CSE 417T: Introduction to Machine Learning

Lecture 6: Bias-Variance Trade-off

Henry Chai

09/13/18

Recall

- Let $B(n,k) =$ the maximum number of dichotomies on n points s.t. no subset of k points is shattered
- If k is a breakpoint for $\mathcal{H}$, then $m_{\mathcal{H}}(n) \leq B(n,k)$
- If $B(n,k)$ is bounded by a polynomial in n and $m_{\mathcal{H}}(n)$ is bounded by $B(n,k)$, then $m_{\mathcal{H}}(n)$ is bounded by a polynomial in n

Bounding $m_{\mathcal{H}}(n)$

- If k is a breakpoint for $\mathcal{H}$, then $m_{\mathcal{H}}(n) \leq B(n,k)$
- $B(n,k) \leq \sum_{i=0}^{k-1} \binom{n}{i} = \sum_{i=0}^{k-1} \frac{n!}{(n-i)!\,i!} \leq n^{k-1} + 1$
- If k is a breakpoint for $\mathcal{H}$, then $m_{\mathcal{H}}(n) \leq n^{k-1} + 1$

Growth Function: Examples

- For 1D positive rays, $k = 2$ is a break point and $m_{\mathcal{H}}(n) = n + 1 \leq n^{2-1} + 1 = n + 1$
- For 1D positive intervals, $k = 3$ is a break point and $m_{\mathcal{H}}(n) = \frac{n^2}{2} + \frac{n}{2} + 1 \leq n^{3-1} + 1 = n^2 + 1$
- For 2D linear separators, $k = 4$ is a break point and $m_{\mathcal{H}}(3) = 8 \leq 3^{4-1} + 1 = 28$

Growth Function: Examples

- For 1D positive rays, $k = 2$ is a break point and $m_{\mathcal{H}}(n) = n + 1 \leq n^{2-1} + 1 = n + 1$
- For 1D positive intervals, $k = 3$ is a break point and $m_{\mathcal{H}}(n) = \frac{n^2}{2} + \frac{n}{2} + 1 \leq n^{3-1} + 1 = n^2 + 1$
- For 2D linear separators, $k = 4$ is a break point and $m_{\mathcal{H}}(4) = 14 \leq 4^{4-1} + 1 = 65$

VC-Dimension

- $d_{VC}(\mathcal{H}) =$ the largest value of n s.t. $m_{\mathcal{H}}(n) = 2^n$
- The VC-dimension is the greatest number of points that can be shattered by $\mathcal{H}$
- If k^* is the smallest breakpoint for $\mathcal{H}$, then $d_{VC}(\mathcal{H}) = k^* - 1$
- $m_{\mathcal{H}}(n) \leq n^{k^*-1} + 1$

VC-Dimension

- $d_{VC}(\mathcal{H}) =$ the largest value of n s.t. $m_{\mathcal{H}}(n) = 2^n$
- The VC-dimension is the greatest number of points that can be shattered by $\mathcal{H}$
- If k^* is the smallest breakpoint for $\mathcal{H}$, then $d_{VC}(\mathcal{H}) = k^* - 1$
- $m_{\mathcal{H}}(n) \leq n^{d_{VC}} + 1$
- $E_{out}(g) \leq E_{in}(g) + O\left(\sqrt{d_{VC}\frac{\log(n)}{n}}\right)$

Sample Complexity

- How many samples do we need in our training data to say that the generalization error is less than ϵ with probability at least $1 - \delta$?
- Set $\sqrt{\frac{8}{n}\log\left(\frac{4\left(1+(2n)^{d_{VC}}\right)}{\delta}\right)} \leq \epsilon$
- Conclude that we need $n \geq \frac{8}{\epsilon^2}\log\left(\frac{4\left(1+(2n)^{d_{VC}}\right)}{\delta}\right)$
- Practical rule of thumb: $n \geq 10 d_{VC}$

Penalty for Model Complexity

- Given n samples, how good can we say our learned hypothesis will do with confidence at least $1 - \delta$?

- Conclude that $E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{n} \log\left(\frac{4\left((2n)^{dvc}+1\right)}{\delta}\right)}$

Approximation Generalization Tradeoff

How well does g generalize?

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

How well does g approximate f?

Approximation Generalization Tradeoff

How well does g generalize?

