Lecture 08

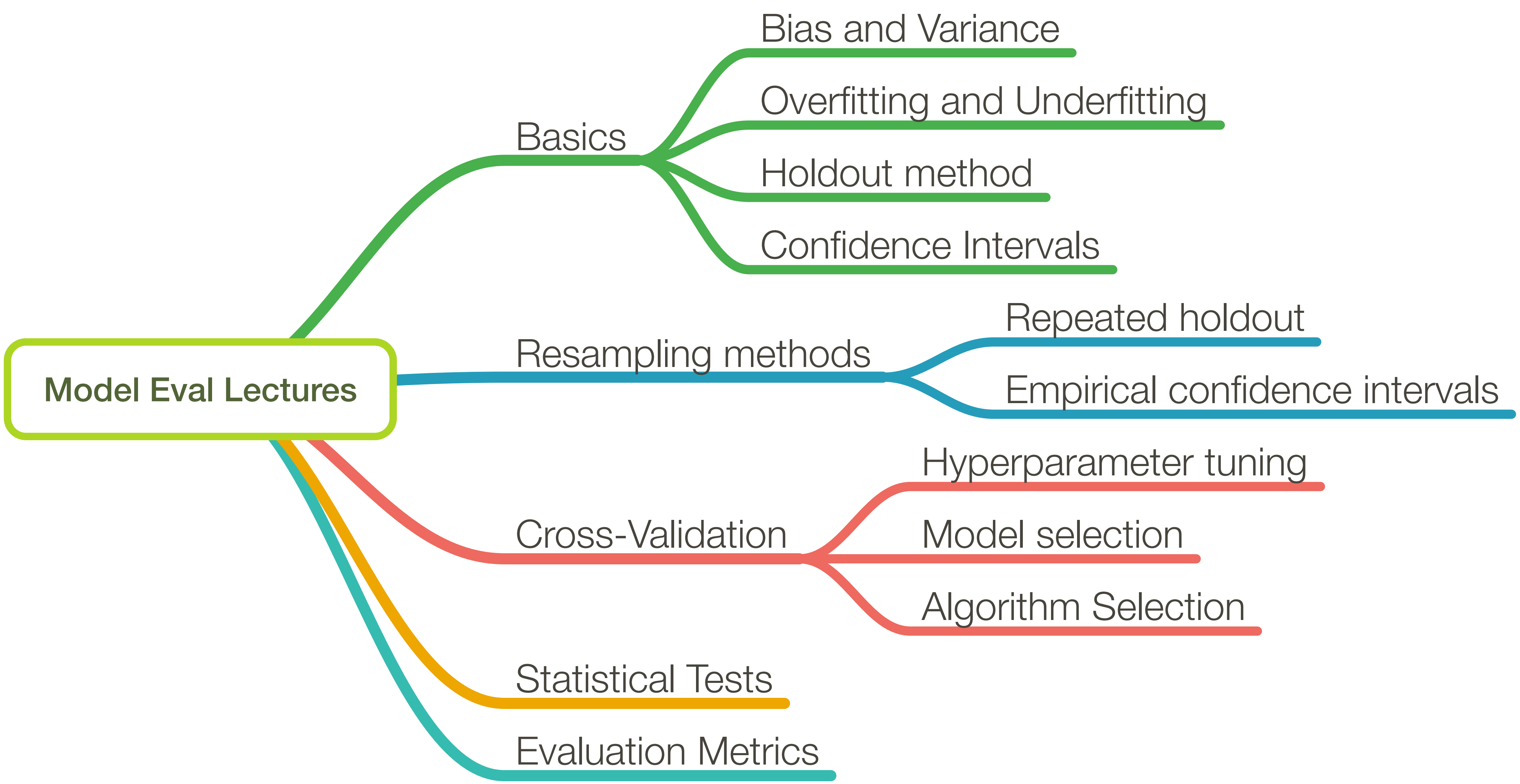
Model Evaluation 1:
Introduction to Overfitting and Underfitting

STAT 479: Machine Learning, Fall 2018

Sebastian Raschka

http://stat.wisc.edu/~sraschka/teaching/stat479-fs2018/

Overview



Model Eval Lectures
Basics
Bias and Variance
Overfitting and Underfitting
Holdout method
Confidence Intervals
Resampling methods
Repeated holdout
Empirical confidence intervals
Cross-Validation
Hyperparameter tuning
Model selection
Algorithm Selection
Statistical Tests
Evaluation Metrics

Overfitting and Underfitting

Overfitting and Underfitting

"Generalization Performance"

- Goal is to fit a model that performs well on unseen inputs, that is, a model that "generalizes" well to unseen data

- A model that performs well on unseen inputs has a good generalization performance

- We say that a model with a good generalization performance has a "high generalization accuracy" or "low generalization error"

Overfitting and Underfitting

Assumptions

- i.i.d. assumption: inputs are independent, and training and test examples are identically distributed (drawn from the same probability distribution)

- The training error or accuracy provides an (optimistically) biased estimate of the generalization performance

- For some random model that has not been fitted to the training set, we expect both the training and test error to be equal

