→ Gradient Descent Variants

→ Polynomial Regression

→ Underfitting & Overfitting

→ Bias - Variance

Repeat until convergence

$$w_j = w_j - \eta \frac{\partial L}{\partial w_j}$$

$$w_j = w_j - \eta \frac{1}{m} \sum_{i=1}^{m} \left(y^{(i)} - \hat{y}^{(i)} \right) x_j^{(i)}$$

$$\hat{y} = w_0 + w_1 x$$

$$= w_0 + \sum_{j=1}^{d=5} w_j x_j$$

$$= w_0 + w_1 x_1 + w_2 x_2 - \cdots - w_5 x_5$$

Repeat until convergence

{

$$w_j = w_j - \eta \frac{\partial L}{\partial w_j} = \frac{1}{m} \sum_{i=1}^{m} (y^{(i)} - \hat{y}^{(i)}) x_j^{(i)}$$

}

$y^{(i)}$ = target

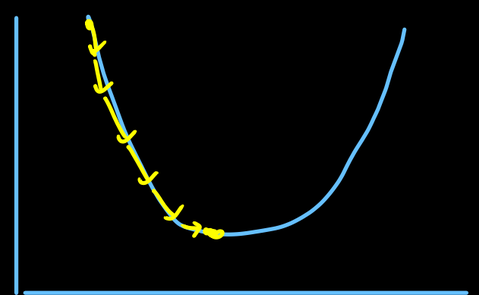
$\hat{y}^{(i)}$ = Prediction

w_i = weight of feature i

η = Learning rate

$\rightarrow \hat{y}^{(i)} = w_0 + w_1 x_1^{(i)}$

	x_1	x_2	x_3	y	
m data points	$x_1^{(1)}$	$x_2^{(1)}$	$x_3^{(1)}$	$y^{(1)}$	← 1st sample
	$x_1^{(2)}$	$x_2^{(2)}$	$x_3^{(2)}$	$y^{(2)}$	← 2nd sample
	$x_1^{(m)}$	$x_2^{(m)}$	$x_3^{(m)}$	$y^{(m)}$	



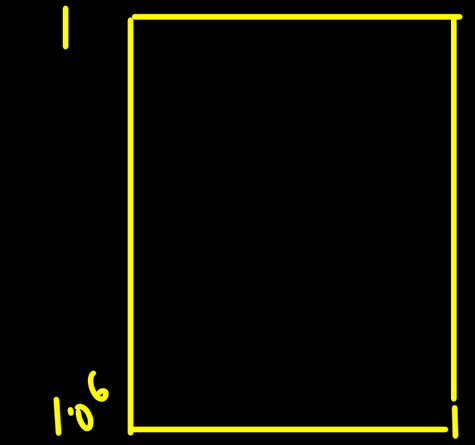
Batch GD

1000
11

for iter in range (iteration):

{

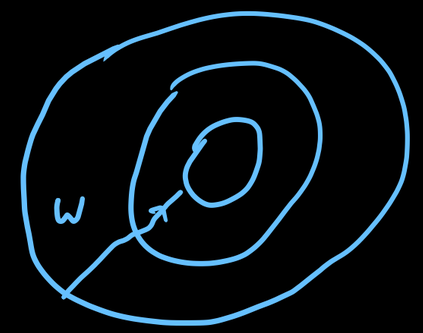
$$w_j = w_j - \eta \frac{1}{m} \sum_{i=1}^{m} \left(y^{(i)} - \hat{y}^{(i)} \right) x_j^{(i)}$$



}

→ Huge memory in RAM

→ Very less updates



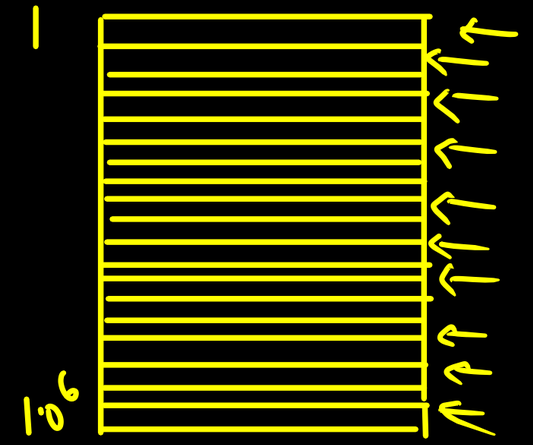
SGD (Stochastic GD)

for iter in range (iteration):

