The bias-variance tradeoff

STAT 4710

September 19, 2023

Where we are

✓ **Unit 1:** R for data mining

Unit 2: Prediction fundamentals

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

What drives test error?

Problem parameters

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

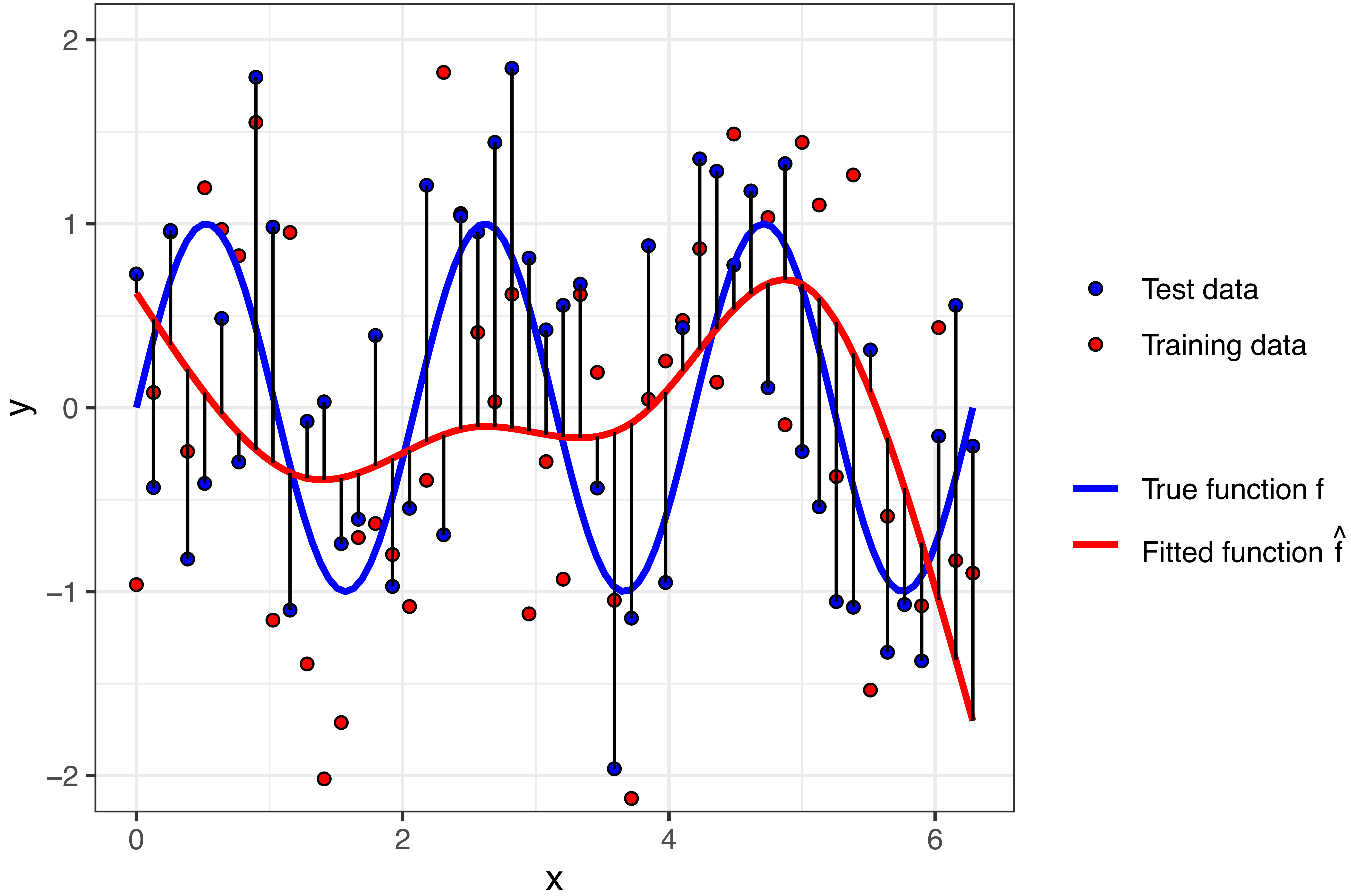
Phenomena

- 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?

The expected test error

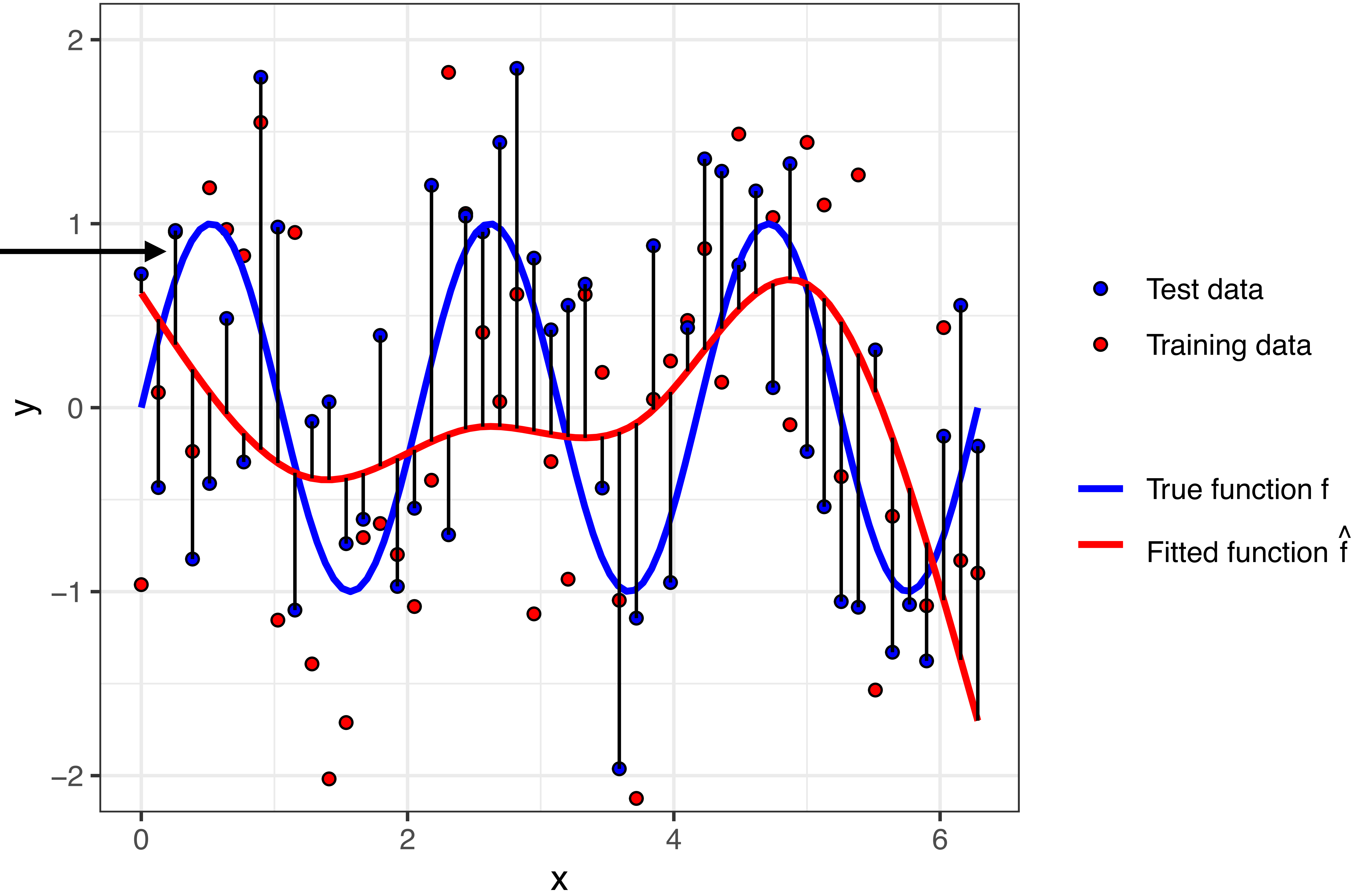
How to quantify performance of a prediction method (e.g. natural spline with df = 5)?



The expected test error

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

How to quantify performance of a prediction method (e.g. natural spline with df = 5)?

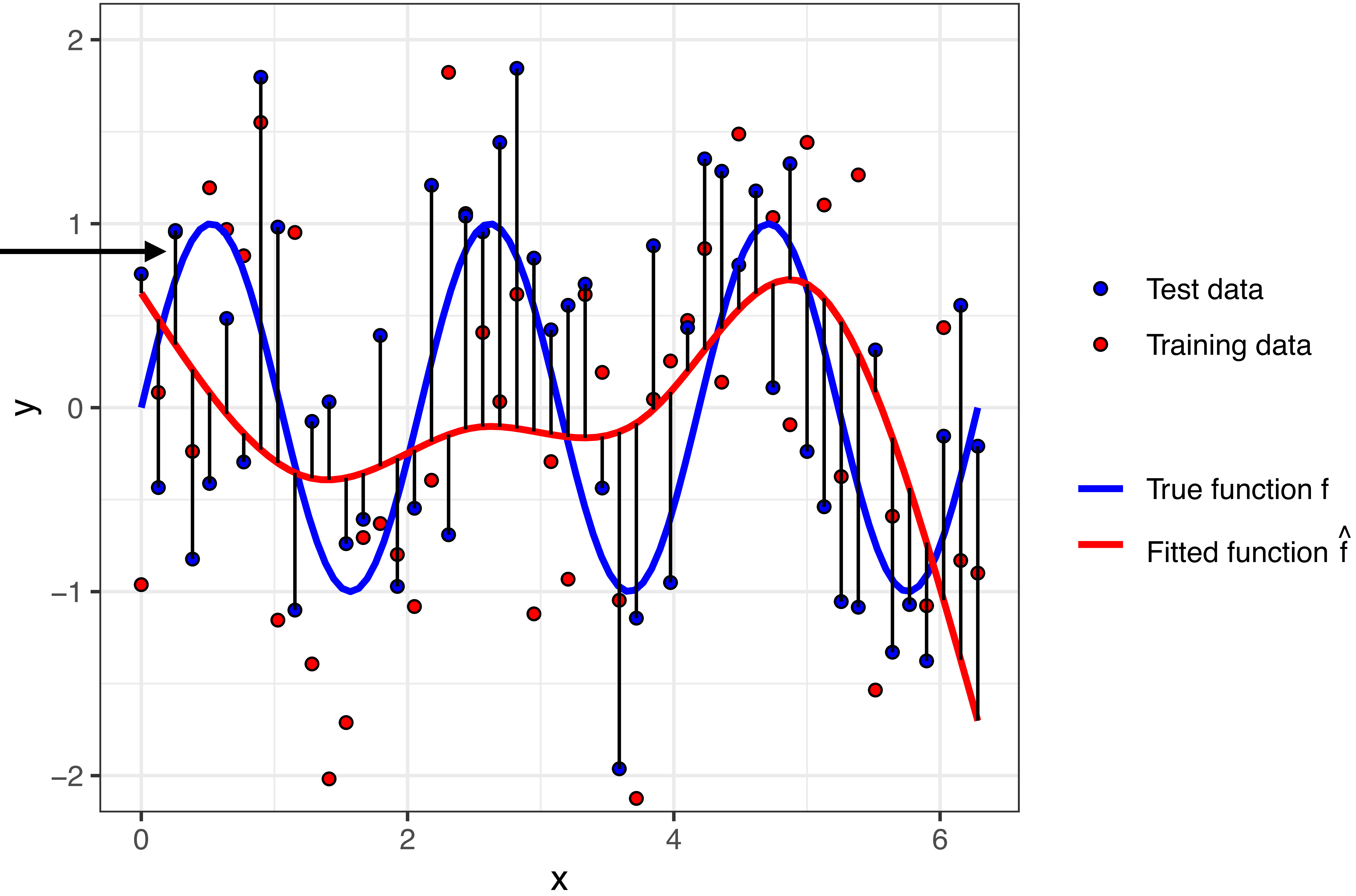


The expected test error

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

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

How to quantify performance of a prediction method (e.g. natural spline with df = 5)?



Define **expected test error (ETE)** as

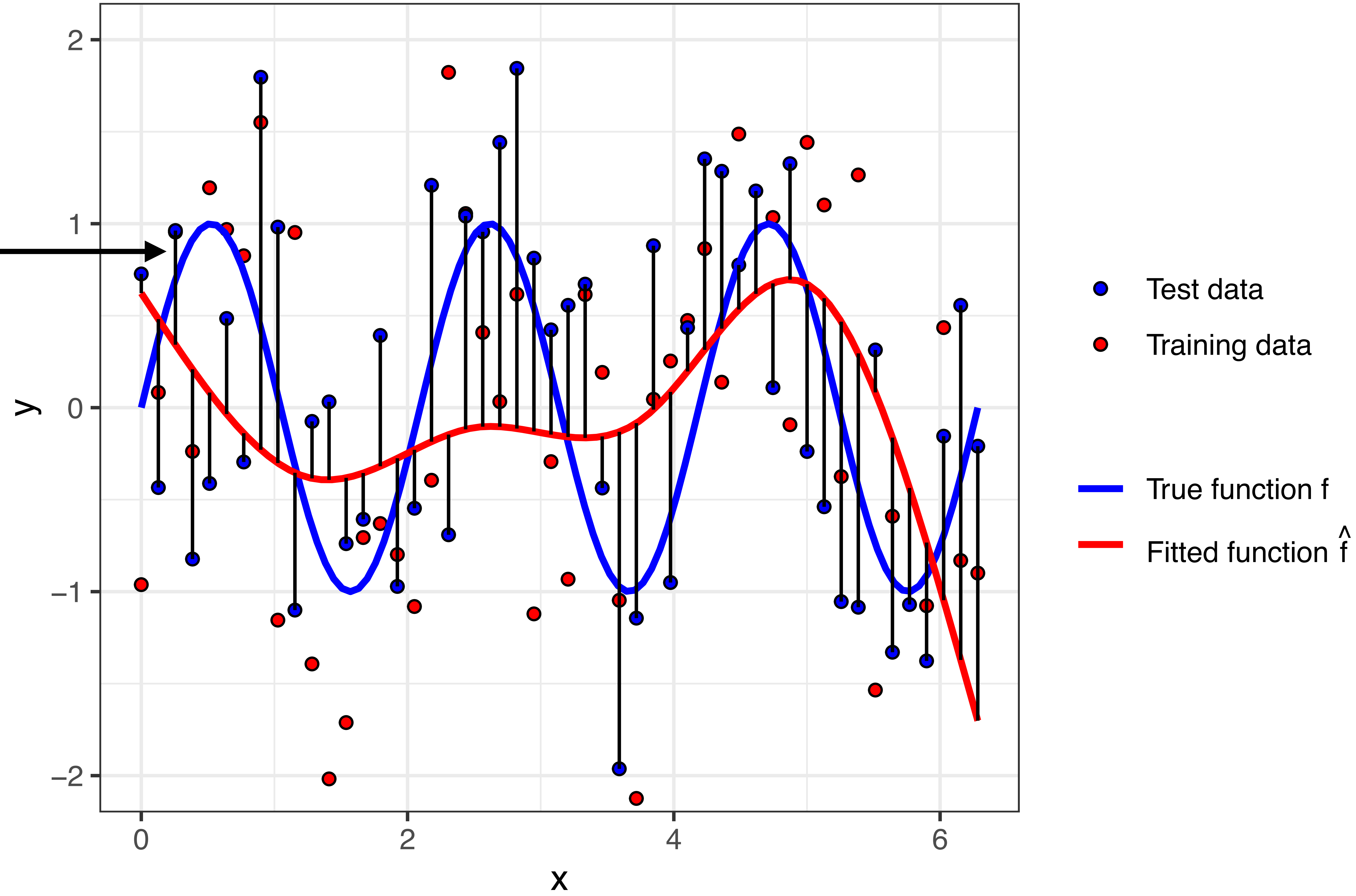
$$\text{ETE} = \mathbb{E}[\text{Test error}]$$

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

The expected test error

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

How to quantify performance of a prediction method (e.g. natural spline with df = 5)?



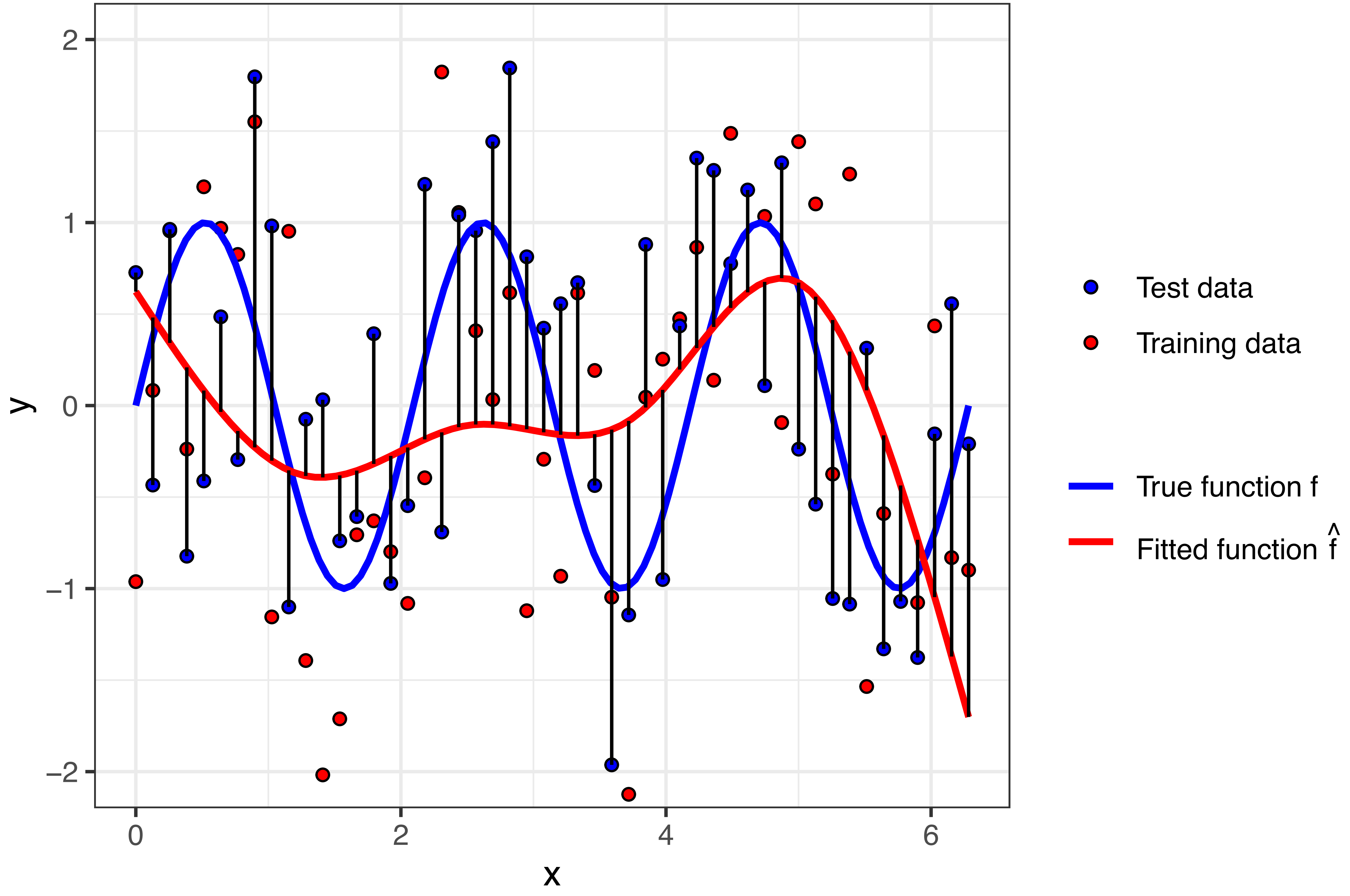
Define **expected test error (ETE)** as

$$\text{ETE} = \mathbb{E}[\text{Test error}]$$

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

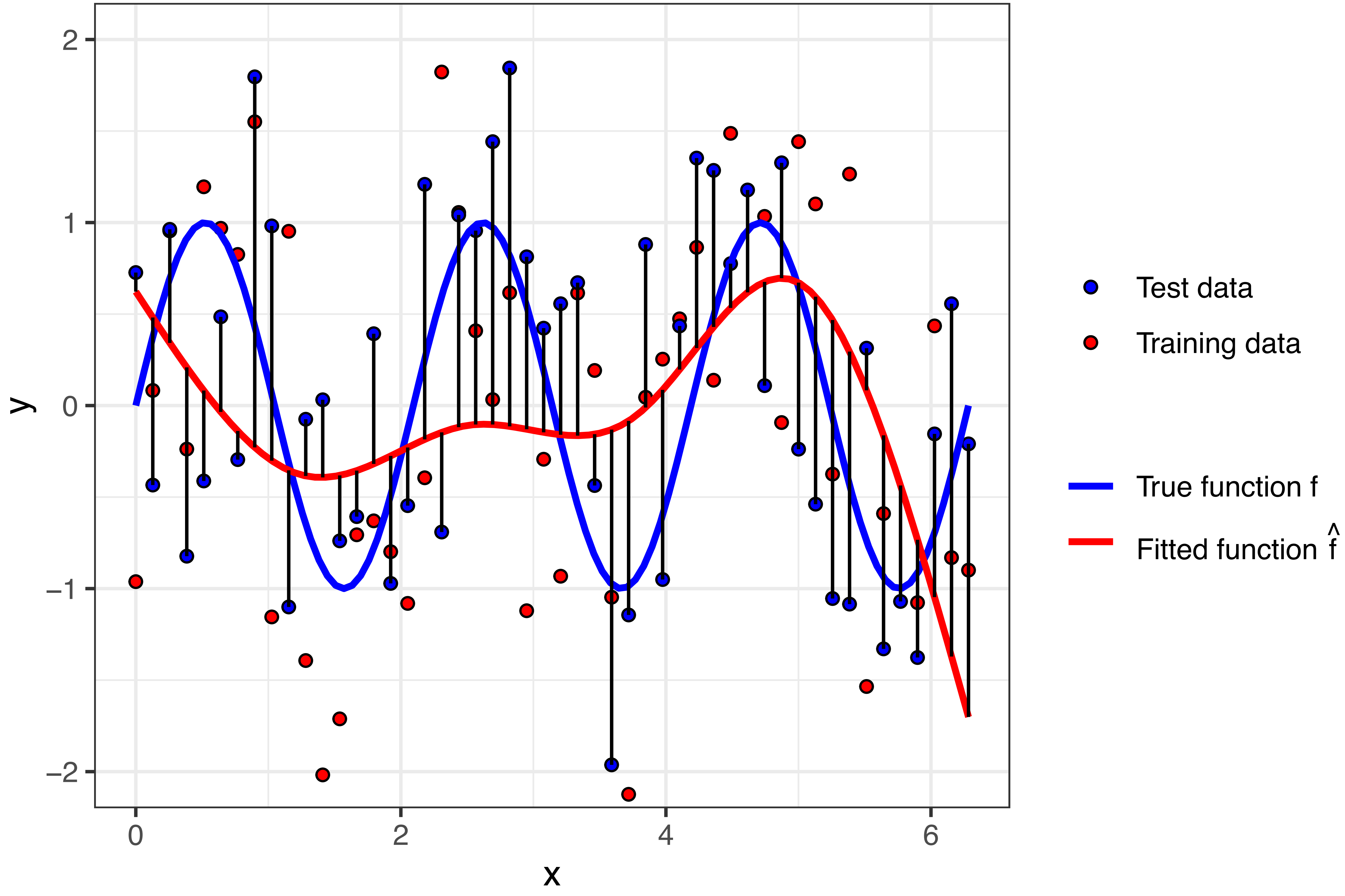
Averaging over randomness in Y^{train} and Y^{test} (think of X^{train} and X^{test} as fixed).

Contribution of randomness in test set



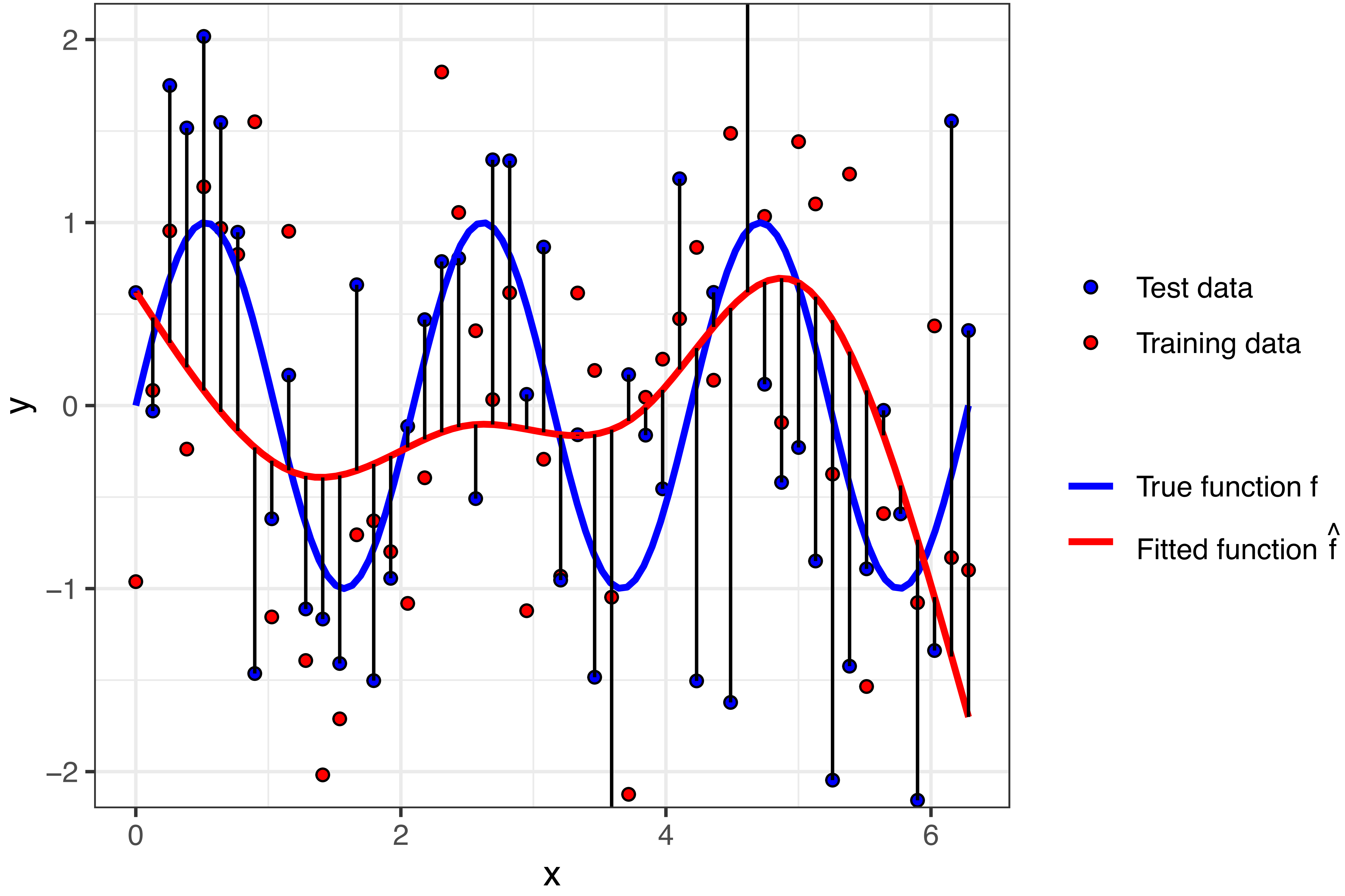
Contribution of randomness in test set

Let's consider different test data sets, while keeping the training data set fixed.



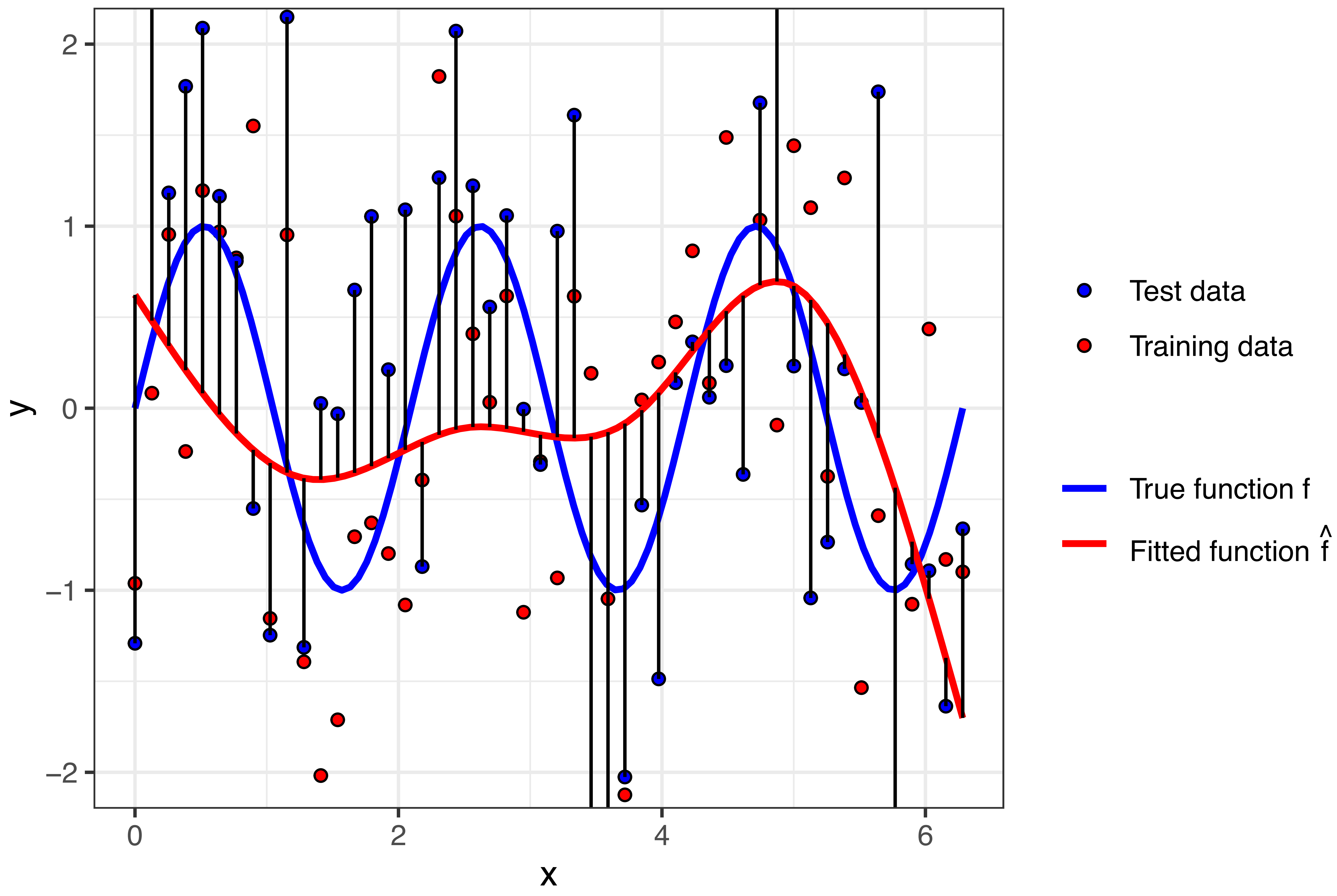
Contribution of randomness in test set

Let's consider different test data sets, while keeping the training data set fixed.



