Assignment 2

Variance and Bias in Machine Learning (Underfitting, Overfitting, and Best Fit Model)

1. Introduction

In Machine Learning, the performance of a predictive model depends primarily on two fundamental sources of error:

- Bias
- Variance

These two components determine how well a model learns patterns from training data and how effectively it generalizes to unseen data.

A model that fails to properly balance bias and variance may suffer from:

- Underfitting (model too simple)
- Overfitting (model too complex)

The process of balancing these two errors is known as the Bias–Variance Tradeoff, which is one of the most important theoretical foundations in supervised learning.

Understanding bias and variance is critical because:

- It helps in selecting appropriate model complexity.
- It guides hyperparameter tuning.
- It explains why models perform differently on training and testing data.
- It forms the basis of regularization techniques.

2. Bias

2.1 Definition

Bias refers to the error introduced by approximating a real-world problem (which may be complex) with a simplified model.

In simpler terms:

Bias measures how far the model's expected predictions are from the true function.

Mathematically:

$$\text{Bias}(x) = \mathbb{E}[\hat{f}(x)] - f(x)$$

Where:

- $\hat{f}(x)$ = predicted function

- f(x) = true underlying function
- $E[\hat{f}(x)]$ = expected prediction over multiple training datasets

Bias arises due to:

- Simplified assumptions
- Incorrect model choice
- Insufficient model flexibility

2.2 High Bias

A high-bias model:

- Is too simple
- Makes strong assumptions about data
- Fails to capture complex relationships
- Produces systematic errors

Characteristics:

- Poor training accuracy
- Poor testing accuracy
- Similar errors across datasets
- Low sensitivity to training data changes

Example:

Using Linear Regression to model highly non-linear data such as:

$$y = x^3 + 2x^2 + 5$$

The linear model cannot capture curvature → high bias.

2.3 Low Bias

A low-bias model:

- Is flexible
- Captures complex patterns
- Approximates the true function well

However, if not controlled, it may result in high variance.

3. Variance

3.1 Definition

Variance refers to how much the model's predictions change when trained on different training datasets.

It measures the sensitivity of the model to fluctuations in training data.

Mathematically:

$$\text{Variance}(x) = \text{Var}(\hat{f}(x))$$

Variance reflects how much the learned function would differ if we trained it on another dataset drawn from the same distribution.

3.2 High Variance

A high-variance model:

- Is too complex
- Fits noise in training data
- Is sensitive to small changes in dataset
- Has unstable predictions

Characteristics:

- Very low training error
- High test error
- Large gap between training and test accuracy

Example:

A deep decision tree without pruning:

- Memorizes training data
- Captures random noise
- Poor generalization

3.3 Low Variance

A low-variance model:

- Produces stable predictions
- Does not fluctuate significantly with new data
- Is robust

However, if too stable and simple → may lead to high bias.

4. Underfitting and Overfitting

4.1 Underfitting (High Bias, Low Variance)

Underfitting occurs when the model is too simple to learn the underlying pattern of the data.

Causes:

- Model complexity too low
- Insufficient features
- Excessive regularization

Characteristics:

- High training error
- High testing error
- Poor pattern learning

Graphical Interpretation:

A straight line attempting to fit curved data.

Conclusion:

Underfitting = High Bias + Low Variance

4.2 Overfitting (Low Bias, High Variance)

Overfitting occurs when the model learns noise along with actual patterns.

Causes:

- Model complexity too high
- Small dataset
- Lack of regularization

Characteristics:

- Very low training error
- High testing error
- Poor generalization

Graphical Interpretation:

Curve passes through all data points in a highly irregular shape.

Conclusion:

Overfitting = Low Bias + High Variance

5. Bias–Variance Tradeoff

The total expected squared error at a point x can be decomposed as:

$$\text{Total Error} = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error}$$

Where:

1. Bias²

Error due to incorrect model assumptions.

2. Variance

Error due to sensitivity to training data.

3. Irreducible Error

Error caused by inherent noise in the data.

