ECE595 / STAT598: Machine Learning I
Lecture 29 Bias and Variance

Spring 2020

Stanley Chan

School of Electrical and Computer Engineering
Purdue University

PURDUE UNIVERSITY

Outline

- Lecture 28 Sample and Model Complexity
- Lecture 29 Bias and Variance
- Lecture 30 Overfit

Today's Lecture:

- From VC Analysis to Bias-Variance
  - Generalization Bound
  - Bias-Variance Decomposition
  - Interpreting Bias-Variance
- Example
  - 0-th order vs 1-st order model
  - Trade off

Generalizing the Generalization Bound

Theorem (Generalization Bound)

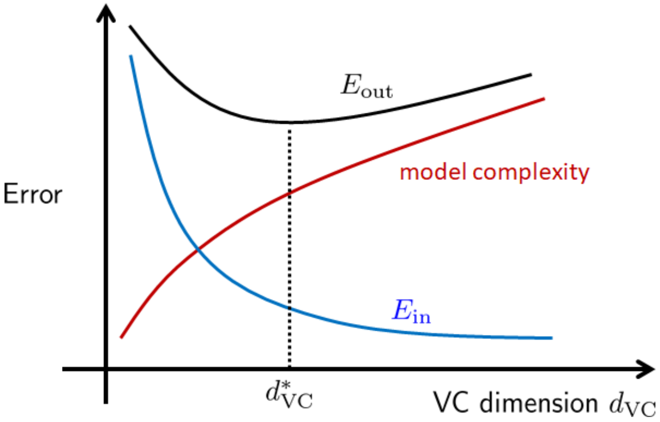
For any tolerance $\delta > 0$

$$E_{\text{out}}(g) \leq E_{\text{in}}(g) + \sqrt{\frac{\mathbf{8}}{N} \log \frac{\mathbf{4}m_{\mathcal{H}}(\mathbf{2N})}{\delta}},$$

with probability at least $1 - \delta$.

- g: final hypothesis
- $m_{\mathcal{H}}(N)$: how complex is your model
- d_{VC}: parameter defining $m_{\mathcal{H}}(N) \leq N^{d_{\text{VC}}} + 1$
- Large d_{VC} = more complex
- So more difficult to train, and hence require more training samples

Trade-off Curve



VC Analysis

- VC analysis is a **decomposition**.
- Decompose E_{out} into E_{in} and ϵ.

$$E_{\text{out}} \leq E_{\text{in}} + \underbrace{\sqrt{\frac{8}{N} \log \frac{4\left((2N)^{d_{\text{VC}}} + 1\right)}{\delta}}}_{=\epsilon}$$

- E_{in} = training error, ϵ = penalty of complex model.
- Bias and variance is another decomposition.
- Decompose E_{out} into
  - How well can $\mathcal{H}$ approximate f?
  - How well can we zoom in a good h in $\mathcal{H}$?
- Roughly speaking we will have

$$E_{\text{out}} = \text{bias} + \textit{variance}$$

From VC Analysis to Bias-Variance

- In **VC analysis** we define the out-sample error as

$$E_{\text{out}}(g) = \mathbb{P}[g(\boldsymbol{x}) \neq f(\boldsymbol{x})]$$

- Let $B = \{g(\boldsymbol{x}) \neq f(\boldsymbol{x})\}$ be the bad event. $B \in \{0, 1\}$.
- Then this is equal to

$$\begin{aligned} E_{\text{out}}(g) &= \mathbb{P}[B = 1] \\ &= 1 \cdot \mathbb{P}[B = 1] + 0 \cdot \mathbb{P}[B = 0] \\ &= \mathbb{E}[B]. \end{aligned}$$

- So $E_{\text{out}}(g)$ can be written as

$$E_{\text{out}}(g) = \mathbb{E}_{\boldsymbol{x}}[\mathbf{1}\{g(\boldsymbol{x}) \neq f(\boldsymbol{x})\}].$$

- Expectation taken over all $\boldsymbol{x} \sim p(\boldsymbol{x})$.

Changing the Error Measure

- In **VC analysis** we define the out-sample error as

$$E_{\text{out}}(g) = \mathbb{E}_{\boldsymbol{x}}\left[\mathbf{1}\{g(\boldsymbol{x}) \neq f(\boldsymbol{x})\}\right]$$

- Expectation of a 0-1 loss.
- In **Bias-variance** analysis we define the out-sample error as

$$E_{\text{out}}(g) = \mathbb{E}_{\boldsymbol{x}}\left[(g(\boldsymbol{x}) - f(\boldsymbol{x}))^2\right].$$

- Expectation of a square loss.
- Square loss is differentiable.

