

DATASCI 347 Machine Learning

Lecture 2: Bias-variance decomposition

Ruoxuan Xiong

Suggested reading: ISL Chapter 2

Lecture plan

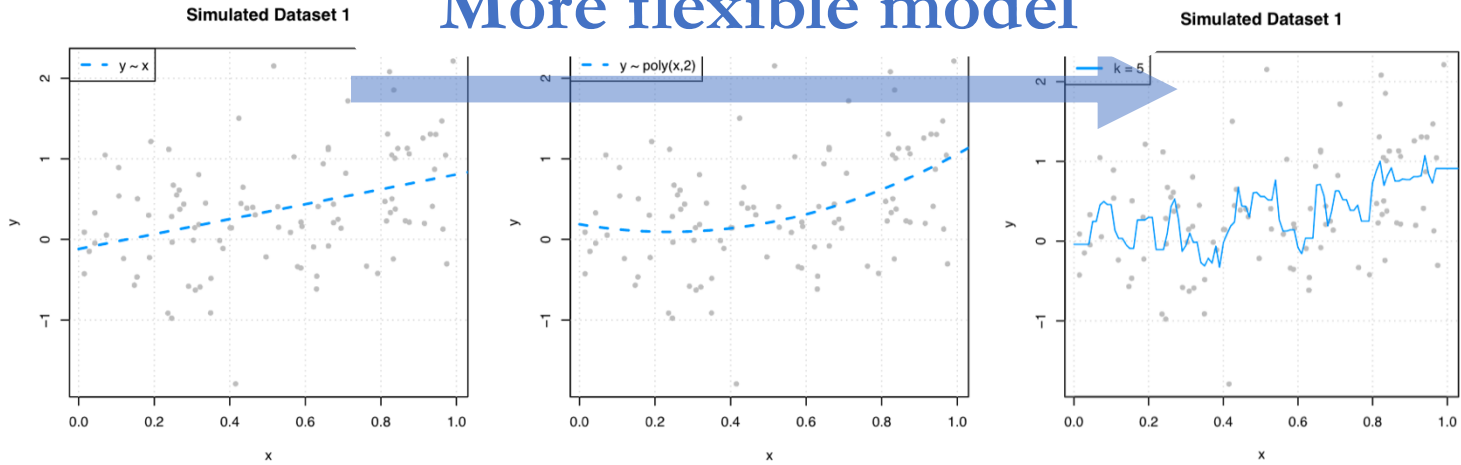
- Bias-variance decomposition for regression problems (MSE)

Training vs. test data and MSE

- **Training data:** data $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ used to fit f
- **Training MSE:** $\text{MSE} = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{f}(X_i))^2$
- **Test data:** data, $(X'_1, Y'_1), (X'_2, Y'_2), \dots, (X'_m, Y'_m)$ previous unseen and *not used* to fit f
- **Test MSE:** $\text{MSE} = \frac{1}{m} \sum_{i=1}^m (Y'_i - \hat{f}(X'_i))^2$
- We care **more about test MSE** than training MSE
 - It measures **how well the model generalizes**
- *A low training MSE does not imply a low test MSE*