{

for i in range (m):

$$w_j = w_j - \eta \left(y^{(i)} - \hat{y}^{(i)} \right) x_j^{(i)}$$

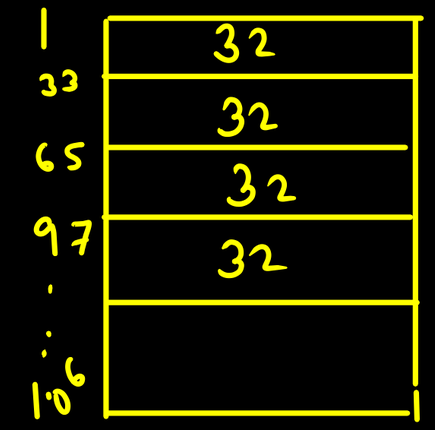


→ Very large update for w_j

Chunks = Batch size

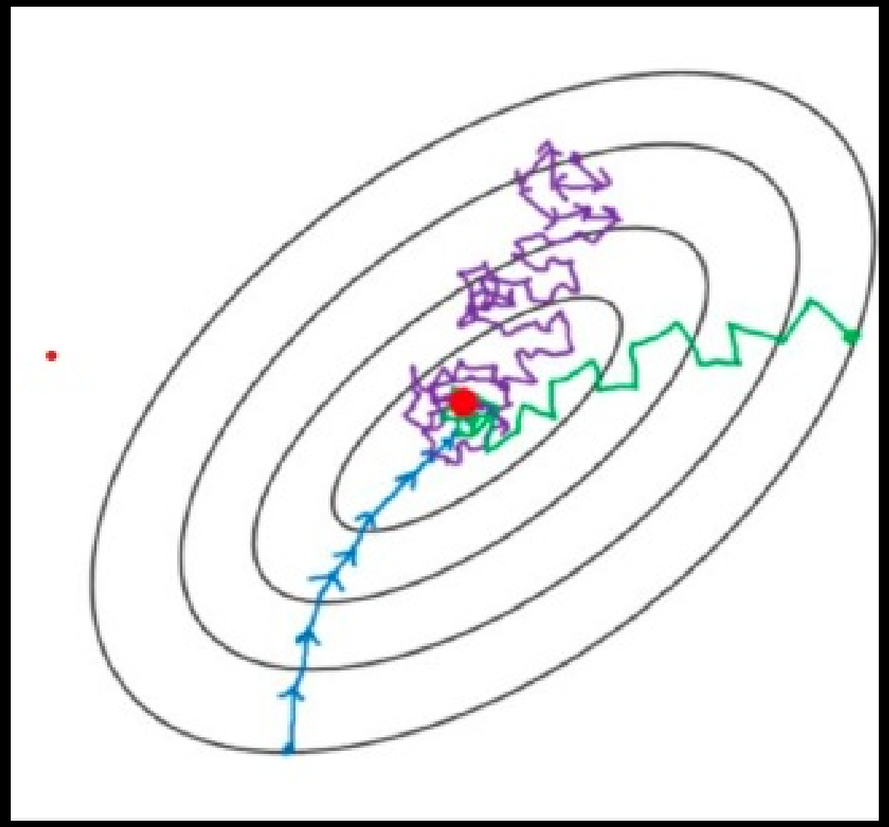
Mini batch

```
for iter in range (iteration):
{
    for i in range (No of batches):
    {
```



$$w_j = w_j - \eta \frac{1}{\text{Batch size}} \sum_{i=1}^{\text{Batch size}} (y^{(i)} - \hat{y}^{(i)}) x_j^{(i)}$$

