Assignment Q1 — Variance and Bias in Machine Learning

VARIANCE AND BIAS IN MACHINE LEARNING

ASSIGNMENT QUESTION 1

(Understanding Overfitting and Underfitting)

PAGE 1

1. INTRODUCTION TO BIAS AND VARIANCE

Question: For the best fit model, should we have:

- Low bias or high variance?
- Low bias or low variance?
- High bias or high variance?
- Low bias or high variance?

Answer: LOW BIAS AND LOW VARIANCE is the ideal combination for the best fit model.

---

2. UNDERSTANDING BIAS

Definition: Bias is the error introduced by approximating a real-world problem with a simplified model. It represents the difference between the average prediction of our model and the correct value we are trying to predict.

Characteristics of High Bias:

- Model makes strong assumptions about the data
- Oversimplifies the relationship between features and target
- Results in **underfitting**
- Poor performance on both training and test data
- Example: Linear regression on non-linear data

Characteristics of Low Bias:

- Model captures complex relationships in data
- Makes fewer assumptions about data structure
- Can fit the training data well
- Better accuracy on training set

---

3. UNDERSTANDING VARIANCE

Definition: Variance is the error introduced by the model's sensitivity to small fluctuations in the training set. It measures how much the predictions vary for different training sets.

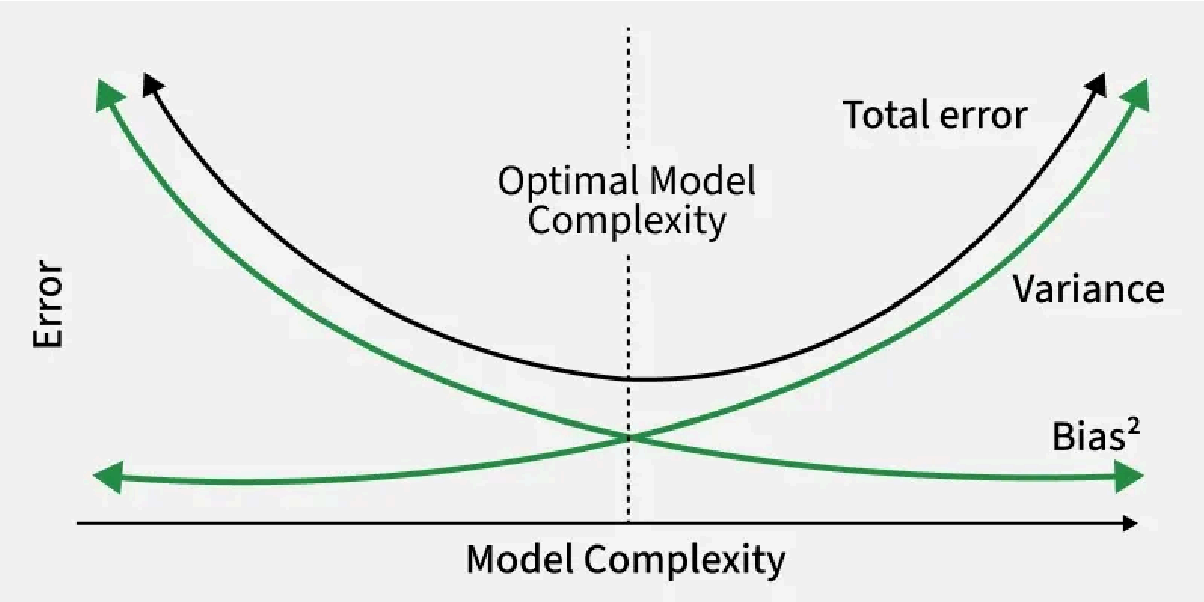
Characteristics of High Variance:

- Model is too complex
- Captures noise in the training data
- Results in **overfitting**
- Excellent performance on training data but poor on test data
- Poor generalization to new, unseen data

Characteristics of Low Variance:

- Model is stable across different training sets
- Less sensitive to fluctuations in training data
- Consistent predictions
- Better generalization

Figure 1: The Bias-Variance Tradeoff Diagram



Bias-Variance Tradeoff

This diagram shows how bias decreases and variance increases as model complexity increases. The optimal model complexity is at the point where total error is minimized.