MSE varies with model flexibility

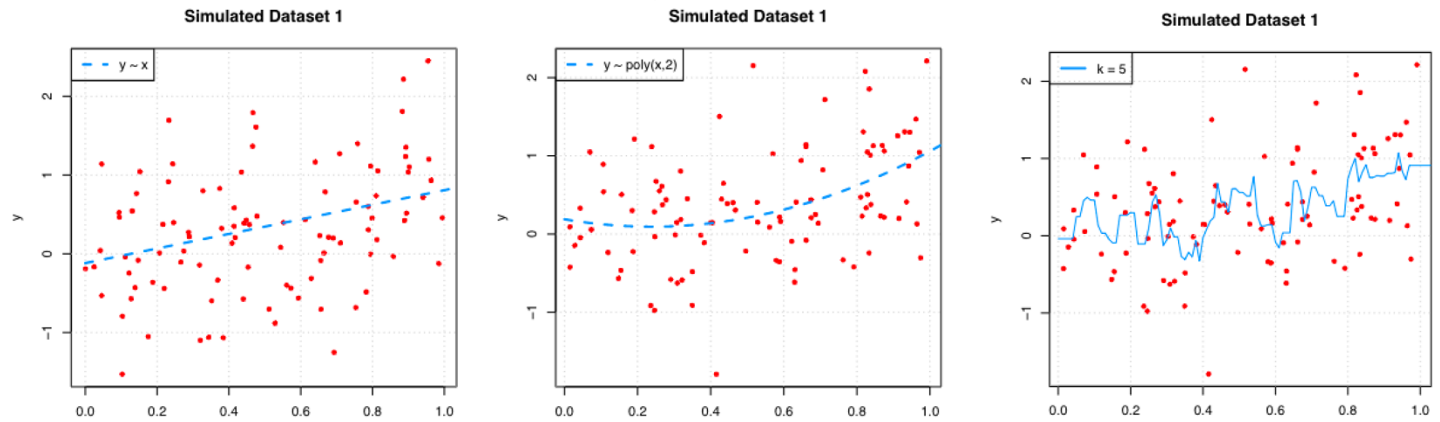
More flexible model



Training MSE = 0.439

Training MSE = 0.425

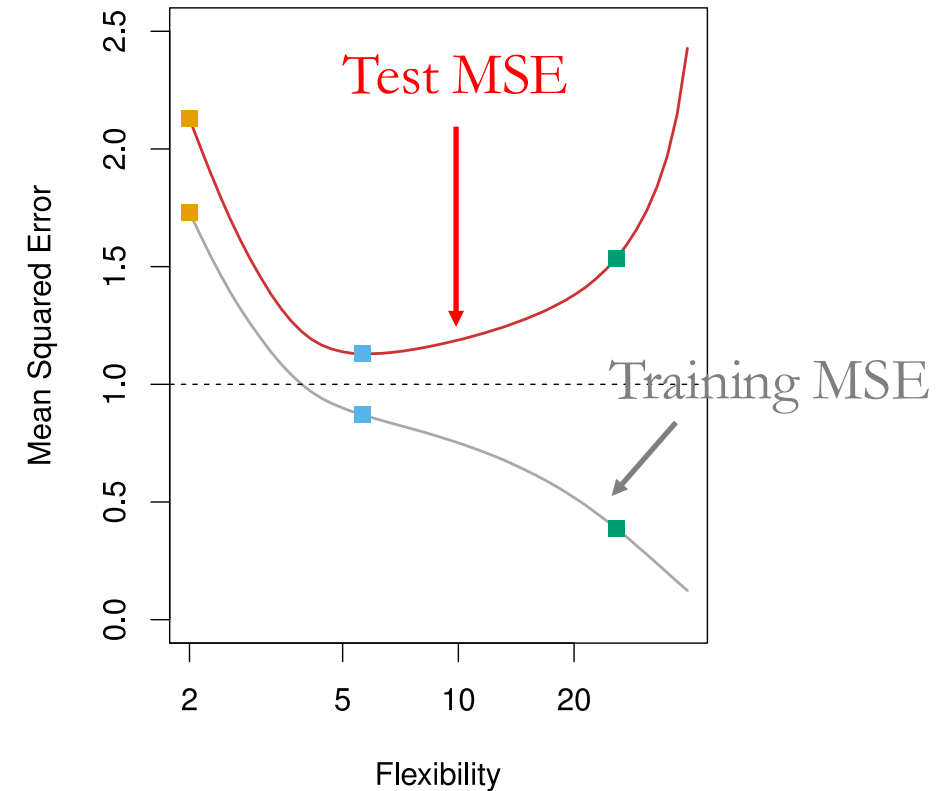
Training MSE = 0.354



Test MSE = 0.533

Test MSE = 0.518

Test MSE = 0.564

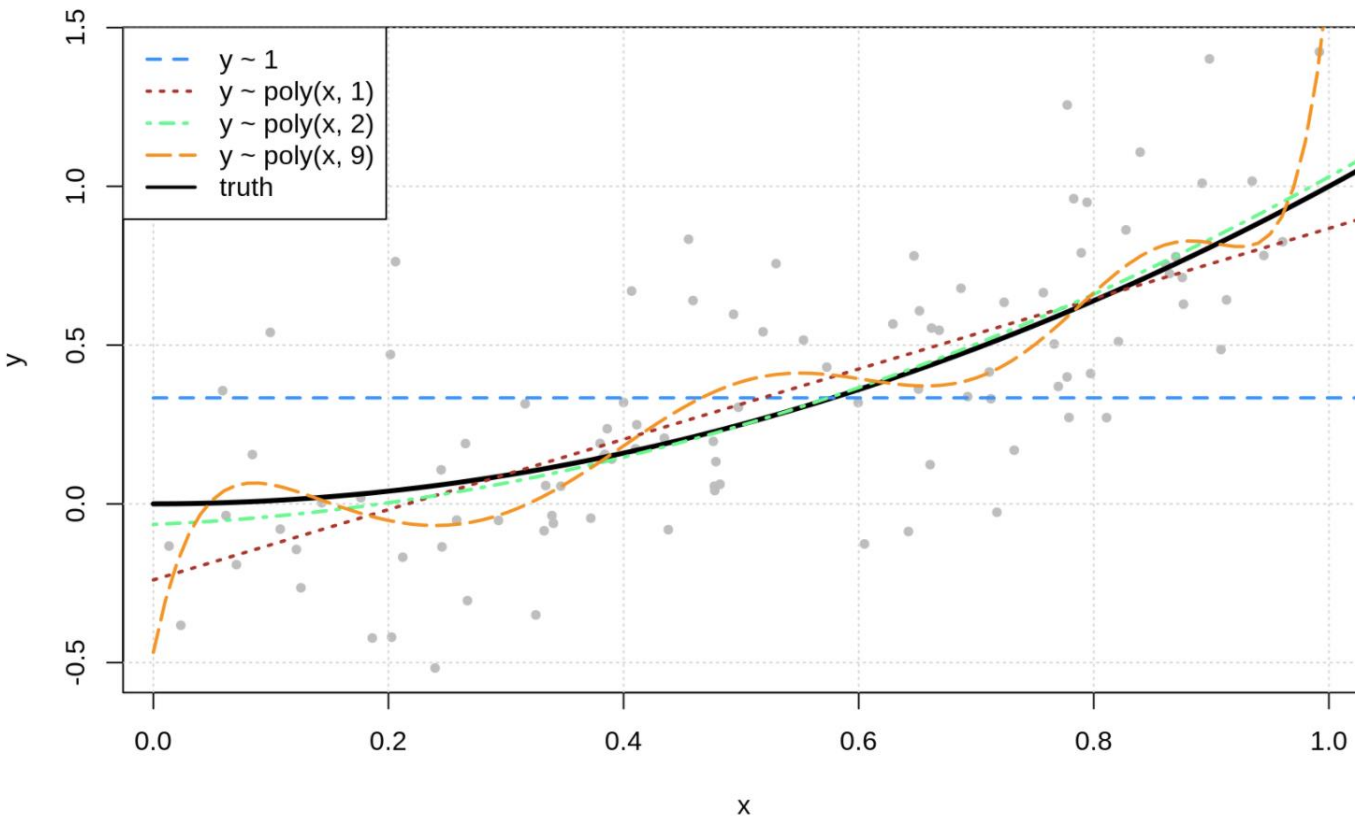


Why does MSE vary with model flexibility?

- Suppose the *true regression function* is $f(x) = x^2$ (x is a scalar)
- We fit *models of increasing flexibility* to approximate this function
 - Constant model: $\hat{f}_0(x) = \hat{\beta}_0$
