

$$\begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ 0 & : & \sigma_n \end{bmatrix}$$

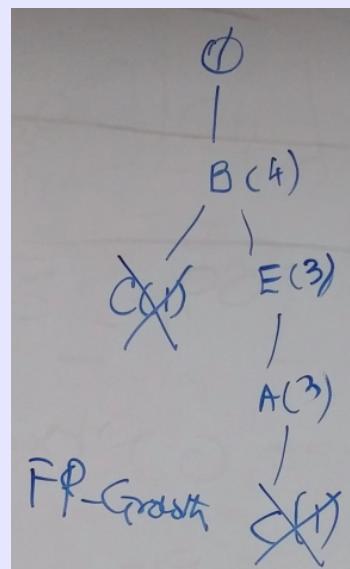
$$X = \sum_{i=1}^{\text{rank}(X)} \sigma_i u_i v_i^T = U \Sigma V^T$$

Annotations:

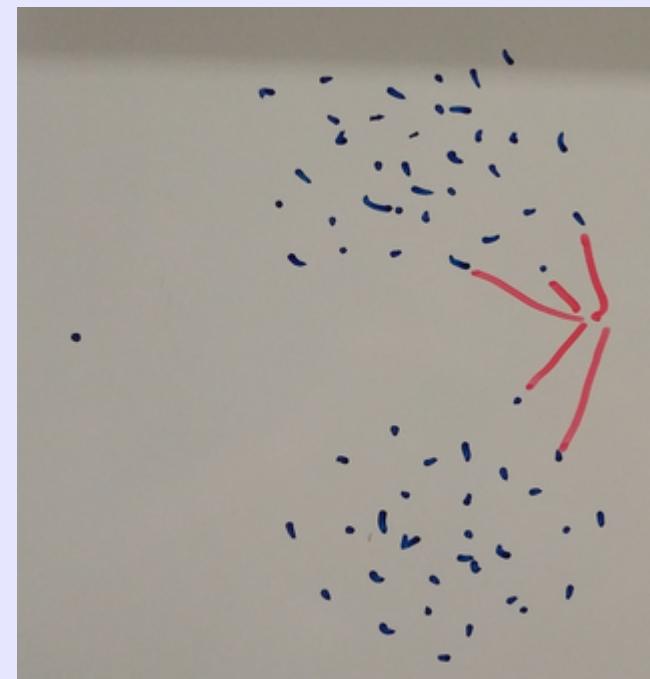
- σ^i \leftarrow i^{th} singular value of X
- u_i \leftarrow i^{th} left singular value of X (i^{th} column of U)
- v_i^T \leftarrow i^{th} right singular vector of X (i^{th} column of V^T)
- Captures the patterns among attributes
- Captures the patterns among the objects

