The bias-variance tradeoff

STAT 471

September 21, 2021

Where we are

Unit 1: Intro to modern data mining

Unit 2: Tuning predictive models

Unit 3: Regression-based methods

Unit 4: Tree-based methods

Unit 5: Deep learning

Lecture 1: Model complexity

Lecture 2: Bias-variance trade-off

Lecture 3: Cross-validation

Lecture 4: Classification

Lecture 5: Unit review and quiz in class

Homework 1 due the following Monday.

Today’s question: What drives test error?

Problem parameters

- Sample size
- Noise level
- Fitted model complexity (number of parameters)
- True model complexity

Phenomena

- Model bias: extent to which model unable to capture the truth
- Overfitting: extent to which the fit is sensitive to noise in training data
- Irreducible error: noise in test points that is impossible to predict

How do all these elements come together?

Expected test error

Given a fitted $\hat{f}$ and a test set $(X_1^{\text{test}}, Y_1^{\text{test}}), \ldots, (X_N^{\text{test}}, Y_N^{\text{test}})$, recall that

$$\text{Test error of } \hat{f} = \frac{1}{N}\sum_{i=1}^{N}(Y_i^{\text{test}} - \hat{Y}_i^{\text{test}})^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i^{\text{test}} - \hat{f}(X_i^{\text{test}}))^2.$$

The test error is a random function of the test set and the training set.

Define the **expected test error (ETE)** of a prediction rule as

$$\text{ETE} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[(Y_i^{\text{test}} - \hat{f}(X_i^{\text{test}}))^2].$$

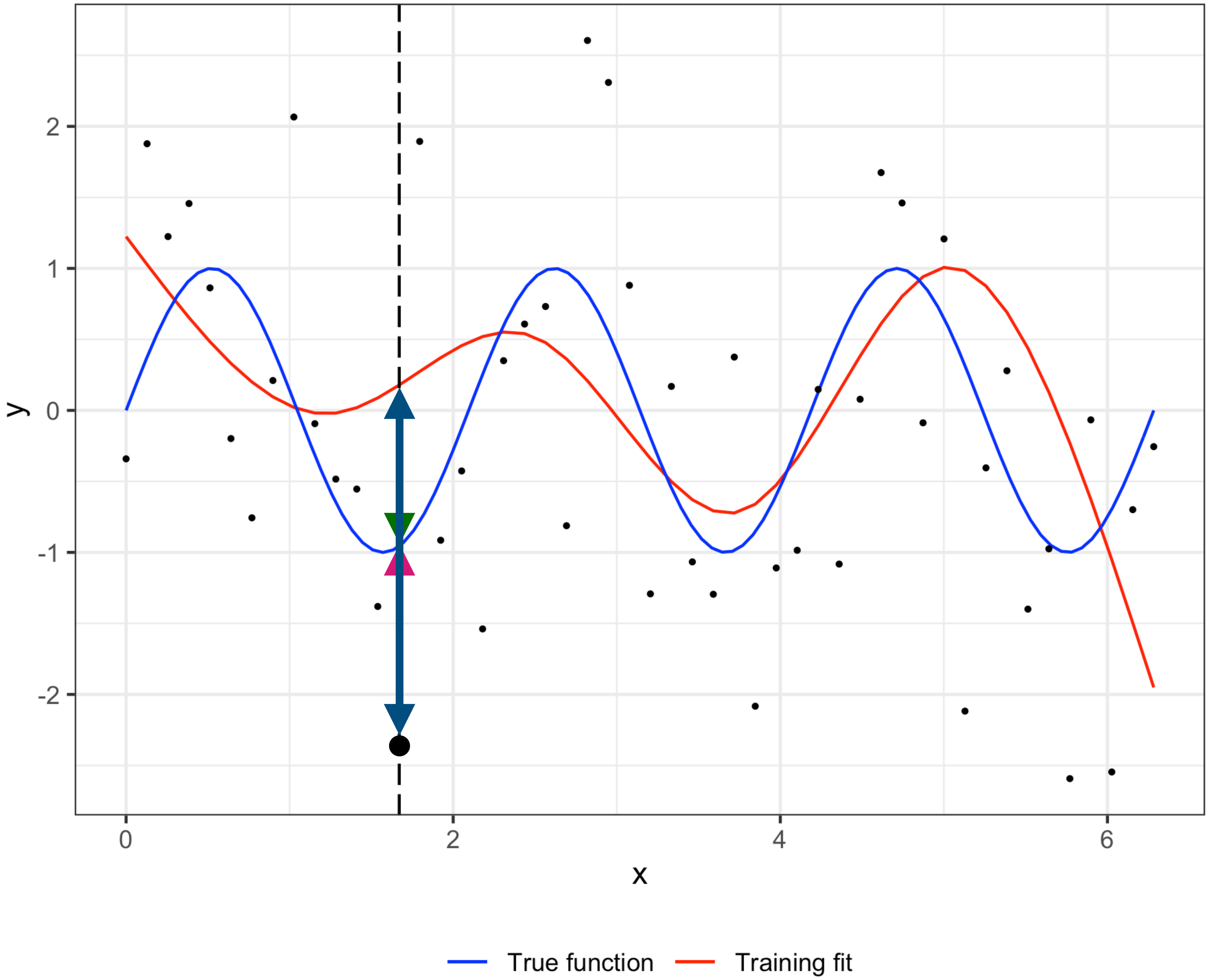
The ETE is easier to understand theoretically.