 - Linear model: $\hat{f}_1(x) = \hat{\beta}_0 + x \cdot \hat{\beta}_1$
 - Quadratic model: $\hat{f}_2(x) = \hat{\beta}_0 + x \cdot \hat{\beta}_1 + x^2 \cdot \hat{\beta}_2$
 - Ninth degree polynomial model: $\hat{f}_9(x) = \hat{\beta}_0 + x \cdot \hat{\beta}_1 + \dots + x^9 \cdot \hat{\beta}_9$
- The model is *more flexible* if the *polynomial degree is larger*

Four fitted models

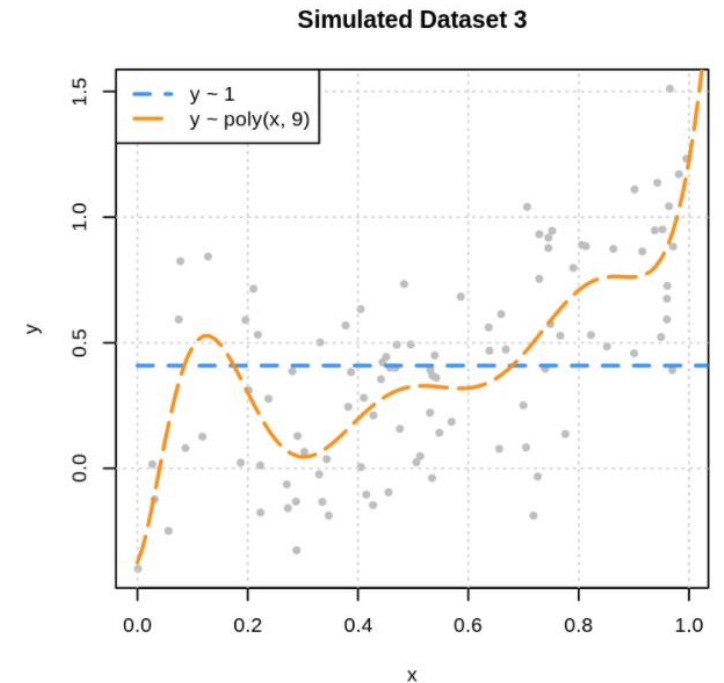
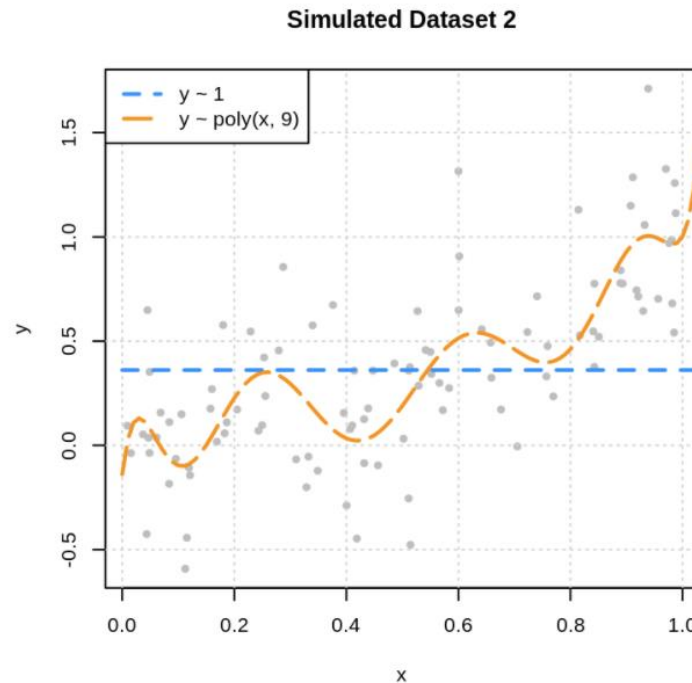
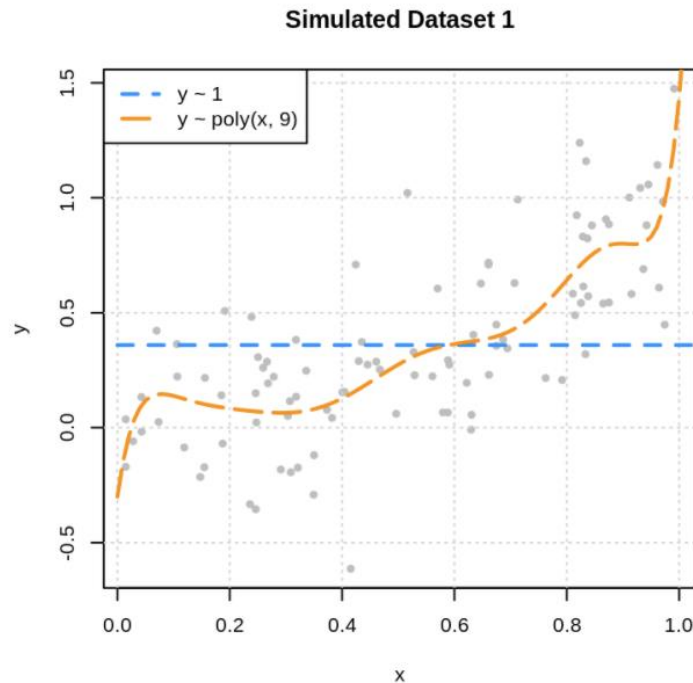
Four Polynomial Models fit to a Simulated Dataset