CS 422: Data Mining
Vijay K. Gurbani, Ph.D.,
Illinois Institute of Technology

Bias-variance tradeoff

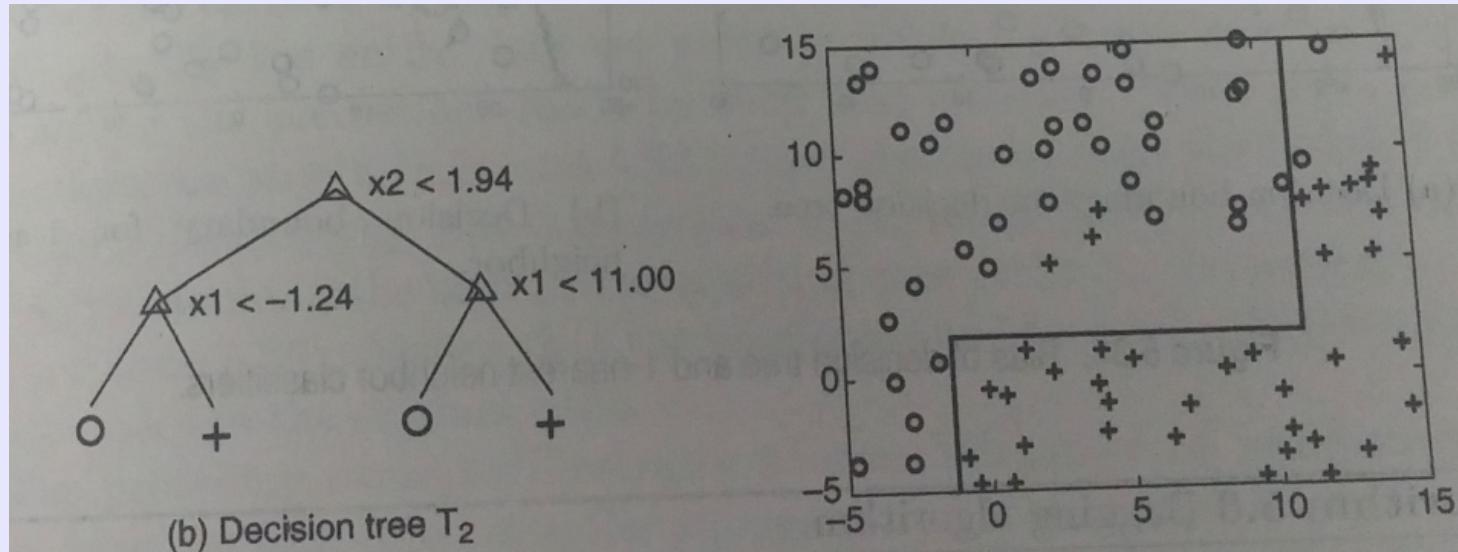


CS 422-04
vgurbani@iit.edu



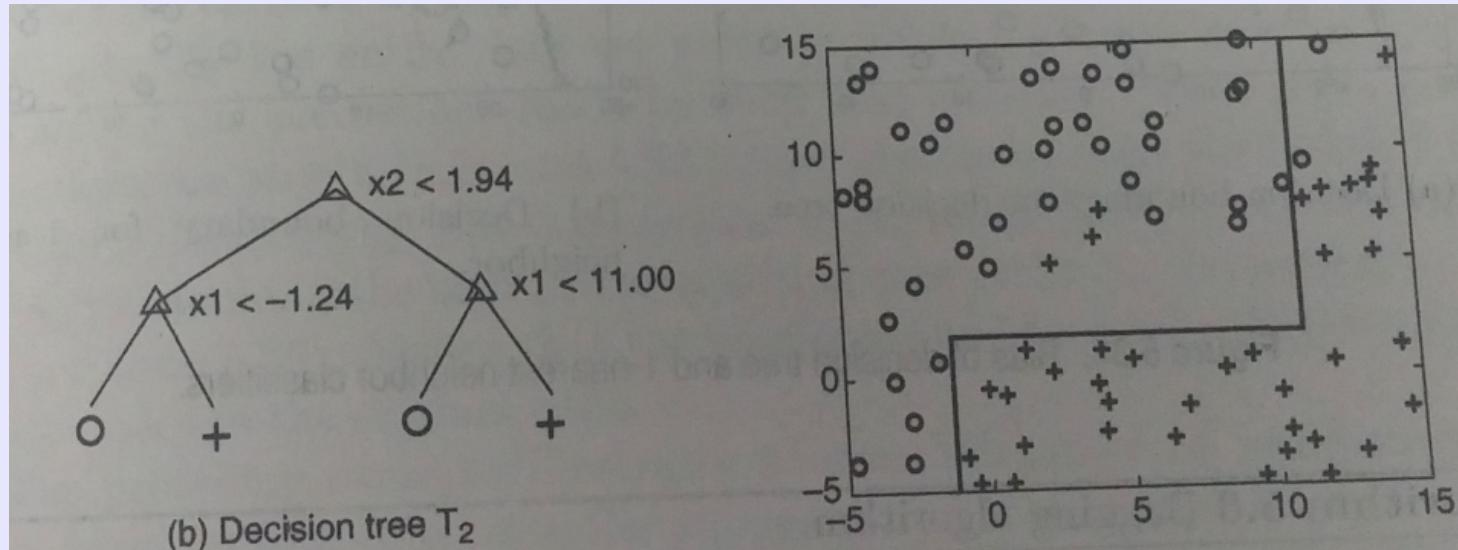
Bias, Variance, Under/Over-fitting

High Bias

