Machine Learning from Data

Lecture 7: Spring 2021

Today's Lecture

- Approximation Vs Generalization
  - VC Dimension
  - Bias-Variance Analysis
  - Learning Curve

Theory of Learning and its relevance.

- We created a link between E_{out} and E_{in}

$$\mathbb{P}\left[\,|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon\,\right] \le 4m_{\mathcal{H}}(2N)e^{-\epsilon^2 N/8}, \qquad \text{for any } \epsilon > 0.$$

$\mathbb{P}[|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon] \le 2|\mathcal{H}|e^{-2\epsilon^2 N}$ $\leftarrow$ finite $\mathcal{H}$

$$\mathbb{P}\left[\,|E_{\text{in}}(g) - E_{\text{out}}(g)| \le \epsilon\,\right] \ge 1 - 4m_{\mathcal{H}}(2N)e^{-\epsilon^2 N/8}, \qquad \text{for any } \epsilon > 0.$$

$\mathbb{P}[|E_{\text{in}}(g) - E_{\text{out}}(g)| \le \epsilon] \ge 1 - 2|\mathcal{H}|e^{-2\epsilon^2 N}$ $\leftarrow$ finite $\mathcal{H}$

$$E_{\text{out}}(g) \le E_{\text{in}}(g) + \sqrt{\frac{8}{N}\log\frac{4m_{\mathcal{H}}(2N)}{\delta}}, \qquad \text{w.p. at least } 1 - \delta.$$

$E_{\text{out}}(g) \le E_{\text{in}}(g) + \sqrt{\frac{1}{2N}\log\frac{2|\mathcal{H}|}{\delta}}$ $\leftarrow$ finite $\mathcal{H}$

$$m_{\mathcal{H}}(N) \le \sum_{i=1}^{k-1}\binom{N}{i} \le N^{k-1} + 1 \qquad k \text{ is a break point.}$$

The VC Dimension

•

Summarize

$$m_{\mathcal{H}}(N) \sim N^{k-1}$$

The tightest bound is obtained with the smallest break point k^*.

Definition [VC Dimension] $d_{\mathrm{VC}} = k^* - 1$.

The VC dimension is the largest N which can be shattered ($m_{\mathcal{H}}(N) = 2^N$).

$N \leq d_{\mathrm{VC}}$: $\mathcal{H}$ *could* shatter your data ($\mathcal{H}$ can shatter some N points).

$N > d_{\mathrm{VC}}$: N is a break point for $\mathcal{H}$; $\mathcal{H}$ cannot possibly shatter your data.

$$\boxed{\begin{aligned} & m_{\mathcal{H}}(N) \leq N^{d_{\mathrm{vc}}} + 1 \sim N^{d_{\mathrm{vc}}} \\ & E_{\mathrm{out}}(g) \leq E_{\mathrm{in}}(g) + O\left(\sqrt{\frac{d_{\mathrm{vc}} \log N}{N}}\right) \end{aligned}}$$

Examples

•

Convex Sets

•

Summarize

	N 1	2	3	4	5	$\cdots$	#Param	d_{VC}
2-D perceptron	2	4	8	14	$\cdots$		3	3
1-D pos. ray	2	3	4	5	$\cdots$		1	1
2-D pos. rectangles	2	4	8	16	$< 2^5$	$\cdots$	4	4
pos. convex sets	2	4	8	16	32	$\cdots$	∞	∞

There are models with few parameters but infinite d_{VC}.

There are models with redundant parameters but small d_{VC}.

Sample Complexity?

Set the error bar at ϵ.

$$\epsilon = \sqrt{\frac{8}{N} \ln \frac{4((2N)^{d_{\mathrm{vc}}} + 1)}{\delta}}$$

Solve for N:

$$N = \frac{8}{\epsilon^2} \ln \frac{4((2N)^{d_{\mathrm{vc}}} + 1)}{\delta} = O\left(d_{\mathrm{VC}} \ln N\right)$$

Example. $d_{\mathrm{VC}} = 3$; error bar $\epsilon = 0.1$; confidence 90% ($\delta = 0.1$).
A simple iterative method works well. Trying $N = 1000$ we get

$$N \approx \frac{1}{0.1^2} \log\left(\frac{4(2000)^3 + 4}{0.1}\right) \approx 21192.$$

We continue iteratively, and converge to $N \approx 30000$.
If $d_{\mathrm{VC}} = 4$, $N \approx 40000$; for $d_{\mathrm{VC}} = 5$, $N \approx 50000$.