Dependency on Training Set

- In VC analysis we define the out-sample error as

$$E_{\text{out}}(g^{(\mathcal{D})}) = \mathbb{E}_{\boldsymbol{x}}\left[\mathbf{1}\{g^{(\mathcal{D})}(\boldsymbol{x}) \neq f(\boldsymbol{x})\}\right]$$

- The final hypothesis depends on $\mathcal{D}$.
- If you use a different $\mathcal{D}$, your g will be different.
- In Bias-variance analysis we define the out-sample error as

$$E_{\text{out}}(g^{(\mathcal{D})}) = \mathbb{E}_{\boldsymbol{x}}\left[(g^{(\mathcal{D})}(\boldsymbol{x}) - f(\boldsymbol{x}))^2\right].$$

- Why did we skip $\mathcal{D}$ in VC analysis?
  - Hoeffding bound is uniform for **all** $\mathcal{D}$
  - So it does not matter which $\mathcal{D}$ you used to generate g
  - Not true for bias-variance

Averaging over all $\mathcal{D}$

- To account for all the possible $\mathcal{D}$'s, compute the expectation and define the expected out-sample error.

$$\mathbb{E}_{\mathcal{D}}\left[E_{\text{out}}(g^{(\mathcal{D})})\right] = \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\boldsymbol{x}}\left[(g^{(\mathcal{D})}(\boldsymbol{x}) - f(\boldsymbol{x}))^2\right]\right].$$

- $E_{\text{out}}(g^{(\mathcal{D})})$: Out-sample error for the particular g found from $\mathcal{D}$
- $\mathbb{E}_{\mathcal{D}}\left[E_{\text{out}}(g^{(\mathcal{D})})\right]$: Out-sample error averaged over all possible $\mathcal{D}$'s
- VC trade-off is a "worst case" analysis
  - Uniform bound on every $\mathcal{D}$
- Bias-variance trade-off is an "average" analysis
  - Average over different $\mathcal{D}$'s

Decomposing $\mathbb{E}_{\text{out}}(g^{(\mathcal{D})})$

- To account for all the possible $\mathcal{D}$'s, compute the expectation and define the expected out-sample error.

$$\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\text{out}}(g^{(\mathcal{D})})\right] = \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\boldsymbol{x}}\left[(g^{(\mathcal{D})}(\boldsymbol{x}) - f(\boldsymbol{x}))^2\right]\right].$$

- Let us do some calculation

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\boldsymbol{x}}\left[(g^{(\mathcal{D})}(\boldsymbol{x}) - f(\boldsymbol{x}))^2\right]\right] \\
&= \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_{\mathcal{D}}\left[(g^{(\mathcal{D})}(\boldsymbol{x}) - f(\boldsymbol{x}))^2\right]\right] \\
&= \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_{\mathcal{D}}\left[g^{(\mathcal{D})}(\boldsymbol{x})^2 - 2g^{(\mathcal{D})}(\boldsymbol{x})f(\boldsymbol{x}) + f(\boldsymbol{x})^2\right]\right] \\
&= \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_{\mathcal{D}}\left[g^{(\mathcal{D})}(\boldsymbol{x})^2\right] - 2\underbrace{\mathbb{E}_{\mathcal{D}}[g^{(\mathcal{D})}(\boldsymbol{x})]}_{\overline{g}(\boldsymbol{x})}f(\boldsymbol{x}) + f(\boldsymbol{x})^2\right].
\end{aligned}$$

The Average $\overline{g}(\boldsymbol{x})$

- The decomposition gives

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

$$= \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_{\mathcal{D}}\left[g^{(\mathcal{D})}(\boldsymbol{x})^2\right] - 2\underbrace{\mathbb{E}_{\mathcal{D}}[g^{(\mathcal{D})}(\boldsymbol{x})]}_{\overline{g}(\boldsymbol{x})}f(\boldsymbol{x}) + f(\boldsymbol{x})^2\right]$$

- We define the term

$$\overline{g}(\boldsymbol{x}) = \mathbb{E}_{\mathcal{D}}[g^{(\mathcal{D})}(\boldsymbol{x})]$$

- The asymptotic limit of the estimate

$$\overline{g}(\boldsymbol{x}) \approx \frac{1}{K}\sum_{k=1}^{K} g^{(\mathcal{D}_k)}(\boldsymbol{x})$$

- $g^{(\mathcal{D}_k)}$ are inside the hypothesis set. But $\overline{g}$ is *not* necessarily inside.