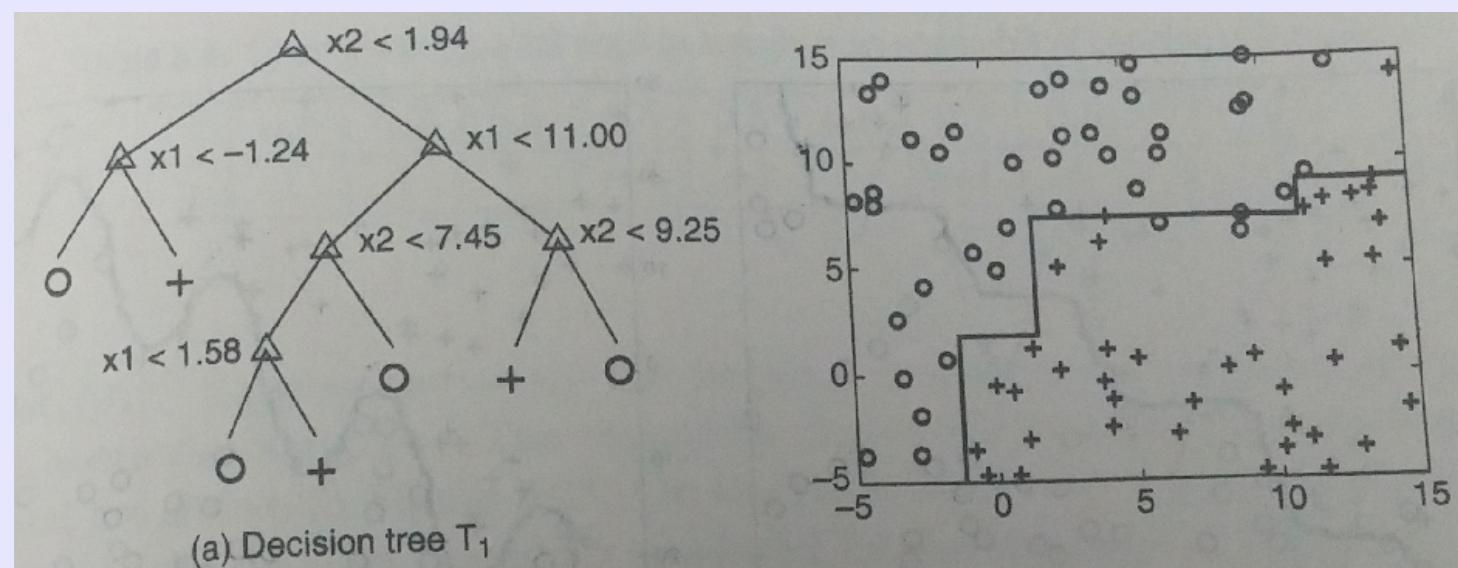


Bias, Variance, Under/Over-fitting

High Bias

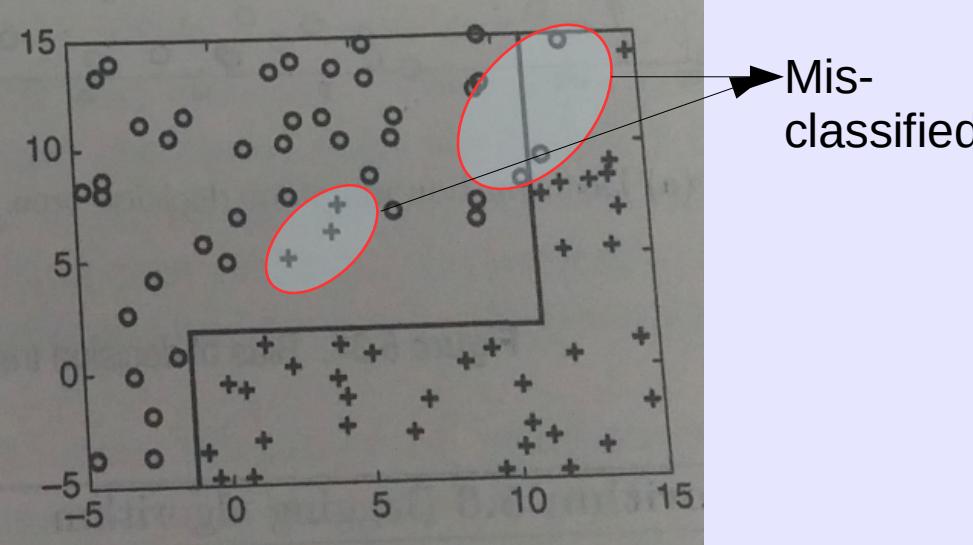
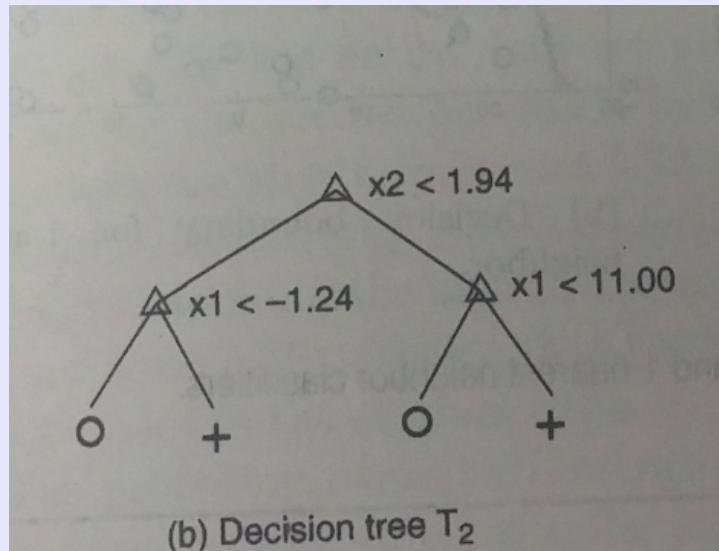


Low Bias

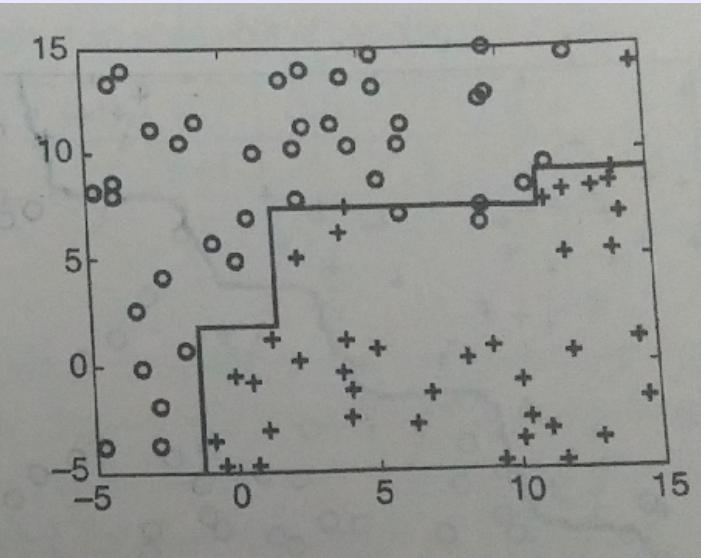
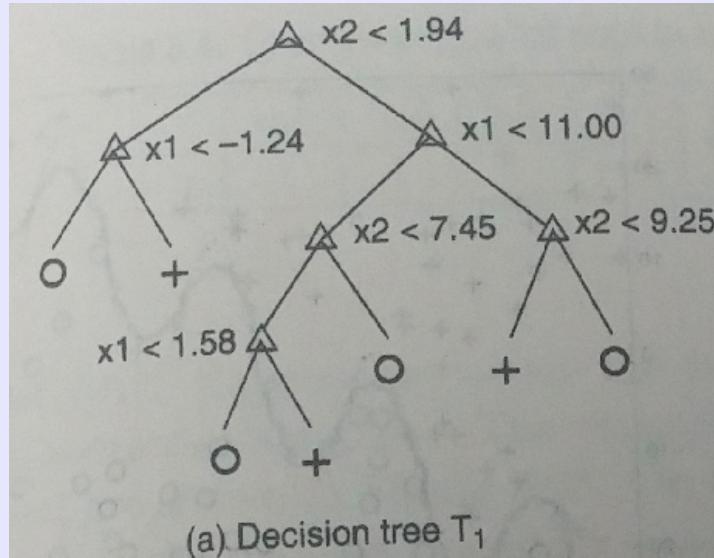


