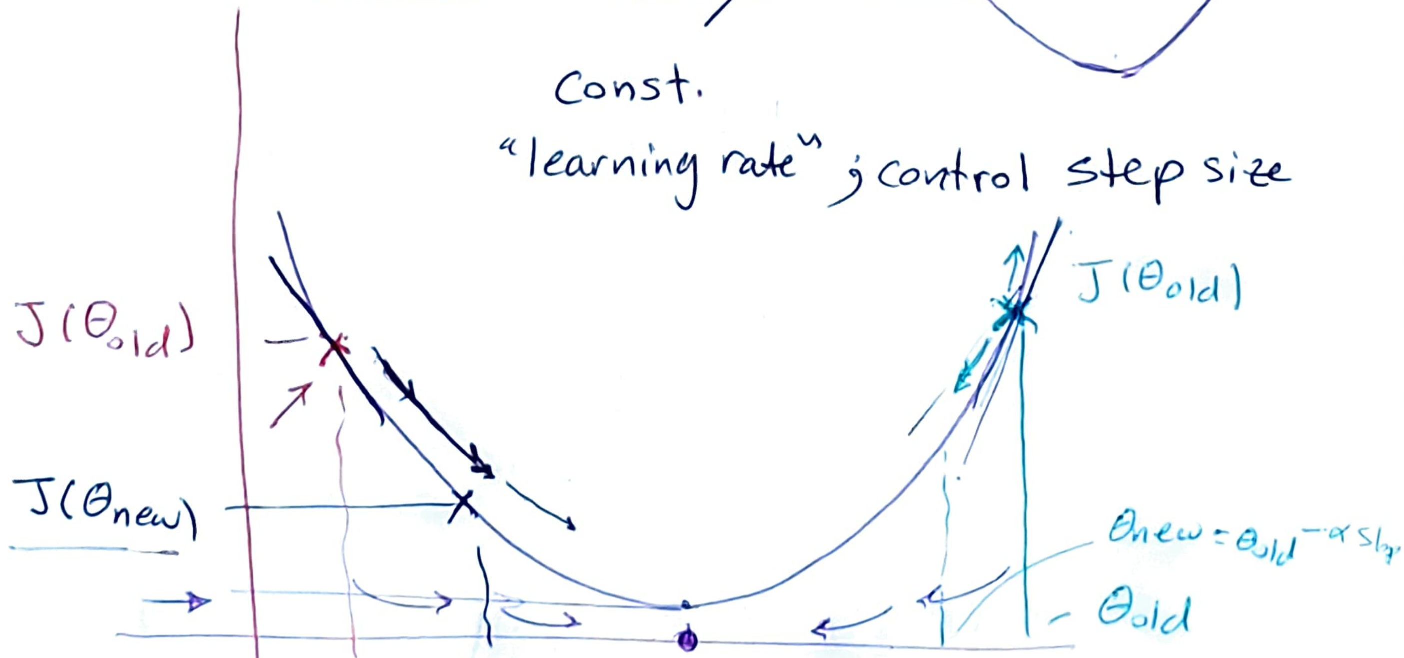
update θ : θ_{new}

$J(\theta_{new}) < J(\theta_{old})$

update θ:

$$\boxed{\theta_{new} = \theta_{old} - \alpha \, grad(J(\theta))}$$

Const.
"learning rate" ; control step size



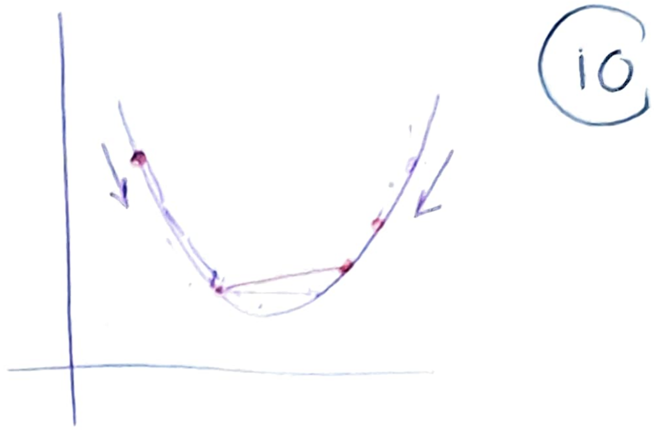
$\theta_{old} < \theta_{new} = -\alpha \, slope(J(\theta_{old}))$

