Types of Gradient Descent

03 August 2025 16:43

Row	y_actual	y_pred	(y_pred - y_actual)²
1	10	12	$(12 - 10)^2 = 4$
2	8	7	$(7 - 8)^2 = 1$
3	15	14	$(14 - 15)^2 = 1$
4	20	18	$(18 - 20)^2 = 4$
5	13	16	$(16 - 13)^2 = 9$

→ How many times should this calculation be made to calculate loss?

↳ n times.

Fewer the n, Faster it is to calculate for all points

How would that impact calculating the loss function?

↳ $\sum (y_{pred} - y_i)^2$

$$w^{t+1} = w^t - \lambda \frac{\delta L}{\delta w}$$

1) Batch Gradient Descent

↳ all squared errors at once

weight updated only once for all n points

If I have 100 points → 100 calculations

⇓

I am updating weights once

Process of going through entire data ⇒ epoch

2) Mini-Batch Gradient Descent

Row	y_actual	y_pred	(y_pred - y_actual)^2
1	10	12	$(12 - 10)^2 = 4$
2	8	7	$(7 - 8)^2 = 1$
3	15	14	$(14 - 15)^2 = 1$
4	20	18	$(18 - 20)^2 = 4$
5	13	16	$(16 - 13)^2 = 9$
100	50	45	$(45-50)^2 = 25$

Rows 1–5 ⇒ Batch 1

⇒ Batch 2

⇒ Batch 20

Dividing 100 points into 20 batches of 5 points each

calculate loss batchwise {only 5 per batch}

⇓

fewer calculations per loss } ⇒ 5

but

Per epoch ← {still, 20 updates!!

More updates to weights for the entire dataset means quicker convergence!!

BGD		MBGD
10 updates	vs	200 updates
will not give optimal result		may give optimal result

} ⇒ in 10 epochs

going through same data

MBGD reaches minima faster ⇒ More updates/epoch

3) Stochastic Gradient Descent

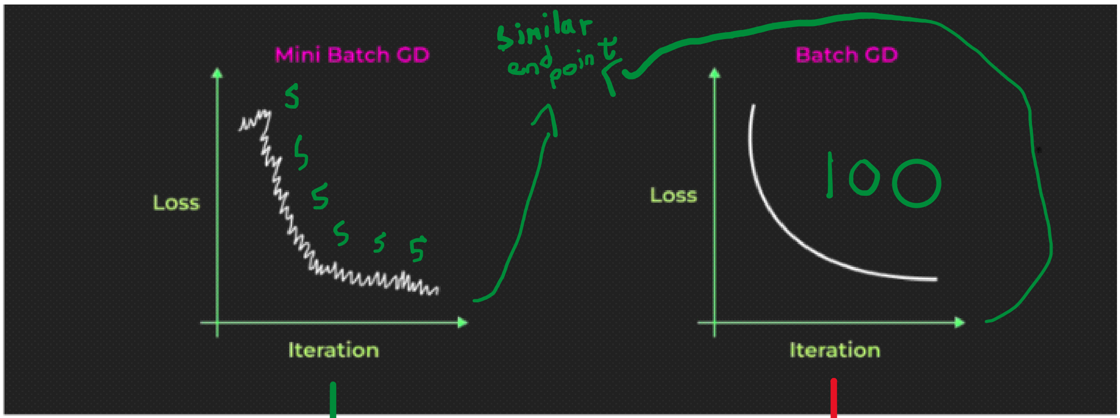
↳ only 1 point per batch, means only 1 calculation per batch for loss, but 100 updates for 1 look at the data! ⇒ 100 updates per epoch!!

{ Intuition → i) More updates = quicker convergence!
ii) Fewer calculations per loss! }

Do MBGD and SGD give the same result as regular/batch GD?

{ Not exactly!! }

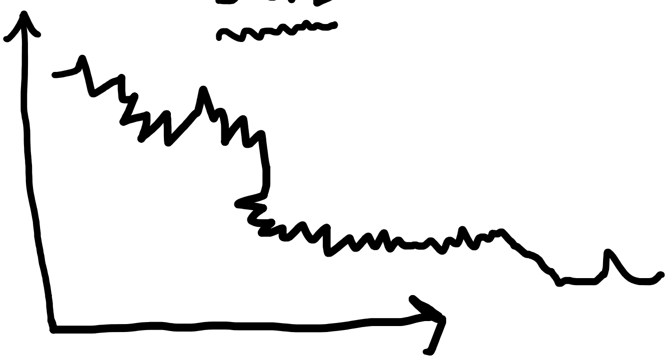
But gives a very close approximation, and reach a good local minima!



