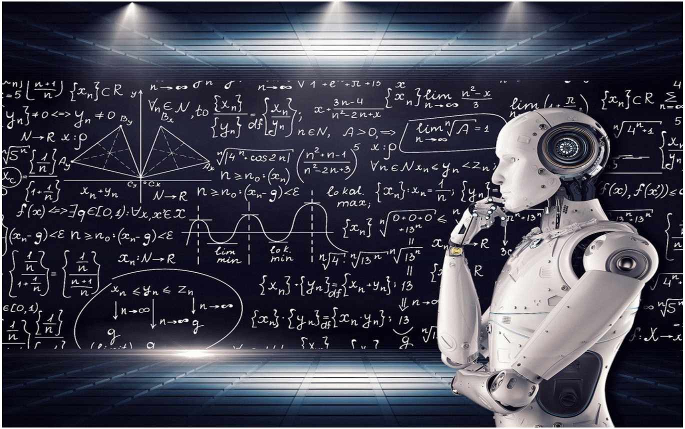
Generalization Error

ML Instruction Team, Fall 2022

CE Department
Sharif University of Technology

Generalization Performance



When can we say the machine has learned?

Assumptions

- Inputs are independent, and training and test examples are identically distributed (i.i.d).
- For some random model that has not been fitted to the training set, we expect both the training and test error to be equal.
- The training error or accuracy provides an (optimistically) biased estimate of the generalization performance.

Terminology

Point estimator θ of some parameter θ

$$\textbf{Bias} = E[\hat{f}] - f \tag{1}$$

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

$$\text{Noise: } \sigma^2 = E\left[\epsilon^2\right] \tag{3}$$

$$\text{target function: } y = \boldsymbol{f}(x) + \epsilon \tag{4}$$

$$\text{predicted target value: } \hat{y} = \hat{f}(x) \tag{5}$$

$$\text{mean squared loss: } MSE = E\left[(y - \hat{y})^2\right] \tag{6}$$

Bias-Variance Decomposition of Squared Error

$$
\begin{aligned}
MSE = E[(y-\hat{y})^2] &= E[(f+\epsilon-\hat{f})^2] && (7)\\
&= E[(f+\epsilon-\hat{f}+E[\hat{f}]-E[\hat{f}])^2] && (8)\\
&= (f-E[\hat{f}])^2 + E[\epsilon^2] + E[(E[\hat{f}]-\hat{f})^2] && (9)\\
&= \textbf{Bias} + \textbf{Var} + \sigma^2 && (10)
\end{aligned}
$$

