Overfitting and Underfitting

These slides are based on the slides of Sebastian rashcka for introduction to deep learning course and the slides of Sharifi Zarchi for the introduction to machine learning course

Generalization Performance

- Want a model to "generalize" well to __________ data
- (Want "high generalization accuracy" or "low generalization error")
- Test sets is a method to estimating the generalization performance on unseen data.

Assumptions

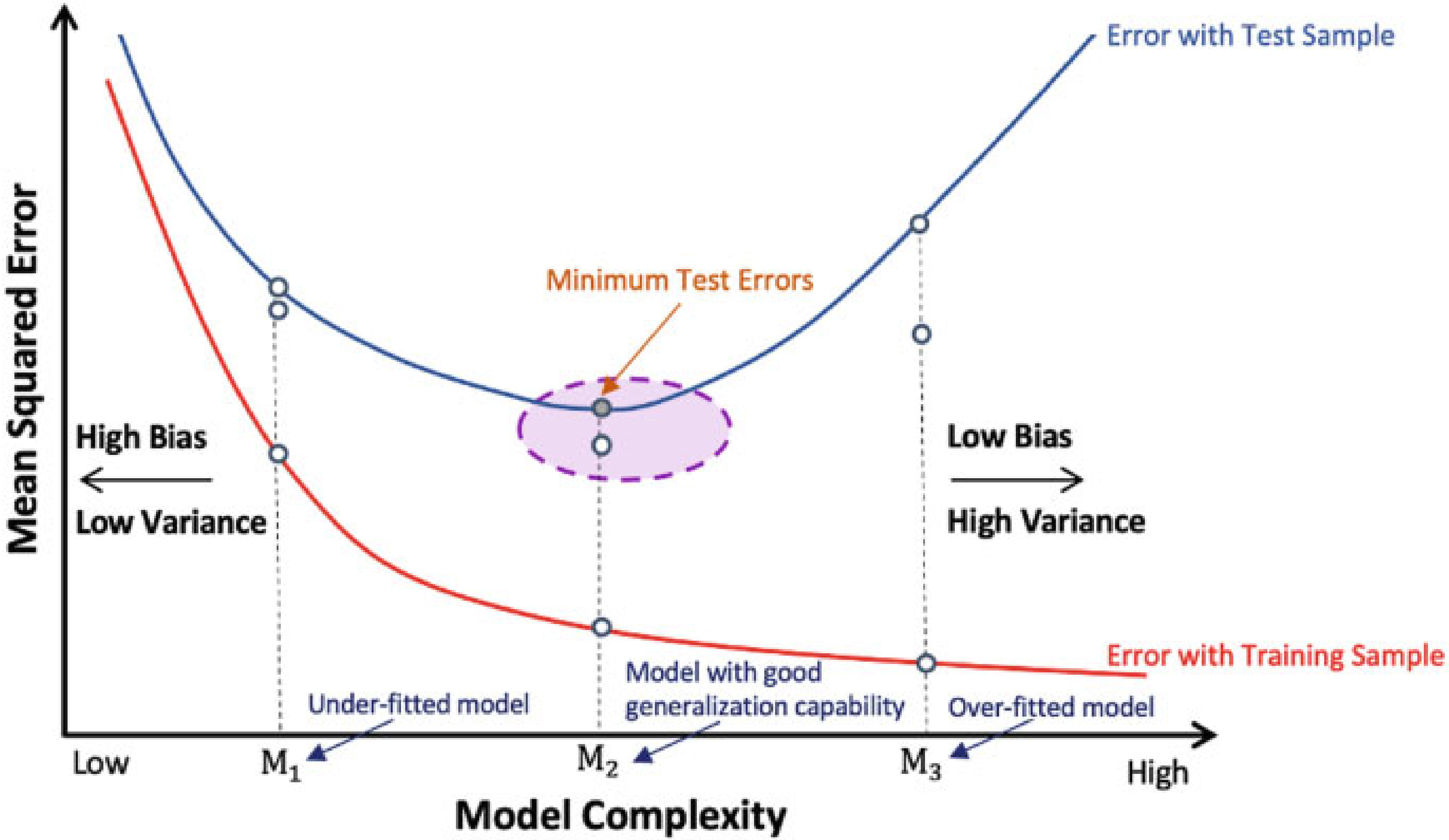
- i.i.d. assumption: training and test examples are independent and identically distributed (drawn from the same joint probability distribution, P(X, y))