Bias, Variance, Under/Over-fitting

High Bias



Low Bias



Bias, Variance, Under/Over-fitting

- Bias is how far a model's predictions are from the target.
 - $\text{bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$
- Variance, on the other hand, is how your model reacts to changes in the training data.
 - $\text{variance}(\hat{\theta}) = E(\hat{\theta} - E(\hat{\theta}))^2$
- As the model becomes more complex, it picks up patterns in the training data, thus making it less generalizable.
 - A different training dataset (drawn from the same distribution) may induce results that vary from before.

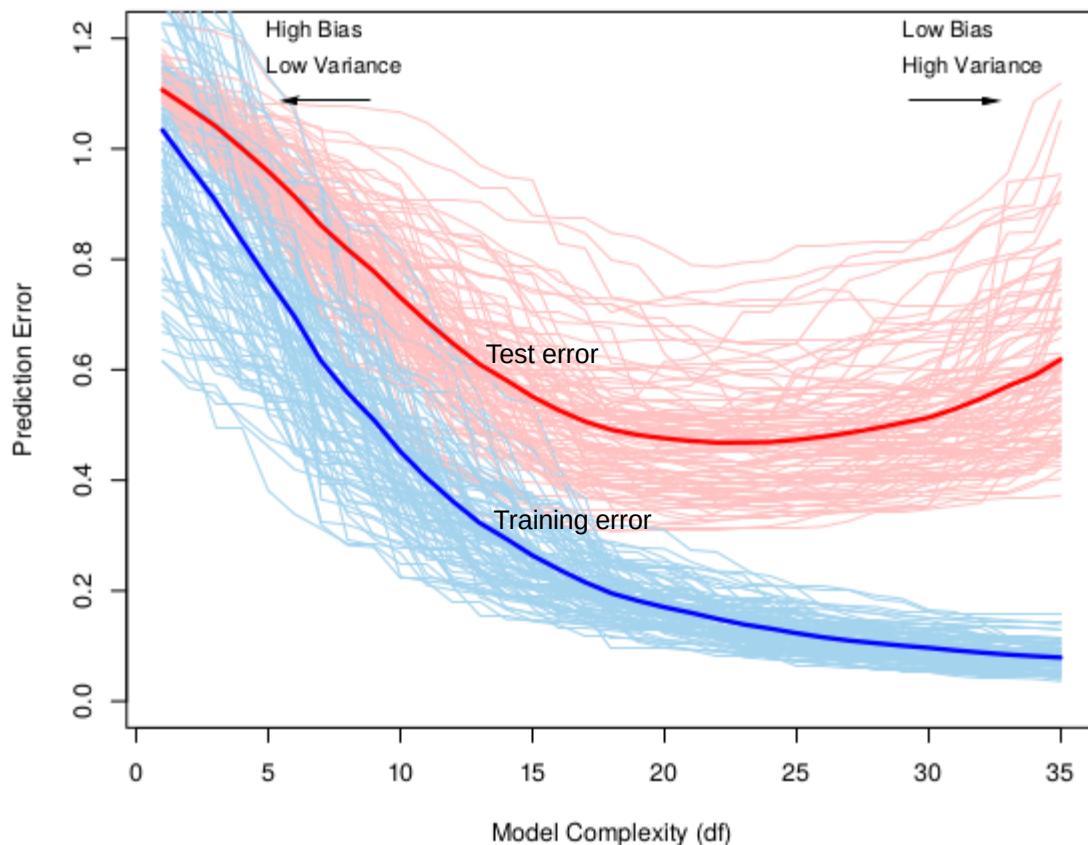
Bias, Variance, Under/Over-fitting

- Let's play a game!

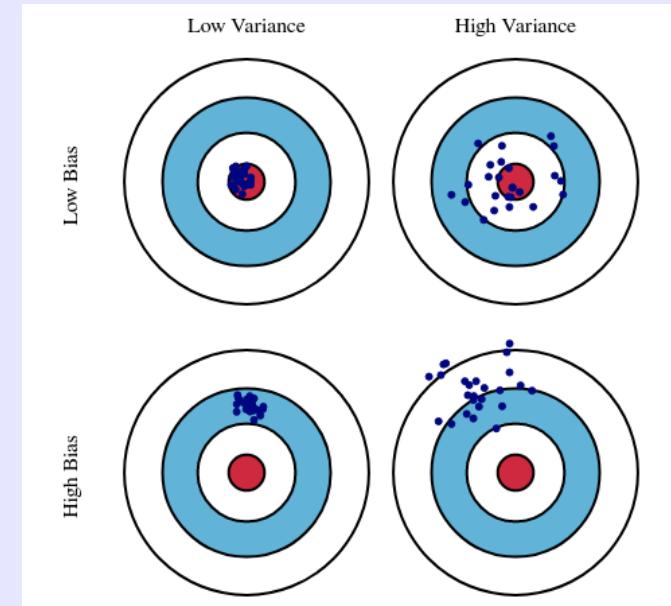


Photo by JESHOOTS.COM on Unsplash (<https://unsplash.com>)

Bias, Variance, Under/Over-fitting



Source: The Elements of Statistical Learning, 2e



A model with many predictive attributes will exhibit **low bias, high variance**.

