A – OVERFITTING

Low bias, High variance

B – UNDERFITTING

High bias, High variance

C – Generalized

Low bias, Low variance

L2 Regularization (Ridge Regression)

L2 regularization, also known as Ridge Regression, is a technique used to prevent overfitting in linear regression by penalizing large coefficients.

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^{n} \theta_j^2$$

Key Ideas:

- It shrinks the coefficients (θ values) toward zero, but not exactly zero
- Helps when there's multicollinearity or too many features
- Larger λ → more regularization → simpler model

EXAMPLE –

Problem Setup:

Suppose you have a linear regression model to predict house price based on area and number of bedrooms:

$$\hat{y} = \theta_0 + \theta_1 \cdot \text{area} + \theta_2 \cdot \text{bedrooms}$$

Let's say, after training without regularization, you get:

- $\theta_1 = 500$
- $\theta_2 = 300$

But the model overfits, meaning it performs well on training data but poorly on test data.

Applying L2 (Ridge) Regularization:

Now apply L2 regularization with a penalty term:

$$J(\theta) = \text{MSE} + \lambda(\theta_1^2 + \theta_2^2)$$

Assume λ=10

So, if:

- $\theta_1 = 500$ → penalty = $10 \cdot 500^2 = 2,500,000$
- $\theta_2 = 300$ → penalty = $10 \cdot 300^2 = 900,000$

That's a huge penalty!
The model is now **forced to reduce θ□ and θ□** to make the cost smaller.

After retraining with L2:

- θ1 becomes **100**
- θ2 becomes **80**

Result:

- Smaller coefficients → less complex model
- Model may lose a bit of accuracy on training data
- But generalizes better on new/test data

L1 Regularization (Lasso Regression)

L1 regularization, also known as Lasso Regression (Least Absolute Shrinkage and Selection Operator), is a technique used to prevent overfitting by penalizing the absolute values of the model's coefficients.

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^{n} |\theta_j|$$

Key Features:

Feature	L1 (Lasso)
Penalty term	Sum of absolute values (
Effect on coefficients	Can shrink some exactly to zero
Feature selection	Yes (can eliminate features)
Useful when...	You want a sparse model