Overfitting and Underfitting

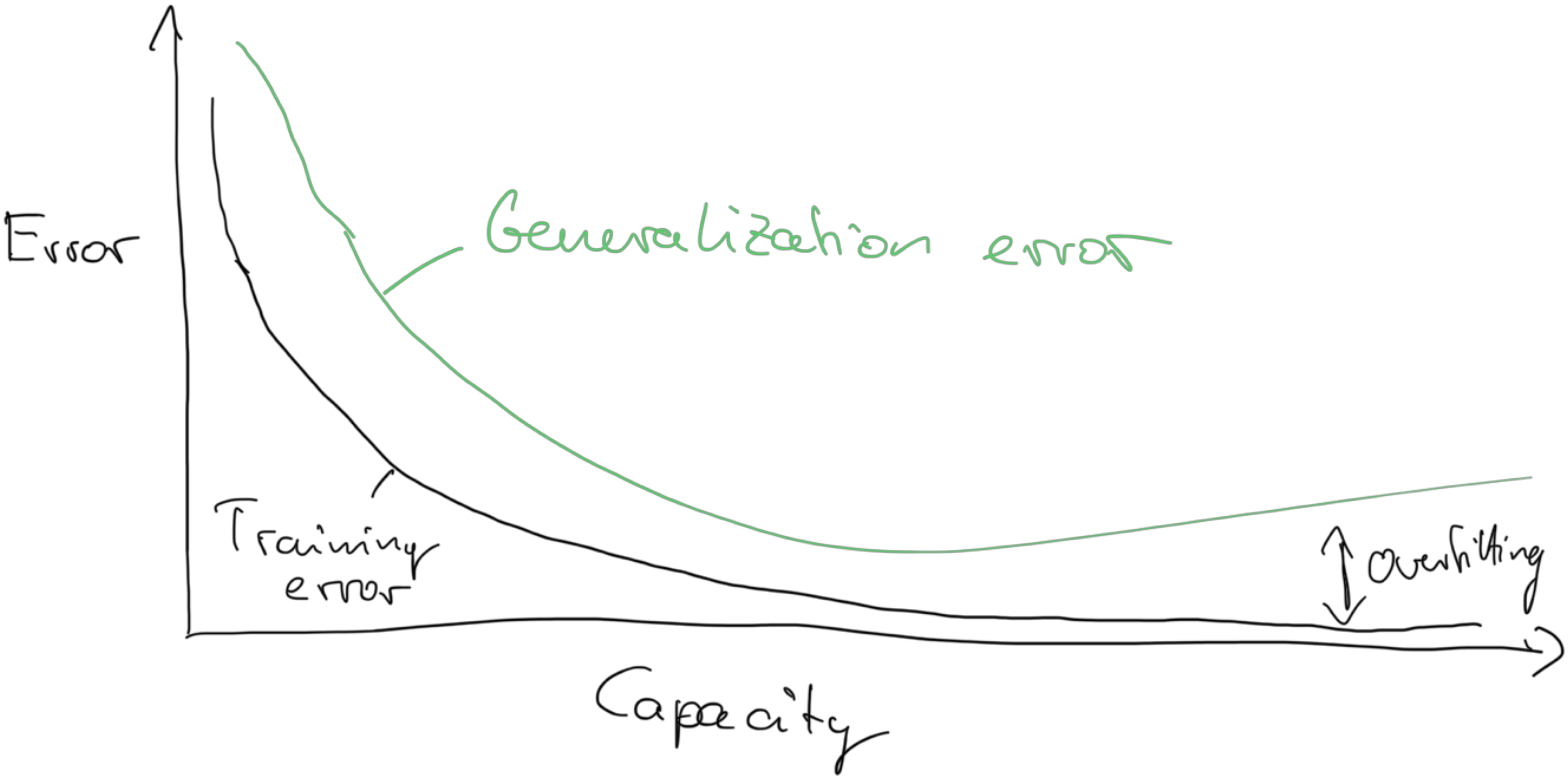
Model Capacity

- Underfitting: both training and test error are large

- Overfitting: gap between training and test error (where test error is higher)

- Generally, the larger the hypothesis space being searched by a learning algorithm, the higher its tendency to overfit (the size of the hypothesis space is related to the so-called "capacity" of a model); vice versa, models with small capacity do not even fit the training set well

Overfitting and Underfitting



Overfitting and Underfitting

Model Capacity

- Underfitting: both training and test error are large

- Overfitting: gap between training and test error (where test error is higher)

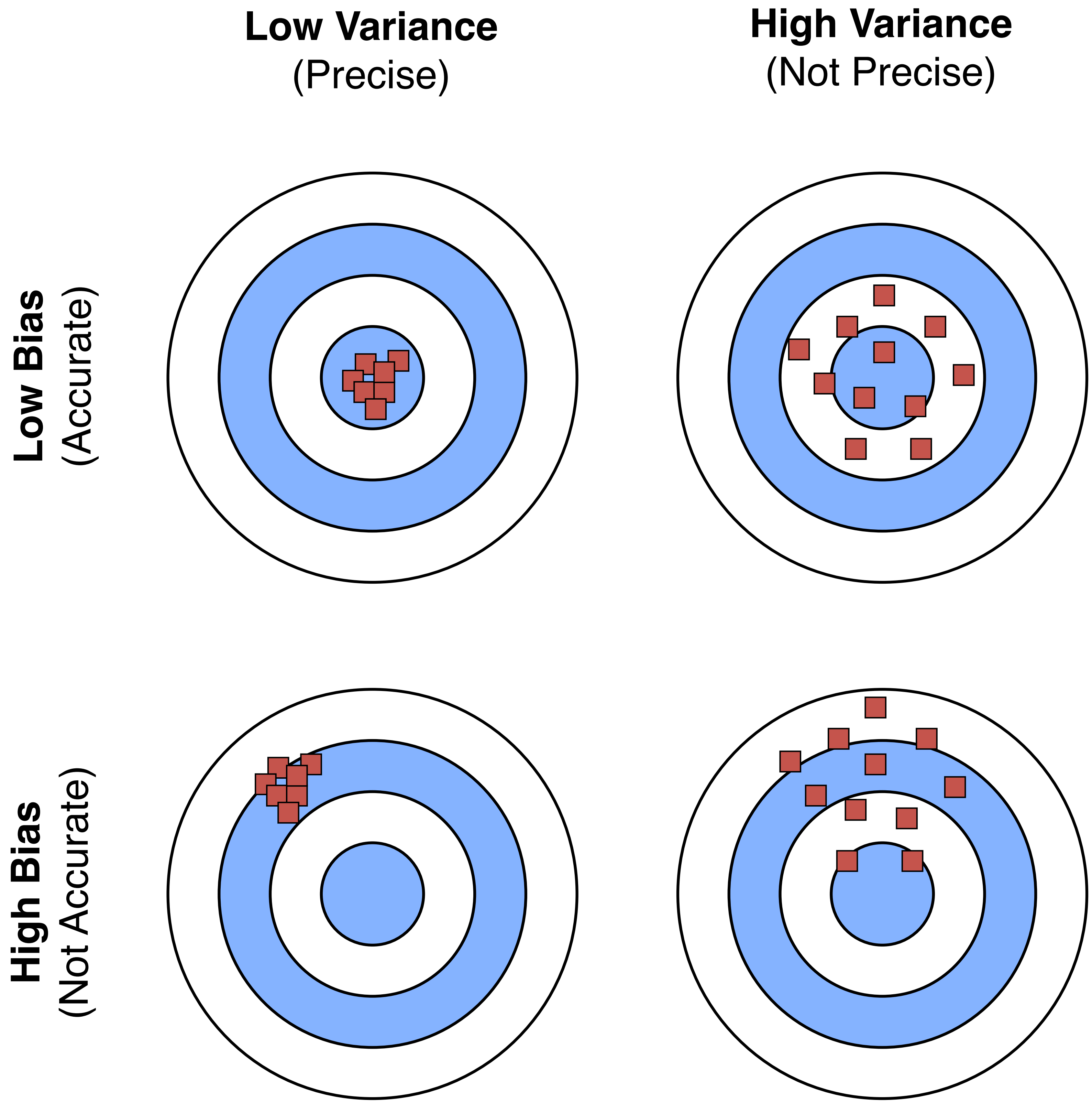
- Generally, the larger the hypothesis space being searched by a learning algorithm, the higher its tendency to overfit (the size of the hypothesis space is related to the so-called "capacity" of a model); vice versa, models with small capacity do not even fit the training set well