Noisier due to taking Fewer points to {calculate loss} {but Faster}

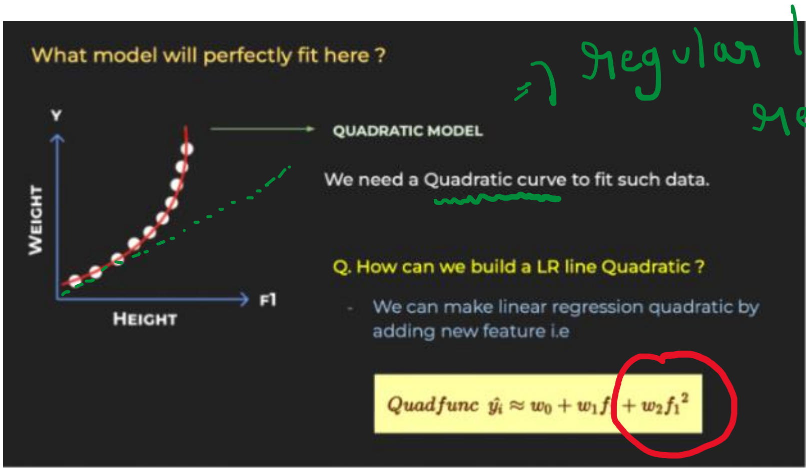
Smoother but a lot slower

SGD



Polynomial Regression

03 August 2025 20:01



⇒ regular linear regression will Fail!

$y = w_1 x + w_0$ → Cannot capture non-linear trends

$y = w_1 x_1 + w_2 x_1^2 + w_0$

⇓

This may capture non linear trend

$$y = w_1 \underset{\downarrow F_1}{x_1} + w_2 \underset{\downarrow F_2}{x_1^2} + w_0$$

degree of Polynomial

x	x^2	y
-2	4	9
-1	1	2
0	0	1
1	1	6
2	4	17

$x^2, x^3, x^4 \ldots\ldots$

Such addition of non-linear feature which make linear Regression model non- linear is called POLYNOMIAL REGRESSION

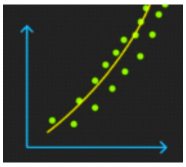
$$\left\{ y = w_1 x_1 + w_2 x_1^2 + w_3 x_1^3 + w_4 x_1^4 + w_5 x_1^5 \ldots\ldots \right\}$$

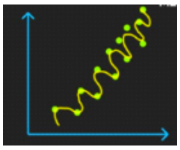
↓

captures non linear trends

Generalization and Occam's Razor

03 August 2025 21:48

 → Model 1

 → Model 2 → Learning noise

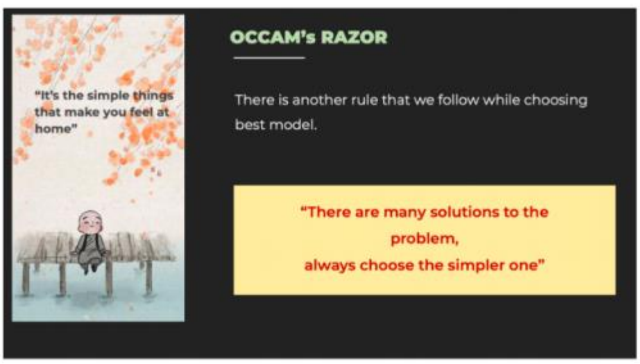
Rote learning !!

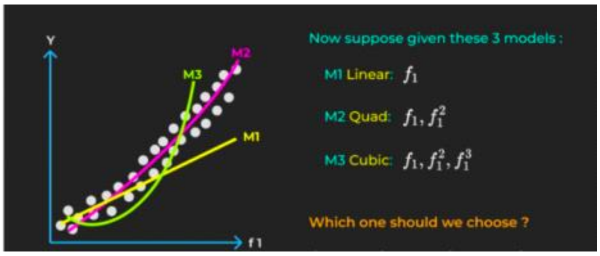
Generalization – Learn meaningful patterns for both training / testing rather than memorize the training data, learning everything including the noise.

Which of models 1 and 2 learns noise ? => 2

Occum's razor → simplest, most easily understandable models should be the preffered one.