grad

→ Convex function

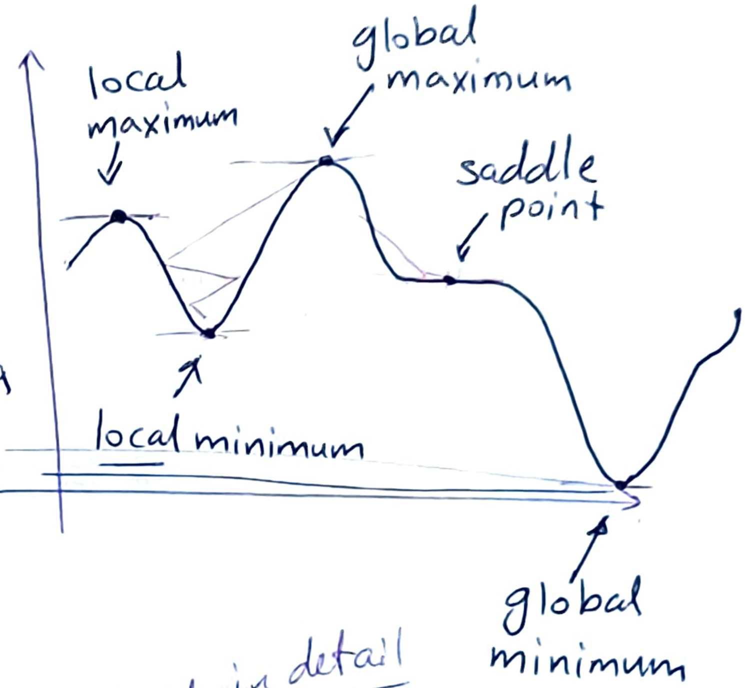
global minimum

"local minimum"



→ non convex

- minimum ; Plural : minima
- maximum ; plural : maxima



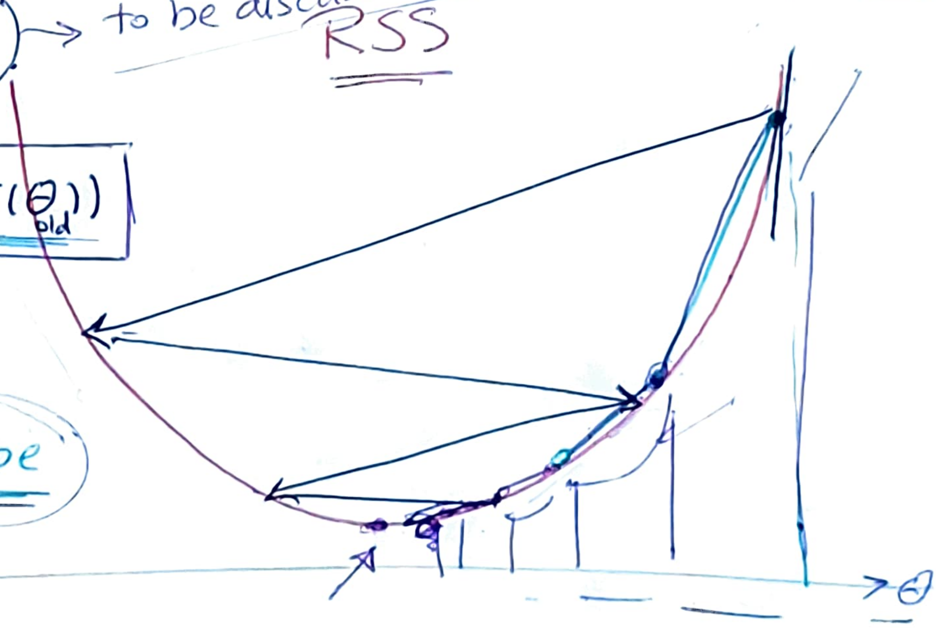
step ?

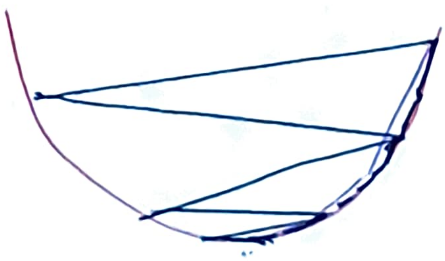
learning rate → to be discussed in detail

