Recap

Ensemble

Bias Variance   [1%] → [DS]

Very conceptual

Bagging

→ Random forest

Boosting

Gradient Boosting   → How
                     → What

Why? →

xgboost

Today's class in one line

Gradient Boosting is performing **Gradient Descent** in **Function Space**

How Gradient Boosting Works?
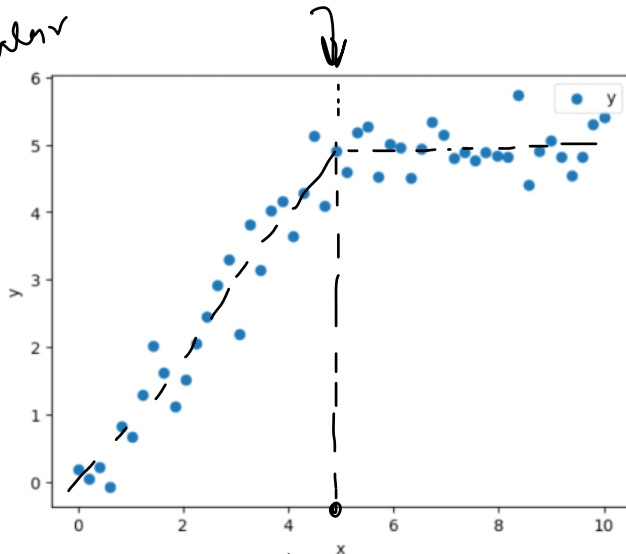
salary

what?

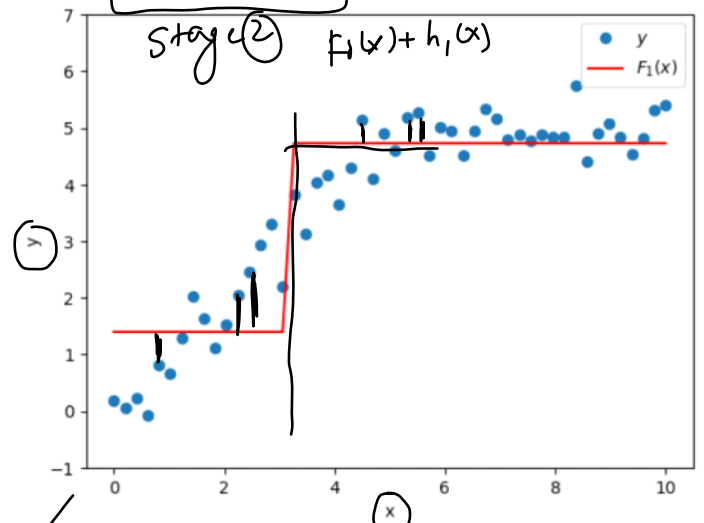| X	Y	model
exp	Salary	$X \to Y$
7	32	
3	18	
1	5	

$$y = f(x)$$

exp   Regression

mean salary

Stage 1

$F_0(x)$

sal

$F_0(x)$

exp

decision tree

residuals

y

$0 \, exp$

$h_1(x)$

$F_0(x) + h_1(x)$

Stage 2 $F_1(x) + h_1(x)$

$y - F_0(x)$ $h_1(x)$

residuals

y

x

$F_2(x) = F_0(x) + h_1(x) + h_2(x)$

$y - F_1(x)$ $h_2(x)$

y $F_2(x)$

residual

$y - F_1(x)$

$h_2(x)$

X

y

$F_2(x)$

$F_3(x) = F_0(x) + h_1(x) + h_2(x) + h_3(x)$

m = 3

$h_3(x)$

$y - F_2(x)$

$h_3(x)$

residuals

m = 3

y

$F_3(x)$

m = 4

$y - F_3(x)$

$h_4(x)$

residuals

m = 4

y

$F_4(x)$

m = 5

$y - F_4(x)$

$h_5(x)$

duals

m = 5

y

$F_5(x)$

m = 5

Legend: $y - F_4(x)$ (diamonds), $h_5(x)$ (dashed)

m = 5

Legend: y, $F_5(x)$

m = 6

Legend: $y - F_5(x)$ (diamonds), $h_6(x)$ (dashed)