Occam's Razor

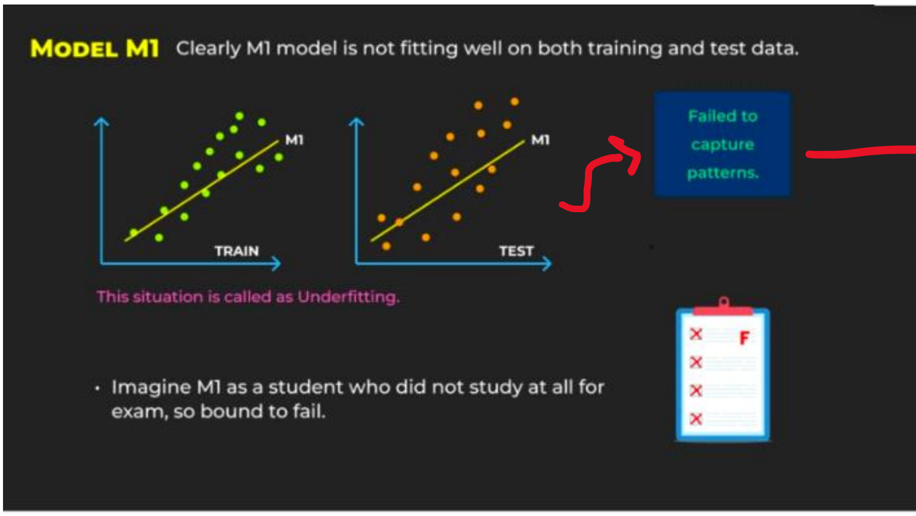




based on generalization
and occums razor.

Overfitting and Underfitting

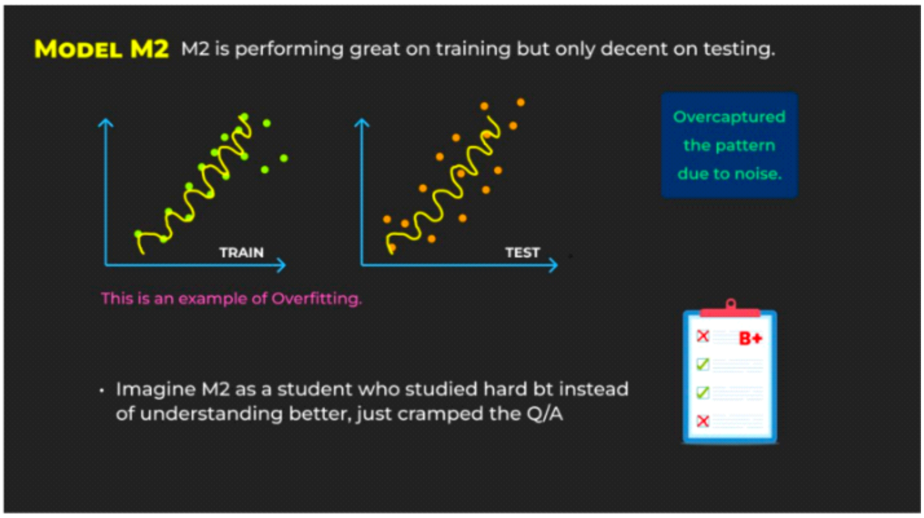
04 August 2025 17:47



in both training and testing

Under Fitting

Bad Fit!!