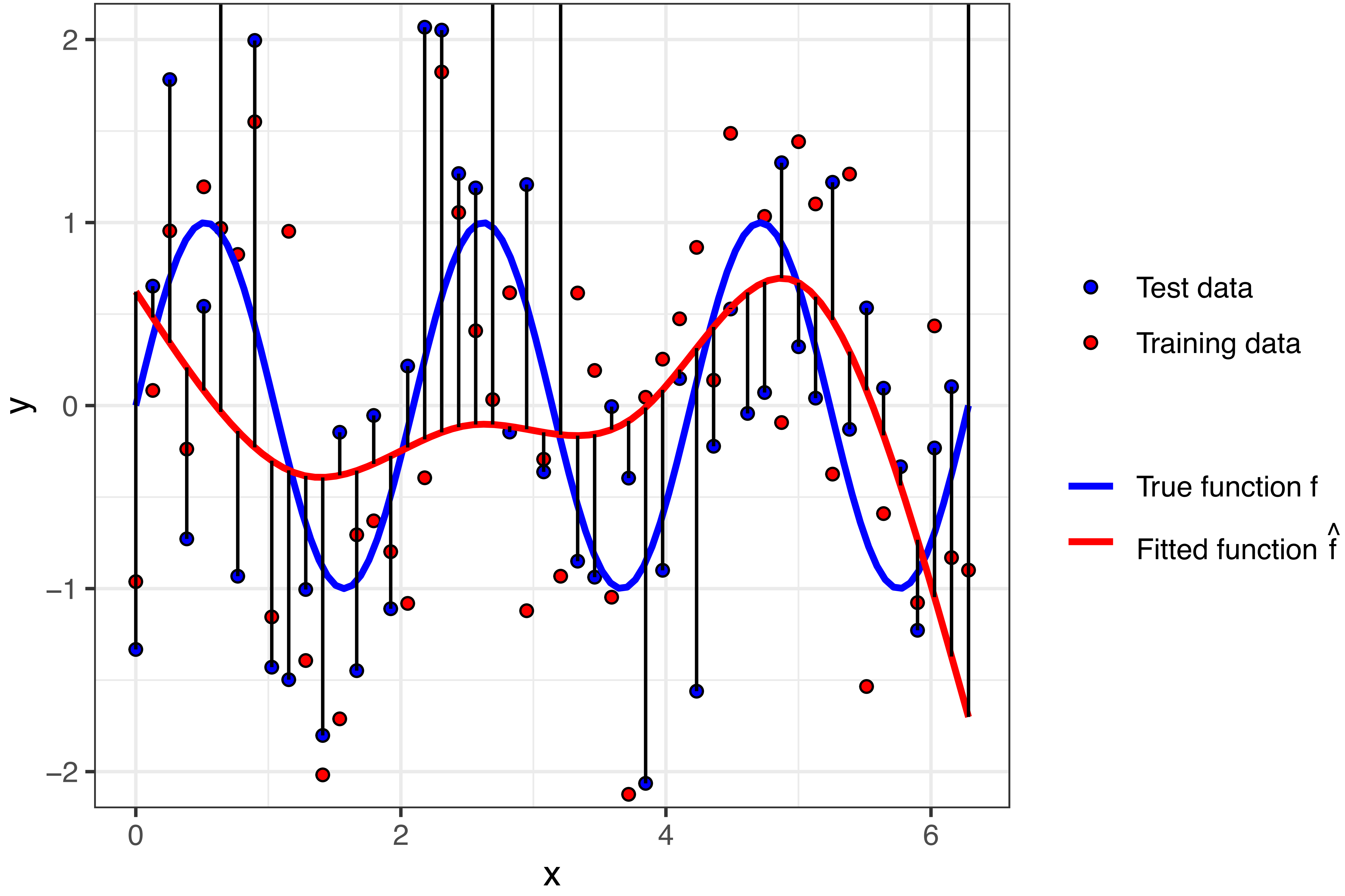
Contribution of randomness in test set

Let's consider different test data sets, while keeping the training data set fixed.

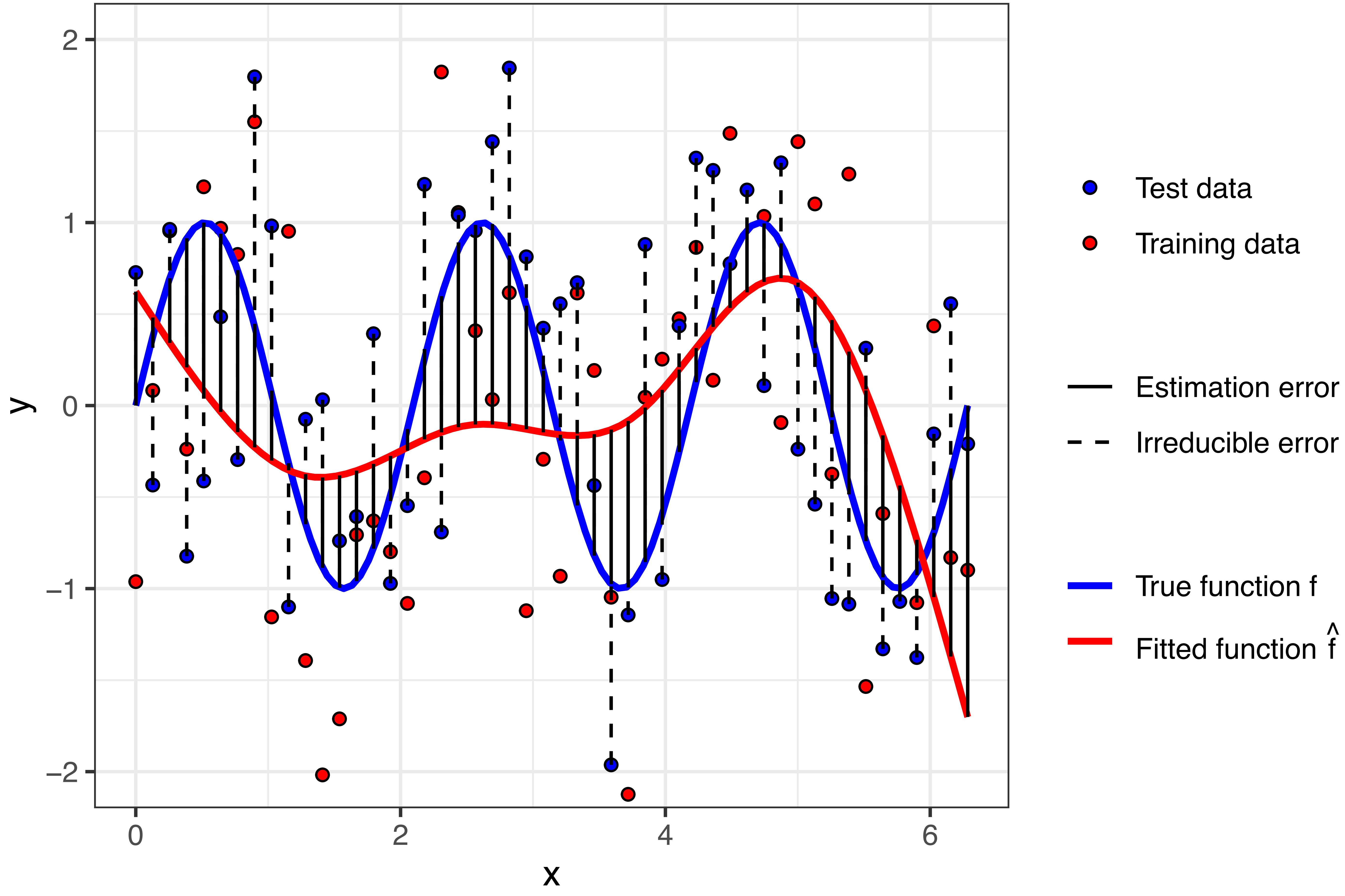


Contribution of randomness in test set

Let's consider different test data sets, while keeping the training data set fixed.

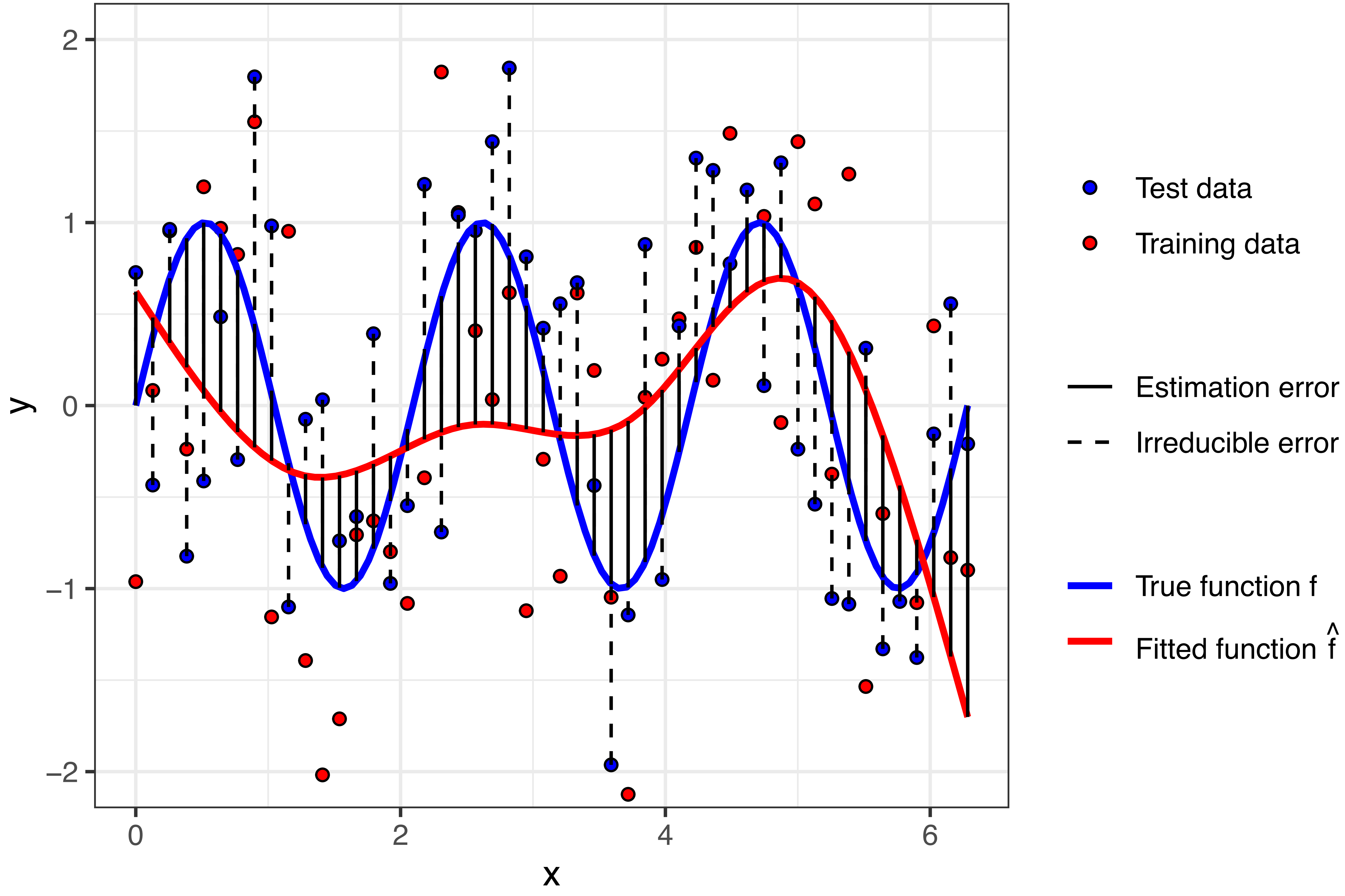


Prediction error = Estimation error + Irreducible error



Prediction error = Estimation error + Irreducible error

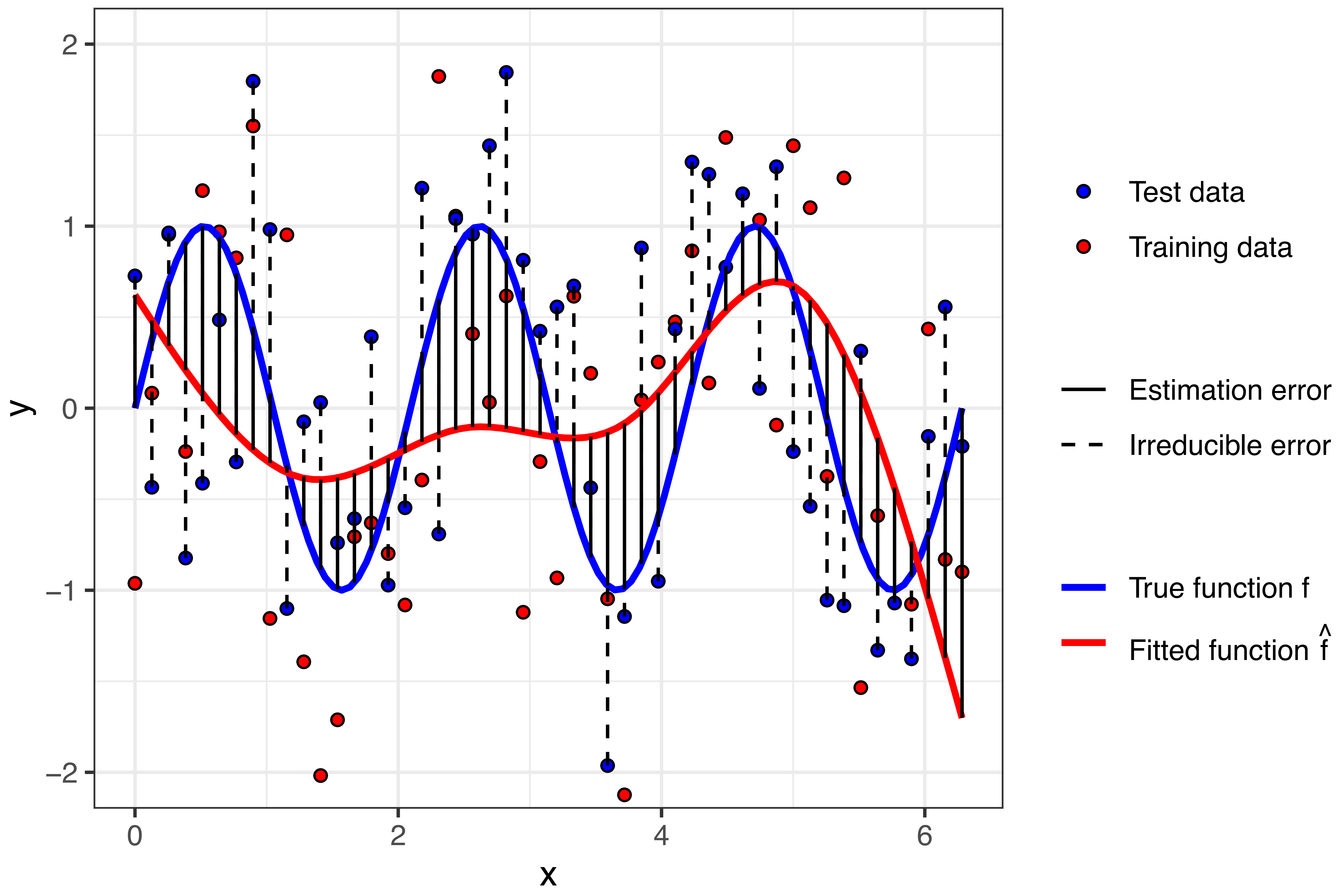
For each test point, part of the prediction error stays fixed (estimation error) and part of it fluctuates (irreducible error).



Prediction error = Estimation error + Irreducible error

For each test point, part of the prediction error stays fixed (estimation error) and part of it fluctuates (irreducible error).

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



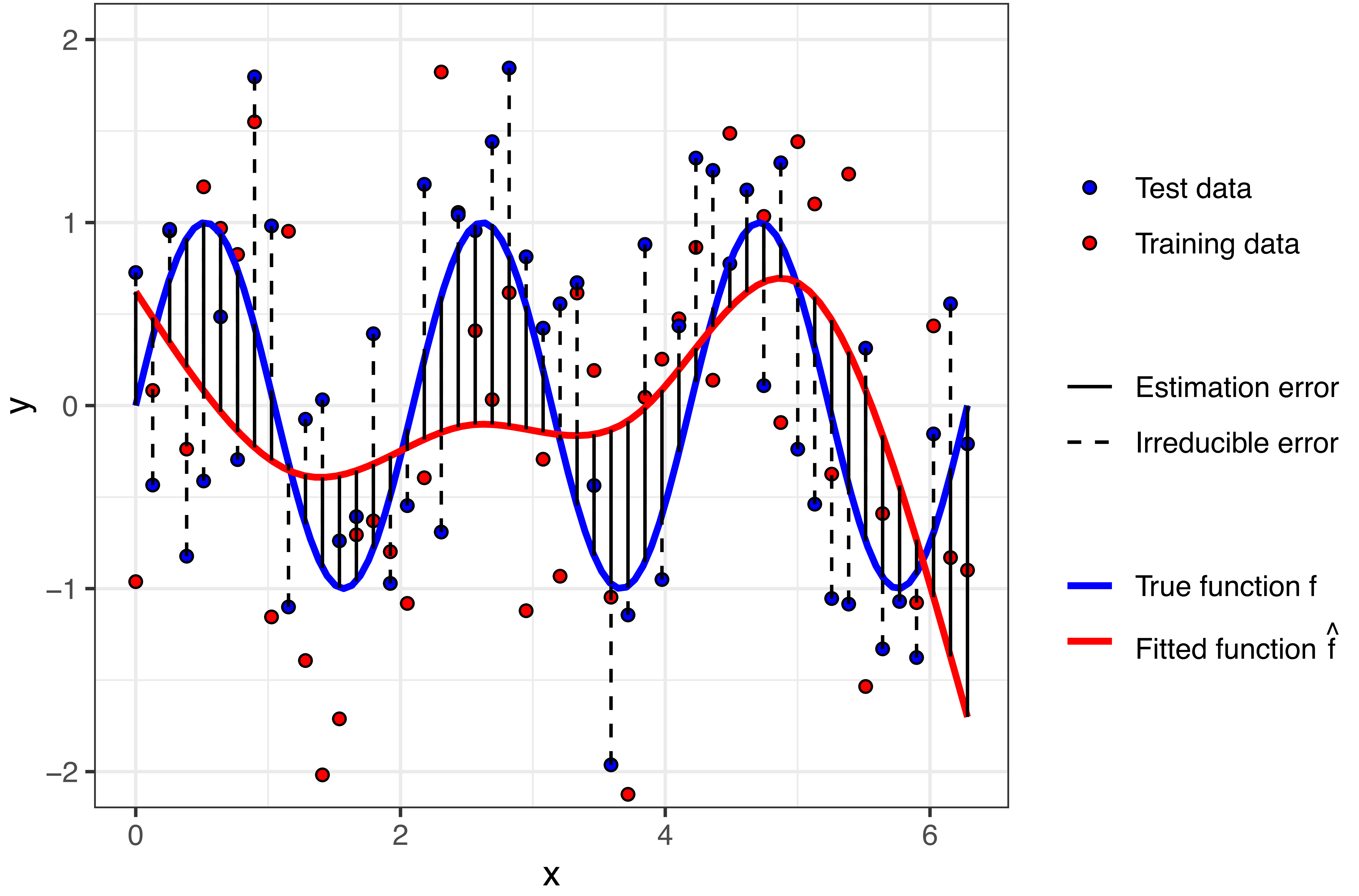
Prediction error = Estimation error + Irreducible error

For each test point, part of the prediction error stays fixed (estimation error) and part of it fluctuates (irreducible error).

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

Then,

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



Prediction error = Estimation error + Irreducible error

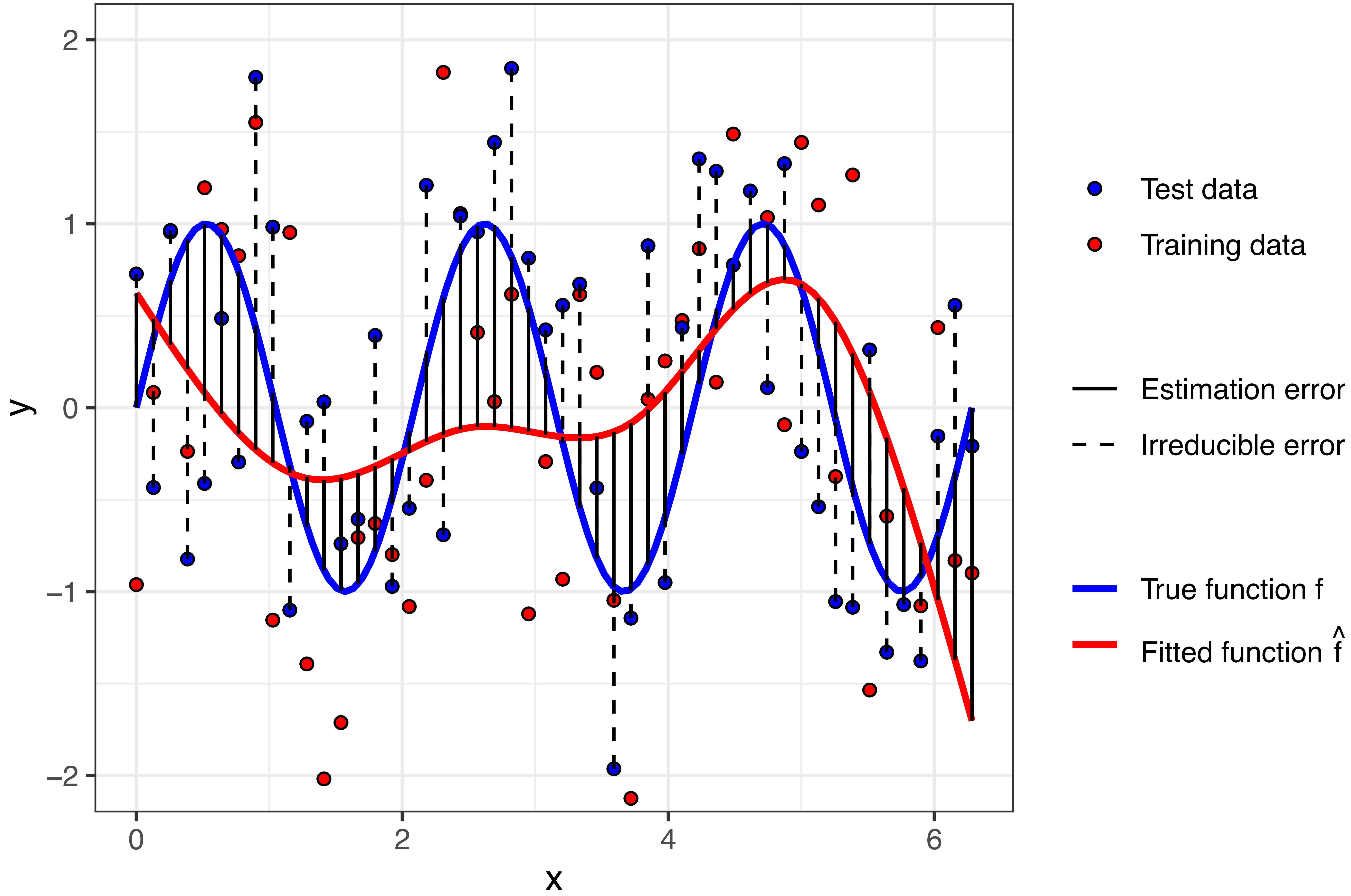
For each test point, part of the prediction error stays fixed (estimation error) and part of it fluctuates (irreducible error).

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

Then,

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

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



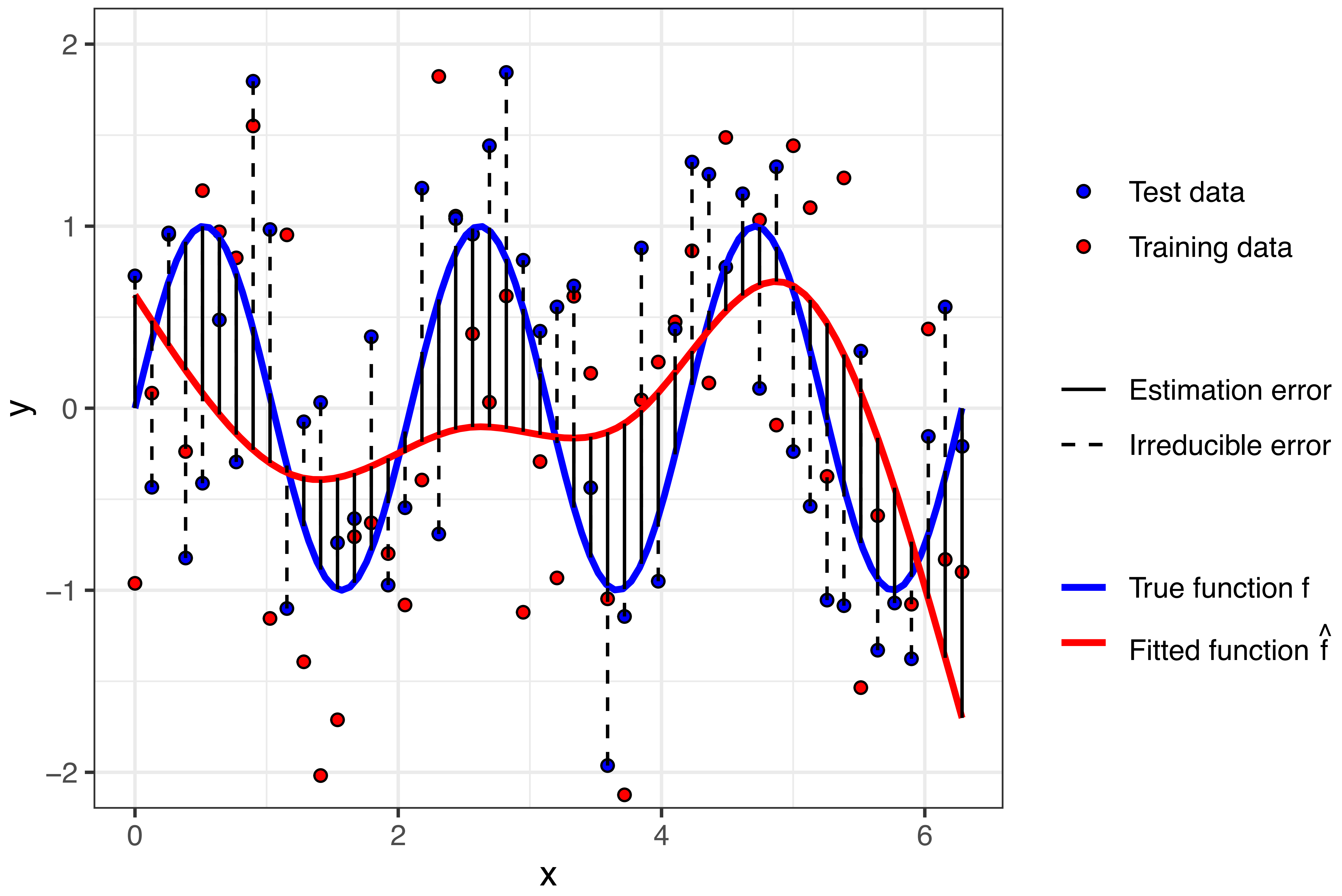
Prediction error = Estimation error + Irreducible error

For each test point, part of the prediction error stays fixed (estimation error) and part of it fluctuates (irreducible error).

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

Then,

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



Prediction error = Estimation error + Irreducible error

For each test point, part of the prediction error stays fixed (estimation error) and part of it fluctuates (irreducible error).

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

Then,

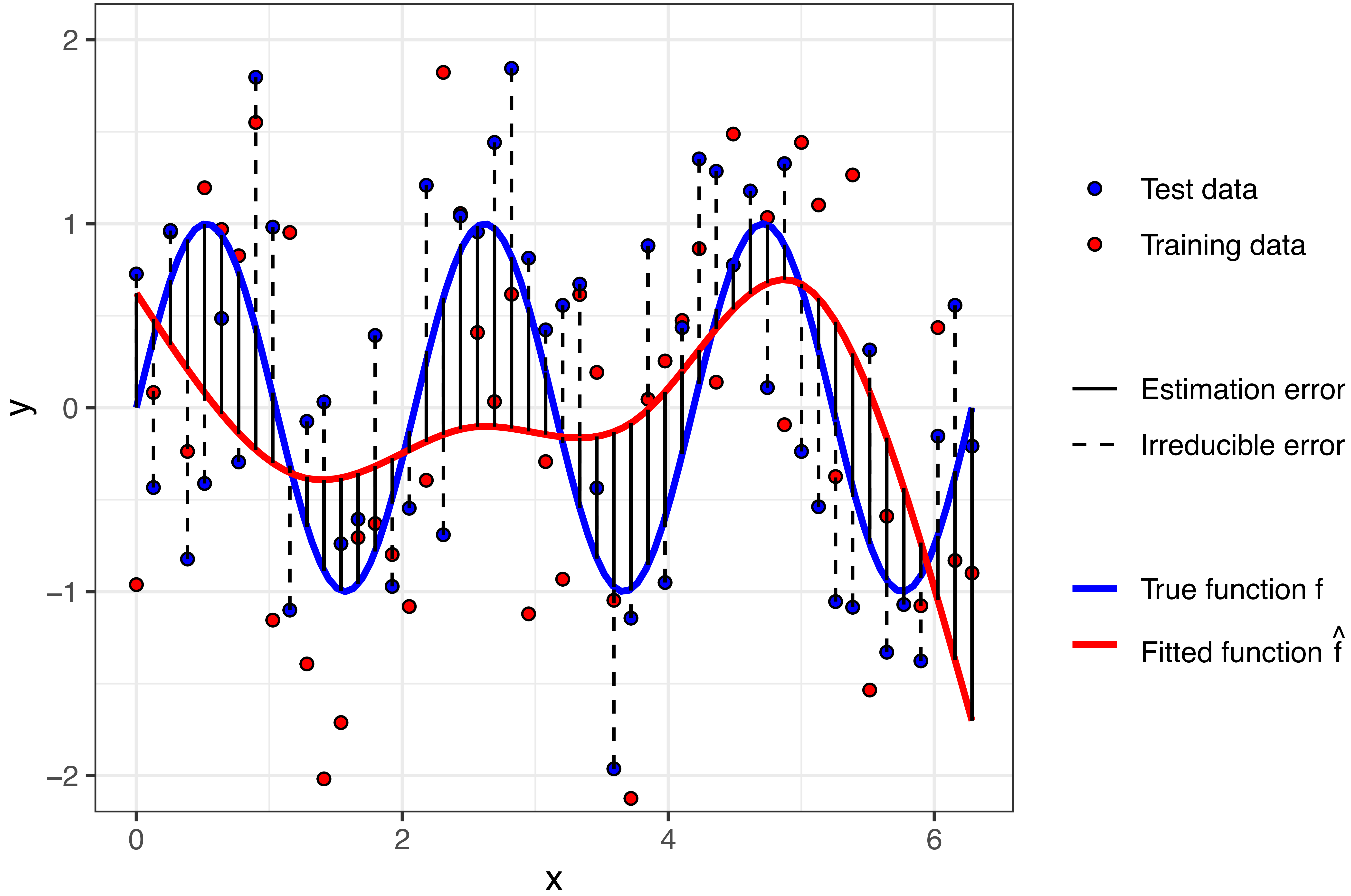
$$\text{ETE}_i = \mathbb{E}[(Y_i^{\text{test}} - \hat{f}(X_i^{\text{test}}))^2]$$

