The Bias-Variance Tradeoff: A Deep Dive

Nipun Batra and teaching staff

IIT Gandhinagar

August 21, 2025

Table of Contents

Understanding the Problem Setup

The Learning Problem: A Real-World Example

Definition: Our Scenario

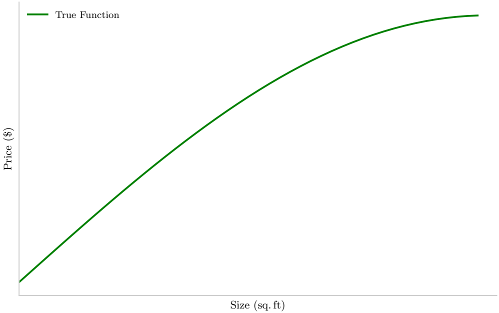
Goal: Predict housing prices based on house area

Example: The True Relationship

Unknown to us: There exists a true function $f_{\theta_{\text{true}}}$ that perfectly relates area to price:

$$y_t = f_{\theta_{\text{true}}}(\mathbf{x}_t)$$

The Learning Problem: A Real-World Example (contd.)



Key Points:

Key Challenge: We never know $f_{\theta_{\text{true}}}$ - we must estimate it from data!

The Three Sources of Prediction Error

Important: Fundamental Question

Why do our predictions fail? What causes the difference between our predictions and reality?

Definition: Three Universal Sources of Error

Every machine learning prediction suffers from:

1. **Noise** - Irreducible randomness in the data
2. **Bias** - Systematic errors from model assumptions
3. **Variance** - Sensitivity to particular training sets

Key Points:

The Tradeoff: We can often reduce bias OR variance, but not both simultaneously!

Preview: Error Decomposition

Example: Preview

Coming up: We'll see exactly how these three components combine mathematically and how to balance them.

Source 1: Noise - The Irreducible Error

Understanding Noise: The Fundamental Limitation

Definition: What is Noise?

Noise represents factors affecting the target that we cannot observe or control

Example: Real-World Noise Sources

In housing prices:

- House condition (hard to measure precisely)
- Neighborhood market dynamics
- Buyer's personal preferences

Noise: Why It's Irreducible

Example: More Noise Sources

Additional factors we cannot control:

- Economic conditions on sale day
- Unmeasurable aesthetic factors
- Random market fluctuations
- Measurement errors in data collection

Important: Key Insight

Irreducible Error: No matter how sophisticated our model, noise cannot be eliminated!

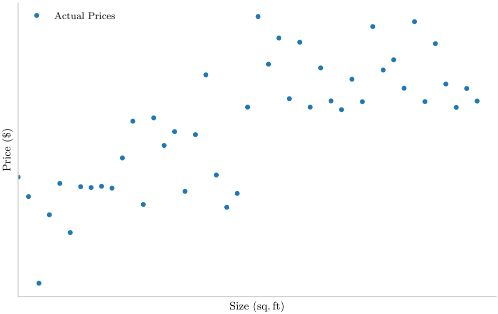
Noise: Mathematical Formulation

Key Points: Under the Noisy conditions

True relationship becomes:

$$y_t = f_{\theta_{\text{true}}}(x_t) + \epsilon_t$$

where $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$ is the noise term



Noise: Mathematical Properties

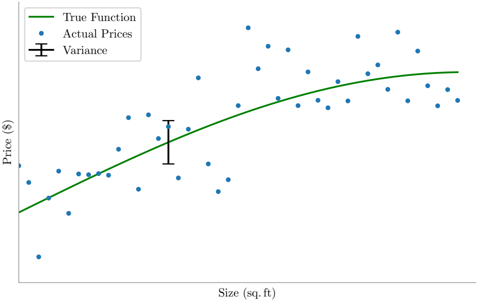
Definition: Key Properties of Noise

- **Zero mean:** $E[\epsilon_t] = 0$ (unbiased)
- **Constant variance:** $\text{Var}(\epsilon_t) = \sigma^2$
- **Independent:** Each observation's noise is independent

Key Points: Why These Properties Matter

- **Zero mean:** Noise doesn't systematically bias our target
- **Constant variance:** Prediction uncertainty is consistent
- **Independence:** One data point's noise doesn't affect others

Visualizing Noise: Data Distribution



Visualizing Noise: Data Distribution (contd.)

