Bias, Variance, Regularization, Overfitting, Underfitting

→ Ridge/Lasso

→ Data

→ Split Data → Train, Test

→ Apply/fit model on train

→ Test model's acc on test

→ Cross Validation

D

Train   Test

↑

Model →

↓

	Acc. Train	Acc. Test	
1	High	High	✓
2	High	Low	→ Overfitting
3	Low	Low	→ Underfitting
4	Low	High	✗

Overfitting :- Train Acc. ↑ , Test Acc. ↓
(low bias) Train Error ↓ , Test Error ↑ (high variance)

Underfitting :- Train Acc. ↓

bias :- Training Error

variance :- Test Error

High Train error, High test error
High bias High variance

ideal case :- low bias, low variance

bias :- Difference b/w Act & predicted value

↳ less when ↳ diff is very less.

↳ Assumptions on your model

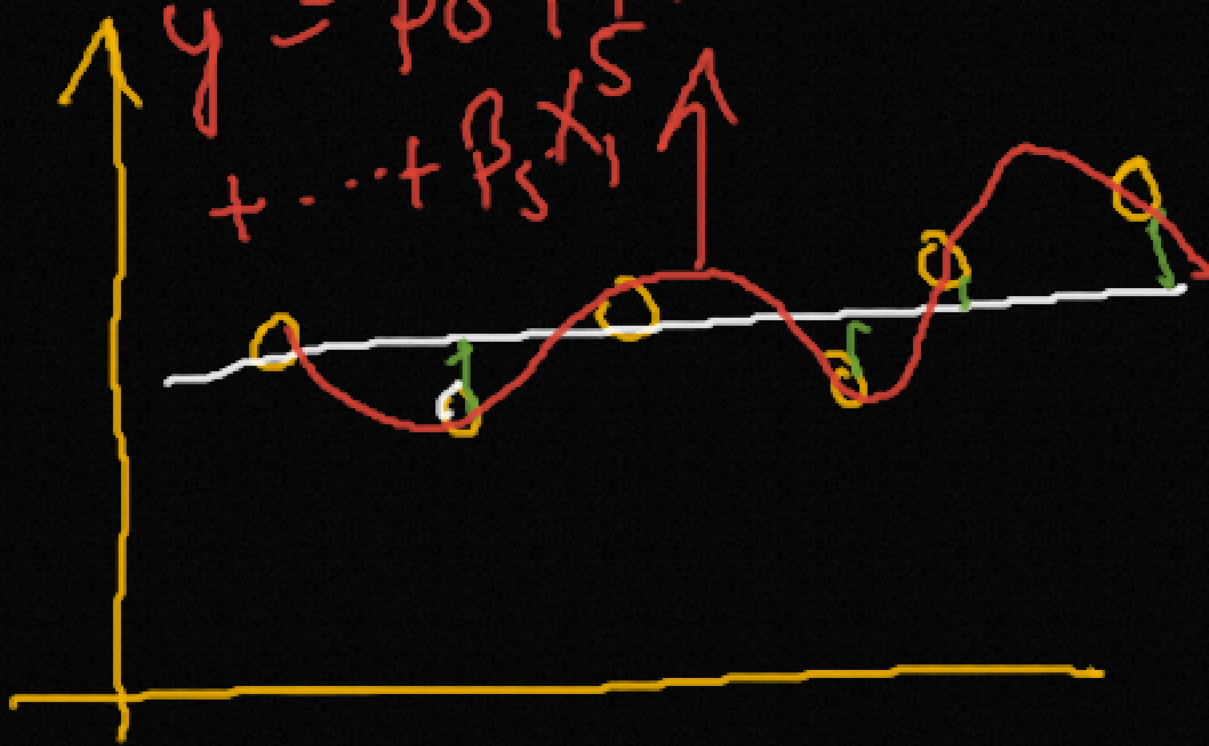
↳ less complex model → high bias

↳ very complex model → low bias (Tr. error)

$\hat{y} = \beta_0 + \beta_1 x_1$ (linear) (Complex)