- Zero predictor model fits poorly
- Linear model is reasonable
- Quadratic model fits much better
- Ninth degree model seems rather wild

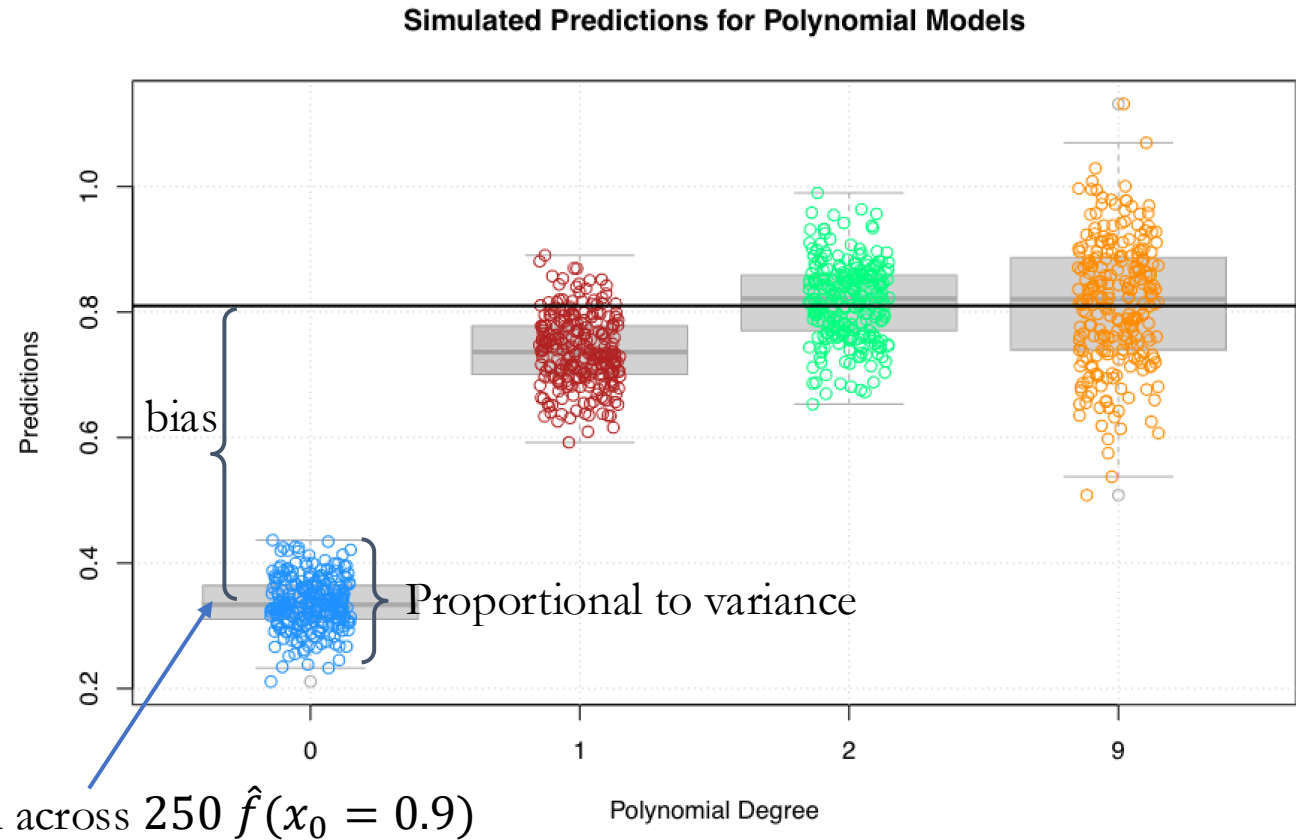
Model fits change across training datasets