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

Decreases as d_{VC} increases

Approximation Generalization Tradeoff

Increases as d_{VC} increases

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

Decreases as d_{VC} increases

Bias-Variance Tradeoff

- Regression with squared error: $\mathcal{Y} = \mathbb{R}$ and $e(h, f, \vec{x}) = \left(f(\vec{x}) - h(\vec{x})\right)^2$
- $g_{\mathcal{D}} \in \mathcal{H} =$ the hypothesis returned by $\mathcal{A}$ when the input training data is $\mathcal{D}$

Bias-Variance Tradeoff

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

- $$\begin{aligned}\mathbb{E}_{\mathcal{D}}[E_{out}(g_{\mathcal{D}})] &= \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\vec{x}}\left[\left(g_{\mathcal{D}}(\vec{x}) - f(\vec{x})\right)^2\right]\right] \\ &= \mathbb{E}_{\vec{x}}\left[\mathbb{E}_{\mathcal{D}}\left[\left(g_{\mathcal{D}}(\vec{x}) - f(\vec{x})\right)^2\right]\right] \\ &= \mathbb{E}_{\vec{x}}\left[\mathbb{E}_{\mathcal{D}}[g_{\mathcal{D}}(\vec{x})^2 - 2g_{\mathcal{D}}(\vec{x})f(\vec{x}) + f(\vec{x})^2]\right] \\ &= \mathbb{E}_{\vec{x}}[\mathbb{E}_{\mathcal{D}}[g_{\mathcal{D}}(\vec{x})^2] - 2\bar{g}(\vec{x})f(\vec{x}) + f(\vec{x})^2]\end{aligned}$$

- where $\bar{g}(\vec{x}) = \mathbb{E}_{\mathcal{D}}[g_{\mathcal{D}}(\vec{x})] \approx \frac{1}{k}\sum_{i=1}^{k} g_{\mathcal{D}_i}(\vec{x})$

Bias-Variance Tradeoff

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

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

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

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

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

Bias-Variance Tradeoff

How variable is g?

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

How well, on average, does g approximate f?

Bias-Variance Tradeoff

How well could g approximate anything?

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

How well, on average, does g approximate f?

Bias-Variance Tradeoff

How well could g approximate noise?

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

How well, on average, does g approximate f?

Bias-Variance Tradeoff

How well could g approximate noise?

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

Decreases as $\mathcal{H}$ becomes more complex

Bias-Variance Tradeoff

Increases as $\mathcal{H}$ becomes more complex

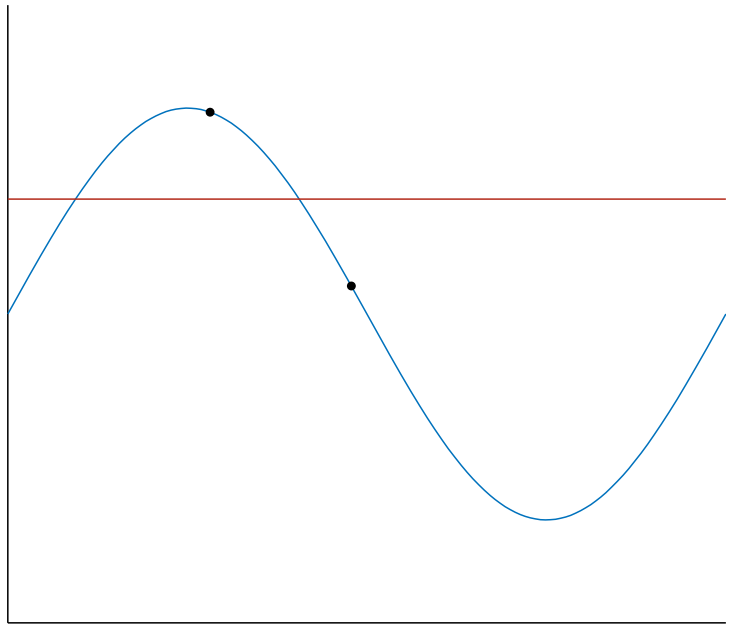
$$\mathbb{E}_{\mathcal{D}}[E_{out}(g_{\mathcal{D}})] = \mathbb{E}_{\vec{x}}\left[\mathbb{E}_{\mathcal{D}}[g_{\mathcal{D}}(\vec{x})^2 - \bar{g}(\vec{x})^2] + \left(\bar{g}(\vec{x}) - f(\vec{x})\right)^2\right]$$

