Machine Learning

Lecture 9: Variance-Bias

Sibei Yang

SIST, ShanghaiTech

Email: yangsb@shanghaitech.edu.cn

[Elements]: 2.9; 5.5; 7.3 [PRML]:3.2

Outline

- Variance-Bias decomposition & tradeoff
- Learning curves

Approximation-generalization tradeoff

- Small E_{out}: good approximation of f out of sample
- More complex better chance of approximating f
- Less complex better chance of generalizing out of sample
- Ideal $\mathcal{H} = \{f\}$
- Bias-variance analysis: decomposing E_{out} into

  A. How well $\mathcal{H}$ can approximate f

  B. How well we can zoom in on a good $h \in \mathcal{H}$

- Applies to real-valued targets and uses squared error

Decomposing 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 $\mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_K$

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

Using the average hypothesis

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

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

$$\left. + \ 2\ \left(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x})\right)\ \left(\bar{g}(\mathbf{x}) - f(\mathbf{x})\right)\right]$$

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

Bias and Variance

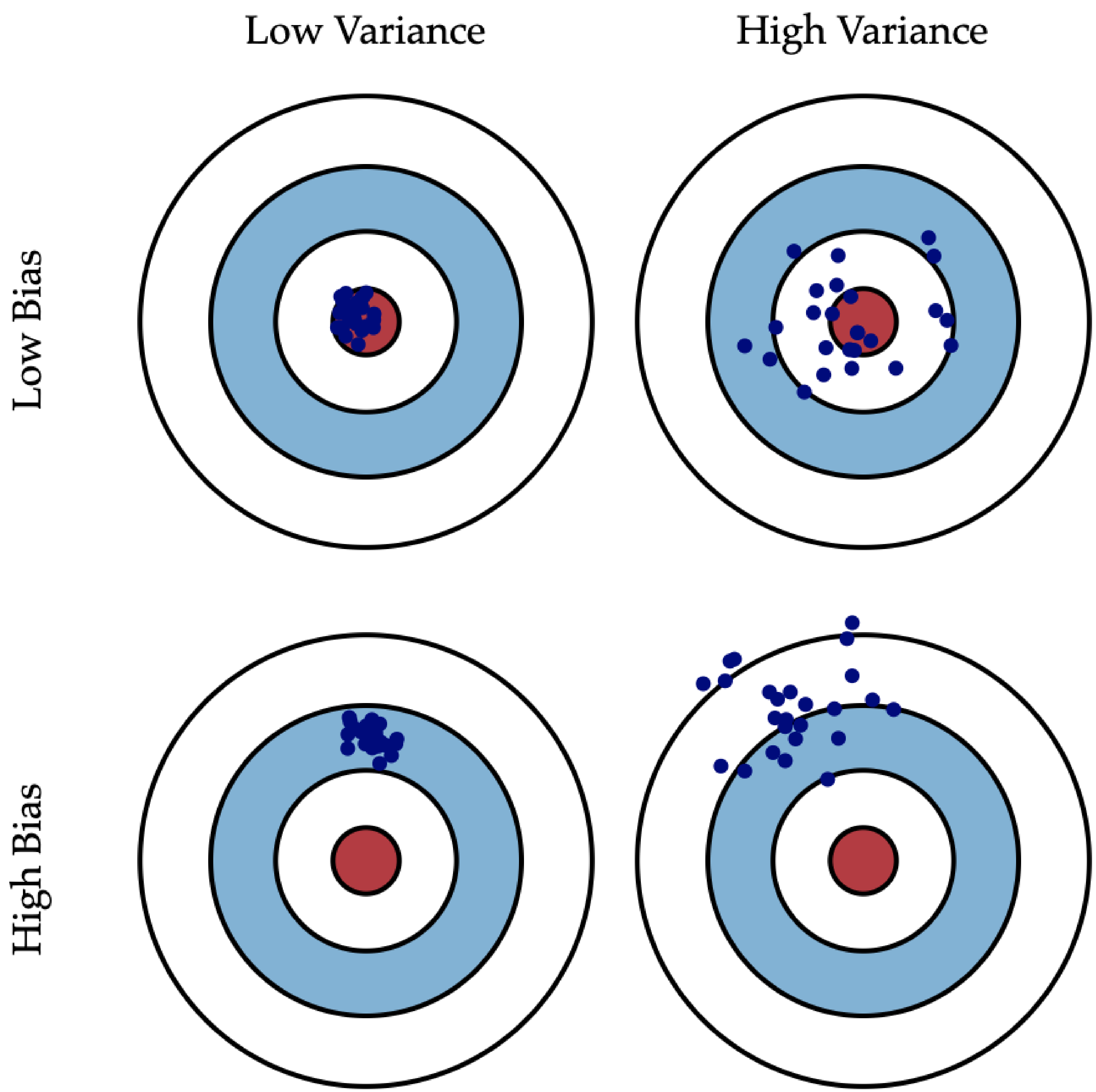
$$\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x})-f(\mathbf{x})\right)^2\right]=\underbrace{\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x})-\bar{g}(\mathbf{x})\right)^2\right]}_{\mathsf{var}(\mathbf{x})}+\underbrace{\left(\bar{g}(\mathbf{x})-f(\mathbf{x})\right)^2}_{\mathsf{bias}(\mathbf{x})}$$