- For some random model that has not been fitted to the training set, we expect **the training error is ______ _____ the test error**

- Underfitting: both the training and test error are _________

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

- Large hypothesis space being searched by a learning algorithm -> high tendency to ______fit

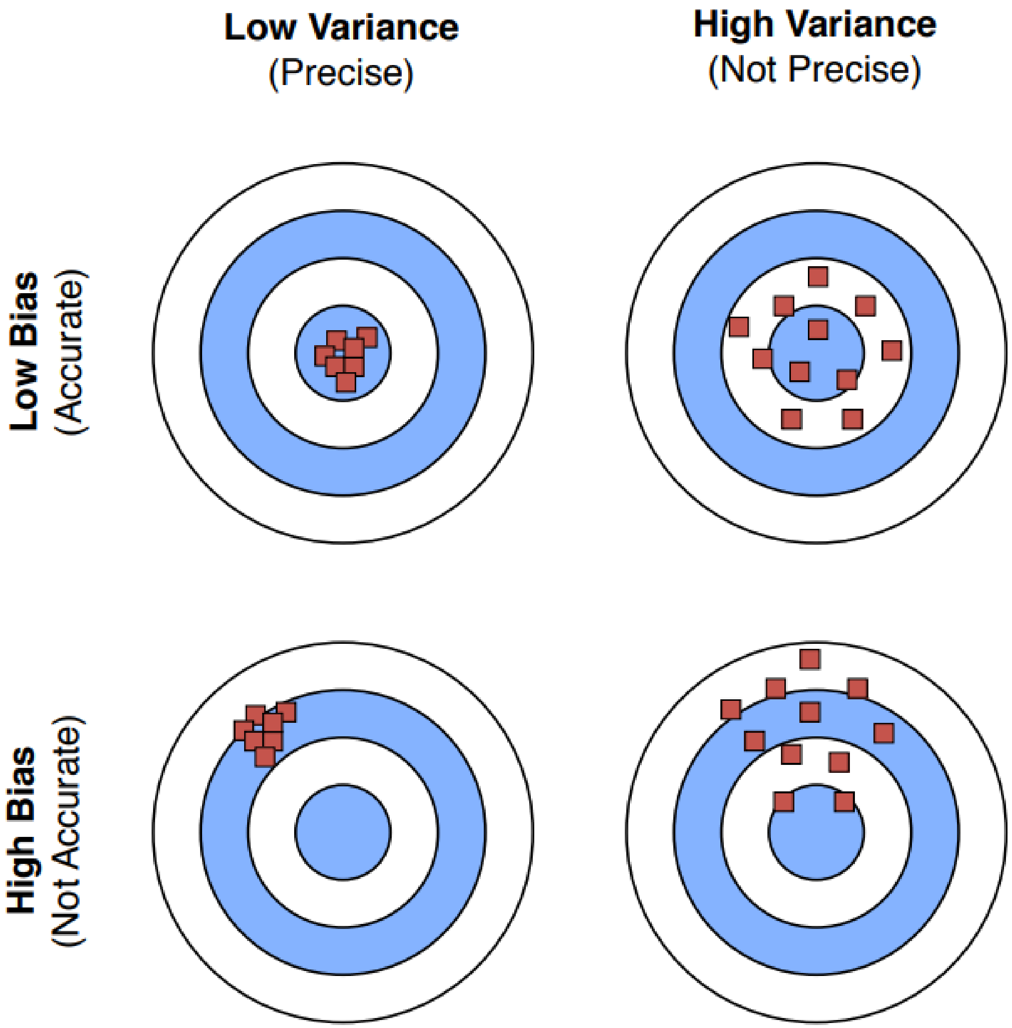


Error with Test Sample
Mean Squared Error
Minimum Test Errors
High Bias
Low Variance
Low Bias
High Variance
Error with Training Sample
Model with good
generalization capability
Under-fitted model
Over-fitted model
Low
M1
M2
M3
High
Model Complexity