Decreases as $\mathcal{H}$ becomes more complex

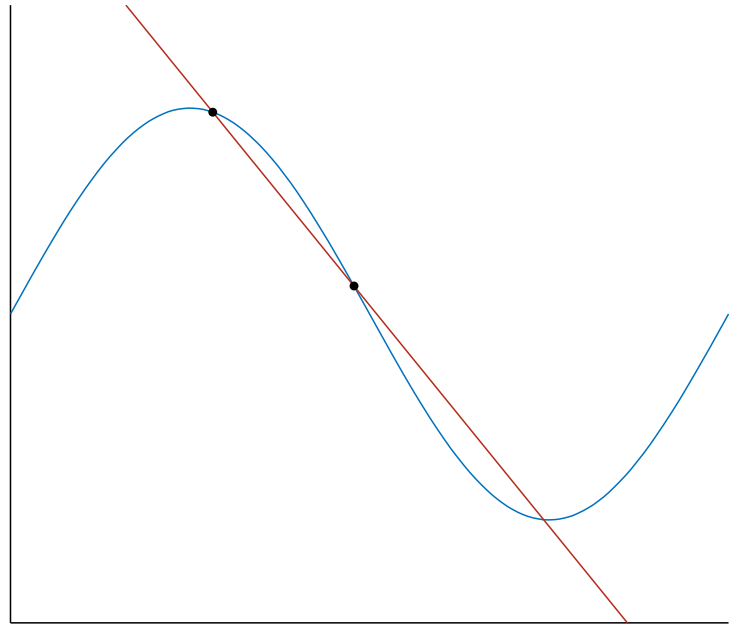
Bias-Variance Tradeoff (Example)

- $\mathcal{X} = \mathbb{R}$ and $\mathcal{P} = \text{Uniform}(0, 2\pi)$
- $y = \sin(x)$
- $\mathcal{D} = \{(x_1, \sin(x_1)), (x_2, \sin(x_2))\}$
- $\mathcal{H}_0 = \{h : h(x) = b\}$ and $\mathcal{H}_1 = \{h : h(x) = ax + b\}$

Bias-Variance Tradeoff (Example)

