CSE 417T
Introduction to Machine Learning

Lecture 6

Instructor: Chien-Ju (CJ) Ho

Recap

Discussion on the VC Bound

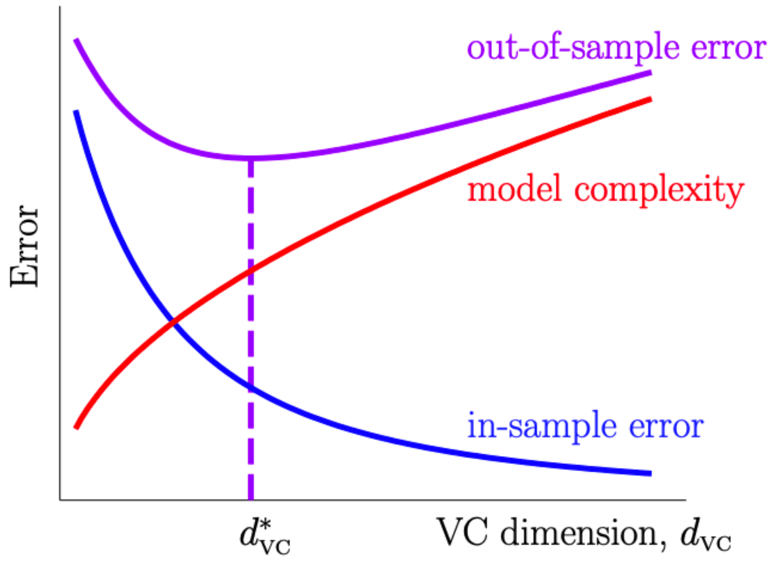
- Think about the high-level tradeoff of choosing d_{VC} and its dependency on N
- The approximation-generalization trade-off

What we want to minimize

How well g generalizes

$$E_{out}(g) \leq E_{in}(g) + O\left(\sqrt{d_{VC}\frac{\ln N}{N}}\right)$$

How well g approximates f in training data



Today’s Lecture

The notes are not intended to be comprehensive. They should be accompanied by lectures and/or textbook.

Let me know if you spot errors.

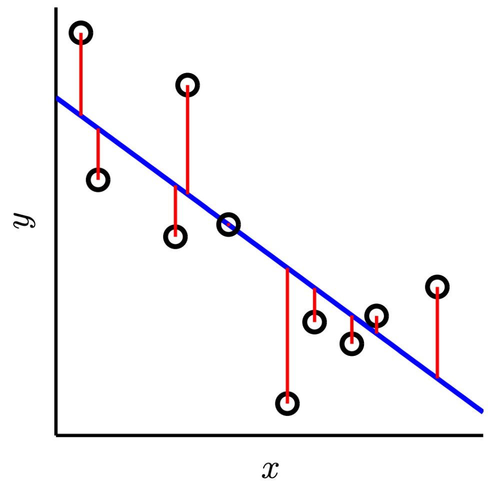
Bias-Variance Decomposition

Another theory of generalization

Real-Value Target and Squared Error

- So far, we focus on binary target function and binary error
  - Binary target function $f(\vec{x}) \in \{-1,1\}$
  - Binary error $e\big(h(\vec{x}), f(\vec{x})\big) = \mathbb{I}[h(\vec{x}) \neq f(\vec{x})]$

- Real-value target functions ["**regression**"] and squared error?
  - Real-value target function $f(\vec{x}) \in \mathbb{R}$
  - Squared error $e\big(h(\vec{x}), f(\vec{x})\big) = \big(h(\vec{x}) - f(\vec{x})\big)^2$



Real-Value Target and Squared Error

- Real-value target functions [called "**regression**"] and squared error?
  - Real-value target function $f(\vec{x}) \in \mathbb{R}$
  - Squared error $e\big(h(\vec{x}), f(\vec{x})\big) = \big(h(\vec{x}) - f(\vec{x})\big)^2$

- Errors:
  - In-sample error: $E_{in}(g) = \frac{1}{N}\sum_{n=1}^{N} e(h(\vec{x}_n), f(\vec{x}_n)) = \frac{1}{N}\sum_{n=1}^{N}\big(h(\vec{x}_n) - f(\vec{x}_n)\big)^2$
  - Out-of-sample error: $E_{out}(g) = \mathbb{E}_{\vec{x}}[e\big(h(\vec{x}), f(\vec{x})\big)] = \mathbb{E}_{\vec{x}}[\big(g(\vec{x}) - f(\vec{x})\big)^2]$

- Theory of generalization: What can we say about $E_{out}(g)$?