Key Points:

Key Observation:

- Data points scatter around the true function
- The spread (variance) is constant: σ^2
- This randomness cannot be removed by better modeling

Important: Implication for ML

Lower bound on error: Any model will have at least σ^2 error due to noise

Source 2: Bias - Systematic Model Limitations

Understanding Bias: Model Flexibility

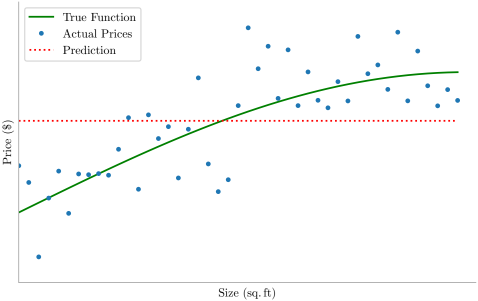
Definition: What is Bias?

Bias measures how well our model class can represent the true function

Example: Extreme Example: Constant Function

Model choice: $\hat{f}(x) = c$ (constant, regardless of house size)
Question: Can this model capture the true price-size relationship?

Bias: Visualizing the Problem



Important:

Obvious Problem: A constant function cannot capture any relationship with house size!

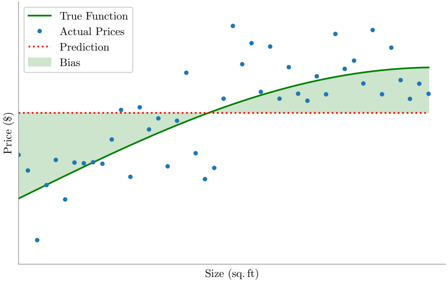
Bias: Fitting a Constant Model (contd.)

Key Points:

Best Constant Fit:

- The optimal constant is the average of all prices
- But this completely ignores the size information!
- Large systematic errors remain

Bias: Visualizing the Systematic Error



Bias: Visualizing the Systematic Error (contd.)

Definition: Bias Definition

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

The systematic difference between truth and average prediction

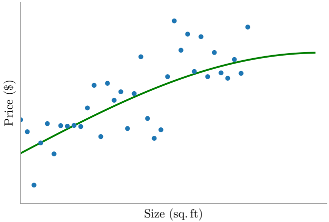
Important: Key Insight

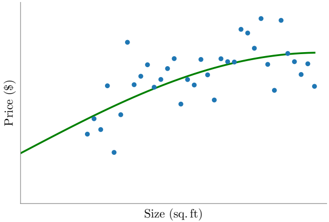
High bias = Underfitting: Model assumptions are too restrictive

Multiple Datasets: Understanding Variability

Key Points:

Crucial Insight: Many different datasets are possible from the same true relationship!





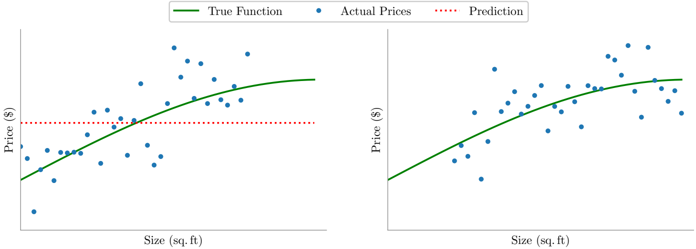
Why Datasets Differ

Example:

Same underlying relationship, different data points due to:

- Random sampling of houses
- Different noise realizations
- Natural variation in the population

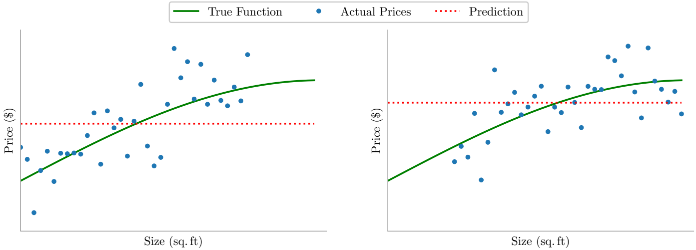
Fitting Models to Different Datasets



Key Points:

Question: If we fit the same model type (constant) to different datasets, what happens?

Different Predictions from Different Datasets



Important:

Key Observation: Even with the same model type, we get different predictions!