- In general, the fitted model depends on the training data
- When we refit the same model on different datasets:
 - The **constant model** $\hat{f}_0(x)$ changes only slightly
 - The **9th-degree polynomial** $\hat{f}_9(x)$ can change dramatically
- This variability across training sets is called the **variance of the model**
 - The variance of $\hat{f}_0(x)$ is smaller than the variance of $\hat{f}_9(x)$



Predicting $f(x_0)$ at a specific point x_0

- We study prediction at a single test input
- Test point: $x_0 = 0.9$
- True value: $f(x_0 = 0.9) = x_0^2 = 0.81$
- We generate **250 independent training datasets**
- For each dataset, we fit polynomial models of degree 0,1,2, and 9, and record the prediction $\hat{f}(x_0 = 0.9)$



Bias and variance of prediction at x_0

- The squared bias

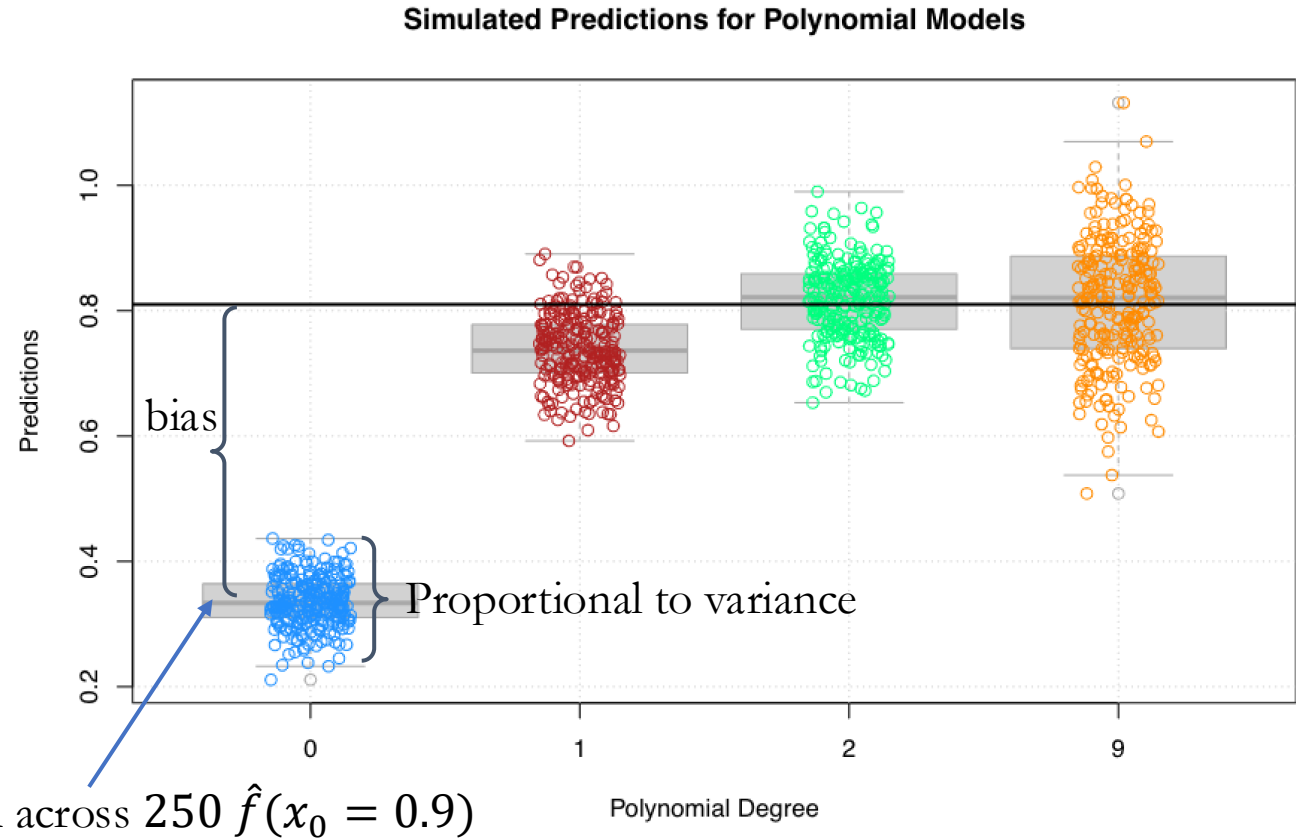
$$\hat{f}_2(x) \approx \hat{f}_9(x) < \hat{f}_1(x) < \hat{f}_0(x)$$