m = 6
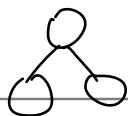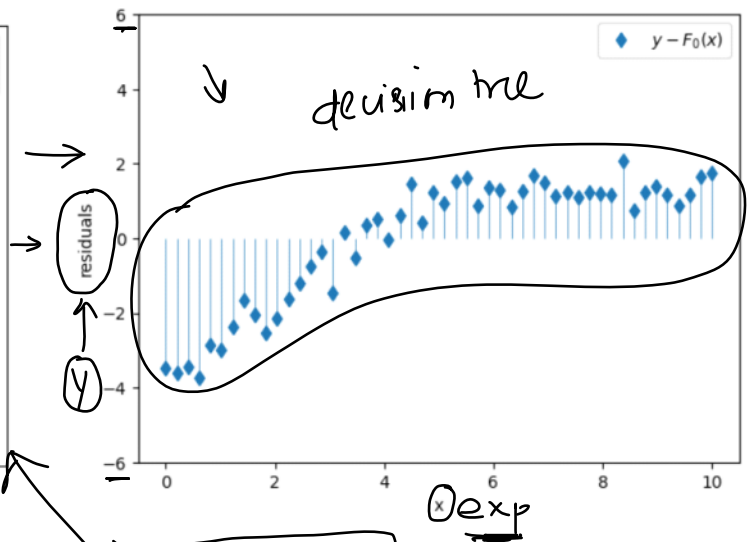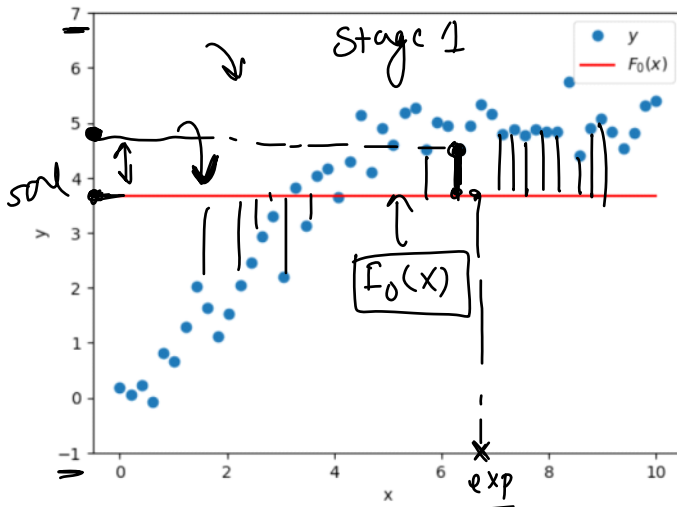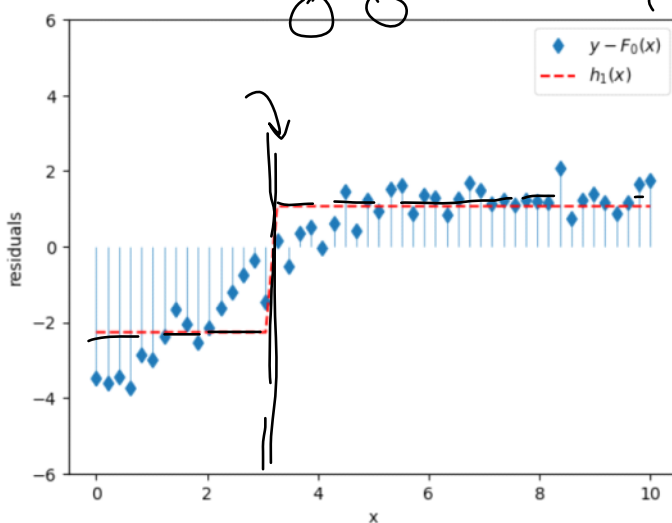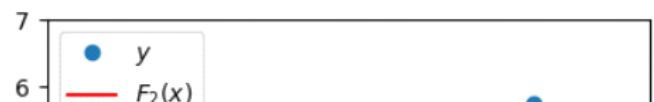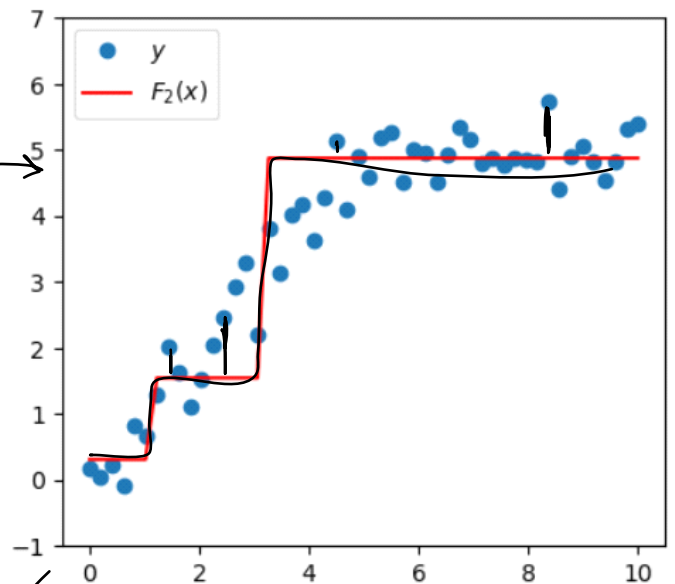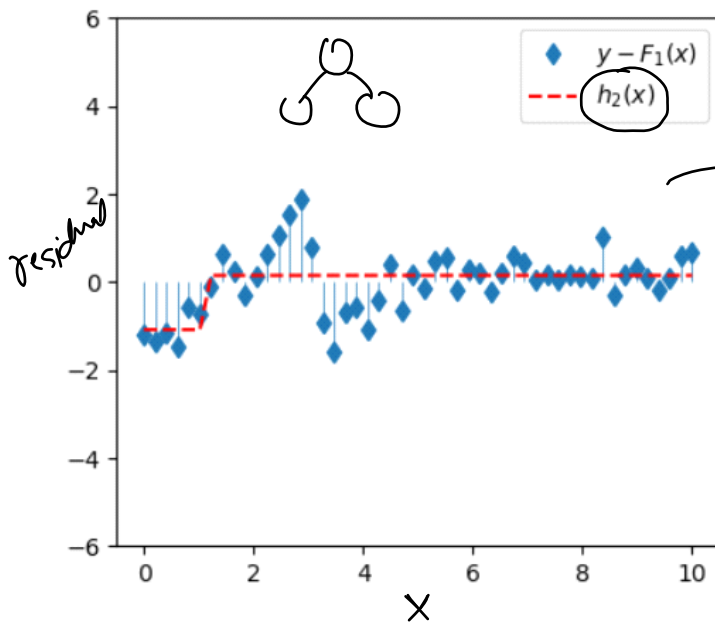
Legend: y, $F_6(x)$
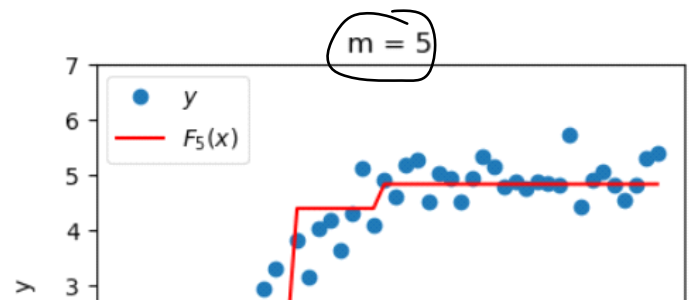