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

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

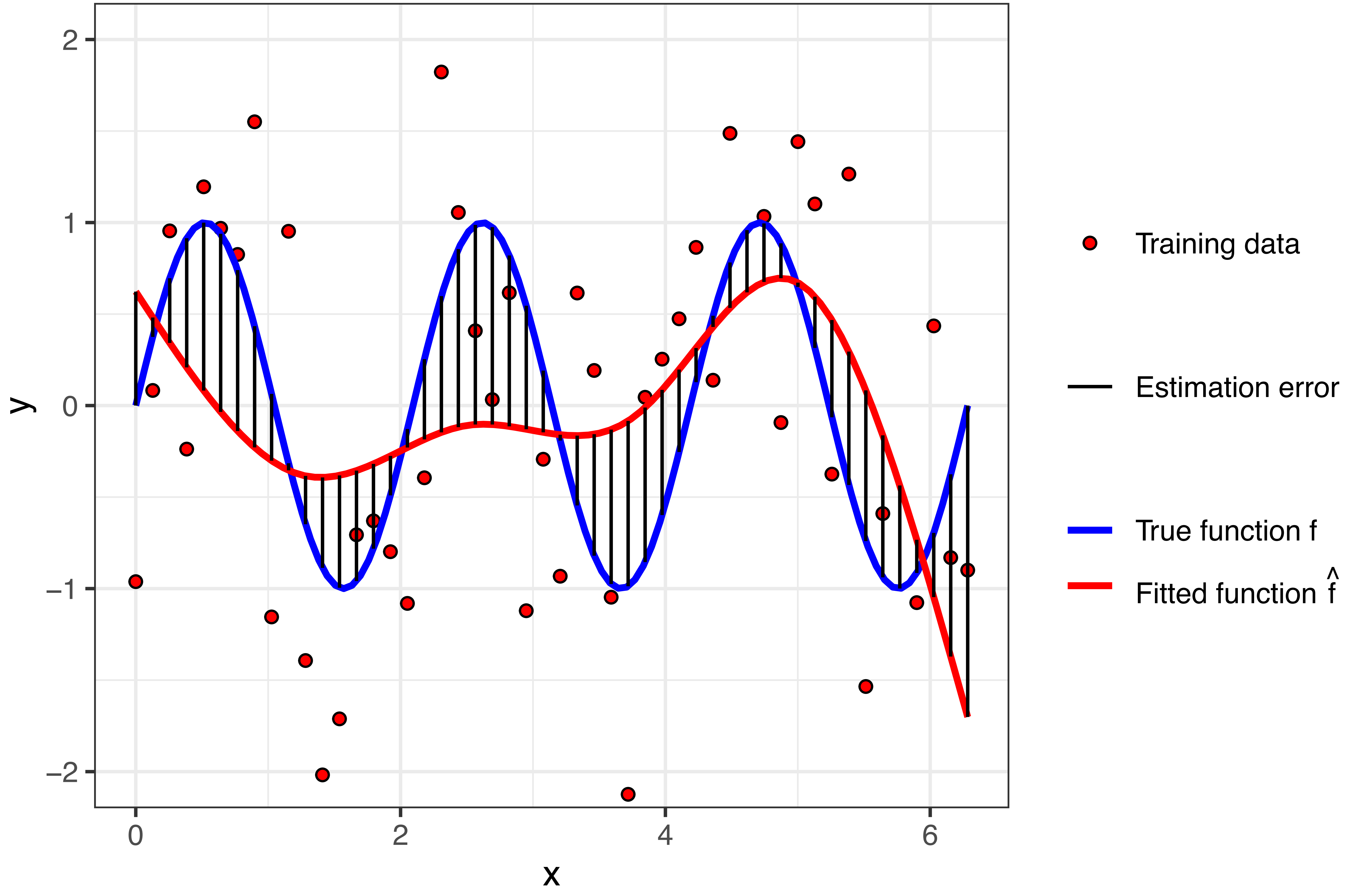
estimation error

irreducible error



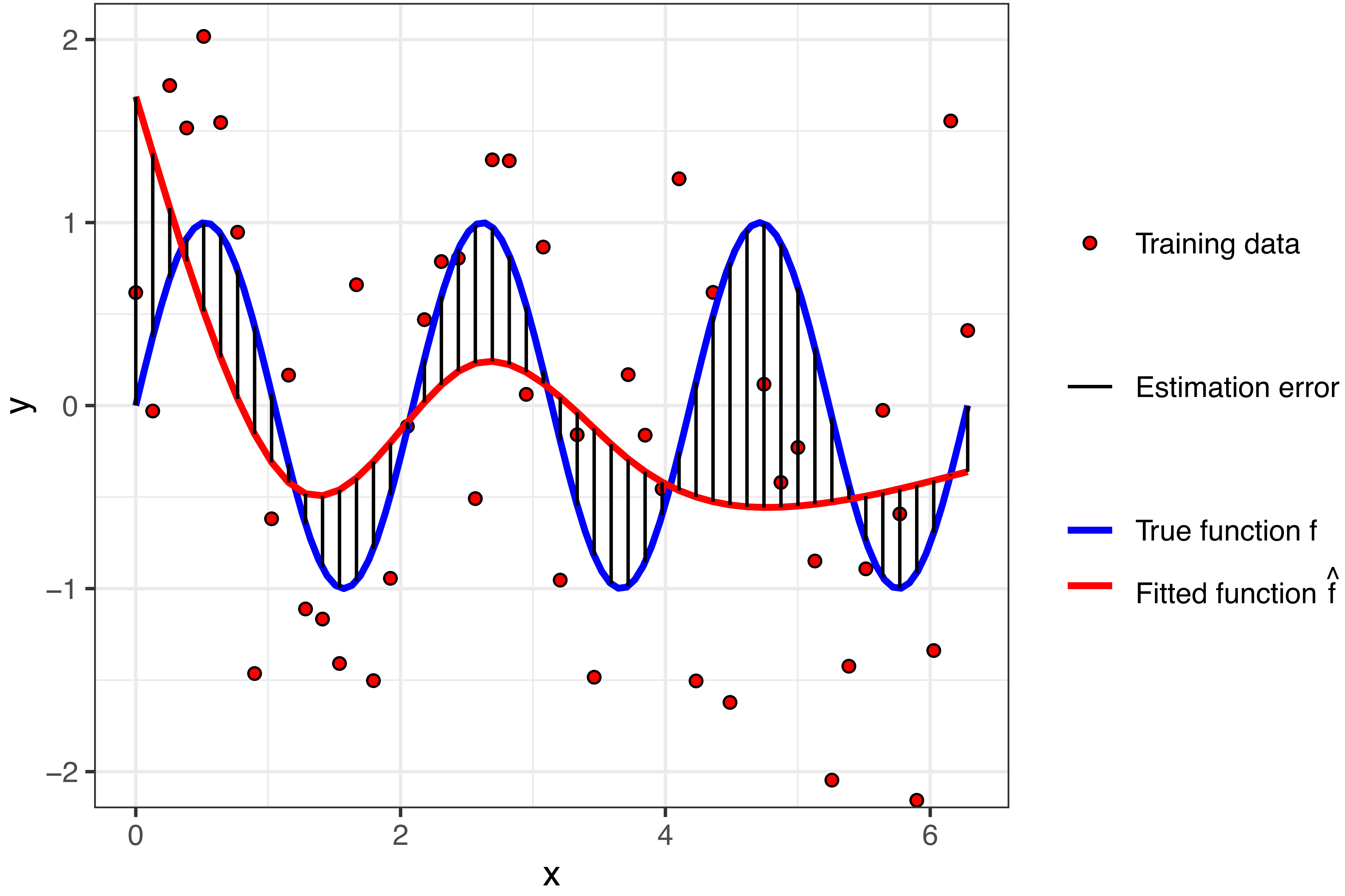
Contribution of randomness in training set

How estimation error $f(X_i^{\text{test}}) - \hat{f}(X_i^{\text{test}})$ varies as function of the training set.



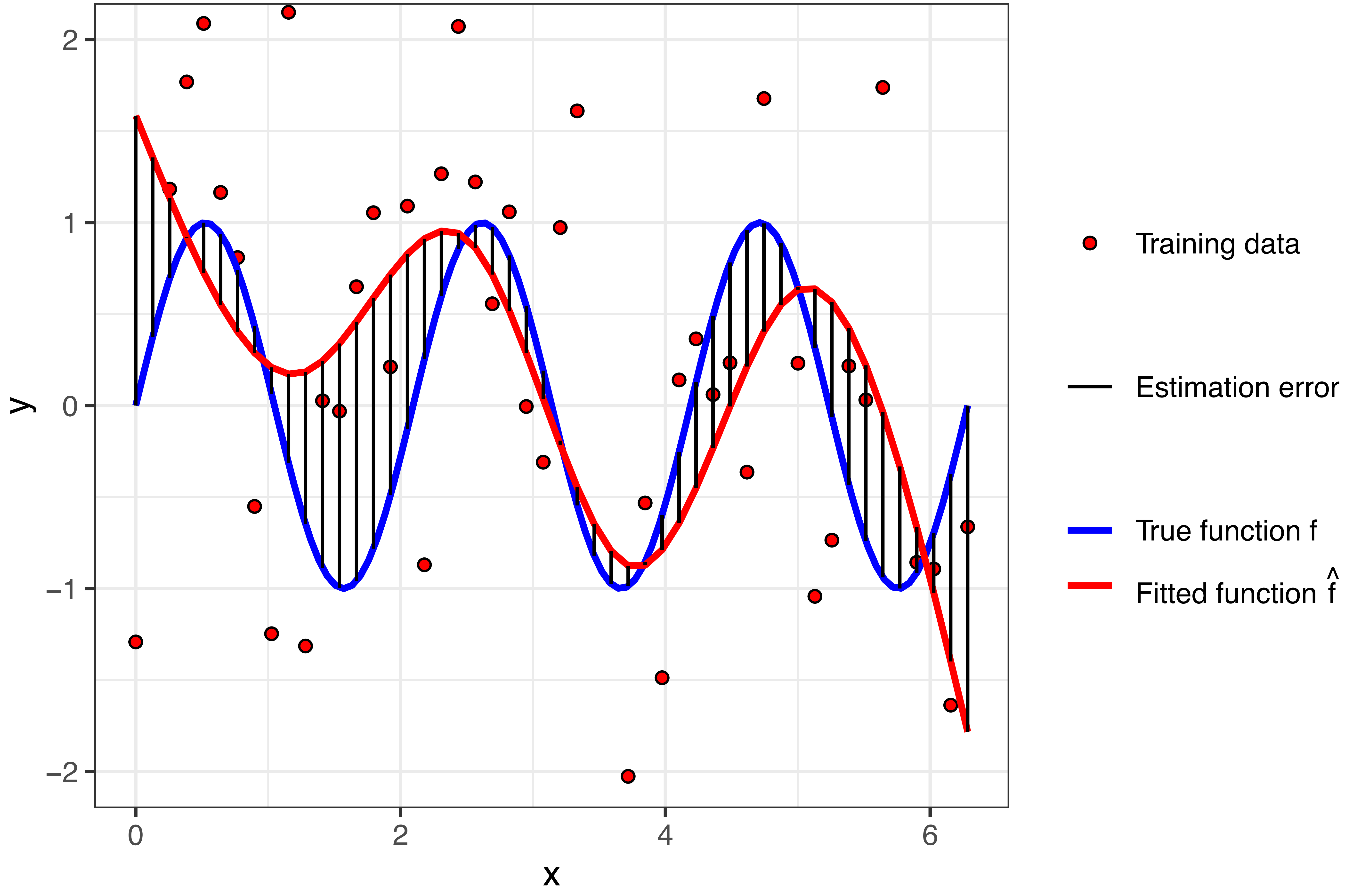
Contribution of randomness in training set

How estimation error $f(X_i^{\text{test}}) - \hat{f}(X_i^{\text{test}})$ varies as function of the training set.



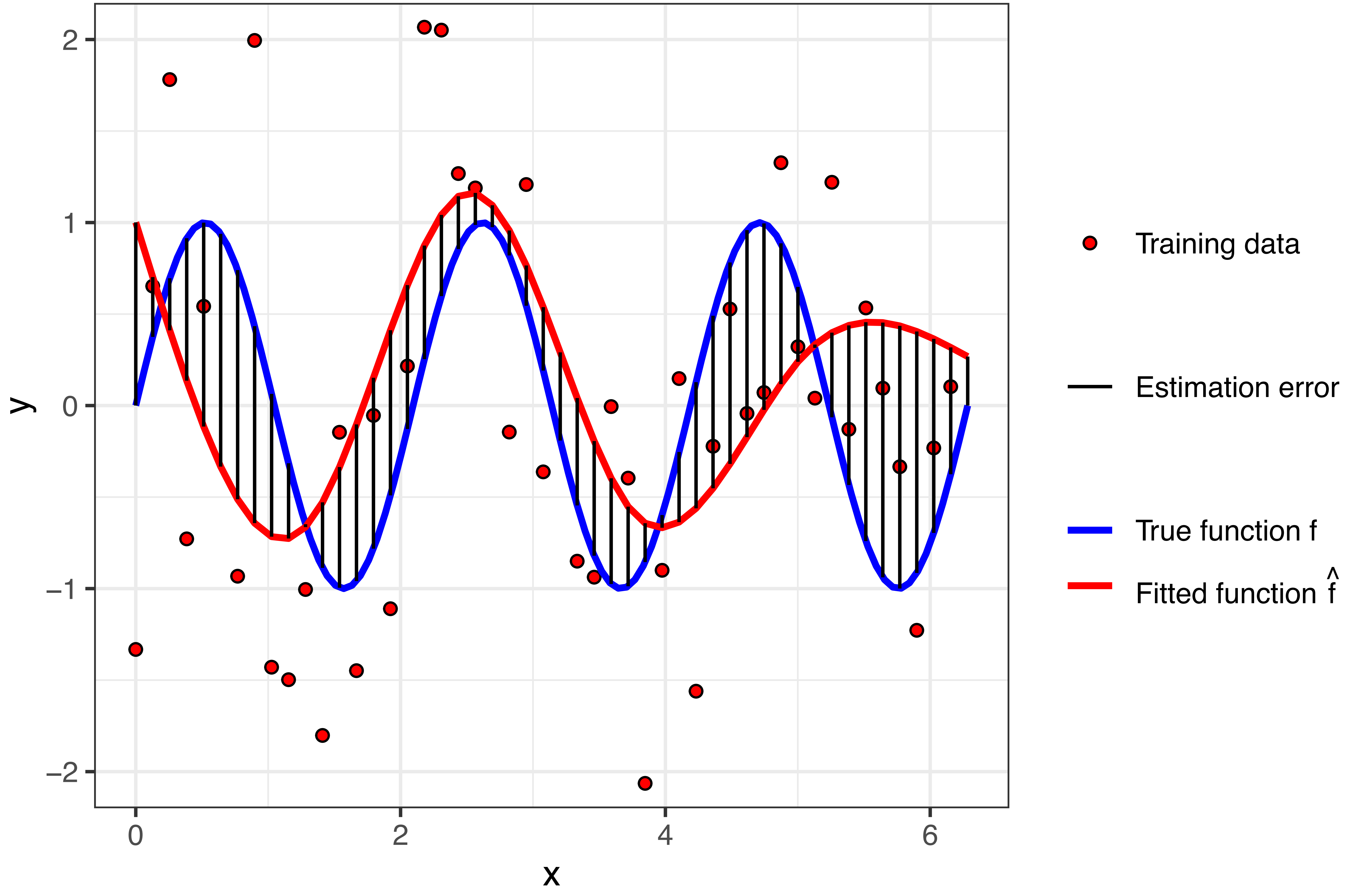
Contribution of randomness in training set

How estimation error $f(X_i^{\text{test}}) - \hat{f}(X_i^{\text{test}})$ varies as function of the training set.

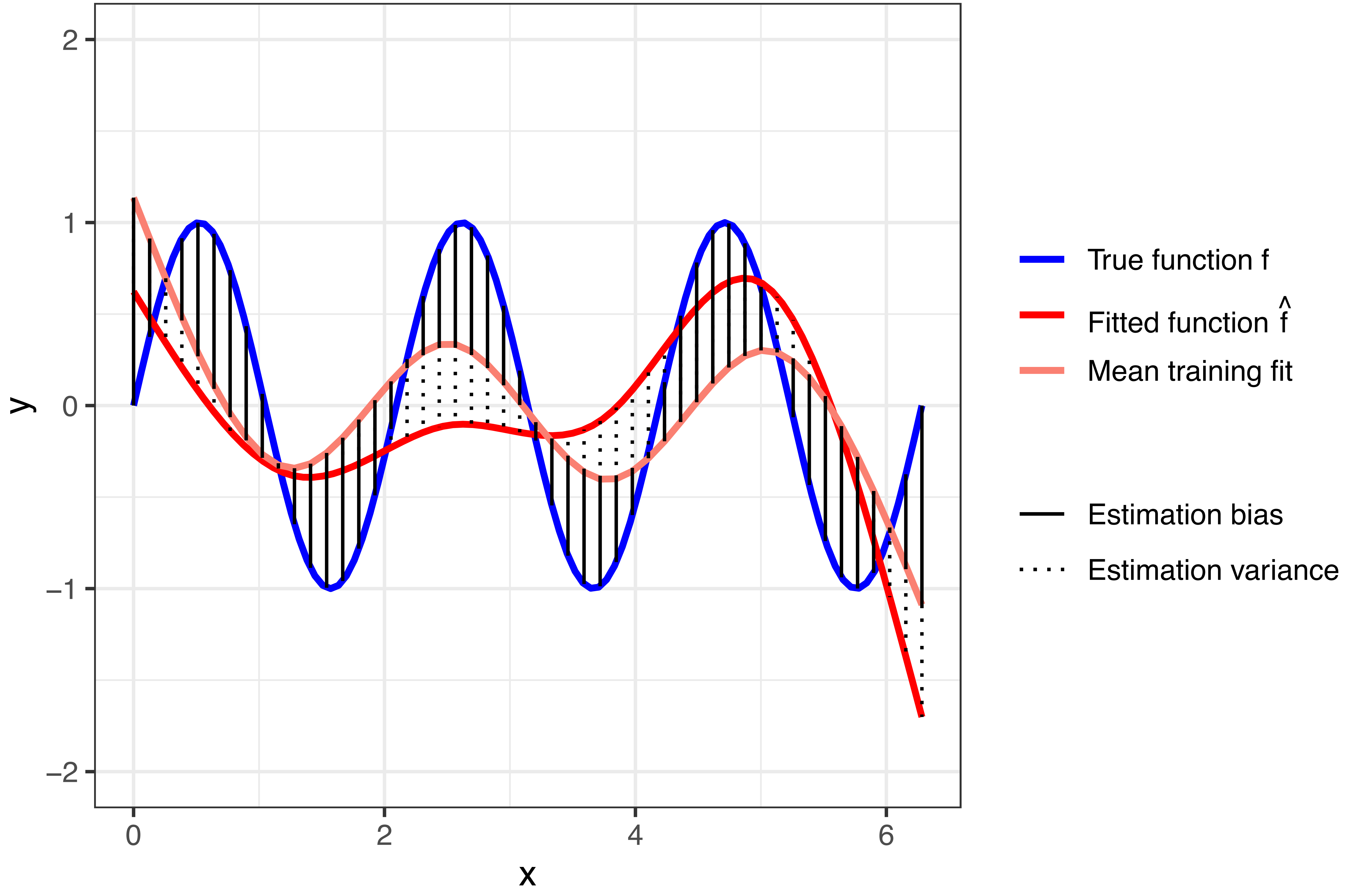


Contribution of randomness in training set

How estimation error $f(X_i^{\text{test}}) - \hat{f}(X_i^{\text{test}})$ varies as function of the training set.

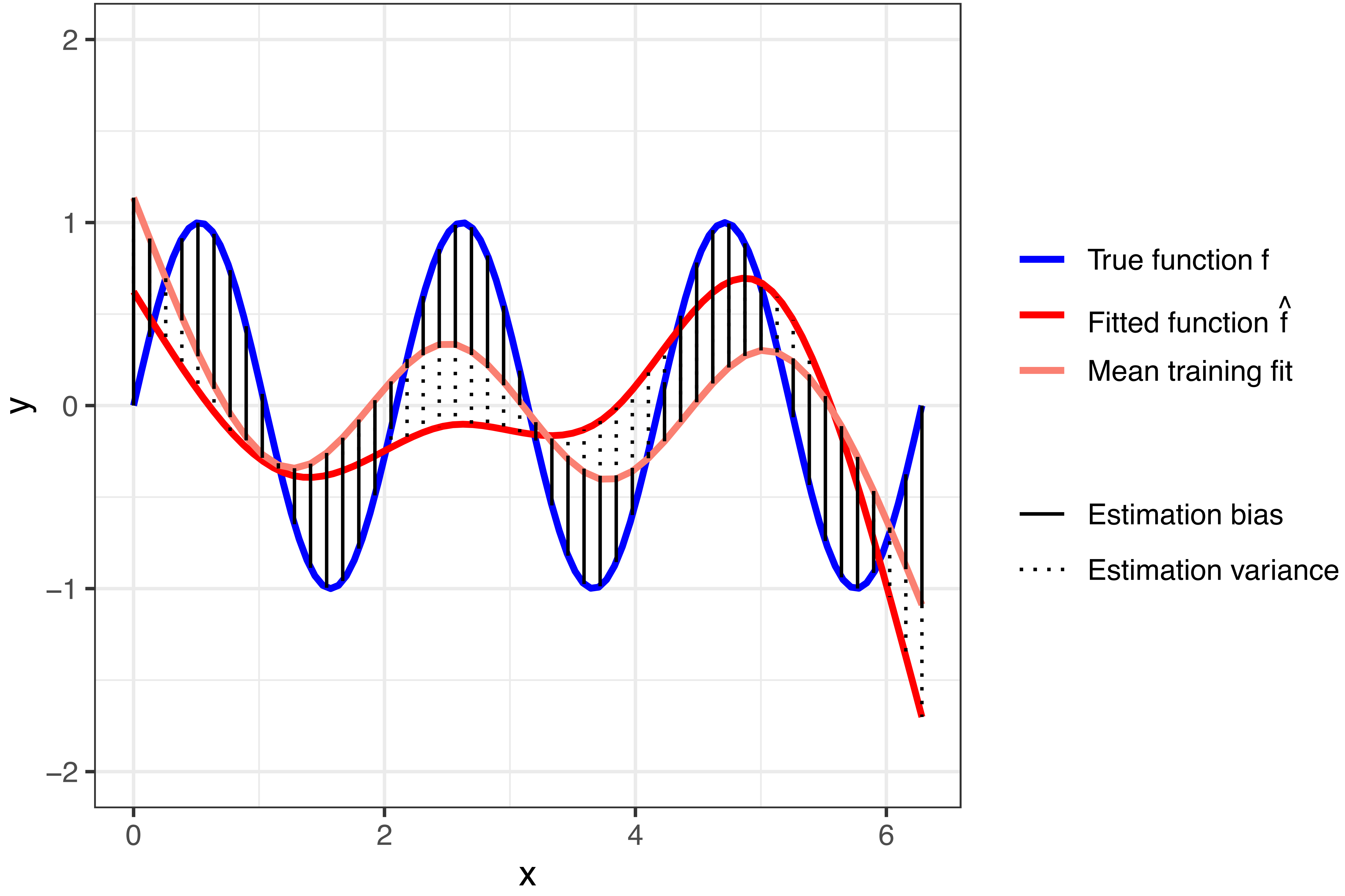


Estimation error = Bias² + Variance



Estimation error = Bias^2 + Variance

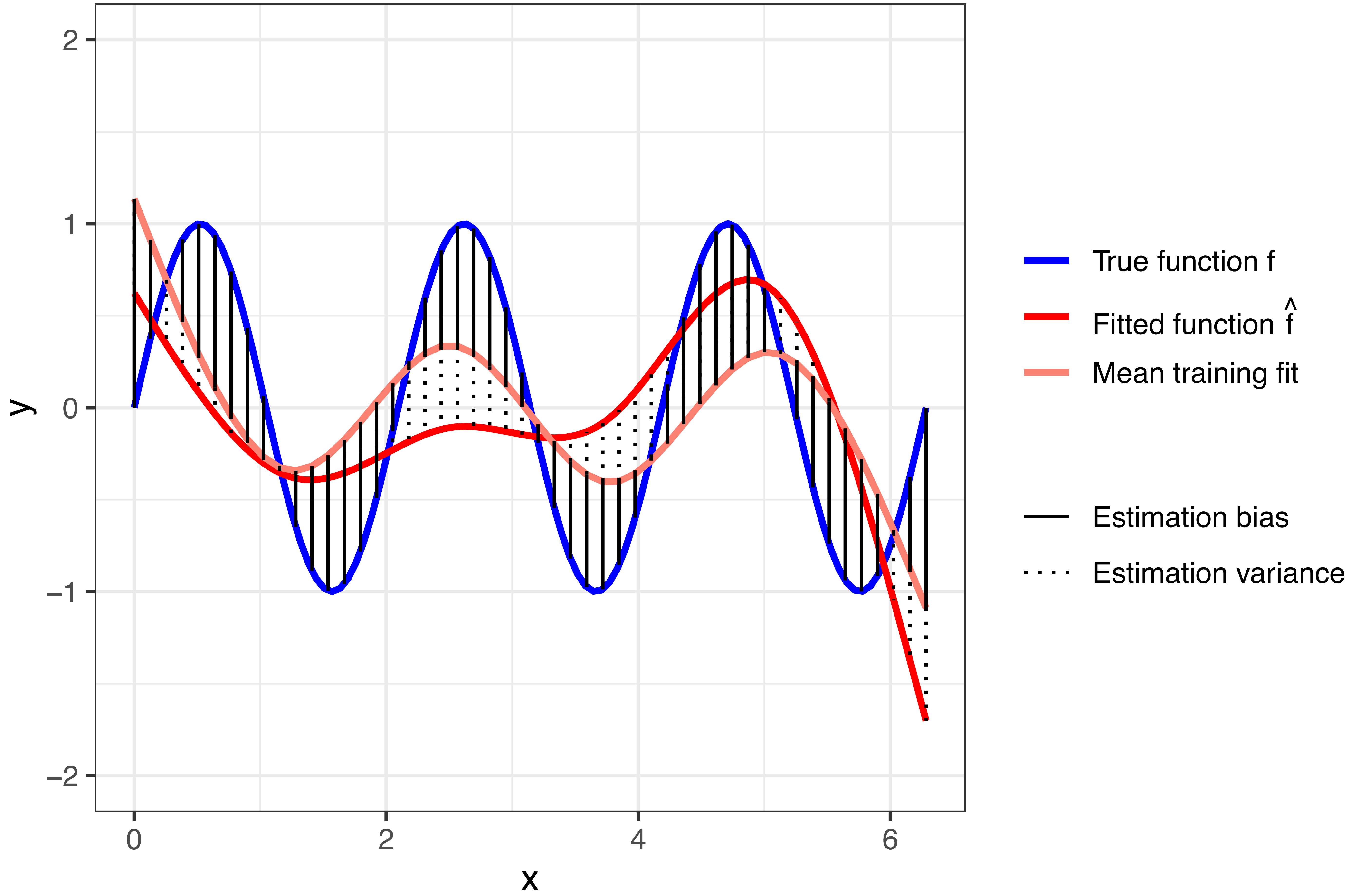
For each test point, part of the estimation error stays fixed (bias) and part of it fluctuates (variance):



Estimation error = Bias^2 + Variance

For each test point, part of the estimation error stays fixed (bias) and part of it fluctuates (variance):

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



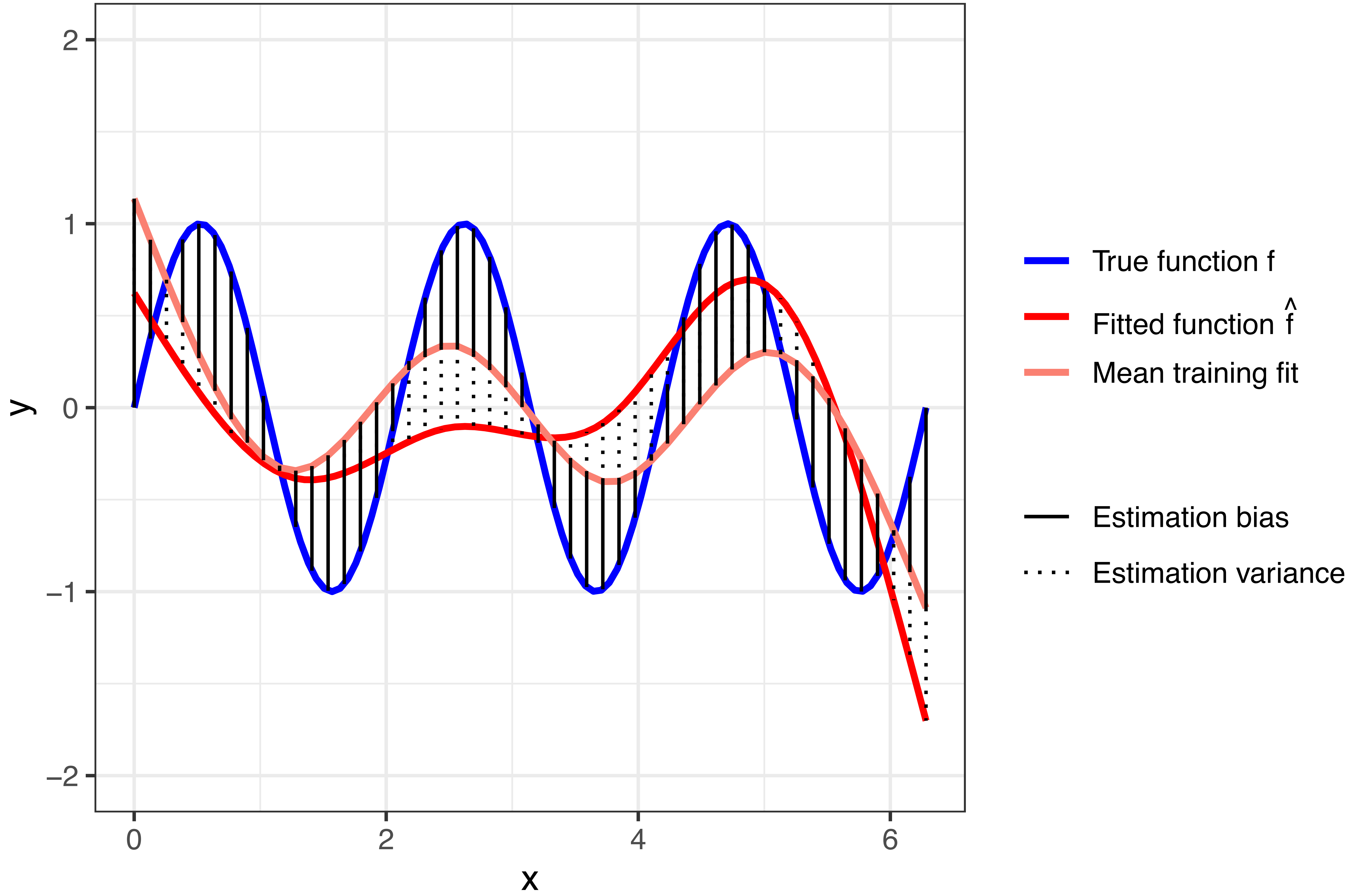
Estimation error = Bias² + Variance

For each test point, part of the estimation error stays fixed (bias) and part of it fluctuates (variance):