A model with too few predictive attributes will exhibit **low variance, but may be quite biased**.

Informally: **bias** is how far a model's predictions are from target (underfitting), and **variance** is the degree to which these predictions vary between model iterations (overfitting).

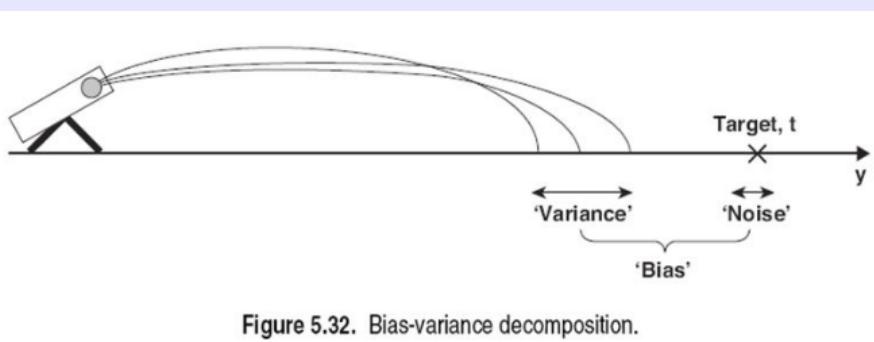


Figure 5.32. Bias-variance decomposition.

Bias, Variance, Under/Over-fitting

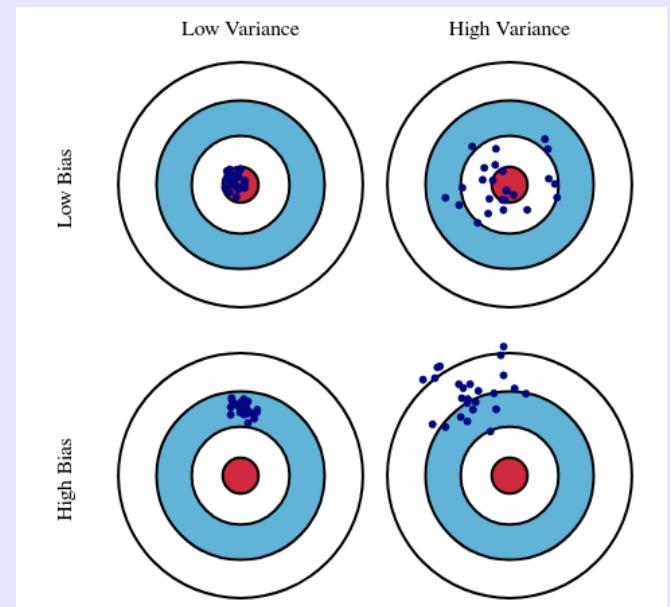
Informally: **bias** is how far a model's predictions are from target (underfitting), and **variance** is the degree to which these predictions vary between model iterations (overfitting).

Mathematically,

$$\text{Err}(\hat{\theta}) = \text{bias}(\hat{\theta})^2 + \text{variance}(\hat{\theta}) + \epsilon, \text{ where } \epsilon \text{ is the irreducible error}$$

Proof follows. Recall that:

- 1) $E[X+Y] = E[X]+E[Y]$ (Linearity of expectation)
- 2) $E[XY] = E[X] * E[Y]$ iff X and Y are iid
- 3) $E[E[X]] = E[X]$



Bias, Variance, Under/Over-fitting

$\text{Err}(\hat{\theta}) = \text{bias}(\hat{\theta})^2 + \text{variance}(\hat{\theta}) + \epsilon$, where ϵ is the irreducible error

The mean squared error of an estimator $\hat{\theta} = f(D; \theta)$ can be defined as:

$$\text{mse}(\hat{\theta}) = E[(\hat{\theta} - \theta)^2 | \theta]$$

Proposition: $\text{mse}(\hat{\theta}) = \text{bias}(\hat{\theta})^2 + \text{variance}(\hat{\theta})$

Bias, Variance, Under/Over-fitting

$\text{Err}(\hat{\theta}) = \text{bias}(\hat{\theta})^2 + \text{variance}(\hat{\theta}) + \epsilon$, where ϵ is the irreducible error

The mean squared error of an estimator $\hat{\theta} = f(D; \theta)$ can be defined as:

$$\text{mse}(\hat{\theta}) = E[(\hat{\theta} - \theta)^2 | \theta]$$

Proposition: $\text{mse}(\hat{\theta}) = \text{bias}(\hat{\theta})^2 + \text{variance}(\hat{\theta})$

Proof: $\text{mse}(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = E[\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta]^2$

Bias, Variance, Under/Over-fitting

$\text{Err}(\hat{\theta}) = \text{bias}(\hat{\theta})^2 + \text{variance}(\hat{\theta}) + \epsilon$, where ϵ is the irreducible error

The mean squared error of an estimator $\hat{\theta} = f(D; \theta)$ can be defined as:

$$\text{mse}(\hat{\theta}) = E[(\hat{\theta} - \theta)^2 | \theta]$$

Proposition: $\text{mse}(\hat{\theta}) = \text{bias}(\hat{\theta})^2 + \text{variance}(\hat{\theta})$

Proof:
$$\begin{aligned} \text{mse}(\hat{\theta}) &= E(\hat{\theta} - \theta)^2 = E[\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta]^2 \\ &= E[(\hat{\theta} - E(\hat{\theta})) + (E(\hat{\theta}) - \theta)]^2 \end{aligned}$$