performs great on training but bad on testing

Over Fitting

Bias - Variance Trade off

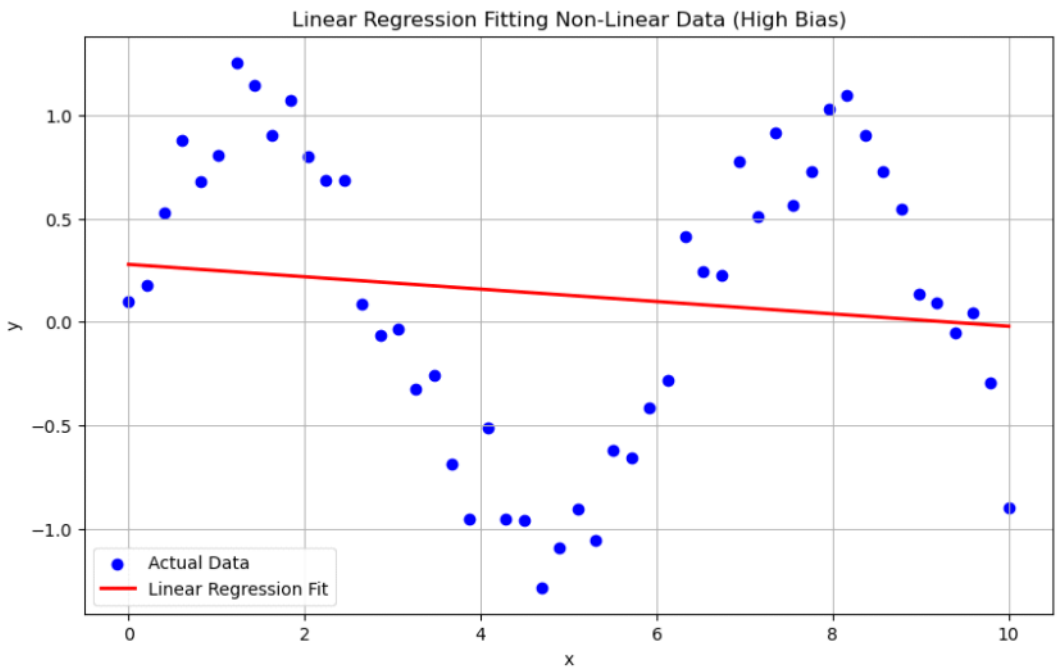
04 August 2025 17:54

High bias = Oversimplification

⇓

Common problem in Classical linear regression.

Why ?



⇒ Underfitting

IF Model / best fit / value of weights change everytime you slightly modify training data, is it a good thing or bad thing ?

⇒ inconsistency

High variance = Overcomplication!

Let us I have 1 training set.
I randomly divide points from the same training set into

Set 1

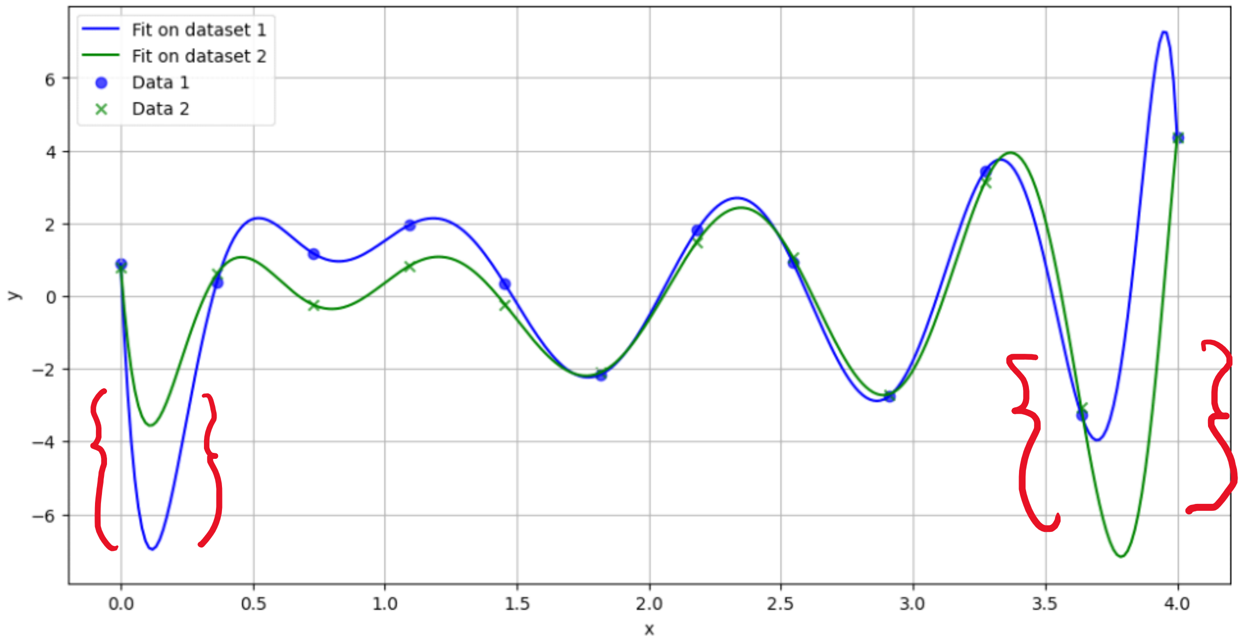
Set 2

I will train models using {polynomial regression} for both datasets. Ideally, I should get similar Models For both sets

