

Approximation-generalization tradeoff

Small E_{out} : good approximation of f out of sample.

More complex $\mathcal{H} \implies$ better chance of **approximating** f

Less complex $\mathcal{H} \implies$ better chance of **generalizing** out of sample

Ideal $\mathcal{H} = \{f\}$ winning lottery ticket 😊

Quantifying the tradeoff

VC analysis was one approach: $E_{\text{out}} \leq E_{\text{in}} + \Omega$

Bias-variance analysis is another: decomposing E_{out} into

1. How well \mathcal{H} can approximate f
2. How well we can zoom in on a good $h \in \mathcal{H}$

Applies to **real-valued targets** and uses **squared error**

Start with E_{out}

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

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

Now, let us focus on:

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

The average hypothesis

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

we define the 'average' hypothesis $\bar{g}(\mathbf{x})$:

$$\bar{g}(\mathbf{x}) = \mathbb{E}_{\mathcal{D}} \left[g^{(\mathcal{D})}(\mathbf{x}) \right]$$

Imagine **many** data sets $\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_K$

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

Using $\bar{g}(\mathbf{x})$

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

Bias and variance

$$\mathbb{E}_{\mathcal{D}} \left[\left(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x}) \right)^2 \right] = \underbrace{\mathbb{E}_{\mathcal{D}} \left[\left(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x}) \right)^2 \right]}_{\text{var}(\mathbf{x})} + \underbrace{\left(\bar{g}(\mathbf{x}) - f(\mathbf{x}) \right)^2}_{\text{bias}(\mathbf{x})}$$

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

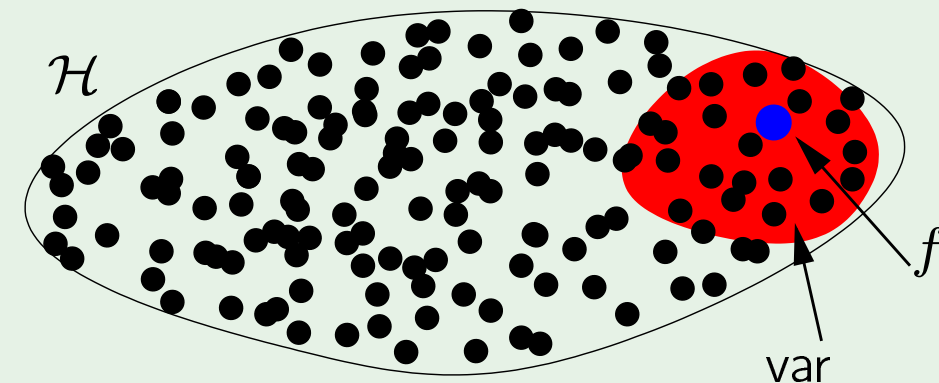
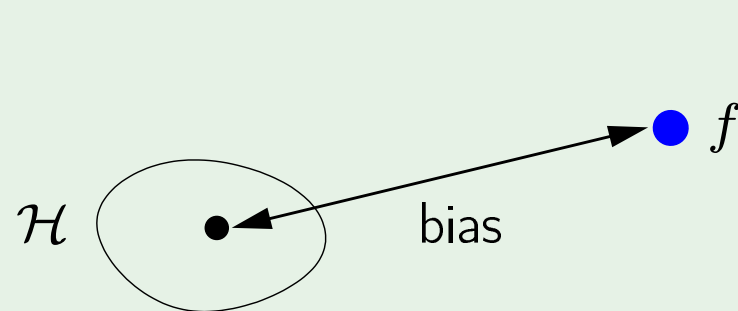
$$= \mathbb{E}_{\mathbf{x}} [\text{bias}(\mathbf{x}) + \text{var}(\mathbf{x})]$$

$$= \text{bias} + \text{var}$$

The tradeoff

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

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



$\mathcal{H} \uparrow$



Example: sine target

f

$$f : [-1, 1] \rightarrow \mathbb{R} \quad f(x) = \sin(\pi x)$$

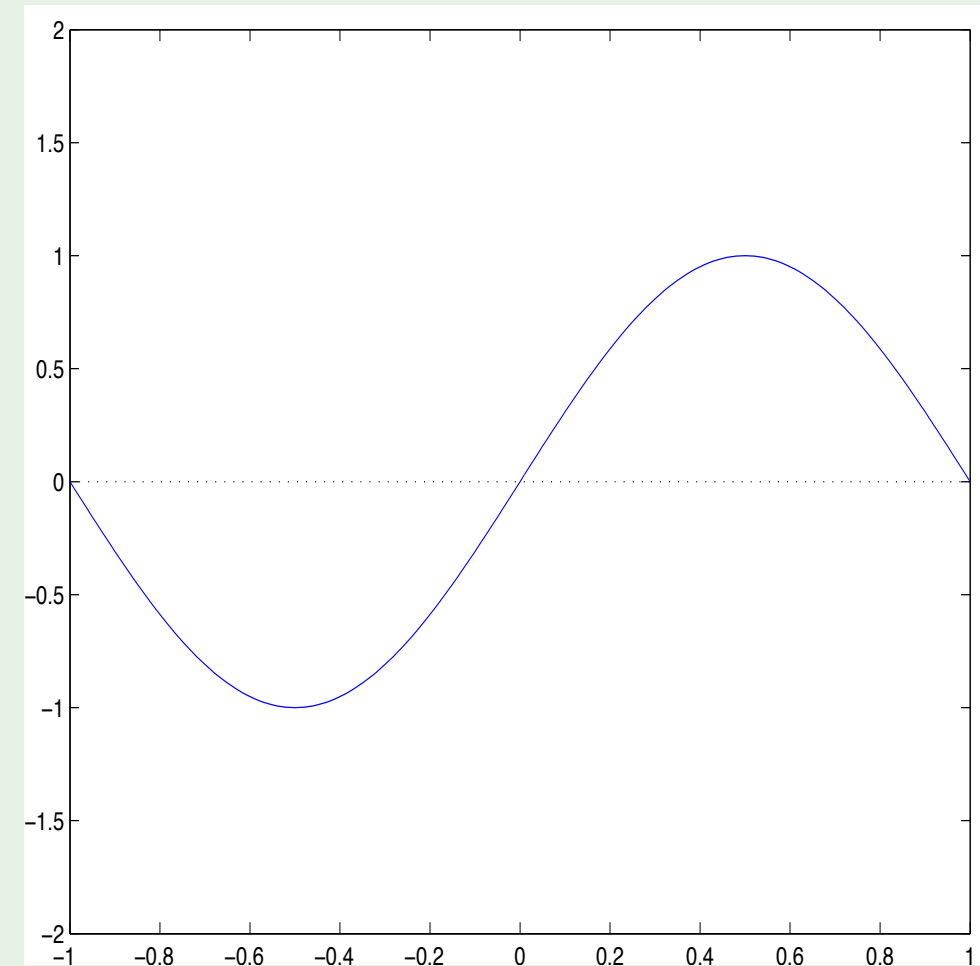
Only two training examples! $N = 2$

Two models used for learning:

$$\mathcal{H}_0: \quad h(x) = b$$

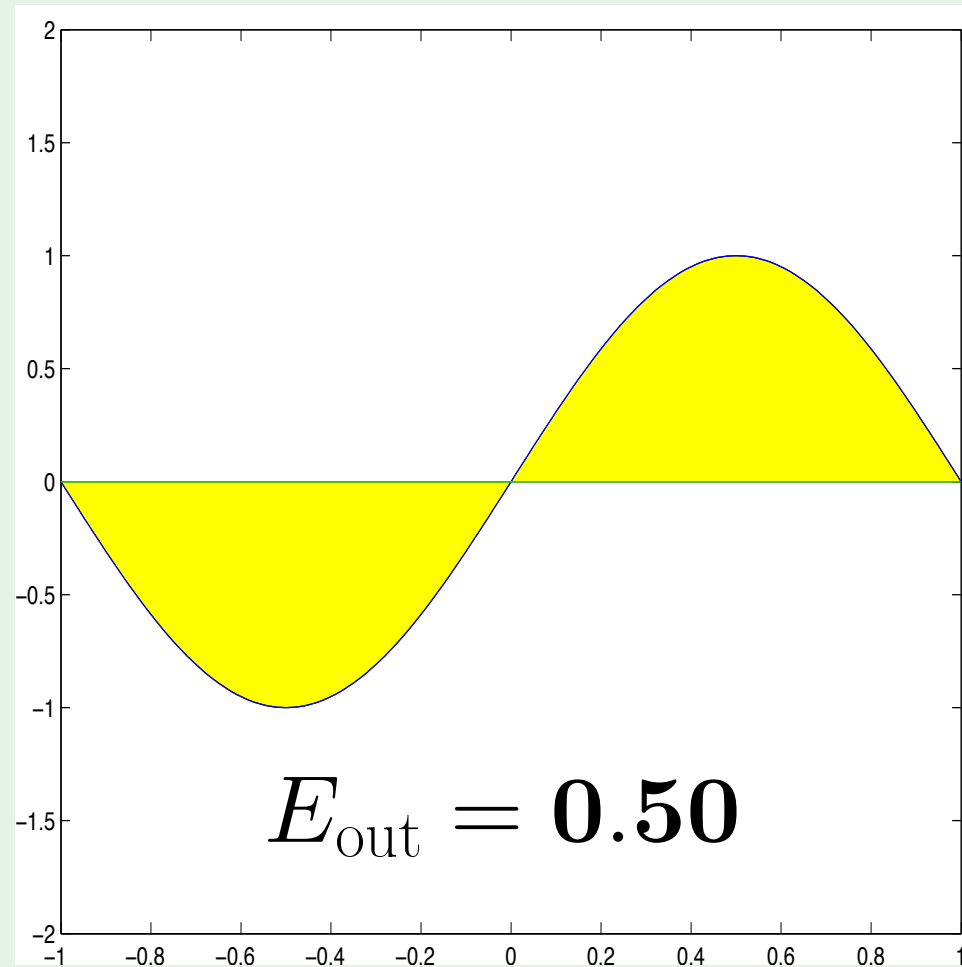
$$\mathcal{H}_1: \quad h(x) = ax + b$$

Which is better, \mathcal{H}_0 or \mathcal{H}_1 ?