Bias, Variance, Under/Over-fitting

$\text{Err}(\hat{\theta}) = \text{bias}(\hat{\theta})^2 + \text{variance}(\hat{\theta}) + \epsilon$, where ϵ is the irreducible error

The mean squared error of an estimator $\hat{\theta} = f(D; \theta)$ can be defined as:

$$\text{mse}(\hat{\theta}) = E[(\hat{\theta} - \theta)^2 | \theta]$$

Proposition: $\text{mse}(\hat{\theta}) = \text{bias}(\hat{\theta})^2 + \text{variance}(\hat{\theta})$

$$\begin{aligned}\text{Proof: } \text{mse}(\hat{\theta}) &= E(\hat{\theta} - \theta)^2 = E[\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta]^2 \\ &= E[(\hat{\theta} - E(\hat{\theta})) + (E(\hat{\theta}) - \theta)]^2 \\ &= E[(\hat{\theta} - E(\hat{\theta}))^2 + 2(\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta}) - \theta) + (E(\hat{\theta}) - \theta)^2]\end{aligned}$$

Bias, Variance, Under/Over-fitting

$\text{Err}(\hat{\theta}) = \text{bias}(\hat{\theta})^2 + \text{variance}(\hat{\theta}) + \epsilon$, where ϵ is the irreducible error

The mean squared error of an estimator $\hat{\theta} = f(D; \theta)$ can be defined as:

$$\text{mse}(\hat{\theta}) = E[(\hat{\theta} - \theta)^2 | \theta]$$

Proposition: $\text{mse}(\hat{\theta}) = \text{bias}(\hat{\theta})^2 + \text{variance}(\hat{\theta})$

$$\begin{aligned}\text{Proof: } \text{mse}(\hat{\theta}) &= E(\hat{\theta} - \theta)^2 = E[\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta]^2 \\ &= E[(\hat{\theta} - E(\hat{\theta})) + (E(\hat{\theta}) - \theta)]^2 \quad \xrightarrow{0} \\ &= E[(\hat{\theta} - E(\hat{\theta}))^2 + 2(\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta}) - \theta) + (E(\hat{\theta}) - \theta)^2]\end{aligned}$$

Bias, Variance, Under/Over-fitting

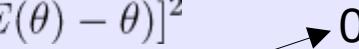
$\text{Err}(\hat{\theta}) = \text{bias}(\hat{\theta})^2 + \text{variance}(\hat{\theta}) + \epsilon$, where ϵ is the irreducible error

The mean squared error of an estimator $\hat{\theta} = f(D; \theta)$ can be defined as:

$$\text{mse}(\hat{\theta}) = E[(\hat{\theta} - \theta)^2 | \theta]$$

Proposition: $\text{mse}(\hat{\theta}) = \text{bias}(\hat{\theta})^2 + \text{variance}(\hat{\theta})$

Proof: $\text{mse}(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = E[\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta]^2$

$$\begin{aligned} &= E[(\hat{\theta} - E(\hat{\theta})) + (E(\hat{\theta}) - \theta)]^2 \\ &= E[(\hat{\theta} - E(\hat{\theta}))^2 + 2(\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta}) - \theta) + (E(\hat{\theta}) - \theta)^2] \\ &= \underbrace{E(\hat{\theta} - E(\hat{\theta}))^2}_{\text{Variance}} + E(E(\hat{\theta}) - \theta)^2 \end{aligned}$$


Bias, Variance, Under/Over-fitting

$\text{Err}(\hat{\theta}) = \text{bias}(\hat{\theta})^2 + \text{variance}(\hat{\theta}) + \epsilon$, where ϵ is the irreducible error

The mean squared error of an estimator $\hat{\theta} = f(D; \theta)$ can be defined as:

$$\text{mse}(\hat{\theta}) = E[(\hat{\theta} - \theta)^2 | \theta]$$

Proposition: $\text{mse}(\hat{\theta}) = \text{bias}(\hat{\theta})^2 + \text{variance}(\hat{\theta})$