- Note that g is learned by some algorithm on the dataset D
  - We'll make the dependency on D explicit and write it as $\boldsymbol{g^{(D)}}$ here.
  - [In VC theory, we consider the worst-case D through the definition of growth function $m_H(N)$]

- $E_{out}\left(g^{(D)}\right) = \mathbb{E}_{\vec{x}}[\left(g^{(D)}(\vec{x}) - f(\vec{x})\right)^2]$
- $\mathbb{E}_D\left[E_{out}\left(g^{(D)}\right)\right]$

$$= \mathbb{E}_D\left[\mathbb{E}_{\vec{x}}\left[\left(g^{(D)}(\vec{x}) - f(\vec{x})\right)^2\right]\right]$$

$$= \mathbb{E}_{\vec{x}}\left[\mathbb{E}_D\left[\left(g^{(D)}(\vec{x}) - \bar{g}(\vec{x}) + \bar{g}(\vec{x}) - f(\vec{x})\right)^2\right]\right]$$

$$= \mathbb{E}_{\vec{x}}\left[\mathbb{E}_D\left[\left(g^{(D)}(\vec{x}) - \bar{g}(\vec{x}) + \bar{g}(\vec{x}) - f(\vec{x})\right)^2\right]\right]$$

$$= \mathbb{E}_{\vec{x}}\left[\mathbb{E}_D\left[\left(g^{(D)}(\vec{x}) - \bar{g}(\vec{x})\right)^2 + \left(\bar{g}(\vec{x}) - f(\vec{x})\right)^2 + 2\left(g^{(D)}(\vec{x}) - \bar{g}(\vec{x})\right)\left(\bar{g}(\vec{x}) - f(\vec{x})\right)\right]\right]$$

Define "expected" hypothesis
$\bar{g}(\vec{x}) = \mathbb{E}_D\left[g^{(D)}(\vec{x})\right]$

- Note that $\mathbb{E}_D\left[\left(g^{(D)}(\vec{x}) - \bar{g}(\vec{x})\right)\left(\bar{g}(\vec{x}) - f(\vec{x})\right)\right] = \left(\bar{g}(\vec{x}) - f(\vec{x})\right)\mathbb{E}_D\left[\left(g^{(D)}(\vec{x}) - \bar{g}(\vec{x})\right)\right] = 0$

Finishing Up

- $\mathbb{E}_D\left[E_{out}\left(g^{(D)}\right)\right]$

$$= \mathbb{E}_{\vec{x}}\left[\mathbb{E}_D\left[\left(g^{(D)}(\vec{x}) - \bar{g}(\vec{x})\right)^2 + \left(\bar{g}(\vec{x}) - f(\vec{x})\right)^2\right]\right]$$

$$= \mathbb{E}_{\vec{x}}\left[\mathbb{E}_D\left[\left(g^{(D)}(\vec{x}) - \bar{g}(\vec{x})\right)^2\right]\right] + \mathbb{E}_{\vec{x}}\left[\left(\bar{g}(\vec{x}) - f(\vec{x})\right)^2\right]$$

$$= \mathbb{E}_{\vec{x}}\left[\text{Variance of } g^{(D)}(\vec{x}) + \text{Bias of } \bar{g}(\vec{x})\right]$$

$$= \text{Variance} + \text{Bias}$$

- Bias-Variance Decomposition

$\bar{g}(\vec{x}) = \mathbb{E}_D\left[g^{(D)}(\vec{x})\right]$

X: a random variable
μ: the mean of X

Variance of X:
$Var(X) = \mathbb{E}[(X-\mu)^2]$

Discussion

- $\mathbb{E}_D\left[E_{out}\left(g^{(D)}\right)\right] = \mathbb{E}_{\vec{x}}\left[\overbrace{\left(\bar{g}(\vec{x}) - f(\vec{x})\right)^2}^{\text{Bias}(\vec{x})}\right] + \mathbb{E}_{\vec{x}}\left[\overbrace{\mathbb{E}_D\left[\left(g^{(D)}(\vec{x}) - \bar{g}(\vec{x})\right)^2\right]}^{\text{Var}(\vec{x})}\right]$

- This is a **conceptual** decomposition
    - Both $\bar{g}$ and f are unknown
    - We can’t really calculate bias and variance in practice

- However, it provides a conceptual guideline in decreasing E_{out}

Example of Bias-Variance Decomposition

- Fitting a sine function
  - $f(x) = \sin(\pi x)$
  - x is drawn uniformly at random from $[0,2]$