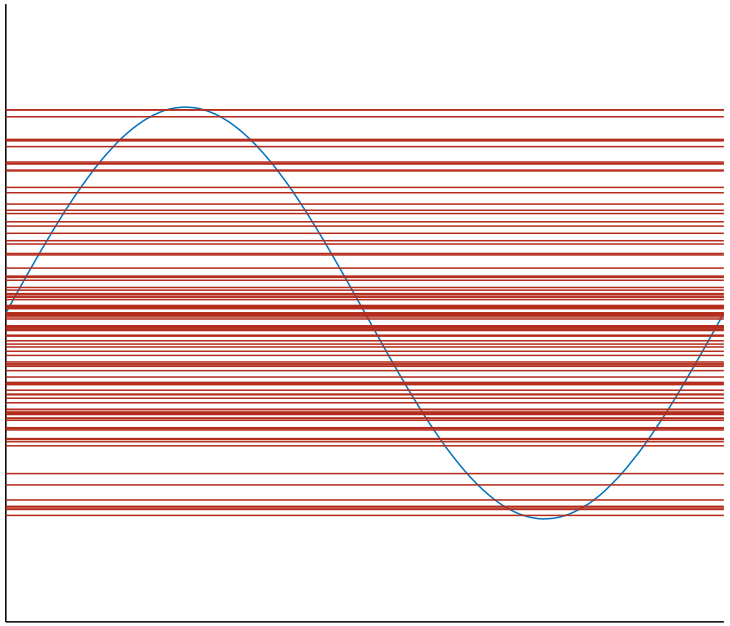


$\mathcal{H}_0$

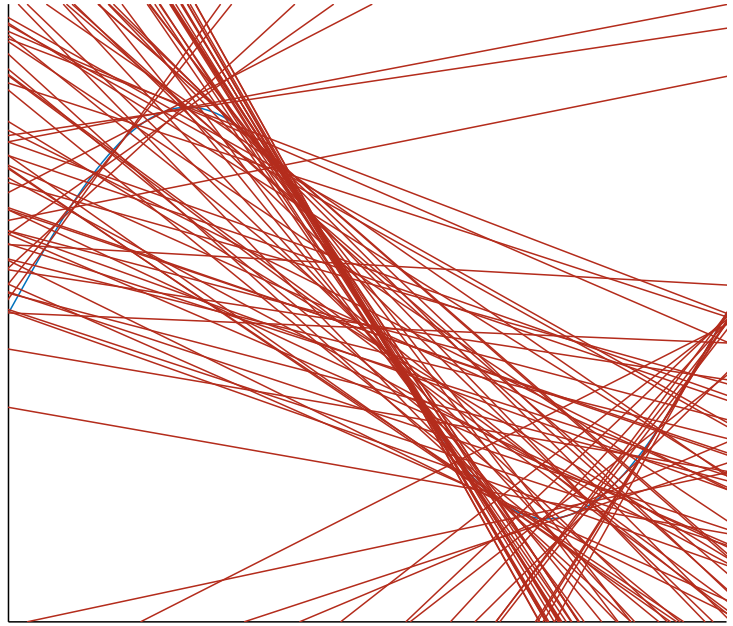


$\mathcal{H}_1$

Bias-Variance Tradeoff (Example)

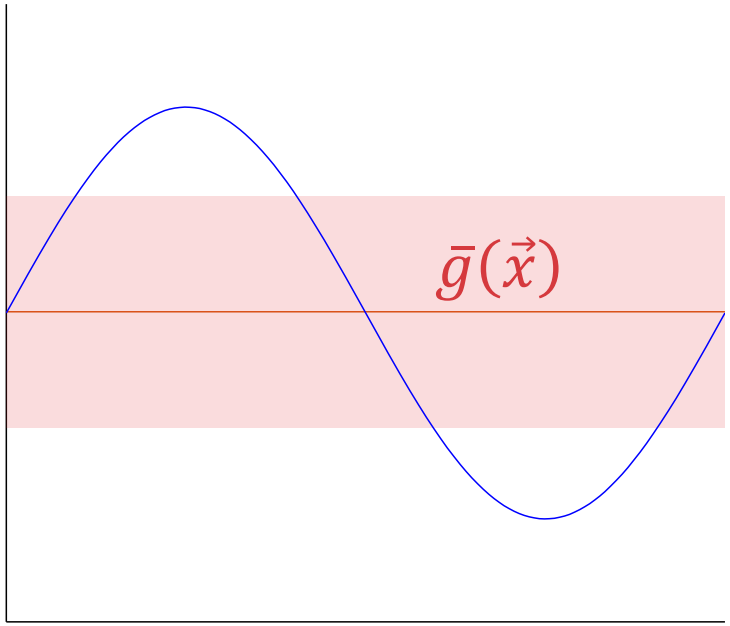


$\mathcal{H}_0$

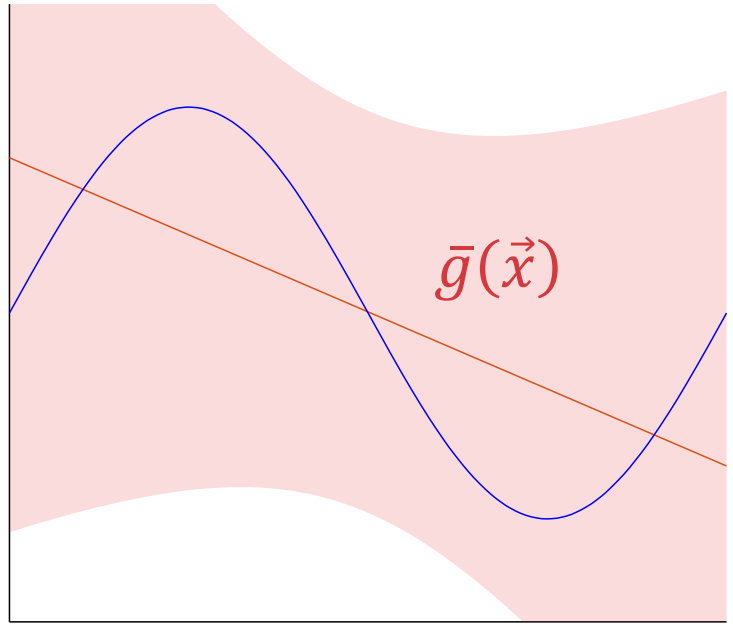


$\mathcal{H}_1$

Bias-Variance Tradeoff (Example)