$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \dots + \beta_5 x_1^5$

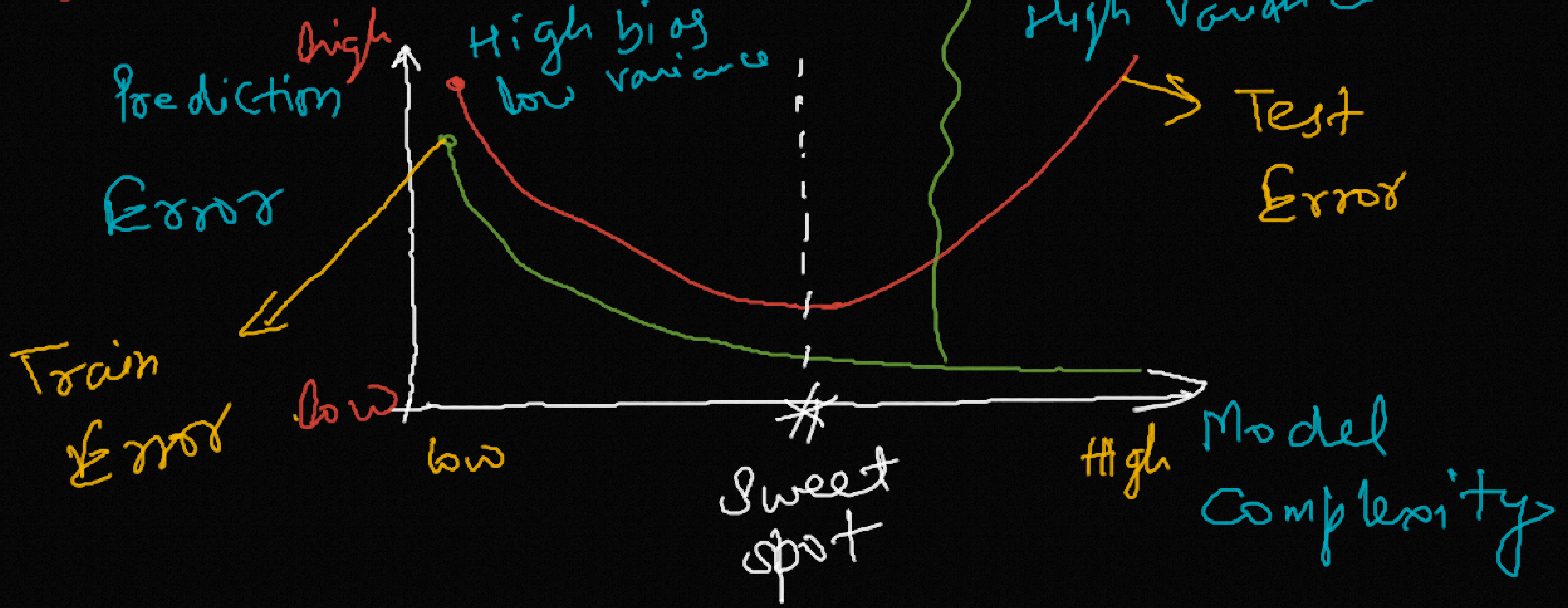
x_1	y_{Act}	y_{pred} (linear)	y_{pred} (Complex)
1	2	1	2
2	4	5	4
3	6	3	6



Variance:- model (M) is ready.

Model		D			Acc.	Error
↓						
$M \rightarrow$	Tr_1	+	Te_1	$\rightarrow$	A_1	E_1
$M \rightarrow$	Tr_2	+	Te_2	$\rightarrow$	A_2	E_2
$M \rightarrow$	Tr_3	+	Te_3	$\rightarrow$	A_3	E_3
$M \rightarrow$	Tr_4	+	Te_4	$\rightarrow$	A_4	E_4
	⋮					
$M \rightarrow$	Tr_n	+	Te_n	$\rightarrow$	A_n	E_n

Bias_Variance Trade off



Complex model
→ low bias
→ high variance

Simple model
→ high bias
→ high/low variance

Test Data

ACC.