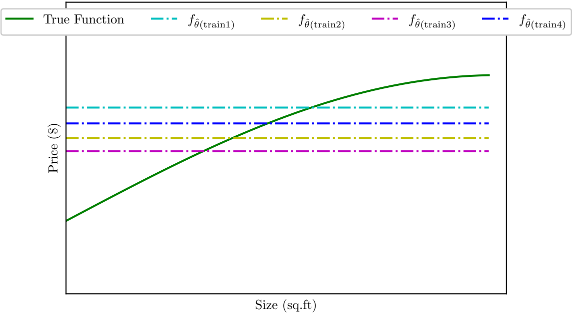
Prediction Variability: Concepts

Definition:

This variability leads us to two concepts:

- **Average prediction:** What happens "on average" across all possible datasets
- **Prediction variance:** How much predictions vary across datasets

Many Datasets: The Full Picture



Key Points:

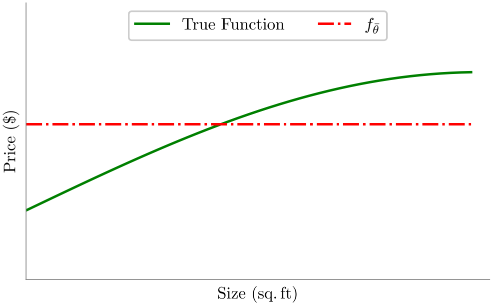
Multiple Datasets: Each gives a slightly different constant fit

Expected Prediction: The Big Question

Example:

The Big Question: What is the "typical" or "expected" prediction our model makes?

The Average Model: Expected Prediction



Expected Prediction: Definition

Definition: Expected Prediction

$$E[\hat{f}(x)] = \text{Average prediction across all possible training sets}$$

Key Points:

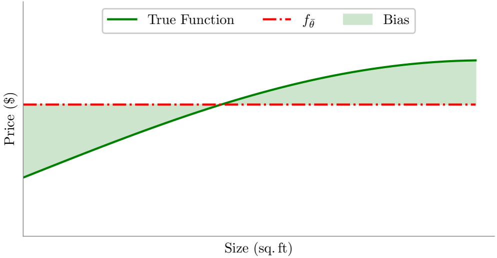
For constant models: The expected prediction is the expected value of the target variable

Bias: The Final Definition

Definition: Bias Formula

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

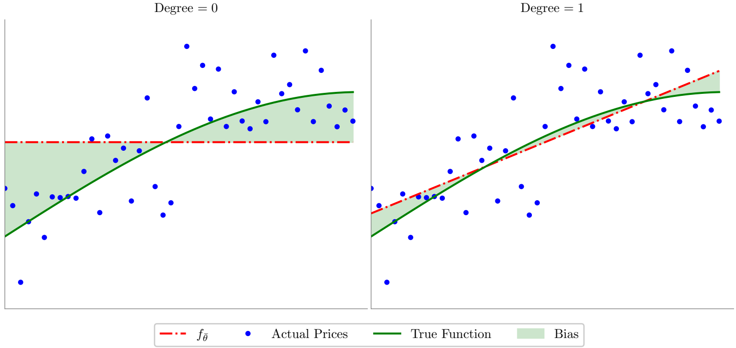
Difference between truth and expected prediction



Model Complexity vs Bias: The Relationship

Key Points:

Universal Pattern: As model complexity increases, model become flexible enough to approximate true function , hence bias decreases



Variance: Sensitivity to Data

From Bias to Variance: The Other Side

Important:

We've seen: High-complexity models have low bias
Question: If low bias is good, why not always use high-complexity models?

Definition: Enter Variance

Variance measures how much predictions change when we train on different datasets

Key Points:

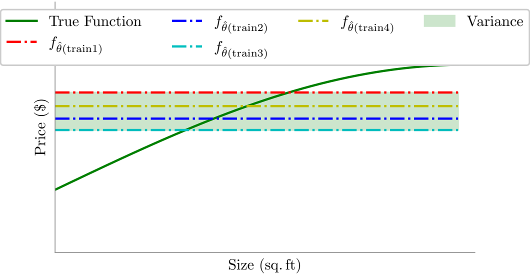
Intuition: Simple models are Stable, consistent predictions, while Complex models are highly sensitive to specific training data

Understanding Variance: Prediction Consistency

Definition: Variance Definition

Variance = How much do predictions vary across different training sets?