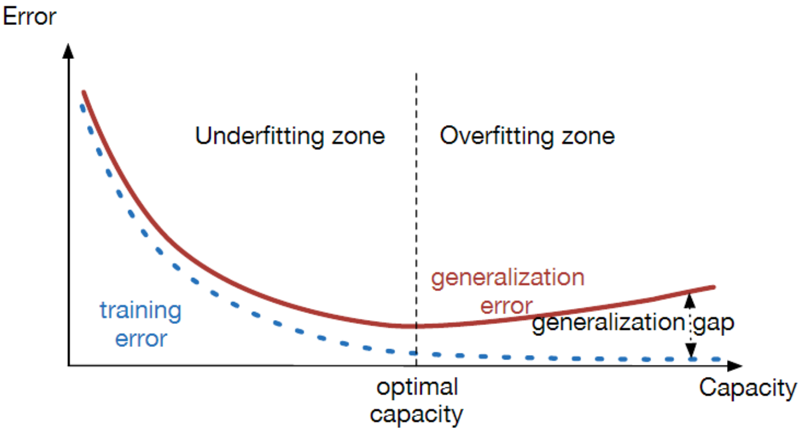
Bias-Variance Decomposition

$$\textbf{Loss} = \textbf{Bias} + \textbf{Variance} + \textbf{Noise}$$

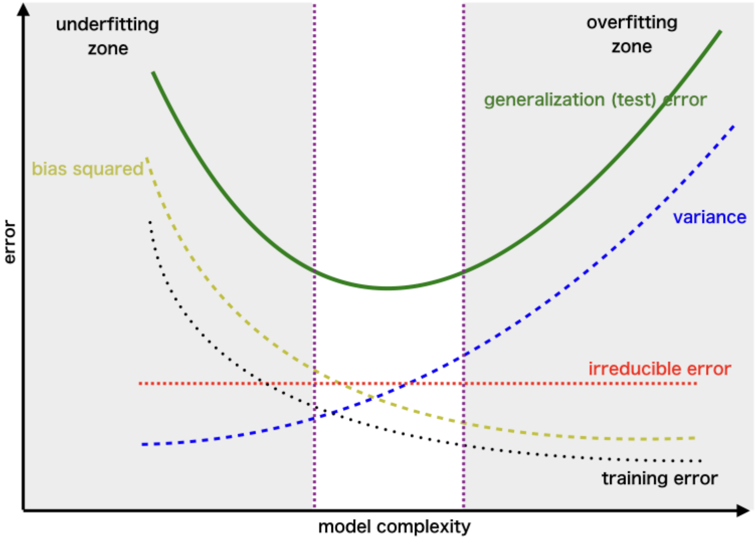
- Decomposition of the loss into bias and variance help us understand learning algorithms, concepts are correlated to underfitting and overfitting
- Helps explain why ensemble methods (last lecture) might perform better than single models

Underfitting VS Overfitting

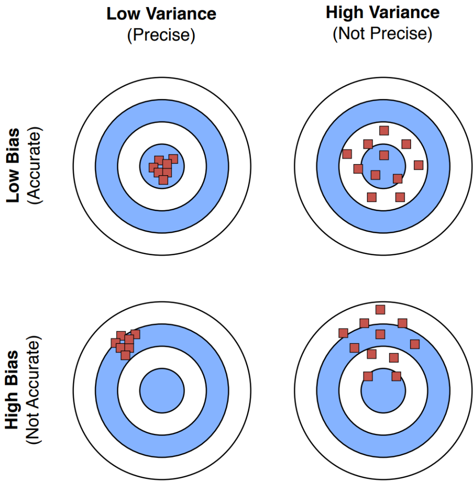
- **Underfitting** : both training and test error are large
- **Overfitting** : gap between training and test error