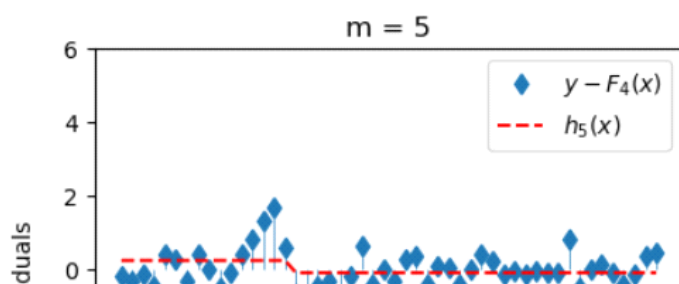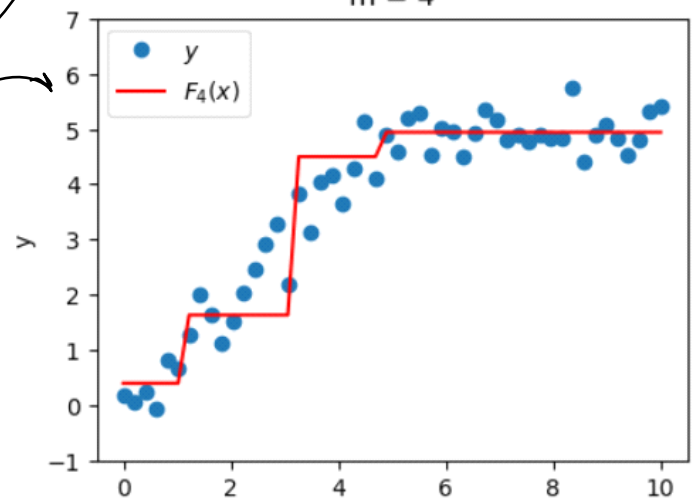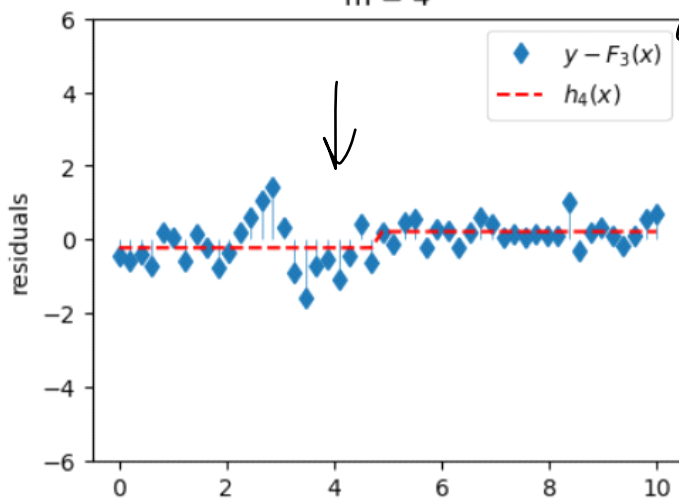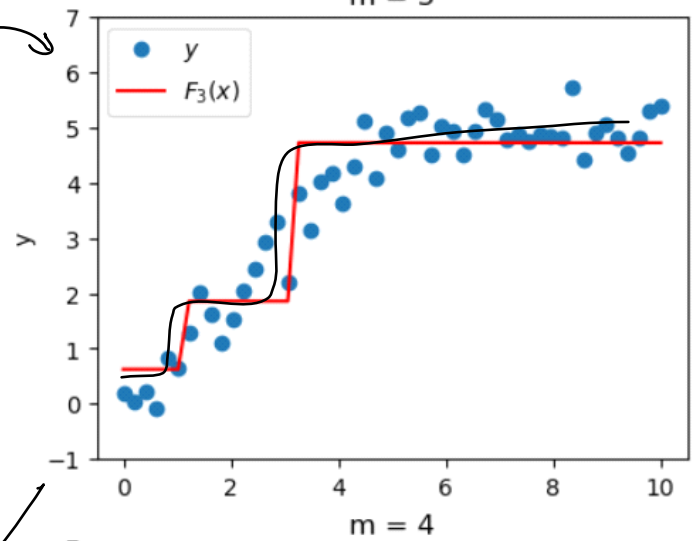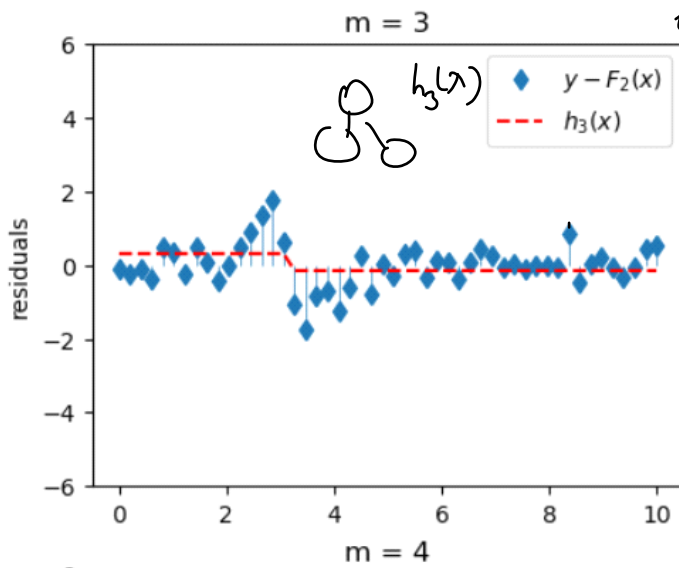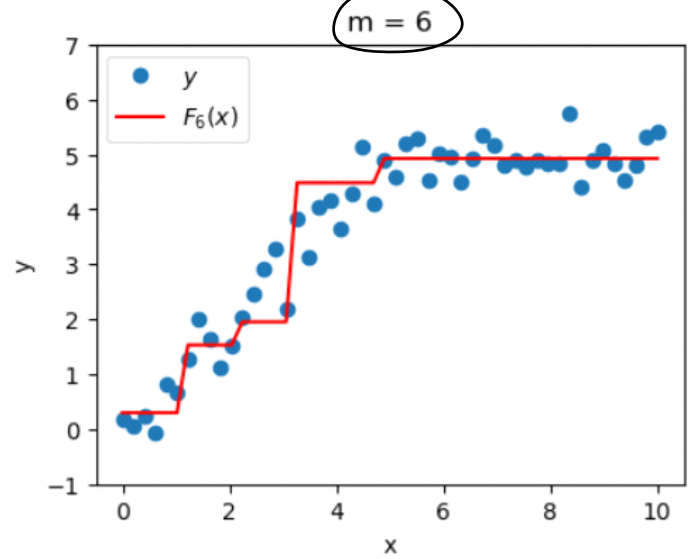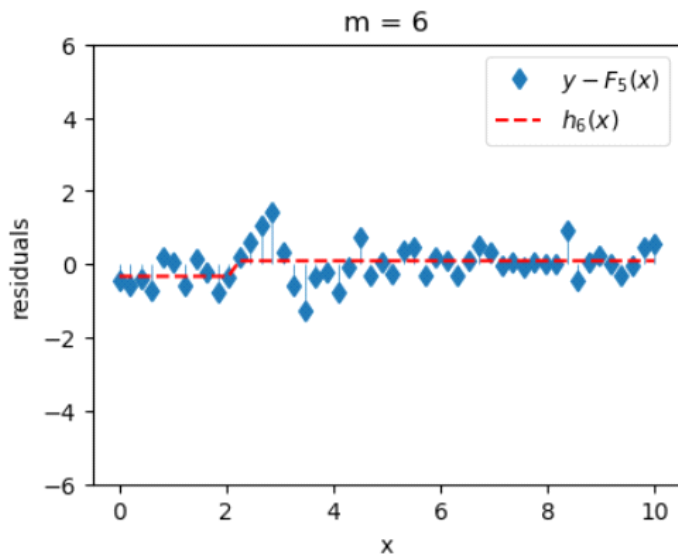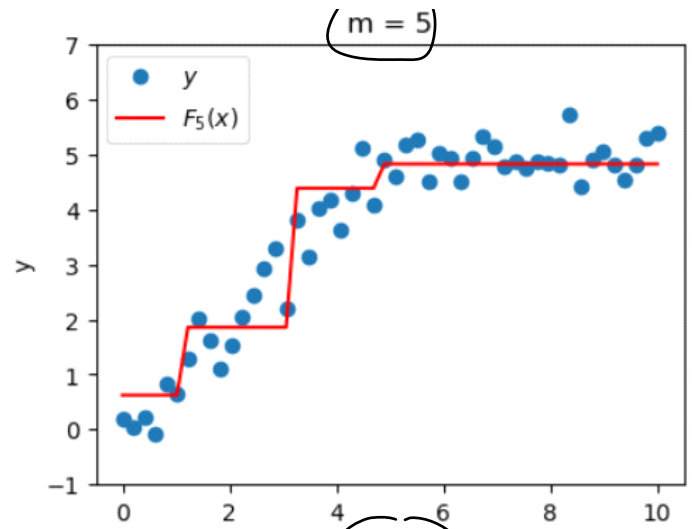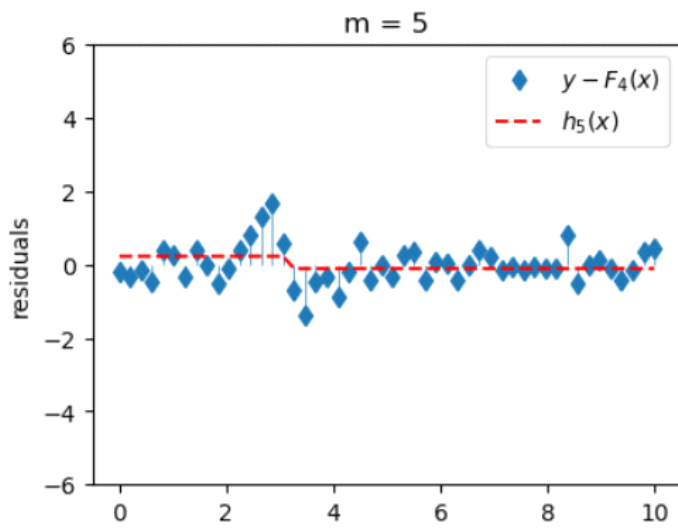
Gradient
Boosting