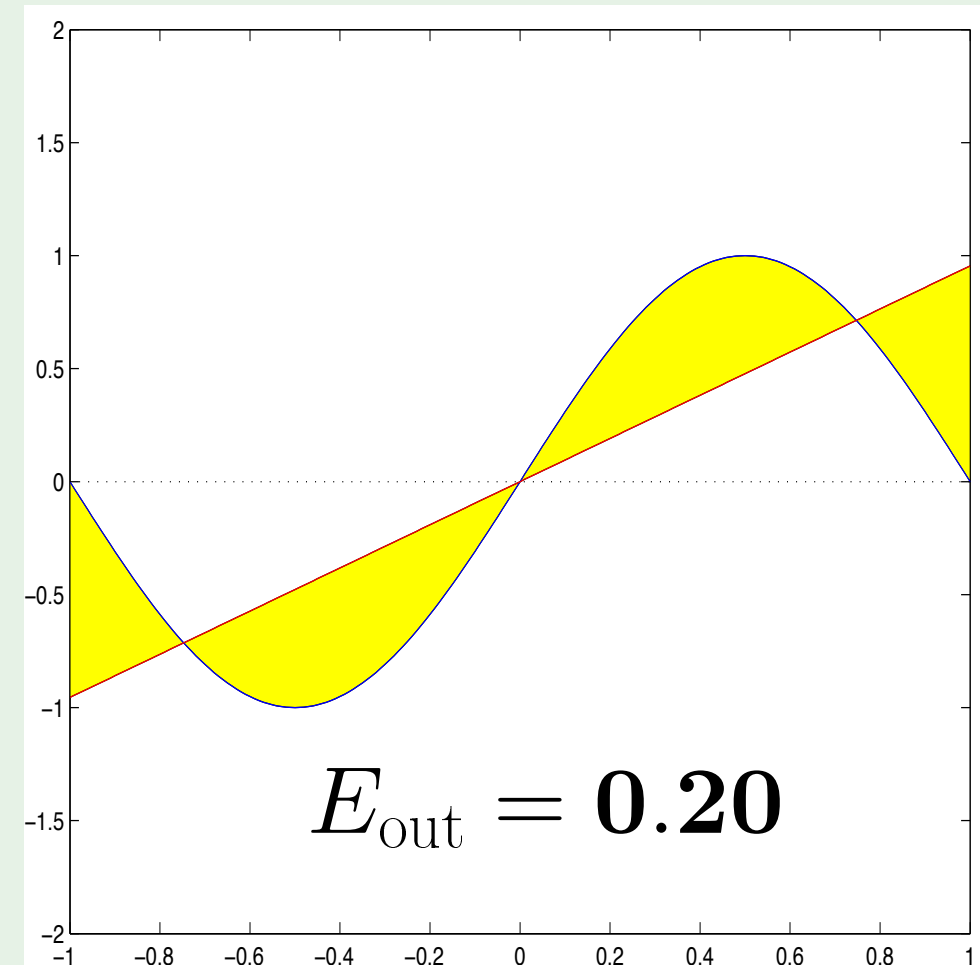


Approximation - \mathcal{H}_0 versus \mathcal{H}_1

\mathcal{H}_0

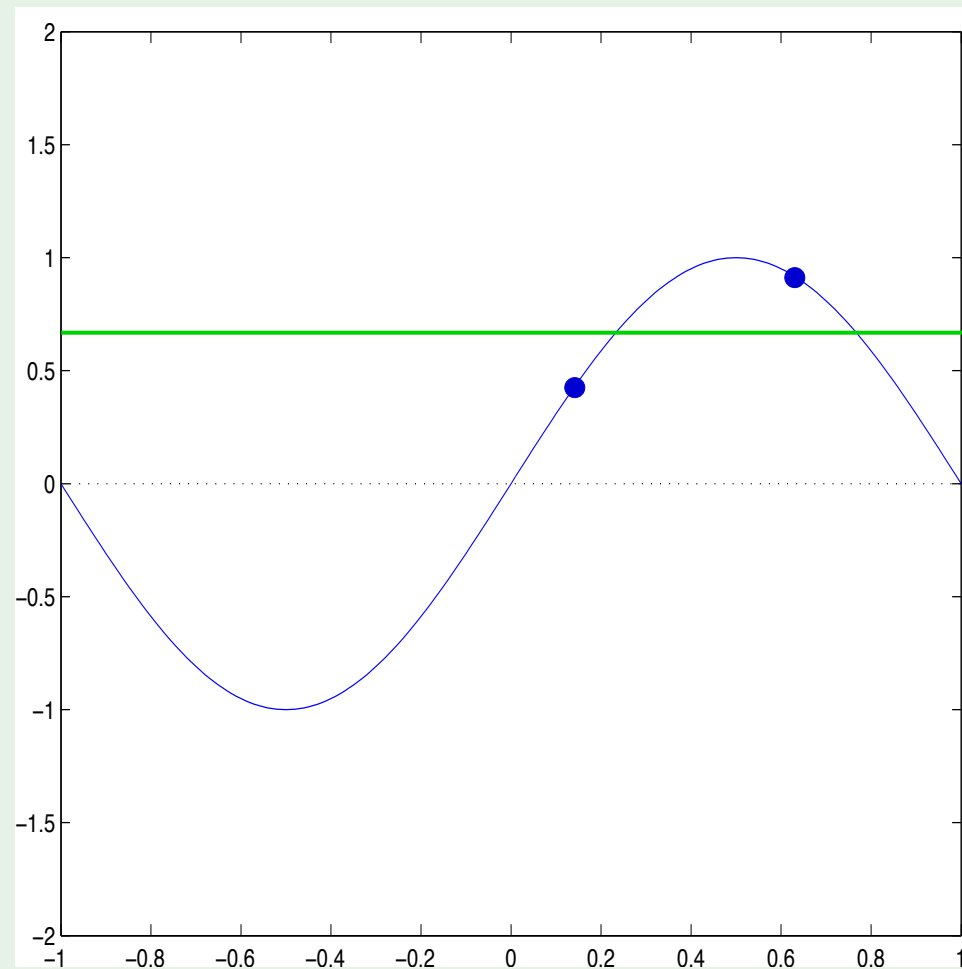


\mathcal{H}_1