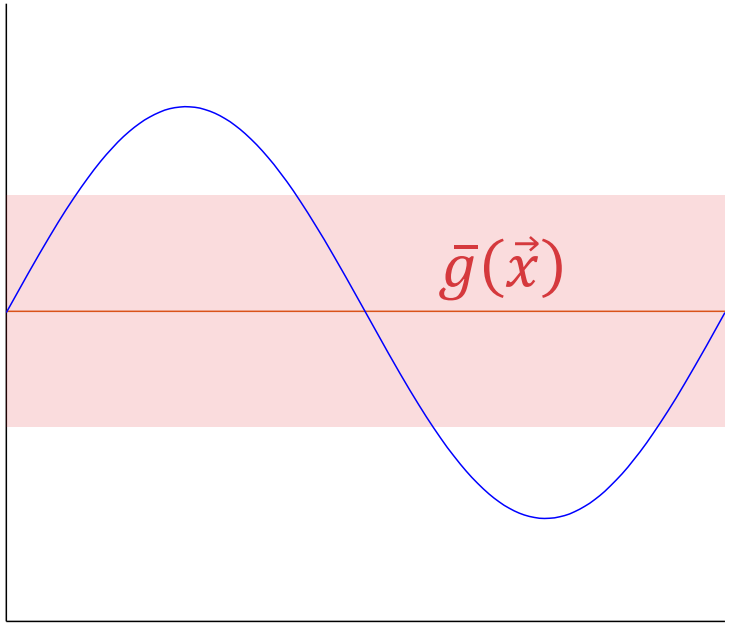


$\mathcal{H}_0$



$\mathcal{H}_1$

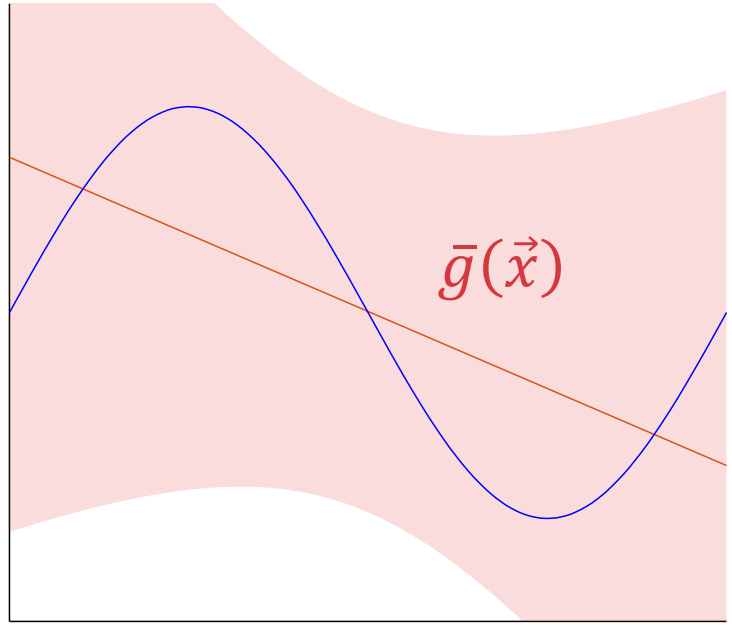
Bias-Variance Tradeoff (Example)