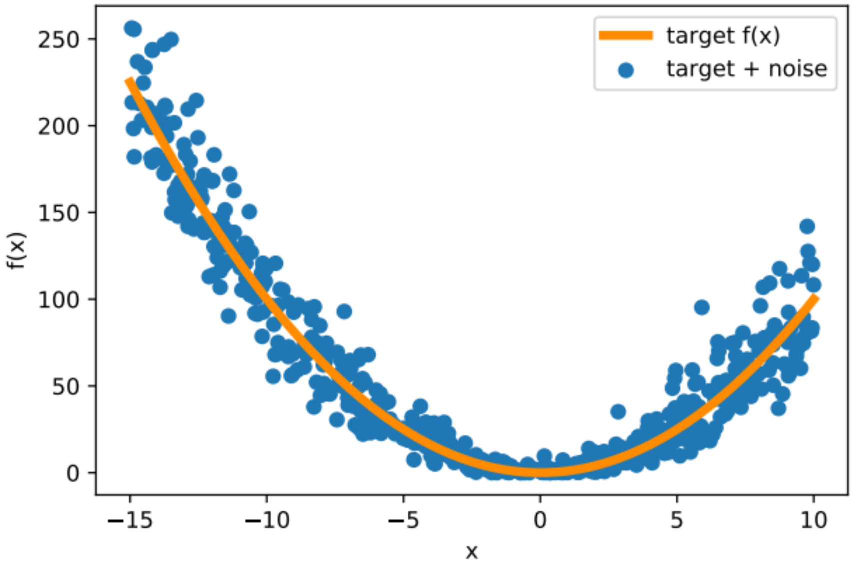
Bios – Variance decomposition

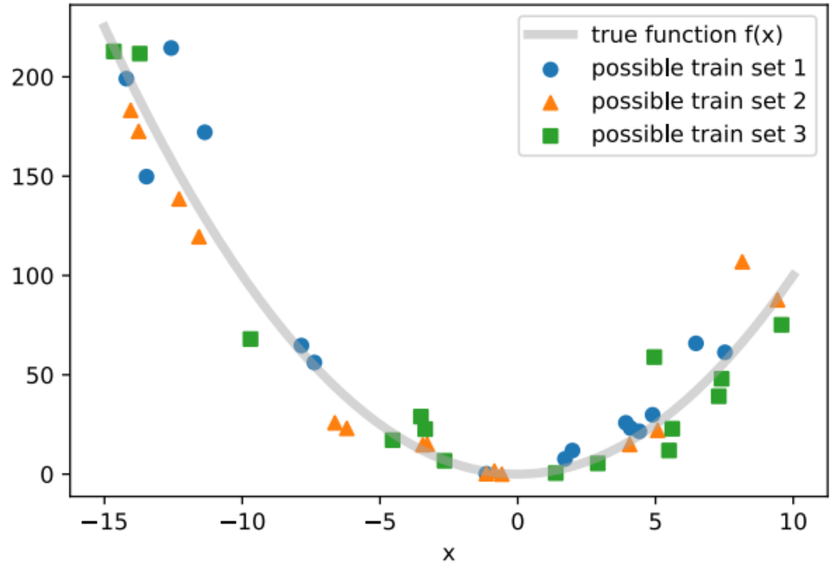
- Decomposition of the loss into bias and variance help us understand learning algorithms, concepts are related to underfitting and overfitting
- Helps explain why ensemble methods might perform better than single models