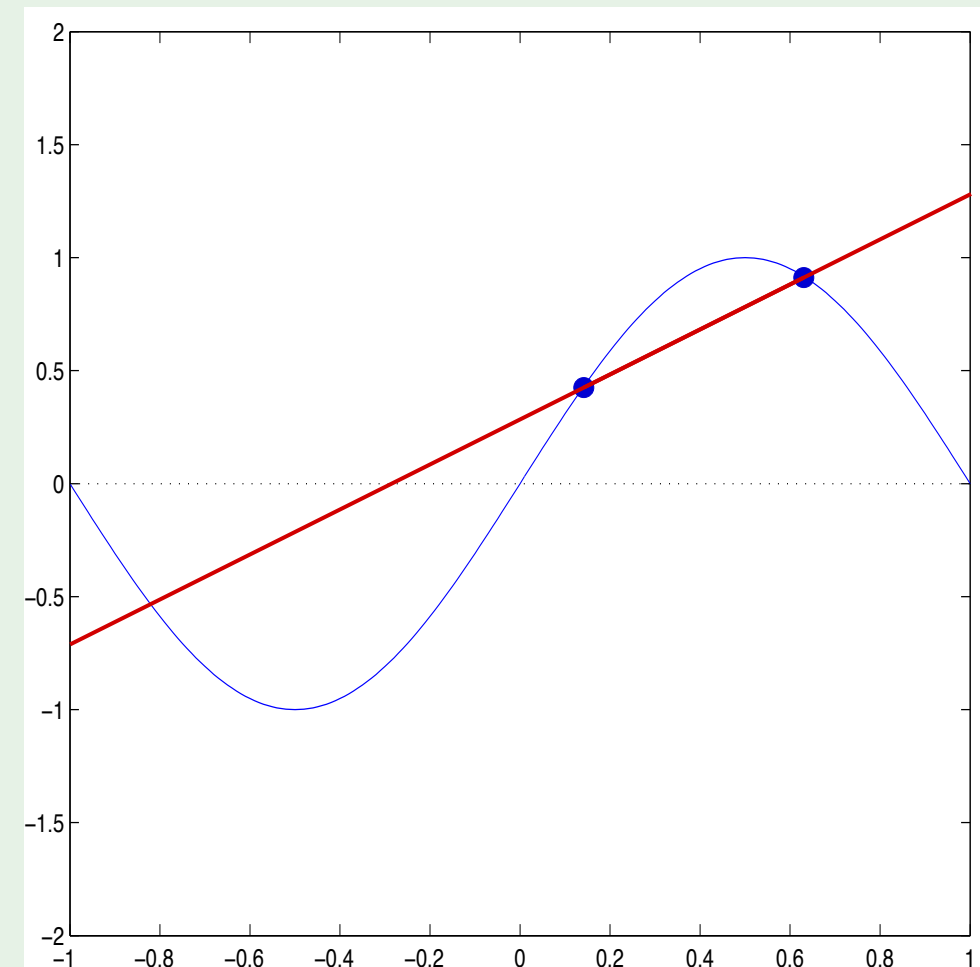


Learning - \mathcal{H}_0 versus \mathcal{H}_1

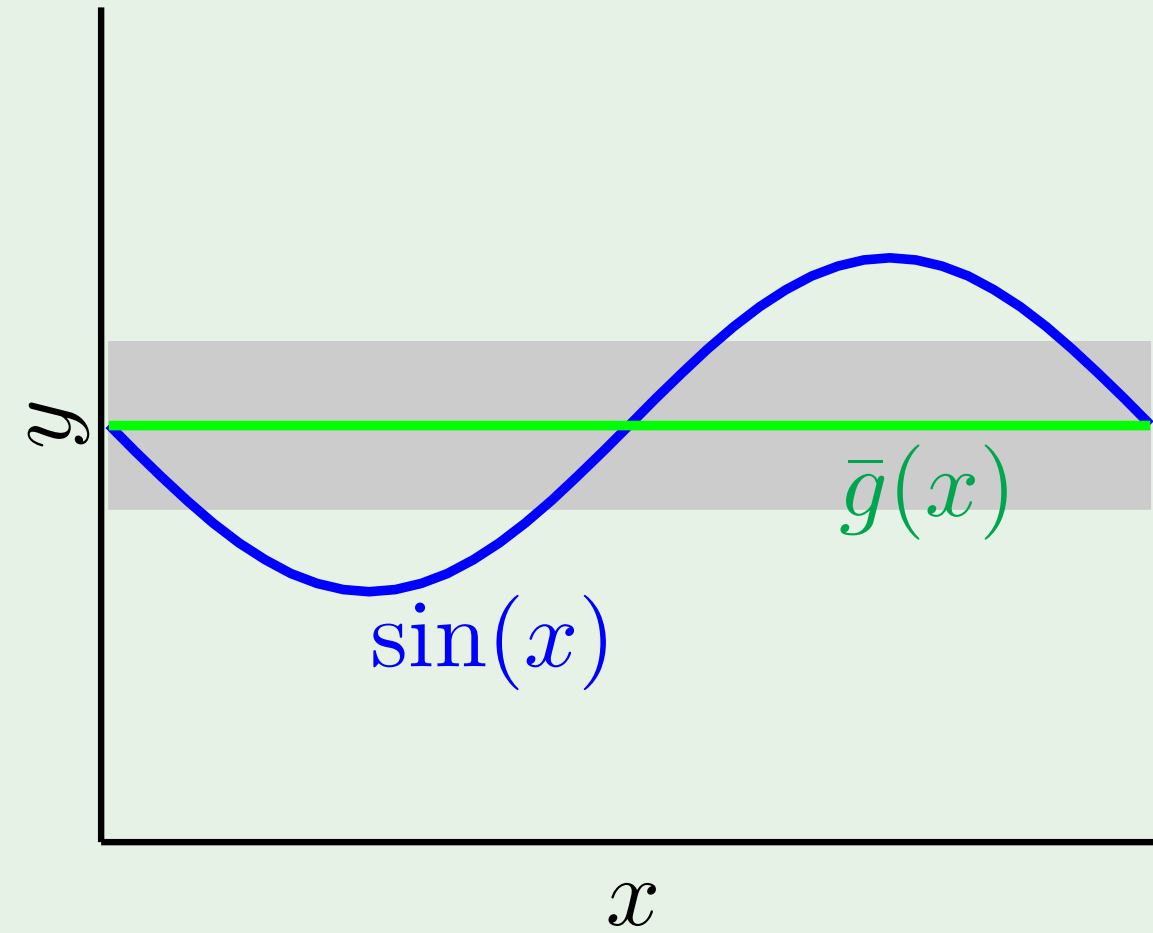
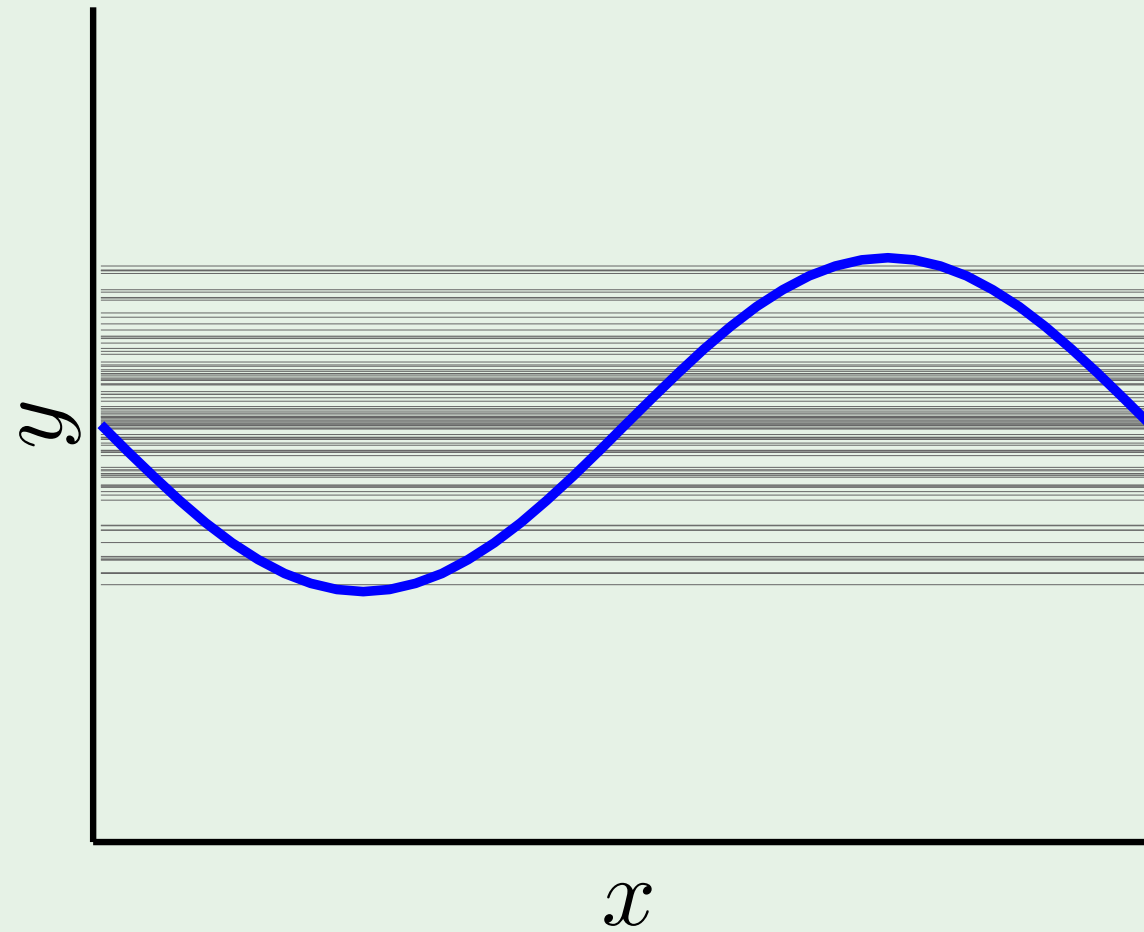
\mathcal{H}_0