- Two hypothesis set
  - H_0: $h(x) = b$
  - H_1: $h(x) = ax + b$

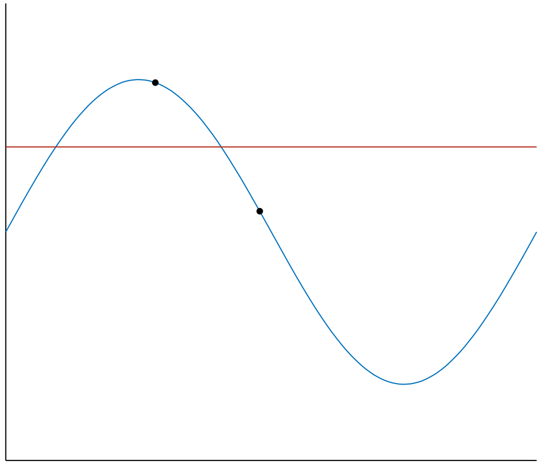
- Assume our algorithm finds g with minimum in-sample error

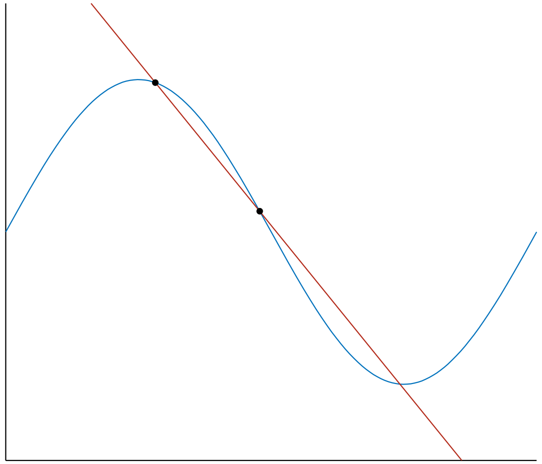
Example of Bias-Variance Decomposition

$H_0: h(x) = b$

$H_1: h(x) = ax + b$

N=2





$$\mathbb{E}_D\left[E_{out}\left(g^{(D)}\right)\right] = \mathbb{E}_{\vec{x}}\left[\underbrace{\left(\bar{g}(\vec{x}) - f(\vec{x})\right)^2}_{\text{Bias}(\vec{x})}\right] + \mathbb{E}_{\vec{x}}\left[\underbrace{\mathbb{E}_D\left[\left(g^{(D)}(\vec{x}) - \bar{g}(\vec{x})\right)^2\right]}_{\text{Var}(\vec{x})}\right]$$

Discussion:

If $N = 2$, would you choose H_0 or H_1? Why?

If $N = 50$, would you choose H_0 or H_1? Why?

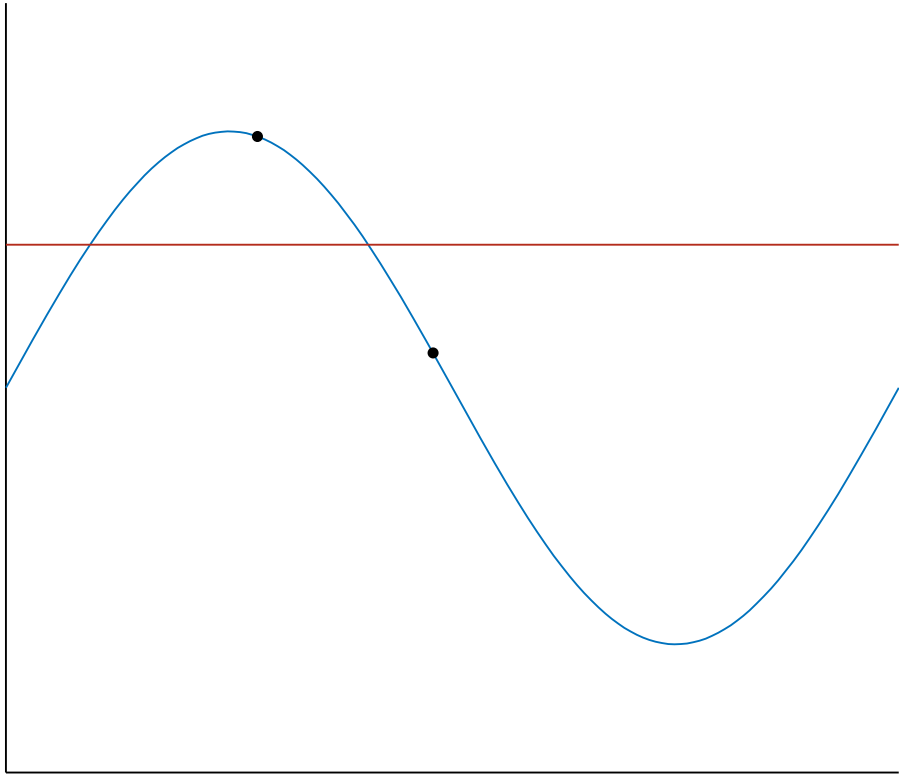
What's the change of biases/variances for H_0/ H_1 from $N = 2$ to $N = 50$.