What →

10

Revision

Legend: y, GBM
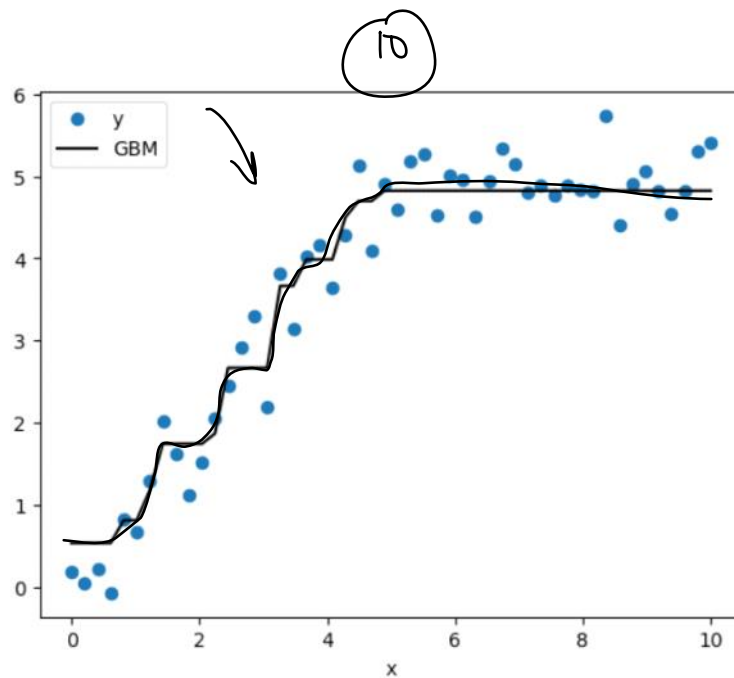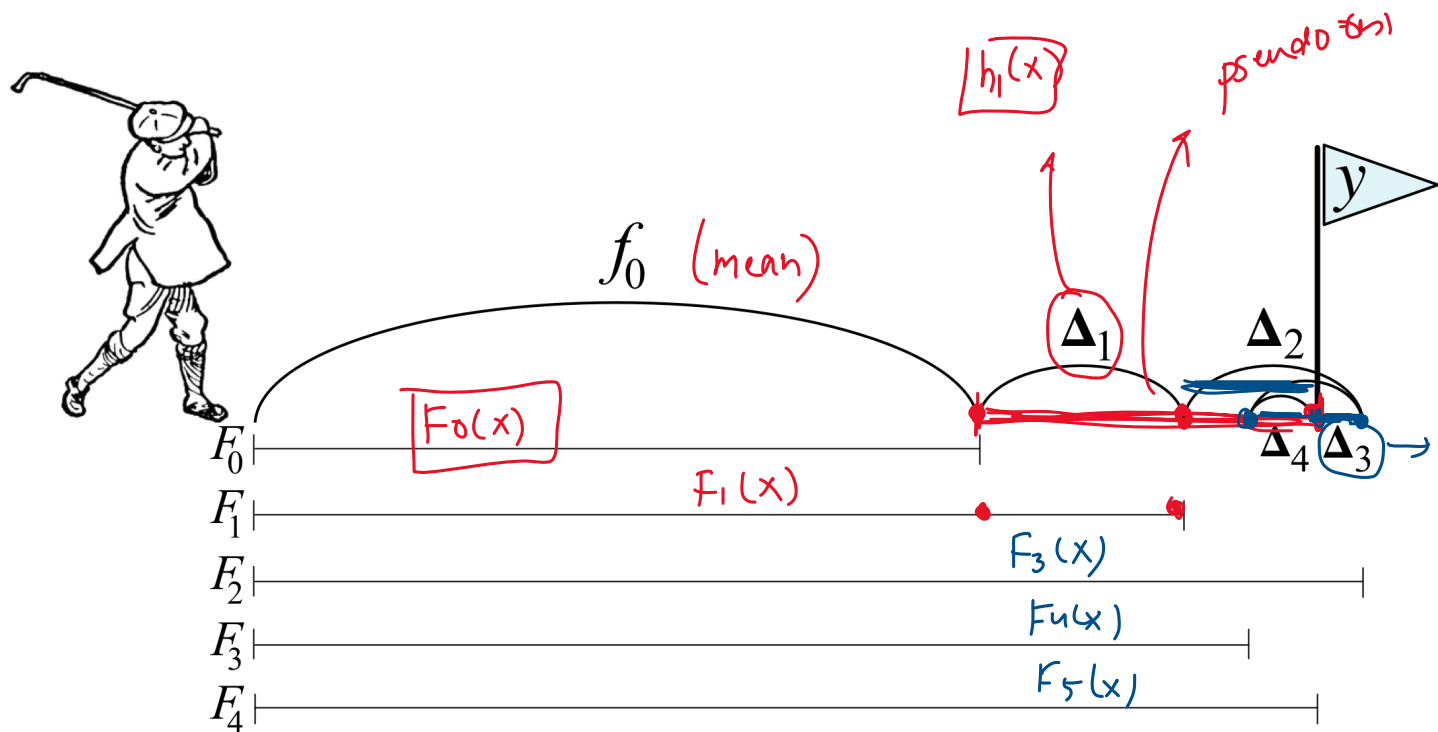
1. Function Space Vs Parameter Space

05 August 2023   16:42

$y = f(x)$ $2x+3$ $f(x) = ex + 2x \rightarrow$ $[\hat{y}_1 = \hat{y}_2] = [\hat{y}_3 = \hat{y}_n = 20$

50000

$\rightarrow$ error

$\begin{array}{l}y \\ \text{salary}\end{array}$ $F_0(x)$ $F_2(x)$ $Y = f(x) = \overline{y} \rightarrow y$ ka mean