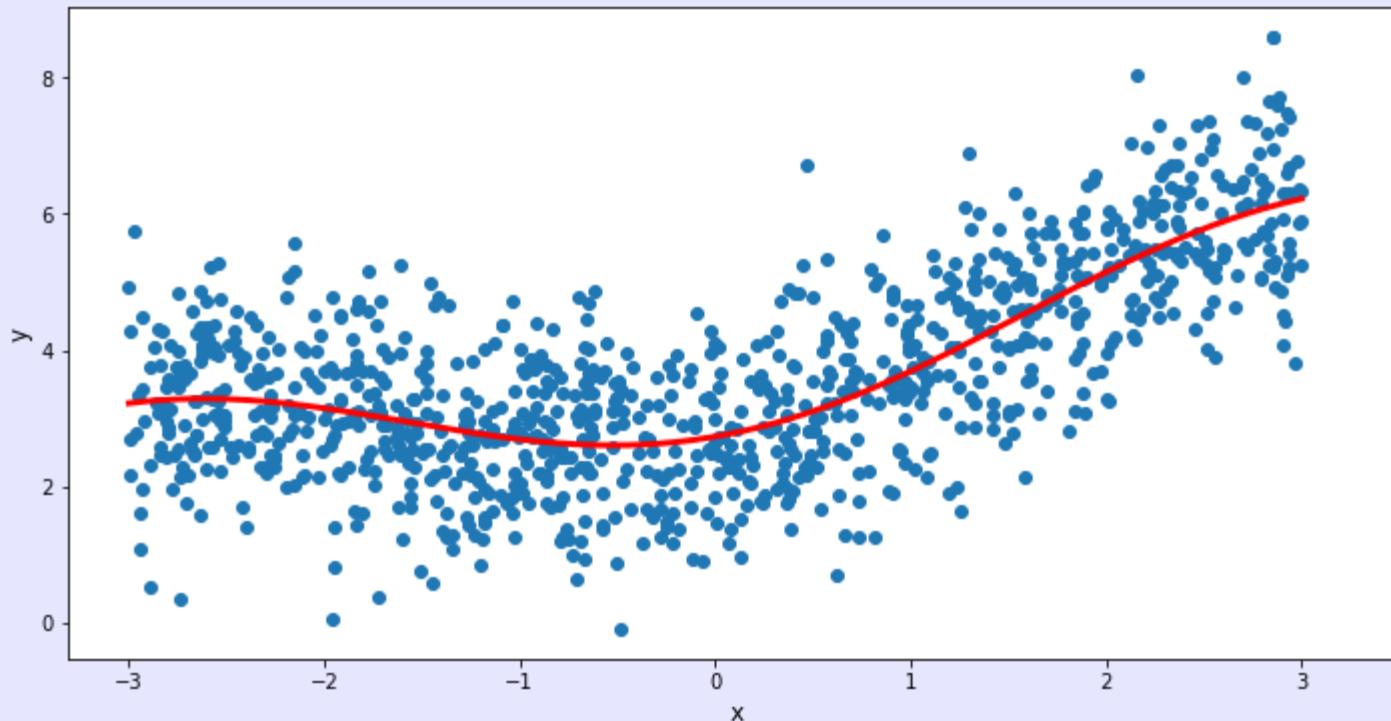
Proof: $\text{mse}(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = E[\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta]^2$

$$\begin{aligned} &= E[(\hat{\theta} - E(\hat{\theta})) + (E(\hat{\theta}) - \theta)]^2 \xrightarrow{0} \\ &= E[(\hat{\theta} - E(\hat{\theta}))^2 + 2(\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta}) - \theta) + (E(\hat{\theta}) - \theta)^2] \\ &= E(\hat{\theta} - E(\hat{\theta}))^2 \quad + E(E(\hat{\theta}) - \theta)^2 \\ &\quad \underbrace{\qquad\qquad\qquad}_{\text{Variance}} \quad + \underbrace{(E(\hat{\theta}) - \theta)^2}_{\text{Bias}^2} \end{aligned}$$



Bias, Variance, Under/Over-fitting

Example: $f(x) = \frac{1}{2}x + \sqrt{\max(x, 0)} - \cos x + 2$



- Noise is $\mathcal{N}(0, 1)$.
- 1,000 points.
- Red curve: true $f(x)$

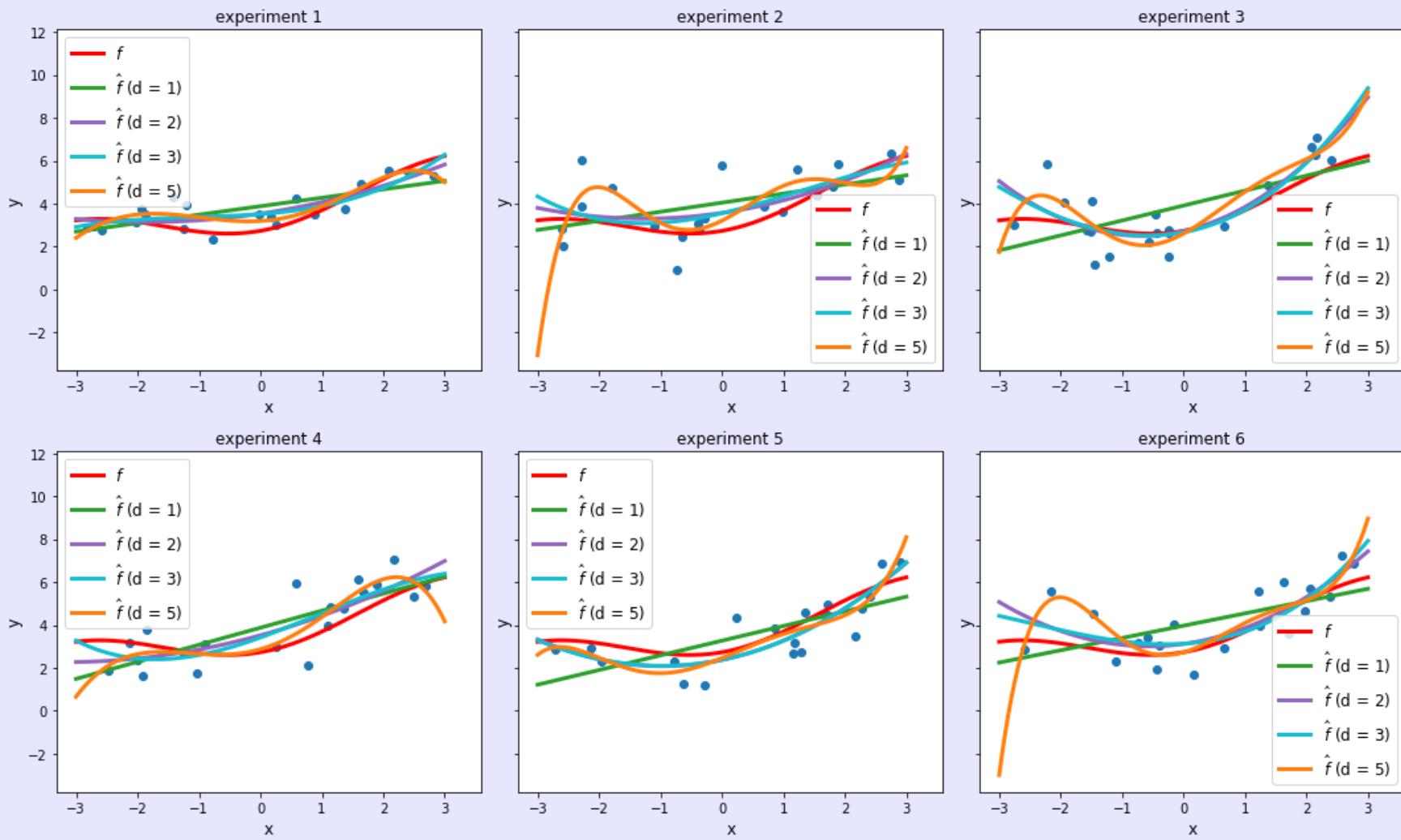
Note: $f(x)$ follows a non-linear pattern due to square root and cosine in its definition.