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

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

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

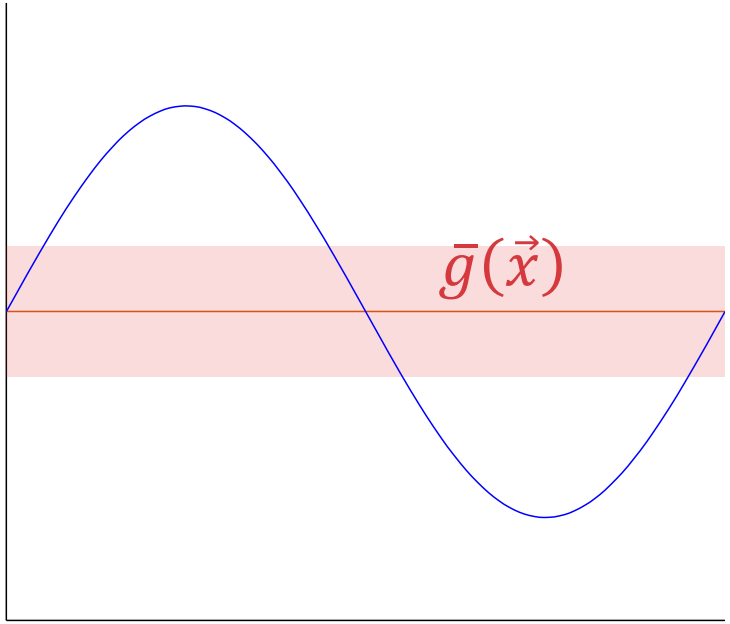


Bias of $\bar{g}(\vec{x}) \approx 0.21$

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

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

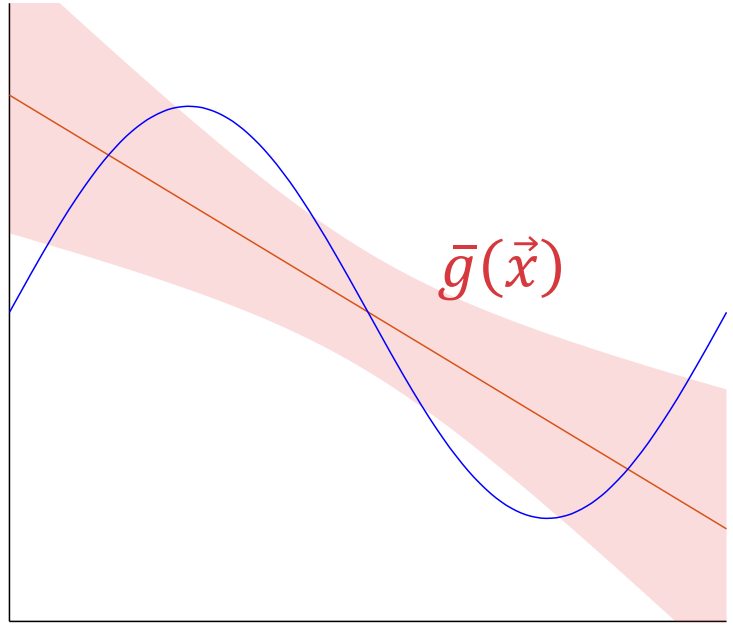
Bias-Variance Tradeoff (Example)



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

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

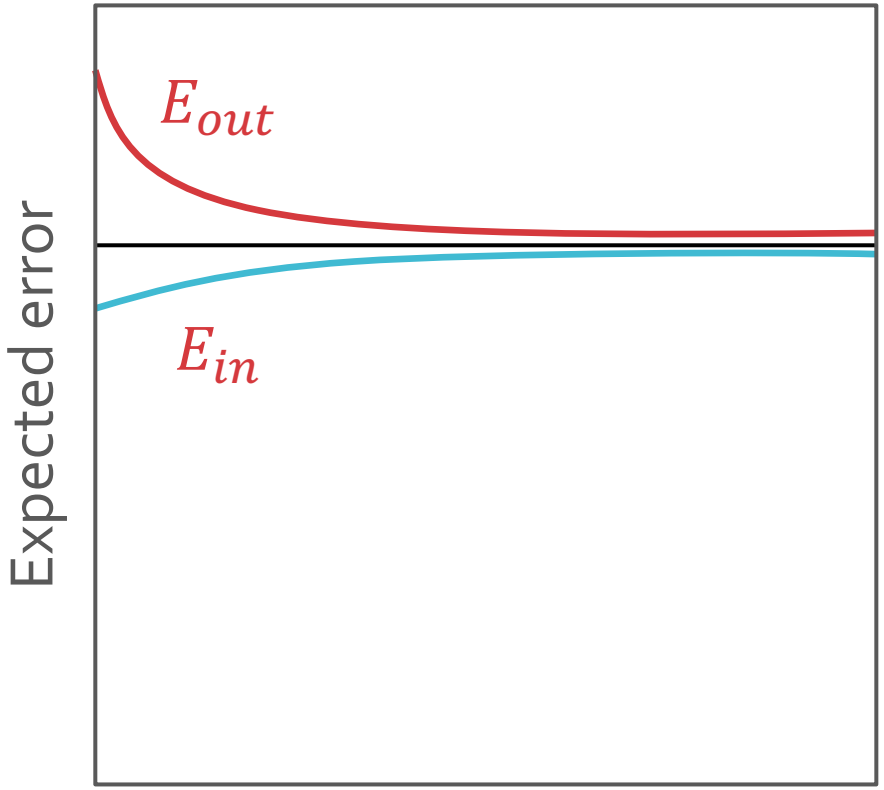
$\mathbb{E}_{\mathcal{D}}[E_{out}(g_{\mathcal{D}})] \approx 0.60$