$\overline{y}$ better $\rightarrow$ improve

$3x+2$ $F_0(x)$ error   loss function

y $F_1(x)$ reduce error $\hat{y}_1 \ \hat{y}_2 \ \hat{y}_3 \ \hat{y}_u \cdots \hat{y}_n = 20$

$\hat{y}_i$

Loss function $\begin{array}{ccc}exp & salary & prem \\ 10 & 10 & 20 \\ 20 & 20 & 20 \\ 30 & 30 & 20\end{array}$ $F_1(x_i) = \hat{y}_i = 3x_i + 2$

30 dim

30 input feat | 1 out $L(Y, \hat{y}) \rightarrow$ number

X   Y $\frac{mse}{logloss}$ error

hinge loss   (100)

Function space $mse \rightarrow sv$ $L(Y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ $\left[\frac{1}{3}\left[(10-20)^2 + (20-20)^2 + (30-20)^2\right]\right]$

$L(Y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

Constans t (data)

$L(Y, \hat{y}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + (y_3 - \hat{y}_3)^2 + \cdots + (y_n - \hat{y}_n)^2$

variables $\rightarrow$ may vary (model)
type function
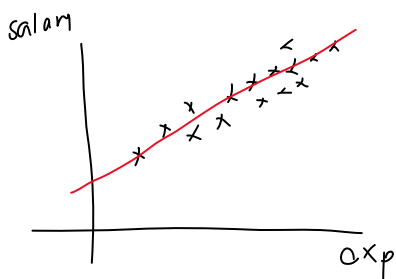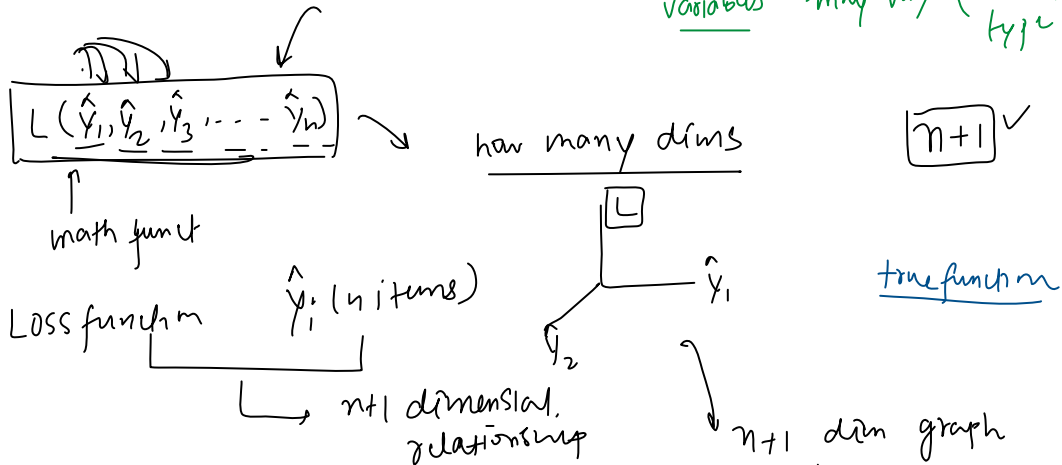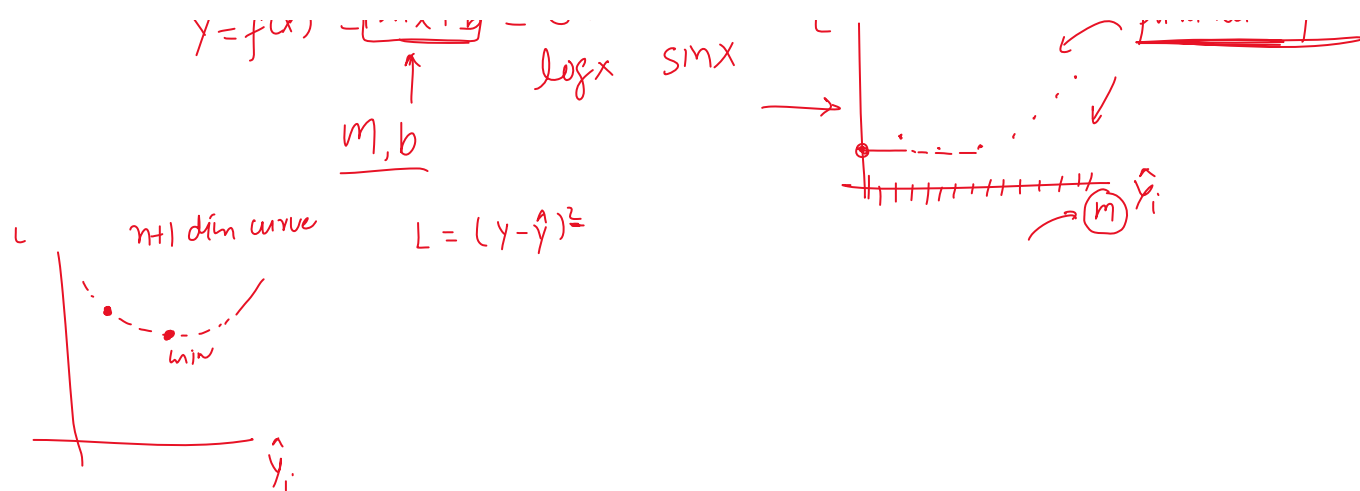
$L(\hat{y}_1, \hat{y}_2, \hat{y}_3 \cdots \hat{y}_n)$ how many dims $\boxed{n+1}$ ✓

math funct L

Loss function $\hat{y}_i$ (n items) $\hat{y}_1$ true function

$\hat{y}_2$

$\rightarrow$ n+1 dimensial relationsing   n+1 dim graph $F_0(x) = \overline{y} = 20$

function space   L   infinite $\hat{y}_i = F_0(x) = \overline{y} = 20$

parametric mc algo   min   function space

linear $\rightarrow$ line $F_0(x) f_1(x) f_3(x)$ $\hat{y}_i$

salary

exp $y = f(x) = \boxed{mx + b} = ex$ parameter space

$logx \quad sinx$

Session 2 Gradient Boosting Perspectives Page 7

$y = f(x)$ $\underbrace{}$ $\log x$ $\sin x$

$\uparrow$

$\underline{m, b}$

$n+1$ dim curve $\qquad L = (y - \hat{y})^2$

L

min

$\hat{y}_i$

(m) $\hat{y}_i$

$\hat{y}_i$

y

$\bar{y}$

$F_0(x)$

x

$n+1$ dim $\boxed{2d}$

$n+1$ dim

$2d$ x

L

$-ve$

L_{min}

$-ve$ $+ve$

$f_0(x)$

$\hat{y}_i$

$\uparrow_0$ $+ve$

slope $+ve$

slope $-ve$

$\underline{slope}$

$\boxed{- slope}$

$L = (y_i - \hat{y}_i)^2$

$\boxed{\dfrac{\partial L}{\partial \hat{y}_i}}$

$\dfrac{\partial L}{\partial F_0(x_i)}$

$F_0(x) + h_1(x)$

$\boxed{\hat{y}_i = F_0(x_i)}$

$\boxed{- slope}$ n slope n slope n numbe

$\boxed{\dfrac{\partial L}{\partial \hat{y}_1}}$ $\boxed{\dfrac{\partial L}{\partial \hat{y}_2}}$ $\boxed{\dfrac{\partial L}{\partial \hat{y}_3}}$ $\cdots$ $\boxed{\dfrac{\partial L}{\partial \hat{y}_n}}$