PAGE 2

4. THE BIAS-VARIANCE TRADEOFF

The bias-variance tradeoff is a fundamental concept in machine learning that describes the tension between two sources of error:

Total Error = Bias^2 + Variance + Irreducible Error

Where:

- **Bias^2**: Systematic error from wrong assumptions
- **Variance**: Error from sensitivity to training data
- **Irreducible Error**: Noise that cannot be eliminated

---

5. OVERFITTING (High Variance, Low Bias)

Definition: Overfitting occurs when a model learns the training data too well, including its noise and random fluctuations, rather than the underlying pattern.

Signs of Overfitting:

- Very high accuracy on training data
- Poor performance on test/validation data
- Large gap between training and test error
- Model is too complex for the amount of data

Causes:

- Model complexity is too high
- Too many features/parameters
- Insufficient training data
- Training for too many epochs

Solutions:

- Use regularization techniques (L1, L2)
- Reduce model complexity
- Increase training data
- Use cross-validation
- Apply early stopping
- Feature selection

---

6. UNDERFITTING (High Bias, Low Variance)

Definition: Underfitting occurs when a model is too simple to capture the underlying pattern in the data.

Signs of Underfitting:

- Poor performance on both training and test data
- High training error

- Model is too simple
- Cannot capture data complexity

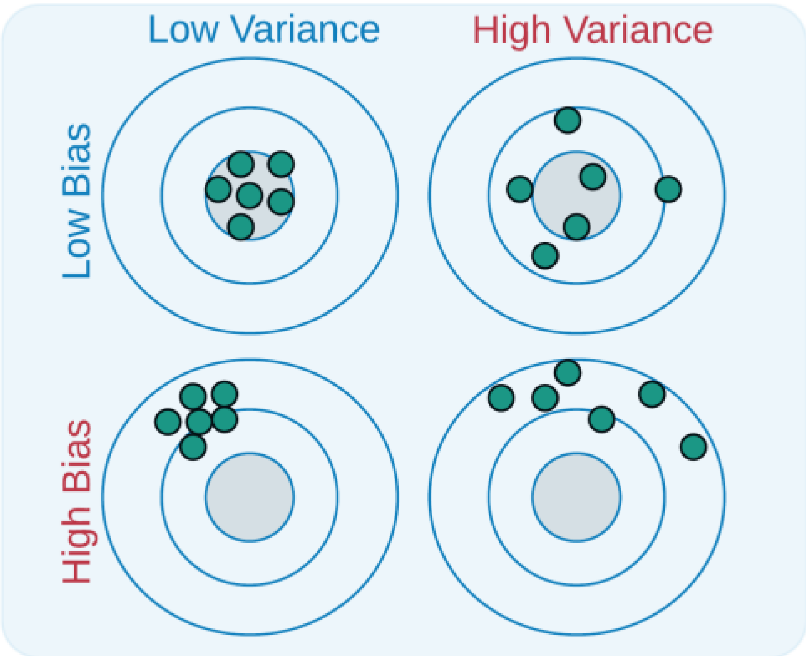
Causes:

- Model is too simple
- Insufficient features
- Too much regularization
- Limited model capacity

Solutions:

- Increase model complexity
- Add more features
- Reduce regularization
- Use more sophisticated algorithms
- Train longer