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

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

$$=\quad \mathsf{bias} \quad + \quad \mathsf{var}$$

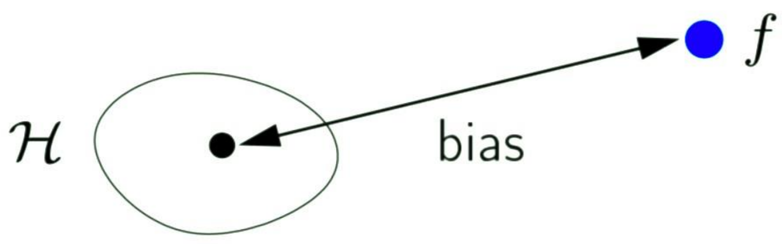
Bias and Variance

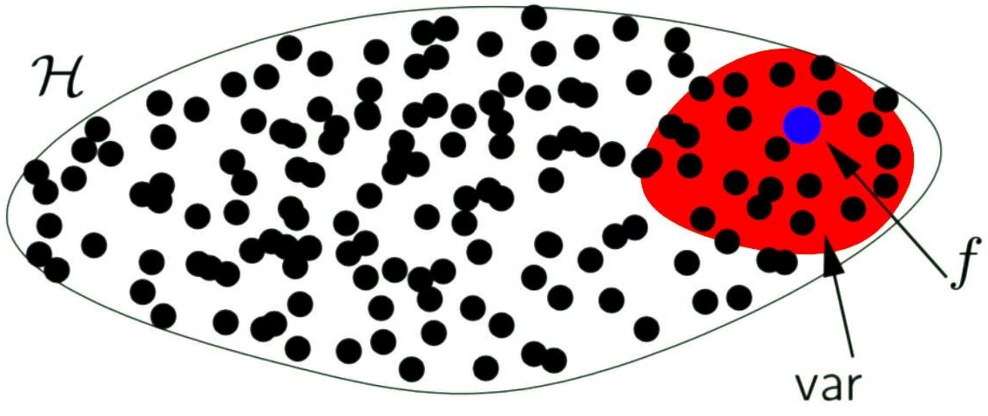


http://scott.fortmann-roe.com/docs/BiasVariance.html

The tradeoff

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





$$\downarrow \qquad \mathcal{H} \uparrow \qquad \uparrow$$

Example: sin target

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

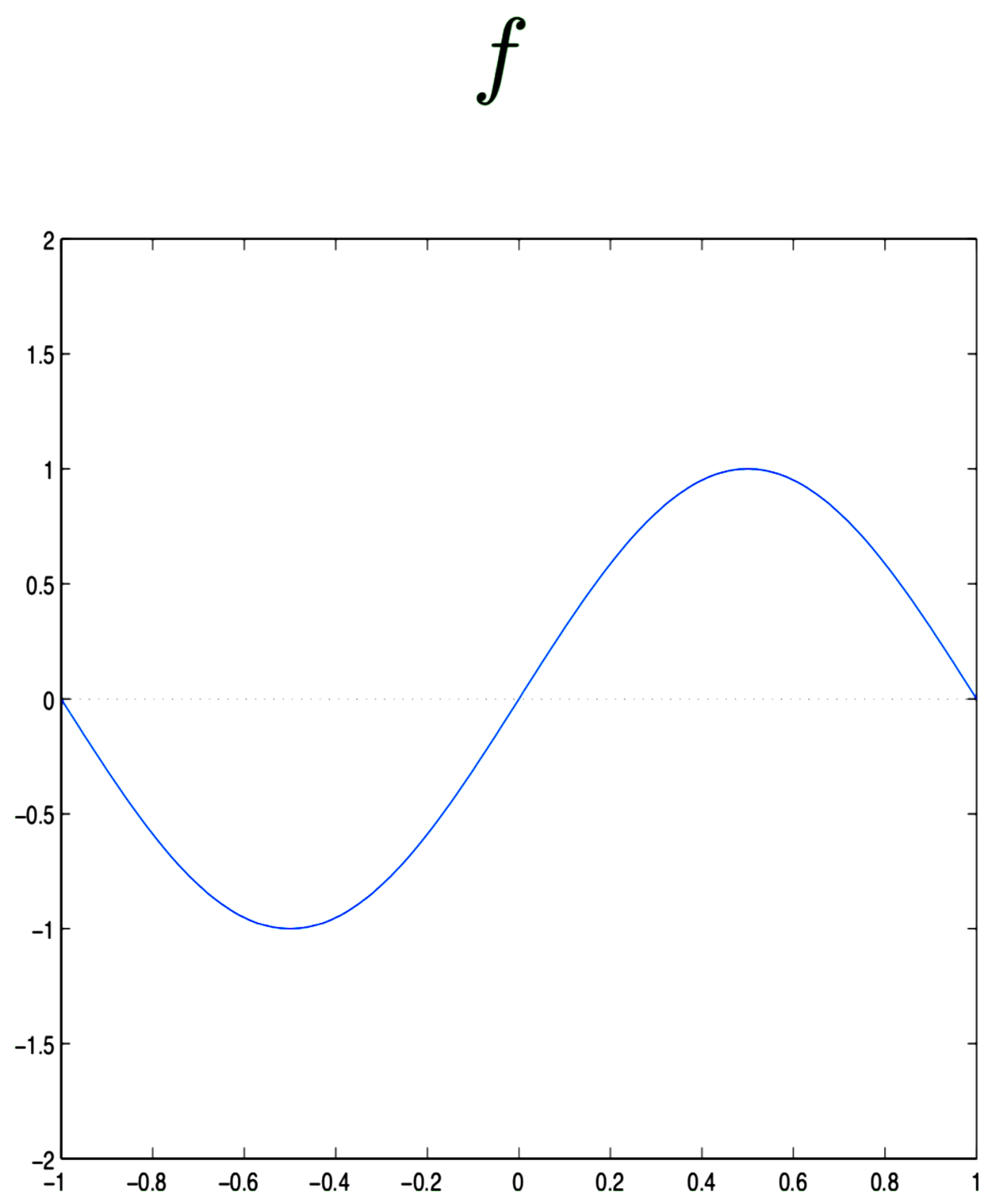
Only two training examples! $\quad N = 2$

Two models used for learning:

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

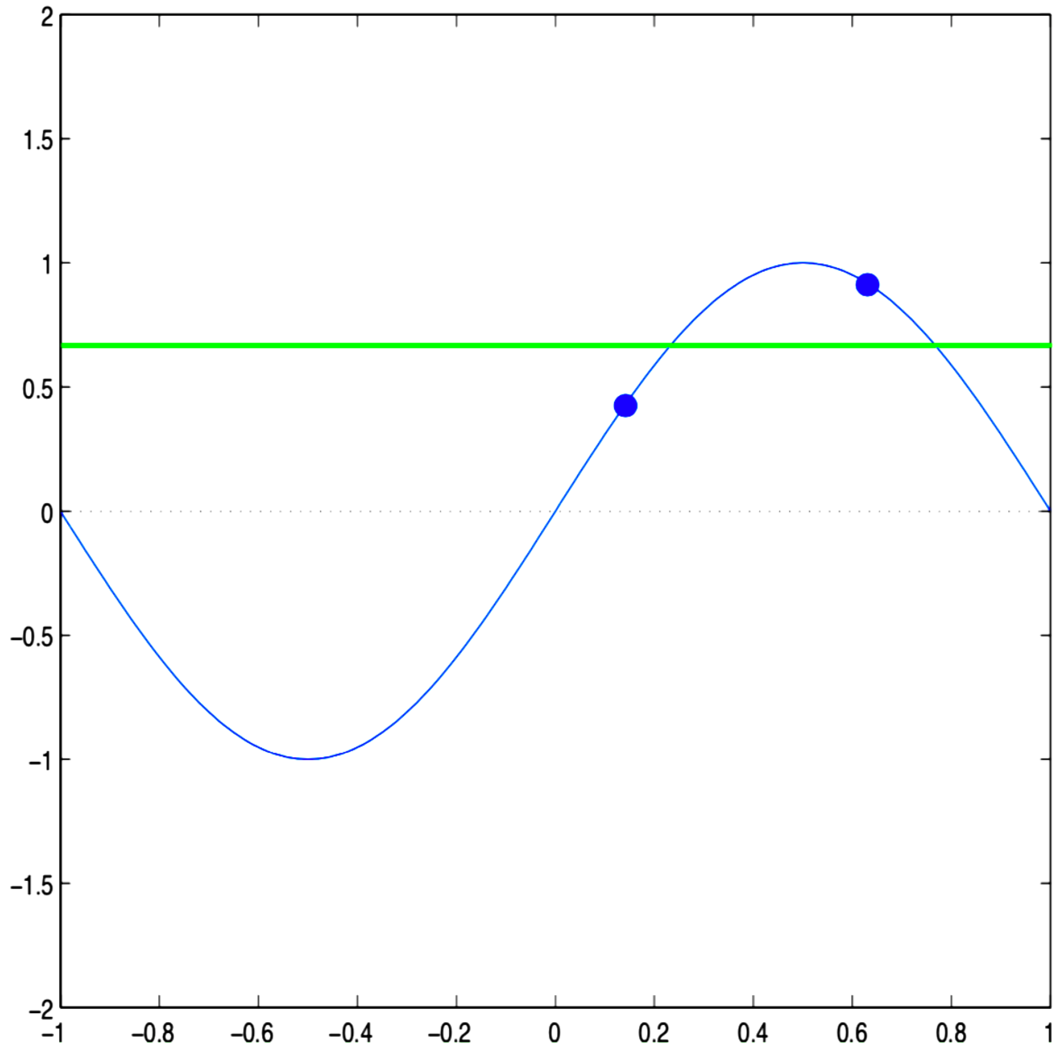
$$\mathcal{H}_1: \quad h(x) = ax + b$$