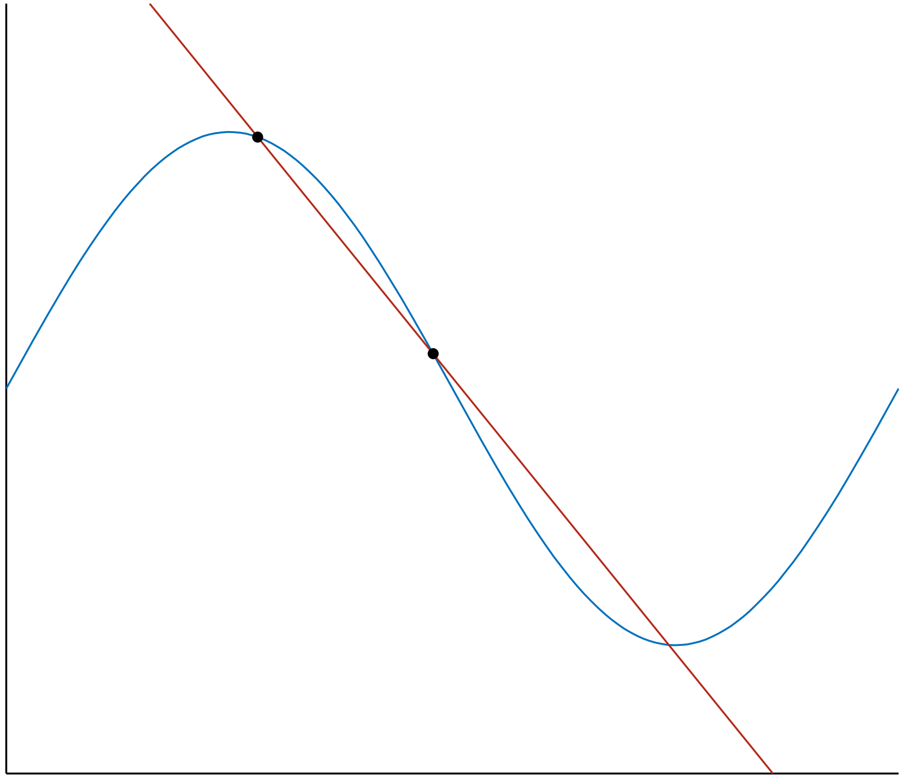
Example of Bias-Variance Decomposition

$H_0: h(x) = b$

$H_1: h(x) = ax + b$

N=2



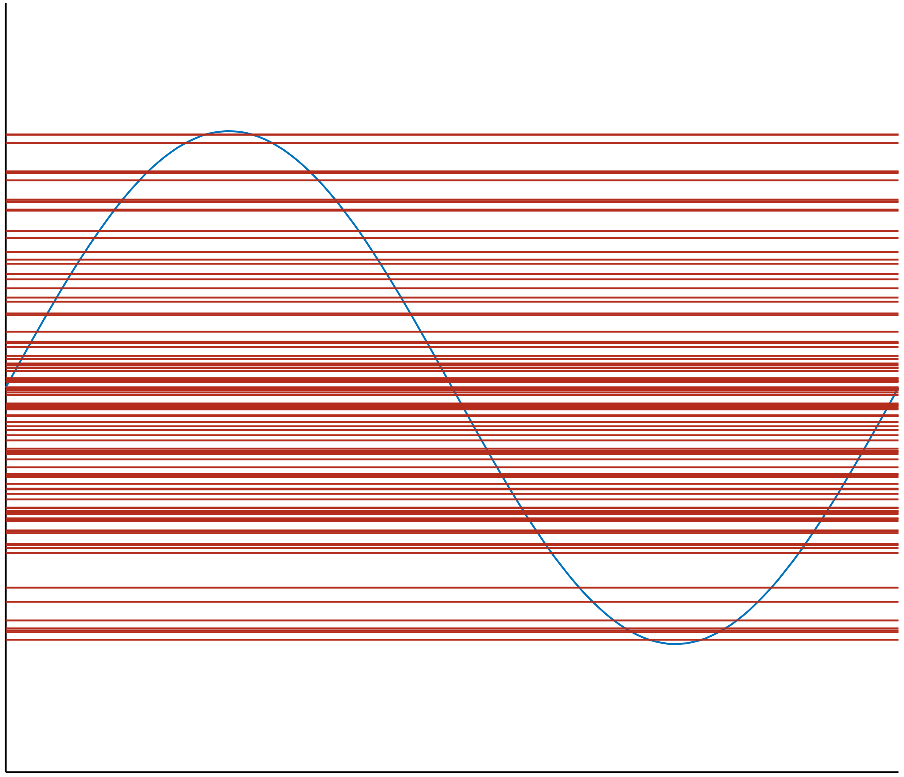


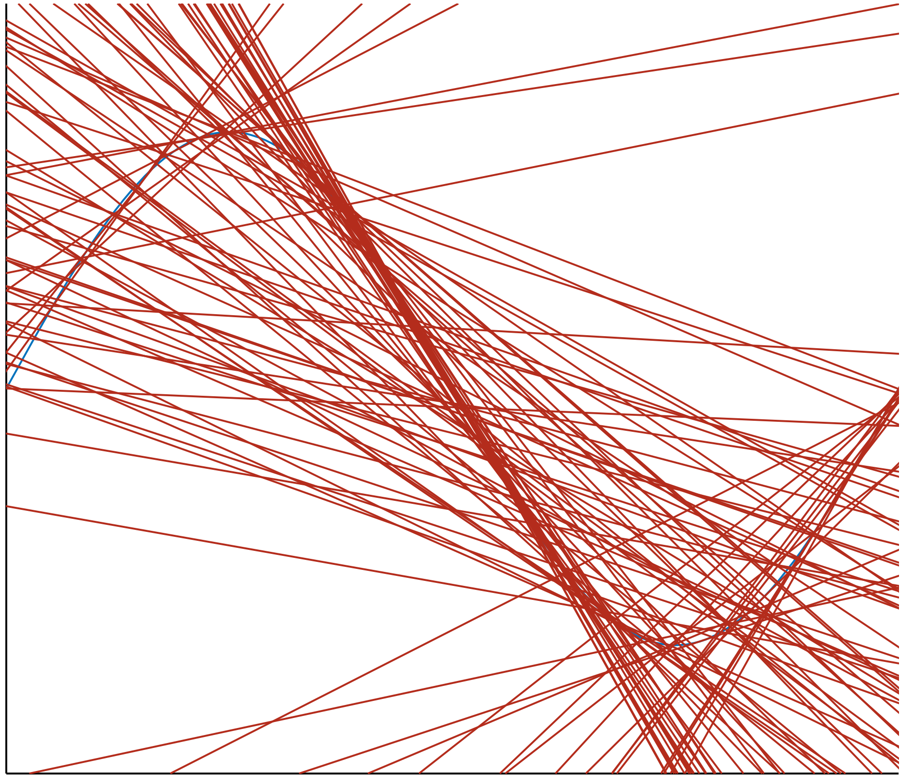
Example of Bias-Variance Decomposition

$H_0: h(x) = b$

$H_1: h(x) = ax + b$

N=2



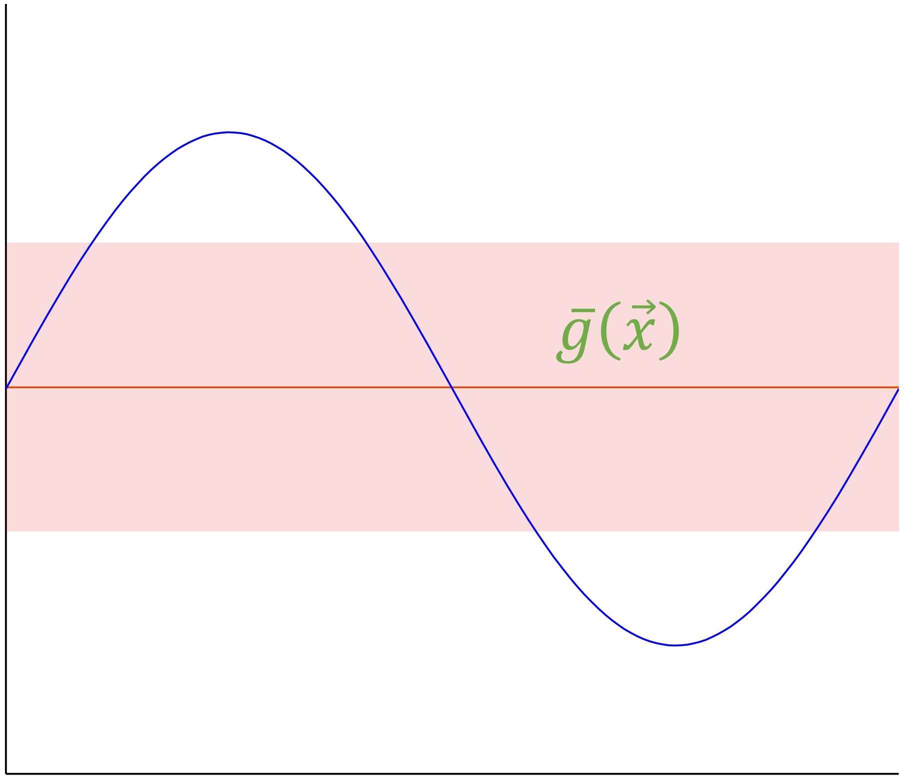


$$\mathbb{E}_D[E_{out}(g^{(D)})] = \mathbb{E}_{\vec{x}}\left[\underbrace{\left(\bar{g}(\vec{x}) - f(\vec{x})\right)^2}_{\text{Bias}(\vec{x})}\right] + \mathbb{E}_{\vec{x}}\left[\underbrace{\mathbb{E}_D\left[\left(g^{(D)}(\vec{x}) - \bar{g}(\vec{x})\right)^2\right]}_{\text{Var}(\vec{x})}\right]$$