$\boxed{m = 1}$

$r_i = -\left[\dfrac{\partial L(y_i, f(x_i))}{\partial F(x_i)} \right]_{f=f_0}$

$\boxed{\dfrac{1}{2}}$
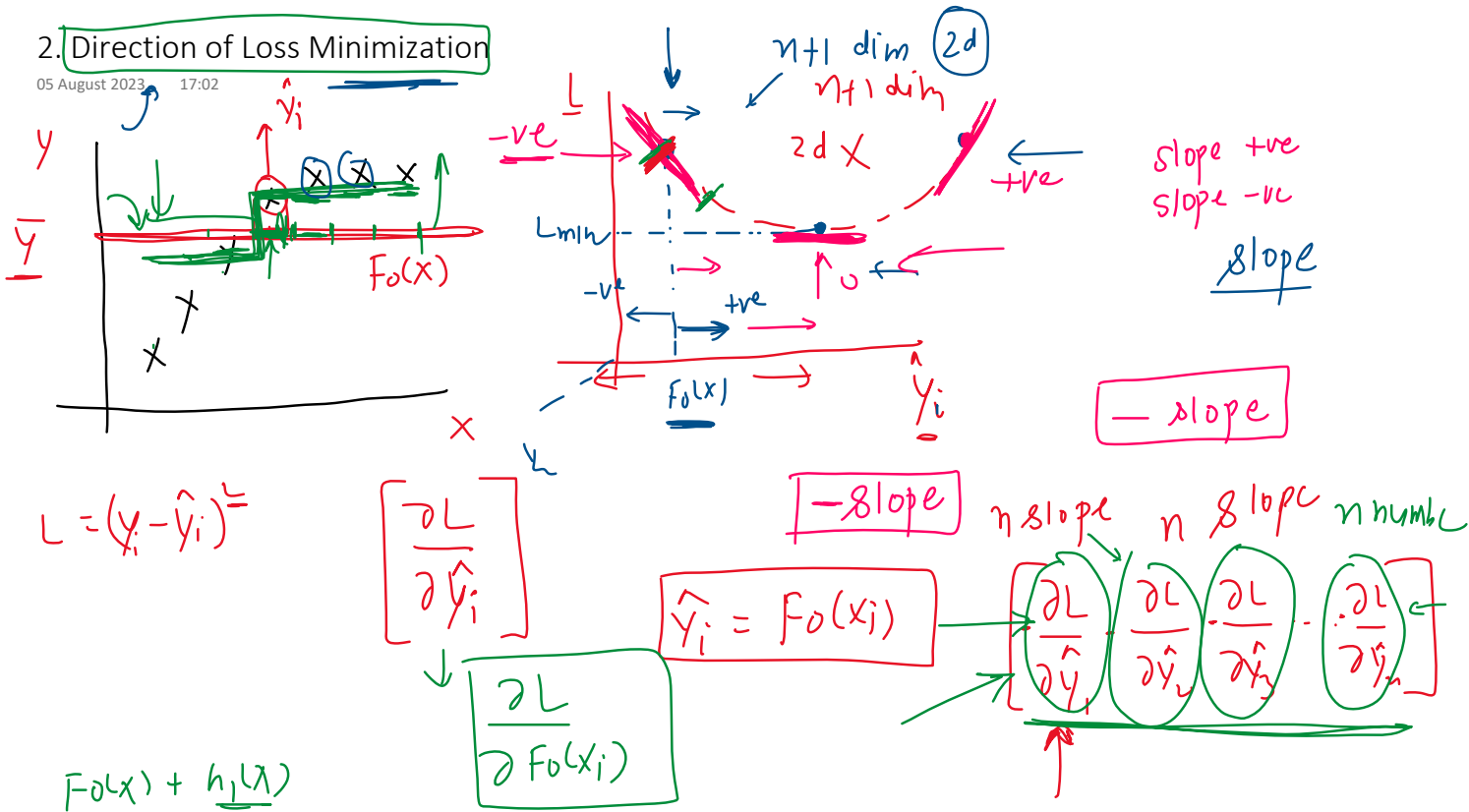
$= \partial \dfrac{\sum\limits_{i=1}^{n} (y_i - f(x_i))^2}{\partial f(x_i)} \bigg]_{f=f_0}$

$\left[y_i - f(x_i) \right]_{f=f_0}$

$\boxed{y_i - f_0(x_i)} \longrightarrow$ slope

$\longrightarrow y_1 - f_0(x_1)$

$\longrightarrow y_2 - f_0(x_2)$

$\longrightarrow y_n - f_0(x_n)$

L $-$ slope

$\boxed{y_i - f_0(x_i)}$ $-ve$

$+ve$

$f_0(x)$

f_0

slope

$\boxed{y_i - f_0(x_i)}$ -ve
+ve

-ve increase

$y_i - \boxed{f_0(x_i)}$ $\boxed{+ve}$ upar

+ve

$\boxed{-ve}$ neeche

$\boxed{\hat{y}_3}$ $\boxed{\dfrac{\partial L}{\partial \hat{y}_3}} = \boxed{-ve}$

$\hat{y}_4 = slope$ $\underline{+ve}$

reduce

increase

$f_0(x)$

→ update →

3. Update the Function

$$X \mid \text{residua} \longrightarrow \frac{\text{weak}}{\text{deci}}$$

$$h_1(x)$$

$F_0(x)$ update

small step

$F_0'(x)$ $\hat{y_i}$

$$\left(\boxed{y_i - f_0(x_i)} + f_0(x_i)\right)$$

$$\left[\frac{\partial L}{\partial \hat{y_i}}\right] \rightarrow \text{direction} \rightarrow \boxed{y_i - f_0(x_i)} \quad \bigcirc y_i$$