- Increasing degree from 2 to 9 does not meaningfully reduce bias

- Variance

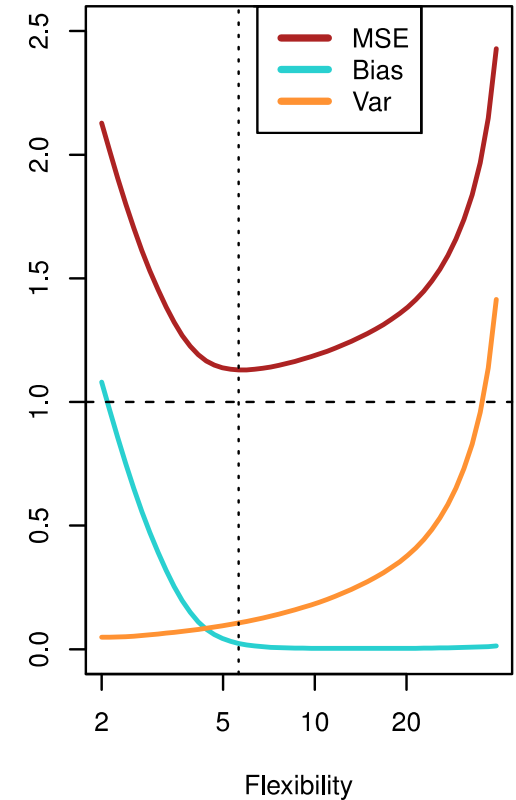
$$\hat{f}_0(x) < \hat{f}_1(x) < \hat{f}_2(x) < \hat{f}_9(x)$$

- Increasing degree increases variance



Combining bias and variance

- The **test mean squared error (MSE)** depends on both **bias** and **variance**
- As model flexibility increases:
 - Bias decreases
 - Variance increases
- Their combined effect produces the **U-shaped test MSE curve**



Analyzing test MSE

- Suppose the true regression function is f
- The response satisfies $Y = f(X) + \varepsilon$, with $E[\varepsilon | X] = 0$
- Let the training dataset be $\mathcal{D} = \{(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)\}$
- The fitted model $\hat{f}$ *depends on training data* $\mathcal{D}$
 - If we change the training data $\mathcal{D}$, we get a different estimate $\hat{f}$
- Prediction at a fixed test point x_0
 - The MSE at x_0 can be decomposed as

$$\text{MSE}(x_0) = \underbrace{E_{Y|X, \mathcal{D}} \left[\left(Y - \hat{f}(X) \right)^2 \mid X = x_0 \right]}_{\text{Expected value over } \mathcal{D} \text{ and } Y|X} = \underbrace{E_{\mathcal{D}} \left[\left(f(x_0) - \hat{f}(x_0) \right)^2 \right]}_{\text{Reducible error}} + \underbrace{V_{Y|X}[Y \mid X = x_0]}_{\text{Irreducible error}}$$

Decomposition of MSE

- The MSE at x_0 can be decomposed as

$$\text{MSE}(x_0) = \underbrace{E_{Y|X, \mathcal{D}} \left[\left(Y - \hat{f}(X) \right)^2 \mid X = x_0 \right]}_{\text{Expected value over } Y|X \text{ and } \mathcal{D}} = \underbrace{E_{\mathcal{D}} \left[\left(f(x_0) - \hat{f}(x_0) \right)^2 \right]}_{\text{Reducible error}} + \underbrace{V_{Y|X} [Y \mid X = x_0]}_{\text{Irreducible error}}$$

- The MSE at x_0 is the **expected prediction error** for
 - a **new response** Y_0
 - at a **fixed input** $X_0 = x_0$
 - using a **model** $\hat{f}$ trained on random data $\mathcal{D}$
- Two sources of *randomness*
 - Noise in the outcome: Y_0 is random because $Y = f(X) + \varepsilon$ and ε is random
 - Randomness in fitted model $\hat{f}$: $\hat{f}$ is random because it depends on randomly sampled $\mathcal{D}$