5.1 Effect of Model Complexity

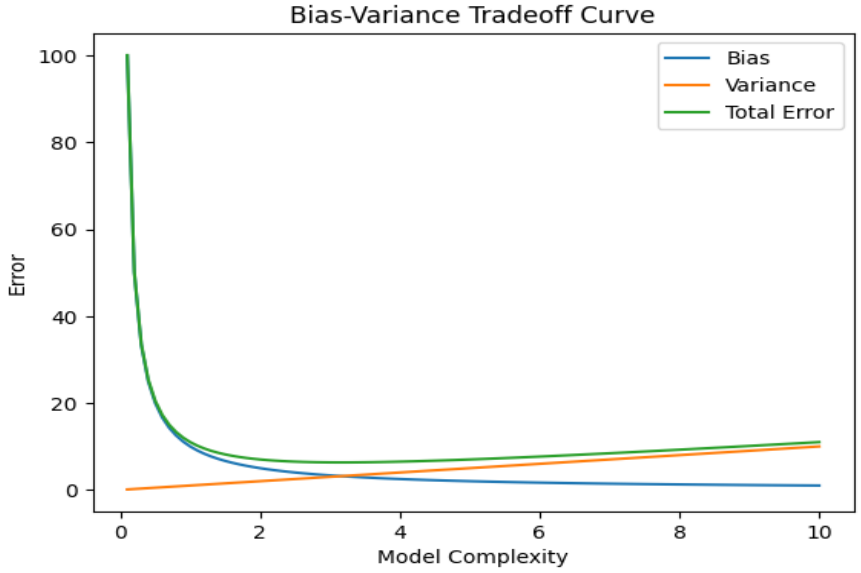
As model complexity increases:

- Bias decreases
- Variance increases

This relationship creates a U-shaped curve for total error.

The optimal model is located at the minimum of the total error curve.

6. Diagram Explanation (Bias–Variance Curve)



- X-axis → Model Complexity
- Y-axis → Error

- Bias curve (decreasing)
- Variance curve (increasing)
- Total error curve (U-shaped)

Interpretation:

- The Bias curve decreases as model complexity increases.
- The Variance curve increases as model complexity increases.
- The Total Error curve forms a U-shape.
- The minimum point of the Total Error curve represents the optimal model complexity.
- Left side → Underfitting region (High Bias).
- Right side → Overfitting region (High Variance).
- Middle point → Best Fit Model (Low Bias + Low Variance).

7. Which is the Best Fit Model?

Bias Level	Variance Level	Result
High	Low	Underfitting
Low	High	Overfitting
High	High	Worst Case
Low	Low	Best Fit

Correct Answer:

Low Bias and Low Variance

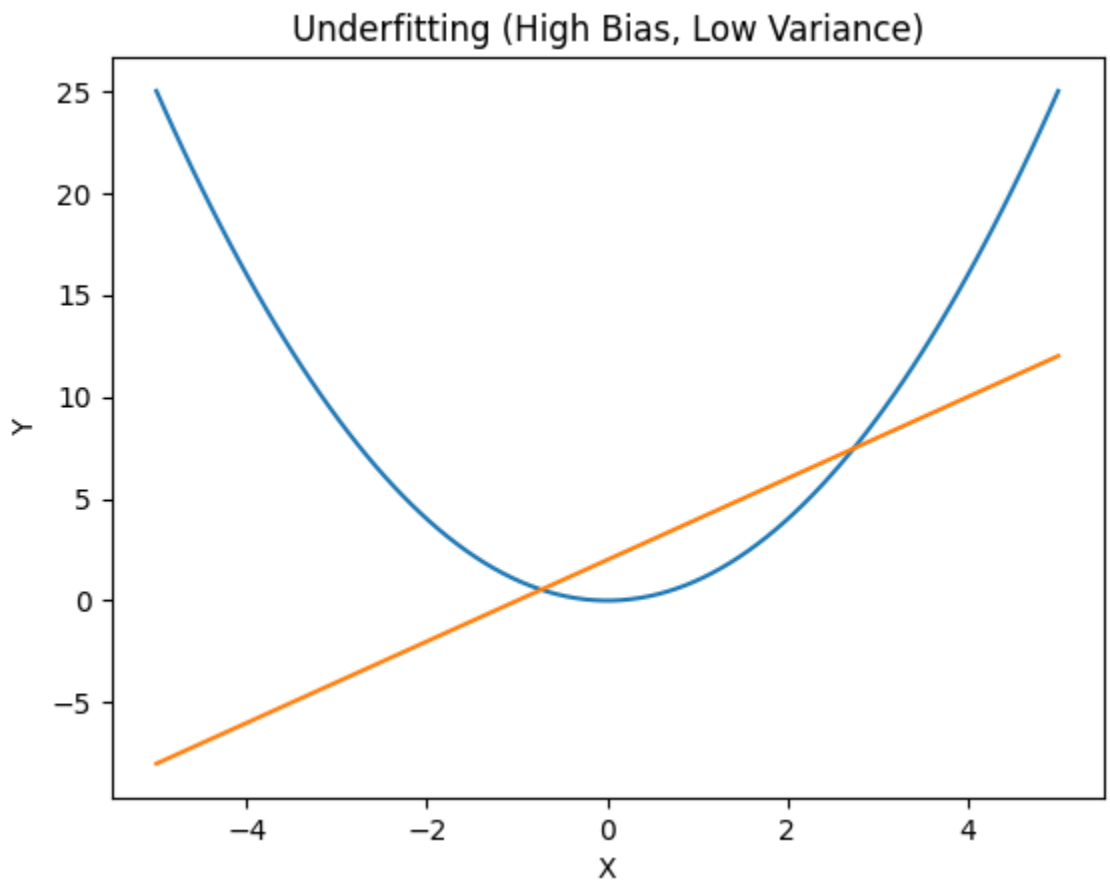
Why?