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

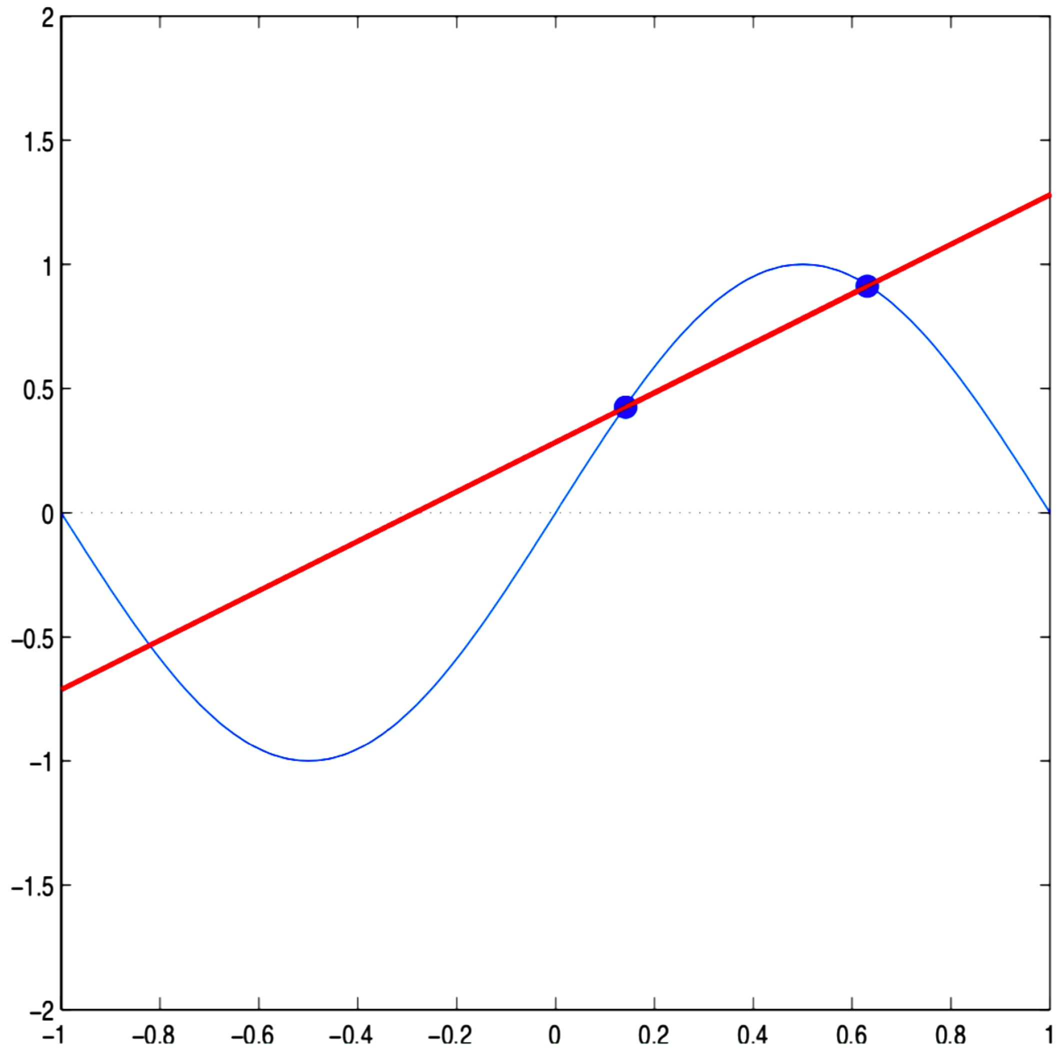


Learning: $\mathcal{H}_0$ versus $\mathcal{H}_1$

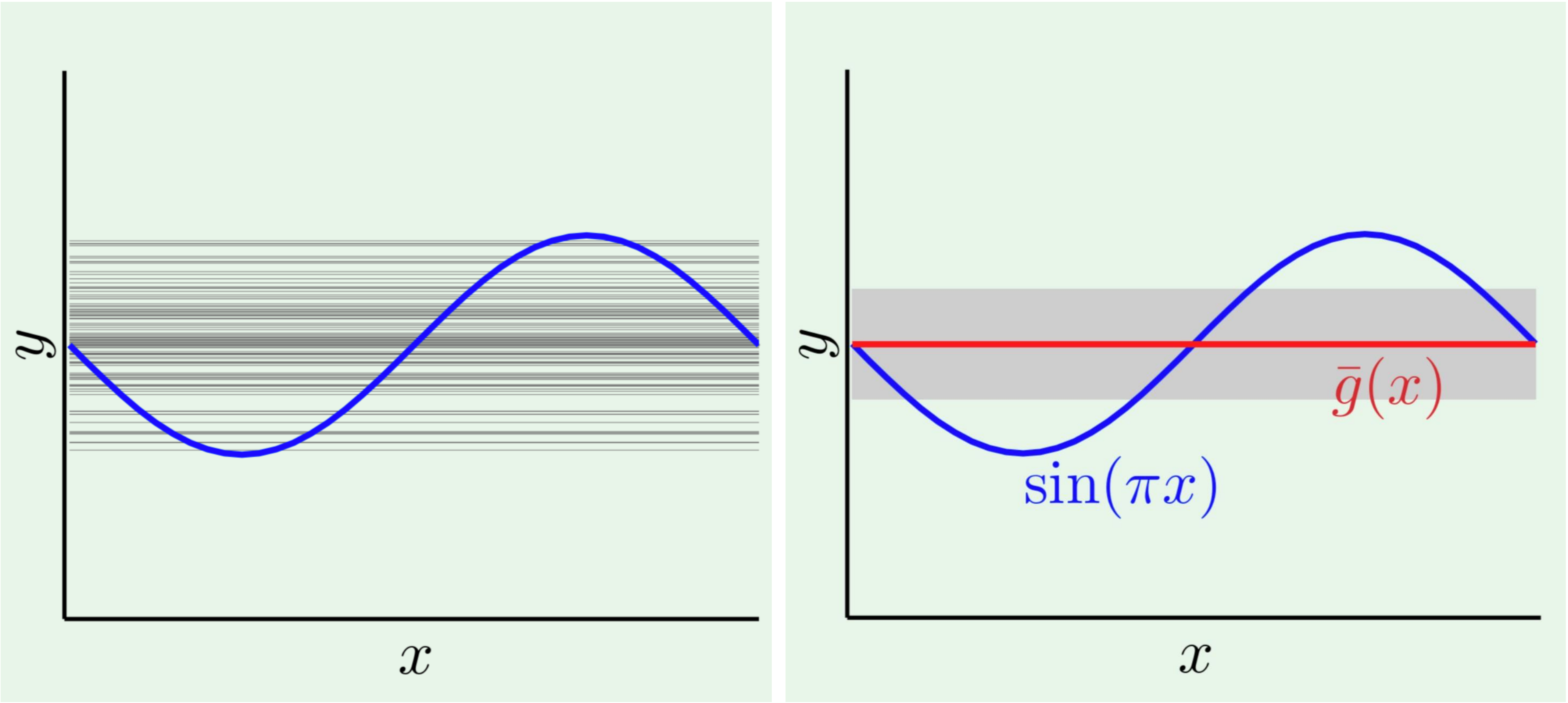