Bias and Variance

- Do some additional calculation

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\text{out}}(g^{(\mathcal{D})})\right] \\
&= \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_{\mathcal{D}}\left[g^{(\mathcal{D})}(\boldsymbol{x})^2\right] - 2\mathbb{E}_{\mathcal{D}}[g^{(\mathcal{D})}(\boldsymbol{x})]f(\boldsymbol{x}) + f(\boldsymbol{x})^2\right] \\
&= \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_{\mathcal{D}}\left[g^{(\mathcal{D})}(\boldsymbol{x})^2\right] - 2\overline{g}(\boldsymbol{x})f(\boldsymbol{x}) + f(\boldsymbol{x})^2\right] \\
&= \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_{\mathcal{D}}\left[g^{(\mathcal{D})}(\boldsymbol{x})^2\right] - \overline{g}(\boldsymbol{x})^2 + \overline{g}(\boldsymbol{x})^2 - 2\overline{g}(\boldsymbol{x})f(\boldsymbol{x}) + f(\boldsymbol{x})^2\right] \\
&= \mathbb{E}_{\boldsymbol{x}}\Big[\underbrace{\mathbb{E}_{\mathcal{D}}\left[g^{(\mathcal{D})}(\boldsymbol{x})^2\right] - \overline{g}(\boldsymbol{x})^2}_{\mathbb{E}_{\mathcal{D}}[(g^{(\mathcal{D})}(\boldsymbol{x})-\overline{g}(\boldsymbol{x}))^2]} + \underbrace{\overline{g}(\boldsymbol{x})^2 - 2\overline{g}(\boldsymbol{x})f(\boldsymbol{x}) + f(\boldsymbol{x})^2}_{(\overline{g}(\boldsymbol{x})-f(\boldsymbol{x}))^2}\Big].
\end{aligned}
$$

- Define two terms

$$
\text{bias}(\boldsymbol{x}) \overset{\text{def}}{=} (\overline{g}(\boldsymbol{x}) - f(\boldsymbol{x}))^2,
$$

$$
\text{var}(\boldsymbol{x}) \overset{\text{def}}{=} \mathbb{E}_{\mathcal{D}}[(g^{(\mathcal{D})}(\boldsymbol{x}) - \overline{g}(\boldsymbol{x}))^2].
$$

Bias and Variance

- The decomposition:

$$\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\text{out}}(g^{(\mathcal{D})})\right]$$

$$= \mathbb{E}_{\mathbf{x}}\Big[\underbrace{\mathbb{E}_{\mathcal{D}}\left[g^{(\mathcal{D})}(\mathbf{x})^2\right] - \overline{g}(\mathbf{x})^2}_{\mathbb{E}_{\mathcal{D}}[(g^{(\mathcal{D})}(\mathbf{x})-\overline{g}(\mathbf{x}))^2]} + \underbrace{\overline{g}(\mathbf{x})^2 - 2\overline{g}(\mathbf{x})f(\mathbf{x}) + f(\mathbf{x})^2}_{(\overline{g}(\mathbf{x})-f(\mathbf{x}))^2}\Big].$$

- Define two terms

$$\text{bias}(\mathbf{x}) \stackrel{\text{def}}{=} (\overline{g}(\mathbf{x}) - f(\mathbf{x}))^2,$$

$$\text{var}(\mathbf{x}) \stackrel{\text{def}}{=} \mathbb{E}_{\mathcal{D}}[(g^{(\mathcal{D})}(\mathbf{x}) - \overline{g}(\mathbf{x}))^2].$$

- Take expectation

$$\text{bias} = \mathbb{E}_{\mathbf{x}}[\text{bias}(\mathbf{x})] = \mathbb{E}_{\mathbf{x}}\left[(\overline{g}(\mathbf{x}) - f(\mathbf{x}))^2\right],$$

$$\text{var} = \mathbb{E}_{\mathbf{x}}[\text{var}(\mathbf{x})] = \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathcal{D}}[(g^{(\mathcal{D})}(\mathbf{x}) - \overline{g}(\mathbf{x}))^2]\right].$$