Figure 2: Visual Representation of Bias and Variance



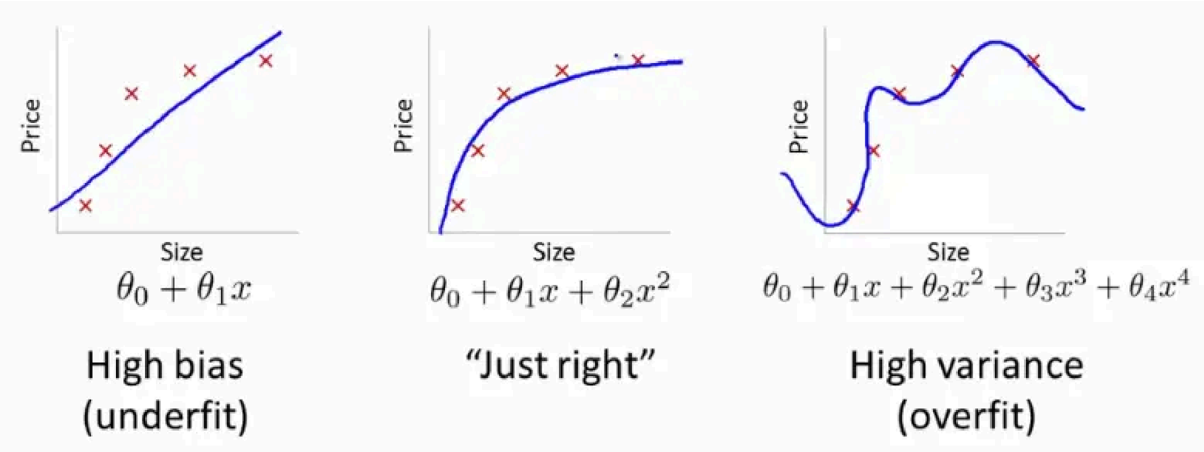
Bias Variance Visualization

This target diagram illustrates four scenarios:

- *Top-Left (Low Bias, Low Variance): Ideal — shots clustered around bullseye*
- *Top-Right (Low Bias, High Variance): Scattered but centered*
- *Bottom-Left (High Bias, Low Variance): Clustered but off-target*

- *Bottom-Right (High Bias, High Variance): Scattered and off-target*

Figure 3: Overfitting vs Underfitting in Regression



Overfitting Underfitting

Left: High bias (underfit) — linear model too simple

Middle: Just right — optimal complexity

Right: High variance (overfit) — polynomial too complex

PAGE 3

7. THE FOUR SCENARIOS EXPLAINED

Scenario 1: Low Bias, Low Variance √ (IDEAL)

- **Best case scenario**
- Model accurately predicts the target
- Consistent predictions across different datasets
- Good generalization to new data
- **This is what we aim for!**

Scenario 2: Low Bias, High Variance

- Model fits training data very well
- Predictions vary significantly with different training sets
- Overfitting occurs
- Poor generalization

Scenario 3: High Bias, Low Variance

- Model consistently makes same type of error
- Too simplistic
- Underfitting occurs
- Cannot capture patterns

Scenario 4: High Bias, High Variance

- Worst case scenario
- Model is both inaccurate and inconsistent
- Neither fits training data nor generalizes

---

8. ACHIEVING THE OPTIMAL BALANCE

To achieve **low bias and low variance**:

1. Choose the Right Model Complexity:

- Not too simple (avoid underfitting)
- Not too complex (avoid overfitting)
- Match complexity to data size and complexity

2. Use Cross-Validation:

- K-fold cross-validation
- Validate on multiple data splits
- Ensure robust performance estimates

3. Regularization Techniques:

- L1 (Lasso) regularization
- L2 (Ridge) regularization
- Dropout (for neural networks)
- Early stopping

4. Feature Engineering:

- Select relevant features
- Remove irrelevant/noisy features
- Create meaningful feature combinations