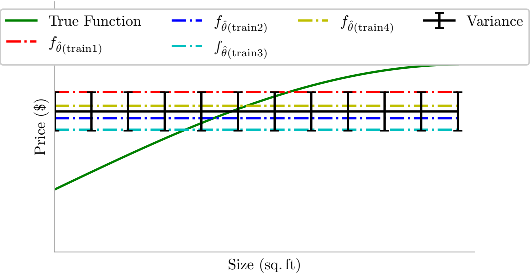
$$\text{Var}(\hat{f}(x)) = E[(\hat{f}(x) - E[\hat{f}(x)])^2]$$



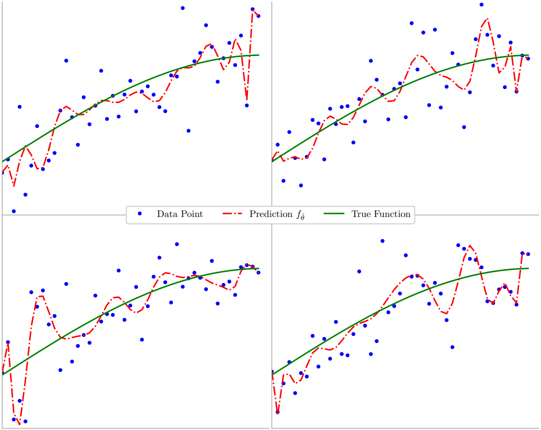
Low Complexity: Low Variance

Key Points:

Simple Models (e.g., linear): Simple model have few parameters to estimate which leads to consistent predictions across different training sets.



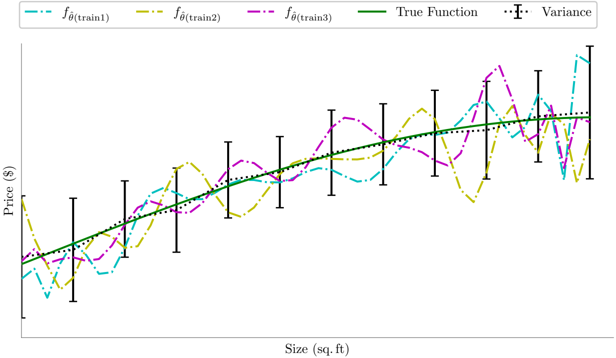
High Complexity: The Variance Problem Emerges



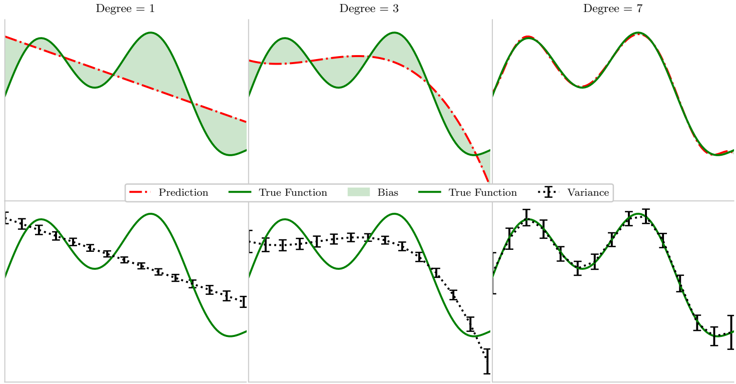
High Complexity: Extreme Variance

Key Points:

Complex Models (e.g., high-degree polynomials): Complex models have many parameters to estimate which leads to dramatic different predictions across different training sets.



The Bias-Variance Tradeoff: The Central Tension



The Bias-Variance Tradeoff: The Central Tension

Important: The Fundamental Tradeoff

- **Simple models:** High bias, low variance
- **Complex models:** Low bias, high variance
- **Optimal complexity:** Balance between the two

Key Points:

Key Insight: We cannot minimize both bias and variance simultaneously!

Mathematical Decomposition

Why Mathematical Analysis Matters

Definition: The Goal

Can we mathematically prove that prediction error can be expressed as a function of bias, variance, and noise?