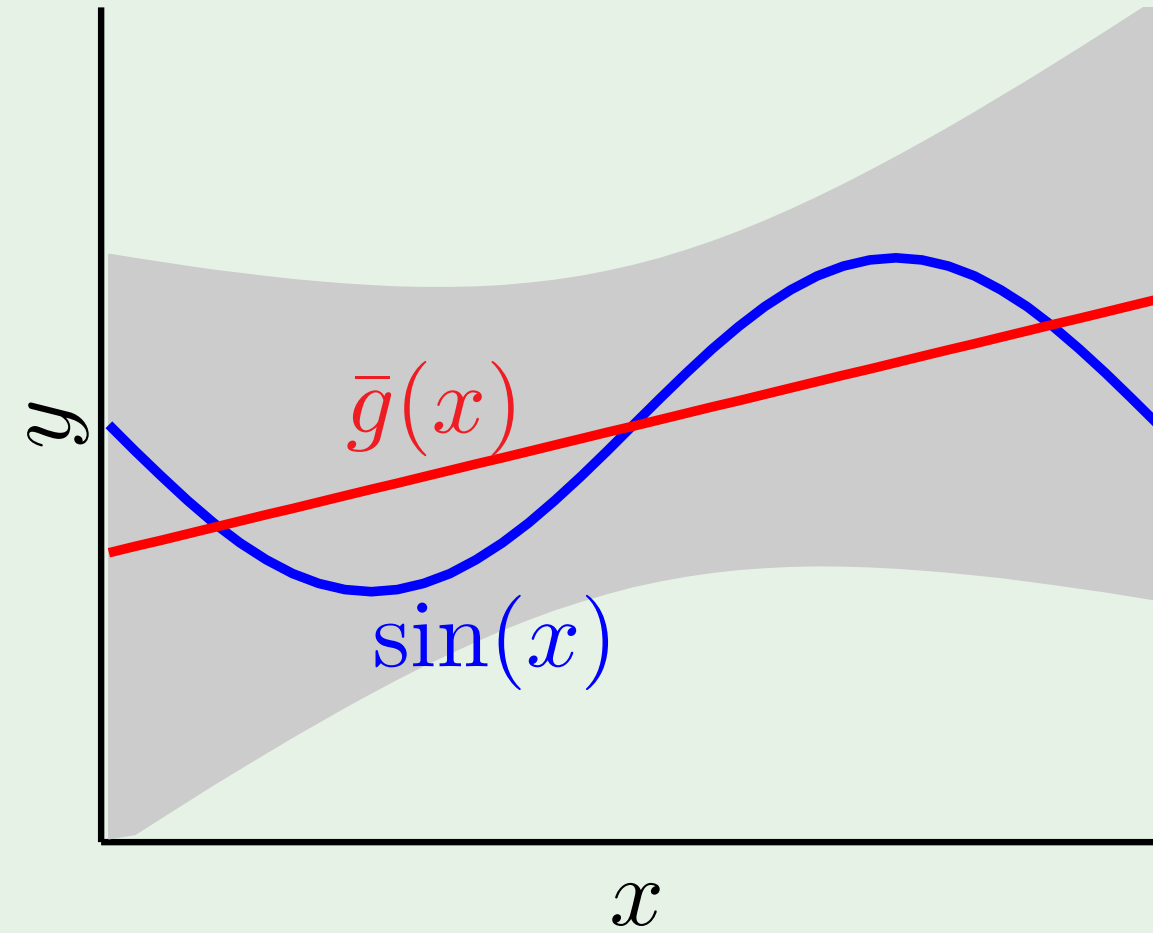
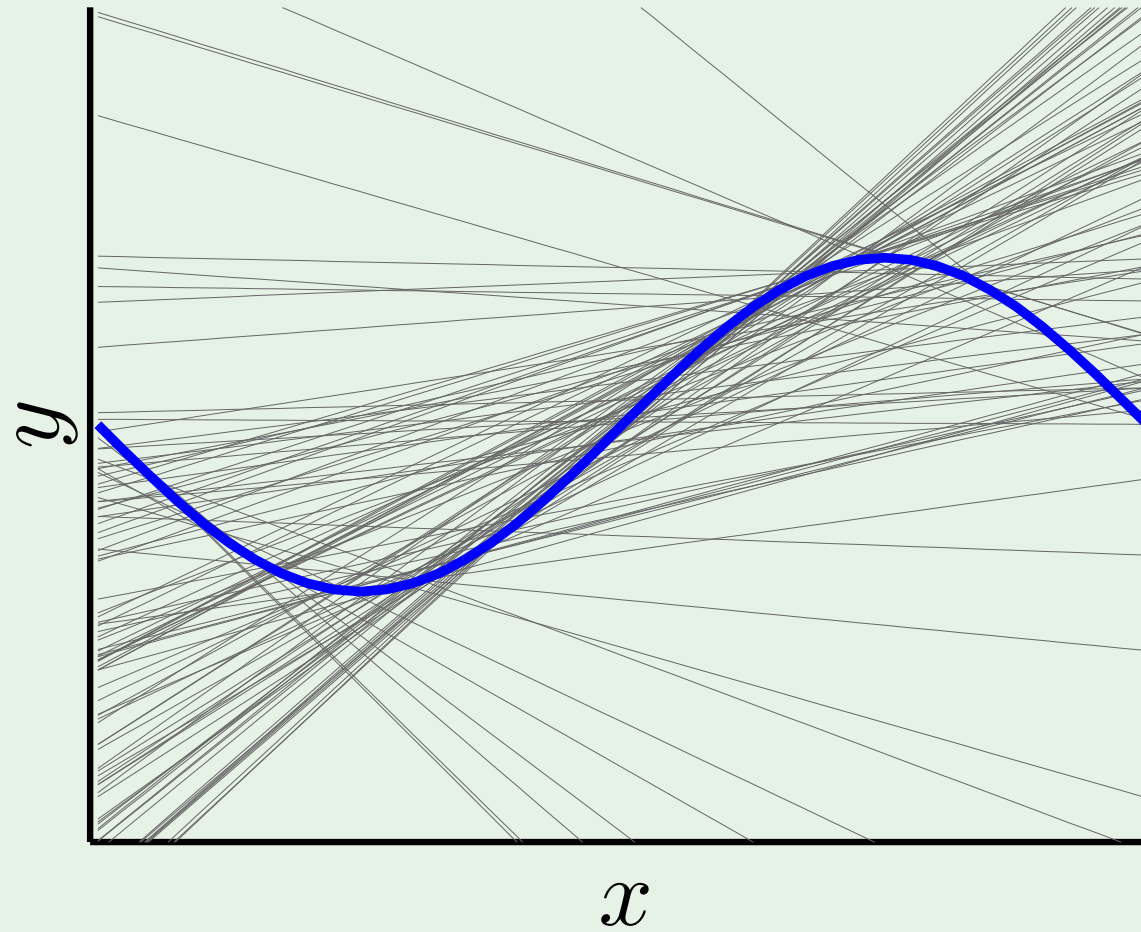
\mathcal{H}_1