when it has a very high degree, high variance

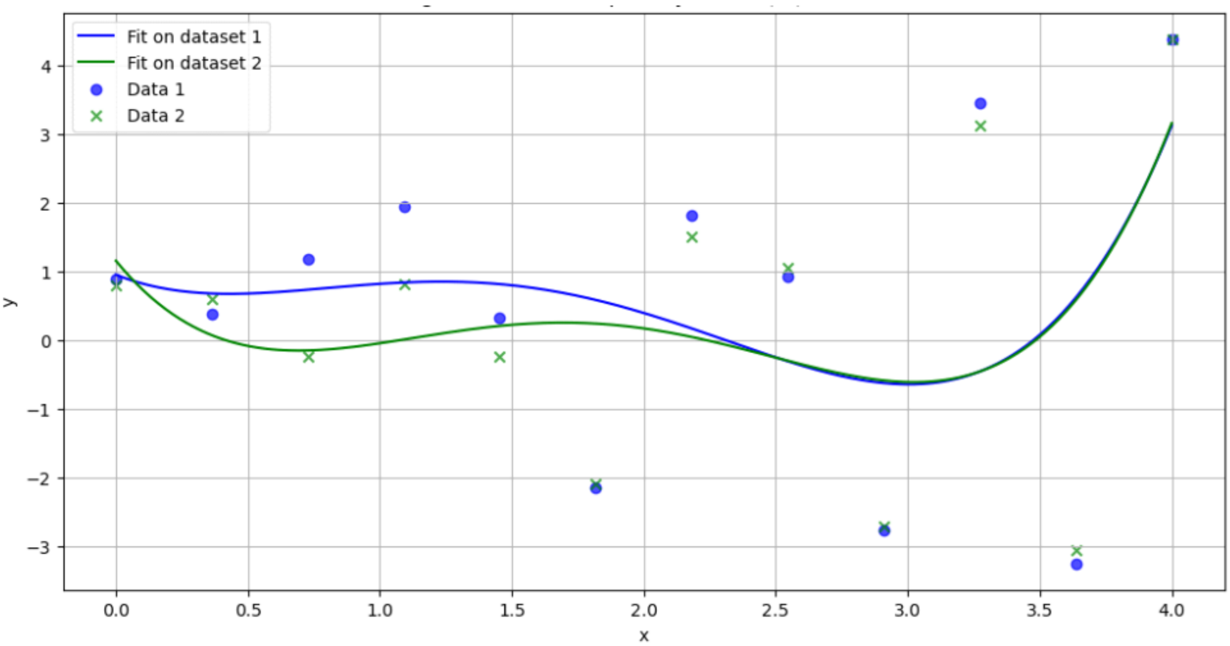
Why?

