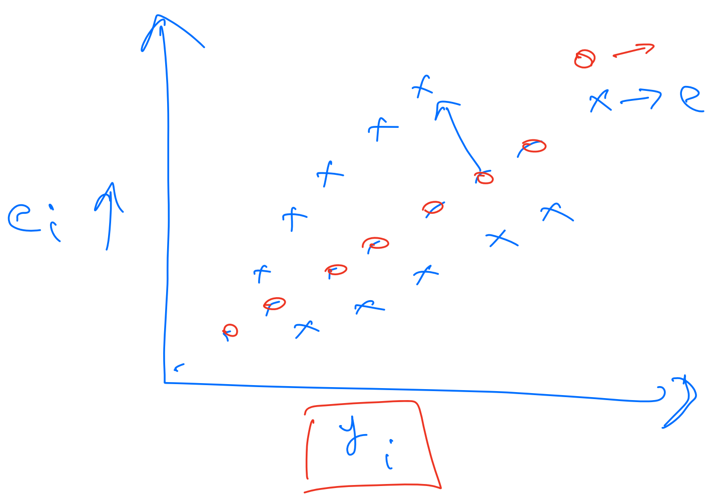
Last Class May 30

1) Recap of Feature Importance, Model Interpretation, Gradient Descent Geometric Intuition ✓
2) 5 – 6 Statistical Assumptions of Linear Regression ✓
   ↳ VIF, MC, etc
3) Practical Scenario for these assumptions ✓
   (Business Needs)
4) Train / Test Split OR Train / Val / Test split ✓

Today's Class — June 1

1) Recap of heteroskedasticity ✓
2) SGD / Mini Batch ✓ (BGD)
3) Polynomial Regression ✓
4) Generalization & Occam's Razor ✓
5) Underfit – Overfit Tradeoff (Bias – Variance) ✓
6) Regularization (L1 & L2) ✓
7) Hyperparameter Tuning. ✓



o → (y_i)

x → e_i

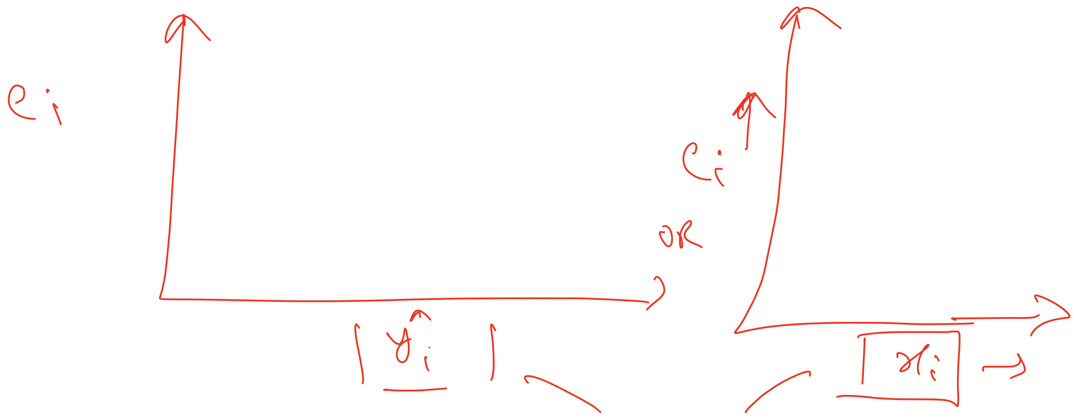
$e_i = y_i - \hat{y}_i$

gt   pred

should we plot

e_i vs y_i ×

or e_i vs $\hat{y}_i$ ✓



$$\hat{y}_i = w_0 + w_1 x_i$$

Case I: $y_i \to$ constant

is LR necessary? × no

Case II: $y_i \to$ changing

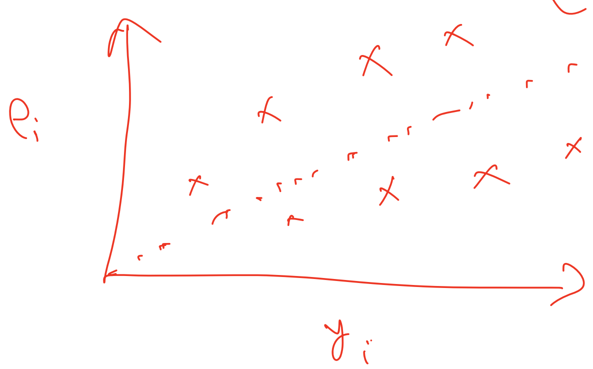
$\boxed{\hat{y}_i \to \text{constant}}$

$e_i = \boxed{y_i} - \hat{y}_i \Rightarrow$ will ei change or increase/decrease with y_i