1. BGD ✓
2. SGD ✓
3. Mini batch GD →

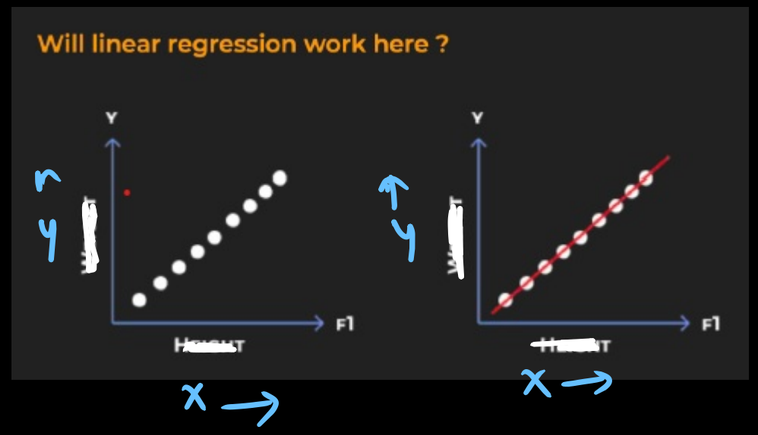


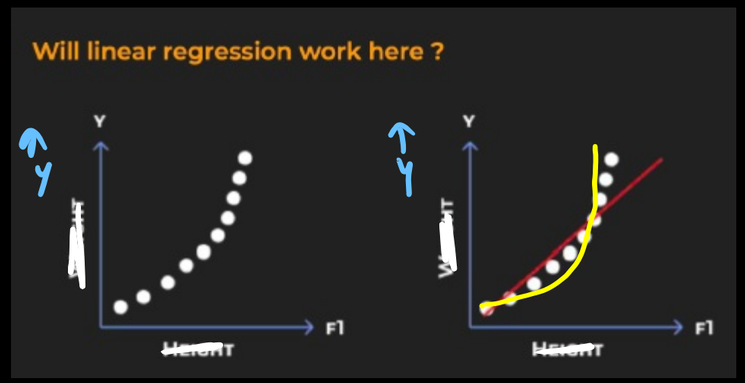
① B — BGD

② G → SGD
V → Mini batch GD

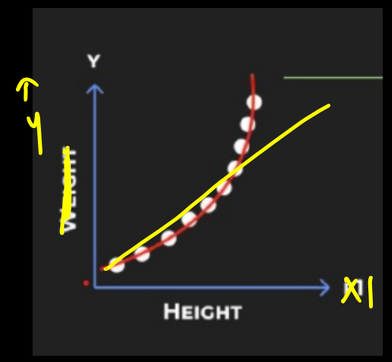
Linear Reg

$$\hat{y} = w_0 + \sum w_j x_j$$





Polynomial Reg



Linear Reg

$$\hat{y} = w_0 + w_1 x_1 \rightarrow \text{Linear}$$

$$\hat{y} = w_0 + \sum_{j=1}^{d} w_j x_j$$

Poly Reg

$$\hat{y} = w_0 + w_{11} x_1 + w_{12} x_1^2$$

- 2 = Quadratic
- 3 = Cubic
- $n > 3 \rightarrow$ Poly

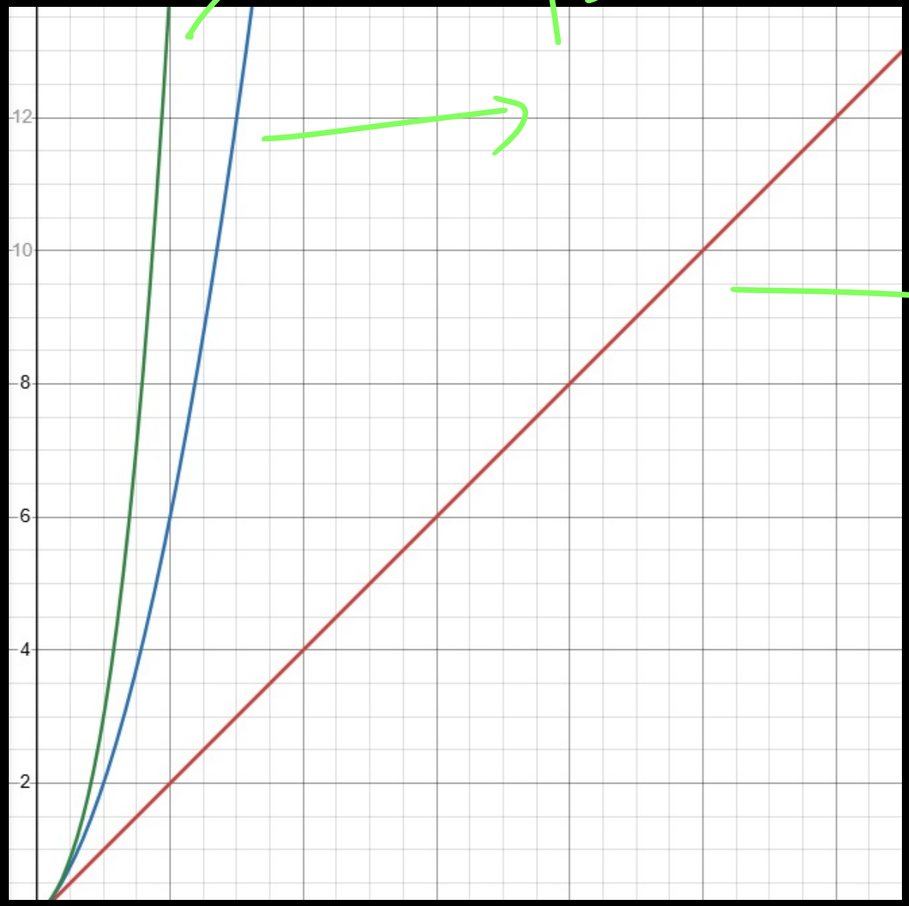
$$\hat{y} = w_0 + w_1 x_1 + w_2 x_2 \cdots w_d x_d$$