($N \propto d_{\mathrm{VC}}$, but gross overestimates)

Practical Rule of Thumb: $N = 10 \times d_{\mathbf{vc}}$

Theory Vs Practice

-

VC Bound Quantifies Approximation Vs. Generalization

$$d_{\text{VC}} \uparrow \implies \text{better chance of } \textbf{approximating } f \ (E_{\text{in}} \approx 0).$$

$$d_{\text{VC}} \downarrow \implies \text{better chance of } \textbf{generalizing} \text{ to out of sample } (E_{\text{in}} \approx E_{\text{out}}).$$

$$E_{\text{out}} \leq E_{\text{in}} + \Omega(d_{\text{VC}}).$$

Bias Variance Trade-off

1. How well *can* the learning approximate f.

   ...as opposed to how well *did* the learning approximate f in-sample (E_{in}).

2. How close can you get to that approximation with a finite data set.

   ...as opposed to how close is E_{in} to E_{out}.

Bias-variance analysis applies to squared errors (classification and regression)

Bias-variance analysis can take into account the *learning algorithm*

Different learning algorithms can have different E_{out} when applied to the same $\mathcal{H}$!

A simple learning problem

•

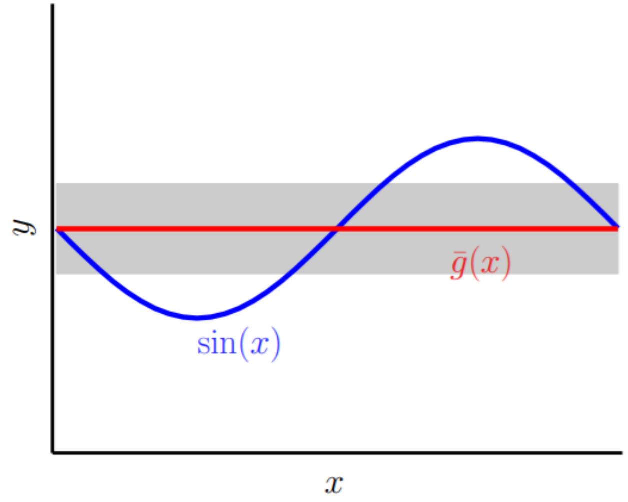
Problem continued…

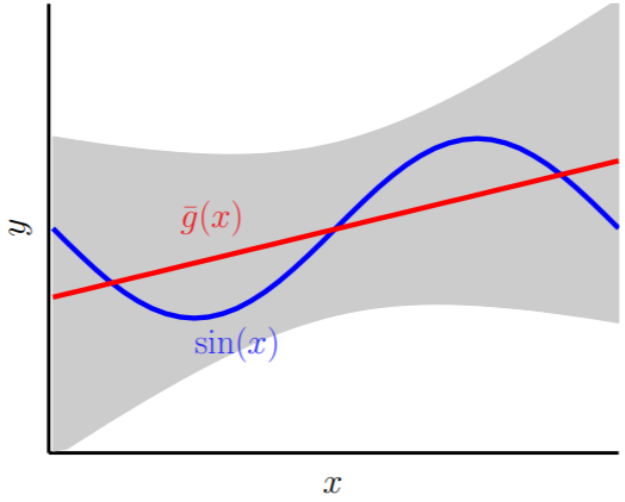
•

Expected Behavior with Datasets

-

Which one is better?