Irreducible error

- The MSE at x_0 can be decomposed as

$$\text{MSE}(x_0) = \underbrace{E_{Y|X, \mathcal{D}} \left[\left(Y - \hat{f}(X) \right)^2 \mid X = x_0 \right]}_{\text{Expected value over } Y|X \text{ and } \mathcal{D}} = \underbrace{E_{\mathcal{D}} \left[\left(f(x_0) - \hat{f}(x_0) \right)^2 \right]}_{\text{Reducible error}} + \underbrace{V_{Y|X}[Y \mid X = x_0]}_{\text{Irreducible error}}$$

- **Irreducible error**: the variance of Y given that $X = x_0$
 - If $Y = f(X) + \varepsilon$ with $E[\varepsilon \mid X] = 0$ and $V[\varepsilon \mid X] = \sigma_\varepsilon^2$, then $V_{Y|X}[Y \mid X = x_0] = \sigma_\varepsilon^2$
 - Irreducible error **can not be reduced** for any $\hat{f}$

Reducible error

- The MSE at x_0 can be decomposed as

$$\text{MSE}(x_0) = \underbrace{E_{Y|X, \mathcal{D}} \left[\left(Y - \hat{f}(X) \right)^2 \mid X = x_0 \right]}_{\text{Expected value over } Y|X \text{ and } \mathcal{D}} = \underbrace{E_{\mathcal{D}} \left[\left(f(x_0) - \hat{f}(x_0) \right)^2 \right]}_{\text{Reducible error}} + \underbrace{V_{Y|X}[Y \mid X = x_0]}_{\text{Irreducible error}}$$

- **Reducible error**: the expected squared error by using $\hat{f}(x_0)$ to estimate $f(x_0)$
 - The randomness comes from training data $\mathcal{D}$ used to obtain $\hat{f}$ (both f and x_0 are fixed)
 - Reducible error **can be reduced** by using a better $\hat{f}$

Bias-variance decomposition of reducible error

- *Reducible error* can be decomposed as the *squared bias* and *variance*

$$E_{\mathcal{D}} \left[\left(f(x_0) - \hat{f}(x_0) \right)^2 \right] = \underbrace{\left(f(x_0) - E_{\mathcal{D}}[\hat{f}(x_0)] \right)^2}_{\text{bias}^2(\hat{f}(x_0))} + \underbrace{E_{\mathcal{D}} \left[\left(\hat{f}(x_0) - E_{\mathcal{D}}[\hat{f}(x_0)] \right)^2 \right]}_{\text{var}(\hat{f}(x_0))}$$

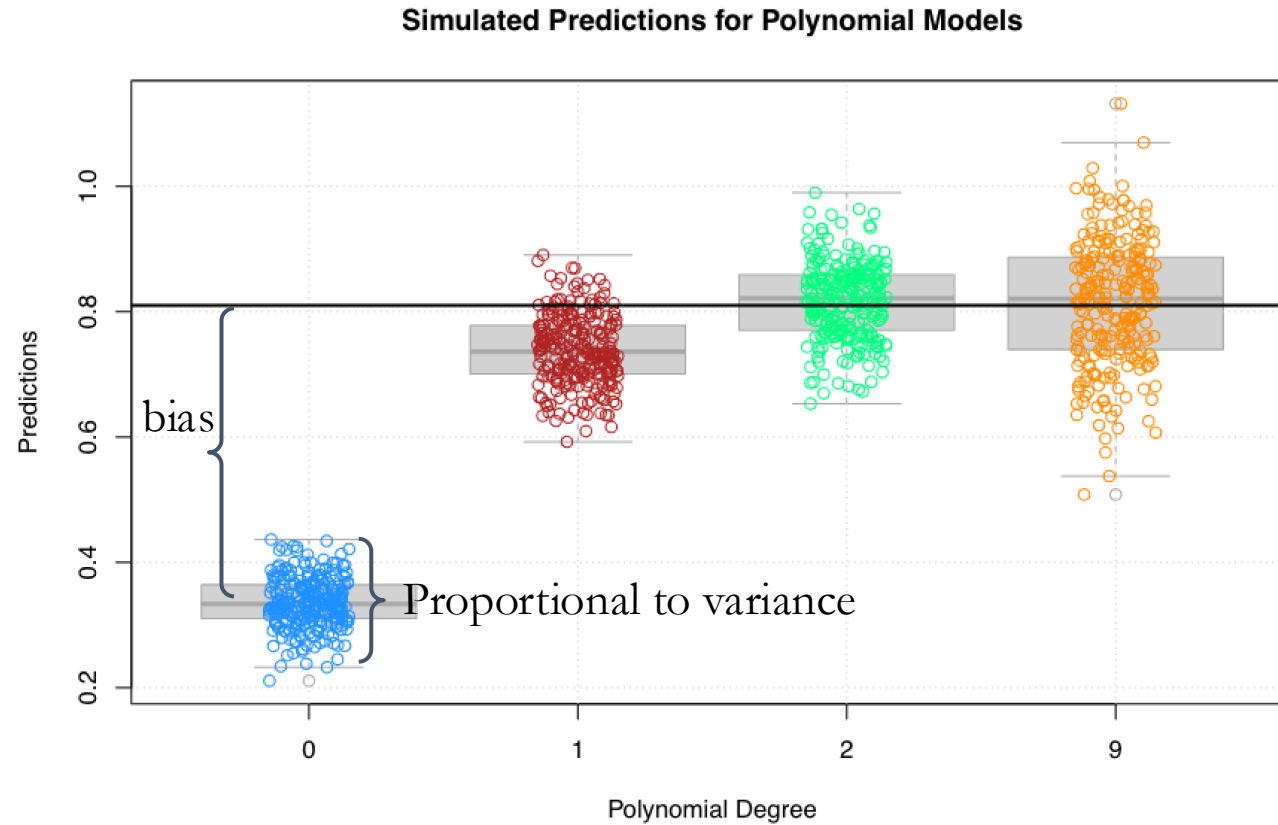
- *Take home exercise: Prove this decomposition*
- Hint: Use the property