Bias-Variance Decomposition and Trade-off

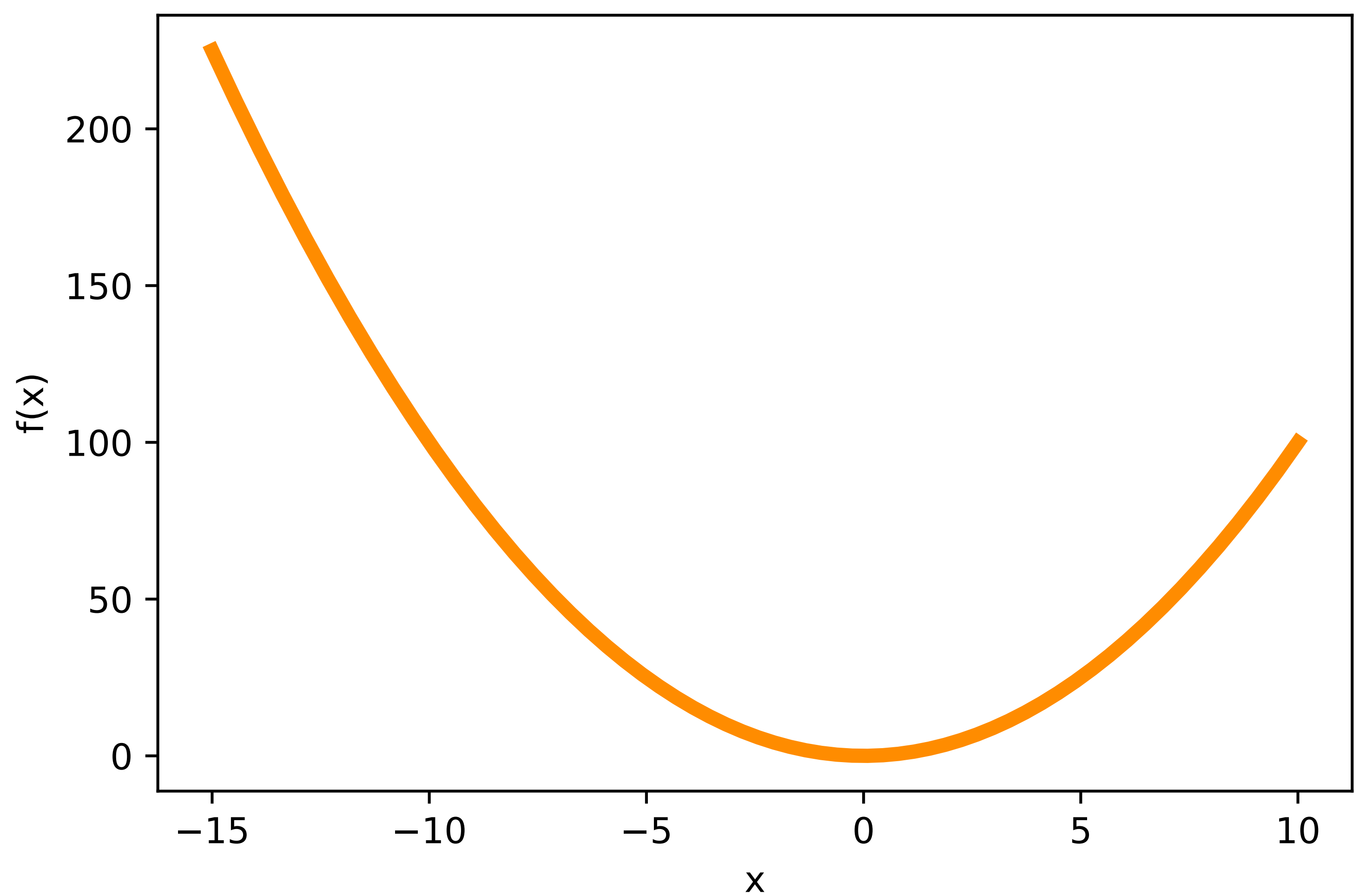
Bias-Variance Decomposition

- Decomposition of the loss into bias and variance help us understand learning algorithms, concepts are correlated to underfitting and overfitting