$$+ w_{12} x_1^2 + w_{22} x_2^2 \cdots w_{d2} x_d^2$$

$$+ w_{13} x_1^3 + w_{23} x_2^3 \cdots w_{d3} x_d^3$$

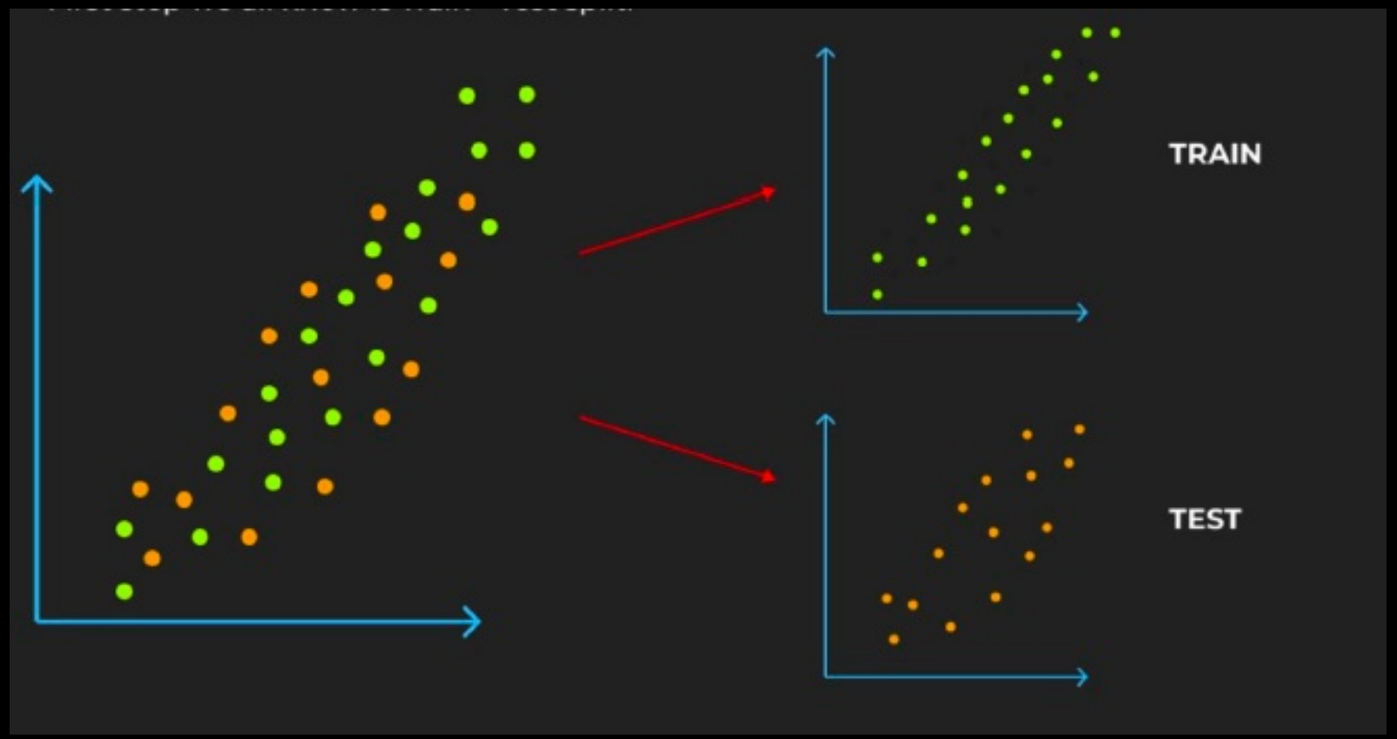
$y = x + x^2 + x^3$

$y = x + x^2$



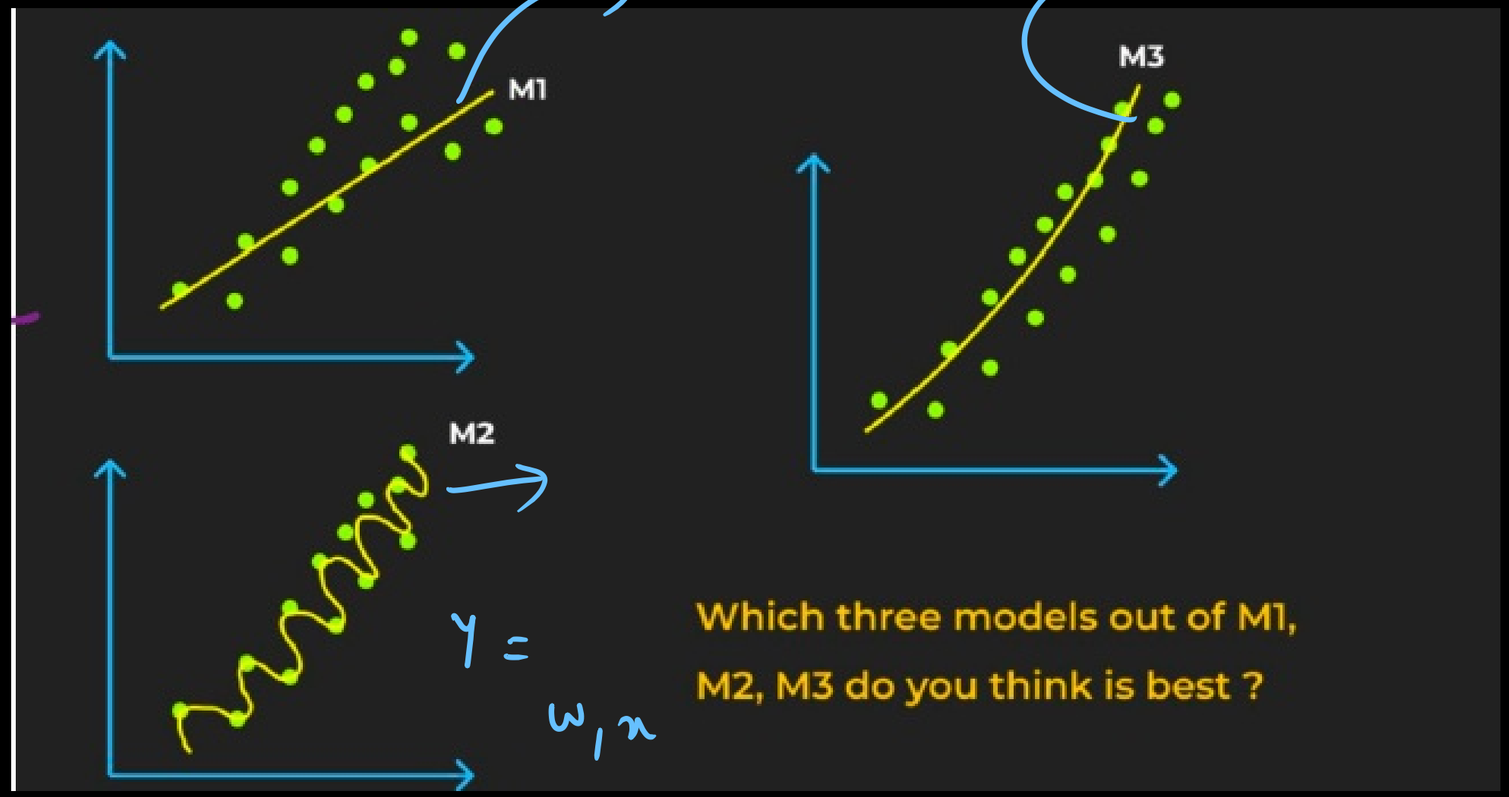
$y = x$

→ Break until 22:07 PM


TRAIN
TEST

$y = w_1 x_1 + w_0$

$y = w_1 x + w_{11} x_1^2 + w_0$



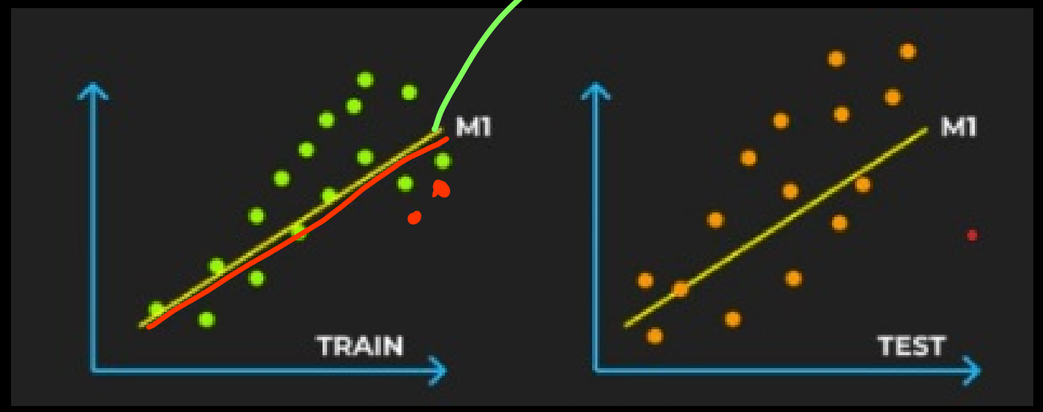
$y = w_1 x$

$+ w_{11} x_1^2$

$+ w_{13} x_1^3 + w_{14} x_1^4$

Underfitting