Bias-Variance Intuition

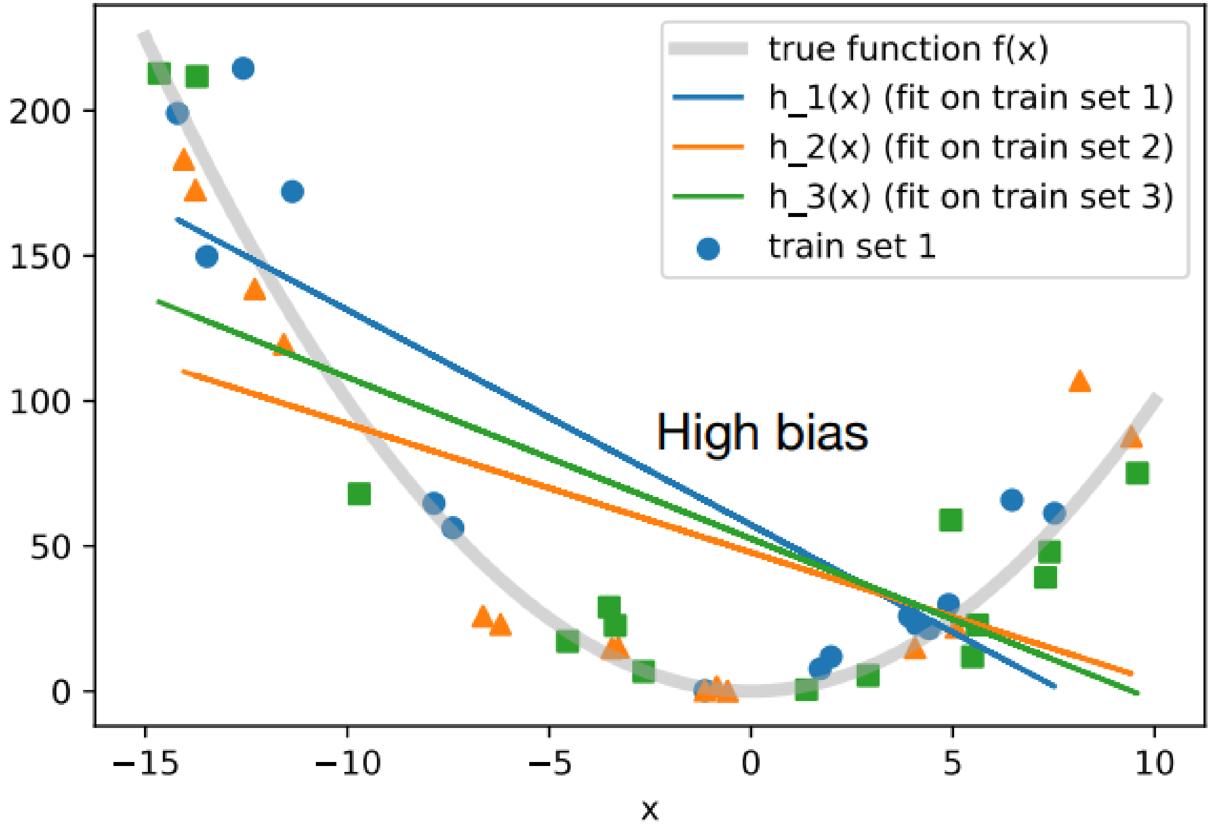


Bias-Variance Intuition





Bias-Variance Intuition

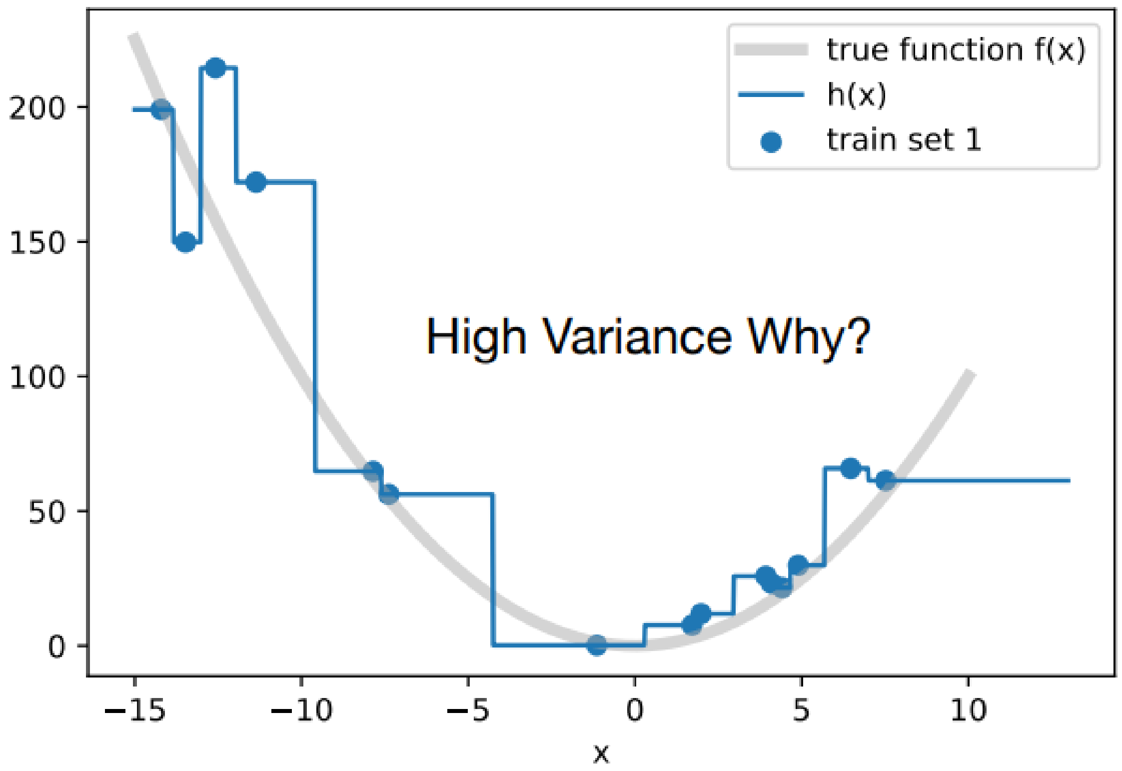


linear regression models

(There are points where the bias is zero ...)