Bias and Variance Decomposition

- The decomposition:

$$\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\text{out}}(g^{(\mathcal{D})})\right] = \mathbb{E}_{\boldsymbol{x}}\Bigg[\underbrace{\mathbb{E}_{\mathcal{D}}\left[g^{(\mathcal{D})}(\boldsymbol{x})^2\right] - \overline{g}(\boldsymbol{x})^2}_{\mathbb{E}_{\mathcal{D}}[(g^{(\mathcal{D})}(\boldsymbol{x})-\overline{g}(\boldsymbol{x}))^2]} + \underbrace{\overline{g}(\boldsymbol{x})^2 - 2\overline{g}(\boldsymbol{x})f(\boldsymbol{x}) + f(\boldsymbol{x})^2}_{(\overline{g}(\boldsymbol{x})-f(\boldsymbol{x}))^2}\Bigg].$$

- This gives

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\text{out}}(g^{(\mathcal{D})})\right] &= \mathbb{E}_{\boldsymbol{x}}[\text{bias}(\boldsymbol{x}) + \text{var}(\boldsymbol{x})] \\ &= \text{bias} + \text{var}\end{aligned}$$

Interpreting the Bias-Variance

- The decomposition:

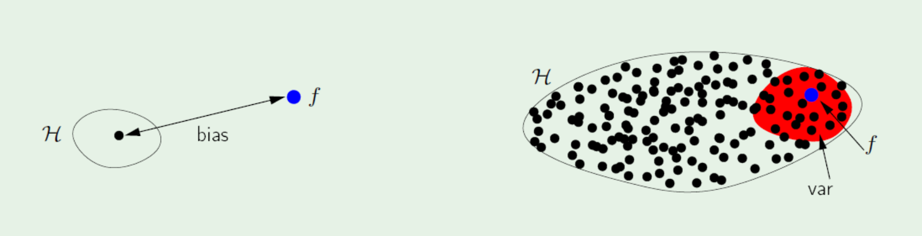
$$\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\text{out}}(g^{(\mathcal{D})})\right]$$
$$= \mathbb{E}_{\boldsymbol{x}}\Bigg[\underbrace{\mathbb{E}_{\mathcal{D}}\left[g^{(\mathcal{D})}(\boldsymbol{x})^2\right] - \overline{g}(\boldsymbol{x})^2}_{\mathbb{E}_{\mathcal{D}}[(g^{(\mathcal{D})}(\boldsymbol{x})-\overline{g}(\boldsymbol{x}))^2]} + \underbrace{\overline{g}(\boldsymbol{x})^2 - 2\overline{g}(\boldsymbol{x})f(\boldsymbol{x}) + f(\boldsymbol{x})^2}_{(\overline{g}(\boldsymbol{x})-f(\boldsymbol{x}))^2}\Bigg].$$

- The two terms:

$$\text{bias}(\boldsymbol{x}) \overset{\text{def}}{=} (\overline{g}(\boldsymbol{x}) - f(\boldsymbol{x}))^2,$$
$$\text{var}(\boldsymbol{x}) \overset{\text{def}}{=} \mathbb{E}_{\mathcal{D}}[(g^{(\mathcal{D})}(\boldsymbol{x}) - \overline{g}(\boldsymbol{x}))^2].$$

- bias($\boldsymbol{x}$): How close is the **average function** $\overline{g}$ to the target
- var($\boldsymbol{x}$): How much **uncertainty** you have around $\overline{g}$

Model Complexity



- The bias and variance are

$$\text{bias}(\boldsymbol{x}) \overset{\text{def}}{=} (\overline{g}(\boldsymbol{x}) - f(\boldsymbol{x}))^2,$$

$$\text{var}(\boldsymbol{x}) \overset{\text{def}}{=} \mathbb{E}_{\mathcal{D}}[(g^{(\mathcal{D})}(\boldsymbol{x}) - \overline{g}(\boldsymbol{x}))^2].$$

- If you have a simple $\mathcal{H}$, then large bias but small variance
- If you have a complex $\mathcal{H}$, then small bias but large variance

Outline

- Lecture 28 Sample and Model Complexity
- Lecture 29 Bias and Variance
- Lecture 30 Overfit