$y_i \uparrow$, $e_i \uparrow$

$\hat{y}_i \Rightarrow$ constant

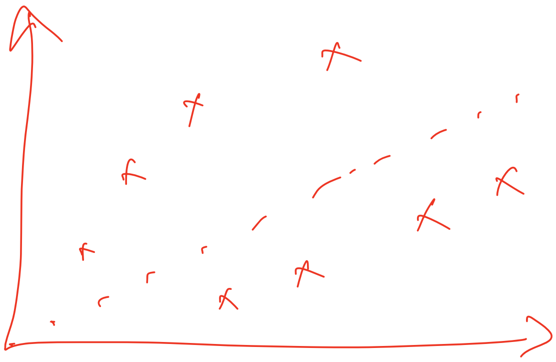
bad model being learnt



Case III: when y_i is also changing & $\hat{y}_i$ is also changing

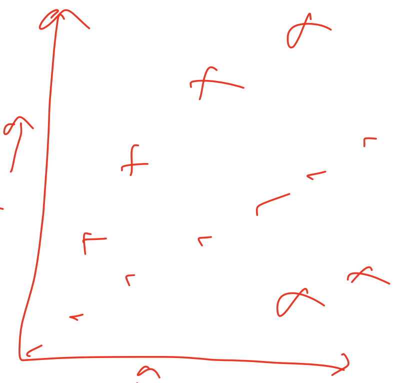
Common Scenario — None of y_i & $\hat{y}_i$ is constant

e_i



y_i

e_i



$\hat{y}_i$

$$y_i = 2x_i + 1$$
$$\hat{y}_i = x_i + 1$$

$$e_i = y_i - \hat{y}_i = \boxed{x_i}$$

$$e_i = f(x_i)$$

e_i vs $\hat{y}_i$

e_i

$\hat{y}_i$

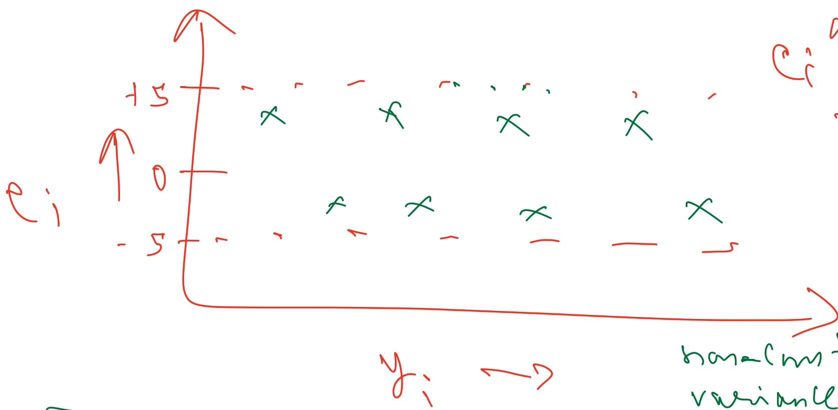
e_i

y_i

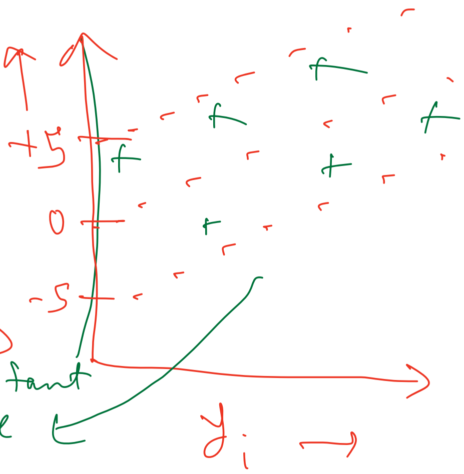
$$y_i = 2x_i + 1$$
$$\hat{y}_i = 2x_i + 3$$