Dissecting the expected test error

Suppose $Y = f(X) + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$.

Then,

$$\begin{aligned}
\text{ETE}_i &= \mathbb{E}[\underline{(Y_i^{\text{test}} - \hat{f}(X_i^{\text{test}}))^2}] \\
&= \mathbb{E}[(f(X_i^{\text{test}}) + \epsilon_i^{\text{test}} - \hat{f}(X_i^{\text{test}}))^2] \\
&= \underline{\mathbb{E}[(f(X_i^{\text{test}}) - \hat{f}(X_i^{\text{test}}))^2]} + \underline{\sigma^2}
\end{aligned}$$



Dissecting the expected test error

$$\text{ETE}_i = \mathbb{E}[(f(X_i^{\text{test}}) - \hat{f}(X_i^{\text{test}}))^2] + \sigma^2$$

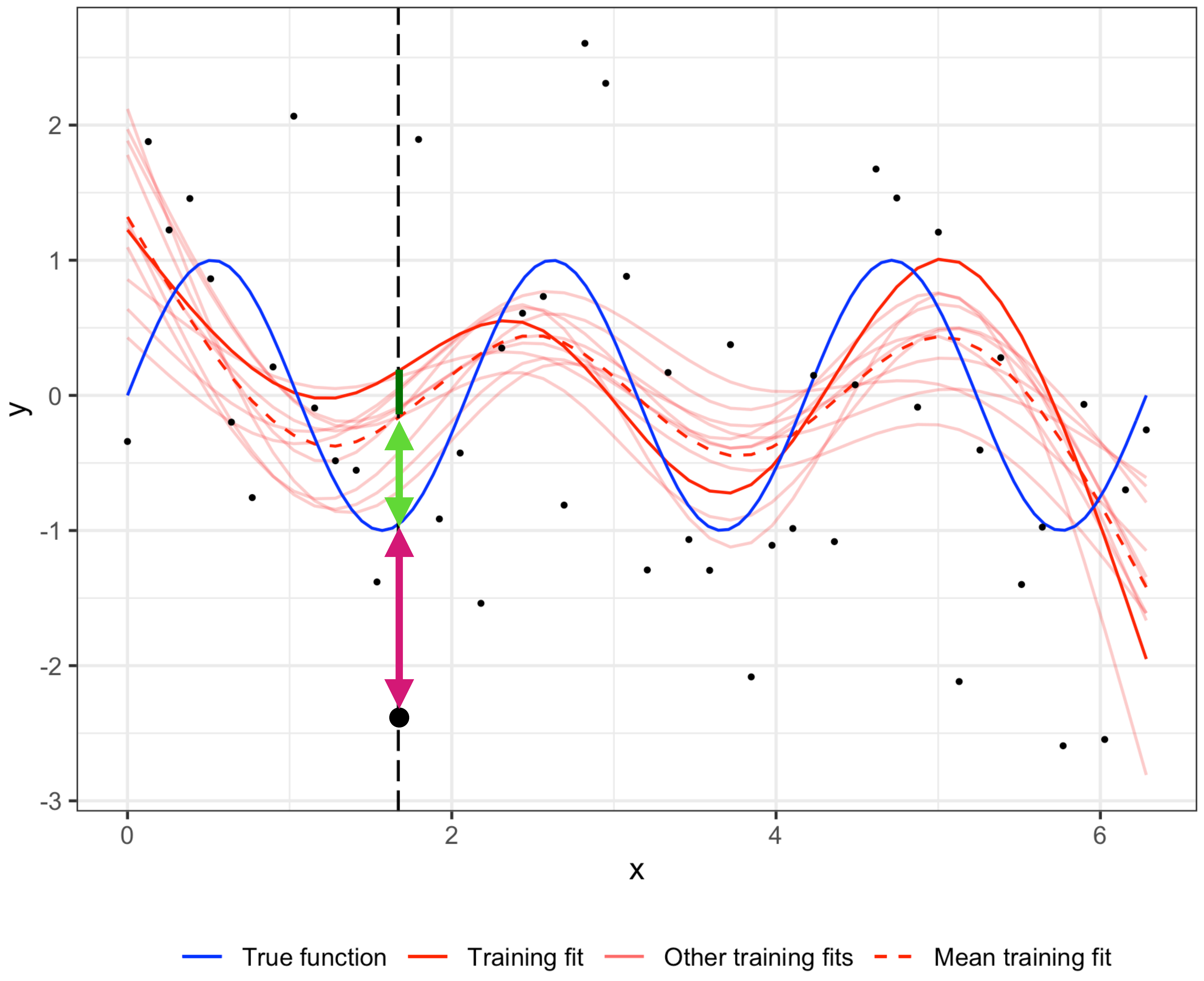
$$= (f(X_i^{\text{test}}) - \text{Ave}(\hat{f}(X_i^{\text{test}}))^2$$

$$+\mathbb{E}[(\text{Ave}(\hat{f}(X_i^{\text{test}})) - \hat{f}(X_i^{\text{test}}))^2]$$

$$+\sigma^2$$

$$= \text{Bias}_i^2 + \text{Variance}_i + \text{Irreducible error}$$

This is the bias-variance decomposition.