Example of Bias-Variance Decomposition

N=2

$H_0: h(x) = b$

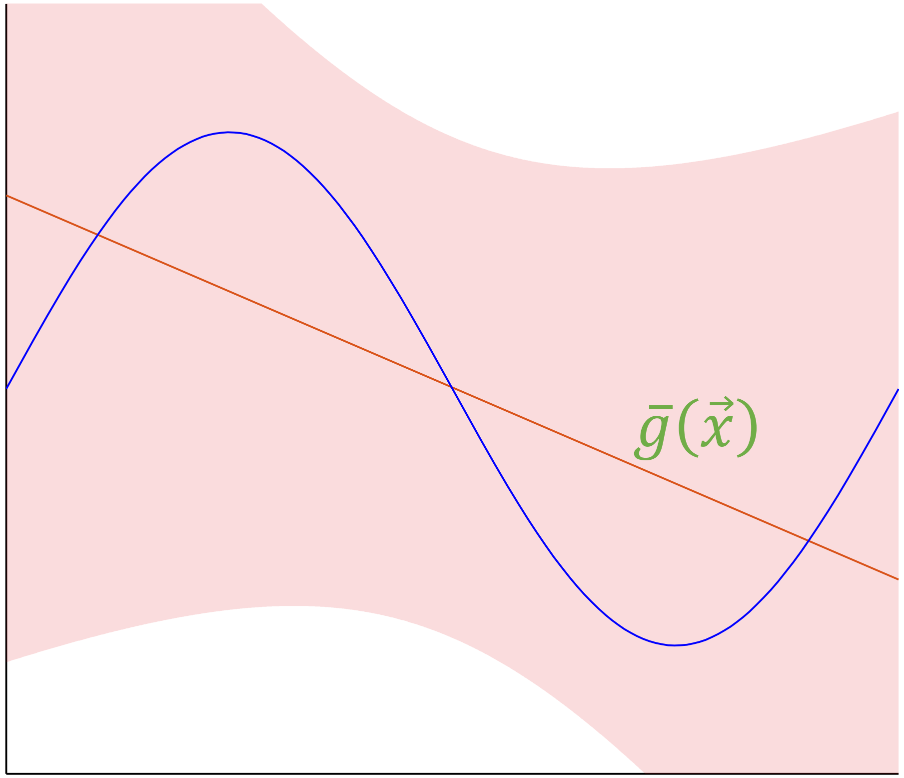


Bias of $\bar{g}(\vec{x}) \approx 0.50$

Variance of $g_{\mathcal{D}}(\vec{x}) \approx 0.25$

$\mathbb{E}_{\mathcal{D}}[E_{out}(g_{\mathcal{D}})] \approx 0.75$

$H_1: h(x) = ax + b$



Bias of $\bar{g}(\vec{x}) \approx 0.21$

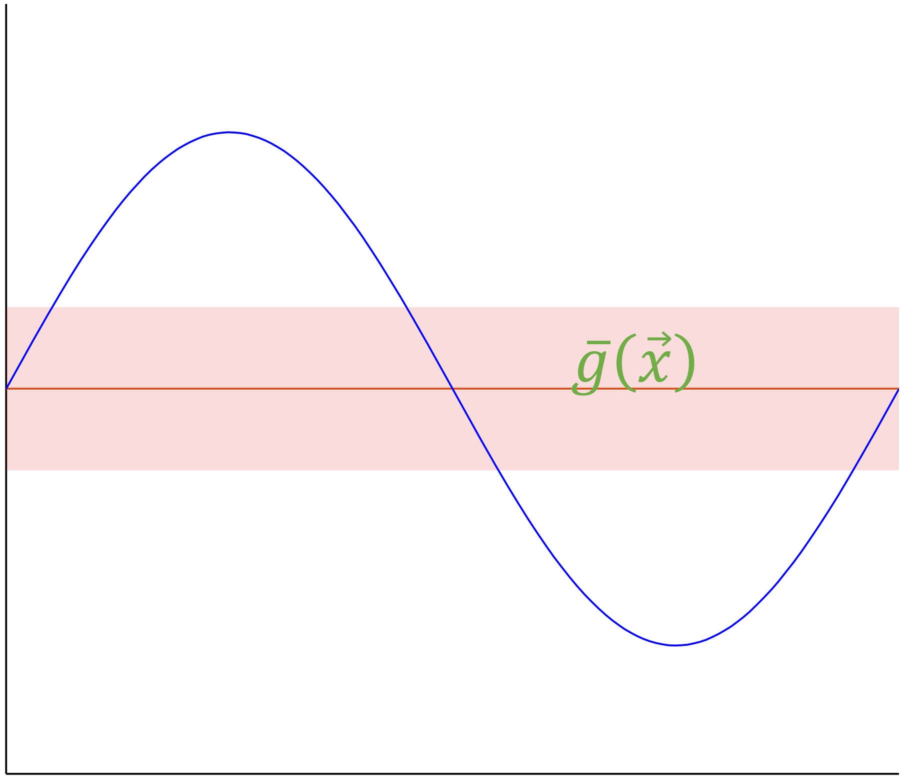
Variance of $g_{\mathcal{D}}(\vec{x}) \approx 1.74$

$\mathbb{E}_{\mathcal{D}}[E_{out}(g_{\mathcal{D}})] \approx 1.95$

$$\mathbb{E}_D[E_{out}(g^{(D)})] = \mathbb{E}_{\vec{x}}\Big[\underbrace{\big(\bar{g}(\vec{x}) - f(\vec{x})\big)^2}_{\text{Bias}(\vec{x})}\Big] + \mathbb{E}_{\vec{x}}\Big[\underbrace{\mathbb{E}_D\Big[\big(g^{(D)}(\vec{x}) - \bar{g}(\vec{x})\big)^2\Big]}_{\text{Var}(\vec{x})}\Big]$$

What if we increase N to 5?