$$e_i = \boxed{-2}$$

constant



Constant variance



non constant variance

$y_i = f(x_i) + \boxed{\varepsilon}$ → random noise

$\hat{y}_i = f_2(x_i) + \boxed{\varepsilon_1}$ → Some other random noise

SGD / Mini Batch GD

→ Batch Gradient Descent

$$\sum_{i=1}^{N} grad(y_i, x_i, w)$$

computed for all data points (x_i, y_i)

$$w_0 = w_0 - \eta \boxed{\frac{\partial L}{\partial w_0}} \rightarrow grad$$

$$w_1 = w_1 - \eta \boxed{\frac{\partial L}{\partial w_1}} \rightarrow grad$$

(x_i, y_i)

↓

$(x_{i1}, x_{i2}, x_{i3}, \ldots x_{id})$

Let's say there are 10,000 (x_i, y_i) datapoints

✓ → time consuming] Batch GD [floating point ↓ 64 bit ; int → 32 bit] $2^{31}-1$ ↓ 0 to $2^{31}-1$

✓ → overflow error

SGD → Stochastic Gradient Descent
(Update after every point (x_i, y_i))

$$\frac{\partial L}{\partial w_j}(x_i, y_i)$$

$$w_j = w_j - \eta \frac{\partial L}{\partial w}\left((x_i, y_j)\right)$$

Mini Batch

$$\frac{\partial L}{\partial w_j}\Big|_{(x_i, y_i)}$$ → batch of 100

Update the weights

$(x_{100}, y_{100}) \rightarrow (x_{200}, y_{200})$

$(x_{201}, y_{201}) \rightarrow (x_{301}, y_{300})$

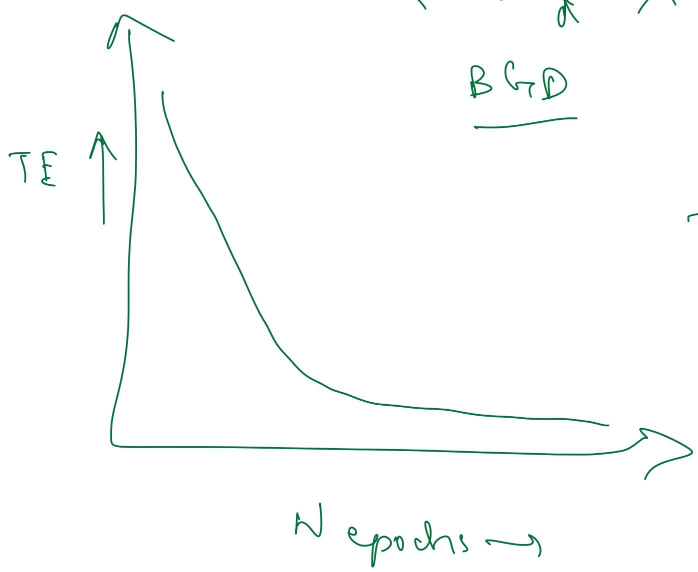
-2^{32} to $2^{31}-1$ → $-2^{32}, \dots$