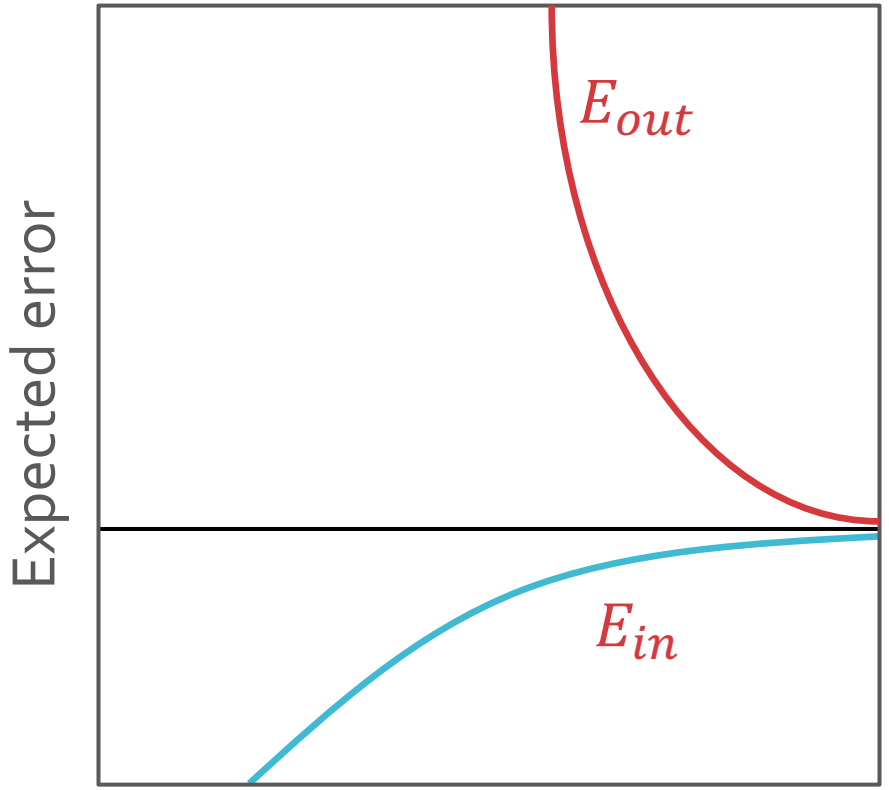


Bias of $\bar{g}(\vec{x}) \approx 0.21$

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

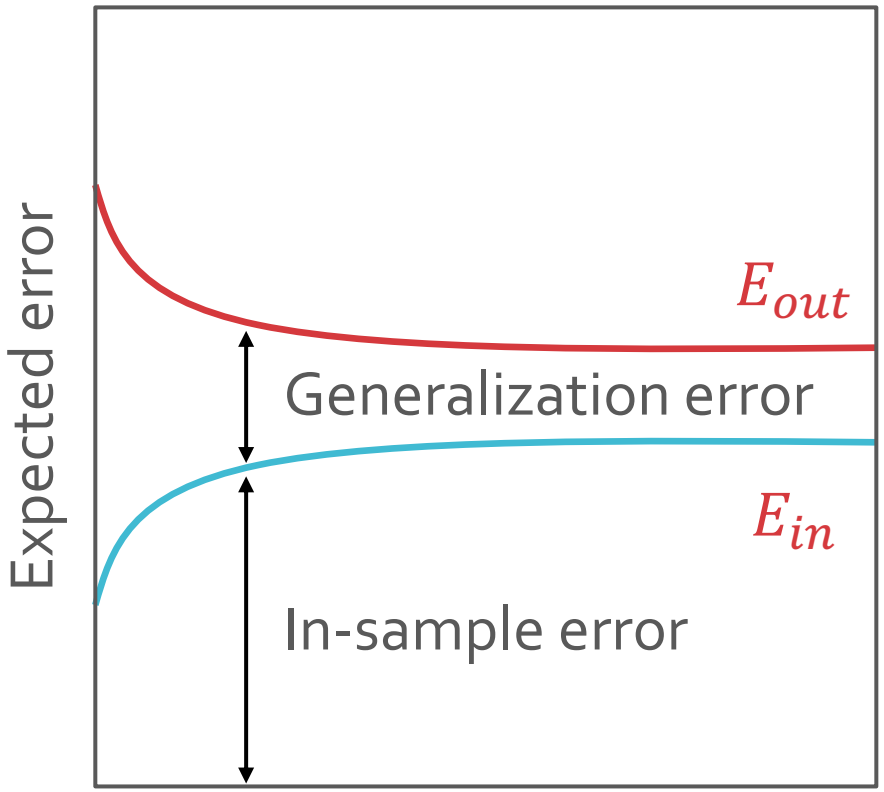
$\mathbb{E}_{\mathcal{D}}[E_{out}(g_{\mathcal{D}})] \approx 0.42$