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

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

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



Estimation error = Bias² + Variance

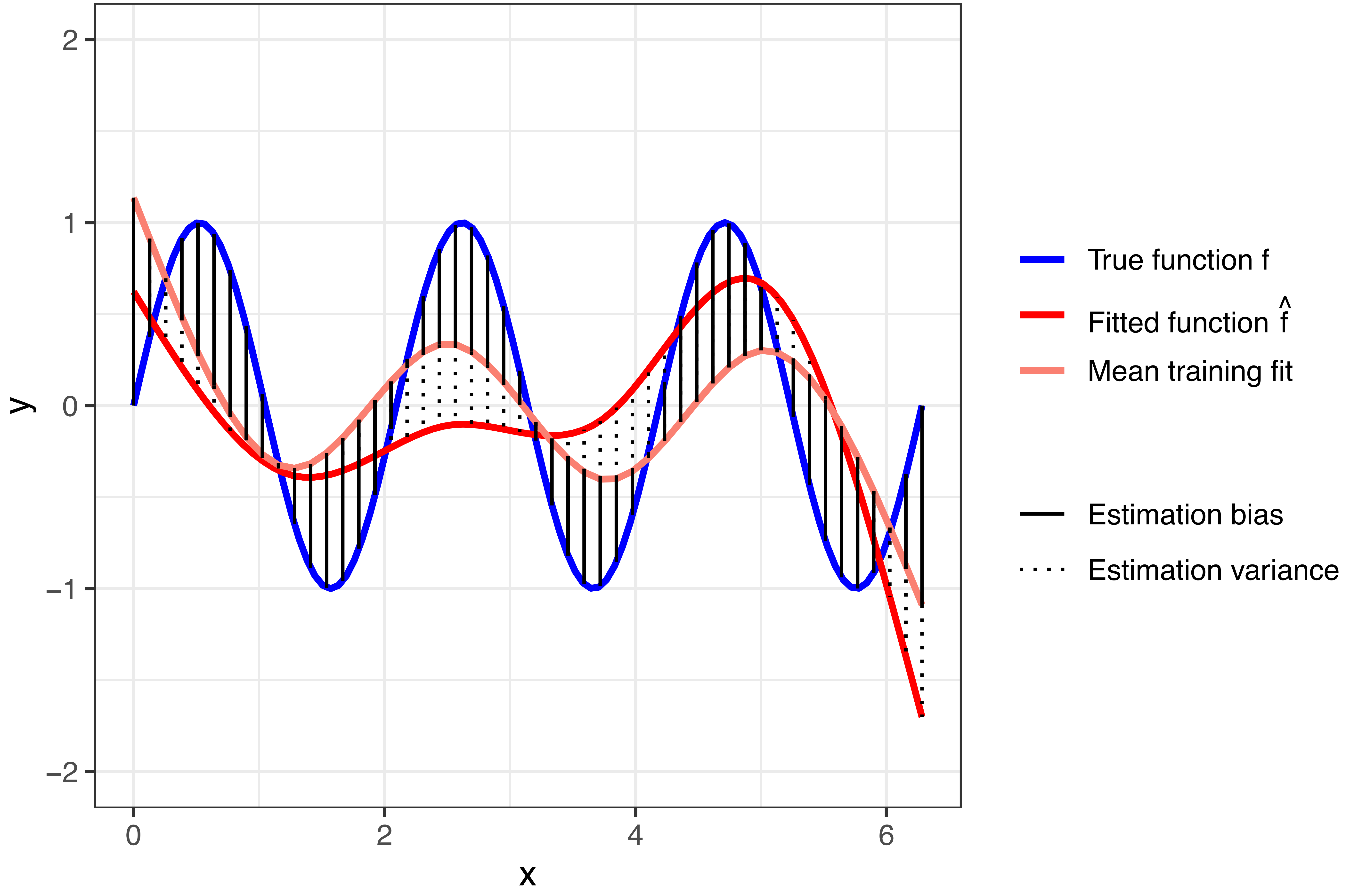
For each test point, part of the estimation error stays fixed (bias) and part of it fluctuates (variance):

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

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

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

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



Estimation error = Bias² + Variance

For each test point, part of the estimation error stays fixed (bias) and part of it fluctuates (variance):

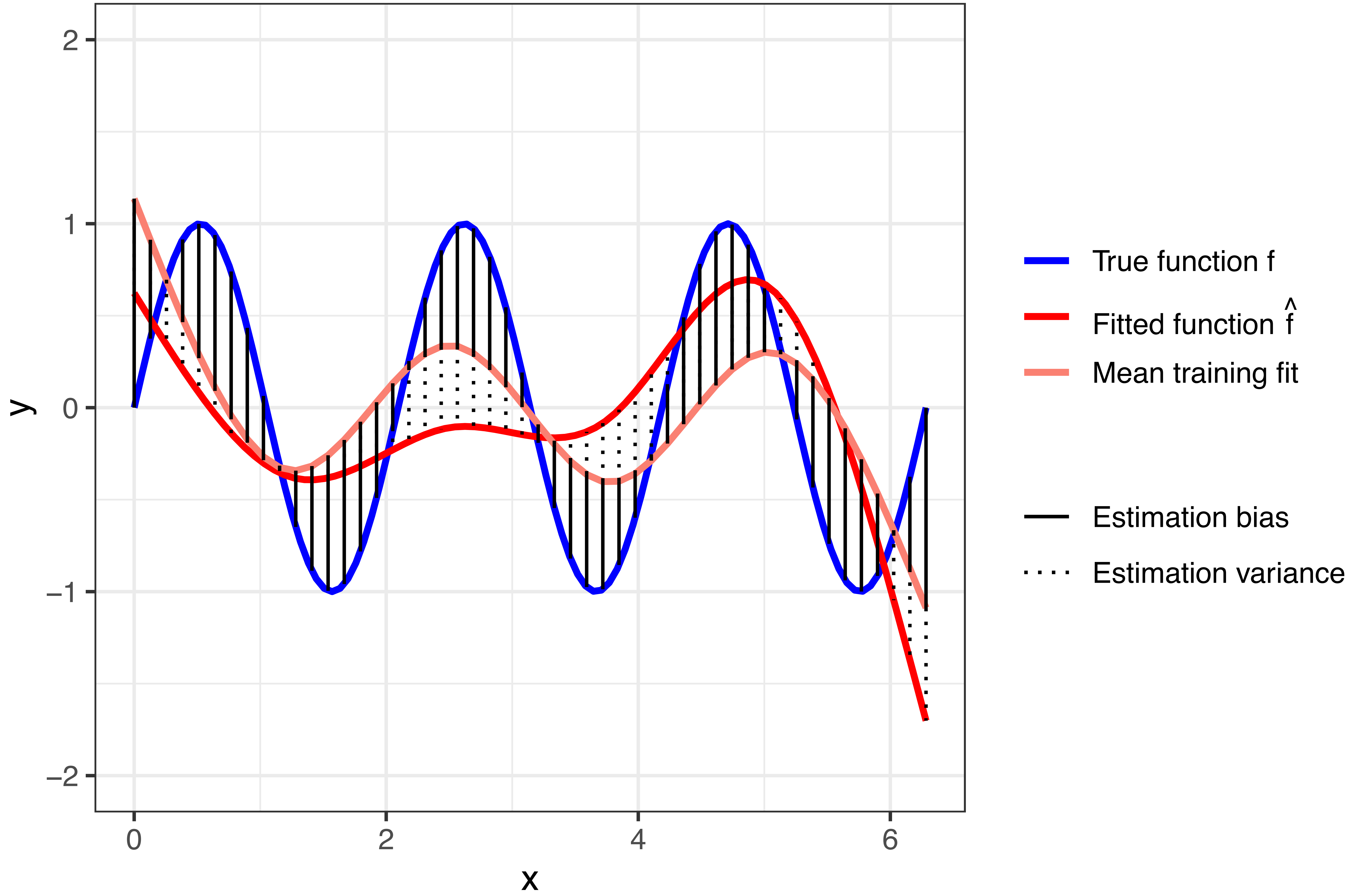
$$\mathbb{E}[(f(X_i^{\text{test}}) - \hat{f}(X_i^{\text{test}}))^2]$$

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

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

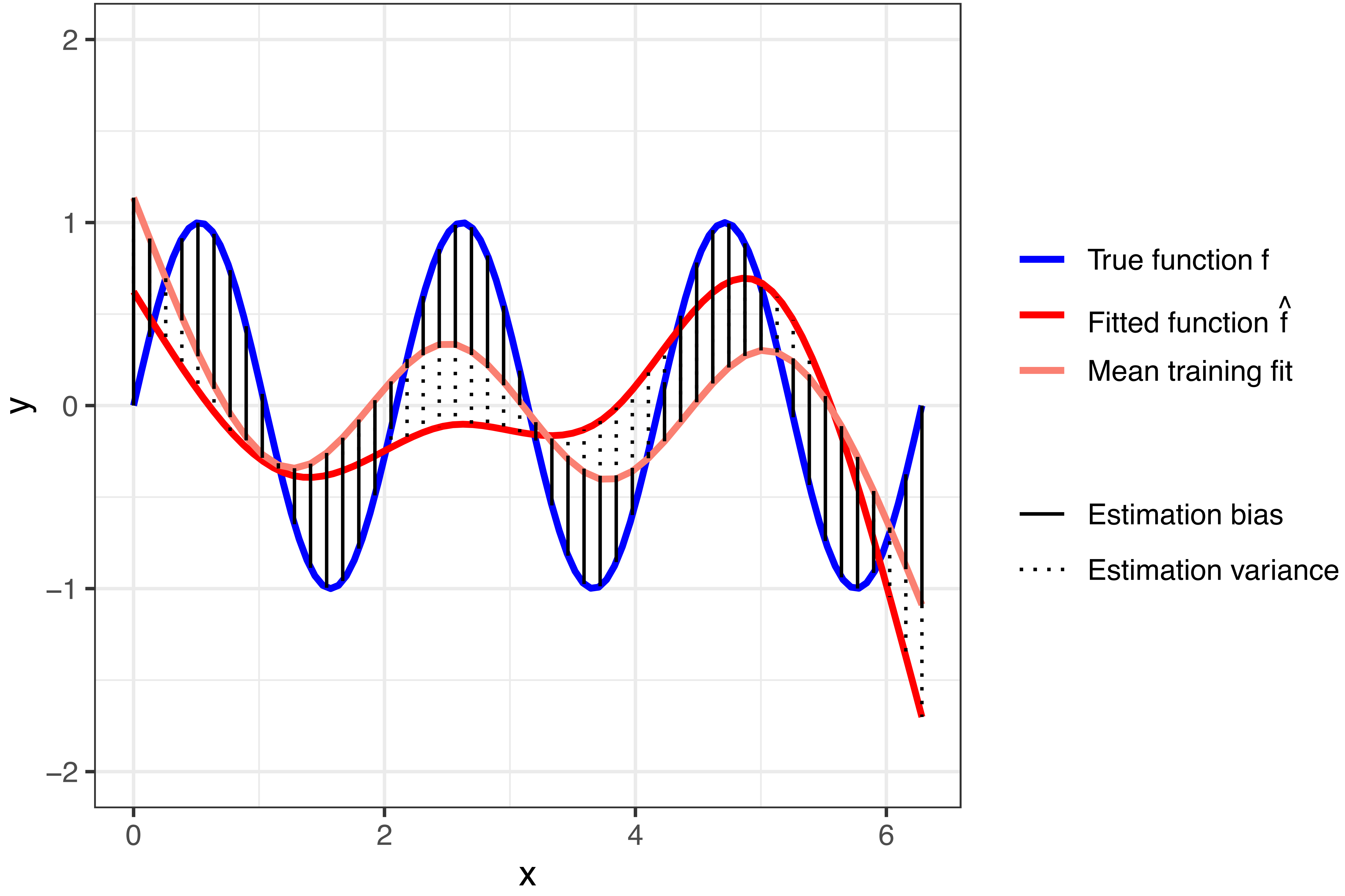
$$= \text{Bias}_i^2 + \text{Variance}_i$$

This is the bias-variance decomposition.



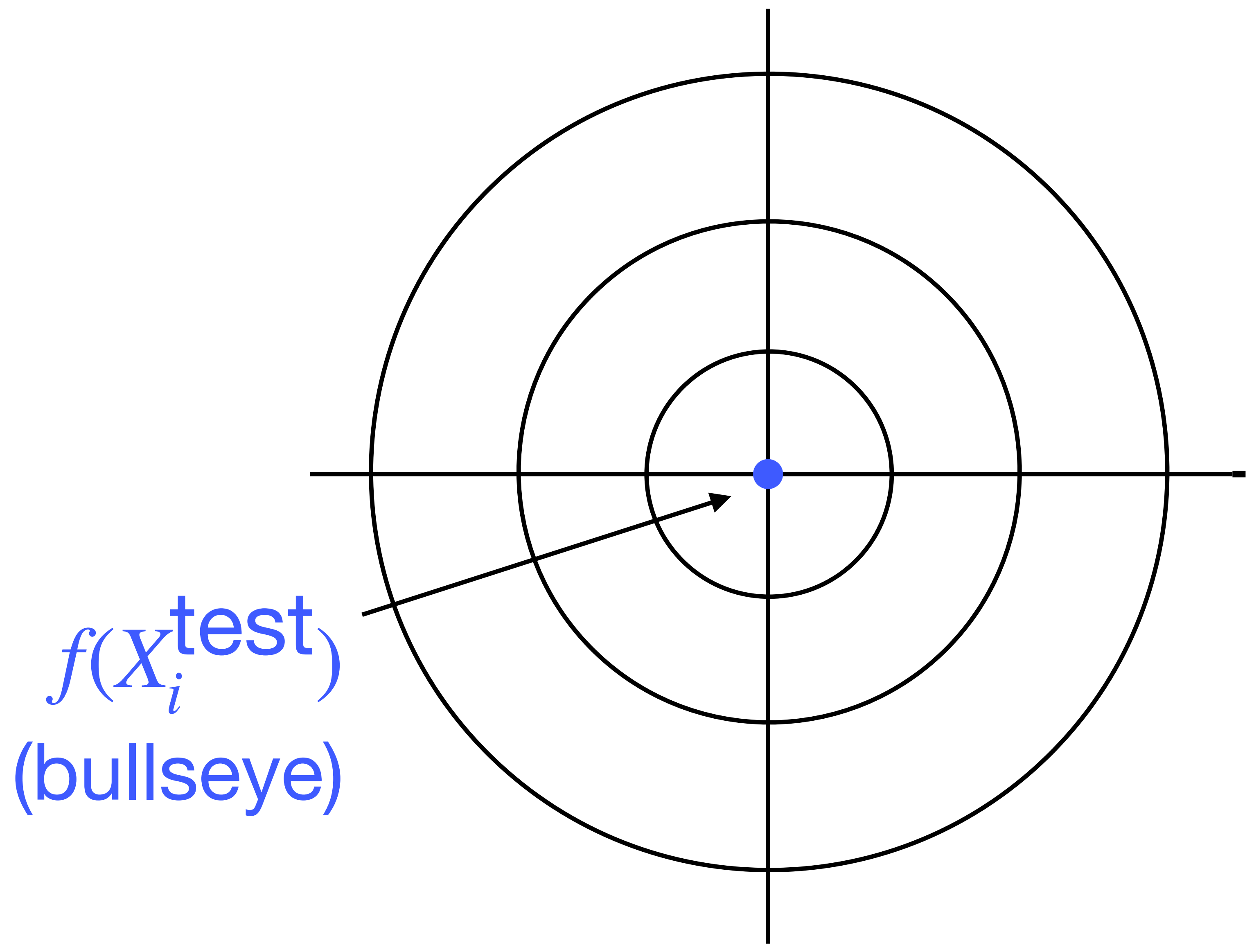
Estimation error = Bias^2 + Variance

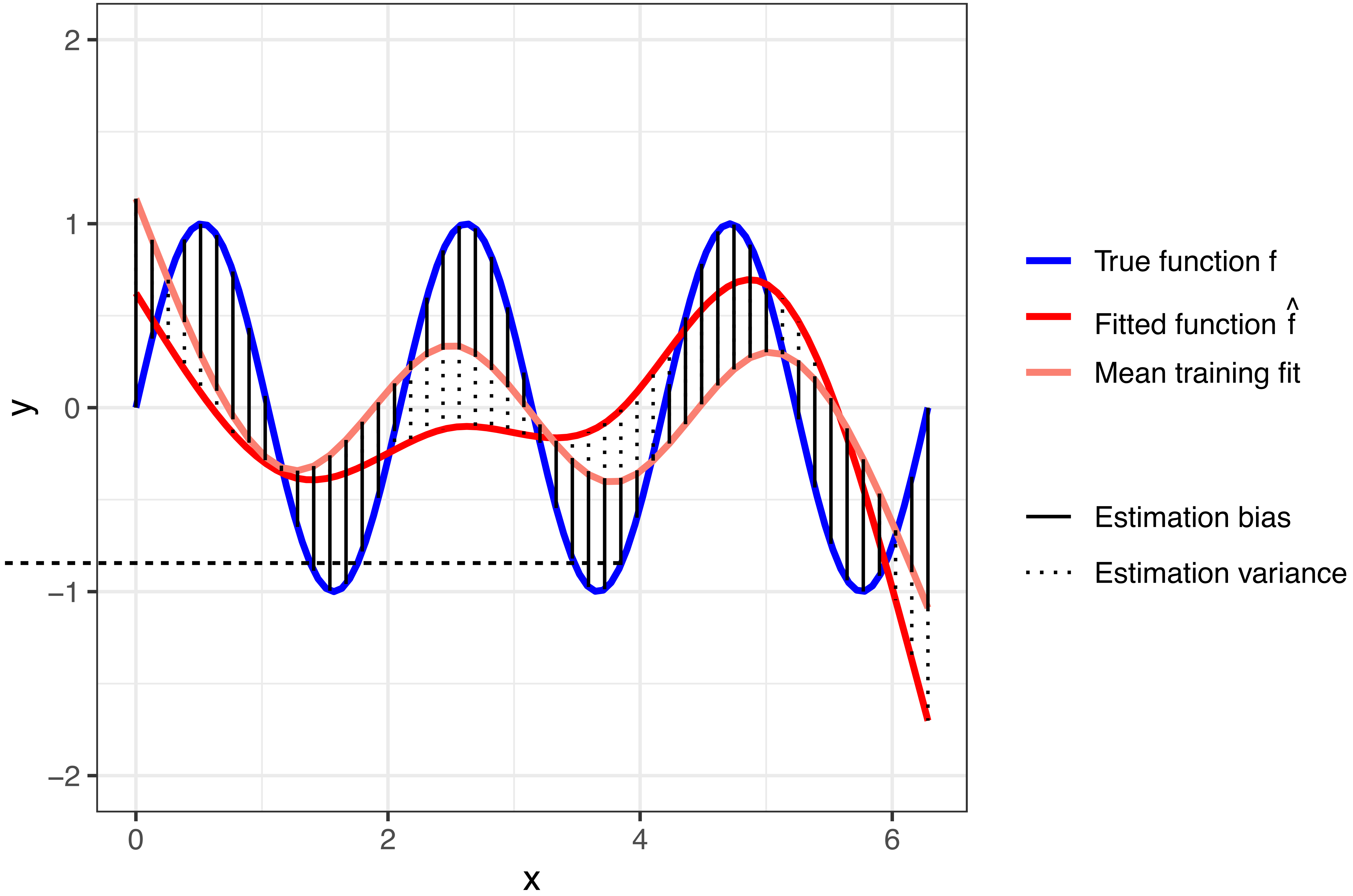
Consider a test point X_i^{test}. Each training data set leads to a prediction $\hat{f}(X_i^{\text{test}})$, which is like throwing a dart at $f(X_i^{\text{test}})$.



Estimation error = Bias^2 + Variance

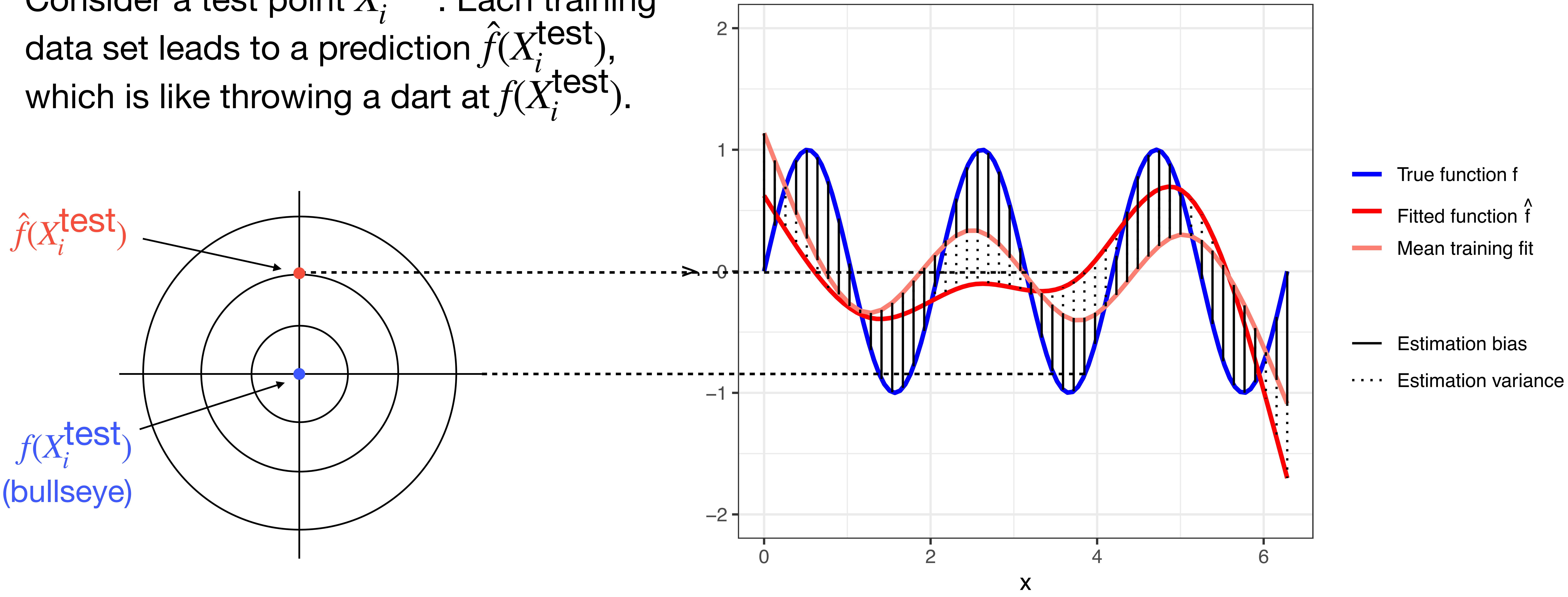
Consider a test point X_i^{test}. Each training data set leads to a prediction $\hat{f}(X_i^{\text{test}})$, which is like throwing a dart at $f(X_i^{\text{test}})$.





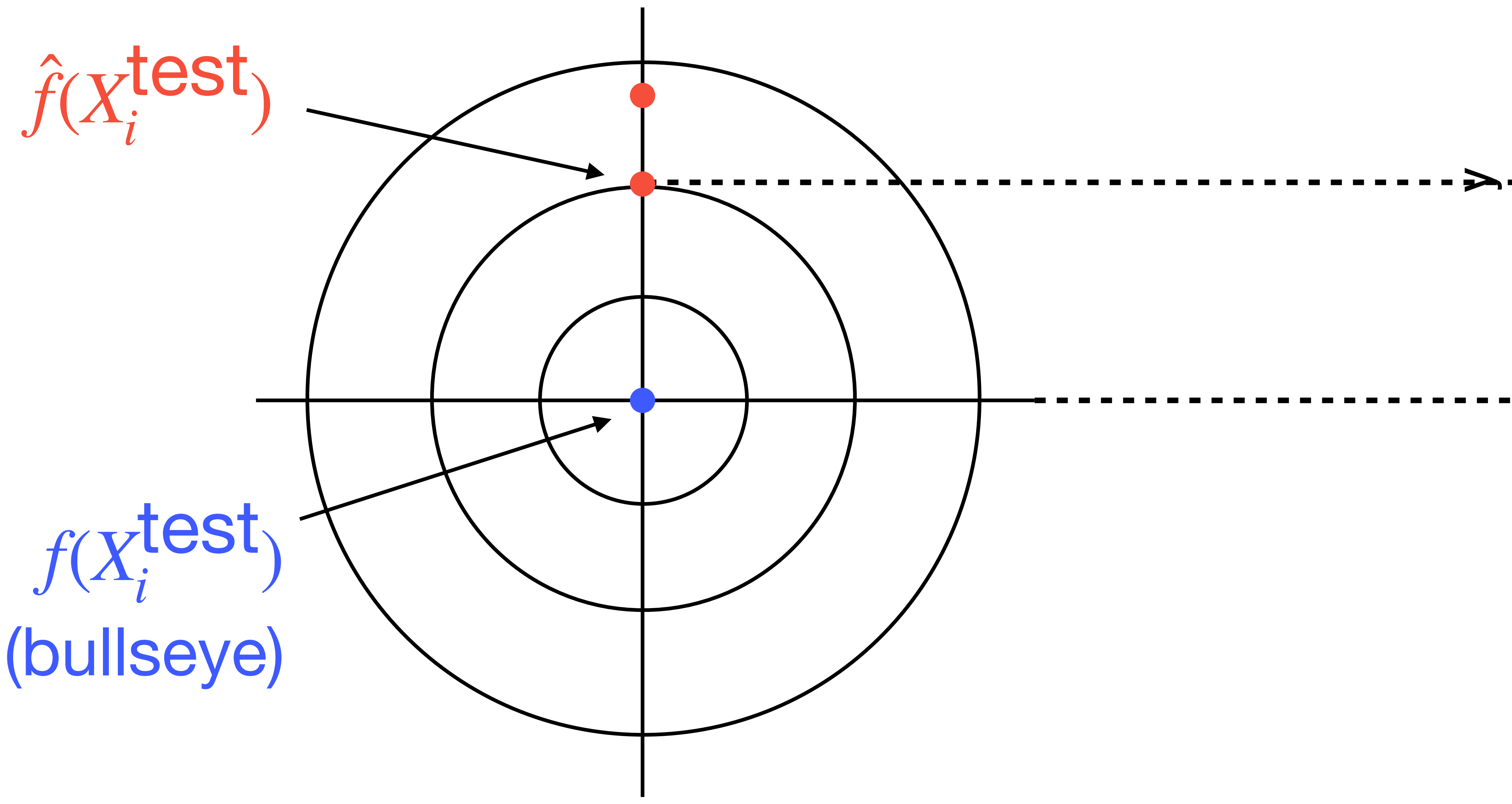
Estimation error = Bias² + Variance

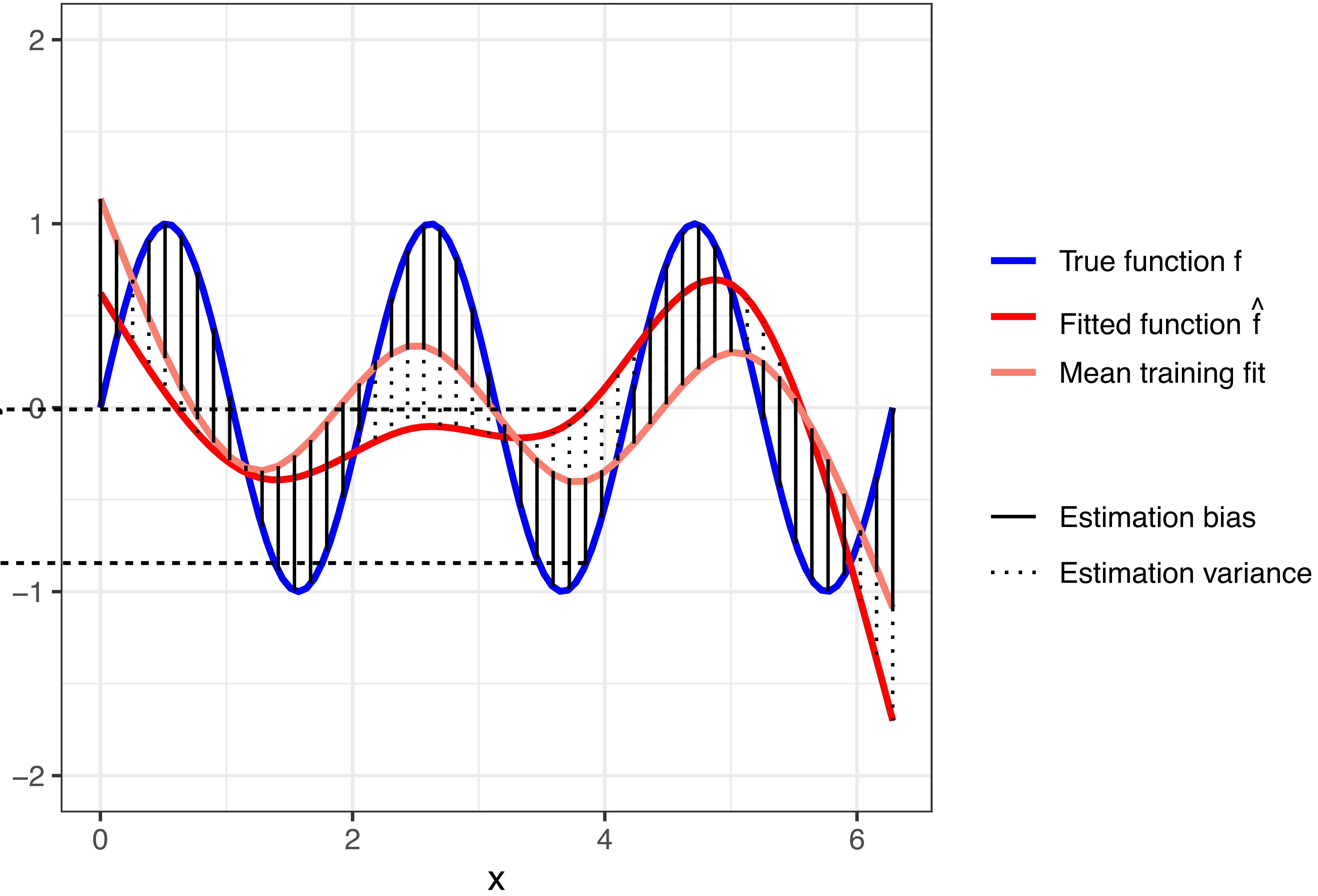
Consider a test point X_i^{test}. Each training data set leads to a prediction $\hat{f}(X_i^{\text{test}})$, which is like throwing a dart at $f(X_i^{\text{test}})$.



Estimation error = Bias² + Variance

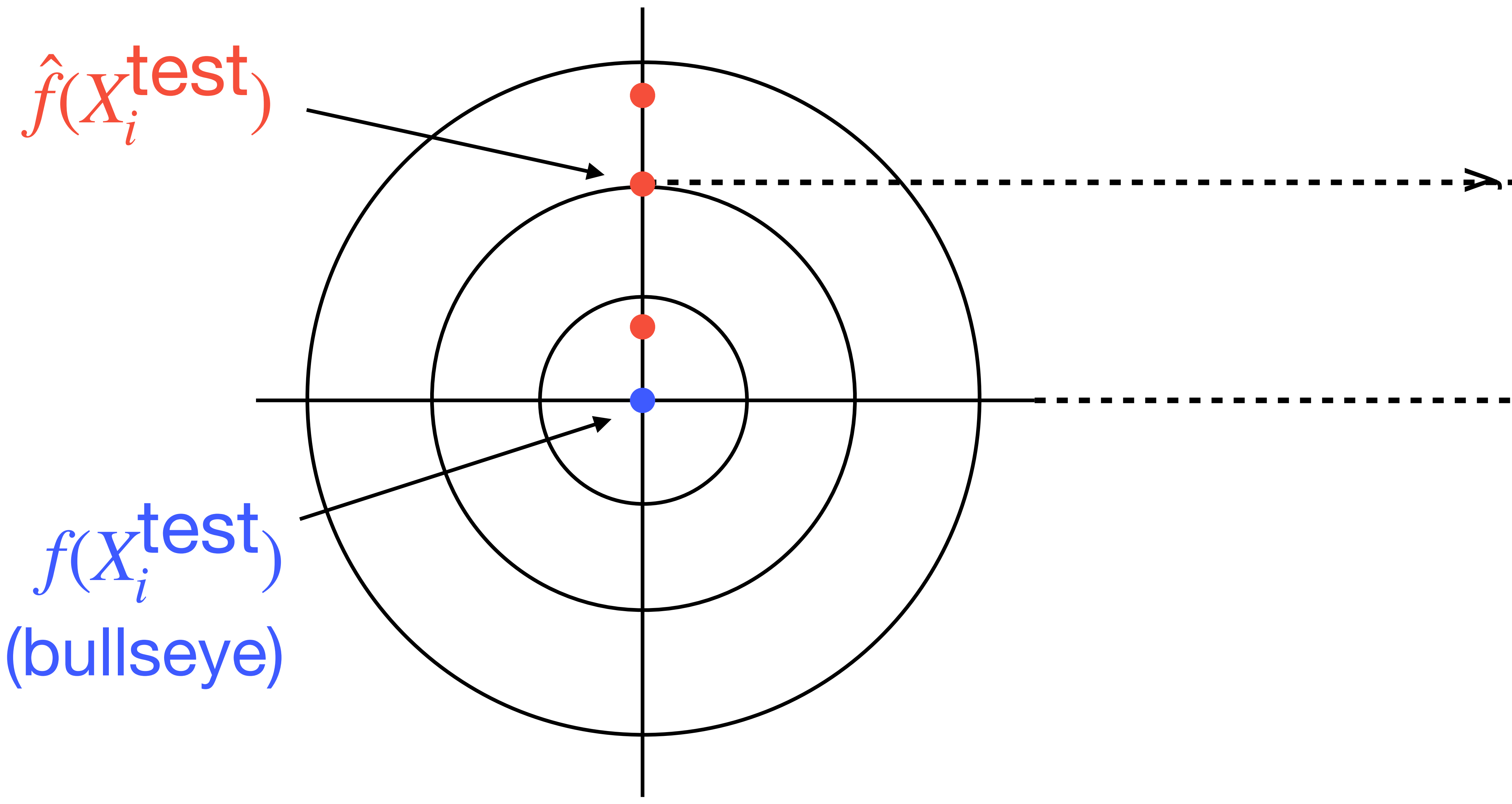
Consider a test point X_i^{test}. Each training data set leads to a prediction $\hat{f}(X_i^{\text{test}})$, which is like throwing a dart at $f(X_i^{\text{test}})$.