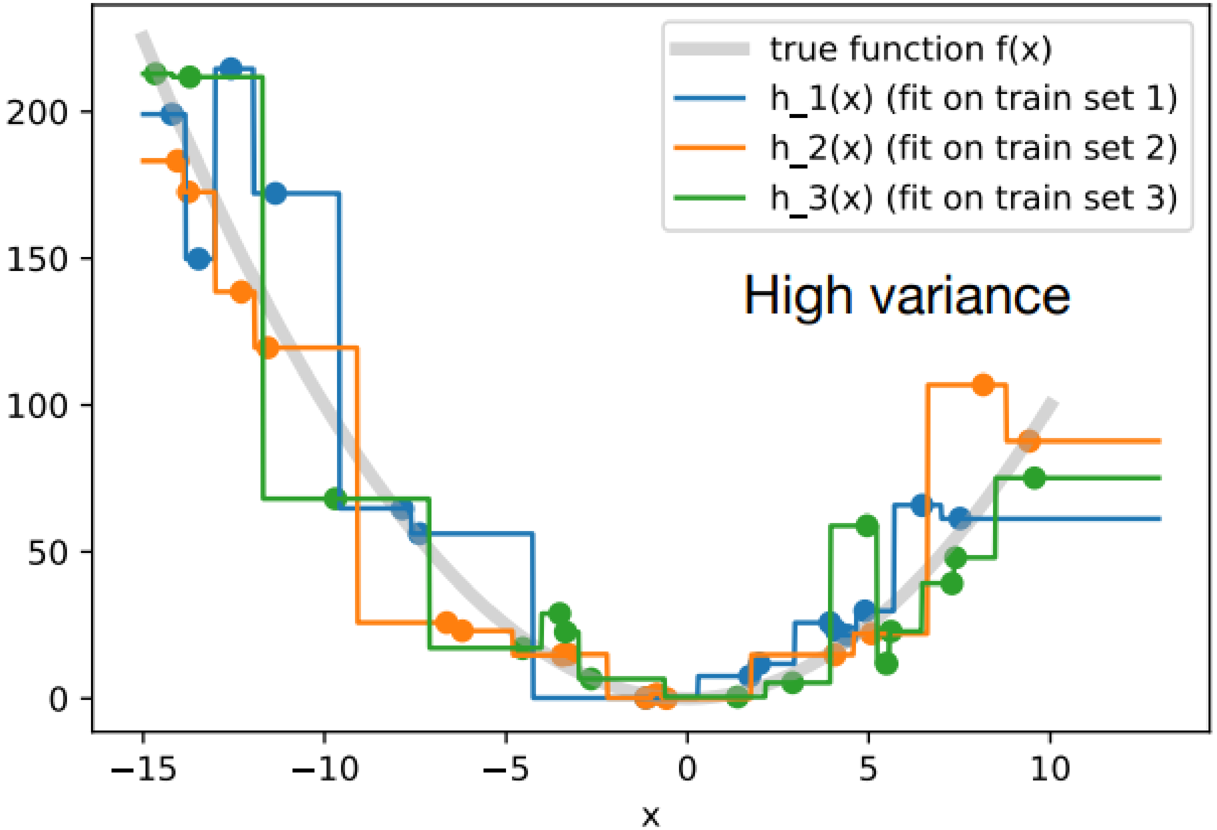
Bias-Variance Intuition

(here, I fit an unpruned decision tree)