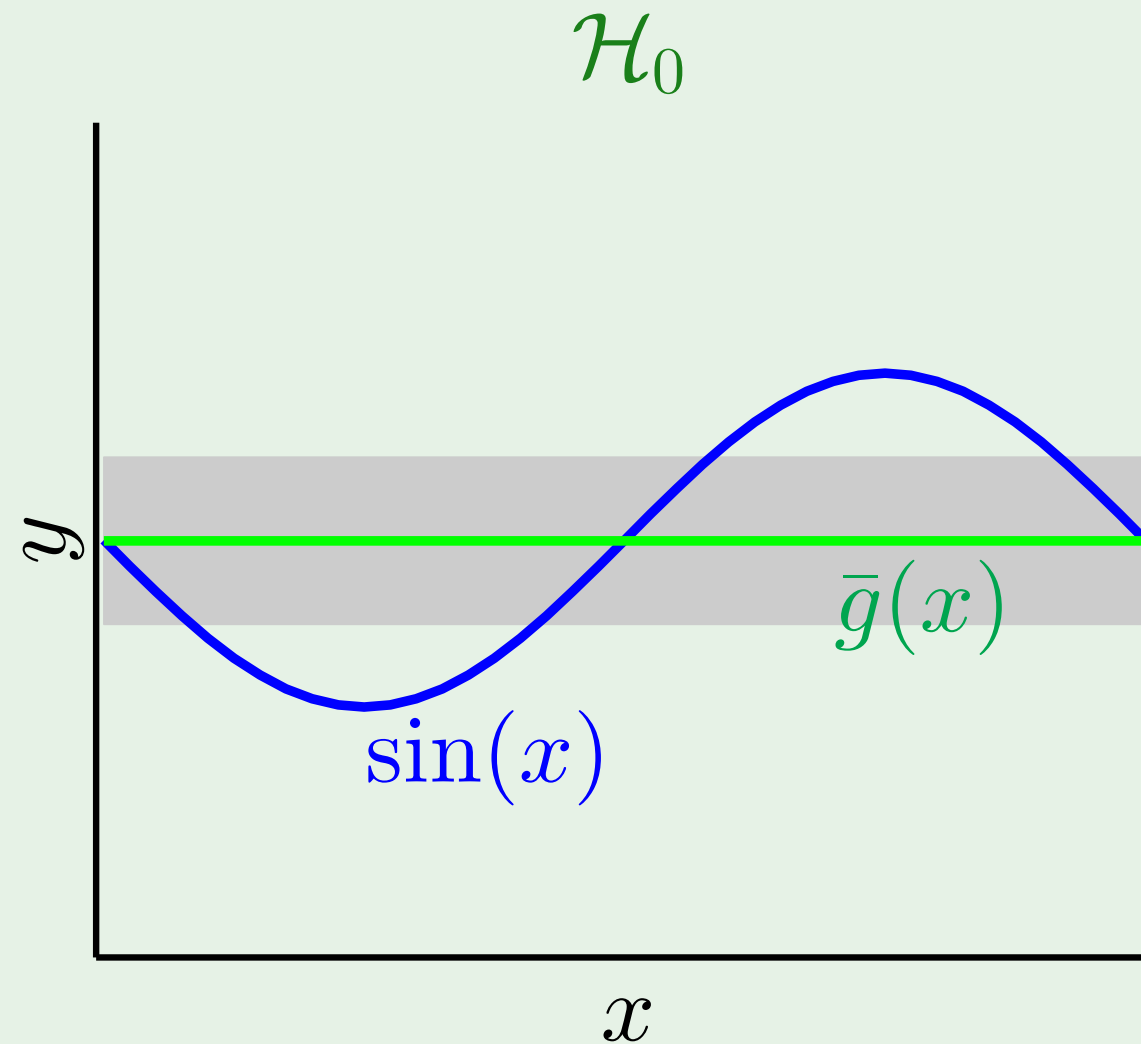
Bias and variance - \mathcal{H}_0



Bias and variance - \mathcal{H}_1

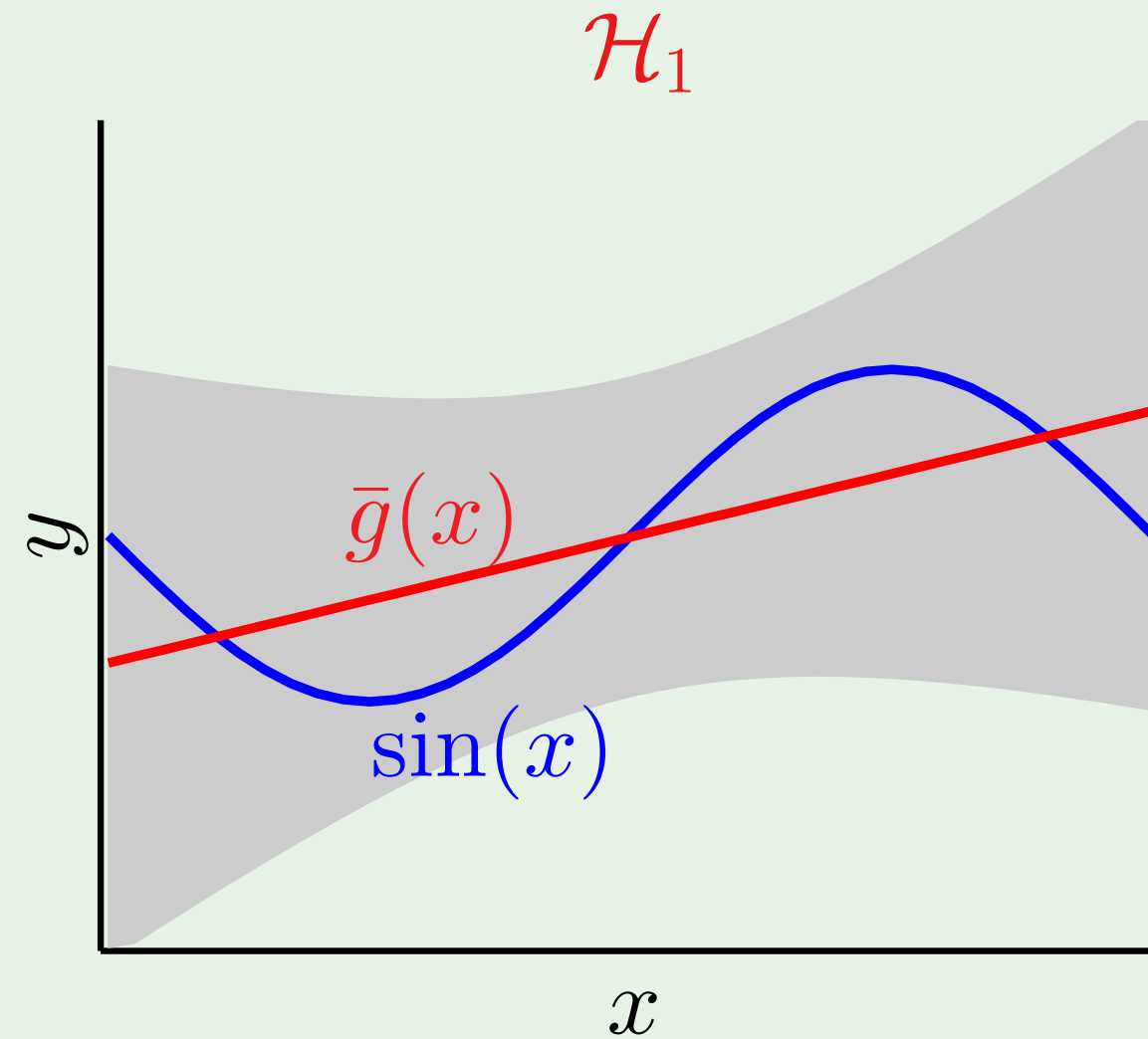


and the winner is ...