$\mathcal{H}_0$



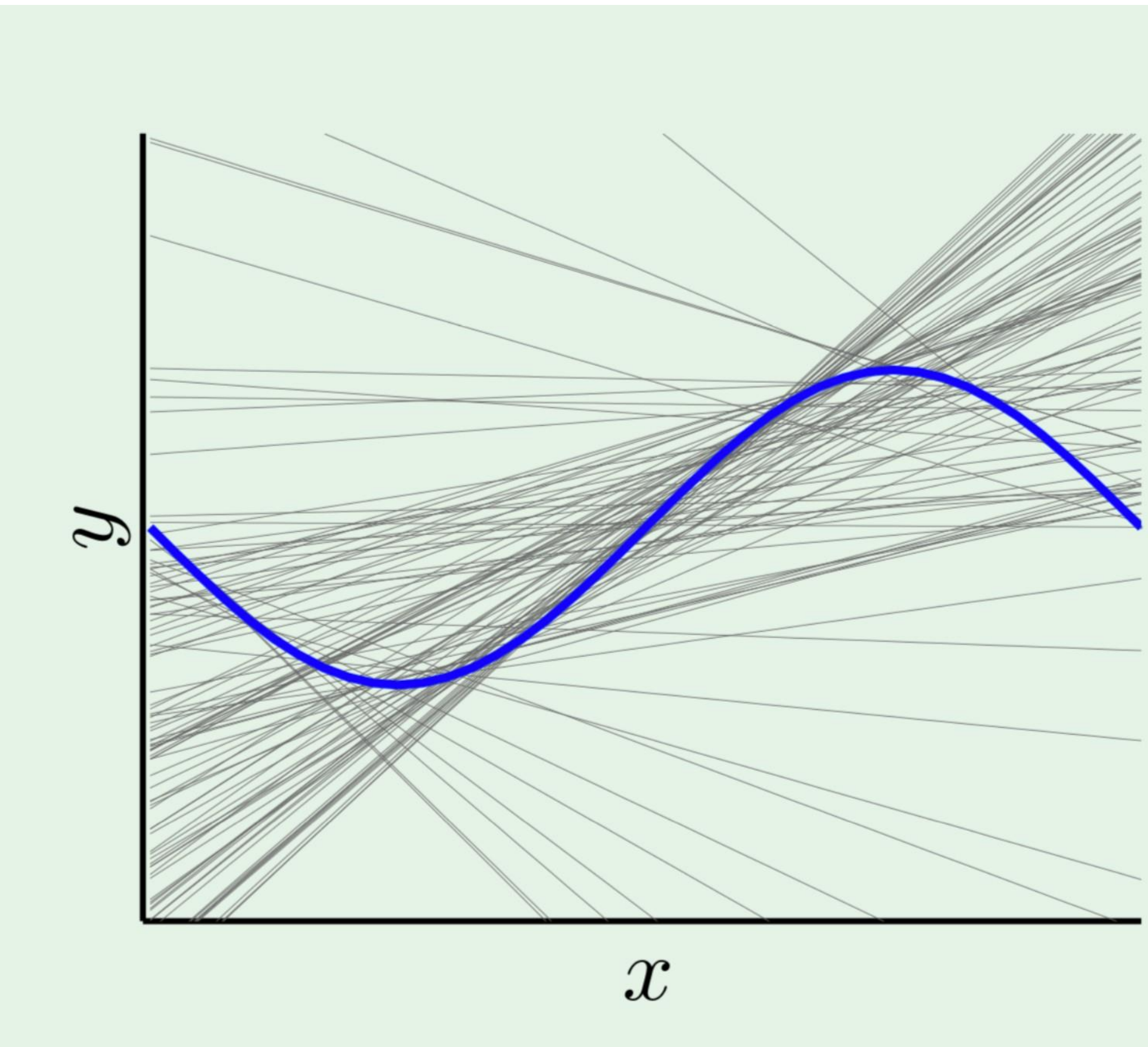
$\mathcal{H}_1$

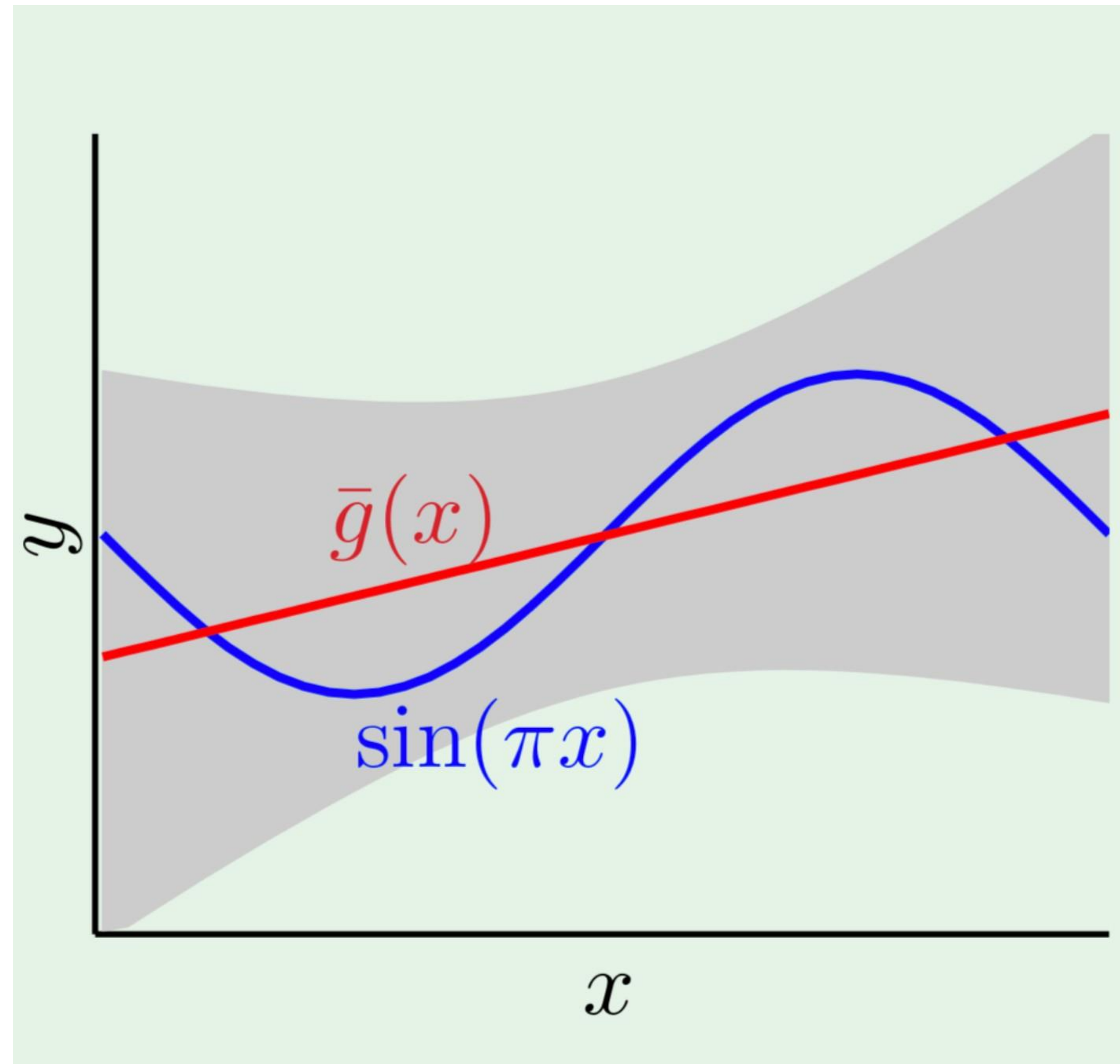