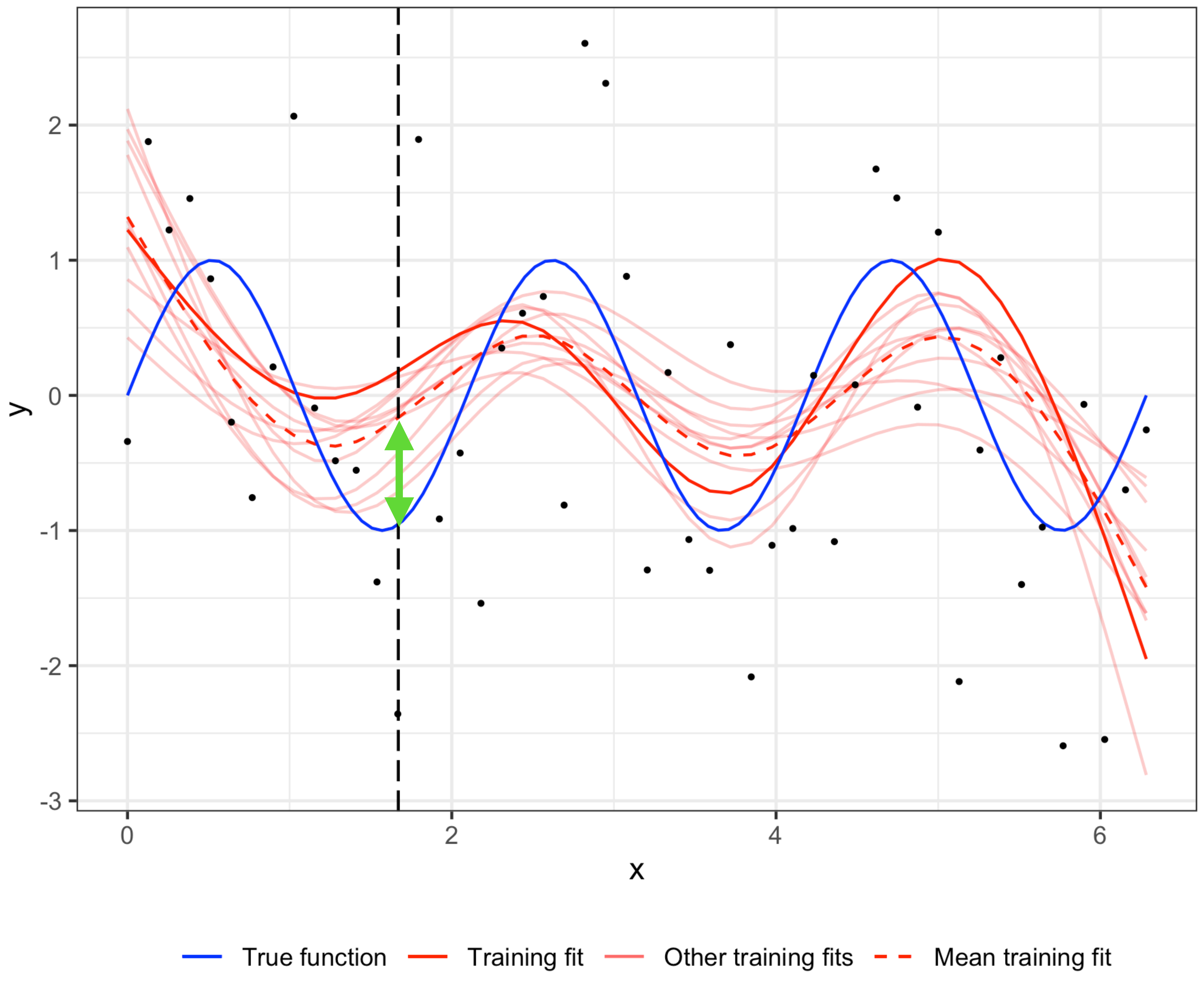


Understanding bias

$$\text{Bias}_i = \text{Ave}(\hat{f}(X_i^{\text{test}})) - f(X_i^{\text{test}})$$

Bias reflects the distance from the average fitted model to the true trend.

Adding model complexity reduces bias.



Understanding variance

$$\text{Variance}_i = \underline{\mathbb{E}[(\hat{f}(X_i^{\text{test}}) - \text{Ave}(\hat{f}(X_i^{\text{test}})))^2]}$$

Variance is the wobbling of the model fit due to the randomness in the training data.