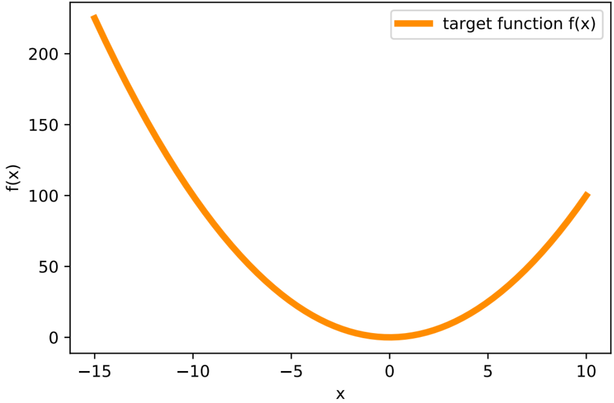
Bias-Variance Trade-off



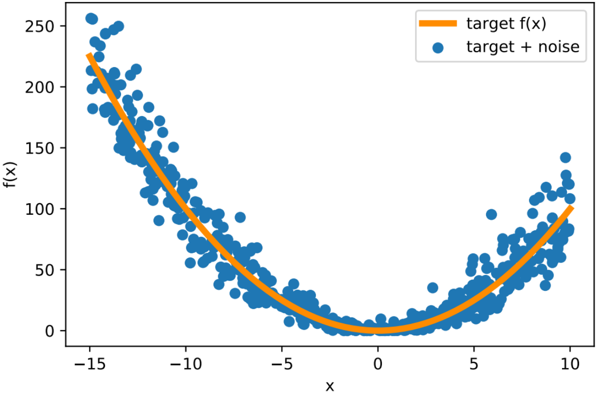
Bias-Variance Trade-off



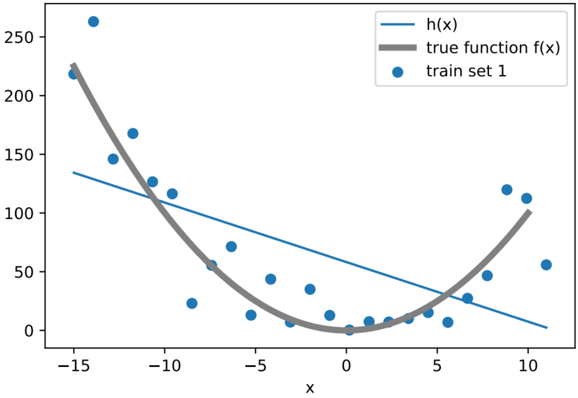
Bias-Variance Trade-off



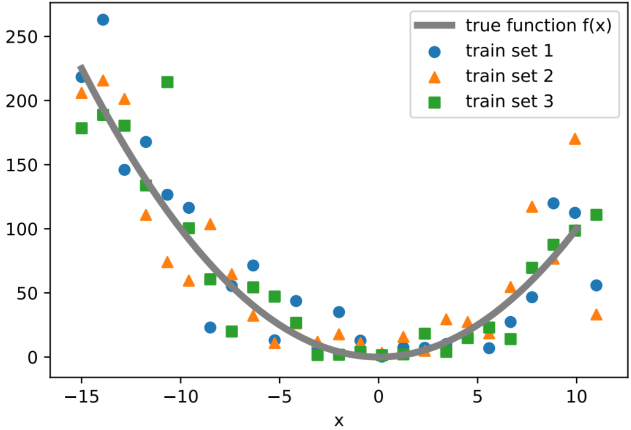
Bias-Variance Trade-off



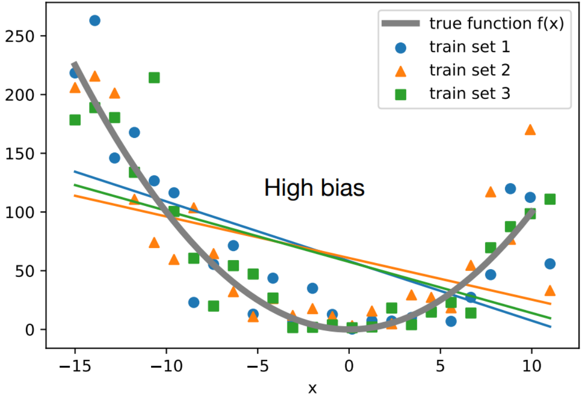
Bias-Variance Trade-off



Bias-Variance Trade-off

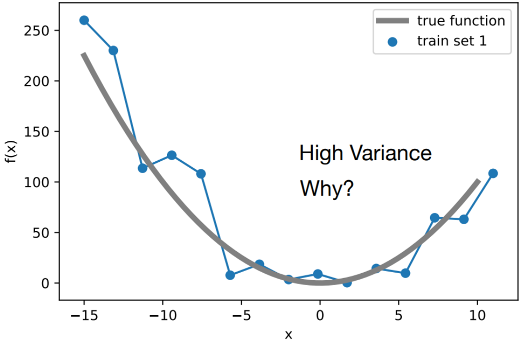


Bias-Variance Trade-off



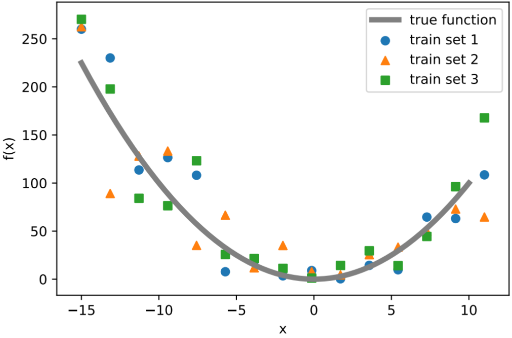
(There are two points where the bias is zero)

Bias-Variance Trade-off



(here, I fit an unpruned decision tree)

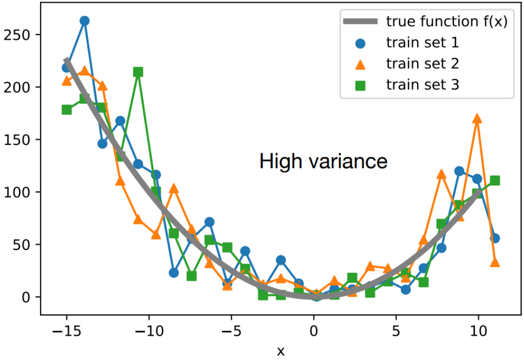
Bias-Variance Trade-off



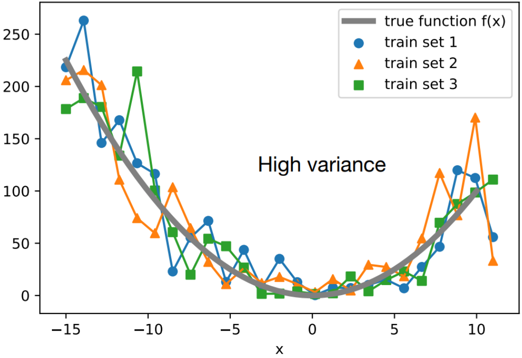
where f(x) is some true (target) function

suppose we have multiple training sets

Bias-Variance Trade-off



Bias-Variance Trade-off



What happens if we take the average?
Does this remind you of something?

Thank You!

Any Question?