bias = **0.50**

var = **0.25**



bias = **0.21**

var = **1.69**

Lesson learned

Match the ‘model complexity’

to the **data resources**, not to the **target complexity**

Outline

- Bias and Variance
- Learning Curves

Expected E_{out} and E_{in}

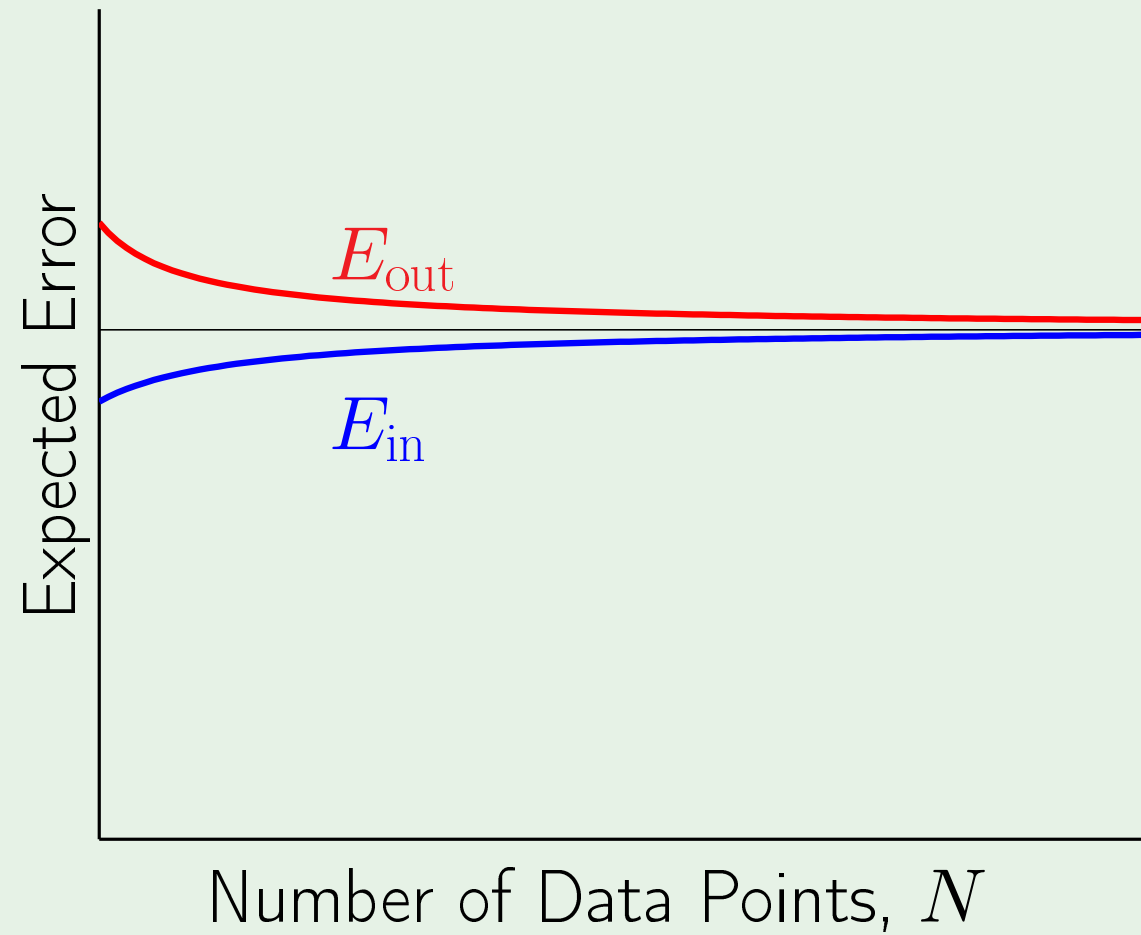
Data set \mathcal{D} of size N

Expected out-of-sample error $\mathbb{E}_{\mathcal{D}}[E_{\text{out}}(g^{(\mathcal{D})})]$

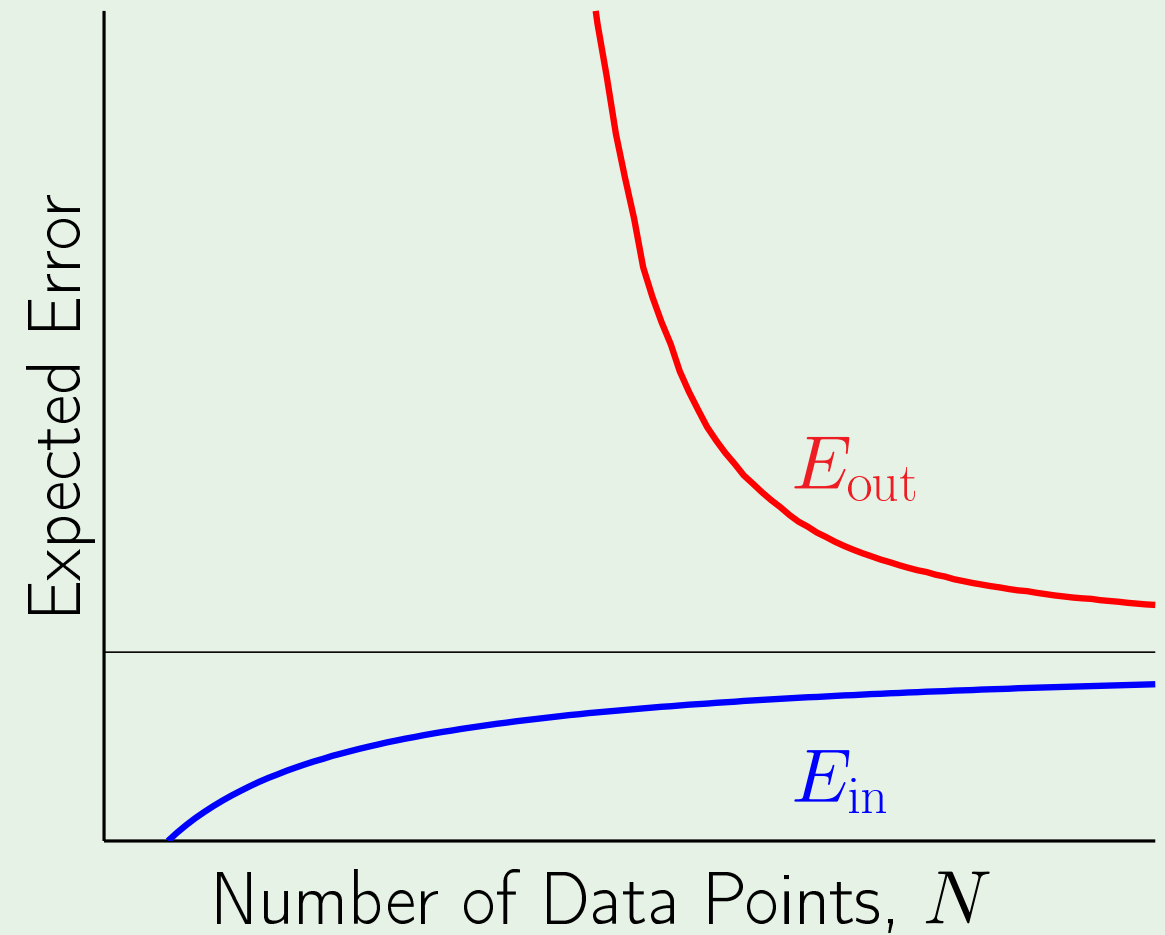
Expected in-sample error $\mathbb{E}_{\mathcal{D}}[E_{\text{in}}(g^{(\mathcal{D})})]$

How do they vary with N ?