$y = w_0 + w_1 x_1$



→ Error $(y - \hat{y})$

↳ High bias & low variance

1) High → Sensitivity w.r.t train data

2) Low

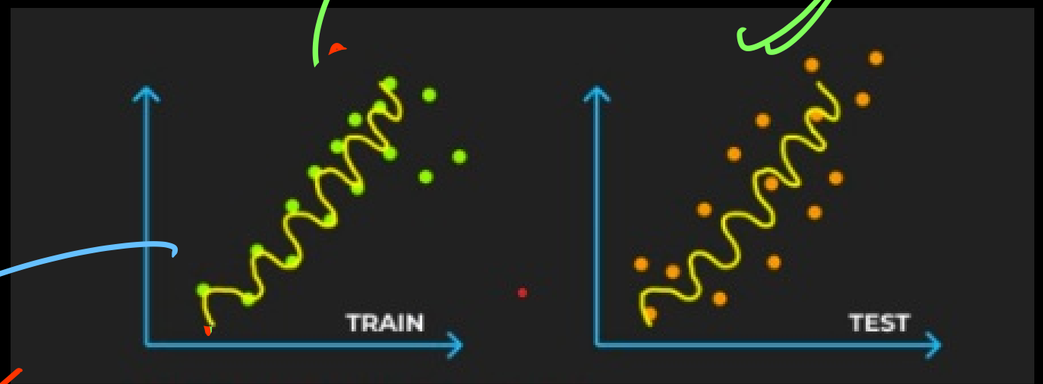
→ Student did not study text book (training data)

✓ high train error

✓ high test error

Overfitting

$$y = w_0 + w_1 x_1 + w_{12} x_1^2 + w_{13} x_1^3 + w_{14} x_1^4$$



Error $(y - \hat{y})$

=

Low bias

Variance High

→ Student has memorized the text book (train data)