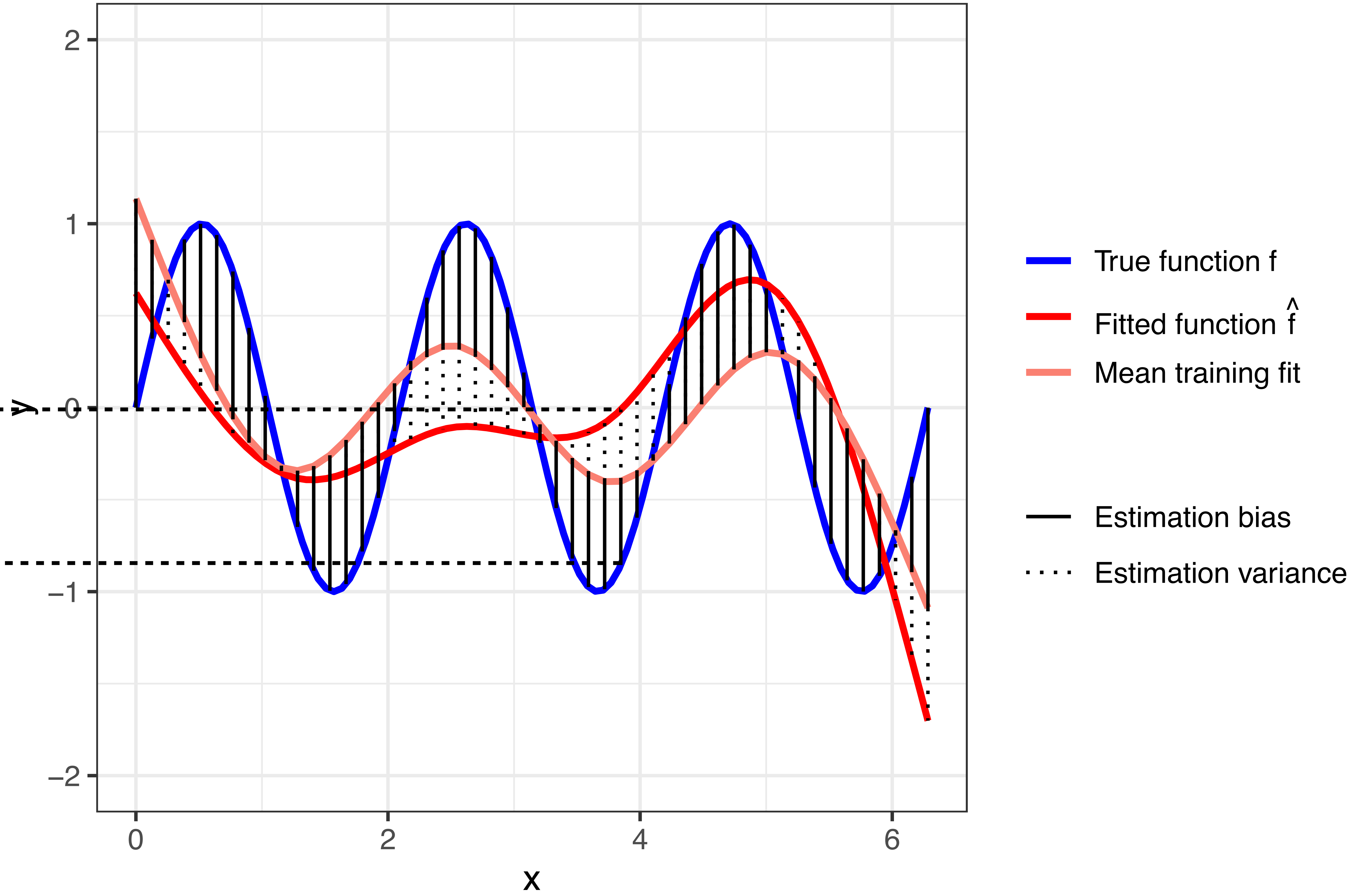




Estimation error = Bias^2 + Variance

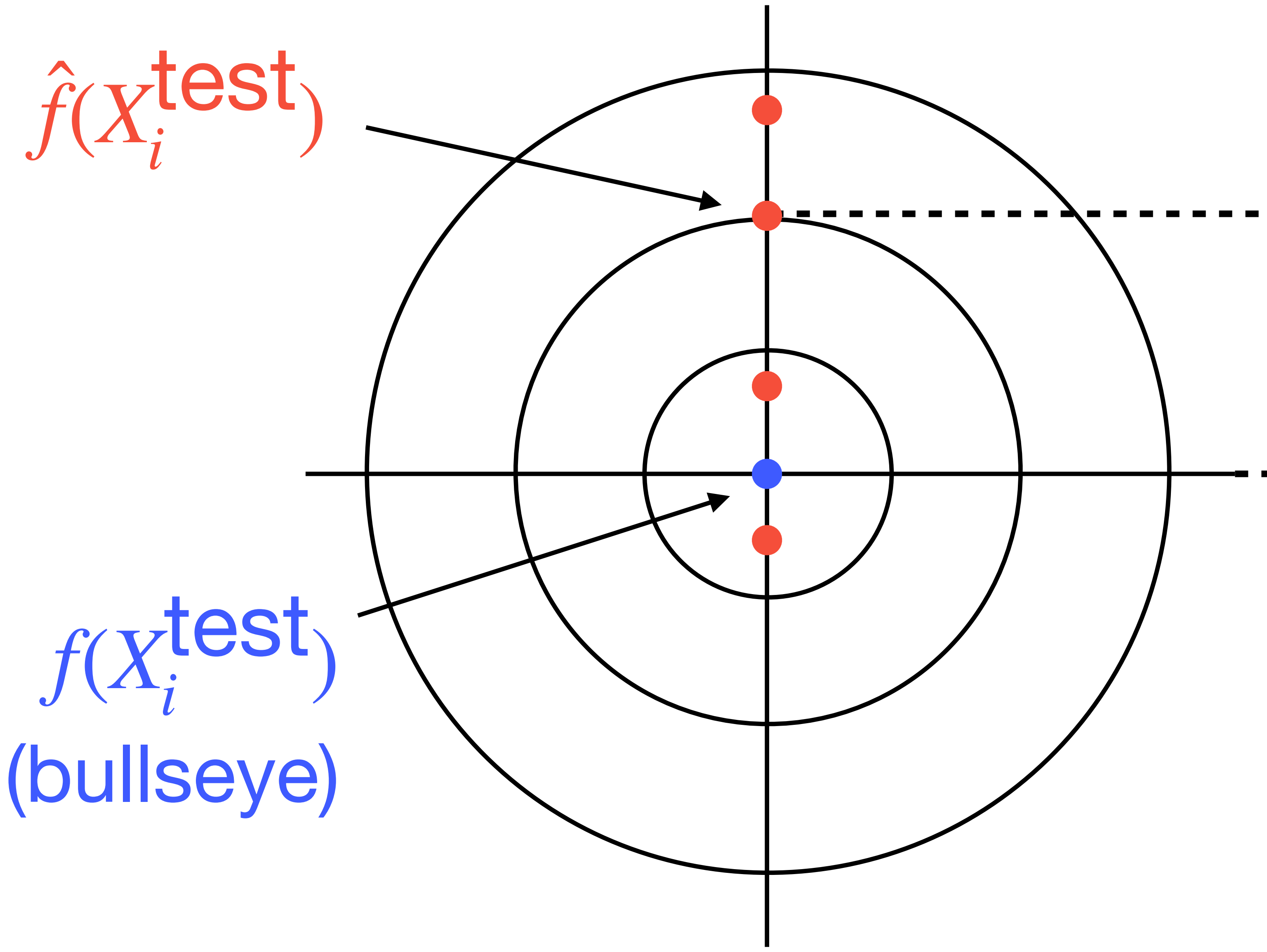
Consider a test point X_i^{test}. Each training data set leads to a prediction $\hat{f}(X_i^{\text{test}})$, which is like throwing a dart at $f(X_i^{\text{test}})$.

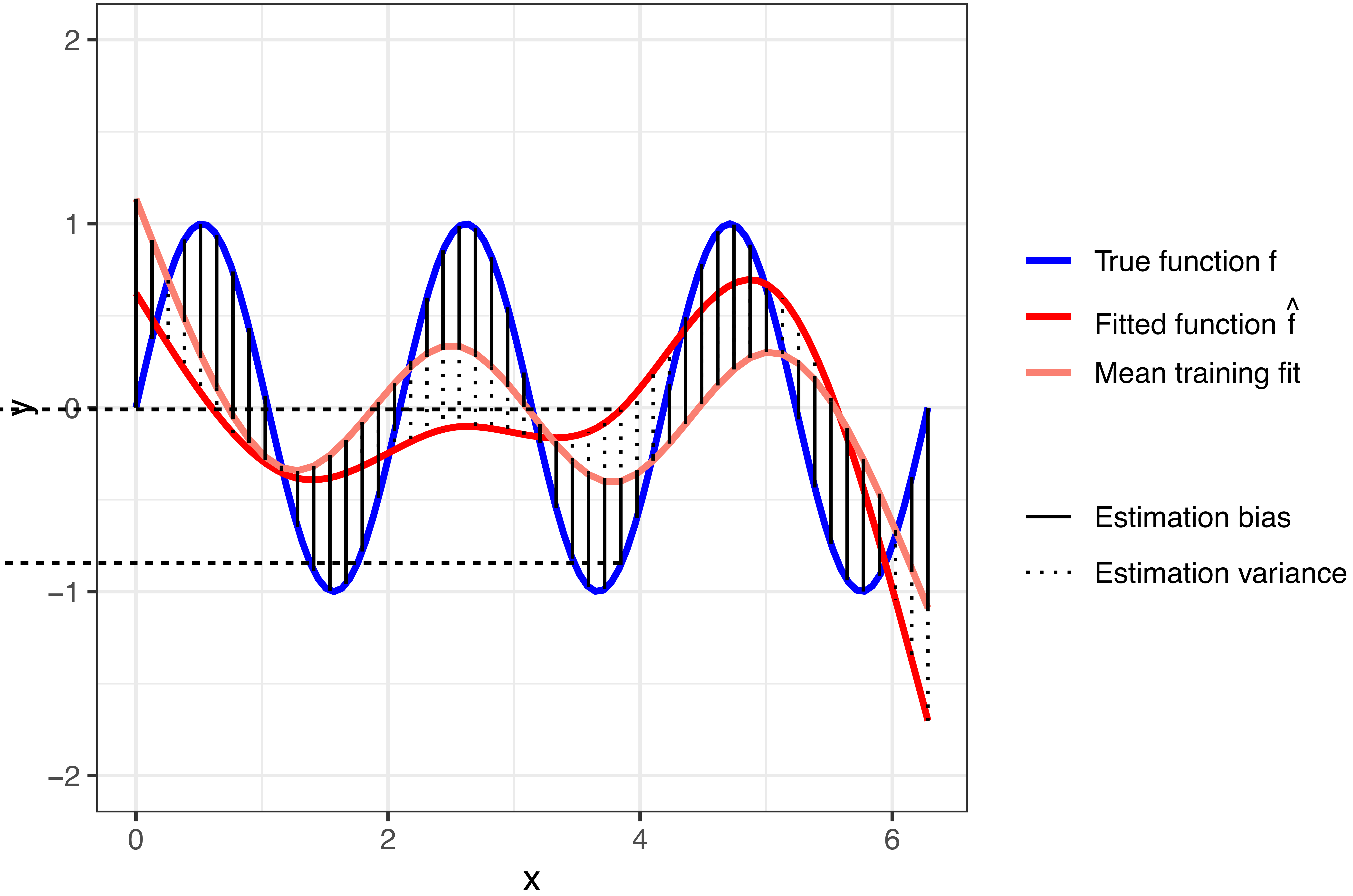




Estimation error = Bias² + Variance

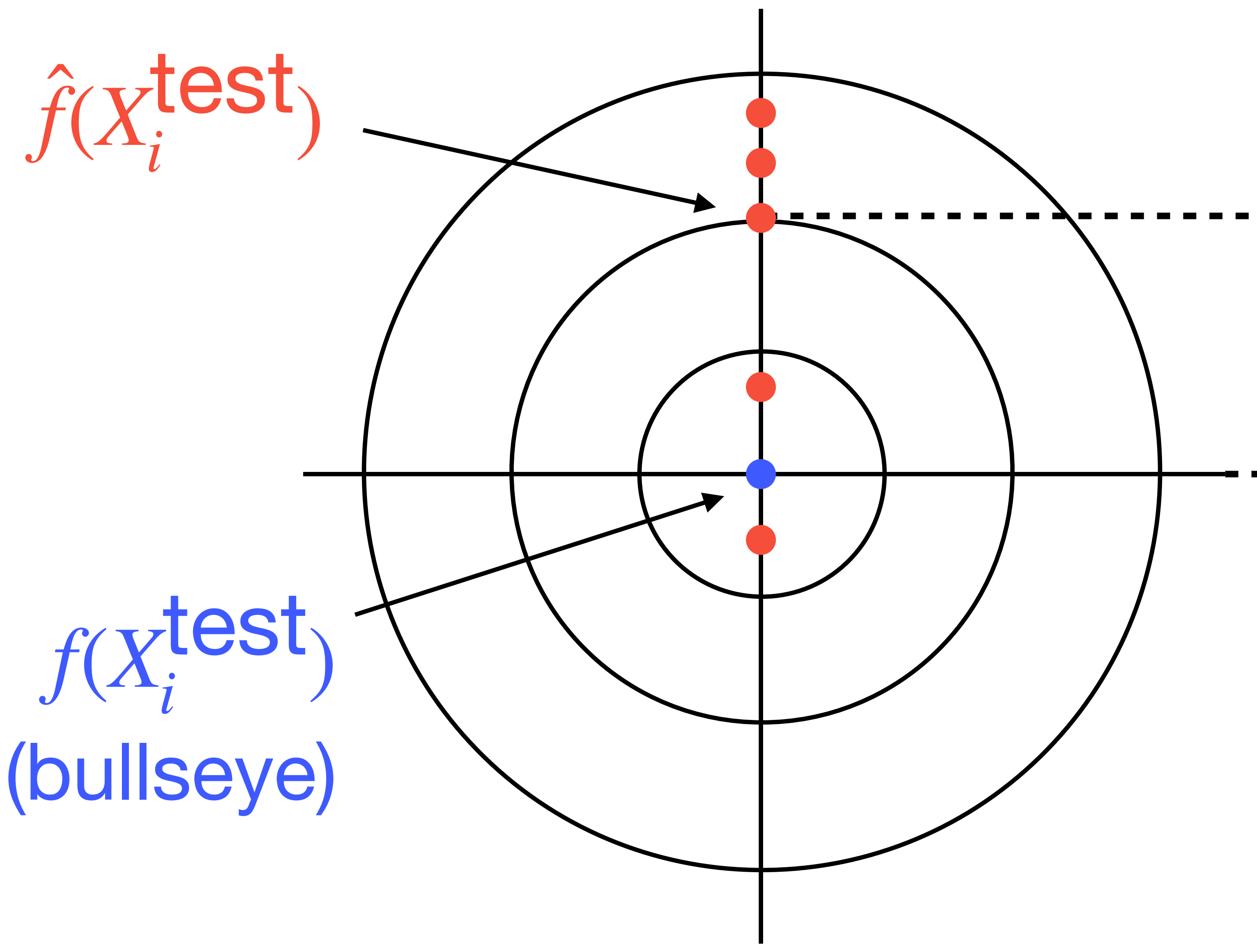
Consider a test point X_i^{test}. Each training data set leads to a prediction $\hat{f}(X_i^{\text{test}})$, which is like throwing a dart at $f(X_i^{\text{test}})$.

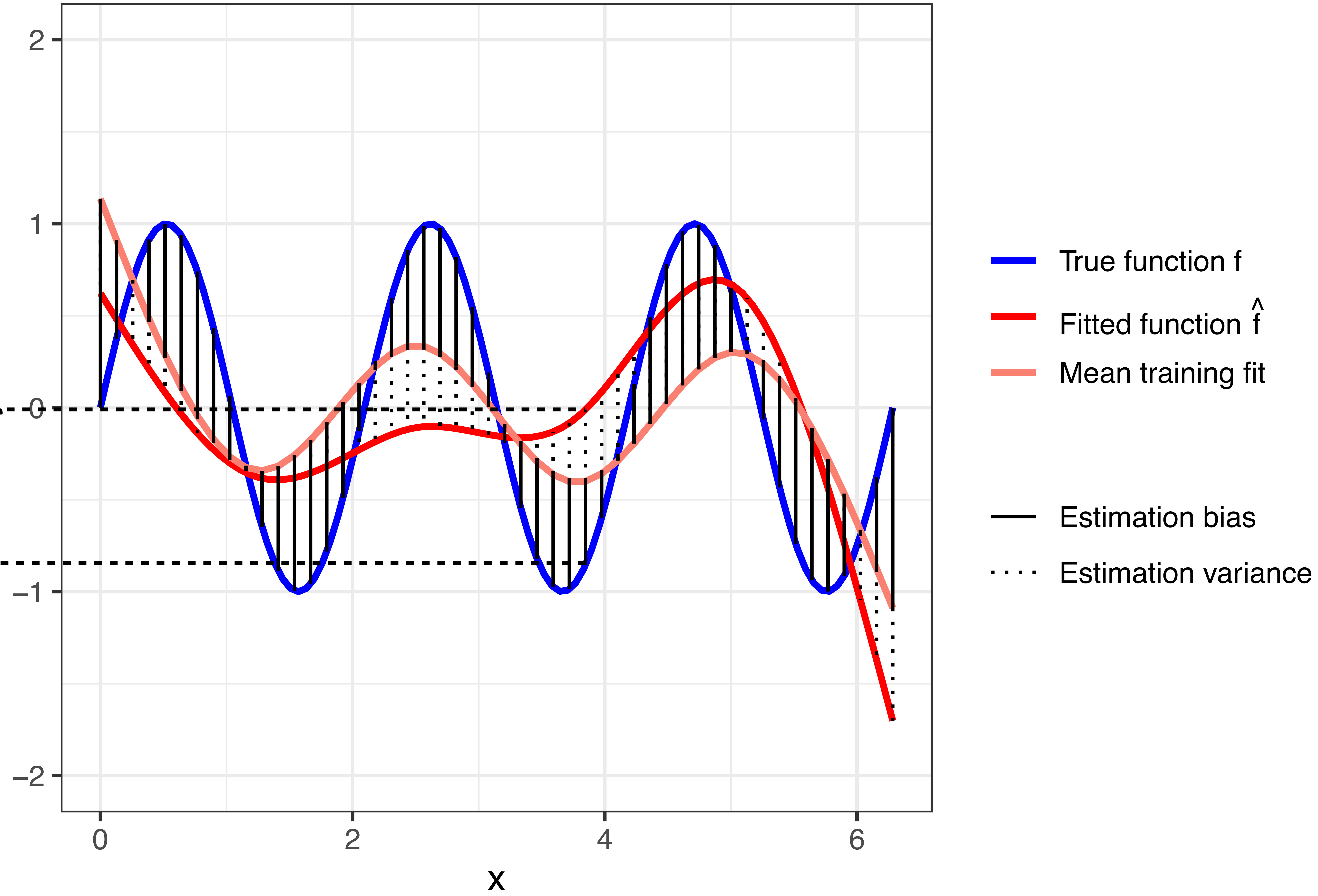




Estimation error = Bias^2 + Variance

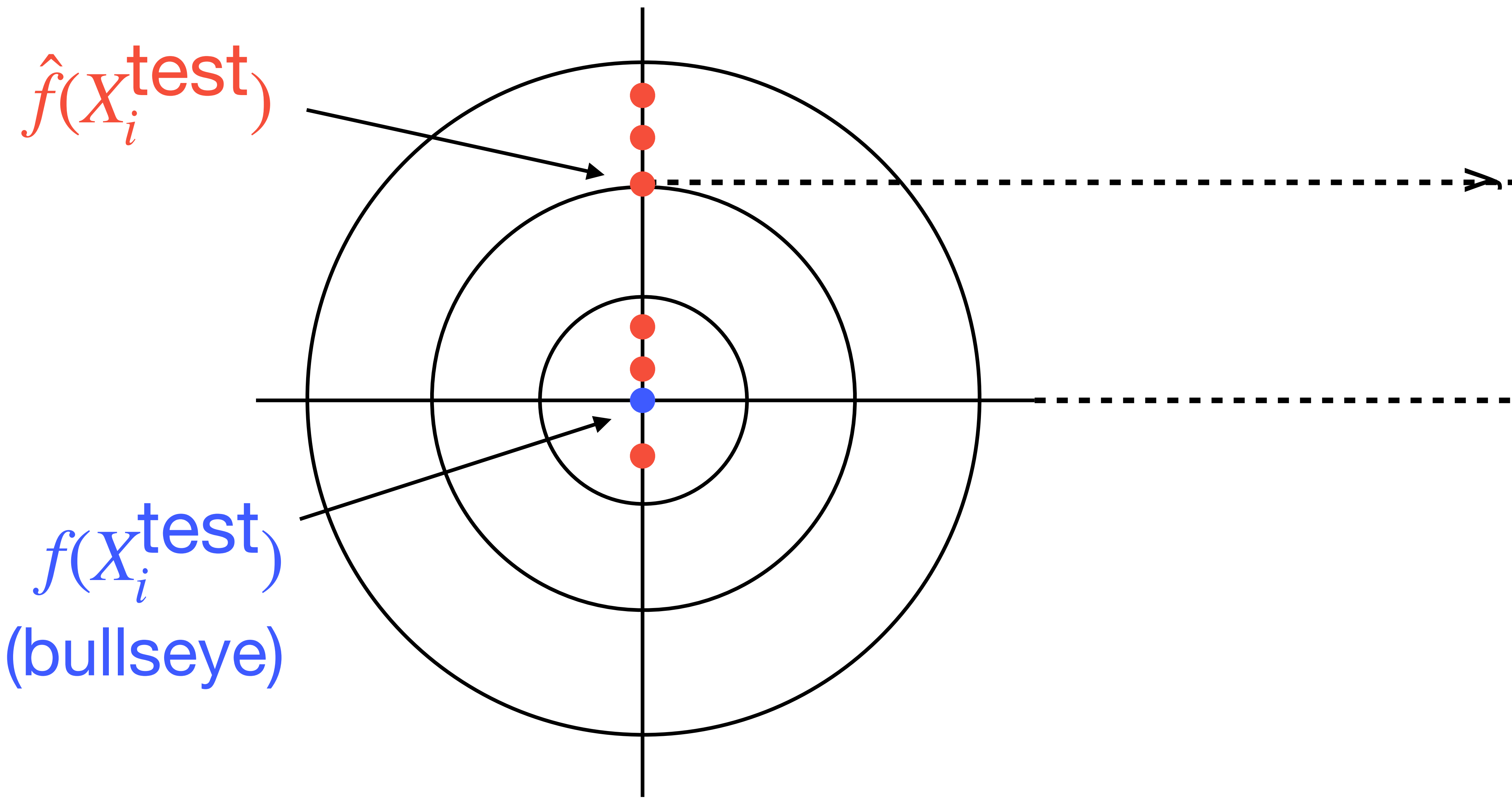
Consider a test point X_i^{test}. Each training data set leads to a prediction $\hat{f}(X_i^{\text{test}})$, which is like throwing a dart at $f(X_i^{\text{test}})$.

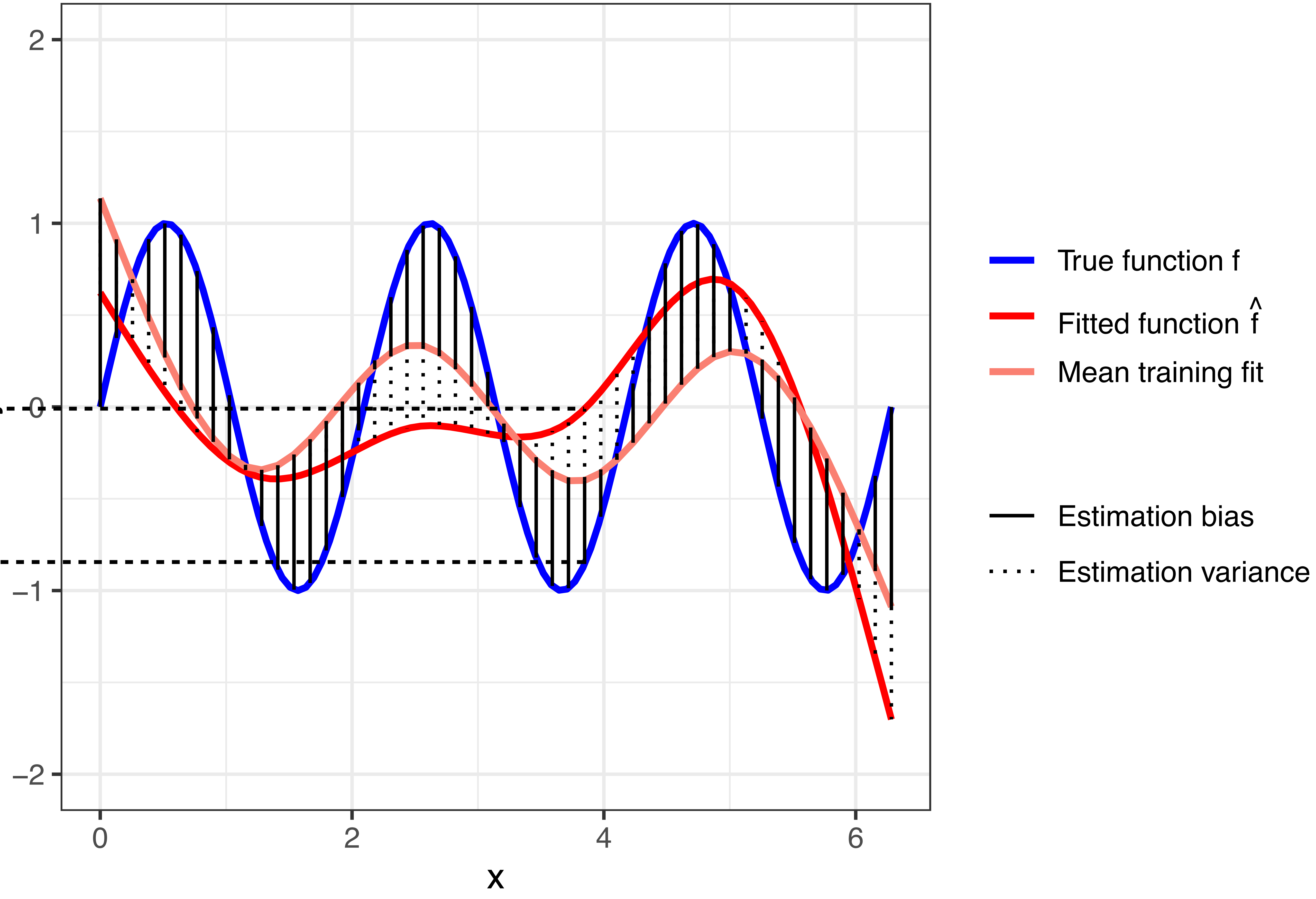




Estimation error = Bias² + Variance

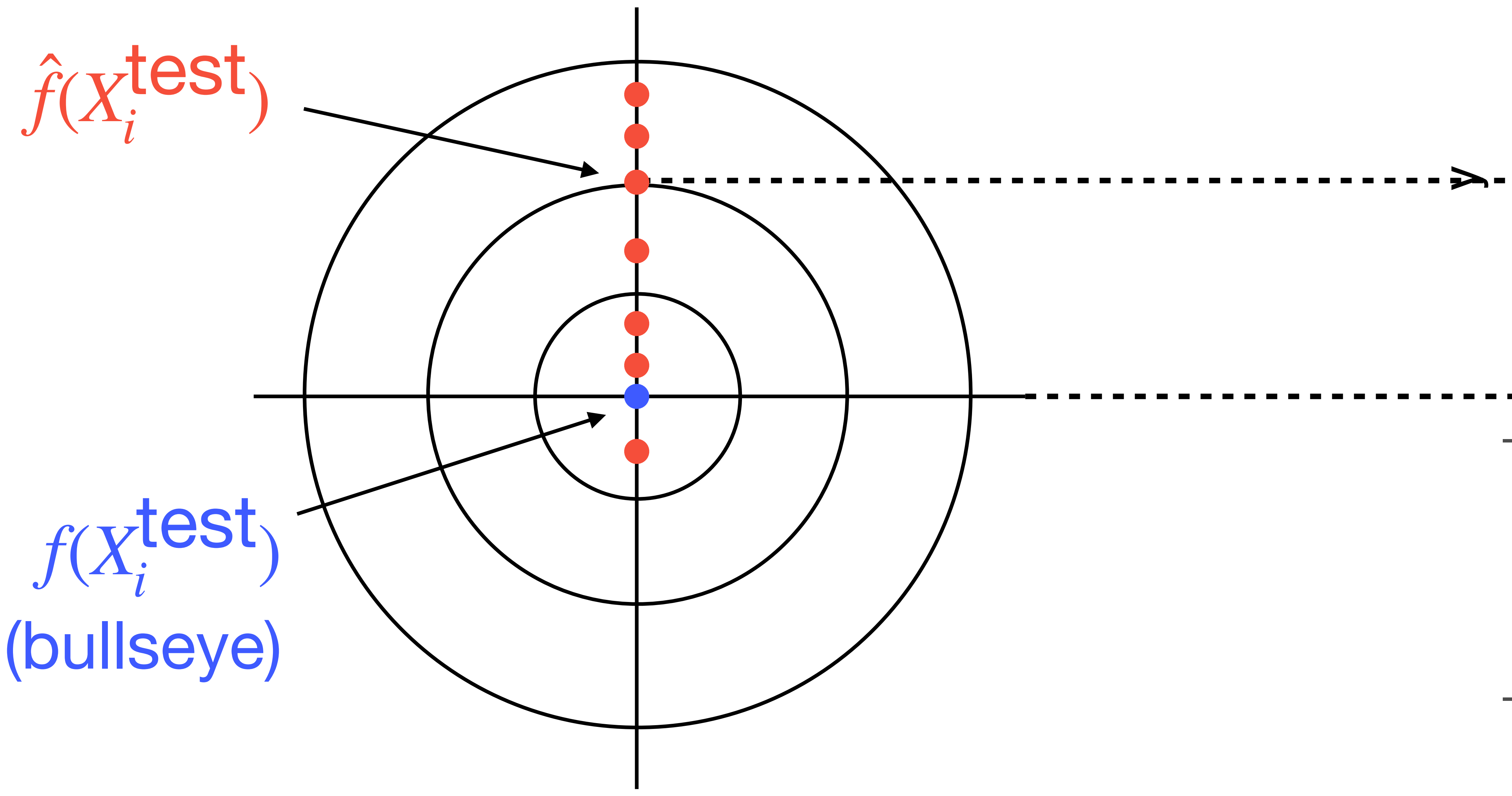
Consider a test point X_i^{test}. Each training data set leads to a prediction $\hat{f}(X_i^{\text{test}})$, which is like throwing a dart at $f(X_i^{\text{test}})$.