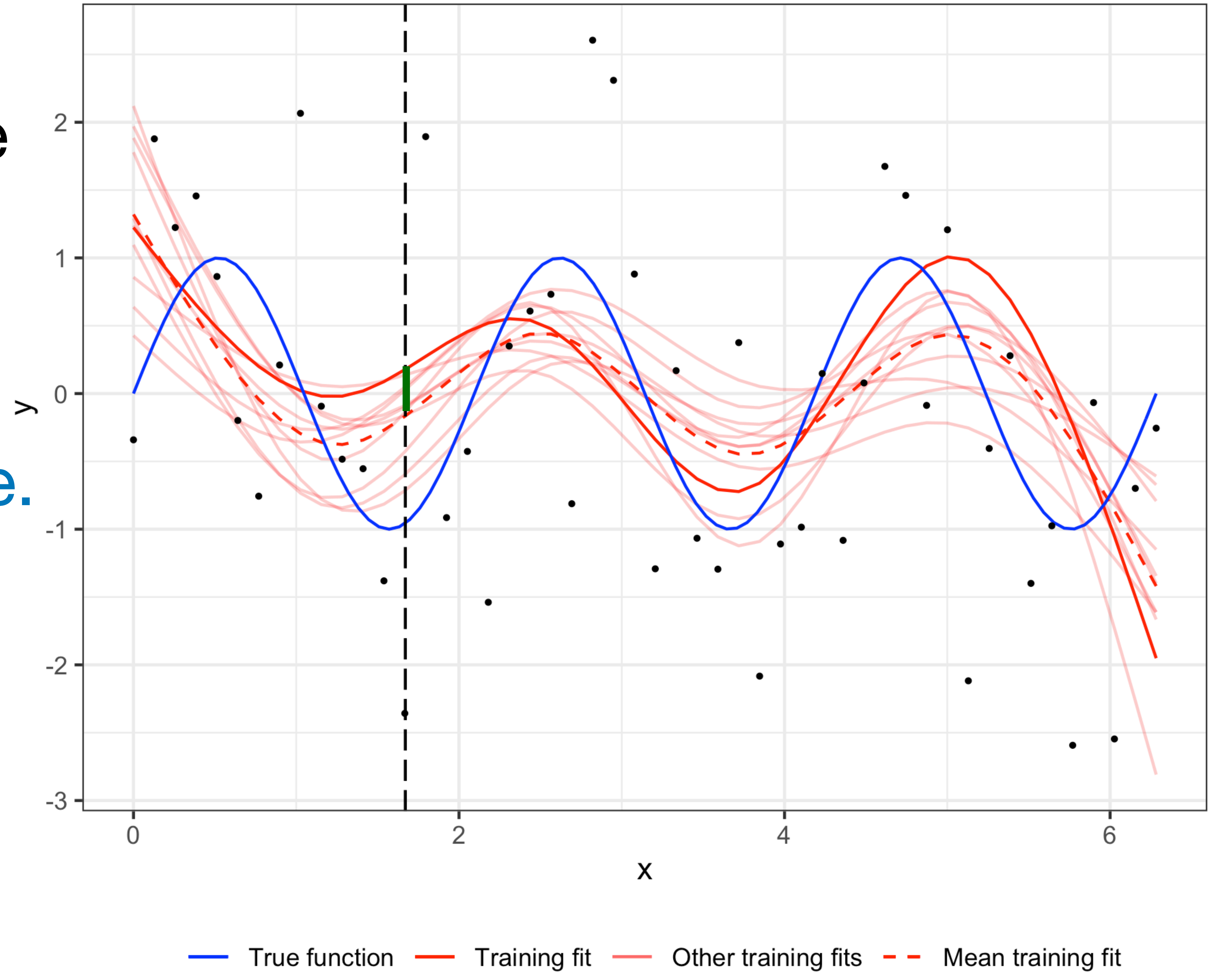
Variance is a consequence of overfitting.

Adding model complexity increases variance.

In linear models,

$$\text{Variance} = \frac{1}{n}\sum_{i=1}^{n}\text{Variance}_i = \frac{\sigma^2 p}{n}$$

(assuming $n = N$ and $X_i^{\text{test}} = X_i^{\text{train}}$)