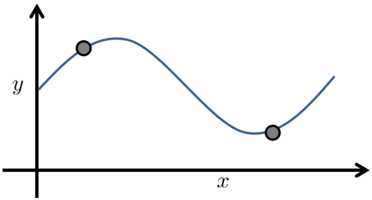
Today's Lecture:

- From VC Analysis to Bias-Variance
  - Generalization Bound
  - Bias-Variance Decomposition
  - Interpreting Bias-Variance
- Example
  - 0-th order vs 1-st order model
  - Trade off

Example

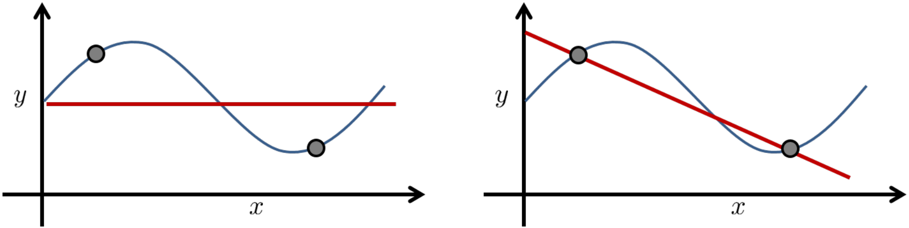
- Consider a $\sin(\cdot)$ function

$$f(x) = \sin(\pi x)$$



- You are only given $N = 2$ training samples
- These two samples are sampled uniformly in $[-1, 1]$.
- Call them (x_1, y_1) and (x_2, y_2)
- Hypothesis Set 0: $\mathcal{M}_0$ = Set of all lines of the form $h(x) = b$;
- Hypothesis Set 1: $\mathcal{M}_1$ = Set of all lines of the form $h(x) = ax + b$.
- Which one fits better?

Example



- If you give me two points, I can tell you the fitted lines
- For $\mathcal{M}_0$:

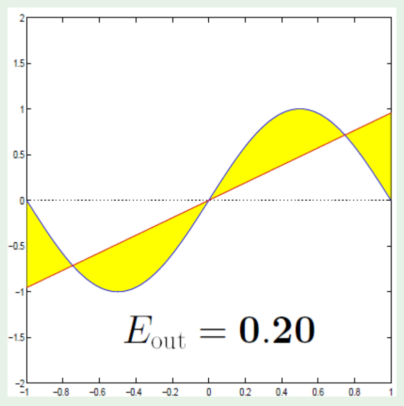
$$h(x) = \frac{y_1 + y_2}{2}.$$

- For $\mathcal{M}_1$:

$$h(x) = \left(\frac{y_2 - y_1}{x_2 - x_1}\right) x + (y_1 x_2 - y_2 x_1).$$

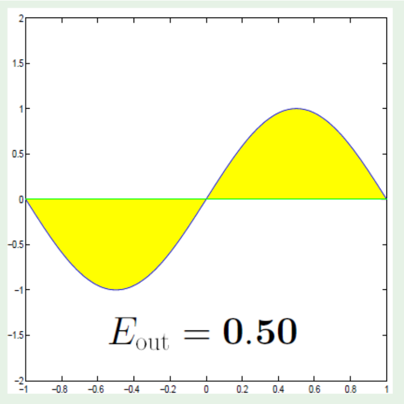
Out-sample Error E_{out}

- If you use $\mathcal{M}_1$
- Then you get this
- $E_{\text{out}} = 0.2$



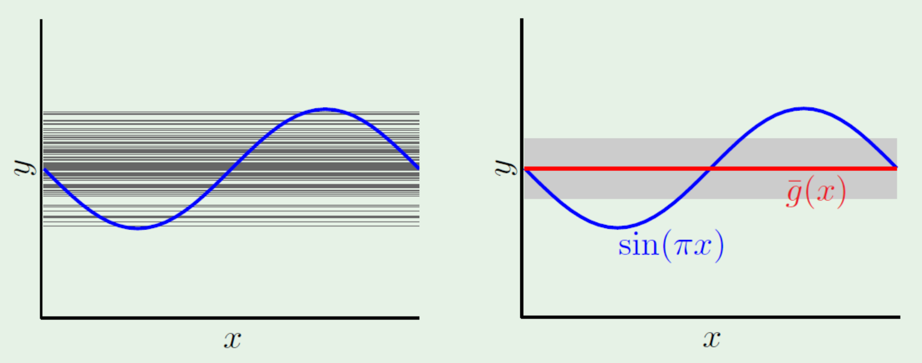
Out-sample Error E_{out}

- If you use $\mathcal{M}_0$
- Then you get this
- $E_{\text{out}} = 0.5$



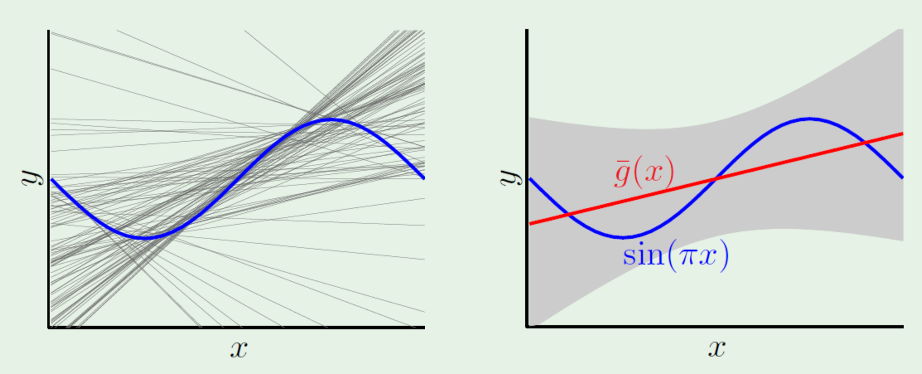
Scan through $\mathcal{D}$

- Now draw a different training set
- Then you have a different curve every time
- Plot them all on the same figure
- Here is what you will get



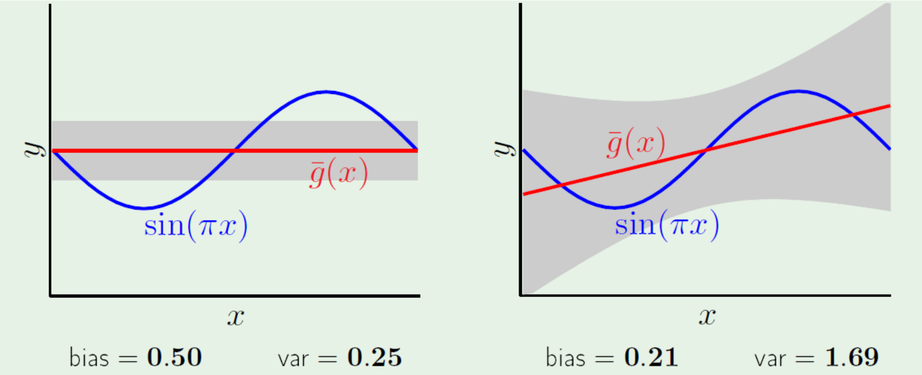
Scan through $\mathcal{D}$

- Now draw a different training set
- Then you have a different curve every time
- Plot them all on the same figure
- Here is what you will get

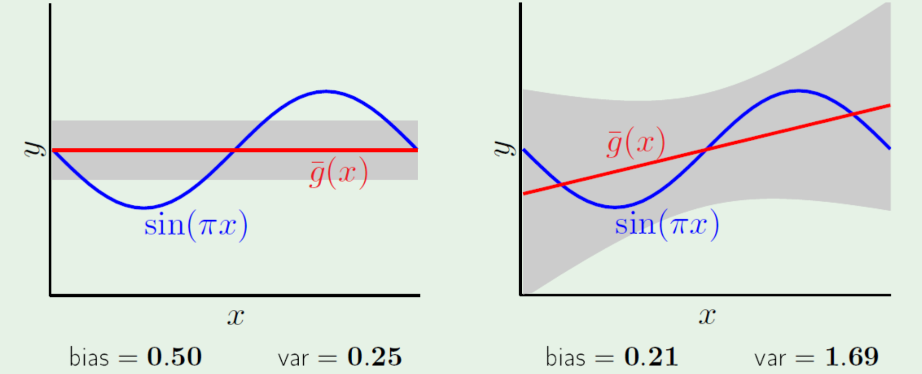


Limiting Case

- Draw infinitely many training sets
- You will have two quantities
- $\overline{g}(x)$: The average line
- $\sqrt{\text{var}(x)}$: The variance