- Low bias → Captures real pattern
- Low variance → Generalizes well
- Balanced complexity
- Strong performance on both training and test data

This represents the ideal machine learning model.

8. Graphical Representation of Bias and Variance

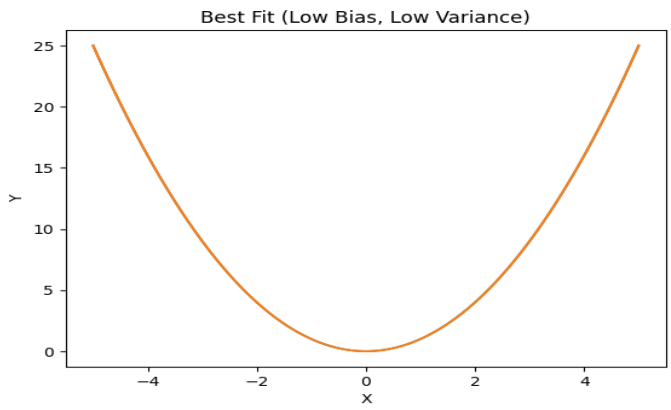
Figure 1: Underfitting (High Bias, Low Variance)



Explanation:

- The straight line fails to capture the curved pattern.
- Model is too simple.
- High training and testing error.
- This represents High Bias.
- The model makes strong assumptions and cannot learn the true pattern.

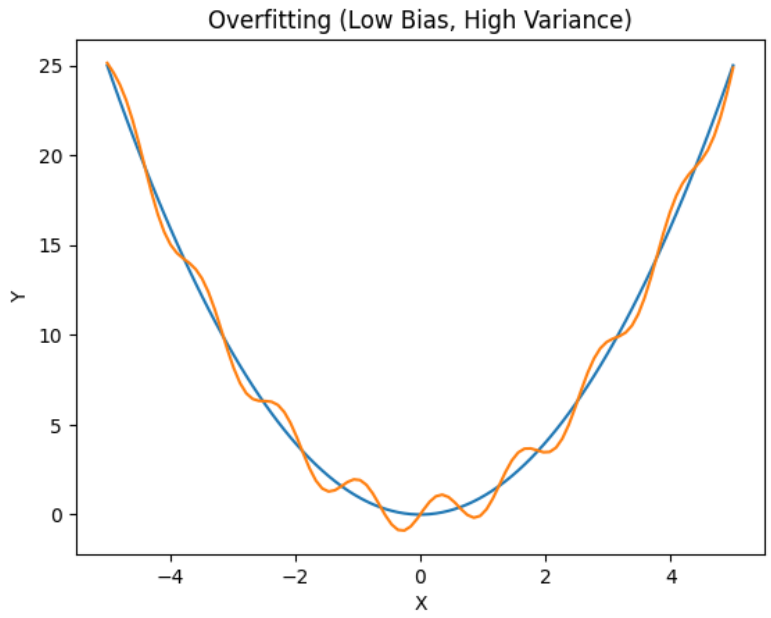
Figure 2: Best Fit (Low Bias, Low Variance)



Explanation:

- The curve perfectly matches the true pattern.
- Low training error.
- Low testing error.
- Good generalization.
- Represents an optimal balance between bias and variance.

Figure 3: Overfitting (Low Bias, High Variance)



Explanation:

- The curve fluctuates excessively.
- Fits noise in the training data.
- Very low training error.
- High testing error.
- Model is too complex.
- Represents High Variance.

9. Practical Methods to Control Bias and Variance

To Reduce Bias:

- Increase model complexity
- Add features
- Use non-linear models

To Reduce Variance:

- Increase training data
- Use regularization (L1, L2)
- Apply cross-validation
- Use pruning (for trees)
- Use ensemble methods (Random Forest)

10. Real-World Applications

Bias-variance analysis is crucial in:

- Medical diagnosis systems
- Stock market prediction
- Recommendation systems
- Image recognition
- Autonomous vehicles

Balancing bias and variance ensures:

- Reliable predictions
- Robust performance
- Better deployment readiness

11. Conclusion

Bias and variance are two fundamental components of prediction error in machine learning.

- High bias leads to underfitting.
- High variance leads to overfitting.
- The optimal model achieves a tradeoff between both.

The best-fit model should have:

- **Low Bias and Low Variance**

Such a model:

- Captures the true underlying pattern
- Avoids learning noise
- Performs well on unseen data
- Achieves optimal generalization