Bias-Variance Intuition

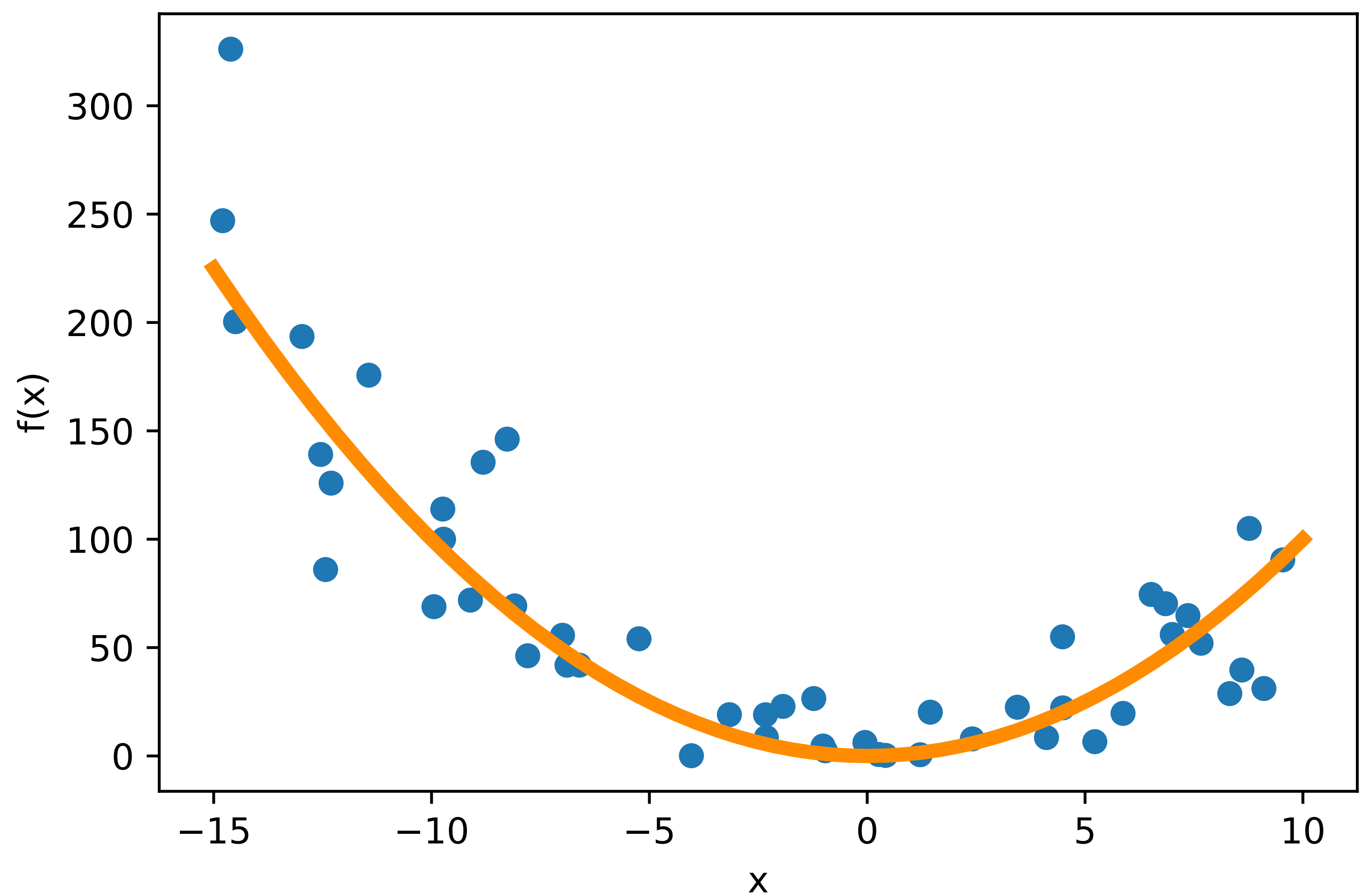


Bias and Variance Example



where f(x) is some true (target) function

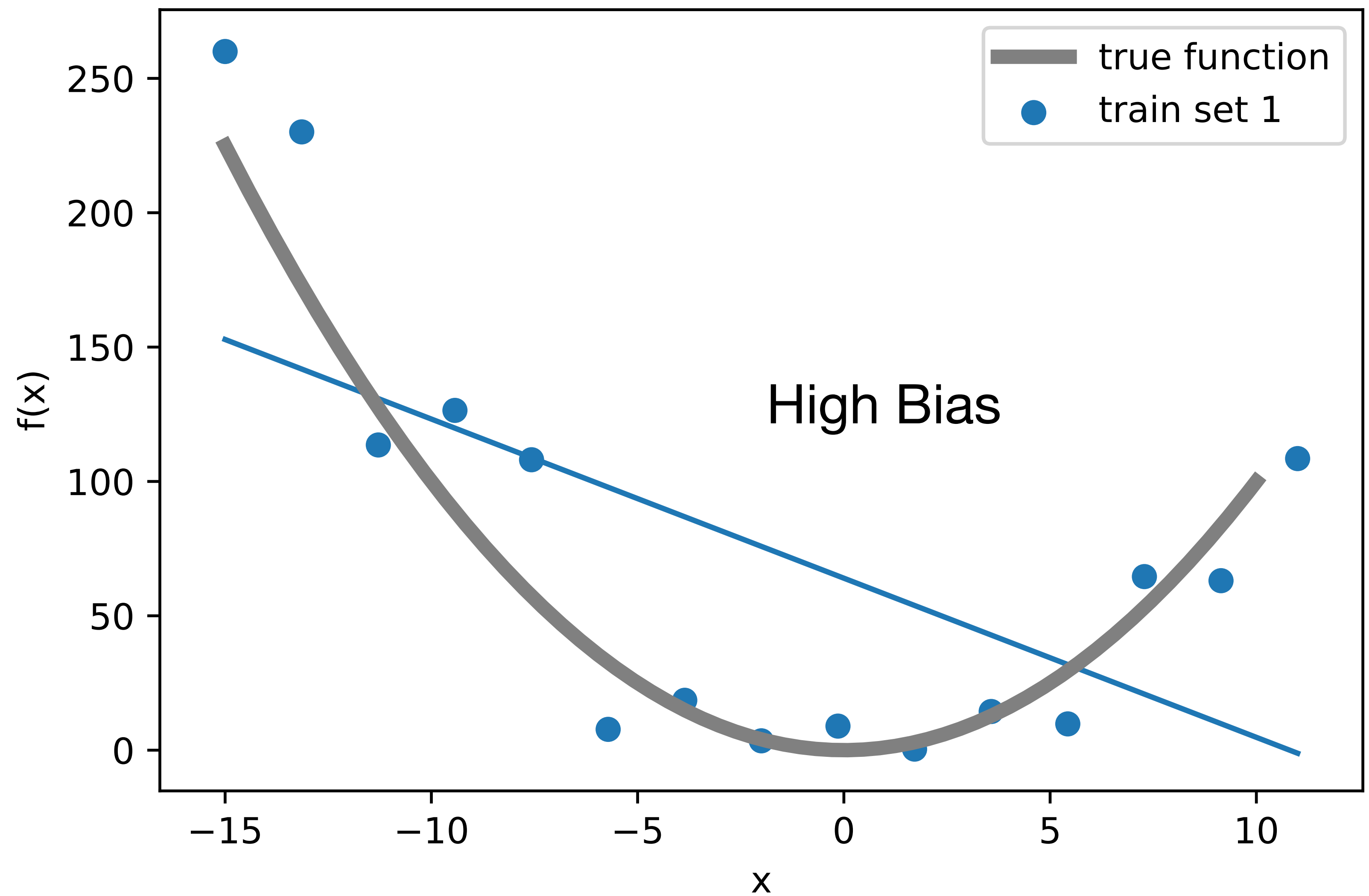
Bias and Variance Example



where f(x) is some true (target) function

the blue dots are a training dataset;