The bias-variance tradeoff

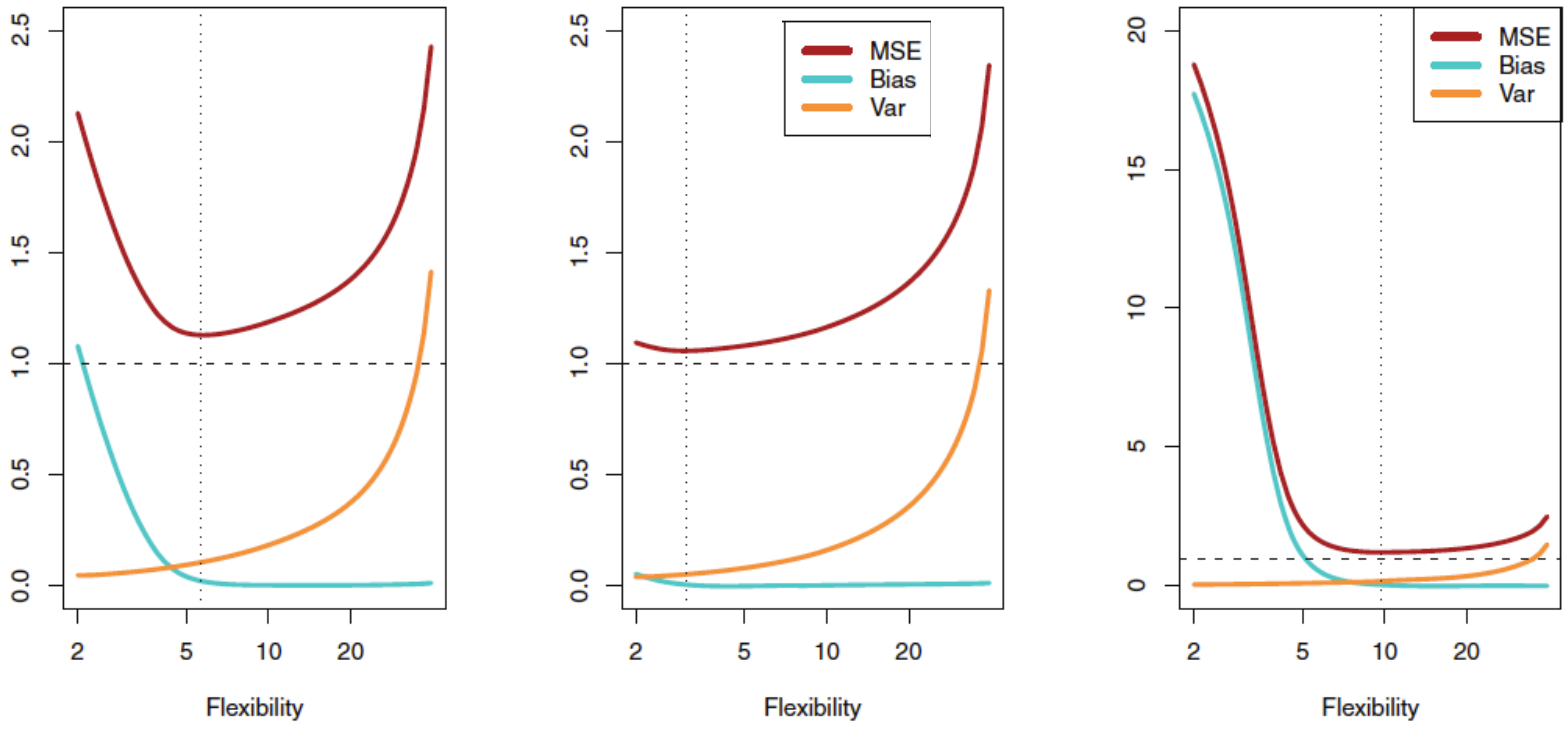
Recall that

$$\text{ETE} = \text{Bias}^2 + \text{Variance} + \text{Irreducible error.}$$

- Adding model complexity reduces bias
- Adding model complexity increases variance

When varying model complexity, there is a tradeoff between bias and variance.

Navigating the bias-variance tradeoff



The shapes of these curves differ based on the problem parameters.

What drives test error?

Problem parameters

- Sample size
- Noise level
- Fitted model complexity (number of parameters)
- True model complexity

Phenomena

- Model bias: extent to which model unable to capture the truth

+

- Overfitting: extent to which the fit is sensitive to noise in training data

+

- Irreducible error: noise in test points that is impossible to predict

= ETE