The curves

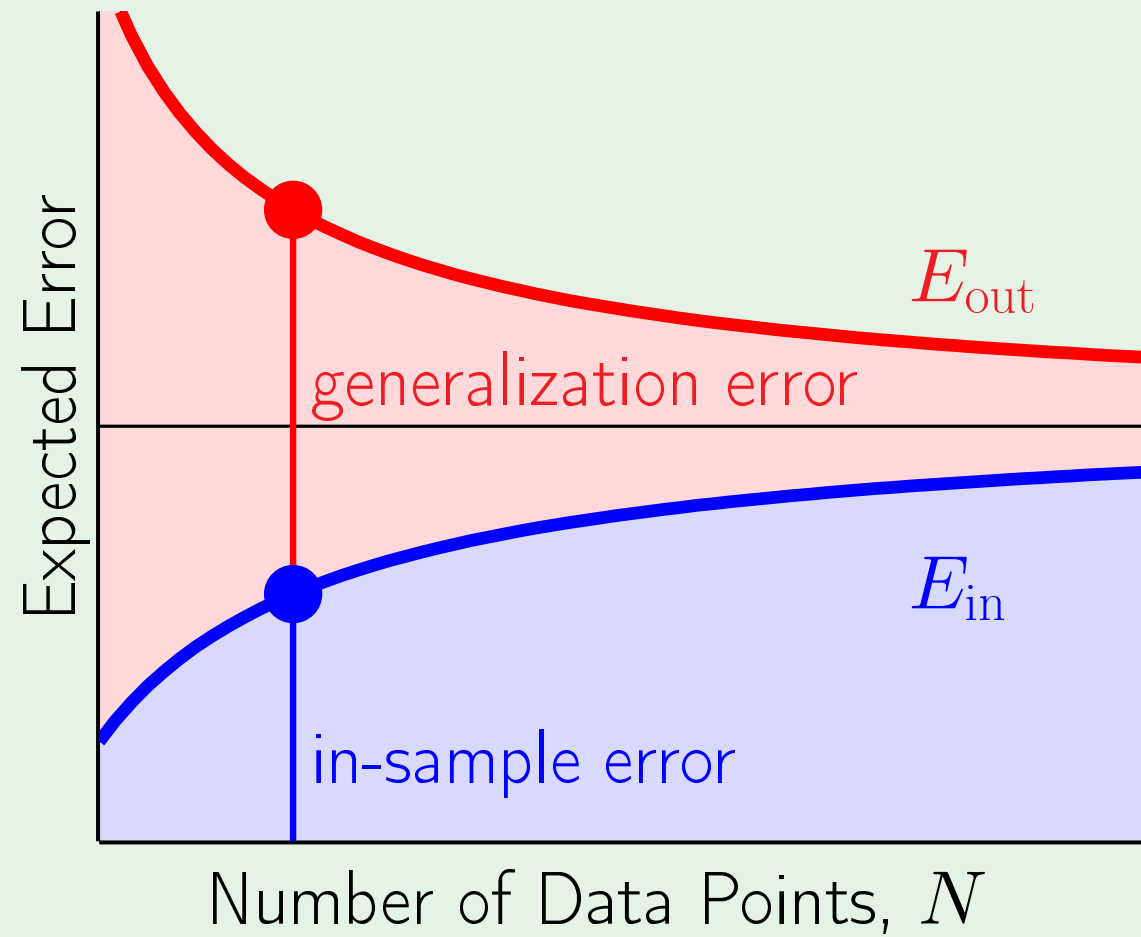


Simple Model

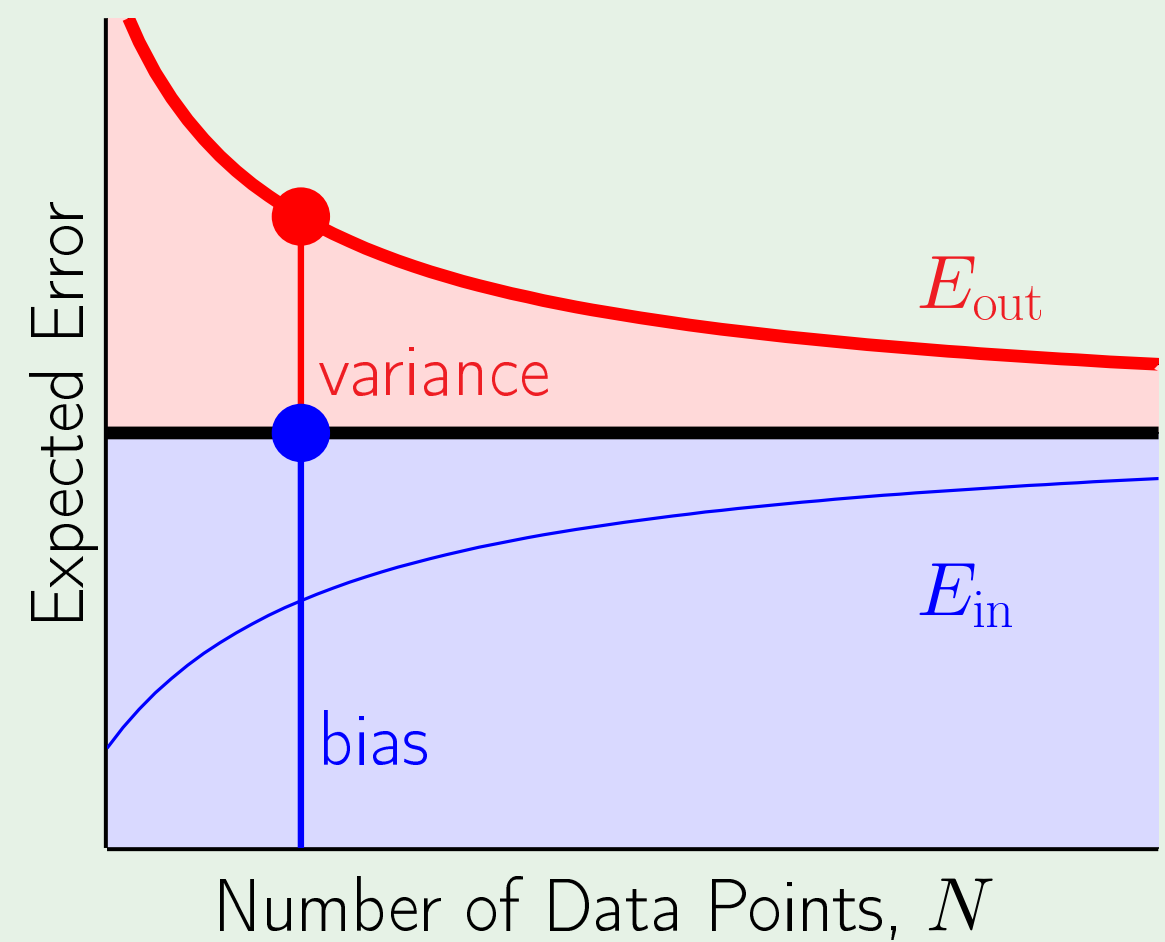


Complex Model

VC versus bias-variance



VC analysis



bias-variance