here, I added some random Gaussian noise

Bias and Variance Example

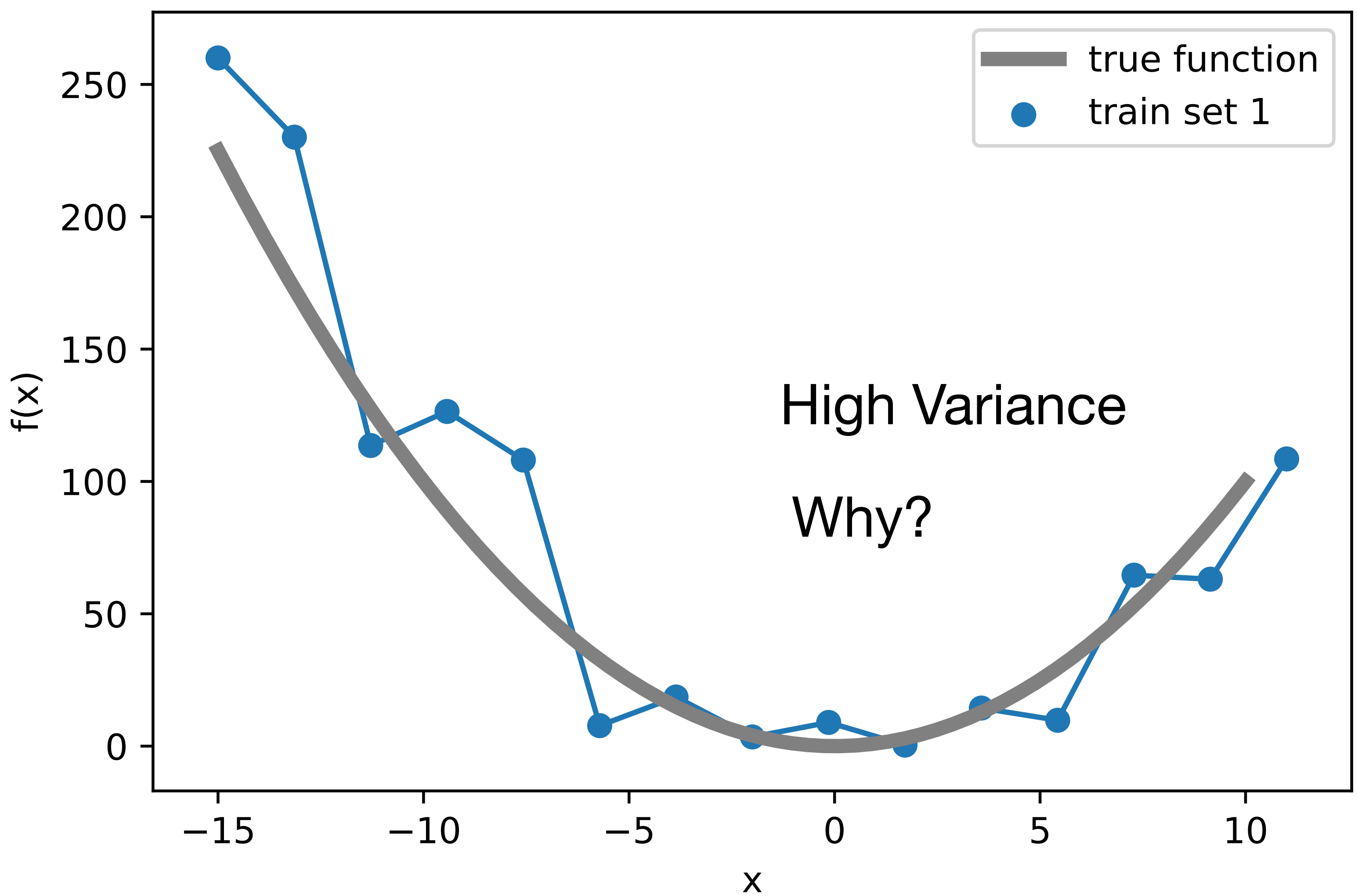


where f(x) is some true (target) function

the blue dots are a training dataset;
here, I added some random Gaussian noise

here, suppose I fit a simple linear model (linear regression)
or a decision tree stump