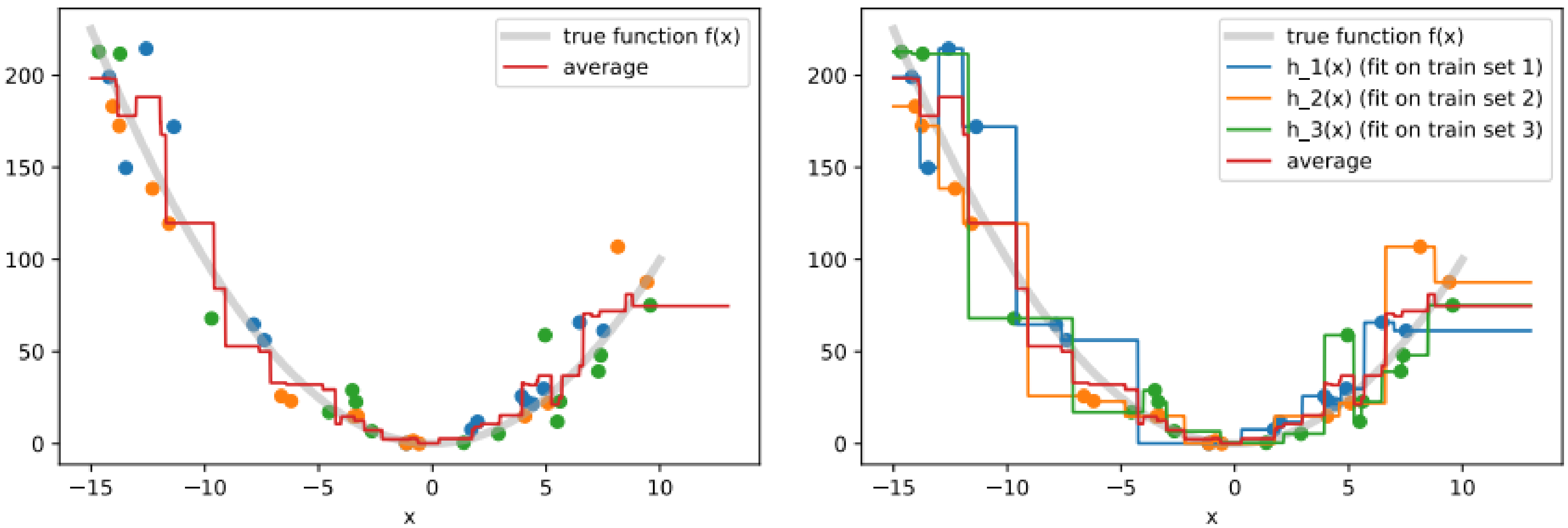


Bias-Variance Intuition

suppose we have multiple training sets



Bias-Variance Intuition



Terminology

Point estimator $\hat{\theta}$ of some parameter θ

(could also be a function, e.g., the hypothesis is

an estimator of some target function)

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

Terminology

General Definition

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

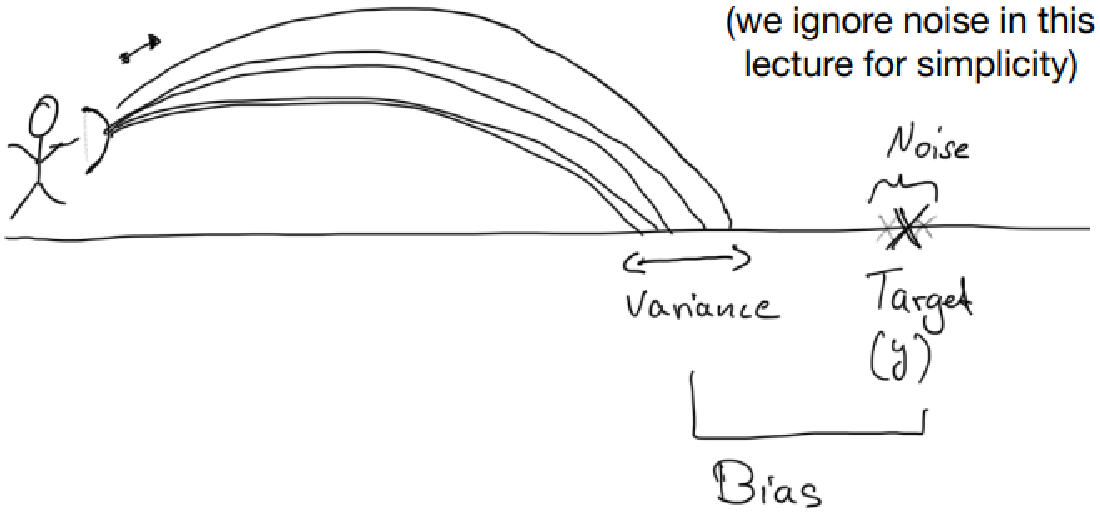
$$\text{Var}[\hat{\theta}] = E[\hat{\theta}^2] - (E[\hat{\theta}])^2$$