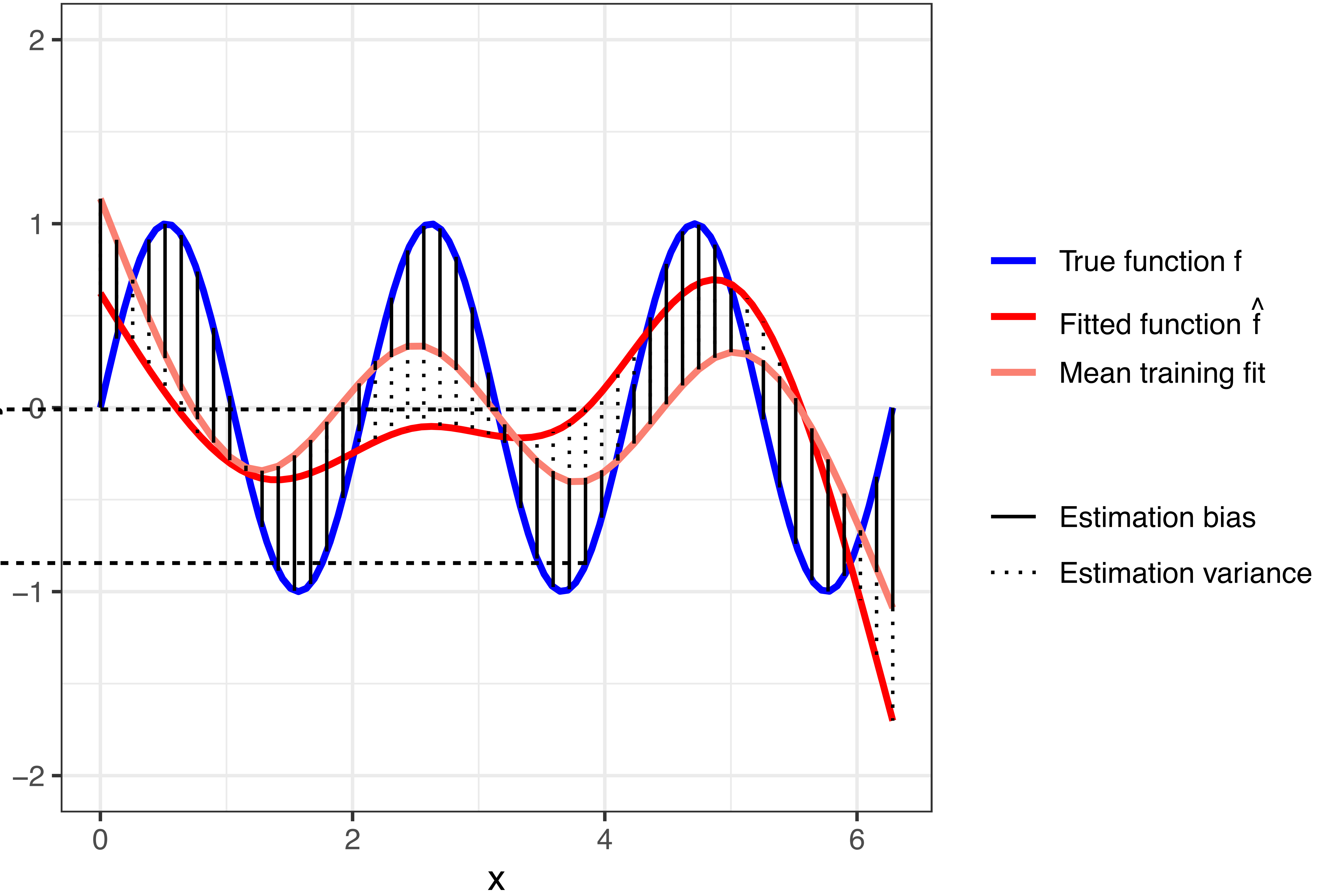




Estimation error = Bias² + Variance

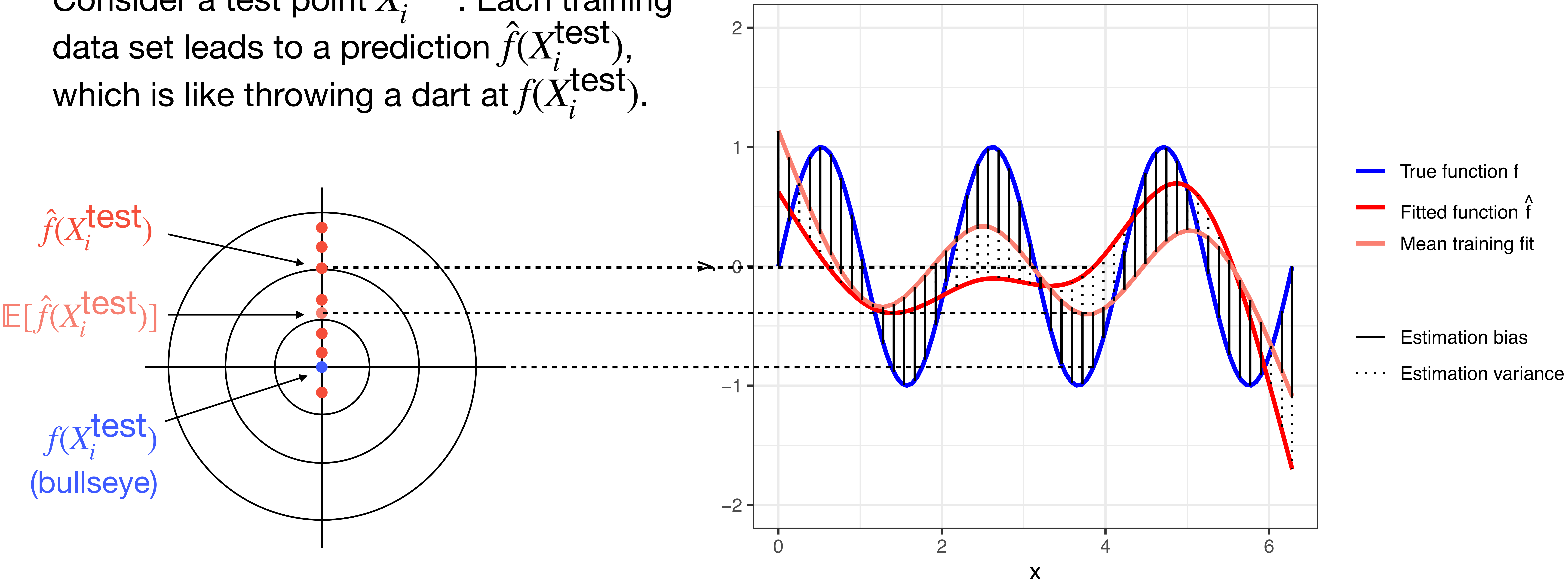
Consider a test point X_i^{test}. Each training data set leads to a prediction $\hat{f}(X_i^{\text{test}})$, which is like throwing a dart at $f(X_i^{\text{test}})$.





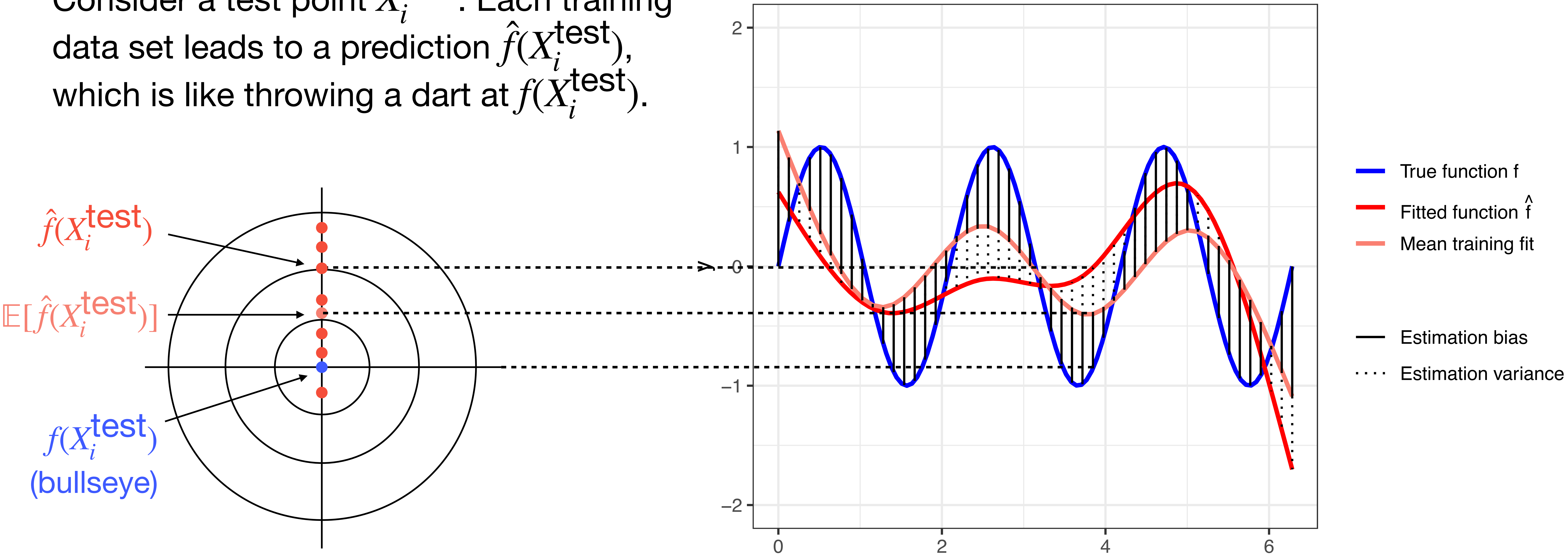
Estimation error = Bias^2 + Variance

Consider a test point X_i^{test}. Each training data set leads to a prediction $\hat{f}(X_i^{\text{test}})$, which is like throwing a dart at $f(X_i^{\text{test}})$.



Estimation error = Bias^2 + Variance

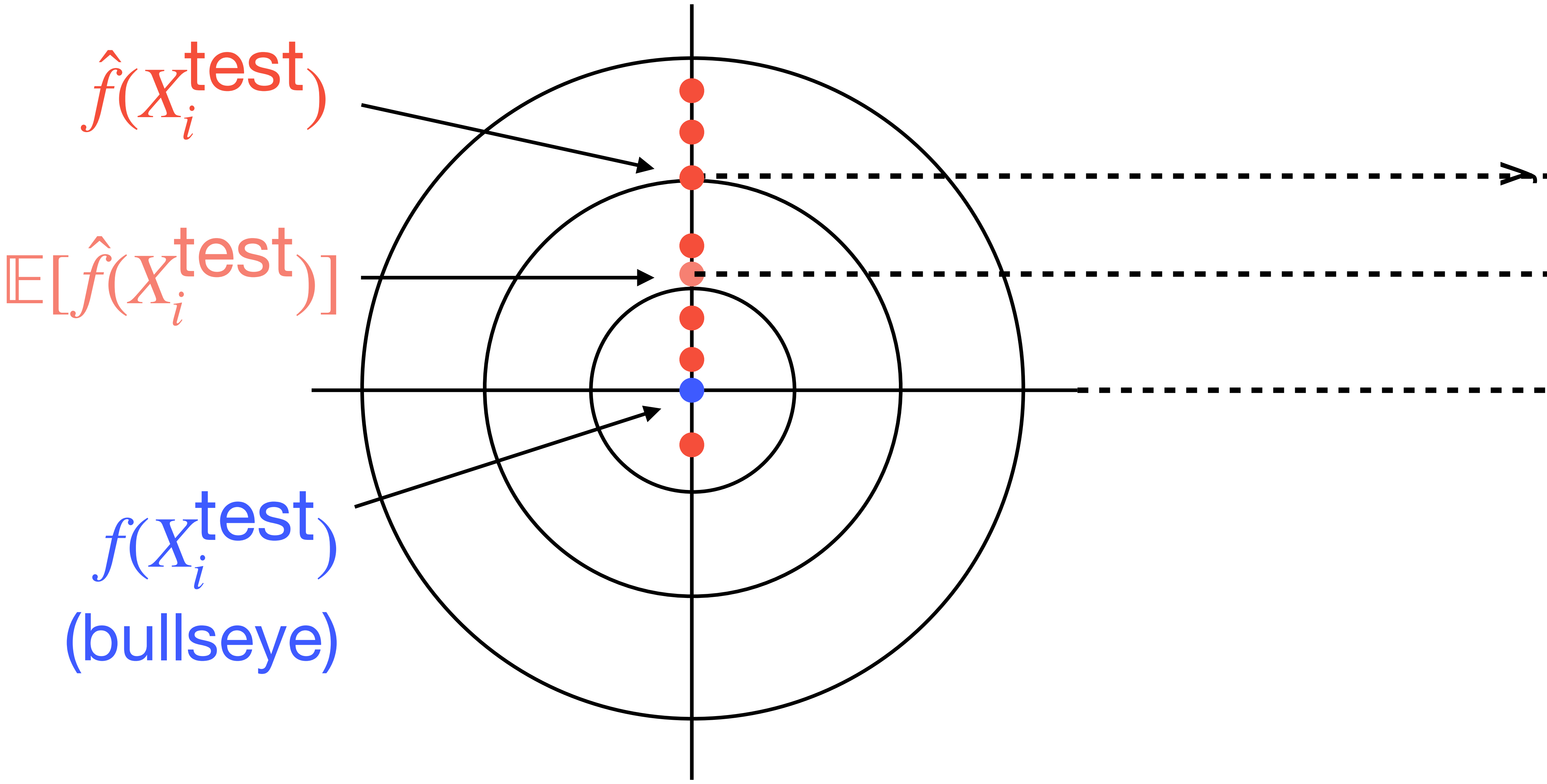
Consider a test point X_i^{test}. Each training data set leads to a prediction $\hat{f}(X_i^{\text{test}})$, which is like throwing a dart at $f(X_i^{\text{test}})$.

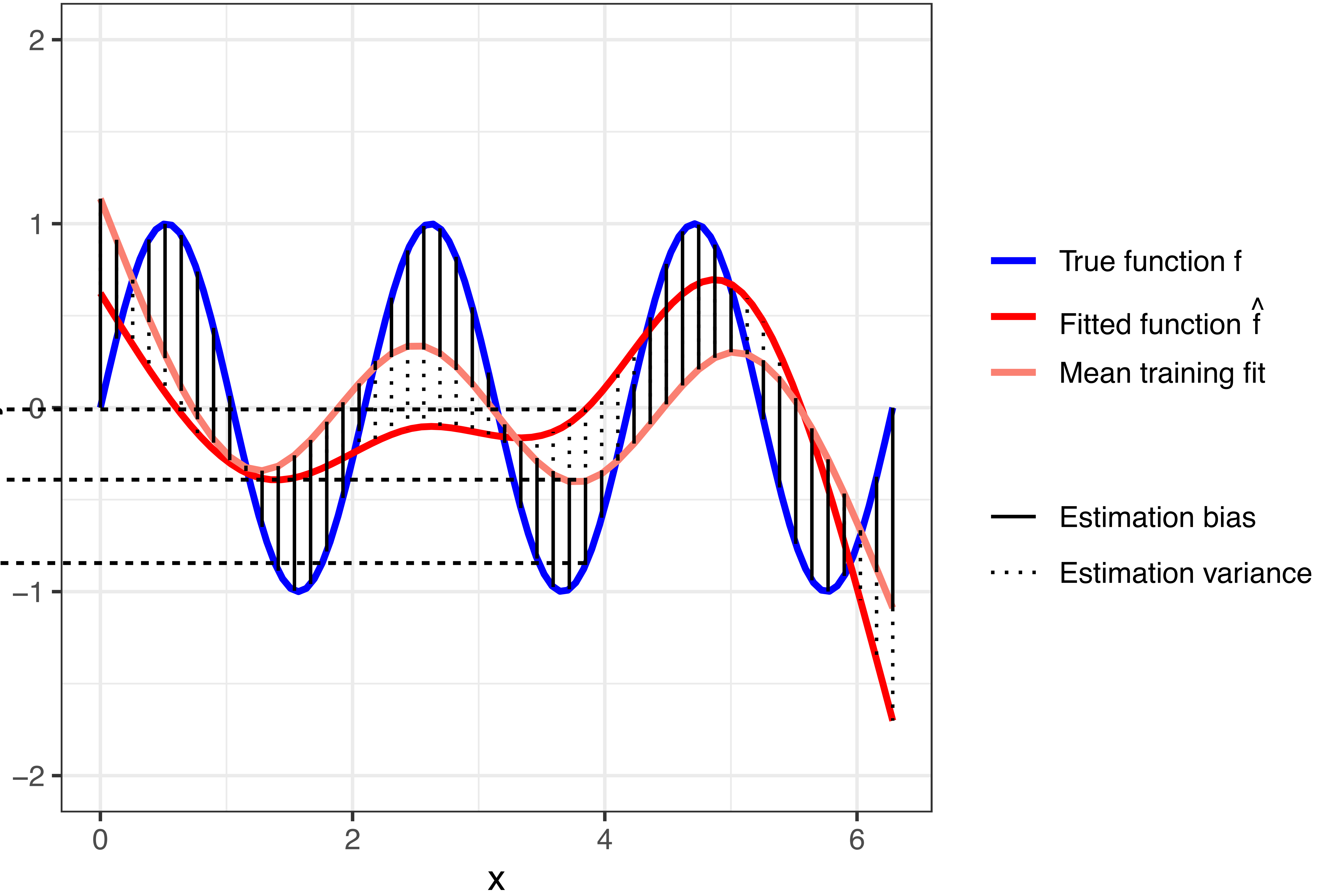


Bias: Aim systematically off in one direction.

Estimation error = Bias² + Variance

Consider a test point X_i^{test}. Each training data set leads to a prediction $\hat{f}(X_i^{\text{test}})$, which is like throwing a dart at $f(X_i^{\text{test}})$.



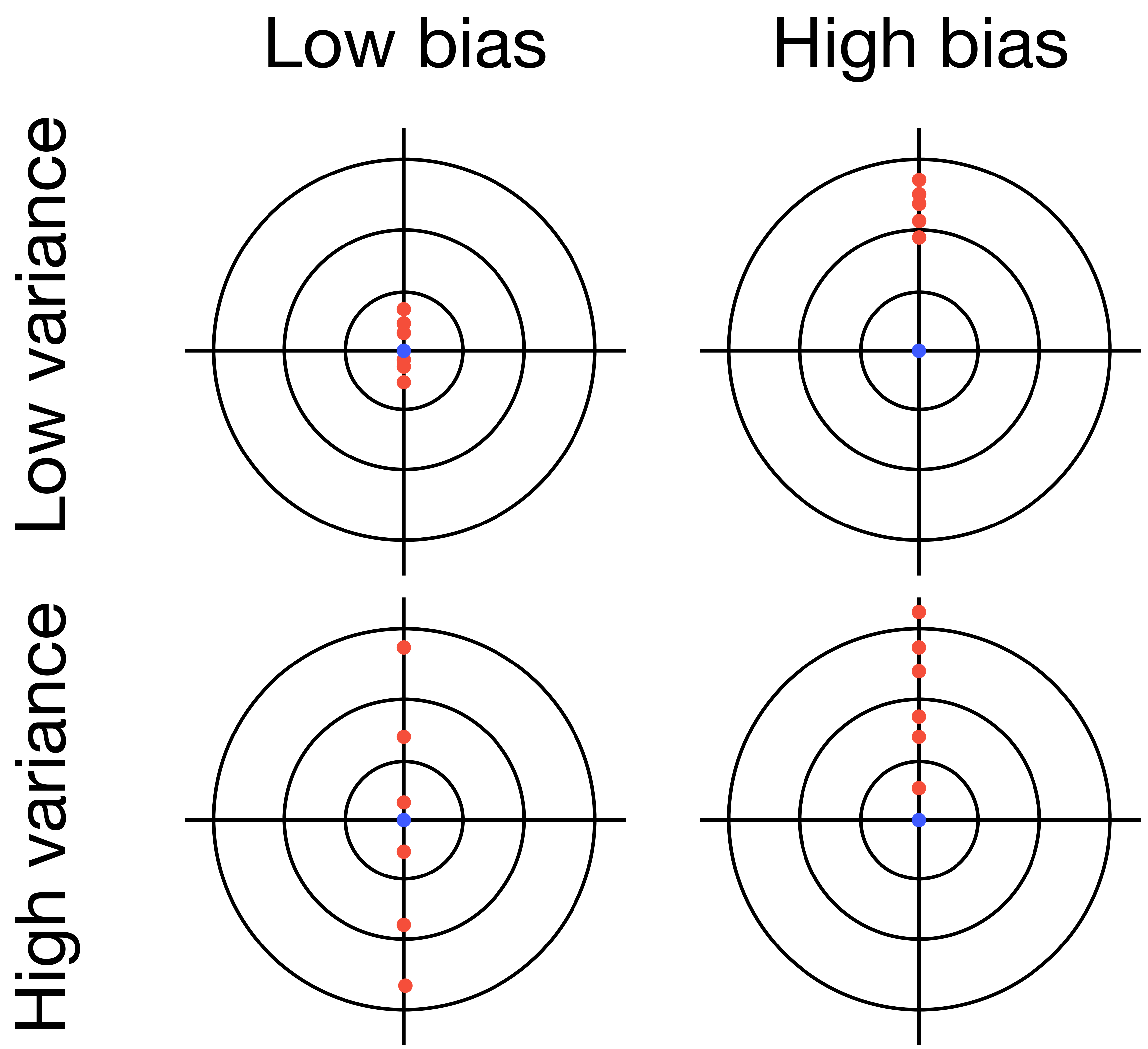


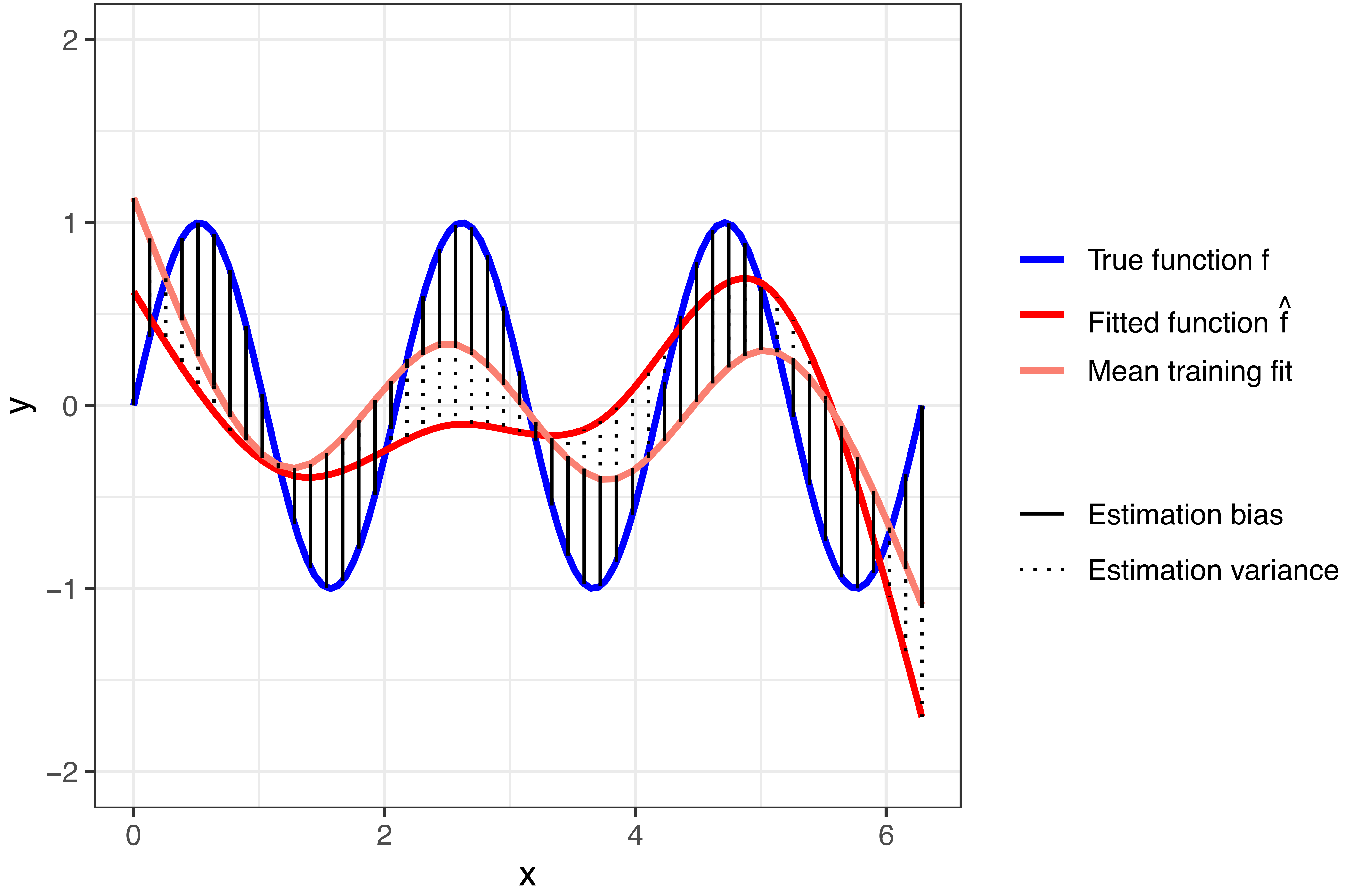
Bias: Aim systematically off in one direction.

Variance: Aim wobbling between throws.

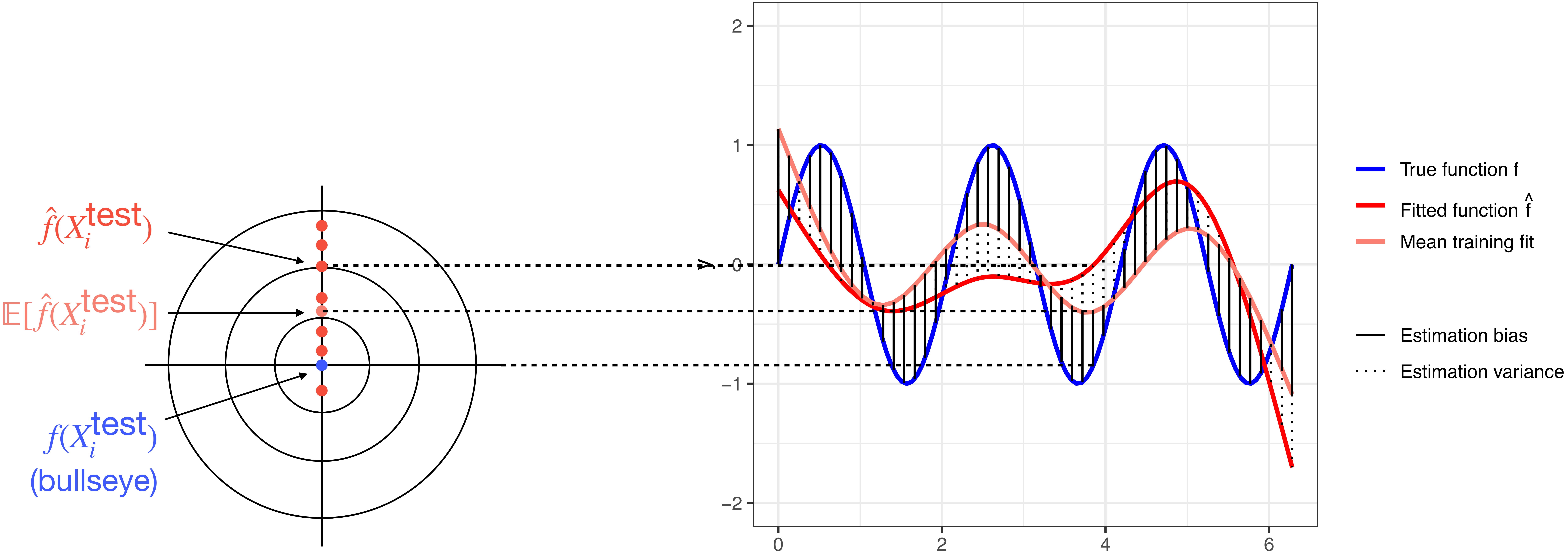
Estimation error = Bias^2 + Variance

Consider a test point X_i^{test}. Each training data set leads to a prediction $\hat{f}(X_i^{\text{test}})$, which is like throwing a dart at $f(X_i^{\text{test}})$.



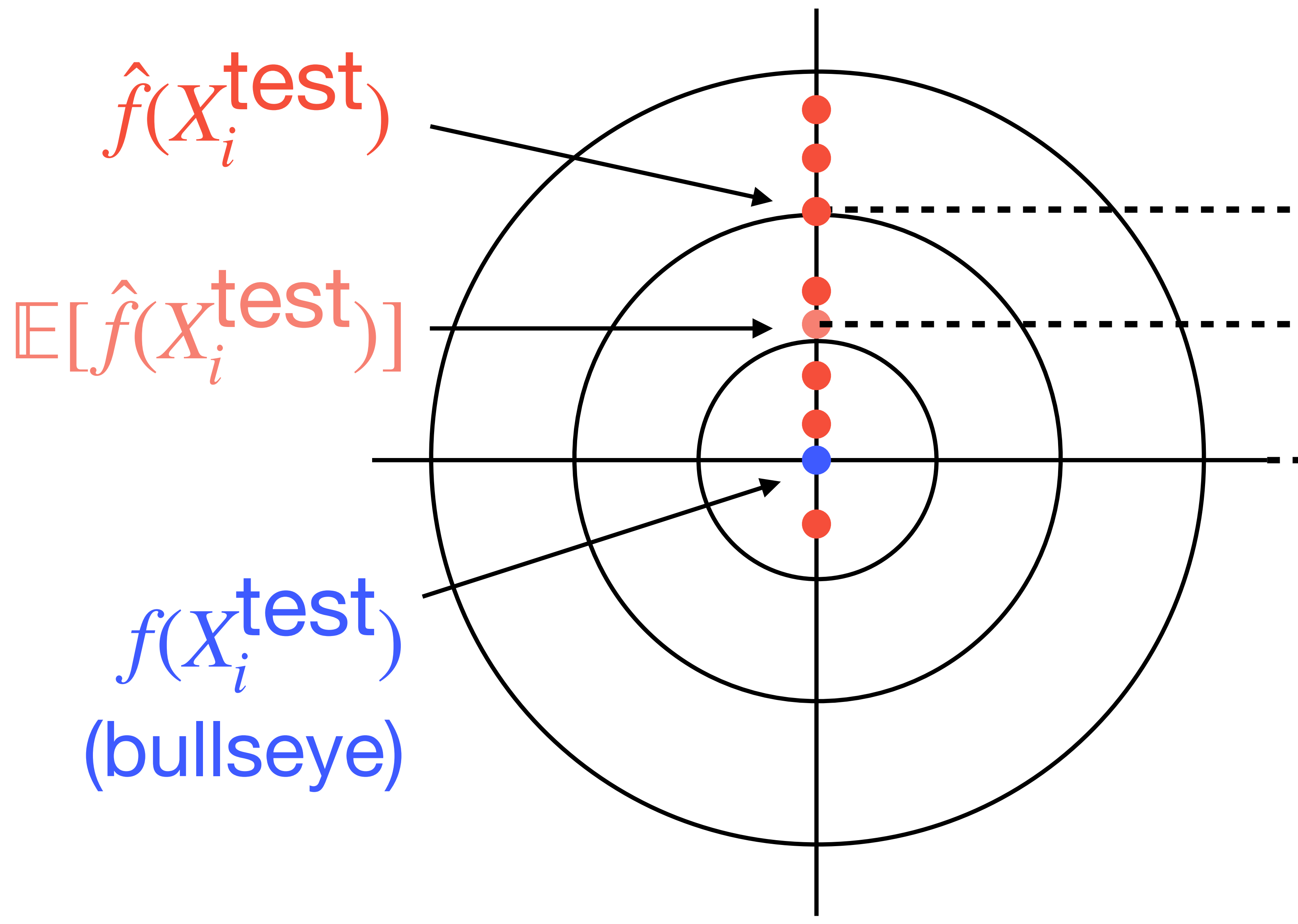


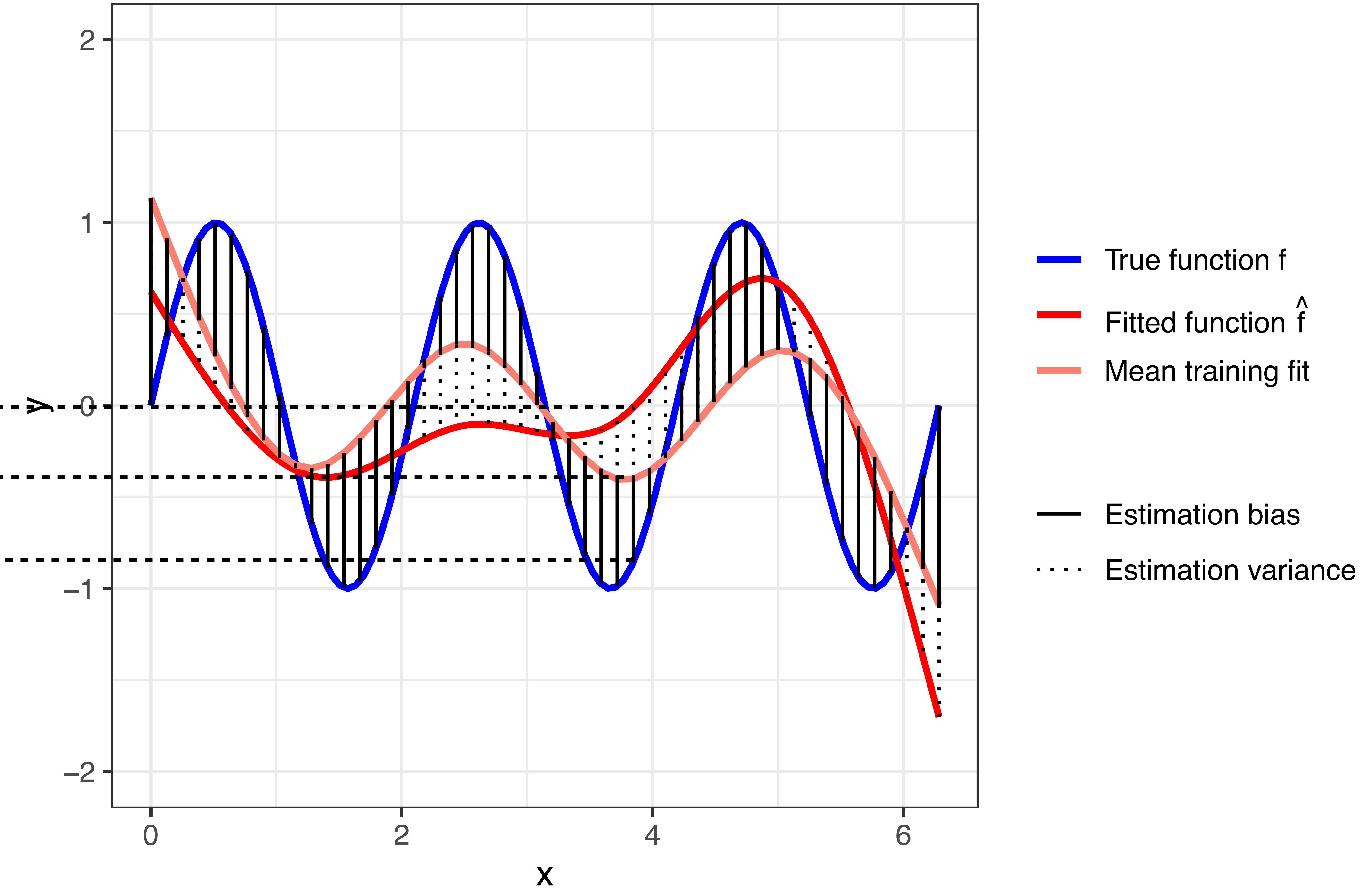
Understanding bias



Understanding bias

$\text{Bias}_i = \mathbb{E}[\hat{f}(X_i^{\text{test}})] - f(X_i^{\text{test}})$, distance from average fitted model to true trend.