$H_0: h(x) = b$

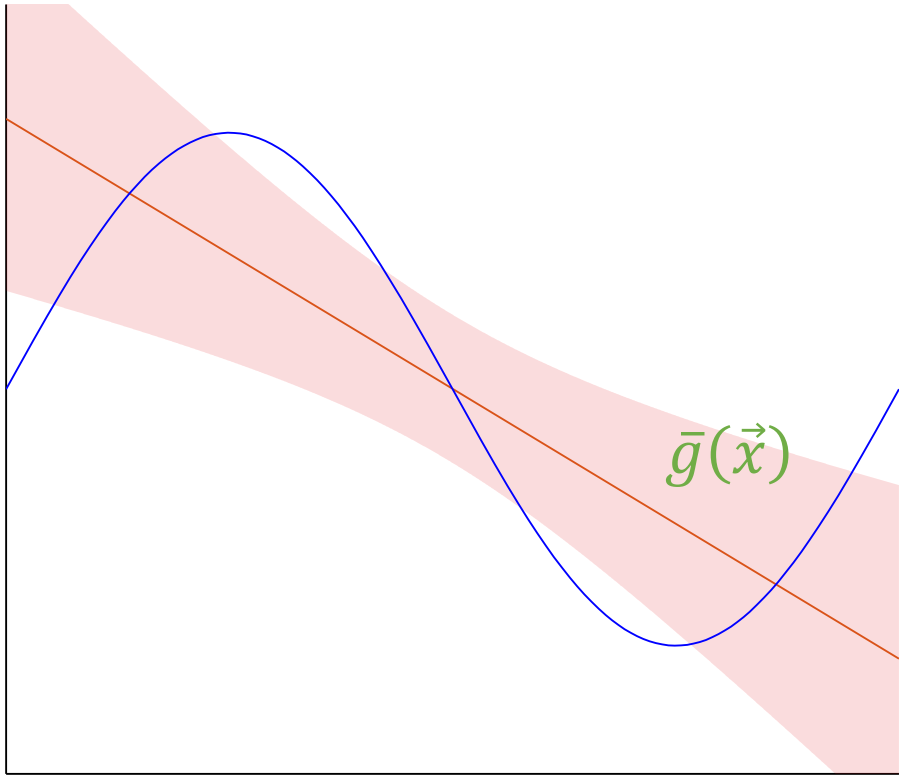


Bias of $\bar{g}(\vec{x}) \approx 0.50$

Variance of $g_{\mathcal{D}}(\vec{x}) \approx 0.10$

$\mathbb{E}_{\mathcal{D}}[E_{out}(g_{\mathcal{D}})] \approx 0.60$

$H_1: h(x) = ax + b$



Bias of $\bar{g}(\vec{x}) \approx 0.21$

Variance of $g_{\mathcal{D}}(\vec{x}) \approx 0.21$

$\mathbb{E}_{\mathcal{D}}[E_{out}(g_{\mathcal{D}})] \approx 0.42$

Discussion

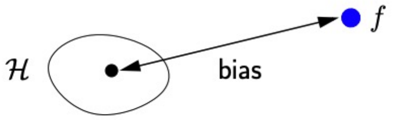
- $\mathbb{E}_D\left[E_{out}\left(g^{(D)}\right)\right] = \mathbb{E}_{\vec{x}}\left[\underbrace{\left(\bar{g}(\vec{x}) - f(\vec{x})\right)^2}_{\text{Bias}(\vec{x})}\right] + \mathbb{E}_{\vec{x}}\left[\underbrace{\mathbb{E}_D\left[\left(g^{(D)}(\vec{x}) - \bar{g}(\vec{x})\right)^2\right]}_{\text{Var}(\vec{x})}\right]$

- Increasing the number of data points N
  - Biases roughly stay the same
  - Variances decrease
  - Expected E_{out} decreases

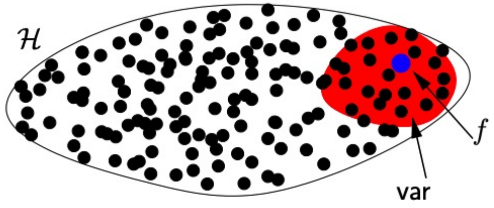
Discussion

- $\mathbb{E}_D\left[E_{out}\left(g^{(D)}\right)\right] = \mathbb{E}_{\vec{x}}\left[\overbrace{\left(\bar{g}(\vec{x}) - f(\vec{x})\right)^2}^{\text{Bias}(\vec{x})}\right] + \mathbb{E}_{\vec{x}}\left[\overbrace{\mathbb{E}_D\left[\left(g^{(D)}(\vec{x}) - \bar{g}(\vec{x})\right)^2\right]}^{\text{Var}(\vec{x})}\right]$