✓ Low train error

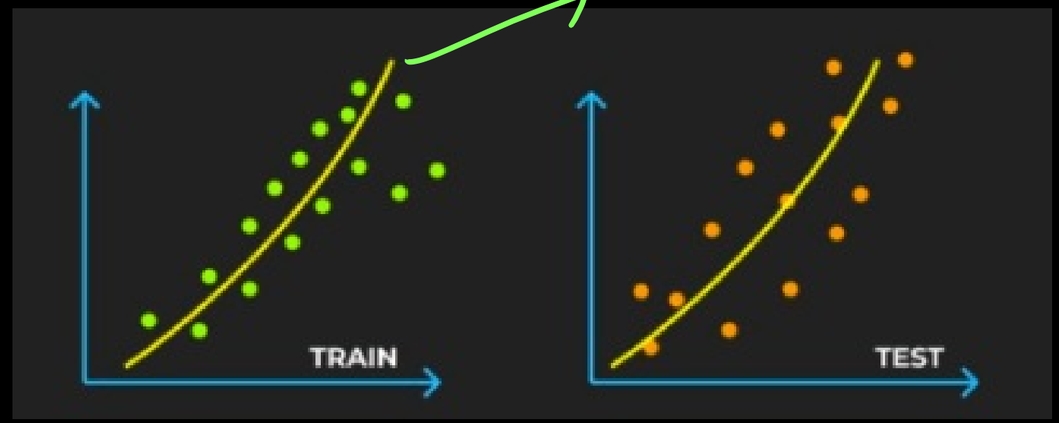
High test error

1) High

2) Low

→ Perfect fit (Best fit)