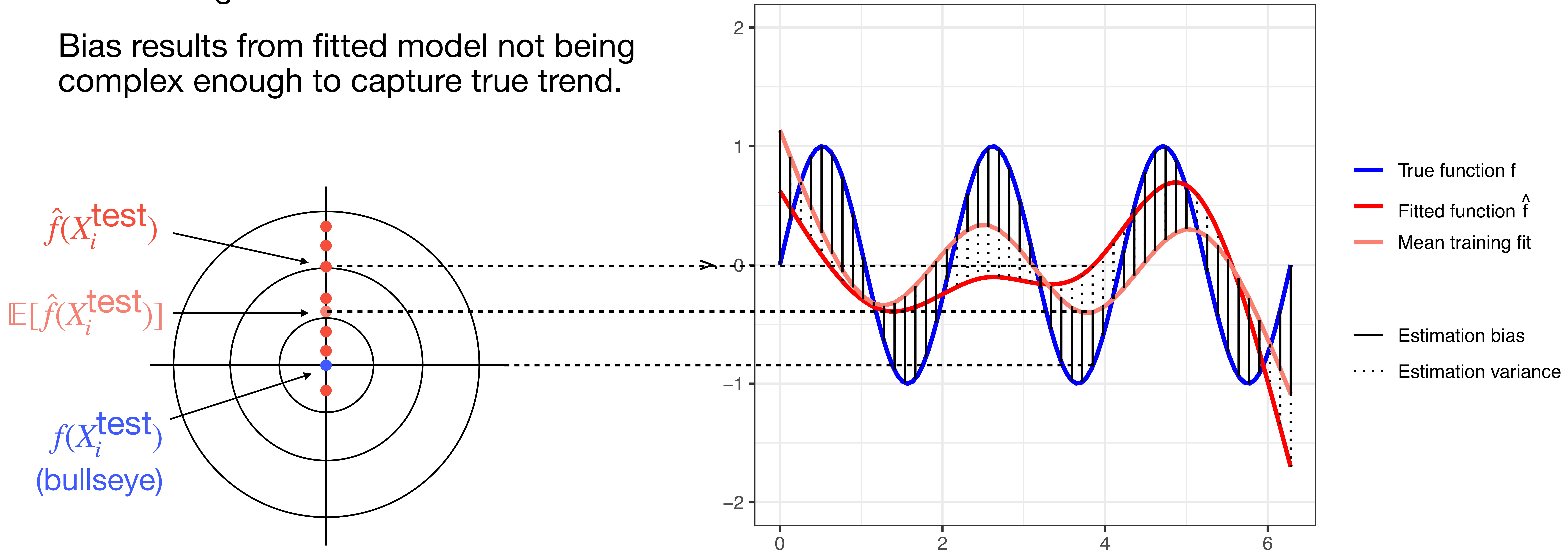




Understanding bias

$\text{Bias}_i = \mathbb{E}[\hat{f}(X_i^{\text{test}})] - f(X_i^{\text{test}})$, distance from average fitted model to true trend.

Bias results from fitted model not being complex enough to capture true trend.

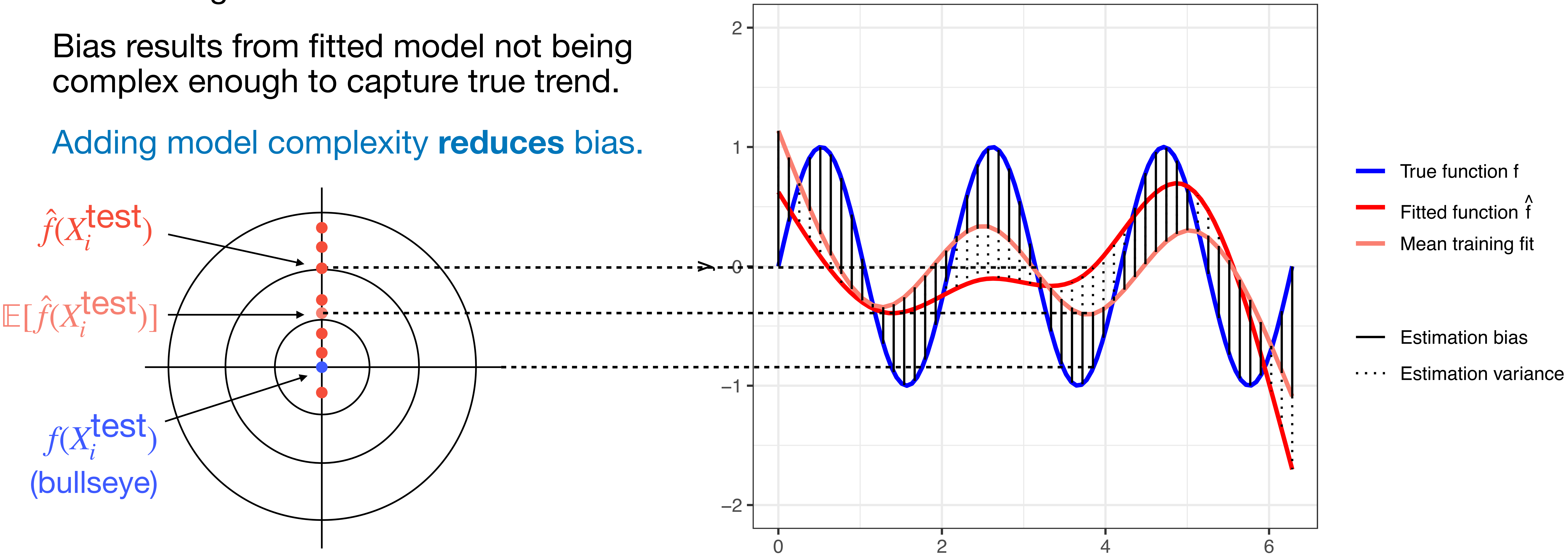


Understanding bias

$\text{Bias}_i = \mathbb{E}[\hat{f}(X_i^{\text{test}})] - f(X_i^{\text{test}})$, distance from average fitted model to true trend.

Bias results from fitted model not being complex enough to capture true trend.

Adding model complexity **reduces** bias.

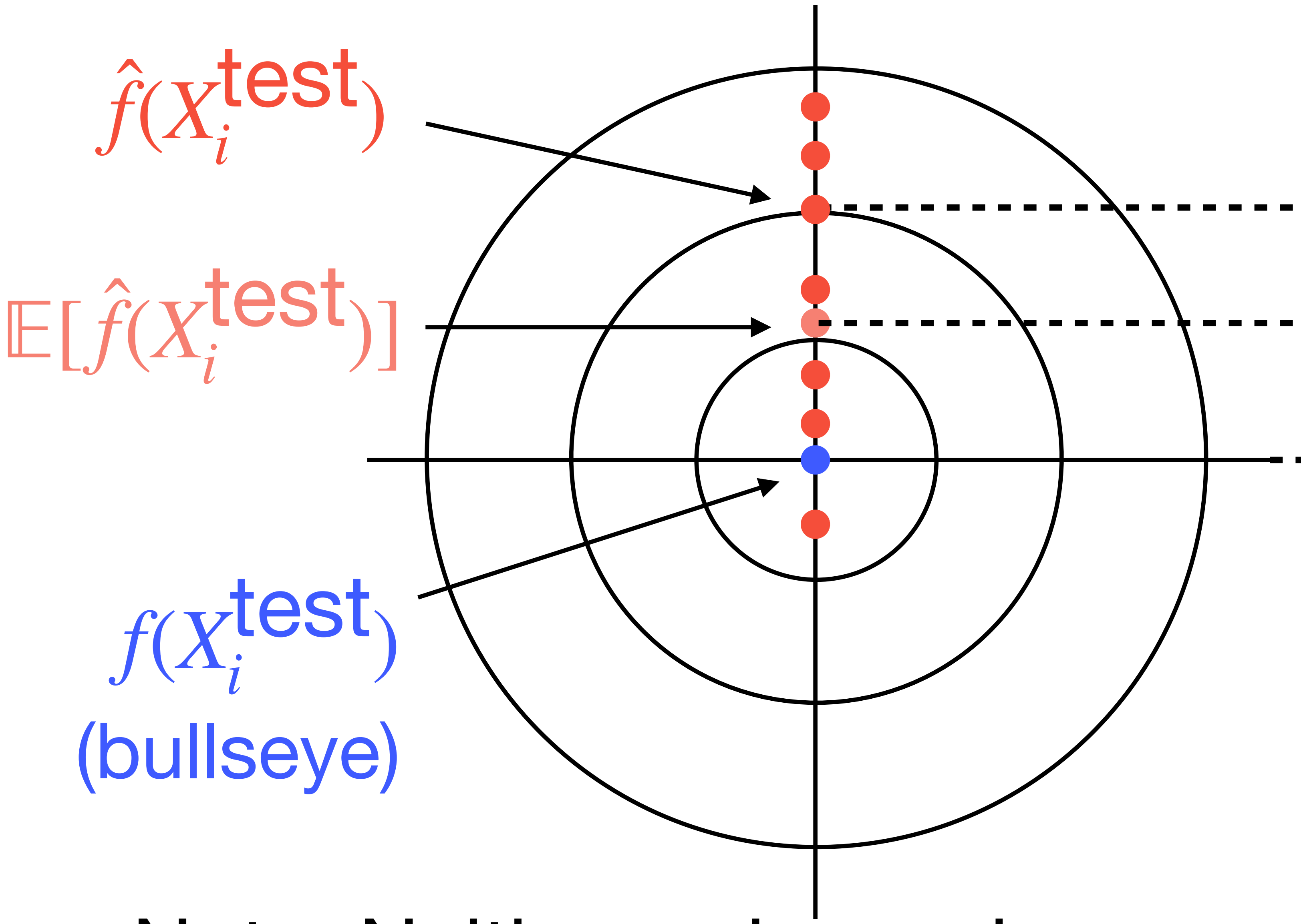


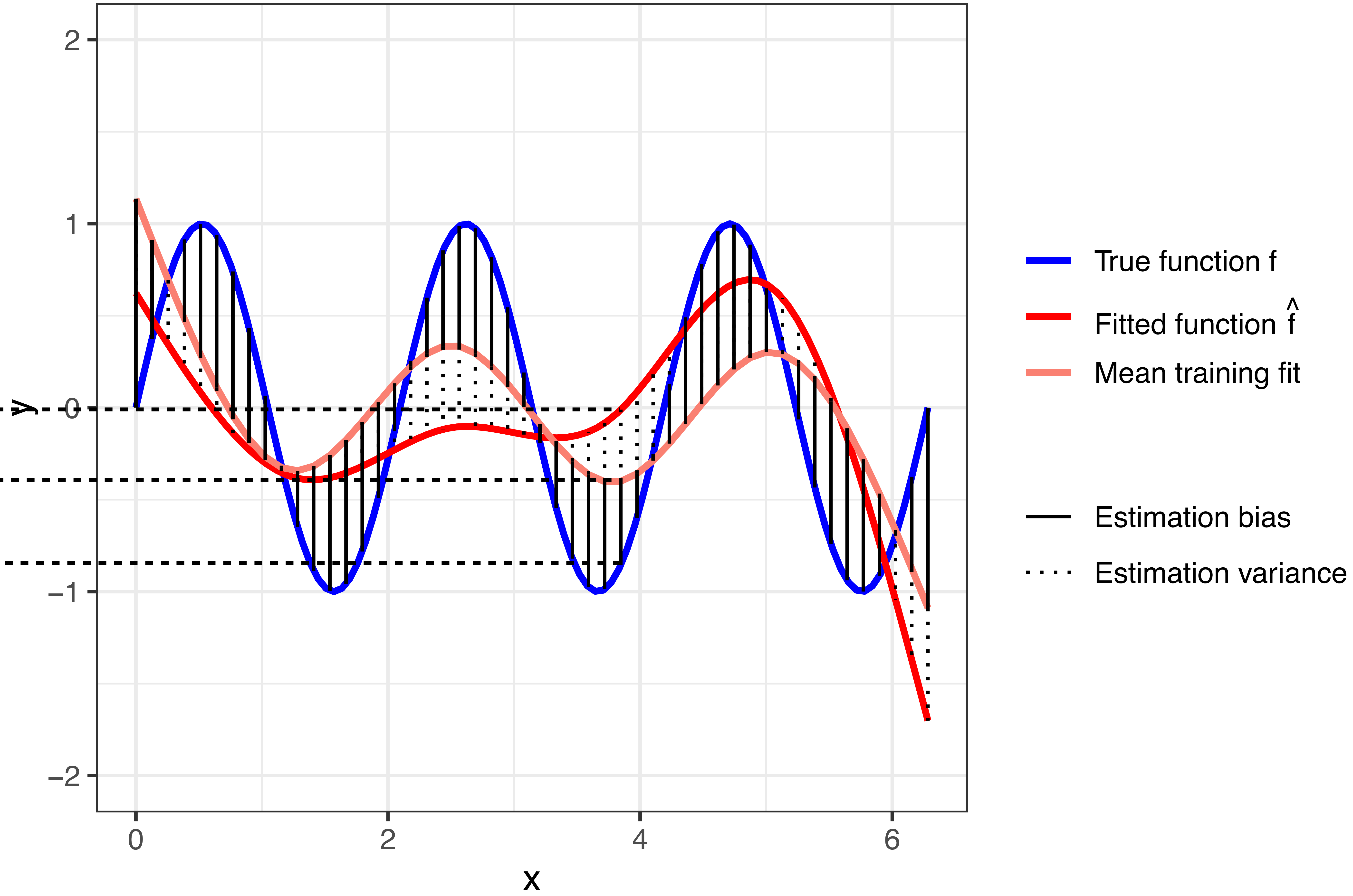
Understanding bias

$\text{Bias}_i = \mathbb{E}[\hat{f}(X_i^{\text{test}})] - f(X_i^{\text{test}})$, distance from average fitted model to true trend.

Bias results from fitted model not being complex enough to capture true trend.

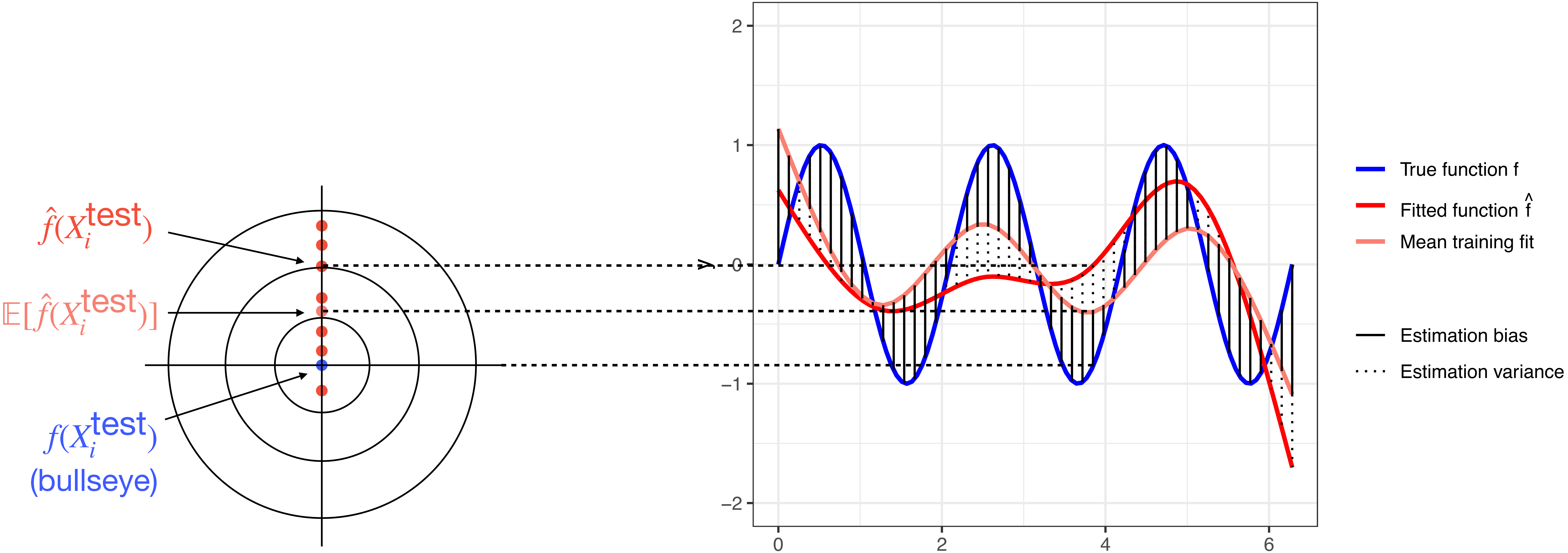
Adding model complexity **reduces** bias.





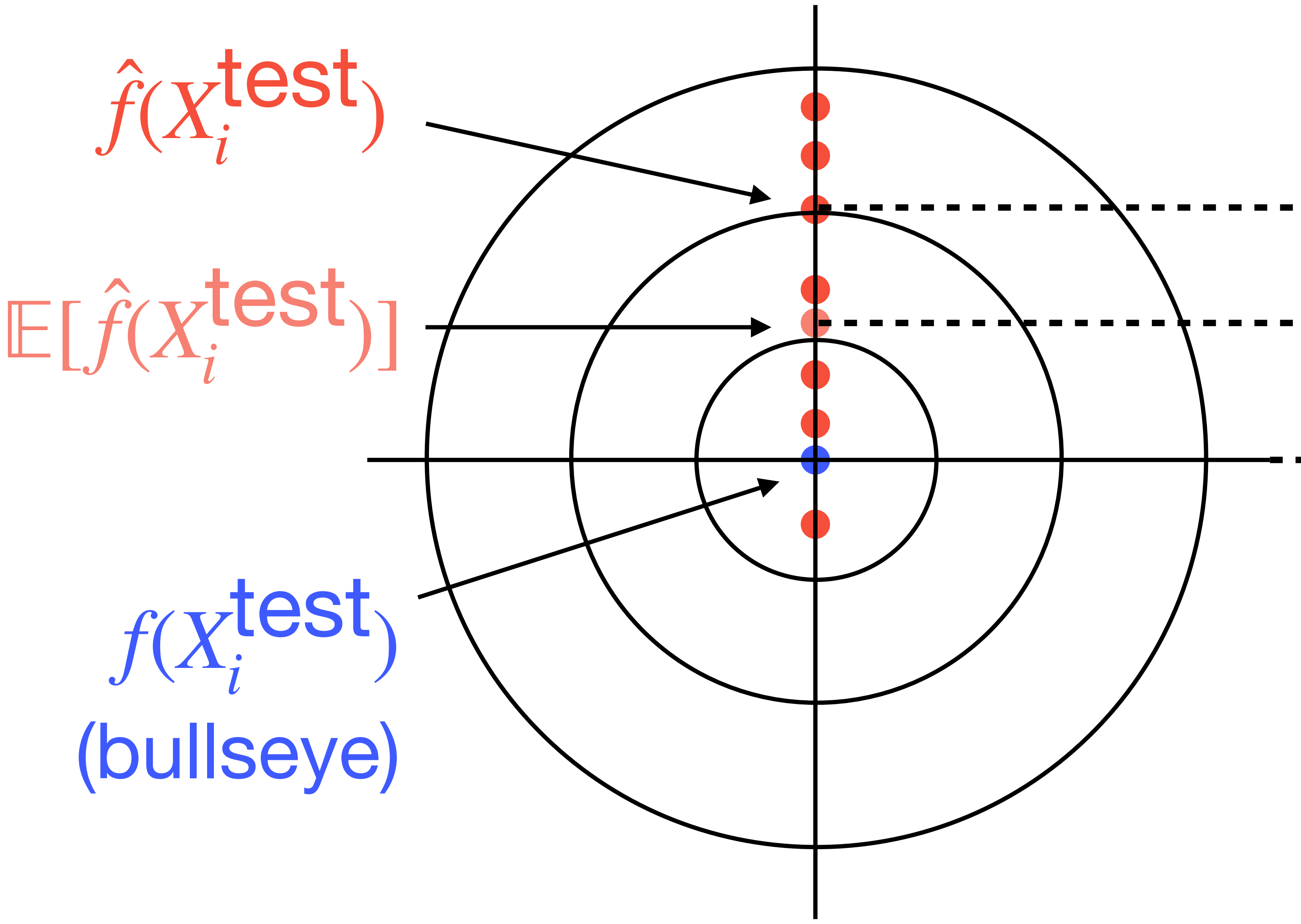
Note: Neither noise variance nor training sample size impacts bias.

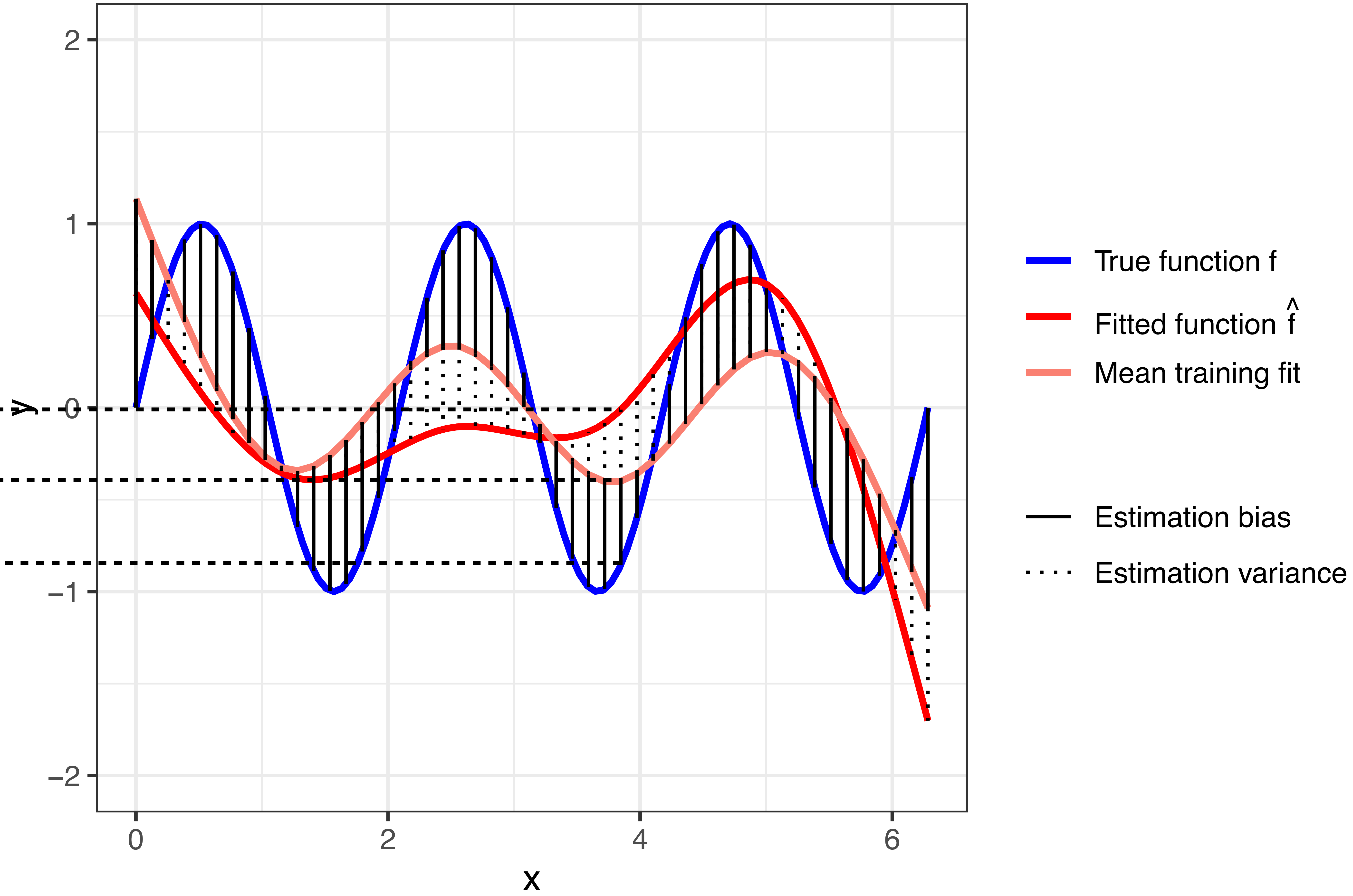
Understanding variance



Understanding variance

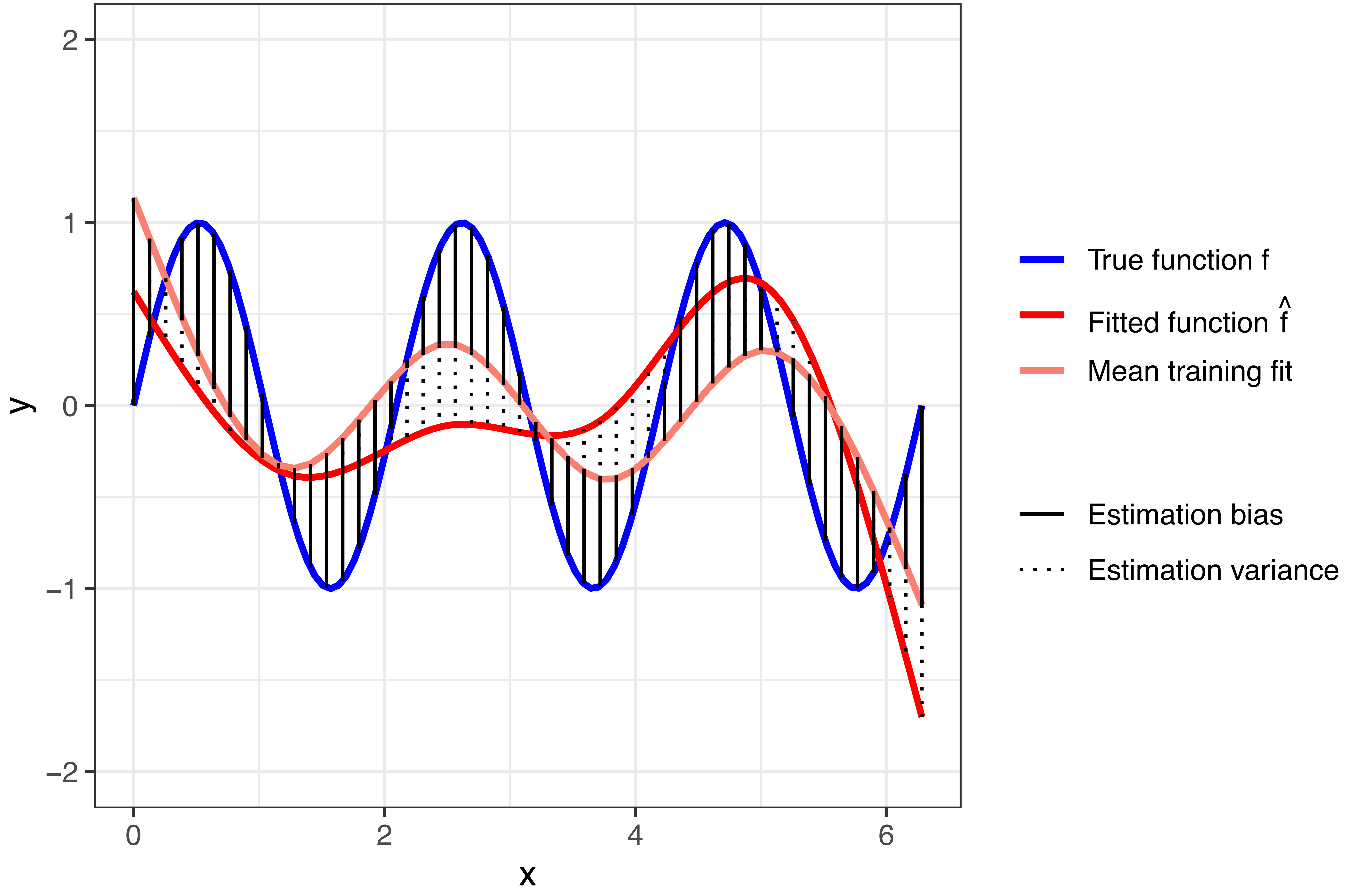
Variance$_i$ $= \mathbb{E}[(\hat{f}(X_i^{\text{test}}) - \mathbb{E}[\hat{f}(X_i^{\text{test}})])^2]$,
the wobbling of the model fit due to the randomness in the training data.