Bias and Variance Example

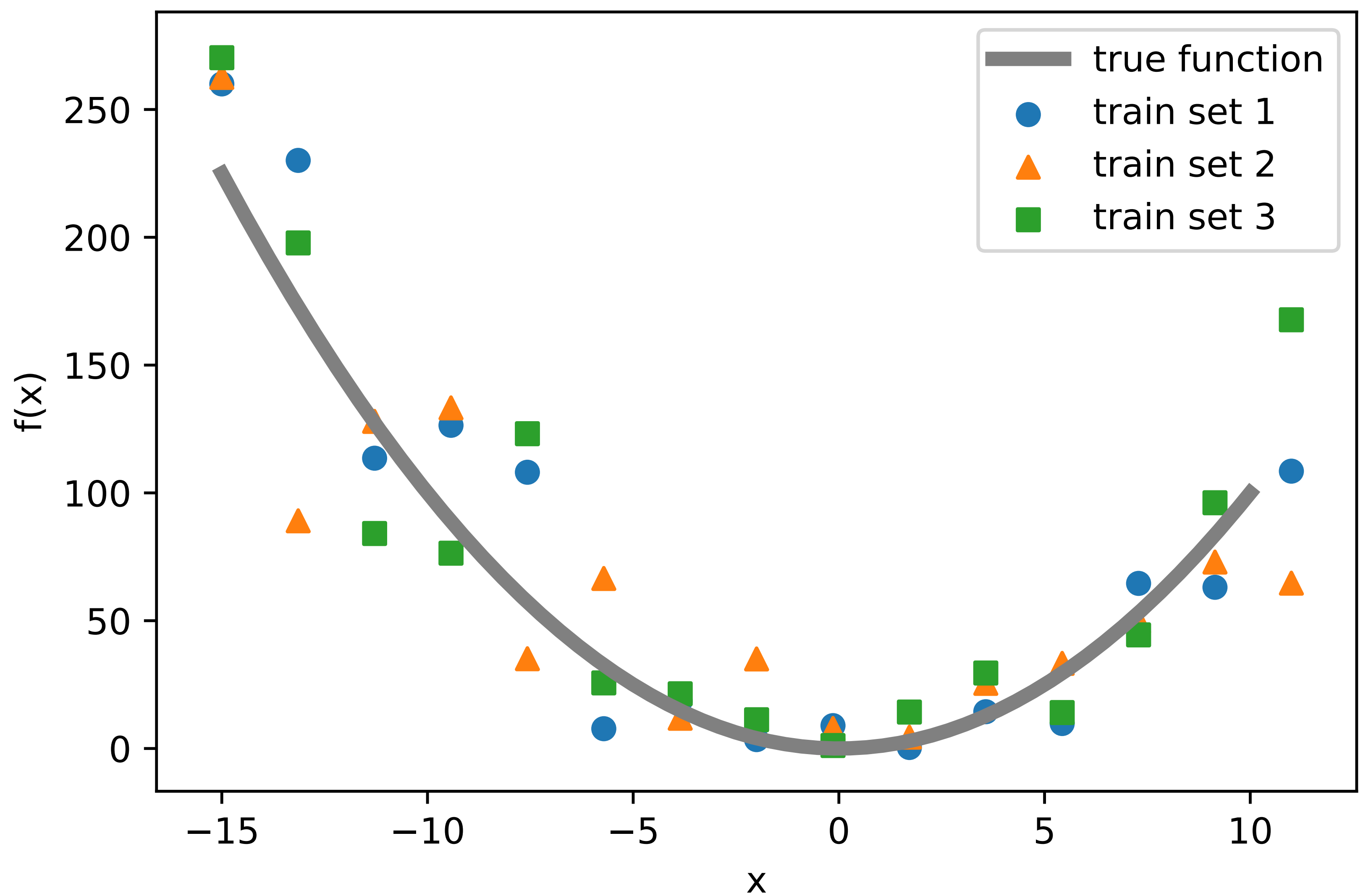


where f(x) is some true (target) function

the blue dots are a training dataset;
here, I added some random Gaussian noise

here, suppose I fit an unpruned decision tree

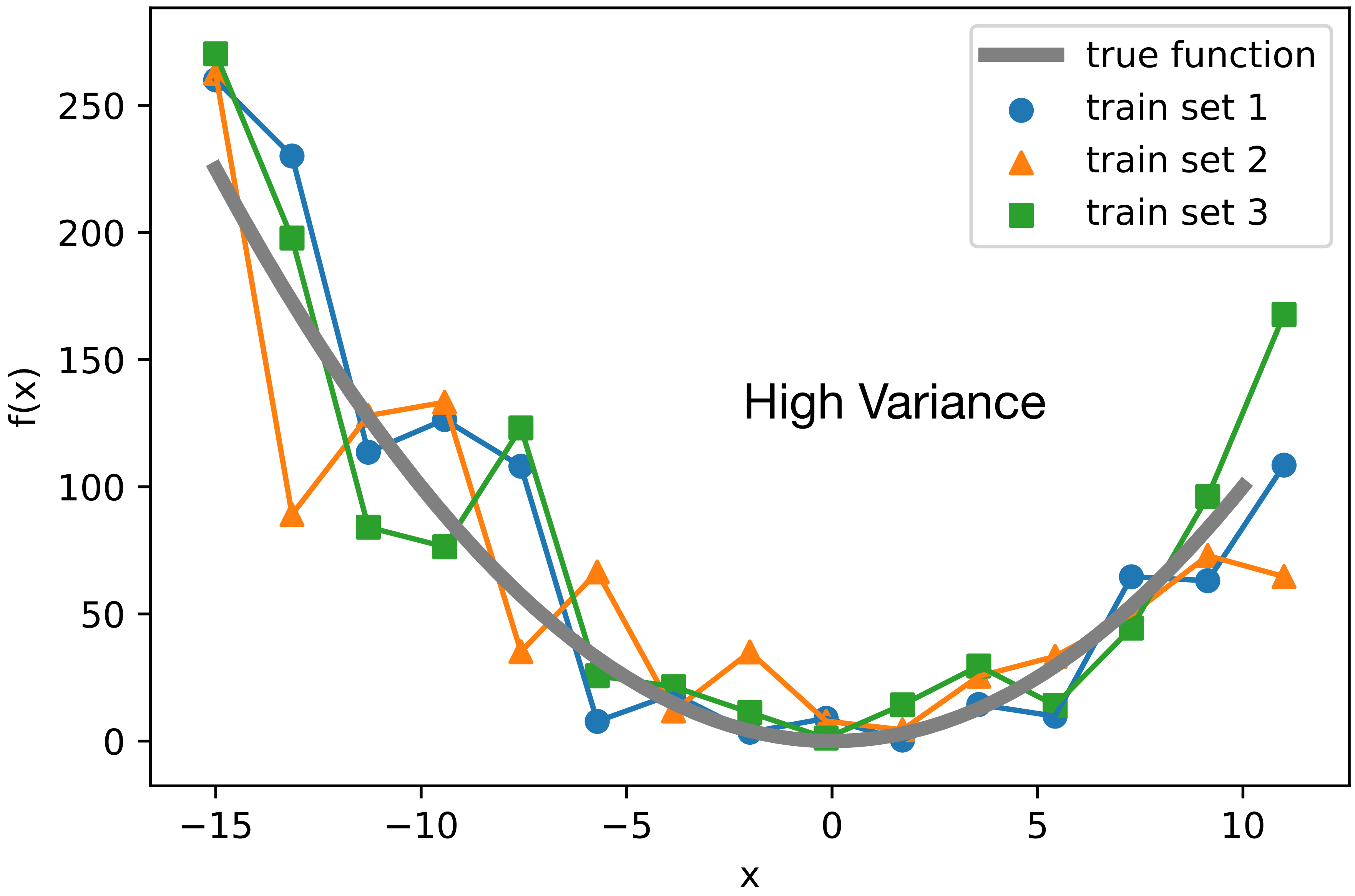
Bias and Variance Example



where f(x) is some true (target) function

suppose we have multiple training sets

Bias and Variance Example



Bias-Variance Decomposition

Point estimator: $\hat{\theta} = f(x^{[1]}, x^{[2]}, \ldots, x^{[n]})$

of some parameter θ

(could also be a function, e.g., the hypothesis is an estimator of some target function)

Bias-Variance Decomposition

Point estimator: $\hat{\theta} = f(x^{[1]}, x^{[2]}, \dots, x^{[n]})$

of some parameter θ

(could also be a function, e.g., the hypothesis is
an estimator of some target function)

Bias $= E[\hat{\theta}] - \theta$

(the expectation is over the training data, i.e, the average estimator from
different training samples)

Bias Example

Normal Distribution: $\mathcal{N}(\mu, \sigma^2)$

Probability density function: $f(x^{[i]}; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \frac{(x^{[i]} - \mu)^2}{\sigma^2} \right)$

Is the sample mean an unbiased estimator of the mean of the Gaussian?

$$\hat{\mu} = \frac{1}{n} \sum_i x^{[i]}$$

Bias Example

Is the sample mean an unbiased estimator of the mean of the Gaussian?

$$\hat{\mu} = \frac{1}{n} \sum_i x^{[i]}$$

$$\begin{aligned} Bias(\hat{\mu}) &= E[\hat{\mu}] - \mu \\ &= E[\frac{1}{n} \sum_i x^{[i]}] - \mu \\ &= \frac{1}{n} \sum_i E[x^{[i]}] - \mu \\ &= \frac{1}{n} \sum_i \mu - \mu \\ &= \mu - \mu = 0 \end{aligned}$$

Variance Example

Is the sample variance as an estimator of the mean of the Gaussian?

$$\hat{\sigma}^2 = \frac{1}{n} \sum_i (x^{[i]} - \hat{\mu})^2$$

$$\begin{aligned} Bias(\hat{\sigma}^2) &= E[\hat{\sigma}^2] - \sigma^2 \\ &= E\left[\frac{1}{n} \sum_i \left(x^{[i]} - \hat{\mu}\right)^2\right] - \sigma^2 \\ &= \ldots \\ &= \frac{m-1}{m}\sigma^2 - \sigma^2 \end{aligned}$$

Variance Example

Is the sample variance as an estimator of the mean of the Gaussian?

$$\hat{\sigma}^2 = \frac{1}{n} \sum_i (x^{[i]} - \hat{\mu})^2$$