Specifically, can we show:

$$\text{error} = E\left[(y - \hat{f}(x))^2\right] = \text{function of bias, variance, and noise}$$

Key Points: Why This Matters

- Understand the fundamental limits of learning
- Make informed model and algorithm choices
- Explicitly balance bias and variance

Bias-Variance Decomposition: The Goal

Definition: What We Want to Prove

$$\text{error} = E\left[(y - \hat{f}(x))^2\right] = \text{function of bias, variance, and noise}$$

Key Points: Strategy

1. Start with squared error at a single point
2. Take expectation over all randomness (training set and noise)
3. Use algebraic tricks to separate terms
4. Identify noise, bias, and variance

Step 1: The Squared Error

Definition: Squared Loss at x

Prediction error: $(y - \hat{f}(x))^2$

Key Points: Taking Expectations

Expected error:

$$E_{\mathcal{D},y}[(y - \hat{f}(x))^2]$$

where:

- $\mathcal{D}$: Random training set
- y: Random target (includes noise)

Step 2: Add and Subtract the True Function

Example: The Trick

Add and subtract $f_{\text{true}}(x)$ inside the square:

$$E[(y - f_{\text{true}}(x) + f_{\text{true}}(x) - \hat{f}(x))^2]$$

Key Points: Earlier seen : Under Noisy conditions

True relationship becomes:

$$y_t = f_{\theta_{\text{true}}}(x_t) + \epsilon_t$$

Definition: Grouping Terms

$$E\left[\underbrace{(y - f_{\text{true}}(x))}_{\epsilon} + \underbrace{(f_{\text{true}}(x) - \hat{f}(x))}_{\text{prediction error}}\right]^2$$

Step 3: Expand the Square

Example: Algebraic Expansion

Let $a = \epsilon$, $b = f_{\text{true}}(x) - \hat{f}(x)$:

$$(a + b)^2 = a^2 + 2ab + b^2$$

So,

$$E[\epsilon^2 + 2\epsilon(f_{\text{true}}(x) - \hat{f}(x)) + (f_{\text{true}}(x) - \hat{f}(x))^2]$$

Key Points: Linearity of Expectation

$$E[\epsilon^2] + 2E[\epsilon(f_{\text{true}}(x) - \hat{f}(x))] + E[(f_{\text{true}}(x) - \hat{f}(x))^2]$$

Step 4: Identify the Three Terms

Definition: Three Terms

- **Term 1:** $E[\epsilon^2]$ (noise)
- **Term 2:** $2E[\epsilon(f_{\text{true}}(x) - \hat{f}(x))]$ (cross-term)
- **Term 3:** $E[(f_{\text{true}}(x) - \hat{f}(x))^2]$ (prediction error)

Key Points: Next Steps

Analyze each term separately to reveal noise, bias, and variance.

Step 5: Analyzing Term 1 (Noise)

Definition: Term 1

$\epsilon = y - f_{\text{true}}(x)$ is the noise.

Recall how variance is defined:

$$\begin{aligned}\text{Var}(\epsilon) &= \mathbb{E}\left[(\epsilon - \mathbb{E}[\epsilon])^2\right] \\ &= \mathbb{E}\left[\epsilon^2 - 2\epsilon\,\mathbb{E}[\epsilon] + (\mathbb{E}[\epsilon])^2\right] \\ &= \mathbb{E}[\epsilon^2] - 2\mathbb{E}[\epsilon]\mathbb{E}[\epsilon] + (\mathbb{E}[\epsilon])^2 \\ &= \mathbb{E}[\epsilon^2] - (\mathbb{E}[\epsilon])^2\end{aligned}$$

So, $\mathbb{E}[\epsilon^2] = \text{Var}(\epsilon) + (\mathbb{E}[\epsilon])^2$.
For our noise, $\mathbb{E}[\epsilon] = 0$ and $\text{Var}(\epsilon) = \sigma^2$, so $\mathbb{E}[\epsilon^2] = \sigma^2 + 0^2 = \sigma^2$.
$\boxed{\text{Term } 1 = \sigma^2}$, **This is the irreducible error (noise)!**

Step 6: Analyzing Term 2 (Cross-Term)

Definition: Term 2

$2E[\epsilon(f_{\text{true}}(x) - \hat{f}(x))]$

Key Points: Key Insight

ϵ (noise) is independent of $\hat{f}(x)$ (model prediction), so:

$$E[\epsilon(f_{\text{true}}(x) - \hat{f}(x))] = E[\epsilon] \cdot E[f_{\text{true}}(x) - \hat{f}(x)] = 0$$

Important: Result

$\boxed{\text{Term } 2 = 0}$

The cross-term vanishes!