$\mathcal{H}_0$

bias $= 0.50$

var $= 0.25$

$E_{\text{out}} = 0.75$ ✓

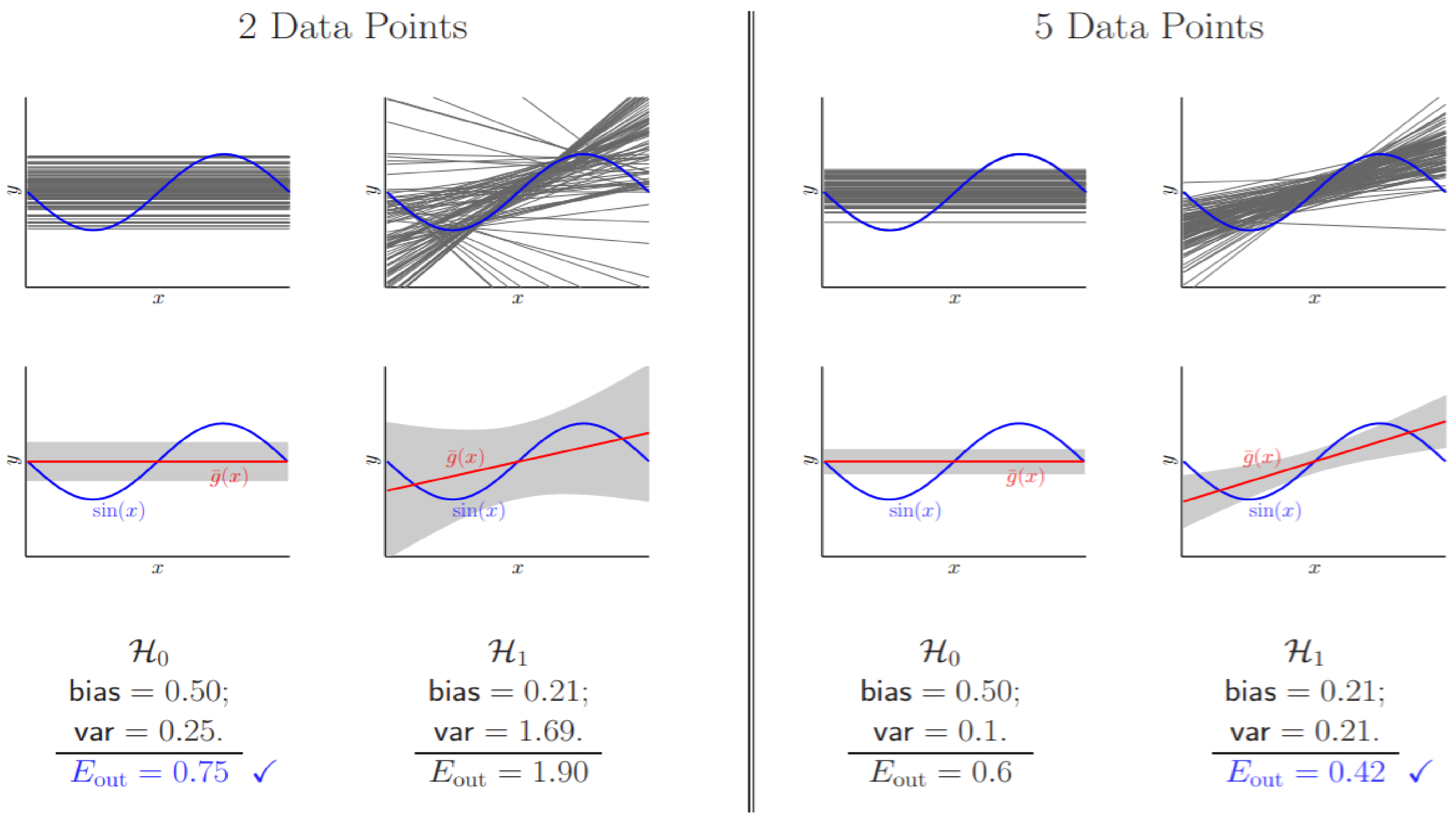
$\mathcal{H}_1$

bias $= 0.21$

var $= 1.69$

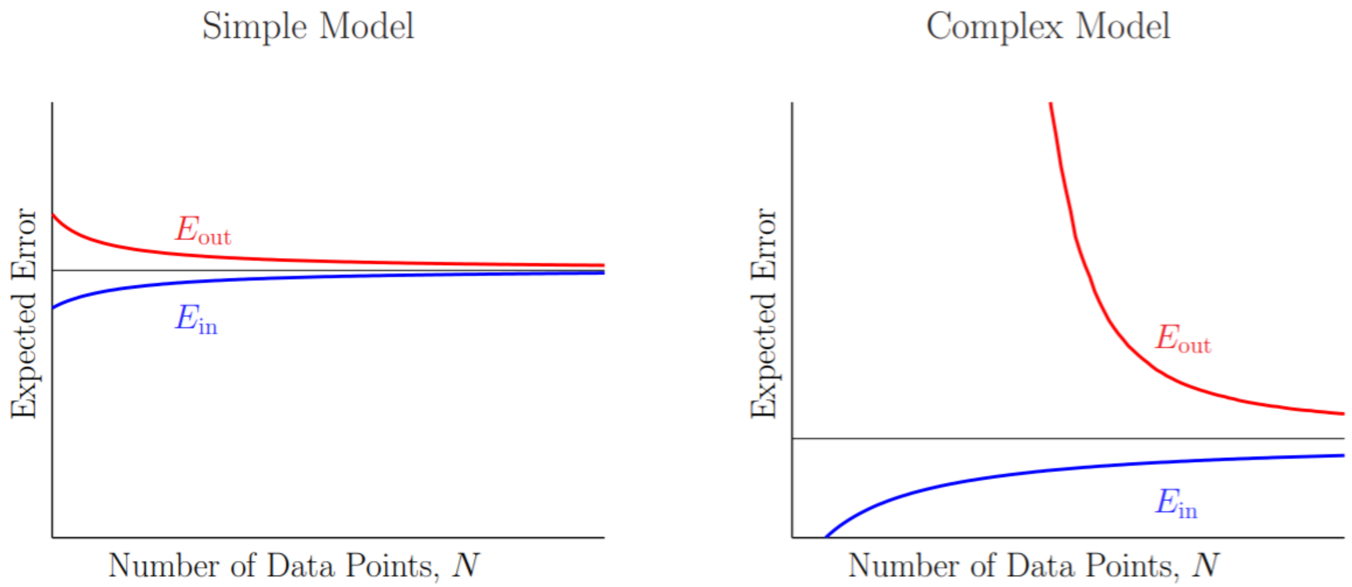
$E_{\text{out}} = 1.90$

Data



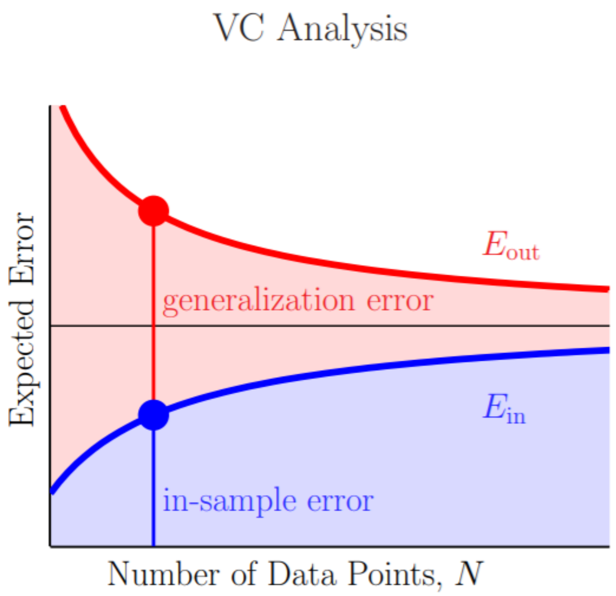
2 Data Points
5 Data Points
y
x
$\sin(x)$
$\bar{g}(x)$
$\mathcal{H}_0$
bias $= 0.50$;
var $= 0.25$.
$E_{\text{out}} = 0.75$ ✓
$\mathcal{H}_1$
bias $= 0.21$;
var $= 1.69$.
$E_{\text{out}} = 1.90$