- Increasing the complexity of H
  - Bias goes down (more likely to approximate f)
  - Variance goes up (The stability of $g^{(D)}$ is worse)



Very small model



Very large model

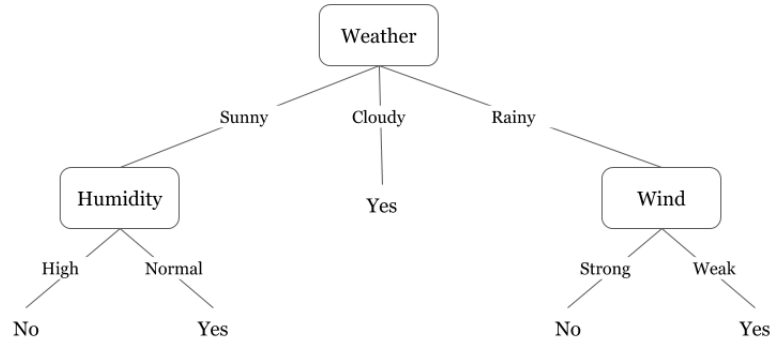
Discussion

- $$\mathbb{E}_D\left[E_{out}\left(g^{(D)}\right)\right] = \mathbb{E}_{\vec{x}}\Big[\underbrace{\left(\bar{g}(\vec{x}) - f(\vec{x})\right)^2}_{\text{Bias}(\vec{x})}\Big] + \mathbb{E}_{\vec{x}}\Big[\underbrace{\mathbb{E}_D\left[\left(g^{(D)}(\vec{x}) - \bar{g}(\vec{x})\right)^2\right]}_{\text{Var}(\vec{x})}\Big]$$

- This is a **conceptual** decomposition
  - Both $\bar{g}$ and f are unknown
  - We can't really calculate bias and variance for practical problems
- However, it provides a conceptual guidelines in decreasing E_{out}

Example

- Will talk about this in details in the 2nd half of the semester
- Decision tree
  - A low bias but high variance hypothesis set
  - Practical performance is not ideal



- Random forest
  - Trying to reduce the variance while not sacrificing bias
  - Idea: Generate many trees randomly and average them