$$E_{\mathcal{D}} \left[\left(f(x_0) - \hat{f}(x_0) \right)^2 \right] = E_{\mathcal{D}} \left[\underbrace{\left(f(x_0) - E_{\mathcal{D}}[\hat{f}(x_0)] \right)}_A + \underbrace{E_{\mathcal{D}}[\hat{f}(x_0)] - \hat{f}(x_0)}_B \right]^2$$

$$E_{\mathcal{D}}[(A + B)^2] = A^2 + \underbrace{2AE_{\mathcal{D}}[B]}_{= 0} + E_{\mathcal{D}}[B^2] = A^2 + E_{\mathcal{D}}[B^2]$$

(A is nonrandom and $E_{\mathcal{D}}[A^2] = A^2$)

Recall our toy example



Summing up the decomposition

- The MSE at x_0 can be decomposed as

$$\text{MSE}(x_0) = \underbrace{\text{bias}^2(\hat{f}(x_0)) + \text{var}(\hat{f}(x_0))}_{\text{Reducible error}} + \underbrace{\sigma_\varepsilon^2}_{\text{Irreducible error}}$$



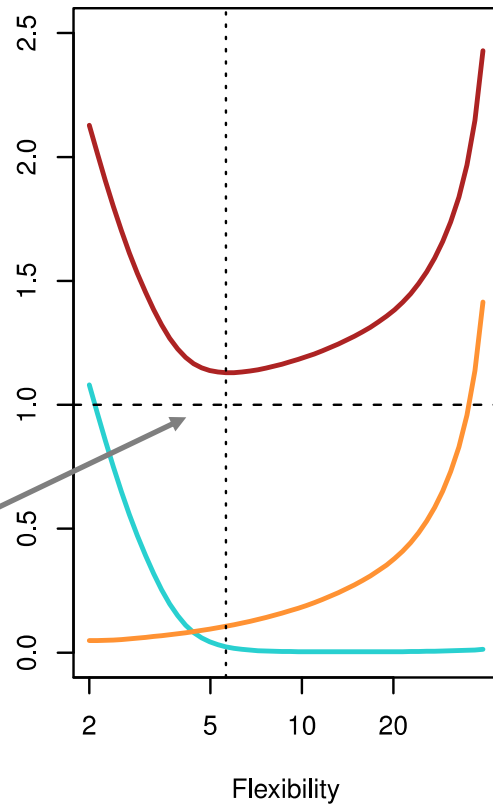
Visualizing the bias-variance tradeoff

$$Y = f(X) + \varepsilon$$
$$\text{Var}(\varepsilon) = 1$$

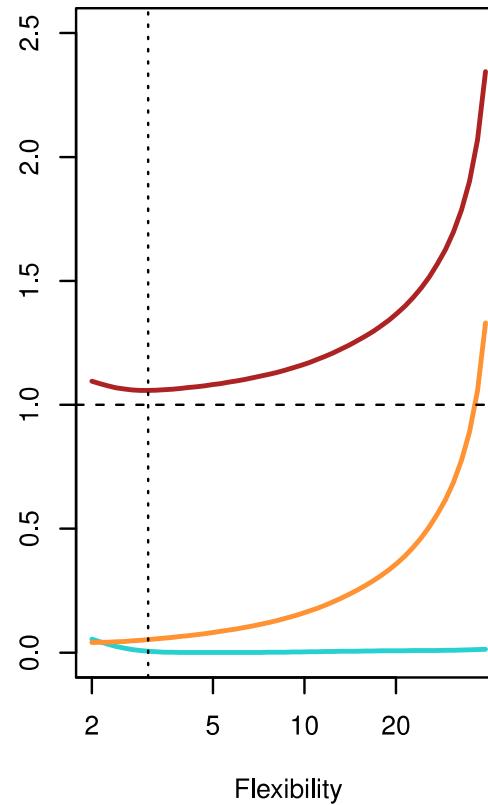
SNR: Signal-to-noise ratio
Signal measured by $f(X)$
Noise measured by ε

Irreducible
error

f is complicated,
low SNR



f is simple,
low SNR



f is complicated,
high SNR