$\mathcal{H}_0$
bias $= 0.50$;
var $= 0.1$.
$E_{\text{out}} = 0.6$
$\mathcal{H}_1$
bias $= 0.21$;
var $= 0.21$.
$E_{\text{out}} = 0.42$ ✓

Learning Curve

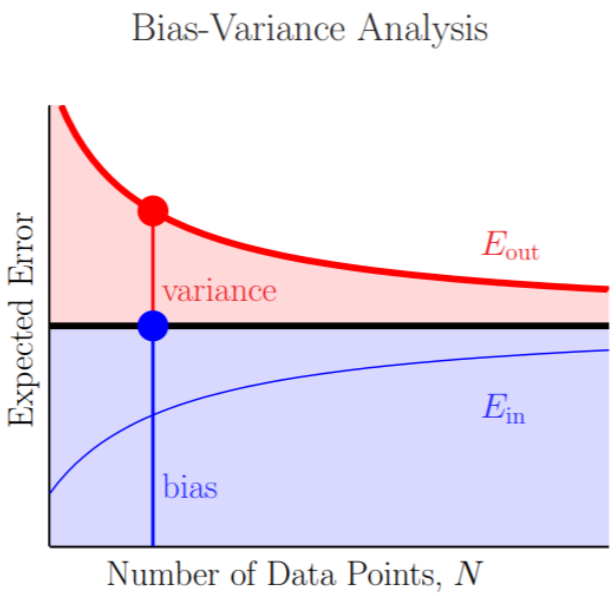


$$E_{out} = \mathbb{E}_{\mathbf{x}}\left[E_{out}(\mathbf{x})\right]$$

Comparison



Pick $\mathcal{H}$ that can generalize and has a good chance to fit the data



Pick $(\mathcal{H}, \mathcal{A})$ to approximate f and not behave wildly after seeing the data

Thanks!