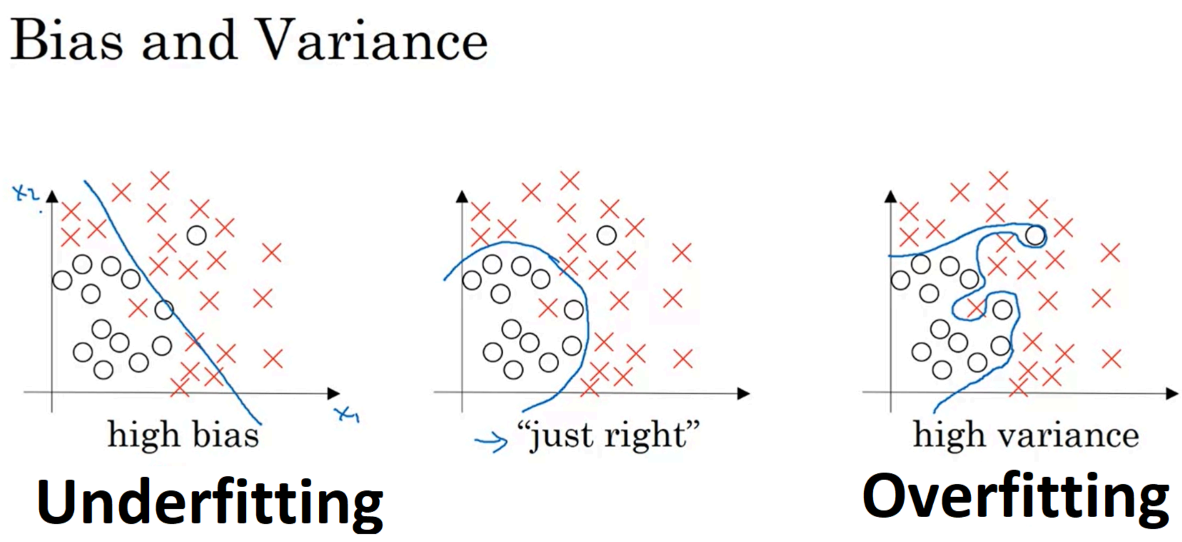
5. Ensemble Methods:

- Bagging (reduces variance)
- Boosting (reduces bias)
- Random Forests
- Gradient Boosting