$$\Theta_{new} = \Theta_{old} - \boxed{\alpha \, grad(J(\Theta_{old}))}$$

step

slope

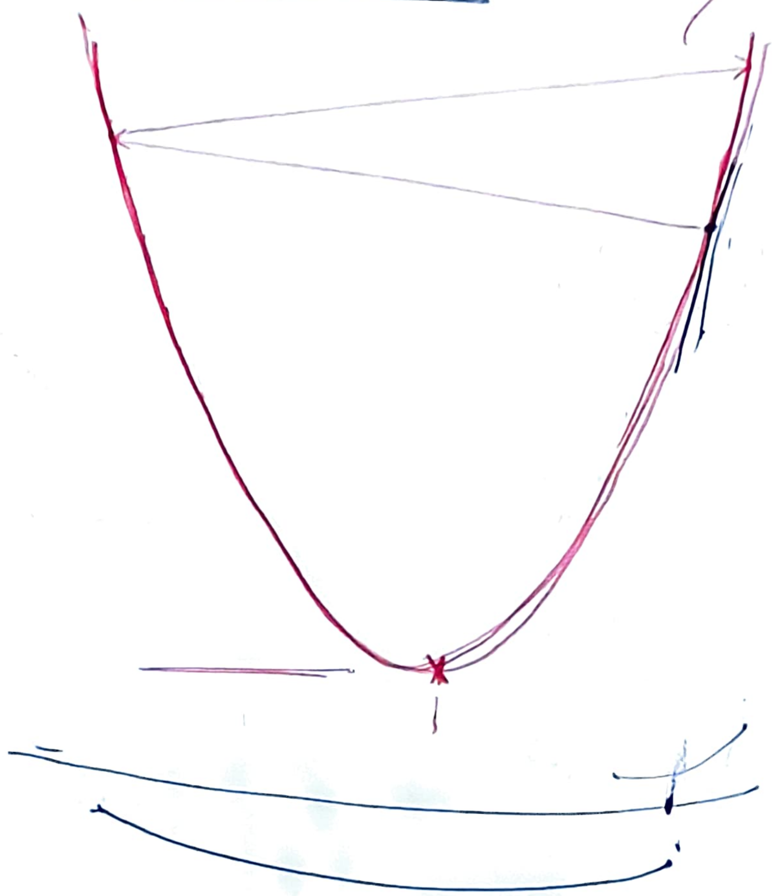




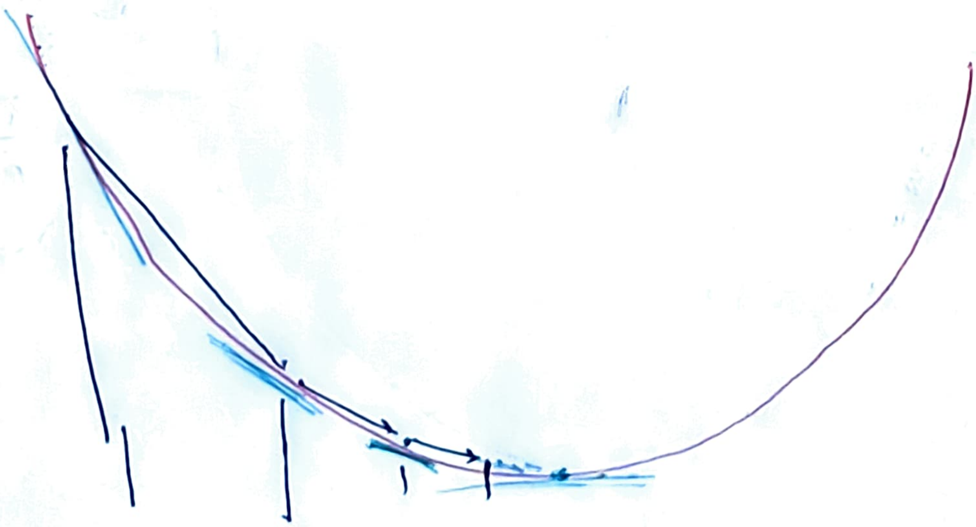
too big learning rate

divergence

not convergence



α : too large



SSR

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

i	X	Y	$\hat{y}$
1	x_1	y_1	$\hat{y}_1$
2	x_2	y_2	$\hat{y}_2$
3			
$\vdots$	$\vdots$	$\vdots$	$\vdots$
m	x_m	y_m	$\hat{y}_m$

$$\|\overrightarrow{error}\| = \|\overrightarrow{Residual}\| = \|\vec{\hat{y}} - \vec{y}\|$$

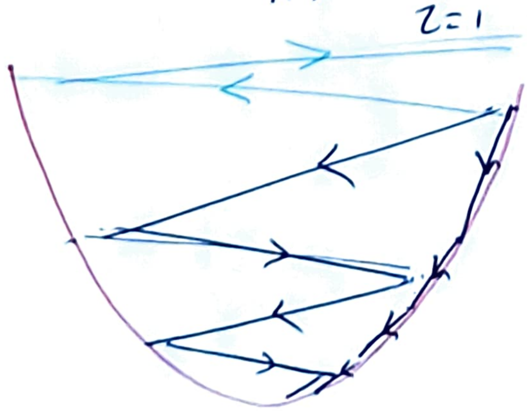
---

l_2-norm of a vector $\Rightarrow$ L_2-minimization

$$\|\overrightarrow{error}\| = \left\| \begin{bmatrix} \hat{y}_1 - y_1 \\ \hat{y}_2 - y_2 \\ \vdots \\ \hat{y}_m - y_m \end{bmatrix} \right\| = \sqrt{\sum_{i=1}^{m} (\hat{y}_i - y_i)^2}$$

Mean squared error **MSE**

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



l_1 - minimization

l_1 - norm of a vector $\vec{e} = \vec{\hat{y}} - \vec{y}$

$= \sum_{i=K} |e_K|$

$\vec{v} = \begin{bmatrix} 1 \\ -1 \\ -5 \\ 2 \\ 4 \end{bmatrix}$

ex l_1-norm $(\vec{v}) = |1| + |-1| + |-5| + |2| + |4|$

$= 13$

MAE : mean absolute error

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

$$= \sum_{i=1}^{m} \left| (\theta x_i - y_i) \right|$$