$$\begin{aligned} Bias(\hat{\sigma}^2) &= E[\hat{\sigma}^2] - \sigma^2 \\ &= E\left[\frac{1}{n} \sum_i \left(x^{[i]} - \hat{\mu}\right)^2\right] - \sigma^2 \\ &= \ldots \\ &= \frac{m-1}{m}\sigma^2 - \sigma^2 \end{aligned}$$

The unbiased estimator is actually:

$$\hat{\sigma}'^2 = \frac{1}{n-1} \sum_i (x^{[i]} - \hat{\mu})^2$$

Bias-Variance Decomposition of Squared Error

General Definition:

$$\textbf{Bias}(\hat{\theta}) = E[\hat{\theta}] - \theta$$

$$\textbf{Var}(\hat{\theta}) = E\left[\hat{\theta}^2\right] - \left(E\left[\hat{\theta}\right]\right)^2$$

$$\textbf{Var}(\hat{\theta}) = E[(E[\hat{\theta}] - \hat{\theta})^2]$$

Intuition:

Bias is the difference between the average estimator from different training samples and the true value. (The expectation is over the training sets.)

The variance provides an estimate of how much the estimate varies as we vary the training data (e.g,. by resampling).

Bias-Variance Decomposition of Squared Error

General Definition:

$\textbf{Bias}(\hat{\theta}) = E[\hat{\theta}] - \theta$

$\textbf{Var}(\hat{\theta}) = E\left[\hat{\theta}^2\right] - \left(E\left[\hat{\theta}\right]\right)^2$

$\textbf{Var}(\hat{\theta}) = E[(E[\hat{\theta}] - \hat{\theta})^2]$

ML terms for Squared Error Loss:

$y = f(x)$ (target, target function)

$\hat{y} = \hat{f}(x) = h(x)$

$SE = (y - \hat{y})^2$

Bias-Variance Decomposition of Squared Error

ML terms for Squared Error Loss:

$y = f(x)$ (target, target function)

$\hat{y} = \hat{f}(x) = h(x)$

$SE = (y - \hat{y})^2$

$$SE = (y - \hat{y})^2$$

$$\begin{aligned}(y - \hat{y})^2 &= (y - E[\hat{y}] + E[\hat{y}] - \hat{y})^2 \\ &= (y - E[\hat{y}])^2 + (E[\hat{y}] - y)^2 + 2(y - E[\hat{y}])(E[\hat{y}] - \hat{y})\end{aligned}$$

Bias-Variance Decomposition of Squared Error

$$SE = (y - \hat{y})^2$$

$$\begin{aligned}(y - \hat{y})^2 &= (y - E[\hat{y}] + E[\hat{y}] - \hat{y})^2 \\ &= (y - E[\hat{y}])^2 + (E[\hat{y}] - y)^2 + 2(y - E[\hat{y}])(E[\hat{y}] - \hat{y})\end{aligned}$$

$$\begin{aligned}E[SE] &= E[(y - \hat{y})^2] \\ E[(y - \hat{y})^2] &= (y - E[\hat{y}])^2 + E[(E[\hat{y}] - \hat{y})^2] \\ &= \textbf{[Bias of the fit]}^2 + \textbf{Variance of the fit}\end{aligned}$$