$y = w_0 + w_1 x_1 + w_{12} x_1^2$



1) High

2) Low

→ Student has studied text book smartly & can apply concept

Low train error Low test error

$y = w_0 + w_1 x$

← SIMPLEST

COMPLEX

	TRAINING	TESTING
UNDERFIT	POOR	POOR
PERFECTLY FIT	GREAT	GREAT
OVERFIT	BEST	~~DECENT~~ Poor

← High train & test

← Low train & high test

$y = w_0 + w_1 x \quad \cdots \quad w_{15} x^5$

Loss

Underfitting

Overfitting

Error

Train Error

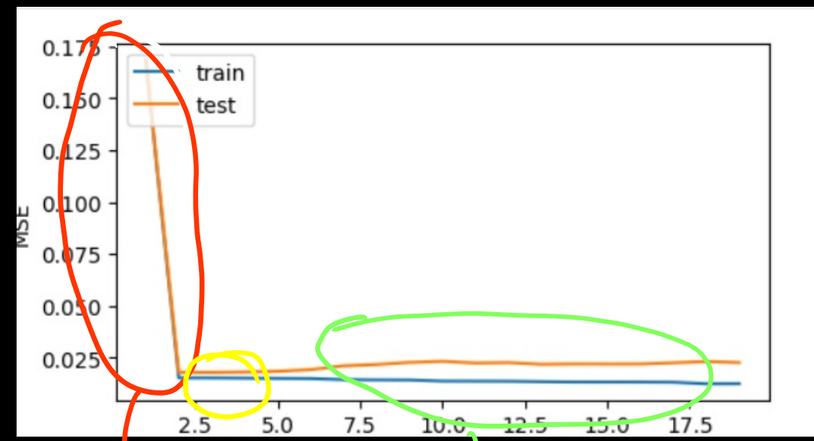
Test Error

1

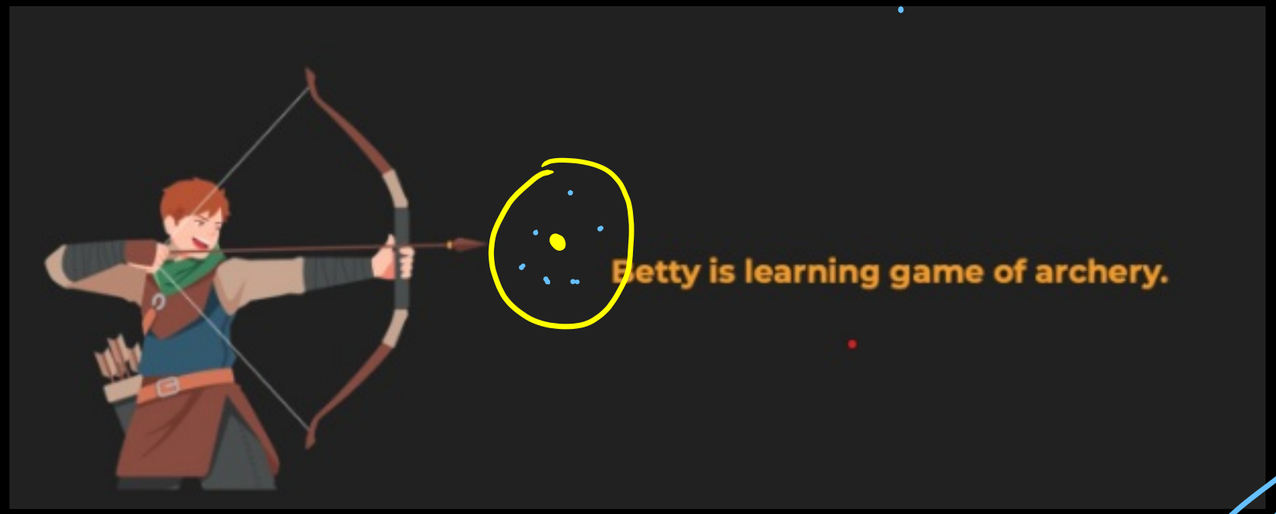
10

d → degree of polynomial

Perf
fit



Bias & Variance



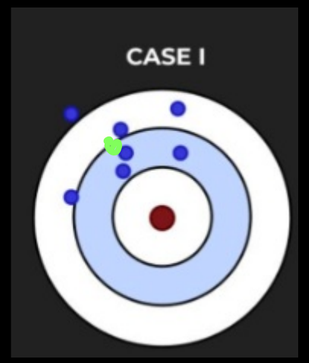
f = Actual target

$\hat{f}$ = Where you hit

Bias = $\text{Mean}(\hat{f}) - f$

Variance = Spread of $\hat{f}$

$= \frac{1}{m} \sum_{i=1}^{m} (\hat{f} - \bar{\hat{f}})^2$