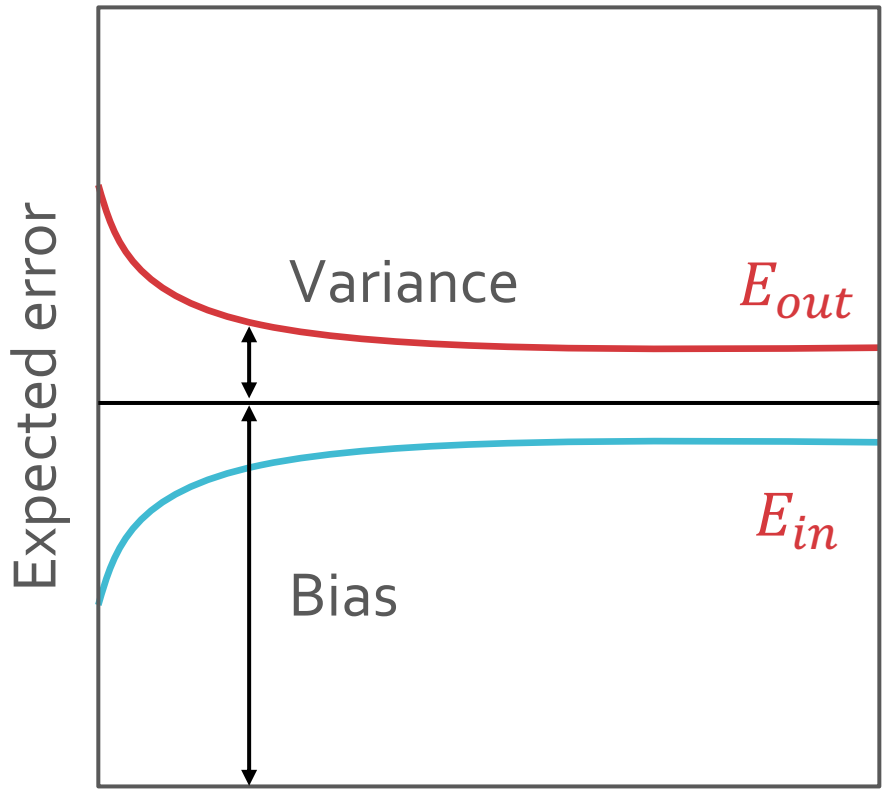




Simple model

Complex model





VC analysis

Bias-Variance analysis

Why $2n$?

$$P\{|E_{in}(g) - E_{out}(g)| > \epsilon\} \leq 4m_{\mathcal{H}}(2n)e^{-\frac{1}{8}\epsilon^2 n}$$

$$\Updownarrow$$

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{n}\log\left(\frac{4m_{\mathcal{H}}(2n)}{\delta}\right)}$$

with probability at least $1 - \delta$

Why $2n$?

- Intuition: $E_{out}(g)$ is difficult to reason about
- Replace $E_{out}(g)$ with $E'_{in}(g)$, the error on a second dataset of size n not used in the training process
- $P\{|E_{in}(g) - E_{out}(g)| > \epsilon\} \leq 2P\{|E_{in}(g) - E'_{in}(g)| > \epsilon\}$

Test Sets

- Instead of bounding $E_{out}(g)$ using $E_{in}(g)$, estimate $E_{out}(g)$ using the error on some test dataset $\mathcal{D}'$
  - $E_{test}(g) =$ error on the test dataset
- If the $\mathcal{D}'$ is not involved in the training process, then we are validating g using $\mathcal{D}'$
  - Therefore, Hoeffding's bound applies!
  - Even better, Hoeffding's bound applies with $m = |\mathcal{H}| = 1$
  - $P\{|E_{test}(g) - E_{out}(g)| > \epsilon\} \leq 2e^{-2\epsilon^2 n'}$ where $n' = |\mathcal{D}'|$

Test Sets

- But at what cost?
- We are given a finite pool of data
- Carving out a test dataset to bound $E_{out}(g)$ leaves fewer data points to train with
- A smaller training dataset generally means the learned g is worse i.e. $E_{test}(g)$ is large
- Practical rule of thumb: 70-80% training, 20-30% testing