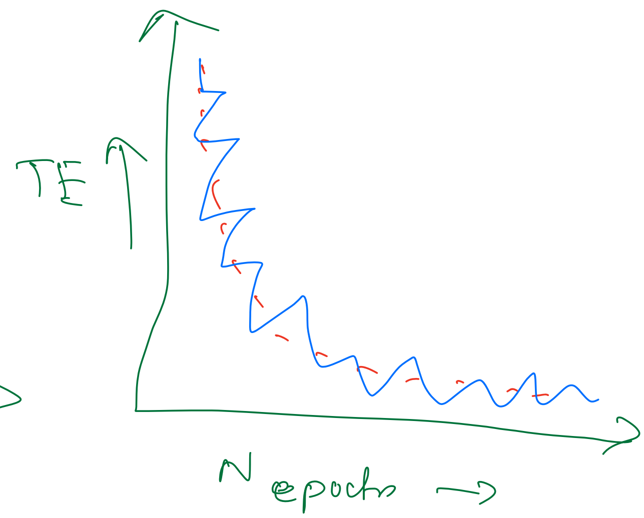
SGD → # of Computations

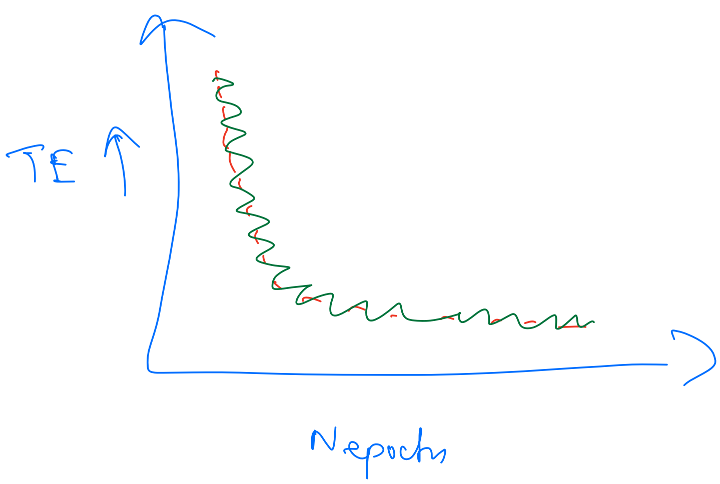
$\boxed{\propto N_d \;,\; \propto N_{epochs}}$