low

↓

Test Error

↓

high → high variance

↓

Underfitting

↓

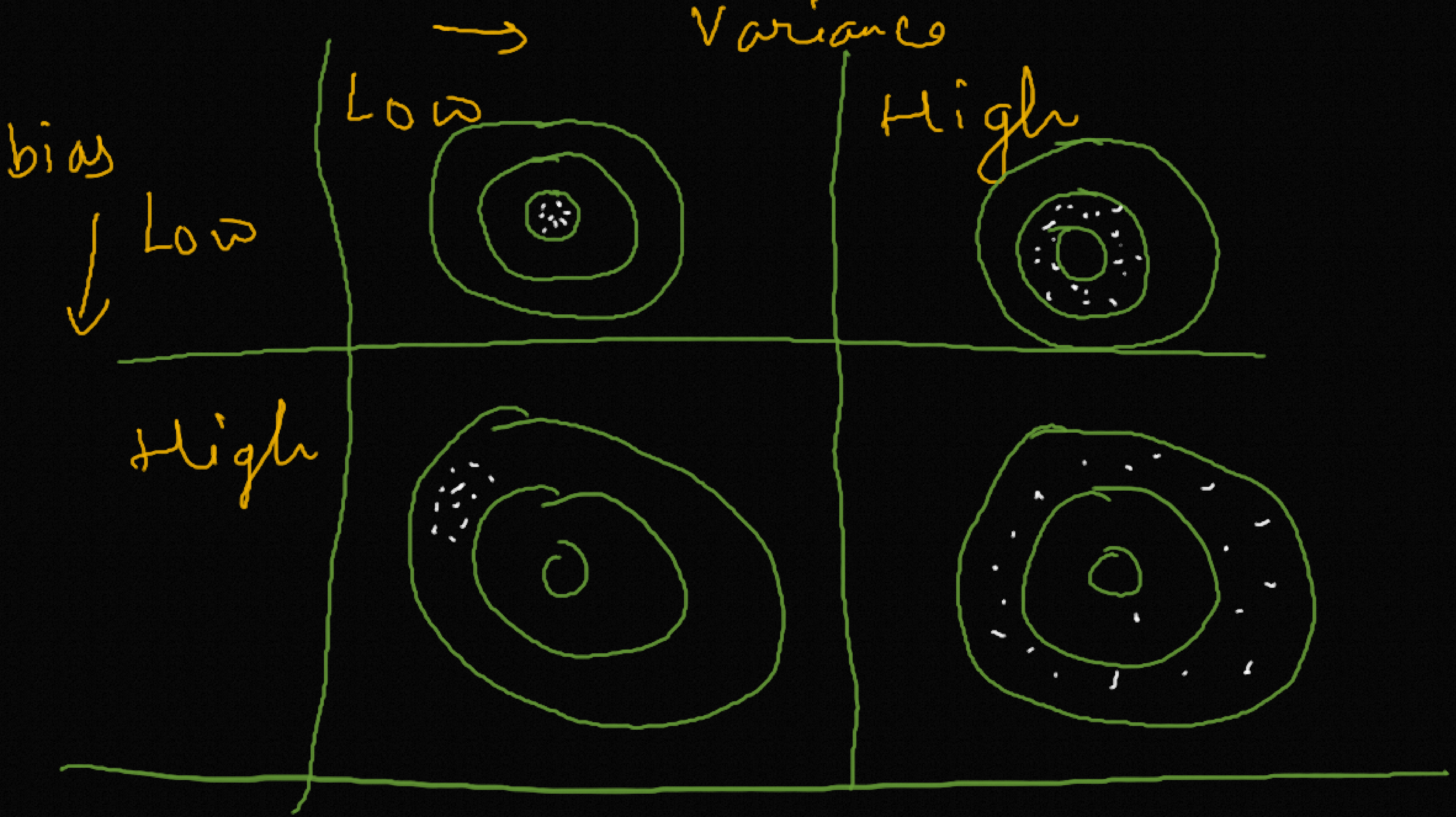
bias is high → model is simple

↑

Overfitting

↑

bias is low → model is complex.


Variance
Low
High
bias
Low
High

Overfitting :- Good Train Acc. , Poor test Acc.

SLR :- GD. $\Rightarrow$ mean squared error (loss fn)

min loss = 0

$$\hat{y} = \beta_0 + \beta_1 X_1 \Rightarrow \hat{y} \approx y$$

Error $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ $\boxed{0}$

Sol^n is **Regularization**.

① L_1 Regularization

→ Lasso Reg.

→ Cost fn $\hat{y}_i = \beta_0 + \beta_1 x_1$

$$(y_i - \hat{y}_i)^2 + \lambda_1 ||\beta_i||$$

(sum of wts)

→ Feature Selection

L1 is used for F.S. bcoz it reduces coefs of irrelevant variables to 0.

→ Sparse Nature

L1 solutions sparse

② L_2 Regularization

→ Ridge Reg. (Name)

→ Cost fn (sum of squares of wts)

$$(y_i - \hat{y}_i)^2 + \lambda_2 ||\beta||^2$$

→ Feature selection

L2 is not used! bcoz it will reduce coeff close to 0 but not exactly 0.

→ Sp. Nature

L2 does not.

L_1	L_2
L_1 is robust towards outliers.	L_2 is not.