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

Terminology

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

Intuition



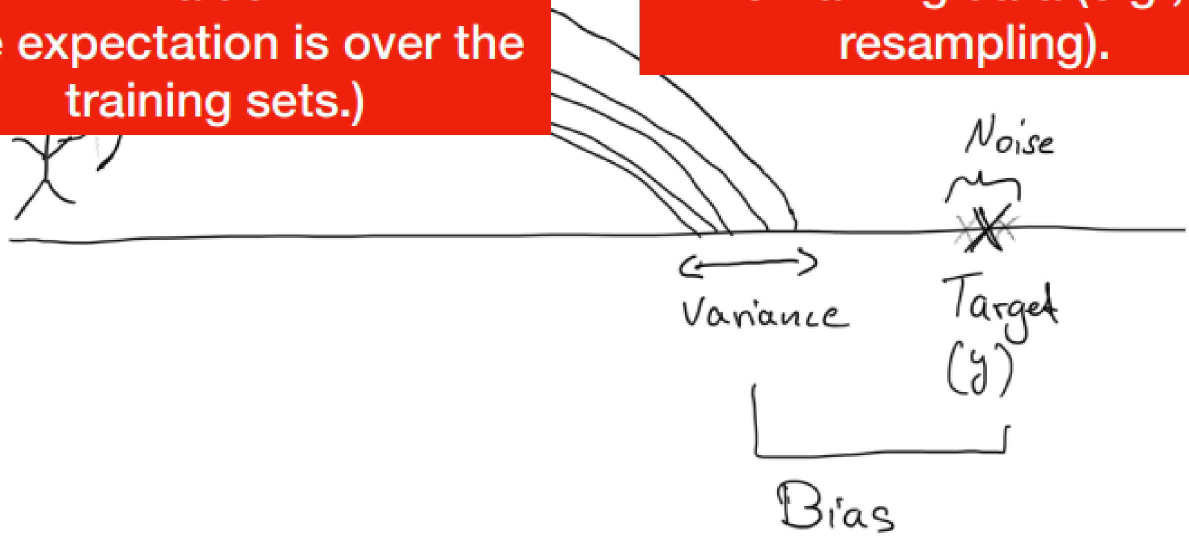
Terminology

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

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

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

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



Bias-Variance of the Squared Error

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

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

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

"ML Notation" for Squared Error Loss

$y = f(x)$ target

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

$S = (y - \hat{y})^2$ squared error

for simplicity, we ignore the noise term

Bias-Variance of the Squared Error

"ML Notation" for Squared Error Loss

$y = f(x)$ target

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

$S = (y - \hat{y})^2$ squared error

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

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

Bias-Variance of the Squared Error

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

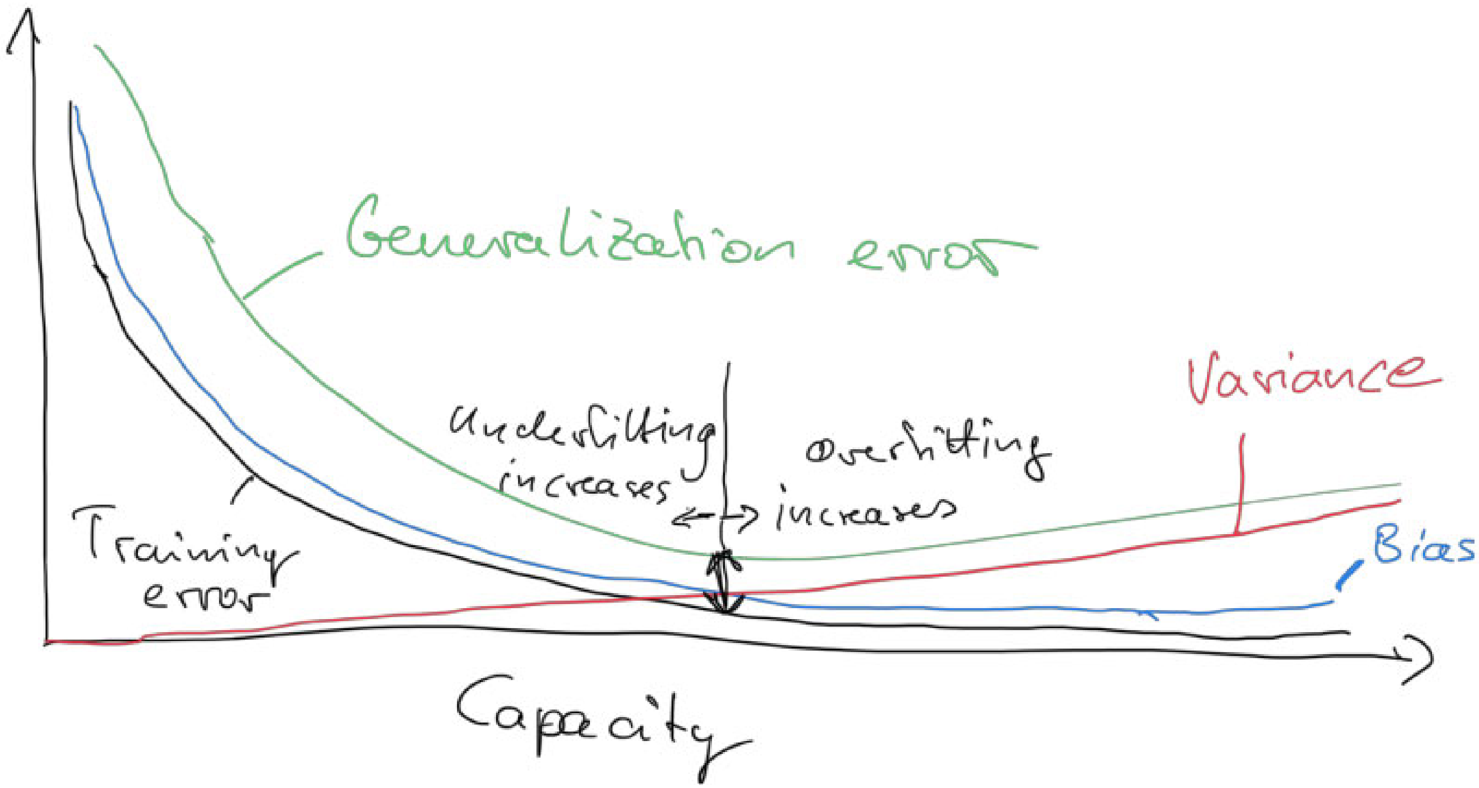
$$\begin{aligned} (y - \hat{y})^2 &= (y - E[\hat{y}] + E[\hat{y}] - \hat{y})^2 \\ &= (y - E[\hat{y}])^2 + (E[\hat{y}] - \hat{y})^2 + 2(y - E[\hat{y}])(E[\hat{y}] - \hat{y}) \end{aligned}$$