$\propto N_d \times N_{epochs}$

BGD







$N_{comp} \rightarrow N_{epochs} \times N_{updates}$

Batch of 100

10,000 datapoints

SGD → $N_{updates} \rightarrow 10{,}000$

MBGD → $N_{updates} \rightarrow \frac{10{,}000}{100}$

$= 100$

$N = 10{,}000$

$MB = 100$

$N_{epochs} = 3$

$$N_{updates} = \frac{N}{MB} \times N_{epochs}$$

$$= \frac{10000}{100} \times 3$$

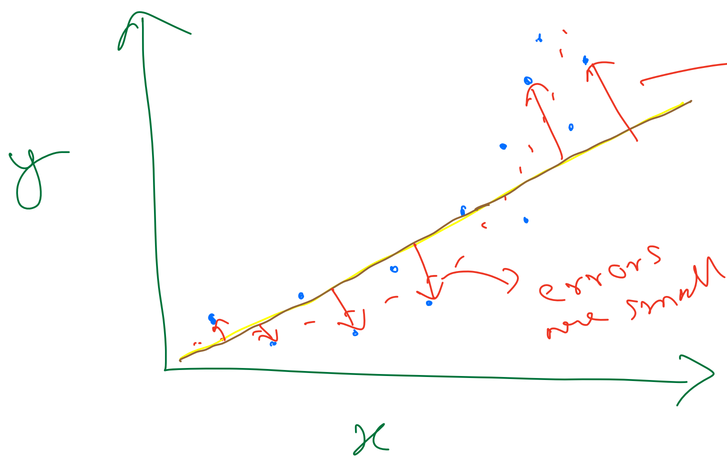
$$= 100 \times 3 = 300$$

Polynomial Regression

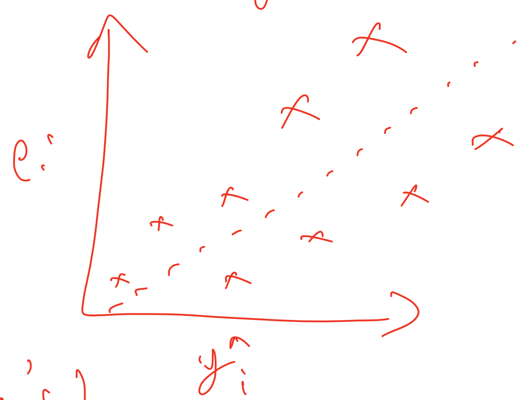
$y = w_0 + w_1 x \rightarrow$ LR

$y = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4$

degree 4 polynomial regression



errors are big



[higher degree of inputs (x_i's)
linear (1st order) degree of weights

$w_0^2, \; w_1^3, \; w_2^4$ ✗

$y = w_0 + w_1 x_1 + w_2 x_2 \rightarrow$ LR is not working

$y = w_0 + w_1 \underline{x_1} + w_2 \underline{x_2} + w_3 x_1^2 + w_4 x_2^2$

$w_1 x_1 + b x_1 \longrightarrow (w_1 + b) x_1$

$x_1, x_2, x_3 \longrightarrow MC \longrightarrow x_1, x_2$

$\boxed{1, x_1, x_2, x_1^2, x_2^2}$

$1, 7, 8, 49, 64$

$x_1^2 \longrightarrow x_1$

MC

$x_1^2 \longrightarrow$

$f_1 \to b x_1 + c$

f_1 & $x_1 \longrightarrow$ collinear

but x_1^2 & $x_1 \longrightarrow$ not collinear

$f2 = \log(f_1)$

$y = (w_0 + w_1 \boxed{x_1} + w_2 \boxed{x_2} + w_3 x_1^2 + w_4 x_2^2)$

$W = [w_0, w_1, w_2, w_3, w_4]$

$x_1 x_2 \longrightarrow$ degree 2 term

$x_1, x_2 \longrightarrow$ No MC | $x_2, x_2^2 \longrightarrow$ No MC

$x_1, x_1^2 \longrightarrow$ NO MC | $x_1, x_2^2 \longrightarrow$

$x_2 \rightarrow \sqrt{x_1}$, $\boxed{x_1 , x_2^2 \longrightarrow MC}$

RSP $\rightarrow$ $\boxed{x_1 \rightarrow \text{Salary}}$ $\boxed{x_2 \rightarrow \text{Number of kids}}$

$y = w_0 + w_1 x_1 + w_2 x_2^2 + \cdot + 3x_1^3 + 4x_2^2$
$+ \ldots$

$[w_0 , w_1 , w_2 , \ldots]$

x_3 = mileage

degree 2 polynomial features

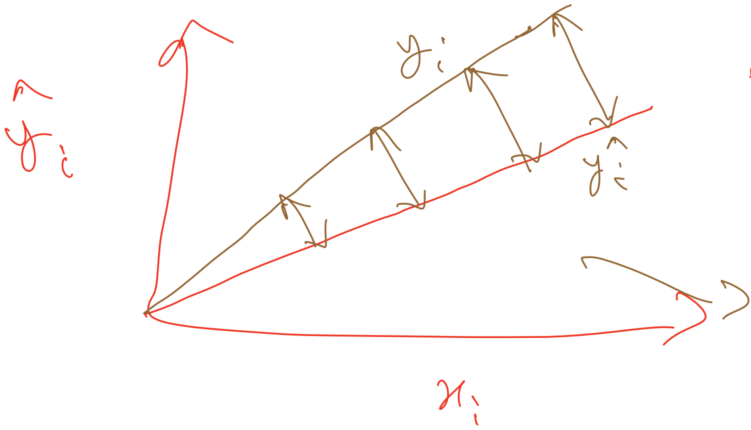
$x_1 , x_2 , x_3 , x_1^2 , x_2^2 , x_3^2 , x_1 x_2 , x_1 x_3 , x_2 x_3$

$[x_1 , x_2 , x_3]$

$x_1 \rightarrow$ 1st feature (salary)
$x_2 \rightarrow$ 2nd feature (kids)
$x_3 \rightarrow$ 3rd feature (mileage

$X = [\quad]$

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}_{N\times 1} \qquad X_{poly} = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 & x_1^4 & x_1^5 \\ 1 & x_2 & x_2^2 & x_2^3 & x_2^4 & x_2^5 \\ - & & - & & - & \\ & & & & & \vdots \end{bmatrix}_{N\times 6}$$



$\hat{y}_i = 1 \cdot x_i + 1$

$y_i = 2x_i + 1$

classic hetero-skedasticity problem

$$v = [1, 2, 3, 4, 5]$$

$$VIF_j = \frac{1}{1 - R2_j}$$

$$f_j = [f_1, f_2, \ldots f_{10}]$$

$j = 9$

$f_9 \to y$