Understanding variance

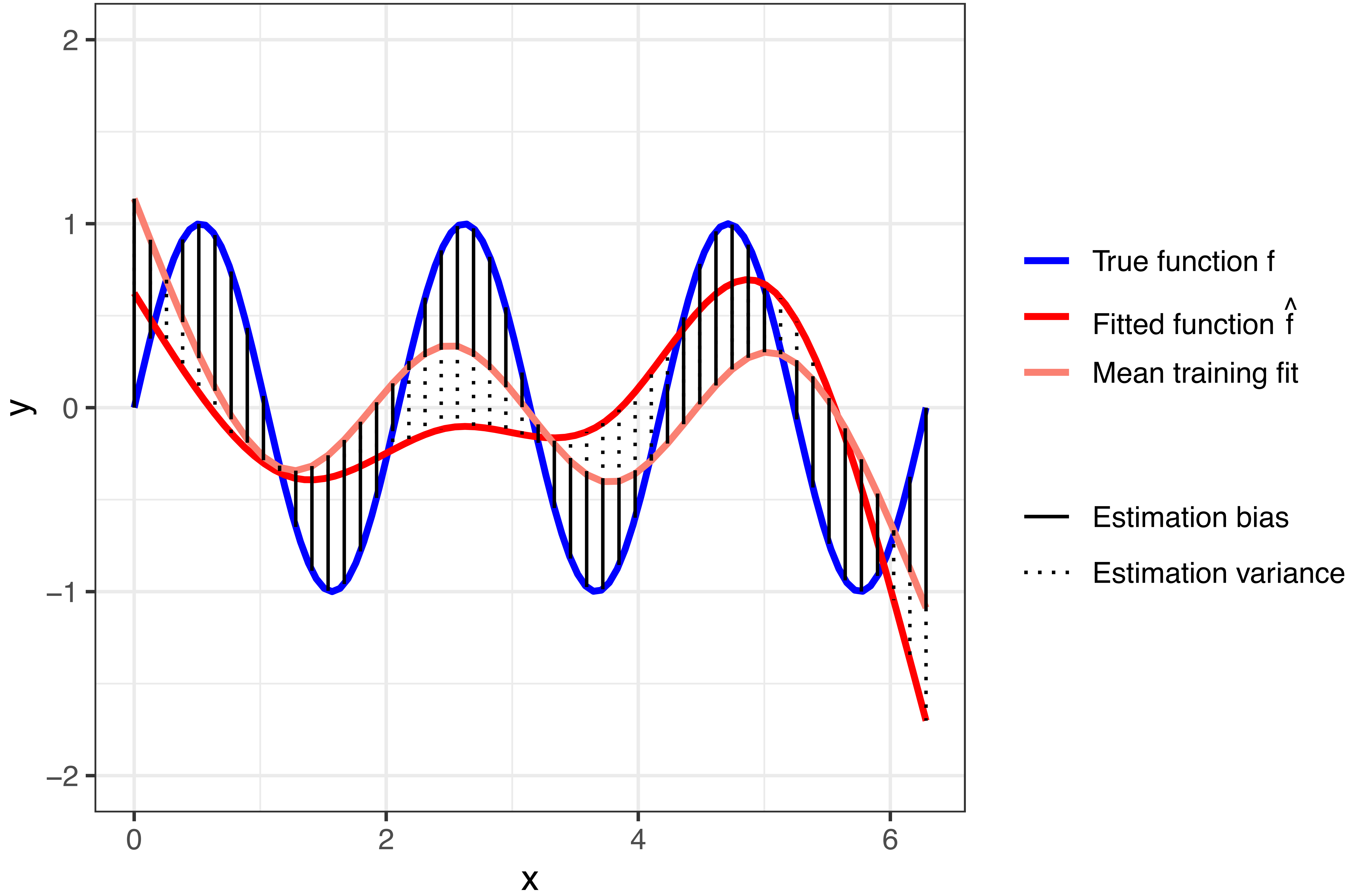
$\text{Variance}_i = \mathbb{E}[(\hat{f}(X_i^{\text{test}}) - \mathbb{E}[\hat{f}(X_i^{\text{test}})])^2]$, the wobbling of the model fit due to the randomness in the training data.



Understanding variance

$\text{Variance}_i = \mathbb{E}[(\hat{f}(X_i^{\text{test}}) - \mathbb{E}[\hat{f}(X_i^{\text{test}})])^2]$, the wobbling of the model fit due to the randomness in the training data.

Variance is due to overfitting; sensitivity to noise in training data causes wobbling.

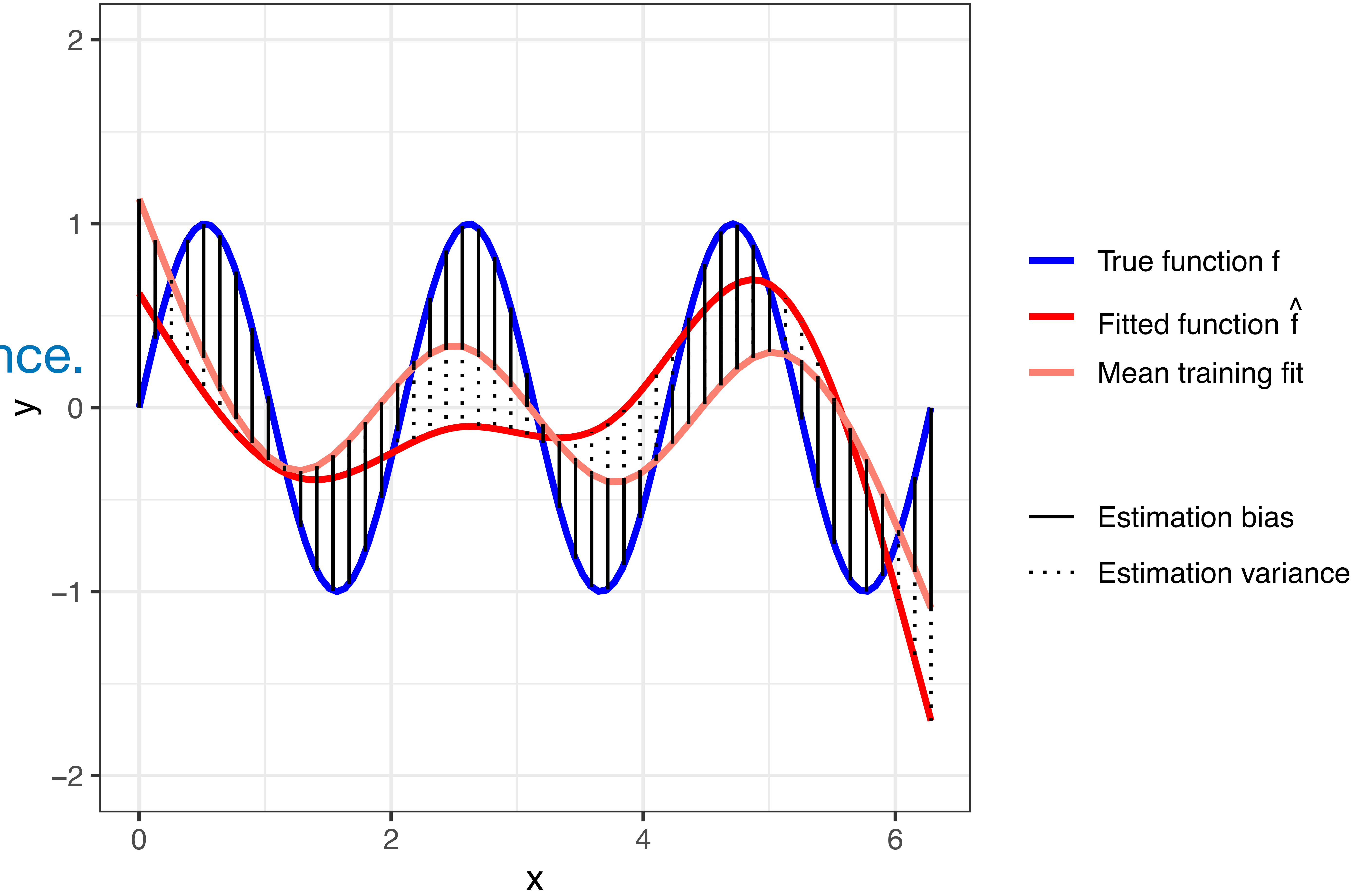


Understanding variance

$\text{Variance}_i = \mathbb{E}[(\hat{f}(X_i^{\text{test}}) - \mathbb{E}[\hat{f}(X_i^{\text{test}})])^2]$, the wobbling of the model fit due to the randomness in the training data.

Variance is due to overfitting; sensitivity to noise in training data causes wobbling.

Adding model complexity **increases** variance.



Understanding variance

$\text{Variance}_i = \mathbb{E}[(\hat{f}(X_i^{\text{test}}) - \mathbb{E}[\hat{f}(X_i^{\text{test}})])^2]$, the wobbling of the model fit due to the randomness in the training data.

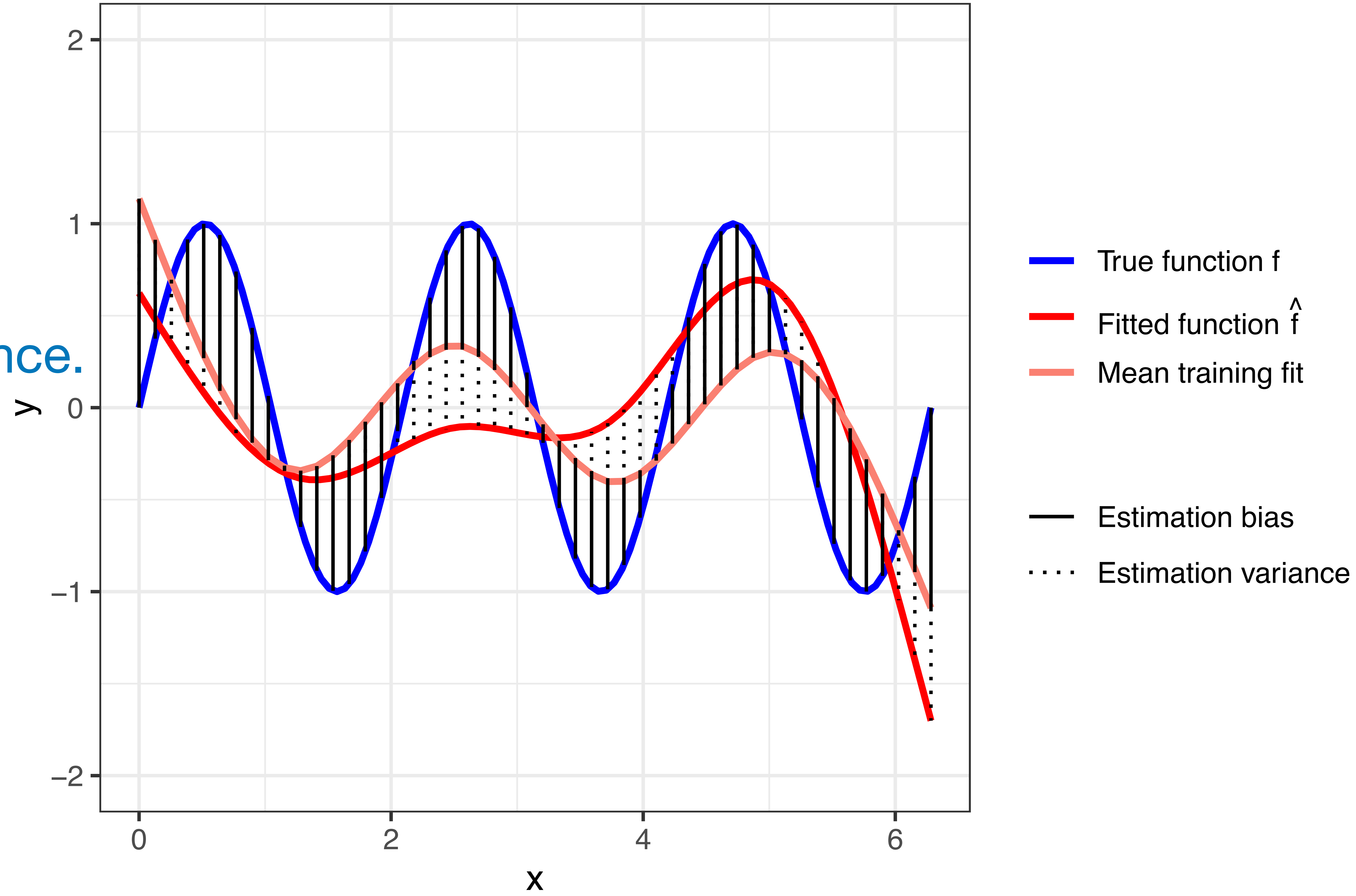
Variance is due to overfitting; sensitivity to noise in training data causes wobbling.

Adding model complexity **increases** variance.

In linear models,

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

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



Understanding variance

$\text{Variance}_i = \mathbb{E}[(\hat{f}(X_i^{\text{test}}) - \mathbb{E}[\hat{f}(X_i^{\text{test}})])^2]$, the wobbling of the model fit due to the randomness in the training data.

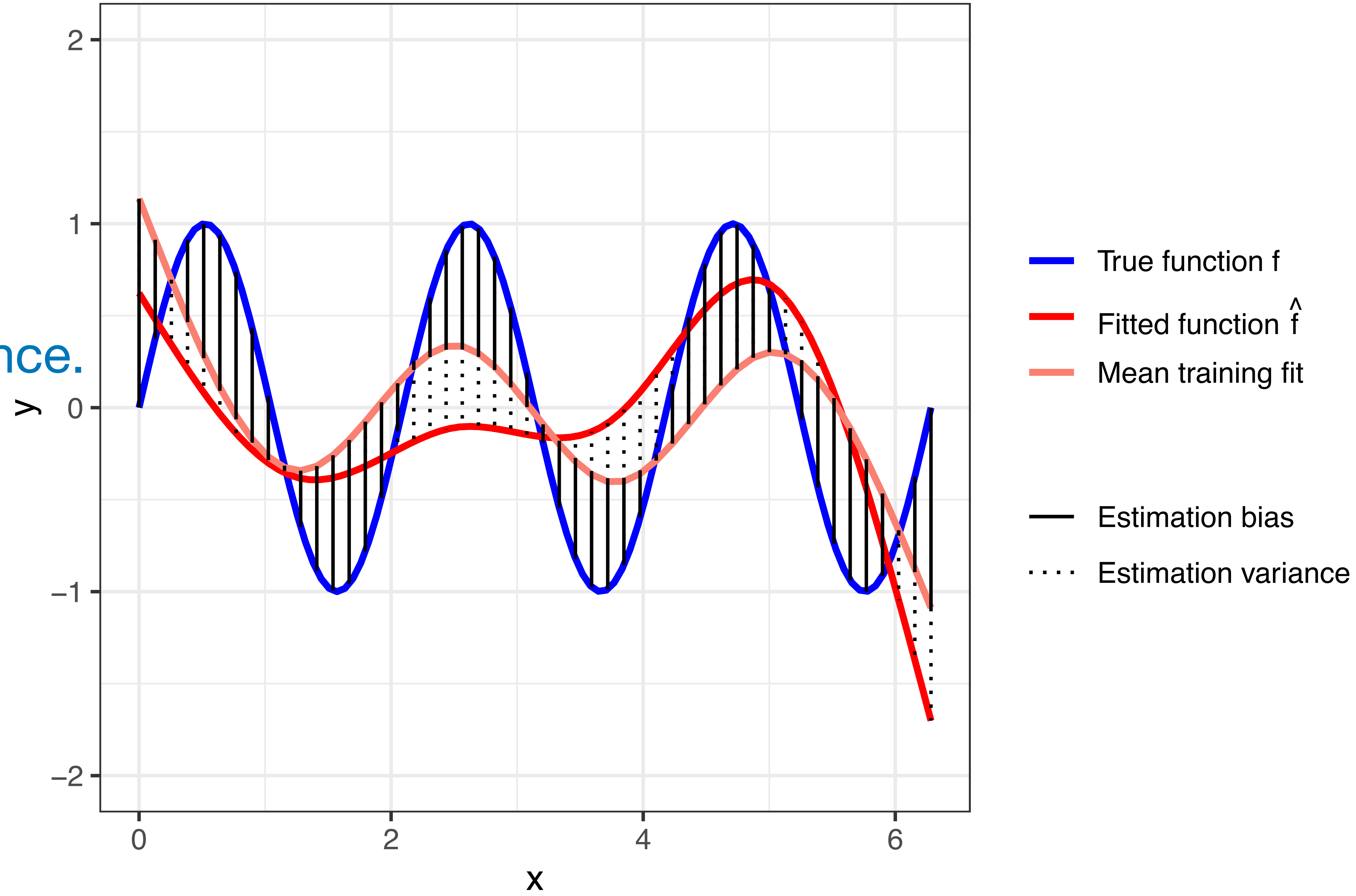
Variance is due to overfitting; sensitivity to noise in training data causes wobbling.

Adding model complexity **increases** variance.

In linear models,

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

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



Note: Complexity of true model does not impact variance.

Putting it all together: The bias-variance tradeoff

Putting it all together: The bias-variance tradeoff

Averaging over i, we get

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

Putting it all together: The bias-variance tradeoff

Averaging over i, we get

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

$$= (\text{Estimation error})_i + \text{Irreducible error}$$

Putting it all together: The bias-variance tradeoff

Averaging over i, we get

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

$$= (\text{Estimation error})_i + \text{Irreducible error}$$

$$= \frac{1}{N}\sum_{i=1}^{N} \text{Bias}_i^2 + \frac{1}{N}\sum_{i=1}^{N} \text{Variance}_i + \sigma^2$$

Putting it all together: The bias-variance tradeoff

Averaging over i, we get

$$\begin{aligned}\text{ETE} &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[(Y_i^{\text{test}} - \hat{f}(X_i^{\text{test}}))^2] \\ &= (\text{Estimation error})_i + \text{Irreducible error} \\ &= \frac{1}{N}\sum_{i=1}^{N}\text{Bias}_i^2 + \frac{1}{N}\sum_{i=1}^{N}\text{Variance}_i + \sigma^2 \\ &= \text{Mean squared bias + Mean variance + Irreducible error.}\end{aligned}$$

Putting it all together: The bias-variance tradeoff

Averaging over i, we get

$$\begin{aligned}
\text{ETE} &= \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[(Y_i^{\text{test}} - \hat{f}(X_i^{\text{test}}))^2] \\
&= (\text{Estimation error})_i + \text{Irreducible error} \\
&= \frac{1}{N}\sum_{i=1}^{N} \text{Bias}_i^2 + \frac{1}{N}\sum_{i=1}^{N} \text{Variance}_i + \sigma^2 \\
&= \text{Mean squared bias + Mean variance + Irreducible error.}
\end{aligned}$$

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

Putting it all together: The bias-variance tradeoff

Averaging over i, we get

$$\begin{aligned}\text{ETE} &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[(Y_i^{\text{test}} - \hat{f}(X_i^{\text{test}}))^2] \\ &= (\text{Estimation error})_i + \text{Irreducible error} \\ &= \frac{1}{N}\sum_{i=1}^{N}\text{Bias}_i^2 + \frac{1}{N}\sum_{i=1}^{N}\text{Variance}_i + \sigma^2 \\ &= \text{Mean squared bias + Mean variance + Irreducible error.}\end{aligned}$$

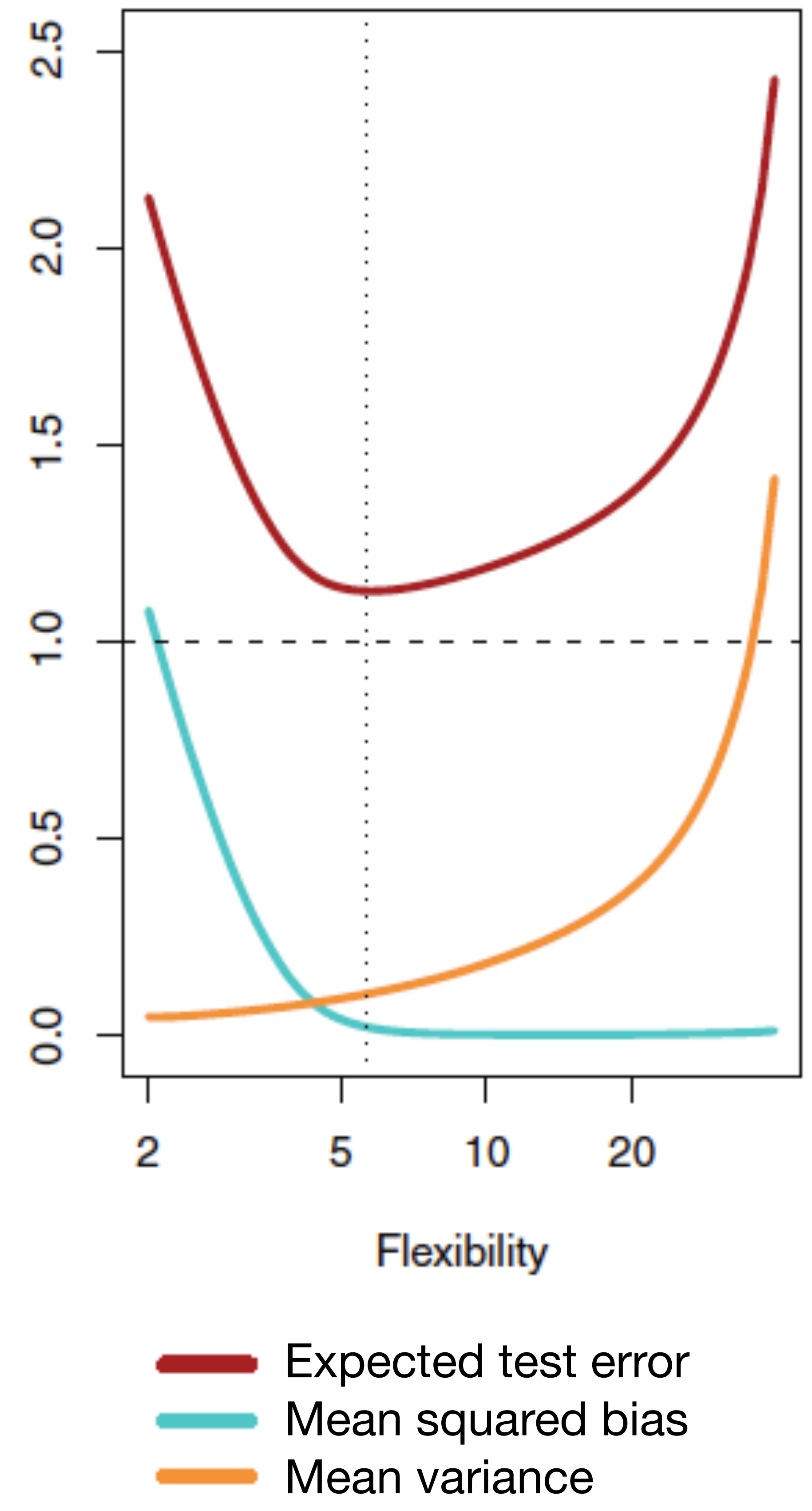
- Adding model complexity reduces bias
- Adding model complexity increases variance

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

Putting it all together: The bias-variance tradeoff

Averaging over i, we get

$$\begin{aligned}\text{ETE} &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[(Y_i^{\text{test}} - \hat{f}(X_i^{\text{test}}))^2] \\ &= (\text{Estimation error})_i + \text{Irreducible error} \\ &= \frac{1}{N}\sum_{i=1}^{N}\text{Bias}_i^2 + \frac{1}{N}\sum_{i=1}^{N}\text{Variance}_i + \sigma^2 \\ &= \text{Mean squared bias + Mean variance + Irreducible error.}\end{aligned}$$



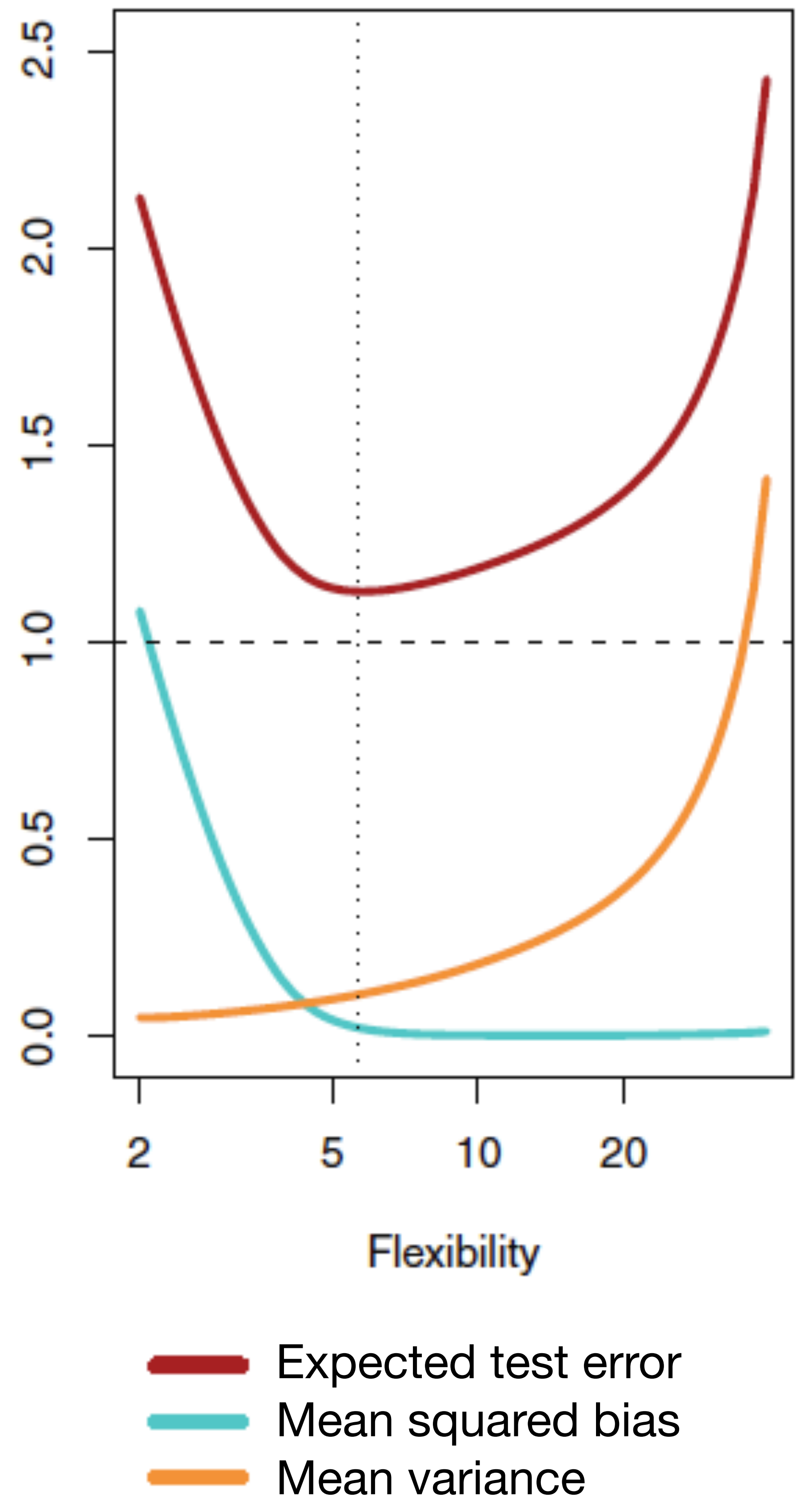
- Adding model complexity reduces bias
- Adding model complexity increases variance

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

Putting it all together: The bias-variance tradeoff

Averaging over i, we get

$$\text{ETE} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[(Y_i^{\text{test}} - \hat{f}(X_i^{\text{test}}))^2]$$
$$= (\text{Estimation error})_i + \text{Irreducible error}$$
$$= \frac{1}{N}\sum_{i=1}^{N} \text{Bias}_i^2 + \frac{1}{N}\sum_{i=1}^{N} \text{Variance}_i + \sigma^2$$
$$= \text{Mean squared bias} + \text{Mean variance} + \text{Irreducible error.}$$

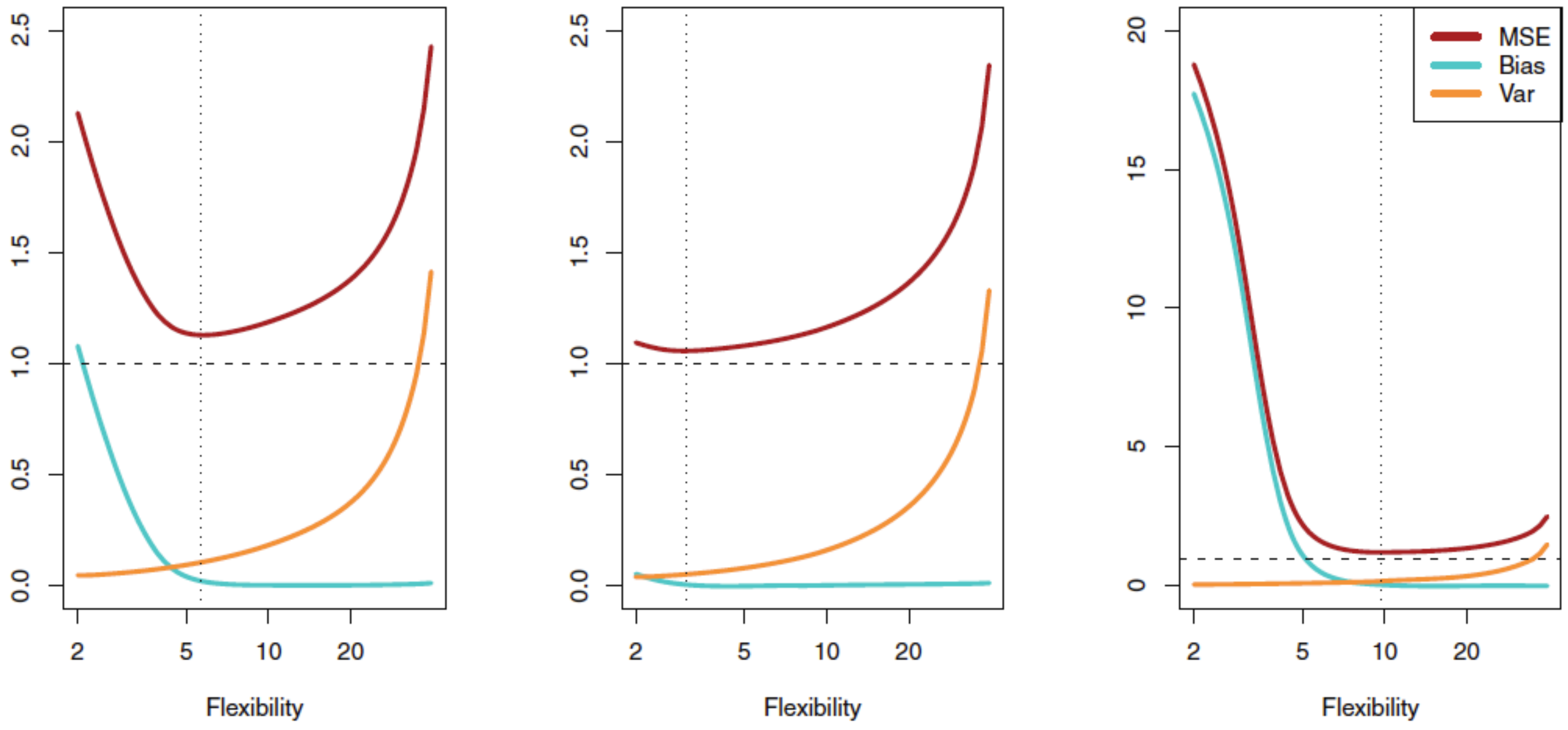


- Adding model complexity reduces bias
- Adding model complexity increases variance

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

Choosing the best predictive model requires balancing the two (Goldilocks principle).

Navigating the bias-variance tradeoff



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

What drives test error?

Problem parameters

- Training 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

What drives test error?

Problem parameters

- Training 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

What drives test error?

Problem parameters

- Training 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

What drives test error?

Problem parameters

- Training 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

What drives test error?

Problem parameters

- Training 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

What drives test error?

Problem parameters

- Training 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