$$m_{new} = \underline{m_{old}} - \boxed{slope}$$

$$F_1(x_i) = F_0(x_i) - slope$$

$$\boxed{F_1(x_i) = F_0(x_i) + \underline{y_i - f_0(x_i)}}$$

$$F_0(x) = \underline{m_{old}}$$

$$F_1(x) = m_{new}$$

massive overfitting

$$\boxed{f_1(x_i)} = f_0(x_i) + y_i - f_0(x_i)$$

$$\boxed{f_1(x_i) = y_i} \longrightarrow y_i \rightarrow \text{data output}$$

$$m_{new} = m_{old} - slope$$

$$\boxed{f_1(x) = y_i}$$

every out → $y_1 \quad y_2 \quad y_3 \quad y_n$

$$\boxed{lr = 1}$$

$$F_1(x) = f_0(x_i) + \boxed{y_i - f_0(x_i)}$$

add x

$$slope = \boxed{y_i - f_0(x_i)} - ⓐ$$

psedu $\simeq a$

$$f_0(x_i) + \boxed{a} \times \rightarrow \underline{predict}$$

$$h_1(x_i) \simeq a$$

$$X \mid \frac{y_i - f_0(x_i) \quad \{a\}}{\text{decision}} \quad a \simeq a$$

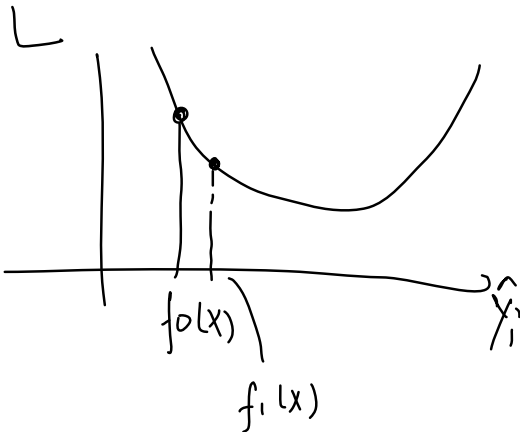weak learne

$$f_1(x) = f_0(x) + h_1(x)$$

right direct

$$f_1(x) = f_0(x) + h_1(x)$$

right direction

$\llcorner$ weak dt

$$x_i \mid \underbrace{y_i - f_0(x_i)}_{\text{residu}}$$

$$f_1(x) = f_0(x) + h_1(x)$$

$\llcorner$

$f_0(x)$