$$E[S] = E\left[(y - \hat{y})^2\right]$$

$$E\left[(y - \hat{y})^2\right] = (y - E[\hat{y}])^2 + E\left[(E[\hat{y}] - \hat{y})^2\right]$$

$$= \text{Bias}^2 + \text{Var}$$

Now, how is this related to overfitting and underfitting?


Generalization error
Variance
underfitting increases
overfitting increases
Bias
Training error
Capacity

Overfitting

- Overfitting means the model works well on training data, but it doesn't generalize well.
- Overfitting occurs when there is too much complexity in the model in comparison to the amount and noise in the training data.

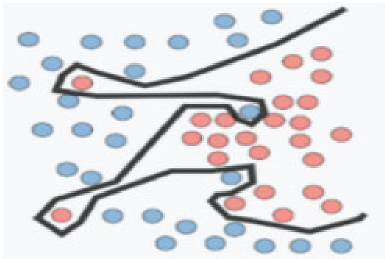


Figure: High Variance, Source

- How to fix this problem?
  - Simplify the model by selecting one with fewer parameters, reducing the number of attributes in the training data, or constraining the model.
  - Gather more training data.
  - Reduce the noise in the training data.

Underfitting

- Underfitting is the opposite of overfitting: it occurs when your model is too simple to learn the underlying structure of the data.

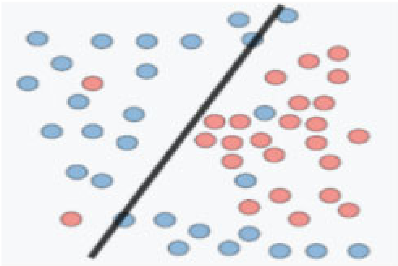


Figure: High Variance, Source

- The main options to fix this problem:
  - Selecting a more powerful model, with more parameters
  - Feeding better features to the learning algorithm (feature engineering)
  - Reducing the constraints on the model (e.g., reducing the regularization hyperparameter)