Step 7: Analyzing Term 3 (Prediction Error)

Definition: Term 3

$E[(f_{\text{true}}(x) - \hat{f}(x))^2]$
This is the mean squared error of the model's prediction.

Key Points: Next Step

Decompose this term into bias and variance using another add-and-subtract trick.

Step 8: Add and Subtract the Expected Prediction

Example: The Trick

Add and subtract $E[\hat{f}(x)]$:

$$E[(f_{\text{true}}(x) - E[\hat{f}(x)] + E[\hat{f}(x)] - \hat{f}(x))^2]$$

Key Points: Grouping

$$(\text{bias}) + (\text{variance deviation})$$

Step 9: Expand and Separate Terms

Example: Expand the Square

Let $\alpha = f_{\text{true}}(x) - E[\hat{f}(x)]$ (bias)
$\beta = E[\hat{f}(x)] - \hat{f}(x)$ (variance deviation)

$$E[(\alpha + \beta)^2] = E[\alpha^2] + 2E[\alpha\beta] + E[\beta^2]$$

Step 10: Analyze Each Term

Definition: Three Terms

- $E[\alpha^2]$ (bias squared)
- $2E[\alpha\beta]$ (cross-term)
- $E[\beta^2]$ (variance)

Step 11: Bias Squared

Key Points: Bias Term

α is deterministic (not random)!

- $f_{\text{true}}(x)$ is a fixed function value
- $E[\hat{f}(x)]$ is the expected prediction (will become a constant after the distribution is defined)

so $E[\alpha^2] = (f_{\text{true}}(x) - E[\hat{f}(x)])^2 = [\text{Bias}(x)]^2$

Important: Result

$$\boxed{E[\alpha^2] = [\text{Bias}(x)]^2}$$

Step 12: Cross-Term

Key Points: Cross-Term

α is constant, so $E[\alpha\beta] = \alpha \cdot E[\beta]$.
But $E[\beta] = E[E[\hat{f}(x)] - \hat{f}(x)] = 0$,the expected deviation of a random variable from its mean is zero, so the cross-term is zero.

Important: Result

$\boxed{2E[\alpha\beta] = 0}$, the cross-term vanishes!

Step 13: Variance Term

Key Points: Variance

$$E[\beta^2] = E[(E[\hat{f}(x)] - \hat{f}(x))^2] = E[(\hat{f}(x) - E[\hat{f}(x)])^2] = \text{Variance}(\hat{f}(x))$$

Important: Result

$$\boxed{E[\beta^2] = \text{Variance}(\hat{f}(x))}$$

Step 14: The Complete Decomposition

Important: Putting It All Together

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

Definition: Component Summary

- σ^2 = **Irreducible error** (noise)
- $[\text{Bias}(x)]^2$ = **Systematic error** (model assumptions)
- $\text{Variance}(\hat{f}(x))$ = **Random error** (training set sensitivity)

The Fundamental Tradeoff

Key Points: The Fundamental Tradeoff

- **Reduce bias:** Use more complex models $\rightarrow$ Increase variance
- **Reduce variance:** Use simpler models $\rightarrow$ Increase bias
- **Optimal complexity:** Minimize bias^2 + variance

Summary and Applications

Summary: The Bias-Variance Tradeoff

Definition: What We've Proven

Every prediction error can be decomposed as:

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

Key Points: Key Takeaways

- **Noise:** Cannot be reduced (irreducible)
- **Bias:** Reduced by increasing model complexity
- **Variance:** Reduced by decreasing model complexity
- **Optimal model:** Balances bias and variance

Bias-Variance Tradeoff: Practical Applications

Important: Practical Applications

- **Model selection:** Choose complexity to minimize total error
- **Ensemble methods:** Reduce variance while maintaining low bias
- **Regularization:** Explicitly control the bias-variance tradeoff
- **Cross-validation:** Estimate the full error decomposition