Result 1 Blue model very different from green

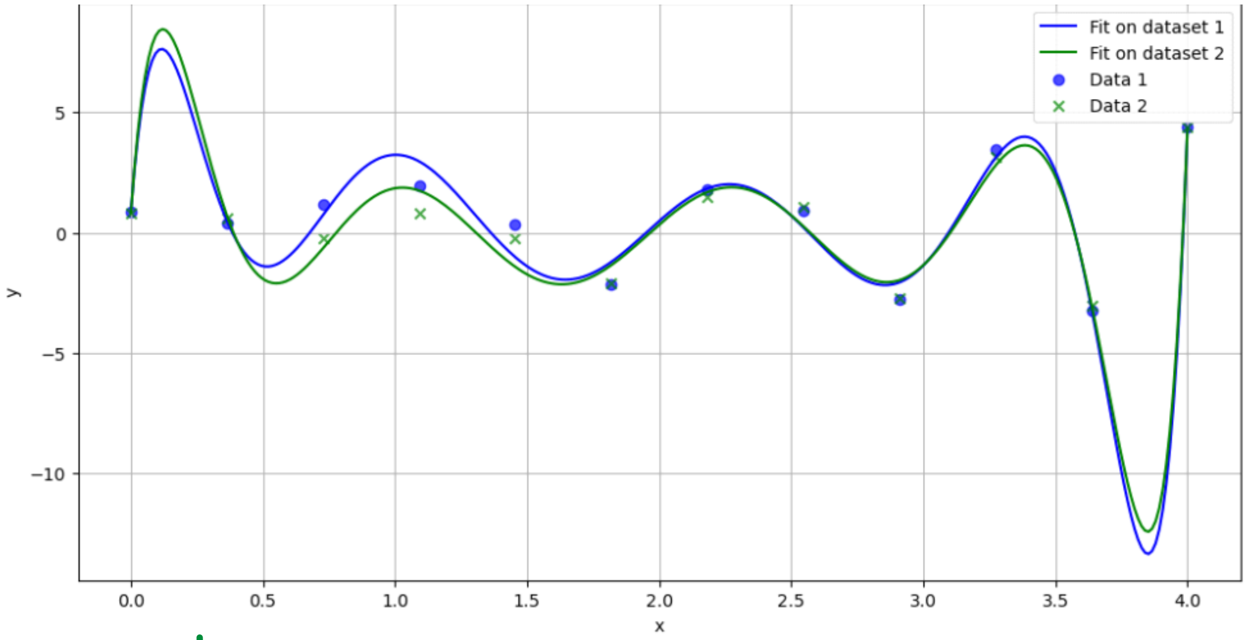


Model

Result 2 } => High bias / under fitting



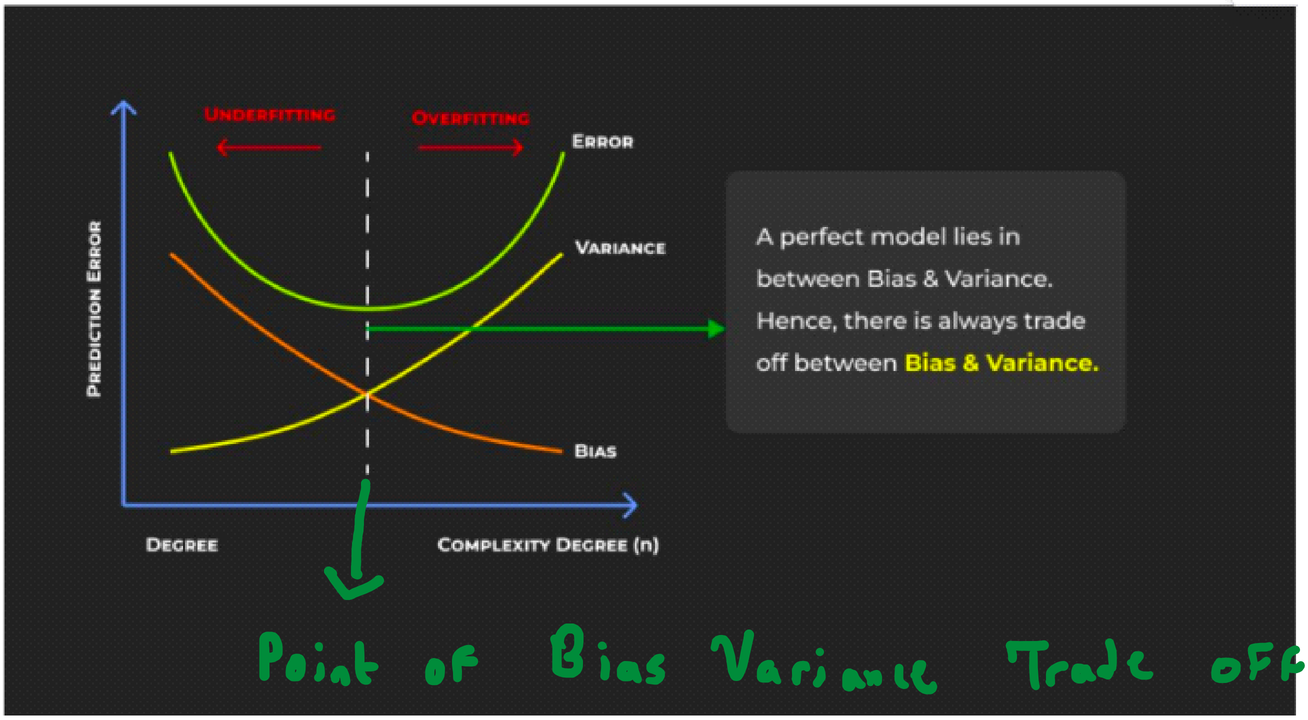
Result 3 => Just right } => Bias / Variance Trade off



Bias Variance Trade off !

Relating Bias-Variance Trade off to model fitting

04 August 2025 18:31



High Complexity => High Variance and high error

Very low complexity => High bias and high error

04 August 2025 23:07

day	No. of customers x_1	x_2 Marketing Spend	y Sales
1			
2			
3			
4			
5			

Create a new feature => M.A of customers