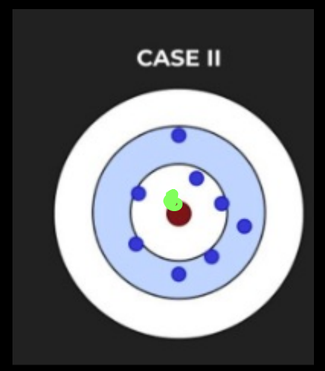


→ Bias-H

→ Var-H

1) High

2) Low



→ Bias − L

→ Var − H

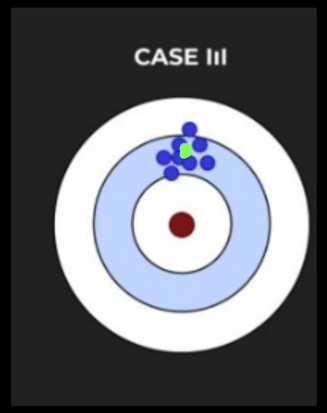
f = Actual target

$\hat{f}$ = Where you hi-

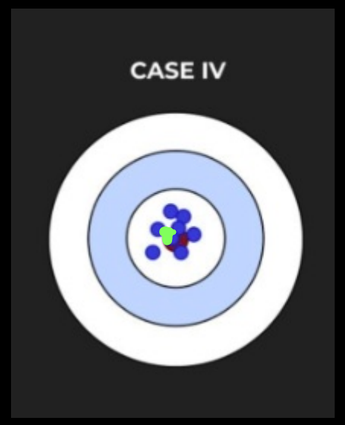
$$\text{Bias} = \text{Mean}(\hat{f}) - f$$

$$\text{Variance} = \text{Spread of } \hat{f}$$

$$= \frac{1}{m} \sum_{i=1}^{m} (\hat{f} - \bar{\hat{f}})^2$$



→ Bias – H 1) High
→ Var – L 2) Low



→ Bias – L
→ Var – L

f = Actual target

$\hat{f}$ = Where you hi-

$$\text{Bias} = \text{Mean}(\hat{f}) - f$$

$$\text{Variance} = \text{Spread of } \hat{f}$$

$$= \frac{1}{m} \sum_{i=1}^{m} (\hat{f} - \bar{\hat{f}})^2$$

Error $(y - \hat{y})$

Sensitivity with test data ↑

Under — High bias & low variance

Overfitting — low bias & high variance