Bias and variance: $\mathcal{H}_0$



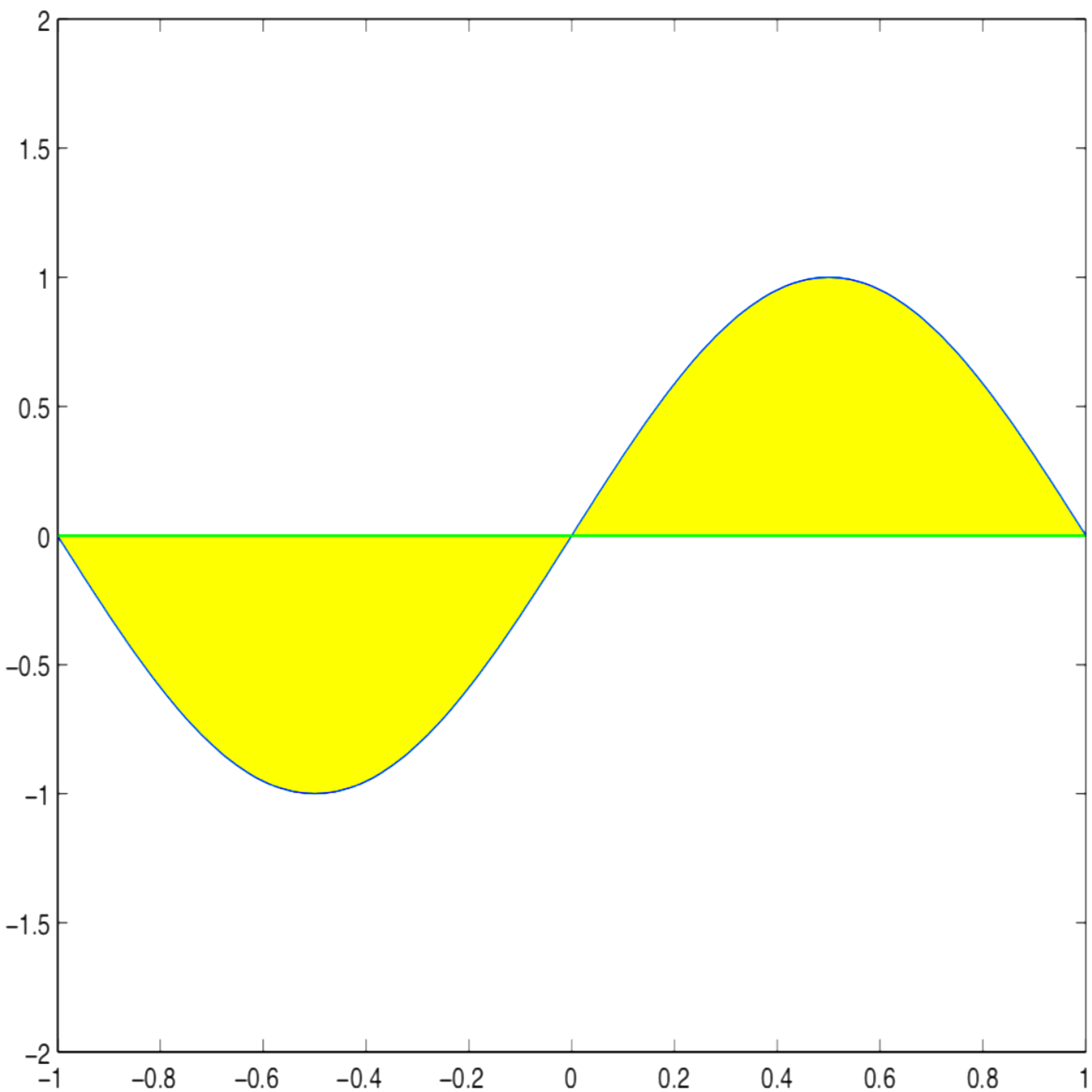
Bias and variance: $\mathcal{H}_1$



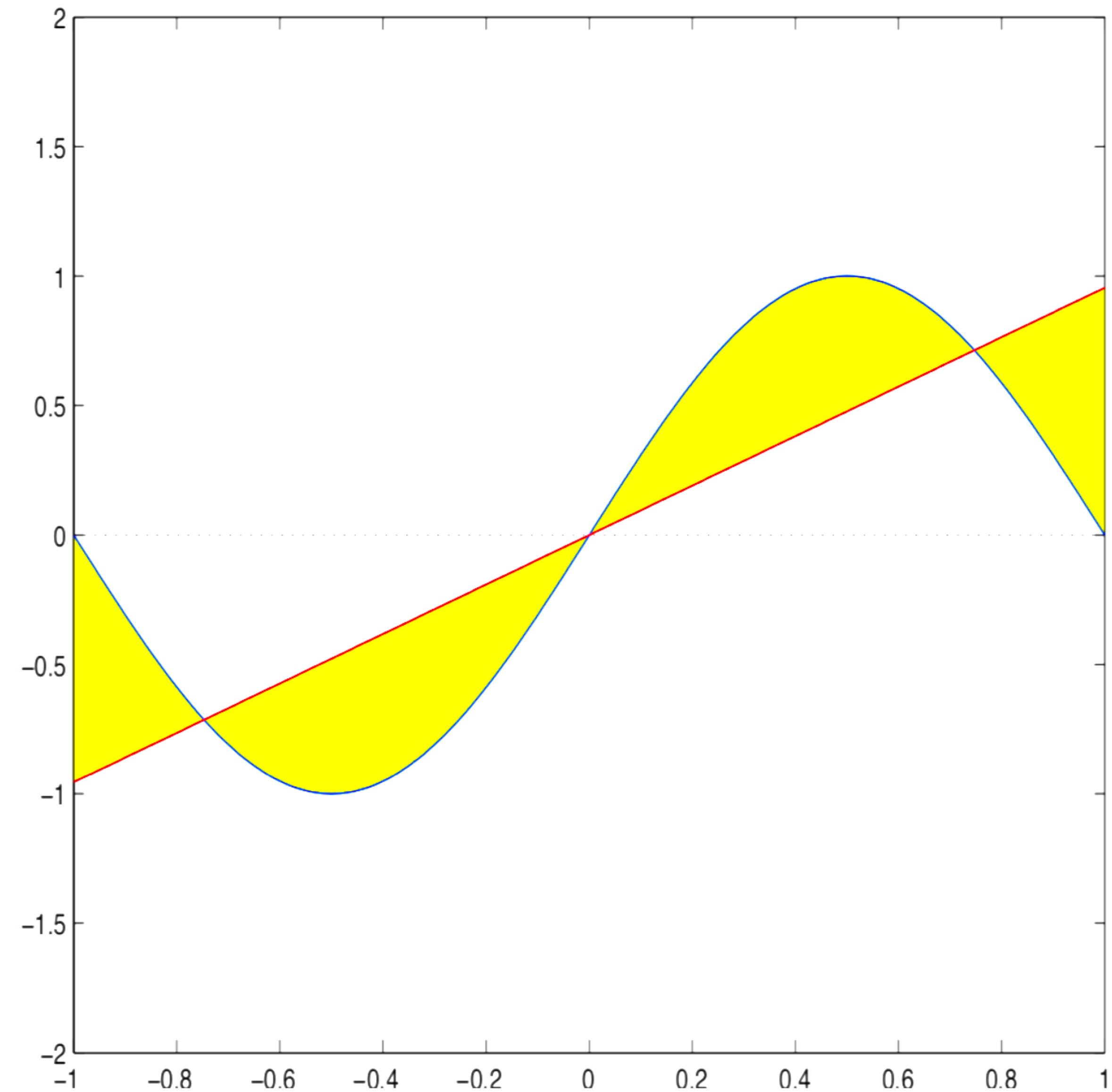