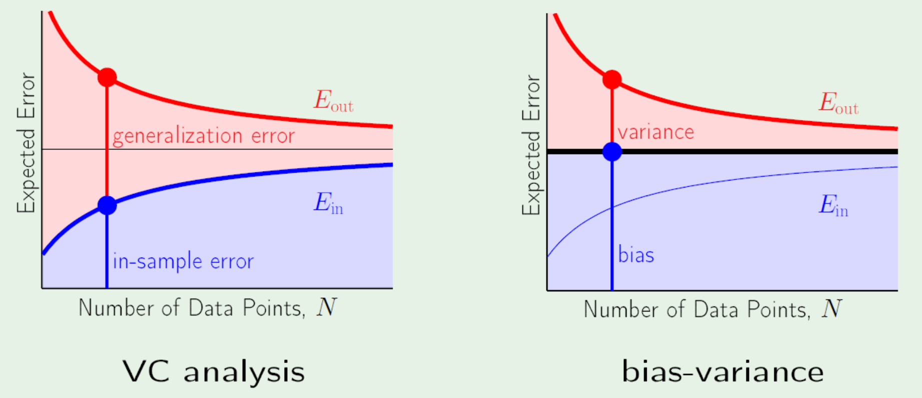
How Come!



- $\overline{g}(\boldsymbol{x})$ is a good **average**.
- But the **error bar** is big!
- Analogy: I have a powerful canon but not very accurate.

Learning Curve

- Expected out-sample error: $E_{\text{out}}(g^{(\mathcal{D})})$
- Expected in-sample error: $E_{\text{in}}(g^{(\mathcal{D})})$
- How do they change with N?



VC vs Bias-Variance

- VC analysis is independent of $\mathcal{A}$
- Bias-variance depends on $\mathcal{A}$
- With the same $\mathcal{H}$, VC always returns the same generalization bound
- Guarantee over all possible choices of dataset $\mathcal{D}$
- Bias-variance: For the same $\mathcal{H}$, you can have different $g^{(\mathcal{D})}$
- Depend on $\mathcal{D}$, you have a different $E_{\text{out}}(g^{(\mathcal{D})})$
- Therefore we take expectation

$$\mathbb{E}_{\mathcal{D}}\left[E_{\text{out}}(g^{(\mathcal{D})})\right]$$

- In practice, bias and variance cannot be computed
- You do not have f
- It is a conceptual tool to design algorithms

Reading List

- Yasar Abu-Mostafa, Learning from Data, chapter 2.2
- Chris Bishop, Pattern Recognition and Machine Intelligence, chapter 3.2
- Duda, Hart and Stork, Pattern Classification, chapter 9.3
- Stanford STAT202 https://web.stanford.edu/class/stats202/content/lec2.pdf
- CMU 10-601 https://www.cs.cmu.edu/~wcohen/10-601/bias-variance.pdf
- UCSD 271A http://www.svcl.ucsd.edu/courses/ece271A/handouts/ML2.pdf

Appendix

Case Study: Linear Regression

- You are given a training dataset
- $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$
- Train a linear regression model

$$
\begin{aligned}
\widehat{\boldsymbol{w}} &= \underset{\boldsymbol{w}}{\text{argmin}} \;\; \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n^T\boldsymbol{w} - y_n)^2 \\
&= \underset{\boldsymbol{w}}{\text{argmin}} \;\; \frac{1}{N}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2
\end{aligned}
$$

- What is the in-sample error?
- What is the out-sample error?

In-Sample Error

- In-sample error is

$$E_{\text{in}}(\widehat{\boldsymbol{w}}) = \frac{1}{N}\|\boldsymbol{X}\widehat{\boldsymbol{w}} - \boldsymbol{y}\|^2$$

- What is $\widehat{\boldsymbol{w}}$?
- Take derivative, setting to zero:

$$\frac{d}{d\boldsymbol{w}}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2 = 2\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}) = \boldsymbol{0}.$$

- Solution is

$$\widehat{\boldsymbol{w}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}.$$

- So In-Sample error is

$$\begin{aligned} E_{\text{in}}(\widehat{\boldsymbol{w}}) &= \frac{1}{N}\|\boldsymbol{X}\widehat{\boldsymbol{w}} - \boldsymbol{y}\|^2 \\ &= \frac{1}{N}\|\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{y}\|^2 \end{aligned}$$