Task: To predict $f(x)$, model the problem with polynomial regression of varying degrees:

$$\hat{f}(x) = w_0 + w_1 x + w_2 x^2 + \dots + w_d x^d$$

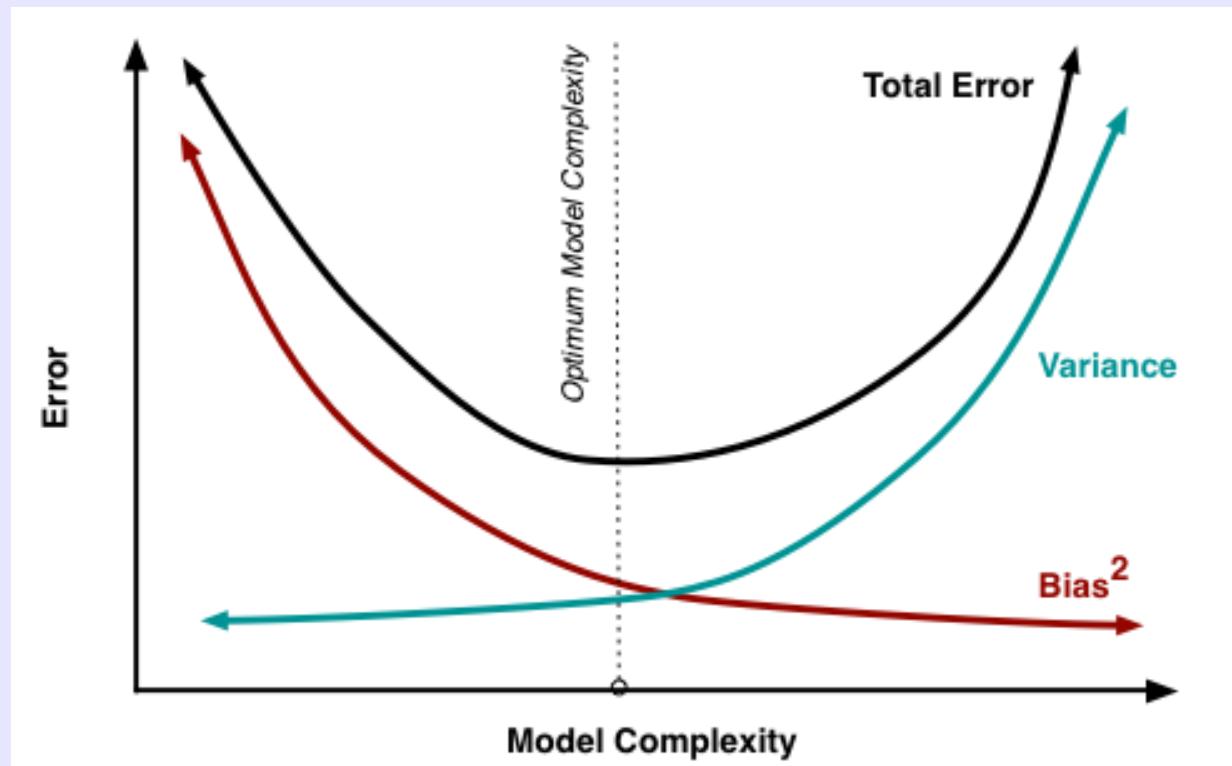
Bias, Variance, Under/Over-fitting

Example: 20 random points chosen for training in 6 different experiments. Different degree regression models ($d = 1, 2, 3, 5$); note the variance: complex models have high variance.



Bias, Variance, Under/Over-fitting

- At its heart, bias and variance is really about dealing with underfitting (bias) and overfitting (variance).



Graphic source: Scott Fortman
<http://scott.fortmann-roe.com/docs/BiasVariance.html>

Bias, Variance, Under/Over-fitting

- Mitigation approaches

- Dimensionality reduction and feature selection can reduce variance by simplifying models.
- Regularization can help reduce variance.
- If you cannot reduce dimensions or engage in feature selection, a larger training set may decrease variance.
- Adding features decreases bias, at the expense of additional variance.

A model with many predictive attributes will exhibit **low bias, high variance**. A model with too few predictive attributes will exhibit **low variance, but may be quite biased**.