Bias-Variance Decomposition of Squared Error

$$SE = (y - \hat{y})^2$$

???

$$\begin{aligned}(y - \hat{y})^2 &= (y - E[\hat{y}] + E[\hat{y}] - \hat{y})^2 \\ &= (y - E[\hat{y}])^2 + (E[\hat{y}] - y)^2 + \boxed{2(y - E[\hat{y}])(E[\hat{y}] - \hat{y})}\end{aligned}$$

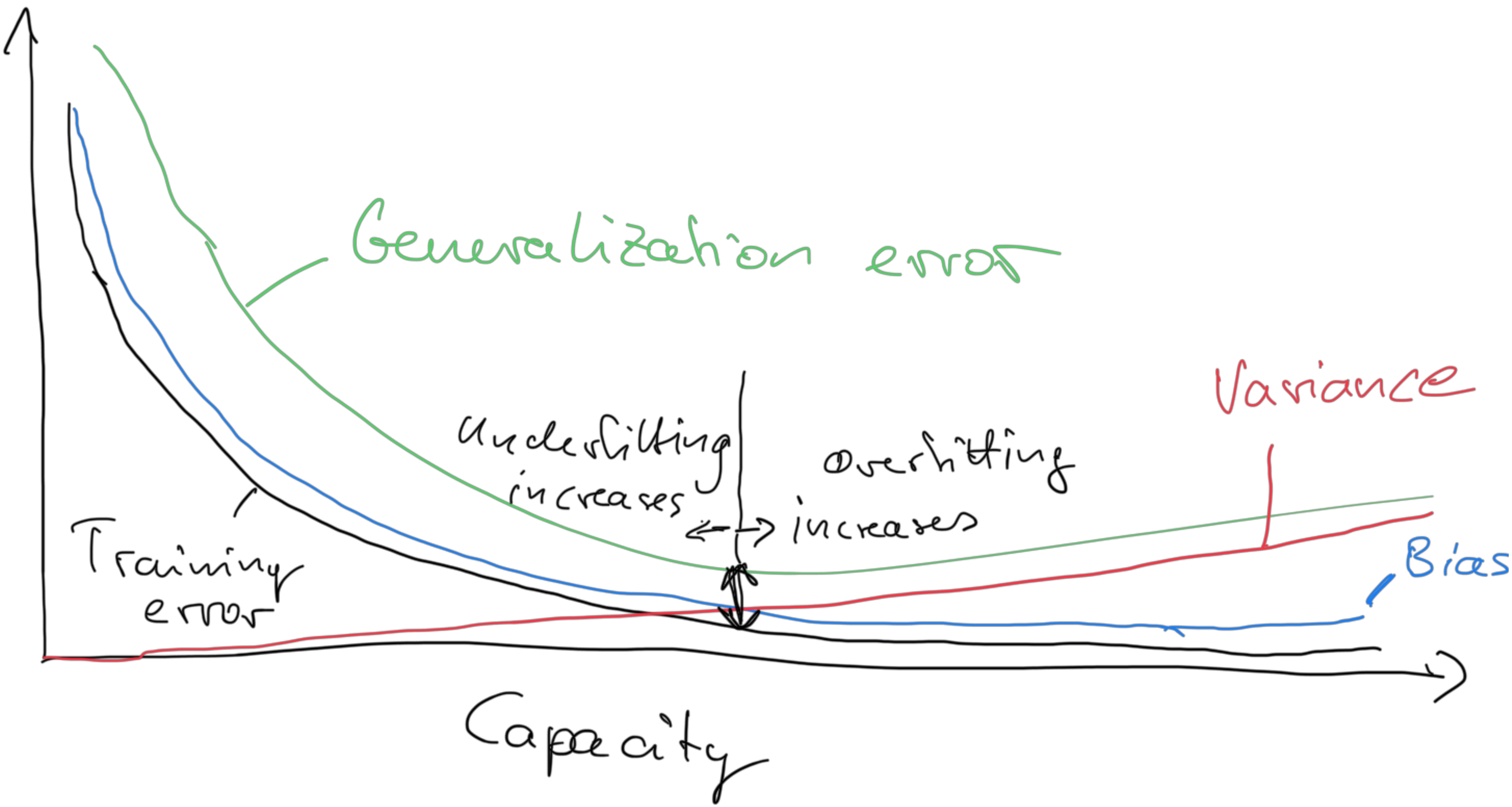
$$\begin{aligned}E[SE] &= E[(y - \hat{y})^2] \\ E[(y - \hat{y})^2] &= (y - E[\hat{y}])^2 + E[(E[\hat{y}] - \hat{y})^2] \\ &= \textbf{[Bias of the fit]}^2 + \textbf{Variance of the fit}\end{aligned}$$

Bias-Variance Decomposition of Squared Error

$$SE = (y - \hat{y})^2$$

$$\begin{aligned}
(y - \hat{y})^2 &= (y - E[\hat{y}] + E[\hat{y}] - \hat{y})^2 \\
&= (y - E[\hat{y}])^2 + (E[\hat{y}] - y)^2 + \overset{???}{\boxed{2(y - E[\hat{y}])(E[\hat{y}] - \hat{y})}}
\end{aligned}$$

$$\begin{aligned}
E[2(y - E[\hat{y}])(E[\hat{y}] - \hat{y})] &= 2E[(y - E[\hat{y}])(E[\hat{y}] - \hat{y})] \\
&= 2(y - E[\hat{y}])E[(E[\hat{y}] - \hat{y})] \\
&= 2(y - E[\hat{y}])(E[E[\hat{y}]] - E[\hat{y}]) \\
&= 2(y - E[\hat{y}])(E[\hat{y}] - E[\hat{y}]) \\
&= 0
\end{aligned}$$


Generalization error
Variance
underfitting increases
overfitting increases
Bias
Training error
Capacity

to be continued ...