Modeling the Input

- Define
$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T.$$
- Can show that $\boldsymbol{H}^k = \boldsymbol{H}$ for any $k > 0$, and $\boldsymbol{H} = \boldsymbol{H}^T$.
- $\text{Tr}(\boldsymbol{H}) = d + 1$.
- Assume $\boldsymbol{y} = \boldsymbol{X}^T\boldsymbol{w}^* + \boldsymbol{\epsilon}$, then
$$\begin{aligned}\widehat{\boldsymbol{y}} &\overset{\text{def}}{=} \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \\ &= \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{w}^* + \boldsymbol{\epsilon}) \\ &= \boldsymbol{X}\boldsymbol{w}^* + \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon} \\ &= \boldsymbol{X}\boldsymbol{w}^* + \boldsymbol{H}\boldsymbol{\epsilon}.\end{aligned}$$
- Residue is
$$\begin{aligned}\widehat{\boldsymbol{y}} - \boldsymbol{y} &= (\boldsymbol{X}\boldsymbol{w}^* + \boldsymbol{H}\boldsymbol{\epsilon}) - (\boldsymbol{X}^T\boldsymbol{w}^* + \boldsymbol{\epsilon}) \\ &= (\boldsymbol{H} - \boldsymbol{I})\boldsymbol{\epsilon}.\end{aligned}$$

In-Sample Error

- In-sample error is

$$\begin{aligned} E_{\text{in}}(\widehat{\boldsymbol{w}}) &= \frac{1}{N}\|\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{y}\|^2 \\ &= \frac{1}{N}\|\widehat{\boldsymbol{y}} - \boldsymbol{y}\|^2 = \frac{1}{N}\boldsymbol{\epsilon}^T(\boldsymbol{H}-\boldsymbol{I})^T(\boldsymbol{H}-\boldsymbol{I})\boldsymbol{\epsilon} \\ &= \frac{1}{N}\boldsymbol{\epsilon}^T(\boldsymbol{H}-\boldsymbol{I})\boldsymbol{\epsilon} \end{aligned}$$

- Take expectation over $\mathcal{D}$ yields

$$\begin{aligned} \mathbb{E}_{\mathcal{D}}\left[E_{\text{in}}(\widehat{\boldsymbol{w}})\right] &= \mathbb{E}\left[\frac{1}{N}\boldsymbol{\epsilon}^T(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{\epsilon}\right] \\ &= \frac{1}{N}\text{Tr}(\boldsymbol{I}-\boldsymbol{H})\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] \\ &= \frac{\sigma^2}{N}\text{Tr}(\boldsymbol{I}-\boldsymbol{H}) = \frac{\sigma^2}{N}(d+1-N) = \sigma^2\left(1-\frac{d+1}{N}\right). \end{aligned}$$

Out-Sample

- We study a simplified case: The out-samples are $(\boldsymbol{x}_1, y'_1), \ldots, (\boldsymbol{x}_N, y'_N)$.
- Assume $\boldsymbol{y}' = \boldsymbol{X}\boldsymbol{w}^* + \boldsymbol{\epsilon}'$.
- E_{out} is

$$E_{\text{out}}(\widehat{\boldsymbol{w}}) = \frac{1}{N}\|\widehat{\boldsymbol{y}} - \boldsymbol{y}'\|^2 = \frac{1}{N}\|\boldsymbol{H}\boldsymbol{\epsilon} - \boldsymbol{\epsilon}'\|^2.$$

- $\mathbb{E}_{\mathcal{D}}[E_{\text{out}}(\widehat{\boldsymbol{w}})]$ is

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}[E_{\text{out}}(\widehat{\boldsymbol{w}})] &= \frac{1}{N}\mathbb{E}_{\mathcal{D}}\left[\boldsymbol{\epsilon}^T\boldsymbol{H}^T\boldsymbol{H}\boldsymbol{\epsilon} + \|\boldsymbol{\epsilon}'\|^2\right] \\
&= \frac{1}{N}\left\{\mathbb{E}_{\mathcal{D}}\left[\boldsymbol{\epsilon}^T\boldsymbol{H}^T\boldsymbol{H}\boldsymbol{\epsilon}\right] + \mathbb{E}_{\mathcal{D}}\left[\boldsymbol{\epsilon}'\boldsymbol{\epsilon}'^T\right]\right\} \\
&= \frac{1}{N}\left\{\sigma^2(d+1) + \sigma^2 N\right\} = \sigma^2\left(1 + \frac{d+1}{N}\right).
\end{aligned}$$