Approximation: $\mathcal{H}_0$ versus $\mathcal{H}_1$

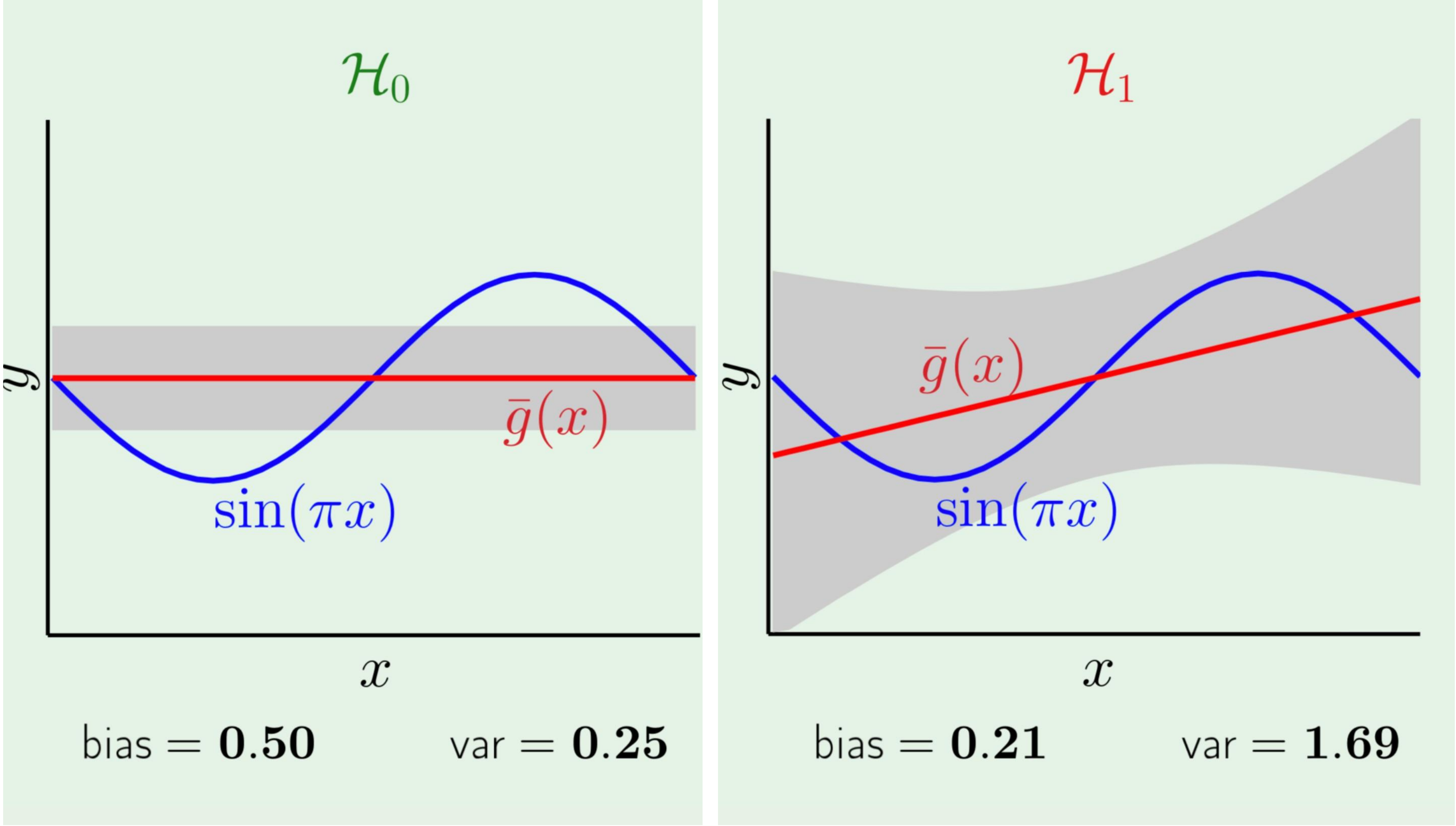
$\mathcal{H}_0$



$\mathcal{H}_1$



The winner is…



Conclusion:

Match the "model complexity" to the data resource, NOT to the target complexity

Expected E_{out} and E_{in}

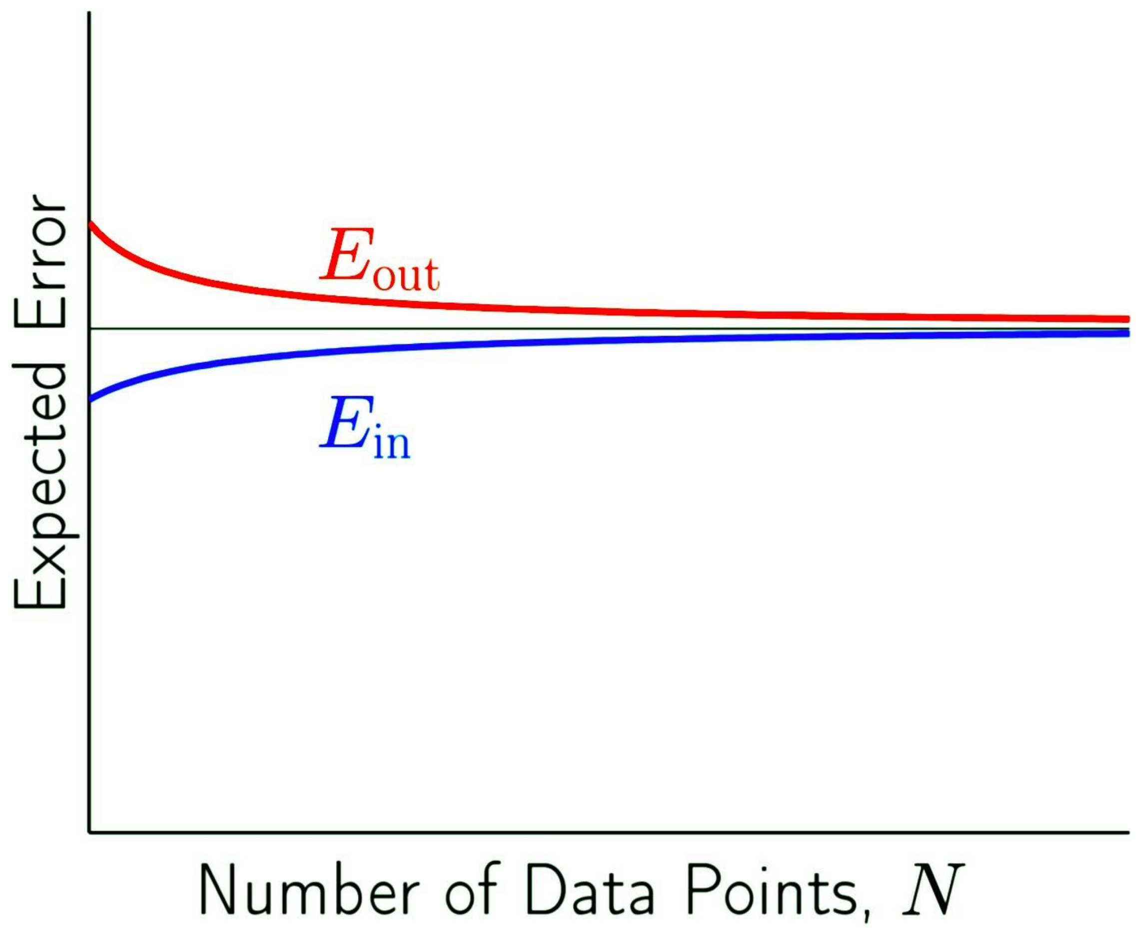
Data set $\mathcal{D}$ of size N

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

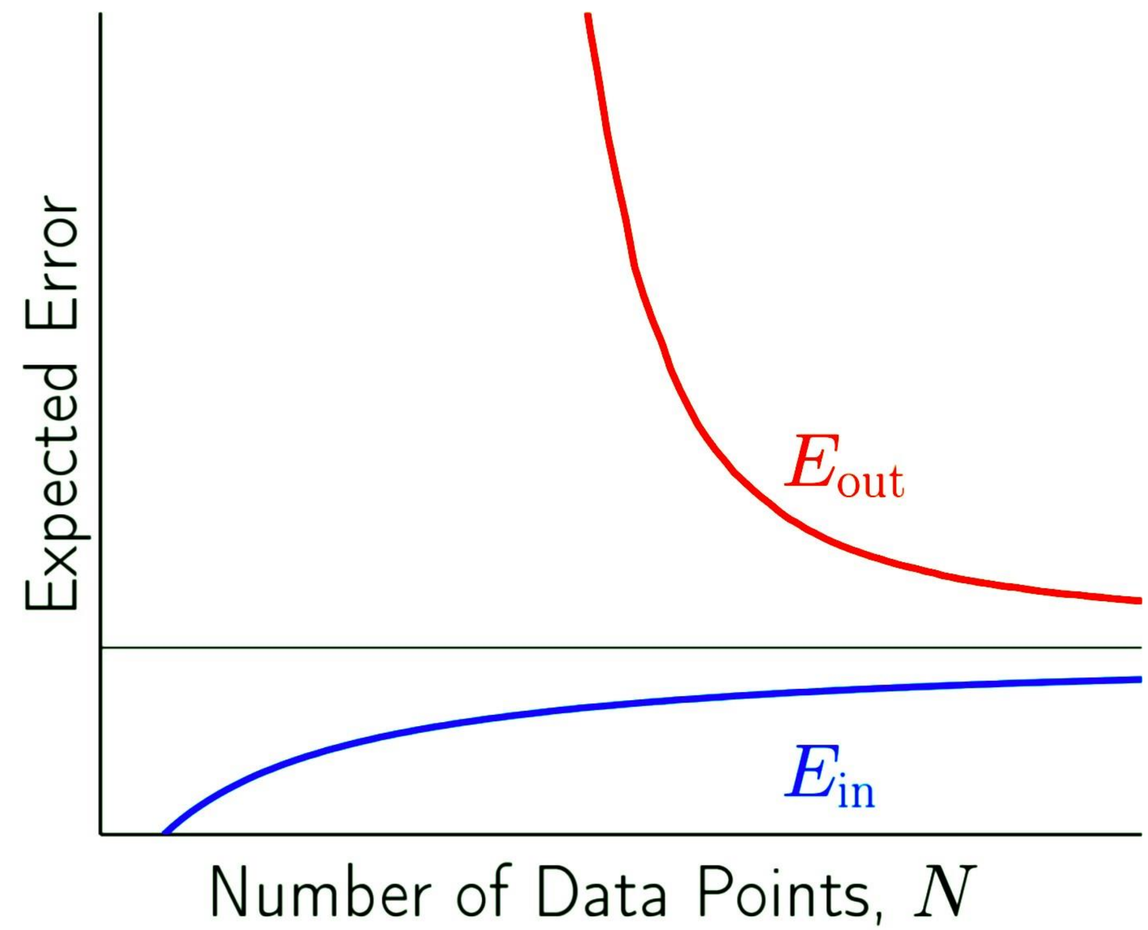
Expected in-sample error $\mathbb{E}_{\mathcal{D}}[E_{\text{in}}(g^{(\mathcal{D})})]$

How do they vary with N?

Learning curves



Simple Model



Complex Model