6. Increase Training Data:

- More data helps reduce variance
- Helps model generalize better

Figure 4: Classification Example — Underfitting, Good Fit, and Overfitting



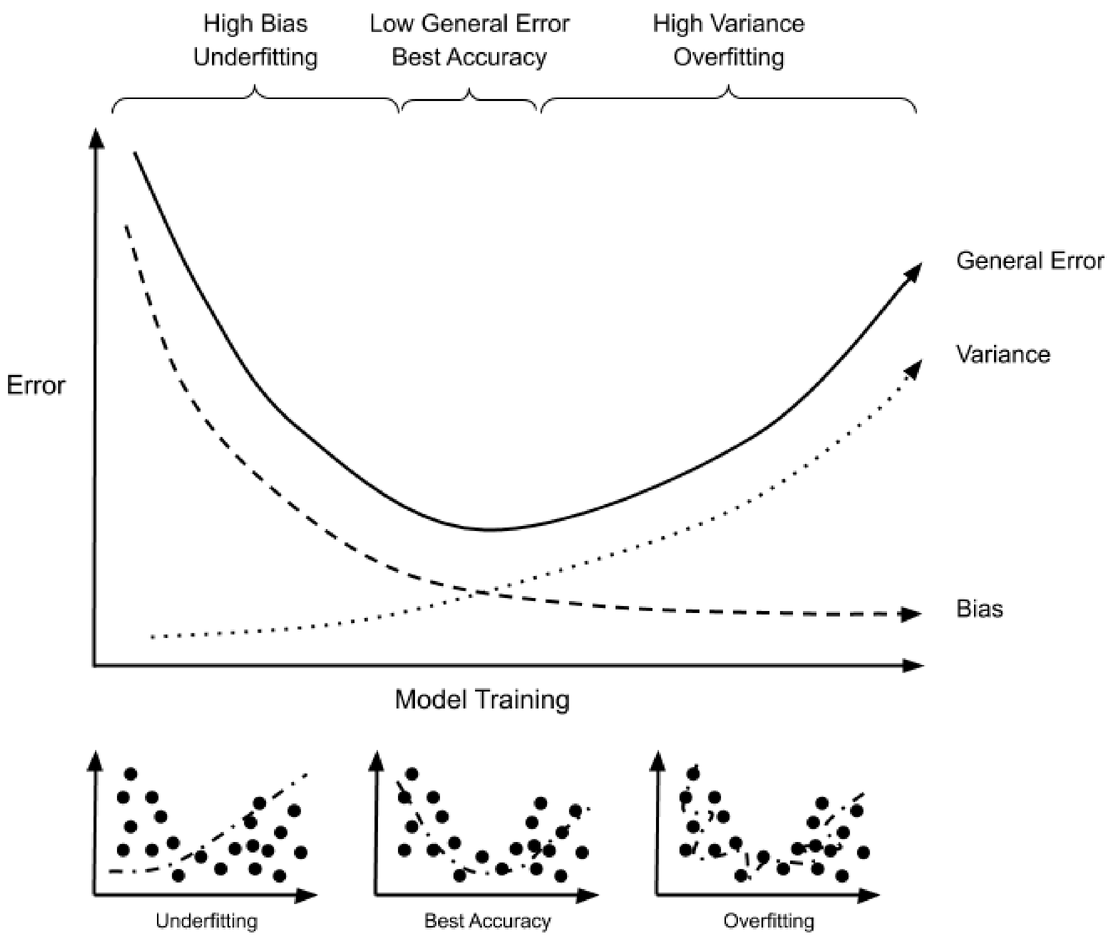
Classification Boundary

Left: High bias — decision boundary too simple

Middle: Just right — appropriate complexity

Right: High variance — boundary captures noise

Figure 5: The Optimal Balance Point



Optimal Balance

The sweet spot is at the minimum of the generalization error curve, where bias and variance are balanced optimally.

9. MATHEMATICAL FORMULATION

The expected prediction error can be decomposed as:

$E[(y - \hat{f}(x))^2] = \text{Bias}[\hat{f}(x)]^2 + \text{Var}[\hat{f}(x)] + \sigma^2$

Where:

- **$\text{Bias}[\hat{f}(x)] = E[\hat{f}(x)] - f(x)$**
  - Difference between average prediction and true value
- **$\text{Var}[\hat{f}(x)] = E[(\hat{f}(x) - E[\hat{f}(x)])^2]$**
  - Variability of predictions
- **σ^2** = Irreducible error (noise)

10. PRACTICAL EXAMPLES

Example 1: Linear Regression

- High Bias: Using linear regression on non-linear data

- Low Bias, High Variance: High-degree polynomial regression
- Low Bias, Low Variance: Appropriate degree polynomial with regularization

Example 2: Decision Trees

- High Bias: Very shallow tree (depth = 1 or 2)
- Low Bias, High Variance: Very deep tree (no pruning)
- Low Bias, Low Variance: Optimally pruned tree with cross-validation

Example 3: Neural Networks

- High Bias: Too few layers/neurons
- Low Bias, High Variance: Very large network without regularization
- Low Bias, Low Variance: Appropriate architecture with dropout and regularization

---

11. SUMMARY TABLE

Scenario	Bias	Variance	Training Error	Test Error	Problem
Underfitting	High	Low	High	High	Model too simple
Overfitting	Low	High	Very Low	High	Model too complex
Optimal	**Low**	**Low**	**Low**	**Low**	**Just right √**
Worst Case	High	High	High	Very High	Both problems

---

12. CONCLUSION

Final Answer: For the best fit model, we should have **LOW BIAS AND LOW VARIANCE**.

This represents the optimal balance where:

- The model is complex enough to capture the underlying patterns in the data (low bias)
- The model is not so complex that it captures noise and fails to generalize (low variance)
- The model performs well on both training and unseen test data
- The total error is minimized

Achieving this balance requires careful model selection, appropriate regularization, cross-validation, and understanding the tradeoff between model complexity and generalization ability.

The bias-variance tradeoff is one of the most important concepts in machine learning, and mastering it is essential for building models that generalize well to new data.

REFERENCES

1. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer.
2. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

3. Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media.
4. Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.