$\hat{y}_i$

$f_1(x)$

4. Iterate

$x \to$ residuals $\to h_2(x)$

$\to F_1(x)$

$F_3(x)$

$F_2(x) \to$ slope / residu

x

$h_3(x)$

$F_2(v) = F_0(x) + h_1(x) + h_2(x)$

L

$F_0(x)$

$F_1(x)$

$\hat{y}_i$

$F_1(x) = F_0(x) + h_1(x)$

weak den

$x | r$ cond
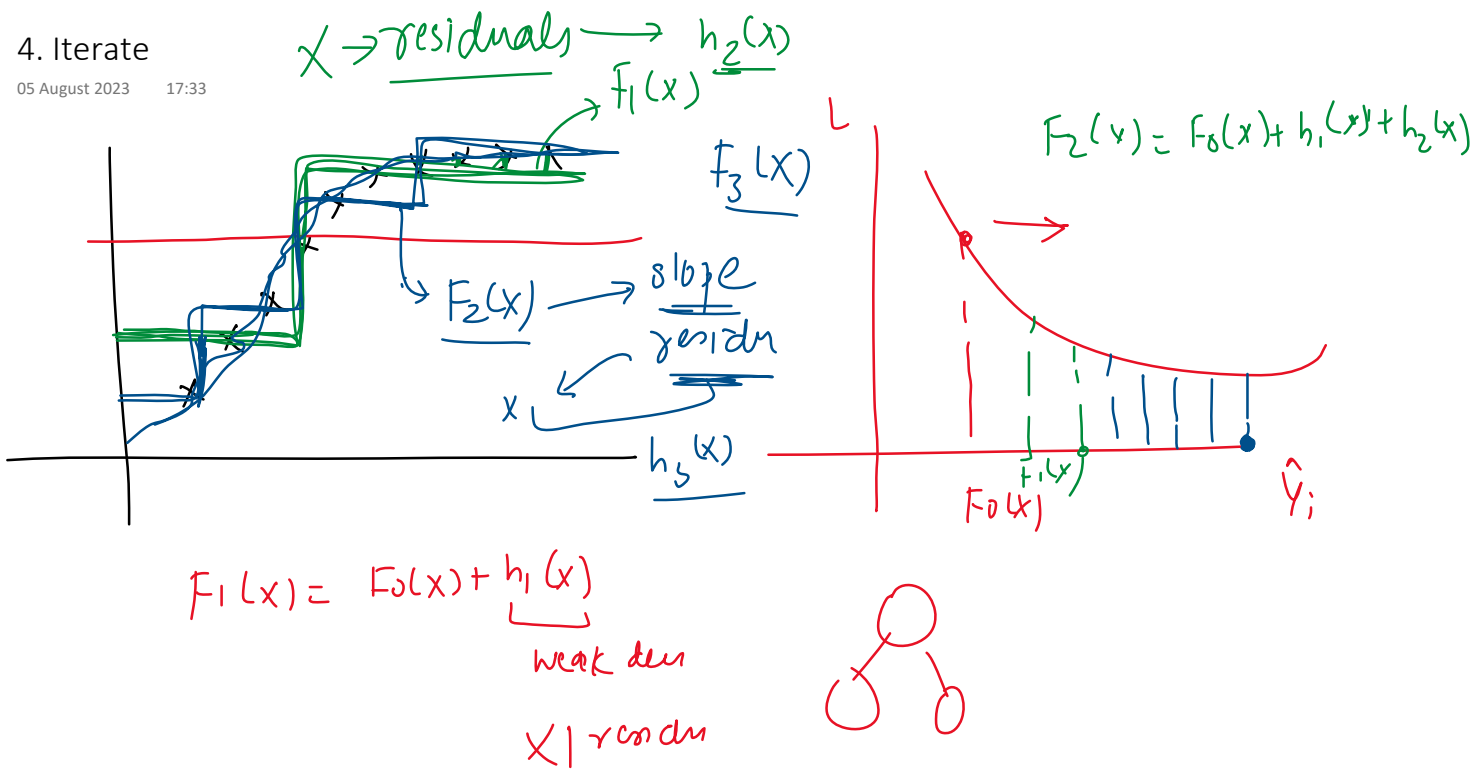
Detour - Gradient Descent

Another Perspective

MSE Loss Function

$(y - F_0)^2$

$(y - F_1)^2$

$(y - F_2)^2$

y

f_0

Δ_1

Δ_2

Δ_4 Δ_3

F_0

F_1

F_2

F_3

F_4
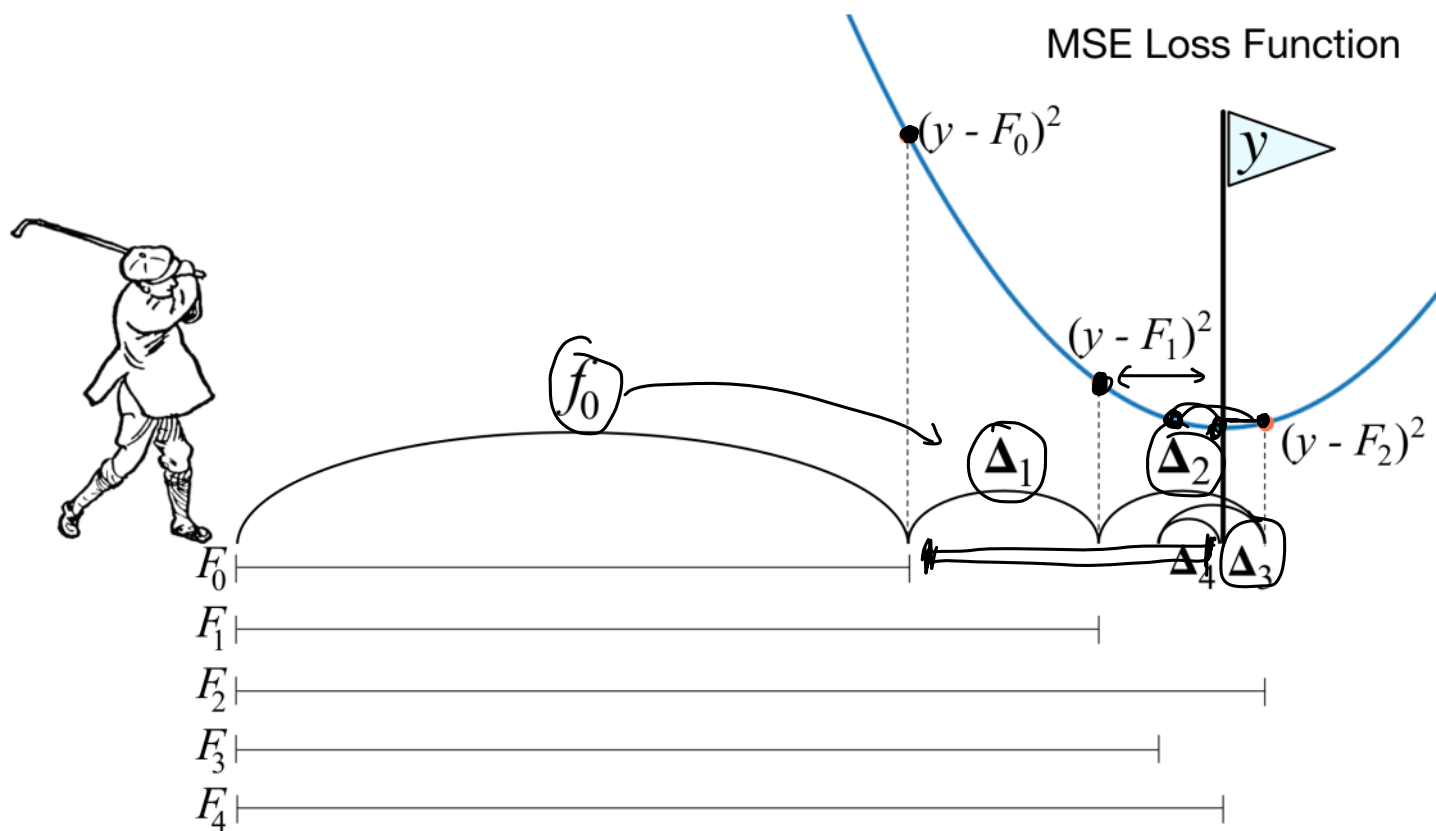
Difference Between Gradient Boosting and Gradient Descent

$$\begin{bmatrix} 1) & \text{Gradient Descent} \rightarrow \text{parameter space} \\ & \text{Gradient Boosting} \rightarrow \text{function space} \end{bmatrix} \boxed{0.1} \rightarrow$$

$$m_\eta = m_0 - \eta \, \underline{\text{slope}}$$

weak predict'n

$$f_m(X) = f_{m-1}(X) + \eta \, h_m(X)$$

→ dcisn

↳ decision

learning rate
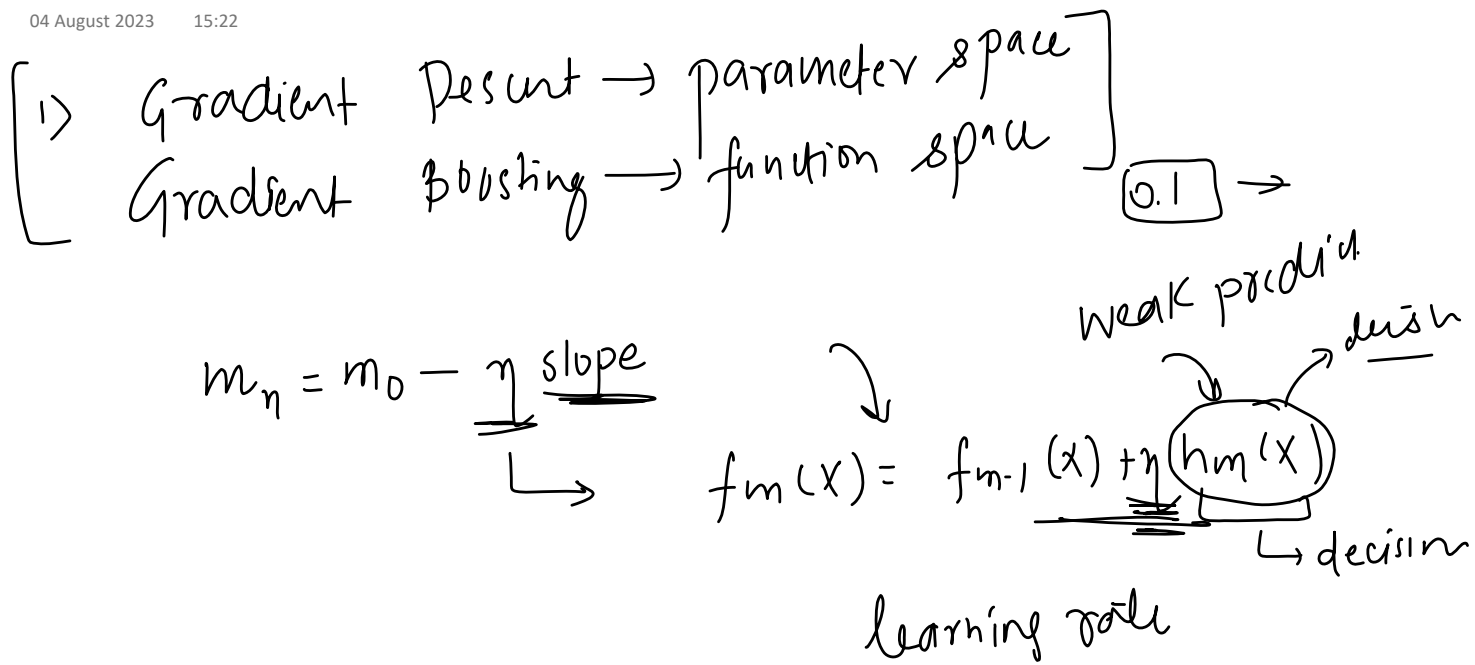
Advantages of Gradient Boosting

05 August 2023   07:52

Let's Appreciate

05 August 2023   18:31